UNIVERSITÉ
PAUL
SABATIER

TOULOUSE III

Introduction à la Statistique Exploratoire

L3 SID

Xavier Gendre

xavier.gendre@math.univ-toulouse.fr

17 décembre 2012

License

Notations

$\sharp E$	Cardinal de l'ensemble E
$\rho_K(x, y)$	Corrélation de Kendall entre les observations des variables couplées x et y
$\rho(x, y)$	Corrélation de Pearson entre les observations des variables couplées x et y
$\rho_S(x, y)$	Corrélation de Spearman entre les observations des variables couplées x et y
$\mathrm{Cov}(x, y)$	Covariance entre les observations des variables couplées x et y
χ^2	Distance du χ^2 à l'indépendance
$\mathbb{R}$	Ensemble des nombres réels
$\mathbb{R}_+$	Ensemble des nombres réels positifs $[0, +\infty[$
$\emptyset$	Ensemble vide
F_x	Fonction de répartition associée aux observations d'une variable x
$\overline{x}$	Moyenne des observations d'une variable x
q_α	Quantile d'ordre $\alpha \in [0, 1]$
${}^tM$	Transposée de la matrice M
$x_{(1)}, \ldots, x_{(n)}$	Version ordonnée des observations $x_1, \ldots, x_n$
$\mathrm{Var}\,(x)\,,\ \sigma_x^2$	Variance des observations d'une variable x

Table des matières

Chapitre 1

Moyenne et variance des observations d'une variable

1.1 Introduction

Lors de l'étude d'un phénomène, nous sommes souvent amenés à observer des *variables* qui lui sont relatives. Ces variables peuvent être de différentes natures (grandeurs physiques, caractéristiques biologiques, ...).

Exemple Si le phénomène est météorologique, les variables d'intérêt peuvent être la température t et la classe c des nuages. Un jour d'été, nous pourrions observer $t_1 = 36°C$ et $c_1 =$ "Cirrus" et, un jour d'hiver, avoir les observations $t_2 = -4°C$ et $c_2 =$ "Cumulus".

Comme l'illustre cet exemple, le type d'une variable est, a priori, quelconque. Dans ce premier chapitre, nous considérerons uniquement des variables à valeurs réelles. Ces variables sont dites *quantitatives* car elles reflètent une idée de grandeur (température, vitesse, âge, ...).

1.2 Moyenne pondérée

Nous supposons que nous avons à notre disposition un nombre entier $n > 0$ d'observations $x_1, \ldots, x_n \in \mathbb{R}$ d'une variable quantitative x.

Définition 1.1. *La* **moyenne** $\overline{x}$ *des observations* $x_1, \ldots, x_n \in \mathbb{R}$ *pondérée par les* **poids** $p_1, \ldots, p_n > 0$ *est définie par*

$$\overline{x} = \frac{1}{p_1 + \cdots + p_n} \sum_{i=1}^{n} p_i x_i \ .$$

La variable x est dite **centrée** *si $\overline{x} = 0$. Les poids sont dits* **normalisés** *si ils vérifient $p_1 + \cdots + p_n = 1$. Dans ce cas, la moyenne $\overline{x}$ devient*

$$\overline{x} = \sum_{i=1}^{n} p_i x_i \ .$$

Il est important de remarquer que, quels que soient les poids $p_1, \ldots, p_n$, il est toujours possible de calculer $\overline{x}$ avec les poids normalisés $\tilde{p}_1, \ldots, \tilde{p}_n$ donnés par

$$\tilde{p}_i = \frac{p_i}{p_1 + \cdots + p_n}, \ i \in \{1, \ldots, n\}.$$

En effet, par définition, nous avons

$$\overline{x} = \frac{1}{p_1 + \cdots + p_n} \sum_{i=1}^{n} p_i x_i = \sum_{i=1}^{n} \frac{p_i}{p_1 + \cdots + p_n} x_i = \sum_{i=1}^{n} \tilde{p}_i x_i$$

et les poids $\tilde{p}_1, \ldots, \tilde{p}_n$ sont normalisés car

$$\sum_{i=1}^{n} \tilde{p}_i = \sum_{i=1}^{n} \frac{p_i}{p_1 + \cdots + p_n} = \frac{p_1 + \cdots + p_n}{p_1 + \cdots + p_n} = 1 \ .$$

Dans la suite, nous considérerons souvent le cas particulier de la *moyenne uniforme* pour laquelle tous les poids sont égaux (*i.e.* toutes les observations ont la même importance). Prenons, par exemple, $p_1 = \cdots = p_n = 1$. Dans ce cas, nous avons $\tilde{p}_1 = \cdots = \tilde{p}_n = 1/n$ et nous retrouvons la moyenne usuelle

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \ .$$

Plus généralement, dire que l'observation x_i "pèse" un poids p_i revient à considérer que la valeur x_i intervient dans le calcul de $\overline{x}$ avec la proportion $\tilde{p}_i$.

Exemple Un lotissement contient $n = 10$ maisons dont 6 mesurent 80 m^2 et 4 mesurent 120 m^2. Afin de calculer la moyenne uniforme des surfaces, une première façon consiste à revenir à la définition 1.1,

$$\frac{1}{10}(\underbrace{80 + \cdots + 80}_{6 \text{ fois}} + \underbrace{120 + \cdots + 120}_{4 \text{ fois}}) = 96 \ m^2 \ .$$

La seconde manière consiste à considérer que les habitations de 80 m^2 représentent 6/10 des maisons du lotissement et que celles de 120 m^2 en représentent 4/10. Ainsi, nous pouvons calculer la surface moyenne en ne considérant plus que le groupe des habitats de 80 m^2 et celui des 120 m^2 pour obtenir

$$\frac{6}{10} \times 80 + \frac{4}{10} \times 120 = 96 \ m^2 \ .$$

Autrement dit, nous avons calculé la moyenne de $x_1 = 80$ et de $x_2 = 120$ pondérée par les poids $\tilde{p}_1 = 6/10$ et $\tilde{p}_2 = 4/10$.

Cette seconde façon de faire le calcul est un exemple de décomposition par groupes du calcul de la moyenne. Elle se généralise par le Théorème 1.1 suivant.

Rappel Soient $G_1, \ldots, G_n$ des parties d'un ensemble E, on dit qu'elles forment une *partition* de E si et seulement si

(i) $G_1 \cup \cdots \cup G_n = E$

et

(ii) $\forall i, j \in \{1, \ldots, n\},\ i \neq j \Rightarrow G_i \cap G_j = \emptyset$.

Théorème 1.1. [Moyenne par groupes] *Soient un entier $N > 0$ et $G_1, \ldots, G_N$ une partition de $\{1, \ldots, n\}$, nous notons $\overline{x}$ la moyenne des observations $x_1, \ldots, x_n$ pondérée par $p_1, \ldots, p_n > 0$ et, pour tout $k \in \{1, \ldots, N\}$, $\overline{x}_k$ la moyenne des x_i pour $i \in G_k$,*

$$\overline{x}_k = \frac{\sum_{i \in G_k} p_i x_i}{\sum_{i \in G_k} p_i} \ .$$

Alors, la moyenne $\overline{x}$ est la moyenne des $\overline{x}_k$ pondérée par les poids $q_k = \sum_{i \in G_k} p_i$,

$$\overline{x} = \frac{1}{q_1 + \cdots + q_N} \sum_{k=1}^{N} q_k \overline{x}_k \ .$$

Démonstration. Il suffit de remarquer que $q_1 + \cdots + q_N = p_1 + \cdots + p_n$ et de remplacer les $\overline{x}_k$ par leurs définitions,

$$\begin{aligned} \frac{1}{q_1 + \cdots + q_N} \sum_{k=1}^{N} q_k \overline{x}_k &= \frac{1}{p_1 + \cdots + p_n} \sum_{k=1}^{N} \frac{q_k}{\sum_{i \in G_k} p_i} \sum_{i \in G_k} p_i x_i \\ &= \frac{1}{p_1 + \cdots + p_n} \sum_{k=1}^{N} \sum_{i \in G_k} p_i x_i \\ &= \frac{1}{p_1 + \cdots + p_n} \sum_{i=1}^{n} p_i x_i = \overline{x} \ . \end{aligned}$$

□

Exercice 1.1. *Montrer que si les poids $p_1, \ldots, p_n > 0$ sont normalisés alors les poids $q_1, \ldots, q_N$ définis dans le Théorème 1.1 sont aussi normalisés.*

Le Théorème 1.1 est particulièrement utile lorsque la variable x ne peut prendre qu'un nombre N fini de valeurs distinctes $y_1, \ldots, y_N$. En effet, dans ce cas, nous pouvons définir les groupes $G_1, \ldots, G_N$ par

$$G_k = \{i \in \{1, \ldots, n\} \text{ tels que } x_i = y_k\} \ , k \in \{1, \ldots, N\} \ . \tag{1.1}$$

Le groupe G_k contient donc tous les indices i tels que l'observation x_i soit égale à la valeur y_k. Par conséquent, nous savons que $\overline{x}_k = y_k$. De plus, $G_1, \ldots, G_N$ forment une partition de $\{1, \ldots, n\}$. Si nous notons n_k l'effectif du groupe G_k, $k \in \{1, \ldots, N\}$, le Théorème 1.1 nous

permet de calculer la moyenne uniforme $\overline{x}$ des observations $x_1, \ldots, x_n$ en fonction des valeurs $y_1, \ldots, y_N$ et des effectifs $n_1, \ldots, n_N$,

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{N} n_k y_k$$

car $q_k = n_k$ et $n_1 + \cdots + n_N = n$.

Exercice 1.2. *Montrer que les groupes $G_1, \ldots, G_N$ définis par (1.1) forment bien une partition de $\{1, \ldots, n\}$.*

Enfin, pour les calculs de moyenne, il est souvent pratique d'utiliser le fait que la moyenne est linéaire.

Proposition 1.1. *Si a et b sont des nombres réels quelconques et que nous considérons les observations $z_i = ax_i + b$, $i \in \{1, \ldots, n\}$, relatives à la variable quantitative $z = ax + b$, alors*

$$\overline{z} = \overline{ax + b} = a \times \overline{x} + b \ .$$

Prenons, de plus, une variable quantitative y et ses n observations $y_1, \ldots, y_n \in \mathbb{R}$. Si nous considérons les observations $z_i = x_i + y_i$, $i \in \{1, \ldots, n\}$, relatives à la variable quantitative $z = x + y$, alors

$$\overline{z} = \overline{x + y} = \overline{x} + \overline{y} \ .$$

Exercice 1.3. *Ecrire la preuve de la Proposition 1.1.*

1.3 Variance

Nous supposons à partir de maintenant que nous disposons d'un nombre entier $n > 0$ d'observations $x_1, \ldots, x_n \in \mathbb{R}$ d'une variable x et de poids $p_1, \ldots, p_n > 0$ normalisés.

Définition 1.2. *La* **variance** $\mathrm{Var}\,(x)$ *des observations $x_1, \ldots, x_n$ est définie par*

$$\mathrm{Var}\,(x) = \sum_{i=1}^{n} p_i (x_i - \overline{x})^2 \ .$$

*Nous noterons aussi $\mathrm{Var}\,(x) = \sigma_x^2$ où $\sigma_x = \sqrt{\mathrm{Var}\,(x)}$ est appelé l'***écart-type***. La variable x est dite* **réduite** *si $\mathrm{Var}\,(x) = 1$.*

La variance est donc la moyenne des carrés des écarts $x_1 - \overline{x}, \ldots, x_n - \overline{x}$ pondérée par $p_1, \ldots, p_n$. Cette quantité mesure la dispersion des observations autour de $\overline{x}$. De plus, elle est toujours positive.

Exercice 1.4. *Montrer que si $\mathrm{Var}\,(x) = 0$ alors toutes les observations $x_1, \ldots, x_n$ sont égales à la moyenne $\overline{x}$.*

La variance est quadratique. En particulier, nous avons la proposition suivante.

Proposition 1.2. *Si a et b sont des nombres réels quelconques et que nous considérons les observations $z_i = ax_i + b$, $i \in \{1, \ldots, n\}$, relatives à la variable quantitative $z = ax + b$, alors*

$$\mathrm{Var}\,(z) = \mathrm{Var}\,(ax + b) = a^2 \times \mathrm{Var}\,(x) \ .$$

Il faut faire attention car la variance n'est pas additive ; c'est-à-dire que, en général, nous n'avons pas $\mathrm{Var}\,(x+y) = \mathrm{Var}\,(x) + \mathrm{Var}\,(y)$.

Exemple Pour $n = 2$, supposons que nous ayons observé $x_1 = -1$, $x_2 = 1$, $y_1 = 0$ et $y_2 = 1$ et que les poids soient $p_1 = p_2 = 1/2$. Dans ce cas, $\overline{x} = 0$ et $\overline{y} = 1/2$. Nous avons alors

$$\mathrm{Var}\,(x) + \mathrm{Var}\,(y) = 1 + \frac{1}{4} = \frac{5}{4} \neq \frac{9}{4} = \mathrm{Var}\,(x+y) \ .$$

Exercice 1.5. *Ecrire la preuve de la Proposition 1.2.*

Pour calculer $\mathrm{Var}\,(x)$, il est parfois utile d'utiliser l'expression donnée par la proposition suivante.

Proposition 1.3. *La variance vaut la moyenne des carrés moins le carré de la moyenne,*

$$\mathrm{Var}\,(x) = \overline{x^2} - \overline{x}^2 \ ,$$

avec $\overline{x^2} = \sum_{i=1}^{n} p_i x_i^2$.

Démonstration. Développons le carré dans la définition de la variance,

$$\begin{aligned} \sum_{i=1}^{n} p_i (x_i - \overline{x})^2 &= \sum_{i=1}^{n} p_i x_i^2 - 2\overline{x} \sum_{i=1}^{n} p_i x_i + \overline{x}^2 \sum_{i=1}^{n} p_i \\ &= \overline{x^2} - 2\overline{x}^2 + \overline{x}^2 \\ &= \overline{x^2} - \overline{x}^2 \ . \end{aligned}$$

□

Exercice 1.6. *Soit un entier $n > 0$. Déduire de la Proposition 1.3 que, pour tout $x_1, \ldots, x_n \in \mathbb{R}$ et pour tout $p_1, \ldots, p_n > 0$ tels que $p_1 + \cdots + p_n = 1$, nous avons*

$$\left(\sum_{i=1}^{n} p_i x_i \right)^2 \leqslant \sum_{i=1}^{n} p_i x_i^2 \ .$$

Il s'agit d'un cas particulier de l'inégalité de Jensen.

Comme pour la moyenne, il est possible de décomposer l'expression de la variance par groupes.

Théorème 1.2. [Variance par groupes] *En utilisant les mêmes hypothèses et notations que dans le Théorème 1.1, nous notons, pour tout $k \in \{1, \ldots, N\}$, σ_k^2 la variance des x_i pour $i \in G_k$,*

$$\sigma_k^2 = \frac{1}{q_k} \sum_{i \in G_k} p_i (x_i - \overline{x}_k)^2 \ .$$

Alors, la variance $\mathrm{Var}\,(x)$ *se décompose en*

$$\mathrm{Var}\,(x) = \mathrm{Var}_{inter}(x) + \mathrm{Var}_{intra}(x) \tag{1.2}$$

avec

$$\mathrm{Var}_{inter}(x) = \sum_{k=1}^{N} q_k(\overline{x}_k - \overline{x})^2 \qquad \textit{(Variance inter-groupe)}$$

et

$$\mathrm{Var}_{intra}(x) = \sum_{k=1}^{N} q_k \sigma_k^2 \ . \qquad \textit{(Variance intra-groupe)}$$

Démonstration. Faisons apparaître les $\overline{x}_k$ dans l'expression de la variance,

$$\begin{aligned}
\mathrm{Var}\,(x) &= \sum_{i=1}^{n} p_i(x_i - \overline{x})^2 \\
&= \sum_{k=1}^{N} \sum_{i \in G_k} p_i \left((x_i - \overline{x}_k) + (\overline{x}_k - \overline{x})\right)^2 \\
&= \sum_{k=1}^{N} \sum_{i \in G_k} p_i(x_i - \overline{x}_k)^2 + 2\sum_{k=1}^{N}(\overline{x}_k - \overline{x}) \sum_{i \in G_k} p_i(x_i - \overline{x}_k) \\
&\quad + \sum_{k=1}^{N}(\overline{x}_k - \overline{x})^2 \sum_{i \in G_k} p_i \\
&= \sum_{k=1}^{N} q_k \sigma_k^2 + 2\sum_{k=1}^{N}(\overline{x}_k - \overline{x}) \sum_{i \in G_k} p_i(x_i - \overline{x}_k) + \sum_{k=1}^{N} q_k(\overline{x}_k - \overline{x})^2 \ .
\end{aligned}$$

Pour conclure, il suffit de remarquer que le terme central est nul,

$$\sum_{i \in G_k} p_i(x_i - \overline{x}_k) = \sum_{i \in G_k} p_i x_i - \overline{x}_k \sum_{i \in G_k} p_i = q_k \overline{x}_k - \overline{x}_k q_k = 0 \ .$$

□

Les deux termes qui apparaissent dans la décomposition (1.2) ne s'interprètent pas de la même façon. La variance inter-groupe $\mathrm{Var}_{inter}(x)$ est la variance des moyennes et elle traduit la dispersion entre les groupes. La variance intra-groupe Var_{intra} est la moyenne des variances et elle correspond à la dispersion dans les groupes.

Chapitre 2

Distribution des observations d'une variable

2.1 Introduction

Une fois que nous disposons des observations d'une variable quantitative, il peut être intéressant de regarder comment ces observations sont réparties. Pour rendre compte visuellement de cette distribution, il existe de nombreuses méthodes graphiques. Nous présentons dans ce chapitre certaines parmi les plus utilisées.

Dans la suite, nous supposons avoir $n > 0$ observations $x_1, \ldots, x_n \in \mathbb{R}$ d'une variable quantitative x et des poids $p_1, \ldots, p_n > 0$ normalisés.

2.2 Histogramme

La première représentation à laquelle nous allons nous intéresser est celle des *histogrammes*. Nous considérons $N+1$ nombres réels $a_0 < a_1 < \cdots < a_N$ tels que toutes les observations soient regroupées dans les N intervalles $[a_0, a_1[, \ldots, [a_{N-1}, a_N[$. Dans la suite, pour tout $k \in \{1, \ldots, N\}$, nous noterons n_k le nombre d'observations x_i présentes dans l'intervalle $[a_{k-1}, a_k[$,

$$n_k = \sharp\{i \in \{1, \ldots, n\} \text{ tels que } x_i \in [a_{k-1}, a_k[\} \ .$$

Nous appelons *fréquence* de $[a_{k-1}, a_k[$ la quantité $f_k = n_k/n$ et *poids* de $[a_{k-1}, a_k[$ la quantité

$$p^{(k)} = \sum_{\substack{i \text{ tel que} \\ x_i \in [a_{k-1}, a_k[}} p_i \ .$$

Pour représenter l'histogramme, nous traçons un rectangle au-dessus de chaque intervalle. Afin de calculer les hauteurs de ces rectangles, il faut distinguer deux cas selon que les intervalles ont tous la même longueur ou non.

2.2.1 Intervalles de même longueur

Nous nous plaçons dans le cas simple où les intervalles $[a_{k-1}, a_k[$ sont de longueur constante ($a_0 - a_1 = \cdots = a_N - a_{N-1}$). Selon le type d'histogramme voulu, la hauteur des rectangles peut être une des valeurs suivantes :

- n_k pour un diagramme des effectifs,
- f_k pour un diagramme des fréquences,
- $p^{(k)}$ pour un diagramme des poids.

Exemple Supposons que nos 20 observations aient toutes le même poids 1/20 et soient réparties de la façon suivante : 5 valeurs dans $[0, 10[$, 3 valeurs dans $[10, 20[$, 5 valeurs dans $[20, 30[$ et 7 valeurs dans $[30, 40[$.

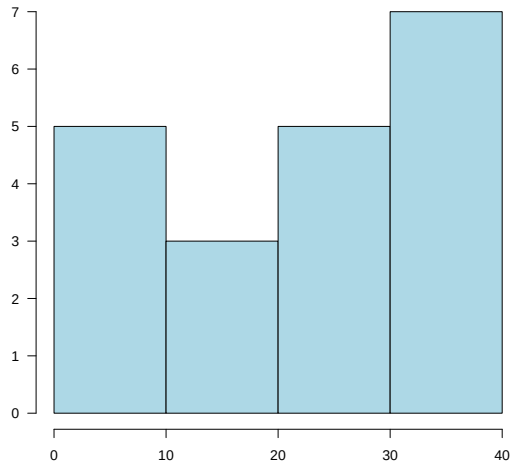


FIGURE 2.1 – Diagramme des effectifs pour des intervalles de même taille.

Exercice 2.1. *Que faut-il modifier sur la Figure 2.1 pour obtenir directement le diagramme des fréquences ?*

2.2.2 Intervalles de longueurs différentes

Nous ne supposons plus que les intervalles soient tous de même longueur. Si nous tracions les rectangles comme précédemment, leurs surfaces seraient faussées et cela donnerait une mauvaise représentation de la distribution des observations.

Il est donc important de normaliser la hauteur des rectangles par la longueur des intervalles :

- $\dfrac{n_k}{a_k - a_{k-1}}$ pour un diagramme des effectifs,
- $\dfrac{f_k}{a_k - a_{k-1}}$ pour un diagramme des fréquences,
- $\dfrac{p^{(k)}}{a_k - a_{k-1}}$ pour un diagramme des poids.

Exemple Reprenons les observations de l'exemple précédent et regroupons les intervalles $[10, 20[$ et $[20, 30[$. Nous avons donc deux intervalles de longueur 10 et un de longueur 20.

Si nous ne renormalisons pas la hauteur des rectangles, la représentation est faussée comme le montre la Figure 2.2. En revanche, la figure 2.3 représente le diagramme des fréquences correctement normalisé.

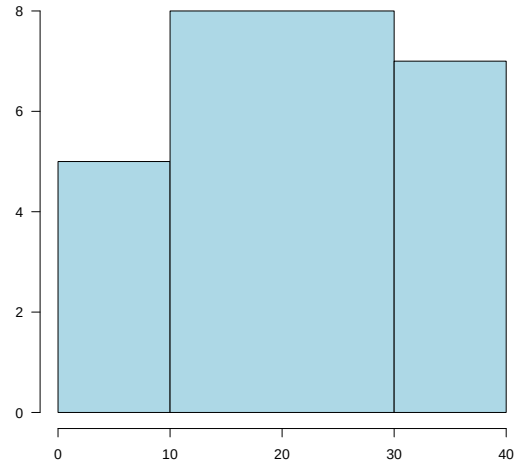


FIGURE 2.2 – Diagramme des effectifs incorrect dans le cas d'intervalles de longueurs différentes (à comparer avec les Figures 2.1 et 2.3).

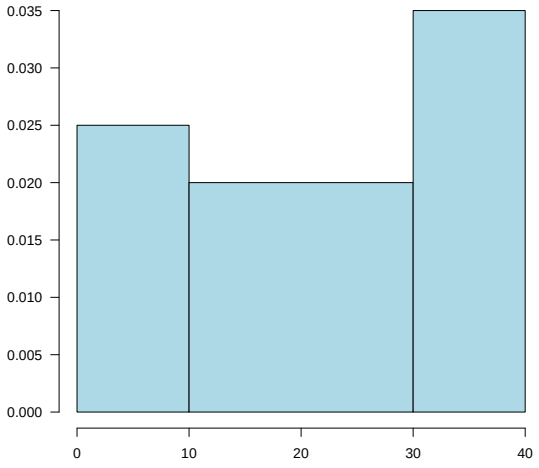


FIGURE 2.3 – Diagramme des fréquences pour des intervalles de longueurs différentes.

Exercice 2.2. *Si les poids $p_1, \ldots, p_n$ sont tous égaux à $1/n$, comment comparer le diagramme des fréquences et celui des poids ?*

2.3 Poids cumulés

Comme dans la section précédente, nous considérons $N+1$ nombres réels $a_0 < a_1 < \cdots < a_N$ tels que toutes les observations soient regroupées dans les N intervalles $[a_0, a_1[, \ldots, [a_{N-1}, a_N[$.

Le *diagramme des poids cumulés* est un histogramme particulier construit de la façon suivante. Pour $k \in \{1, \ldots, N\}$, la hauteur h_k du rectangle relatif à l'intervalle $[a_{k-1}, a_k[$ est la somme des poids de toutes les observations inférieures à a_k,

$$h_k = \sum_{\substack{i \text{ tel que} \\ x_i \leqslant a_k}} p_i \ .$$

De plus, il est courant de superposer à cet histogramme une courbe linéaire par morceaux reliant (a_{k-1}, h_{k-1}) à (a_k, h_k) pour k allant de 1 à N (par convention, on pose $h_0 = 0$). Cette courbe est donc croissante et prend des valeurs de 0 à 1 puisque les poids sont normalisés. Elle illustre la façon dont les observations seraient réparties si cette répartition était uniforme sur chaque intervalle.

Exemple Reprenons encore les 20 observations de même poids 1/20 des exemples de la section précédente. Les figures 2.4 et 2.5 montrent le diagramme des poids cumulés pour deux choix d'intervalles différents. Nous remarquons, en particulier, que la longueur des intervalles ne modifie pas la façon de tracer ces diagrammes contrairement aux histogrammes.

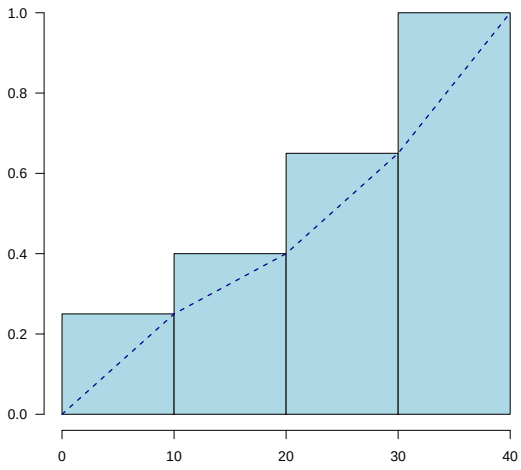


FIGURE 2.4 – Diagramme des poids cumulés avec $[0, 10[$, $[10, 20[$, $[20, 30[$ et $[30, 40[$.

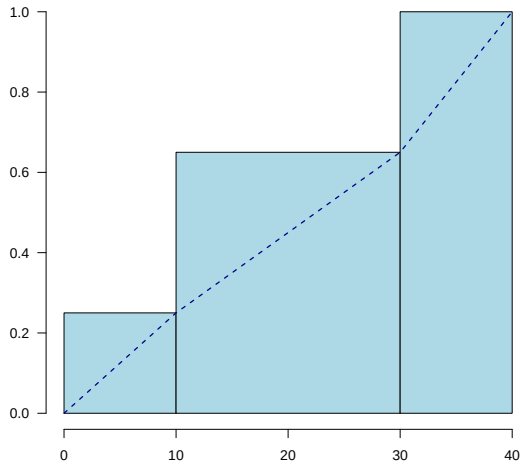


FIGURE 2.5 – Diagramme des poids cumulés avec $[0, 10[$, $[10, 30[$ et $[30, 40[$.

2.4 Fonction de répartition et quantiles

La *fonction de répartition* peut être vue comme un diagramme des poids cumulés particulier pour lequel il n'y aurait qu'une unique observation dans chaque intervalle. Il s'agit d'une fonction F_x constante par morceaux et croissante de 0 à 1 définie pour tout $t \in \mathbb{R}$ par

$$F_x(t) = \sum_{\substack{i \text{ tel que} \\ x_i \leqslant t}} p_i \ .$$

Cette fonction fait donc un saut en chaque point x_i. Pour la représenter, il peut être pratique de considérer la *version ordonnée* des observations. Cette version est une permutation des observations, notée $x_{(1)}, \ldots, x_{(n)}$, choisie de telle sorte que nous ayons

$$x_{(1)} \leqslant \cdots \leqslant x_{(n)} \ .$$

Nous savons alors que la fonction de répartition F_x vaut 0 sur $]-\infty, x_{(1)}[$, qu'elle fait un saut à chaque point $x_{(i)}$ et qu'elle vaut 1 sur $[x_{(n)}, +\infty[$.

Exemple Considérons que nous avons les $n = 5$ observations suivantes :

$$x_1 = 3 \ , \ x_2 = -1 \ , \ x_3 = 4 \ , \ x_4 = 3 \ , \ x_5 = 0 \ .$$

La version ordonnée de ces observation est donc

$$x_{(1)} = x_2 = -1 \ , \ x_{(2)} = x_5 = 0 \ , \ x_{(3)} = x_1 = 3 \ , \ x_{(4)} = x_4 = 3 \ , \ x_{(5)} = x_3 = 4 \ .$$

Il faut noter que le choix de l'ordre de $x_{(3)}$ et $x_{(4)}$ est arbitraire puisque les données x_1 et x_4 sont égales. Si nous considérons maintenant que les poids $p_1, \ldots, p_5$ sont tous égaux à $1/5$, alors la fonction de répartition est donnée par la Figure 2.6.

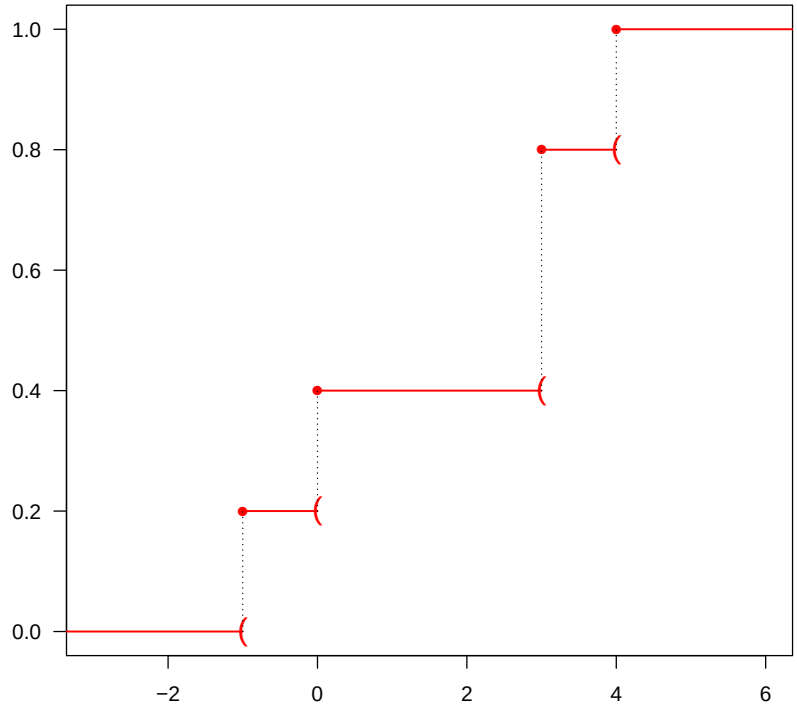


FIGURE 2.6 – Fonction de répartition de $x_1 = 3,\ x_2 = -1,\ x_3 = 4,\ x_4 = 3,\ x_5 = 0$ avec des poids uniformes.

Définition 2.1. *Soit $\alpha \in [0,1]$, nous appelons* **quantile** *d'ordre α tout nombre $q_\alpha \in \mathbb{R}$ tel que nous ayons une proportion α des observations qui soit inférieure ou égale à q_α.*

Ainsi, si $\alpha = i/n$ pour un $i \in \{1,\dots,n\}$, alors $q_\alpha = q_{i/n} = x_{(i)}$. Par convention, si $\alpha \in [0, 1/n[$, nous poserons que le quantile d'ordre α vaut $q_\alpha = -\infty$ car il n'y a aucune observation avant $x_{(1)}$. Enfin, si $\alpha \in [(i-1)/n, i/n[$ pour un $i \in \{1,\dots,n\}$, alors il existe $\theta \in [0,1]$ tel que $\alpha = (i - 1 + \theta)/n$ et nous interpolons entre $q_{(i-1)/n} = x_{(i-1)}$ et $q_{i/n} = x_{(i)}$ pour obtenir le quantile d'ordre α,

$$q_\alpha = x_{(i-1)} + \theta \left(x_{(i)} - x_{(i-1)} \right) \ . \tag{2.1}$$

Notons que cette définition par interpolation est bien compatible avec la convention $q_\alpha = -\infty$ lorsque $\alpha \in [0, x_{(1)}[$.

Un des avantages de cette définition par interpolation est que les quantiles sont directement lisibles sur le graphe de la fonction de répartition. Il faut cependant faire attention aux doublons dans les observations (comme x_1 et x_4 dans l'exemple). Pour cela, nous notons $p_{(i)}$ le poids de l'observation $x_{(i)}$ et nous introduisons les quantités $F^{(i)}$ pour $i \in \{1,\dots,n\}$,

$$F^{(i)} = p_{(1)} + \cdots + p_{(i)} \leqslant F_x(x_{(i)}) \ .$$

Il suffit alors de tracer la courbe linéaire par morceaux joignant les points $(x_{(i-1)}, F^{(i-1)})$ et $(x_{(i)}, F^{(i)})$ et de lire q_α sur l'axe des abscisses comme étant l'antécédent de α sur cette courbe (voir Figure 2.7). Si il n'y a aucun doublon parmi les observations, cette courbe relie simplement les points de saut de F_x. Par contre, si il y a des doublons, certaines parties de la courbes deviennent des segments verticaux comme cela se voit sur la Figure 2.7 entre $x_{(3)}$ et $x_{(4)}$.

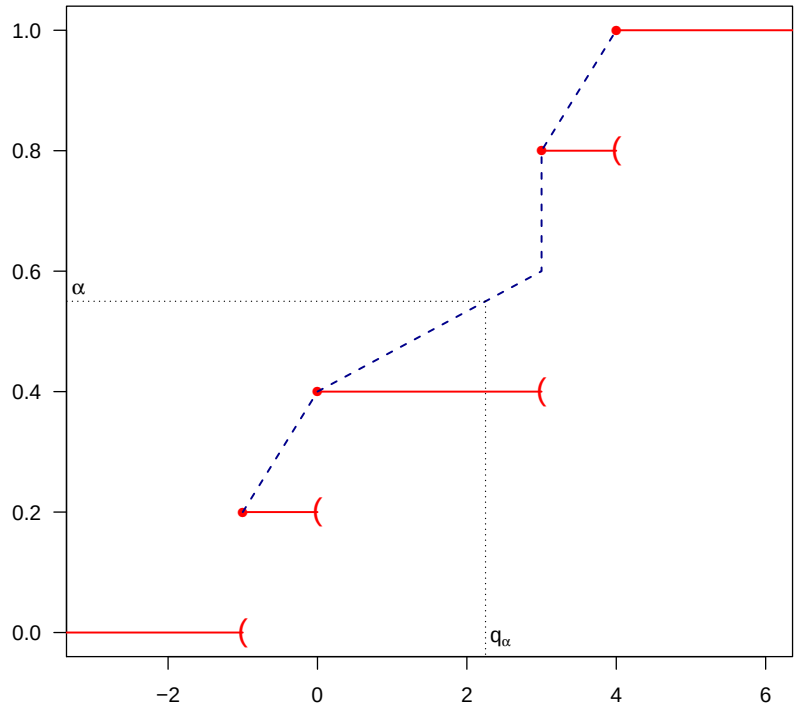


FIGURE 2.7 – Lecture du quantile q_α d'ordre α grâce à la fonction de répartition.

Définition 2.2. *La* **médiane** *est le quantile* $q_{0.5}$ *d'ordre* 50%. *De plus, nous appelons* **quartiles** *les quantiles* $q_{0.25}$ *et* $q_{0.75}$ *à* 25% *et* 75%.

Exemple Reprenons les valeurs de l'exemple précédent. Pour calculer la médiane, il faut trouver le quantile d'ordre 50%. Or, nous savons que $0.5 = 2.5/5 = (3 - 1 + 0.5)/5$ (*i.e.* $i = 3$ et $\theta = 0.5$ dans (2.1)). Donc, la médiane vaut

$$q_{0.5} = x_{(2)} + 0.5\left(x_{(3)} - x_{(2)}\right) = 0 + 0.5(3 - 0) = 1.5\ .$$

De la même façon, nous savons que $0.25 = 1.25/5 = (2 - 1 + 0.25)/5$ et $0.75 = 3.75/5 = (4 - 1 + 0.75)$ et nous calculons les quartiles,

$$q_{0.25} = x_{(1)} + 0.25\left(x_{(2)} - x_{(1)}\right) = -1 + 0.25(0 + 1) = -0.75$$

et

$$q_{0.75} = x_{(3)} + 0.25\left(x_{(4)} - x_{(3)}\right) = 3 + 0.75(3-3) = 3 \ .$$

Ces résultats se retrouvent graphiquement comme l'illustre la Figure 2.8.

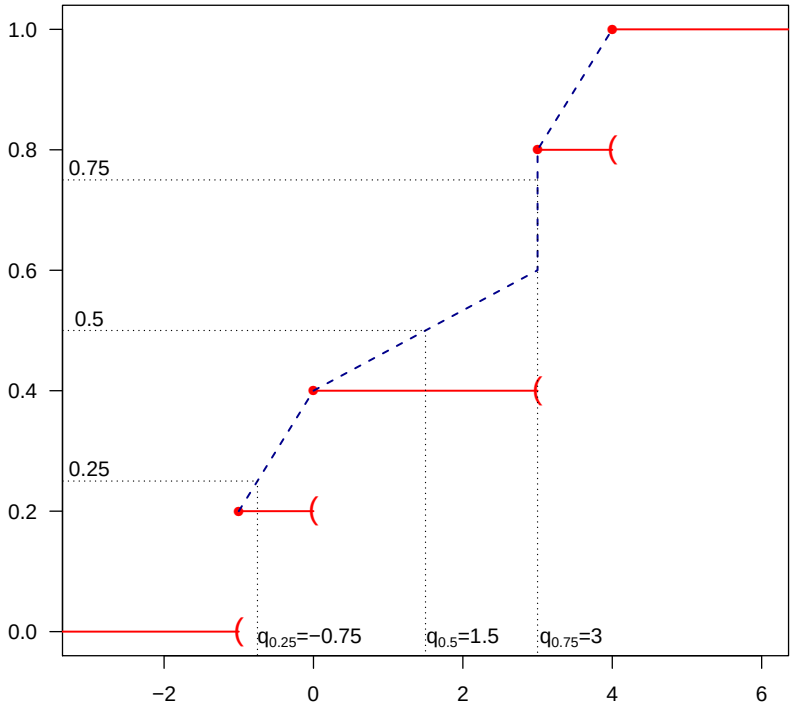


FIGURE 2.8 – Lecture de la médiane et des quartiles de l'exemple.

Exercice 2.3. *Considérons que nous avons les $n = 5$ observations suivantes :*

$$x_1 = 3 \ , \ x_2 = -1 \ , \ x_3 = 5 \ , \ x_4 = 2 \ , \ x_5 = -3 \ .$$

Tracer la fonction de répartition F_x et calculer la médiane et les quartiles. Retrouver ensuite graphiquement ces résultats.

2.5 Boîte à moustaches (box plot)

La *boîte à moustaches* (ou *box plot* en anglais) est un graphe synthétique et très utilisé en pratique pour représenter la distribution des observations d'une variable quantitative. Le corps de ce graphe fait apparaître la médiane, les deux quartiles et l'inter-quartile $IQ = q_{0.75} - q_{0.25}$. Nous ajoutons des "moustaches" pour représenter les données en dehors de l'inter-quartile. Les extrémités des moustaches peuvent avoir des significations différentes selon les situations (voir Figure 2.9) :

- elles peuvent indiquer les valeurs minimales et maximales des observations,
- elles peuvent représenter le maximum entre la plus petite valeur et $q_{0.25} - 1.5 \times IQ$ ainsi que le minimum entre la plus grande valeur et $q_{0.75} + 1.5 \times IQ$ (dans ce cas, les données en dehors sont dites *exceptionnelles* ou *outliers*),
- elles peuvent prendre d'autres valeurs, il faut donc prendre garde à cela pour interpréter un box plot.

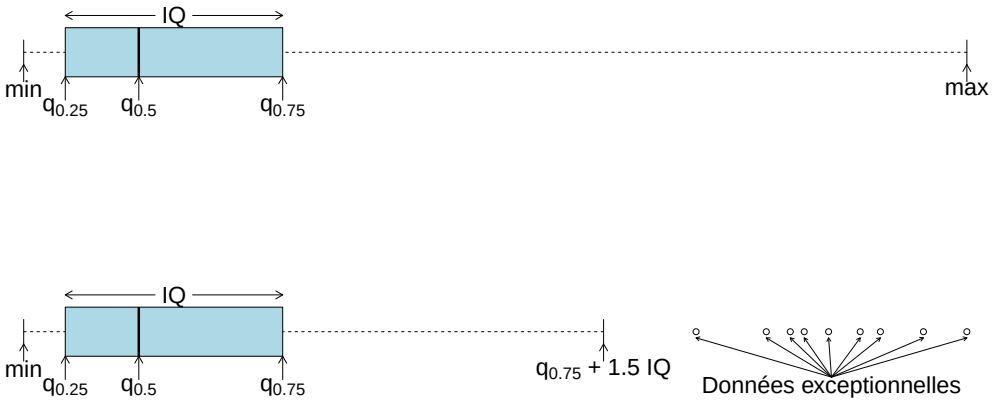


FIGURE 2.9 – Les moustaches peuvent indiquer les valeurs minimales et maximales (au dessus) ou être limitées à $q_{0.25} - 1.5 \times IQ$ et $q_{0.75} + 1.5 \times IQ$ (au dessous).

2.6 Diagramme quantile-quantile (q-q plot)

En plus de nos observations $x_1, \ldots, x_n$ de la variable x, nous supposons dans cette section que nous disposons aussi d'un nombre entier $m > 0$ d'observations $y_1, \ldots, y_m$ d'une variable quantitative y qui n'a, a priori, aucun rapport avec x. De plus, pour $\alpha \in [0, 1]$, nous noterons $q_\alpha(x)$ et $q_\alpha(y)$ les quantiles d'ordre α pour les variables x et y respectivement.

Le *diagramme quantile-quantile* (ou *q-q plot* en anglais) est un outil graphique qui permet de comparer les distributions des deux jeux d'observations. Par exemple, en pratique, les données $x_1, \ldots, x_n$ peuvent être observées pour un phénomène à étudier et les données $y_1, \ldots, y_m$ peuvent être calculées à partir d'un modèle théorique décrivant ce phénomène. Le diagramme quantile-quantile permet alors de vérifier graphiquement la validité du modèle en comparant la distribution observée avec la distribution théorique.

Pour construire le diagramme quantile-quantile, il faut considérer un nombre entier $K > 0$ de valeurs $0 \leqslant \alpha_1 \leqslant \cdots \leqslant \alpha_K \leqslant 1$ (plus K sera grand, plus notre diagramme pourra être précis). Le diagramme s'obtient finalement en traçant la courbe linéaire par morceaux qui passe par les points $(q_{\alpha_k}(x), q_{\alpha_k}(y))$ pour $k \in \{1, \ldots, K\}$.

Si cette courbe est "proche" de la première diagonale (la droite "$y = x$"), alors nous pourrons conclure que les distributions de x et de y sont similaires. Si la courbe est proche d'une droite qui n'est pas la première diagonale, alors nous pourrons proposer une transformation

affine (type "$y = ax + b$") pour rendre les distributions comparables. Enfin, si le diagramme ne fait pas apparaître une droite, nous ne pouvons pas conclure, a priori.

Exemple Nous considérons deux machines A et B qui sont utilisées pour remplir des sachets de 25 grammes d'un médicament. Elles ont toutes les deux été testées $n = 1000$ fois chacune. Le tableau suivant donne le relevé de ces mille tests, en indiquant, pour chaque machine, le nombre de fois où elle a donné un sachet dont le poids est indiqué en première ligne :

Poids	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
A	10	10	10	50	30	10	20	40	80	450	150	10	40	0	90
B	0	0	10	10	10	20	40	162	0	488	10	140	50	30	30

Prenons $\alpha_1 = 0.1$, $\alpha_2 = 0.2$, ... , $\alpha_{10} = 1.0$ et calculons les quantiles associés : $q_{\alpha_1}(A) = 20$, $q_{\alpha_2}(A) = 24$, $q_{\alpha_3}(A) = 25$, ... , $q_{\alpha_7}(A) = 25$, $q_{\alpha_8}(A) = 26$, $q_{\alpha_9}(A) = 28$, $q_{\alpha_{10}}(A) = 30$ et $q_{\alpha_1}(B) = 23$, $q_{\alpha_2}(B) = 23$, $q_{\alpha_3}(B) = 25$, ... , $q_{\alpha_7}(B) = 25$, $q_{\alpha_8}(B) = 27$, $q_{\alpha_9}(B) = 28$, $q_{\alpha_{10}}(B) = 30$. Nous traçons donc le diagramme à partir des points $(q_{\alpha_1}(A), q_{\alpha_1}(B)) = (20, 23)$, $(q_{\alpha_2}(A), q_{\alpha_2}(B)) = (24, 23)$, ... La Figure 2.10 montre le diagramme obtenu et suggère que les deux machines n'admettent par les mêmes distributions au vu de l'écart à la première diagonale.

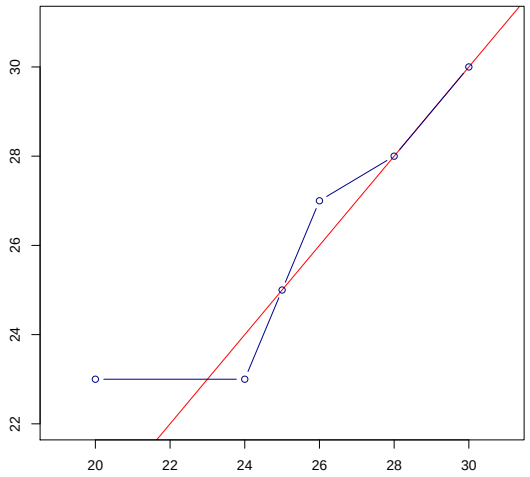


FIGURE 2.10 – Diagramme quantile-quantile de l'exemple.

Chapitre 3

Observations de deux variables couplées

3.1 Introduction

La dernière section du chapitre précédent présente un exemple de comparaison entre deux jeux de données relatifs à des variables x et y. Lorsque ces jeux de données sont issus d'une même expérience, nous parlerons d'*observations couplées*. Dans ce cas, le statisticien peut naturellement se poser nombre de questions sur leurs liens potentiels : existe-t-il une relation entre x et y ? Si oui, de quelle nature ? Si non, comment le justifier ? ...

Dans ce chapitre, nous considérons donc deux jeux de même taille $n > 0$ d'observations couplées $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$ relatifs à des variables quantitatives x et y respectivement. Ces observations seront pondérées par les poids $p_1, \ldots, p_n > 0$ normalisés.

3.2 Covariance et corrélation linéaire

Un première façon d'établir un lien entre les variables x et y consiste à regarder si les observations ont tendance à varier dans le même sens.

Définition 3.1. *La* **covariance** *entre les observations de x et celles de y est définie par*

$$\mathrm{Cov}(x, y) = \sum_{i=1}^{n} p_i (x_i - \overline{x})(y_i - \overline{y}) \ .$$

Le signe de la covariance a une signification importante. En effet, la covariance $\mathrm{Cov}(x, y)$ aura tendance à être positive si, pour de nombreux $i \in \{1, \ldots, n\}$, nous avons $x_i \geqslant \overline{x}$ et $y_i \geqslant \overline{y}$ ou bien si nous avons $x_i \leqslant \overline{x}$ et $y_i \leqslant \overline{y}$. Autrement dit, nous aurons

- $\mathrm{Cov}(x, y) > 0$ si les variables x et y ont tendance à varier dans le même sens,
- $\mathrm{Cov}(x, y) < 0$ si les variables x et y ont tendance à varier en sens inverse.

Lorsque la covariance est proche de 0, il n'est pas possible de l'interpréter directement.

Exemple Si x est la température extérieure et si y est le volume de crème glacée acheté, les observations de x et de y auront tendance à varier dans le même sens (plus il fait chaud, plus il y a de glaces consommées) et la covariance sera positive. Si x est la température extérieure

et si y est la consommation de gaz pour le chauffage, les observations de x et de y évolueront en sens inverse (plus il fait chaud, moins nous chauffons les maisons) et la covariance sera négative.

Proposition 3.1. *Soient a un nombre réel quelconque et $z_1, \dots, z_n \in \mathbb{R}$ les observations d'une variable quantitative z, la covariance vérifie les propriétés suivantes :*

- **Bilinéarité** *:*

$$\mathrm{Cov}(ax, y) = a\mathrm{Cov}(x, y) \qquad et \qquad \mathrm{Cov}(x, ay) = a\mathrm{Cov}(x, y) \ ,$$

$$\mathrm{Cov}(x, y + z) = \mathrm{Cov}(x, y) + \mathrm{Cov}(x, z)$$

et

$$\mathrm{Cov}(x + z, y) = \mathrm{Cov}(x, y) + \mathrm{Cov}(z, y) \ ,$$

- **Symétrie** *:*

$$\mathrm{Cov}(x, y) = \mathrm{Cov}(y, x) \ ,$$

- **Positivité** *:*

$$\mathrm{Cov}(x, x) \geqslant 0 \ .$$

Démonstration. La covariance est symétrique par définition,

$$\mathrm{Cov}(x, y) = \sum_{i=1}^{n} p_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} p_i (y_i - \overline{y})(x_i - \overline{x}) = \mathrm{Cov}(y, x) \ .$$

Pour avoir sa bilinéarité, par symétrie, il suffit de montrer la linéarité en x. La Proposition 1.1 donne

$$\mathrm{Cov}(ax, y) = \sum_{i=1}^{n} p_i (ax_i - \overline{ax})(y_i - \overline{y}) = a \sum_{i=1}^{n} p_i (x_i - \overline{x})(y_i - \overline{y}) = a\mathrm{Cov}(x, y)$$

et

$$\begin{aligned} \mathrm{Cov}(x, y + z) &= \sum_{i=1}^{n} p_i (x_i - \overline{x}) \left((y_i - \overline{y}) + (z_i - \overline{z}) \right) \\ &= \sum_{i=1}^{n} p_i (x_i - \overline{x})(y_i - \overline{y}) + \sum_{i=1}^{n} p_i (x_i - \overline{x})(z_i - \overline{z}) \\ &= \mathrm{Cov}(x, y) + \mathrm{Cov}(x, z) \ . \end{aligned}$$

Remarquons enfin que la covariance d'une variable avec elle-même est sa variance,

$$\mathrm{Cov}(x, x) = \mathrm{Var}\,(x) \geqslant 0 \ .$$

□

Comme nous l'avons vu au Chapitre 1, la variance n'est pas additive. Cependant, il est possible de développer la variance d'une somme de variables quantitatives en faisant intervenir la covariance.

Proposition 3.2. *Nous avons*

$$\mathrm{Var}\,(x+y) = \mathrm{Var}\,(x) + 2\mathrm{Cov}(x,y) + \mathrm{Var}\,(y)\ .$$

Démonstration. Il suffit de développer la somme qui définit la variance de $x+y$,

$$\begin{aligned}\mathrm{Var}\,(x+y) &= \sum_{i=1}^{n} p_i\left((x_i-\overline{x})+(y_i-\overline{y})\right)^2\\ &= \sum_{i=1}^{n} p_i(x_i-\overline{x})^2 + 2\sum_{i=1}^{n} p_i(x_i-\overline{x})(y_i-\overline{y}) + \sum_{i=1}^{n} p_i(y_i-\overline{y})^2\\ &= \mathrm{Var}\,(x) + 2\mathrm{Cov}(x,y) + \mathrm{Var}\,(y)\ .\end{aligned}$$

□

Pour calculer la covariance $\mathrm{Cov}(x,y)$, il est souvent pratique d'utiliser le résultat suivant.

Proposition 3.3. *La covariance vaut la moyenne des produits moins le produit des moyennes,*

$$\mathrm{Cov}(x,y) = \overline{xy} - \overline{x}\times\overline{y}$$

avec $\overline{xy} = \sum_{i=1}^{n} p_i x_i y_i$.

Démonstration. Décomposons la somme de la définition de la covariance,

$$\begin{aligned}\mathrm{Cov}(x,y) &= \sum_{i=1}^{n} p_i(x_i-\overline{x})(y_i-\overline{y})\\ &= \sum_{i=1}^{n} p_i x_i y_i - \overline{x}\sum_{i=1}^{n} p_i y_i - \overline{y}\sum_{i=1}^{n} p_i x_i + \overline{x}\times\overline{y}\sum_{i=1}^{n} p_i\\ &= \overline{xy} - 2\overline{x}\times\overline{y} + \overline{x}\times\overline{y}\\ &= \overline{xy} - \overline{x}\times\overline{y}\ .\end{aligned}$$

□

En dehors de son signe, la valeur de la covariance $\mathrm{Cov}(x,y)$ ne donne pas beaucoup plus d'informations car elle est dépendante de l'échelle des variables x et y. Pour contourner cela, il faut normaliser les observations et considérer la quantité suivante.

Définition 3.2. *La* **corrélation** *(ou* **coefficient de corrélation linéaire de Pearson***) entre les observations de x et de y est définie par*

$$\rho(x,y) = \frac{\mathrm{Cov}(x,y)}{\sqrt{\mathrm{Var}\,(x)}\sqrt{\mathrm{Var}\,(y)}} = \frac{\mathrm{Cov}(x,y)}{\sigma_x\sigma_y}\ .$$

Il est évident que le signe de la corrélation s'interprète toujours comme celui de la covariance. En particulier, il n'y a toujours pas d'interprétation directe lorsque la corrélation est proche de zéro. Cependant, il est possible d'interpréter la valeur de la corrélation car celle-ci est une quantité bornée et indépendante de l'échelle dans laquelle sont exprimées les observations de x et de y.

Théorème 3.1. *La corrélation est bornée par* 1 *en valeur absolue,*

$$-1 \leqslant \rho(x,y) \leqslant 1 \ .$$

Démonstration. Il est possible de voir les observations de x et de y comme deux vecteurs $(x_1-\overline{x},\dots,x_n-\overline{x})'$ et $(y_1-\overline{y},\dots,y_n-\overline{y})'$ de $\mathbb{R}^n$. Ainsi, par la Proposition 3.1, la covariance $\mathrm{Cov}(x,y)$ entre x et y est un produit scalaire (*i.e.* une forme bilinéaire symétrique définie positive, voir le cours d'algèbre linéaire) entre ces vecteurs. De plus, la norme associée à ce produit scalaire est $\sqrt{\mathrm{Var}\,(x)}$. La preuve du théorème est donc une conséquence directe de l'Inégalité de Cauchy-Schwarz que nous re-démontrons ici.

Soit $t \in \mathbb{R}$, nous considérons la variable quantitative $z = x+ty$. En utilisant les Propositions 1.2, 3.1 et 3.2, nous obtenons

$$\mathrm{Var}\,(z) = \mathrm{Var}\,(x+ty) = \mathrm{Var}\,(x) + 2t\mathrm{Cov}(x,y) + t^2\mathrm{Var}\,(y) \ .$$

La variance de z est donc un polynôme du second degré en t. Comme $\mathrm{Var}\,(z) \geqslant 0$ pour tout $t \in \mathbb{R}$, nous savons que ce polynôme a au plus une racine réelle et que son discriminant Δ est négatif,

$$\Delta = 4\mathrm{Cov}(x,y)^2 - 4\mathrm{Var}\,(x)\,\mathrm{Var}\,(y) \leqslant 0 \ .$$

Autrement dit,

$$\begin{aligned} \mathrm{Cov}(x,y)^2 \leqslant \mathrm{Var}\,(x)\,\mathrm{Var}\,(y) &\iff |\mathrm{Cov}(x,y)| \leqslant \sqrt{\mathrm{Var}\,(x)}\sqrt{\mathrm{Var}\,(y)} \\ &\iff |\rho(x,y)| \leqslant 1 \ . \end{aligned}$$

□

La valeur de $\rho(x,y)$ nous renseigne donc sur l'importance du lien potentiel entre x et y. Plus particulièrement, nous avons que plus $|\rho(x,y)|$ est proche de 1, plus la relation affine entre les variables x et y est avérée comme nous allons le voir dans la section suivante.

Exercice 3.1. *Montrer que si les points observés* $(x_1,y_1),\dots,(x_n,y_n)$ *sont sur une droite d'équation "*$y = ax+b$*" alors* $|\rho(x,y)| = 1$*. Réciproquement, montrer que si* $|\rho(x,y)| = 1$ *alors les points observés* $(x_1,y_1),\dots,(x_n,y_n)$ *sont tous alignés le long d'une droite dont on donnera l'équation selon que* $\rho(x,y) = 1$ *ou que* $\rho(x,y) = -1$*. (Utiliser l'exercice 1.4 et la preuve du Théorème 3.1)*

3.3 Régression linéaire

Dans toute cette section, nous supposerons que les poids $p_1,\dots,p_n$ sont uniformes, *i.e.* $p_1 = \cdots = p_n = 1/n$.

Lorsque nous cherchons à établir une relation entre deux variables quantitatives x et y, une première approche simple consiste à regarder si il existe une relation affine (*i.e.* de la forme $y = ax+b$ avec $a,b \in \mathbb{R}$) entre elles. Bien entendu, en pratique, il est presque toujours impossible d'établir une telle relation de façon exacte entre les observations de x et celles de y. Cependant, nous pouvons chercher la droite qui explique "au mieux" y par rapport à x.

Cette procédure s'appelle la *régression linéaire* et elle se formalise comme ce qui suit. Nous cherchons deux nombres réels a et b tels que l'erreur commise en expliquant les observations y_i par $ax_i + b$, $i \in \{1, \ldots, n\}$, soit la plus petite possible au sens des moindres carrés. Autrement dit, nous cherchons $a, b \in \mathbb{R}$ tels que l'erreur moyenne

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \tag{3.1}$$

soit minimale (voir Figure 3.1). Les valeurs de a et de b telles que cette erreur soit minimale donnent l'équation $y = ax + b$ de la *droite de régression*.

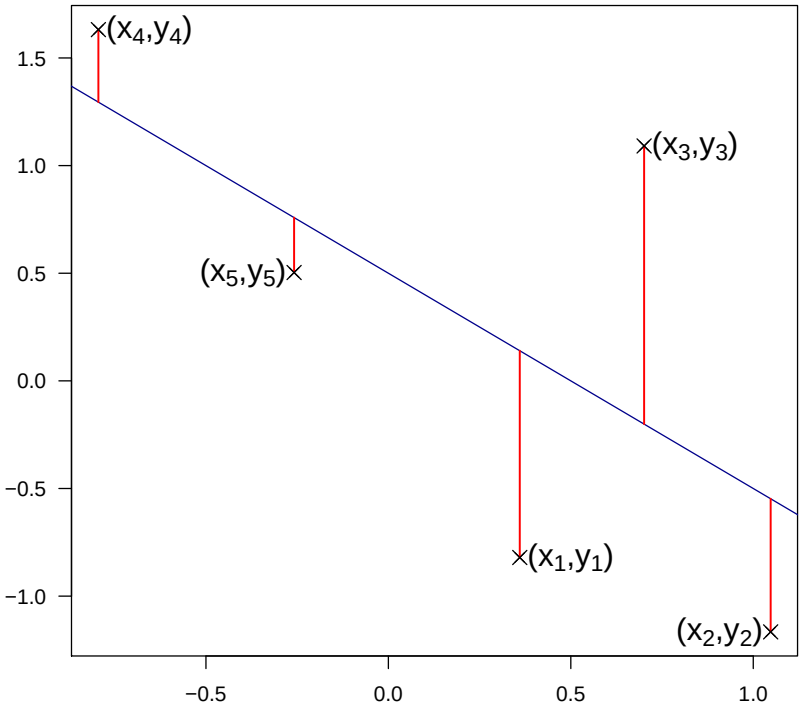


FIGURE 3.1 – Erreurs (en rouge) commises entre les observations et la droite $y = 0.5 - x$.

Théorème 3.2. *Si les variables x et y sont centrées et réduites alors l'erreur (3.1) est minimale pour $a = \mathrm{Cov}(x, y)$ et $b = 0$. Dans ce cas, l'équation de la droite de régression est donc*

$$y = \mathrm{Cov}(x, y) \times x \ .$$

Démonstration. Nous commençons par développer l'erreur (3.1) en utilisant le fait que $\overline{x} =$

$\overline{y} = 0$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i - (ax_i + b))^2 &= \frac{1}{n}\sum_{i=1}^{n}((y_i - ax_i) - b)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i)^2 - \frac{2b}{n}\sum_{i=1}^{n}(y_i - ax_i) + \frac{1}{n}\sum_{i=1}^{n} b^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i)^2 - 2b(\overline{y} - a\overline{x}) + b^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i)^2 + b^2 \ .
\end{aligned}$$

La quantité que nous cherchons à minimiser est la somme de deux termes positifs, l'un dépendant de a et l'autre de b. Nous obtenons donc directement que $b = 0$. Pour déterminer a, nous continuons à développer cette quantité en utilisant que $\overline{x^2} = \mathrm{Var}\,(x) = 1$ et $\overline{y^2} = \mathrm{Var}\,(y) = 1$,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i)^2 &= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{2a}{n}\sum_{i=1}^{n} y_i x_i + \frac{a^2}{n}\sum_{i=1}^{n} x_i^2 \\
&= \overline{y^2} - 2a\overline{xy} + a^2\overline{x^2} \\
&= a^2 - 2\overline{xy} \times a + 1 \ .
\end{aligned}$$

Nous minimisons donc ce polynôme du second degré en $a = \overline{xy} = \mathrm{Cov}(x, y)$. □

Le résultat du théorème 3.2 se généralise à des variables quantitatives x et y qui ne sont plus supposées être centrées réduites de la façon suivante.

Corollaire 3.1. *Dans le cas général, l'erreur* (3.1) *est minimale pour* $a = \mathrm{Cov}(x, y)/\mathrm{Var}\,(x)$ *et* $b = \overline{y} - a\overline{x}$ *et l'équation de la droite de régression est donnée par*

$$y = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}\,(x)}x + \left[\overline{y} - \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}\,(x)}\overline{x}\right] \ .$$

Démonstration. Considérons les variables quantitatives x' et y' obtenues en centrant et en réduisant x et y respectivement,

$$x' = \frac{x - \overline{x}}{\sigma_x} \qquad \text{et} \qquad y' = \frac{y - \overline{y}}{\sigma_y} \ .$$

Pour ces variables, le Théorème 3.2 nous donne l'équation de la droite de régression,

$$\begin{aligned}
y' = \mathrm{Cov}\,(x', y') \times x' &\iff \frac{y - \overline{y}}{\sigma_y} = \mathrm{Cov}\left(\frac{x - \overline{x}}{\sigma_x}, \frac{y - \overline{y}}{\sigma_y}\right) \times \frac{x - \overline{x}}{\sigma_x} \\
&\iff y - \overline{y} = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y} \times \frac{\sigma_y}{\sigma_x} \times (x - \overline{x}) \\
&\iff y = \underbrace{\frac{\mathrm{Cov}(x, y)}{\sigma_x^2}}_{=a} x + \underbrace{\overline{y} - \frac{\mathrm{Cov}(x, y)}{\sigma_x^2}\overline{x}}_{=b} \ .
\end{aligned}$$

□

Nous retrouvons dans ce résultat le fait que si $\mathrm{Cov}(x, y) > 0$, alors les variables varient dans le même sens et inversement (voir Figures 3.2 et 3.3).

Exercice 3.2. *Le Corollaire 3.1 donne la droite de régression de y par rapport à x d'équation $y = ax + b$. Si $a \neq 0$, nous pouvons en déduire que $x = a'y + b'$ avec $a' = 1/a$ et $b' = -b/a$. Calculer l'équation de la droite de régression de x sur y (i.e. trouver $\bar{a}, \bar{b} \in \mathbb{R}$ avec $x = \bar{a}y + \bar{b}$) et comparer-la avec $x = a'y + b'$. Conclure que les droites de régression de y sur x et de x sur y ne sont pas les mêmes.*

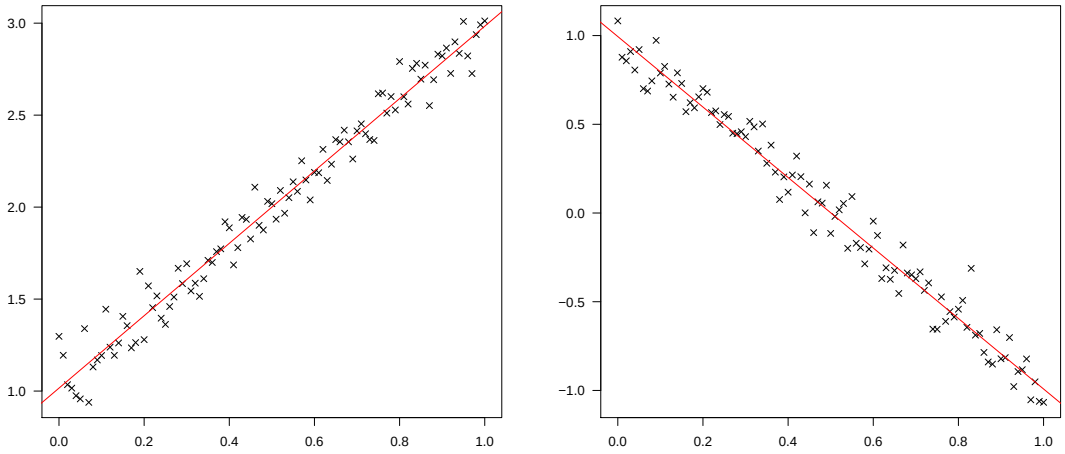


FIGURE 3.2 – Tracés des observations (x_i, y_i) et de la droite de régression (en rouge) associée. A gauche, $\rho(x, y) = 0.9855$ et à droite, $\rho(x, y) = -0.9863$.

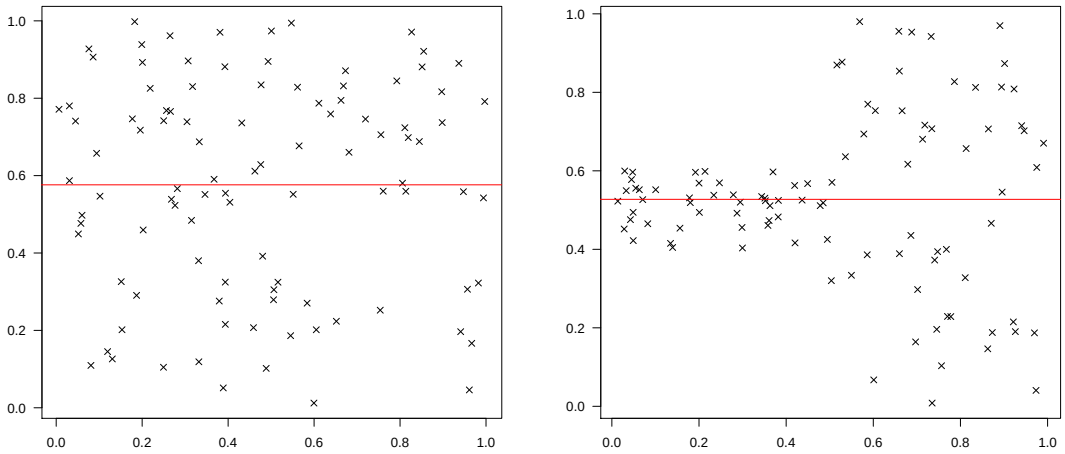


FIGURE 3.3 – Tracés des observations (x_i, y_i) et de la droite de régression (en rouge) associée. Dans les deux cas, $\rho(x, y) = 0$ et nous voyons qu'il n'est pas possible d'interpréter cette valeur car les variables x et y peuvent être indépendantes (à gauche) ou liées (à droite).

3.4 Corrélations de rang

Dans toute cette section, nous supposerons que les poids $p_1, \dots, p_n$ sont uniformes, *i.e.* $p_1 = \cdots = p_n = 1/n$.

Le coefficient de corrélation de Pearson $\rho(x, y)$ est un bon indicateur de l'alignement des observations et donc de l'existence d'une relation affine entre les variables x et y. Cependant, ces variables peuvent être liées sans pour autant que cette relation soit affine. Les coefficients de corrélation que nous allons introduire dans cette section mesurent l'existence d'une relation entre x et y sans en préciser la nature. Ces deux coefficients sont basés sur les rangs des observations dans leurs versions ordonnées.

3.4.1 Corrélation de Spearman

Pour chaque $i \in \{1, \dots, n\}$, nous définissons r_i comme le *rang* de l'observation x_i dans la version ordonnée $x_{(1)} \leqslant \cdots \leqslant x_{(n)}$. De même, s_i est le rang de y_i parmi les $y_{(1)} \leqslant \cdots \leqslant y_{(n)}$. En cas d'égalité entre plusieurs observations, les rangs de celles-ci sont tous pris égaux à la valeur moyenne des rangs concernés. Dans la suite, les rangs $r_1, \dots, r_n$ et $s_1, \dots, s_n$ seront traités comme les observations des variables de rang r et s respectivement.

Exemple Supposons que nous ayons observé

$$x_1 = 4.2 \ , \ x_2 = 3.1 \ , \ x_3 = 5.1 \ , \ x_4 = 3.1 \text{ et } x_5 = 1.3 \ .$$

La version ordonnée de nos observations est donc

$$x_{(1)} = 1.3 \ , \ x_{(2)} = 3.1 \ , \ x_{(3)} = 3.1 \ , \ x_{(4)} = 4.2 \text{ et } x_{(5)} = 5.1 \ .$$

Dans ce classement, le rang de x_1 est $r_1 = 4$, celui de x_3 est $r_3 = 5$ et celui de x_5 est $r_5 = 1$. Puisque x_2 et x_4 sont égaux et que leurs rangs auraient dû valoir 2 ou 3, nous posons $r_2 = r_4 = (2 + 3)/2 = 2.5$. Au final, nous avons donc

$$r_1 = 4 \ , \ r_2 = 2.5 \ , \ r_3 = 5 \ , \ r_4 = 2.5 \text{ et } r_5 = 1 \ .$$

Définition 3.3. *Le* **coefficient de corrélation de Spearman** *entre les observations des variables couplées x et y est le coefficient de corrélation linéaire entre les rangs r et s,*

$$\rho_S(x, y) = \rho(r, s) \ .$$

Par définition, la corrélation de Spearman est toujours comprise entre -1 et 1 et sa valeur s'interprète de façon similaire à celle de la corrélation de Pearson. En effet, si $|\rho_S(x, y)|$ est proche de 1, la répartition des points observés $(x_1, y_1), \dots, (x_n, y_n)$ est proche de la courbe d'une fonction monotone (croissante si $\rho_S(x, y) > 0$ et décroissante si $\rho_S(x, y) < 0$). Nous pourrons alors conclure à l'existence d'une relation monotone entre les variables x et y (voir Figure 3.4). Comme pour la corrélation de Pearson, si $\rho_S(x, y)$ est proche de 0, il n'y a, a priori, aucune interprétation.

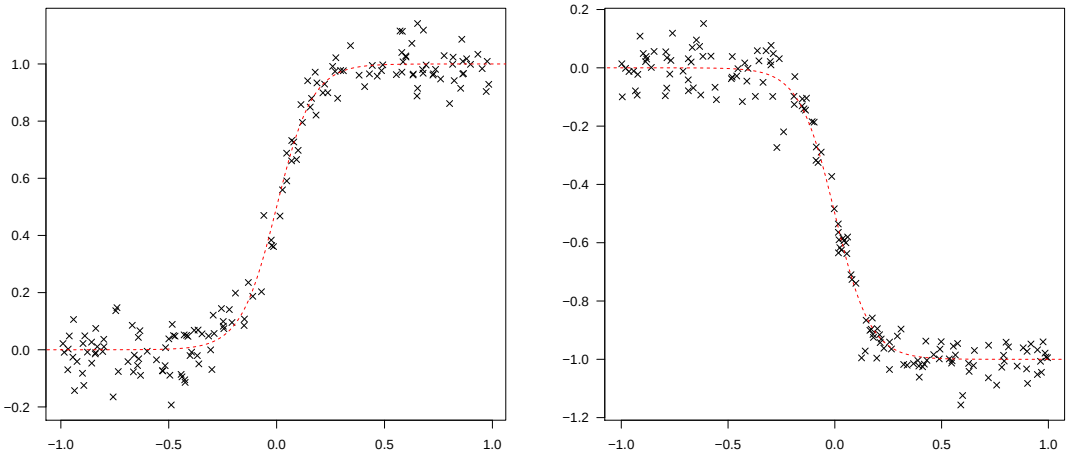


FIGURE 3.4 – Pour ces observations (x_i, y_i), la corrélation de Spearman vaut $\rho_S(x, y) = 0.8787$ (à gauche) et $\rho_S(x, y) = -0.8982$ (à droite).

Exemple Pour x, nous reprenons les observations de l'exemple précédents et, pour y, nous observons

$$y_1 = 3.8 \ , \ y_2 = -0.6 \ , \ y_3 = -1.2 \ , \ y_4 = -3.5 \text{ et } y_5 = -3.5 \ .$$

Ainsi, les rangs sont donnés par

$$s_1 = 5 \ , \ s_2 = 4 \ , \ s_3 = 3 \ , \ s_4 = 1.5 \text{ et } s_5 = 1.5 \ .$$

La corrélation de Spearman entre x et y vaut donc

$$\rho_S(x, y) = \rho(r, s) = \frac{\mathrm{Cov}(r, s)}{\sqrt{\mathrm{Var}\,(r)}\sqrt{\mathrm{Var}\,(s)}} = \frac{1.05}{\sqrt{1.9 \times 1.9}} \simeq 0.553 \ .$$

Les rangs r et s vérifient certaines propriétés qui évitent de refaire systématiquement les mêmes calculs lorsque nous cherchons à obtenir la corrélation de Spearman.

Proposition 3.4. *La moyenne des rangs est toujours égale à*

$$\overline{r} = \overline{s} = \frac{n+1}{2} \ .$$

Si il n'y a aucun ex-æquo ni parmi les observations de x, ni parmi celles de y, alors,

$$\mathrm{Var}\,(r) = \mathrm{Var}\,(s) = \frac{n^2 - 1}{12}$$

et

$$\rho_S(x, y) = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (r_i - s_i)^2 \ .$$

Démonstration. Par construction, il est évident que

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} s_i = \sum_{k=1}^{n} k = \frac{n(n+1)}{2} \ .$$

Donc, nous avons $\overline{r} = \overline{s} = (n+1)/2$.

L'absence d'égalité parmi les observations de x et parmi celles de y implique que les rangs $r_1, \ldots, r_n$ et $s_1, \ldots, s_n$ sont des permutations des entiers de 1 à n. Par conséquent,

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} s_i^2 = \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} \ .$$

La variance de r se calcule alors facilement,

$$\begin{aligned} \mathrm{Var}\,(r) &= \overline{r^2} - \overline{r}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} r_i^2 - \frac{(n+1)^2}{4} \\ &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= \frac{(n+1)(n-1)}{12} = \frac{n^2-1}{12} \ , \end{aligned}$$

et de même pour $\mathrm{Var}\,(s) = (n^2-1)/12$.

Pour le calcul de la corrélation de Spearman et parce que $\overline{r} = \overline{s}$, nous avons

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(r_i - s_i)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left((r_i - \overline{r}) - (s_i - \overline{s})\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(r_i - \overline{r})^2 - \frac{2}{n}\sum_{i=1}^{n}(r_i - \overline{r})(s_i - \overline{s}) + \frac{1}{n}\sum_{i=1}^{n}(s_i - \overline{s})^2 \\ &= \mathrm{Var}\,(r) - 2\mathrm{Cov}(r,s) + \mathrm{Var}\,(s) \\ &= \frac{n^2-1}{6} - 2\mathrm{Cov}(r,s) \ . \end{aligned}$$

Ainsi, nous obtenons $\mathrm{Cov}(r,s) = \dfrac{n^2-1}{12} - \dfrac{1}{2n}\displaystyle\sum_{i=1}^{n}(r_i - s_i)^2$ et donc

$$\rho_S(x,y) = \rho(r,s) = \frac{\mathrm{Cov}(r,s)}{\sqrt{\mathrm{Var}\,(r)}\sqrt{\mathrm{Var}\,(s)}} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(r_i - s_i)^2 \ .$$

□

Rappel Pour un entier $n > 0$, la somme des nombres entiers de 1 à n se calcule facilement grâce à la remarque suivante qui consiste à faire la somme de deux façons différentes (de 1 à

n et de n à 1),

$$\begin{aligned}\sum_{k=1}^{n} k &= \frac{1}{2}\left(\sum_{k=1}^{n} k + \sum_{k=1}^{n}(n+1-k)\right)\\ &= \frac{1}{2}\sum_{k=1}^{n}(k+n+1-k)\\ &= \frac{1}{2}\sum_{k=1}^{n}(n+1) = \frac{n(n+1)}{2}\ .\end{aligned}$$

La somme des carrés des n premiers nombres entiers se calcule en remarquant que

$$\sum_{k=1}^{n}(k+1)^3 = \sum_{k=1}^{n} k^3 - 1 + (n+1)^3 \quad \text{et} \quad \sum_{k=1}^{n}(k+1)^3 = \sum_{k=1}^{n} k^3 + 3\sum_{k=1}^{n} k^2 + 3\sum_{k=1}^{n} k + \sum_{k=1}^{n} 1\ .$$

En identifiant les deux égalités et en utilisant le résultat précédent, nous obtenons,

$$\begin{aligned}(n+1)^3 - 1 = 3\sum_{k=1}^{n} k^2 + \frac{3n(n+1)}{2} + n \quad &\Longleftrightarrow \quad \sum_{k=1}^{n} k^2 = \frac{2(n+1)^3 - 2 - 3n(n+1) - 2n}{6}\\ &\Longleftrightarrow \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}\ .\end{aligned}$$

3.4.2 Corrélation de Kendall

Nous supposerons dans cette sous-section qu'il n'y a aucun ex-æquo ni parmi les observations de x, ni parmi celles de y.

Nous présentons maintenant une autre mesure de corrélation basée sur les rangs. Pour cela, nous introduisons la notion de *variation concordante*. Soient $i, j \in \{1, \ldots, n\}$ tels que $i < j$, nous disons qu'il y a une variation concordante entre i et j si

$$\text{"}r_i < r_j \text{ et } s_i < s_j\text{"} \qquad \text{ou} \qquad \text{"}r_i > r_j \text{ et } s_i > s_j\text{"}\ .$$

Remarquons que, en l'absence d'ex-æquo parmi les observations, toutes les paires $i < j$ sont soit concordantes, soit non-concordantes. Nous notons R le nombre de variations concordantes parmi tous les choix $i < j$ possibles.

Définition 3.4. *Le* **coefficient de corrélation de Kendall** *entre les observations des variables couplées x et y est défini par*

$$\rho_K(x,y) = \frac{4R}{n(n-1)} - 1\ .$$

Exercice 3.3. *Montrer que*

$$0 \leqslant R \leqslant \frac{n(n-1)}{2}$$

et en déduire que $-1 \leqslant \rho_K(x,y) \leqslant 1$.

Comme pour les autres corrélations, celle de Kendall ne s'interprète pas lorsqu'elle est proche de 0. Si $|\rho_K(x,y)|$ est proche de 1, nous pouvons en déduire que les écarts $x_i - x_j$ et $y_i - y_j$, pour $i < j$, sont liés. Cette situation suggère donc l'existence d'une relation entre les variables x et y.

Exemple Prenons les observations suivantes avec $n = 5$,

$$x_1 = 4.2 \ , \ x_2 = 3.1 \ , \ x_3 = 5.1 \ , \ x_4 = 2.1 \ , \ x_5 = 1.3$$

et

$$y_1 = 3.4 \ , \ y_2 = -0.6 \ , \ y_3 = -1.2 \ , \ y_4 = -3.5 \ , \ y_5 = 3.8 \ .$$

Nous avons les rangs suivants,

$$r_1 = 4 \ , \ r_2 = 3 \ , \ r_3 = 5 \ , \ r_4 = 2 \ , \ r_5 = 1$$

et

$$s_1 = 4 \ , \ s_2 = 3 \ , \ s_3 = 2 \ , \ s_4 = 1 \ , \ s_5 = 5 \ .$$

Les paires $(i,j) \in \{1,\dots,n\}^2$ telles que $i < j$ qui sont concordantes sont

$$(1,2) \ , \ (1,4) \ , \ (2,4) \text{ et } (3,4) \ .$$

Nous avons donc $R = 4$ et

$$\rho_K(x,y) = \frac{4 \times 4}{5 \times 4} - 1 = -0.2 \ .$$

3.5 Distance du χ^2 à l'indépendance

Toutes les méthodes que nous avons présentées dans les sections précédentes étaient relatives à des observations de variables quantitatives. Cependant, toutes les variables ne peuvent pas être représentées comme des mesures de grandeurs physiques. Certaines variables, dites *qualitatives*, ne peuvent prendre qu'un nombre fini d'états (appelés aussi des modes). Ces états ne sont pas, en général, des mesures.

Exemple Supposons que nous observions la couleur des yeux de plusieurs personnes. La variable relative à ces observations est qualitative car elle ne peut prendre que des valeurs parmi {BLEU, MARRON, VERT}. En particulier, il n'est ni possible de faire des calculs avec ces observations, ni de les ordonner pour étudier les corrélations vues précédemment.

Soit un entier $n > 0$, nous considérons que nous disposons des observations $u_1,\dots,u_n$ et $v_1,\dots,v_n$ de deux variables qualitatives couplées u et v respectivement. De plus, la variable u ne peut prendre que des valeurs dans $\{\alpha_1,\dots,\alpha_r\}$ et la variable v ne peut prendre que des valeurs dans $\{\beta_1,\dots,\beta_s\}$ avec $r, s > 0$ deux nombres entiers.

Contrairement aux sections précédentes, nous allons présenter une méthode pour justifier l'absence de relation entre les variables u et v, *i.e.* l'indépendance. Pour cela, nous allons considérer les effectifs suivants, pour tout $i \in \{1,\dots,r\}$ et tout $j \in \{1,\dots,s\}$,

$$n_{i,j} = \sharp\{k \in \{1,\dots,n\} \text{ tels que } (u_k, v_k) = (\alpha_i, \beta_j)\} \ ,$$

$$n_{i\bullet} = \sum_{j=1}^{s} n_{i,j} = \sharp\{k \in \{1,\dots,n\} \text{ tels que } u_k = \alpha_i\}$$

et

$$n_{\bullet j} = \sum_{i=1}^{r} n_{i,j} = \sharp\{k \in \{1,\dots,n\} \text{ tels que } v_k = \beta_j\} \ .$$

Ces effectifs sont généralement représentés dans une *table de contingence* avec ses marges qui contiennent les effectifs sommés $n_{i\bullet}$ et $n_{\bullet j}$ (voir Figure 3.5). Bien sûr, les effectifs en ligne et en colonne sont reliés par

$$n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{i,j} = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{s} n_{\bullet j} \ .$$

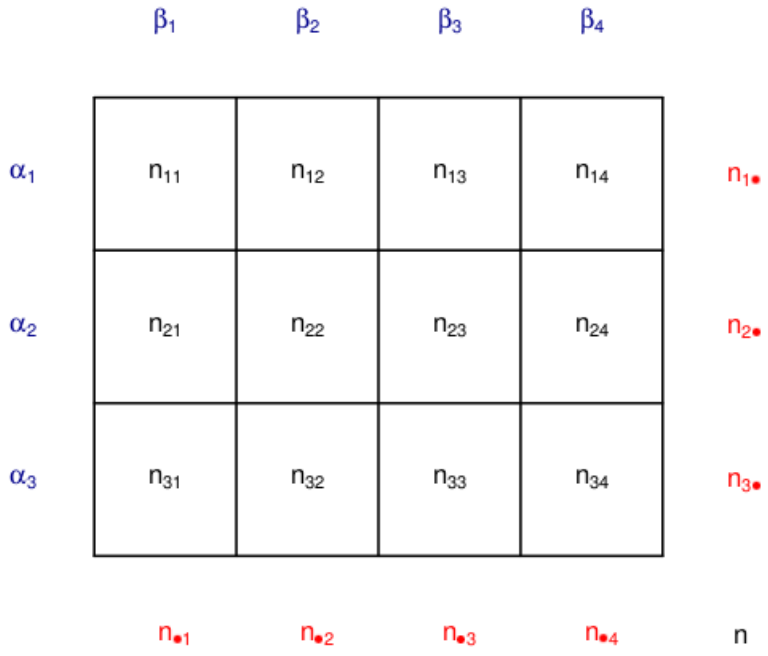


FIGURE 3.5 – Table de contingence pour $r = 3$ et $s = 4$ avec ses marges (en rouge) et les modes (en bleu).

L'objectif de la méthode que nous présentons ici va être de comparer cette table de contingence observée avec une table de contingence théorique qui correspond à ce que nous aurions dû observer en cas d'indépendance entre les variables u et v. Si ces deux tables sont assez "proches", alors nous pourrons accepter l'idée d'indépendance entre nos variables.

Considérons $i \in \{1,\dots,r\}$ et $j \in \{1,\dots,s\}$, quelle doit être la valeur théorique $n^*_{i,j}$ du nombre d'observations de la paire (α_i, β_j) en cas d'indépendance ? Pour répondre à cette question, il faut comprendre ce que l'indépendance implique en terme d'effectifs d'observations. Prenons la ligne des $n_{i\bullet}$ observations de α_i. Si le fait d'avoir observé $u = \alpha_i$ n'influence pas la

valeur prise par la variable v, alors les $n_{\bullet j}$ observations de β_j se répartissent le long de la ligne des α_i avec les proportions $n_{i\bullet}/n$. Il faut noter que cet argument est symétrique entre u et v et que nous pouvons aussi en déduire que les $n_{i\bullet}$ observations de α_i se répartissent le long de la colonne de β_j avec les proportions $n_{\bullet j}/n$. Ainsi, nous obtenons les effectifs théoriques sous l'hypothèse d'indépendance,

$$n^*_{i,j} = \frac{n_{i\bullet} \times n_{\bullet j}}{n} \ .$$

Définition 3.5. *La* **distance du χ^2 à l'indépendance** *est définie par*

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{i,j} - n^*_{i,j})^2}{n^*_{i,j}} \ .$$

La quantité χ^2 n'est pas une distance au sens mathématique (elle n'est pas symétrique, ...) mais elle traduit l'idée de proximité entre les tables de contingence observée et théorique. Plus χ^2 sera proche de 0, plus l'hypothèse d'indépendance sera acceptable.

De plus, il est important de noter la normalisation par $n^*_{i,j}$ des termes de la somme définissant χ^2. Leurs présences assurent que les cellules des tables de contingence sont traitées de la même façon si elles contiennent un effectif important ou faible. En effet, un écart $n_{i,j} - n^*_{i,j} = 1$ aura plus d'importance si $n^*_{i,j} = 1$ que si $n^*_{i,j} = 1000$, par exemple.

Afin de donner un ordre de grandeur pour χ^2, nous présentons ici le calcul de la distance dans un cas particulier qui correspond à la situation extrême où u et v ne sont absolument pas indépendantes. Pour cela, nous considérons que les variables ont les mêmes modes (*i.e.* $r = s$ et, pour tout $i \in \{1, \dots, r\}$, $\alpha_i = \beta_i$) et que les observations sont toujours égales (*i.e.* pour tout $i \in \{1, \dots, n\}$, $u_i = v_i$). La table de contingence des effectifs observés $n_{i,j}$ est donc donnée par

$$[n_{i,j}]_{1 \leqslant i,j \leqslant r} = \begin{bmatrix} n/r & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & n/r \end{bmatrix} \ .$$

Pour tout $i, j \in \{1, \dots, r\}$, nous avons $n_{i\bullet} = n_{\bullet j} = n/r$ et donc $n^*_{i,j} = n/r^2$. Ainsi, la table de contingence théorique est donnée par

$$\left[n^*_{i,j}\right]_{1 \leqslant i,j \leqslant r} = \begin{bmatrix} n/r^2 & n/r^2 & \cdots & n/r^2 \\ n/r^2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & n/r^2 \\ n/r^2 & \cdots & n/r^2 & n/r^2 \end{bmatrix} \ .$$

Nous pouvons maintenant calculer χ^2,

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{r}\sum_{j=1}^{r}\frac{(n_{i,j}-n_{i,j}^*)^2}{n_{i,j}^*} = \sum_{i=1}^{r}\left\{\frac{(n_{i,i}-n_{i,i}^*)^2}{n_{i,i}^*} + \sum_{\substack{j=1\\ j\neq i}}^{r}\frac{(n_{i,j}-n_{i,j}^*)^2}{n_{i,j}^*}\right\} \\
&= \sum_{i=1}^{r}\left\{\frac{(n/r-n/r^2)^2}{n/r^2} + \sum_{\substack{j=1\\ j\neq i}}^{r}\frac{(0-n/r^2)^2}{n/r^2}\right\} = \sum_{i=1}^{r}\left\{\frac{n(r-1)^2}{r^2} + \sum_{\substack{j=1\\ j\neq i}}^{r}\frac{n}{r^2}\right\} \\
&= r\times\left\{\frac{n(r-1)^2}{r^2} + \frac{n(r-1)}{r^2}\right\} = n(r-1)\ .
\end{aligned}$$

Plus généralement, nous avons le résultat suivant qui nous permet de donner l'ordre de grandeur de χ^2 en cas d'absence totale d'indépendance.

Proposition 3.5. *Nous avons*

$$\chi^2 \leqslant n\times\sqrt{(r-1)(s-1)}\ .$$

Démonstration. Admis. □

Chapitre 4

Observations de plusieurs variables couplées

4.1 Introduction

Dans le chapitre précédent, nous avons introduit quelques outils permettant de discuter de l'existence d'une relation entre deux variables. Parmi ces outils, certains peuvent être généralisés pour considérer les relations potentielles entre un nombre arbitraire de variables. Nous nous restreindrons dans ce chapitre à l'étude d'un jeu de données relatif à n observations de p variables quantitatives couplées.

Ainsi, dans la suite de ce chapitre, nous considérerons p variables quantitatives $x^1, \dots, x^p$ et n vecteurs observés $(x_1^1, \dots, x_1^p)', \dots, (x_n^1, \dots, x_n^p)' \in \mathbb{R}^p$ pondérés par des poids $p_1, \dots, p_n > 0$ normalisés. Pour $i \in \{1, \dots, n\}$ et $j \in \{1, \dots, p\}$, nous notons donc x_i^j la $i^{\text{ème}}$ observation de la $j^{\text{ème}}$ variable.

Afin de garder des notations simples, pour tout $j \in \{1, \dots, p\}$, nous noterons x^j pour désigner la $j^{\text{ème}}$ variable ou pour désigner le vecteur de ses observations $(x_1^j, \dots, x_n^j)' \in \mathbb{R}^n$ selon le contexte. De même, pour tout $i \in \{1, \dots, n\}$, nous noterons $x_i = (x_i^1, \dots, x_i^p)' \in \mathbb{R}^p$ pour désigner le vecteur des $i^{\text{ème}}$ observations de chaque variable.

4.2 Matrices de covariance et de corrélation

Les notions de covariance et de corrélation linéaire que nous avons vues étaient définies entre deux variables. Elles se généralisent naturellement à p variables en considérant toutes les paires de variables (x^i, x^j) pour $i, j \in \{1, \dots, p\}$. Les valeurs obtenues sont alors présentées sous forme matricielle.

Définition 4.1. *La* **matrice de covariance** $\Sigma = (\Sigma_{ij})_{1 \leqslant i,j \leqslant p}$ *est la matrice carrée de taille*

$p \times p$ dont les entrées sont données par $\Sigma_{ij} = \mathrm{Cov}(x^i, x^j)$ pour $i, j \in \{1, \dots, p\}$.

$$\Sigma = \begin{bmatrix} \mathrm{Var}\left(x^1\right) & \mathrm{Cov}(x^1, x^2) & \cdots & \mathrm{Cov}(x^1, x^{p-1}) & \mathrm{Cov}(x^1, x^p) \\ \mathrm{Cov}(x^2, x^1) & \mathrm{Var}\left(x^2\right) & \cdots & \mathrm{Cov}(x^2, x^{p-1}) & \mathrm{Cov}(x^2, x^p) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathrm{Cov}(x^{p-1}, x^1) & \mathrm{Cov}(x^{p-1}, x^2) & \cdots & \mathrm{Var}\left(x^{p-1}\right) & \mathrm{Cov}(x^{p-1}, x^p) \\ \mathrm{Cov}(x^p, x^1) & \mathrm{Cov}(x^p, x^2) & \dots & \mathrm{Cov}(x^p, x^{p-1}) & \mathrm{Var}\left(x^p\right) \end{bmatrix}$$

Par définition, la diagonale de Σ contient les variances des variables. De plus, grâce à la Proposition 3.1, nous savons que cette matrice est symétrique, *i.e.* pour tout $i, j \in \{1, \dots, p\}$, $\Sigma_{ij} = \Sigma_{ji}$.

Définition 4.2. *La* **matrice de corrélation** $\mathfrak{C} = (\mathfrak{C}_{ij})_{1 \leqslant i,j \leqslant p}$ *est la matrice carrée de taille $p \times p$ dont les entrées sont données par $\mathfrak{C}_{ij} = \rho(x^i, x^j)$ pour $i, j \in \{1, \dots, p\}$.*

$$\mathfrak{C} = \begin{bmatrix} 1 & \rho(x^1, x^2) & \cdots & \rho(x^1, x^{p-1}) & \rho(x^1, x^p) \\ \rho(x^2, x^1) & 1 & \cdots & \rho(x^2, x^{p-1}) & \rho(x^2, x^p) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho(x^{p-1}, x^1) & \rho(x^{p-1}, x^2) & \cdots & 1 & \rho(x^{p-1}, x^p) \\ \rho(x^p, x^1) & \rho(x^p, x^2) & \dots & \rho(x^p, x^{p-1}) & 1 \end{bmatrix}$$

Cette matrice n'a que des 1 sur sa diagonale et est également symétrique car $\rho(x^i, x^j) = \rho(x^j, x^i)$ pour tout $i, j \in \{1, \dots, p\}$.

En pratique, nous utiliserons principalement la matrice de covariance. Une des raisons de ce choix est la simplicité de l'écriture matricielle de Σ. Pour illustrer cela, nous considérons la *matrice des données centrées* X. Cette matrice est de taille $n \times p$ et, pour tout $i \in \{1, \dots, n\}$ et $j \in \{1, \dots, p\}$, l'entrée X_{ij} vaut $x_i^j - \overline{x^j}$ où $\overline{x^j}$ est la moyenne des $x_1^j, \dots, x_n^j$ pondérée par les poids $p_1, \dots, p_n$.

$$X = \begin{bmatrix} x_1^1 - \overline{x^1} & x_1^2 - \overline{x^2} & \cdots & x_1^p - \overline{x^p} \\ x_2^1 - \overline{x^1} & x_2^2 - \overline{x^2} & \cdots & x_2^p - \overline{x^p} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 - \overline{x^1} & x_1^n - \overline{x^2} & \cdots & x_n^p - \overline{x^p} \end{bmatrix}$$

Si l'on considère que les données ont été obtenues en observant p variables sur n individus, chaque ligne de X est relative à un individu et chaque colonne de X est relative à une variable. Nous introduisons aussi la *matrice des poids* W qui est la matrice diagonale de taille $n \times n$ donnée par

$$W = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & p_n \end{bmatrix}.$$

Proposition 4.1. *La matrice de covariance s'écrit*

$$\Sigma = {}^tXWX\ .$$

Démonstration. Puisque ${}^tX$ est de taille $p \times n$, W de taille $n \times n$ et X de taille $n \times p$, la matrice ${}^tXWX$ est bien de taille $p \times p$. De plus, pour tout $i, j \in \{1, \dots, p\}$, nous avons

$$\begin{aligned}\left({}^tXWX\right)_{ij} &= \sum_{k=1}^{n} {}^tX_{ik}(WX)_{kj} = \sum_{k=1}^{n} {}^tX_{ik} \sum_{\ell=1}^{n} W_{k\ell} X_{\ell j} \\ &= \sum_{k=1}^{n} W_{kk} X_{ki} X_{kj} = \sum_{k=1}^{n} p_k \left(x_k^i - \overline{x^i}\right)\left(x_k^j - \overline{x^j}\right) \\ &= \mathrm{Cov}(x^i, x^j) = \Sigma_{ij} \ .\end{aligned}$$

□

Nous avons vu que le coefficient de corrélation linéaire $\rho(x, y)$ nous permet de mesurer la "proximité" des points $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$ à une droite. Lorsque nous disposons d'un nombre p de variables, en général, il n'est plus possible de décrire la relation entre toutes les variables en terme de simples droites (du moins, quand $p > 2$) et il nous faut faire intervenir la quantité suivante qui généralise la notion de corrélation linéaire.

Définition 4.3. *Considérons une variable couplée y supplémentaire et ses n observations $y_1, \dots, y_n \in \mathbb{R}$. Le* **coefficient R de corrélation multiple** *de y avec les variables $x^1, \dots, x^p$ est la corrélation maximale entre y et toutes les combinaisons linéaires des x^j, $j \in \{1, \dots, p\}$,*

$$R = \sup\left\{\rho(y, a_1x^1 + \cdots + a_px^p) \text{ avec } (a_1, \dots, a_p)' \in \mathbb{R}^p\right\} \ .$$

Exercice 4.1. *A priori, la définition donne $R \in [-1, 1]$. En comparant $\rho(y, a_1x^1 + \cdots + a_px^p)$ et $\rho(y, -a_1x^1 - \cdots - a_px^p)$ pour un vecteur $(a_1, \dots, a_p)' \in \mathbb{R}^p$ arbitraire, montrer que nous avons toujours $0 \leqslant R \leqslant 1$.*

Comme pour la corrélation linéaire, le coefficient de corrélation multiple ne s'interprète pas lorsqu'il est proche de zéro. Quand R est proche de 1, cela suggère que les points $(y_1, x_1^1, \dots, x_1^p)$, $\dots$, $(y_n, x_n^1, \dots, x_n^p) \in \mathbb{R}^{p+1}$ sont "proches" d'un sous-espace linéaire de dimension p. Autrement dit, cela suggère l'existence de $a_1, \dots, a_p, b \in \mathbb{R}$ tels que les points observés soient décrits de façon "acceptable" par la relation

$$y = a_1x^1 + \cdots + a_px^p + b \ .$$

Dans le cas particulier où $p = 1$, nous avons $R = |\rho(y, x^1)|$ et nous retrouvons bien la même interprétation : quand R est proche de 1, les points $(y_1, x_1^1), \dots, (y_n, x_n^1)$ du plan $\mathbb{R}^2$ sont "proches" d'une droite, *i.e.* d'un sous-espace de dimension $p = 1$.

Exemple Dans le cas $p = 2$, il est possible de donner une interprétation géométrique simple du coefficient R. En effet, dans ce cas, nous cherchons si les points $(y_1, x_1^1, x_2^2), \dots, (y_n, x_n^1, x_n^2) \in \mathbb{R}^3$ sont "proches" d'un plan (*i.e.* un sous-espace de dimension $p = 2$). Pour les données centrées, nous avons vu que la covariance est un produit scalaire dont la norme associée est l'écart-type (voir la démonstration du Théorème 3.1). De plus, nous avons

$$\mathrm{Cov}(y, a_1x^1 + a_2x^2) = \sqrt{\mathrm{Var}\,(y)} \times \sqrt{\mathrm{Var}\,(a_1x^1 + a_2x^2)} \times \rho(y, a_1x^1 + a_2x^2)$$

et donc la corrélation $\rho(y, a_1x^1 + a_2x^2)$ joue le rôle du cosinus de l'angle entre $y - \overline{y}$ et $a_1(x^1 - \overline{x^1}) + a_2(x^2 - \overline{x^2})$ (faire l'analogie avec le produit scalaire usuel entre deux vecteurs u et v, $u \cdot v = \|u\| \times \|v\| \times \cos(\widehat{u, v})$).

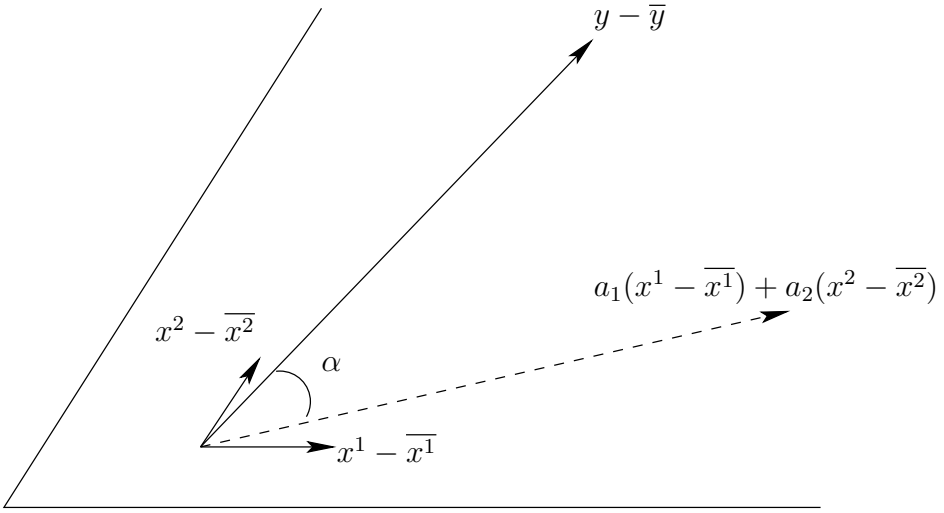


FIGURE 4.1 – La corrélation $\rho(y, a_1x^1 + a_2x^2)$ joue le rôle du $\cos(\alpha)$ où α est l'angle formé par $y - \overline{y}$ et $a_1(x^1 - \overline{x^1}) + a_2(x^2 - \overline{x^2})$.

A l'aide de cette analogie, nous comprenons que le choix de $(a_1, a_2) \in \mathbb{R}^2$ qui rend $\rho(y, a_1x^1 + a_2x^2)$ maximal est celui tel que l'angle formé par $y - \overline{y}$ et $a_1(x^1 - \overline{x^1}) + a_2(x^2 - \overline{x^2})$ soit minimal (car cela maximise le $\cos(\alpha)$). En d'autres termes, $\rho(y, a_1x^1 + a_2x^2)$ est maximal pour $(a_1, a_2) \in \mathbb{R}^2$ tels que $a_1(x^1 - \overline{x^1}) + a_2(x^2 - \overline{x^2})$ soit égal à la projection orthogonale π_y de $y - \overline{y}$ sur le plan engendré par $x^1 - \overline{x^1}$ et $x^2 - \overline{x^2}$. Ainsi, nous obtenons que $R = \rho(y, \pi_y)$.

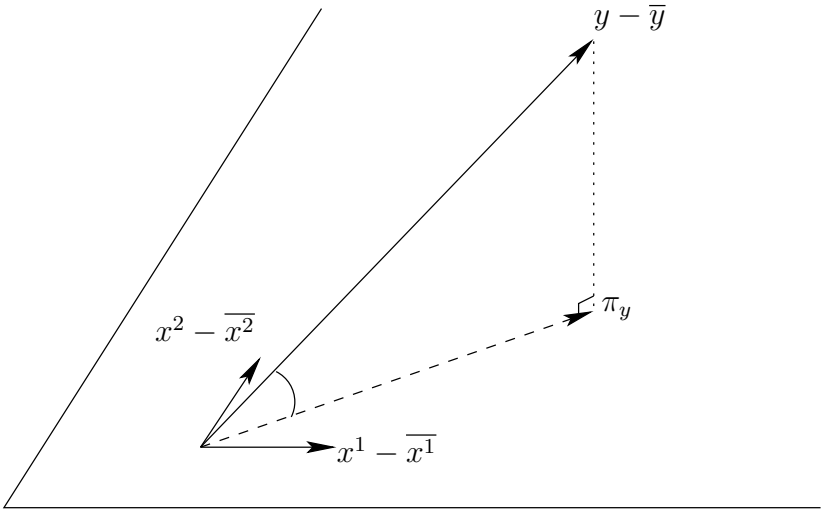


FIGURE 4.2 – La corrélation $\rho(y, a_1x^1 + a_2x^2)$ maximale est atteinte pour $(a_1, a_2) \in \mathbb{R}^2$ tels que $a_1x^1 + a_2x^2$ soit la projection orthogonale π_y de $y - \overline{y}$ sur le plan engendré par $x^1 - \overline{x^1}$ et $x^2 - \overline{x^2}$ et $R = \rho(y, \pi_y)$.

Le résultat obtenu dans cet exemple se généralise à $p > 2$ par des arguments similaires et mène à la proposition suivante.

Proposition 4.2. *Considérons le vecteur centré* $Y = (y_1 - \overline{y}, \ldots, y_n - \overline{y})' \in \mathbb{R}^n$ *et notons* $\mathrm{Im}(X)$ *l'espace engendré par les colonnes de* X*. Si* $\dim(\mathrm{Im}(X)) = p$*, la projection orthogonale* π_y *de* Y *sur* $\mathrm{Im}(X)$ *est donnée par*

$$\pi_y = X({}^tXX)^{-1}\,{}^tXY \ .$$

De plus, nous avons $R = \rho(y, \pi_y)$*.*

Démonstration. Nous allons montrer que si $\dim(\mathrm{Im}(X)) = p$ alors $M = X({}^tXX)^{-1}\,{}^tX$ est la projection orthogonale sur $\mathrm{Im}(X)$. Par définition, pour tout $z \in \mathbb{R}^p$,

$$Mz = X \times (({}^tXX)^{-1}\,{}^tXz) \in \mathrm{Im}(X) \ .$$

De plus, M est idempotente car

$$\begin{aligned} M^2 &= X({}^tXX)^{-1}\,{}^tX \times X({}^tXX)^{-1}\,{}^tX \\ &= X({}^tXX)^{-1}({}^tXX)({}^tXX)^{-1}\,{}^tX \\ &= X({}^tXX)^{-1}\,{}^tX = M \ . \end{aligned}$$

La matrice M étant symétrique, il s'agit bien de la projection orthogonale sur $\mathrm{Im}(X)$. Enfin, pour montrer que $R = \rho(y, \pi_y)$, il suffit de raisonner comme dans l'exemple pour obtenir que la corrélation est maximale lorsque $(a_1, \ldots, a_p)' \in \mathbb{R}^p$ est tel que $a_1x^1 + \cdots + a_px^p = \pi_y$. □

4.3 Inertie

La variance des observations d'une variable quantitative est une mesure de la dispersion de ces observations par rapport à leur moyenne. Pour étendre cette notion à des observations dans $\mathbb{R}^p$, une première idée consiste à faire la somme des variances des coordonnées.

Définition 4.4. *L'***inertie standard** *des* n *observations* $(x_1^1, \ldots, x_1^p), \ldots, (x_n^1, \ldots, x_n^p) \in \mathbb{R}^p$ *est définie par*

$$I = \sum_{j=1}^{p} \mathrm{Var}\left(x^j\right) \ .$$

L'inertie standard étend l'idée de dispersion autour de la moyenne et le rôle de cette moyenne est joué par le vecteur g des moyennes des variables, appelé *centre de gravité*,

$$g = \begin{pmatrix} \overline{x^1} \\ \vdots \\ \overline{x^p} \end{pmatrix} \in \mathbb{R}^p \ .$$

En effet, il est possible de réécrire l'inertie standard de la façon suivante,

$$\begin{aligned} I &= \sum_{j=1}^{p} \mathrm{Var}\left(x^j\right) = \sum_{j=1}^{p}\sum_{i=1}^{n} p_i(x_i^j - \overline{x^j})^2 \\ &= \sum_{i=1}^{n} p_i \sum_{j=1}^{p} (x_i^j - g_j)^2 \\ &= \sum_{i=1}^{n} p_i d_2^2(x_i, g) \end{aligned}$$

où d_2 est la distance euclidienne usuelle sur $\mathbb{R}^p$. Cette écriture nous amène à considérer la définition suivante de l'inertie basée sur une distance d quelconque sur $\mathbb{R}^p$.

Rappel Une *distance* sur $\mathbb{R}^p$ est une fonction $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$ qui vérifie les points suivants, pour $x, y, z \in \mathbb{R}^p$,

1. $d(x, y) = d(y, x)$ (Symétrie),
2. $d(x, x) = 0 \Leftrightarrow x = 0$ (Séparation),
3. $d(x, z) \leqslant d(x, y) + d(y, z)$ (Inégalité triangulaire).

Définition 4.5. *Etant donnée une distance d sur $\mathbb{R}^p$, nous définissons l'***inertie par rapport à** *d des n observations $(x_1^1, \ldots, x_1^p), \ldots, (x_n^1, \ldots, x_n^p) \in \mathbb{R}^p$ par*

$$I_d = \sum_{i=1}^{n} p_i d^2(x_i, g) \ .$$

La distance d est une distance sur l'espace des variables et permet de définir une notion d'inertie I_d mesurant la dispersion des observations en un sens particulier lié au choix de d, par exemple, en donnant une importance différente à chaque variable (voir les exemples de la Section 4.4).

Afin de décrire ce qui induit de la dispersion dans nos données, nous serons amenés à considérer le rôle que joue chaque vecteur d'observations x_i dans le calcul de l'inertie.

Définition 4.6. *Pour $i \in \{1, \ldots, n\}$, nous appelons* **contribution à l'inertie** *I_d de l'individu i la proportion*

$$\frac{p_i d^2(x_i, g)}{I_d} \in [0, 1] \ .$$

Dans le cas de l'inertie standard I, cette contribution vaut donc

$$\frac{p_i}{I} \sum_{j=1}^{p} (x_i^j - \overline{x^j})^2 \ .$$

4.4 Changement de distance

Nous avons vu que, en réécrivant l'inertie standard I, nous étions amenés à considérer la définition plus générale de l'inertie I_d par rapport à une distance d. Nous allons maintenant nous intéresser à une classe particulière de distances construite en suivant la remarque suivante, pour $i \in \{1, \ldots, n\}$,

$$d_2^2(x_i, g) = \sum_{j=1}^{p} (x_i^j - g_j)^2 = {}^t X_i X_i$$

où $X_i = (x_i^1 - \overline{x^1}, \ldots, x_i^p - \overline{x^p}) \in \mathbb{R}^p$ est le $i^{\text{ème}}$ vecteur ligne de la matrice X des données centrées. Etant donnée une matrice symétrique définie positive M (voir les rappels de la Section 4.5) de taille $p \times p$, nous considérons la distance d_M sur $\mathbb{R}^p$ définie par, pour tout $x, y \in \mathbb{R}^p$,

$$d_M^2(x, y) = {}^t(x - y) M (x - y) \ .$$

Exercice 4.2. *A l'aide des rappels de la Section 4.5, vérifier que d_M est bien une distance sur $\mathbb{R}^p$.*

Ainsi, pour tout $i \in \{1, \dots, n\}$, nous avons $d_M^2(x_i, g) = {}^tX_i M X_i$ et nous notons I_M l'inertie par rapport à d_M. Nous présentons quelques choix de M dans la suite de cette section. La notion d'inertie généralisant celle de variance, nous disposons, en particulier, d'un résultat analogue au Théorème 1.2.

Théorème 4.1. [Inertie par groupes] *Soient $N > 0$ et $G_1, \dots, G_N$ une partition de $\{1, \dots, n\}$. Pour tout $k \in \{1, \dots, N\}$, nous considérons*

$$q_k = \sum_{i \in G_k} p_i \qquad et \qquad g^{(k)} = \begin{pmatrix} \overline{x_k^1} \\ \vdots \\ \overline{x_k^p} \end{pmatrix} \in \mathbb{R}^p$$

où $\overline{x_k^j}$ est la moyenne des x_i^j pour $i \in G_k$,

$$\overline{x_k^j} = \frac{1}{q_k} \sum_{i \in G_k} p_i x_i^j, \ j \in \{1, \dots, p\} \ .$$

Etant donnée une matrice symétrique définie positive M, l'inertie I_M se décompose en

$$I_M = I_M^{inter} + I_M^{intra}$$

avec

$$I_M^{inter} = \sum_{k=1}^{N} q_k d_M^2(g^{(k)}, g) \qquad \textit{(Inertie inter-groupe)}$$

et

$$I_M^{intra} = \sum_{k=1}^{N} q_k I_M^{(k)} \qquad \textit{(Inertie intra-groupe)}$$

où $I_M^{(k)}$ est l'inertie par rapport à d_M des observations $x_i = (x_i^1, \dots, x_i^p)'$ pour $i \in G_k$,

$$I_M^{(k)} = \frac{1}{q_k} \sum_{i \in G_k} p_i d_M^2(x_i, g^{(k)})$$

Démonstration. Voir la feuille du TD 4. □

Exercice 4.3. *Vérifier que dans le cas de l'inertie standard I, nous retrouvons bien*

$$I = \sum_{j=1}^{p} \mathrm{Var}_{inter}(x^j) + \sum_{j=1}^{p} \mathrm{Var}_{intra}(x^j) \ .$$

4.4.1 Distance euclidienne

Si M est la matrice identité Id_p de taille p, alors la distance d_M est la distance euclidienne classique sur $\mathbb{R}^p$,

$$d^2_{\mathrm{Id}_p}(x_i, g) = {}^tX_i \mathrm{Id}_p X_i = {}^tX_i X_i = d_2^2(x_i, g),\ i \in \{1, \dots, n\}.$$

Ce cas correspond à ce que nous avons vu dans la section précédente et à l'inertie standard,

$$I_{\mathrm{Id}_p} = I = \sum_{i=1}^{n} p_i d_2^2(x_i, g) = \sum_{j=1}^{p} \mathrm{Var}\left(x^j\right).$$

Par analogie avec la contribution d'un individu à l'inertie, il est possible ici de définir l'influence de chaque variable sur l'inertie.

Définition 4.7. *Pour $j \in \{1, \dots, p\}$, nous appelons* **contribution à l'inertie standard** *I de la variable j la proportion*

$$\frac{\mathrm{Var}\left(x^j\right)}{I} \in [0, 1].$$

4.4.2 Distance des variables réduites

Afin de mettre toutes les variables à la même échelle, nous pouvons considérer leurs versions réduites, *i.e.* considérer les observations centrées réduites $(x_i^j - \overline{x^j})/\sqrt{\mathrm{Var}\,(x^j)}$. En prenant M comme la matrice $p \times p$ diagonale des inverses des variances,

$$M = \begin{bmatrix} 1/\mathrm{Var}\left(x^1\right) & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1/\mathrm{Var}\left(x^p\right) \end{bmatrix},$$

cela revient à travailler avec avec la distance d_M qui est telle que, pour tout $i \in \{1, \dots, n\}$,

$$d^2_M(x_i, g) = {}^tX_i M X_i = \sum_{j=1}^{p} \frac{(x_i^j - \overline{x^j})^2}{\mathrm{Var}\,(x^j)}\ .$$

L'inertie par rapport à d_M est donc égale à

$$\begin{aligned} I_M &= \sum_{i=1}^{n} p_i\, {}^tX_i M X_i = \sum_{i=1}^{n} p_i \sum_{j=1}^{p} \frac{(x_i^j - \overline{x^j})^2}{\mathrm{Var}\,(x^j)} \\ &= \sum_{j=1}^{p} \frac{1}{\mathrm{Var}\,(x^j)} \sum_{i=1}^{n} p_i (x_i^j - \overline{x^j})^2 = \sum_{j=1}^{p} \frac{\mathrm{Var}\left(x^j\right)}{\mathrm{Var}\,(x^j)} \\ &= p\ . \end{aligned}$$

L'inertie I_M est constante et chaque variable y contribue avec une proportion $1/p$.

Exercice 4.4. *Faire un parallèle entre l'utilisation de la distance des variables réduites et la matrice de corrélation $\mathfrak{C}$.*

4.4.3 Distance de Mahalanobis

Plus généralement, pour normaliser les observations des variables, le statisticien Mahalanobis proposa, en 1936, de prendre $M = \Sigma^{-1}$, l'inverse de la matrice de covariance. Bien entendu, ce choix n'est possible que lorsque Σ est une matrice inversible. Cela donne, pour $i \in \{1, \dots, n\}$,

$$d^2_{\Sigma^{-1}}(x_i, g) = {}^tX_i \Sigma^{-1} X_i \ .$$

De plus, nous avons le même résultat que pour la distance des variables réduites, à savoir que l'inertie $I_{\Sigma^{-1}}$ est constante et égale à p,

$$\begin{aligned}
I_{\Sigma^{-1}} &= \sum_{i=1}^{n} p_i \, {}^tX_i \Sigma^{-1} X_i = \sum_{i=1}^{n} p_i \sum_{j=1}^{p} \sum_{k=1}^{p} \left(\Sigma^{-1}\right)_{jk} X_{ij} X_{ik} \\
&= \sum_{i=1}^{n} p_i \sum_{j=1}^{p} \sum_{k=1}^{p} \left(\Sigma^{-1}\right)_{jk} (x_i^j - \overline{x^j})(x_i^k - \overline{x^k}) \\
&= \sum_{j=1}^{p} \sum_{k=1}^{p} \left(\Sigma^{-1}\right)_{jk} \mathrm{Cov}(x^j, x^k) = \sum_{j=1}^{p} \sum_{k=1}^{p} \left(\Sigma^{-1}\right)_{jk} \Sigma_{kj} \\
&= \sum_{j=1}^{p} \left(\Sigma^{-1}\Sigma\right)_{jj} = \sum_{j=1}^{p} \mathrm{Id}_{p_{jj}} = p \ .
\end{aligned}$$

Par contre, la contribution des variables à l'inertie n'est plus nécessairement égale à $1/p$ et ne s'exprime pas simplement.

L'intérêt du choix de la distance de Mahalanobis provient de considérations théoriques. En effet, l'utilisation de Σ^{-1} permet de décrire les variables x^j comme étant des sommes de variables réduites et non corrélées. En particulier, les quantités $d^2_M(x_i, g)$ s'expriment alors en terme de sommes des contributions isolées de ces nouvelles variables.

Exemple Prenons $p = 2$, $\alpha, \beta, \gamma \in \mathbb{R}$ et supposons qu'il existe deux variables réduites et non corrélées u et v (*i.e.* Cov(u,v)=0) telles que $x^1 = \alpha u$ et $x^2 = \beta u + \gamma v$. Il est facile de montrer que $\mathrm{Var}\left(x^1\right) = \alpha^2$, $\mathrm{Cov}(x^1, x^2) = \alpha\beta$ et $\mathrm{Var}\left(x^2\right) = \beta^2 + \gamma^2$. En supposant que $\alpha \neq 0$ et $\gamma \neq 0$, nous avons donc la matrice de covariance

$$\Sigma = \begin{bmatrix} \alpha^2 & \alpha\beta \\ \alpha\beta & \beta^2 + \gamma^2 \end{bmatrix}$$

qui est inversible (son déterminant vaut $\alpha^2\gamma^2 > 0$) et dont l'inverse est

$$\Sigma^{-1} = \frac{1}{\alpha^2\gamma^2} \begin{bmatrix} \beta^2 + \gamma^2 & -\alpha\beta \\ -\alpha\beta & \alpha^2 \end{bmatrix} \ .$$

Pour $i \in \{1, \dots, n\}$, nous avons donc

$$\begin{aligned}
d^2_{\Sigma^{-1}}(x_i, g) &= {}^tX_i \Sigma^{-1} X_i \\
&= \Sigma^{-1}_{11} X^2_{i1} + \left(\Sigma^{-1}_{12} + \Sigma^{-1}_{21}\right) X_{i1} X_{i2} + \Sigma^{-1}_{22} X^2_{i2} \\
&= \frac{\beta^2 + \gamma^2}{\alpha^2\gamma^2}(x^1_i - \overline{x^1})^2 - \frac{2\alpha\beta}{\alpha^2\gamma^2}(x^1_i - \overline{x^1})(x^2_i - \overline{x^2}) + \frac{\alpha^2}{\alpha^2\gamma^2}(x^2_i - \overline{x^2})^2 \\
&= \frac{1}{\alpha^2}(x^1_i - \overline{x^1})^2 \\
&\quad + \frac{\beta^2}{\alpha^2\gamma^2}(x^1_i - \overline{x^1})^2 - \frac{2\beta}{\alpha\gamma^2}(x^1_i - \overline{x^1})(x^2_i - \overline{x^2}) + \frac{1}{\gamma^2}(x^2_i - \overline{x^2})^2 \\
&= \left(\frac{x^1_i - \overline{x^1}}{\alpha}\right)^2 + \left(\frac{\beta(x^1_i - \overline{x^1})}{\alpha\gamma} - \frac{x^2_i - \overline{x^2}}{\gamma}\right)^2 \\
&= (u_i - \overline{u})^2 + \left(\frac{\beta(u_i - \overline{u})}{\gamma} - \frac{\beta(u_i - \overline{u}) + \gamma(v_i - \overline{v})}{\gamma}\right)^2 = (u_i - \overline{u})^2 + (v_i - \overline{v})^2 \ .
\end{aligned}$$

Nous obtenons donc bien que les quantités $d^2_{\Sigma^{-1}}(x_i, g)$ s'écrivent comme la somme des contributions des observations u_i et v_i séparément.

4.5 Matrices symétriques définies positives et diagonalisation

Nous avons vu dans ce chapitre que les matrices symétriques définies positives jouent un rôle important lorsque nous souhaitons étudier un ensemble de p variables couplées. L'objet de cette section est de faire certains rappels d'algèbre linéaire sur les propriétés de ces matrices que nous utiliserons dans la suite du cours.

4.5.1 Matrices symétriques

Définition 4.8. *Considérons une matrice carrée M de taille $p \times p$. Nous dirons que M est* **symétrique** *si*

$$\forall i, j \in \{1, \dots, p\},\ M_{ij} = M_{ji} \ .$$

Ces matrices ont de bonnes propriétés comme nous le verrons par la suite. De plus, c'est pour M symétrique que nous avons défini la distance d_M telle que, pour tout $i \in \{1, \dots, n\}$,

$$d^2_M(x_i, g) = {}^tX_i M X_i \ .$$

Puisque nous n'utilisons que cette propriété de la distance d_M, il est facile de voir que nous pouvons considérer n'importe quelle matrice M' de taille $p \times p$ et qu'il est toujours possible de se ramener à une matrice symétrique M telle que, pour tout $i \in \{1, \dots, n\}$,

$$d^2_{M'}(x_i, g) = {}^tX_i M' X_i = {}^tX_i M X_i = d^2_M(x_i, g) \ .$$

En effet, si M' est une matrice $p \times p$ quelconque, nous avons, pour tout $v = (v_1, \ldots, v_p)' \in \mathbb{R}^p$,

$$\begin{aligned} {}^tvM'v &= \sum_{j=1}^{p}\sum_{k=1}^{p} M'_{jk} v_j v_k \\ &= \sum_{j=1}^{p}\left\{ M'_{jj} v_j^2 + \sum_{j<k} \left(M'_{jk} + M'_{kj}\right) v_j v_k \right\} . \end{aligned}$$

Ainsi, pour symétriser, il nous suffit de considérer la matrice M symétrique définie par, pour tout $j, k \in \{1, \ldots, p\}$,

$$M_{jk} = \frac{M'_{jk} + M'_{kj}}{2}$$

et nous obtenons

$$\begin{aligned} d^2_M(x_i, g) &= {}^tX_i M X_i \\ &= \sum_{j=1}^{p}\left\{ M_{jj} X_{ij}^2 + \sum_{j<k} \left(M_{jk} + M_{kj}\right) X_{ij} X_{ik} \right\} \\ &= \sum_{j=1}^{p}\left\{ \frac{M'_{jj} + M'_{jj}}{2} X_{ij}^2 + \sum_{j<k} \left(\frac{M'_{jk} + M'_{kj}}{2} + \frac{M'_{kj} + M'_{jk}}{2} \right) X_{ij} X_{ik} \right\} \\ &= \sum_{j=1}^{p}\left\{ M'_{jj} X_{ij}^2 + \sum_{j<k} \left(M'_{jk} + M'_{kj}\right) X_{ij} X_{ik} \right\} \\ &= {}^tX_i M' X_i = d^2_{M'}(x_i, g) . \end{aligned}$$

Exemple Prenons la matrice carrée suivante

$$M' = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix} .$$

Pour tout $v = (v_1, v_2)' \in \mathbb{R}^2$, nous avons

$$\begin{aligned} \begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} &= 2v_1^2 + 3v_1 v_2 - v_1 v_2 + 2v_2^2 \\ &= 2v_1^2 + 2v_1 v_2 + 2v_2^2 \\ &= \begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \end{aligned}$$

avec $1 = \dfrac{3 + (-1)}{2}$. Il est donc équivalent de considérer M' et sa version symétrisée

$$M = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} .$$

4.5.2 Matrices définies positives

Définition 4.9. *Considérons une matrice carrée M de taille $p \times p$. Nous dirons que M est* **définie positive** *si*

$$\forall v \in \mathbb{R}^p, \ {}^t v M v \geqslant 0 \ .$$

Cette propriétés implique en particulier que nous avons bien ${}^t X_i M X_i \geqslant 0$, pour tout $i \in \{1, \ldots, n\}$.

Exemple La matrice

$$M = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \ .$$

est symétrique définie positive. En effet, pour tout $(x, y)' \in \mathbb{R}^2$, nous avons

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 2x^2 - 2xy + 2y^2 = x^2 + y^2 + (x-y)^2 \geqslant 0 \ .$$

La matrice

$$M = \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} \ .$$

est symétrique mais pas définie positive. En effet, prenons $(1, 1)' \in \mathbb{R}^2$, nous avons

$$\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = -2 < 0 \ .$$

4.5.3 Diagonalisation des matrices symétriques

Une des propriétés particulièrement intéressantes des matrices symétriques est le résultat suivant.

Théorème 4.2. *Toute matrice symétrique se diagonalise dans une base orthonormée.*

Démonstration. Voir le cours d'algèbre linéaire. □

Rappel Soit E un espace vectoriel euclidien de dimension d muni d'un produit scalaire $\langle \cdot, \cdot \rangle$. Une base $\{b_1, \ldots, b_d\}$ de E est dite *orthonormée* si, pour tout $i, j \in \{1, \ldots, d\}$ avec $i \neq j$, $\langle b_i, b_j \rangle = 0$ et $\langle b_i, b_i \rangle = 1$.

Afin d'obtenir les valeurs propres et les vecteurs propres d'une matrice symétrique M, il faut procéder par étapes. Nous donnons ici un exemple à partir de la matrice symétrique définie positive suivante

$$M = \begin{bmatrix} 7 & -4 \\ -4 & 13 \end{bmatrix} \ .$$

Exercice 4.5. *Vérifier que M est bien définie positive.*

Rappel Soit M une matrice de taille $p \times p$, nous disons que $\lambda \in \mathbb{R}$ est une *valeur propre* de M associée au *vecteur propre* $v \in \mathbb{R}^p$ si et seulement si $Mv = \lambda v$.

Recherche des valeurs propres

Les valeurs propres $\lambda_1, \ldots, \lambda_p$ de M sont les racines du polynôme caractéristique de M, *i.e.* il faut résoudre en $\lambda \in \mathbb{R}$

$$\det(M - \lambda \mathrm{Id}_p) = 0 \ .$$

Calculons ce déterminant dans notre cas,

$$\begin{aligned} \det(M - \lambda \mathrm{Id}_2) &= \det \begin{pmatrix} 7-\lambda & -4 \\ -4 & 13-\lambda \end{pmatrix} \\ &= (7-\lambda)(13-\lambda) - 16 = \lambda^2 - 20\lambda + 75 \ . \end{aligned}$$

Nous obtenons donc que les deux valeurs propres de M sont

$$\lambda_1 = \frac{20+10}{2} = 15 \qquad \text{et} \qquad \lambda_2 = \frac{20-10}{2} = 5 \ .$$

Remarquons que les deux valeurs propres de M sont positives. Cette remarque est plus généralement vraie pour toutes les valeurs propres d'une matrice définie positive.

Proposition 4.3. *Soit M une matrice symétrique de taille $p \times p$, si M est définie positive alors toutes ses valeurs propres sont positives.*

Démonstration. Soit $\lambda \in \mathbb{R}$ une valeur propre de M et $v = (v_1, \ldots, v_p)' \in \mathbb{R}^p \setminus \{0\}$ un vecteur propre non-nul associé. Par définition, nous savons que ${}^t v M v \geqslant 0$. De plus, nous avons

$${}^t v M v = \lambda \, {}^t v v = \lambda \sum_{i=1}^{p} v_i^2 \geqslant 0 \ .$$

Nous en déduisons que $\lambda \geqslant 0$. □

Recherche des vecteurs propres

Par définition, pour $i \in \{1, \ldots, p\}$, un vecteur propre v_i associé à la valeur propre λ_i est tel que $M v_i = \lambda_i v_i$. Pour le trouver, il faut donc résoudre en $x \in \mathbb{R}^p$

$$(M - \lambda_i \mathrm{Id}_p) x = 0 \ .$$

Pour notre exemple, il nous faut trouver $x = (x_1, x_2)' \in \mathbb{R}^2$ tel que

- pour $\lambda_1 = 15$, nous avons

$$\begin{bmatrix} -8 & -4 \\ -4 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \Longleftrightarrow \left\{ \begin{array}{rcl} -8x_1 - 4x_2 &=& 0 \\ -4x_1 - 2x_2 &=& 0 \end{array} \right. \Longleftrightarrow -2x_1 = x_2 \ .$$

  En prenant $x_1 = 1$, nous obtenons un vecteur propre $v_1 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$.

- pour $\lambda_2 = 5$, nous avons

$$\begin{bmatrix} 2 & -4 \\ -4 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \Longleftrightarrow \left\{ \begin{array}{rcl} 2x_1 - 4x_2 &=& 0 \\ -4x_1 + 8x_2 &=& 0 \end{array} \right. \Longleftrightarrow x_1 = 2x_2 \ .$$

  En prenant $x_2 = 1$, nous obtenons un vecteur propre $v_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

Diagonalisation et interprétations

Soit $i \in \{1, \dots, p\}$, nous notons désormais $u_i \in \mathbb{R}^p$ le vecteur propre *normalisé* associé à la valeur propre $\lambda_i > 0$, *i.e.* u_i est tel que $u_{i1}^2 + \cdots + u_{ip}^2 = 1$. Ce vecteur propre s'obtient à partir d'un vecteur propre v_i non-nul associé à λ_i en considérant

$$u_i = \frac{v_i}{\|v_i\|}$$

avec $\|v_i\|^2 = v_{i1}^2 + \cdots + v_{ip}^2$.

Les vecteurs $u_1, \dots, u_p$ forment une base de $\mathbb{R}^p$. De plus, la matrice M étant symétrique, nous pouvons considérer que cette base est orthonormée. La matrice de changement de base allant de la base canonique à celle donnée par les u_i est la matrice P de taille $p \times p$ dont les colonnes sont les u_i,

$$P = \begin{bmatrix} u_{11} & u_{21} & \dots & u_{p1} \\ \vdots & \vdots & & \vdots \\ u_{1p} & u_{2p} & \dots & u_{pp} \end{bmatrix} .$$

En tant que matrice de changement de base, P est inversible. En outre, la base canonique et la base formée par les u_i étant orthonormées, la matrice P est orthogonale (*i.e.* $P^{-1} = {}^tP$). Cette dernière remarque facilite le calcul de l'inverse P^{-1} puisqu'il suffit simplement de considérer la matrice transposée ${}^tP$.

Considérons la matrice diagonale D obtenue à partir des valeurs propres $\lambda_1, \dots, \lambda_p$,

$$D = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_p \end{bmatrix} .$$

Nous avons donc obtenu la diagonalisation de la matrice M,

$$M = PDP^{-1} = PD\,{}^tP .$$

Reprenons les vecteurs v_1 et v_2 de notre exemple. Ils mènent aux vecteurs propres normalisés

$$u_1 = \begin{bmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{bmatrix} \qquad \text{et} \qquad u_2 = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix} .$$

Notons que ces vecteurs forment bien une base orthonormée de $\mathbb{R}^2$ car $\|u_1\|^2 = \|u_2\|^2 = 1$ et ${}^tu_1u_2 = 0$. La matrice de changement de base P vaut donc

$$P = \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}$$

et il est facile de vérifier qu'elle est inversible ($\det(C) = 1$) et orthogonale,

$$P^{-1} = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} = {}^tP .$$

La diagonalisation de M s'écrit donc

$$M = P \begin{bmatrix} 15 & 0 \\ 0 & 5 \end{bmatrix} {}^tP .$$

En dimension $p = 2$, il est possible d'illustrer cette diagonalisation par un dessin comme celui de la Figure 4.3. La courbe C_M représente l'ellipse associée à la matrice symétrique positive M et les droites engendrées par les vecteurs propres u_1 et u_2 sont les deux axes de symétrie de C_M. De plus, remarquons que le demi-grand axe le long de u_1 vaut $\lambda_1 = 15$ et celui le long de u_2 vaut $\lambda_2 = 5$.

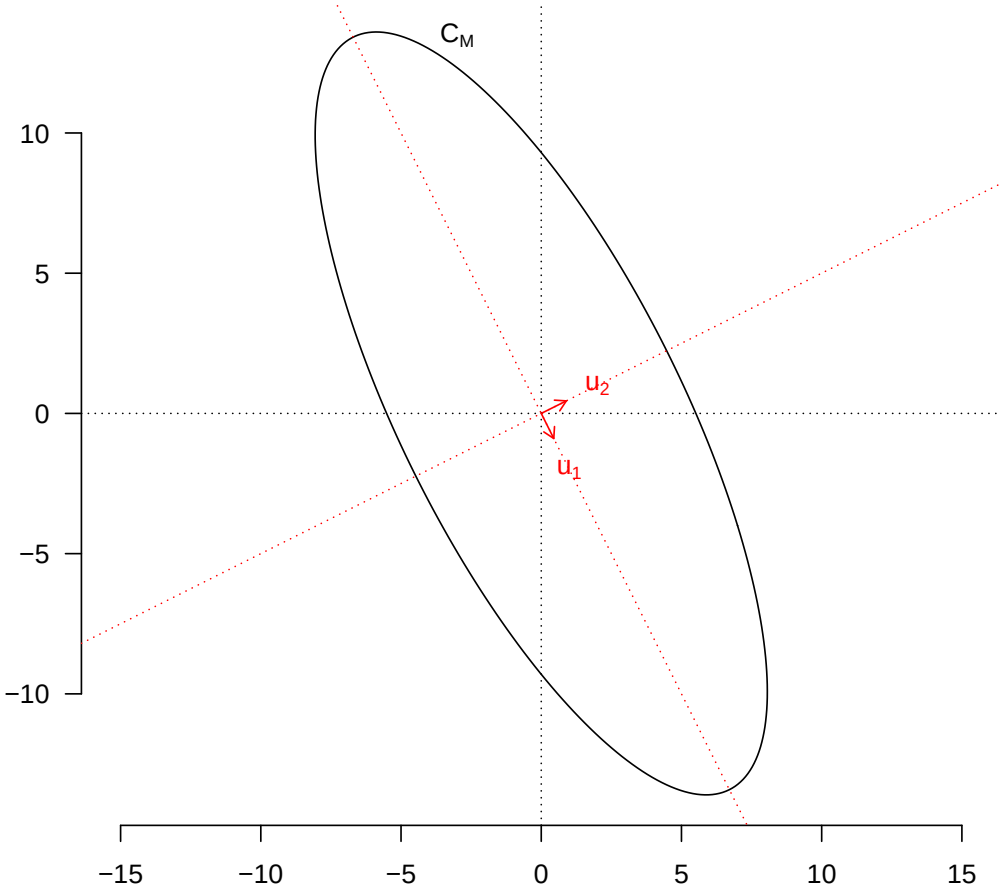


FIGURE 4.3 – Ellipse associée à la matrice symétrique définie positive M et la base orthonormée de diagonalisation $\{u_1, u_2\}$.

Chapitre 5

Analyse en composantes principales

5.1 Introduction

Comme dans le chapitre précédent, nous considérons ici p variables quantitatives couplées $x^1, \dots, x^p$ pour lesquelles nous disposons de n observations $x_1 = (x_1^1, \dots, x_1^p)', \dots, x_n = (x_n^1, \dots, x_n^p)' \in \mathbb{R}^p$ pondérées par les poids $p_1, \dots, p_n > 0$ normalisés. Afin de simplifier les notations de ce chapitre, nous supposerons que ces observations sont centrées, *i.e.* $\overline{x^j} = 0$ pour tout $j \in \{1, \dots, p\}$. Nous avons donc à notre disposition la matrice de taille $n \times p$ des données centrées,

$$X = \begin{bmatrix} x_1^1 - \overline{x^1} & \dots & x_1^p - \overline{x^p} \\ \vdots & \ddots & \vdots \\ x_n^1 - \overline{x^1} & \dots & x_n^p - \overline{x^p} \end{bmatrix} = \begin{bmatrix} x_1^1 & \dots & x_1^p \\ \vdots & \ddots & \vdots \\ x_n^1 & \dots & x_n^p \end{bmatrix},$$

et la matrice diagonale de taille $n \times n$ des poids,

$$W = \begin{bmatrix} p_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & p_n \end{bmatrix}.$$

La matrice X contient donc les observations, chaque colonne correspondant aux n observations d'une variable et chaque ligne correspondant aux p observations pour un individu donné.

Nous nous intéressons à la question suivante : comment donner une "bonne" représentation graphique de ces données $x_1, \dots, x_n \in \mathbb{R}^p$? Si $p = 2$, il est possible de tracer le nuage de points associé aux observations dans le plan. Si $p > 2$, cette représentation n'est plus faisable. Pour représenter les observations, nous allons chercher à construire un plan sur lequel projeter les observations tout en essayant de conserver au maximum la structure des données.

5.2 Composantes principales

Le point de départ de notre étude est le simple résultat suivant.

Proposition 5.1. *La matrice de covariance $\Sigma = {}^tXWX$ est symétrique et définie positive.*

Démonstration. La symétrie de Σ est évidente. Pour la définie positivité, nous considérons $v \in \mathbb{R}^p$ et nous avons

$$ {}^t v \Sigma v = {}^t v \, {}^t X W X v = {}^t (Xv) W (Xv) = \sum_{i=1}^{n} p_i (Xv)_i^2 \geqslant 0 \ . $$

□

Exercice 5.1. *Déduire directement de cette proposition que la matrice de corrélation $\mathfrak{C}$ est symétrique et définie positive.*

En particulier, ce résultat et le Théorème 4.2 impliquent que Σ est diagonalisable en base orthonormée. De plus, grâce à la Proposition 4.3, nous pouvons considérer $\lambda_1 \geqslant \cdots \geqslant \lambda_p \geqslant 0$ et des vecteurs $u^1, \ldots, u^p \in \mathbb{R}^p$ orthonormés tels que $\Sigma = PDP^{-1} = PD\,{}^t P$ avec

$$ D = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_p \end{bmatrix} \qquad \text{et} \qquad P = \begin{bmatrix} u_1^1 & \ldots & u_1^p \\ \vdots & & \vdots \\ u_p^1 & \ldots & u_p^p \end{bmatrix} \ . $$

Les colonnes de P sont les vecteurs propres $u^j = (u_1^j, \ldots, u_p^j)' \in \mathbb{R}^p$, pour $j \in \{1, \ldots, p\}$, et elles permettent de définir p nouvelles variables $c^1, \ldots, c^p$ comme des combinaisons linéaires des variables $x^1, \ldots, x^p$,

$$ c^j = \sum_{k=1}^{p} u_k^j x^k = u_1^j x^1 + \cdots + u_p^j x^p, \ j \in \{1, \ldots, p\} \ . $$

Les c^j sont des variables "virtuelles" et sont appelées les *variables principales*. Pour $i \in \{1, \ldots, n\}$ et $j \in \{1, \ldots, p\}$, la $i^{\text{ème}}$ observation de la variables c^j est donc donnée par

$$ c_i^j = \sum_{k=1}^{p} u_k^j x_i^k = \sum_{k=1}^{p} P_{kj} X_{ik} = (XP)_{ij} \ . $$

Comme nous l'avons fait pour les observations initiales, nous pouvons considérer le vecteur C^j des n observations de la variable c^j, $j \in \{1, \ldots, p\}$,

$$ C^j = \begin{pmatrix} c_1^j \\ \vdots \\ c_n^j \end{pmatrix} \in \mathbb{R}^n \ . $$

Etant donné que les variables $x^1, \ldots, x^p$ sont supposées centrées, il est important de remarquer qu'il en va de même pour les variables $c^1, \ldots, c^p$ puisque ces dernières en sont des combinaisons linéaires.

Définition 5.1. *Les vecteurs $C^1, \ldots, C^p$ sont appelés les* **composantes principales**. *La matrice C de taille $n \times p$ dont les colonnes sont les composantes principales est la* **matrice des composantes principales**,

$$ C = XP = \begin{bmatrix} c_1^1 & \ldots & c_1^p \\ \vdots & & \vdots \\ c_n^1 & \ldots & c_n^p \end{bmatrix} \ . $$

La matrice C doit être considérée de la même manière que la matrice des données centrées X. En effet, chaque ligne correspond aux p observations des variables $c^1, \dots, c^p$ pour un individu donné et chaque colonne correspond aux n observations d'une des variables principales. Le principal avantage à considérer C plutôt que X réside dans la structure de covariance.

Proposition 5.2. *La matrice de covariance des variables principales est la matrice diagonale D. Autrement dit, pour tout $j, j' \in \{1, \dots, p\}$ avec $j \neq j'$, nous avons $\mathrm{Var}\left(c^j\right) = \lambda_j$ et $\mathrm{Cov}(c^j, c^{j'}) = 0$.*

Démonstration. Considérons $j, j' \in \{1, \dots, p\}$ et, grâce à la bilinéarité de la covariance, calculons

$$\begin{aligned}
\mathrm{Cov}(c^j, c^{j'}) &= \mathrm{Cov}\left(\sum_{k=1}^{p} u_k^j x^k, \sum_{k'=1}^{p} u_{k'}^{j'} x^{k'}\right) \\
&= \sum_{k=1}^{p}\sum_{k'=1}^{p} u_k^j u_{k'}^{j'} \mathrm{Cov}\left(x^k, x^{k'}\right) \\
&= \sum_{k=1}^{p}\sum_{k'=1}^{p} P_{kj} P_{k'j'} \Sigma_{kk'} = \sum_{k=1}^{p} P_{kj} (\Sigma P)_{kj'} \\
&= ({}^t P \Sigma P)_{jj'} = D_{jj'} \ .
\end{aligned}$$

□

Exercice 5.2. *Si certaines des valeurs propres de Σ sont nulles, que cela signifie-t-il pour les variables initiales $x^1, \dots, x^p$?*

5.3 Représentation graphique

5.3.1 Plan principal

Pour $i \in \{1, \dots, n\}$, la $i^{\text{ème}}$ ligne de la matrice C donnent les coordonnées $C_i = (c_i^1, \dots, c_i^p)' \in \mathbb{R}^p$ du $i^{\text{ème}}$ individu dans le repère des composantes principales. Les observations des variables principales sont dans $\mathbb{R}^p$ et, pour $p > 2$, nous ne pouvons toujours pas les représenter simplement. Cependant, par construction, nous avons classé les variables principales par variance décroissante,

$$\lambda_1 = \mathrm{Var}\left(c^1\right) \geqslant \lambda_2 = \mathrm{Var}\left(c^2\right) \geqslant \cdots \geqslant \lambda_p = \mathrm{Var}\left(c^p\right) \geqslant 0 \ .$$

Autrement dit, les deux premières composantes principales correspondent aux deux directions dans lesquelles la "dispersion" des données est la plus importante. C'est dans ce plan engendré par C^1 et C^2, appelé *plan principal*, que nous représenterons nos données.

5.3.2 Représentation des individus

Pour $i \in \{1, \dots, n\}$, les coordonnées du $i^{\text{ème}}$ individu dans le plan principal sont donc données par les deux premiers éléments (c_i^1, c_i^2) de la $i^{\text{ème}}$ ligne de la matrice C.

La question naturelle de la qualité de cette représentation se pose. Comme l'illustre la figure 5.1, nous allons raisonner de la même façon que pour le coefficient de corrélation multiple R. C'est-à-dire que nous allons considérer l'angle θ_i formé entre le vecteur $c_i \in \mathbb{R}^p$ et la

représentation du $i^{\text{ème}}$ individu dans le plan principal dont les coordonnées sont (c_i^1, c_i^2). Plus cet angle sera petit, meilleure sera la représentation du $i^{\text{ème}}$ individu dans le plan principal, *i.e.* plus c_i sera proche du plan principal. Pour $i \in \{1, \dots, n\}$, nous mesurons donc la qualité de la représentation du $i^{\text{ème}}$ individu par la quantité

$$\cos^2 \theta_i = \frac{(c_i^1)^2 + (c_i^2)^2}{(c_i^1)^2 + \cdots + (c_i^p)^2} \ . \tag{5.1}$$

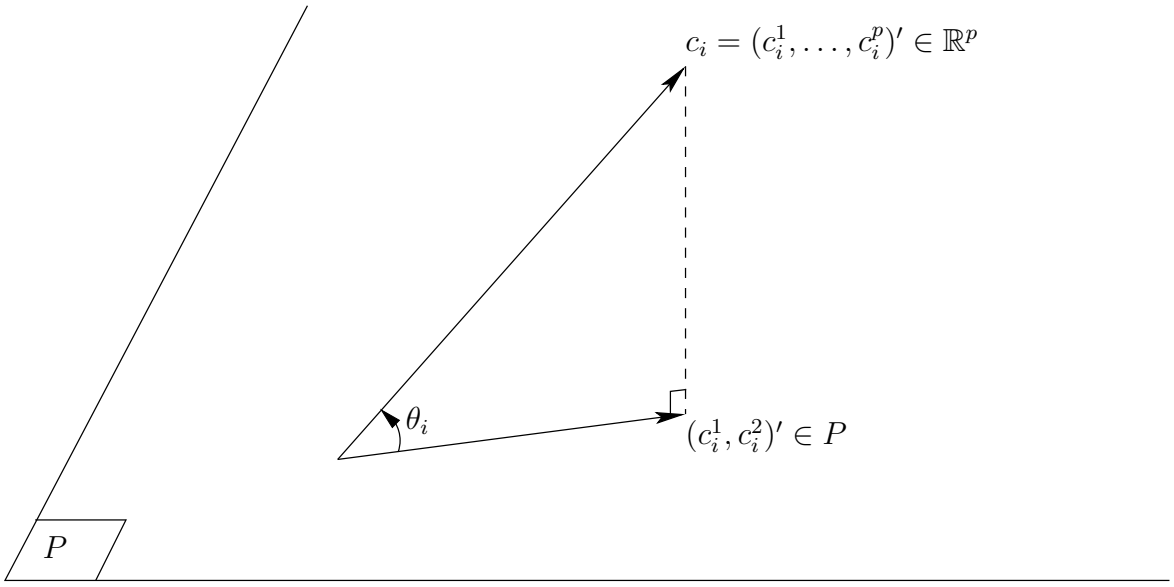


FIGURE 5.1 – L'angle θ_i formé entre le vecteur $c_i \in \mathbb{R}^p$ et la représentation du $i^{\text{ème}}$ individu dans le plan principal P.

Plus le cosinus carré de (5.1) sera proche de 1, plus la représentation du $i^{\text{ème}}$ individu sera bonne. Pour faire apparaître cette qualité sur le graphique, la taille du point représentant le $i^{\text{ème}}$ individu peut être proportionnelle à $\cos^2 \theta_i$ (voir Figure 5.2).

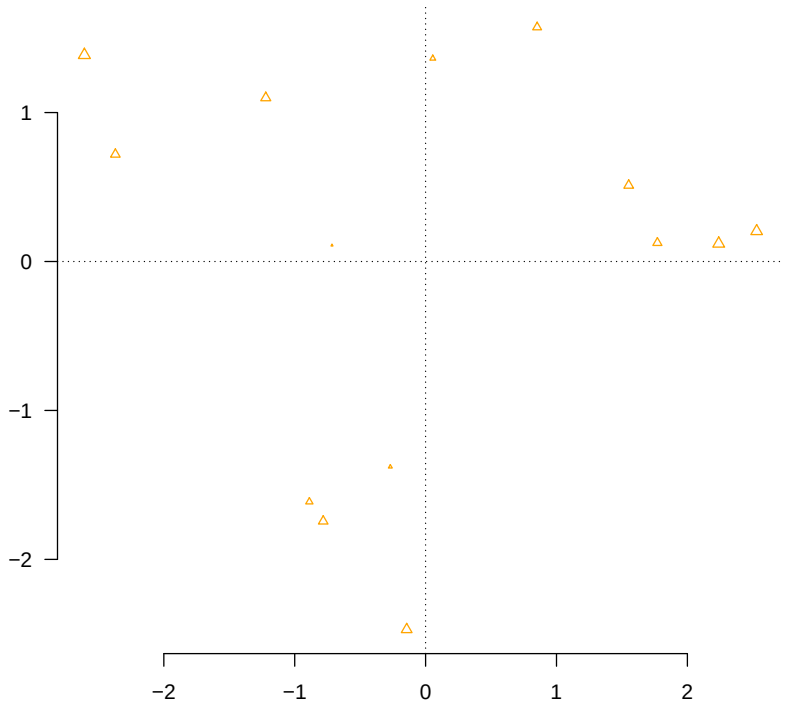


FIGURE 5.2 – Nuage de $n = 14$ points représenté dans le plan principal avec des tailles proportionnelles à la qualité de la représentation.

5.3.3 Interprétation des axes

Afin de comprendre les axes du plan principal, il nous faut savoir quel rôle joue chaque variable x^j dans la construction des variables principales. Le résultat suivant donne une expression des corrélations linéaires entre variables initiales et variables principales.

Proposition 5.3. *Prenons $j, j' \in \{1, \dots, p\}$, nous avons*

$$\rho\left(x^j, c^{j'}\right) = \frac{\sqrt{\lambda_{j'}}}{\sqrt{\mathrm{Var}\left(x^j\right)}} \times u_j^{j'} \ .$$

Démonstration. Avant de calculer cette corrélation, nous faisons la remarque suivante : étant donné que la matrice P est orthogonale, nous savons que

$$C = XP \iff X = CP^{-1} = C\,{}^tP \ .$$

Ainsi, pour tout $i \in \{1, \dots, n\}$ et $j \in \{1, \dots, p\}$, la $i^{\text{ème}}$ observation centrée de la variable x^j vaut

$$x_i^j = (C\,{}^tP)_{ij} = \sum_{k=1}^{p} C_{ik}\,{}^tP_{kj} = \sum_{k=1}^{p} c_i^k u_j^k \ . \tag{5.2}$$

Autrement dit, nous avons la variable $x^j = \sum_{k=1}^{p} c^k u_j^k$ et la covariance, pour $j' \in \{1, \dots, p\}$,

$$\mathrm{Cov}\left(x^j, c^{j'}\right) = \mathrm{Cov}\left(\sum_{k=1}^{p} c^k u_j^k, c^{j'}\right) = \sum_{k=1}^{p} \mathrm{Cov}\left(c^k, c^{j'}\right) u_j^k = \lambda_{j'} u_j^{j'}$$

où la dernière égalité découle de la Proposition 5.2. Nous avons donc la corrélation linéaire

$$\rho\left(x^j, c^{j'}\right) = \frac{\mathrm{Cov}\left(x^j, c^{j'}\right)}{\sqrt{\mathrm{Var}\left(x^j\right)} \times \sqrt{\mathrm{Var}\left(c^{j'}\right)}} = \frac{\lambda_{j'} u_j^{j'}}{\sqrt{\mathrm{Var}\left(x^j\right)} \times \sqrt{\lambda_{j'}}} = \frac{\sqrt{\lambda_{j'}}}{\sqrt{\mathrm{Var}\left(x^j\right)}} \times u_j^{j'} \ .$$

□

Etant donné que nous nous limitons ici au plan principal, chaque variable x^j est à mettre en relation avec c^1 et c^2. Pour chaque $j \in \{1, \dots, p\}$, nous considérons donc le point $\mathcal{P}_j$ donné par ses coordonnées

$$\left(\rho(x^j, c^1), \rho(x^j, c^2)\right) = \left(\frac{\sqrt{\lambda_1} u_j^1}{\sqrt{\mathrm{Var}\left(x^j\right)}}, \frac{\sqrt{\lambda_2} u_j^2}{\sqrt{\mathrm{Var}\left(x^j\right)}}\right) \ .$$

Ces points sont tous dans le disque unité $\mathcal{D} = \{(x, y) \in \mathbb{R}^2 \text{ tel que } x^2 + y^2 \leqslant 1\}$. En effet, grâce à la Proposition 5.3, nous avons

$$\rho(x^j, c^1)^2 + \rho(x^j, c^2)^2 \leqslant \sum_{k=1}^{p} \rho(x^j, c^k)^2 = \sum_{k=1}^{p} \frac{\lambda_k (u_j^k)^2}{\mathrm{Var}\left(x^j\right)} = \frac{1}{\mathrm{Var}\left(x^j\right)} \sum_{k=1}^{p} \lambda_k (u_j^k)^2 = 1$$

car, en utilisant encore (5.2) et la Proposition 5.2, puisque les variables principales sont décorrélées,

$$\mathrm{Var}\left(x^j\right) = \mathrm{Var}\left(\sum_{k=1}^{p} c^k u_j^k\right) = \sum_{k=1}^{p} (u_j^k)^2 \mathrm{Var}\left(c^k\right) = \sum_{k=1}^{p} \lambda_k (u_j^k)^2 \ .$$

Le *cercle des corrélations* est le graphique représentant les points $\mathcal{P}_j$, $j \in \{1, \dots, p\}$, sous forme de vecteurs d'origine nulle ainsi que le cercle unité (voir l'exemple de la Figure 5.3). Soit $j \in \{1, \dots, p\}$, plus le point $\mathcal{P}_j$ sera proche du cercle, plus la variable x^j associée aura de l'influence sur les axes du plan principal et y sera bien représentée.

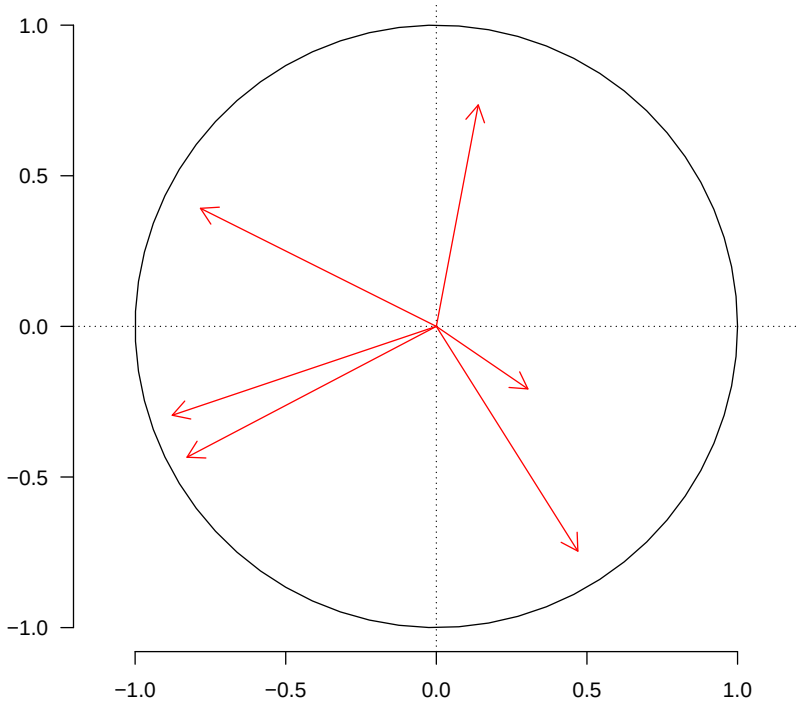


FIGURE 5.3 – Cercle des corrélations de $p = 6$ variables.

5.4 Inertie

L'objectif initial était de donner une "bonne" représentation des données $x_1, \dots, x_n \in \mathbb{R}^p$ dans le plan qui puisse rendre compte de la structure de ces observations. Dans le cadre de ce chapitre, nous utiliserons l'inertie standard I comme critère de qualité pour cette représentation,

$$I = I(x^1, \dots, x^p) = \sum_{j=1}^{p} \mathrm{Var}\left(x^j\right) \ .$$

Ce choix est, en particulier, motivé par le résultat suivant.

Proposition 5.4. *L'ACP conserve l'inertie standard,*

$$I(x^1, \dots, x^p) = I(c^1, \dots, c^p) = \sum_{k=1}^{p} \lambda_k \ .$$

Démonstration. La Proposition 5.2 donne

$$I(c^1, \dots, c^p) = \sum_{k=1}^{p} \mathrm{Var}\left(c^k\right) = \sum_{k=1}^{p} \lambda_k \ .$$

De plus, grâce à (5.2) et au fait que les variables principales sont décorrélées, nous calculons facilement l'inertie standard associée aux variables $x^1, \dots, x^p$,

$$\begin{aligned} I(x^1, \dots, x^p) &= \sum_{j=1}^{p} \mathrm{Var}\left(x^j\right) = \sum_{j=1}^{p} \mathrm{Var}\left(\sum_{k=1}^{p} c^k u_j^k\right) \\ &= \sum_{j=1}^{p}\sum_{k=1}^{p} (u_j^k)^2 \times \mathrm{Var}\left(c^k\right) \\ &= \sum_{k=1}^{p} \lambda_k \underbrace{\sum_{j=1}^{p} (u_j^k)^2}_{=1} = \sum_{k=1}^{p} \lambda_k \end{aligned}$$

car $u^1, \dots, u^p$ est une base orthonormée de $\mathbb{R}^p$. □

5.4.1 Qualité globale

Il est possible de considérer la contribution des variables principales à l'inertie standard I, pour $j \in \{1, \dots, k\}$,

$$\frac{\mathrm{Var}\left(c^j\right)}{I} = \frac{\lambda_j}{\lambda_1 + \dots + \lambda_p} \ .$$

Plus cette contribution sera grande, plus la dispersion dans la direction de la composante principale associée sera importante. Par construction, les deux composantes principales qui ont les plus grandes contributions sont C^1 et C^2 et nous utiliserons ces deux contributions pour quantifier la qualité de la représentation dans le plan principal.

Définition 5.2. *La* **part d'inertie** *expliquée par le plan principal est*

$$r_2 = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \dots + \lambda_2} \ .$$

La quantité r_2 mesure la qualité globale de la représentation dans le plan principal, *i.e.* plus r_2 sera proche de 1, meilleure sera cette représentation. Il est possible de se faire une idée de r_2 grâce à un graphique représentant la décroissance des valeurs propres $\lambda_1 \geqslant \dots \geqslant \lambda_p$ appelé *éboulis des valeurs propres* (voir Figure 5.4).

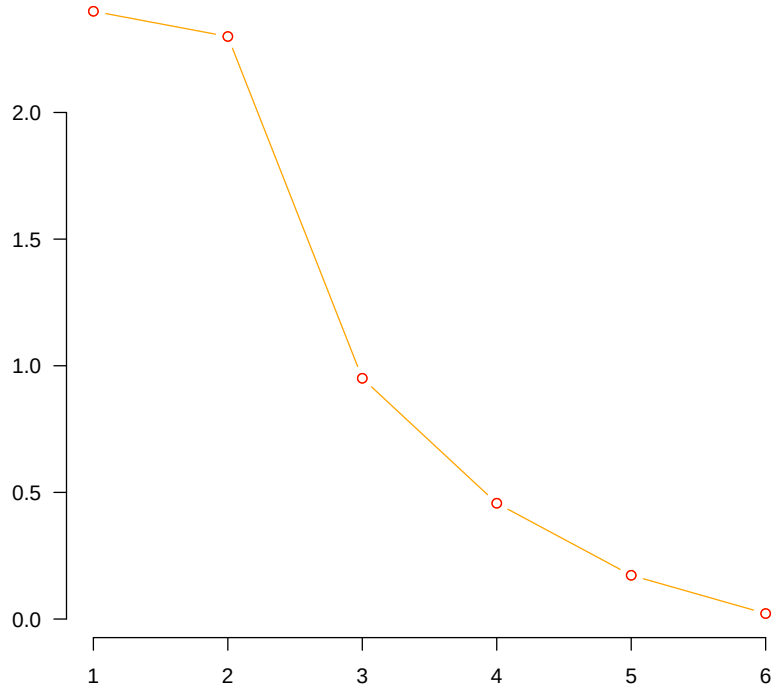


FIGURE 5.4 – Eboulis de $p = 6$ valeurs propres.

5.4.2 Changement de distance

Etant donnée une matrice S définie positive de taille $p \times p$, il est possible d'adapter les variables considérées à un problème donné en considérant les "nouvelles" données issues de la matrice $X_S = XS$.

Exemple Afin de mettre à la même échelle les variables $x^1, \ldots, x^p$, nous pouvons être amenés à considérer leurs versions réduites. Cela correspond à modifier le problème en prenant

$$S = \begin{bmatrix} \frac{1}{\sqrt{\mathrm{Var}\,(x^1)}} & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \frac{1}{\sqrt{\mathrm{Var}\,(x^p)}} \end{bmatrix} .$$

En effet, pour $i \in \{1, \ldots, n\}$ et $j \in \{1, \ldots, p\}$, il est facile de voir que le $i^{\text{ème}}$ observation de la $j^{\text{ème}}$ variable de X_S est donnée par

$$(X_S)_{ij} = (XS)_{ij} = \sum_{k=1}^{p} X_{ik} S_{kj} = \frac{x_i^j - \overline{x^j}}{\sqrt{x^j}} .$$

Par linéarité, les colonnes de X_S demeurent centrées. Pour $i \in \{1, \dots, n\}$, le $i^{\text{ème}}$ vecteur ligne de X_S, notée $X_{S,i} \in \mathbb{R}^p$, contient donc les observations centrées des "nouvelles" variables relatives au $i^{\text{ème}}$ individu. Ainsi, pour $j \in \{1, \dots, p\}$, nous avons la $i^{\text{ème}}$ observation de la $j^{\text{ème}}$ variable modifiée,

$$(X_{S,i})_j = (XS)_{ij} = \sum_{k=1}^{p} X_{ik} S_{kj} = \sum_{k=1}^{p} {}^tS_{jk} (X_i)_k = ({}^tSX_i)_j \ .$$

Autrement dit, $X_{S,i} = {}^tSX_i$ où X_i est le $i^{\text{ème}}$ vecteur ligne de X. Le carré de la distance euclidienne entre le $i^{\text{ème}}$ individu et le centre de gravité est donc donné par

$${}^t(X_{S,i}) X_{S,i} = {}^t\left({}^tSX_i\right) {}^tSX_i = {}^tX_i \left(S\,{}^tS\right) X_i = {}^tX_i M X_i$$

avec $M = S\,{}^tS$ qui est, par construction, une matrice symétrique définie positive de taille $p \times p$.

En transformant les données X en X_S, nous sommes donc naturellement amenés à travailler avec les quantités $d^2_M(x_i, g) = {}^tX_i M X_i$ et donc avec l'inertie I_M par rapport à la distance d_M sur $\mathbb{R}^p$. En faisant l'ACP à partir de X_S au lieu de X, c'est l'inertie I_M qui sera conservée et nous pourrons construire un nouveau plan principal de façon à conserver au maximum l'inertie I_M au lieu de l'inertie standard I. Cela donne une nouvelle procédure d'ACP dont l'interprétation géométrique dépendra du choix de la matrice M.

Annexe A

Exemple d'ACP

Nous illustrons ici l'analyse en composantes principales sur un cas concret pour des données issues du jeu vidéo "The Elder Scrolls V : Skyrim" développé par Bethesda Game Studios et édité par Bethesda Softworks. Il s'agit d'un jeu de rôle dans lequel le joueur a, entre autres choses, la possibilité d'utiliser des arcs (et des arbalètes) pour mener ses quêtes à bien. Les caractéristiques des ces arcs sont les suivantes [1] :

Nom	Poids	Valeur	Dégât	Vitesse
Long Bow	5	30	6	1
Hunting Bow	7	50	7	0.9375
Orcish Bow	9	150	10	0.8125
Nord Hero Bow	7	200	11	0.875
Dwarven Bow	10	270	12	0.75
Elven Bow	12	470	13	0.6875
Glass Bow	14	820	15	0.625
Ebony Bow	16	1440	17	0.5625
Daedric Bow	18	2500	19	0.5
Dragonbone Bow	20	2725	20	0.75
Crossbow	14	120	19	1
Enhanced Crossbow	15	200	19	1
Dwarven Crossbow	20	350	22	1
Enhanced Dwarven Crossbow	21	550	22	1

Nous disposons donc de $n = 14$ arcs représentés par $p = 4$ variables qui sont le poids de l'arc, sa valeur, les dégâts qu'il inflige et la vitesse à laquelle il tire les flèches. L'étude qui va suivre est réalisée à l'aide du logiciel libre R [2] et le fichier des données `skyrim_bows` est disponible sur la page de l'auteur :

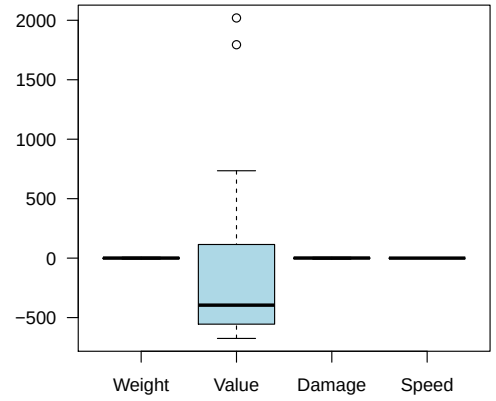
```
> data <- read.csv(file=
+       "http://www.math.univ-toulouse.fr/~xgendre/ens/l3sid/skyrim_bows")
```

Nous commençons par regarder l'allure générale des données centrées contenues dans la matrice X en affichant les boîtes à moustaches relatives aux 4 variables. L'échelle des variations

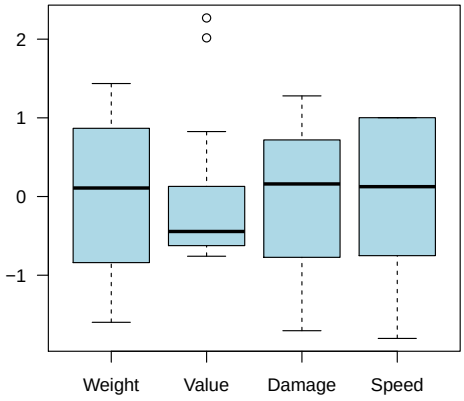
1. Source : http://www.uesp.net/wiki/Skyrim:Weapons#Archery
2. Voir http://www.r-project.org/

de la variable `Value` est nettement plus grande que celles des autres variables. Afin de ne pas concentrer notre étude uniquement sur cette variable, nous choisissons de normaliser les variables (*i.e.* nous travaillons avec la distance des variables réduites).

```
# Données centrées
> X <- scale(data[,2:5],scale=F)
> boxplot(X)
```




```
# Données centrées réduites
> X <- scale(data[,2:5])
> boxplot(X)
```

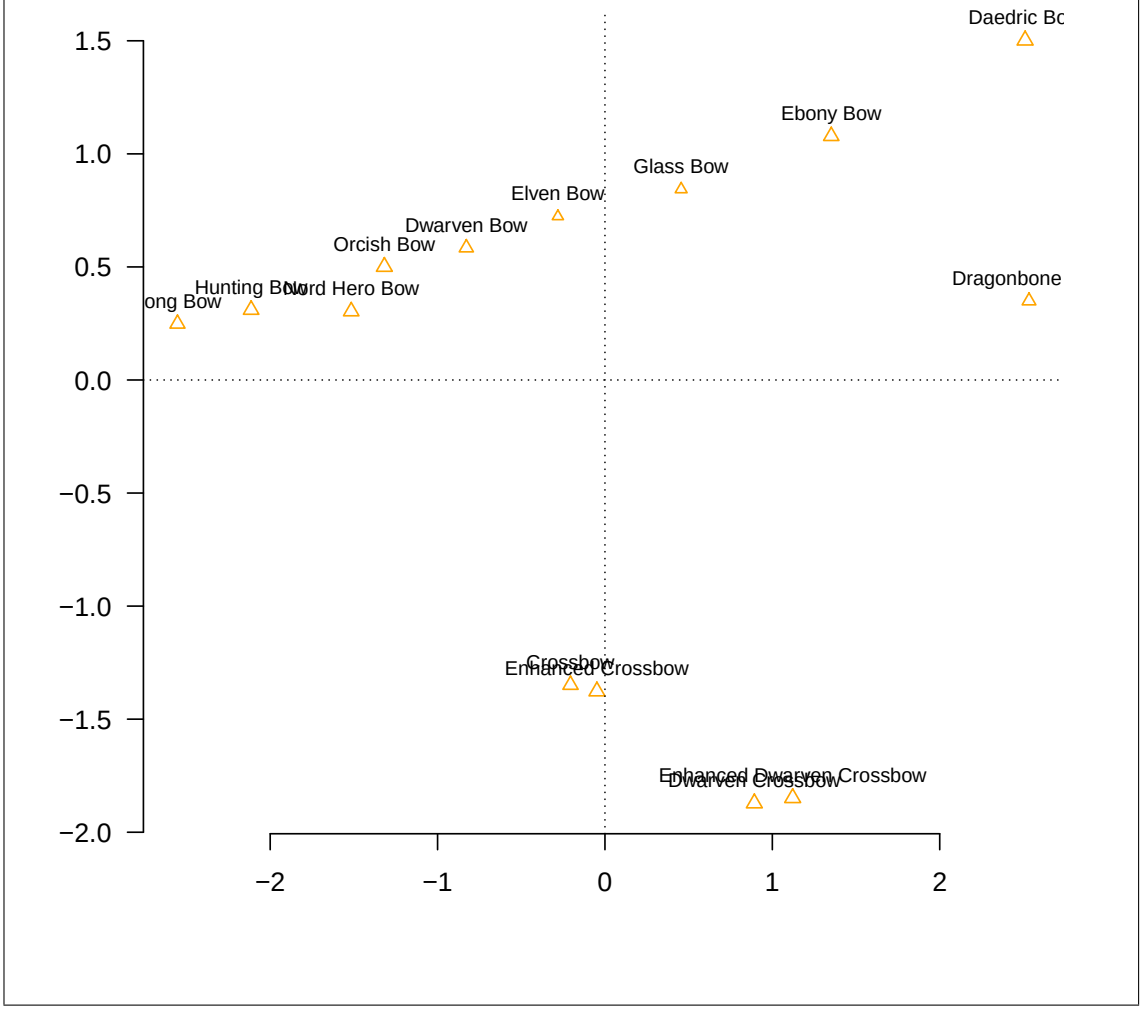


Nous supposons que les arcs sont tous pondérés de la même façon, *i.e.* $p_1 = \cdots = p_{14} = 1/14$. La matrice de covariance $\Sigma = {}^tXX/14$ se calcule facilement ainsi que ses valeurs et vecteurs propres. En particulier, nous obtenons la matrice des composantes principales $C = XP$ dont les deux premières colonnes contiennent les coordonnées des arcs dans le plan principal.

```
> ACP <- eigen(t(X) %*% X / 14)
> C <- X %*% ACP$vectors
> C
                              [,1]       [,2]        [,3]        [,4]
Long Bow                -2.55391815  0.2486141  0.67448976  0.05993000
Hunting Bow             -2.11378134  0.3102861  0.38825928  0.21054795
Orcish Bow              -1.31788780  0.5009160 -0.14036157  0.09991514
Nord Hero Bow           -1.51591815  0.3036447  0.10393262 -0.30871374
Dwarven Bow             -0.82823147  0.5847674 -0.37104812 -0.02758948
Elven Bow               -0.28118170  0.7233640 -0.50605134  0.09667827
Glass Bow                0.45522681  0.8433071 -0.56540530  0.07510570
Ebony Bow                1.35200475  1.0786370 -0.39791838  0.01400782
Daedric Bow              2.51012547  1.5020046  0.13923517 -0.11150169
Dragonbone Bow           2.53347110  0.3498669  1.06951370  0.01798606
Crossbow                -0.20575118 -1.3485746 -0.10187148 -0.30729279
Enhanced Crossbow       -0.04809313 -1.3765868 -0.05656160 -0.17940223
Dwarven Crossbow         0.89250072 -1.8717587 -0.19117039  0.12500263
Enhanced Dwarven Crossbow  1.12143407 -1.8484878 -0.04504234  0.23532638
```

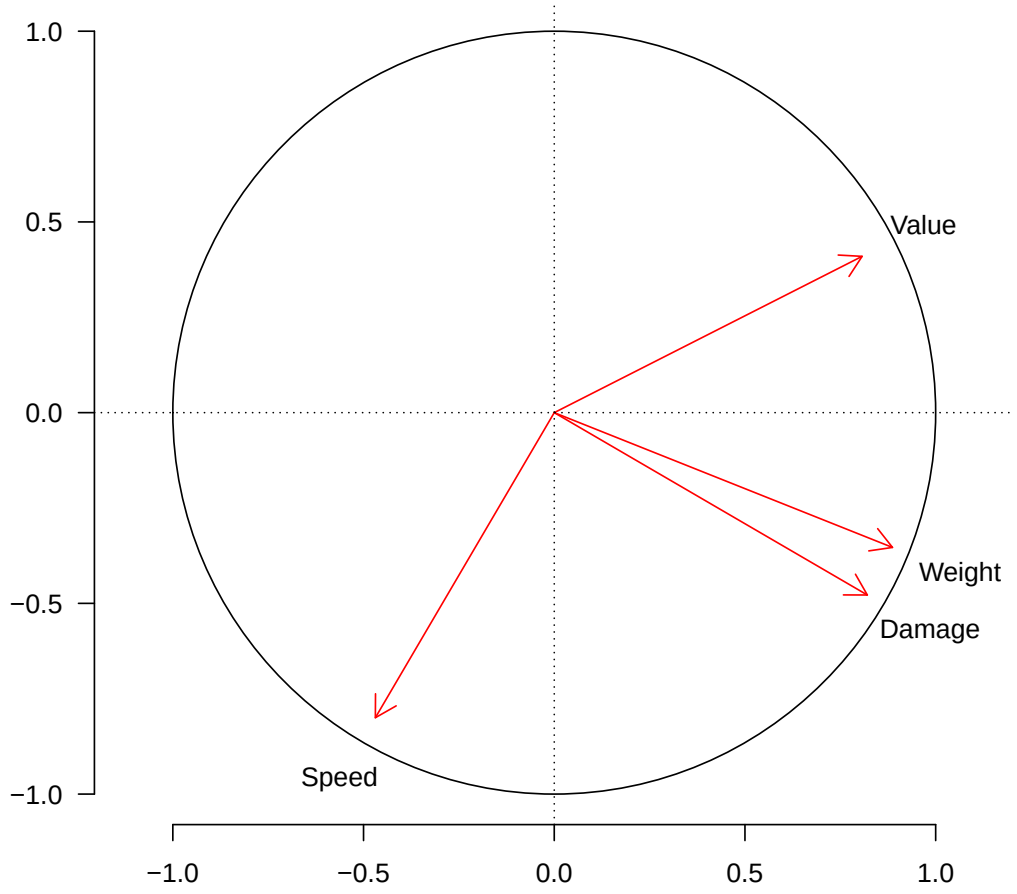
Afin de rendre compte graphiquement de la qualité de représentation de chaque individu dans le plan principal, nous calculons les $\cos^2 \theta_i$ (voir (5.1)) et nous représentons le $i^{\text{ème}}$ individu par un symbole dont la taille est proportionnelle à ce cosinus carré.

```
> cos2theta <- rowSums(C[,1:2]^2)/rowSums(C^2)
> plot(C[,1:2],cex=cos2theta); abline(h=0); abline(v=0)
> text(C[,1:2],labels=rownames(C),pos=3)
```



La première remarque que nous pouvons faire à l'aide de cette représentation est qu'elle permet de discriminer les arcs (*bow* en anglais) et les arbalètes (*crossbow* en anglais). En effet, les arcs ont tous des ordonnées positives alors que celles des arbalètes sont négatives. Pour aller plus loin dans l'interprétation des axes, nous considérons le cercle des corrélations.

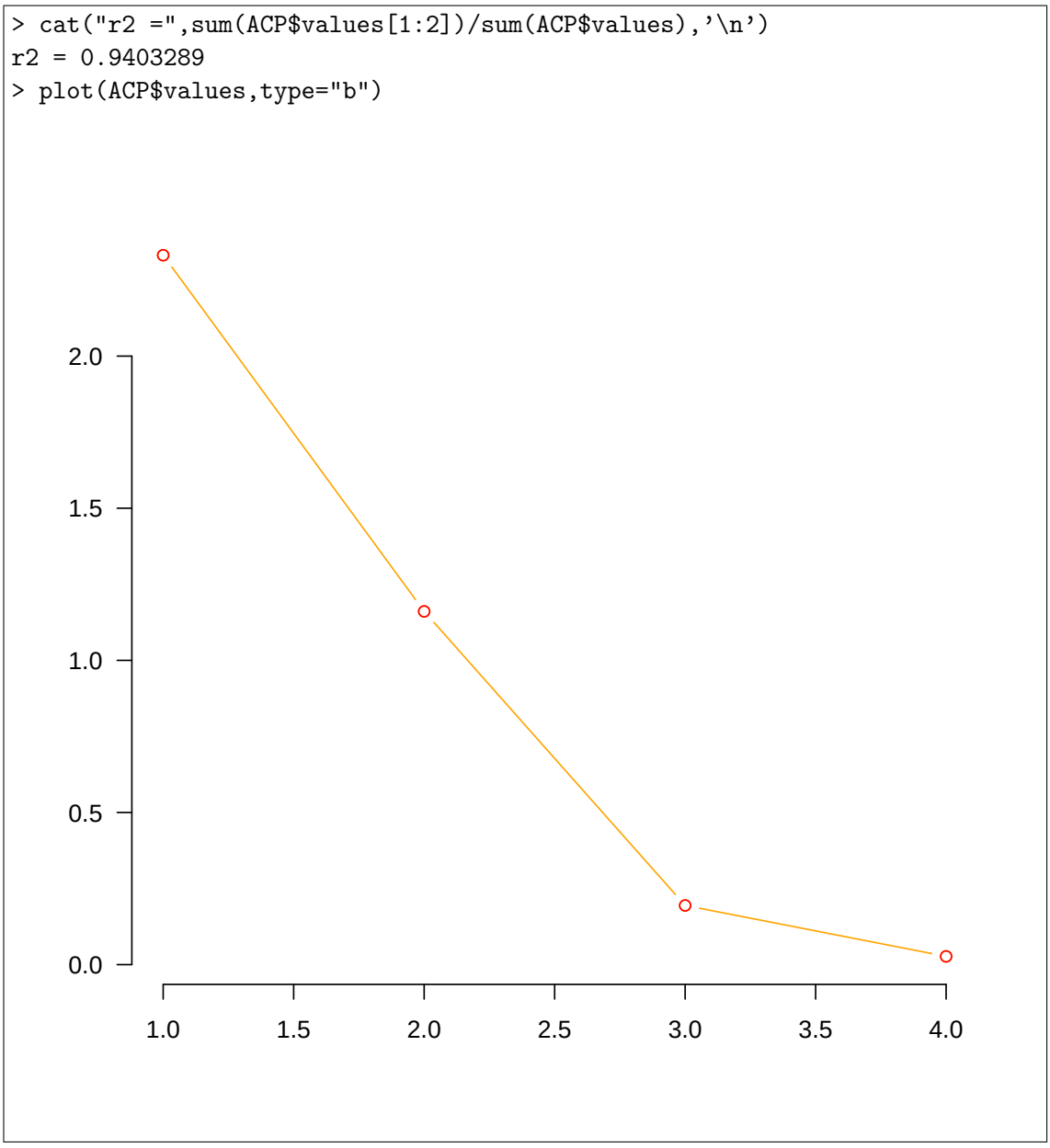
```
> rho <- ACP$vectors[,1:2] %*% diag(sqrt(ACP$values[1:2]))
             [,1]       [,2]
Weight  0.8867973 -0.3533824
Value   0.8071258  0.4097793
Damage  0.8208823 -0.4780940
Speed  -0.4690172 -0.7997447
> symbols(x=0,y=0,circles=1,inches=F)
> arrows(0,0,rho[,1],rho[,2])
```



Nous lisons que la première variable principale est essentiellement corrélée positivement avec les variables `Value`, `Weight` et `Damage`. Ainsi, les arcs dispendieux, lourds et faisant beaucoup de dégâts (*e.g. Dragonbone Bow*) auront tendance à être à droite sur le plan principal (*i.e.* à avoir de grandes abscisses). La deuxième variable principale est, quant à elle, surtout corrélée négativement avec la variable `Speed` : les arcs les plus rapides (principalment les arbalètes) auront donc tendance à se trouver en bas du plan principal (*i.e.* à avoir des ordonnées négatives).

Enfin, pour quantifier la qualité globale de la représentation et, donc, valider nos analyses, nous calculons la part d'inertie r_2 expliquée par le plan principal que nous illustrons à l'aide de l'éboulis des valeurs propres.

```
> cat("r2 =",sum(ACP$values[1:2])/sum(ACP$values),'\n')
r2 = 0.9403289
> plot(ACP$values,type="b")
```



Index