UNIVERSITE
PAUL
SABATIER
TOULOUSE III

Publications du Laboratoire
de
Statistique et Probabilités

Analyse des Données Multidimensionnelles

Méthodes factorielles

André Carlier

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

Avant-Propos Ce document est divisé en quatre chapitres. Un premier chapitre introductif contient des définitions générales en statistique univariée, bivariée et multivariée et développe la notion de liaison. Le second chapitre est un chapitre de mathématiques utilisées en analyse multidimensionnelles. Il s'agit principalement d'algèbre linéaire et plus spécialement des propriétés des opérateurs dans les espaces euclidiens. Enfin les chapitres 3 et 4 développent des méthodes classiques en analyse multivariée, l'analyse en composantes principales, (chapitre 3) et l'analyse des correspondances simples et multiples (chapitre 4).

Chapitre 1

Définitions générales

1.1 Définitions élémentaires

Définition 1 *Une* **Unité Statistique** *(u.s.) est un* **individu** *ou* **objet** *sur lequel on effectue des mesures ou observations. Les unités statistiques sont numérotées de 1 à n (on note $I = \{1, \ldots, n\}$ cet ensemble d'indices) et affectées de* **poids** *p_i ($p_i > 0$), mesurant leur importance relativement aux autres. Si les p_i sont de somme 1, on dit que les poids sont* **normalisés**. *On le supposera dans la suite.*

Remarque 1.1.1 *L'ensemble des individus pourra être un échantillon (une partie) d'une population plus grande. Sous des hypothèses basées sur la théorie du calcul des probabilités, il sera possible de déduire d'observations sur l'échantillon des conclusions applicables à l'ensemble de la population. C'est l'objet de la* **statistique inférentielle**. *L'ensemble des observations pourra aussi concerner toute la population. On parle alors d'échantillon exhaustif. L'objectif est alors de décrire les données. C'est le but de la* **statistique descriptive**.

Définition 2 *On appelle* **variable** *un ensemble de n observations de même type effectuées sur les n individus.*

Définition 3 *On dit qu'une variable est* **quantitative** *quand elle prend ses valeurs dans l'ensemble des réels. Si elle prend ses valeurs dans un ensemble dont le nombre d'éléments est fini, on dit qu'elle est* **qualitative** *(on dit aussi* **catégorielle** *ou* **nominale** *). Dans le cadre du modèle linéaire, on parle de* **facteurs**. *L'ensemble des valeurs d'une variable qualitative est appelé l'ensemble des* **modalités** *de la variable; pour un facteur, on parle de l'ensemble des* **niveaux** *du facteur. Si l'ensemble des modalités possède une structure d'ordre, on parle de variable* **ordinale** *ou* **qualitative ordonnée**.

On indice les variables par j ($j = 1, \ldots, p$), et on note par $\mathbf{y}^j$ la j-ème variable, et par y_i^j la valeur de la variable $\mathbf{y}^j$ pour l'individu i. La variable quantitative $\mathbf{y}^j$ est identifiée au vecteur de $\mathbf{R}^n$ de coordonnées y_i^j , que l'on représenté en colonne. On appelle **espace des variables** l'espace $\mathbf{R}^n$ dont les éléments sont des variables.

La suite de ce chapitre concerne exclusivement les variables quantitatives.

Définition 4 (Moyenne empirique d'une variable) *La moyenne empirique d'une variable $\mathbf{y}^j$ est définie par:*

$$\bar{y^j} = \sum_{i \in I} p_i y_i^j$$

Propriété caractéristique:

$$\sum_{i \in I} p_i (y_i^j - \bar{y^j}) = 0$$

Autre propriété La moyenne est une forme linéaire sur $\mathbf{R}^n$.

Définition 5 *Soit $\mathbf{1}_n$ le vecteur de $\mathbf{R}^n$ dont toutes les coordonnées sont égales à 1, alors:*

$$\mathbf{x} = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_i - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} = \mathbf{y} - \bar{y}\mathbf{1}_n$$

est appelée **variable centrée de y**. *Une variable est dite* centrée *ssi elle est de moyenne nulle.*

1.2 Covariance empirique de deux variables

Définition 6 *La* **covariance empirique** *de $\mathbf{y}^1$ et $\mathbf{y}^2$ est une application de $\mathbf{R}^n \times \mathbf{R}^n$ dans $\mathbf{R}$ définie par :*

$$cov(\mathbf{y}^1, \mathbf{y}^2) = \sum_{i \in I} p_i (y_i^1 - \bar{y^1})(y_i^2 - \bar{y^2}).$$

Elle s'obtient encore par l'une des expressions suivantes:

$$\begin{aligned} cov(\mathbf{y}^1, \mathbf{y}^2) &= \sum_{i \in I} p_i y_i^1 y_i^2 - \bar{y^1}\bar{y^2}, \\ &= 1/2 \sum_{i \in I} \sum_{i' \in I} p_i p_{i'} (y_i^1 - y_{i'}^1)(y_i^2 - y_{i'}^2), \\ &= \sum_{i \in I} p_i (y_i^1 - \bar{y^1})(y_i^2). \end{aligned} \tag{1.1}$$

Propriété 1 *La covariance possède les propriétés suivantes :*

- *C'est une forme bilinéaire,*
  - *linéarité à droite: pour tous réels α, β, pour toutes variables $\mathbf{y}^1$ et $\mathbf{y}^2$:*

    $$cov(\mathbf{y}, \alpha\, \mathbf{y}^1 + \beta\, \mathbf{y}^2) = \alpha\; cov(\mathbf{y}, \mathbf{y}^1) + \beta\; cov(\mathbf{y}, \mathbf{y}^2),$$

  - *linéarité à gauche: s'obtient de la même manière par permutation.*
- *symétrique,*
  $cov(\mathbf{y}^1, \mathbf{y}^2) = cov(\mathbf{y}^2, \mathbf{y}^1)$
- *et positive*
  $cov(\mathbf{y}, \mathbf{y}) \geq 0.$
- $cov(\mathbf{y}^1 + \beta\mathbf{1}_n, \mathbf{y}^2) = cov(\mathbf{y}^1, \mathbf{y}^2)$ *(la covariance ne change pas si on ajoute à une variable une constante).*

1.3 Variance empirique et écart-type empirique

Définition 7 (Variance empirique) *La variance empirique de* $\mathbf{y}$ *est:*

$$var(\mathbf{y}) = cov(\mathbf{y}, \mathbf{y}) = \sum_{i \in I} p_i (y_i^j - \bar{y^j})^2$$

.

L'écart-type empirique *est* $\sigma_y = \sqrt{var(\mathbf{y})}$.

Propriétés:

- $var(\mathbf{y}) = 0 \Leftrightarrow \forall i \in I, \quad y_i = \bar{y}$ (la variable est constante),
- $var(\mathbf{y}) = \sum_{i \in I} p_i y_i^2 - \bar{y}^2,$

- $var(\mathbf{y}) = 1/2 \sum_{i \in I} \sum_{i' \in I} p_i p_i'(y_i - y_i')^2$, (moyenne quadratique pondérée des écarts entre deux observations)
- $\forall \alpha$ et $\beta \in \mathbf{R}$:
$$var(\alpha \mathbf{y} + \beta 1_n) = var(\alpha \mathbf{y}) = \alpha^2 var(\mathbf{y}).$$

Remarque 1.3.1 *La transformation $\mathbf{y} \to \mathbf{y} + \beta \mathbf{1}_n$ correspond à un changement de l'origine de l'échelle des mesures, et la transformation $\mathbf{y} \to \alpha \mathbf{y}$ correspond à un changement d'unité.*

Définition 8 *On appelle variable* **réduite** *ou* **centrée réduite** *associée à* $\mathbf{y}$ *la variable $z = (z_i)_{i \in I}$ telle que:*
$$z_i = \frac{(y_i - \bar{y})}{\sigma_y}$$

Propriété 2 $\bar{z} = 0$ *;* $var(z) = 1$ *.*

1.4 Coefficient de corrélation linéaire empirique

Définition 9 *Le coefficient de corrélation linéaire empirique de $\mathbf{y}^1$ et $\mathbf{y}^2$ est défini par:*
$$r(\mathbf{y}^1, \mathbf{y}^2) = \frac{cov(\mathbf{y}^1, \mathbf{y}^2)}{\sqrt{var(\mathbf{y}^1) var(\mathbf{y}^2)}}$$

Propriété 3 *Il possède les propriétés suivantes:*

- *Domaine de variation:* $r(\mathbf{y}^1, \mathbf{y}^2) \in [-1, +1]$
- *Symétrie* $r(\mathbf{y}^1, \mathbf{y}^2) = r(\mathbf{y}^2, \mathbf{y}^1)$
- $r(\mathbf{y}, \mathbf{y}) = 1$ *et* $r(\mathbf{y}^1, \mathbf{y}^2) = \pm 1 \Leftrightarrow \exists \alpha$ *et* $\beta \ \mid \ \forall i \in I : y_i^1 = \alpha y_i^2 + \beta$ *avec* $signe(\alpha) = signe(r(\mathbf{y}^1, \mathbf{y}^2))$ *.*
- *Propriétés d'invariance:*
$\forall \alpha, \beta, \alpha', \beta' \in \mathbf{R},\ \alpha > 0\ \alpha' > 0\ \ r(\alpha \mathbf{y}_1 + \beta \mathbf{1}_n, \alpha' \mathbf{y}_2 + \beta' \mathbf{1}_n) = r(\mathbf{y}^1, \mathbf{y}^2).$

Remarque 1.4.1 *Le coefficient de corrélation linéaire ne mesure la liaison que lorsque celle-ci est de type linéaire (voir figure 1.1) des exemples de cas non-linéaires). Il suppose aussi une "bonne" répartition des observations (voir sur la même figure des exemples avec valeurs dites "aberrantes"). On peut éliminer ce type de problème en utilisant le coefficient de corrélation des rangs: on remplace les valeurs observées de chaque variable par leurs rangs et on ne s'intéresse qu'à l'ordre des observations.*

1.5 Interprétation géométrique de quelques indices statistiques

On munit l'espace des variables $\mathbf{R}^n$ du produit scalaire défini par:
$$< \mathbf{x}, \mathbf{x}' > = \sum_{i \in I} p_i x_i x_i',$$

et on note $\| \ \|$ la norme associée: $\|\mathbf{x}\| = < \mathbf{x}, \mathbf{x} >^{1/2}$. et $\mathbf{x} = \mathbf{y} - \overline{\mathbf{y}} \mathbf{1}_n$ la variable centrée de $\mathbf{y}$.
On peut écrire:

- $\overline{\mathbf{y}} = < \mathbf{y}, \mathbf{1}_n >$
- $\mathbf{y}$ est centré $\Leftrightarrow \mathbf{y} \perp \mathbf{1}_n$. Donc $\mathbf{1}_n^{\perp}$, orthogonal dans $\mathbf{R}^n$ de la droite engendrée par $\mathbf{1}_n$, est le sous-espace des variables centrées de $\mathbf{R}^n$.
- $\mathbf{y} = \overline{\mathbf{y}} \mathbf{1}_n + (\mathbf{y} - \overline{\mathbf{y}} \mathbf{1}_n)$. Cette décomposition orthogonale montre que la variable centrée de $\mathbf{y}$ s'obtient en projetant orthogonale-ment $\mathbf{y}$ sur le sous-espace des variables centrées $\mathbf{1}_n^{\perp}$;
- $var(\mathbf{y}) = \|\mathbf{y} - \bar{y} \mathbf{1}_n\|^2 = \|\mathbf{x}\|^2$;

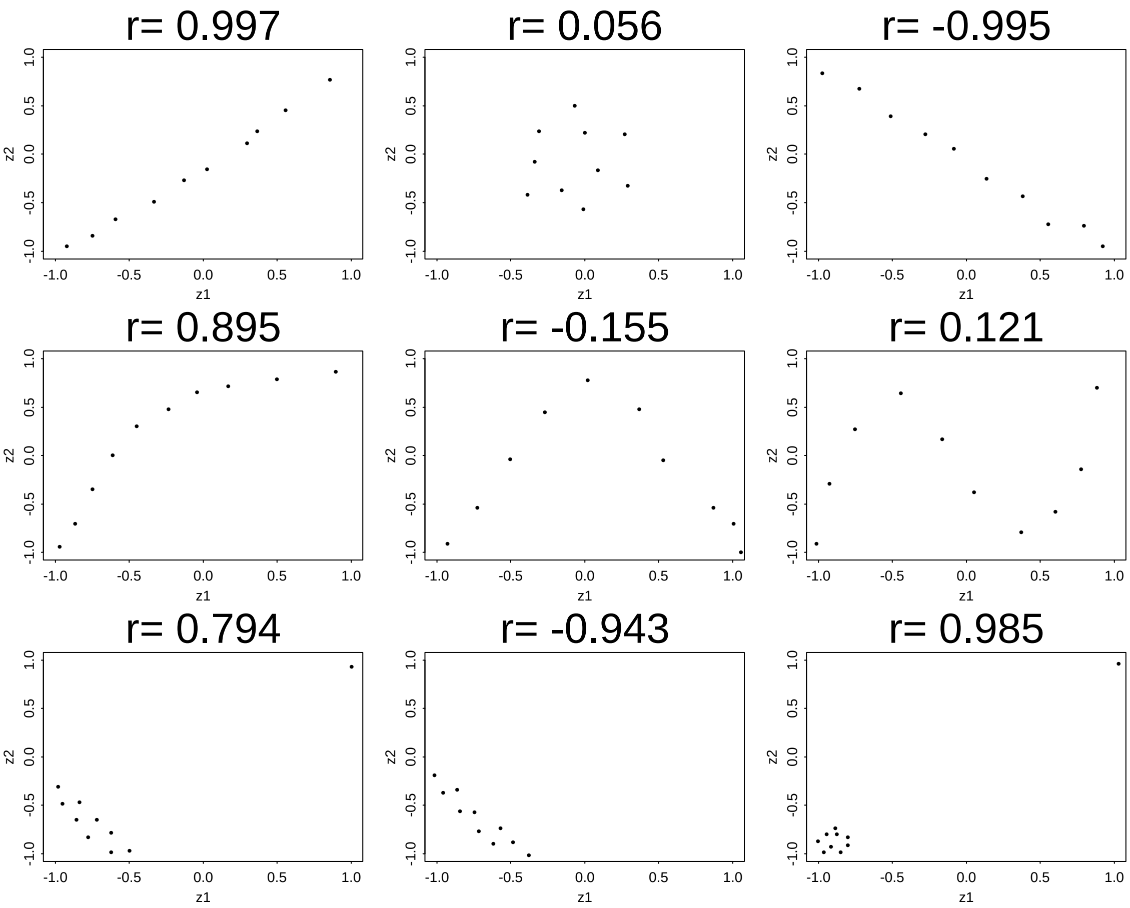


FIG. 1.1 - *Le coefficient de corrélation linéaire dans différents cas de figures*

$\sigma_y = \|\mathbf{y} - \bar{y}\mathbf{1}_n\|$: l'écart-type d'une variable s'interprète comme la longueur de la variable centrée de $\mathbf{y}$

- $cov(\mathbf{y}^1, \mathbf{y}^2) =< \mathbf{y}^1 - \bar{y}^1\mathbf{1}_n, \mathbf{y}^2 - \bar{y}^2\mathbf{1}_n >=< \mathbf{x}^1, \mathbf{x}^2 >$;
- $r(\mathbf{y}^1, \mathbf{y}^2) = \frac{<\mathbf{y}^1 - \bar{y}^1\mathbf{1}_n, \mathbf{y}^2 - \bar{y}^2\mathbf{1}_n>}{\|\mathbf{y}^1 - \bar{y}^1\mathbf{1}_n\|\|\mathbf{y}^2 - \bar{y}^2\mathbf{1}_n\|} = \frac{<\mathbf{x}^1, \mathbf{x}^2>}{\|\mathbf{x}^1\|\|\mathbf{x}^2\|} = cos(\mathbf{x}^1, \mathbf{x}^2)$: la coefficient de corrélation de deux variables s'interprète comme le cosinus des variables centrées associées.

1.6 La matrice de variances-covariances

Soit p variables quantitatives $\{\mathbf{y}^j, j = 1, \ldots, p\}$ où $\mathbf{y}^j$ est le vecteur colonne de $\mathbf{R}^n$ d'éléments y_i^j; on note $D = diag(p_i)$ la matrice diagonale d'ordre n d'éléments p_i. Enfin Y et X sont les matrices $n \times p$ de colonnes $\mathbf{y}^j$ et $\mathbf{x}^j$ respectivement.

Définition 10 *La matrice $n \times p$ contenant les variables $\mathbf{y}^j$ en colonne est appelé* **tableau de données***; la matrice X, contenant en colonne les variables centrées, est le* **tableau centré***;*

Définition 11 *La matrice Γ définie par $\Gamma_{jk} = cov(\mathbf{y}^j, \mathbf{y}^k)$ est appelée* **matrice de variance-covariance empirique** *des variables $\mathbf{y}^1, \ldots, \mathbf{y}^p$.*

On a comme expressions matricielles:

- $\bar{y^j} = \mathbf{y}^{j\prime} D\mathbf{1}_n = \mathbf{1}_n' D\mathbf{y}^j$;
- $cov(\mathbf{y}^j, \mathbf{y}^k) = (\mathbf{y}^j - \bar{y^j}\mathbf{1}_n)' D(\mathbf{y}^k - \bar{y^k}\mathbf{1}_n) = \mathbf{x}^{j\prime} D\mathbf{x}^k$
- $var(\mathbf{y}^j) = (\mathbf{y}^j - \bar{y^j}\mathbf{1}_n)' D(\mathbf{y}^j - \bar{y^j}\mathbf{1}_n) = \mathbf{x}^{j\prime} D\mathbf{x}^j$
- $\Gamma = X' DX$
- Soit $\mathbf{a} = (a_j)_{j \in J}$ et $\mathbf{b} = (b_j)_{j \in J}$ deux vecteurs de $\mathbf{R}^p$. Alors $Y\mathbf{a} = \sum_{j \in J} a_j \mathbf{y}^j$, $Y\mathbf{b} = \sum_{j \in J} b_j \mathbf{y}^j$ et $cov(\sum_{j \in J} a_j \mathbf{y}^j, \sum_{j \in J} b_j \mathbf{y}^j) = \mathbf{a}\Gamma\mathbf{b}$

Propriété 4 *La matrice* Γ *est* **symétrique** *(*$\Gamma' = \Gamma$*) et* **positive** *(pour tout vecteur* $\mathbf{u}$ *de* $\mathbf{R}^p$*,* $\mathbf{u}'\Gamma\mathbf{u} \geq 0$*).*

Propriété 5 *La matrice* Γ *est définie si et seulement si les variables* $\mathbf{y}^j$ $(j = 1, \ldots, p)$ *sont, une fois centrées, linéairement indépendantes.*

1.7 Liaison variable quantitative - qualitative

On suppose que les individus sont munis de poids $p_i > 0$, de somme 1. La variable qualitative, notée $\mathcal{X}$, a q modalités et définit une partition sur l'ensemble des individus. On note G_k l'ensemble des indices du groupe k, $\Theta_k = \sum_{i \in G_k} p_i$ le poids total du groupe k, $\mathbf{y}$ la variable quantitative, et $\overline{y}$ et $\overline{y}_k$ respectivement la moyenne générale de $\mathbf{y}$ et la moyenne de $\mathbf{y}$ dans le groupe k.

1.7.1 Variance inter et intra classe

On définit la variance inter-classe de $\mathbf{y}$, notée $var_B(\mathbf{y})$, comme étant la variance pondérée des moyennes de classe et la variance intra-classe de $\mathbf{y}$, notée $var_W(\mathbf{y})$, comme la moyenne des variances à l'intérieur des classes, soit :

$$var_B(\mathbf{y}) = \sum_{k=1}^{K} \Theta_k (\overline{y}_k - \overline{y})^2 \quad \text{et} \quad var_W(\mathbf{y}) = \sum_{k=1}^{K} \Theta_k \sum_{i \in G_k} \frac{p_i}{\Theta_k} (y_i - \overline{y}_k)^2$$

On a la décompositon suivante :

$$var_(\mathbf{y}) = var_B(\mathbf{y}) + var_W(\mathbf{y}).$$

1.7.2 Le rapport de corrélation empirique

C'est un indice mesurant la liaison entre une variable qualitative et une variable quantitative.

Définition 12 *La mesure de liaison est donnée par le* rapport de corrélation empirique*, rapport de la variance inter-classe sur la variance totale :*

$$\eta^2 = \frac{var_B(\mathbf{y})}{var(\mathbf{y})} \tag{1.2}$$

Ce rapport a les propriétés suivantes :

- Il est égal à 1 si la variable $\mathbf{y}$ est constante dans les classes.
- Il est égal à 0 si les moyennes des classes sont toutes égales.
- Il varie entre 0 et 1 et représente la proportion de variation de $\mathbf{y}$ "expliquée" par la variable qualitative.
- Sous l'hypothèse d'égalité des moyennes de classe, d'indépendance des observations et de distribution gaussienne des résidus, la quantité :

  $$\frac{\eta^2}{1-\eta^2} * \frac{n-q}{q-1} = \frac{\frac{var_B(\mathbf{y})}{q-1}}{\frac{var_W(\mathbf{y})}{n-q}}$$

  suit une loi de Fisher a $q-1$ et $n-q$ degrés de liberté.

Cette dernière propriété vient du fait que cette statistique est le rapport du carré des normes de deux vecteurs : la projection du vecteur $\mathbf{y}$ sur le sous-espace engendré par les indicatrices centrées de la variable explicative et la projection du même vecteur sur l'orthogonal dans le sous-espace des variables centrées. On divise les numérateurs et dénominateurs respectivement par la dimension de ces deux espaces (le nombre de degrés de liberté). Alors le rapport obtenu si les deux sous-espaces sont choisis arbitrairement doit tendre vers 1 si n est grand et suit une loi de Fisher.

1.8 Liaison entre variables qualitatives

Ici l'indice de liaison s'interprète comme une distance entre deux distributions. On se donne deux variables qualitatives de modalités I et J. Leur table de contingence contient dans chaque cellule (i, j) les effectifs des individus appartenant aux catégories i de I et j de J. Prenons l'exemple de la répartition des étudiants étrangers en France:

etetr	DROI	SCEC	LETT	SCIE	MEDE	PHAR	ODON	PLUR	IUT
euro	2058	1427	12119	2507	1503	216	78	172	499
asie	1414	1629	7940	4265	3098	619	248	784	348
afri	8696	10297	15185	14092	5298	2875	537	429	1844
amer	726	899	6839	1521	725	37	19	71	89

TAB. 1.1 - *Nombre d'étudiants étrangers selon les études suivies et la nationalité*

La statistique du $\chi 2$ est une mesure de liaison entre ces deux variables; c'est une mesure de distance entre:

- la distribution ajustée sous hypothèse d'indépendance,
- la distribution observée.

On désigne par n_{ij} l'effectif de la case (i, j) de la table de contingence, par $n_{i.}$ l'effectif de la ligne i, par $n_{.j}$ celui de la colonne j et par n l'effectif total de la table. Enfin I et J désignent respectivement le nombre de lignes et de colonnes du tableau. Sous hypothèse d'indépendance, le tableau des "effectifs théoriques" a pour éléments $N_{ij} = n_{i.}n_{.j}/n$ et la statistique du $\chi 2$ s'écrit :

$$\begin{aligned} \chi 2 &= \sum_i \sum_j \frac{(n_{ij} - N_{ij})^2}{N_{ij}} \\ &= \sum_i \sum_j n \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{n_{i.}n_{.j}} \end{aligned} \quad (1.3)$$

Sous l'hypothèse d'indépendance, cette statistique suit une loi du $\chi 2$ à $(I-1)(J-1)$ degrés de liberté. Dans notre exemple, le $\chi 2$ vaut $14551,8$ pour 3×8 degrés de liberté. Au seuil $\alpha = 0.05$, on rejette l'hypothèse H_0 d'indépendance si le $\chi 2$ observé dépasse la valeur critique $\chi 2_{0.05}$, telle que sous H_0 $\Pr(\chi 2 > \chi 2_{0.05}) = 1 - 0.05$. Pour 24 d.d.l. cette valeur seuil est de $36,4$. Elle est largement dépassée, donc la liaison est hautement significative (pour des détails sur cette notion, voir le paragraphe 4.2).

1.9 Exercices

Exercice 1.9.1 Exercice sur le coefficient de qualité R^2 d'une régression On considère deux variables fixées $\mathbf{x}$ et $\mathbf{y}$ de $\mathbf{R}^n$ et on cherche à modéliser $\mathbf{y}$ en fonction de $\mathbf{x}$. Pour cela, on considère un sous-espace vectoriel $\mathcal{F}$ de fonctions réelles, dont on suppose qu'il contient la fonction constante $\mathbf{1}$. On désigne par $\mathbf{x}$ le vecteur :

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \dots \\ x_i \\ \dots \\ x_n \end{pmatrix} \text{ et par } f(\mathbf{x}) \text{ le vecteur } \begin{pmatrix} f(x_1) \\ \dots \\ f(x_i) \\ \dots \\ f(x_n) \end{pmatrix}.$$

On cherche f dans $\mathcal{F}$ tel que :

$$y_i = f(x_i) + e_i \text{ avec } \sum_{i=1}^{n} e_i^2 \text{minimum. } (1)$$

1. On munit $\mathbf{R}^n$ de la métrique euclidienne usuelle. Montrer que $V = \{f(\mathbf{x}),\ f \in \mathcal{F}\}$ est un sous-espace de $\mathbf{R}^n$, et que, si P_V désigne l'opérateur de projection orthogonale sur

V, la solution au problème (1) consiste à prendre $f(\mathbf{x}) = P_V(y)$. En déduire que $\bar{\mathbf{e}} = 0$ et que :

- le coefficient de corrélation linéaire $r(f(\mathbf{x}), \mathbf{e})$ est nul.
- on a la relation entre les variances empiriques :

$$var(\mathbf{y}) = var(f(\mathbf{x})) + var(\mathbf{e}).$$

2. On pose $R^2 = var(f(\mathbf{x}))/var(\mathbf{y})$ et on considère cet indice comme un coefficient de qualité de la régression.

   (a) Dans quel cas a-t'on $R^2 = 0$, $R^2 = 1$?.

   (b) Montrer que $R^2 = r^2(f(\mathbf{x}), \mathbf{y})$.

3. Lorsque $\mathcal{F}$ est l'ensemble des fonctions linéaires affines réelles, à quoi est égal le coefficient R^2?

Exercice 1.9.2 Extrait de l'examen de Septembre 1994 (Analyse des données du magistère)
Diana Murray envoie, sur le réseau des usagers de Splus, le message suivant (traduction libre) :

```
J'ai travaillé avec des séries  chronologiques et j'ai observé quelque
chose  de  curieux  que  je  ne  comprends  pas.  Cela a rapport  avec
l'inversion  d'une  matrice  de  corrélation  obtenue  à partir  d'une
matrice constituée de variables en colonnes.  J'ai trouvé que  lorsque
j'ai plus de variables que d'observations et quand j'essaye d'inverser
la matrice de corrélation, j'ai un message d'erreur qui  me dit que la
matrice  n'est pas  inversible.  Cela apparait même  lorsque j'utilise
comme séries chronologiques des nombres aléatoires.
Dans le premier  exemple,  j'ai 100 variables et  101 observations. La
matrice  de  corrélation  est  inversible.  Si je ne met que 100 ou 99
observations dans les variables,  la  matrice de corrélation n'est pas
inversible....
Je peux imaginer que la colinéarité peut arriver avec  des matrices de
corrélation de grandes dimensions,  mais je ne  comprends pas pourquoi
cela devrait dépendre si fortement du nombre des observations....

I would greatly appreciate any insight or  guidance  that one may have
offer !
```

Que pouvez vous répondre à Diana?

Chapitre 2

Algèbre linéaire : pré-requis, rappels, compléments

2.1 Applications linéaires et opérateurs

2.1.1 Application linéaire et matrice associée

Remarque 2.1.1 *Tous les espaces vectoriels considérés ici sont des espaces vectoriels sur R de dimension finie.*

Sont supposées connus les notions et propriétés suivantes (certaines des définitions sont cependant rappellées) :

- Application linéaire g d'un espace vectoriel E dans un espace vectoriel F (on désigne dans la suite par $\mathcal{L}_{E,F}$ l'espace vectoriel des applications linéaires de E dans F).
- Forme linéaire : application linéaire d'un espace vectoriel E dans $\mathbf{R}$ ($F = \mathbf{R}$).
- Endomorphisme ou opérateur ($F = E$) : application linéaire d'un espace vectoriel dans lui-même (noté dans la suite $\mathcal{L}_E$). Un endomorphisme particulier est l'application identité (noté id_E).
- Noyau $Ker(g)$ et image $Im(g)$ d'une application linéaire g :
  $Ker(g) = \{x \in E | \, g(x) = 0_F\}$; $Im(g) = \{y \in F | \, \exists x \in E \text{ tel que } y = g(x)\}$
- Rang d'une application g (noté $rang(g)$) : $rang(g) = dim(Im(g))$.
- Matrice associée à g dans les bases $\mathcal{E} = \{e_j, j = 1, \ldots, p\}$ et $\mathcal{F} = \{f_i, i = 1, \ldots, n\}$: on la note
  $A = mat\,(g, \mathcal{E}, \mathcal{F})$. C'est une matrice $n \times p$ qui a pour colonne j le vecteur des coordonnées de $g(e_j)$ dans la base $\mathcal{F}$:
  Si on a $g(e_j) = \sum_{i=1}^{n} a_{ij} f_i$, alors la matrice associée A s'écrit :

$$A = \left(\begin{array}{cc|c|cc} a_{11} & \ldots & a_{1j} & \ldots & a_{1p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{i1} & \ldots & a_{ij} & \ldots & a_{ip} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nj} & \ldots & a_{np} \end{array} \right)$$

  On note $\mathcal{M}_{n,p}$ l'espace vectoriel des matrices à n lignes et p colonnes). La matrice de l'opérateur identité est une matrice diagonale (id.est. dont les éléments extradiagonaux sont nuls) ayant pour éléments diagonaux des 1. On la note I (ou encore I_n pour indiquer qu'elle elle est carrée d'ordre n).
- Propriété : l'image d'une matrice A est le sous-espace vectoriel engendré par ses colonnes.

Notations Si $\{u_1, \ldots, u_p\}$ est une famille de vecteurs, $[u_1, \ldots, u_p]$ désigne le sous-espace engendré par cette famille. Donc si $U = (u_1 | \ldots | u_p)$, alors $Im(U) = [u_1, \ldots, u_p]$. On notera encore $\mathbf{R}.u$ la droite vectorielle $[u]$.

2.1.2 L'isomorphisme canonique entre $(E, \mathcal{E})$ et R^p

Définition 13 *Si E est un espace vectoriel de dimension finie p, munie d'une base $\mathcal{E}$, on appelle* isomorphisme canonique *entre E et R^p l'application qui associe à un vecteur x de E le vecteur de ses p coordonnées dans la base $\mathcal{E}$.*

Conséquences Lorsqu'une seule base de référence est définie sur E, on pourra noter de la même manière un élément de E et le vecteur de ses coordonnées dans la base $\mathcal{E}$, élément de R^p. Cette facilité sera souvent utilisée dans la suite. De la même manière, si les bases de référence de E et F sont uniques, on pourra noter de la même manière l'application linéaire g et sa matrice associée A. Ce sera encore une pratique usuelle dans la suite.

2.1.3 Calcul matriciel

Propriété 6 *Matrice associée à un composé d'applications (produit matriciel)*
Soit $B = (b_{kj})_{k=1,\ldots,q,\ j=1,\ldots,p}$ la matrice d'une application linéaire $g \in \mathcal{L}_{E,F}$, soit $A = (a_{ik})_{i=1,\ldots,n,\ k=1,\ldots,q}$ la matrice d'une application linéaire $f \in \mathcal{L}_{F,G}$. Alors la matrice $C = (c_{ij})_{i=1,\ldots,n,\ j=1,\ldots,p}$ de l'application linéaire composée $f \circ g$ est telle que :

$$c_{ij} = \sum_{k=1}^{q} a_{ik} b_{kj}.$$

On l'appelle matrice produit de A par B et on écrit $C = AB$.

$$\begin{tikzcd}
E \arrow[r, "g", "B"'] \arrow[rr, bend left, "f \circ g"] & F \arrow[r, "f", "A"'] & G
\end{tikzcd}$$

On remarque que le produit n'est défini que si A a autant de colonnes que B a de lignes. Si A est une matrice $n \times q$ et B une matrice $q \times p$, alors C est une matrice $n \times p$.

Produit par des matrices diagonales

- Prémultiplier une matrice A par une matrice diagonale D revient à multiplier chaque ligne i de A par l'élément diagonal de rang i.
- Postmultiplier une matrice A par une matrice diagonale D revient à multiplier chaque colonne j de A par l'élément diagonal de rang j.

Produit par bloc Il est souvent pratique de partitionner une matrice en blocs. Par exemple, une matrice $n \times p$ peut être considérée comme un ensemble de n lignes empilées verticalement ou comme un ensemble de p colonnes accolées horizontalement. On suppose dans la suite que les deux matrices A et B définies ci-dessus sont partitionnées en blocs délimitées par des lignes horizontales ou verticales. On indice maintenant non plus les éléments mais les blocs. Ainsi dans la suite, A_{ik} désigne une matrice, située dans la i-ème ligne-bloc et dans la j-ième colonne-bloc de A. On partitionne ensuite la matrice B comme sur le schéma ci-dessous, de sorte que B_{kj} ait autant de lignes que A_{ik} a de colonnes (donc le découpage en lignes-bloc de B doit être fait comme le découpage en colonnes-bloc de A).

$$A = \left(\begin{array}{c|c|c|c|c}
A_{11} & \ldots & A_{1k} & \ldots & A_{1q} \\ \hline
\ldots & \ldots & \ldots & \ldots & \ldots \\ \hline
A_{i1} & \ldots & A_{ik} & \ldots & A_{iq} \\ \hline
\ldots & \ldots & \ldots & \ldots & \ldots \\ \hline
A_{n1} & \ldots & A_{nk} & \ldots & A_{nq}
\end{array}\right) \quad B = \left(\begin{array}{c|c|c|c|c}
B_{11} & \ldots & B_{1j} & \ldots & B_{1p} \\ \hline
\ldots & \ldots & \ldots & \ldots & \ldots \\ \hline
B_{k1} & \ldots & B_{kj} & \ldots & B_{kp} \\ \hline
\ldots & \ldots & \ldots & \ldots & \ldots \\ \hline
B_{q1} & \ldots & B_{qj} & \ldots & B_{qp}
\end{array}\right)$$

Alors les produits matriciels $A_{ik}B_{kj}$ sont définis quelque soit i et j, et en notant :

$$C_{ij} = \sum_{k=1}^{q} A_{ik} B_{kj},$$

on obtient en regroupant les blocs C_{ij} (avec i indice de ligne bloc, et j indice de colonne bloc) une matrice C partitionnée égale au produit matriciel AB.

Comme cas particuliers importants, on peut considérer le cas où A (matrice $n \times p$) et B (matrice $m \times q$) ne sont pas partitionnées en ligne mais chacune de leurs colonnes, notée respectivement $\mathbf{a}_i$ et $\mathbf{b}_j$, constitue un bloc.

$$A = (\mathbf{a}_1 | \dots, | \mathbf{a}_p) \quad B = (\mathbf{b}_1 | \dots, | \mathbf{b}_q)$$

- Si $p = q$, le produit $C = AB'$ est alors le bloc unique C_{11} tel que $C = \sum_{i=1}^{p} \mathbf{a}_i \mathbf{b}_i'$ (C est une somme de matrices de rang 1). De plus, si $\Lambda = diag(\lambda_1, \dots, \lambda_p)$, alors $A\Lambda B' = \sum_{i=1} \lambda_i \mathbf{a}_i \mathbf{b}_i'$.
- Si $m = n$, le produit $A'B$ a pour bloc $C_{ij} = \mathbf{a}_i' \mathbf{b}_j$ (matrice 1×1 identifiée à un réel). On retrouve la définition du produit matriciel.

Exercice 2.1.1 Montrer que la somme de k matrices de rang 1 de mêmes dimensions est une matrice de rang inférieure ou égale à k.

2.1.4 Automorphisme ou opérateur régulier:

- Définition : un opérateur $g \in \mathcal{L}_E$ est dit *régulier* s'il définit une bijection de E dans E (un tel endomorphisme est alors appelé *automorphisme*).
- Matrice inverse : soit I la matrice identité et A une matrice carrée. On appelle *inverse de* A, et on note A^{-1} une matrice telle que $AA^{-1} = A^{-1}A = I$; une matrice admettant un inverse est dite *inversible*.
- Caractérisations d'un opérateur régulier :
  1. $g \in \mathcal{L}_E$ est régulier $\Leftrightarrow Ker(g) = \{0_E\}$,
  2. $g \in \mathcal{L}_E$ est régulier $\Leftrightarrow rang(g) = dim E$,
  3. $g \in \mathcal{L}_E$ est régulier $\Leftrightarrow det(A) \neq 0$ (où A une matrice associée à g dans une base quelconque),
  4. $g \in \mathcal{L}_E$ est régulier $\Leftrightarrow$ si A est une matrice associée à cet opérateur dans une base quelconque, A est inversible.

2.1.5 Opérateur de projection

Définition 14 (Opérateur idempotent) *P idempotent* $\Leftrightarrow P^2 = P$

Propriété 7 *P est idempotent*

$$\begin{aligned} &\Leftrightarrow \quad P_{|Im(P)} = id_{|Im(P)} \\ &\Rightarrow \quad E = Im(P) \oplus Ker(P) \end{aligned}$$

Définition 15 *On appelle* opérateur de projection *un opérateur idempotent (P projette sur $Im(P)$ parallèlement à $Ker(P)$).*

2.1.6 L'application trace

- Trace d'une matrice : on appelle trace d'une matrice carrée A et on note $trace(A)$ la somme des éléments diagonaux de A.
- Propriétés :
  - L'application *trace* est une forme linéaire sur l'ensemble des matrices carrées d'ordre p;
  - Si A est une matrice $p \times n$ et B une matrice $n \times p$, alors $trace(AB) = trace(BA)$;
  - Si A et P sont des matrices carrées régulières d'ordre p, $trace(P^{-1}AP) = trace(A)$
- Conséquences de la dernière propriété : on définit la trace d'un opérateur comme étant la trace d'une des quelconques matrices qui la représente. L'application *trace* est donc aussi une forme linéaire sur $\mathcal{L}_E$.

Propriété 8 *P opérateur de projection* $\Rightarrow trace(P) = rang(P)$

2.1.7 Décomposition spectrale d'un opérateur

- **Définitions**
  - Valeurs et vecteurs propres (réelles) d'un opérateur : on dit que $\mathbf{u}$ est un *vecteur propre* de l'opérateur f de $\mathcal{L}_E$ s'il est non nul et s'il existe $\lambda \in \mathbf{R}$ tel que $f(\mathbf{u}) = \lambda \mathbf{u}$. Le réel λ est appelé *valeur propre* de f.
  - Polynôme caractéristique : soit A la matrice associé à f dans une base de référence donnée; λ est une valeur propre de f si et seulement si

$$Ker(f - \lambda id_E) \neq \{0_E\},$$

    donc encore si et seulement si $det(A - \lambda I) = 0$. La fonction de λ $det(A - \lambda I)$ est un polynôme de degré n en λ appelé *polynôme caractéristique de f*.
  - Espace propre associé à la valeur propre λ : c'est le noyau (non réduit à $\{0_E\}$) de $Ker(f - \lambda id_E)$.
  - Ordre de multiplicité algébrique d'une valeur propre λ : c'est l'ordre de multiplicité de λ dans le polynôme caractéristique.
  - Ordre de multiplicité géométrique d'une valeur propre λ : c'est la dimension de l'espace propre associé.
  - Spectre d'un opérateur diagonalisable : c'est la suite de ses valeurs propres distinctes (on le note $Sp(A)$).
  - Opérateur diagonalisable : un opérateur est diagonalisable si il existe une base de E formée de vecteurs propres de f (dans cette base, la matrice de cet opérateur est diagonale).
  - Projecteur propre : c'est un projecteur sur un sous-espace propre, admettant comme noyau la somme des autres sous-espaces propres.
- **Propriétés**
  - Conditions nécessaires et suffisantes pour qu'un opérateur soit diagonalisable. Un opérateur est diagonalisable si et seulement si il satisfait l'une des conditions équivalentes suivantes :
    1. Toute valeur propre de f a même ordre de multiplicité algébrique et géométrique,
    2. La somme des dimensions des espaces propres de f est égale à la dimension de l'espace E,
  - Une condition suffisante pour qu'un opérateur soit diagonalisable est qu'il ait des valeurs propres distinctes.
  - Calcul d'un projecteur propre : Le projecteur propre associé à la valeur propre λ s'écrit :

$$P_\lambda = \prod_{\lambda' \in Sp(A),\ \lambda' \neq \lambda} \frac{(f - \lambda' id_E)}{\lambda - \lambda'}$$

  - Décomposition spectrale d'un opérateur diagonalisable
    - En fonction des projecteurs propres :

$$A = \sum_{\lambda \in sp(A)} \lambda \ P_\lambda ,$$

      où $sp(A)$ et P_λ sont respectivement le spectre de A et le projecteur propre associé à la valeur propre λ.
    - En fonction de projecteurs de rang 1 : soit $\{\mathbf{u}_j, j = 1, \ldots, p\}$ une base de vecteurs propres de A, $\{\lambda_j, j = 1, \ldots, p\}$ la suite des valeurs propres associées. Alors A s'écrit :

$$A = \sum_{j=1,\ldots,p} \lambda_j \quad P_{u_j} .$$

      Dans cette expression, P_{u_j} désigne le projecteur tel que $Im(P_{u_j}) = [u_j]$ et $Ker(P_{u_j}) = [u_{j'}, j' \neq j]$.

- Équation (matricielle) aux valeurs propres.

$$AU = U\Lambda \ ,$$

- Trace d'un opérateur diagonalisable: la trace d'un opérateur diagonalisable est égale à la somme de ses valeurs propres.
- Calcul des puissances d'un opérateur diagonalisable:

$$A^r = \sum_{\lambda \in sp(A)} \lambda^r P_\lambda \ , \ \ \forall r \in \mathbf{N}$$

  Si les valeurs propres sont strictement positives, cette propriété reste vraie pour tout $r \in \mathbf{Z}$ et $r \in \mathbf{Q}$.
- Un opérateur diagonalisable est inversible si et seulement si ses valeurs propres sont non nulles; l'inverse d'un tel opérateur s'obtient en remplaçant dans sa décomposition spectrale ses valeurs propres par leurs inverses (c'est encore l'application de la propriété précédente pour $r = -1$, ce qui montre qu'elle est vraie dans $\mathbf{Z}$ si l'opérateur est inversible).
- L'image d'un opérateur diagonalisable est engendrée par ses vecteurs propres associés à des valeurs propres non nulles. Son rang est égal au nombre de valeurs propres non nulles (en comptant k fois une valeur propre d'ordre de multiplicité k).
- Un opérateur f est un opérateur de projection si et seulement si il est diagonalisable et a un spectre égal à $\{0, 1\}$.

2.2 Notions euclidiennes

2.2.1 Définitions et propriétés

1. *Produit scalaire :* g est un produit scalaire sur E si c'est une application de $E \times E$ dans $\mathbf{R}$:

   - bilinéaire,
   - symétrique $(\forall x, y \in E, g(x, y) = g(y, x))$
   - définie $(\forall x \in E, g(x, x) = 0 \Longrightarrow x = 0)$
   - et positive $(\forall x \in E, g(x, x) \geq 0)$.

   On notera dans la suite $< x, y >$ le produit scalaire de x et y.

2. *Espace euclidien :* un espace euclidien est un espace vectoriel de dimension finie munie d'un produit scalaire.

3. *Norme euclidienne :* la norme euclidienne associée au produit scalaire $< , >$ de E est notée $\|\ \|$ et est définie par

$$\forall x \in E, \ \ \|x\| = < x, x >^{1/2}$$

4. *Distance euclidienne :* la distance euclidienne associée au produit scalaire $< , >$ de E est définie par $d(x, y) = \|x - y\|$, où $\|\ \|$ est la norme définie ci-dessus.

5. *Matrice de distance :* soit $\mathcal{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_p\}$ une base de E. On appelle *matrice de distance* ou *matrice associée au produit scalaire* $<, >$ dans la base $\mathcal{E}$ et on note $M = mat(<, > , \mathcal{E})$ la matrice M définie par :

$$(M)_{ij} = < e_i, e_j > .$$

6. *Expression matricielle du produit scalaire*; soit x et y deux vecteurs de E, identifiés aux vecteurs de $\mathbf{R}^p$ de leurs coordonnées dans la base $\mathcal{E}$. Le produit scalairez $< x, y >$ a pour expression matricielle:

$$< x, y > \ = x'My.$$

7. Propriété : Une base $\mathcal{E}$ est orthogonale ssi $< e_i, e_j >= 0$ si $i \neq j$ et est orthonormale ssi $< e_i, e_j >= \delta_{ij}$. On en déduit la relation suivante entre propriétés de la base $\mathcal{E}$ et propriétés de la matrice associée M :
   - $\mathcal{E}$ est une base orthonormale $\Leftrightarrow$ $M = I_p$ matrice identité d'ordre p,
   - $\mathcal{E}$ est une base orthogonale $\Leftrightarrow$ M est une matrice diagonale.
8. Soit $U = (\mathbf{u}_1, \dots, \mathbf{u}_r)$ une matrice constituée en colonnes de r vecteurs $\mathbf{u}_1, \dots, \mathbf{u}_r$ de $\mathbf{R}^p$. Alors

   Les vecteurs forment une famille orthonormés $\Leftrightarrow$ $U'MU = I_r$.
9. On appelle *produit scalaire usuel* le produit scalaire défini par la matrice identité dans la base de référence.

Notations Un espace euclidien est noté $(E, < , >)$, ou si une base de référence est définie : $(E,\, \mathcal{E},\, M)$ avec $M = mat(< , > , \mathcal{E})$. Lorsque $E = R^p$ et $\mathcal{E}$ est sa base canonique, on écrit : (R^p, M) au lieu de $(R^p, \mathcal{E}, M)$. La norme associée est notée $\| \ \|_M$.

2.2.2 Propriétés

1. Inégalité de Cauchy-Schwartz :

$$\forall x, y \in E, \quad | < x, y > | \ \leq \ \|x\|.\|y\| \ .$$

   Cette inégalité est à la base de la définition de la notion de *cosinus* dans un espace euclidien.
2. Inégalité de Bessel : Soit $\mathcal{E} = \{e_1, e_2, \dots, e_p\}$ une famille de vecteurs orthonormés. On a :

$$\forall x \in E, \quad \sum_{i=1}^{p} < e_i, x >^2 \leq \|x\|^2$$

3. Si de plus $\mathcal{E}$ est une base, cette inégalité devient une égalité et on a :

$$\forall x \in E, \quad x = \sum_{i=1}^{p} < e_i, x > e_i.$$

   Cette égalité permet un calcul facile des coordonnées d'un vecteur dans une base orthonormée.
4. Application : la trace d'un opérateur A dans la base orthonormale $\mathcal{E}$ s'écrit :

$$trace(A) = \sum_{j=1}^{p} < Ae_j, e_j >$$

2.2.3 Application adjointe d'une application linéaire

Dans ce paragraphe on considère les espaces euclidiens $(E,\, \mathcal{E},\, M)$ et $(F,\, \mathcal{F},\, D)$ et leurs produits scalaires respectifs $< , >_M$ et $< , >_D$.

Propriété 9 *Étant donnée une application f de $\mathcal{L}_{E,F}$, on montre qu'il existe une application unique f^* de $\mathcal{L}_{F,E}$ telle que :*

$$< f(x),\, y >_D = < x, f^*(y) >_M \quad \forall x \in E,\, \forall y \in \ F \ .$$

Définition 16 – *On appelle* application adjointe *l'application f^* définie dans la proposition précédente.*

– *La* matrice adjointe *de $X = mat(f,\, \mathcal{E},\, \mathcal{F})$ est la matrice de l'application adjointe, notée X^* ou $adj(X, M, N)$. On a donc :*

$$X^* = mat(f^*,\, \mathcal{F},\, \mathcal{E}) = M^{-1} X' D \ .$$

Remarque 2.2.1 *Lorsque les produits scalaires sont usuels dans E et F, alors la matrice adjointe est égale à la matrice transposée.*

Propriété 10 *1.* $Imf^* = (Ker(f))^\perp$

2. $Imf = (Ker(f^*))^\perp$

3. $(fg)^* = g^* f^*$

4. $rang(f) = rang(f^*) = rang(f^* \circ f) = rang(f \circ f^*)$

2.2.4 Propriétés des opérateurs d'un espace euclidien

Définition 17 (Opérateur auto-adjoint) *Un opérateur f de l'espace euclidien $(E, \mathcal{E}, M)$ est auto-adjoint ssi il est identique à son adjoint.*

Propriété 11 *On note $X = mat(f, \mathcal{E})$. L'opérateur f est auto-adjoint ssi la matrice MX est symétrique (on dit encore ssi X est M-symétrique).*

Propriété 12 *f auto-adjoint $\Rightarrow E = Im(f) \overset{\perp}{\oplus} Ker(f)$*

Définition 18 (Opérateur positif) *Un opérateur f de l'espace euclidien $(E, \mathcal{E}, M)$ est positif ssi* $\forall x \in E \quad < f(x), x >_M \geq 0$

Définition 19 (Opérateur défini) *Un opérateur f de l'espace euclidien $(E, \mathcal{E}, M)$ est défini ssi* $\forall x \in E,\ x \neq 0 \quad < f(x), x >_M \neq 0$.

Propriété 13 *Si f est une application linéaire d'un espace euclidien E dans un espace euclidien F, alors $f^* \circ f$ et $f \circ f^*$ sont des opérateurs auto-adjoints et positifs (respectivement de $\mathcal{L}_E$ et de $\mathcal{L}_F$).*

Définition 20 (Opérateur de projection orthogonale (ou "orthoprojecteur")) *L'opérateur P est un opérateur de projection orthogonale ssi il est idempotent et auto-adjoint .*

Propriété 14

$$P \text{ est un orthoprojecteur } \Leftrightarrow \begin{cases} E = Im(P) \overset{\perp}{\oplus} Ker(P) \\ P_{Im(P)} = id_{Im(P)} \end{cases}$$

Expression matricielle d'un orthoprojecteur Soit W un sous-espace vectoriel de l'espace euclidien $(\mathbf{R}^p, M)$ engendré par les colonnes de la matrice X ($W = Im(X)$). L'opérateur P_W de projection orthogonale sur W a pour matrice (notée encore P_W) :

$$P_W = X(X'\,MX)^{-1}\,X'\,M$$

Premier cas particulier: X est le vecteur unicolonne x (donc $W = \mathbf{R}.x$) :

$$P_W = \frac{1}{\|x\|^2} x x'\,M$$

Second cas particulier: $X = U$ a pour colonnes des vecteurs deux à deux orthonormés dans $\mathbf{R}^p$ $\quad P_U = UU'M$

Propriété 15 *Si W_1 et W_2 sont deux sous-espaces de E, d'orthoprojecteurs associés P_1 et P_2, alors*

$$\begin{array}{cc} & W_1 \text{ et } W_2 \text{ sont orthogonaux} \\ \Leftrightarrow & P_1 \circ P_2 = 0 \\ \Leftrightarrow & P_2 \circ P_1 = 0. \\ \Leftrightarrow & P_1 + P_2 \text{ est l'orthoprojecteur sur } W_1 \oplus W_2 \end{array}$$

Propriété 16 *Si W_1 et W_2 sont deux sous-espaces de E, d'orthoprojecteurs associés P_1 et P_2, et tels que $W_1 \subset W_2$ alors $P_1 \circ P_2 = P_2 \circ P_1 = P_1$.*

2.2.5 Isométrie

Définition 21 *On dit que l'application linéaire f ($f \in \mathcal{L}_{E,F}$) est une* isométrie *de l'espace euclidien $(E, <,>_E)$ dans l'espace euclidien $(F, <,>_F)$ ssi*

–
$$\forall (x,y) \in E \times E \quad < f(x), f(y) >_F = < x, y >_E \ . \tag{2.1}$$

– *f est surjective*

Une condition équivalente à la condition (2.1) s'écrit :

$$\forall x \in E \quad \|f(x)\|_F = \|x\|_E \ .$$

Les définitions et propriétés qui suivent éclairent la notion d'*inverse généralisée*. Elles peuvent être étudiées à titre d'exercice.

Définition 22 *On dit que f est une* isométrie partielle *ssi :*

$$\forall x, y \in Ker(f)^{\perp} \quad < f(x), f(y) >_F = < x, y >_E \ .$$

Propriété 17 (Propriétés caractéristiques) *f est partiellement isométrique $\Leftrightarrow$*

1. *f^* est partiellement isométrique,*
2. *$f \circ f^*$ est l'orthoprojecteur sur $Im(f)$,*
3. *$f^* \circ f$ est l'orthoprojecteur sur $Im(f^*)$,*
4. *$f^* \circ f \circ f^* = f^*$*
5. *$f \circ f^* \circ f = f$*

Ces deux dernières propriétés définissent des inverses généralisés externes et internes de f.

2.2.6 Décomposition spectrale dans un espace euclidien

Propriété 18 *On a les propriétés suivantes :*

1. *Les valeurs propres des opérateurs positifs sont positives.*
2. *Les espaces propres distincts d'un opérateur autoadjoint sont orthogonaux.*
3. *Tout opérateur autoadjoint est diagonalisable dans une base orthonormée.*
4. *On suppose que $(E, \mathcal{E}, M)$ est un espace euclidien de dimension p et A est un opérateur autoadjoint et positif de E. De plus, $\{\mathbf{u}_s, s = 1, \ldots, r\}$ est une base de vecteurs propres de A, indicés de sorte que leur valeurs propres associées soient rangées par ordre décroissant, et r est le rang de A. La décomposition spectrale de A s'écrit :*

$$A = \sum_{j=1}^{r} \lambda_j u_j u_j' \, M$$

Soit U la matrice $n \times r$ dont les colonnes sont les vecteurs u_j, $\Lambda = diag(\lambda_j, \, j = 1, \ldots, r)$ la matrice diagonale des valeurs propres non nulles, alors :

$$A = U \Lambda U' \, M \ .$$

2.2.7 Un produit scalaire sur l'espace des matrices $n \times p$

Soit $X \in \mathcal{M}_{n,p}$ la matrice à n lignes et p colonnes, considérée comme application linéaire de $(\mathbf{R}^p, M)$ dans $(\mathbf{R}^n, D)$.

Définition 23 (Produit scalaire et norme sur $\mathcal{M}_{n,p}$) *On définit le produit scalaire sur $\mathcal{M}_{n,p}$ par la relation :*

$$< X, Y >_{tr} = trace(XMY' \, D) \ .$$

La norme associée s'écrit:

$$\|X\|_{tr} = (trace(XMX'D))^{1/2}$$

On peut vérifier que cette forme bilinéaire est symétrique définie positive.

En posant $Z = XM^{1/2}$ et en notant z^j les colonnes de Z , on a $\|X\|_{tr}^2 = trace(Z'DZ) = \sum_{j=1}^{p} \|z^j\|_D^2$, avec $\mathbf{z}_j$ j-ième colonne de Z.

Un cas particulier Ce produit scalaire est une généralisation du produit scalaire usuel entre matrices vectorisées : si $M = I_p$ et $D = I_n$, alors :

$$< X, Y >_{tr} = trace(X'Y) = \sum_{i=1}^{n} \sum_{j=1}^{p} x_i^j y_i^j$$

$$< X, Y >_{tr} = < vec(X), vec(Y) >_{n \times m} ,$$

où vec est l'opérateur de vectorisation qui transforme une matrice $n \times m$ en un vecteur de $R^{n \times m}$, et où $<,>_{n \times m}$ est le produit scalaire usuel de cet espace. Dans ce cas particulier, le carré de la distance entre deux matrices s'écrit :

$$d^2(X, Y) = \|X - Y\|_{tr}^2 = \sum_i \sum_j (x_i^j - y_i^j)^2$$

La propriété suivante permet de construire des orthoprojecteurs particuliers dans l'espace des matrices.

Propriété 19 *Soit Q un orthoprojecteur de $(\mathbf{R}^p, M)$, et P un orthoprojecteur de $(\mathbf{R}^n, D)$ Alors les endomorphismes de $(\mathcal{M}_{n,p}, <,>_{tr})$ définis par:*

$$X \longrightarrow XQ'$$

$$X \longrightarrow PX$$

$$X \longrightarrow PXQ' ,$$

sont des orthoprojecteurs. Les sous-espaces sur lesquels ils projettent sont respectivement $\{X \in \mathcal{M}_{n,p} \mid Im(X') \subset Im(Q)\}$, $\{X \in \mathcal{M}_{n,p} \mid Im(X) \subset Im(P)\}$ et l'intersection de ces deux sous-espaces.

Quelques expressions particulières de $trace(XMX'D)$.

On note $(x^j, j = 1, \ldots, p)$ les colonnes de X et $(x_i, i = 1, \ldots, n)$ les transposés [1] des lignes de X.

- Si $M = diag(m_{jj}, j = 1, \ldots, p)$, alors $trace(X'DXM) = \sum_{j=1}^{p} m_{jj} \|x^j\|_D^2$
- Si $D = diag(p_i, i = 1, \ldots, n)$, alors $trace(X'DXM) = \sum_{i=1}^{n} p_i \|x_i\|_M^2$
- Si $M = diag(m_{jj}, j = 1, \ldots, p)$, et $D = diag(p_i, i = 1, \ldots, n)$ alors $trace(X'DXM) = \sum_{i=1}^{n} \sum_{j=1}^{p} p_i m_{jj} (x_i^j)^2$

2.3 Décomposition aux valeurs singulières et approximation matricielle

2.3.1 Décomposition aux valeurs singulières

On considère les espaces euclidiens $(\mathbf{R}^p, M)$ et $(\mathbf{R}^n, D)$ et X une matrice $n \times p$ de rang r de $\mathcal{M}_{n,p}$.

Théorème 1 *Il existe deux familles de vecteurs orthonormés $\{\mathbf{v}_s, \quad s = 1, \ldots, r\}$ dans $(\mathbf{R}^p, M)$ et $\{\mathbf{u}_s, \ s = 1, \ldots, r\}$ dans $(\mathbf{R}^n, D)$, et une suite de réels strictement positifs rangés par ordre décroissant $\{\lambda_s, s = 1, \ldots, r\}$ tels que :*

$$X = \sum_{s=1}^{r} \lambda_s \mathbf{u}_s \mathbf{v}_s' \tag{2.2}$$

C'est une expression en termes vectoriels de la décomposition aux valeurs singulières (DVS) d'une matrice quelconque. L'écriture en termes matriciels de ce théorème est la suivante :

1. Par convention, tous les vecteurs de R^p ou R^n sont écrits en colonne

Expression matricielle de la décomposition aux valeurs singulières Il existe une matrice U $n \times r$, une matrice V $p \times r$ et une matrice diagonale à éléments positifs Λ tels que :

$$X = U\Lambda V' \quad avec \;\; U'DU = V'MV = I_r \tag{2.3}$$

Les matrices U et V on pour colonnes les vecteurs $\mathbf{u}_s$ et $\mathbf{v}_s$, définis ci-dessus. On a encore les propriétés suivantes :

1. Les vecteurs $\{\mathbf{u}_s, \;\; s = 1, \ldots, r\}$ forment une base orthonormée de $Im(X)$ dans $(\mathbf{R}^n, D)$, et les vecteurs $\{\mathbf{v}_s, \;\; s = 1, \ldots, r\}$ une base orthonormée de $Im(X')$ dans $(\mathbf{R}^p, M)$.

2. Equation aux valeurs propres :

$$X' DXM = V\Lambda^2 V' M = \sum_{s=1}^{r} (\lambda_s)^2 \mathbf{v}_s \mathbf{v}_s' M \tag{2.4}$$

et

$$XMX' D = U\Lambda^2 U' D = \sum_{s=1}^{r} (\lambda_s)^2 \mathbf{u}_s \mathbf{u}_s' D \tag{2.5}$$

On reconnaît les décompositions spectrales des opérateurs $X' DXM$ et de l'"opérateur d'Escoufier" $XMX' D$. Ici les valeurs propres sont notées $(\lambda_s)^2$, et λ_s désigne une valeur singulière.

Remarque La DVS de X nécessite la donnée de X, M et D. Pour une matrice X donnée, il existe donc une infinité de telles décompositions associées à chaque couple (M,D).

Démonstration On démontrera préalablement, à titre d'exercice, que $Im(X') = Im(X' DX)$ (On peut utiliser pour cette démonstration une relation entre le rang et la dimension du noyau d'une application linéaire).
On part du résultat : s'il est vrai, on a nécessairement :

$$X'DX = V\Lambda U'DU\Lambda V'$$

soit :

$$X'DXM = V\Lambda^2 V'M \text{ (Décomposition spectrale de } X'DXM)$$

Donc (premier résultat) : Les vecteurs singuliers à droite de X sont nécessairement les vecteurs propres M-orthonormés de la matrice M-symétrique $X'DXM$ (les vecteurs colonnes de V étant orthonormés, on a bien $V'MV = I_r$).
De plus, en multipliant $X = U\Lambda V'$ à droite par MV, on a nécessairement :

$$XMV = U\Lambda \text{ soit } U = XMV\Lambda^{-1}.$$

Il reste à démontrer que le triplet (U, Λ, V) satisfait à l'équation (2.3).
On a bien $U'DU = \Lambda^{-1}V'MX'DXMV\Lambda^{-1} = I_r$. D'autre part :

$$U\Lambda V' = XMVV'.$$

Or, une conséquence du fait que les colonnes de V soient les vecteurs propres M-orthonormés de $X'DXM$ est que $VV'M$ est l'orthoprojecteur sur $Im(V) = Im(X'DXM)$. M étant régulier, on a encore $Im(X'DXM) = Im(X'DX)$ et d'après le résultat cité au début de la démonstration, $Im(X'DX) = Im(X')$. Donc $VV'MX' = X'$ ($VV'M$ est l'identité sur $Im(X')$), d'où le résultat.

Propriété 20 (Formule de reconstitution) *La relation 2.2 s'écrit élément par élément :*

$$x_i^j = \sum_{s=1}^{r} \lambda_s u_{si} v_{sj} \quad \forall (i,j) \in I \times J$$

Dans cette relation, les matrices $\mathbf{u}_s\mathbf{v}'_s(s = 1, \ldots, r)$ constituent une famille d'éléments orthonormés de l'espace $\mathcal{M}_{n,p}$. On en déduit que :

$$\|X\|^2_{tr} = \sum_{s=1}^{r} (\lambda_s)^2$$

Propriété 21 (Formules de transitions) *De l'équation 2.2, on déduit que :*

$$\mathbf{u}_s = \frac{1}{\lambda_s} XM\mathbf{v}_s \quad \forall s = 1, \ldots, r \tag{2.6}$$

$$\mathbf{v}_s = \frac{1}{\lambda_s} X'\, D\mathbf{u}_s \quad \forall s = 1, \ldots, r \tag{2.7}$$

On en déduit l'existence d'une base orthonormée $\{\mathbf{v}_s, s = 1, \ldots, p\}$ de $(\mathbf{R}^p, M)$ et $\{\mathbf{u}_s, s = 1, \ldots, n\}$ de $(\mathbf{R}^n, D)$ telle que:

$$XM\mathbf{v}_s = \lambda_s\mathbf{u}_s \quad s = 1, \ldots, r$$

$$X'\, D\mathbf{u}_s = \lambda_s\mathbf{v}_s \quad s = 1, \ldots, r$$

$$XM\mathbf{v}_s = 0 \quad s > r$$

$$X'\, D\mathbf{u}_s = 0 \quad s > r$$

2.3.2 Approximation de faible rang d'une matrice $n \times p$

Lemme 1 *Soit $\{(\lambda_s)^2,\ s = 1, \ldots, r\}$ r réels positifs rangés par ordre décroissant, k un entier inférieur ou égal à r, et soit $\mathcal{X} = \{x_s,\ s = 1, \ldots, r,\ \ 0 \le x_s \le 1,\ \ \sum_{s=1}^{r} x_s \le k\}$ une partie de $\mathbf{R}^r$. Alors :*

1. $\max_{x\in\mathcal{X}} \sum_{s=1}^{r} (\lambda_s)^2 x_s = \sum_{s=1}^{k} \lambda_s^2$ *et*
2. $\arg\max_{x\in\mathcal{X}} \sum_{s=1}^{r} (\lambda_s)^2 x_s = \mathbf{x}_0 = (\overbrace{1, \ldots, 1, \ldots, 1}^{k}, \overbrace{0, \ldots, 0}^{r-k})$
3. *De plus, $\mathbf{x}_0$ est solution unique* $\Longleftrightarrow \lambda_k^2 > \lambda_{k+1}^2$.

Démonstration: On pose :

$$A = \sum_{s=1}^{r} (\lambda_s)^2 x_s = \sum_{s=1}^{k} (\lambda_s)^2 x_s + \sum_{s=k+1}^{r} (\lambda_s)^2 x_s.$$

Pour démontrer que A admet pour maximum $\sum_{s=1}^{k} (\lambda_s)^2$, on montre que $\sum_{s=1}^{k} (\lambda_s)^2 - A$ est strictement positif si $x \neq x_0$ (sauf dans le cas où $(\lambda_k)^2 = (\lambda_{k+1})^2$, où seule la positivité est conservée) et est nul pour $\mathbf{x} = \mathbf{x}_0$. En effet :

$$\begin{aligned}
\sum_{s=1}^{k} (\lambda_s)^2 - A &= \sum_{s=1}^{k} (\lambda_s)^2 (1 - x_s) - \sum_{s=k+1}^{r} (\lambda_s)^2 x_s \\
&\ge \lambda_k^2 (k - \sum_{s=1}^{k} x_s) - \lambda_{k+1}^2 \left(\sum_{s=k+1}^{r} x_s \right) \quad \text{(à cause de la décroissance des } (\lambda_s)^2\text{)} \\
&\ge \sum_{s=k+1}^{r} x_s \left(\lambda_k^2 - \lambda_{k+1}^2\right) \text{ car } \sum_{s=1}^{r} x_s \le k.
\end{aligned}$$

Cette quantité est donc strictement positive si au moins un x_s de rang supérieur à k est strictement positif et si $\lambda_k^2 > \lambda_{k+1}^2$. Donc sous cette dernière hypothèse, le maximum est atteint pour x_0 uniquement et vaut $\sum_{s=1}^{k} (\lambda_s)^2$.

Définition et notations

Soit X, D et M définis comme ci-dessus et $r = rang(X)$. Pour k donné ($k \leq r$) on cherche une matrice $\widehat{X}^k$ de rang k qui soit la plus proche de X. La DVS de X au sens des métriques M et D étant l'expression (2.3), on note U_k et V_k les matrices ayant comme k-ième colonne les vecteurs $\mathbf{u}_k$ et $\mathbf{v}_k$ respectivement, et on pose $\Lambda_k = diag(\lambda_s,\ s = 1, \ldots, k)$.

Théorème 2 (Eckart-Young 1936) *Parmi l'ensemble des matrices de $\mathcal{M}_{n,p}$, la matrice $\widehat{X}^k = U_k \Lambda_k V_k'$est la matrice de rang k la plus proche de X au sens de la distance $\| \ \|$. Le minimum atteint est :*

$$\|X - \widehat{X}^k\|^2_{tr} = \sum_{s=k+1}^{r} (\lambda_s)^2$$

$\widehat{X}^k$ est appelé approximation de rang k de X.

Démonstration On note par Y une matrice $n \times p$ de rang inférieur ou égal à k et par P_Y l'orthoprojecteur sur $Im(Y)$ dans $(\mathbf{R}^n, D)$. On montre alors que :

1. $\|X - Y\| \geq \|X - P_Y(X)\|$ et $rang(P_Y(X)) \leq rang(Y)$. En conséquence, dans ce problème de minimisation, $P_Y(X)$ "fait au moins aussi bien" que Y. Puisque P_Y ne dépend que de $Im(Y)$ (noté W) et pas directement de Y, il suffit de chercher un sous-espace W de dimension inférieure ou égale à k qui minimise $\|X - P_W(X)\|$, puis de prendre $Y = P_W(X)$.
2. $\|X\|^2 = \|P_W(X)\|^2 + \|X - P_W(X)\|^2$, donc pourra chercher W maximisant $\|P_W(X)\|^2$.
3. En remplaçant X par sa décomposition spectrale $X = \sum_{s=1}^{r} \lambda_s \mathbf{u}_s \mathbf{v}_s'$, on cherche donc W de dimension inférieure ou égale à k qui maximise :
   $\|P_W(X)\|^2 = \| \sum_{s=1}^{r} \lambda_s P_W(\mathbf{u}_s)\, \mathbf{v}_s' \|^2$.
   La quantité à maximiser s'écrit $\|P_W(X)\|^2 = \sum_{s=1}^{r} (\lambda_s)^2 \|P_W(\mathbf{u}_s)\|^2_D$
4. Puis on utilise le lemme 1 en posant $x_s = \|P_W(\mathbf{u}_s)\|^2_D$ et en montrant que x_s est compris entre 0 et 1 et que $\sum_{s=1}^{r} x_s \leq trace(P_W) \leq k$.
5. On déduit que le maximum est atteint si on peut prendre $x_1 = x_2 = \ldots = x_k = 1$ et $x_{k+1} = .. = x_r = 0$. Or on a

$$\begin{aligned} x_s = 1 \quad &\Longleftrightarrow \mathbf{u}_s \in Im(Y) \text{ et} \\ x_s = 0 \quad &\Longleftrightarrow \mathbf{u}_s \perp Im(Y) \end{aligned}$$

   donc on déduit que $Y = U_k$ est solution, donc la matrice Y de rang inférieur ou égal à k la plus proche de X est $P_{U_k} X = \sum_{s=1}^{k} \lambda_s \mathbf{u}_s \mathbf{v}_s'$ CQFD.

On note $P_{<k} = U_k U_k'\, D$ et $Q_{<k} = V_k V_k'\, M$ les orthoprojecteurs de $(\mathbf{R}^n, D)$ et $(\mathbf{R}^p, M)$ qui projettent respectivement sur $Im(U_k)$ et $Im(V_k)$.

Propriété 22 *On a les identités suivantes :*

1. $\widehat{X}^k = XQ'_{<k} = P_{<k}X$
2. *Pour tout sous-espace T de E et W de F de dimension inférieure ou égale à k, si Q_T et P_W désignent les orthoprojecteurs sur ces sous-espaces, on a :*

$$\|XQ_T'\|^2_{tr} = trace(Q_T X'\, DXM) \leq \|\widehat{X}^k\|^2_{tr} \tag{2.8}$$

$$\|P_W X\|^2_{tr} = trace(P_W XMX'\, D) \leq \|\widehat{X}^k\|^2_{tr} \tag{2.9}$$

Ainsi le maximum des expressions 2.8 ou 2.9 sur l'ensemble des sous espaces T de E et W de F de dimension inférieure ou égale à k est obtenu respectivement pour $T = Im(V_k)$ et $W = Im(U_k)$. On a alors : $P_W = P_{<k} = U_k U_k' D$ et $Q_T = Q_{<k} = V_k V_k' M$.

Remarque La matrice $\widehat{X}^k = XQ'_{<k}$ est obtenue à partir de X en remplaçant ses lignes par leurs projections sur $Im(V_k)$ dans $(\mathbf{R}^p, M)$. Cette même matrice, égale à $P_{<k}X$, s'obtient encore à partir de X en remplaçant ses colonnes par leurs projections sur $Im(U_k)$ dans $(\mathbf{R}^n, D)$. On verra qu'en *Analyse en composantes principales*, ces sous-espaces prennent le nom de *sous-espace principal des lignes* ($Im(V_k)$) ou de *sous-espace principal des colonnes* ($Im(U_k)$).

2.4 Exercices sur le chapitre 2

Exercice 2.4.1 *(Examen de Septembre 1996 : extrait)*On considère la décomposition aux valeurs singulières (DVS) d'une matrice X pour deux métriques M et D :

$$X = \sum_{s=1}^{r} \lambda_s \mathbf{u}^s \, {}^t\mathbf{v}_s$$

où les λ_s sont strictement positifs et rangés par ordre décroissant, et où les familles de vecteurs $\{\mathbf{u}^s, s = 1, \ldots, r\}$ et $\{\mathbf{v}^s, s = 1, \ldots, r\}$ sont orthornormées au sens de D et M respectivement. En déduire la décomposition aux valeurs singulières de ${}^tX\,D\,X$ puis celle de DXM. Précisez pour ces deux décompositions les métriques concernées. En déduire les vecteurs et les valeurs propres de ${}^tX\,D\,X\,M$

Exercice 2.4.2 On donne la matrice :

$$A = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 2 \end{pmatrix}.$$

1. Calculer son polynôme caractéristique et déterminer ses valeurs propres.
2. Montrer que les vecteurs :

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 1 \\ -\sqrt{2} \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 1 \\ 1 \\ \sqrt{2} \end{pmatrix}, \mathbf{u}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

   sont vecteurs propres de A.
3. Ecrire la décomposition spectrale de A ; calculer A^2, $A^{1/2}$; peut-on calculer A^{-1} ?
4. A est-elle la matrice d'un produit scalaire?
5. En remarquant que A est une matrice symétrique, réécrire la décomposition spectrale de A suivant les deux expressions de la propriété 24.4 (p.23).

Exercice 2.4.3 On donne la matrice X suivante:

$$X = \begin{pmatrix} 1 & 1 & 3 \\ -1 & -1 & 3 \\ 1 & 1 & -3 \\ -1 & -1 & -3 \\ 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}$$

et les métriques définies par les matrices $M = I_3$ dans $\mathbf{R}^3$ et $D = \frac{1}{6}I_6$ dans $\mathbf{R}^6$. Effectuer la décomposition aux valeurs singulières de X (calculer ses vecteurs singuliers à gauche et à droite et ses valeurs singulières). Donner les approximations de rang un, deux et trois de X.

Exercice 2.4.4 Soit $\mathcal{M}_{n,p}$ l'espace vectoriel des matrices $n \times p$, D et M des matrices définissant une métrique euclidienne dans les espaces $\mathbf{R}^n$ et $\mathbf{R}^p$ respectivement. On considére le produit scalaire de la trace défini par :

$$X \in \mathcal{M}_{n,p}, Y \in \mathcal{M}_{n,p} \longrightarrow < X \; , \; Y >_{tr} = trace(X'DYM)$$

Soit $\{\mathbf{e}_i\}_{i=1,\ldots,n}$ la base canonique de $\mathbf{R}^n$, $\{\mathbf{f}_j\}_{j=1,\ldots,p}$ la base canonique de $\mathbf{R}^p$, $\mathbf{x}$ et $\mathbf{y}$ deux vecteurs de $\mathbf{R}^p$. On note

$$\mathbf{x} \otimes \mathbf{y} = \mathbf{x}\mathbf{y}' \in \mathcal{M}_{n,p}.$$

- Quelle est la matrice $\mathbf{e}_i \otimes \mathbf{f}_j$?
- De quel rang est la matrice $\mathbf{x} \otimes \mathbf{y}$? Une matrice X qui peut s'écrire sous la forme $\mathbf{x} \otimes \mathbf{y}$ est appelée *tenseur élémentaire*.
- Montrer que toute matrice X est combinaison linéaire de tenseurs élémentaires.

- Caractériser le nombre minimum de tenseurs élémentaires intervenant dans la combinaison linéaire décomposant une matrice X (on justifiera le résultat).
- On définit l'application bilinéaire sur les couples de tenseurs élémentaires par:

$$\left.\begin{array}{l}(\mathbf{x}_1,\ \mathbf{y}_1) \in \mathbf{R}^n \times \mathbf{R}^p \\ (\mathbf{x}_2,\ \mathbf{y}_2) \in \mathbf{R}^n \times \mathbf{R}^p\end{array}\right) \longrightarrow < \mathbf{x}_1 \otimes \mathbf{y}_1\ ,\ \mathbf{x}_2 \otimes \mathbf{y}_2 > = < \mathbf{x}_1\ ,\ \mathbf{x}_2 >_D < \mathbf{y}_1\ ,\ \mathbf{y}_2 >_M$$

  (on pourra vérifier la cohérence de cette définition). Montrer que cette définition, étendue par linéarité à l'ensemble des matrices $n \times p$, redonne le produit scalaire de la trace.

2.5 Travaux dirigés : Quelques exercices sur les espaces euclidiens

Exercice 2.5.1 *Soit X une matrice réelle $n \times p$, telle que $Ker(X) = \{0_{R^p}\}$. On munit l'espace $\mathbf{R}^n$ d'une structure euclidienne, définie par la matrice M dans la base canonique de $\mathbf{R}^n$. Soient $\mathbf{y}$ et $\mathbf{e}$ deux vecteurs de $\mathbf{R}^n$ et α un vecteur de $\mathbf{R}^p$ tels que :*

$$\mathbf{y} = X\alpha + e$$

1. *Donner une condition nécessaire et suffisante pour que $X\alpha$ soit la projection orthogonale de $\mathbf{y}$ sur $Im(X)$ (on notera P_X cet opérateur de projection orthogonale).*
2. *Montrer que $X'MX$ est inversible. En déduire une expression de P_X.*

Exercice 2.5.2 *Soit $\mathcal{M}_p$ l'ensemble des matrices carrées d'ordre p. On considère l'application trace définit sur l'ensemble de toutes les matrices carrées, réunion de tous les ensembles $\mathcal{M}_p$ (p=1,...,k,...), par :*

$$A = (a_{ij})_{i,j=1,\dots,p} \in \mathcal{M}_p \longrightarrow trace(A) = \sum_{i=1}^{p} a_{ii}$$

On montrera les propositions qui suivent:

1. *C'est une forme linéaire sur $\mathcal{M}_p$.*
2. *Si A et B sont des matrices quelconques (non forcément carrées) telles que le produit AB existe et est carré, alors le produit BA existe et on a : $trace(AB) = trace(BA)$.*
3. *Deux matrices équivalentes de $\mathcal{M}_p$ (ou matrices d'un même opérateur linéaire dans deux bases différentes de $\mathbf{R}^p$) ont même trace. En déduire la possibilité de définir la trace d'un opérateur par la trace d'une de ses matrices associées.*
4. *Soit A est un opérateur d'un espace euclidien de dimension p, et $\mathcal{E} = \{\mathbf{e}_j, j = 1, \dots, p\}$ une base orthonormée de cet espace. Montrer que $trace(A)$ s'écrit :*

$$trace(A) = \sum_{j=1}^{p} < A(\mathbf{e}_j), \mathbf{e}_j >$$

5. *On suppose maintenant que A est un ortoprojecteur. Montrer que l'on a :*

$$trace(A) = \sum_{j=1}^{p} ||A(\mathbf{e}_j)||^2$$

Exercice 2.5.3 *Soit f une application de $\mathbf{R}^p$ dans $\mathbf{R}^n$, de matrice associée X dans les bases canoniques de $\mathbf{R}^p$ et de $\mathbf{R}^n$. On munit $\mathbf{R}^p$ et $\mathbf{R}^n$ des structures euclidiennes M^{-1} et D.*

1. *En notant $\mathcal{M}_{p,n}$ l'espace vectoriel des matrices $n \times p$, et X^* la matrice adjointe de la matrice X, montrer que : l'application :*

$$(X, Y) \in \mathcal{M}_{p,n} \times \mathcal{M}_{p,n} \longrightarrow trace(XY^*)$$

   est un produit scalaire sur $\mathcal{M}_{p,}$.

2. *Montrer que si P est la matrice d'un orthoprojecteur de $\mathbf{R}^n$ et Q celle d'un orthoprojecteur de $\mathbf{R}^p$, les applications :*
   (a) $X \in \mathcal{M}_{p,n} \longrightarrow PX \in \mathcal{M}_{p,n}$,
   (b) $X \in \mathcal{M}_{p,n} \longrightarrow XQ' \in \mathcal{M}_{p,n}$,
   (c) $X \in \mathcal{M}_{p,n} \longrightarrow PXQ' \in \mathcal{M}_{p,n}$,

   sont des orthoprojecteurs dans $\mathcal{M}_{p,n}$ sur des sous-espaces que l'on précisera.

   En raisonnant en termes de matrices, de quoi sont constituées les colonnes de PX ? les lignes de XQ' ?

Chapitre 3

L'analyse en composantes principales

3.1 Introduction

Dans les trente dernières années, la révolution apportée par l'informatique a (entre autres) rendu nécessaire le développement d'outils adaptés aux traitements de grandes masses d'informations. Il a ainsi fallu développer des outils permettant une appréhension rapide de l'information contenue dans de "grands tableaux de données". C'est l'objet de l'*Analyse des Données Multidimensionnelles.* Ces outils se sont développés d'abord à partir de méthodes préexistantes à cette période, (analyse en composante principales, analyse des correspondances), mais dont l'utilisation restait confidentielle vu le manque de moyens de calcul, puis en utilisant pleinement la structure et la puissance de calcul de l'ordinateur (réseaux de neurones, méthodes de rééchantillonage,...).

Dans ce cours, nous verrons dans un premier temps deux méthodes factorielles : l'analyse en composantes principales et l'analyse des correspondances. Ce sont des méthodes descriptives fournissant des résumés numériques et graphiques permettant de synthétiser l'information la plus importante —en un sens qu'il conviendra de contrôler— contenue dans un tableau de données. *Si nous ne parlons ici que de méthodes multidimensionnelles, nous pensons nécessaires de faire précéder les traitements multivariés de traitements univariés destinés à vérifier la cohérence et la pertinence des données vis à vis des problèmes posés.*

Dans un premier temps, nous définissons le vocabulaire usuel de l'analyse des données multidimensionnelles. On considère les définitions données dans le premier chapitre comme connues. Puis, nous présentons les méthodes d'analyse en composantes principales et d'analyse des correspondances.

3.2 Définitions usuelles en statistique multidimensionnelle

Le tableau de données

C'est un tableau rectangulaire, noté Y, dont les lignes sont associés aux individus, indicés par $\mathbf{I} = \{1, \ldots, i, \ldots, n\}$, i indice inférieur, et dont les colonnes sont associées aux variables et indicés par $\mathbf{J} = \{1, \ldots, j, \ldots, p\}$, j indice supérieur. Ainsi, l'élément (i, j) de Y, noté y_i^j, est égal à la valeur prise par la variable j pour l'individu i. Sa i-ème ligne (une fois transposée) est notée $\mathbf{y}_i$, et sa j-ème colonne $\mathbf{y}^j$ (par convention, *tous les vecteurs sont représentés en colonne*). Le vecteur $\mathbf{y}_i$ est le i-ème vecteur individu (qui contient l'ensemble des mesures ou valeurs des variables pour cet individu), et le vecteur $\mathbf{y}^j$ est la j-ème variable. Enfin, chaque individu est muni d'un poids p_i strictement positif, représentant le poids associé à l'individu i. Ce poids est en particulier utilisé pour calculer des moyennes pondérées. On suppose que la somme des poids est égale à 1.

$$Y = \left(\begin{array}{c|c|c|c|c} y_1^1 & \dots & y_1^j & \dots & y_1^p \\ \dots & \dots & \dots & \dots & \dots \\ y_i^1 & \dots & y_i^j & \dots & y_i^p \\ \dots & \dots & \dots & \dots & \dots \\ y_n^1 & \dots & y_n^j & \dots & y_n^p \end{array} \right) = \left(\, \mathbf{y}^1 \mid \dots \mid \mathbf{y}^j \mid \dots \mid \mathbf{y}^p \, \right) = \left(\begin{array}{c} {}^t\mathbf{y}_1 \\ \dots \\ {}^t\mathbf{y}_i \\ \dots \\ {}^t\mathbf{y}_n \end{array} \right)$$

Espace des individus, nuage des individus

L'*espace des individus* est l'espace $\mathbf{R}^p$ qui contient les transposés des vecteurs lignes $\mathbf{y}_i$, de la matrice Y . L'ensemble des vecteurs $\mathbf{y}_i$ est appelé *nuage des individus.*

Espace des variables, nuage des variables

L'*espace des variables*, défini précédemment, est l'espace $\mathbf{R}^n$. L'ensemble des variables $\mathbf{y}^j$, éléments de $\mathbf{R}^n$, est appelé le *nuage des variables.*

Barycentre du nuage des individus

Le vecteur $\mathbf{g} = \sum_{i=1}^n p_i \mathbf{y}_i$ est appelé *barycentre* du nuage des individus. Sa j-ième coordonnée est égale à $g_j = \sum_{i=1}^n p_i y_i^j = \overline{y}^j$, moyenne pondérée de la variable $\mathbf{y}^j$ (on rappelle que les poids sont positifs et de somme 1). On appelle encore $\mathbf{g}$ l'*individu moyen* (les variables ont pour cet individu leurs valeurs moyennes).

Tableau centré

C'est le tableau des données centrées, ayant en position (i,j) $x_i^j = (y_i^j - \overline{y}^j)$, écart à la moyenne de la variable j pour l'individu i. Le tableau centré a en colonnes les variables centrées ($\mathbf{x}^j = \mathbf{y}^j - \overline{y}^j \mathbf{1}_n$). Ses lignes (transposées) sont les vecteurs individus centrés $\mathbf{x}_i = \mathbf{y}_i - \mathbf{g}$.

$$Y = \left(\begin{array}{ccccc} y_1^1 & \dots & y_1^j & \dots & y_1^p \\ \dots & \dots & \dots & \dots & \dots \\ y_i^1 & \dots & y_i^j & \dots & y_i^p \\ \dots & \dots & \dots & \dots & \dots \\ y_n^1 & \dots & y_n^j & \dots & y_n^p \end{array} \right) \qquad \text{Tableau non centré}$$

$$t(\mathbf{g}) = \left(\, \overline{y}^1 \quad \dots \quad \overline{y}^j \quad \dots \quad \overline{y}^p \, \right) \qquad \text{Barycentre}$$

$$X = \left(\begin{array}{ccccc} y_1^1 - \overline{y^1} & \dots & y_1^j - \overline{y}^j & \dots & y_1^p - \overline{y^p} \\ \dots & \dots & \dots & \dots & \dots \\ y_i^1 - \overline{y^1} & \dots & y_i^j - \overline{y}^j & \dots & y_i^p - \overline{y^p} \\ \dots & \dots & \dots & \dots & \dots \\ y_n^1 - \overline{y^1} & \dots & y_n^j - \overline{y}^j & \dots & y_n^p - \overline{y^p} \end{array} \right) \qquad \text{Tableau centré}$$

$$= \left(\, \mathbf{y}^1 - \overline{y^1}\mathbf{1}_n \mid \dots \mid \mathbf{y}^j - \overline{y}^j \mathbf{1}_n \mid \dots \mid \mathbf{y}^p - \overline{y^p}\mathbf{1}_n \, \right) = \left(\begin{array}{c} {}^t\mathbf{y}_1 - {}^t\mathbf{g} \\ \dots \\ {}^t\mathbf{y}_i - {}^t\mathbf{g} \\ \dots \\ {}^t\mathbf{y}_n - {}^t\mathbf{g} \end{array} \right)$$

3.2.1 Définition d'une distance entre individus

Un des objectifs de l'analyse des données étant de décrire les proximités entre les individus (existe-t'il des groupes d'individus semblables qui se différencient d'autres groupes d'individus semblables ?) nous avons besoin de mesurer la similarité de deux vecteurs individus. Deux individus seront dit proches s'ils ont à peu près les mêmes valeurs des variables (si le tableau est un tableau de notes, deux élèves proches auront à peu près les mêmes notes). Concrètement, on choisit une matrice $M = (m_{jk})$, carrée d'ordre p, symétrique définie positive, pour définir une métrique euclidienne dans $\mathbf{R}^p$. La distance entre deux individus est alors mesurée par la distance euclidienne de leurs vecteurs lignes associés. Le choix de M dépend de l'idée que ce fait l'utilisateur de la notion de dissemblance ou de ressemblance entre les individus. Les distances les plus souvent utilisées sont :

- La *distance usuelle* définie comme la racine de la somme des carrés des différences des coordonnées :
$$d(\mathbf{y}_i, \mathbf{y}_k) = \left(\sum_{j=1}^{p} (y_i^j - y_k^j)^2 \right)^{1/2}.$$
Sa matrice M associée est l'identité.
- La *distance pondérée* définie comme :
$$d(\mathbf{y}_i, \mathbf{y}_k) = \left(\sum_{j=1}^{p} m_{jj} (y_i^j - y_k^j)^2 \right)^{1/2}.$$
où les coefficients m_{jj}, strictement positifs, pondèrent l'influence de la j-ième variable. Sa matrice associée est $M = diag(m_{jj}, j = 1, \ldots, p)$.

Cette distance euclidienne va nous permettre de définir des notions d'angle et d'orthogonalité dans $\mathbf{R}^p$ ainsi que des projecteurs orthogonaux. Nous notons $\|\,.\,\|_M$ la norme associée.

Dans la suite, nous supposerons que la matrice M est choisie diagonale. Il existe pourtant des cas où une matrice non diagonale est utile. Dans ce cas certaines des propriétés citées dans la suite ne sont plus valables. Elles seront précisées dans la suite.

3.2.2 Le produit scalaire de l'espace des variables

On rappelle qu'on a défini au chapitre I une métrique dans l'espace des variables $\mathbf{R}^n$ appelée *métrique des poids*, à partir du produit scalaire :

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \ldots \\ y_i \\ \ldots \\ y_n \end{pmatrix}, \; \mathbf{z} = \begin{pmatrix} z_1 \\ \ldots \\ z_i \\ \ldots \\ z_n \end{pmatrix} \longrightarrow < \mathbf{y}, \mathbf{z} > = \sum_{i=1}^{n} p_i \; y_i \; z_i$$

Les poids sont strictement positifs (condition nécessaire pour la définition du produit scalaire), et normalisés ($\sum_i p_i = 1$). La matrice D étant définie par $D = diag(p_i, i = 1, \ldots, n)$, l'expression matricielle du produit scalaire est ici:

$$< \mathbf{y}, \mathbf{z} > =^t \mathbf{y} D \mathbf{z}.$$

Dans cet espace les indices *variance*, *covariance*, *coefficient de corrélation* ont des interprétations géométriques données au chapitre 1 : la *longueur* d'une variable centrée est son écart-type, et le *cosinus* de deux variables centrées est leur coefficient de corrélation linéaire. Ces propriétés sont à la base de l'interprétation des représentations de variables en analyse en composantes principales.

3.2.3 L'inertie, une mesure de variance multidimensionnelle

On a vu la notion de variance d'une variable $\mathbf{y}$, qui est une mesure de dispersion unidimensionnelle :
$$var(\mathbf{y}) = \sum_i p_i (y_i - \overline{y})^2$$

L'inertie d'un tableau de données est définie comme la *mesure de dispersion multidimensionnelle*, suivante :
$$In(Y) = \sum_i p_i \|\mathbf{y}_i - \mathbf{g}\|_M^2.$$

On appellera dans la suite $\|\mathbf{y}_i - \mathbf{g}\|_M^2$ (carré de la distance d'un individu à l'individu moyen) l'*originalité de l'individu i*. Si M est diagonale, l'inertie est encore une somme pondérée des variances des variables $\mathbf{y}^j$:
$$In(Y) = \sum_j m_{jj} \; var(\mathbf{y}^j).$$

En notant X la matrice centrée associée à Y, on a encore :

$$In(Y) = \text{trace}({}^t XDXM)$$

C'est donc le carré de la norme de la matrice X au sens des métriques M dans $\mathbf{R}^p$ et D dans $\mathbf{R}^n$.

L'inertie est au coeur des méthodes descriptives en analyse des données : c'est cette mesure de variation qui est à expliquer par les facteurs. Un élément important de l'analyse consiste à bien doser ses constituants (les variables) éventuellement en les pondérant correctement. Il sera toujours possible de contrôler l'*influence* de chaque variable dans cette mesure globale. Cette influence, appelée *contribution* de la variable à l'inertie totale est simplement mesurée par $m_{jj} var(\mathbf{y}^j)/In(Y)$ (en général exprimée en pourcentage). On contrôlera de même l'influence des individus.

3.3 Objectifs de l'analyse en composantes principales

L'objet de l'analyse en composantes principales est de *décrire* (en simplifiant) et de *résumer* les données, sous une forme numérique et graphique.

La possibilité de résumer ou simplifier les données suppose l'existence d'une *redondance partielle* dans les données, soit encore l'existence de *liaisons* (non déterministes) entre les variables.

On dira que deux variables sont liées si la connaissance de la valeur de l'une d'entre elles pour un individu donne des informations sur les valeurs probables de l'autre variable.

Prenons des exemples :

- L'assertion *Mr Martin mesurait 1m90 et pesait 45kg* est étonnante. Ces deux variables étant liées, l'information "mesurait 1m90" rend la seconde information très peu vraisemblable ou exceptionnelle.
- De même, l'assertion *Mr Martin était manœuvre et gagnait moins de 70KF par an* est très vraisemblable. Il existe une forte liaison entre ces variables, ce qui rend l'information partiellement redondante.

Le résumé d'un tableau de données va consister en :

- une description, éventuellement *simplifiée*, des liaisons entre les variables,
- une description, éventuellement *simplifiée*, des proximités entre individus,
- et une interprétation simultanée individus-variables.

3.4 Les trois étapes de l'analyse en composantes principales

L' analyse en composantes principales opère en trois temps :

3.4.1 Préparation des données et construction de la matrice X

- Un premier examen consiste à représenter une matrice de diagrammes de dispersion (cf. Figure 3.1, pour l'exemple présenté au paragraphe 3.9), pour voir s'il y a lieu de transformer les variables pour rendre leurs liaisons linéaires. En effet, leurs liaisons seront décrites dans l'analyse par des coefficients de corrélation linéaire, qui ne sont valides que si cette liaison est de type linéaire.
- Ensuite, les variables du tableau à résumer sont centrées. Puis, selon la nature des données, les variables sont éventuellement réduites. On verra (cf. "mesure d'influence et interprétation des facteurs") que l'influence des variables dans l'analyse se mesure par leurs variances : si les données sont réduites et si $M = I_p$ ($m_{jj} = 1$ pour tout j), toutes les variables ont la même influence dans l'analyse. En conséquence, on réduit les variables si elles sont de nature différentes ou si on estime que leurs influences respectives dans l'analyse ne doit pas dépendre de leurs variabilité. Dans le cas des notes scolaires, la réponse ne va pas de soit : parce qu'une matière est à forte variance, faut-il qu'elle ait une influence plus grande dans l'analyse?

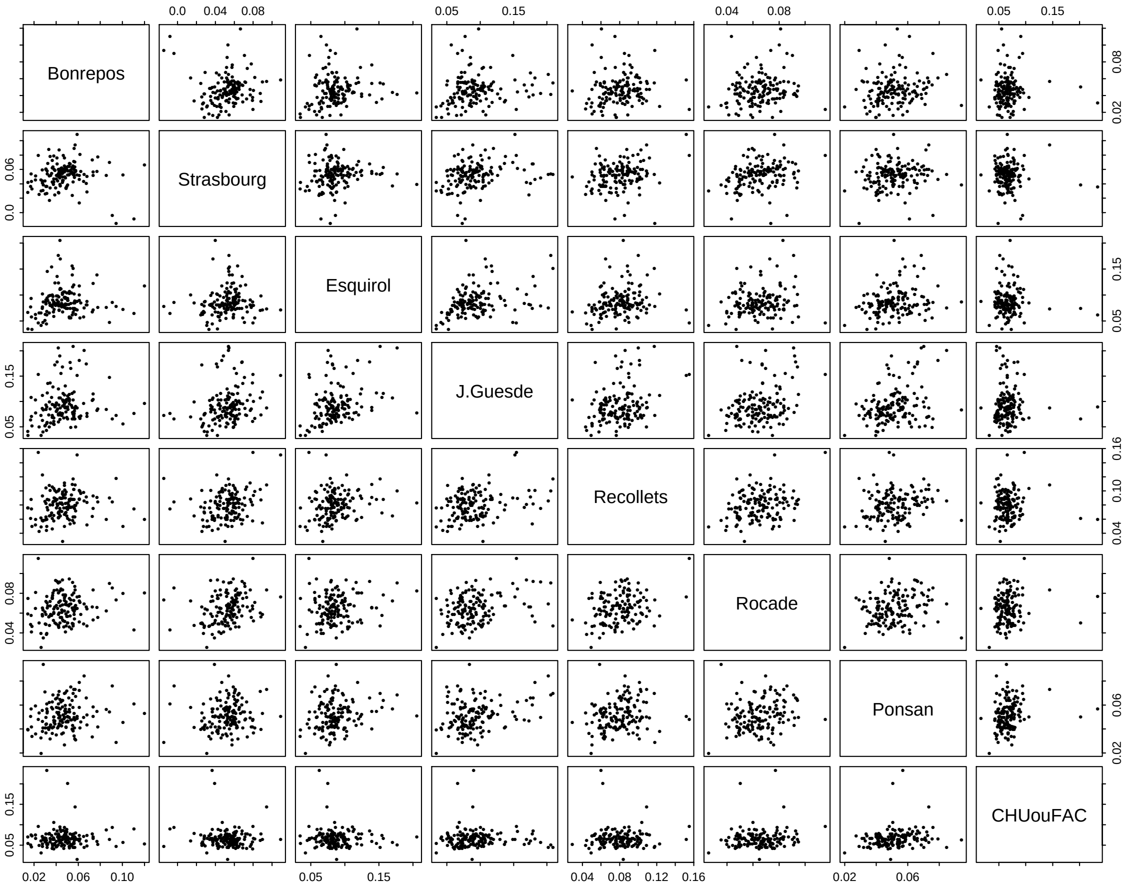


FIG. 3.1 - *Durées inter-arrêt Matabiau-Rangueil : matrice de diagrammes de dispersion*

Dans la suite, le tableau qui résulte de ces transformations est noté X.

Considérons le tableau des notes scolaires donné Tableau 3.1(a), et qui est noté Y, et la matrice centrée associée, notée X (Tableau 3.1(b)). On va effectuer une analyse en composantes principales de ce tableau.

note	m	s	f	l	dm
JE	6.0	6.0	5.0	5.5	8.0
AL	8.0	8.0	8.0	8.0	9.0
AN	6.0	7.0	11.0	9.5	11.0
MO	14.5	14.5	15.5	15.0	8.0
DI	14.0	14.0	12.0	12.5	10.0
AD	11.0	10.0	5.5	7.0	13.0
PI	5.5	7.0	14.0	11.5	10.0
BR	13.0	12.5	8.5	9.5	12.0
EV	9.0	9.5	12.5	12.0	18.0

	m	s	f	l	dm
JE	-3.67	-3.83	-5.22	-4.56	-3.00
AL	-1.67	-1.83	-2.22	-2.06	-2.00
AN	-3.67	-2.83	0.78	-0.56	0.00
MO	4.83	4.67	5.28	4.94	-3.00
DI	4.33	4.17	1.78	2.44	-1.00
AD	1.33	0.17	-4.72	-3.06	2.00
PI	-4.17	-2.83	3.78	1.44	-1.00
BR	3.33	2.67	-1.72	-0.56	1.00
EV	-0.67	-0.33	2.28	1.94	7.00

TAB. 3.1 - *Tableau de notes scolaires (a) et tableau centré (b)*

3.4.2 Approximation de faible rang de la matrice X

Etant donné la matrice X, de rang r à n lignes et p colonnes, on sait, d'après le théorème d'Eckart-Young, construire la meilleure approximation de rang donné de cette matrice. On sait trouver des matrices X_1, X_2,..., X_k de mêmes dimensions que X telle que :

$$X = X_1 + X_2 + \ldots + X_k + \ldots + X_r,$$

et telle que la meilleure approximation de rang k s'écrive :

$$\widehat{X}^k = X_1 + X_2 + \ldots + X_k.$$

En effet, si la décomposition aux valeurs singulières de X s'écrit: $X = \sum_{s=1}^{r} \lambda_s \mathbf{u}_s^t \mathbf{v}_s$, alors la meilleure approximation de rang k ($k \leq r$) de X est $\widehat{X}^k = \sum_{s=1}^{k} \lambda_s \mathbf{u}_s^t \mathbf{v}_s$, soit encore, en notant :

$$\begin{array}{rcl} U_k & = & (\mathbf{u}_1 | \ldots | \mathbf{u}_s | \ldots | \mathbf{u}_k) \\ V_k & = & (\mathbf{v}_1 | \ldots | \mathbf{v}_s | \ldots | \mathbf{v}_k) \\ \Lambda_k & = & diag(\lambda_1, \ldots, \lambda_k) \\ P_{<k} & & \text{orthoprojecteur sur } Im(U_k) \text{ dans } (\mathbf{R}^n, D) \\ Q_{<k} & & \text{orthoprojecteur sur } Im(V_k) \text{ dans } (\mathbf{R}^p, M), \end{array}$$

cette approximation s'écrit :

$$\widehat{X}^k = U_k \Lambda_k^t V_k = X^t Q_{<k} = P_{<k} X. \tag{3.1}$$

On construira d'abord la meilleure approximation de rang 2, puis, si l'approximation n'est pas assez bonne, on construira des approximations de rangs supérieurs.

Définitions

- u_s est le s-ième *facteur réduit* ou *composante principale réduite*,
- v_s est le s-ième *vecteur principal*, l'axe qu'il engendre est le s-ième *axe principal*
- $W_k = [u_1, \ldots, u_k]$ est le k-ième *sous-espace principal de l'espace des variables*,
- $Z_k = [v_1, \ldots, v_k]$ est le k-ième *sous-espace principal de l'espace des individus* (pour $k = 1$ on parle d'*axe principal*, pour $k = 2$ de *plan principal*).

Propriété

L'équation (3.1) montre que :

1. l'approximation d'ordre k du i-ème vecteur individu $\widehat{\mathbf{x}}_i^k$, i-ème ligne de $\widehat{X}^k$, est obtenue par projection orthogonale du vecteur individu $\mathbf{x}_i$ sur le sous-espace principal Z_k de $\mathbf{R}^p$.
$$\widehat{\mathbf{x}}_i^k = Q_k(\mathbf{x}_i),$$
2. l'approximation d'ordre k de la j-ième variable centrée, notée $\widehat{\mathbf{x}^j}^k$, est obtenue par projection orthogonale de la variable centrée $\mathbf{x}^j$ sur le sous-espace principal W_k de $\mathbf{R}^n$.
$$\widehat{\mathbf{x}^j}^k = P_k(\mathbf{x}^j)$$

Définitions

Le nuage $\{Q_k(\mathbf{x}_i), i = 1, \ldots, n\}$ obtenu par projection du nuage initial $\{\mathbf{x}_i, i = 1, \ldots, n\}$ est appelé *nuage projeté* des individus.

On pourra vérifier à titre d'exercice que le projeté d'un nuage centré est un nuage centré.

De même, le nuage $\{P_k(\mathbf{x}^j), j = 1, \ldots, p\}$ obtenu par projection du nuage des variables $\{\mathbf{x}^j, j = 1, \ldots, p\}$ est appelé *nuage projeté* des variables. Ce n'est pas en général un nuage centré.

Propriété

En notant u_{is} et v_{js} respectivement la i-ème coordonnées de $\mathbf{u}_s$ et la j-ème coordonnées de $\mathbf{v}_s$ et $\widehat{\mathbf{x}_i^j}^k$ l'élément (i, j) de $\widehat{X}^k$ (appelé *approximation d'ordre* k de x_i^j), on a des écritures équivalentes à l'équation 3.1, qui donnent la décomposition des lignes de $\widehat{X}^k$, de ses colonnes, ou de ses éléments :

$$\widehat{\mathbf{x}_i}^k = \sum_{s=1}^k (\lambda_s u_{is})\mathbf{v}_s \quad \text{décomposition des lignes} \tag{3.2}$$

$$\widehat{\mathbf{x}^j}^k = \sum_{s=1}^k (\lambda_s v_{js})\mathbf{u}_s \quad \text{décomposition des colonnes} \tag{3.3}$$

$$\widehat{\mathbf{x}_i^j}^k = \sum_{s=1}^k \lambda_s u_{is} v_{js} \quad \text{décomposition des éléments.} \tag{3.4}$$

La première équation donne une décomposition de $\widehat{\mathbf{x}_i}^k$ dans la base orthonormée des vecteurs principaux $\{\mathbf{v}_s, s = 1, \ldots, r\}$.

Définition

Dans cette décomposition de $\widehat{\mathbf{x}_i}^k$ (ou de $\mathbf{x}_i$ pour $k = r$), la coordonnée $\lambda_s u_{is}$ de $\widehat{\mathbf{x}_i}^k$ est appelée *s-ième composante principale* de l'individu i. Le vecteur $\lambda_s \mathbf{u}_s$ de $\mathbf{R}^n$ est une variable centrée de variance $(\lambda_s)^2$. Ce vecteur, contenant les coordonnées des individus sur le s-ième vecteur de base $\mathbf{v}_s$, est le s-ième *facteur non réduit* ou la s-ième *composante principale.*

La seconde équation donne une décomposition de $\widehat{\mathbf{x}^j}^k$ dans la base orthonormée des facteurs réduits $\{\mathbf{u}_s, s = 1, \ldots, r\}$.

Enfin, la troisiéme décomposition est connue sous le nom de formule de reconstitution des données

Définition

Dans cette décomposition de $\widehat{\mathbf{x}^j}^k$ (ou de $\mathbf{x}^j$ pour $k = r$), la coordonnée $\lambda_s v_{js}$ est appelée *saturation*[1].

Ces deux décompositions seront à la base de la représentation graphique de ces approximations. Enfin, la dernière équation est connue sous le nom de *formule de reconstitution des données.* Elle donne, pour $k = r$, une formule permettant de reconstituer x_i^j.

Propriétés

1. On note Q_Z l'orthoprojecteur sur le sous-espace Z de $\mathbf{R}^p$ et P_W l'orthoprojecteur sur le sous-espace W de $\mathbf{R}^n$. Les sous-espaces principaux Z_k (dans $\mathbf{R}^p$) et W_k (dans $\mathbf{R}^n$) possèdent en particulier les propriétés suivantes : parmi les sous-espaces Z de dimension inférieures ou égale à k,

   (a) Z_k est le sous-espace qui maximise l'inertie $\|X^t Q_Z\|^2 = \sum_{i=1}^n p_i \|Q_Z(\mathbf{x}_i)\|_M^2$ du nuage projeté. Le maximum de cette inertie est égal à la somme des carrés des k premières valeurs singulières.

   (b) Z_k est le sous-espace à distance minimum du nuage centré des individus :

   $$Z_k = arg \min_{Z \subset \mathbf{R}^p, dim(Z) \leq k} \sum_{i=1}^n p_i \|\mathbf{x}_i - Q_Z(\mathbf{x}_i)\|^2.$$

   Le minimum est égal à la somme des carrés des $r - k$ dernières valeurs singulières.

2. Décomposition associée de l'inertie :

$$\begin{aligned} \|X\|^2 &= \sum_{s=1}^r \lambda_s^2 \\ &= \|\widehat{X}^k\|^2 + \|X - \widehat{X}^k\|^2 \\ &= \|\widehat{X}^k\|^2 + \sum_{s=k+1}^r \lambda_s^2 \end{aligned} \tag{3.5}$$

1. On déduit de l'équation 3.3 que $var(\mathbf{x}^j) = \sum_{s=1}^r (\lambda_s v_{js})^2$ et donc que $(\lambda_s v_{js})^2$ représente l'importance prise par le facteur s dans la variable $\mathbf{x}^j$, d'où le nom de *saturation.* du facteur s dans la variable j.

Cette décomposition se visualise pour les différentes valeurs de k à l'aide de l'*éboulis des valeurs propres* (ou carrés des valeurs singulières λ_s^2) donné figure 3.2 : on y donne l'éboulis de l'exemple pédagogique, celui de l'exemple des bus présenté à la fin du chapitre, ainsi que, pour ce dernier exemple, les boites à moustaches des composantes principales $\lambda_s \mathbf{u}_s$. Cette dernière représentation est une autre façon de visualiser l'importance des facteurs (on rappelle que $var(\lambda_s \mathbf{u}_s) = (\lambda_s)^2$).

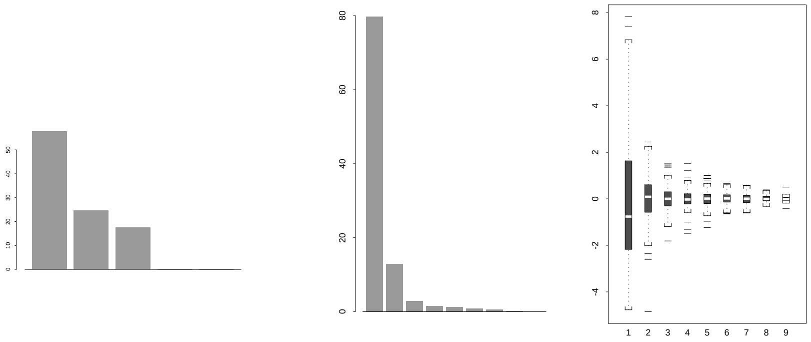


FIG. 3.2 - *Éboulis des valeurs propres et boites à moustaches des facteurs (deux exemples*

$X =$

	m	s	f	l	dm
JE	-3.51	-3.78	-5.39	-4.77	-0.46
AL	-1.62	-1.73	-2.40	-2.14	-0.20
AN	-3.61	-2.91	0.67	-0.44	0.29
MO	4.90	4.87	4.94	4.80	0.32
DI	4.51	4.04	1.72	2.39	-0.05
AD	1.11	0.22	-4.49	-2.94	-0.62
PI	-4.07	-2.85	3.53	1.50	0.67
BR	3.33	2.52	-1.67	-0.36	-0.40
EV	-1.02	-0.37	3.08	1.96	0.44

$+$

	m	s	f	l	dm
JE	-0.16	-0.05	0.17	0.21	-2.54
AL	-0.05	-0.10	0.18	0.08	-1.80
AN	-0.06	0.08	0.11	-0.12	-0.29
MO	-0.07	-0.20	0.34	0.14	-3.32
DI	-0.18	0.13	0.06	0.05	-0.95
AD	0.22	-0.05	-0.23	-0.12	2.62
PI	-0.10	0.02	0.25	-0.06	-1.67
BR	0.00	0.15	-0.05	-0.20	1.40
EV	0.35	0.04	-0.80	-0.02	6.56

TAB. 3.2 - *Approximation de rang 2 de X et tableau des résidus $X - \widehat{X}^2$*

Le tableau 3.2 donne l' approximation $\widehat{X}^2$, de la matrice X, ainsi que la matrice des résidus $X - \widehat{X}^2$. Des indices mesurant la qualité des approximations sont présentés au prochain paragraphe.

3.4.3 Représentation des approximations

L'équation (3.2) donne la décomposition de l'approximation du vecteur individu dans une base orthonormée. Ce sont ces coordonnées qui sont utilisées pour représenter les vecteurs individus approchés. Pour $k = 2$, on effectue cette représentation sur le plan (cf. Figure 3.3 à gauche) : les coordonnées des individus sont égales à $(\lambda_1 u_{i1}, \lambda_2 u_{i2})$. On montre que cette représentation respecte au mieux les distances inter-individus, telles qu'elles peuvent être calculées sur les tableaux X ou Y avec la distance M. Pour $k = 3$, on effectue une seconde représentation sur les axes 1 et 3, avec comme coordonnées $(\lambda_1 u_{i1}, \lambda_3 u_{i3})$.
De même, d'après l'équation (3.3), les approximations des variables centrées ont pour coordonnées $(\lambda_s v_{js}, s = 1, \ldots, k)$ dans une base orthonormée. Cette représentation respecte au mieux les angles et les longueurs des variables centrées, telles qu'elles peuvent être calculées

sur le tableau X avec le produit scalaire des poids. Pour $k = 2$, sur la figure 3.3 à droite, les vecteurs variables ont pour coordonnées $(\lambda_1 v_{j1}, \lambda_2 v_{j2})$.

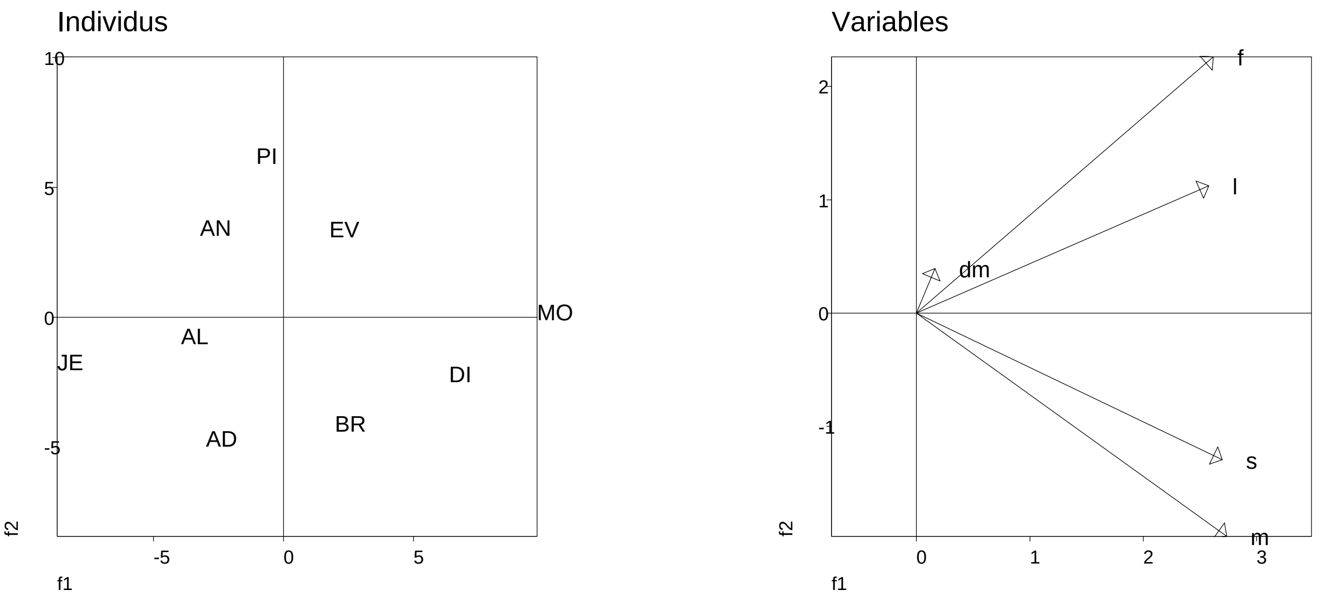


FIG. 3.3 - *Représentation des individus et des variables: plan 1-2*

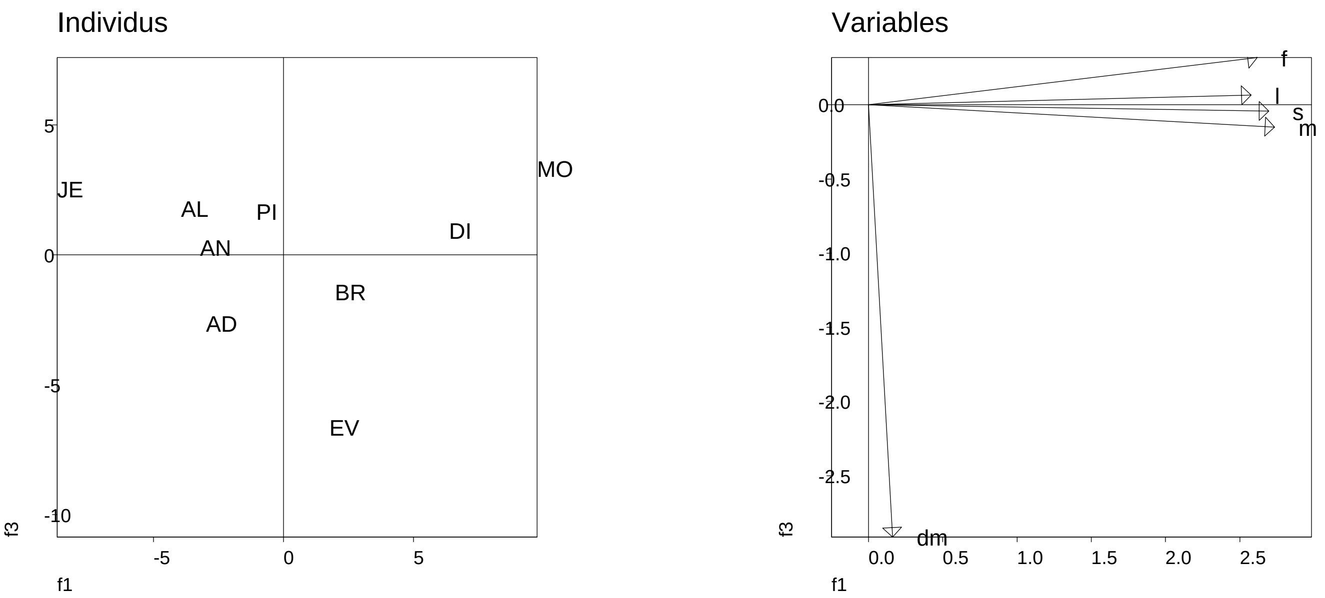


FIG. 3.4 - *Représentation des individus et des variables: plan 1-3*

La représentation des approximations des vecteurs individus est la représentation du nuage projeté des individus sur le sous-espace principal Z_k. De même, celle des approximations des vecteurs variables est la représentation du nuage projeté des variables sur W_k. Ces deux représentations seront appelées représentations isométriques (par opposition à d'autres qui ne le seront pas). Elles permettent donc de faire une analyse de proximités des vecteurs individus, et une analyse de liaison des variables.

3.5 Autres indices d'aide à l'interprétation

La propriété de "double d'orthogonalité de l'ACP a pour conséquence les deux décompositions additives de l'inertie totale décrite dans le paragraphe suivant. De ces décompositions découlent les indices d'aide à l'interprétation exposés dans cette partie.

3.5.1 Double décomposition de l'inertie

Facteurs	1	...	s	...	k	Sous-total	...	r	Total
Variable									
1	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
j	...	...	$m_{jj}(\lambda_s v_{js})^2$	...	...	$m_{jj} var(\widehat{\mathbf{x}^j}^k)$	...	...	$m_{jj} var(\mathbf{x}^j)$
...	...	...	...	...	...	...	...	...	...
J	...	...	...	...	...	...	...	...	...
Total	λ_1^2	...	λ_s^2	...	λ_k^2	$\|\widehat{X}^k\|^2$	...	λ_r^2	$\|X\|^2$

TAB. 3.3 - *La décomposition "variables $\times$ facteurs" de l'inertie*

Facteurs	1	...	s	...	k	Sous-total	...	r	Total
Individu									
1	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
i	...	...	$p_i(\lambda_s u_{is})^2$	...	...	$p_i\|\widehat{\mathbf{x}}_i^k\|^2$	...	...	$p_i\|\mathbf{x}_i\|^2$
...	...	...	...	...	...	...	...	...	...
n	...	...	...	...	...	...	...	...	...
Total	λ_1^2	...	λ_s^2	...	λ_k^2	$\|\widehat{X}^k\|^2$	...	λ_r^2	$\|X\|^2$

TAB. 3.4 - *La décomposition "individus $\times$ facteurs "de l'inertie*

L'orthonormalité des bases de facteurs réduits $\{\mathbf{u}_s\}$ ou de vecteurs principaux $\{\mathbf{v}_s\}$ et les équations (3.2) et (3.3) sont à la base des deux décompositions additives présentées ici sous la forme des tableaux 3.3 et 3.4. La donnée d'un tableau est équivalente à celle de trois équations (la somme partielle d'une ligne est donnée dans la colonne "sous-total", la somme d'une ligne est donnée dans la dernière colonne, et la somme des éléments d'une colonne est donnée dans la dernière ligne). On présente dans un premier temps les indices de qualité, et dans un second les mesures d'influence ou contribution.

Exercice Écrire, puis démontrer les trois équations dérivées de chacun des tableaux.

3.5.2 Indices de qualité

Ces tableaux permettent de calculer les mesures d'influence (contributions) et de qualité (cosinus carré) définies dans ce paragraphe. On en déduit les indices suivant :

- *Qualité globale*: La somme partielle des éléments de la dernière ligne divisée par la somme globale (inertie totale) est la proportion d'inertie expliquée par les k premiers facteurs.
  $$\|\widehat{X}^k\|^2/\|X\|^2.$$
  On représente le plus souvent le diagramme en bâtons des carrés des valeurs singulières $(\lambda_s)^2$ ou "éboulis des valeurs propres" (cf. figure 3.2). Sa surface totale représente l'inertie à expliquerla surface des k premières barres l'inertie expliquée.
- *Qualité de l'approximation d'un vecteur individu ou variable*: la somme partielle d'une ligne (colonne "sous-total") divisée par le total de la ligne est un indice de qualité de l'explication des variables ou des individus par les premiers facteurs. On a ainsi:
  - *Qualité de la représentation d'une variable :* La proportion de variance expliquée de la variable j est le carré du *coefficient de corrélation multiple de la variable avec les composantes principales* $\mathbf{u}_s$ (noté $r(\mathbf{x}^j, (\mathbf{u}_1, \ldots, \mathbf{u}_k))^2$. C'est encore le R^2 de la régression multiple de $\mathbf{x}^j$ en fonction des variables $\mathbf{u}_1, \ldots, \mathbf{u}_k$.
    $$r^2(\mathbf{x}^j, (\mathbf{u}_1, \ldots, \mathbf{u}_k)) = var(\widehat{\mathbf{x}^j}^k)/var(\mathbf{x}^j)$$
    Multipliés par 1000, ces indices figurent, pour l'exemple des notes scolaires, dans le tableau 3.5. Ainsi, on s'aperçoit que la variable `dm` n'est expliquée qu'à 2,1% sur le plan principal, les quatre autres l'étant à plus de 99%.

	f1	f2	f3	f4	f5	TOT
m	657	**998**	1000	1000	1000	1000
s	812	**999**	999	1000	1000	1000
f	568	**991**	999	999	999	1000
l	839	**998**	999	1000	1000	1000
dm	3	21	1000	1000	1000	1000
TOT	577	824	1000	1000	1000	1000

TAB. 3.5 - *Indices de qualité des variables*

– *Qualité de la représentation des individus* L'originalité d'un individu a été défini comme le carré de sa distance à l'individu moyen. L'indice donne la proportion de l'*originalité* de l'individu pris en compte par le résumé. Cette proportion est le carré du cosinus du vecteur $\mathbf{x}_i$ avec sa projection $\widehat{\mathbf{x}}_i^k$ sur Z_k

$$cos^2(\mathbf{x}_i, \widehat{\mathbf{x}_i}^k) = ||\widehat{\mathbf{x}_i}^k||^2/||\mathbf{x}_i||^2.$$

Transformé en millièmes, c'est le tableau 3.6 dans l'exemple des notes scolaires. Dans notre exemple, seul Evelyne n'a pas son originalité correctement expliquée sur le plan principal de l'espace des individus (à 25% seulement).

Exercice Écrire $r^2(\mathbf{x}^j, (\mathbf{u}_1, \ldots, \mathbf{u}_k))$ en fonction des carrés de corrélation simples $r^2(\mathbf{x}^j, \mathbf{u}_s)$.

	Axe1	Axes1a2	Axes1a3	Axes1a4	Axes1a5	TOT
JE	889	923	1000	1000	1000	1000
AL	804	830	1000	1000	1000	1000
AN	460	994	998	999	999	1000
MO	894	894	999	999	999	1000
DI	877	979	999	999	1000	1000
AD	236	814	999	999	1000	1000
PI	26	933	1000	1000	1000	1000
BR	174	910	997	1000	1000	1000
EV	53	**251**	999	999	999	1000
TOT	577	824	1000	1000	1000	1000

TAB. 3.6 - *Indices de qualité des individus*

3.5.3 Mesures d'influences ou contributions

Les mesures d'influence sont donnés dans les Tableaux 3.7 et 3.8. Ce sont des tableaux de fréquences colonne (donnés en millièmes) calculés sur les Tableaux 3.3 et 3.4.

– *Contribution des variables ou des individus à l'inertie totale*

  On étudie la décomposition de l'inertie totale comme somme des éléments de la dernière colonne des tableaux 3.3 et 3.4. On obtient les mesures d'influence des individus ou variables dans l'analyse, qui sont appelées contributions des individus — ou des variables — à l'inertie totale. L'importance de la variable j est mesurée par le produit de m_{jj} par sa variance. C'est pourquoi lorsque les variables ont des unités différentes il est nécessaire de les réduire. Si de plus la métrique est usuelle ($M = I_p$), les contributions des variables deviennent alors toutes égales. De la même façon, la contribution d'un individu est $p_i||\mathbf{x}_i||_M^2$. Un individu dont la contribution dans l'analyse est très forte risque de la perturber. On pourra être amené à "expulser" ce gêneur (*outlier* ou *individu aberrant*), par exemple en lui affectant un poids nul. Ces contributions figurent dans l'avant-dernière colonne (d'intitulé "TOT") des tableaux 3.7 et 3.8.

  Ces contributions s'utilisent de la façon suivante. Quand un individu a une forte contribution, il est intéressant de contrôler si cette forte contribution est due à un fort poids ou à une forte originalité. De la même façon, une variable de forte contribution pourra l'être parce que le coefficient m_{jj} est grand ou parce qu'elle a une forte variance. Les tableaux de contributions contiennent dans leurs dernières colonnes des valeurs proportionnelles aux variances des variables (colonne "VARI") ou à l'originalité des individus (colonne "ORI"). Ces valeurs sont égales aux contributions des variables ou des individus si $M = I_p$ (resp. $D = 1/nI_n$).

	f1	f2	f3	f4	f5	TOT	VARI
m	265	321	3	83	328	233	233
s	257	138	0	306	299	183	183
f	242	423	12	155	168	246	246
l	235	104	1	455	206	162	162
dm	1	13	985	1	0	177	177
TOT	1000	1000	1000	1000	1000	1000	1000

TAB. 3.7 - *Contributions des variables aux facteurs*

	f1	f2	f3	f4	f5	TOT	ORI
JE	298	27	84	114	155	193	193
AL	61	5	42	42	21	44	44
AN	41	110	1	152	4	51	51
MO	374	0	144	154	114	241	241
DI	160	43	12	26	398	105	105
AD	35	199	90	3	247	85	85
PI	4	357	37	68	21	97	97
BR	15	152	25	302	1	51	51
EV	12	107	565	139	39	133	133
TOT	1000	1000	1000	1000	1000	1000	1000

TAB. 3.8 - *Contributions des individus aux facteurs*

– *Contributions des facteurs aux variables*: Le total de la ligne, situé en dernière colonne, se décompose additivement en quantités associées à chaque facteur, soit après simplification par m_{jj} ou p_i :

$$\begin{aligned} var(\mathbf{x}^j) &= \sum_{s=1}^{r}(\lambda_s v_{js})^2 \\ ||\mathbf{x}_i||_M^2 &= \sum_{s=1}^{r}(\lambda_s u_{is})^2 \end{aligned}$$

On retrouve la signification d'une saturation, dont le carrée est la contribution du facteur à l'explication de la variance d'une variable. Une notion analogue est la contribution d'un facteur à l'originalité de chaque individu. Ces indices permettent de répondre à la question : quels sont les facteurs qui permettent d'interpréter une variable ou un individu. Ils ne sont pas ici donnés directement ici, mais uniquement sous forme cumulés, dans les tableaux 3.5 et 3.6.

Exercice Montrer que les contributions relatives des facteurs aux variables ou aux individus sont des carrés de coefficients de corrélation ou de cosinus :

$$\begin{aligned} \frac{(\lambda_s v_{js})^2}{var(\mathbf{x}^j)} &= r(\mathbf{u}_s, \mathbf{x}^j)^2 \quad \text{Contribution du facteur } s \text{ à } \mathbf{x}^j \\ \frac{(\lambda_s u_{is})^2}{||\mathbf{x}_i||_M^2} &= cos(\mathbf{v}_s, \mathbf{x}_i)^2 \quad \text{Contribution du facteur } s \text{ à } \mathbf{x}_i. \end{aligned}$$

- *Contributions des variables et des individus aux facteurs*: la variance $(\lambda_s)^2$ d'un facteur (dernière ligne du tableau 3.7 ou 3.8) est due à la somme de quantités positives associées à chaque variable ou à chaque individu :

$$(\lambda_s)^2 = \sum_{j=1}^{p} m_{jj}(\lambda_s v_{js})^2$$
$$(\lambda_s)^2 = \sum_{i=1}^{n} p_i(\lambda_s u_{is})^2.$$

On mesure l'importance d'un individu ou d'une variable dans un facteur par le rapport de cette quantité à $(\lambda_s)^2$.

Remarque : Si la métrique est usuelle ou si les poids sont équipondérés, ces contributions se lisent directement sur les graphiques isométriques (les contributions sont proportionnelles aux carrés des coordonnées des vecteurs individus ou variables. Dans ce cas la consultation des contributions aux facteurs peut être évitée.

Dans notre exemple, l'importance d'une variable dans la détermination d'un facteur est donné par le tableau des *contributions des variables aux facteurs* (tableau 3.7, 5 premières colonnes). Ainsi, on voit que les deux premiers facteurs ne sont pas influencés par la variable `dm`, et que le troisième facteur n'est caractérisé pratiquement que par cette variable. L'importance des individus pour la détermination des facteurs est donné dans le tableau 3.8, 5 premières colonnes. Ainsi, Evelyne a une influence déterminante dans la détermination du facteur 3.

3.6 Schéma d'interprétation d'une analyse en composantes principales

L'interprétation d'une ACP demande une certaine pratique. La multitude des précautions et indices permettant de contrôler une interprétation ne doit pas empècher l'analyste d'aller à l'essentiel, puis de moduler ses affirmations en s'aidant de certains indices disponibles. En aucun cas une interprétation ne consiste en une accumulation de lectures d'indices...

3.6.1 Phase préliminaire ou choix de codage et des métriques

On commence par vérifier si une étude des liaisons linéaires est justifiée ou si une si des transformations préliminaires de variables sont souhaitables. Pour ce faire, on affiche la matrice des diagrammes de dispersion des variables deux à deux (fonction `pairs` de Splus). Il faut ensuite se poser la question du choix du codage, de la métrique et des poids.

Choix du codage et de la métrique Ces choix se font de façon concommittantes et s'appuyent sur la décomposition de l'inertie

$$In(Y) = \sum_{j=1}^{p} m_{jj} var(\mathbf{y}^j).$$

qui détermine l'influence d'une variable dans l'analyse. On a cependant les règles suivantes :

- Lorsque les variables sont d'unités différentes, il est *nécessaire* de les réduire.
- Lorsqu'elles ont même unité, on peut encore les réduire si on estime que toutes les variables doivent avoir la même influence. Sinon, l'importance des variables peut être choisie proportionnelle à une mesure de leur variabilité. Ainsi, l'influence d'une variable peut-être égale à sa variance ($m_{jj} = 1$ et variables non réduites ou $m_{jj} = var(\mathbf{y}^j)$ et variables réduites) ou à leur écart-type ($m_{jj} = 1/\sigma_j$ et variables non-réduites ou $m_{jj} = \sigma_j$ et variables réduites). Des considérations d'invariance par regroupement de variables peut aider à guider ce choix (voir le problème d'examen du Magistère en 1995).

Les choix les plus fréquents sont $M = I_p$ avec variables non-réduites.

Choix des poids Faut-il donner à chaque individu le même poids? Quand les individus sont "égaux" entre eux, ce choix s'impose. Mais lorsque les individus représentent des aggrégats (communes, catégories, ...), il peut être nécessaire de pondérer ces aggrégats de sorte que la moyenne pondérée des variables sur les aggrégats soit égale à la moyenne de la variable sur la population entière (en clair, on donne à chaque aggrégat un poids proportionnel aux individus qu'il représente).

3.6.2 Contrôle a posteriori des "constituants du mélange multidimensionnnel"

Les choix étant effectués, on en vérifie la validité, en effectuant un contrôle a posteriori qui utilise les contributions des variables et des individus à l'inertie totale. Rappellons que l'inertie est un "mélange" des variances des variables et des "originalités" des individus ($\|\mathbf{y}_i - \mathbf{g}\|^2$. Il est important de vérifier si les contributions des différentes variables à l'inertie sont suffisamment homogènes, ou si leur hétérogénéité est justifiée par la nature des données. Il faut procéder de même pour les individus. Si un individu est un outlier, il risque d'entrainer à lui seul la détermination d'un facteur. Si les contributions semblent trop hétérogènes, on peut relancer le programme après avoir changé les options (codage, poids, métrique), éventuellement après avoir "expulsé" un individu trop atypique.

3.6.3 Examen du nombre de facteurs à retenir

Une méthode empirique consiste à observer l'éboulis des valeurs propres, représentant par un histogramme les pourcentages d'inertie expliquée. L'inertie totale est représentée par la surface de l'histogramme, l'inertie conservée par la surface des barres associées aux facteurs conservés (les k premiers).

```
*********************************************          45%
****************************************               40%
******                                                  6%
****                                                    4%
***                                                     3%
**                                                      2%
```

L'objet est de conserver le maximum d'inertie avec le minimum de facteurs. On cherche l'endroit de la rupture de pente de l'histogramme (ici entre 2 et 3). Ansi on voit que dans cette analyse, on ne perd pas beaucoup d'information en négligeant les facteurs d'ordre supérieurs à 2.

L'idée statistique sous jacente est qu'il faut séparer le signal (ce qui contient l'information pertinente) du bruit. Des méthodes récentes, utilisant la théorie des pertubations et le bootstrap peuvent donner des indications précieuses. Une méthode due à Ph. Besse (cf. Besse 1993) mesure, par une approximation analytique, l'instabilité du sous-espace de dimension k en présence de fluctuations d'échantillonages. On cherche un sous-espace principal associé à un minimum local de cet indicateur. Cette méthode est implémentée dans la fonction Splus `rsp` qui fournit pour chaque dimension la valeur de cette instabilité. Ses minima locaux sont des dimensions qu'il est possible de choisir. Lorsque la dimension proposée est grande (supérieure à 6 par exemple), il sera souvent impossible d'interpréter tous les axes. Mais le choix de la dimension dépend aussi des objectifs de l'analyse (décrire ou résumer pour une analyse ultérieure).

3.6.4 Interprétation des facteurs et de la représentation des variables

Représentation des variables

On utilise pour cette interprétation la représentation des variables projetées sur leur sous-espace principal. On peut y lire les variances (carrés des longueurs des représentations des variables) et corrélation (par les cosinus qu'elles forment) des variables approximées.

On peut lire numériquement l'importance de chaque variable ($m_{jj}v_{jk}^2$) dans la détermination d'un facteur (contributions des variables aux facteurs). *Un facteur s'interprète relativement aux variables qui y ont le plus contribuées.* Ce sont donc ces quantités qu'il faut consulter pour interprèter un facteur. Insistons sur le fait que le plus souvent, $m_{jj} = 1$, et les plus grandes valeurs de $m_{jj} v_{jk}^2$ sont celles des variables de plus grandes coordonnées. Dans ce cas, cette étude peut donc se faire graphiquement, "d'un coup d'oeil".

Symétriquement, on peut s'intéresser à une variable particulière. Pour savoir sur quels facteurs une variable s'interprète, on regarde dans le tableau des corrélations variables-facteurs (qui sont encore des contributions des facteurs aux variables). Enfin, on peut vérifier la qualité du résumé d'une variable sur un sous-espace dans le tableau des corrélations multiples "variable - sous-espace".

Cercle des corrélations

On peut encore représenter les projections des variables réduites (ou vecteurs unitaires portés par les variables), et on a les propriétés suivantes :

- Dans cette représentation, toutes les projections se font à l'intérieur d'un cercle de rayon 1. C'est pourquoi on l'appelle "*Cercle des corrélation*" (voir exemple Figure 3.11). Les coordonnées des variables sur les axes sont les corrélations variables facteurs (cf. équation (3.3)).
- La longueur d'une variable est son coefficient de corrélation multiple "variable - sous-espace de représentation", et les coordonnées d'une variable sont des corrélations variables-facteurs.
- Si une variable est sur le cercle, elle appartient au sous-espace de représentation. Dans ce cas, la longueur de la projection de toute autre variable sur celle-ci est égale au coefficient de corrélation entre cette autre variable avec la première.
- Losque les variables ont été préalablement réduites, cette représentation se confond avec la précédente.

3.6.5 Interprétation de la représentation des individus :

L'interprétation est symétrique de celle des variables. Si les individus sont nombreux ou anonymes, on peut ignorer cette interprétation. Dans ce cas, on s'interesse à des barycentres d'individus (qui ont telle ou telle caractéristique) et non aux individus eux-mêmes (cf. 3.7.1).

3.6.6 Représentation simultanée

L'interprétation la mieux adaptée à une représentation simultanée est celle du biplot, avec ou sans axes gradués. Elle fait l'objet d'une étude ultérieure.

L'interprétation classique se base sur

- la notion de corrélation entre variables et facteurs :

$$r(\mathbf{y}^j, \mathbf{u}_k) = \frac{\lambda_k v_{jk}}{\sigma_j},$$

  obtenues graphiquement à partir des coordonnées des variables en les divisant par σ_j (ou par 1, cas du cercle des corrélations).
- et sur le fait que la coordonnée d'un individu i sur le s-ième axe principal est la valeur $\lambda_k \mathbf{u}_{i,k}$ du facteur (non réduit) k pour l'individu i.

Les deux ingrédients ($r(\mathbf{y}^j, \mathbf{u}_k)$ et $\lambda_k \mathbf{u}_{i,k}$) étant lues graphiquement, il s'agit donc de faire une interprétation *directement* à partir du graphique. Le principe en est le suivant :

En règle général, si un facteur a une forte corrélation positive avec une variable, une grande valeur de ce facteur pour un individu sera associée à une forte valeur de cette variable pour ce même individu. Si cette corrélation est fortement négative, les conclusions sont inverses. Par conséquent, un individu ayant *une grande coordonnée sur le s-ième axe principal* a tendance a avoir de *fortes valeurs positives pour les variables fortement corrélée positivement avec le s-ième facteur* et de *fortes valeurs négatives pour les variables fortement corrélée négativement avec ce facteur.* Ce raisonnement est un des moyens de faire la jonction entre les interprétations du nuage des individus et du nuage des variables (interprétation simultanée).

Une autre façon de comprendre l'interprétation simultanée est basée sur la formule de reconstitution des données (cf ; (3.4), qui s'écrit à l'ordre $k = r$ (reconstitution parfaite) :

$$x_i^j = \sum_{s=1}^{r} \frac{1}{\lambda_s} (\lambda_s u_{is})(\lambda_s v_{js})$$

On reconnait dans la formule les coordonnées des individus et des variables. Lorsque les coordonnées sur l'axe k des individus et des variables sont grandes et de même signe, alors la k-iéme dimension apporte une contribution positive à x_i^j. En fait ces contributions sont en général hiérachisées (les plus grandes sont sur les premiers axes), ce qui fait qu'on interprète d'abord sur le plan principal (2 premiers facteurs), puis on corrige avec les apports des autres dimensions considérées. Mais lorsqu'on obtient sur le deuxième axe une contribution à x_i^j opposée à celle obtenue sur le premier axe, il est difficile de conclure graphiquement. On verra que la technique du biplot permet de faire cette interprétation plan par plan et non pas axe par axe.

Facteur taille, facteur forme

Définition On appelle facteur *taille* un facteur dont les corrélations avec l'ensemble des variables sont toutes de mêmes signes. Un individu qui a un grand facteur taille est un individu qui a *en moyenne* de grandes valeurs pour l'ensemble des variables. On appelle facteur *forme* ou facteur d'*opposition* un facteur qui a des corrélations fortement positives avec un groupe de variables et des corrélations fortement négatives avec un autre groupe de variables. Un individu qui prend une grande valeur pour ce facteur prendra en moyenne de fortes valeurs positives pour les variables du premier groupe et de fortes valeurs négatives pour les variables du second groupe.

Dans l'exemple des notes scolaires, le premier facteur est un facteur taille qui oppose les bons élèves au mauvais (ici il s'agirait plutôt d'un facteur *niveau*). C'est la cause de variation la plus importante du tableau. Le second facteur par ordre d'importance est un facteur d'opposition ; il oppose les élèves qui sont situés en haut sur le graphique (les littéraires) à ceux qui sont situés en bas (les scientifiques).

3.7 Elements illustratifs

3.7.1 Variables et individus illustratifs ; régression orthogonale

On peut ajouter à une analyse déjà effectuée des variables ou des individus *supplémentaires* ou illustratifs. Les variables (ou les individus) supplémentaires subissent préalablement les mêmes traitements (centrages, réduction) que les variables (ou individus) initiaux. On peut alors les projeter sur les sous-espaces principaux, pour analyser leur position par rapport aux éléments "actifs" (éléments qui ont contribués à l'analyse) de même type.

Régression orthogonale: Pour les variables, cette pratique revient à regresser les variables illustratives sur les variables actives (ou sur un sous-espace principal inclus dans le sous-espace qu'elles engendrent), et s'appelle "régression orthogonale". En effet, considérons une nouvelle variable centrée $\mathbf{x}^{sup}$ et la base orthonormée de $\mathbf{R}^n$ $\{\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \ldots, \mathbf{u}_n\}$, obtenue en complétant la famille orthonormée des facteurs réduits. Si on décompose la nouvelle variable $\mathbf{x}^s$ sur cette base, et si $k \leq r$, on obtient:

$$\mathbf{x}^{sup} = (\sum_{s=1}^{k} \alpha_s \mathbf{u}_s) + (\sum_{s=k+1}^{n} \alpha_s \mathbf{u}_s).$$

Le premier terme entre parenthèses est la projection de $\mathbf{x}^s$ sur le sous-espace principal $Im(U_k)$. en d'autres termes, c'est le vecteur des valeurs ajustées de la régression de $\mathbf{x}^{sup}$ sur les nouvelles variables non corrélées $\mathbf{u}_s$), $s = 1, \ldots, k$. Si $k = r$, c'est encore le vecteur des valeurs ajustées de la régression de $\mathbf{x}^{sup}$ sur les variables initiales de l'ACP (car $Im(X) = Im(U_r)$). L'intérêt est multiple. Lorsque les variables initiales sont linéairement indépendantes, le calcul du projecteur sur le sous-espace engendré pose problème. Mais on montre de plus que cette régression, lorsqu'on n'utilise pas tous les facteurs, est *robuste* (les coefficients estimés et les valeurs ajustées ne dépendent pas trop de petites variations sur les variables).

On peut ensuite calculer les corrélations entre les variables illustratives et les facteurs et les coefficients de corrélation multiple entre variables illustratives et sous-espace factoriels. Ces derniers coefficients sont à comparer au R^2 de la régression multiple: ils lui sont égaux si on utilise tous les facteurs.

Barycentres Plutôt que d'analyser un à un la position des individus, on représente des barycentres d'individus qui ont la même caractéristique. Etant donnée une variable catégorielle, on peut lui associer une partition de la population : un groupe est l'ensemble des individus qui ont la même modalité. On peut alors voir s'il existe des interactions entre les facteurs représentés et la variable qualitative en représentant les barycentres des groupes. Si la variable catégorielle est indépendante d'un facteur, on peut considérer que les éléments d'un groupe sont tirés au hasard et la moyenne du facteur pour ce groupe ne doit pas être très différente de la moyenne générale du facteur (c'est à dire 0). Mais cela dépend aussi du nombre d'éléments du groupe (à cause de la loi des grands nombres).

3.8 La représentation du biplot

On présente ici une autre représentation des approximations, qui en permet une lecture plus précise et plus rapide. On s'intéresse ici à la représentation d'un tableau de rang 2. Mais dans la pratique, cette technique sera étendue à des matrices de rang supérieures.

3.8.1 Tableau de rang deux et facteurs

Si X est une matrice $n \times p$ de rang 2, il existe deux *nouvelles variables* de $\mathbf{R}^n$ (jouant le rôle de *facteurs*), telles que toute variable soit combinaison linéaire de ces deux facteurs. Prenons l'exemple du tableau 3.9 : on suppose avoir mesuré l'intelligence de n élèves (c'est un vieux rêve des psychomètres), ainsi que leurs aptitudes manuelles. On leur a ensuite fait subir quatre tests.

	Intell.	Apt.man
Alfred	17	7
Albert	20	19
Alain	18	11
Antoine	11	8
Aristide	15	17
Ahmed	4	15
Andre	3	18
Alceste	9	15
Aime	11	18
Arsene	6	11

	Test1	Test2	Test3	Test4
Alfred	89	120	45	75
Albert	137	195	96	99
Alain	105	145	62	83
Antoine	68	95	43	52
Aristide	111	160	83	77
Ahmed	61	95	64	31
Andre	66	105	75	30
Alceste	81	120	69	51
Aime	98	145	83	62
Arsene	57	85	50	35

TAB. 3.9 - *Toute colonne d'un des tableaux est combinaison linéaire des colonnes de l'autre tableau*

On peut vérifier dans cet exemple, que les notes des tests s'obtiennent comme combinaison linéaire des notes d'intelligence et d'aptitude manuelle. Ainsi, le score obtenu au test 1 est quatre fois la note d'intelligence plus trois fois l'aptitude manuelle. De même, pour les autres tests, ces nombres sont de $(5, 5)$, $(1, 4)$, et $(4, 1)$. Il est donc possible, sans perte d'information, de comprimer l'information contenu dans le tableau de droite en donnant les notes d'intelligence et d'aptitude manuelle, puis en donnant les coefficients permettant de passer de ces notes à celles des tests (soit ici $20 + 8$ nombres au lieu de 40).

Matriciellement, si X est la matrice 10×4 des notes de test, si U est la matrice 10×2 dont les colonnes sont les notes d'intelligence et d'aptitude manuelle, et si V est la matrice 4×2 contenant les coefficients

$$V = \begin{pmatrix} 4 & 3 \\ 5 & 5 \\ 1 & 4 \\ 4 & 1 \end{pmatrix},$$

alors on a :

$$X = \begin{pmatrix} 89 & 120 & 45 & 75 \\ 137 & 195 & 96 & 99 \\ 105 & 145 & 62 & 83 \\ 68 & 95 & 43 & 52 \\ 111 & 160 & 83 & 77 \\ 61 & 95 & 64 & 31 \\ 66 & 105 & 75 & 30 \\ 81 & 120 & 69 & 51 \\ 98 & 145 & 83 & 62 \\ 57 & 85 & 50 & 35 \end{pmatrix} = \begin{pmatrix} 17 & 7 \\ 20 & 19 \\ 18 & 11 \\ 11 & 8 \\ 15 & 17 \\ 4 & 15 \\ 3 & 18 \\ 9 & 15 \\ 11 & 18 \\ 6 & 11 \end{pmatrix} \begin{pmatrix} 4 & 5 & 1 & 4 \\ 3 & 5 & 4 & 1 \end{pmatrix} = U\,{}^tV.$$

En notant $\mathbf{u}_s$ la s-ième colonne de U et $\mathbf{v}_s$ la s-ième colonne de V, on a $X = \sum_{s=1}^{2} \mathbf{u}_s\,{}^t\mathbf{v}_s$

Propriétés

D'une façon générale, si X est une matrice $n \times p$ de rang k, il existe k vecteurs linéairement indépendants $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$ de $\mathbf{R}^n$ et k vecteurs linéairement indépendants $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ de $\mathbf{R}^p$ tels que :

$$X = \sum_{s=1}^{k} \mathbf{u}_s\,{}^t\mathbf{v}_s$$

3.8.2 Le biplot classique

On écrit la matrice X $n \times p$ de rang 2 sous la forme $X = A\,{}^tB$ avec $A = (a_{ik})$ $n \times 2$ et $B = (b_{kj})$ $p \times 2$, soit encore :

$$x_i^j = a_{i1}b_{j1} + a_{i2}b_{j2}.$$

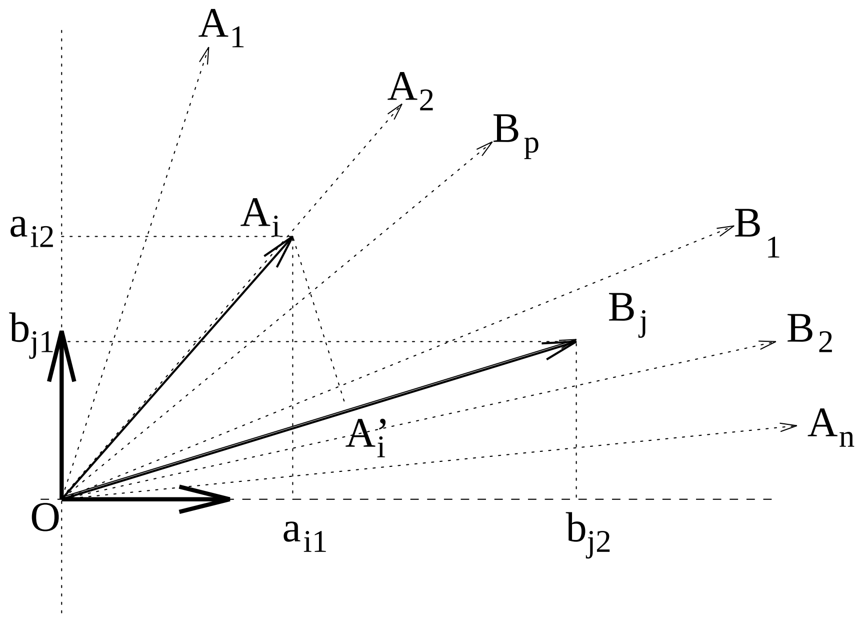


FIG. 3.5 - *Lecture du biplot*

La figure 3.5 est un *biplot*[2] composé de deux famille de vecteurs, les vecteur OA_i, de coordonnées dans une base orthonormée (a_{i1}, a_{i2}) et les vecteurs OB_j , de coordonnées (b_{j1}, b_{j2}). Les points A_i et B_j sont appelés respectivement marqueurs des individus i et des variables j. Cette figure représente les données dans le sens suivant : l'élément $x_i^j = a_{i1}b_{j1} + a_{i2}b_{j2}$ de X est le produit scalaire usuel dans $\mathbf{R}^2$ des vecteurs OA_i et OB_j. En notant A'_i la projection orthogonale de A_i sur l'axe joignant l'origine O à B_j, on voit que x_i^j est le produit des longueurs algébriques des vecteurs OA'_i et OB_j, soit :

2. Le biplot est en général présenté pour une matrice de rang r quelconque. Il se représente alors dans un espace de dimension r. Ici on se restreint à des biplots "bidimensionnels" qui sont les seuls à permettre une lecture facile.

$$x_i^j = |OA_i'||OB_j|.$$

En conséquence,

- lorsque l'angle $(OA_i,\ OB_j)$ est aigu, alors la valeur x_i^j de la variable j pour l'individu i est positive,
- lorsque l'angle $(OA_i,\ OB_j)$ est obtus, x_i^j est négatif,
- lorsque l'angle $(OA_i,\ OB_j)$ est droit, x_i^j est nul,
- enfin, plus un marqueur d'individu A_i est éloigné dans la direction d'une flèche OB_j, plus la valeur de la variable j pour cet individu i est grande.

Quelques propriétés

- Toute matrice de rang 2 admet une telle représentation plane.
- Il existe une infinité de telles représentations (autant que de décompositions de X sous la forme AB^t).
- Une matrice de rang 4 est la somme de deux matrices de rang 2 ; elle peut se représenter à l'aide de 2 biplots bidimensionnels. La valeur d'une variable sur un individu s'obtient en sommant les valeurs lues sur chaque graphique. En général, on s'arrange pour que le premier biplot donne l'information la plus significative, et le second *précise* ou apporte des *corrections* aux valeurs lues sur le premier.

Exemple Pour obtenir la figure 3.6, on a ajouté deux colonnes à X en lui accolant les deux colonnes (ou variables) Intelligence et Aptitude manuelle. Par construction, la nouvelle matrice X est encore de rang 2. On s'est alors donné une décomposition de X de la forme $X = A^tB$ et on effectué la représentation des deux nuages sur deux graphiques différents pour ne pas surcharger la figure (ce n'est pas la façon classique d'opérer). On peut vérifier que la lecture graphique des éléments de la matrice X est possible.

Exercice Retrouver les valeurs de deux éléments de la matrice X donnée Tableau 3.9.

On peut aussi retrouver sur la figure les coefficients de la décomposition des notes de tests selon les des deux variables initiales : ainsi la note du test1 s'écrit : $test1 = 4 * Intell. + 3 * Apt.man.$

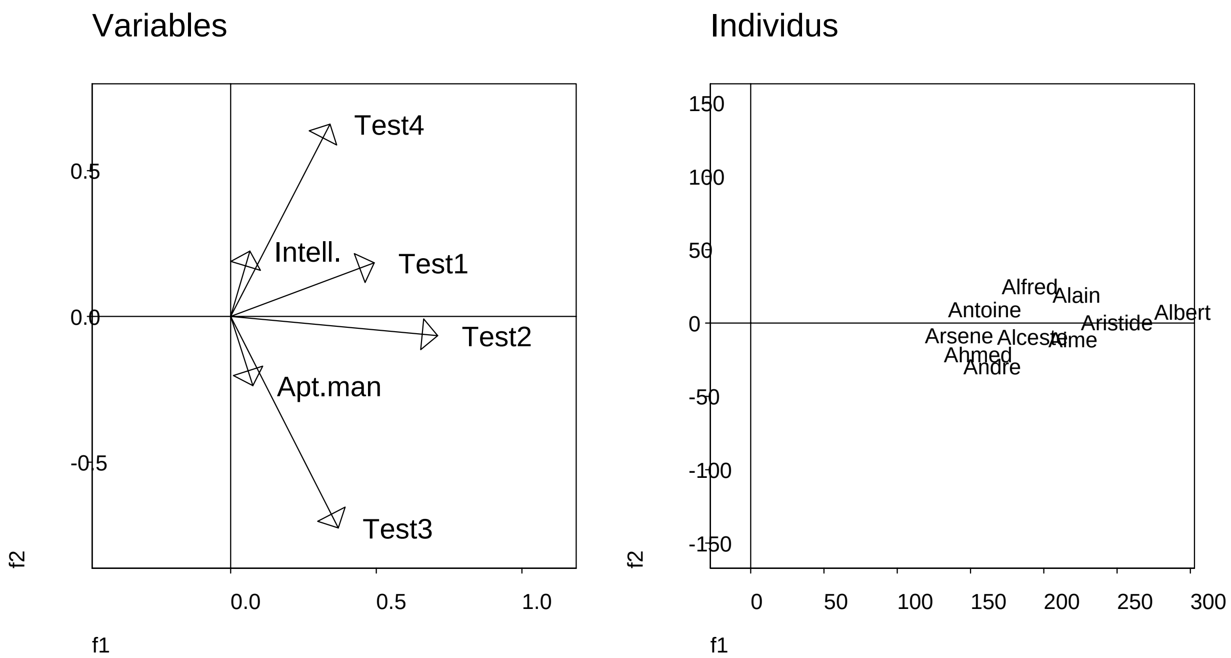


FIG. 3.6 - *Biplot de Gabriel sur les notes de test*

Le choix de la décomposition

Dans une logique interne à l'analyse en composantes principales, la décomposition de l'approximation d'ordre k de X sous la forme $\widehat{X}^k = A^tB$ est issue directement de la décomposition aux valeurs singulières et s'obtient par un certain choix de factorisation dans l'équation 3.1. On utilisera le plus souvent l'une des factorisations suivantes (on prend ici $k = 2$) :

$$\widehat{\mathbf{x}_i^j}^2 = \sum_{s=1}^{2}(\lambda_s u_{is})(v_{js}) \tag{3.6}$$

$$\widehat{\mathbf{x}_i^j}^2 = \sum_{s=1}^{2}(u_{is})(\lambda_s v_{js}) \tag{3.7}$$

$$\widehat{\mathbf{x}_i^j}^2 = \sum_{s=1}^{2}(\sqrt{\lambda_s}u_{is})(\sqrt{\lambda_s}v_{js}) \tag{3.8}$$

Ainsi, avec la première factorisation, on a $a_{is} = \lambda_s u_{is}$ et $b_{js} = v_{js}$. Les propriétés de ces différents choix seront étudiées au §3.8.4 et 3.8.5.

3.8.3 Le biplot gradué

La figure 3.7 est un biplot gradué. Les flèches ont été remplacées par des axes gradués de mêmes directions ; les marqueurs des individus ont la même représentation. En projetant les marqueurs des individus orthogonalement sur les axes associés aux variables (ou tests) on peut lire directement les notes des tests. Cette lecture est semblable à celle du diagramme de dispersion classique : mais ici les axes sont obliques et de nombre quelconque. La justification de cette propriété est conséquence des propriétés suivantes du produit scalaire :

- D'après l'équation :
  $$x_i^j = |OA_i'||OB_j|,$$
  x_i^j ne dépend que de la projection de A_i sur la droite OB_j (cf. Figure 3.5).
- De plus, pour j fixé, x_i^j est proportionnel à la longueur algébrique $|OA_i'|$.

Il est donc possible de graduer l'axe OB_j pour lire les valeurs de la variable j et obtenir ainsi une lecture directe des approximations des éléments x_i^j. Ces valeurs sont lues en projetant orthogonalement les marqueurs des individus i sur l'axe gradué j. On peut aussi changer l'origine de l'axe gradué, en rajoutant par exemple la moyenne de la variable. La valeur lue sur le graphique est alors l'approximation de la variable non centrée. On peut encore, si la variable était réduite, remultiplier les valeurs approchées par l'écart-type. On retrouve alors des approximations des valeurs situées dans le tableau initial. Remarquons que le biplot gradué permet en particulier de lire l'étendue de la variation de chaque variable. Sur la figure 3.7, on remarque que l'axe j a même direction que la flèche j de la figure 3.6, le nuage des individus restant inchangé.

Remarques

1. Il est bien sûr possible d'inverser les rôles de i et de j et de représenter un axe par individu. Comme on s'intéresse le plus souvent à comparer les valeurs d'une même variable pour différents individus, c'est plutôt la première représentation qui est effectuée.
2. Dans ce biplot, un des ensemble est représenté par des marqueurs, l'autre par des axes. La propriété qui suit, et qui concerne des conservations de distance et d'angles, ne s'appliquera pleinement que lorsque le nuage considéré sera représenté par des marqueurs.

3.8.4 Le biplot isométrique ligne

Parmi l'infinité des biplots possibles, certains possèdent des propriétés supplémentaires intéressantes. Le biplot issu de la factorisation donnée équation (3.6) présente la particularité d'avoir comme représentation des individus la représentation classique du nuage des

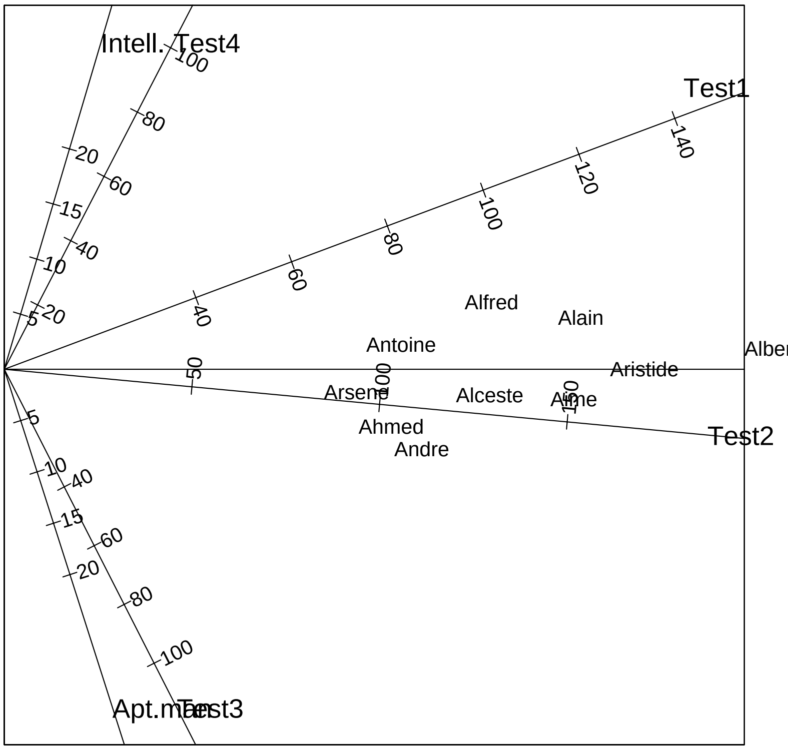


FIG. 3.7 - *Biplot gradué sur les notes de test (graphes superposés*

individus de l'ACP (cf. § 3.4.3). On l'appelle *biplot isométrique ligne* parce qu'il représente isométriquement le nuage $\{_{<k}(\mathbf{x}_i), i = 1, \ldots, n\}$, projection du nuage des individus.

Un tel biplot permet d'effectuer une *analyse des proximités entre individus*, donc d'étudier quels sont les individus qui se ressemblent, quels sont ceux qui s'opposent. Il conserve aussi les angles entre vecteurs individus.

Remarque La représentation du biplot isométrique ligne sera constituée d'une représentation "classique" (la représentation du nuage projeté des lignes) et d'une qui l'est moins (la représentation des variables). Il ne sera pas possible sur ce dernier graphique d'effectuer une lecture graphique des corrélations ou des écart-types des variables.

3.8.5 Le biplot isométrique colonne

Le biplot issu de la factorisation donnée équation (3.7) a pour représentation des variables la représentation classique de l'ACP (cf. § 3.4.3). On l'appelle *biplot isométrique colonne* parce qu'il représente isométriquement le nuage $\{P_{<k}(\mathbf{x}^j), j = 1, \ldots, p\}$, projection du nuage des variables sur le sous-espace principal. Il s'agit de la même manière d'un biplot préservant les distances et les angles entre les vecteurs colonnes. L'intérêt de cette propriété est directement lié à l'interprétation géométrique de l'écart-type et du coefficient de corrélation de variables centrées.

Ainsi, les variables étant centrées, le biplot isométrique colonne représente les variables par des vecteurs dont les angles ont pour cosinus leurs coefficients de corrélation. A titre d'exemple, effectuons l'ACP du tableau des tests (Tableau 3.9). On donne Tableau 3.10 ce tableau après centrage ; il a encore un rang ègal à 2.
La matrice des corrélations (Tableau 3.11) montre que toutes les variables sont corrélées positivement et que les notes aux tests 1 et 2 sont très corrélées.

La représentation des individus donnée figure 3.8 est celle du biplot isométrique colonne. Les cosinus des angles sont égaux aux coefficients de corrélation. Comme les variables sont corrélées positivement, tous les angles sont aigus, et l'angle entre les marqueurs des tests 1 et 2 est petit (cosinus proche de 1). De plus, la longueur des vecteurs est égal à l'écart-type des variables (on pourra vérifier que les écart-types des quatre variables sont respectivement de 25.69, 34.56, 17.65, 23.75).

En résumé, parmi les trois biplots associés à l'ACP,

– le biplot isométrique ligne a une représentation des lignes qui est la représentation

	Test1	Test2	Test3	Test4
Alfred	1.7	-6.5	-22.0	15.5
Albert	49.7	68.5	29.0	39.5
Alain	17.7	18.5	-5.0	23.5
Antoine	-19.3	-31.5	-24.0	-7.5
Aristide	23.7	33.5	16.0	17.5
Ahmed	-26.3	-31.5	-3.0	-28.5
Andre	-21.3	-21.5	8.0	-29.5
Alceste	-6.3	-6.5	2.0	-8.5
Aime	10.7	18.5	16.0	2.5
Arsene	-30.3	-41.5	-17.0	-24.5

TAB. 3.10 - *Tests : notes centrées*

	Test1	Test2	Test3	Test4
Test1	1.00	0.99	0.67	0.94
Test2	0.99	1.00	0.77	0.88
Test3	0.67	0.77	1.00	0.38
Test4	0.94	0.88	0.38	1.00

TAB. 3.11 - *Notes aux tests : la matrice des corrélations*

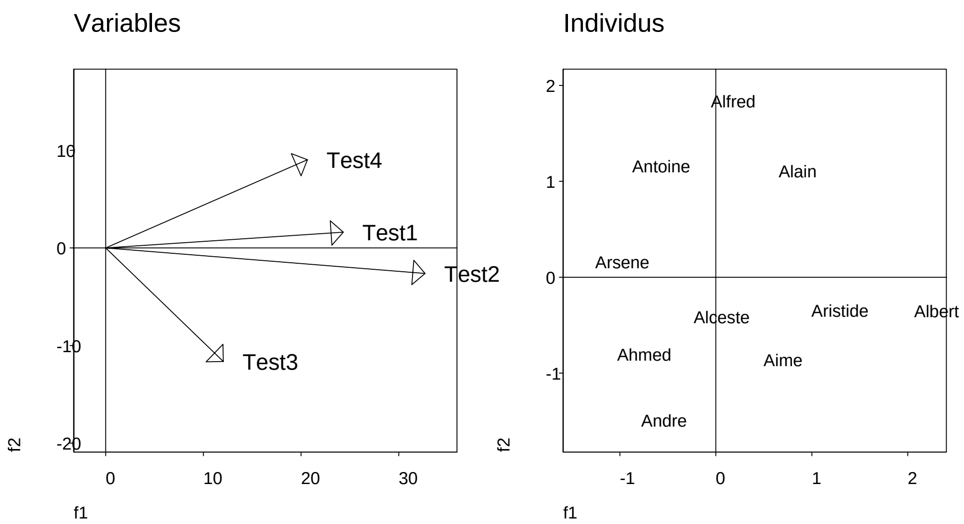


FIG. 3.8 - *Le biplot isométrique colonne*

classique de l'ACP (représentation du nuage projeté des lignes), mais celle des colonnes étant non classique)

- le biplot isométrique colonne a une représentation des variables qui est la représentation classique de l'ACP (représentation du nuage projeté des variables), mais celle des lignes est non classique[3]
- la dernière représentation (dite "à l'hollandaise") a deux représentations non isométrique et est utile dans des contextes particuliers.

Remarque Lorsque les valeurs propres associées aux deux axes de représentation sont de même ordre, ces différentes représentations sont visuellement très proches, et on pourra se contenter de n'en utiliser qu'une seule.

3.9 Exemple d'ACP : Étude des durées des trajets des bus de la ligne no 2 (Matabiau-Rangueil)

L'Agence d'Urbanisme de l'Agglomération Toulousaine a mesuré[4] pour la SEMVAT les horaires de passage du bus no 2 à 9 carrefours, et ceci pendant une journée complète du printemps 1991 et une journée "équivalente" du printemps 93. Le tableau de données initial contient les heures de passage des bus allant de la gare Matabiau à Rangueil (Facultés ou Centre Hospitalo-Universitaire) et de Rangueil à la gare Matabiau aux 9 carrefours *Matabiau, Bonrepos, Strasbourg, Esquirol, J.Guesde, Recollets, Rocade, Ponsan, CHUouFAC*.
Ici, on ne considère que le tableau donnant les heures de passage des bus dans le sens Matabiau vers Rangueil, l'année 1991. On peut d'abord visualiser les données initiales (cf. Figure 3.9). Le trajet d'un bus est représenté par une ligne joignant les points de coordonnées (k, t_k) (numéro du carrefour, heure de passage au carrefour). La pente d'une ligne est d'autant plus grande que le bus avance lentement.

On a ensuite construit le tableau des temps de trajet entre deux stations. Pour visualiser ces intervalles, on utilise les boxplots ou "boites à moustaches" (cf. figure 3.10 à gauche). Les durées de parcours Ponsan-CHU et Ponsan-Fac étant différente, on retrace le même graphe en différenciant les deux parcours (partie droite de la même figure). On constate que le trajet aboutissant au CHU dure environ 1mn ($\sim$ 2/100 heures) de plus que celui dont le terminus est la fac.

Pour mieux appréhender les différences entre les durées des différents parcours, on peut en faire une analyse multidimensionnelle. Ces données étant quantitatives, on réalise une analyse en composantes principales (ACP). Il n'y a pas de raison de donner aux différents bus des poids différents donc l'analyse sera équipondérée. L'analyse des distributions des durées (cf. figure 3.10) montrent que les plus grandes variances correspondent à des médianes élevées. Une analyse non réduite va donner plus d'importance aux variables de grandes variances. Une analyse réduite tiendra compte de façon uniforme de l'ensemble des tronçons (mais l'autre choix se justifie aussi). Les résultats de cette ACP sont illustrés par un biplot isométrique colonne.

La représentation du plan 1-2 est donnée figure 3.11. On reconnaît un facteur taille (axe 1) qui oppose les bus lents (à droite) aux bus rapides (à gauche). Le second facteur est un facteur forme qui oppose le tronçon "Matabiau-Bonrepos" aux tronçons "Strasbourg-Esquirol", "Rocade-Ponsan", "J.Guesde-Recollets" "Esquirol-J.Guesde". Les trajets longs pour le deuxième groupe et plutôt courts sur Matabiau-Bonrepos se retrouvent sur la partie positive de l'axe 2 (haut de la figure). Les trajets ayant la caractéristique inverse sont les trajets de la partie négative de l'axe 2 (sur la figure 3.11, la classe 4). Pour préciser cette interprétation peu satisfaisante, une classification utilisant la même distance a été effectuée. On obtient le classement donné figure 3.11. La répartition des durées de trajets par classe est donnée figure 3.12 (à gauche). On voit que la classification est quasiment unidimensionnelle (excepté les recouvrements des classes 3, 4 et 5). Dans la partie droite, on donne les heures de départ des bus en abscisse et le numéro de classe en ordonnée. Le rapprochement de ces deux derniers graphes permet d'y voir plus clair. La classe 1 est la classe fluide représentative des bus circulant tôt le matin. La classe 2 est encore assez rapide. C'est une classe du matin (en tout cas, avant 14h30), mais qui ne contient ni la classe rapide précédente, ni la période

3. On montre qu'il s'agit de la représentation des individus avec la métrique de *Mahalanobis*, métrique définie par l'inverse de la métrique de variance-covariance empirique

4. Source : Agence d'Urbanisme de l'Agglomération Toulousaine pour la SEMVAT.

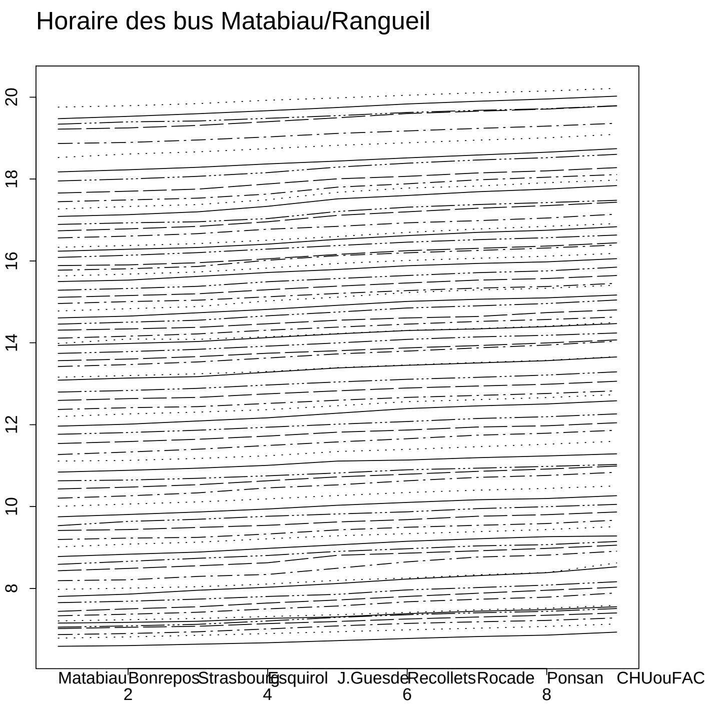


FIG. 3.9 - *Les heures de passage des bus aux différents arrêts.*

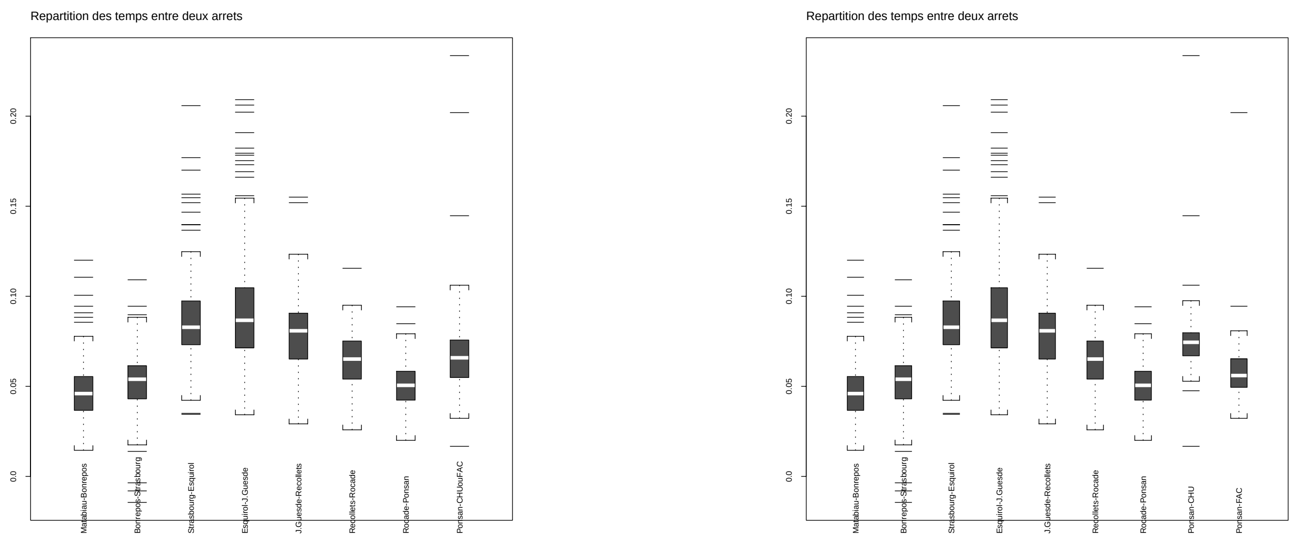


FIG. 3.10 - *Répartition des durées des trajets (Fac et CHU confondus, puis Fac et CHU séparés).*

8h-9h. La classe 3 est plutôt moyenne, et se répartit tout au long de la journée. On la décrit succinctement comme le complémentaire des autres classes (tous les horaires de 7h15 à 19h30, mais une représentation peu fournie de 9h à 13h et surtout de 15h30 à 19h). La classe 4 est assez lente et concerne surtout quelques bus entre 8h et 9h. Ces bus sont particulièrement lents sur le trajet Matabiau-Bonrepos. Enfin la classe 5 est la classe la plus lente, qui regroupe les bus du soir (15h30-19h).

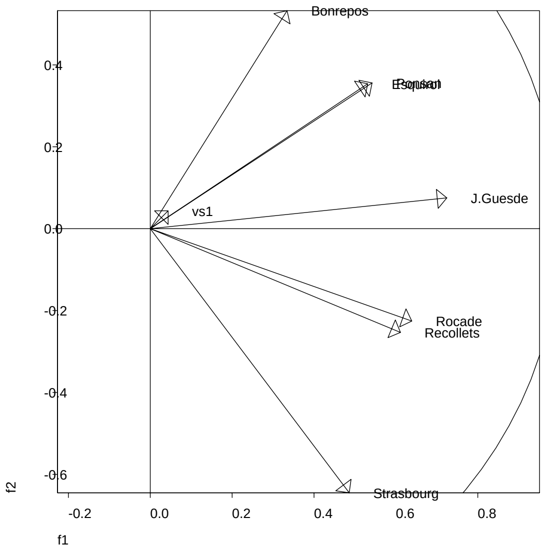


FIG. 3.11 - ACP *réduite : biplot avec représentation des classes (plan 1-2)*

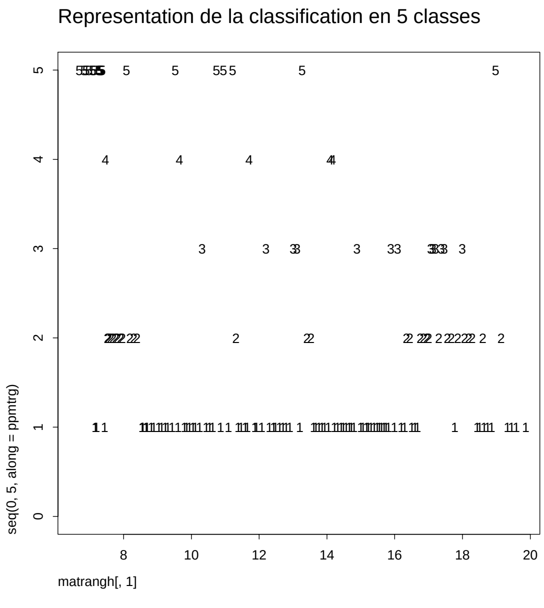


FIG. 3.12 - *Durées de parcours dans les classes et horaire de départ des bus par classe (5 classes)*

Contrôle des contributions Les contributions des variables aux facteurs sont données dans le Tableau 3.12. Elles montrent qu'un seul parcours, Matabiau-Bonrepos, contribue à près des 3/4 de la variance du second facteur. Ce résultat pose problème. En effet l'axe 2 est "du" à la seule variable Matabiau-Bonrepos, et plus particulièrement aux valeurs qu'elle prend sur les individus de la classe 4. Après discussion avec les initiateurs de cette campagne de mesure, il ressort que le trajet Matabiau-Bonrepos est particulièrement court, et sa durée dépend essentiellement de la couleur d'un feu. Cette variable jouerait un faible rôle si les variables n'étaient pas réduites. Mais la réduction des variables donne aux petits écarts de durée observés sur cette variable une grande importance. Il est donc justifié de recommencer l'analyse en supprimant cette variable ou en ne réduisant pas les variables. C'est le deuxième choix que nous mettons en œuvre dans une deuxième étape.

	f1	f2	f3	f4	f5	f>5	TOT	VARI
Matabiau-Bonrepos	75	736	104	2	4	41	143	95
Bonrepos-Strasbourg	111	0	229	579	10	38	143	87
Strasbourg-Esquirol	134	95	367	1	239	79	143	219
Esquirol-J.Guesde	183	20	45	60	1	368	143	124
J.Guesde-Recollets	167	27	125	184	52	209	143	280
Recollets-Rocade	165	62	119	165	22	223	143	121
Rocade-Ponsan	166	60	12	10	672	42	143	74
Total	1000	1000	1000	1000	1000	1000	1000	1000

TAB. 3.12 - *Contribution des variables aux facteurs et à l'inertie totale*

On obtient le biplot 3.13 :

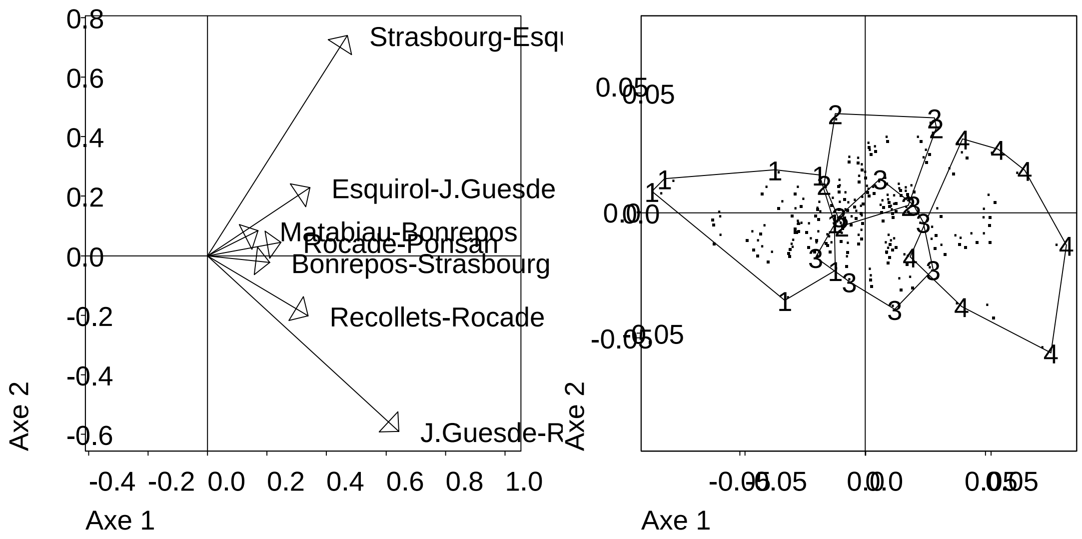


FIG. 3.13 - ACP *non réduite : biplot avec représentation des classes (plan 1-2)*

Les nouvelles contributions des variables aux facteurs sont données tableau 3.13. On voit que les contributions des variables au premier facteur sont le reflet de leur variance : le premier facteur traduit surtout la durée des trajets au centre ville. Le second facteur oppose deux trajet particulier : Strasbourg-Esquirol et J.Guesde-Recollets, ce qui peut-être du à des gênes de types différents (marché versus rentrée universitaire ?). Ici, seul un spécialiste ou un utilisateur attentif pourrait expliquer la fait qu' à une heure donnée un parcours lent sur un tronçon devient plus rapide sur un autre tronçon. Cette opposition correspond aux classes 2 et 3. Les bus de la classe 3 (7h30-8h30, 11h30-12h, vers 14h) sont plus rapides sur Strasbourg-Esquirol que sur J. Guesde-Recollet, et les bus de la classe 2 (9h-11h, 13h30-16h sauf 14h, et 17-19h (sauf 18h) ont la caractéristique inverse. La classe 1 regroupe les bus rapides (avant 7h45, et une partie des trajets en matinée). La classe 4 les bus lents (surtout entre 16 et 18h). Cette dernière analyse est plus satisfaisante, ne mettant plus en relief des artefacts. Cependant, en interprétant 2 axes, on n'a expliqué que 65% de l'inertie totale (cf. l'éboulis des valeurs propres, figure 3.15). Il serait sans doute possible de pousser plus loin l'interprétation.

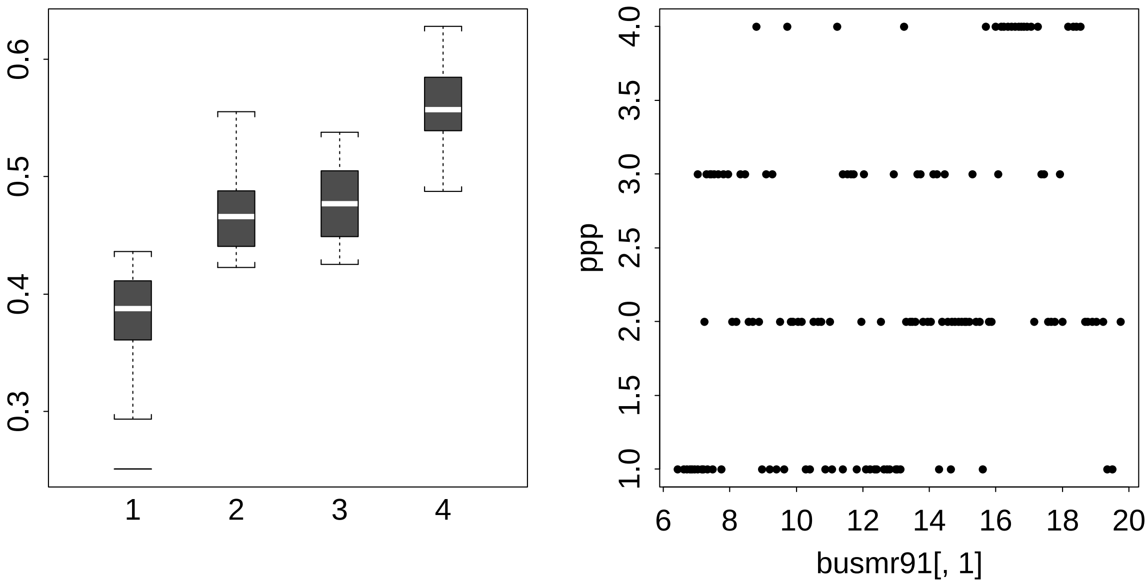


FIG. 3.14 - *Durées de parcours dans les classes et horaire de départ des bus par classe (4 classes)*

	f1	f2	f3	f4	f5	f>5	TOT	VARI
Matabiau-Bonrepos	29	7	301	420	144	55	95	95
Bonrepos-Strasbourg	44	1	93	228	433	111	87	87
Strasbourg-Esquirol	221	549	146	34	2	27	219	219
Esquirol-J.Guesde	118	53	131	246	94	182	124	124
J.Guesde-Recollets	413	348	162	5	20	27	280	280
Recollets-Rocade	114	41	159	42	301	189	121	121
Rocade-Ponsan	61	2	9	25	6	408	74	74
Total	1000	1000	1000	1000	1000	1000	1000	1000

TAB. 3.13 - ACP *non réduite : contributions des variables aux facteurs*

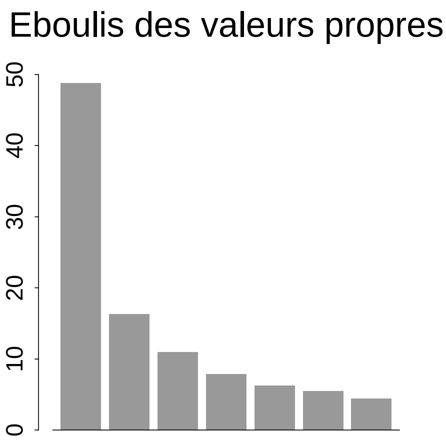
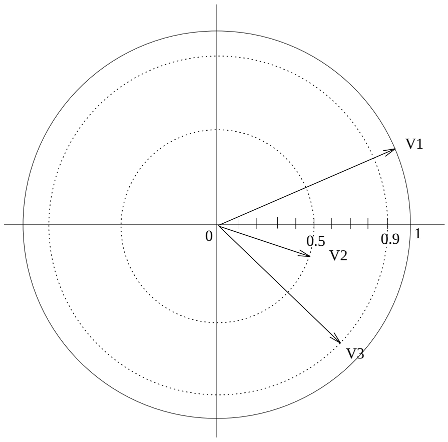
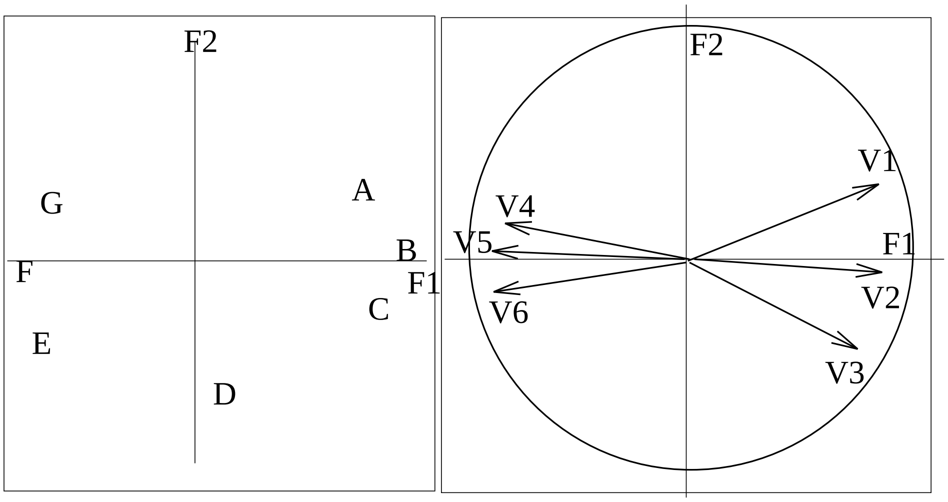


FIG. 3.15 - *BUS,* ACP *non réduite - Eboulis des valeurs propres*

Exercice 3.9.1 *(Extrait de l'examen de Septembre 1996)*
On représente 3 variables dans un cercle de corrélation. Quelle est la proportion de variance expliquée de chacune de ces variables? Que peux-t'on dire des corrélations existantes entre ces variables? (soyez précis).



Exercice 3.9.2 *(Extrait de l'examen de Septembre 1996)*On donne les représentations isométriques des individus et des variables dans une acp. Interprétez le premier facteur, le second. Que peux-t'on dire des valeurs des variables sur les individus?



Problème d'examen : choix de métrique en analyse en composantes principales

Examen du magistère d'économiste statisticien, Toulouse, 1995, partie théorique.

Les données et le problème

L'Agence d'Urbanisme de l'Agglomération Toulousaine a mesuré pour la SEMVAT les horaires de passage du bus no 2 à 9 carrefours, et ceci pendant une journée complète du printemps 1991 et une journée "équivalente" du printemps 93. Ces bus vont de la gare Matabiau à Rangueil —Facultés ou Centre Hospitalo-Universitaire (C.H.U.)— et de Rangueil —Facultés ou C.H.U.— à la gare Matabiau (à l'extrémité sud, le trajet se dédouble en deux branches). On décide de travailler sur les données de 1993 et, pour rendre les données homogènes, on supprime les données relatives à l'extrémité dédoublée. On calcule à partir des horaires de passage les durées des trajets sur chacun des tronçons (intervalle entre deux carrefours où ont été effectuées des mesures), dans un sens et dans l'autre. On veut étudier la fluidité de la circulation en fonction du tronçon et de l'heure de départ et pour ce faire, on effectue l'acp du tableau $Y = (y_i^j)$ contenant les durées de parcours sur les tronçons j pour les rotations i (une rotation i concerne donc deux bus partant à peu près à la même heure des deux extrémités du parcours). Les longueurs en mètres des 14 tronçons sont les suivantes :

--------	MatBon	BonStra	StraEsq	EsqJuG	JuGRec	RecRoc	RocIut	IutRoc	RocRec	RecJuG	JuGEsq	EsqStra	StraBon	BonMat
MatBon	2.476	0.67	0.65	1.10	0.231	0.404	0.2810	-0.190	0.10	-0.113	-0.108	-0.025	-0.342	-0.0411
BonStra	0.665	2.42	0.24	1.28	1.002	0.709	0.2260	-0.143	0.30	-0.104	-0.050	0.347	-0.478	0.2275
StraEsq	0.649	0.24	6.98	3.13	0.980	0.598	0.6806	0.809	-0.23	0.376	-0.357	0.608	0.589	0.2201
EsqJuG	1.100	1.28	3.13	12.27	1.974	1.612	1.3014	0.824	0.36	0.872	-0.186	0.124	-0.674	-0.5026
JuGRec	0.231	1.00	0.98	1.97	3.780	0.814	0.4256	0.320	0.36	-0.087	0.469	-0.065	0.317	0.2803
RecRoc	0.404	0.71	0.60	1.61	0.814	2.107	0.3563	0.179	0.48	0.141	0.421	0.119	-0.086	0.0594
RocIut	0.281	0.23	0.68	1.30	0.426	0.356	1.4784	0.238	0.38	0.151	-0.213	-0.082	0.231	-0.0053
IutRoc	-0.190	-0.14	0.81	0.82	0.320	0.179	0.2383	1.684	0.24	0.571	0.030	0.132	0.118	-0.0692
RocRec	0.104	0.30	-0.23	0.36	0.358	0.476	0.3766	0.245	3.50	0.992	1.076	0.531	0.204	0.2471
RecJuG	-0.113	-0.10	0.38	0.87	-0.087	0.141	0.1509	0.571	0.99	3.094	0.322	-0.041	0.517	-0.1445
JuGEsq	-0.108	-0.05	-0.36	-0.19	0.469	0.421	-0.2129	0.030	1.08	0.322	3.032	0.319	0.604	0.0494
EsqStra	-0.025	0.35	0.61	0.12	-0.065	0.119	-0.0821	0.132	0.53	-0.041	0.319	2.961	0.764	0.0294
StraBon	-0.342	-0.48	0.59	-0.67	0.317	-0.086	0.2310	0.118	0.20	0.517	0.604	0.764	4.153	0.2420
BonMat	-0.041	0.23	0.22	-0.50	0.280	0.059	-0.0053	-0.069	0.25	-0.144	0.049	0.029	0.242	2.0751

TAB. 3.14 - *Matrice des variances-covariances*

Matabiau-Bonrepos	Bonrepos-Strasbourg	Strasbourg-Esquirol	Esquirol-J.Guesde
480	610	850	875
J.Guesde-Recollets	Recollets-Rocade	Rocade-Ponsan	
970	1320	970	
Ponsan-Rocade	Rocade-Recollets	Recollets-J.Guesde	J.Guesde-Esquirol
970	1320	970	875
Esquirol-Strasbourg	Strasbourg-Bonrepos	Bonrepos-Matabiau	
850	610	90	

Le problème posé est celui du codage et de la métrique utilisée.

1. Considérations pratiques (3pts)

   (a) (1.5 pts) Si on effectue une acp réduite avec métrique identité, que vaut l'inertie totale ? Quelle est la contribution de chaque variable à l'inertie totale ? Quel avantage et inconvénient présente ce codage ? (Comparer les contributions à l'importance des tronçons mesurées par leur longueur).

   (b) (1pt) On effectue une acp non réduite avec métrique identité. On donne Figure 3.16 la représentation par des boites à pattes des distributions des variables. On donne aussi Tableau 3.14 la matrice de variance-covariance (multipliée par 10^4) et la valeur de l'inertie totale ($52.002\ 10^{-4}$). Quelle est la contribution relative des 2 variables les plus influentes à l'inertie totale ?

   (c) (0.5pt) Ce dernier codage semble t'il donner plus d'importance aux trajets courts ou longs ? (comparer brièvement avec le codage précédent).

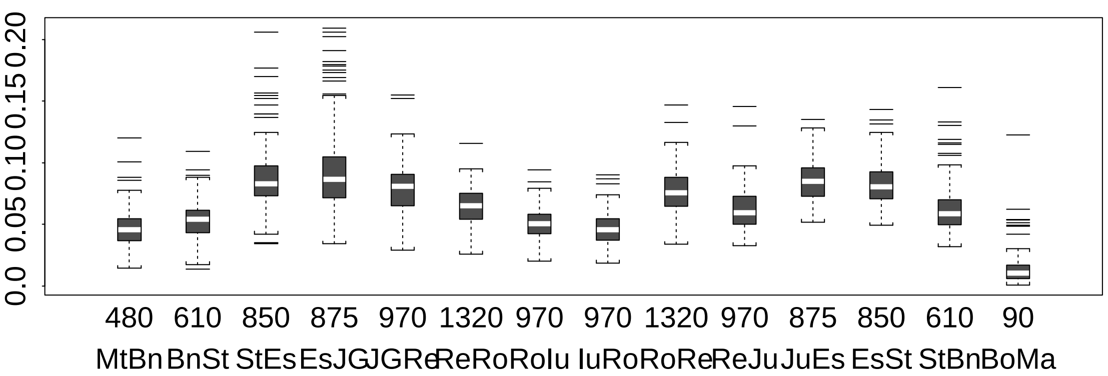


FIG. 3.16 - *Boxplots des durées de trajet et longueurs des trajets*

2. Problème théorique (9pts) : on souhaite trouver une métrique qui donne à chaque tronçon une importance qui dépende de façon "raisonnable" de la longueur du tronçon.

   (a) (1.5pt) Dans un premier temps, on utilise la métrique identité, et on suppose que deux tronçons successifs, indicés par j_0 et j_1 ont les mêmes caractéristiques (longueur, encombrement,...) et sont parcourus par chaque bus globalement à la même vitesse. Par conséquent, les variables "durée de parcours" $\mathbf{x}^{j_0}$ et $\mathbf{x}^{j_1}$ sont égales. Dans une acp non-réduite, on note $C(j_0) + C(j_1)$ la contribution de ces deux variables à l'inertie totale.

On réunit ces deux tronçons en un seul, et la nouvelle variable $\mathbf{x}^j$ est donc la somme des deux précédentes. Dans une nouvelle acp non réduite, quelle est, en fonction de $C(j_0)+C(j_1)$, la contribution $C(j)$ de cette nouvelle variable à l'inertie totale? Ceci confirme t'il la réponse donnée à la question 1c?

(b) (5.5pts) On suppose qu'un trajet homogène de longueur l_j est découpé en deux parties de longueurs l_{j0} et l_{j1} et on note α_0 et α_1 deux nombres positifs tels que $l_{j0} = \alpha_0 l_j$ et $l_{j1} = \alpha_1 l_j$ (on a donc $\alpha_0 + \alpha_1 = 1$). La variable $\mathbf{x}^j$ donnant la durée de parcours sur le tronçon j, l'homogénéité du trajet implique que la durée des parcours des tronçons j_0 et j_1 est respectivement $\mathbf{x}^{j0} = \alpha_0 \mathbf{x}^j$ et $\mathbf{x}^{j1} = \alpha_1 \mathbf{x}^j$. On cherche à faire une acp non-réduite avec une métrique $M = diag(m_j, j = 1,p)$ qui ait comme propriété de ne pas modifier les résultats de l'acp si on regroupe des variables proportionnelles. Ainsi on ne fait pas dépendre les contribution des variables du découpage du trajet : l'expérimentateur peut définir son découpage en recherchant des segments homogènes, sans se soucier de leurs longueurs.

On note m_j l'élément (j,j) de la métrique M avant découpage et m_{j_0} et m_{j_1} les éléments (j_0, j_0) et (j_1, j_1) de la nouvelle matrice M après découpage. Les autres éléments de M et les autres variables sont inchangées. Dans la suite, on suppose que les variables ont été centrées.

i. L'espace des variables $\mathbf{R}^n$ étant muni de la métrique des poids, donnez le critère de l'acp caractérisant les sous-espaces principaux W_k de l'espace des variables $(dim(W_k) = k)$ par l'intermédiaire des opérateurs de projection orthogonale dans $\mathbf{R}^n$ d'image W_k (on notera un tel projecteur P_k). On exprimera le critère en fonction du tableau X des données centrées, puis comme une somme de termes dus à chaque variable (c'est cette dernière forme qui sera utilisée par la suite).

ii. On suppose que la métrique M est définie par l'intermédiaire d'une fonction f définie sur $\mathbf{R}^n$et à valeurs positives par

$$m_j = f(\mathbf{x}^j)$$

(l'élément m_j de la métrique M est une fonction de la variable $\mathbf{x}^j$ de même indice). Montrer qu'une condition de l'invariance du critère par découpage du tronçon j s'écrit :

$$(\alpha_0)^2 f(\alpha_0 \mathbf{x}^j) + (\alpha_1)^2 f(\alpha_1 \mathbf{x}^j) = f(\mathbf{x}^j).$$

Comme cette propriété doit être vraie quelque soit la variable $\mathbf{x}^j$ considérée, et quelque soit le découpage, elle s'écrit encore :

$$(\alpha_0)^2 f(\alpha_0 \mathbf{x}) + (\alpha_1)^2 f(\alpha_1 \mathbf{x}) = f(\mathbf{x}) \;\; \forall \mathbf{x} \in \mathbf{R}^n, \;\; \forall \alpha_0 > 0, \;\; \forall \alpha 1 > 0 \;\; \alpha_0 + \alpha_1 = 1.$$

iii. Montrer que les fonctions réelles positives f définie sur $\mathbf{R}^n$ telles que $f(\alpha \mathbf{x}) = (1/\alpha) f(\mathbf{x})$ $(\alpha > 0)$ sont solutions (on peut montrer que c'est une condition nécessaire et suffisante, sous des conditions de continuité, mais cela n'est pas demandé). En particulier vérifier qu'on peut prendre $f(\mathbf{x}) = 1/\sigma(\mathbf{x})$ avec $\sigma_{\mathbf{x}}$ écart-type de $\mathbf{x}$. On appelle une telle acp "acp semi-réduite".

iv. Quelles sont les invariances que l'on peut déduire de l'invariance du critère de l'acp?

(c) (2pts) On donne une représentation d'un biplot issu de l'acp semi-réduite de ce tableau, après une répartition des individus (ou rotations) en classes. Et on donne aussi la répartition des horaires de départ par classe. Effectuer une brève interprétation de ces graphes.

Exercice 3.9.3 *(Examen de Septembre 1996 : extrait)*Dans l'acp sur les vins de Cahors (où on dispose de mesures effectuées sur des vins de différents millésimes, avec différents porte-greffes et différents clones) on a représenté les barycentres d'interaction *année* $\times$ *porte-greffe*, ainsi que les barycentres des années (voir exemple ci-contre). Quelle propriété vérifie un point "année" par rapport aux barycentres d'interaction *année* $\times$ *porte-greffe* faisant intervenir cette même année? (démontrer brièvement votre réponse).

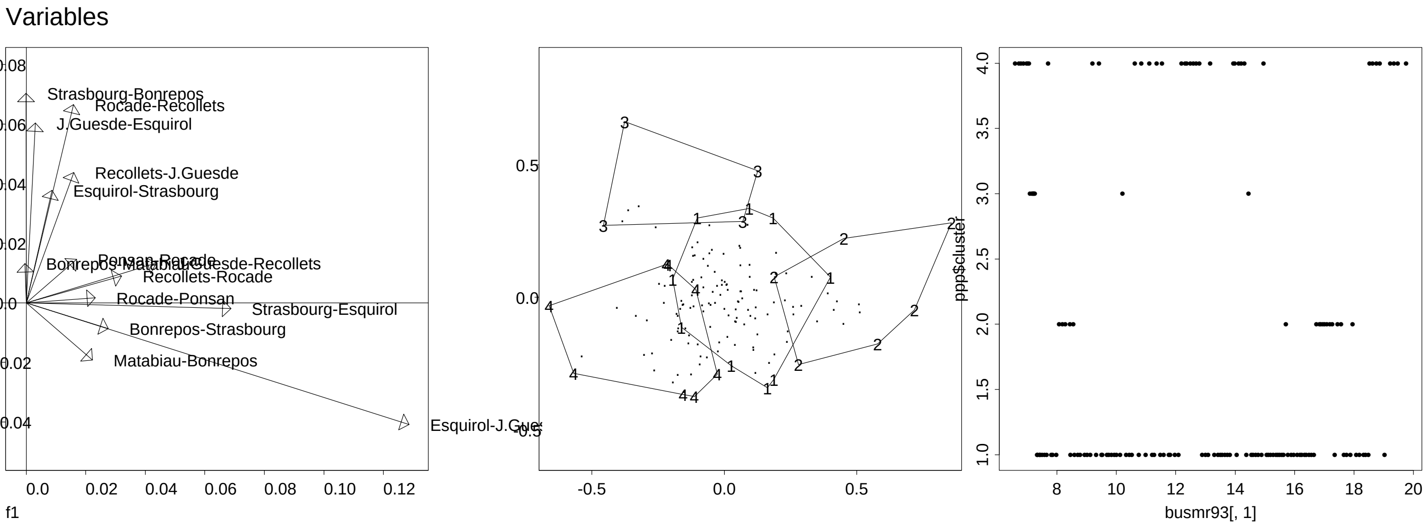
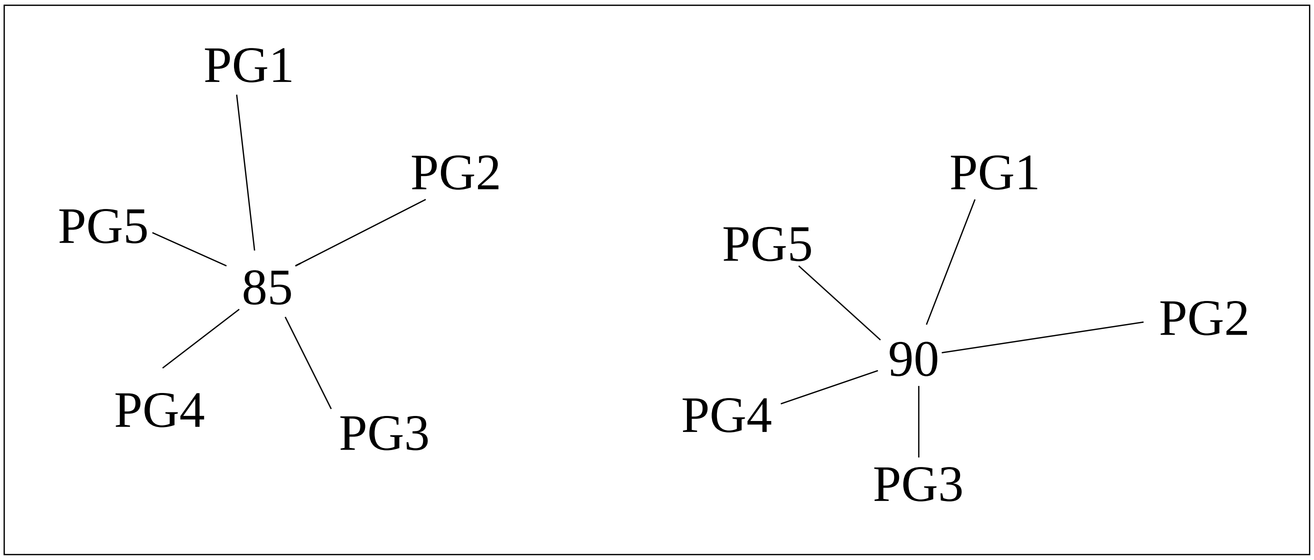


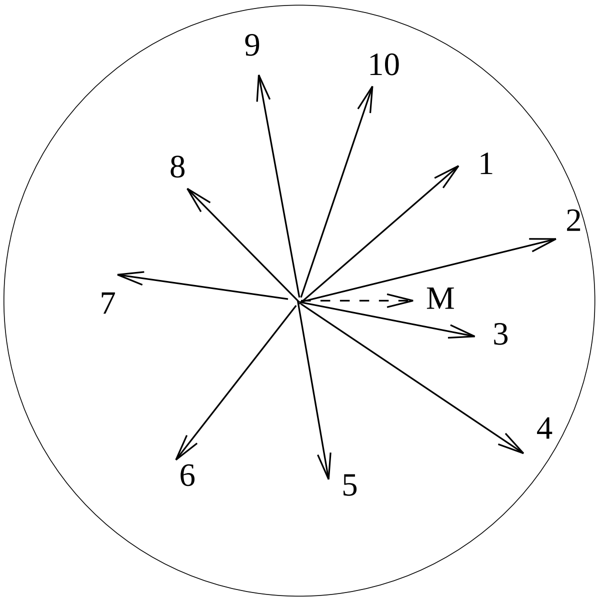
FIG. 3.17 - *Biplot isométrique-ligne avec classes d'individus et distribution des heures de départ par classe*



Exercice 3.9.4 *(Examen de Septembre 1996 : extrait)*On fait une acp avec métrique identité et variables réduites et une autre avec la métrique $M = diag(1/(\sigma_1)^2, \ldots, 1/(\sigma_j)^2, \ldots, 1/(\sigma_p)^2)$ et sur variables non réduites. Quelle est l'inertie totale de ces deux analyses. Que peux-t'on dire de la représentation des individus de ces deux analyses? (on justifiera mathématiquement la réponse).

Exercice 3.9.5 *(Examen de Septembre 1996 : extrait)* On a un tableau de notes d'élèves que l'on soumet à une analyse des correspondances. Quels inconvénients peut avoir une telle pratique? (Soyez précis et donner deux ou trois exemples).

Question 3.9.1 *(Examen de Septembre 1996 : extrait)*Dix dégustateurs de 30 vins différents donnent pour chaque vin une note comprise entre 0 et 20. On affecte ensuite à chaque vin sa note moyenne, qui représente donc une mesure de la qualité générale du vin. On fait une acp du tableau des notes, et on rajoute la variable moyenne (M) en supplémentaire. On obtient le cercle des corrélations ci-contre. Quels commentaires pourriez vous faire sur les dégustateurs et sur la note de qualité moyenne du vin à l'organisme viticole qui a mené cette dégustation?



Seconde partie : interprétation d'une analyse en composantes principales

(10 points ; dans les interprétations, on restera concis avec le souci d'aller à l'essentiel)

On considère le tableau[5] donnant des mesures effectuées sur du vin jeune, provenant de parcelles de vignobles plantées avec 5 types de porte-greffe et, pour chaque porte-greffe, 6 clones différents. Les vendanges sur lesquelles portent ces mesures on été effectuées en 1983, 1985, 1988, 1989 et 1990. On obtient ainsi un tableau ayant $5 \times 6 \times 5$ lignes et 13 colonnes, trois fixant le niveau des facteurs contrôlés, et 9 donnant les valeurs des mesures effectuées. Ces mesures sont

Anth:	Anthocyanes (en mg/l)
D520:	DO 520 = indice de réfraction de la couleur rouge
D420:	DO 420 = indice de réfraction de la couleur jaune
I.C:	Intensité Colorante = DO 520 + DO 420
Nuance:	Nuance = DO 420/DO 520
I.Folin:	Indice de Folin
Degré:	Degré alcoolique du vin
Ext.sec:	Extrait sec (en g/l)
Rdt:	Rendement (moyenne au pied en Kg/souche)
A.T:	Acidité totale
A.V:	Acidité volatile

1. D'après la représentation de l'éboulis des valeurs propres, (figure 3.18), combien de facteurs semble-t'il nécessaire d'interpréter ?
2. Donner une **brève** interprétation des 3 premiers facteurs (on trouvera les coefficients de qualité et les contributions des variables en fin d'énoncé).
3. Après examen de la représentation des individus, on effectue des représentations isométriques du nuage des individus sur les plans 1-2 et 1-3, en entourant les points représentant une même année (figure 3.21). Caractériser de façon synthétique les années à partir des variables mesurées.
4. On donne (figure 3.22) aussi les barycentres des années, et les barycentres des groupes d'individus définis par la valeur des facteurs année et clone (on les appellera les *barycentres d'interaction*). L'écart-type de la variable degré étant $0,574$, quelle est l'approximation de la différence moyenne de degré observée entre les années 1983 et 1990 (on précisera le calcul effectué). Comparer avec la différence observée entre 1983 et 1985.
5. On examine l'interaction année $\times$ clone. L'effet du clone vous parait-il important par rapport à l'effet année? Est-il possible de caractériser les clones indépendamment des années ? Lorsque cela semble possible, effectuer cette caractérisation.

5. SOURCE: Institut Technique du Vin (Gaillac)

6. On examine maintenant l'interaction année $\times$ porte-greffe (figure 3.23). L'effet du porte greffe vous parait-il important ? En examinant les représentations sur les plans 1-2 et 1-3, qu'est ce qui différencie les porte-greffes lors de l'année 1985 ?

7. Si on arrête l'interprétation au troisième facteur, quelle précaution oratoire faut-il prendre ?

Carrés des corrélations variables facteurs

	f1	f2	f3	f4	f5	f>5	TOT
Anth	818	70	4	1	4	103	1000
D520	742	195	20	0	8	35	1000
D420	669	286	3	0	3	40	1000
I.C	728	240	12	0	4	16	1000
Nuance	61	158	413	35	115	217	1000
I.Folin	110	514	0	69	1	306	1000
Degré	437	223	33	85	0	221	1000
Ext.sec	121	638	4	96	3	139	1000
Rdt	96	29	551	67	80	177	1000
A.T	44	326	137	401	24	68	1000
A.V	82	129	285	32	455	18	1000
TOT	355	255	133	71	63	122	1000

Carrés des corrélations multiples variables sous-espaces principaux

	f1	f1a2	f1a3	f1a4	f1a5	f>5	TOT
Anth	818	889	893	894	897	1000	1000
D520	742	938	957	958	965	1000	1000
D420	669	954	957	957	960	1000	1000
I.C	728	968	980	980	984	1000	1000
Nuance	61	219	632	667	783	1000	1000
I.Folin	110	624	625	693	694	1000	1000
Degré	437	660	693	778	779	1000	1000
Ext.sec	121	759	762	858	861	1000	1000
Rdt	96	125	676	743	823	1000	1000
A.T	44	370	507	908	932	1000	1000
A.V	82	210	496	527	982	1000	1000
TOT	355	611	744	815	878	1000	1000

Contribution des variables aux facteurs

	f1	f2	f3	f4	f5	f>5	TOT	VARI
Anth	209	25	3	1	5	77	91	994
D520	190	70	13	0	11	26	91	0
D420	171	102	2	1	4	30	91	0
I.C	186	85	8	0	6	12	91	0
Nuance	16	56	283	44	166	162	91	0
I.Folin	28	183	0	88	1	228	91	5
Degré	112	79	23	108	1	165	91	0
Ext.sec	31	227	3	122	4	104	91	1
Rdt	25	10	377	85	114	132	91	0
A.T	11	116	94	511	35	51	91	0
A.V	21	46	195	40	654	13	91	0
	1000	1000	1000	1000	1000	1000	1000	1000

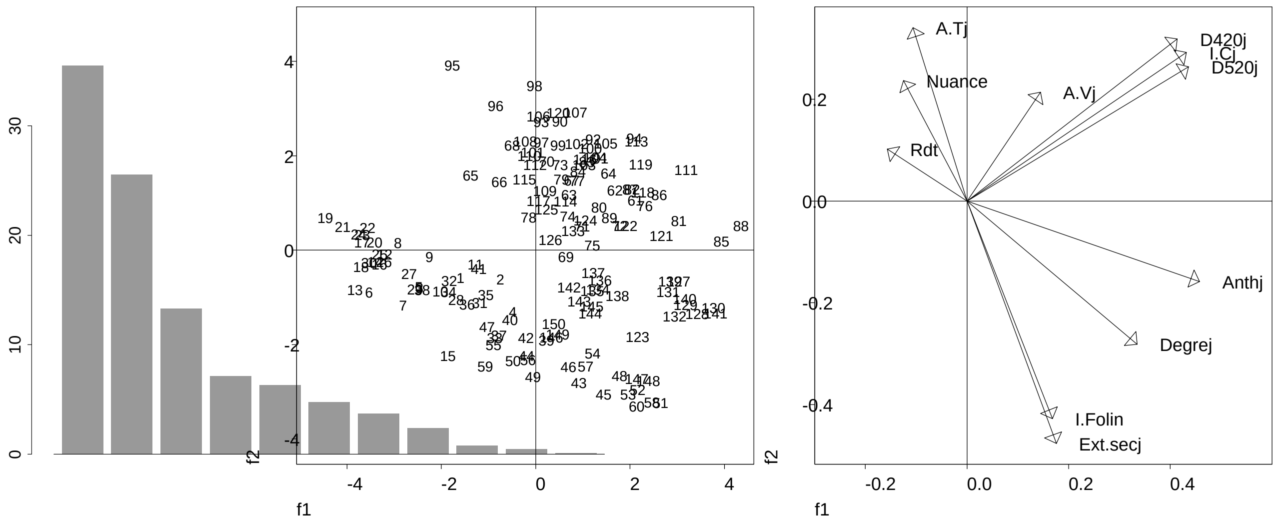


FIG. 3.18 - *Biplot isométrique-ligne (plan 1-2)*

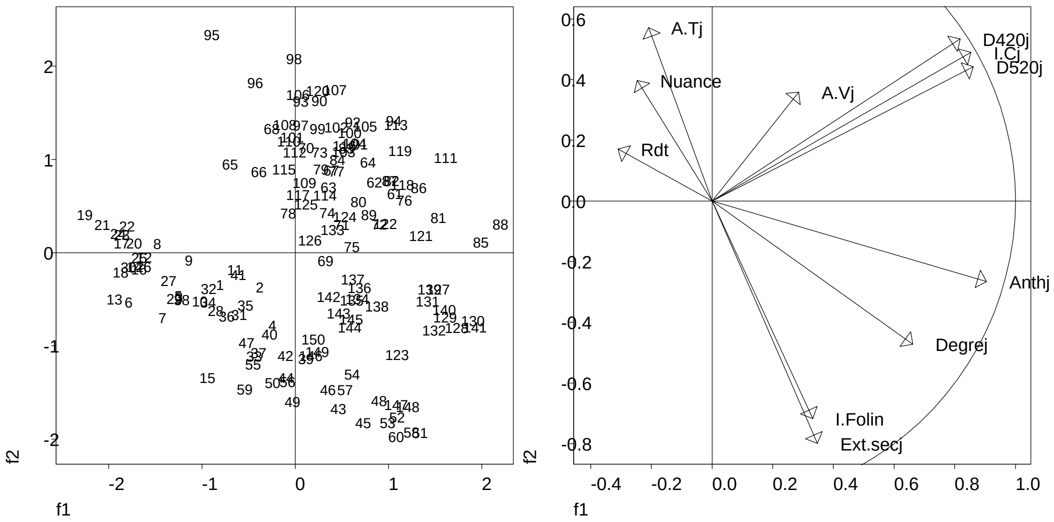


FIG. 3.19 - *Biplot isométrique-colonne (plan 1-2)*

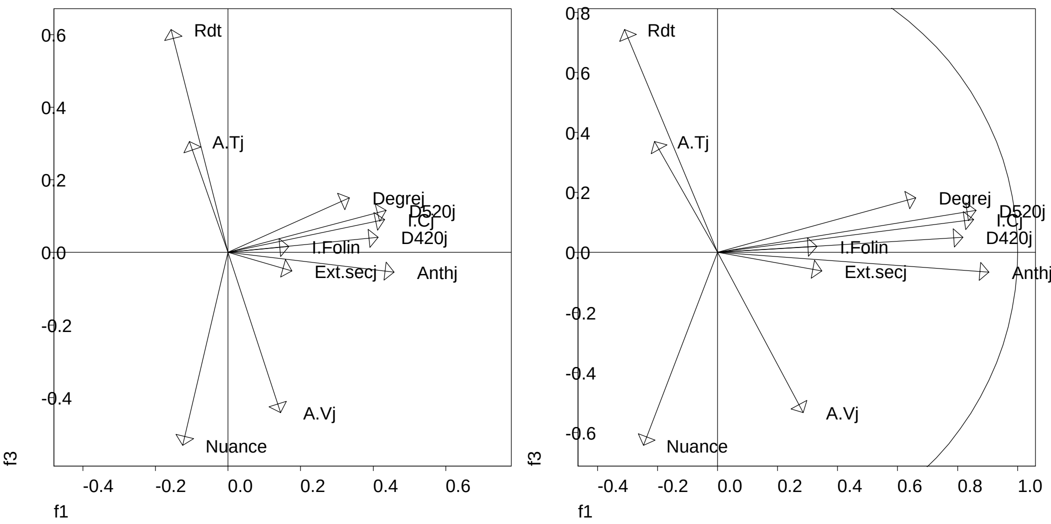


FIG. 3.20 - *Représentation des variables (1-3) Non-isométrique et isométrique*

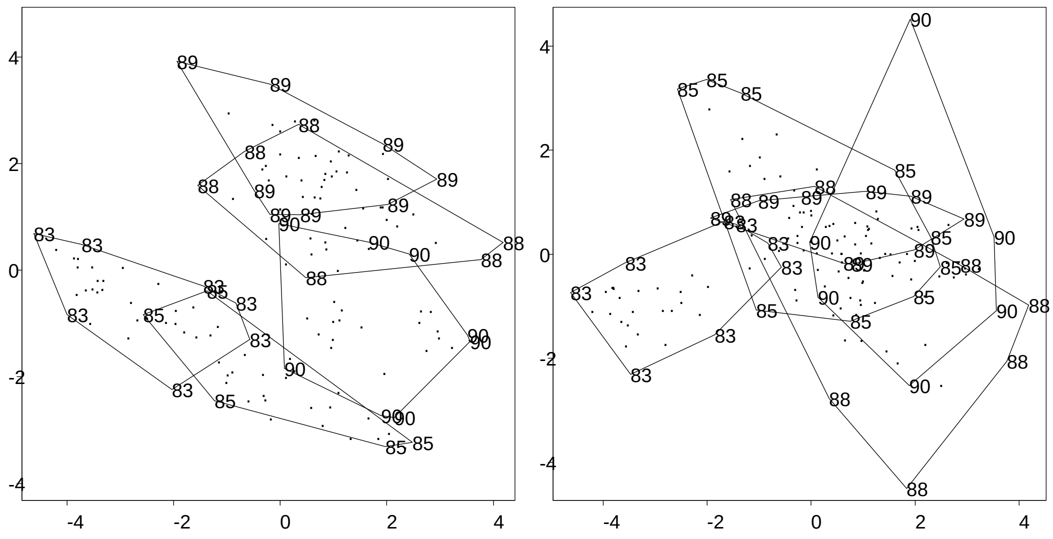


FIG. 3.21 - *Iso-ligne (1-2) et (1-3) : représentation des classes*

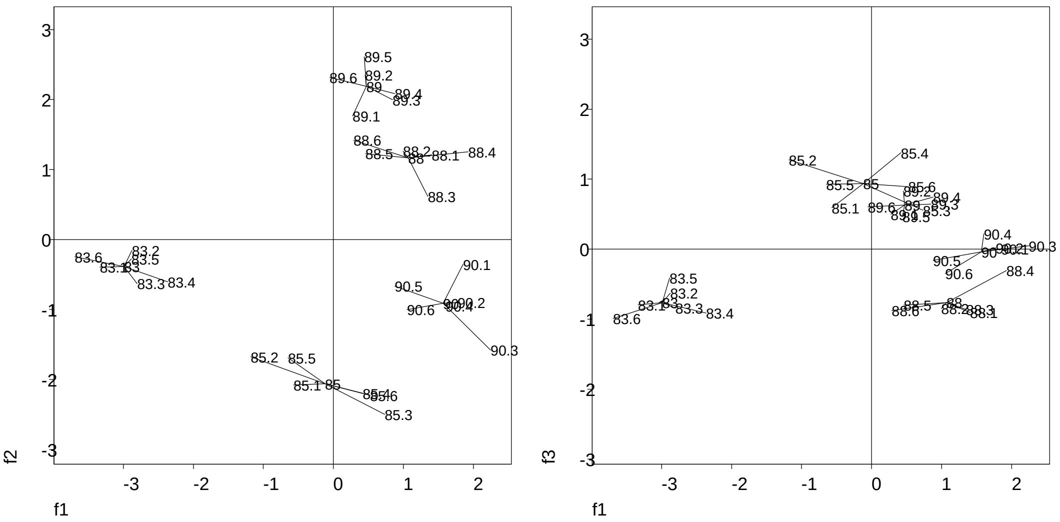


FIG. 3.22 - *Iso-ligne (1-2) et (1-3) : représentation des l'interaction Année-Clone*

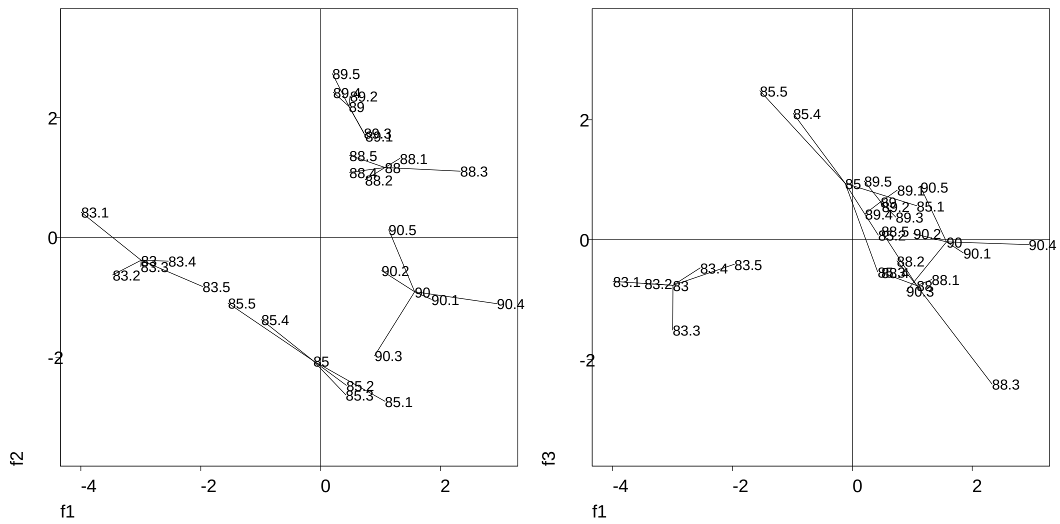


FIG. 3.23 - *Iso-ligne (1-2) et (1-3) : représentation des l'interaction Année-Porte-Greffe*

Chapitre 4

Le traitement des données qualitatives

4.1 Présentation des données

Un tableau de données qualitatives (ou *catégorielles*) se présente sous la forme :

- d'un tableau individus variables dans lequel les variables sont codées par des entiers "de 1 à k" (k = nombre de modalités) ou en alphanumérique (par exemple, pour la variable SEXE, "m" et "f"),
- d'un *tableau de contingence* (cas de deux variables),
- ou d'un *tableau de Burt* (cas de plus de deux variables).

Tableau individus variables

C'est un tableau présenté sous la forme usuelle, ayant à l'intersection de la ligne i et de la colonne j, la catégorie prise par la variable j pour l'individu i. On donne, Tableau 4.1, une partie d'un tableau individus x variables[1], donnant pour 27 races de chien 7 caractéristiques:

```
taille (-, +, ++),
poids (-, +, ++),
vélocité (-, +, ++),
intelligence (-, +, ++),
affection (-, +),
agressivité (-, +),
fonction (compagnie, chasse, garde)
```

Tableau d'indicatrices

C'est un tableau constitué en colonnes des indicatrices des modalités de p variables qualitatives. Il a autant de lignes que d'individus et un nombre de colonnes égal au nombre total de modalités de l'ensemble des variables. Il a la propriété d'avoir comme somme des éléments d'une quelconque de ses lignes le nombre p de variables qualitatives. On donne tableau 4.2 un exemple de tableau d'indicatrices.

Tableau de contingence

On peut résumer la liaison entre deux variables par un tableau de contingence. Ainsi la liaison `Taille x Poids` est décrite dans le tableau 4.3 qui montre qu'il y a 7 races de faible taille et de faible poids, 10 de grande taille et de poids moyen, etc..
Un exemple de table de contingence d'assez grandes dimensions est donné Tableau 4.5.

1. Données communiquées par M. Tennenhaus à G. Saporta et se trouvant dans Saporta (Probabilités, Analyse des Données et Sta- tistique, Technip 1990)

	Tail	Poid	Velo	Inte	Affe	Agre	Fonc
Bcrn	3	2	3	3	2	2	3
Bsst	1	1	1	1	1	2	2
BrgA	3	2	3	3	2	2	3
Boxr	2	2	2	2	2	2	1
Bldg	1	1	1	2	2	1	1
BlMs	3	3	1	3	1	2	3
Cnch	1	1	2	3	2	1	1
Chhh	1	1	1	1	2	1	1
...	..	..	..	..	..	..	..
...	..	..	..	..	..	..	..
Pkns	1	1	1	1	2	1	1
Pntr	3	2	3	3	1	1	2
StBn	3	3	1	2	1	2	3
Sttr	3	2	3	2	1	1	2
Tckl	1	1	1	2	2	1	1
TrNv	3	3	1	2	1	1	3

TAB. 4.1 - *Un exemple de tableau individus x variables*

	Tail1	Tail2	Tail3	Poid1	Poid2	Poid3	Velo1	Velo2	Velo3
Bcrn	0	0	1	0	1	0	0	0	1
Bsst	1	0	0	1	0	0	1	0	0
BrgA	0	0	1	0	1	0	0	0	1
Boxr	0	1	0	0	1	0	0	1	0
Bldg	1	0	0	1	0	0	1	0	0
BlMs	0	0	1	0	0	1	1	0	0
Cnch	1	0	0	1	0	0	0	1	0
Chhh	1	0	0	1	0	0	1	0	0
Cckr	0	1	0	1	0	0	1	0	0
Clly	0	0	1	0	1	0	0	0	1
Dlmt	0	1	0	0	1	0	0	1	0
Dbrm	0	0	1	0	1	0	0	0	1
...	...	...	...	...	...	...	...	...	...

TAB. 4.2 - *Un exemple de tableau de d'indicatrices pour trois variables qualitatives*

	Poid-	Poid+	Poid++
Tail-	7	0	0
Tail+	1	4	0
Tail++	0	10	5

TAB. 4.3 - *Un exemple de tableau de contingence*

Notations On désigne par N la table de contingence, I et J le nombre de lignes et de colonnes de N, $\mathbf{I} = \{1, \ldots, I\}$ et $\mathbf{J} = \{1, \ldots, J\}$ les ensembles indiçant les lignes et les colonnes de N, et par n_{ij} l'élément (i, j) de N. On note enfin

$$\begin{array}{ll} n_{.j} = \sum_{i \in \mathbf{I}} n_{ij} & n_{i.} = \sum_{j \in \mathbf{J}} n_{ij}, \\ n = \sum_{i \in \mathbf{I}} \sum_{j \in \mathbf{J}} n_{ij} & \\ f_{ij} = n_{ij}/n & F = (f_{ij})_{i \in \mathbf{I}, j \in \mathbf{J}} \end{array}$$

Tableau de Burt

C'est une "super matrice" dont l'élément (j, k) est le tableau de contingence croisant les variables qualitatives j et k. On donne tableau 4.4 le tableau de Burt croisant les 4 premières variables du tableau 4.1. Remarquons qu'un tableau de Burt ne permet d'analyser que les liaisons deux à deux entre les variables (on peut estimer la probabilité de l'évènement ($X = i$ et $Y = j$), mais pas celle de l'évènement ($X = i$ et $Y = j$ et $Z = k$))

	Tail1	Tail2	Tail3	Poid1	Poid2	Poid3	Velo1	Velo2	Velo3	Inte1	Inte2	Inte3
Tail1	7	0	0	7	0	0	5	2	0	3	3	1
Tail2	0	5	0	1	4	0	1	4	0	0	4	1
Tail3	0	0	15	0	10	5	4	2	9	5	5	5
Poid1	7	1	0	8	0	0	6	2	0	3	4	1
Poid2	0	4	10	0	14	0	0	6	8	3	6	5
Poid3	0	0	5	0	0	5	4	0	1	2	2	1
Velo1	5	1	4	6	0	4	10	0	0	4	5	1
Velo2	2	4	2	2	6	0	0	8	0	1	5	2
Velo3	0	0	9	0	8	1	0	0	9	3	2	4
Inte1	3	0	5	3	3	2	4	1	3	8	0	0
Inte2	3	4	5	4	6	2	5	5	2	0	12	0
Inte3	1	1	5	1	5	1	1	2	4	0	0	7

TAB. 4.4 - *Tableau de Burt sur les quatre premières variables*

Propriété On remarque que si Z est le tableau d'indicatrices de p variables, B le tableau de Burt de ces mêmes variables, alors $B = Z'Z$.

4.2 Notion de liaison entre variables qualitatives

Avant d'étudier l'analyse des correspondances qui nous permettra d'analyser les liaisons entre deux variables catégorielles, nous présentons la notion de liaison entre variables qualitatives. Cette notion utilise les mêmes notions probabilistes de "non indépendance" que celles utilisées dans le cadre quantitatif.

4.2.1 Définition

On dit que deux variables sont liées si la connaissance de la valeur de l'une d'entre elles pour un individu modifie la distribution attendue de l'autre.

Exemple: On donne (cf. Tableau 4.5) le tableau donnant la répartition des catégories socio-professionnelles (C.S.P.) des parents en fonction des études supérieures suivies par leurs enfants, ainsi que la signification des identificateurs de lignes et de colonnes:

	dro	sci	let	med	pha	iut
age	4814	8941	8719	1583	959	558
oua	463	1118	1224	153	33	83
ind	3610	2608	3365	1634	728	110
art	2870	5091	5966	1583	562	261
gco	2777	2608	3212	1736	794	136
pco	4444	6209	7954	2655	1125	313
lib	9813	7823	9637	8985	3143	120
ing	4444	8195	7648	3880	1307	151
adm	14071	13535	18050	7147	2134	235
mes	833	993	1071	817	182	21
mso	555	621	765	357	66	16
tec	2129	3974	4436	1021	298	235
cam	7776	10058	13308	3931	1125	407
ins	1666	5464	6425	1481	579	141
emb	5184	8071	10861	2399	529	0
con	1481	2856	3059	613	132	177
ous	2314	4098	4436	766	182	516
min	278	869	918	153	17	37
mar	185	373	306	51	0	21
man	741	1863	1989	255	33	83
ser	741	993	1377	204	33	63
arm	2592	3104	3671	817	198	136

TAB. 4.5 - *Origine socio-professionnelle des étudiants français: effectifs*

	dro	sci	let	med	pha	iut	Tot
age	18.8	35.0	34.1	6.2	3.7	2.2	100
oua	15.1	36.4	39.8	5.0	1.1	2.7	100
ind	29.9	21.6	27.9	13.6	6.0	0.9	100
art	17.6	31.2	36.5	9.7	3.4	1.6	100
gco	24.7	23.2	28.5	15.4	7.0	1.2	100
pco	19.6	27.4	35.0	11.7	5.0	1.4	100
lib	24.8	19.8	24.4	22.7	8.0	0.3	100
ing	17.3	32.0	29.8	15.1	5.1	0.6	100
adm	25.5	24.5	32.7	13.0	3.9	0.4	100
mes	21.3	25.4	27.3	20.9	4.6	0.5	100
mso	23.3	26.1	32.1	15.0	2.8	0.7	100
tec	17.6	32.9	36.7	8.4	2.5	1.9	100
cam	21.2	27.5	36.4	10.7	3.1	1.1	100
ins	10.6	34.7	40.8	9.4	3.7	0.9	100
emb	19.2	29.8	40.2	8.9	2.0	0.0	100
con	17.8	34.3	36.8	7.4	1.6	2.1	100
ous	18.8	33.3	36.0	6.2	1.5	4.2	100
min	12.2	38.2	40.4	6.7	0.7	1.6	100
mar	19.8	39.9	32.7	5.4	0.0	2.2	100
man	14.9	37.5	40.1	5.1	0.7	1.7	100
ser	21.7	29.1	40.4	6.0	1.0	1.8	100
arm	24.6	29.5	34.9	7.8	1.9	1.3	100
Tot	21.0	28.3	33.7	12.0	4.0	1.1	100

TAB. 4.6 - *Origine socio-professionnelle des étudiants français: pourcentages lignes*

age	Agriculteurs exploitants
oua	Ouvrier agricole
ind	Industriel
art	Artisan
gco	Moyen et gros commerçant
pco	Petit commerçant
lib	Profession libérale
ing	Ingénieur
adm	Cadre de l'administration
mes	Profession médicale et salariée
mso	Profession médicale et sociale
tec	Technicien
cam	Cadre administratif moyen
ins	Instituteur
emb	Employés de bureau
con	Contremaitre
ous	Ouvrier spécialisé
min	Mineur
mar	Pècheur
man	Manoeuvre
ser	Personnel de service
arm	Armée , Police

dro	Droit
sci	Sciences
let	Lettres
med	Medecine
pha	Pharmacie
IUT	IUT

Si le fait de connaître la profession du père d'un étudiant donne une information (non déterministe) sur ses probabilités de faire tel ou tel type d'étude, alors ces deux variables sont dites liées. Avant de connaître l'origine socio-professionnelle d'un étudiant, la probabilité de faire tel ou tel type d'études est celle de l'ensemble de la population étudiante (dernière ligne du tableau 4.6). Une fois l'information de l'origine connue, la distribution attendue est celle définie par le profil ligne associé.

Donc, en ce qui concerne la "vraie" distribution de probabilité (inconnue mais dont une estimation nous est donnée par les fréquences relatives observées) la non liaison équivaut à l'égalité de tous les profils lignes entre eux. Si tous les profils lignes sont égaux entre eux, toutes les lignes sont proportionnelles et donc aussi proportionnelles à la marge ligne. Donc tous les profils lignes sont égaux au profil marginal ligne (situé dans la dernière ligne du tableau 4.5). On en déduit que, pour tout indices i et j :

$$\Pr(j \mid i) = \Pr(j),$$

ce qui donne les écritures équivalentes suivantes:

$$\Pr(i \text{ et } j) / \Pr(i) = \Pr(j),$$

$$\Pr(i \text{ et } j) = \Pr(i)\Pr(j),$$

$$\Pr(i \text{ et } j) / \Pr(i)\Pr(j) = 1.$$

Un tableau de contingence associé à deux variables liées est donc un tableau où, pour certains couples (i, j), $\Pr(i \text{ et } j)/\Pr(i)\Pr(j)$ *s'écarte significativement de* 1.

4.2.2 Une mesure globale de la liaison

Même si la probabilité théorique vérifie la propriété d'indépendance, une probabilité observée, calculée sur un échantillon tiré suivant cette loi, ne vérifiera pas exactement les conditions d'indépendance. On sait que sous l'hypothèse d'indépendance, la quantité :

$$\chi^2 = \sum_{i,j} \left[\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \right]$$

suit une loi du χ^2 à $(I-1)(J-1)$ degrés de liberté (avec I et J nombre de modalités des variables). En utilisant les fréquences f_{ij}, estimateurs des probabilités $\Pr(i \text{ et } j)$, cette quantité, qui s'écrit encore:

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

s'interprète comme une mesure de distance de la distribution observée à l'indépendance:

– Si le χ^2 est grand, on rejette l'hypothèse d'indépendance et on conclue qu'il y a liaison, due à des écarts significatifs entre $f_{ij}/(f_i f_j)$ et 1 pour au moins certains couples (i, j).

– Si le χ^2 est petit, on accepte l'hypothèse de non liaison, et il n'y a rien à expliquer.

Dans l'hypothèse où une liaison a été détectée, l'analyse des correspondances a pour but de décrire et résumer la liaison entre les deux variables. Elle réalise des approximations graphiques de la matrice dont l'élément (i, j) est $(f_{ij}/(f_i f_j) - 1)$, appelés *indices de dépendance entre i et j*.

4.3 L'Analyse (Factorielle) des Correspondances (AFC)

Dérivée de l'analyse en composantes principales, l'analyse des correspondances traite de *variables qualitatives*. Elle a pour but de *résumer* et de *décrire* les liaisons existantes entre deux variables qualitatives. Elle permet de mettre en évidence les "attirances" et les "répulsions" entre catégories i et j.

4.3.1 Notations et définitions

Les observations de ces deux variables sont comptabilisées dans une table de contingence $N = (n_{ij})$. On utilise les notations précédentes, et on note respectivement $F = 1/n\ N$ le tableau des fréquences, $D_I = diag(f_{i.}, i \in \mathbf{I})$ et $D_J = diag(f_{.j}, j \in \mathbf{J})$ les matrices diagonale des poids des lignes et des poids des colonnes. On présente ici l'AFC en traitant les lignes et les colonnes de N de façon symétrique.

Définition 24 L'AFC d'une table de contingence s'obtient en réalisant l'ACP du tableau $D_I^{-1} F D_J^{-1}$, d'éléments $(f_{ij}/(f_{i.} f_{.j}))$ avec les métrique D_I dans $\mathbf{R}^I$ (espace des variables de cette ACP) et D_J dans $\mathbf{R}^J$ (espace des individus). Dans la suite, on se référera à cette ACP en écrivant "ACP des indices de liaisons".

Les poids des individus sont donc les poids des lignes $f_{i.}$ et les poids des variables pour la métrique de $\mathbf{R}^J$ sont $f_{.j}$.

4.3.2 Propriétés

1. La moyenne de chaque variable du tableau vaut 1. Le tableau centré a pour élément (i, j) l'indice de liaison $f_{ij}/(f_{i.} f_{.j}) - 1$. Ce tableau est doublement centré (en colonnes par construction, en lignes car $\sum_j f_{.j}(f_{ij}/(f_{i.} f_{.j}) - 1) = 0$). L'AFC va générer des approximations de ces indices de liaison qui seront lues graphiquement sur les représentations biplot.

2. L'inerties totale de cette ACP a pour valeur le Φ^2 de Pearson de la table de contingence, soit encore le χ^2 de cette table divisée par n :

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

3. La D.V.S. que réalise l'AFC s'écrit :

$$\frac{f_{ij}}{f_{i.} f_{.j}} - 1 = \sum_{s=1} \lambda_s u_{is} v_{js} \quad \left\{ \begin{array}{lll} \text{avec} & \{\mathbf{u}_s,\ s = 1 \ldots, r\} & \text{famille } D_I\text{-orthonormée} \\ \text{et} & \{\mathbf{v}_s,\ s = 1 \ldots, r\} & \text{famille } D_J\text{-orthonormée.} \end{array} \right. \tag{4.1}$$

4. Les vecteurs $\mathbf{u}_s,\ s = 1\dots,r$ sont centrés (soit $\sum_i p_i u_{is} = 0$, propriétés de l'ACP). De même les vecteurs $\mathbf{v}_s,\ s = 1\dots,r$ sont centrés, car combinaisons linéaires des vecteurs lignes centrés (ceci est une conséquence des formules de transition et du double centrage du tableau d'indices de liaison). En notant $\mathbf{1}_I$ et $\mathbf{1}_J$ les vecteurs de $\mathbf{R}^I$ et de $\mathbf{R}^J$ dont toutes les coordonnées sont égales à 1, cela s'écrit matriciellement ${}^t\mathbf{u}_s D_I \mathbf{1}_I = {}^t\mathbf{v}_s D_I \mathbf{1}_I = 0$.

5. En conséquence de la propriété précédente et en posant $\mathbf{u}_0 = \mathbf{1}_I$ et $\mathbf{v}_0 = \mathbf{1}_J$, les famille de vecteurs $\{\mathbf{u}_s,\ s = 0,1,\dots,r\}$ et $\{\mathbf{v}_s,\ s = 0,1,\dots,r\}$ sont respectivemnt D_I et D_J orthonormées.

6. Les valeurs singulières sont inférieures ou égales à 1 (cf. exercice 4.3.1).

7. En posant $\lambda_0 = 1$, la DVS précédente peut se réécrire comme la DVS du tableau non centré $D_I^{-1}FD_J^{-1}$ de la façon suivante :

$$\frac{f_{ij}}{f_{i.}f_{.j}} = \sum_{s=0}^{r} \lambda_s u_{is} v_{js} \quad \left\{ \begin{array}{lll} \text{avec} & \{\mathbf{u}_s,\ s = 0\dots,r\} & \text{famille } D_I\text{-orthonormée} \\ \text{et} & \{\mathbf{v}_s,\ s = 0\dots,r\} & \text{famille } D_J\text{-orthonormée.} \end{array} \right. \tag{4.2}$$

(il suffit de remarquer que l'élément (i,j) de 1 $\mathbf{1}_I\ {}^t\mathbf{1}_J$ est égal à 1). Dans cette DVS, le premier triplet singulier $(1, \mathbf{1}_I, \mathbf{1}_J)$ apparaît comme un artefact et ne doit pas être utilisé dans les représentations.

8. La représentation isométrique des lignes de l'ACP des indices de liaison est identique à celle obtenue dans l'ACP des profils lignes $(f_{ij}/f_{i.},\ i = 1,\dots,I)$ munis des poids $f_{i.}$ avec la métrique D_J^{-1} (dite *métrique du* χ^2).

9. La représentation isométrique des colonnes de l'ACP des indices de liaison est identique à celle obtenue dans l'ACP des profils colonnes $(f_{ij}/f_{.j},\ j = 1,\dots,J)$- jouant ici le rôle d'individus - munis des poids $f_{.j}$ avec la métrique D_I^{-1}.

10. Formules de transition : elles s'écrivent de la façon suivante :

$$\lambda_s v_{js} = \sum_{i\in\mathbf{I}} \frac{f_{ij}}{f_{.j}} u_{is} \tag{4.3}$$

$$\lambda_s u_{is} = \sum_{j\in\mathbf{J}} \frac{f_{ij}}{f_{i.}} v_{js} \tag{4.4}$$

Preuve : avec les notations utilisés pour présenter l'ACP, les formules de transition sont :

$$\begin{aligned} \lambda_s \mathbf{u}_s &= XM\mathbf{v}_s \\ \lambda_s \mathbf{v}_s &= {}^tXD\mathbf{u}_s\ . \end{aligned}$$

Appliquées à cette ACP particulière, on obtient :

$$\begin{aligned} \lambda_s \mathbf{u}_s &= (D_I^{-1}FD_J^{-1})D_J\mathbf{v}_s \\ \lambda_s \mathbf{v}_s &= (D_J^{-1}\,{}^tFD_I^{-1})D_I\mathbf{u}_s\ . \end{aligned}$$

Exercice 4.3.1 *On appelle matrice stochastique une matrice d'éléments positifs ou nuls tels que la somme des éléments des lignes soit égale à 1.*

(a) Monter que les vecteur $\{\mathbf{u}_s,\ s = 0,1,\dots,r\}$ sont vecteurs propres de la matrice $D_I^{-1}FD_J^{-1}\,{}^tF$.

(b) Montrer que la matrice $D_I^{-1}FD_J^{-1}\,{}^tF$ est stochastique.

(c) Montrer que si $A = (a_{ij})$ est carrée, la plus grande valeur propre λ de A vérifie l'inégalité $|\lambda| < \sum_j |a_{ij}|$ (pour tout i fixé).

(d) En déduire que les valeurs singulières de l'AFC sont toutes inférieures ou égales à 1.

4.3.3 Les représentations usuelles de l'AFC

On utilise trois représentations en analyse des correspondances. Les deux représentations dites *barycentriques* et la représentation *simultanée*. On considère l'espace $\mathbf{R}^r$ muni de la métrique usuelle et sa base canonique orthonormée. Dans cet espace on note

- M_i un point (ou *marqueur de la ligne* i) de coordonnées $(\lambda_s u_{is},\ s = 1, \ldots, r)$ et $\mathcal{M}_I$ le nuage de ces I points,
- N_j un point de coordonnées $(\lambda_s v_{js},\ s = 1, \ldots, r)$ et $\mathcal{N}_J = \{N_j,\ j \in \mathbf{J}\}$.
- M'_i un point de coordonnées $(u_{is},\ s = 1, \ldots, r)$ et $\mathcal{M}'_I = \{M_i,\ i \in \mathbf{I}\}$.
- N'_j un point de coordonnées $(v_{js},\ s = 1, \ldots, r)$. et $\mathcal{N}'_J = \{N'_j,\ j \in \mathbf{J}\}$.

La représentation simultanée.

Définition 25 On appelle *représentation simultanée* la représentation sur un même graphe des nuages $\mathcal{M}_I$ et $\mathcal{M}_J$.

On reconnait dans la représentation du nuage $\mathcal{M}_I$ (resp. $\mathcal{N}_J$) la représentation isométrique des lignes (resp. des colonnes) de l'ACP des indices de liaison. ces deux représentations sont donc simplement superposées. On justifiera cette pratique à partir des propriétés barycentriques du paragraphe suivant. On peut, à partir de cette représentation, étudier les proximités entre profil lignes et entre profil colonnes. Par contre, il ne sera pas licite d'interpréter les distances entre un profil ligne et un profil colonne.

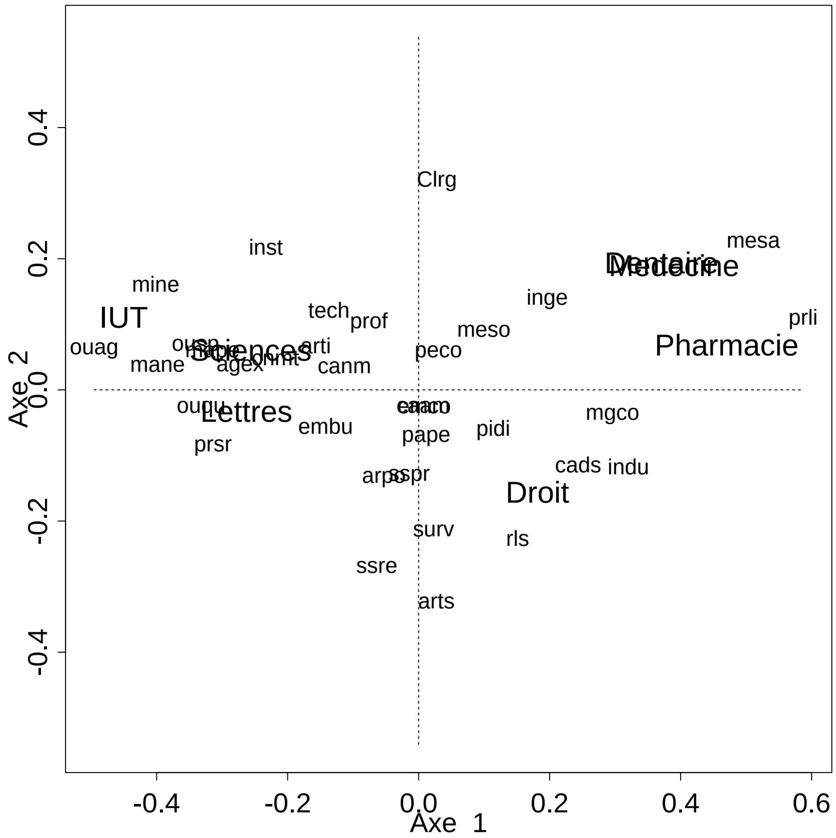


FIG. 4.1 - *Origine socio-professionnelle des étudiants : représentation simultanée sur le plan 1-2*

Les représentations barycentriques

Définition 26 la *première représentation barycentrique* de l'analyse des correspondances s'obtient par superposition de

- la représentation du nuage de points $\{M_i,\ i \in \mathbf{I}\}$ (représentation isométrique des lignes).
- et de celle du nuage $\{N'_j, j \in \mathbf{J}\}$ (représentation non isométrique)

La *seconde représentation barycentrique* de l'analyse des correspondances s'obtient par superposition de

- la représentation du nuage $\{M'_i, i \in \mathbf{I}\}$ (représentation non isométrique)
- et de celle du nuage de points $\{N_j,\ j \in \mathbf{J}\}$) (représentation isométrique des colonnes).

On a alors les propriétés suivantes :

(a) Dans la première représentation barycentrique, tout point M_i est barycentre du nuage de points $\{N'_j, j \in \mathbf{J}\}$ pour le système de poids $(f_{ij}/f_{i.},\ j \in \mathbf{J})$, constitué des éléments du i-ième profil ligne de la table de contingence. On donne figure 4.3 un exemple d'une telle représentation.

Remarque 4.3.1 La différence d'échelle entre les deux nuages (ici $\lambda_1 \simeq 0.25$, $\lambda_2 \simeq 0.11$) implique que le nuage des barycentres est concentré autour de l'origine et rend cette représentation est peu lisible. Il pourra être utile de faire subir au nuage des barycentres une homothétie pour lui donner une échelle semblable à celle du nuage $\mathcal{N}_J$; les points perdent alors leurs propriétés de barycentres. Une autre solution consiste à représenter les deux nuages sur deux graphiques adjacents.

(b) Dans la seconde représentation, tout point N_j est barycentre du nuage de points $\{M'_i, i \in \mathbf{I}\}$ pour le système de poids $(f_{ij}/f_{.j},\ i \in \mathbf{I})$, constitué des éléments du j-ième profil colonne de la table de contingence.

Ces *relations barycentriques* sont conséquences directes des formules de transition. Elles sont à la base d'une des méthodes d'interprètation de l'analyse des correspondances (voir un exemple Figures 4.6 et 4.7). Le principe en est le suivant (première représentation barycentrique). *Si un barycentre M_i de $\mathcal{M}_I$ est excentré dans une certaine direction, c'est qu'il a été "attiré" par des points du nuage $\mathcal{N}'_J$ situés dans cette direction. Par conséquent les poids des éléments j associés à ces points, éléments du profil ligne i, ont tendance à être plus forts que la moyenne : il y a une attraction entre cette catégorie i et les catégories j associées à ces points.* Par extension, on utilise ces propriétés pour l'interprétation de la représentation simultanée. En effet, les formes des nuages sont approximativement conservées quand qu'on passe de $\mathcal{M}_I$ à $\mathcal{M}'_I$, surtout lorsque les valeurs propres associées au plan d'interpétation sont à peu près égales.

Remarque 1 On raisonne ici directement dans l'espace $\mathbf{R}^s$. Les représentations seront effectuées dans des plans de $\mathbf{R}^s$ (par exemple, plan 1-2, plan 1-3, etc.). Mais les relations barycentriques se conservent par projection et restent vraies dans ces sous-espaces.

4.3.4 L'équivalence distributionnelle

Propriété 23 On ne change pas les résultats d'une AFC en remplaçant deux lignes ou deux colonnes proportionnelles par leur somme.

C'est une propriété de stabililité de l'analyse relativement au processus de catégorisation qui a permis de définir les modalités des deux variables qualitatives.

Dans l'exemple de l'étude de la liaison entre type d'étude et profession des parents, la définition des catégories socio-économiques est assez arbitraire. Par exemple, il est possible soit de subdiviser la catégorie "ouvrier" en de nombreuses sous-catégories (comme c'est le cas dans notre exemple) soit de regrouper ces sous-catégories en une seule catégorie "ouvrier". Cette propriété nous assure que si le comportement face aux études des enfants des différentes sous-catégories est identique, le fait de diviser ou non cette catégorie n'aura pas d'influence sur l'analyse. On a une propriété identique concernant les types d'étude : la division de "Droit Sciences-éco" en deux types d'études "Droit" et "Sciences-éco" ne change pas les résultats de l'analyse si ces deux types d'études ont le même recrutement.

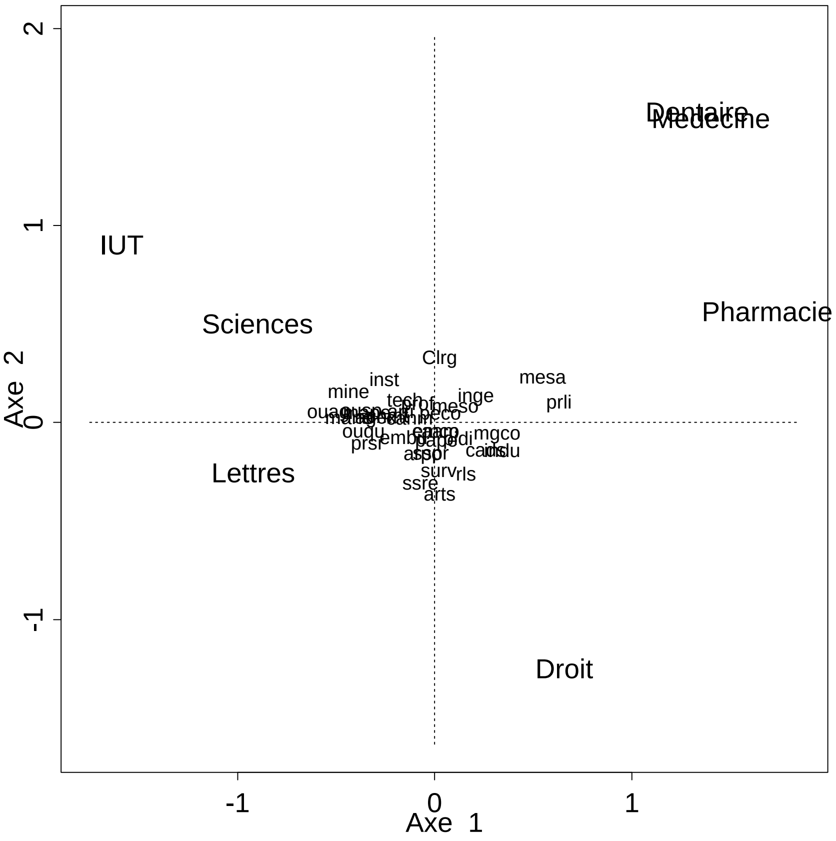


FIG. 4.2 - *Origine socio-professionnelle des étudiants : première représentation barycentrique sur le plan 1-2*

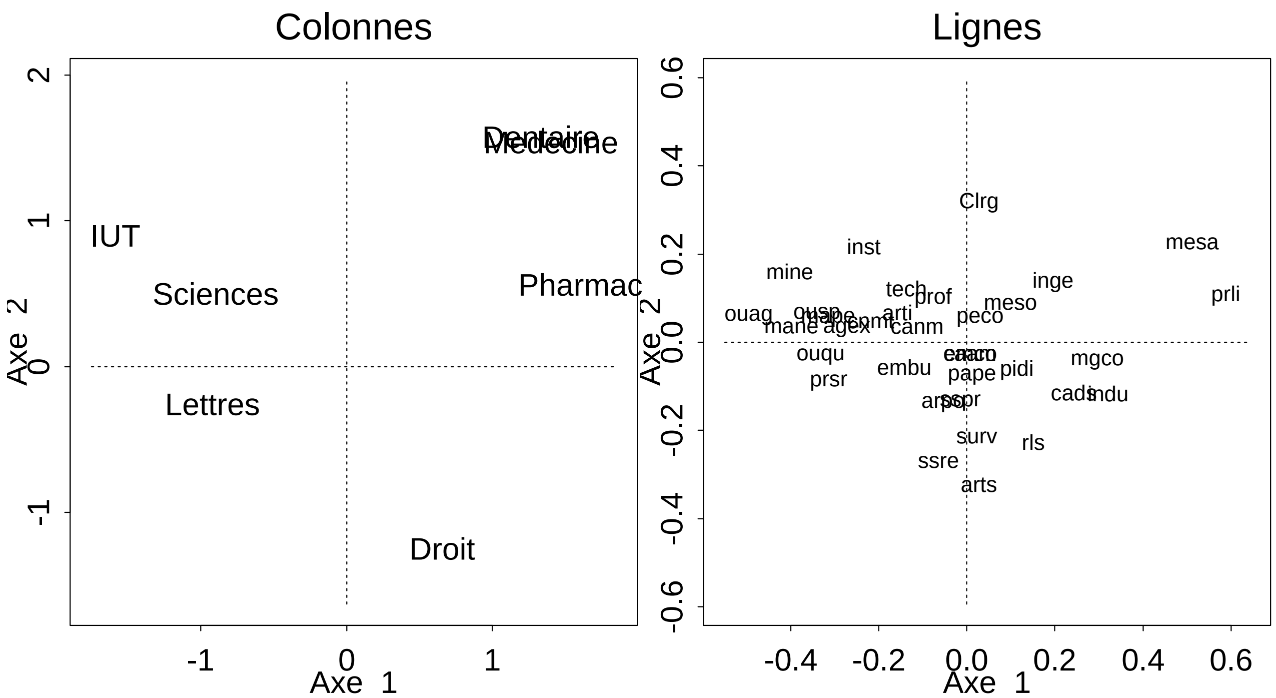


FIG. 4.3 - *Origine socio-professionnelle des étudiants : la même représentation sur deux graphiques séparés*

4.3.5 L'interprétation d'une analyse des correspondances

Avant toute interprétation, il faut vérifier que le $\chi 2$ de la table de contingence N est suffisamment grand pour que la liaison entre les deux variables qualitatives soit jugée significative. Rappelons que le $\chi 2$ calculé à partir de la table suit, sous l'hypothèse de non-liaison, une loi du $\chi 2$ à $(I-1)(J-1)$ degrés de liberté. Une fois cette vérification faite, on effectue sur la représentation simultanée les interprétations suivantes:

- L'analyse de proximité des lignes : quels sont les profils lignes qui se ressemblent, quels sont ceux qui sont éloignés deux à deux ? Deux profils lignes i et i' sont semblables s'ils "s'associent de la même façon à J" ($\Pr[i|j] \simeq \Pr[i'|j]$ pour tout j). Le barycentre, origine des axes, représente la situation d'indépendance. Un point proche du barycentre s'associe à J comme l'ensemble de la population (on a $Pr[j|i] \simeq Pr[j]$).

- L'analyse symétrique des proximités des colonnes.

- L'analyse simultanée, qui utilise les propriétés barycentriques. Cette pratique n'est pas pleinenement justifiable, mais elle tient compte du fait que la transformation qui fait passer de $\mathcal{M}_I$ à $\mathcal{M}'_I$ ou de $\mathcal{N}_J$ à $\mathcal{N}'_J$ ne consiste qu'à éloigner les points du barycentre, de façon presque homothétique, en laissant les points dans le même cadran (dans le plan $1-2$, la transformation est vraiment homotéthique si $\lambda_1 = \lambda_2$). Cette interprétation peut se formuler de la façon suivante :

  Propriété 24 Deux marqueurs i et j qui s'éloignent du barycentre dans la même direction ont tendance à être associés positivement ($Pr[j|i] > Pr[j]$) ; deux marqueurs i et j qui s'éloignent du barycentre dans une direction opposée ont tendance à être associés négativement ($Pr[j|i] < Pr[j]$). Deux marqueurs i et j orthogonaux correspondent à des catégories en situation d'indépendance ($Pr[j|i] \simeq Pr[j]$). Enfin, plus un marqueur est éloigné de l'origine, plus l'association avec les points éloignés dans la même direction que lui est forte.

  Ces affirmation restent approximatives si on interprète le premier plan principal et doivent être corrigées par les analyses effectuées sur les plans ultérieurs.

- L'interprétation des contributions : quelles sont les lignes et les colonnes qui ont une forte contribution à l'inertie totale ? Ces fortes contributions sont-elles le témoin d'une forte originalité, d'un fort poids, ou de ces deux propriétés simultanément réunies ?

- L'interprétation des facteurs : on peut chercher à leur donner un sens. Un facteur s'interprète en fonction des lignes et colonnes qui y ont la plus forte contribution.

- On utilise ensuite les coefficients de qualité (cosinus carrés cumulés) : quand on interprète k facteurs, quelles sont les catégories i et j bien représentées ? quelles sont celles dont l'originalité n'est pas ou peu prise en compte ?

4.4 AFC et biplot

La technique de lecture du biplot s'applique avec intérêt en AFC. Nous distinguons les biplots classiques "points-et-flêches" des biplots callibrés.

4.4.1 Le biplot classique

Les biplots isométriques lignes et isométriques colonne sont respectivement les premières et secondes représentations barycentriques de l'AFC. On obtient ainsi une deuxième technique d'interprétation, qui a l'avantage d'être précise. Cette précision dépend de la finesse du graphique, et permet de lire graphiquement les approximations des indices de liaisons. Le produit scalaire des vecteurs OM_i et ON'_j sur le plan 1-2 est donc égal à l'approximation de rang 2 de l'indice $f_{ij}/(f_{i.}f_{.j}) - 1$.

La quantité $f_{ij}/(f_{i.}f_{.j}) - 1$ est une estimation du taux $Pr[j|i]/Pr[j] - 1$ (ou du rapport obtenu symétriquement par permutation de i et j) qui mesure l'apport de la connaissance de i sur l'occurrence de j. Ainsi, si i est la catégorie "profession libérale" et si j est la catégorie "médecine", un indice égal à 0.1 signifie que, par rapport à l'ensemble de la population, un enfant de profession libérale a 10% de plus de chances de faire médecine. Un indice égal à

1 signifie que le surplus de chance est de 100%. Un indice égal à 0.6 singifie que les chances sont de 40% inférieures à celles d'un étudiant moyen.

Remarque 4.4.1 Du dernier point ci-dessus, on peut voir qu'il sera plus facile d'avoir des représentations de points excentriques (loin du barycentre) avec des catégories de petite taille : un tel indice peut atteindre 5 avec de petites catégories (ex : fréquence dans l'ensemble de la population : 1%, fréquence dans la catégorie : 6%, soit un indice de $0,06/0,01-1=5$). Pour une catégorie très présente (par exemple 30% dans la population) on ne pourra avoir qu'un indice plus faible (par exemple si la fréquence dans une catégorie donnée atteint 45%, $0,45/0,30-1=0,5$, alors que le surcroit de présence dans la catégorie atteint 15%).

Remarque 4.4.2 Le biplot symétrique présente l'intérêt d'avoir pour les deux nuages une échelle comparable. Mais il est constitué de deux représentations non isométriques

4.4.2 Le biplot callibré

On a vu qu'avec un biplot callibré sur le plan 1-2 réalisé après une ACP, on pouvait graduer un axe j de façon à lire, en projetant un marqueur i sur cet axe, les valeurs approchées à l'ordre 2 de la variable centrée et éventuellement réduite pour l'individu i. Par transformation linéaire des graduations, on peut aussi représenter les valeurs approchées non centrées et non réduites. De la même façon, en AFC, on peut graduer l'axe j pour lire les approximations des indices de liaison $f_{ij}/(f_{i.}f_{.j})-1$. En faisant subir aux graduations la transformation linéaire $y \longrightarrow 100 f_{.j} * (y+1)$, on peut lire directement les valeurs approchées des approximations du pourcentage $100 * f_{ij}/f_{i.}$ de la population de la catégorie j dans la catégorie i. Cette pratique sera réservée au plan 1-2, pour analyser l'approximation de rang 2. Sur le plan 3-4, on effectuera la transformation sans décentrage. On y lira les corrections en pourcentage à effectuer sur les valeurs lues sur le plan 1-2. Dans notre exemple , on ne reproduit que le biplot callibré sur le plan 1-2 (Figure 4.4). On lira sur l'axe j (type d'étude j) le pourcentage d'étudiants issus de la CSP i. On voit que, en approximation de rang 2, le pourcentage d'étudiants faisant médecine est de 23% chez les professions libérale, mais n'est que de 5% chez les marins, personnels de service, ouvriers spécialisés. Le pourcentage d'étudiants faisant pharmacie varie de 1% (marins, personnels de service, ouvriers) à près de 9% (professions libérales). Par contre 40% des enfants d'ouvrier font des lettres, alors que seulement 25% des enfants de professions libérales suivent cette filière. On peut comparer ces valeurs à celles figurant dans la tableau 4.6.

4.5 Exemples et problèmes

4.5.1 Etude de la liaison entre études suivies et catégories socio-économiques

Nous donnons pour terminer les résultats d'une analyse des correspondances effectuées sur le tableau 4.5.

```
Pourcentage d'inertie expliquée
            f1 f2 f3 f4  f5 tot
inertie exp 75 12  9  2   2 100
inertie cum 75 87 96 98 100 100
```

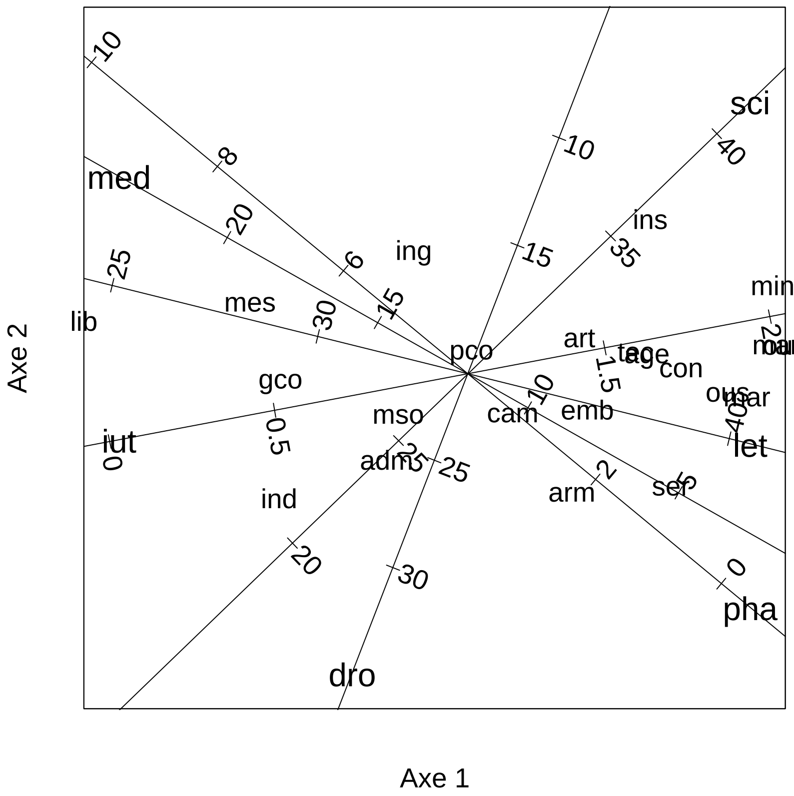


FIG. 4.4 - *Biplot callibré: pourcentage, par catégorie sociale, d'étudiants faisant des études en ...*

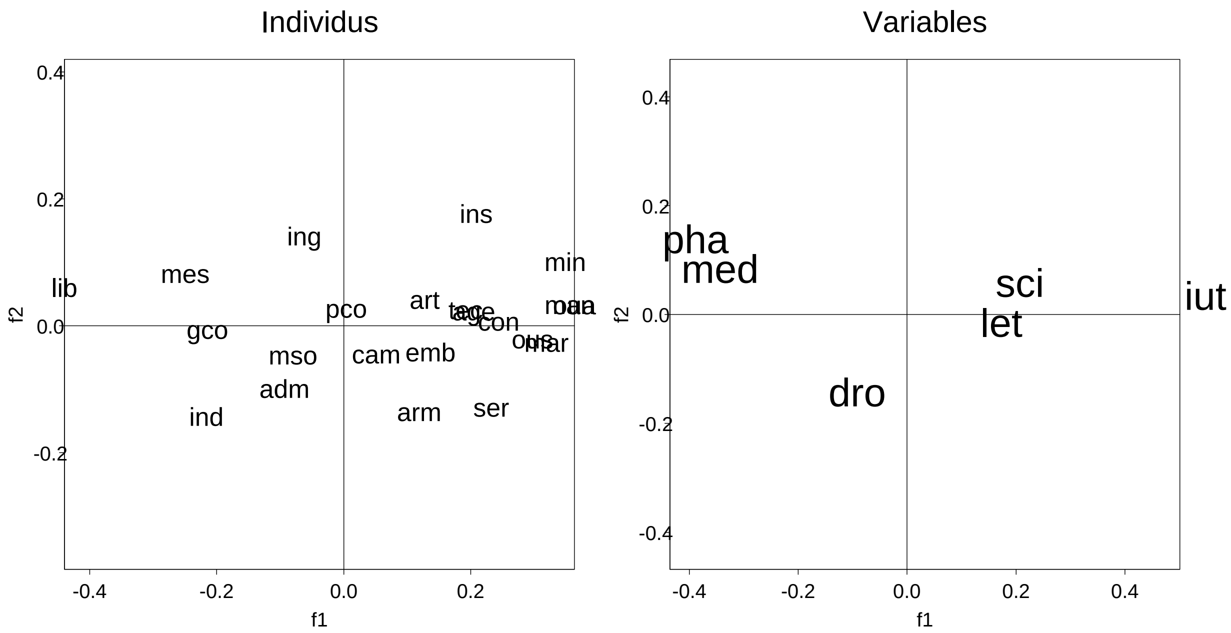


FIG. 4.5 - *Représentation simultanée sur graphiques disjoints*

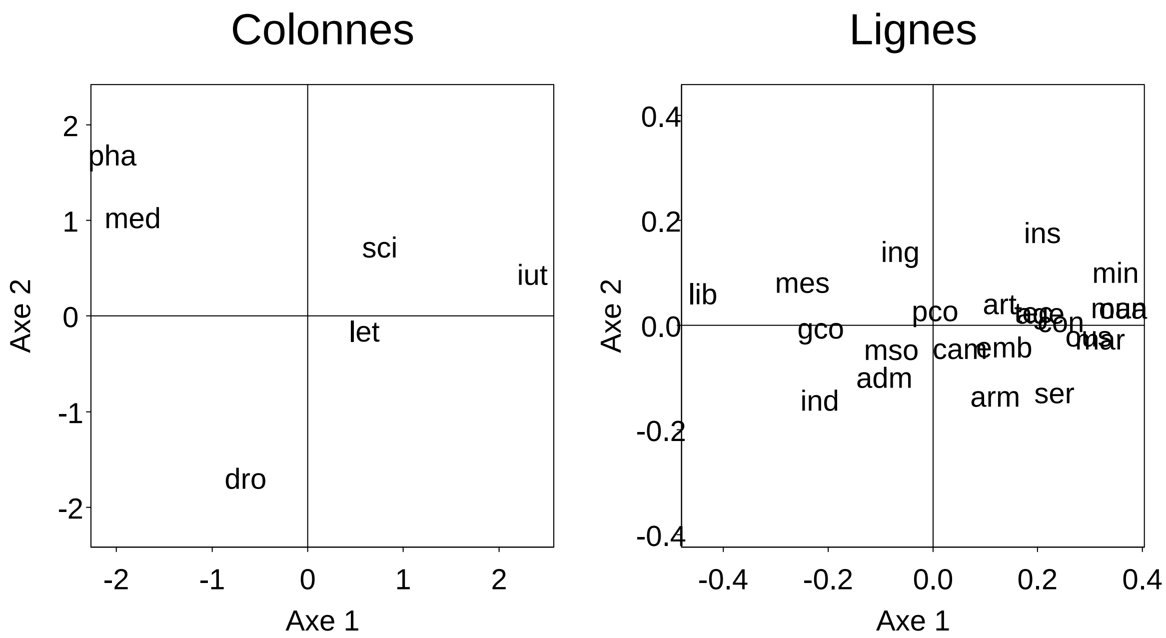


Fig. 4.6 - *Biplot isométrique ligne*

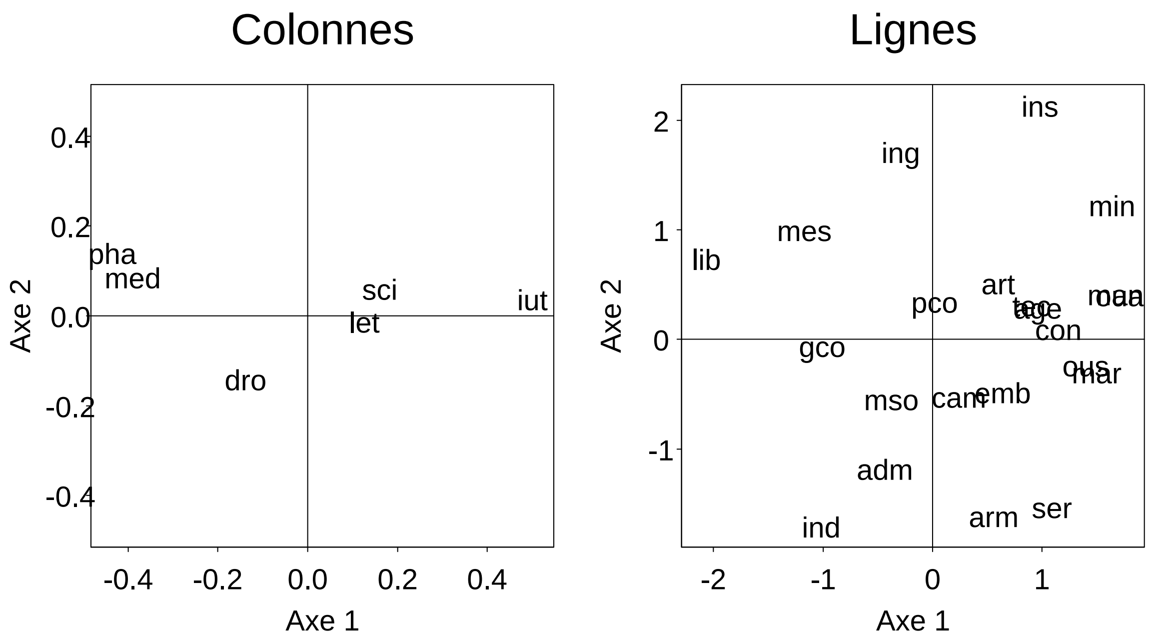


Fig. 4.7 - *Biplot isométrique colonne*

	f1	f2	f3	f4	f5	f>5	TOT
age	731	10	161	96	2	0	1000
oua	957	8	25	9	0	0	1000
ind	615	264	71	49	2	0	1000
art	870	97	7	3	23	0	1000
gco	821	0	119	25	35	0	1000
pco	3	171	85	7	734	0	1000
lib	976	19	3	2	0	0	1000
ing	134	704	17	22	123	0	1000
adm	437	484	73	0	6	0	1000
mes	740	82	4	100	73	0	1000
mso	355	113	82	258	191	0	1000
tec	945	18	27	10	1	0	1000
cam	441	314	92	102	52	0	1000
ins	498	361	108	0	32	0	1000
emb	416	36	547	0	0	0	1000
con	941	1	27	11	19	0	1000
ous	606	3	341	50	0	0	1000
min	883	76	19	10	12	0	1000
mar	745	5	43	1	207	0	1000
man	969	9	8	1	13	0	1000
ser	741	223	0	27	9	0	1000
arm	415	526	3	4	53	0	1000
TOT	753	115	95	20	18	0	1000

TAB. 4.7 - *Contributions des facteurs à l'explication des C.S.P.*

	f1	f2	f3	f4	f5	f>5	TOT
age	731	741	902	998	1000	1000	1000
oua	957	965	991	1000	1000	1000	1000
ind	615	879	950	998	1000	1000	1000
art	870	968	974	977	1000	1000	1000
gco	821	821	940	965	1000	1000	1000
pco	3	174	259	266	1000	1000	1000
lib	976	995	998	1000	1000	1000	1000
ing	134	838	855	877	1000	1000	1000
adm	437	921	994	994	1000	1000	1000
mes	740	822	827	927	1000	1000	1000
mso	355	468	551	809	1000	1000	1000
tec	945	962	989	999	1000	1000	1000
cam	441	755	847	948	1000	1000	1000
ins	498	859	968	968	1000	1000	1000
emb	416	452	1000	1000	1000	1000	1000
con	941	942	969	981	1000	1000	1000
ous	606	609	950	1000	1000	1000	1000
min	883	959	978	988	1000	1000	1000
mar	745	749	793	793	1000	1000	1000
man	969	978	987	987	1000	1000	1000
ser	741	964	964	991	1000	1000	1000
arm	415	941	943	947	1000	1000	1000
TOT	753	868	963	982	1000	1000	1000

TAB. 4.8 - *Qualité de la représentation des lignes par les sous-espaces principaux*

	f1	f2	f3	f4	f5	f>5	TOT
age	67	6	118	337	7	5	69
oua	25	1	5	9	0	15	20
ind	35	99	32	106	4	19	43
art	17	12	1	2	19	104	14
gco	33	0	37	38	60	5	30
pco	0	8	5	2	217	103	5
lib	479	61	12	40	1	0	370
ing	6	213	6	38	246	161	35
adm	30	217	40	1	18	1	52
mes	15	11	1	78	65	34	16
mso	1	2	2	26	22	33	2
tec	28	3	6	11	2	12	22
cam	6	28	10	53	31	134	10
ins	43	204	74	2	119	286	65
emb	32	18	329	0	1	8	57
con	31	0	7	14	28	0	25
ous	68	2	304	214	0	12	85
min	17	10	3	7	10	4	15
mar	6	0	3	0	71	19	6
man	39	2	3	1	22	3	31
ser	12	23	0	16	6	18	12
arm	9	78	1	3	51	23	17
TOT	1000	1000	1000	1000	1000	1000	1000

TAB. 4.9 - *Contribution des lignes aux facteurs et à l'inertie totale*

	f1	f2	f3	f4	f5	f>5	TOT
dro	478	489	23	4	5	0	1000
sci	830	118	2	14	36	0	1000
let	805	8	125	11	52	0	1000
med	932	46	2	18	2	0	1000
pha	808	85	33	46	29	0	1000
iut	376	2	594	23	4	0	1000
TOT	753	115	95	20	18	0	1000

TAB. 4.10 - *Qualité de représentation des colonnes par les facteurs*

	f1	f2	f3	f4	f5	f>5	TOT
dro	478	967	991	995	1000	1000	1000
sci	830	948	950	964	1000	1000	1000
let	805	812	937	948	1000	1000	1000
med	932	978	979	998	1000	1000	1000
pha	808	892	925	971	1000	1000	1000
iut	376	378	972	996	1000	1000	1000
TOT	753	868	963	982	1000	1000	1000

TAB. 4.11 - *Qualité de représentation des colonnes par les sous-espaces principaux des variables*

	f1	f2	f3	f4	f5	f>5	TOT
dro	89	596	35	31	40	210	140
sci	161	150	3	102	302	283	146
let	119	7	147	60	331	337	111
med	403	129	6	301	40	120	325
pha	168	115	55	365	256	40	157
iut	60	2	755	141	30	11	120
TOT	1000	1000	1000	1000	1000	1000	1000

TAB. 4.12 - *Contribution des colonnes aux facteurs et à l'inertie totale*

4.5.2 Structure d'âge des communes de l'agglomération toulousaine

On effectue l'analyse des correspondances du tableau [2] à 2 entrées donnant la répartition de la population en classes d'âge et sexe dans les communes de l'Agglomération Toulousaine (mais dans laquelle la population de 20 à 75 ans forme une seule classe).

Le pourcentage d'inertie expliquée par les premiers facteurs est donné Tableau 4.13.

	f1	f2	f3	f4	f5	tot
inertie exp	85	10	4	1	1	100
inertie cum	85	94	98	99	100	100

Tab. 4.13 - afc *Âges x Communes: pourcentages expliqués*

Deux facteurs expliquent bien la liaison entre les communes et la répartition des classes d'âge. On obtient le biplot isométrique colonne donné figure 4.8.

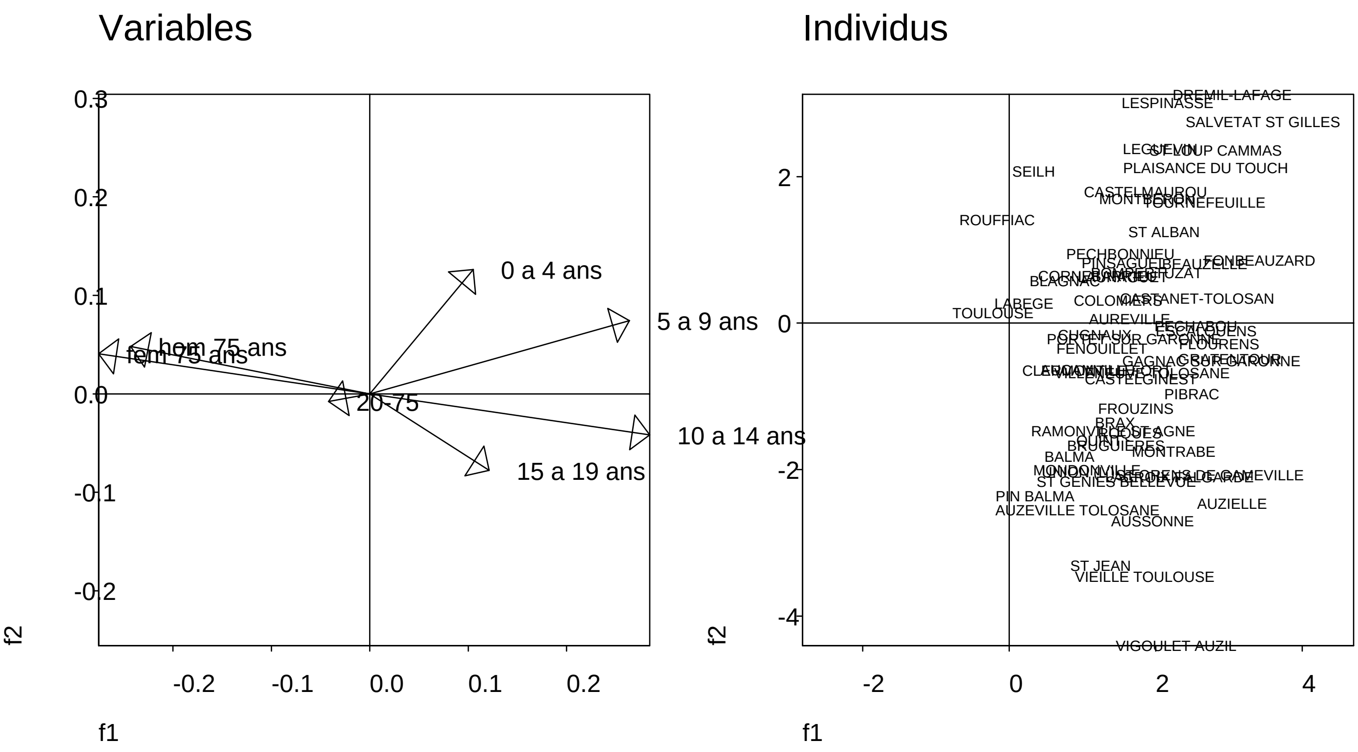


Fig. 4.8 - *Analyse des correspondances : biplot 1 -2 (isom. colonne).*

Les contributions et critères de qualité des communes et des classes d'âge/sexe sont donnés tableaux 4.14 et 4.15. Les contributions sont une mesure combinée de l'*originalité* du profil et de l'importance de la commune. Ils permettent de voir que Toulouse a une très forte contribution (ce qui est du à sa taille et à son profil atypique). D'autres communes ont une forte contribution : Tournefeuille, et dans une moindre mesure St Orens, La Salvetat, Plaisance du Touch. Les contributions des facteurs à l'explication des communes indiquent quels facteurs permettent d'expliquer l'originalité des communes. Ainsi Labège est une commune plutôt moyenne, mais dont l'originalité a trait au facteur 3 (ratio 0-10ans 11-20ans atypique). Souvent, les communes dont l'explication se fait principalement par des facteurs de rang supérieur à 2 sont des communes proches du profil moyen (ex: Seilh).

L'analyse du biplot 1-2 (figure 4.8) montre que le premier facteur oppose les communes jeunes aux communes âgées, les classes marquant le plus le phénomène étant les classes d'âge 5-9 ans et 10-14 ans (exemple Toulouse 7,4 % de plus de 75 ans, Auzielle 1 %). Le deuxième facteur oppose les classes d'âge jeunes entre elles : les communes situées sur la partie positive de l'axe vertical ont une plus forte proportion d'enfants en bas âge (ex : Dremil Lafage 19 % de 0 à 9 ans), l'inverse se produisant du coté opposé (Vigoulet Auzil 6.4 %). Enfin, l'axe 3 est fortement influencé par la catégorie "femmes de plus de 75 ans", s'opposant à "Enfants

2. Source: INSEE, Étude effectuée pour l'Agence d'Urbanisme de l'Agglomération Toulousaine

	f1	f2	f3	f4	f5	f>5	TOT
AUCAMVILLE	1	2	17	2	1	14	2
AUREVILLE	1	0	6	7	3	0	1
AUSSONNE	12	47	7	1	9	3	15
AUZEVILLE TOLOSANE	0	21	54	4	76	67	5
AUZIELLE	12	11	2	32	9	0	11
....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....
LAUNAGUET	5	2	1	0	0	4	5
LEGUEVIN	16	38	1	4	1	0	18
LESPINASSE	5	21	16	5	10	5	7
MONDONVILLE	0	9	0	21	5	21	1
MONTBERON	4	7	4	68	9	0	4
MONTRABE	11	12	6	2	0	1	10
PECHABOU	6	0	0	5	23	50	6
PECHBONNIEU	2	3	2	4	9	2	2
PIBRAC	42	9	113	1	26	2	41
PIN BALMA	0	6	43	32	16	0	3
PINSAGUEL	4	3	9	1	14	19	4
PLAISANCE DU TOUCH	39	73	20	1	3	0	41
POMPERTUZAT	2	1	33	1	22	3	3
PORTET SUR GARONNE	3	1	24	1	37	4	4
QUINT	4	13	7	32	2	2	6
RAMONVILLE ST AGNE	2	41	137	8	34	32	11
ROQUES	6	10	7	145	46	28	8
ROUFFIAC	1	3	79	20	10	2	4
ST ALBAN	19	11	29	91	27	42	19
ST GENIES BELLEVUE	0	11	0	2	0	3	1
ST JEAN	8	126	31	8	26	0	21
ST LOUP CAMMAS	8	12	2	142	15	4	9
ST ORENS DE GAMEVILLE	33	67	8	64	34	2	36
SALVETAT ST GILLES	40	52	11	1	0	6	39
SEILH	0	6	14	0	2	58	1
TOULOUSE	347	12	0	0	0	2	295
TOURNEFEUILLE	90	73	0	0	0	52	83
UNION (L')	4	78	0	45	20	12	11
VIEILLE TOULOUSE	1	17	7	5	3	49	3
VIGOULET AUZIL	3	29	0	11	2	30	6
VILLENEUVE TOLOSANE	5	6	4	0	9	5	5
TOT	1000	1000	1000	1000	1000	1000	1000

TAB. 4.14 - *Contributions des communes aux facteurs.*

	f1	f2	f3	f4	f5	f>5	TOT
AUCAMVILLE	501	129	325	7	4	34	1000
AUREVILLE	722	0	203	54	20	1	1000
AUSSONNE	680	298	17	1	4	1	1000
AUZEVILLE TOLOSANE	20	409	401	6	105	59	1000
AUZIELLE	872	93	7	21	6	0	1000
BALMA	339	562	2	5	83	8	1000
BEAUZELLE	961	17	17	0	1	3	1000
BLAGNAC	436	221	180	0	20	143	1000
BRAX	733	114	16	11	123	4	1000
BRUGUIERES	563	293	77	4	61	0	1000
CASTANET-TOLOSAN	983	6	7	0	1	2	1000
CASTELGINEST	879	55	39	7	4	16	1000
CASTELMAUROU	532	191	238	4	20	14	1000
CLERMONT LE FORT	76	115	90	80	368	271	1000
COLOMIERS	916	14	61	8	1	1	1000
CORNEBARRIEU	709	221	8	3	1	59	1000
CUGNAUX	963	6	2	2	27	0	1000
DREMIL-LAFAGE	791	180	0	23	6	0	1000
ESCALQUENS	993	0	2	2	3	0	1000
FENOUILLET	842	28	33	60	32	5	1000
FLOURENS	967	2	7	4	20	0	1000
FONBEAUZARD	985	12	0	1	1	1	1000
FROUZINS	852	90	8	45	4	2	1000
GAGNAC SUR GARONNE	898	12	4	4	31	52	1000
GRATENTOUR	984	5	0	9	0	2	1000
LABEGE	19	4	807	0	95	75	1000
LACROIX FALGARDE	751	225	1	0	22	0	1000
LAUNAGUET	939	50	7	1	0	3	1000
LEGUEVIN	783	214	1	2	1	0	1000
LESPINASSE	623	279	80	5	10	3	1000
MONDONVILLE	146	640	1	116	26	70	1000
MONTBERON	684	153	34	114	14	0	1000
MONTRABE	866	110	22	1	0	1	1000
PECHABOU	928	0	1	7	27	38	1000
PECHBONNIEU	780	134	35	14	32	5	1000
PIBRAC	871	21	103	0	4	0	1000
PIN BALMA	12	231	620	92	44	0	1000
PINSAGUEL	812	65	78	2	22	20	1000
PLAISANCE DU TOUCH	806	175	18	0	1	0	1000
POMPERTUZAT	487	22	431	2	54	5	1000
PORTET SUR GARONNE	699	14	217	1	64	5	1000
QUINT	668	238	45	44	2	2	1000
RAMONVILLE ST AGNE	130	367	463	6	22	13	1000
ROQUES	659	118	32	136	40	16	1000
ROUFFIAC	147	74	723	36	16	2	1000
ST ALBAN	831	57	56	36	10	10	1000
ST GENIES BELLEVUE	201	776	0	13	2	8	1000
ST JEAN	328	604	56	3	9	0	1000
ST LOUP CAMMAS	730	128	8	121	12	2	1000
ST ORENS DE GAMEVILLE	786	185	8	14	7	0	1000
SALVETAT ST GILLES	859	130	10	0	0	1	1000
SEILH	1	406	390	0	12	191	1000
TOULOUSE	996	4	0	0	0	0	1000
TOURNEFEUILLE	911	86	0	0	0	3	1000
UNION (L')	278	675	0	30	12	5	1000
VIEILLE TOULOUSE	303	523	85	12	7	70	1000
VIGOULET AUZIL	464	495	1	14	3	23	1000
VILLENEUVE TOLOSANE	834	119	29	0	13	5	1000
TOT	846	98	37	8	7	4	1000

TAB. 4.15 - *Contributions des facteurs à l'explication du profil des communes.*

	f1	f2	f3	f4	f5	f>5	TOT
0 a 4 ans	36	446	85	369	1	13	80
5 a 9 ans	257	176	46	297	102	66	239
10 a 14 ans	290	54	171	12	289	128	260
15 a 19 ans	68	239	6	230	267	118	85
20-75	78	23	120	59	11	7	73
hom 75 ans	81	27	141	22	200	508	80
fem 75 ans	191	36	431	12	131	160	183
TOT	1000	1000	1000	1000	1000	1000	1000

TAB. 4.16 - *Contributions des classes à la détermination des facteurs.*

	f1	f2	f3	f4	f5	f>5	TOT
0 a 4 ans	379	546	39	35	0	1	1000
5 a 9 ans	907	72	7	9	3	1	1000
10 a 14 ans	945	20	24	0	8	2	1000
15 a 19 ans	672	276	3	20	22	6	1000
20-75	901	31	61	6	1	0	1000
hom 75 ans	853	33	66	2	17	28	1000
fem 75 ans	883	19	88	0	5	4	1000
TOT	846	98	37	8	7	4	1000

TAB. 4.17 - *Contributions des facteurs à l'explication des profils de classe.*

	f1	f2	f3	f4	f5	f>5	TOT
0 a 4 ans	379	925	965	999	999	1000	1000
5 a 9 ans	907	979	986	996	999	1000	1000
10 a 14 ans	945	965	990	990	998	1000	1000
15 a 19 ans	672	949	952	972	994	1000	1000
20-75	901	931	992	999	1000	1000	1000
hom 75 ans	853	886	952	954	972	1000	1000
fem 75 ans	883	903	991	991	996	1000	1000
TOT	846	944	981	989	996	1000	1000

TAB. 4.18 - *Qualité de la représentation des profils de classe.*

de moins de 4 ans". Ainsi Rouffiac est une commune âgée, avec peu d'adolescents, mais qui a une classe d'âge de 0-4 ans plus faible, relativement aux communes du même type, que celle des 5-9 ans.

Pour préciser cette interprétation, on fait une classification automatique des communes, en utilisant la méthode de Ward, des poids proportionnels à l'effectif des communes, et en utilisant la distance du Chi-deux (l'ensemble de ces choix rend cette classification automatique tout à fait cohérente avec les "choix" de l'analyse des correspondances).

```
Groupe 1
AUCAMVILLE          AUREVILLE          BLAGNAC        CLERMONT LE FORT
COLOMIERS           CORNEBARRIEU       CUGNAUX        FENOUILLET
LAUNAGUET           PECHBONNIEU        PINSAGUEL      PORTET SUR GARONNE
VILLENEUVE TOLOSANE

Groupe 2 :
CASTANET-TOLOSAN   CASTELMAUROU      DREMIL-LAFAGE    GAGNAC SUR GARONNE
LEGUEVIN           LESPINASSE        MONTBERON        PECHABOU
PLAISANCE DU TOUCH POMPERTUZAT       ST ALBAN         ST LOUP CAMMAS
SALVETAT ST GILLES TOURNEFEUILLE

Groupe 3 :
BALMA              MONDONVILLE       QUINT            RAMONVILLE ST AGNE
ST GENIES BELLEVUE ST JEAN           UNION (L')       VIEILLE TOULOUSE

Groupe 4 :
AUZIELLE           BEAUZELLE         ESCALQUENS       FLOURENS
FONBEAUZARD        GRATENTOUR        PIBRAC

Groupe 5 :
AUSSONNE           BRAX              BRUGUIERES       CASTELGINEST
FROUZINS           LACROIX FALGARDE  MONTRABE         ROQUES
ST ORENS DE GAMEVILLE                VIGOULET AUZIL

Groupe 6 :
AUZEVILLE TOLOSANE LABÈGE            PIN BALMA        ROUFFIAC
SEILH

Groupe 7 :
TOULOUSE
```

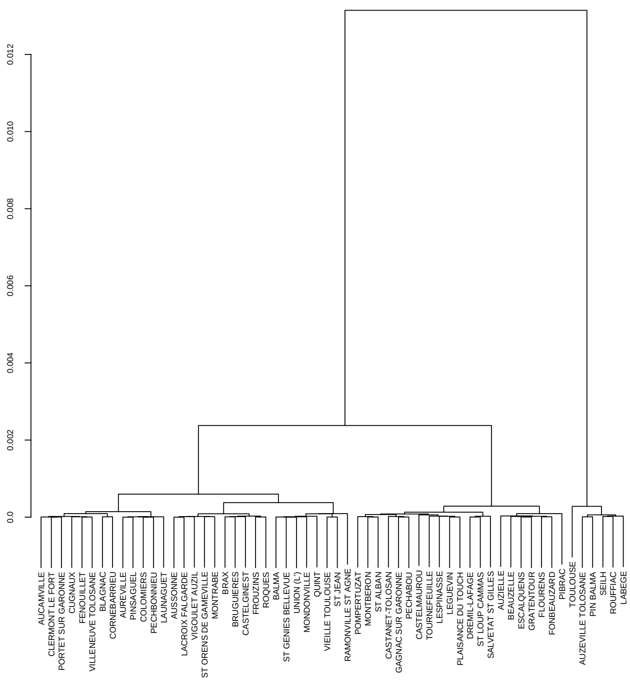


FIG. 4.9 - *Arbre hiérarchique issu de la classification des communes.*

Les groupes de communes obtenus sont donnés ci-dessus (on a choisi une classification en 7 classes, au vu de l'arbre hiérarchique, donné figure 4.9). On donne (figure 4.10) la

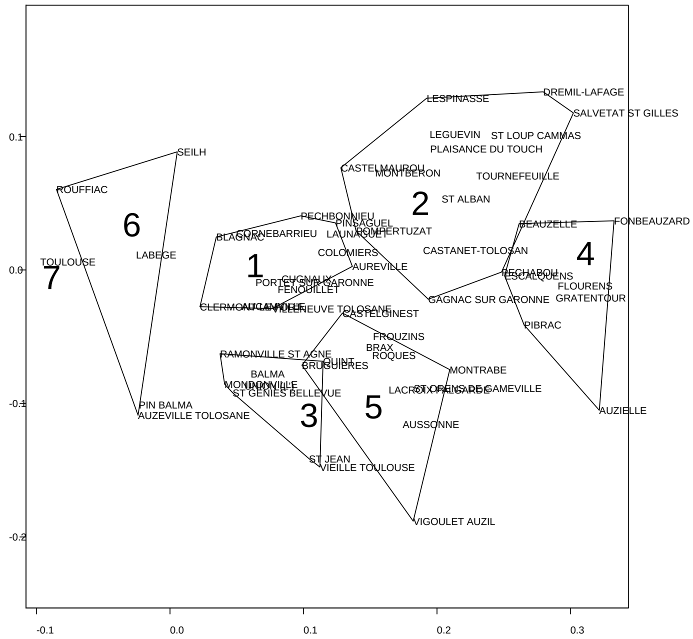


Fig. 4.10 - *Représentation des classes de communes sur le plan 1-2.*

représentation des classes sur le plan factoriel 1-2 (représentation des individus dans le biplot avec cette fois le contour des classes). On voit la parfaite séparation des classes, due à la qualité du résumé sur le plan (94% de l'inertie expliquée). La description des structures d'âge des communes de l'agglomération est facilitée par la classification (on utilise dans cette description principalement la représentation des variables de la figure 4.8 et le plan associé des individus de la figure 4.10.

- La classe 4 contient les communes les plus jeunes, avec une représentation équilibrée des classes d'enfants en bas âge et adolescents.
- Les classes 2 et 5 sont assez jeunes (2 l'étant un peu plus que 5). Mais la classe 5 contient une plus forte proportion d'adolescents que la classe 2, qui elle en revanche a une plus forte proportion d'enfants en bas âge.
- Encore moins jeunes sont les classes 1 et 3 (la classe 1 ayant plus d'enfants en bas âge et la classe 3 plus d'adolescents).
- Enfin la classe 6 est une classe âgée et la ville de Toulouse est la commune de l'agglomération de loin la plus âgée.

Interprétation d'une analyse des correspondances simple (problème d'examen)

Examen du magistère d'économiste statisticien, Toulouse, 1995, partie pratique.

Les données

On considère le tableau à 3 entrées donnant le nombre d'étudiants en France par type d'étude, sexe et profession des parents (1968-69). Les entrées sont les suivantes :

Première entrée : profession des parents (c.s.p., 34 items):

agex	Agriculteurs exploitants
ouag	Ouvrier agricole
indu	Industriel
arti	Artisan
pape	Patron pêcheur
mgco	Moyen et gros commerçant
peco	Petit commerçant
prli	Profession libérale
rls	Profession littéraire et scientifique
inge	Ingénieur
cads	Cadre de l'administration
prof	Professeur
mesa	Profession médicale et salariée
meso	Profession médicale et sociale
tech	Technicien
caam	Cadre administratif moyen
inst	Instituteur
pidi	Profession intellectuelle diverse
surv	Surveillant des établissements d'enseignement
embu	Employés de bureau
emco	Employés de commerce
cnmt	Contremaître
ouqu	Ouvrier qualifié
ousp	Ouvrier spécialisé
mine	Mineur
mape	Pêcheur
mane	Manoeuvre
prsr	Personnel de service
arts	Artiste
Clrg	Clergé
arpo	Armée , police
canm	Catégories non mentionnées
sspr	Sans profession
ssre	Sans réponse

Deuxième entrée : sexe

M	masculin
F	féminin

Troisième entrée : type d'étude suivi
Droit Sciences Lettres Médecine Dentaire Pharmacie IUT
(Source Ministère de l'Éducation Nationale).

1. (1pt) Si on faisait une analyse des correspondances multiple de ce tableau, quelles seraient les dimensions du tableau de Burt à construire?
2. (1pt) On construit un tableau à deux entrées en empilant verticalement les tableaux obtenus pour chaque sexe, et ayant pour lignes les c.s.p. et pour colonne les types d'étude. On obtient un tableau de contingence à deux entrées de 68 lignes et 7 colonnes, sur lequel on effectue une AFC. Les noms des lignes sont constitués de la lettre M ou F pour indiquer le sexe et du nom abrégé de la profession des parents. A quelles probabilités conditionnelles mène l'étude des profils lignes de ce tableau? (resp. colonnes?)
3. (6pts) On donne un certain nombre de résultats issus de l'AFC du dernier tableau. Les interpréter en répondant aux questions suivantes :
   (a) Donner les numéros des tableaux contenant les indices de qualité de la représentation des lignes ou des colonnes sur les sous-espaces principaux.
   (b) Comment s'interprète le premier facteur? Quel semble être la première cause de choix des études?
   (c) Quels choix d'études opposent les deux sous-nuages visibles Figure 4.11, à droite?
   (d) Sur cette même figure, quels choix d'études opposent les catégories situées aux deux extrémités de chaque sous-nuage (par exemple Fprli/Fouag; Mprli/Mouag)?
   (e) Par une lecture numérique du biplot, donner et interpréter les deux plus grandes valeurs approchées lisibles sur le graphique (une étiquette est centrée sur le point qu'elle représente ; on utilisera l'aggrandissement donné Figure 4.12).
   (f) Si on voulait étudier plus particulièrement la répartition des étudiants en droit, sur quel plan devrait-on faire cette étude?
   (g) Quels précautions faut-il prendre si on arrête l'interprétation aux deux premiers facteurs?

```
Pourcentage de inertie expliquée
```

	f1	f2	f3	f4	f5	tot
inertie exp	62	25	7	3	2	100
inertie cum	62	87	94	97	100	100

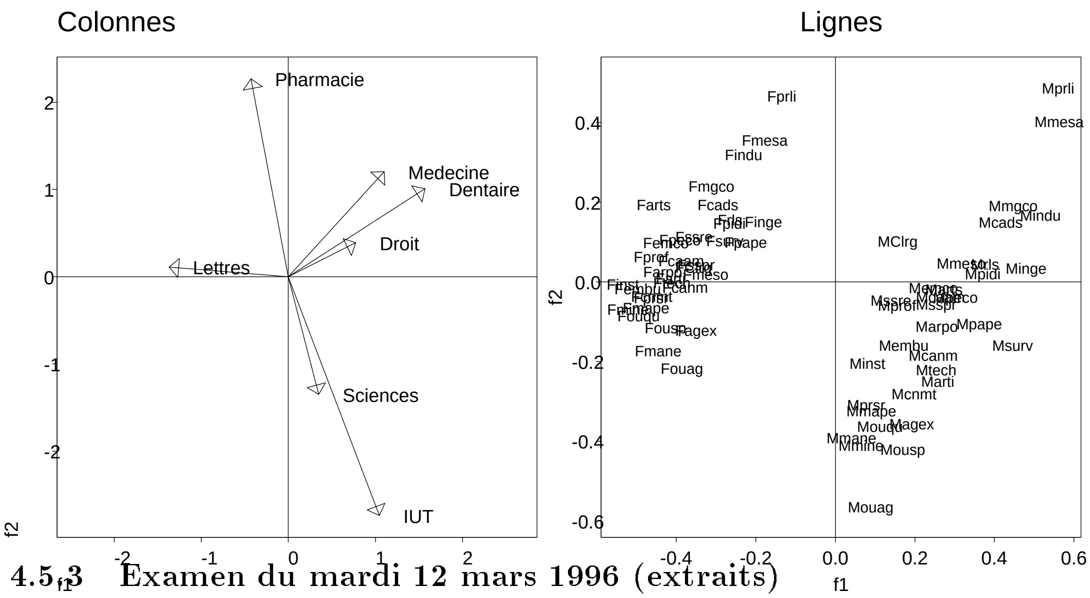


FIG. 4.11 - *Biplot isométrique-ligne (axes 1 et 2)*

4.5.3 Examen du mardi 12 mars 1996 (extraits)

On veillera à justifier les réponses.

Partie théorique

(10 points)

Les questions sont découpées de sorte qu'il est toujours possible, pour résoudre une question, d'utiliser les résultats d'une question précédente (même non démontrée)

On considère deux ensembles de variables centrées constituant les colonnes de deux matrices $X = (x^1, \ldots, x^p)$ $(n \times p)$ et $Y = (y^1, \ldots, y^q)$ $(n \times q)$. L'espace $\mathbf{R}^n$ est muni de la métrique des poids associée à la matrice $D = 1/nI_n$ (individus équipondérés), et la norme associée est notée $\|\ \|_D$. On note P_X et P_Y les orthoprojecteurs de $\mathbf{R}^n$ qui projettent respectivement sur $Im(X)$ et $Im(Y)$. On s'intéresse à la quantité $\mathcal{L}(X,Y) = trace(P_X P_Y)$. On va montrer que c'est une mesure de liaison entre les sous-espaces $Im(X)$ et $Im(Y)$ (on suppose dans la suite que X et Y sont de pleins rangs).

1. Montrer que $\mathcal{L}(X,Y)$ est une mesure symétrique en X et Y.

2. Rappeler l'expression de la trace d'un opérateur de l'espace euclidien $\mathbf{R}^n$ à partir de la donnée d'une base orthonormée (B.O.N.). Appliquer cette formule à l'opérateur $P_X P_Y$ en choisissant une B.O.N. (u_s) dont les premiers vecteurs engendrent $Im(Y)$.

3. En déduire que $trace(P_X P_Y) = \sum_{s=1}^{q} \|P_X(u_s)\|_D^2$ (justifier). Que vaut $\mathcal{L}(X,Y)$ si $Im(X) \perp Im(Y)$?, si $Im(Y) \subset Im(X)$?

4. Montrer que si $X = (x)$ et $Y = (y)$ (X et Y sont des matrices unicolonnes), alors $\mathcal{L}(X,Y) = cos(x,y)^2$.

5. On considère deux variables qualitatives $\mathcal{X}$ et $\mathcal{Y}$ et les matrices d'indicatrices associées X et Y (on a donc $x^i = 1_{[\mathcal{X}=i]}$, indicatrice de la modalité i de $\mathcal{X}$, et de même $y^j = 1_{[\mathcal{Y}=j]}$). On utilise les notations du cours d'analyse des correspondances : si N est le tableau de contingence croisant $\mathcal{X}$ et $\mathcal{Y}$, n est l'effectif total, alors $F = (f_{ij}) = N/n$, $f_{i.} = \sum_j f_{ij}$, $f_{.j} = \sum_i f_{ij}$, $D_I = diag(f_{i.})$; $D_J = diag(f_{.j})$. Montrer que $F = X'DY$, $D_I = X'DX$ et $D_J = Y'DY$.

6. On note $\|\ \|$ la norme matricielle définie sur l'espace des matrices $p \times q$ par $A \longrightarrow \|A\|^2 = trace(A'D_I A D_J)$. En déduire que pour ces matrices X et Y, $\mathcal{L}(X,Y) = \|D_i^{-1} F D_J^{-1}\|^2$. Montrer que le Φ^2 de Pearson s'écrit $\|D_i^{-1} F D_J^{-1} - E\|^2$, où E est la matrice $p \times q$ ne contenant que des 1. En déduire que $\mathcal{L}(X,Y) = 1 + \Phi^2$.

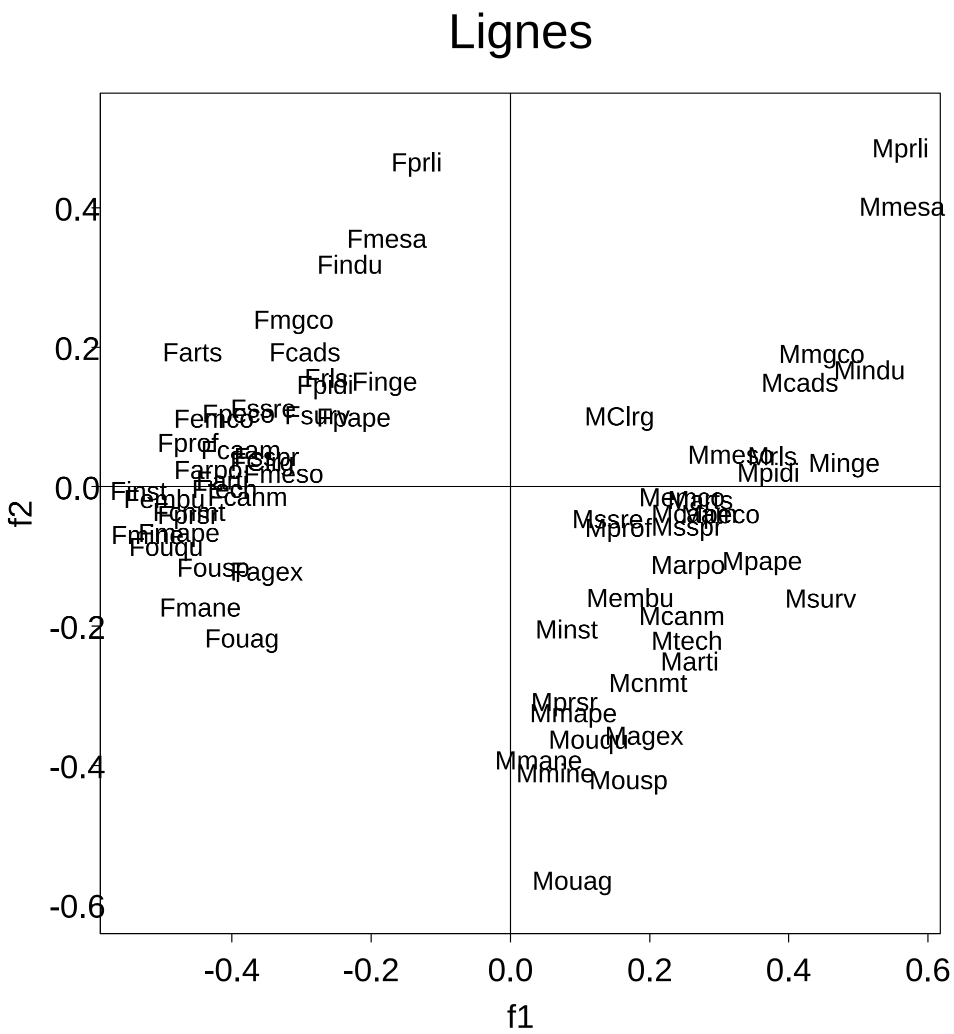


FIG. 4.12 - *Biplot isométrique-ligne (axes 1 et 2) : individus seuls (aggrandissement)*

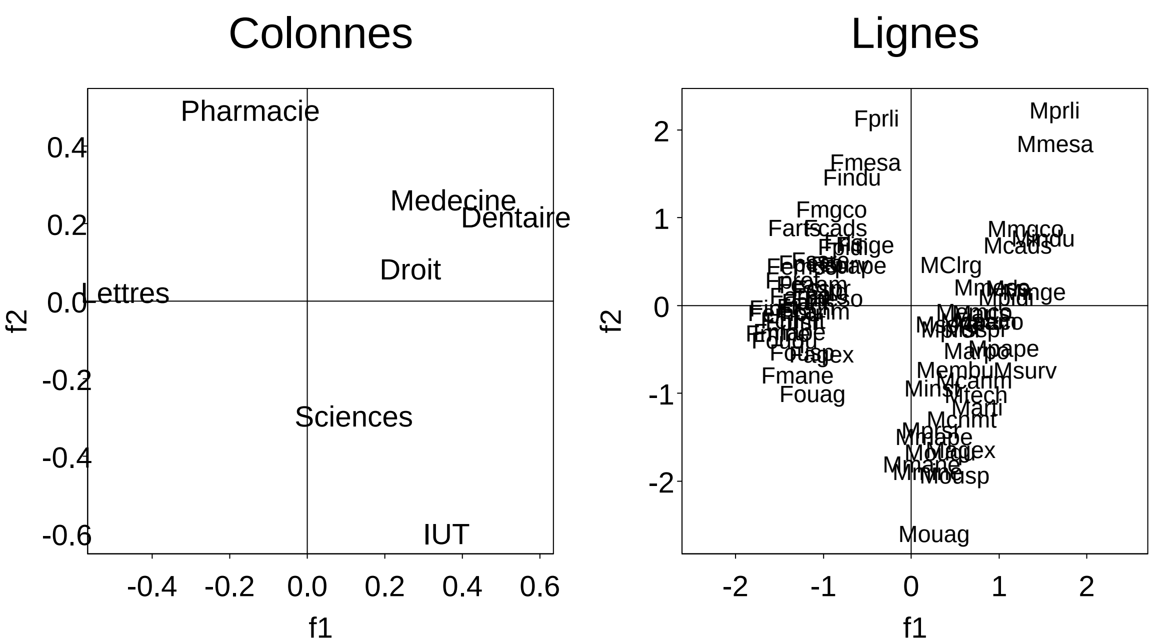


FIG. 4.13 - *Biplot isométrique-colonne (axes 1 et 2)*

	f1	f2	f3	f4	f5	Axes >5	TOT
Magex	209	711	14	62	0	4	1000
Mouag	21	853	21	91	13	0	1000
Mindu	810	87	72	10	20	0	1000
Marti	458	429	36	48	27	2	1000
Mpape	667	56	56	3	117	101	1000
Mmgco	778	143	7	35	27	10	1000
Mpeco	840	13	45	27	27	48	1000
Mprli	554	420	20	3	3	0	1000
Mrls	588	8	273	1	123	7	1000
Minge	880	5	57	33	24	1	1000
Mcads	790	104	95	10	0	0	1000
Mprof	345	47	235	225	146	4	1000
Mmesa	526	271	71	78	43	11	1000
Mmeso	838	19	70	68	5	0	1000
Mtech	494	368	127	1	4	5	1000
Mcaam	849	17	12	76	44	3	1000
Minst	47	294	475	91	78	15	1000
Mpidi	880	3	43	0	72	2	1000
Msurv	397	51	131	161	236	25	1000
Membu	507	431	51	9	2	0	1000
Memco	858	3	9	4	99	27	1000
Mcnmt	321	646	21	0	2	10	1000
Mouqu	83	864	1	32	18	1	1000
Mousp	121	738	27	57	55	2	1000
Mmine	20	780	165	0	20	15	1000
Mmape	66	845	54	0	5	29	1000
Mmane	10	956	24	2	0	7	1000
Mprsr	46	705	36	86	121	6	1000
Marts	346	2	597	7	34	14	1000
MClrg	83	35	406	387	34	55	1000
Marpo	595	111	214	44	29	7	1000
Mcanm	580	324	13	14	66	4	1000
Msspr	696	34	219	0	43	9	1000
Mssre	170	18	608	31	173	0	1000
Fagex	808	95	40	11	46	0	1000
Fouag	753	237	9	0	0	0	1000
Findu	286	552	6	157	0	0	1000
Farti	992	1	4	1	0	1	1000
Fpape	612	124	0	250	0	14	1000
Fmgco	552	331	22	94	1	0	1000
Fpeco	903	67	8	14	1	7	1000
Fprli	60	722	73	134	11	1	1000
Frls	635	226	19	20	97	3	1000
Finge	386	275	185	66	76	12	1000
Fcads	687	298	6	3	4	2	1000
Fprof	894	18	67	19	1	1	1000
Fmesa	171	697	66	1	55	10	1000
Fmeso	766	3	21	42	146	22	1000
Ftech	965	0	12	19	3	0	1000
Fcaam	955	19	2	14	10	0	1000
Finst	905	0	88	6	1	0	1000
Fpidi	602	185	112	19	78	3	1000
.....	....	....	....	....	....	....	....
.....	....	....	....	....	....	....	....
Fcanm	968	1	6	16	8	1	1000
Fsspr	748	12	119	70	50	1	1000
Fssre	623	64	236	55	21	1	1000
TOT	622	252	66	35	22	3	1000

TAB. 4.19 - *Contribution des facteurs aux lignes*

	Axe 1	Axes1a2	Axes1a3	Axes1a4	Axes1a5	Axes >5	TOT
Magex	209	921	935	996	996	1000	1000
Mouag	21	874	895	987	1000	1000	1000
Mindu	810	897	969	980	1000	1000	1000
Marti	458	887	923	971	998	1000	1000
Mpape	667	723	779	782	899	1000	1000
Mmgco	778	921	928	963	990	1000	1000
Mpeco	840	854	898	926	952	1000	1000
Mprli	554	974	994	997	1000	1000	1000
Mrls	588	596	869	870	993	1000	1000
Minge	880	885	942	975	999	1000	1000
Mcads	790	895	989	1000	1000	1000	1000
Mprof	345	392	626	851	996	1000	1000
Mmesa	526	797	867	946	989	1000	1000
Mmeso	838	856	927	994	1000	1000	1000
Mtech	494	862	989	991	995	1000	1000
Mcaam	849	866	877	953	997	1000	1000
Minst	47	341	816	907	985	1000	1000
Mpidi	880	883	926	926	998	1000	1000
Msurv	397	448	579	739	975	1000	1000
Membu	507	938	989	998	1000	1000	1000
Memco	858	861	870	874	973	1000	1000
Mcnmt	321	967	988	988	990	1000	1000
Mouqu	83	947	949	981	999	1000	1000
Mousp	121	859	886	943	998	1000	1000
Mmine	20	800	965	965	985	1000	1000
Mmape	66	911	965	965	971	1000	1000
Mmane	10	967	991	993	993	1000	1000
Mprsr	46	751	787	873	994	1000	1000
Marts	346	347	944	952	986	1000	1000
MClrg	83	118	525	912	945	1000	1000
Marpo	595	707	920	964	993	1000	1000
Mcanm	580	903	916	930	996	1000	1000
Msspr	696	729	949	949	991	1000	1000
Mssre	170	188	796	827	1000	1000	1000
Fagex	808	903	943	954	1000	1000	1000
Fouag	753	991	999	1000	1000	1000	1000
Findu	286	837	843	1000	1000	1000	1000
Farti	992	993	997	999	999	1000	1000
Fpape	612	736	736	986	986	1000	1000
Fmgco	552	883	905	999	1000	1000	1000
Fpeco	903	970	978	993	993	1000	1000
Fprli	60	782	855	988	999	1000	1000
Frls	635	861	880	900	997	1000	1000
Finge	386	662	846	912	988	1000	1000
Fcads	687	985	991	994	998	1000	1000
Fprof	894	911	978	997	999	1000	1000
Fmesa	171	868	934	935	990	1000	1000
Fmeso	766	769	790	832	978	1000	1000
Ftech	965	965	978	997	1000	1000	1000
Fcaam	955	974	976	990	1000	1000	1000
Finst	905	905	993	999	1000	1000	1000
Fpidi	602	788	900	919	997	1000	1000
.....	....	....	....	....	....	....	....
.....	....	....	....	....	....	....	....
Fcanm	968	969	975	992	999	1000	1000
Fsspr	748	760	879	949	999	1000	1000
Fssre	623	687	924	979	999	1000	1000
TOT	622	874	940	975	997	1000	1000

TAB. 4.20 - *Contributions cumulées des facteurs aux lignes*

	f1	f2	f3	f4	f5	f>5	ORI
Magex	11	96	7	60	0	34	13
Mouag	0	30	3	23	5	9	27
Mindu	27	7	22	6	19	21	24
Marti	12	27	8	22	19	16	11
Mpape	0	0	0	0	2	0	14
Mmgco	20	9	2	16	19	16	19
Mpeco	22	1	11	13	19	16	8
Mprli	118	221	40	10	18	132	41
Mrls	2	0	8	0	11	2	18
Minge	47	1	29	31	36	33	19
Mcads	89	29	100	21	0	70	16
Mprof	3	1	21	38	39	6	5
Mmesa	13	17	17	35	31	16	44
Mmeso	2	0	2	3	0	2	9
Mtech	8	14	19	0	2	10	10
Mcaam	25	1	3	40	37	19	6
Minst	1	15	94	34	46	13	10
Mpidi	5	0	2	0	12	4	11
Msurv	2	1	6	13	31	3	36
Membu	9	18	8	3	1	11	4
Memco	7	0	1	1	23	5	5
Mcnmt	4	19	2	0	1	8	9
Mouqu	3	70	0	19	17	20	11
Mousp	5	69	10	38	58	23	17
Mmine	0	15	12	0	4	5	16
Mmape	0	3	1	0	0	1	9
Mmane	0	22	2	0	0	6	12
Mprsr	0	8	2	7	16	3	10
Marts	1	0	9	0	1	1	16
MClrg	0	0	7	13	2	1	22
Marpo	8	4	26	10	11	8	8
Mcanm	16	22	3	7	50	17	8
Msspr	9	1	28	0	16	8	7
Mssre	5	1	184	18	156	20	8
Fagex	32	9	15	8	52	25	11
Fouag	4	3	0	0	0	4	14
Findu	5	23	1	47	0	10	14
Farti	25	0	1	1	0	16	13
Fpape	0	0	0	1	0	0	6
Fmgco	8	12	3	25	0	9	13
Fpeco	29	5	2	8	1	20	12
Fprli	5	143	55	191	25	50	22
Frls	1	1	0	0	3	1	8
Finge	6	10	27	18	32	9	6
Fcads	33	35	3	3	6	30	9
Fprof	26	1	18	10	1	18	18
Fmesa	1	10	4	0	9	4	13
Fmeso	2	0	0	2	9	1	10
Ftech	18	0	2	6	2	12	13
Fcaam	43	2	1	11	13	28	11
Finst	36	0	33	5	1	25	23
Fpidi	2	2	4	1	7	2	9
.....	....	....	....	....	....	....	....
.....	....	....	....	....	....	....	....
Fcanm	29	0	2	9	7	19	11
Fsspr	13	1	19	21	24	11	12
Fssre	26	7	92	41	24	26	15
TOT	1000	1000	1000	1000	1000	1000	1000

TAB. 4.21 - *Contribution des lignes aux facteurs*

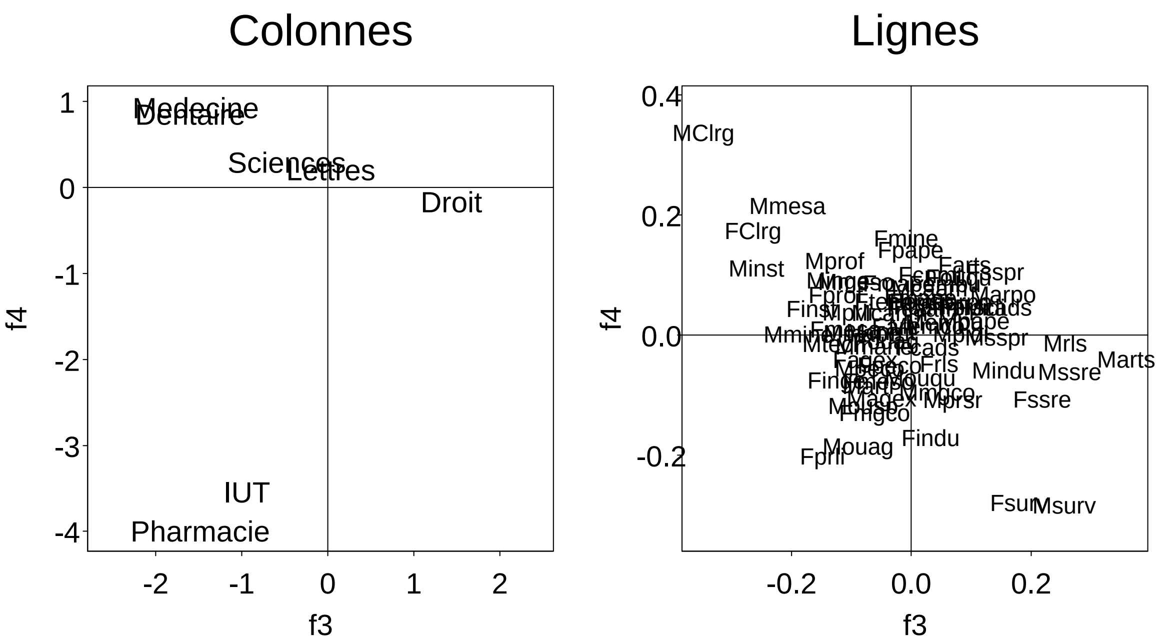


FIG. 4.14 - *Biplot isométrique-ligne (axes 3 et 4)*

	f1	f2	f3	f4	f5	f>5	TOT
Droit	680	70	249	2	0	0	1000
Sciences	131	791	26	6	47	0	1000
Lettres	993	3	0	1	3	0	1000
Medecine	571	277	118	23	10	1	1000
Dentaire	677	113	75	11	0	123	1000
Pharmacie	51	577	65	247	59	0	1000
IUT	199	553	17	127	104	0	1000
TOT	622	252	66	35	22	3	1000

TAB. 4.22 - *Contribution des facteurs aux colonnes*

	f1	f1a2	f1a3	f1a4	f1a5	f>5	TOT
Droit	680	750	998	1000	1000	1000	1000
Sciences	131	922	948	953	1000	1000	1000
Lettres	993	996	996	997	1000	1000	1000
Médecine	571	847	965	988	999	1000	1000
Dentaire	677	790	865	876	877	1000	1000
Pharmacie	51	628	693	941	1000	1000	1000
IUT	199	752	769	896	1000	1000	1000
TOT	622	874	940	975	997	1000	1000

TAB. 4.23 - *Contributions cumulées des facteurs aux colonnes*

	f1	f2	f3	f4	f5	f>5	TOT	ORI
Droit	156	40	537	6	0	0	143	47
Sciences	27	405	50	21	271	3	129	50
Lettres	608	4	0	15	46	3	381	102
Médecine	150	179	291	109	75	73	163	114
Dentaire	25	10	26	8	0	920	23	196
Pharmacie	6	169	72	522	197	1	74	194
IUT	28	193	23	320	411	0	88	296
	1000	1000	1000	1000	1000	1000	1000	1000

TAB. 4.24 - *Contribution des colonnes aux facteurs*

7. X étant défini comme précédemment, vérifier que $1 \in Im(X)$ et que si P_1 est l'orthoprojecteur sur la droite des constantes, $P_X - P_1$, noté P_X^0, est le projecteur sur le sous-espace obtenu à partir de $Im(X)$ par centrage. En appliquant cette propriété à Y, on note encore $P_Y^0 = P_Y - P_1$ et on note $Im(X^0)$ et $Im(Y^0)$ le sous-espaces image de ces deux projecteurs.

8. On considère la Décomposition aux Valeurs Singulières (DVS) de $P_X^0 P_Y^0$, pour les métriques D et D^{-1}, qui s'écrit :

$$P_X^0 P_Y^0 = \sum_{s=1}^{r} \lambda_s \xi_s \eta_s' D,$$

avec $\xi_s' D \xi_k = (\eta_s' D) D^{-1} (D \eta_k) = \delta_{sk}$ (δ de Kroonecker). Exprimer $P_X^{0\,\prime}$ à partir de P_X^0. En déduire que la DVS de $P_Y^0 P_X^0$ (pour les mêmes métriques) s'écrit :

$$P_Y^0 P_X^0 = \sum_{s=1}^{r} \lambda_s \eta_s \xi_s' D,$$

9. Montrer que $\xi_s \in Im(X)$, $\eta_s \in Im(Y)$ et qu'il existe u_s unique dans $\mathbf{R}^p$ et v_s unique dans $\mathbf{R}^q$ tels que $\xi_s = X u_s$ et $\eta_s = Y v_s$.

10. Montrer que $\{u_s, s = 1, \ldots, r\}$ et $\{v_s, s = 1, \ldots, r\}$ sont des familles orthonormées au sens de D_I et D_J respectivement.

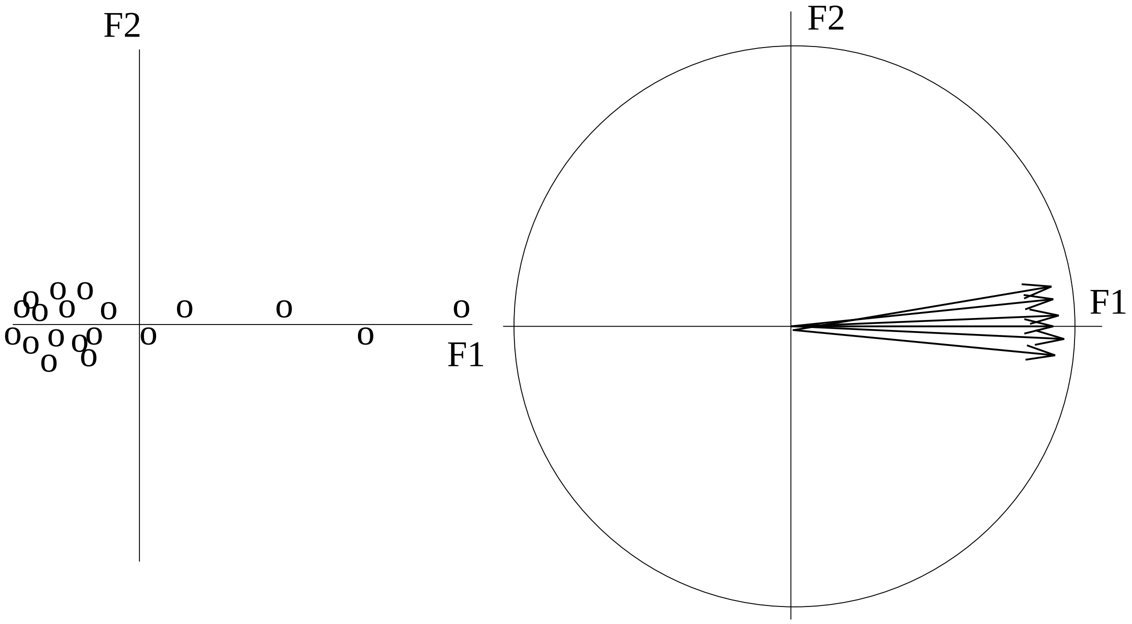
11. Montrer que $P_X^0 P_Y^0 = P_X P_Y - P_1$. En remplaçant P_X^0 et P_Y^0 par leurs expressions en fonction de X, Y, et D, déduire de la DVS de $P_X^0 P_Y^0$ celle de l'AFC de N, qui s'écrit :

$$D_I^{-1} F D_J^{-1} - E = \sum_{s=1}^{r} \lambda_s u_s v_s' \qquad \text{pour les métriques } D_I \text{ et } D_J.$$

En déduire les relations existant entre la DVS de $P_X^0 P_Y^0$ et l'analyse des correspondances de N.

Exercice 4.5.1 *(Examen du 17 Septembre 1996 : extrait)*

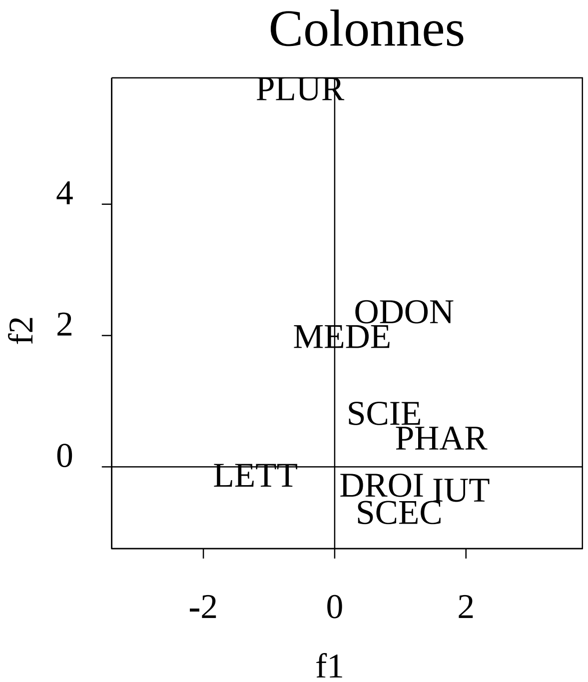
On a fait l'ACP d'une table de contingence contenant les effectifs des populations des communes périphériques de Toulouse (en ligne) classées par catégories socio-professionnelles du chef de ménage (en colonne). On a les représentations isométriques suivantes:

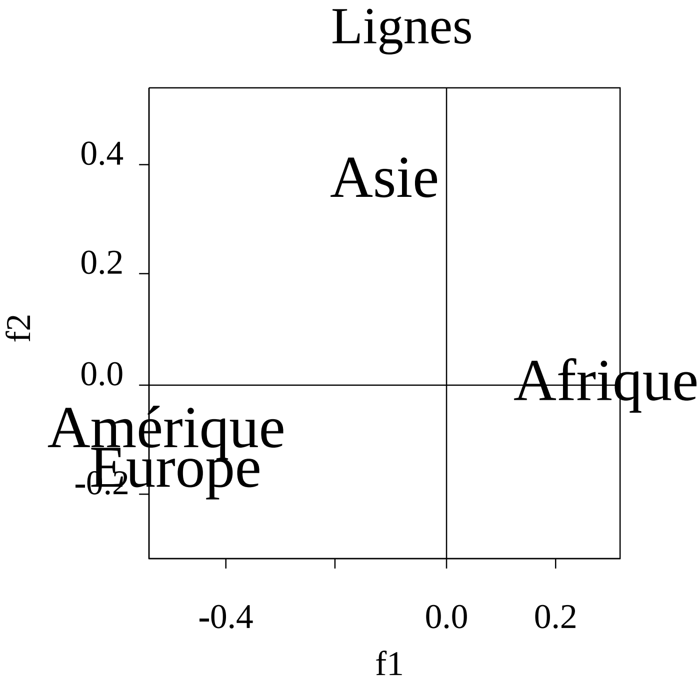


Que dire de cette analyse? Quelle interprétation a le premier facteur. Que proposeriez vous de faire par la suite et pourquoi?

Exercice 4.5.2 *(Examen du 17 Septembre 1996 : extrait)* On dispose du nombre d'étudiants étrangers en France dans les différents cycles d'études de l'enseignemnet supérieur (droit, sciences éco, lettres, sciences, médecine, odontologie, pluridisciplinaire, IUT) en fonction de leurs origines (Europe, Asie, Afrique, Amérique). On fait l'AFC de ce tableau. Les deux premiers facteurs expliquent 99% de l'inertie. Quel est le nombre maximum de facteurs que l'on peut avoir ici? (Justifier.) On obtient les résultats suivants sur le premier plan principal (représentation biplot isométrique ligne ; on suppose qu'un label est centré sur le marqueur qu'il représente). Faîtes une interprétation rapide de ces résultats. Quels sont les indice que l'on peut lire numériquement sur ce dessin en utilisant les points PLUridisciplinaire et Asie? Faites cette lecture numérique. Sachant que les étudiants pluridisciplinaires représentent 1,31% de l'effectif total des étudiants, quelle est le pourcentage de ce type d'étude chez les étudiants d'origine asiatique (à la date où ont été relevées ces mesures).

Sachant que les étudiants originaires d'Asie représentent 18,3% de l'ensemble des étudiants étrangers, quelle proportion des étudiants pluridisciplinaires représentent-ils? (Ces quantités seront données en approximation de rang 2).





4.6 L'Analyse (Factorielle) des correspondances multiples (AFCM)

L'analyse des correspondances multiples est une extension de l'AFC à plus de deux variables qualitatives. Elle est, aux variables qualitatives, ce que l'ACP est aux variables quantitatives. Elle fournit des facteurs qui résument "au mieux" les liaisons existantes entre les variables deux à deux. Elle peut aussi servir d'étape intermédiaire avant une analyse discriminante (méthode DISQUAL par exemple), car elle fournit une base orthonormée d'un ensemble de variables indicatrices. Pour étudier ses relations avec l'AFC, nous nous limitons dans un premier temps au cas de deux variables qualitatives, puis nous l'"étendons à K variables.

4.6.1 Cas de deux variables qualitatives

Soit $\mathcal{X}$ et $\mathcal{Y}$ deux variables qualitatives ayant respectivement m_1 et m_2 modalités. Ces modalités sont notées $i = 1, \ldots, m_1$ et $j = 1, \ldots, m_2$. Une série d'observations de ces variables sur n individus peut se représenter par trois types de tableau:

- Un tableau de contingence $N = (n_{ij})$, de dimension $m_1 \times m_2$ avec n_{ij} égal au nombre d'individus ayant la modalité i de la première variable et la modalité j de la deuxième variable.

- Un tableau d'indicatrices Z de dimension $n \times (m_1 + m_2)$;
- Le tableau de Burt $B = Z'Z$, tableau carré de dimension $(m_1 + m_2) \times (m_1 + m_2)$.

L'AFC effectuée sur le tableau N étant l'AFC classique, on étudie dans un premier temps, les relations entre les éléments singuliers produits par les AFC effectuées sur ces trois tableaux.

Dans la suite on utilise le vocabulaire de la DVS et on appelle vecteurs singuliers à gauche les facteur réduits et vecteurs singuliers à droite les vecteurs propres. Les résultats peuvent être disposés sous forme de tableau de la façon suivante :

Tableau analysé	Valeur singulière	vecteurs singuliers
N	$\lambda_s,\ s = 1, \ldots, r$	à gauche $\mathbf{u}_s$ à droite $\mathbf{v}_s$
B	$\frac{1+\lambda_s}{2},\ s = 1, \ldots, r,$ $\frac{1-\lambda_s}{2}, s = 1, \ldots, r$ $1/2$ $1/2$	à gauche $\begin{pmatrix} \mathbf{u}_s \\ \mathbf{v}_s \end{pmatrix}$, $\begin{pmatrix} \mathbf{u}_s \\ -\mathbf{v}_s \end{pmatrix}$, $\begin{pmatrix} \sqrt{2}\mathbf{u}_s \\ 0 \end{pmatrix}$ avec $\mathbf{u}_s \perp Im(D_I^{-1}F)$ $\begin{pmatrix} 0 \\ \sqrt{2}\mathbf{v}_s \end{pmatrix}$ avec $\mathbf{v}_s \perp Im(D_J^{-1}F')$ à droite idem (par symétrie)
Z	$\mu_s = \sqrt{\frac{1+\lambda_s}{2}},\ s = 1, \ldots, r$ $\mu_s = \sqrt{\frac{1-\lambda_s}{2}},\ s = 1, \ldots, r$ et $\sqrt{1/2}$	à droite $\begin{pmatrix} \mathbf{u}_s \\ \mathbf{v}_s \end{pmatrix}$, $\begin{pmatrix} \mathbf{u}_s \\ -\mathbf{v}_s \end{pmatrix}$, (mêmes vecteurs que ceux associés à $1/2$ ci-dessus) **à gauche** variables prenant la valeur $\frac{u_{is}+v_{js}}{2\mu_s}$ pour l'individu ayant les modalités i de la première variable et j de la seconde variable

Remarque 4.6.1 *Les rangs de Z et $B = Z'Z$ sont égaux, mais au moins de deux fois supérieurs à celui de N. Des valeurs propres artificielles apparaissent donc dans ces analyses et doivent être écartées. Il s'agit des valeurs propres inférieures ou égales à $(1/2)^2$ (B) ou $1/2$ (Z). Les valeurs propres à considérer sont strictement supérieures à cette borne.*

Concrètement, les nuages de points issus de ces différentes analyses ont des formes semblables. Il est cependant à noter que dans l'AFC de N, la représentation est simultanée, et que dans les AFC de B et de Z, on ne représente que le nuage des $m_1 + m_2$ colonnes. Le nuage apparaît plus contracté sur l'axe horizontal quand on passe de l'analyse de N à B puis de celle de B à Z. L'inertie expliquée par les axes est aussi décroissante dans cet ordre. Cependant les différentes représentations sont d'autant plus ressemblante que les valeurs propres associées aux axes sont peu différentes. On donne figure 4.15 les représentations simultanées associées à ces trois analyses.

Démonstration On reprend les notations de l'AFC. On note $F = 1/n\ N$ le tableau de fréquence d'éléments $f_{ij} = n_{ij}/n$, $D_I = diag(f_{i.}, i = 1, \ldots, m_1)$ et $D_J = diag(f_{.j}, j = 1, \ldots, m_2)$. On définit les tableaux de fréquence F_B et F_Z respectivement associés à B et à

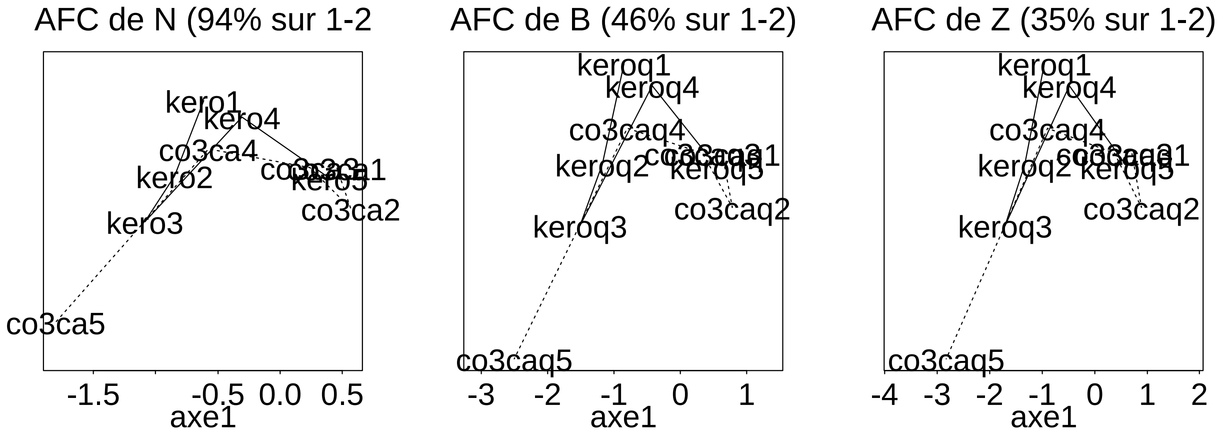


FIG. 4.15 - *Comparaison des* AFC *des tableaux N, B, et Z*

Z. La matrice diagonale des poids des lignes ou des colonnes de B est noté D_B. Les matrices des poids des lignes et des colonnes de Z sont $1/n\ I_n$ et D_B. Ces matrices s'écrivent par bloc sous la forme suivante :

$$B = \left(\begin{array}{c|c} nD_I & N \\ \hline N' & nD_J \end{array}\right) \ ; \ F_B = \frac{1}{4}\left(\begin{array}{c|c} D_I & F \\ \hline F' & D_J \end{array}\right) \ ; \ D_B = \frac{1}{2}\left(\begin{array}{c|c} D_I & 0 \\ \hline 0 & D_J \end{array}\right) ; F_Z = \frac{1}{2n} Z.$$

Dans la suite, on dérive l'afc de la DVS du tableau d'indices de liaison non centrés (cf. propriété 7, paragraphe 4.3.2).

Relation entre l'AFC de N et l'AFC de B
L'AFC de N s'obtient par la DVS de $D_I^{-1} F D_J^{-1}$ pour les métriques D_I et D_J qui s'écrit:

$$D_I^{-1} F D_J^{-1} = \sum_{s=0}^{r} \lambda_s \mathbf{u}_s \mathbf{v}_s' \ , \left\{ \begin{array}{ll} \text{avec } \{\mathbf{u}_s\} & \text{famille } D_I\text{-orthonormée dans } \mathbf{R}^{m_1} \\ \text{et } \{\mathbf{v}_s\} & \text{famille } D_J\text{-orthonormée dans } \mathbf{R}^{m_2} \end{array} \right. \qquad (4.5)$$

On complète *si nécessaire*[3]les familles $\{\mathbf{u}_s\}$ et $\{\mathbf{v}_s\}$ pour obtenir des bases orthonormées des espaces vectoriels $\mathbf{R}^{m_1}$ et $\mathbf{R}^{m_2}$. Alors la matrice identité I_{m_1} (resp. I_{m_2}) est la somme des orthoprojecteurs sur les sous-espaces engendrés par les vecteurs de la base orthonormée $\{\mathbf{u}_s\}$ de $\mathbf{R}^{m_1}$ (resp. $\{\mathbf{v}_s\}$ de $\mathbf{R}^{m_2}$). Ces deux décompositions sont équivalentes aux deux équations qui suivent et s'obtiennent, à partir de celles-ci, par simplification et multiplication à droite par D_I (resp. D_J).

$$\begin{aligned} D_I^{-1} &= \sum_{s=0}^{r} \frac{1+\lambda_s}{2} \mathbf{u}_s \mathbf{u}_s' + \sum_{s=0}^{r} \frac{1-\lambda_s}{2} \mathbf{u}_s \mathbf{u}_s' + \sum_{s=r+1}^{m_1} \mathbf{u}_s \mathbf{u}_s' \qquad (4.6) \\ D_J^{-1} &= \sum_{s=0}^{r} \frac{1+\lambda_s}{2} \mathbf{v}_s \mathbf{v}_s' + \sum_{s=0}^{r} \frac{1-\lambda_s}{2} \mathbf{v}_s \mathbf{v}_s' + \sum_{s=r+1}^{m_2} \mathbf{v}_s \mathbf{v}_s' \qquad (4.7) \end{aligned}$$

Enfin l'équation suivante :

$$D_I^{-1} F D_J^{-1} = \sum_{s=0}^{r} \frac{1+\lambda_s}{2} \mathbf{u}_s \mathbf{v}_s' - \sum_{s=0}^{r} \frac{1-\lambda_s}{2} \mathbf{u}_s \mathbf{v}_s' \qquad (4.8)$$

se simplifie pour donner l'équation équivalente (4.5). On a donc déduit de l'AFC de N les trois équations (4.6), (4.7), et (4.8).
L'AFC de B peut s'obtenir par la DVS de $D_B^{-1} F_B D_B^{-1}$ pour les métriques D_B et D_B. Les membres de droite des trois équations (4.6), (4.7), et (4.8) sont trois des quatre blocs de la matrice $D_B^{-1} F_B D_B^{-1}$ (le quatrième s'obtenant par transposition du troisième). En réécrivant

3. Si cette complétion est nécessaire, on voit apparaître des valeurs propres égales à $(1/2)^2$ (AFC de B) où $1/2$ (AFC de Z). Dans le cas contraire, ces valeurs propres n'apparaissent pas

les décompositions de ces blocs obtenus dans les trois équations précédentes, on obtient la relation :

$$\begin{aligned} D_B^{-1} F_B D_B^{-1} &= \sum_{s=0}^{r} \frac{1+\lambda_s}{2} \begin{pmatrix} \mathbf{u}_s \\ \mathbf{v}_s \end{pmatrix} \begin{pmatrix} \mathbf{u}_s \\ \mathbf{v}_s \end{pmatrix}' + \sum_{s=0}^{r} \frac{1-\lambda_s}{2} \begin{pmatrix} \mathbf{u}_s \\ -\mathbf{v}_s \end{pmatrix} \begin{pmatrix} \mathbf{u}_s \\ -\mathbf{v}_s \end{pmatrix}' \\ &+ \sum_{s=r+1}^{m_1} \frac{1}{2} \begin{pmatrix} \sqrt{2}\mathbf{u}_s \\ 0 \end{pmatrix} \begin{pmatrix} \sqrt{2}\mathbf{u}_s \\ 0 \end{pmatrix}' + \sum_{s=r+1}^{m_2} \frac{1}{2} \begin{pmatrix} 0 \\ \sqrt{2}\mathbf{v}_s \end{pmatrix} \begin{pmatrix} 0 \\ \sqrt{2}\mathbf{v}_s \end{pmatrix}' \end{aligned} \tag{4.9}$$

qui n'est autre que la DVS cherchée de B, à condition de réordonner les triplets singuliers (on peut vérifier que les conditions d'orthonormalité sont bien vérifiées). Ceci achève la démonstration des liaisons existantes entre l'AFC de N et de B.

Relation entre l'AFC de Z et l'AFC de B
L'AFC de Z s'obtient par la DVS de $X = n\, F_Z D_B^{-1} = 1/2\; Z D_B^{-1}$, pour les métriques $1/nI_n$ et D_B^{-1}. Cette DVS a les mêmes vecteurs propres à droite et des valeurs singulières égales aux racines de celles de la DVS (4.9). Ses vecteurs singuliers à gauche s'obtiennent par la formule de transition. Ce résultat s'obtient en remarquant que si $X = n\, F_Z D_B^{-1}$ est la matrice dont on fait la DVS dans l'AFC de Z, alors

$$X'(1/nI_n)X = D_B^{-1}(n/(4n^2)Z'Z))D_B^{-1} = D_B^{-1} F_B D_B^{-1}$$

est celle dont on fait la DVS dans l'AFC de B. Remarquons que la formule de transition a ici une forme particulièrement simple, puisqu'il y a exactement deux termes non nuls et égaux à $1/2$ dans le profil ligne i. Ainsi, la valeur du facteur non-réduit pour un individu i est la demi-somme de la valeur du facteur réduit de même rang pour les deux modalités qu'il atteint.

Remarque 4.6.2 *De $B = Z'Z$, on déduit que la matrice $D_B^{-1} F_B D_B^{-1}$ est semi-définie positve et que ses vecteurs singuliers à droite et à gauche sont les mêmes. Les deux représentations isométriques des lignes et des colonnes sont donc identiques. La représentation simultanée se réduit à une unique représentation.*

4.6.2 Cas de plus de deux variables qualitatives

Lorsqu'on dispose de plus de deux variables qualitatives, il reste possible de faire l'AFC sur le tableau de Burt B ou sur le tableau d'indicatrices accolées Z, mais bien sûr plus sur le tableau N. On appelle Analyse des Correspondances Multiples l'une de ces deux analyses. La démonstration effectuée précédemment sur les liens entre ces deux AFC reste valable pour plus de deux variables. Pratiquement,

- on effectue l'analyse des correspondances sur tableau B (économie de place mémoire et de temps),
- on étudie les propriétés de l'analyse principalement sur le tableau Z (la simplicité de la construction de la matrice Z facilite cette étude).
- on interprète l'analyse comme une AFC de B ou de Z.

Remarque 4.6.3 *Dans les faits, si la représentation des individus est naturelle dans l'AFC de Z, elle peut encore être effectuée dans l'AFC de B par le biais d'individus supplémentaires. En fait le nombre d'individus dans ce type d'analyse (souvent quelques centaines, voir milliers) interdit de les représenter effectivement, autrement que par l'intermédiaire de barycentres (cf. § 4.6.4).*

4.6.3 Étude des propriétés de l'AFC du tableau Z

Pour formuler ces propriétés, on utilise le vocable *question* au lieu de variable qualitative et celui de réponse au lieu de *modalités*.

Propriétés barycentriques

Une première propriété concerne les sous ensembles de modalités définis par une même question. Les deux suivantes sont relatives aux représentations barycentriques définies § ??.

- Le barycentre des marqueurs des réponses d'une même question, pondérées par les effectifs des réponses, est le barycentre global (donc l'origine dans les représentations). Ainsi, lorsqu'une variable a deux modalités, les marqueurs sont aux extrémités d'un segment de droite contenant l'origine, les longueurs des segments étant inversement proportionnels aux effectifs de réponses.
- Dans la représentation ayant les réponses comme barycentre, une réponse est au barycentre des individus qui l'ont donnée,
- Dans la représentation ayant les individus comme barycentre, un individu est au barycentre des réponses qu'il a données.

Ces propriétés restent vraies dans l'AFC du tableau B.

Distances et contributions

On note n_{j_k} le nombre d'individus ayant donné la réponse j_k ($j_k = 1, \ldots, m_k$) à la question k ($k = 1, \ldots, K$) et $m = \sum_{k=1}^{K} m_k$ le nombre total de réponses (nombre de colonnes de la matrice Z, et ordre de la matrice carrée B). On considère l'AFC comme la double ACP (cf. § 24) des profils lignes et des profils colonnes. Dans le tableau Z, remarquons que le nombre total de 1 par lignes est égal à K. Ainsi dans l'AFC sur tableau Z, le poids des lignes est constant et un profil ligne est constitué de $m - K$ zéros et de K $1/K$. La métrique étant définie par la matrice $diag(1/n_{j_k}, j = 1, \ldots, m_k, k = 1, \ldots, K)$, on en déduit les valeurs des distances inter-individuelles. De l'AFC considérée comme une ACP sur les profils colonnes, on déduit le calcul de la distance entre deux réponses, et les contributions des réponses, puis des questions, à l'inertie totale.

La distance entre deux individus i et i' est

$$d^2(i,i') = \sum_{k \in \Delta_{i,i'}} \frac{1}{m^2} \left(\frac{1}{n_{j_k}} + \frac{1}{n_{j'_k}} \right)$$

avec $\Delta_{i,i'} \subset \{1, \ldots, K\}$, ensemble des questions auxquels i et i' ont donné des réponses différentes $j_k \neq j'_k$. Deux individus sont d'autant plus éloignés qu'ils ont donné à de nombreuses questions des réponses différentes et rares.

La distance entre deux modalités k de la question j et k' de la question j' est

$$d^2(j,j') = \frac{n}{n_{j_k} n_{j'_{k'}}} \delta_{k,k'}$$

avec $\delta_{k,k'}$, nombre d'individus ayant répondu k à la question j mais pas k' à la question j', ou ayant répondu k' à la question j', mais pas k à la question j. Deux réponses sont d'autant plus éloignées qu'elles sont rares et qu'elles sont données par des individus differents.

Contribution d'une modalité à l'inertie: la contribution de la modalité k de la question j à l'inertie est égale à

$$\frac{1}{K}\left(1 - \frac{n_{j_k}}{n}\right)$$

Cette contribution est d'autant plus importante que la réponse est rare.

Contribution d'une question à l'inertie: la contribution d'une question j à l'inertie est égale à $\frac{1}{K}(m_k - 1)$. La contribution d'une question est proportionnelle au nombre de réponses possibles diminué d'une unité. L'inertie totale est égale à $m/K - 1$.

Conséquences : on déduit de tous ces indices :

- Que la présence de modalités rares risque de gêner l'analyse. Lorsque cela est possible, il faut tenter de choisir des modalités avec des effectifs comparables. On ne garde des modalités à faible effectif que lorsque l'occurrence d'une telle modalité est un événement important.
- Que la présence de questions avec des nombres de modalité très différents revient à donner à ces questions une grande importance. Il faut donc si possible choisir des ensembles de réponses d'effectifs assez semblables.

4.6.4 Interprétation d'une AFCM

Choix de l'ensemble de variables actives

Une AFCM tendant à résumer les liaisons existantes entre les couples de variables, il peut être utile de mesurer ces liaisons, par exemple en calculant les χ^2 entre les paires de variables. Si ces χ^2 ne sont pas significatifs, il est probable qu'il n'y a pas de liaison à résumer. Si une variable n'est liée à aucune autre, elle pourra être supprimée de l'analyse (bien que ce point de vue puisse être contesté). De façon générale, on aura intérêt à choisir des variables sensées être liées à un phénomène d'intérêt et qui respectent donc une certaine homogénéité. Si le choix est bien fait, le résumé donné par l'AFCM sur les premiers facteurs sera très lié au phénomène d'intérêt, ce qui pourra être mis en évidence par l'adjonction de variables illustratives.

Interprétation d'un facteur

Un facteur s'interprète en fonction des questions et des réponses qui y ont le plus contribué (la contribution d'une question est la somme des contributions de ses réponses). Il s'agit toujours, comme en AFC, de facteurs marquant des oppositions.

Interprétation du biplot et de la représentation simultanée

Remarque préliminaire : lorsqu'on fait l'AFC sur tableau B, la DVS associée est symétrique. Les deux représentations constituant un biplot contiennent donc la même information. Plutôt que d'effectuer deux représentations, il est conseillé d'effectuer la représentation du biplot conservant cette symétrie (cf. § 3.8.2, expression (3.8)). Ainsi, l'interprétation du biplot pourra se faire entre deux modalités à l'intérieur d'un graphe unique, puisque les deux graphes sont identiques.

D'une façon générale, la méthode d'interprétation est semblable à celle d'une AFC. Les quantités approximées par le biplot entre une modalité k d'une question j et k' d'une question j' est le rapport de la fréquence de l'événement conjoint sur le produit des fréquences marginales :

$$\frac{\Pr[\text{ répondre } k \text{ à } j \text{ et } k' \text{ à } j']}{\Pr[\text{ répondre } k \text{ à } j]\,\Pr[\text{ répondre } k' \text{ à } j']} \tag{4.10}$$

Donc si le produit scalaire défini par deux modalités est grand et positif, on en déduira que ces modalités ont tendance à être liées positivement (attirance). Si ce produit scalaire est grand en valeur absolue, mais négatif, ces modalités ont tendance encore à être liées, mais cette fois par la présence d'une “répulsion” entre les deux modalités.

Lorsqu'il ne s'agit pas d'une représentation de type biplot, l'interprétation se fait de la même façon, mais avec moins de précision dans l'estimation au jugé des produits scalaires.

Utilisation de variables illustratives

On enrichit l'interprétation d'une AFCM en rajoutant des variables *supplémentaires* ou *illustratives*. Une réponse étant au barycentre des individus qui l'ont donnée, il est facile de représenter par ce moyen des réponses supplémentaires. Si ces réponses sont suffisamment éloignées du barycentre, on en déduit que cette réponse est liée aux variables actives, et plus précisément, au résumé qui en est fait par les premiers facteurs.

4.7 Exemple

4.7.1 Données et premiers graphiques

On utilise les données sur les chiens présentés § 4.1. On prend comme variables actives l'ensemble des variables caractérisant les chiens excepté la variable "fonction" (compagnie, chasse garde) qui sera mise en variable illustrative. On donne les représentations de l'éboulis des valeurs propres (4.16), et des modalités dans l'afc de B (biplot) et de Z (cf. figures 4.17 et 4.18. Ici, deux facteurs suffisent à interpréter l'analyse (forte accélération de la diminution des valeurs propres après le rang 2). On voit de plus que les graphiques des modalités issus des analyses sur B et Z sont d'aspect très semblables. Dans l'interprétation d'une AFCM, on ne se soucie pas de l'origine précise du graphique, excepté lorsqu'on en fait une interprétation quantitative, par le biplot (réponse à la question "Quelles approximations sont lisibles sur ce graphique"). On donne aussi la représentation des individus dans l'AFC de Z (figure 4.18, partie droite).

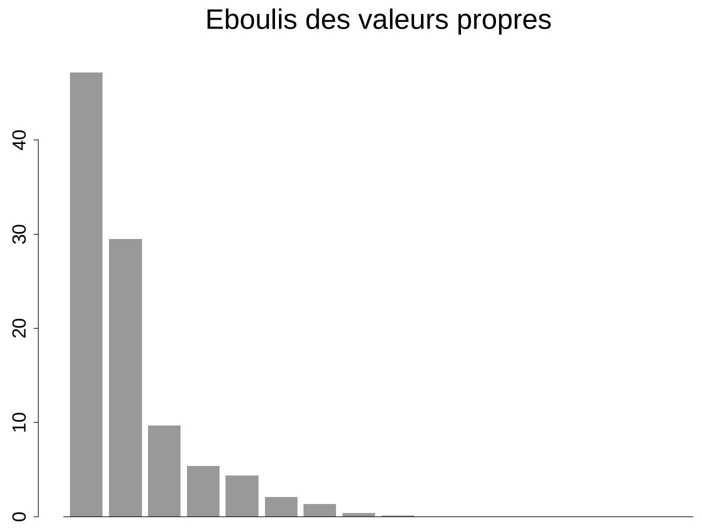


FIG. 4.16 - *Éboulis des valeurs propres (afc de B)*

4.7.2 Interprétation des facteurs et analyse simultanée

Pour interpréter un facteur, on regarde les contributions des modalités et des questions à ce facteur (Tableau 4.25), ainsi que les représentations des modalités. On voit que le premier facteur est très influencé par les variables `taille`, `poids`, `affectivité` et qu'il oppose les races légères, petites ou moyennes, et affectueuses aux races rapides, de grande taille et peu affectueuses. Le deuxième facteur est influencé par le poids, la rapidité, la taille et l'intelligence. Il oppose les chiens lents, lourds ou petits, aux chiens moyens (taille, vitesse et poids). L'analyse du plan 1-2 montre les liens existants entre les modalités, qui sont parfois des Lapalissades : les petits chiens ne pèsent *souvent* pas lourds, les gros chiens sont *souvent* peu affectueux, agressifs et rapides, les chiens de taille moyenne sont *souvent* moyennement rapides et de poids moyen. Le terme "souvent" signifie ici "plus souvent que l'ensemble des races". Il faut éviter (même si ici ces relations sont ici quasi systématiques) de déduire des conclusions péremptoires de telles analyses. Sauf quand on utilise des moyens de lecture quantitatifs du type biplot, on ne peut donner que des donnant le signe et les importances relatives des ratios approximés par l'AFC. (4.10) (cf. § ??, propriété 4.4).

4.7.3 Variables illustratives

On donne enfin la représentation de modalités supplémentaires, celles de la variable `fonction` (cf. figure 4.19). On voit que les chiens de chasse sont plus souvent intelligents, de poids moyen et rapides, que les chiens de compagnie sont plutôt petits, légers, et affectueux, et que les chiens de garde sont plutôt gros, lourds et agressifs. Pour savoir si la liaison entre

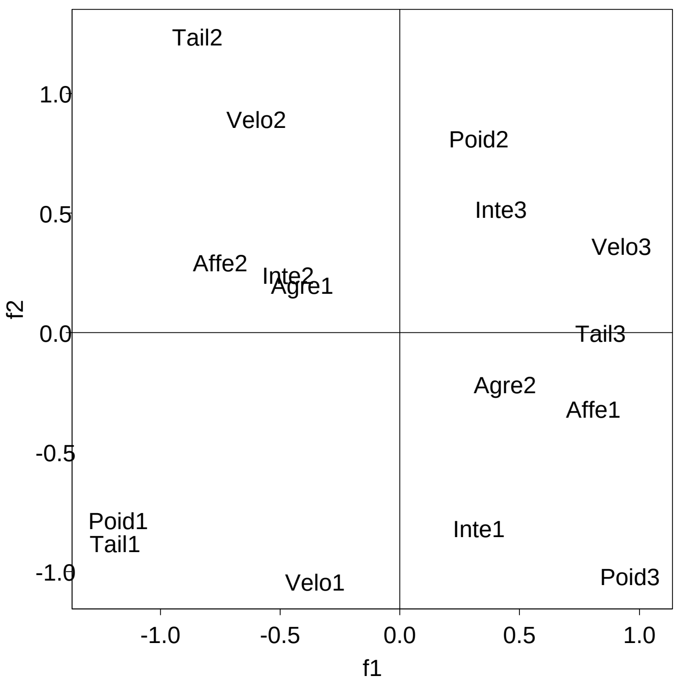


FIG. 4.17 - *Représentation des modalités dans l'afc de B (biplot)*

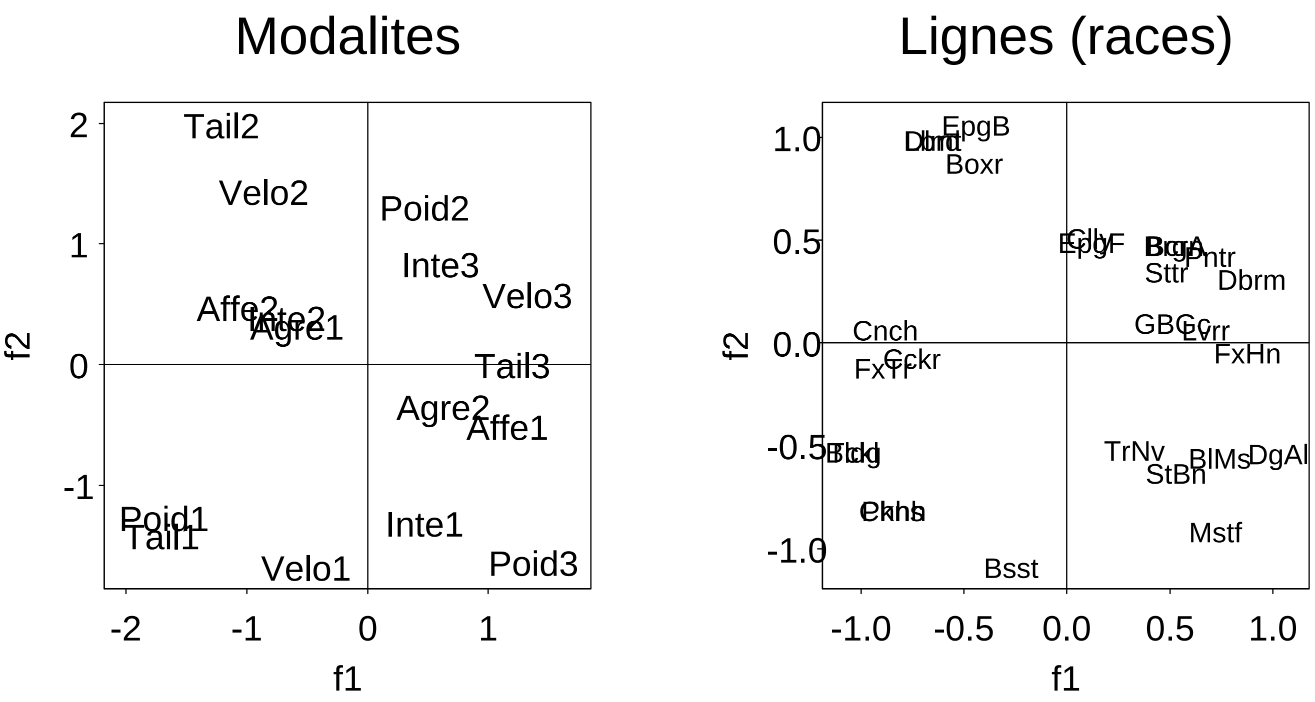


FIG. 4.18 - *Représentation des modalités dans l'afc de Z*

cette nouvelle variable est significative, on peut remarquer que, dans cette représentation non isométrique, les facteurs sont centrés et réduits, donc de variance 1. La valeur moyenne d'un facteur sur n_{j_k} individus choisis au hasard suit donc une loi de moyenne 0 et de variance $1/n_{j_k}$. Ici les effectifs des trois modalités sont respectivement : compagnie (10), chasse (9), et garde (8). Les facteurs ont donc un écart-type de l'ordre de 1/3. Les coordonnées de `compagnie` et `chasse` étant supérieures à deux écart-types, ces liaisons apparaissent comme significatives.

	f1	f2	f3	f4	f5	f¿5	TOT	ORI
Tail1	126	87	87	0	1	29	94	114
Tail2	45	122	135	38	22	99	77	131
Tail3	133	0	0	15	10	9	64	36
Taill	304	207	222	53	33	137	235	281
Poid1	140	79	42	9	0	16	95	100
Poid2	19	148	14	13	21	9	56	34
Poid3	58	83	207	5	66	58	77	131
Poids	217	310	263	27	87	83	228	265
Velo1	16	174	36	0	45	44	66	56
Velo2	36	102	35	12	94	171	62	66
Velo3	98	19	141	14	4	91	70	66
Veloc	152	295	213	26	143	306	198	181
Inte1	11	85	21	4	416	47	53	56
Inte2	33	11	132	91	34	109	42	30
Inte3	16	30	104	212	201	81	50	60
Intel	60	126	257	307	651	237	145	146
Affe1	108	21	11	27	12	55	62	40
Affe2	100	19	10	25	11	51	58	35
Affec	208	40	21	52	23	106	120	75
Agre1	30	9	13	258	29	62	35	21
Agre2	32	10	14	277	32	67	38	25
Agres	62	19	27	535	61	129	73	46
	1000	1000	1000	1000	1000	1000	1000	1000

TAB. 4.25 - *Contribution des modalités et des questions aux facteurs et à l'inertie totale*

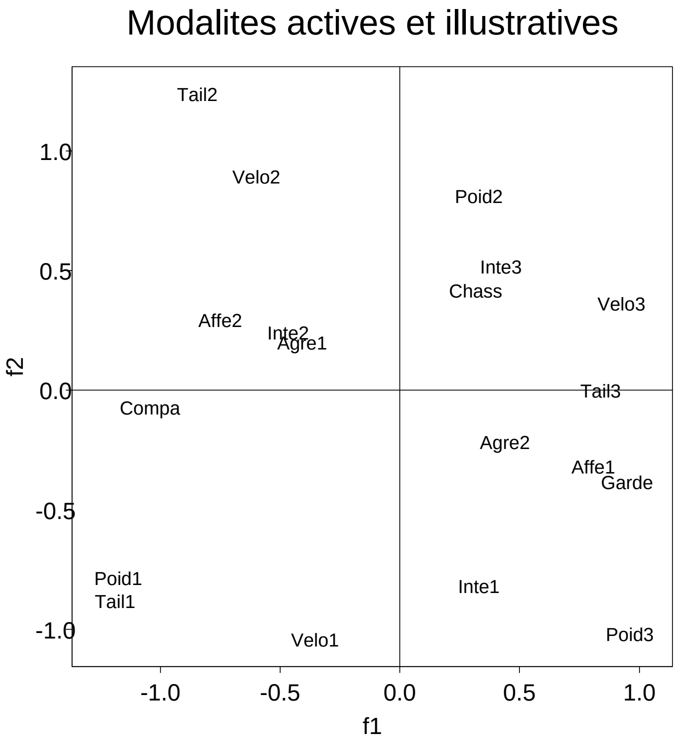


FIG. 4.19 - *Représentation des modalités et modalités illustratives dans l'afc de Z*

Bibliographie

[1] M. Baumgartner (1994) "Une introduction à Splus" École Fédérale Polytechnique de Lausanne. (disponible sous `/app/doc/stat/Sintro.ps` au CICT).

[2] R.A. Becker, J.M. Chambers and A.R. Wilks (1988) *The New S Langage* Wadsworth, Pacific Grove, California

[3] F. Cailliez et Jean-Pierre Pagès (1976) *Introduction à l'Analyse des Données* Société de Mathématiques Appliquées et de Sciences Humaines, 9, rue Duban 75016 Paris

[4] A. Carlier et A. Croquette (1992) "Introduction aux langages S et Splus" (disponible sous `/app/doc/stat/cours_s.ps` au CICT).

[5] A. Carlier et A. Croquette (1994) Nouvelles fonctions pour l'anlayse multidimensionnelle sous Splus : la bibliothèque `multidim`. (disponible sous /users/DAS/carlier/SPLUS/TEX/plaquette.ps au CICT).

[6] J.M. Chambers, T.J. Hastie (1992) *Statistical Model in S* Wadsworth & Brooks/Cole, Pacific Grove, California

[7] Gabriel, K.R. (1971) "The biplot graphical display of matrices with application to principal component analysis", *Biometrika*, 58(3)

[8] Greenacre, M.J. (1984). *Theory and applications of correspondence analysis.* London : Academic Press.

[9] Lebart, L., Morineau, A. & Piron, M. (1995). *Statistique exploratoire multidimensionnelles*, Dunod, Paris.

[10] Mardia, K.V. et Bibby, J.M. (1979) *Multivariate Analysis.* Academic Press, New York, NY.

[11] W.N. Venables & B.D. Ripley (1994) *Modern Applied Statistics with Splus* Statistical Computing, Springer Verlag. , 453-467.

Table des matières