# MATH 444 – Statistique multivariée

## Dr. Jean-Marie Helbling

Notes de cours par
**Léo Belzile**
leo.belzile@epfl.ch

VERSION DU 2 MAI 2018
HIVER 2015 ET 2016, EPFL

*Écrire un courriel à l'auteur si vous trouvez une coquille.*

*Ces notes n'ont été que partiellement révisées et devraient être consultées avec prudence.*

# Table des matières

# Chapitre 1
# Introduction

Pour motiver le cours, on regarde quelques jeux de données qui seront utilisés dans le cours.

## 1.1 Jeux de données

- Crabes :
  - distribution de données
  - test de différences de moyennes (même longueur de carapace chez les mâles et les femelles).
  - séparer au mieux les groupes pour classer de nouveaux cas (MANOVA).
- Examens
  - analyse canonique : liens entre les types d'examen (livre ouvert ou fermé, par exemple)
  - représentation la plus simple possible ; par régression, ou projection
  - représentation en dimension inférieure (analyse en composante principale)
- Composition chimique du sang (BMDP)
  - discrimination
  - régression (dépendance du cholestérol et de l'âge)
  - données manquantes
  - aberrances
- Ventes
  - lien entre les notes d'examen et les ventes
  - formation de groupes plus ou moins homogènes (agglomération, ou "clustering")

## 1.2 Organisation des données

Les données seront stockées sous forme matricielle de dimension $n \times p$. On considère un échantillon de taille $n$ sur lequel on mesure $p$ variables.

$$\mathbf{X} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ X_{21} & \ddots & X_{2p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}$$

et les observations sont indépendantes et identiquement distribuées (iid). On dénote cette matrice $\mathbf{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_n^\top)^\top$ où le vecteur colonne $\boldsymbol{X}_i$ est la $i^{\text{e}}$ observation. Sur ce tableau de données, on calcule les statistiques descriptives suivantes :

- la moyenne $\bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p)^\top = n^{-1}\mathbf{X}^\top \mathbf{1}_n$, où $\bar{X}_1$ correspond à la moyenne de la 1$^{\text{e}}$variable sur tous les individus (moyenne de la 1$^{\text{e}}$colonne de $\mathbf{X}$).

- les variances :

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2, \qquad j = 1, \ldots, p$$

la variance de la variable $j$ sur tous les individus.
- les covariances

$$S_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k), \qquad j \neq k$$

ce qui donne la matrice de variance-covariance

$$\begin{aligned}
\boldsymbol{S} &= \begin{pmatrix} S_1^2 & \cdots & S_{jk} & \cdots & S_{1p} \\ \vdots & \ddots & \ddots & \cdots & \vdots \\ \vdots & S_{kj} & \ddots & \cdots & \vdots \\ S_{p1} & \cdots & \cdots & \cdots & S_p^2 \end{pmatrix} \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top. \\
&= \frac{1}{n-1}\left(\mathbf{X}^\top\mathbf{X} - n\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\top\right) \\
&= \frac{1}{n-1}\left(\mathbf{X}^\top\mathbf{X} - \frac{1}{n}\mathbf{X}^\top\mathbf{1}\mathbf{1}^\top\mathbf{X}\right) \\
&= \frac{1}{n-1}\mathbf{X}^\top\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)\mathbf{X}
\end{aligned}$$

Noter que $\mathbf{I} - \mathbf{1}\mathbf{1}^\top/n$ est une matrice idempotente, et de ce fait son rang est égal à sa trace. On a donc

$$\operatorname{tr}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right) = n - \frac{1}{n}n = n - 1.$$

- la matrice des corrélations, avec élément hors-diagonal $r_{jk} = S_{jk}/\sqrt{S_j^2 S_k^2}$.

$$\begin{aligned}
\boldsymbol{R} &= \begin{pmatrix} 1 & \cdots & r_{1p} \\ \vdots & \ddots & \vdots \\ r_{p1} & \cdots & 1 \end{pmatrix} \\
&= \left(\operatorname{diag}\left(\boldsymbol{S}\right)\right)^{-\frac{1}{2}} \boldsymbol{S} \left(\operatorname{diag}\left(\boldsymbol{S}\right)\right)^{-\frac{1}{2}}
\end{aligned}$$

- notions de variance généralisée : $\det(\boldsymbol{S})$ ou $\det(\boldsymbol{R})$, $\operatorname{tr}(\boldsymbol{S})$ ou valeurs propres.

**Note**

Les valeurs propres de $\boldsymbol{S}$ sont positives puisque $\boldsymbol{S}$ est symétrique (et définie positive).

## 1.3 Rappel de probabilités

Soit $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$, un vecteur aléatoire de dimension $p$, avec $\boldsymbol{\mu} = \mathsf{E}(\boldsymbol{X}) = (\mathsf{E}(X_1), \ldots, \mathsf{E}(X_p))^\top$ et

$$\boldsymbol{\Sigma}_X = \mathsf{Cov}(\boldsymbol{X}) = \mathsf{E}\left((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^\top\right) = \mathsf{E}\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top\right).$$

**Proposition 1.1 (Transformations linéaires)**
Si $\boldsymbol{Z} = \mathbf{C}\boldsymbol{X}$ et $\boldsymbol{W} = \mathbf{D}\boldsymbol{X}$, alors

$$\begin{aligned} \mathsf{E}(\boldsymbol{Z}) &= \boldsymbol{\mu}_{\boldsymbol{Z}} = \mathbf{C}\boldsymbol{\mu}_{\boldsymbol{X}}, \\ \mathsf{Cov}(\boldsymbol{Z}) &= \boldsymbol{\Sigma}_{\boldsymbol{Z}} = \mathbf{C}\boldsymbol{\Sigma}_{\boldsymbol{X}}\mathbf{C}^\top \end{aligned}$$

et

$$\mathsf{Cov}(\boldsymbol{Z}, \boldsymbol{W}) = \mathbf{C}\boldsymbol{\Sigma}_{\boldsymbol{X}}\mathbf{D}^\top.$$

Par exemple, $\mathsf{Cov}\left(\mathbf{a}^\top\boldsymbol{X}, \mathbf{b}^\top\boldsymbol{X}\right) = \mathbf{a}^\top\boldsymbol{\Sigma}_{\boldsymbol{X}}\mathbf{b}$.

**Définition 1.2 (Fonction de répartition (multivariée))**
La fonction de répartition $F : \mathbb{R}^d \to [0, 1]$ est définie pour tout $x_1, \ldots, x_d \in \mathbb{R}$ par

$$F(X_1, \ldots, X_n) = \mathsf{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_d \leq x_d).$$

**Propriétés de $F$**
$F$ caractérise le comportement aléatoire de $\boldsymbol{x}$. Toute fonction de répartition $F$ satisfait les propriétés suivantes.

(i)

$$\lim_{x_1 \to \infty} \ldots \lim_{x_d \to \infty} F(x_1, \ldots, x_d) = 1;$$

(ii)

$$\lim_{x_i \to -\infty} F(x_1, \ldots, x_d) = 0, \quad \forall\, x_i$$

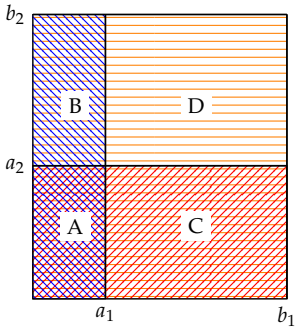
puisque l'intersection de $\emptyset$ avec n'importe lequel autre ensemble est $\emptyset$, de probabilité nulle.

(iii) $F$ est non-décroissante et continue à droite dans chacun de ces $d$ arguments. Par example, on peut considérer fixer $x_2, \ldots, x_d \in \mathbb{R}$, de telle sorte que $t \mapsto F(t, x_2, \ldots, x_d)$ est non-décroissante et continue à droite.

(iv) $F$ est $d-$monotone. Dans le cas $d = 2$, cela signifie que pour $x_1 < x_1^*$ et $x_2 < x_2^*$

$$F(x_1^*, x_2^*) - F(x_1, x_2^*) - F(x_1^*, x_2) + F(x_1, x_2) \geq 0$$

FIGURE 1 – $2-$monotonicité illustré par le biais d'ensemble




L'image illustre $\mathsf{P}(A \cup B \cup C \cup D) - \mathsf{P}(A \cup C) - \mathsf{P}(A \cup B) + \mathsf{P}(A)$ puisqu'il s'agit d'ensembles disjoints. En terme d'aire, c'est $(A + B + C + D) - (A + C) - (A + B) + A = D$ et ainsi la probabilité $\mathsf{P}(a_1 < X_1 < b_1, a_2 < X_2 < b_2) = D$ est positive.

Le cas $d = 3$ donne

$$\begin{aligned} F(x_1^*, x_2^*, x_3^*) - F(x_1, x_2^*, x_3^*) + F(x_1, x_2, x_3^*) - F(x_1, x_2, x_3) \\ - F(x_1^*, x_2, x_3^*) + F(x_1, x_2^*, x_3) \\ - F(x_1^*, x_2^*, x_3) + F(x_1^*, x_2, x_3) \geq 0 \end{aligned}$$

et en dimension $d$ arbitraire

$$\sum_{\substack{c_i \in \{x_i, x_i^*\} \\ i \in \{1, \ldots, d\}}} (-1)^{v(c)} \times F(c_1, \ldots, c_d) \geq 0$$

où $v(c)$ est le nombre de constantes $c_i$ sans étoiles. Inversement, toute fonction $F : \mathbb{R}^d \to [0, 1]$ qui satisfait (i)–(iv) est une fonction de répartition valide, c'est-à-dire qu'il existe une mesure de probabilité qui engendre $F$. La preuve est la même que dans le cas univarié ; l'élément clé est que $(x_1, x_1^*] \times (x_2, x_2^*] \times \cdots \times (x_d, x_d^*]$ sont préciséments des générateurs de la tribu borélienne sur $\mathbb{R}^d$.

Les propriétés (i)–(iii) sont suffisantes en une dimension, mais la propriété essentielle pour la construction d'une fonction de répartition multivariée est la (iv). On dénomme **loi marginale** la loi correspondant à une sous-composante de dimension $d - q$ du $d$-vecteur $\boldsymbol{X}$, pour $1 \leq q \leq d$.

# Chapitre 2
# Loi normale multivariée et ses dérivées

## 2.1 La loi normale

**Définition 2.1 (Loi normale univariée)**
On caractérise la loi normale par sa densité

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

où $(x-\mu)^2/\sigma^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$. On dénote par $X \sim \mathcal{N}(\mu, \sigma^2)$.

**Définition 2.2 (Loi normale multivariée)**
La densité de la loi multinormale est

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

**Note**
La loi multinormale est entièrement caractérisée par ses deux premiers moments, respectivement $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$. Son utilisation pour certaines quantités d'intérêt (la moyenne arithmétique) est justifiée asymptotiquement par le théorème central limite.

**Proposition 2.3**
Si $\boldsymbol{\Sigma}$ est positive définie, alors $\mathsf{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ et $\mathsf{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

**Preuve** Puisque $\boldsymbol{\Sigma}$ est positive définie, il existe une matrice orthogonale $\boldsymbol{P}$ telle que $\boldsymbol{P}^\top \boldsymbol{\Sigma} \boldsymbol{P} = \boldsymbol{\Lambda}$, où $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ est une matrice diagonale de valeurs propres. Posons $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{Y} + \boldsymbol{\mu}$ et calculons la densité de $\boldsymbol{Y}$. Le jacobien de la transformation est $\mathbf{J} = \left|\partial \boldsymbol{X}_i / \partial \boldsymbol{y}_j\right|_{i,j} = |\boldsymbol{P}|$ et l'on note que $|\boldsymbol{\Sigma}| = |\boldsymbol{P}|^2 \prod_{i=1}^p \lambda_i$. La densité de $\boldsymbol{Y}$ est égale à $f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(\boldsymbol{P}\boldsymbol{y} + \boldsymbol{\mu})|\mathbf{J}|$. Dès lors,

$$\begin{aligned} f(\boldsymbol{y}) &= \frac{1}{(2\pi)^{p/2}\sqrt{\lambda_1 \cdots \lambda_p}} \exp\left(-\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{P}^\top \boldsymbol{\Sigma}^{-1} \mathbf{P}\mathbf{y}\right) \\ &= \frac{1}{(2\pi)^{p/2}\sqrt{\lambda_1 \cdots \lambda_p}} \exp\left(-\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{y}\right) \\ &= f(y_1) f(y_2) \cdots f(y_p) \end{aligned}$$

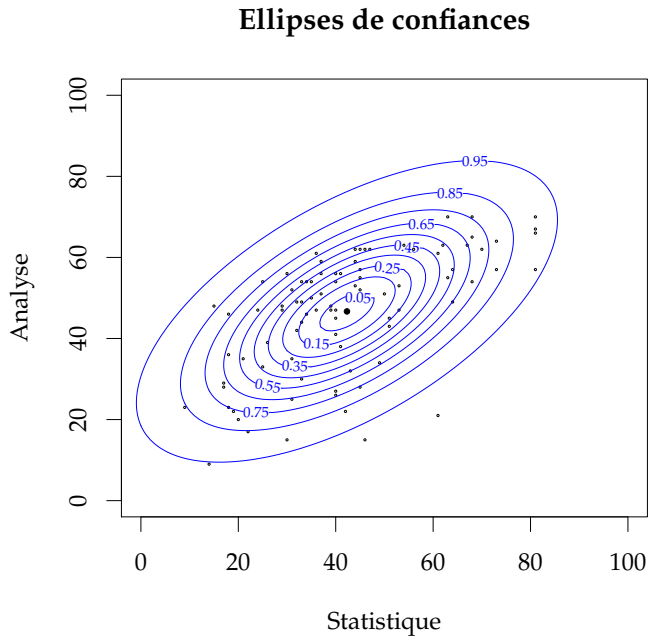
puisque $\boldsymbol{\Sigma}^{-1} = (\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top)^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^\top$. Ainsi, $\boldsymbol{Y} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Lambda})$ et en plus, $Y_i \sim \mathcal{N}(0, \lambda_i)$. Par linéarité, $\mathsf{E}(\boldsymbol{X}) = \boldsymbol{P}\mathsf{E}(\boldsymbol{Y}) + \boldsymbol{\mu} = \boldsymbol{\mu}$. L'on a également

$$\mathsf{Var}(\boldsymbol{X}) = \mathsf{Var}(\boldsymbol{P}\boldsymbol{Y} + \boldsymbol{\mu}) = \mathsf{Var}(\boldsymbol{P}\boldsymbol{Y}) = \boldsymbol{P}\mathsf{Var}(\boldsymbol{Y})\boldsymbol{P}^\top = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top = \boldsymbol{\Sigma}.$$

■

Pour la loi normale multivariée, les courbes de densité constante sont définies par $c^2 = (\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$. Elles représentent des ellipses de centre $\boldsymbol{\mu}$ et d'axes $\pm c\sqrt{\lambda_i}\mathbf{v}_i$ où $\lambda_i$ est une valeur propre de $\boldsymbol{\Sigma}$ avec vecteur propre $\mathbf{v}_i$. La forme dépend de $\boldsymbol{\Sigma}$ et $c$. [1]

FIGURE 2 – Ellipses de confiance pour la distribution marginale {analyse, statistique} du jeu de données examens




**Proposition 2.4**
La forme quadratique

$$\boldsymbol{Q} = (\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \chi^2_p$$

suit une loi de khi-carré avec $p$ degrés de liberté.

**Définition 2.5 (Fonction caractéristique)**
On définit la fonction caractéristique d'un vecteur aléatoire comme

$$\Phi_{\boldsymbol{X}}(\mathbf{t}) \coloneqq \mathsf{E}\left(\exp(\mathrm{i}\mathbf{t}^\top \boldsymbol{X})\right) = \int \exp(\mathrm{i}\mathbf{t}^\top \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

1. Il en résulte qu'il est facile de calculer la probabilité pour une loi multinormale d'être dans une ellipsoïde, contrairement par exemple à un parallélépipède.

pour tout $\mathbf{t} \in \mathbb{R}^p$. Ses propriétés sont

- Existence : la fonction caractéristique existe toujours et elle satisfait $\Phi_{\boldsymbol{X}}(\mathbf{0}) = 1$ et $|\Phi_{\boldsymbol{X}}(\mathbf{t})| \leq 1$.
- Unicité : deux vecteurs aléatoires dotés de la même fonction caractéristiques si et seulement si ils ont la même distribution.
- Inversion : si $\Phi_{\boldsymbol{X}}(\mathbf{t})$ est absolument intégrable, alors la densité de $\boldsymbol{X}$ est donnée par

$$f(\mathbf{x}) = \frac{1}{(2\pi)^p} \int \exp(-\mathrm{i}\mathbf{t}^\top \boldsymbol{x}) \Phi_{\boldsymbol{X}}(\mathbf{t}) \, \mathrm{d}\mathbf{t}$$

- Indépendance : la fonction caractéristique du vecteur $\boldsymbol{X} = (\boldsymbol{X}_1 \quad \boldsymbol{X}_2)$ est telle que $\Phi_{\boldsymbol{X}}(\mathbf{t}) = \Phi_{\boldsymbol{X}_1}(\mathbf{t}) \Phi_{\boldsymbol{X}_2}(\mathbf{t})$ si et seulement si $\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2$.
- Moments :

$$\mathsf{E}\left(X_1^{j_1} X_p^{j_p}\right) = \mathrm{i}^{-(j_1+\cdots+j_p)} \left[ \frac{\partial^{j_1+\cdots+j_p}}{\partial t_1^{j_1} \partial t_p^{j_p}} \Phi_{\boldsymbol{X}}(\mathbf{t}) \right]_{\mathbf{t}=0}$$

- La fonction caractéristique du $(d-q)$-vecteur marginal issu d'un $d$-vecteur $\boldsymbol{X}$ pour $1 \leq q \leq d$ est $\Phi_{\boldsymbol{X}}\left(\begin{smallmatrix} \mathbf{t}_1 \\ \mathbf{0}_q \end{smallmatrix}\right)$
- Si $\boldsymbol{X}$ et $\boldsymbol{Y}$ sont indépendants, alors $\Phi_{\boldsymbol{X}+\boldsymbol{Y}}(\mathbf{t}) = \Phi_{\boldsymbol{X}}(\mathbf{t}) \Phi_{\boldsymbol{Y}}(\mathbf{t})$

**Théorème 2.6 (Cramér-Wold)**
La distribution d'un vecteur aléatoire $\boldsymbol{X}$ est complètement déterminée par l'ensemble de toutes les lois univariées formées des combinaisons linéaires $\mathbf{a}^\top \boldsymbol{X}$, pour tout $\mathbf{a} \in \mathbb{R}^p$.

**Corollaire 2.7**
$\boldsymbol{X}$ suit une loi multinormale si $\mathbf{a}^\top \boldsymbol{X}$ est (multi)normale pour tout $\mathbf{a} \in \mathbb{R}^p$.

**Proposition 2.8 (Fonction caractéristique de la loi multinormale)**
La fonction caractéristique de la loi multinormale est donnée par

$$\exp\left(\mathrm{i}\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top \boldsymbol{\Sigma} \mathbf{t}\right)$$

**Preuve** On procède au calcul en complétant le carré :

$$\begin{aligned}
\Phi_{\boldsymbol{X}}(\mathbf{t}) &= \mathsf{E}\left(\exp(\mathrm{i}\mathbf{t}^\top \boldsymbol{X})\right) \\
&= (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \int \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \exp\left(\mathrm{i}\mathbf{t}^\top \boldsymbol{x}\right) \mathrm{d}\boldsymbol{x} \\
&\overset{\boldsymbol{y}=\boldsymbol{x}-\boldsymbol{\mu}}{=} (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \int \exp\left(-\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}\right) \exp\left(\mathrm{i}\mathbf{t}^\top (\boldsymbol{y}+\boldsymbol{\mu})\right) \mathrm{d}\boldsymbol{y} \\
&= (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \int \exp\left(-\frac{1}{2}(\boldsymbol{y} - i\boldsymbol{\Sigma}\mathbf{t})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - i\boldsymbol{\Sigma}\mathbf{t})\right) \mathrm{d}\boldsymbol{y} \\
&\quad \times \exp\left(\mathrm{i}\mathbf{t}^\top \boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^\top \boldsymbol{\Sigma} \mathbf{t} i^2\right)
\end{aligned}$$

puisque

$$\begin{aligned}-\frac{1}{2}\boldsymbol{y}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}+\mathrm{i}\mathbf{t}^\top\boldsymbol{y}+\mathrm{i}\mathbf{t}^\top\boldsymbol{\mu} &= -\frac{1}{2}\left(\boldsymbol{y}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y}-2\mathrm{i}\mathbf{t}^\top\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}\right)+\mathrm{i}\mathbf{t}^\top\boldsymbol{\mu}\\ &= -\frac{1}{2}(\boldsymbol{y}-i\boldsymbol{\Sigma}\mathbf{t})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-i\boldsymbol{\Sigma}\mathbf{t})+\mathrm{i}\mathbf{t}^\top\boldsymbol{\mu}+\frac{1}{2}\mathbf{t}^\top\boldsymbol{\Sigma}\mathbf{t}i.\end{aligned}$$

■

**Corollaire 2.9**

Si $\boldsymbol{X}\sim\mathcal{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ et qu'on partitionne le vecteur en $\boldsymbol{X}_1,\boldsymbol{X}_2$, alors

$$\boldsymbol{X}=\begin{pmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{pmatrix}\sim\mathcal{N}_p\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix},\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}\right)$$

**Note**

En supposant que $\boldsymbol{X}$ suit une loi normale, les sous-vecteurs $\boldsymbol{X}_1,\boldsymbol{X}_2$ de dimension respective $q, d-q$ sont indépendants[2] (dénoté $\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2$) si et seulement si $\boldsymbol{\Sigma}_{12}=\boldsymbol{\Sigma}_{21}^\top=\mathbf{O}_{q,d-q}$.

**Proposition 2.10 (Distribution de la moyenne)**

La distribution de la moyenne arithmétique de multinormales est $\bar{\boldsymbol{X}}\sim\mathcal{N}_p(\boldsymbol{\mu},n^{-1}\boldsymbol{\Sigma})$.

**Preuve** On écrit $\bar{\boldsymbol{X}}=n^{-1}\sum_{i=1}^n\boldsymbol{X}_i$ avec $\boldsymbol{X}_i\overset{\text{iid}}{\sim}\mathcal{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Puisque $\bar{\boldsymbol{X}}$ est une combinaison linéaire des $\boldsymbol{X}_i$, on a par linéarité $\mathsf{E}\left(\bar{\boldsymbol{X}}\right)=\boldsymbol{\mu}$ et $\mathsf{Var}\left(\bar{\boldsymbol{X}}\right)=n^{-2}n\boldsymbol{\Sigma}=n^{-1}\boldsymbol{\Sigma}$. ■

**Proposition 2.11**

Si $\boldsymbol{X}\sim\mathcal{N}_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ et $\boldsymbol{Y}=\mathbf{C}\boldsymbol{X}+\mathbf{d}$, où $\mathbf{C}$ est une matrice non-stochastique $q\times p$ et $\mathbf{d}$ un vecteur colonne de dimension $q$. Alors $\boldsymbol{Y}\sim\mathcal{N}_q\left(\mathbf{C}\boldsymbol{\mu}+\mathbf{d},\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top\right)$

**Corollaire 2.12**

Si $\mathbf{C}=\begin{pmatrix}\mathbf{I}_s & \mathbf{O}_{s,q-s}\\ \mathbf{O}_{q-s,s} & \mathbf{O}_{q-s,q-s}\end{pmatrix}$ pour $1\leq s\leq q$ et $\mathbf{d}=\mathbf{0}_q$, alors on obtient les distributions marginales, lesquelles sont normales. Si on partitionne $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2)^\top$ de dimensions $s, p-s$ respectivement, alors l'utilisation de la matrice $\mathbf{C}=\begin{pmatrix}\mathbf{I}_s & \mathbf{O}_{s,p-s}\\ -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \mathbf{I}_{p-s}\end{pmatrix}$ dans le précédent résultat donne

$$\boldsymbol{Y}\equiv\begin{pmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1\end{pmatrix}\sim\mathcal{N}_p\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1\end{pmatrix},\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \mathbf{O}_{s,p-s}\\ \mathbf{O}_{p-s,s} & \boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\end{pmatrix}\right)$$

et donc la distribution conditionnelle est $\boldsymbol{X}_2\mid\boldsymbol{X}_1\sim\mathcal{N}_{p-s}(\boldsymbol{\mu}_2+\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu}_1),\boldsymbol{\Sigma}_{22.1})$, où $\boldsymbol{\Sigma}_{22.1}:=\boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$. En effet, sachant $\boldsymbol{X}_1$, $-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1$ est une constante et on peut déduire $\mathsf{E}\left(\boldsymbol{X}_2\right)$ et $\mathsf{Var}\left(\boldsymbol{X}_2\right)$ de l'expression pour $\boldsymbol{X}_2-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1$.

2. La normalité des marges n'est pas une condition suffisante ; la preuve se fait par le biais de la fonction caractéristique.

## 2.2 Estimation des paramètres $\mu$ et $\Sigma$

Dans cette section, on dérive le maximum de vraisemblance des paramètres de la loi multinormale. Soit un échantillon $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. La log-vraisemblance $\ell = \log(L)$ est donnée par

$$\begin{aligned}\ell(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det\left(\Sigma^{-1}\right) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}) \\ &= -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det\left(\Sigma^{-1}\right) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^\top\right)\end{aligned}$$

On s'intéresse à la trace et tout particulièrement à la somme des termes centrés. L'on a pour cette dernière

$$\sum_{i=1}^{n}(\boldsymbol{X}_i \pm \bar{\boldsymbol{X}} - \boldsymbol{\mu})(\boldsymbol{X}_i \pm \bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top = \sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top + n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top$$

puisque les termes croisés $\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top$ sont nuls. La log-vraisemblance par rapport à $\boldsymbol{\mu}$ est proportionnelle à

$$\ell(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \overset{\boldsymbol{\mu}}{\propto} -\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\left(\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top + n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top\right)\right)$$

et en simplifiant l'écriture, on a que

$$\begin{aligned}\ell(\boldsymbol{X}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = &-\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det\left(\boldsymbol{\Sigma}^{-1}\right) \\ &- \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{XX}\right) - \frac{n}{2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}).\end{aligned}$$

Le maximum de vraisemblance de la moyenne découle immédiatement de la dernière expression et on trouve $\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$. En dérivant par rapport à la matrice de précision, on obtient

$$\frac{\partial \ell}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{n}{2}\boldsymbol{\Sigma} - \frac{1}{2}\boldsymbol{S}_{XX}^\top$$

et en égalant à zéro, on obtient $\widehat{\boldsymbol{\Sigma}} = n^{-1}\boldsymbol{S}_{XX}$. On procède maintenant au calcul du biais des estimateurs du maximum de vraisemblance (EMV) :

$$\begin{aligned}\mathsf{E}\left(\widehat{\boldsymbol{\mu}}\right) &= \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\right) = \frac{n}{n}\boldsymbol{\mu} = \boldsymbol{\mu} \\ \mathsf{E}\left(\widehat{\boldsymbol{\Sigma}}\right) &= \frac{1}{n}\mathsf{E}\left(\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top\right)\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\mathsf{E}\left(\sum_{i=1}^{n}(\boldsymbol{X}_i-\boldsymbol{\mu})(\boldsymbol{X}_i-\boldsymbol{\mu})^\top - n(\boldsymbol{\mu}-\bar{\boldsymbol{X}})(\boldsymbol{\mu}-\bar{\boldsymbol{X}})^\top\right)\\
&= \frac{1}{n}n\boldsymbol{\Sigma} - \mathsf{Var}\left(\bar{\boldsymbol{X}}\right)\\
&= \frac{n-1}{n}\boldsymbol{\Sigma}
\end{aligned}
$$

et ce dernier estimateur a un biais de $n^{-1}\boldsymbol{\Sigma}$. On peut prendre

$$
S := \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{X}_i-\bar{\boldsymbol{X}})^\top
$$

comme estimateur de $\boldsymbol{\Sigma}$.

## 2.3 Lois dérivées de la loi normale multivariée

### Définition 2.13 (Loi de Wishart)

La distribution de Wishart, dénotée $\mathcal{W}$ est une distribution sur les matrices aléatoires positives définies et une généralisation multivariée de la distribution de $\chi^2$.

On considère $n$ vecteurs aléatoires indépendants $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ avec $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma})$. Alors

$$
\boldsymbol{W} = \sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^\top = \mathbf{X}^\top\mathbf{X} \sim \mathcal{W}_p(n,\boldsymbol{\Sigma},\boldsymbol{M})
$$

où $\boldsymbol{M} = (\boldsymbol{\mu}_1^\top,\dots,\boldsymbol{\mu}_n^\top)^\top$ est une matrice $p\times n$. La loi est dite centrée si $\boldsymbol{M}=\boldsymbol{O}$ et on écrit alors $\mathcal{W}_p(n,\boldsymbol{\Sigma})$.

### Proposition 2.14 (Propriétés de la loi de Wishart)

1. $\mathsf{E}\left(\boldsymbol{W}\right) = n\boldsymbol{\Sigma} + \boldsymbol{M}\boldsymbol{M}^\top$
2. Si $\mathbf{c}$ est un vecteur non aléatoire, alors $\mathbf{c}^\top\boldsymbol{W}\mathbf{c} \dot{\sim} \sigma_{\mathbf{c}}^2\chi_n^2$ avec $\sigma_{\mathbf{c}}^2 = \mathbf{c}^\top\boldsymbol{\Sigma}\mathbf{c}$. Si $\mathbf{C}$ est une matrice (non stochastique) de rang $q$ et que $\boldsymbol{W}\sim\mathcal{W}_p(n,\boldsymbol{\Sigma})$, alors
$$
\mathbf{C}\boldsymbol{W}\mathbf{C}^\top \sim \mathcal{W}_q(n,\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top).
$$
3. $\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X}\sim\mathcal{W}_p(\mathrm{rang}(\mathbf{A}),\boldsymbol{\Sigma},\cdot)$ si et seulement si $\mathbf{A}^2=\mathbf{A}$. De plus, la distribution $\mathcal{W}_p$ est centrée si $\mathbf{A}\boldsymbol{M}=\mathbf{0}$.
4. $\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X}$ et $\boldsymbol{X}^\top\mathbf{B}\boldsymbol{X}$ de lois Wishart sont indépendants si et seulement si $\mathbf{A}\mathbf{B}=\mathbf{0}$
5. $\boldsymbol{X}^\top\mathbf{B}$ et $\boldsymbol{X}^\top\mathbf{A}\boldsymbol{X}$ de lois Wishart sont indépendants si et seulement si $\mathbf{B}^\top\mathbf{A}=\mathbf{0}$.
6. Si $\boldsymbol{W}\sim\mathcal{W}_p(n,\boldsymbol{\Sigma})$, alors $|\boldsymbol{W}|/|\boldsymbol{\Sigma}| \sim \chi_n^2\chi_{n-1}^2\cdots\chi_{n-p+1}^2$ où les $p$ variables aléatoires khi-carrées sont indépendantes.
7. Si $\boldsymbol{W}\sim\mathcal{W}_p(n,\boldsymbol{\Sigma})$ alors pour tout vecteur $\mathbf{a}$ fixé, on a
$$
\frac{\mathbf{a}^\top\boldsymbol{\Sigma}^{-1}\mathbf{a}}{\mathbf{a}^\top\boldsymbol{W}^{-1}\mathbf{a}} \sim \chi_{n-p+1}^2.
$$

on obtient $\mathbf{1}^\top \boldsymbol{W}\mathbf{1} \sim \chi^2_{n-p+1}$.

**Preuve**

1. Écrire

$$\begin{aligned}\boldsymbol{W} &= \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top = \sum_{i=1}^n (\boldsymbol{X}_i - \boldsymbol{\mu}_i + \boldsymbol{\mu}_i)(\boldsymbol{X}_i - \boldsymbol{\mu}_i + \boldsymbol{\mu}_i)^\top \\ &= \sum_{i=1}^n (\boldsymbol{X}_i - \boldsymbol{\mu}_i)(\boldsymbol{X}_i - \boldsymbol{\mu}_i)^\top + \sum_{i=1}^n \boldsymbol{\mu}_i (\boldsymbol{X}_i - \boldsymbol{\mu}_i)^\top + \sum_{i=1}^n (\boldsymbol{X}_i - \boldsymbol{\mu}_i)\boldsymbol{\mu}_i^\top + \sum_{i=1}^n \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top\end{aligned}$$

et il suffit maintenant de calculer l'espérance. On obtient

$$\begin{aligned}\mathsf{E}(\boldsymbol{W}) &= \sum_{i=1}^n \mathsf{E}\left((\boldsymbol{X}_i - \boldsymbol{\mu}_i)(\boldsymbol{X}_i - \boldsymbol{\mu}_i)^\top\right) + \sum_{i=1}^n \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top \\ &= n\boldsymbol{\Sigma} + \boldsymbol{M}\boldsymbol{M}^\top\end{aligned}$$

2. On a $\boldsymbol{W} = \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top$ avec $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. En multipliant par une matrice $\mathbf{C}$, on obtient la forme quadratique

$$\mathbf{C}\boldsymbol{W}\mathbf{C}^\top = \sum_{i=1}^n \mathbf{C}\boldsymbol{X}_i (\mathbf{C}\boldsymbol{X}_i)^\top = \sum_{i=1}^n \boldsymbol{Z}_i \boldsymbol{Z}_i^\top.$$

Or, $\boldsymbol{Z}_i \sim \mathcal{N}_q(\mathbf{C}\boldsymbol{\mu}_i, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top)$, ce qui implique que $\mathbf{C}\boldsymbol{W}\mathbf{C}^\top \sim \mathcal{W}_q(n, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top, \cdot)$.
Dans le cas vectoriel,

$$\mathbf{c}^\top \boldsymbol{W}\mathbf{c} = \sum_{i=1}^n (\mathbf{c}^\top \boldsymbol{X}_i)(\boldsymbol{X}_i^\top \mathbf{c}) = \sum_{i=1}^n Y_i^2 = \sigma_{\mathbf{c}}^2 \sum_{i=1}^n \left(\frac{Y_i}{\sigma_{\mathbf{c}}}\right)^2 = \sigma_{\mathbf{c}}^2 \chi_n^2$$

avec $Y_i \sim \mathcal{N}\left(\mathbf{c}^\top \boldsymbol{\mu}_i, \sigma_{\mathbf{c}}^2\right)$ où $\sigma_{\mathbf{c}}^2 := \mathbf{c}^\top \boldsymbol{\Sigma}\mathbf{c}$.
3. On prouve un énoncé indirect : posons $\boldsymbol{Z} = \boldsymbol{X}\mathbf{c}$ pour un vecteur constant $\mathbf{c} \neq \mathbf{0}$. Alors $\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X} \sim \mathcal{W}_p(r, \boldsymbol{\Sigma}, \cdot)$ avec $r = \operatorname{rang}(\mathbf{A}) = \operatorname{tr}(\mathbf{A})$ si et seulement si $\boldsymbol{Z}^\top \mathbf{A}\boldsymbol{Z} \sim \sigma_{\mathbf{c}}^2 \chi_r^2$, pour tout $\mathbf{c}$.

La nécessité découle de la proposition (point 2). On a $\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X} \sim \mathcal{W}_p(r, \boldsymbol{\Sigma}, \cdot)$, ce qui implique que $\boldsymbol{Z}^\top \mathbf{A}\boldsymbol{Z} = \mathbf{c}^\top \boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}\mathbf{c} \sim \sigma_{\mathbf{c}}^2 \chi_r^2$. Pour la suffisance, on a $\boldsymbol{Z}^\top \mathbf{A}\boldsymbol{Z}/\sigma_{\mathbf{c}}^2 \sim \chi_r^2$ et $\boldsymbol{Z} \sim \mathcal{N}(\cdot, \sigma_{\mathbf{c}}^2 \mathbf{I}_p)$. On a $\mathbf{A}$ est idempotente de rang $r$ que $\mathbf{A} = \sum_{i=1}^r \mathbf{a}_i \mathbf{a}_i^\top$ avec $\mathbf{a}_i$ un vecteur propre orthonormé de $\mathbf{A}$. Ainsi,

$$\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X} = \sum_{i=1}^r \boldsymbol{X}^\top \mathbf{a}_i \mathbf{a}_i^\top \boldsymbol{X} = \sum_{i=1}^r \boldsymbol{U}_i \boldsymbol{U}_i^\top.$$

Puisque les $\mathbf{a}_i$ sont orthonormaux, $\boldsymbol{Y}\boldsymbol{U}_i \sim \mathcal{N}_p(\mathbf{a}_i^\top \boldsymbol{M}, \boldsymbol{\Sigma})$ sont indépendants, ce qui implique que $\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X}$ suit une loi Wishart $\mathcal{W}_p(r, \boldsymbol{\Sigma}, \cdot)$ qui est centrée si $\mathbf{A}\boldsymbol{M} = \mathbf{0}$.

4. Nécessité de la condition : $\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X} \sim \mathcal{W}_p(r, \boldsymbol{\Sigma}, \cdot)$ et $\boldsymbol{X}^\top \mathbf{B}\boldsymbol{X} \sim \mathcal{W}_p(s, \boldsymbol{\Sigma}, \cdot)$ sont indépendantes. Alors $\boldsymbol{Z}^\top \mathbf{A}\boldsymbol{Z}$ et $\boldsymbol{Z}^\top \mathbf{B}\boldsymbol{Z}$ sont indépendantes (fonctions de deux lois indépendantes). Par point 3, on a $\boldsymbol{Z}^\top \mathbf{A}\boldsymbol{Z}/\sigma_{\mathbf{c}}^2 \sim \chi_r^2$ et $\boldsymbol{Z}^\top \mathbf{B}\boldsymbol{Z}/\sigma_{\mathbf{c}}^2 \sim \chi_s^2$ indépendants pour $\mathbf{c}$ quelconque, donc $\mathbf{AB} = \mathbf{0}$.

Suffisance de la condition : par point 3, $\mathbf{A}$ et $\mathbf{B}$ sont idempotentes de rang $r$ et $s$, respectivement. On a donc $\mathbf{A} = \sum_{i=1}^r \mathbf{a}_i \mathbf{a}_i^\top$ et $\mathbf{B} = \sum_{j=1}^s \mathbf{b}_j$ où $\mathbf{a}_i$ et $\mathbf{b}_j$ sont des vecteurs propres. De plus, $\mathbf{a}_i^\top \mathbf{b}_j = \mathbf{a}_i^\top \mathbf{A}^\top \mathbf{b}_j = \mathbf{0}$ pour tout $i, j$ par hypothèse. On a donc $\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X} = \sum \boldsymbol{U}_i^\top \boldsymbol{U}_i$ avec $\boldsymbol{X}^\top \mathbf{a}_i$ et $\boldsymbol{X}^\top \mathbf{B}\boldsymbol{X} = \sum \boldsymbol{V}_j \boldsymbol{V}_j^\top$ avec $\boldsymbol{V}_j = \boldsymbol{X}^\top \mathbf{b}_j$ indépendants.

5. Pour la nécessité de la condition, on suppose que $\boldsymbol{X}^\top \mathbf{B} \sim \mathcal{N}_p(\cdot, \cdot)$ et $\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X} \sim \mathcal{W}_p(n, \boldsymbol{\Sigma}, \cdot)$ sont indépendants. En utilisant le même raisonnement que dans la preuve précédente, on a $\boldsymbol{Z}^\top \mathbf{B} = \mathbf{c}^\top \boldsymbol{X}^\top \mathbf{c} \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\mathbf{c}}^2 \mathbf{B}^\top \mathbf{B}\right)$ et $\boldsymbol{Z}^\top \mathbf{A}\boldsymbol{Z} \sim \sigma_{\mathbf{c}}^2 \chi_r^2$. De plus, $\boldsymbol{Z}^\top \mathbf{B}$ et $\boldsymbol{Z}^\top \mathbf{A}\boldsymbol{Z}$ sont indépendants comme fonctions de vecteurs aléatoires indépendants.

Pour la suffisance, on a $\boldsymbol{Z}^\top \mathbf{B} \sim \mathcal{N}_p(\cdot, \cdot)$ et $\boldsymbol{Z}^\top \mathbf{A}\boldsymbol{Z} \sim \mathcal{W}_p(n, \boldsymbol{\Sigma}, \cdot)$ sont indépendants pour $\mathbf{c}$ quelconque implique $\boldsymbol{X}^\top \mathbf{B} \sim \mathcal{N}_p(\cdot, \cdot)$ et $\boldsymbol{X}^\top \mathbf{A}\boldsymbol{X} \sim \mathcal{W}_p(n, \boldsymbol{\Sigma}, \cdot)$. Comme $\boldsymbol{Z} \sim \mathcal{N}(\cdot, \sigma_{\mathbf{c}}^2 \boldsymbol{I})$ on a $\boldsymbol{Z}^\top \mathbf{B} \sim \mathcal{N}(\cdot, \sigma_{\mathbf{c}}^2 \mathbf{B}^\top \mathbf{B})$ et $\boldsymbol{Z}^\top \mathbf{B}\mathbf{B}^\top \boldsymbol{Z}/(\mathbf{B}^\top \mathbf{B} \sim \sigma_{\mathbf{c}}^2 \chi_1^2$ avec $\mathbf{B}\mathbf{B}^\top \mathbf{A} = 0$. Multiplier par $\mathbf{B}^\top$, on obtient que $\mathbf{b}^\top \mathbf{A} = \mathbf{0}$ puisque $\mathbf{B}^\top \mathbf{B}$ est constante. $\mathbf{B}$ est perpendiculaire à tous les $\mathbf{A}$ implique que $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_r$ tels que définis précédemment et $\boldsymbol{X}^\top \mathbf{B}$ sont indépendants.
6. Par induction. Le cas $p = 1$ est vrai car $W \sim \sigma^2 \chi_n^2$. Supposons que l'énoncé est vrai pour le cas $p \in \mathbb{N}$ et que $\boldsymbol{W} \sim \mathcal{W}_{p+1}(n, \boldsymbol{\Sigma}, \cdot)$. On écrit $\boldsymbol{W}$ comme

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{11} & \boldsymbol{W}_{12} \\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}_{11} & \mathbf{0}_p \\ \boldsymbol{W}_{21} & 1 \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & \boldsymbol{W}_{11}^{-1} \boldsymbol{W}_{12} \\ \mathbf{0}_p^\top & \boldsymbol{W}_{22} - \boldsymbol{W}_{21} \boldsymbol{W}_{11}^{-1} \boldsymbol{W}_{12} \end{pmatrix}$$

pour $\boldsymbol{W}_{22}$ une matrice $1 \times 1$ et par hypothèse $|\boldsymbol{W}_{11}| \sim |\boldsymbol{\Sigma}_{11}| \chi_n^2 \cdots \chi_{n-p+1}^2$. Or $\boldsymbol{W}_{22.1} = \boldsymbol{W}_{22} - \boldsymbol{W}_{21} \boldsymbol{W}_{11}^{-1} \boldsymbol{W}_{12} \sim \mathcal{W}_1(n - p, \boldsymbol{\Sigma}_{22.1})$ et indépendant de $(\boldsymbol{W}_{11}, \boldsymbol{W}_{12})$. La décomposition $|\boldsymbol{W}| = |\boldsymbol{W}_{11}||\boldsymbol{W}_{22.1}|$ et $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11}||\boldsymbol{\Sigma}_{22.1}|$ assure le résultat.

■

**Définition 2.15 (Loi $\mathcal{T}^2$ de Hotelling)**
On fait le parallèle avec la loi normale univariée : si $X \sim \mathcal{N}(\mu, \sigma^2)$ et $W \sim \sigma^2 \chi_n^2$ sont des variables aléatoires indépendantes, alors

$$T = \frac{\frac{X - \mu}{\sigma}}{\sqrt{\frac{W}{n\sigma^2}}} \sim \mathcal{T}_n,$$

soit une loi de Student-$t$. En général, on a la statistique $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$ pour un

échantillon, ce qui implique que

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)}{\sqrt{\frac{S^2(n-1)}{\sigma^2(n-1)}}} \sim \mathcal{T}_{n-1}$$

et $T^2 = n(X-\mu)^2/W$ a une loi qui est proportionnelle à une loi de Fisher $\mathcal{F}(1,n)$.

La loi $\mathcal{T}^2$ de Hotelling est une généralisation multivariée de cette statistique.

$$T^2 = n(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{W}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \mathcal{T}^2_{p,n}$$

avec $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ et $\boldsymbol{W} \sim \mathcal{W}_p(n, \boldsymbol{\Sigma})$ indépendants. On dénote par $\mathcal{T}^2_{p,n}$ la distribution.

**Proposition 2.16**
La loi de Hotelling $T^2 \sim \mathcal{T}^2_{p,n}$ est équivalente à une loi de Fisher, soit

$$\frac{n-p+1}{p}\frac{T^2}{n} \sim \mathcal{F}_{p,n-p+1}$$

**Preuve** Esquisse : poser $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{\mu}$, ainsi

$$\frac{T^2}{n} = \boldsymbol{Y}^\top \boldsymbol{W}^{-1}\boldsymbol{Y} = \frac{\boldsymbol{Y}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}}{\frac{\boldsymbol{Y}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}}{\boldsymbol{Y}^\top\boldsymbol{W}^{-1}\boldsymbol{Y}}} \equiv \frac{G}{H}$$

où $G$ et $H$ correspondent au numérateur et au dénominateur, respectivement. Posons maintenant $\boldsymbol{Y} = \mathbf{a}$ fixé. La loi conditionnelle de $H$ étant donné $\boldsymbol{Y} = \mathbf{a}$ est $\chi^2_{n-p+1}$ par la propriété 7 de la loi de Wishart et donc ne dépend pas de $\boldsymbol{Y}$ (car la distribution ne dépend pas de $\mathbf{a}$). Ainsi $H$ est indépendant de $\boldsymbol{Y}$ et donc de $G$ (car $\boldsymbol{\Sigma}$ est connu). La forme quadratique $G = \boldsymbol{Y}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$ suit une loi $\chi^2_p$ et le ratio $G/H$ est le ratio de lois $\chi^2_p$ sur $\chi^2_{n-p+1}$. On obtient donc $(n-p+1)G/(pH) \sim \mathcal{F}_{p,n-p+1}$. ■

**Remarque (Lois elliptiques)**
On peut généraliser les résultats obtenus avec la loi normale à une classe de vecteurs aléatoires dont la densité est de la forme générale $f(\boldsymbol{x}) \propto g((\boldsymbol{x}-\mu)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\mu))$. Les principaux exemples sont la loi multinormale et la loi de Student-$t$.

**Définition 2.17 (Lois elliptiques)**
$\boldsymbol{X}$ suit une loi elliptique s'il admet la représentation stochastique $\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + R\mathbf{A}\boldsymbol{\Theta}$, où

- $\boldsymbol{\mu}$ est un vecteur de localisation,
- $R$ est une variable aléatoire positive,
- $\mathbf{A}\mathbf{A}^\top$ est la décomposition de Cholesky de la matrice de variance $\boldsymbol{\Sigma}$ et
- $\boldsymbol{\Theta} \sim \mathcal{S}_d$ est une variable uniforme sur la $d-1$-sphère.

Les variables elliptiques sont dotées de symétrie radiale, qui associe la fonction de répartition centrée à sa fonction de survie. La covariance nulle entre deux marges implique leur indépendance seulement si la loi est multinormale, d'où sa prépondérance dans les modèles graphiques. Notez que selon la forme du générateur $g$, il est possible que certains moments de la distribution n'existent pas. On peut dériver des propriétés à partir de la forme quadratique $R^2 \stackrel{\mathrm{d}}{=} (\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$.

Tableau 1 – Quelques lois elliptiques

| Copule | $F_{R^2}$ | $g(t)$ |
|---|---|---|
| Gaussienne | $\chi^2(d)$ | $c_d e^{-\frac{1}{2}u}$ |
| Student-$t$ | $\mathcal{F}(d,\nu)/d$ | $c_d(1+t/\nu)^{-(d+\nu)/2}$ |
| II$^{\text{e}}$ fn de Pearson | $\mathcal{B}(d/2,\nu+1)$ | $c_d(1-u)^\nu \mathbf{1}_{\nu>-1,-1\leq u\leq 1}$ |
| VII$^{\text{e}}$ fn de Pearson | $m\cdot\mathcal{B}_{\text{II}}(d/2,b-d/2)$ | $c_d(1+t/m)^{-b}\mathbf{1}_{b>d/2,m>0}$ |
| Kotz symétrique | $\Gamma(a+d/2-1,\nu)$ | $c_d u^{b-1}e^{-\nu u}\mathbf{1}_{\nu>0,2a-d>2}$ |
| Laplace | | $c_d e^{-\lvert u\rvert}$ |
| Logistique | | $c_d e^{-u}/(1+e^{-u})^2$ |

## 2.4 Vérification de la normalité multivariée

Cette dernière implique que les marges sont normales. On se penche d'abord sur les tests de normalité univariée.

### 2.4.1. Normalité univariée

Il existe des méthodes informelles (diagrammes quantiles-quantiles ou diagrammes probabilité-probabilité) et des tests formels. Au lieu d'effectuer un test formel d'adéquation, on peut procéder à une vérification informelle à l'aide de graphiques appelés diagrammes Q–Q (quantiles–quantiles) et P–P (probabilités–probabilités). Posons $F$ la fonction de répartition correspondant à la density $f$

**Définition 2.18 (Diagramme quantile–quantile)**
Un graphique des rangs des observations, $x_{[i]}$ versus $a_i$ avec $a_i = F^{-1}(p_i)$ et $p_i = (i-c)/(n-2c+1)$, un estimé empirique de la probabilité cumulative des observations. Typiquement, on choisit $c=0$ (pseudo-observations) ou $c=1/3$ qui donne une fonction empirique approximativement sans biais médian, puisque alors $p_i \approx \mathsf{med}(F(x_{[i]}))$. C'est un paramètre tel que $0\leq c\leq 1$, et $F^{-1}$ représente la fonction inverse de $F$. Si la représentation suit une droite, les données sont susceptibles de provenir de $F$. Noter que le diagramme Q–Q magnifie les observations extrêmes. L'abscisse peut être standardisé (dans le cas d'un diagramme Q–Q normal, on parle alors de l'échelle $\mathcal{N}(0,1)$) ou pas.

**Définition 2.19 (Diagramme probabilités–probabilités)**
Une représentation graphique de $z_{[i]} = F((x_{[i]}-\widehat{\mu})/\widehat{\sigma})$ versus $p_i$ tel que défini ci-

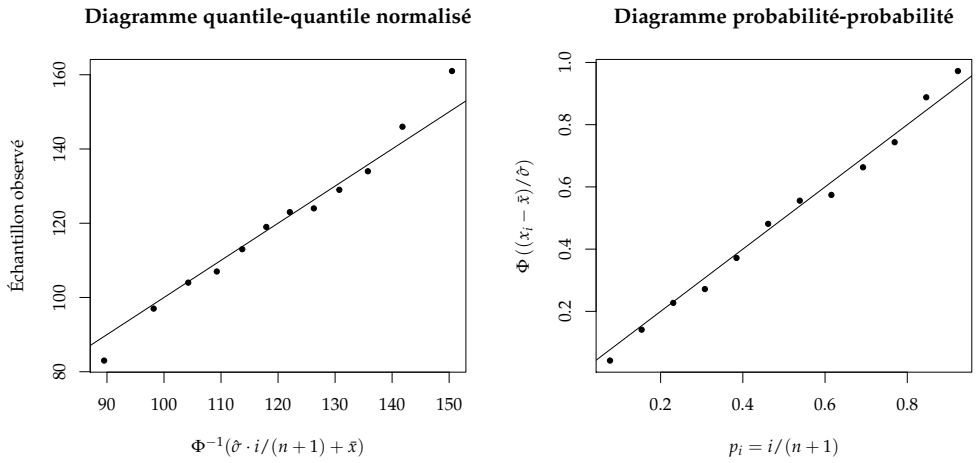
dessus. Si la représentation est une droite à $45^{\text{deg}}$, on peut raisonnablement penser que les données proviennent de $F$.

**Exemple 2.1**

On considère un jeu de données du gain de poids (masse) de femelles rats sous un régime à forte teneur protéique.

Tableau 2 – Données de l'expérience et positions des points des diagrammes Q–Q et P–P normaux standardisés

| $i$ | $x_i$ | $p_i = \frac{i}{n+1}$ | $\Phi^{-1}(p_i)\widehat{\sigma} + \widehat{\mu}$ | $\Phi\left(\frac{x_i - \widehat{\mu}}{\widehat{\sigma}}\right)$ |
|---|---|---|---|---|
| 1 | 83 | 0.077 | 89.50 | 0.0418 |
| 2 | 97 | 0.154 | 98.18 | 0.1141 |
| 3 | 104 | 0.231 | 104.25 | 0.2272 |
| 4 | 107 | 0.308 | 109.25 | 0.2717 |
| 5 | 113 | 0.385 | 113.73 | 0.3717 |
| 6 | 119 | 0.462 | 117.93 | 0.4813 |
| 7 | 123 | 0.538 | 122.07 | 0.5558 |
| 8 | 124 | 0.615 | 126.27 | 0.5742 |
| 9 | 129 | 0.692 | 130.75 | 0.6630 |
| 10 | 134 | 0.769 | 135.75 | 0.7436 |
| 11 | 146 | 0.846 | 141.82 | 0.8879 |
| 12 | 161 | 0.923 | 150.50 | 0.9724 |



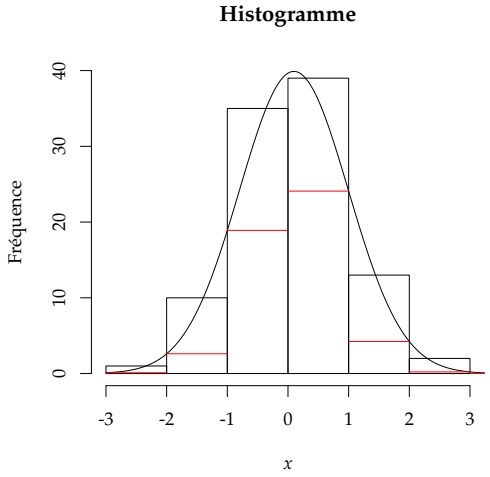

**Proposition 2.20 (Test du $\chi^2$ d'adéquation)**

Le test d'adéquation du $\chi^2$ est basé sur $\mathscr{H}_0 : X \sim F_{\boldsymbol{\theta}}$. Soit $\mathbf{X}$ un échantillon $n \times 1$ d'observations indépendantes et $\nu$ le nombre de paramètres. On peut partitionner l'espace en $k$ intervalles disjoints $A_1, \ldots, A_k$ ; dénotons par $e_j$ la fréquence théorique de $X$ dans $A_i$ et $o_j$ la fréquence observée. Cela revient à tracer un histogramme dans lequel $o_j$ est

le nombre d'observations dans la classe $j$. On a

$$X^2 = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j} \xrightarrow{d} \chi^2_{k-\nu-1}$$

quand $n \to \infty$. L'approximation est bonne pour $n$ large (au bas mot 50) et $e_j > 5$. Dans




le cas de la loi normale, on doit calculer $\mu$ et $\sigma$, ce qui correspond au cas $\nu = 2$.

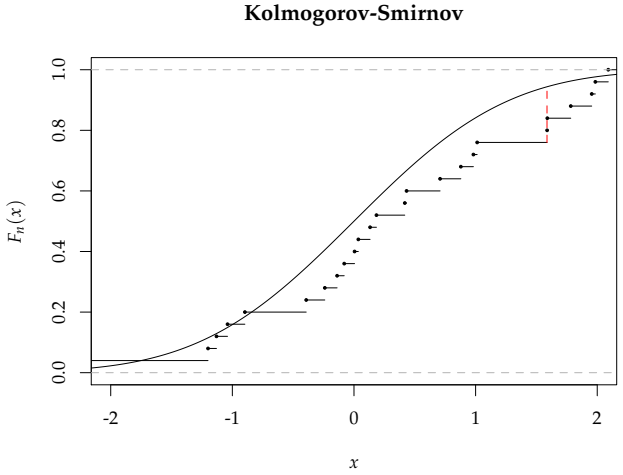
Si les décomptes $e_j$ sont trop petits, il convient de former de plus gros groupes. Si on dessine l'histogramme, porter attention à faire toujours tous les intervalles de longueur égale. Le nombre d'observations pour l'application du test est important, et il est préférable de faire un diagramme Q–Q dans les cas où la taille de l'échantillon est faible.

**Proposition 2.21 (Test d'adéquation de Kolmogorov-Smirnov)**
Comparaison de la fonction de répartition théorique et empirique. On considère la plus grand distance, soit

$$\sup_x |F_n(x) - F(x)| .$$

Il est important de noter que tous les paramètres doivent être spécifiés, soit exactement connus.

Il existe aussi des tests spécifiquement construits pour la loi normale contrairement aux deux précédents qui sont des tests omnibus, comme les statistiques Shapiro–Wilks et Anderson–Darling $A_n^2$. On donne en exemple la première :




**Proposition 2.22 (Test de Shapiro–Wilks)**
Le test de Shapiro–Wilks a comme statistique

$$W = \frac{\sum_{i=1}^{n} a_{n-i+1}^{(n)} \left( X_{[n-i+1]} - X_{[i]} \right)^2}{\sum_{i=1}^{n} (X_{[i]} - \bar{X})^2}$$

où $X_{[i]}$ est la $i$e statistique d'ordre (en ordre croissant). Par exemple, $X_{[1]}$ dénote la plus petite observation. Les valeurs critiques tout comme la loi de $W$, sont tabulées.

## 2.4.2. Normalité multivariée

**Proposition 2.23 (Distance de Mahalanobis)**
On rappelle que $(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2_p$ si $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. On peut utiliser pour $n$ grand l'approximation

$$d_i^2 = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top \boldsymbol{S}^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \dot{\sim} \chi^2_p,$$

soit la distance de Mahalanobis entre $\boldsymbol{X}_i$ et sa moyenne. Le diagramme Q–Q des $d_i^2$ versus les quantiles de la loi $\chi^2$ est un test graphique de multinormalité. On n'a en revanche pas de données indépendantes, et le résultat distributionnel est approximatif.

**Proposition 2.24 (Test de Mardia (1970))**
C'est un test basé sur le coefficient d'asymétrie (*skewness*) et le kurtosis de la loi normale multivariée. On définit les moments centrés pour $\boldsymbol{X}, \boldsymbol{Y}$ indépendants comme

$$\gamma_{1p} = \mathsf{E}\left( \left( \boldsymbol{X} - \boldsymbol{\mu} \right)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \right)^3 \Big)$$

$$\gamma_{2p} = \mathsf{E}\left(\left(\boldsymbol{X}-\boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right)^2\Big).$$

En pratique, on utilise

$$d_i^2 = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top \boldsymbol{S}^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})$$
$$d_{ij} = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\top \boldsymbol{S}^{-1}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})$$

avec les estimateurs

$$\widehat{\gamma}_{1p} = \frac{1}{n^2}\sum_i\sum_j d_{ij}^3$$
$$\widehat{\gamma}_{2p} = \frac{1}{n}\sum_i d_i^4$$

Mardia a montré que, pour $\nu = p(p+1)(p+2)/6$,

$$\frac{n\widehat{\gamma}_{1p}}{6} \dot{\sim} \chi^2_\nu$$
$$\widehat{\gamma}_{2p} \dot{\sim} \mathcal{N}\left(p(p+2), \frac{8p(p+2)}{n}\right)$$

Le logiciel R calcule et rapporte les deux statistiques, et indique le rejet dès qu'une des deux pointe dans ce sens.

**Proposition 2.25 (Test de multinormalité basé sur la distance d'énergie)**
La distance d'énergie entre deux vecteurs aléatoires $\boldsymbol{X}, \boldsymbol{Y}$ de dimension $d$ est définie par

$$\mathcal{E}(\boldsymbol{X},\boldsymbol{Y}) = 2\mathsf{E}|\boldsymbol{X}-\boldsymbol{Y}|_d - \mathsf{E}|\boldsymbol{X}-\boldsymbol{X}'|_d - \mathsf{E}|\boldsymbol{Y}-\boldsymbol{Y}'|_d$$

pour $\boldsymbol{X}'$ ($\boldsymbol{Y}'$) copies iid de $\boldsymbol{X}$ ($\boldsymbol{Y}$) et $|\cdot|_d$ dénote la norme euclidienne. On peut montrer différentes propriétés liées à cette distance :

1. Elle est invariante par rapport aux rotations
2. $\mathcal{E}(\boldsymbol{X},\boldsymbol{Y})^{1/2}$ définit un métrique sur l'ensemble des lois à $d$-variables et $\mathcal{E}(\boldsymbol{X},\boldsymbol{Y})$ est zéro si et seulement si $\boldsymbol{X}$ et $\boldsymbol{Y}$ sont identiquement distribuées.
3. Si $d = 1$ on a $\mathcal{E}(X,Y) = 2\int (F(x)-G(x))^2 dx$ où $F$ et $G$ sont les fonctions de répartition de $X$ et $Y$ respectivement.
4. Pour une loi $\boldsymbol{Z} \sim \mathcal{N}_d(\boldsymbol{0}_d, \mathbf{I}_d)$ on a $\mathsf{E}|\boldsymbol{Z}-\boldsymbol{Z}'|_d = 2\left(\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\right)$.

Pour appliquer ces résultats à un test de normalité on commence par standardiser l'échantillon $\boldsymbol{Y} = \boldsymbol{y}$ via $\boldsymbol{x} = \boldsymbol{S}^{-1/2}(\boldsymbol{y}-\bar{\boldsymbol{y}})$, et on le compare à une variable $\boldsymbol{Z} \sim \mathcal{N}_d(\boldsymbol{0}_d, \mathbf{I}_d)$. On calcule la statistique

$$n\mathcal{E} = n\left(\frac{2}{n}\sum_{i=1}^n \mathsf{E}\,|\boldsymbol{x}_i - \boldsymbol{Z}|_d - 2\left(\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\right) - \left(\frac{1}{n^2}\right)\sum_{i=1}^n\sum_{k=1}^n |\boldsymbol{x}_i - \boldsymbol{x}_k|_d\right)$$

Le terme $\mathsf{E}|\boldsymbol{x}_i - \boldsymbol{Y}|_d$ est calculable puisque $|\boldsymbol{a}-\boldsymbol{Y}|_d^2$ suit une loi $\chi^2$ non centrée, dont le

coefficient de non centralité dépend de $\boldsymbol{a}$ et le nombre de degrés de libertés est aléatoire. En pratique, on utilise une troncation de la série et la distribution de la statistique est obtenue par auto-amorçage (*bootstrap*) paramétrique ou via une approximation asymptotique. Le test est consistent sous toutes les alternatives fixes.

### 2.4.3. Transformations améliorant la normalité

Il existe des techniques basées sur la notion de robustesse. On considère ici une transformation, comme par exemple $\sqrt{x}, \log(x), x^{-1}, \text{logit}(x), \frac{1}{2}\log((1+x)/(1-x))$, etc. La dernière est la transformation de Fisher. On a également le cas de la transformation de Box-Cox, soit

**Proposition 2.26 (Transformation de Box-Cox)**
On considère la transformation

$$X^{(\lambda)} = \begin{cases} \frac{X^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log(X) & \text{si } \lambda = 0. \end{cases}$$

Pour éviter de prendre le logarithme ou une puissance d'un nombre négatif, on peut considérer $X + c$. Pour estimer $\lambda$, on utilise la log-vraisemblance profilée : pour une séquence de valeurs $\lambda$, on obtient les maximums de vraisemblance de la loi normale univariée en utilisant comme données $x_i^\lambda$ pour $i = 1, \ldots, n$ et $\lambda$ fixe. On peut ainsi calculer un intervalle de confiance pour $\lambda$. Pour tester le paramètre, on peut faire un test de rapport de vraisemblance,

$$\mathcal{L} = \frac{\ell_0^*}{\ell_{01}^*} = \frac{\sup_{\boldsymbol{\theta}\in\Omega_0} \ell}{\sup_{\boldsymbol{\theta}\in\Omega_0\sqcup\Omega_1} \ell}$$

et $-2\mathcal{L} \dot{\sim} \chi^2_{s-r}$, où $s = \dim(\Omega_0 \sqcup \Omega_1)$ et $r = \dim(\Omega_0)$.

# Chapitre 3
# Inférences relatives aux paramètres de lois normales

## 3.1 Distribution de $\bar{X}$ et $S$

Soit $\mathbf{X}^\top = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ où $\boldsymbol{X}_i \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Alors,

1. la loi de la moyenne arithmétique est $\bar{\boldsymbol{X}} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$
2. la loi de la variance empirique est $(n-1)\boldsymbol{S} \sim \mathcal{W}_p(n-1, \boldsymbol{\Sigma})$. On a effectivement

$$(n-1)\boldsymbol{S} = \mathbf{X}^\top \left( \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \right) \mathbf{X}$$

où $\mathbf{A} \coloneqq \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ avec $\operatorname{rang}(\mathbf{A}) = n-1$. On a $\mathbf{B} = n^{-1}\mathbf{1}$. Qui est plus, $\bar{\boldsymbol{X}}$ et $\boldsymbol{S}$ sont indépendants. On vérifie que $\mathbf{A} = \mathbf{A}^2$ et que $\mathbf{B}^\top \mathbf{A} = 0$. Ainsi, $(n-1)\boldsymbol{S} \sim \mathcal{W}_p(n-1, \boldsymbol{\Sigma})$ (et indépendant de $\bar{\boldsymbol{X}}$).
3. $T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \mathcal{T}^2_{p,n-1}$. On réécrit cette expression sous la forme

$$\begin{aligned} T^2 &= \frac{n-1}{n-1} n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \\ &= (n-1)(\sqrt{n}\bar{\boldsymbol{X}} - \sqrt{n}\boldsymbol{\mu})^\top ((n-1)\boldsymbol{S})^{-1}(\sqrt{n}\bar{\boldsymbol{X}} - \sqrt{n}\boldsymbol{\mu}) \\ &\sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})^\top \times \mathcal{W}_p(n-1, \boldsymbol{\Sigma}) \times \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}) \end{aligned}$$

suit une loi de Hotelling $\mathcal{T}^2_{p,n-1}$ ou $\frac{(n-1)p}{n-p}\mathcal{F}_{p,n-p}$.

On rappelle le test du rapport des vraisemblances (LRT en anglais) pour un échantillon $\mathbf{X}^\top = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ dépendant du paramètre $\boldsymbol{\theta}$. On veut tester $\mathscr{H}_0 : \boldsymbol{\theta} \in \Omega_0$ contre $\mathscr{H}_1 : \boldsymbol{\theta} \in \Omega_1$. On construit $\lambda(\mathbf{x}) = L_0^*/L_{01}^*$ où $L_0^* = \sup_{\boldsymbol{\theta}\in\Omega_0} L$ et $L_{01}^* = \sup_{\boldsymbol{\theta}\in\Omega_0\cup\Omega_1} L$ pour une vraisemblance $L$.

La région de rejet de $\mathscr{H}_0$ ($\alpha$ étant fixé) est définie par $R : \{\mathbf{x} : \lambda(\mathbf{x}) < c\}$ où $c$ est tel que $\sup_{\boldsymbol{\theta}\in\Omega_0} \mathsf{P}_{\boldsymbol{\theta}}(\mathbf{X} \in R) = \alpha$. On recourt souvent à une approximation.

**Théorème 3.1**
Soit $\Omega_0 \cup \Omega_1$ région de $\mathbb{R}^d$ avec $r = \dim(\Omega_0)$. Alors, sous des conditions suffisantes de régularité et pour chaque $\boldsymbol{\theta} \in \Omega_0$, alors la statistique

$$-2\log(\lambda) \overset{\cdot}{\sim} \chi^2_{d-r}.$$

autour de $\widehat{\boldsymbol{\theta}}$, pour obtenir

## 3.2 Tests relatifs à une moyenne

On s'intéresse à l'hypothèse $\mathscr{H}_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus $\mathscr{H}_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ ou plus généralement $\mathbf{R}\boldsymbol{\mu} = \mathbf{r}$ versus $\mathbf{R}\boldsymbol{\mu} \neq \mathbf{r}$ où l'on assume que $\boldsymbol{\Sigma}$ est connue. Dans ce cas,

$$\ell = C - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{XX}\right) - \frac{n}{2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$$

avec

$$\sup_{\boldsymbol{\mu}=\boldsymbol{\mu}_0} \ell = C - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{XX}\right) - \frac{n}{2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$$

$$\sup_{\boldsymbol{\mu}\in\mathbb{R}^p} \ell = C - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_{XX}\right)$$

pour une constante d'intégration $C \in \mathbb{R}$. Le maximum est atteint pour $\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ et donc

$$-2\log(\lambda) = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \sim \chi^2_p$$

ou plus généralement

$$-2\log(\lambda) = n(\mathbf{R}\bar{\boldsymbol{X}} - \mathbf{r})^\top \left(\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}^\top\right)^{-1}(\mathbf{R}\bar{\boldsymbol{X}} - \mathbf{r}) \sim \chi^2_q$$

où $q$ est le rang de $\mathbf{r}$.

On considère maintenant le cas $\boldsymbol{\Sigma}$ inconnue : la log-vraisemblance est

$$\ell = C - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{n}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\left(n^{-1}\boldsymbol{S}_{XX} + (\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top\right)\right).$$

Sous $\mathscr{H}_0$, on a

$$\sup_{\boldsymbol{\mu}=\boldsymbol{\mu}_0,\boldsymbol{\Sigma}} \ell = C - \frac{n}{2}\log\left(\left|n^{-1}\boldsymbol{S}_{XX} + (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top\right|\right) - \frac{np}{2}$$

avec $\widehat{\boldsymbol{\Sigma}}_0 = n^{-1}\boldsymbol{S}_{XX} + (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top$. Sous l'hypothèse alternative $\mathscr{H}_1$,

$$\sup_{\boldsymbol{\mu}\in\mathbb{R}^p} \ell = C - \frac{n}{2}\log(|\boldsymbol{S}^*|) - \frac{np}{2}$$

avec $\widehat{\boldsymbol{\Sigma}} = n^{-1}\boldsymbol{S}_{XX}$ et $\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$

$$\begin{aligned}
-2\log(\lambda) &= -n\log(|\boldsymbol{S}^*|) + n\log\left|\boldsymbol{S}^* + (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top\right| \\
&= -n\log(|\boldsymbol{S}^*|) + n\log\left|\boldsymbol{S}^*\left(\mathbf{I} + \boldsymbol{S}^{*-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top\right)\right| \\
&= n\log\left|\mathbf{I} + \boldsymbol{S}^{*-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top\right| \\
&= n\log\left(1 + (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{*-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)\right)
\end{aligned}$$

en utilisant le résultat A7.10 de l'annexe, qui dit $|\mathbf{B} + \mathbf{C}\mathbf{C}^\top| = |\mathbf{B}|(1 + \mathbf{C}^\top\mathbf{B}^{-1}\mathbf{C})$. On calcule $d = p + p(p+1)/2$ et $r = p(p+1)/2$. Cette expression plus petite qu'une

constante $c$ ne mène pas au rejet de l'hypothèse nulle $\mathscr{H}_0$ si

$$\begin{aligned} & n \log\left(1 + (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{*-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)\right) < c \\ \Leftrightarrow \quad & (n-1)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{*-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) < c' \end{aligned}$$

et donc

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \sim \mathcal{T}^2_{p,n-1}$$

Pour un test de restriction générale linéaire, on obtient

$$n(\mathbf{R}\bar{\boldsymbol{X}} - \mathbf{r})^\top (\mathbf{R}\boldsymbol{S}\mathbf{R}^\top)^{-1}(\mathbf{R}\bar{\boldsymbol{X}} - \mathbf{r}) \sim \mathcal{T}^2(\operatorname{rang}(\mathbf{R}), n-1).$$

**Remarque (Intervalles simultanés)**
On prend $\mathbf{a} \in \mathbb{R}^p$ et on pose $Y_{\mathbf{a}} = \mathbf{a}^\top \boldsymbol{X} \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ avec $\mathsf{Var}\,(Y_{\mathbf{a}}) = \mathbf{a}^\top \boldsymbol{\Sigma}\mathbf{a}$ et $\mathsf{E}\,(Y_{\mathbf{a}}) = \mathbf{a}^\top \boldsymbol{\mu}$. On utilise la statistique

$$\frac{\bar{y} - \mu_Y}{\sqrt{S_Y^2/n}} \to t_n^2 = n\frac{\mathbf{a}^\top(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \mathbf{a}}{\mathbf{a}^\top \boldsymbol{S}^* \mathbf{a}}$$

pour tout $\mathbf{a}$, et on a un quotient de deux formes quadratiques. On peut donc chercher le maximum $\max_{\mathbf{a}} t_n^2$ et l'on applique le théorème de Cauchy–Schwartz. On rappelle que

$$(\mathbf{z}^\top \mathbf{y})^2 \leq (\mathbf{z}^\top \mathbf{z})(\mathbf{y}^\top \mathbf{y}) \tag{3.1}$$

avec égalité si et seulement si $\mathbf{z} \propto \lambda \mathbf{y}$. On pose ici $\mathbf{z} = \boldsymbol{\Sigma}^{1/2}\mathbf{a}$ et $\mathbf{y} = \boldsymbol{\Sigma}^{-1/2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$. On obtient

$$\max_{\mathbf{a}} t_n^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \boldsymbol{S}^{*-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$$

Si l'on fait l'intervalle de confiance pour toutes les directions possibles et que l'on considère leur enveloppe, on obtient l'ellipse. Dans la direction $X_1$, on a, avec probabilité $(1-\alpha)$, $\bar{X}_1 \pm t_{n-1,1-\alpha/2} S_n/\sqrt{n}$ . En dimension $p$, la probabilité de couvrir le rectangle est $(1-\alpha)^p$ et $\delta = \mathsf{P}\,(\text{rejeter } \mathscr{H}_0) = 1 - (1-\alpha)^p$. Ceci illustre la nécessité de faire du multivariée. La correction de Bonferroni consiste à choisir $\alpha$ tel que $\delta = 0.05$. Ce n'est pas nécessairement la meilleure région (c'est l'ellipse), mais on obtient un risque global de 0.05.

Pour l'intervalle de confiance, on a

$$\max_{\mathbf{a}} = \frac{\mathbf{a}^\top \mathbf{b}}{\mathbf{a}^\top \boldsymbol{\Sigma}\mathbf{a}} = \mathbf{b}^\top \boldsymbol{\Sigma}\mathbf{b}$$

pour $\mathbf{a} = \lambda \boldsymbol{\Sigma}^{-1}\mathbf{b}$. L'intervalle de Bonferroni est dérivé en utilisant pour $m$ test un seuil

de signifiance de $\alpha_i = \alpha/m$, afin que le risque global soit de valeur $\alpha$ au maximum.

**Exemple 3.1**
On prend $\mathbf{a}_1 = (1,\ldots,0)^\top, \mathbf{a}_2 = (0,1,0,\ldots,0)^\top$, etc. pour calculer les intervalles de confiance successivement. On a

$$\bar{X}_i \pm \sqrt{\frac{S_i^2}{n}} t_{n-1,1-\alpha/2}$$

un intervalle de confiance pour $\mu_i$. La probabilité de couvrir la vraie valeur est $1-\alpha$. En supposant l'indépendance entre les composantes, on obtient que la probabilité que les intervalles couvrent les $\mu_i$ est $1-(1-\alpha)^p$. Prenons $p = 6$ en exemple, on a pour $\alpha = 0.05$ que $\delta = 0.26$.

**Exemple 3.2**
Prenons un exemple numérique concret : on considère un échantillon de $n = 10$ observations tirées d'une loi bivariée normale de moyenne arithmétique $\bar{\boldsymbol{X}} = (2.5, 3)^\top$ et de matrice de variance-covariance empirique $\boldsymbol{S} = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. On s'intéresse aux intervalles de confiance pour $\boldsymbol{\mu}$.

Considérons d'abord l'ellipse correspondant aux intervalles de confiance simultanée. La matrice de précision est $\boldsymbol{S}^{-1} = \left(\begin{smallmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{smallmatrix}\right)$. Dans ce cas,

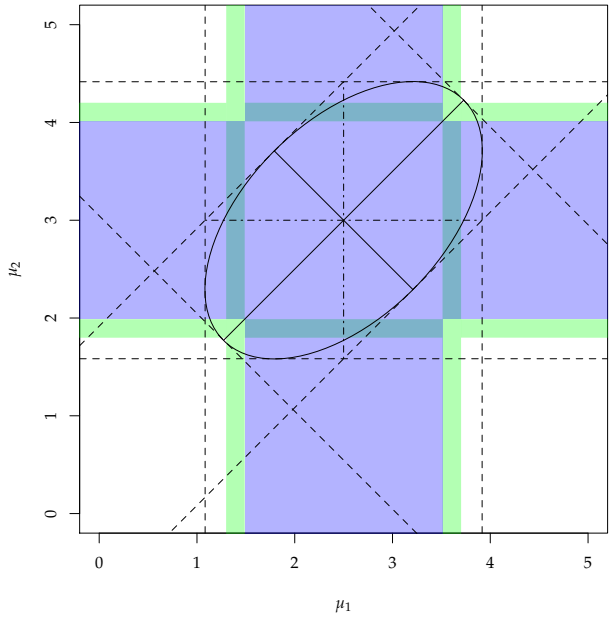
$$(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \leq 1.0035 = \left(\frac{1}{10}\right)\left(\frac{18}{8}\right)\mathcal{F}_{2,8}(0.95) = \frac{1}{n}\frac{(n-1)p}{(n-p)}\mathcal{F}_{p,n-p}(1-\alpha).$$

On procède au calcul des valeurs propres et vecteurs propres de $\boldsymbol{S}^{-1}$. Le polynôme caractéristique est $(2/3-\lambda)^2 - 1/9$ et en égalant ce dernier à zéro, on obtient $\lambda_1 = 1, \lambda_2 = 1/3$. Les vecteurs propres sont donc $\boldsymbol{v}_1 = (1,-1)^\top/\sqrt{2}$ et $\boldsymbol{v}_2 = (1,1)^\top/\sqrt{2}$ et l'équation de l'ellipse est $x_1^2 + x_2^2/3 = 1.0035$. On peut isoler des droites tangentes le long de cet ellipse, avec $b^2 = \mathbf{a}^\top \boldsymbol{S}\mathbf{a} \times \frac{(n-1)p}{n(n-p)}\mathcal{F}_{p,n-p}(1-\alpha)$.

$$\begin{aligned}
\mathbf{a} = (1,1)^\top: &\quad \mathbf{a}^\top\bar{\boldsymbol{X}} = \frac{10}{2}, \quad \mathbf{a}^\top\boldsymbol{S}\mathbf{a} = 6, \quad \mathbf{a}\bar{\boldsymbol{X}} \pm b = \frac{10}{2} \pm \sqrt{\frac{27}{20}4.46} = \frac{10}{2} \pm 2.4538 \\
\mathbf{a} = (0,1)^\top: &\quad \mathbf{a}^\top\bar{\boldsymbol{X}} = 3, \quad \mathbf{a}^\top\boldsymbol{S}\mathbf{a} = 2, \quad \mathbf{a}\bar{\boldsymbol{X}} \pm b = 3 \pm \sqrt{\frac{9}{20}4.46} = 3 \pm 1.4167 \\
\mathbf{a} = (1,0)^\top: &\quad \mathbf{a}^\top\bar{\boldsymbol{X}} = \frac{5}{2}, \quad \mathbf{a}^\top\boldsymbol{S}\mathbf{a} = 2, \quad \mathbf{a}\bar{\boldsymbol{X}} \pm b = \frac{5}{2} \pm \sqrt{\frac{9}{20}4.46} = \frac{5}{2} \pm 1.4167 \\
\mathbf{a} = (1,-1)^\top: &\quad \mathbf{a}^\top\bar{\boldsymbol{X}} = -\frac{1}{2}, \quad \mathbf{a}^\top\boldsymbol{S}\mathbf{a} = 2, \quad \mathbf{a}\bar{\boldsymbol{X}} \pm b = -\frac{1}{2} \pm \sqrt{\frac{9}{20}4.46} = -\frac{1}{2} \pm 1.4167
\end{aligned}$$

ce qui donne $3.046 \leq \mu_1 + \mu_2 \leq 7.954, 1.0833 \leq \mu_1 \leq 3.9167, 1.5833 \leq \mu_2 \leq 4.4167$ et finalement $-1.9167 \leq \mu_1 - \mu_2 \leq 0.9167$. Si on se restreignait à l'analyse univariée, on obtiendrait des intervalles de confiance marginaux. Si $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, on a $\mathbf{a}^\top\bar{\boldsymbol{X}} \sim$

FIGURE 3 – Intervalles de confiance marginaux et simultanés pour le jeu de données factices. Les intervalles marginaux sont en bleu,les intervalles de Bonferroni en vert et les projections de l'ellipse de confiance conjointe en pointillés.




$\mathcal{N}\left(\mathbf{a}^\top \boldsymbol{\mu}, \mathbf{a}^\top \boldsymbol{\Sigma}\mathbf{a}/n\right)$ et donc

$$\sqrt{n}\frac{\mathbf{a}^\top \bar{\boldsymbol{X}} - \mathbf{a}^\top \boldsymbol{\mu}}{\sqrt{\mathbf{a}^\top \boldsymbol{S}\mathbf{a}}} \sim \mathcal{T}_{n-1}$$

et on choisissant $\mathbf{a} = (1,0)^\top$ ou $\mathbf{a} = (0,1)^\top$. On a donc

$$\begin{aligned}
\mu_1 : \sqrt{10}\frac{2.5 - \mu_1}{\sqrt{2}} \leq 1.833 \qquad &\rightarrow \qquad 1.68 \leq \mu_1 \leq 3.32 \\
\mu_2 : \sqrt{10}\frac{3 - \mu_1}{\sqrt{2}} \leq 1.833 \qquad &\rightarrow \qquad 2.18 \leq \mu_2 \leq 3.82,
\end{aligned}$$

pour un niveau de confiance de 95%, comparativement aux intervalles de Bonferroni, qui sont de la forme $\bar{X}_i \pm \mathcal{T}_{n-1}(\alpha/2m)\sqrt{S_i/n}$ et légèrement plus larges, avec cette fois-ci $1.3 \leq \mu_1 \leq 3.7$ et $1.8 \leq \mu_2 \leq 4.2$. Ces deux intervalles sont illustrés par des bandes bleus et vertes, respectivement, tandis que les axes de l'ellipse sont représentés par des lignes pleines et les intervalles simultanés à la frontière de l'ellipse par des lignes pointillées.

## 3.3 Tests relatifs à une variance

Essentiellement, on traitera de 3 hypothèses nulles $\mathscr{H}_0$, à savoir

$$\mathscr{H}_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0, \qquad \mathscr{H}_0 : \boldsymbol{\Sigma} = k\boldsymbol{\Sigma}_0, \qquad \mathscr{H}_0 : \boldsymbol{\Sigma}_{12} = 0.$$

Pour chacune d'elle on part de la log-vraisemblance d'une multinormale, écrite comme

$$\ell = C - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{n}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}^*\right) - \frac{n}{2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu})$$

On doit maximiser cette quantité sous $\mathscr{H}_0$ et sous $\mathscr{H}_0 \sqcup \mathscr{H}_1$. On obtiendra pour chaque maximisation $\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ et donc

$$\begin{aligned}\sup_{\mathscr{H}_0} \ell &= C - \frac{n}{2}\log(|\widehat{\boldsymbol{\Sigma}}_0|) - \frac{n}{2}\operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{S}^*\right)\\ \sup_{\mathscr{H}_0 \sqcup \mathscr{H}_1} \ell &= C - \frac{n}{2}\log(|\widehat{\boldsymbol{\Sigma}}_{01}|) - \frac{n}{2}\operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}_{01}^{-1}\boldsymbol{S}^*\right).\end{aligned}$$

Puisque $\widehat{\boldsymbol{\Sigma}}_{01} = \boldsymbol{S}^*$, on aura

$$-2\log(\lambda) = npA_0 - np\log(G_0) - np,$$

où $A_0$ est la moyenne arithmétique des valeurs propres de $\widehat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{S}^*$ et $G_0$ la moyenne géométrique de ces mêmes valeurs propres. On peut maintenant calculer dans chacun des cas $\widehat{\boldsymbol{\Sigma}}_0$.

**Exemple 3.3**

Sous l'hypothèse nulle $\mathscr{H}_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ versus $\mathscr{H}_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$ on a $\widehat{\boldsymbol{\Sigma}}_0 = \boldsymbol{\Sigma}_0$ ce qui aura comme conséquence que $A_0$ et $G_0$ sont les moyennes arithmétiques et géométriques de $\boldsymbol{\Sigma}_0^{-1}\boldsymbol{S}^*$ et la statistique suit une loi $\chi^2$ avec $\frac{1}{2}p(p+1)$ degrés de liberté.

**Exemple 3.4**

Sous l'hypothèse nulle $\mathscr{H}_0 : \boldsymbol{\Sigma} = k\boldsymbol{\Sigma}_0$ versus $\mathscr{H}_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$, on a que $\widehat{k} = \operatorname{tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{S}^*)/p$ et on obtient finalement

$$-2\log(\lambda) = np\log\left(\frac{A_0}{G_0}\right) \dot{\sim} \chi^2_{\frac{1}{2}p(p+1)-1}.$$

où $A_0$ et $G_0$ sont à nouveau les moyennes arithmétiques et géométriques de $\boldsymbol{\Sigma}_0^{-1}\boldsymbol{S}^*$.

Dans le cas où la matrice $\boldsymbol{\Sigma}_0 = \mathbf{I}_p$ on fait ce qu'on appelle un test de sphéricité. On peut aussi généraliser ce test et tester $\boldsymbol{R} = \mathbf{I}_p$ où $\boldsymbol{R}$ est la matrice des corrélations. Dans ce cas, on aura $A_0 = p/p = 1$ puisque $\operatorname{diag}(\boldsymbol{R}) = \mathbf{1}_p$, tandis que $G_0 = \left(\prod_i \lambda_i^*\right)^{1/p}$. La statistique devient alors

$$-2\log(\lambda) = -n\log(|\widehat{\boldsymbol{R}}|) \dot{\sim} \chi^2_{\frac{1}{2}p(p+1)-p}$$

et suit approximativement une loi khi-deux avec $\frac{1}{2}p(p-1)$ degrés de liberté. Box (1949) a démontré que l'approximation est meilleure si on remplace $n$ par $n - 1 - \frac{1}{6}(2p+5)$.

**Exemple 3.5 (Test pour la nullité d'une sous-matrice de variance)**
La situation suivante resurgira lors de l'analyse canonique. On veut tester l'hypothèse $\mathscr{H} : \boldsymbol{\Sigma}_{12} = 0$, où $\boldsymbol{\mu}$ est inconnue, contre l'alternative $\boldsymbol{\Sigma}_{12} \neq 0$. Sous $\mathscr{H}_0$, la vraisemblance peut se scinder en deux et on aura

$$\widehat{\boldsymbol{\Sigma}}_0 = \begin{pmatrix} \boldsymbol{S}_{11}^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{22}^* \end{pmatrix}, \qquad \widehat{\boldsymbol{\mu}} = \begin{pmatrix} \widehat{\boldsymbol{\mu}}_1 \\ \widehat{\boldsymbol{\mu}}_2 \end{pmatrix} = \bar{\boldsymbol{X}}.$$

On obtiendra ainsi

$$\widehat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{S}^* = \begin{pmatrix} \boldsymbol{S}_{11}^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S}_{22}^* \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{S}_{11}^* & \boldsymbol{S}_{12}^* \\ \boldsymbol{S}_{21}^* & \boldsymbol{S}_{22}^* \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \boldsymbol{S}_{11}^{*\,-1}\boldsymbol{S}_{12}^* \\ \boldsymbol{S}_{22}^{*\,-1}\boldsymbol{S}_{21}^* & \mathbf{I} \end{pmatrix}$$

avec $\operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{S}^*\right) = p$ et

$$\det\left(\widehat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{S}^*\right) = \frac{|\boldsymbol{S}^*|}{|\boldsymbol{S}_{11}^*||\boldsymbol{S}_{22}^*|} = \frac{|\boldsymbol{S}_{22}^* - \boldsymbol{S}_{21}^*\boldsymbol{S}^{*\,-1}_{11}\boldsymbol{S}_{12}^*|}{|\boldsymbol{S}_{22}^*|}$$

et la statistique du rapport de vraisemblance est

$$\begin{aligned} -2\log(\lambda) &= n(p_1+p_2) - n\log\left(\frac{|\boldsymbol{S}^*|}{|\boldsymbol{\Sigma}_0|}\right) - n(p_1+p_2) \\ &= -n\log\left|\mathbf{I} - \boldsymbol{S}_{22}^{*\,-1}\boldsymbol{S}_{21}^*\boldsymbol{S}_{11}^{*\,-1}\boldsymbol{S}_{12}^*\right| \\ &= -n\log\left(\prod_{i=1}^{k}(1-\widehat{\lambda}_i)\right) \\ &\dot{\sim} \chi^2_m \end{aligned}$$

avec

$$m = \frac{1}{2}(p_1+p_2)(p_1+p_2+1) - \frac{1}{2}p_1(p_1+1) - \frac{1}{2}p_2(p_2+1) = p_1p_2$$

et $1-\widehat{\lambda}_i$ est une valeur propre de $\mathbf{I} - \boldsymbol{S}_{22}^{*\,-1}\boldsymbol{S}_{21}^*\boldsymbol{S}_{11}^{*\,-1}\boldsymbol{S}_{12}^*$ ou $\mathbf{I} - \boldsymbol{S}_{11}^{*\,-1}\boldsymbol{S}_{12}^*\boldsymbol{S}_{22}^{*\,-1}\boldsymbol{S}_{21}^*$. On a cette décomposition en fonction de $\boldsymbol{S}_{11}^*$ ou de $\boldsymbol{S}_{22}^*$, ce qui fait que $k = \min(p_1, p_2)$. Si on travaille plutôt avec la matrice $\mathbf{R} = \mathbf{D}^{-1/2}\boldsymbol{S}^*\mathbf{D}^{-1/2}$ on constate après quelques calculs que la statistique ne change pas. Idem si l'on utilise l'estimateur sans biais de la variance, $\boldsymbol{S}$, plutôt que le maximum de vraisemblance.

On peut montrer que le ratio du déterminant de deux lois Wishart est distribuée exactement comme $\Lambda(p_2, n-1-p_1, p_1)$, qui est une loi de Wilks $\Lambda(a,b,c)$. Cette dernière est liée à la loi $\mathcal{F}$ de Fisher pour certaines valeurs des paramètres. Bartlett a proposé

une approximation qui est

$$-\left(b-\frac{1}{2}(a-c+1)\right)\log(\Lambda(a,b,c))\sim\chi^2_{ac}.$$

Ici, $ac=p_1p_2=m$. Sous $\mathscr{H}_0:\boldsymbol{\Sigma}_{12}=\mathbf{0}$, cette approximation donne

$$-\left(n-\frac{1}{2}(p_1+p_2+3)\right)\sum_{i=1}^{k}\log\left(1-\widehat{\lambda}_i\right)\sim\chi^2_{p_1p_2}.$$

Ce test pourrait servir par exemple à vérifier une possible association entre types d'examens (livre ouvert ou fermé) dans le jeu de données `examens`.

## 3.4 Comparaison de deux ou plusieurs populations normales

On considère le test d'égalité des moyennes de populations. Cette situation est une généralisation de l'ANOVA (analyse de variance), appelée MANOVA.

Dans le cas univarié, on compare les moyennes des deux populations $\mu_1-\mu_2=d_0$, à l'aide du test de Student

$$\frac{\bar{X}^{(1)}-\bar{X}^{(2)}-d_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim\mathcal{T}_\nu$$

où $\bar{X}^{(k)}$ et $S_k$ sont respectivement la moyenne et la variance du groupe $k\in\{1,2\}$, et

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

est la matrice de variance pour les observations groupées.

**Proposition 3.2 (Analyse de variance multivariée : égalité de moyennes)**
Soit un échantillon de $n$ observations issues de $K$ populations multinormales. Les données tirées de la population $k$ seront dénotées $\boldsymbol{X}^{(k)}$ et leurs moyennes et variance empiriques par $\bar{\boldsymbol{X}}^{(k)}$ et $\boldsymbol{S}_k$. On teste l'hypothèse $\mathscr{H}_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\cdots=\boldsymbol{\mu}_K$ en supposant $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_K$ inconnue, contre l'alternative où il y a au moins une différence dans les moyennes. On procède au rapport des vraisemblances

$$\ell=C-\frac{1}{2}\sum_{k=1}^{K}\left(n_k\log|\boldsymbol{\Sigma}|+n_k\,\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{S}_k^*+\boldsymbol{d}_k\boldsymbol{d}_k^\top\right)\right)\right)$$

où $\boldsymbol{d}_k=\bar{\boldsymbol{X}}^{(k)}-\boldsymbol{\mu}_k$, $n=\sum_{k=1}^K n_k$ et

$$\boldsymbol{S}_k^*=\frac{1}{n_k}\sum_{j=1}^{n_k}(\boldsymbol{X}_j^{(k)}-\bar{\boldsymbol{X}}^{(k)})(\boldsymbol{X}_j^{(k)}-\bar{\boldsymbol{X}}^{(k)})^\top.$$

Sous $\mathscr{H}_0$, le maximum de la log-vraisemblance est obtenu pour $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S}_T^*$,[3] et

$$\widehat{\boldsymbol{\mu}}_k = \bar{\boldsymbol{X}} = n^{-1} \sum_{k=1}^{K} \sum_{j=1}^{n_k} \boldsymbol{X}_j^{(k)}.$$

. Sous $\mathscr{H}_0 \cup \mathscr{H}_1$, $\boldsymbol{\mu}_k = \bar{\boldsymbol{X}}^{(k)}$, et l'estimateur conjoint de la variance est $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{K} n_k \boldsymbol{S}_k^* = \boldsymbol{S}_p^*$.[4] On va dénoter $\widehat{\boldsymbol{W}} = \sum_{k=1}^{K} n_k \boldsymbol{S}_k^* = n\boldsymbol{S}_p^*$ la variance intra-groupes, $\widehat{\boldsymbol{T}} = n\boldsymbol{S}_T^*$ et

$$\widehat{\boldsymbol{B}} = \sum_{k=1}^{K} n_k (\bar{\boldsymbol{X}}^{(k)} - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}^{(k)} - \bar{\boldsymbol{X}})^\top,$$

la variance inter-groupe. On peut décomposer la variance totale via $\widehat{\boldsymbol{T}} = \widehat{\boldsymbol{W}} + \widehat{\boldsymbol{B}}$. Cette dernière est facile à dériver si l'on note qu'en additionnant et en soustrayant $\bar{\boldsymbol{X}}^{(k)}$ dans chaque terme,

$$\begin{aligned}
\widehat{\boldsymbol{T}} &= \sum_{k=1}^{K} \sum_{j=1}^{n_k} (\boldsymbol{X}_j^{(k)} - \bar{\boldsymbol{X}})(\boldsymbol{X}_j^{(k)} - \bar{\boldsymbol{X}})^\top \\
&= \sum_{k=1}^{K} \sum_{j=1}^{n_k} (\boldsymbol{X}_j^{(k)} - \bar{\boldsymbol{X}}^{(k)})(\boldsymbol{X}_j^{(k)} - \bar{\boldsymbol{X}}^{(k)})^\top + \sum_{k=1}^{K} \sum_{j=1}^{n_k} (\bar{\boldsymbol{X}}^{(k)} - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}^{(k)} - \bar{\boldsymbol{X}})^\top \\
&= \sum_{k=1}^{K} \sum_{j=1}^{n_k} (\boldsymbol{X}_j^{(k)} - \bar{\boldsymbol{X}}^{(k)})(\boldsymbol{X}_j^{(k)} - \bar{\boldsymbol{X}}^{(k)})^\top + \sum_{k=1}^{K} n_k (\bar{\boldsymbol{X}}^{(k)} - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}^{(k)} - \bar{\boldsymbol{X}})^\top \\
&= \widehat{\boldsymbol{W}} + \widehat{\boldsymbol{B}}.
\end{aligned}$$

Les termes croisés s'annulent puisque

$$\sum_{k=1}^{K} \left( \sum_{j=1}^{n_k} (\boldsymbol{X}_j^{(k)} - \bar{\boldsymbol{X}}^{(k)}) \right) (\bar{\boldsymbol{X}}^{(k)} - \bar{\boldsymbol{X}})^\top = 0.$$

On obtient la statistique

$$-2\log(\lambda) = 2\left( -\frac{n}{2} \log\left(|\widehat{\boldsymbol{W}}|\right) + \frac{n}{2} \log\left(|\widehat{\boldsymbol{T}}|\right) \right)$$

et donc

$$\lambda^{2/n} = \left| \widehat{\boldsymbol{T}}^{-1} \widehat{\boldsymbol{W}} \right| = \frac{|\widehat{\boldsymbol{W}}|}{|\widehat{\boldsymbol{B}} + \widehat{\boldsymbol{W}}|} = \left| \mathbf{I} + \widehat{\boldsymbol{W}}^{-1} \widehat{\boldsymbol{B}} \right|^{-1} \sim \Lambda(p, n-K, K-1)$$

laquelle loi de Wilks est la distribution exacte si les hypothèses distributionnelles de départ sont satisfaites (variance conjointe et multinormalité), et suit asymptotiquement

3. Le sous-indice $T$ est utilisé pour dénoter la variance totale, car on peut considérer les observations comme provenant du même échantillon.

4. L'indice $p$ est en référence au terme anglais *pooled*, tandis que $\boldsymbol{W}$ dénote *within-group variance* et $\boldsymbol{B}$ *between-group variance*.

une loi khi-deux. L'approximation de Bartlett est donnée par

$$-\left(-n-\frac{1}{2}(K+p+2)\right)\log(\Lambda)\,\dot{\sim}\,\chi^2_{p(K-1)}.$$

**Exemple 3.6**
Le cas $K=2$ permet d'obtenir une expression parlante. On calcule

$$\widehat{\boldsymbol{B}}=n_1(\bar{\boldsymbol{X}}^{(1)}-\bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}^{(1)}-\bar{\boldsymbol{X}})^\top+n_2(\bar{\boldsymbol{X}}^{(2)}-\bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}^{(2)}-\bar{\boldsymbol{X}})^\top$$

et $\bar{\boldsymbol{X}}=(n_1\bar{\boldsymbol{X}}^{(1)}+n_2\bar{\boldsymbol{X}}^{(2)})/(n_1+n_2)$. En mettant les termes en évidence,

$$\widehat{\boldsymbol{B}}=\frac{n_1n_2}{n_1+n_2}(\bar{\boldsymbol{X}}^{(1)}-\bar{\boldsymbol{X}}^{(2)})(\bar{\boldsymbol{X}}^{(1)}-\bar{\boldsymbol{X}}^{(2)})^\top$$

On procède maintenant au calcul de la statistique ; soit

$$\begin{aligned}\left|\mathbf{I}+\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{B}}\right|&=\left|\mathbf{I}+\frac{n_1n_2}{n_1+n_2}\widehat{\boldsymbol{W}}^{-1}(\bar{\boldsymbol{X}}^{(1)}-\bar{\boldsymbol{X}}^{(2)})(\bar{\boldsymbol{X}}^{(1)}-\bar{\boldsymbol{X}}^{(2)})^\top\right|\\&=1+\frac{n_1n_2}{n_1+n_2}(\bar{\boldsymbol{X}}^{(1)}-\bar{\boldsymbol{X}}^{(2)})^\top\widehat{\boldsymbol{W}}^{-1}(\bar{\boldsymbol{X}}^{(1)}-\bar{\boldsymbol{X}}^{(2)})\end{aligned}$$

en utilisant un des résultats du résumé d'algèbre linéaire.

Dès lors, avec $\boldsymbol{S}_p=(n_1\boldsymbol{S}_1^*+n_2\boldsymbol{S}_2^*)/(n_1+n_2-2)$ on a

$$\frac{n_1n_2}{n_1+n_2}(\bar{\boldsymbol{X}}^{(1)}-\bar{\boldsymbol{X}}^{(2)})^\top\boldsymbol{S}_p^{-1}(\bar{\boldsymbol{X}}^{(1)}-\bar{\boldsymbol{X}}^{(2)})\,\dot{\sim}\,\frac{(n_1+n_2-2)p}{n_1+n_2-p-1}\mathcal{F}_{p,n_1+n_2-p-1}$$

On rappelle que dans le cas d'une seule population, on avait obtenu en supposant $\mathscr{H}_0:\mu=\mu_0$ contre l'alternative $\mathscr{H}_1:\mu\neq\mu_0$, la statistique $n(\bar{X}-\mu_0)^\top S^{-1}(\bar{X}-\mu_0)$.

### 3.4.1. Interactions

On considère une ANOVA à deux voies, exprimée comme un modèle linéaire de la forme

$$\begin{aligned}X_i^{(kg)}&=\mu+\alpha_k+\beta_g+\varepsilon_i^{(kg)}\\X_i^{(kg)}&=\mu+\alpha_k+\beta_g+\alpha_k:\beta_g+\varepsilon_i^{(kg)},\end{aligned}$$

pour $k=1,\dots,K; g=1,\dots,G$ et $i=1,\dots n$ et $\varepsilon_i^{(kg)}\overset{\text{ind}}{\sim}\mathcal{N}(0,\sigma^2)$. Ce modèle est appelé modèle **d'analyse de variance** (ou ANOVA). Le premier modèle implique que l'effet quantitatif des facteurs indexés par $k$ et $g$ est purement additif. À l'inverse, si les effets pour ces facteurs sont plus subtiles, il faudra utiliser un modèle avec interaction, laquelle est représentée par le terme $\alpha_k:\beta_g$.

Par analogie avec le cas précédent, on a dans le cas multivarié les résultats suivants :

**Proposition 3.3 (MANOVA à une voie)**
Soit le modèle

$$\boldsymbol{X}_i^{(k)} = \boldsymbol{\mu} + \boldsymbol{\alpha}_k + \boldsymbol{\varepsilon}_i^{(k)}$$

avec $\boldsymbol{\varepsilon}_i^{(k)} \overset{\text{ind}}{\sim} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ pour $j = 1, \ldots, n_k$ et $k = 1, \ldots, K$. On impose la contrainte $\sum_{k=1}^K n_k \boldsymbol{\alpha}_k = \mathbf{0}$. L'hypothèse sous considération est $\mathscr{H}_0 : \boldsymbol{\alpha}_k = \mathbf{0}$. On peut réécrire

$$\boldsymbol{X}_j^{(k)} = \bar{\boldsymbol{X}} + (\bar{\boldsymbol{X}}^{(k)} - \bar{\boldsymbol{X}}) + (\boldsymbol{X}_j^{(k)} - \bar{\boldsymbol{X}}^{(k)})$$

et calculer $\widehat{\boldsymbol{T}} = \widehat{\boldsymbol{B}} + \widehat{\boldsymbol{W}}$, qui donnent $\lambda^{2/n} = |\widehat{\boldsymbol{W}}|/|\widehat{\boldsymbol{B}} + \widehat{\boldsymbol{W}}| \sim \Lambda(p, n-K, K-1)$. Asymptotiquement, on aura

$$-\left(n - 1 - \frac{K+p}{2}\right) \log(\Lambda) \sim \chi^2_{p(K-1)}$$

sous l'hypothèse nulle.

On a utilisé le critère de Wilks jusqu'à maintenant, mais il y a d'autre critères. Par exemple, p

- $\Lambda$ de Wilks : $\left|\widehat{\boldsymbol{W}}\right| / \left|\widehat{\boldsymbol{B}} + \widehat{\boldsymbol{W}}\right|$
- Roy : basé sur la plus grande valeur propre de $\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{B}}$
- Lawley-Hotelling : $t = \text{tr}(\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{B}})$
- Pillai : $v = \text{tr}\left(\widehat{\boldsymbol{B}}(\widehat{\boldsymbol{W}} + \widehat{\boldsymbol{B}})^{-1}\right)$, réputé moins puissant, mais plus robuste

Si $k = 2$, les quatre critères sont équivalents.

**Proposition 3.4 (MANOVA à deux voies avec interactions)**
En supposant qu'on a le même nombre d'observations $n$ dans chaque groupe, on pose

$$\boldsymbol{X}_i^{(kg)} = \boldsymbol{\mu} + \boldsymbol{\alpha}_k + \boldsymbol{\beta}_g + \boldsymbol{\alpha}_k : \boldsymbol{\beta}_g + \boldsymbol{\varepsilon}_i^{(kg)}$$

avec $\boldsymbol{\varepsilon}_i^{(kg)} \overset{\text{ind}}{\sim} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ pour $k = 1, \ldots, K; g = 1, \ldots, G$ et $i = 1, \ldots n$. L'estimateur avec les données s'exprime sous la forme

$$\boldsymbol{X}_i^{(kg)} = \bar{\boldsymbol{X}} + (\bar{\boldsymbol{X}}^{(k\bullet)} - \bar{\boldsymbol{X}}) + (\bar{\boldsymbol{X}}^{(\bullet g)} - \bar{\boldsymbol{X}}) + (\bar{\boldsymbol{X}}^{(kg)} - \bar{\boldsymbol{X}}^{(k\bullet)} - \bar{\boldsymbol{X}}^{(\bullet g)} + \bar{\boldsymbol{X}}) + (\boldsymbol{X}_i^{(kg)} - \bar{\boldsymbol{X}}^{(kg)}).$$

et on décompose la matrice de variance totale en plusieurs composantes,

$$\widehat{\boldsymbol{T}} = \widehat{\boldsymbol{L}} + \widehat{\boldsymbol{C}} + \widehat{\boldsymbol{I}} + \widehat{\boldsymbol{W}}$$

pour respectivement les effets lignes, colonnes, interactions et la matrice de variance

intra-groupe. Les matrices sont définies de façon usuelle, c'est-à-dire

$$\begin{aligned}
\widehat{\boldsymbol{T}} &= \sum_{k=1}^{K}\sum_{g=1}^{G}\sum_{i=1}^{n}(\boldsymbol{X}_i^{(kg)} - \bar{\boldsymbol{X}})(\boldsymbol{X}_i^{(kg)} - \bar{\boldsymbol{X}})^\top \\
\widehat{\boldsymbol{L}} &= nG\sum_{k=1}^{K}(\bar{\boldsymbol{X}}^{(k\bullet)} - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}^{(k\bullet)} - \bar{\boldsymbol{X}})^\top \\
\widehat{\boldsymbol{C}} &= nK\sum_{g=1}^{G}(\bar{\boldsymbol{X}}^{(\bullet g)} - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}^{(\bullet g)} - \bar{\boldsymbol{X}})^\top \\
\widehat{\boldsymbol{I}} &= n\sum_{k=1}^{K}\sum_{g=1}^{G}(\bar{\boldsymbol{X}}^{(kg)} - \bar{\boldsymbol{X}}^{(k\bullet)} - \bar{\boldsymbol{X}}^{(\bullet g)} + \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}^{(kg)} - \bar{\boldsymbol{X}}^{(k\bullet)} - \bar{\boldsymbol{X}}^{(\bullet g)} + \bar{\boldsymbol{X}})^\top \\
\widehat{\boldsymbol{W}} &= \sum_{k=1}^{K}\sum_{g=1}^{G}\sum_{i=1}^{n}(\boldsymbol{X}_i^{(kg)} - \bar{\boldsymbol{X}}^{(kg)})(\boldsymbol{X}_i^{(kg)} - \bar{\boldsymbol{X}}^{(kg)})^\top.
\end{aligned}$$

Les hypothèses suivantes peuvent être intéressantes (les résultats sont valides sous $\mathscr{H}_0$) :

- $\mathscr{H}_0 : \boldsymbol{\alpha}_k : \boldsymbol{\beta}_g = \mathbf{0}$ (interactions nulles), calculée avec

$$\frac{\left|\widehat{\boldsymbol{W}}\right|}{\left|\widehat{\boldsymbol{I}} + \widehat{\boldsymbol{W}}\right|} \sim \Lambda(p, KG(n-1), (K-1)(G-1))$$

- $\mathscr{H}_0 : \boldsymbol{\alpha}_k = \mathbf{0}$ (effets lignes nuls), calculée avec

$$\frac{\left|\widehat{\boldsymbol{W}}\right|}{\left|\widehat{\boldsymbol{L}} + \widehat{\boldsymbol{W}}\right|} \sim \Lambda(p, KG(n-1), K-1)$$

- $\mathscr{H}_0 : \boldsymbol{\beta}_g = \mathbf{0}$ (effets colonnes nuls), calculée avec

$$\frac{\left|\widehat{\boldsymbol{W}}\right|}{\left|\widehat{\boldsymbol{C}} + \widehat{\boldsymbol{W}}\right|} \sim \Lambda(p, KG(n-1), G-1)$$

Si les interactions ne sont pas significatives, et qu'on élimine $\boldsymbol{\alpha}_k : \boldsymbol{\beta}_g$ du modèle pour tester la présence d'effets lignes et colonnes, alors $\widehat{\boldsymbol{W}} \mapsto \widehat{\boldsymbol{I}} + \widehat{\boldsymbol{W}}$. On doit toujours inclure les effets principaux pour que le modèle soit sensé.

**Proposition 3.5 (Test d'égalité des variances et statistique de Box)**
Soit $\mathscr{H}_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_K$ contre l'alternative $\mathscr{H}_1 : \boldsymbol{\Sigma}_k \neq \boldsymbol{\Sigma}_j$ pour au moins un $k \neq j$. Sous $\mathscr{H}_0 \cup \mathscr{H}_1$, on a $\widehat{\boldsymbol{\mu}}_k = \bar{\boldsymbol{X}}^{(k)}$, $\widehat{\boldsymbol{\Sigma}}_k = \boldsymbol{S}_k^*$ et sous $\mathscr{H}_0 : \widehat{\boldsymbol{\mu}}_k = \bar{\boldsymbol{X}}^{(k)}$ et $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S}_p^* = \widehat{\boldsymbol{W}}/n$.

La statistique est

$$-2\log(\lambda) = \sum_{k=1}^{K} n_k \log\left(\frac{|\boldsymbol{S}_p^*|}{|\boldsymbol{S}_k^*|}\right) \dot{\sim} \chi^2_{p(p+1)(K-1)/2}$$

La statistique de Box (1949) est

$$M = \gamma \sum_{k=1}^{K} (n_k - 1) \log\left(\frac{|\boldsymbol{S}_p|}{|\boldsymbol{S}_k|}\right)$$

avec

$$\gamma = 1 - \frac{2p^3 + 3p - 1}{6(p+1)(K-1)} \left(\sum_{k=1}^{K} \frac{1}{n_k - 1} - \frac{1}{n - K}\right).$$

une meilleure approximation à distance finie que la distribution asymptotique.

# Chapitre 4
# Analyse en composantes principales

L'analyse en composantes principales (ACP) est une technique d'algèbre linéaire qui remonte à Fisher et Hotelling. La motivation est multiple : d'une part, elle sert à la réduction de la dimension, si par exemple les données vivent dans un sous-espace de taille plus faible. D'autre part, elle permet d'interpréter les données de façon succincte, permettant parfois une meilleur perception du problème. Cela peut être utile pour combiner des variables mesurées sur des échelles différentes, par exemple les résultats aux différents épreuves d'un décathlon. Le quotient intellectuel est un exemple d'analyse en composantes principales (ACP).

Faire une ACP reviendra à diagonaliser la matrice de variance covariance ou la matrice des corrélations et en extraire vecteurs propres orthonormés et valeurs propres pour l'analyse.

**Définition 4.1 (Analyse en composante principale)**
Soit $\boldsymbol{X}$ un vecteur aléatoire de dimension $p \times 1$ avec $\mathsf{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ et $\mathsf{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. La transformation $\boldsymbol{C} = \boldsymbol{P}^\top(\boldsymbol{X} - \boldsymbol{\mu})$ où $\boldsymbol{P}$ est une matrice orthogonale $p \times p$ telle que $\boldsymbol{P}^\top \boldsymbol{\Sigma} \boldsymbol{P} = \boldsymbol{\Lambda}$ diagonale (où sans perte de généralité, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$). La $i^{\text{e}}$ composante principale est alors définie par $C_i = \boldsymbol{p}_i^\top(\boldsymbol{X} - \boldsymbol{\mu})$ où $\boldsymbol{p}_i$ est le $i^{\text{e}}$ vecteur propre correspondant à la valeur propre $\lambda_i$.

En multipliant une observation centrée par $\boldsymbol{P}^\top$, on la transforme en un vecteur de longueur $p$ dont les composantes sont les composantes principales. En fait, cela revient à exprimer l'observation dans une autre base.

**Proposition 4.2 (Propriétés de l'ACP)**

1. $\mathsf{E}(C_i) = 0$
2. $\mathsf{Var}(C_i) = \lambda_i$ et $\mathsf{Cov}(C_i, C_j) = 0$ pour tout $i \neq j$. En effet,

$$\mathsf{Var}(\boldsymbol{C}) = \mathsf{Var}\left(\boldsymbol{P}^\top(\boldsymbol{X} - \boldsymbol{\mu})\right) = \boldsymbol{P}^\top \boldsymbol{\Sigma} \boldsymbol{P} = \boldsymbol{\Lambda}.$$

3. $\sum_{i=1}^{p} \mathsf{Var}(C_i) = \mathrm{tr}(\boldsymbol{\Sigma})$, appelée inertie totale. On a

$$\sum_{i=1}^{p} \mathsf{Var}(C_i) = \mathrm{tr}(\boldsymbol{\Lambda}) = \mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{P}^\top\boldsymbol{P}) = \mathrm{tr}(\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top) = \mathrm{tr}(\boldsymbol{\Sigma})$$

4. Pour tout $\mathbf{a}$ tel que $\mathbf{a}^\top\mathbf{a} = 1$, $\mathsf{Var}(\mathbf{a}^\top\boldsymbol{X})$ est maximale pour $\mathbf{a} = \boldsymbol{p}_1$ et vaut $\lambda_1$. En utilisant les multiplicateurs de Lagrange, on cherche à maximiser $\mathsf{Var}(\mathbf{a}^\top\boldsymbol{X}) = \mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}$ et l'on a

$$\max_{\mathbf{a}} \mathcal{L} = \max_{\mathbf{a}} \left(\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a} - \xi\left(\mathbf{a}^\top\mathbf{a} - 1\right)\right) \quad \Rightarrow \quad \begin{cases} \frac{\partial \mathcal{L}}{\partial \mathbf{a}} = 2\boldsymbol{\Sigma}\mathbf{a} - 2\xi\mathbf{a} & = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi} = -(\mathbf{a}^\top\mathbf{a} - 1) & = 0 \end{cases}$$

qui donne $\mathbf{\Sigma a} = \lambda \mathbf{a}$ et $\mathbf{a}^\top \mathbf{a} = 1$.

5. Pour tout vecteur $\mathbf{a}$ tel que $\mathbf{a}^\top \mathbf{a} = 1$ et $\mathbf{a}^\top \boldsymbol{p}_i = 0$ pour $i = 1, \dots, j-1$, alors $\text{Var}\left(\mathbf{a}^\top \boldsymbol{X}\right)$ est maximale pour $\mathbf{a} = \boldsymbol{p}_j$ et vaut $\lambda_j$.

**Remarque**

1. Si $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, alors la 1$^{\text{e}}$ composante principale s'interprète comme l'axe principal de l'ellipsoïde de concentration.
2. La droite passant par $\boldsymbol{\mu}$ de direction $\boldsymbol{p}_1$ est la droite de plus grande variance. C'est aussi la droite qui minimise l'espérance du carré de la distance de $\boldsymbol{X}$ à cette droite.
3. $\sum_{i=1}^k \lambda_i / \sum_{i=1}^p \lambda_i$ représente la proportion de l'inertie totale expliquée par les $k$ premières variables principales.
4. Les composantes principales ne sont pas invariantes par rapport à des changements d'échelle (car il n'existe pas de relations entre les valeurs propres).

## Composantes principales et échantillon

Dans le cas d'un échantillon, remplacer $\boldsymbol{\mu}$ et $\mathbf{\Sigma}$ par les estimateurs usuels $\bar{\boldsymbol{X}}$ et $\boldsymbol{S}$ (ou $\boldsymbol{S}^*$). Si, par exemple, les unités des variables $\boldsymbol{X}$ sont des sec., kg, km et un taux, il est difficile d'analyser l'unité résultante, puisque

$$C_1 = a_1 X_1 + a_2 X_2 + a_3 X_3 + a_4 X_4$$

L'idée dans ces cas est de centrer et réduire (sans unités) et faire l'ACP après, ce qui revient à faire une ACP sur $\boldsymbol{R}$ au lieu de $\boldsymbol{S}$.

## 4.1 Inférence en ACP

**Théorème 4.3**

Soit $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ un échantillon de loi normale $\mathcal{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ et les estimateurs usuels $\bar{\boldsymbol{X}}, \boldsymbol{S}$. Pour $\mathbf{\Sigma}$ définie positive avec $\lambda_1 \neq \lambda_2 \neq \cdots \neq \lambda_p$, on a asymptotiquement

1.
$$\widehat{\boldsymbol{\lambda}} \sim \mathcal{N}_p\left(\boldsymbol{\lambda}, \frac{2}{n-1}\mathbf{\Lambda}^2\right) \quad \text{pour} \quad \boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)^\top$$

2.
$$\widehat{\boldsymbol{p}}_i \sim \mathcal{N}_p\left(\boldsymbol{p}_i, \frac{1}{n-1}\boldsymbol{V}_i\right) \quad \text{avec} \quad \boldsymbol{V}_i = \lambda_i \sum_{j \neq i} \frac{\lambda_j}{(\lambda_j - \lambda_i)^2} \boldsymbol{p}_j \boldsymbol{p}_j^\top.$$

3. On a également
$$\text{Cov}\left((\boldsymbol{p}_i)_r, (\boldsymbol{p}_j)_s\right) = -\lambda_i \lambda_j \frac{(\boldsymbol{p}_i)_r, (\boldsymbol{p}_j)_s}{(\lambda_i - \lambda_j)^2} \frac{1}{n-1}.$$

4. $\{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p\}$ est indépendant de $\{\widehat{\boldsymbol{p}}_1, \ldots, \widehat{\boldsymbol{p}}_p\}$.

On trouvera une démonstration dans un article d'Anderson (1963). Si l'on modifie les données (pour améliorer la normalité), on a $\log(\widehat{\lambda}_i) \dot{\sim} \mathcal{N}\big(\log(\lambda_i), \frac{2}{n-1}\big)$.

L'analyse en composantes principales sert davantage de technique d'exploration des données plutôt que de technique inférentielle.

### 4.1.1. Tests d'hypothèse

On esquisse deux tests utilisés dans le cadre de l'ACP : le premier a pour hypothèse nulle l'égalité des dernières $p-k$ valeurs propres. Le deuxième test sert à vérifier si la proportion d'inertie pour les premières valeurs est égale à une constante, soit $\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i = \psi_0$ donnée.

**Proposition 4.4 (Test d'isotropie partielle)**
Soit un test d'égalité des $p-k$ dernières valeurs propres, avec

$$\begin{aligned}\mathscr{H}_0 &: \lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_p \\ \mathscr{H}_1 &: \{\exists i, j \in \{k+1, \ldots, p\} : \lambda_i \neq \lambda_j, i \neq j\}.\end{aligned}$$

Cette hypothèse est utile dans la mesure où si un critère de sélection nous mène à inclure la $k+1^{\text{e}}$ composante, alors toutes les autres devraient être incluses dans la mesure où $\mathscr{H}_0$ ne peut être rejetée. Souvent, on débute avec $k = 0$.

Le test du ratio de vraisemblance est de la forme usuelle, à savoir

$$-2\log(\lambda) = np(A_0 - \log(G_0) - 1)$$

où $A_0$ et $G_0$ sont les moyennes arithmétiques et géométriques des valeurs propres de $\widehat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{S}^*$ et comme à l'accoutumée, $\widehat{\boldsymbol{\Sigma}}_0$ est l'estimateur de $\boldsymbol{\Sigma}$ sous $\mathscr{H}_0$. On peut montrer que, sous $\mathscr{H}_0$, $\widehat{\boldsymbol{\Sigma}}_0$ et $\boldsymbol{S}^*$ ont les mêmes valeurs propres $\widehat{\lambda}_i^*$ pour $i = 1, \ldots, k$. Donc les valeurs propres de $\widehat{\boldsymbol{\Sigma}}_0$ sont $(\widehat{\lambda}_1^*, \ldots, \widehat{\lambda}_k^*, \widehat{\lambda}^*, \ldots, \widehat{\lambda}^*)$ avec $\widehat{\lambda}^* = \sum_{j=k+1}^{p} \widehat{\lambda}_j^* / (p-k)$. Les valeurs propres de $\widehat{\boldsymbol{\Sigma}}_0^{-1}\boldsymbol{S}^*$ sont ainsi

$$\left(1, \ldots, 1, \frac{\widehat{\lambda}_{k+1}^*}{\widehat{\lambda}^*}, \ldots, \frac{\widehat{\lambda}_p^*}{\widehat{\lambda}^*}\right)$$

et donc $A_0 = 1$ et

$$G_0 = \left[\frac{(\widehat{\lambda}_{k+1}\widehat{\lambda}_{k+2}^* \cdots \widehat{\lambda}_p^*)^{\frac{1}{p-k}}}{\widehat{\lambda}^*}\right]^{\frac{p-k}{p}} = \left(\frac{g_0}{\widehat{\lambda}^*}\right)^{\frac{p-k}{p}}.$$

Sous $\mathscr{H}_0$, on a $k$ valeurs propres distinctes et 1 valeur propre commune, et $k$ vecteurs propres avec $p$ composantes, mais $k(k+1)/2$ restrictions découlant de l'orthogonalité, soit $k+1+kp-k(k+1)/2$ paramètres. Sous $\mathscr{H}_0 \sqcup \mathscr{H}_1$, il y a $p(p+1)/2$ contraintes, d'où

$$-2\log(\lambda) = -n(p-k)\log\left(\frac{g_0}{\widehat{\lambda}^*}\right) \dot{\sim} \chi^2_{(p-k+2)(p-k-1)/2}.$$

**Proposition 4.5 (Test de proportion d'inertie)**
Soit un test d'hypothèse que les $p-k$ composantes représentent moins qu'une certaine proportion de la variabilité totale,

$$\mathscr{H}_0 : \frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\cdots+\lambda_p} = \psi, \qquad \mathscr{H}_1 : \frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\cdots+\lambda_p} \neq \psi.$$

Si $\widehat{\lambda}_i$ sont les valeurs propres de $\boldsymbol{S}$, alors $\widehat{\psi} = (\widehat{\lambda}_1+\cdots+\widehat{\lambda}_k)/(\widehat{\lambda}_1+\cdots+\widehat{\lambda}_p)$. En appliquant l'équation A.10.2, on montre que

$$\widehat{\psi} \dot{\sim} \mathcal{N}(\psi, \tau^2)$$

avec

$$\tau^2 = \frac{2\operatorname{tr}(\boldsymbol{\Sigma}^2)}{(n-1)\left(\operatorname{tr}(\boldsymbol{\Sigma})\right)^2}(\psi^2 - 2\alpha\psi + \alpha^2), \qquad \alpha = \frac{\lambda_1^2+\cdots+\lambda_k^2}{\lambda_1^2+\cdots+\lambda_p^2}.$$

On peut ainsi déterminer un intervalle de confiance asymptotique pour $\psi$.

Si l'on utilise les variables standardisées et la matrice de corrélation $\boldsymbol{R}$, on peut tester l'égalité de toutes les valeurs propres, laquelle égalité tient si et seulement si la matrice de corrélation est $\mathbf{I}_p$. Le test devient

$$-\left(n - \frac{2p+11}{6}\right)\log(\det(\boldsymbol{R})) \dot{\sim} \chi^2_{p(p-1)/2}.$$

En revanche, tester que les $p-k$ valeurs propres de la matrice de corrélation sont égales pose de grosses difficultés. Une proposition de Bartlett est d'utiliser la statistique donnée par $-(n-1)(p-k)\log(g_0/\lambda^*)$, où $g_0, \lambda^*$ sont relatifs aux valeurs propres de $\boldsymbol{R}$. Cette statistique est à comparer au $1-\alpha$ quantile d'une distribution $\chi^2$ avec $(p-k+2)(p-k-1)/2$ degrés de liberté. Notez que cette distribution n'est pas la distribution asymptotique et mène à un test conservateur.

## 4.2 Interprétation et qualité d'une ACP

1. **Corrélations entre les composantes principales et les variables initiales** :

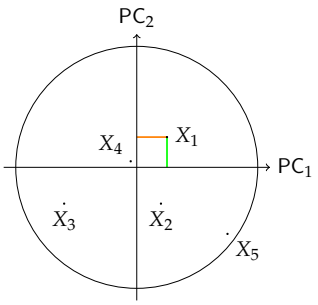
$$\begin{aligned}\mathsf{Cov}\left(C_\iota, X_j\right) = \mathsf{Cov}\left(\boldsymbol{p}_\iota^\top \boldsymbol{X}, X_j\right) &= \mathsf{Cov}\left(\sum_{k=1}^{p}(p_\iota)_k X_k, X_j\right) \\ &= \sum_{k=1}^{p}(p_\iota)_k \mathsf{Cov}\left(X_k, X_j\right) = \lambda_\iota (p_\iota)_j\end{aligned}$$

puisque $\boldsymbol{\Sigma}\boldsymbol{p}_\iota = \lambda_\iota \boldsymbol{p}_\iota$ et $\mathsf{Cov}\left(X_k, X_j\right) = \sigma_{kj}$. De plus, $\mathsf{Var}\left(C_\iota\right) = \lambda_\iota$ et $\mathsf{Var}\left(X_j\right) = \sigma_{jj}$, ce qui donne

$$\mathsf{Cor}\left(C_\iota, X_j\right) = \frac{\lambda_\iota (p_\iota)_j}{\sqrt{\lambda_\iota}\sqrt{\sigma_{jj}}}$$

et si on commence avec $\boldsymbol{R}$ au départ, on a $\sqrt{\lambda_\iota}(p_\iota)_j$ puisque les variance sont égales à 1. On représente graphiquement cet éléments dans le cercle des corrélations, ici de rayon unitaire. Si la normalisation des variances n'est pas la même pour chaque variable, on aura une ellipse puisque la variance de chaque axe sera différente.

FIGURE 4 – Cercle des corrélations




La variable $X_4$ est très mal représentée dans le plan des deux composantes principales puisque leur corrélation est presque nulle. À l'inverse, une portion importante de la variabilité du point $X_5$ est expliquée par les deux premières composantes.
**Interpréter les oppositions sur chaque axe :** on a $\lambda_\iota = \sum_{j=1}^{p} \mathsf{Cor}\left(C_\iota, \boldsymbol{X}_j\right)^2$ et on appelle parfois contribution de la variable $j$ à l'axe $\iota$

$$\frac{\mathsf{Cor}\left(C_\iota, X_j\right)^2}{\lambda_\iota} = (p_\iota)_j^2$$

si l'on travaille à partir de la matrice de corrélations. Ce cercle des corrélations est le pendant pour les variables de la projection des variables. Dans le *biplot*, les deux graphiques sont superposés. Attention : la représentation doit s'utiliser avec précaution. On ne peut pas interpréter la proximité des points (individus), mais leur direction

puisque souvent on a pas les mêmes unités et les échelles des deux représentations sont différentes.

2. **Effet de "taille" :** Si $\sigma_{ij} > 0$ pour tout $i$ et $j$, on peut montrer qu'alors tous les coefficients du premier vecteur propre sont de même signe (théorème de Frobenius). $C_1 = \mathbf{a}_1^\top \boldsymbol{X}$ est une sorte de moyenne pondérée (toutes les variables n'ont pas le même poids). Ainsi, la 1$^{\text{e}}$ composante principale mesure souvent un effet de "taille". On peut aussi remarque que la 2$^{\text{e}}$ composante principale, $C_2$, représente fréquemment un effet de "forme".

3. **Importance des individus** : dans le cas d'un échantillon $\mathbf{X}$ de taille $n \times p$, on a $C_{1i} = \boldsymbol{p}_1^\top (\boldsymbol{X}_{i\bullet} - \bar{\boldsymbol{X}})$, la projection du $i^{\text{e}}$ individu sur la première composante principale. La variance des points projetés est $\frac{1}{n}\sum_{i=1}^n C_{ii}^2 = \widehat{\lambda}_\iota$ et $n^{-1}C_{ii}^2/\widehat{\lambda}_\iota$ est la contribution de l'individu $i$ à la composante $\iota$.

**Théorème 4.6 (Frobenius)**

Soit $\mathbf{A}$ une matrice symétrique avec $a_{jk} > 0$, alors les composantes du premier vecteur propre de $\mathbf{A}$ sont de même signe.

4. **Variables et individus supplémentaires :** On essaie parfois de garder des variables et/ou des individus en réserve pour valider ou confirmer son interprétation. Ceci peut être une aide.

5. **Qualité des représentations :**
   - Mesure globale : pourcentage d'inertie expliquée par les composantes retenues
   - Qualité de la représentation des individus : $\mathsf{CO}_\iota = C_{ii}^2 / \sum_{k=1}^p C_{ki}^2$ et $\mathsf{QLT} = \mathsf{CO}_1 + \mathsf{CO}_2$. L'indicateur QLT est la qualité de la représentation d l'individu $i$ sur les deux premières composantes.

6. **Nombre d'axes (facteurs, composantes) à retenir :**

(a) Critère de Kaiser : en travaillant sur $\boldsymbol{R}$, ne retenir que les composantes avec valeurs propres plus grandes que 1.

(b) Critère de Jolliffe : prendre les composantes jusqu'à obtenir plus de 80% de l'inertie totale

(c) Critère du coude (critère graphique) : on trace l'index des valeurs propres contre $\lambda_i$

(d) *Scree test* de Cattell : calculer $\lambda_1 - \lambda_2 = \varepsilon_1$, $\lambda_2 - \lambda_3 = \varepsilon_2, \ldots$ et calculer par la suite $\varepsilon_1 - \varepsilon_2 = \delta_1$, $\varepsilon_2 - \varepsilon_3 = \delta_2$. Ne retenir que les valeurs propres $\lambda_1, \ldots, \lambda_{k+1}$ avec $\delta_1, \ldots, \delta_k \geq 0$.

## 4.3 ACP, décomposition en valeurs singulières et régression

Soit $\mathbf{X}$ la matrice $n \times p$ formée des observations $\mathbf{X} = (X_1^\top \cdots X_n^\top)^\top$, avec $X_i \in \mathbb{R}^p$ centrées (donc assumant que la moyenne des colonnes est égale à zéro). On considère la sélection de $k < p$ combinaisons linéaires $\mathbf{XA} \in \mathbb{R}^k$ qui, dans un sens, représentent le mieux les données originales.

La décomposition en valeurs singulière de $\mathbf{X}$ est

$$\mathbf{X} = \boldsymbol{UDV}^\top$$

où
- $\boldsymbol{U}$ est une matrice orthonormale $n \times p$; les colonnes de $\boldsymbol{U}$ engendre l'espace des colonnes de $\mathbf{X}$
- $\boldsymbol{V}$ est une matrice orthogonale $p \times p$; les colonnes de $\boldsymbol{V}$ engendre l'espace des lignes de $\mathbf{X}$
- $\boldsymbol{D}$ est une matrice diagonale $p \times p$ dont les éléments $d_i$ sont les valeurs singulières de $\mathbf{X}$, avec $d_1 \geq d_2 \geq \cdots d_p \geq 0$ et $\text{rang}(\mathbf{X}) = \text{rang}(\boldsymbol{D})$.

Une autre formulation de la décomposition singulière prend $\boldsymbol{D}$ comme matrice $n \times p$ contenant le carré des valeurs singulières, à savoir les valeurs propres de $\mathbf{XX}^\top$ et de $\mathbf{X}^\top\mathbf{X}$, $\boldsymbol{U}$ la matrice $n \times n$ des vecteurs singuliers gauches, égale à la matrice de vecteurs propres de $\mathbf{XX}^\top$ et $\boldsymbol{V}^\top$ matrice $p \times p$ de vecteurs singuliers droits, soit les vecteurs propres de $\mathbf{X}^\top\mathbf{X}$.

Les composantes principales sont alors les colonnes de $\mathbf{XV}$. On a

$$(n-1)\boldsymbol{S} = \mathbf{X}^\top\mathbf{X} = \boldsymbol{VD}^\top\boldsymbol{U}^\top\boldsymbol{UDV}^\top = \boldsymbol{VD}^2\boldsymbol{V}^\top$$

puisque $\boldsymbol{U}$ est orthonormale et $\boldsymbol{D}$ est diagonale. Dans l'ACP, $\boldsymbol{S} = \boldsymbol{P\Lambda P}^\top$. On a donc $\boldsymbol{P} = \mathbf{V}$ et $\boldsymbol{\Lambda} = (n-1)^{-1}\boldsymbol{D}^2$.

### Régression

On rappelle quelques notions de régression linéaire. Le but de cette section est d'utiliser la décomposition en valeurs singulières (et l'ACP) pour mieux comprendre les estimateurs à rétrécissement. On conclura avec deux méthodes qui utilisent une transformation de la matrice de régresseurs $\mathbf{X}$ pour la régression.

Dans la régression par moindre carrés, on postule un modèle linéaire de la forme $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, où $\boldsymbol{\varepsilon}$ est une matrice d'aléas indépendants et de moyenne nulle. L'estimateur $\widehat{\boldsymbol{\beta}}$ qui minimise l'erreur moyenne quadratique est la solution du problème

$$\begin{aligned}
\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \quad &\Rightarrow \quad \widehat{\boldsymbol{\beta}}_{\text{MC}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} && \text{si Var}(\boldsymbol{\varepsilon}) \propto \mathbf{I}_n \\
\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \quad &\Rightarrow \quad \widehat{\boldsymbol{\beta}}_{\text{MCP}} = (\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{y} && \text{si Var}(\boldsymbol{\varepsilon}) = \sigma\mathbf{I}_n
\end{aligned}$$

pour $\sigma$ connu. Le dernier cas (données hétéroscédastiques) est appelé moindres carrés pondérés.

Nous considérerons subséquemment les seuls moindres carrés ordinaires, dont l'estimé sera dénoté $\widehat{\boldsymbol{\beta}}$. Les valeurs ajustées sont $\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}_{\mathbf{X}}\boldsymbol{y} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ et par construction, est orthogonal à l'estimé des résidus $\widehat{\boldsymbol{\varepsilon}} = (\mathbf{I} - \mathbf{H}_{\mathbf{X}})\boldsymbol{y}$.

Si $\mathbf{X} = \boldsymbol{U}\boldsymbol{D}\mathbf{V}^\top$, on obtient

$$\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{D}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top)^{-1}\boldsymbol{V}\boldsymbol{D}^\top\boldsymbol{U}^\top\boldsymbol{y} = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{y}$$

où $\boldsymbol{U}^\top\boldsymbol{y}$ sont les coordonnées de $\boldsymbol{y}$ dans la base orthonormale $\boldsymbol{U}$. Évidemment, le résultat découle du fait que $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^\top = \mathbf{I}_p$, $\boldsymbol{D} = \boldsymbol{D}^\top$ et $(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top)^{-1} = \boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^\top$. Puisque $\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\boldsymbol{V}\boldsymbol{V}^\top\widehat{\boldsymbol{\beta}}$, les coefficients par rapport à la matrice de régresseurs $\mathbf{X}\boldsymbol{V}$ sont orthogonaux et peuvent être obtenus par des régressions simples successives. L'exclusion de colonnes de $\mathbf{X}\boldsymbol{V}$ n'affectent pas les estimés.

## Régression avec pénalité et estimateurs à rétrécissement

**Proposition 4.7 (Régression *ridge*)**
Les méthodes de rétrécissement introduisent du biais dans l'estimateur de $\boldsymbol{\beta}$ au profit d'une réduction de la variance. Certaines pénalités permettent même de faire de la sélection de variables. La régression de crête (*ridge*) ajoute une pénalité sur la norme $l_2$ du vecteur de coefficient, via

$$\min_{\boldsymbol{\beta}}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \theta\boldsymbol{\beta}^\top\boldsymbol{\beta} \quad \Rightarrow \quad \widehat{\boldsymbol{\beta}}_{\mathrm{r}} = (\mathbf{X}^\top\mathbf{X} + \theta\mathbf{I}_p)^{-1}\mathbf{X}^\top\boldsymbol{y}$$

si les erreurs sont homoscédastiques et indépendantes. L'estimateur *ridge* s'exprime dans la base de vecteurs propres via

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\mathrm{r}} &= (\mathbf{X}^\top\mathbf{X} + \theta\mathbf{I}_p)^{-1}\mathbf{X}^\top\boldsymbol{y} \\
&= (\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top + \theta\mathbf{I}_p)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{y} \\
&= \boldsymbol{V}(\boldsymbol{D}^2 + \theta\mathbf{I}_p)^{-1}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{y} \\
&= \sum_{j=1}^{p} \boldsymbol{v}_j \frac{d_j}{d_j^2 + \theta}\boldsymbol{u}_j^\top\boldsymbol{y},
\end{aligned}$$

d'où

$$\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{r}} = \sum_{j=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2 + \theta}\boldsymbol{u}_j^\top\boldsymbol{y}.$$

Le coefficient $d_j^2/(d_j^2 + \theta)$ est un coefficient de rétrécissement.

**Proposition 4.8 (Régression LASSO)**
Une des pénalités les plus connues est celle de la régression *lasso*, qui donne le problème

d'optimisation quadratique

$$\min_{\boldsymbol{\beta}}(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})^\top(\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta})+\theta\sum_{j=1}^{p}|\beta_j|.$$

La solution est obtenable par le biais d'algorithmes dédiés qui permettent de trouver l'ensemble des solutions en fonction de $\theta$, et ce au même coût de calcul que pour la régression *ridge*. La nature de la pénalité pousse les estimés pour certains régresseurs à zéro, ce qui crée un modèle plus parcimonieux. Là où la pénalité de crête rétrécie de façon proportionnelle les estimateurs de moindre carré, $\widehat{\beta}_j^{\text{r}}=\widehat{\beta}_j/(1+\theta)$, la régression lasso donne $\widehat{\beta}_j^{\text{lasso}}=\text{sign}(\widehat{\beta}_j)\min\{|\widehat{\beta}_j|-\theta,0\}$. Mentionnons finalement la pénalité du filet élastique, $\theta\sum_{j=1}^{p}(\alpha\beta_j^2+(1-\alpha)|\beta_j|$, un compromis entre lasso et *ridge*.

Les pénalités présentées jusqu'à maintenant ont un désavantage : elles atténuent le signal même si les coefficients sont larges.

**Proposition 4.9 (Régression SCAD)**
La pénalité SCAD, de paramètres $a>2$ et $\theta>0$ est une spline quadratique avec noeuds à $\theta$ et $a\theta$ de la forme

$$\text{p}(a,\theta)=\begin{cases}\theta|\beta_j| & \text{si } |\beta_j|\leq\theta\\ -\frac{\beta_j^2-2a\theta|\beta|+\theta^2}{2(a-1)} & \text{si } \theta<|\beta_j|\leq a\theta\\ \frac{(a+1)\theta^2}{2} & \text{si } |\beta_j|>a\theta,\end{cases}$$

qui donne comme solution si les régresseurs sont orthogonaux

$$\widehat{\beta}_j^{\text{scad}}=\begin{cases}\text{sign}(\widehat{\beta}_j)\min\{|\widehat{\beta}_j|-\theta,0\} & \text{si } |\widehat{\beta}_j|<2\theta\\ \{(a-1)\widehat{\beta})j-\text{sign}(\widehat{\beta}_j))a\theta\}/(a-2) & \text{si } 2\theta<|\widehat{\beta}_j|\leq a\theta\\ \widehat{\beta}_j & \text{si } |\widehat{\beta}_j|>a\theta.\end{cases}$$

Un algorithme de descente par coordonnée est nécessaire pour obtenir une solution. Comme pour le lasso et la régression ridge, les paramètres optimaux pour la pénalité sont souvent dérivés par validation croisée. En revanche, les erreurs standards rapportées par le progiciel ne prennent pas en compte la sélection de variable et les tests d'hypothèse basés sur ces dernières ne sont pas valides.

## Régression avec dérivés de la matrice de régresseurs

Soit $\mathbf{X}$ un tableau $n\times p$ non-stochastique centré et réduit (dont la moyenne des colonnes est de 0 et leur variance de 1) et un vecteur réponse $\boldsymbol{y}$ tel que $\bar{y}=0$, avec le modèle

$$\boldsymbol{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon},\qquad \boldsymbol{\varepsilon}\sim\mathcal{N}_n(\mathbf{0}_n,\sigma^2\mathbf{I}_n).$$

On considère un changement de base de $\boldsymbol{X}$ en prenant $\boldsymbol{z}_i = \boldsymbol{X}\boldsymbol{p}_i$, où $i = 1, \ldots, k \leq p$ Une régression par moindre carrés minimise la distance euclidienne verticale entre les observations et la droite de régression, tandis qu'une régression avec les composantes principales minimisera la distance perpendiculaire à la droite.

**Proposition 4.10 (Régression sur composantes principales)**
On fait la régression de $\boldsymbol{y}$ sur les vecteurs colonnes $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$, lesquels sont orthogonaux par construction. Les estimés sont donnés par

$$\widehat{\boldsymbol{y}}_{\text{rcp}} = \bar{y}\mathbf{1} + \sum_{i=1}^{k} \widehat{\gamma}_i \boldsymbol{z}_i, \qquad \widehat{\gamma}_i = (\boldsymbol{z}_i^\top \boldsymbol{z}_i)^{-1} \boldsymbol{z}_i^\top \boldsymbol{y}.$$

On appelle matrice de chargement la matrice des vecteurs propres de $\boldsymbol{\Sigma}$, divisés par la racine des valeurs propres respectives. Si $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\boldsymbol{P}^\top$, la matrice de chargement est $\mathbf{L} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}$. Si $\mathbf{L}_{1:k} = (\mathbf{l}_1, \ldots, \mathbf{l}_k)$ dénote la sous-matrice de rang $k$, la meilleure approximation de la matrice de covariance (ou de corrélation) est donnée par $\mathbf{L}_{1:k}\mathbf{L}_{1:k}^\top$. Dans le cadre de la décomposition spectrale, $\boldsymbol{U}\boldsymbol{D}$ est une matrice de scores pour la régression. [5]

Le rétrécissement se fait à l'aide du choix de $k$. En pratique, il est peu courant d'utiliser la régression en composantes principales puisque cette dernière tente de représenter de façon parsimonieuse la variation en $\mathbf{X}$ sans considération de son pouvoir explicatif. Toutes les variables de $\mathbf{X}$ sont également nécessaires pour les prédictions. Pour illustrer cette réalité, supposons un instant que la variance $\sigma^2$, les coefficients $\gamma$ et les valeurs propres $\boldsymbol{\lambda}$ sont connus.

**Proposition 4.11 (Choix optimal de $k$ pour RCP)**
Pour minimiser l'erreur moyenne quadratique, on conserve une composante principale si le carré du biais est plus important que la réduction de variance. L'omission d'une composante $i$ introduit un biais de $\gamma_i$ (puisque l'estimateur $\widehat{\gamma}_i$ est sans biais pour $\gamma_i$ et que les régresseurs $\boldsymbol{z}_i$ sont orthogonaux). La réduction de la variance due à l'omission d'une composante est $\sigma^2(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}_{ii} = \sigma^2/((n-1)\lambda_i)$.

Les variables à conserver ne sont donc pas nécessairement les premières composantes et le choix optimal pour $k$ dépend des paramètres $\gamma$, lesquels sont inconnus en pratique.

## Régression par moindres carrés partiels (PLS)

Considérons $\boldsymbol{y} \in \mathbb{R}^n$ et $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ des tableaux centrés (chaque colonne de $\boldsymbol{X}$ de même que $\boldsymbol{y}$ ont moyenne zéro). La régression par moindre carrés partiels est une technique analogue à la régression en composantes principales qui utilise des combinaisons linéaires de $\boldsymbol{X}$ et permet de traiter le cas où $p \gg n$. Cette technique est particulièrement

5. Notez au passage que R utilise le terme `loadings` pour dénoter les axes principaux (un abus de langage).

utilisée en chimiométrie. La régression des moindres carrés partiels construit itérativement une base de $k \leq \min(n-1, p)$ vecteurs de poids $\mathbf{W}_k = (\mathbf{w}_1, \ldots, \mathbf{w}_k)$ pour les colonnes de $\boldsymbol{X}$, produisant une matrice de scores $n \times k$ donnée par l'expression $\mathbf{T} = \boldsymbol{X}\mathbf{W}$.

La première direction est obtenue en maximisant la covariance $\mathrm{Cov}\,(\boldsymbol{X}\mathbf{w}, \boldsymbol{y})$, ou bien encore $\mathrm{Cor}\,(\boldsymbol{X}\mathbf{w}, \boldsymbol{y})\,\sqrt{\mathrm{Var}\,(\boldsymbol{X}\mathbf{w})}$ pour $\mathbf{w} \in \mathbb{R}^p$ sujet à $\|\mathbf{w}\| = 1$. Ce problème d'optimisation contrainte peut être résolu à l'aide des multiplicateurs de Lagrange et on trouve que $\mathbf{w}_1$ est le vecteur propre associé à la plus grande valeur propre de $\boldsymbol{X}^\top \boldsymbol{y}\boldsymbol{y}^\top \boldsymbol{X}$, soit $\mathbf{w}_1 = \boldsymbol{X}^\top \boldsymbol{y} / \|\boldsymbol{X}^\top \boldsymbol{y}\|$. Pour les directions subséquentes, le coefficient $\mathbf{w}_i$ est la solution du problème $\max_{\mathbf{w}} \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{X}\mathbf{w})$ sujet aux contraintes $\|\mathbf{w}\| = 1$ et $\mathbf{w}^\top \boldsymbol{S}\mathbf{w}_l = 0$ pour $l = 1, \ldots, i$. La régression PLS effectue un compromis entre l'analyse en composantes principales (qui maximise la variance) et la régression multiple (l'analyse canonique dans le cas où $\boldsymbol{Y}$ est $n \times m$), qui maximise la corrélation.

Mathématiquement, les moindres carrés partiels reviennent à minimiser

$$\widehat{\boldsymbol{\beta}}_i = \underset{\boldsymbol{\beta} \in \mathcal{K}_i(\mathbf{A}, \mathbf{b})}{\arg\min}\, (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

avec $\mathbf{A} = n^{-1}\boldsymbol{X}^\top \boldsymbol{X}, \mathbf{b} = n^{-1}\boldsymbol{X}^\top \boldsymbol{y}$, où $\mathcal{K}_i(\mathbf{A}, \mathbf{b})$ est un espace de Krylov, c'est-à-dire l'espace engendré par $\mathrm{vect}\{\mathbf{A}^{j-1}\mathbf{b}\}_{j=1}^{i}$.

Pour la résolution, on exprime le système de régression

$$\begin{aligned} \boldsymbol{X} &= \mathbf{T}\mathbf{P}^\top + \mathbf{E} \\ \boldsymbol{Y} &= \mathbf{U}\mathbf{Q}^\top + \mathbf{F} \end{aligned}$$

où $\mathbf{T}, \mathbf{U}$ sont les matrices $n \times k$ de score, $\mathbf{P}$ et $\mathbf{Q}$ sont des matrices de chargements de taille respective $p \times k$ et $1 \times k$, $\mathbf{E}$ et $\mathbf{F}$ des matrice de résidus, $\mathbf{c}^\top \in \mathbb{R}^k$ est un vecteur de contributions et finalement $\mathbf{W}$ est une matrice de poids de taille $n \times k$.[6]

La résolution du problème des moindres carrés partiels passe par des algorithmes itératifs, donc l'algorithme NIPALS. Ce dernier alterne entre régression et projection et on calcule itérativement les colonnes du système. Par exemple, la matrice $\mathbf{T}_l$ à l'étape $l$ est formée par la concaténation des colonnes $\mathbf{t}_1, \ldots, \mathbf{t}_l$, lesquels $\mathbf{t}$ sont calculés récursivement. Des généralisation multivariées existent et diffèrent légèrement de la présentation ci-dessous.

**Algorithme 4.1 (Moindres carrés partiels itératif non-linéaire (NIPALS) — univarié)**
Soit $\mathbf{X}$ et $\boldsymbol{y}$ centrés et le nombre de composantes PLS $k$ :

**Initialisation:** $\mathbf{E}_0 = \boldsymbol{X}$ et $\mathbf{u}_0 = \mathbf{F}_0 = \boldsymbol{y}$

1: **pour** $i \leftarrow 1, \ldots, k$

6. Dans le cas multivarié avec $\mathbf{Y} \in \mathbb{R}^{n \times m}$, on cherche à maximiser $\mathrm{Cov}\,(\boldsymbol{X}\mathbf{w}, \boldsymbol{Y}\mathbf{c})$ avec comme contrainte additionnelle $\|\mathbf{c}\| = 1$, etc. C'est donc désormais une matrice de vecteurs orthogonaux $\mathbf{C}$ $k \times m$ que l'on cherche.

2: $\mathbf{w}_i \leftarrow \mathbf{E}_{i-1}^\top \mathbf{F}_{i-1}/(\mathbf{F}_{i-1}^\top \mathbf{F}_{i-1})$ ▷ régression
3: $\mathbf{w}_i \leftarrow \mathbf{w}_i/\|\mathbf{w}_i\|$
4: $\mathbf{t}_i \leftarrow \mathbf{E}_{i-1}\mathbf{w}_i/\mathbf{w}_i^\top \mathbf{w}_i$
5: $\mathbf{p}_i \leftarrow \mathbf{E}_{i-1}^\top \mathbf{t}_i/\mathbf{t}_i^\top \mathbf{t}_i$ ▷ régression
6: $\mathbf{E}_i \leftarrow \mathbf{E}_{i-1} - \mathbf{t}_i\mathbf{p}_i^\top$ ▷ projection
7: $\mathrm{c}_i \leftarrow \mathbf{F}_{i-1}^\top \mathbf{t}_i/\mathbf{t}_i^\top \mathbf{t}_i$ ▷ régression
8: $\mathbf{u}_i \leftarrow \mathbf{F}_{i-1}/\mathrm{c}_i$
9: $\mathbf{F}_i \leftarrow \mathbf{F}_{i-1} - \mathbf{t}_i\mathrm{c}_i^\top$ ▷ projection

Dans le cas de données manquantes, les quantités sont calculées avec les données complètes. Si aucune donnée n'est manquante, on peut remplacer les deux premières étapes (lignes 2-3) par $\mathbf{w}_i \leftarrow \mathbf{E}_{i-1}^\top \mathbf{F}_{i-1}/\|\mathbf{E}_{i-1}^\top \mathbf{F}_{i-1}\|$ et 4 par $\mathbf{t}_i \leftarrow \mathbf{E}_{i-1}\mathbf{w}_i$.

Les valeurs ajustées après $k$ itérations sont calculées en utilisant la récursion

$$\widehat{\boldsymbol{y}}_{(k)} = \bar{y} + \sum_{i=1}^{k} \mathbf{t}_i\widehat{\alpha}_i, \qquad \widehat{\alpha}_i = \frac{\mathbf{u}_i^\top \mathbf{t}_i}{\mathbf{t}_i^\top \mathbf{t}_i}.$$

On peut réexprimer les coefficients de la régression en fonction de la matrice $\boldsymbol{X}$ de départ pour le modèle de régression $\boldsymbol{y} = \boldsymbol{X}\mathbf{B} + \mathbf{F}$. On a alors

$$\mathbf{B}_{(k)} = \mathbf{W}_k(\mathbf{P}_k^\top \mathbf{W}_k)^{-1}\mathbf{c}_k.$$

Si $k = p$, l'estimé $\mathbf{B}_{(k)}$ est égal à celui de la régression des moindres carrés, tout comme dans la régression en composantes principales. Si la matrice de régresseurs $\boldsymbol{X}$ est orthogonale, les moindres carrés partiels retournent les estimateurs des moindres carrés après la première itération ; aucune étape subséquente ne modifie les estimés initiaux. Les erreurs à l'étape $h$, $\widehat{\mathbf{f}}_h$ sont recalculées comme $\widehat{\mathbf{f}}_h = \boldsymbol{y} - \widehat{\boldsymbol{y}}_{(h)}$.

On peut calculer, comme dans le cas de l'ACP, la proportion de variance expliquée pour $\boldsymbol{y}$ et $\boldsymbol{X}$ avec $h$ chargements :

$$\begin{aligned}
\mathsf{pVar}_h(\boldsymbol{y}) &= \frac{\mathsf{Var}\left(\widehat{\boldsymbol{y}}_{(h)}\right)}{\mathsf{Var}\left(\boldsymbol{y}\right)} = \frac{\mathsf{SC}_h(\boldsymbol{y})}{\boldsymbol{y}^\top \boldsymbol{y}} \\
\mathsf{pVar}_h(\boldsymbol{X}) &= \frac{\mathbf{t}_h^\top \mathbf{t}_h \mathbf{p}_h^\top \mathbf{p}_h}{\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X})} = \frac{\mathsf{SC}_h(\boldsymbol{X})}{\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X})}.
\end{aligned}$$

Puisque les scores sont orthogonaux, la proportion cumulative des $k$ premières variables PLS est la somme des contributions $\sum_{h=1}^{k} \mathsf{pVar}_h$.

Il peut être difficile de choisir le nombre de composantes pour la régression PLS. Les principaux critères proviennent de la validation croisée : on peut utiliser la méthode du canif (*jacknife*) en laissant de côté une observation et en estimant le modèle sans cette

dernière. La somme des carrés des résidus non-standardisée est comme à l'accoutumée

$$\mathsf{RSS}_h = \widehat{f}_h^\top \widehat{f}_h = \sum_{i=1}^{n} \left( y_i - \widehat{y}_{(h)i} \right)^2$$

tandis que la somme du carré des résidus prédits par le modèle estimé, dénotée $\mathsf{PRESS}_h$ est

$$\mathsf{PRESS}_h = \sum_{i=1}^{n} \left( y_i - \widehat{y}_{(h)i}^{-i} \right)^2$$

La statistique $\mathsf{PRESS}_k$ n'existe pas nécessairement (si par exemple $k = n$). On définit quelques statistiques

$$Q^2(h) = 1 - \frac{\mathsf{PRESS}_h}{\mathsf{RSS}_{h-1}}$$

$$Q^2_{\text{cum}}(h) = 1 - \prod_{i=1}^{h} \frac{\mathsf{PRESS}_h}{\mathsf{RSS}_{h-1}}$$

$$R^2_{\mathsf{VC}}(h) = 1 - \frac{\mathsf{PRESS}_h}{\mathsf{RSS}_0}$$

avec $\mathsf{RSS}_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2$. En ce qui a trait aux critères de décisions, on a plusieurs choix

- Ajouter le chargement $h$ si $Q^2_{\text{cum}}(h) > 0.5 Q^2_{\text{cum}}(h-1)$
- Ajouter le chargement $h$ si $\sqrt{\mathsf{PRESS}_h} \leq \gamma \sqrt{\mathsf{RSS}_{h-1}}$ avec $\gamma = 0.95$ pour $n < 100$ et $\gamma = 1$ si $n \geq 100$.

# Chapitre 5
# Analyse canonique

Dans un modèle de régression multiple, on pose comme modèle pour une observation $Y = \mathbf{a}^\top \boldsymbol{X}$ tel que $\mathsf{Cor}\left(Y, \mathbf{a}^\top \boldsymbol{X}\right)$ est maximale. Dans le cas où $\boldsymbol{Y}$ est un vecteur réponse, l'**analyse canonique** consiste à poser $\mathbf{b}^\top \boldsymbol{Y}$ expliquée par $\mathbf{a}^\top \boldsymbol{X}$ de telle sorte que $\mathsf{Cor}\left(\mathbf{b}^\top \boldsymbol{Y}, \mathbf{a}^\top \boldsymbol{X}\right)$ soit maximale.

## 5.1 Définition et propriétés de l'analyse canonique

Soit $(\boldsymbol{X}, \boldsymbol{Y})^\top$ un vecteur de composantes $\boldsymbol{X}$ de dimension $p_1$ et $\boldsymbol{Y}$ de dimension $p_2$, avec $p := p_1 + p_2$, de moyenne $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)^\top =: \boldsymbol{\mu}$ et matrice de variance covariance $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$. On a

$$\begin{aligned} \mathsf{Cov}\left(\mathbf{a}^\top \boldsymbol{X}, \mathbf{b}^\top \boldsymbol{Y}\right) &= \mathbf{a}^\top \boldsymbol{\Sigma}_{12} \mathbf{b} \\ \mathsf{Var}\left(\mathbf{a}^\top \boldsymbol{X}\right) &= \mathbf{a}^\top \boldsymbol{\Sigma}_{11} \mathbf{a} \\ \mathsf{Var}\left(\mathbf{b}^\top \boldsymbol{Y}\right) &= \mathbf{b}^\top \boldsymbol{\Sigma}_{22} \mathbf{b}. \end{aligned}$$

Ainsi,

$$\mathsf{Cor}\left(\mathbf{a}^\top \boldsymbol{X}, \mathbf{b}^\top \boldsymbol{Y}\right) = \frac{\mathbf{a}^\top \boldsymbol{\Sigma}_{12} \mathbf{b}}{\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma}_{11} \mathbf{a} \mathbf{b}^\top \boldsymbol{\Sigma}_{22} \mathbf{b}}} \tag{5.2}$$

Maximiser cette corrélation donnera une valeur positive (pourquoi ?).[7] De plus, il y aura un problème de normalisation de $\mathbf{a}$ et $\mathbf{b}$. On imposera des contraintes pour maximiser équation (5.2), soit $\mathbf{a}^\top \boldsymbol{\Sigma}_{11} \mathbf{a} = 1$ et $\mathbf{b}^\top \boldsymbol{\Sigma}_{22} \mathbf{b} = 1$. On résoudra le problème d'optimisation à l'aide des multiplicateurs de Lagrange

$$\mathcal{L} = \mathbf{a}^\top \boldsymbol{\Sigma}_{12} \mathbf{b} - \frac{1}{2}\delta(\mathbf{a}^\top \boldsymbol{\Sigma}_{11} \mathbf{a} - 1) - \frac{1}{2}\gamma(\mathbf{b}^\top \boldsymbol{\Sigma}_{22} \mathbf{b} - 1).$$

En dérivant, on obtient

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}} = \boldsymbol{\Sigma}_{12} \mathbf{b} - \delta \boldsymbol{\Sigma}_{11} \mathbf{a} = 0 \qquad \Rightarrow \quad \boldsymbol{\Sigma}_{12} \mathbf{b} = \delta \boldsymbol{\Sigma}_{11} \mathbf{a} \tag{5.3}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}} = \boldsymbol{\Sigma}_{21} \mathbf{a} - \delta \boldsymbol{\Sigma}_{22} \mathbf{b} = 0 \qquad \Rightarrow \quad \boldsymbol{\Sigma}_{21} \mathbf{a} = \gamma \boldsymbol{\Sigma}_{22} \mathbf{b}. \tag{5.4}$$

De la première expression, on obtient en inversant les sous-matrices de covariance $\boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \mathbf{b} = \delta \mathbf{a}$ des expressions pour $\mathbf{M}_1$ et $\mathbf{M}_2$, définies comme

$$\begin{aligned} \mathbf{M}_1 \mathbf{a} &:= \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \mathbf{a} = \delta\gamma \mathbf{a} = \lambda \mathbf{a} \\ \mathbf{M}_2 \mathbf{b} &:= \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \mathbf{b} = \delta\gamma \mathbf{b} = \lambda \mathbf{b} \end{aligned}$$

7. Si on trouve $\mathbf{a}$ telle que la corrélation est négative, prendre $-\mathbf{a}$ donnera une valeur positive.

Que vaut $\text{Cor}\left(\mathbf{a}^\top \boldsymbol{X}, \mathbf{b}^\top \boldsymbol{Y}\right)$ au maximum ? En multipliant par $\mathbf{a}^\top$ et par $\mathbf{b}^\top$ respectivement les équations (5.3) et (5.4), on obtient $\delta = \gamma = \mathbf{a}^\top \boldsymbol{\Sigma}_{12}\mathbf{b}$. On a donc

$$\max_{\mathbf{a},\mathbf{b}} \frac{\mathbf{a}^\top \boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma}_{11}\mathbf{a}\mathbf{b}^\top \boldsymbol{\Sigma}_{22}\mathbf{b}}} = \sqrt{\lambda_{\max}}.$$

La valeur maximale de la corrélation est donc $\sqrt{\lambda_{\max}}$ où $\sqrt{\lambda_{\max}}$ est la plus grande valeur propre de $\mathbf{M}_1$ et $\mathbf{M}_2$. Les valeurs propres non nulles sont en quantité égales, soit $k = \min\{p_1, p_2\}$. Noter que l'on a rencontré ces matrices lors du test d'hypothèse $\mathscr{H}_0 : \boldsymbol{\Sigma}_{12} = \mathbf{0}$. Le problème dans le développement ci-dessus est que ni $\mathbf{M}_1$, ni $\mathbf{M}_2$ ne sont symétriques. Il est difficile dans ce cas de trouver les valeurs propres et les vecteurs propres. En pratique, il n'y a pas de différence que l'on travaille sur $\boldsymbol{R}$ ou $\boldsymbol{S}$ ; en écrivant $\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$ pour $\boldsymbol{D}$ la matrice diagonale des variances que l'on sépare, on obtient les mêmes valeurs propres. On résout le problème en introduisant les matrices

$$\begin{aligned}\mathbf{N}_1 &= \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2} = \mathbf{K}\mathbf{K}^\top \\ \mathbf{N}_2 &= \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2} = \mathbf{K}^\top\mathbf{K}\end{aligned}$$

où $\mathbf{K} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. On montre que $\mathbf{M}_1, \mathbf{M}_2, \mathbf{N}_1, \mathbf{N}_2$ ont les mêmes valeurs propres non-nulles $\lambda_1 > \lambda_2 > \cdots > \lambda_k$. On dénote par $\boldsymbol{\alpha}_i$ (respectivement $\boldsymbol{\beta}_i$) les vecteurs propres de $\mathbf{N}_1$ (respectivement $\mathbf{N}_2$). Soit $\mathbf{a}$ vecteur propre de $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ et $\mathbf{b}$ vecteur propre de $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \mathbf{M}_2$. Alors, on peut lier les deux par le biais des relations $\boldsymbol{\alpha}_i^\top\boldsymbol{\alpha}_i = 1, \boldsymbol{\alpha}_i^\top\boldsymbol{\alpha}_j = 0, \mathbf{a}_i = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\alpha}_i$ et $\mathbf{a}_i^\top\boldsymbol{\Sigma}_{11}\mathbf{a}_i = 1$. Pareillement, on a $\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_j = 1, \boldsymbol{\beta}_i^\top\boldsymbol{\beta}_j = 0, \mathbf{b}_j = \boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\beta}_j$ et $\mathbf{b}_j^\top\boldsymbol{\Sigma}_{22}\mathbf{b}_j = 1$. Notez que $\|\mathbf{a}_i\| \neq 1$. $\sqrt{\lambda_i}$ est appelé le $i^{\text{e}}$ **coefficient de corrélation canonique**, $\mathbf{a}_i^\top\boldsymbol{X}$ et $\mathbf{b}_i^\top\boldsymbol{Y}$ sont appelés les $i^{\text{e}}$ **variables canoniques**.

**Proposition 5.1 (Propriétés)**
Posons $\eta_i = \mathbf{a}_i^\top\boldsymbol{X}$ et $\phi_i = \mathbf{b}_i\boldsymbol{Y}$ ; alors,

1. $\mathsf{E}\left(\eta_i\right) = \mathbf{a}_i^\top\boldsymbol{\mu}_1$ et $\mathsf{E}\left(\phi_i\right) = \mathbf{b}_i^\top\boldsymbol{\mu}_2$
2. $\text{Var}\left(\eta_i\right) = \mathbf{a}_i^\top\boldsymbol{\Sigma}_{11}\mathbf{a}_i = \boldsymbol{\alpha}_i^\top\boldsymbol{\alpha}_i = 1$ et $\text{Var}\left(\phi_i\right) = \mathbf{b}_i^\top\boldsymbol{\Sigma}_{22}\mathbf{b}_i = \boldsymbol{\beta}_i^\top\boldsymbol{\beta}_i = 1$
3. $\text{Cor}\left(\eta_i, \eta_j\right) = \text{Cor}\left(\phi_i, \phi_j\right) = 0$ si $i \neq j$ et $\text{Cor}\left(\eta_i, \phi_i\right) = \mathbf{a}_i^\top\boldsymbol{\Sigma}_{12}\mathbf{b}_i = \sqrt{\lambda_i}$. Ceci représente la corrélation de la $i^{\text{e}}$variable canonique en $\boldsymbol{X}$ et de la $i^{\text{e}}$variable canonique en $\boldsymbol{Y}$. On a également

$$\text{Cor}\left(\eta_i, \phi_j\right) = \text{Cov}\left(\eta_i, \phi_j\right) = \mathbf{a}_i^\top\boldsymbol{\Sigma}_{12}\mathbf{b}_j = \boldsymbol{\alpha_i}^\top\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\beta}_j = \boldsymbol{\alpha_i}^\top\mathbf{K}\boldsymbol{\beta}_j$$

On sait que $\boldsymbol{\beta}_j$ (respectivement $\boldsymbol{\alpha}_i$) est vecteur propre de $\mathbf{K}^\top\mathbf{K}$ ($\mathbf{K}\mathbf{K}^\top$). On pose donc $\mathbf{K}^\top\mathbf{K}\boldsymbol{\beta}_j = \lambda_j\boldsymbol{\beta}_j$, ce qui implique que $\mathbf{K}\mathbf{K}^\top\mathbf{K}\boldsymbol{\beta}_j = \mathbf{K}\mathbf{K}^\top d\boldsymbol{\alpha}_j = \lambda_j d\boldsymbol{\alpha}_j$. En notant que $\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j = d^{-2}\boldsymbol{\beta}_j^\top\mathbf{K}^\top\mathbf{K} = d^{-2}\lambda_j\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_j$, on vérifie que $d = \sqrt{\lambda_j}$ et il en découle que $\text{Cor}\left(\eta_i, \phi_j\right) = \boldsymbol{\alpha}_i^\top\sqrt{\lambda_j}\boldsymbol{\alpha}_j = 0$.

## Analyse canonique sur échantillon

On remplace $\boldsymbol{\Sigma}_{ij}$ par $\boldsymbol{S}_{ij}$ ou $\boldsymbol{S}_{ij}^*$. La division par $n-1$ ou $n$ est sans importance puisque ces constantes s'annulent dans le produit $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Posons $\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$; on a

$$\begin{aligned}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21} &= \left(\boldsymbol{D}_1^{1/2}\boldsymbol{R}_{11}\boldsymbol{D}_1^{1/2}\right)^{-1}\left(\boldsymbol{D}_1^{1/2}\boldsymbol{R}_{12}\boldsymbol{D}^{1/2}\right)\left(\boldsymbol{D}_2^{1/2}\boldsymbol{R}_{22}\boldsymbol{D}_2^{1/2}\right)^{-1}\left(\boldsymbol{D}_2^{1/2}\boldsymbol{R}_{21}\boldsymbol{D}_1^{1/2}\right)\\ &= \boldsymbol{D}_1^{-1/2}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\boldsymbol{D}_1^{1/2}\end{aligned}$$

et donc

$$\begin{aligned}\boldsymbol{D}_1^{-1/2}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\boldsymbol{D}_1^{1/2}\boldsymbol{v} &= \lambda\boldsymbol{v}\\ \boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\boldsymbol{D}_1^{1/2}\boldsymbol{v} &= \lambda\boldsymbol{D}_1^{1/2}\boldsymbol{v}.\end{aligned}$$

Notez que la valeur propre maximale est la même si on utilise $\boldsymbol{R}$ ou $\boldsymbol{S}$. Ainsi, le résultat de l' analyse canonique sur $\boldsymbol{R}$ ou sur $\boldsymbol{S}$ est identique.

## 5.2 Inférence en analyse canonique

On présente quelques tests d'hypothèse associés à l'analyse canonique.

(a) $\mathscr{H}_0 : \boldsymbol{\Sigma}_{12} = \boldsymbol{0}$ contre $\mathscr{H}_1 : \boldsymbol{\Sigma}_{12} \neq \boldsymbol{0}$. On a traité ce test précédemment ; dans le présent contexte, si $\boldsymbol{S}_{12} \approx \boldsymbol{0}$, ça ne vaut pas la peine de faire une analyse canonique. La ratio des log-vraisemblances est

$$-2\log(\lambda) = -n\log\left(\prod_{i=1}^{k}(1-\lambda_i)\right)$$

avec $\lambda_i$ valeur propre de $\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}\boldsymbol{S}_{12}$ et la statistique de Wilks donnée par $\lambda^{2/n} \dot{\sim} \Lambda(p_2, n-1-p_1, p_1)$. On peut utiliser l'approximation de Bartlett, soit

$$V = -\left(n - \frac{1}{2}(p_1+p_2+3)\right)\log\left(\prod_{i=1}^{k}(1-\lambda_i)\right) \overset{\mathscr{H}_0}{\dot{\sim}} \chi^2_{p_1p_2}$$

(b) $\mathscr{H}_0$ coefficients de corrélation canonique sont non nuls ($r_i = 0$), où $r_i^2 = \lambda_i$. La statistique

$$V_j = -\left(n - \frac{1}{2}(p_1+p_2+3)\right)\log\left(1-r_j^2\right)$$

Le nombre de degrés de liberté est $(p_1-1)(p_2-1)$ pour la statistique $V - V_1$ pour tester l'hypothèse voulant que les variables canoniques 2 à $k$ ne soient pas significatives, $(p_1-2)(p_2-2)$ pour la statistique $V - V_1 - V_2$ pour tester l'hypothèse nulle que les variables canoniques 3 à $k$ ne sont pas significatives, etc.

### Interprétation et généralisation

Soit les matrices $\mathbf{X}$ et $\mathbf{Y}$, assumées centrées. Les vecteurs $\mathbf{a}_j$ et $\mathbf{b}_j$ correspondent aux $j^{\text{e}}$ vecteurs de corrélation canonique et $\mathbf{Xa}_j$ et $\mathbf{Yb}_j$ représentent les scores des $n$ individus sur la $j^{\text{e}}$variable canonique, tandis que $\eta_j = \mathbf{a}_j^\top \boldsymbol{x}$ et $\phi_j = \mathbf{b}_j^\top \boldsymbol{y}$ représentent la valeur particulière des scores pour $(\boldsymbol{x}, \boldsymbol{y})$. Côté représentation graphique de $(\eta_1, \phi_1)$, on a la droite correspondante de pente $\sqrt{\widehat{\lambda}_i}$, soit

$$\widehat{\phi}_i = \sqrt{\widehat{\lambda}_i}\eta_i.$$

Si la corrélation canonique est près de 1, alors la relation est presque linéaire de pente 1. On peut tracer également $(\widehat{\eta}_1, \widehat{\eta}_2)$ versus $(\widehat{\phi}_1, \widehat{\phi}_2)$; il y a forte ressemblance entre les nuages de points si les corrélations canoniques sont fortes.

## 5.3 Généralisation : analyse canonique et tableaux de contingence

Tableau 3 – Exemple de tableau de contingence

| | | Cheveux clairs | Cheveux foncés |
|---|---|---|---|
| Yeux | bleus | 4 | 0 |
| | verts | 1 | 2 |
| | bruns | 0 | 3 |

On fabrique un tableau $\mathbf{X}$ et $\mathbf{Y}$ à partir du tableau de contingence créé précédemment, soit

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{X}^\top\mathbf{Y} = \begin{pmatrix} 4 & 0 \\ 1 & 2 \\ 0 & 3 \end{pmatrix}$$

Pour faire une analyse canonique, on doit calculer la matrice $\boldsymbol{S} = \left(\begin{smallmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{smallmatrix}\right)$. En revanche, il y a des dépendances ici : la somme des colonnes de $\mathbf{X}$ et $\mathbf{Y}$ égale $\mathbf{1}_n$ et de ce fait $\boldsymbol{S}_{11}$ et $\boldsymbol{S}_{22}$ calculées en recentrant les données ne seront pas inversibles. Les solutions sont les suivantes

1. Enlever une colonne à $\mathbf{X}$ et à $\mathbf{Y}$ et effectuer l'analyse canonique sur le résultat par

la suite.

2. Effectuer l'analyse canonique sur

$$\widetilde{S} = \frac{1}{n}\begin{pmatrix} X^\top X & X^\top Y \\ Y^\top X & Y^\top Y \end{pmatrix}.$$

   Cela donnera une valeur propre nulle, à ne pas considérer. Cela revient (en ne centrant pas) à trouver une inverse généralisé et conduit à la même solution pour les valeurs propres non triviales.

Ce résultat correspond exactement à la technique que l'école française de la statistique appelle l'analyse des correspondances. Elle est due à un algébriste du nom de Benzécri et sert principalement lors de questionnaires.

**Exemple 5.1 (Origine sociale des étudiants)**
Une enquête de l'INSEE recense la branche d'étude de 10000 étudiants en relation avec la catégorie socioprofessionnelle du père. Le résultat de l'enquête est comptabilisé dans le tableau de contingence.

Tableau 4 – Origine socioprofessionnelle et domaine d'étude d'étudiants, 1975-1976

| | DROI | SEC | LETT | SCIE | MEDD | PHAR | PLUR | IUT | Total |
|---|---|---|---|---|---|---|---|---|---|
| EXPA | 80 | 36 | 134 | 99 | 65 | 28 | 11 | 58 | 511 |
| SALA | 6 | 2 | 15 | 6 | 4 | 1 | 1 | 4 | 39 |
| PATR | 168 | 74 | 312 | 137 | 208 | 53 | 21 | 62 | 1035 |
| LICS | 470 | 191 | 806 | 400 | 876 | 164 | 45 | 79 | 3031 |
| CADM | 236 | 99 | 493 | 264 | 281 | 56 | 36 | 87 | 1552 |
| EMPL | 145 | 52 | 281 | 133 | 135 | 30 | 20 | 54 | 850 |
| OUVR | 166 | 64 | 401 | 193 | 127 | 23 | 28 | 129 | 1131 |
| SERV | 16 | 6 | 27 | 11 | 8 | 2 | 2 | 8 | 80 |
| AUTR | 305 | 115 | 624 | 247 | 301 | 47 | 42 | 90 | 1771 |
| Total | 1592 | 639 | 3093 | 1490 | 2005 | 404 | 206 | 571 | 10000 |

Les colonnes sont respectivement DROI (droit), SEC (sciences économiques), LETT (Lettres), SCIE (sciences), MEDD (médecine et dentaire), PHAR (pharmacie) et PLUR (pluridisciplinaire). Les catégories pour les lignes sont respectivement EXPA (exploitant agricole), SALA (salarié agricole), PATR (patron), LICS (profession libérale, cadre supérieur), CADM (cadre moyen), EMPL (employé), OUVR (ouvrier), SERV (personnel de service) et AUTR (autres).

La question d'intérêt est : quelles études poursuivent les enfants des diverses catégories socio-professionnelles ?

On peut analyser les proportions marginales

Le tableau des profils associés aux lignes (colonnes) permet de voir les proportions.

La moyenne des profils-lignes (l'effectif pondéré par la somme des effectifs marginaux

Tableau 5 – Proportion marginale associée

(a) Lignes

| | |
|---|---|
| EXPA | 0.051 |
| SALA | 0.004 |
| PATR | 0.103 |
| LICS | 0.303 |
| CADM | 0.155 |
| EMPL | 0.085 |
| OUVR | 0.113 |
| SERV | 0.008 |
| AUTR | 0.177 |

(b) Colonnes

| | |
|---|---|
| DROI | 0.16 |
| SEC | 0.06 |
| LETT | 0.31 |
| SCIE | 0.15 |
| MEDD | 0.20 |
| PHAR | 0.04 |
| PLUR | 0.02 |
| IUT | 0.06 |

Tableau 6 – Profil associés aux lignes

| | DROI | SEC | LETT | SCIE | MEDD | PHAR | PLUR | IUT |
|---|---|---|---|---|---|---|---|---|
| EXPA | 0.1566 | 0.0705 | 0.2622 | 0.1937 | 0.1272 | 0.0548 | 0.0215 | 0.1135 |
| SALA | 0.1538 | 0.0513 | 0.3846 | 0.1538 | 0.1026 | 0.0256 | 0.0256 | 0.1026 |
| PATR | 0.1623 | 0.0715 | 0.3014 | 0.1324 | 0.2010 | 0.0512 | 0.0203 | 0.0599 |
| LICS | 0.1551 | 0.0630 | 0.2659 | 0.1320 | 0.2890 | 0.0541 | 0.0148 | 0.0261 |
| CADM | 0.1521 | 0.0638 | 0.3177 | 0.1701 | 0.1811 | 0.0361 | 0.0232 | 0.0561 |
| EMPL | 0.1706 | 0.0612 | 0.3306 | 0.1565 | 0.1588 | 0.0353 | 0.0235 | 0.0635 |
| OUVR | 0.1468 | 0.0566 | 0.3546 | 0.1706 | 0.1123 | 0.0203 | 0.0248 | 0.1141 |
| SERV | 0.2000 | 0.0750 | 0.3375 | 0.1375 | 0.1000 | 0.0250 | 0.0250 | 0.1000 |
| AUTR | 0.1722 | 0.0649 | 0.3523 | 0.1395 | 0.1700 | 0.0265 | 0.0237 | 0.0508 |

des lignes), est le profil marginal des colonnes :

$$\sum_{i=1}^{r} \frac{n_{ij}}{n_{i\bullet}} \left( \frac{n_{i\bullet}}{n} \right) = \frac{n_{\bullet j}}{n}$$

On peut diagonaliser le produit de ces deux tableaux de profils ; la décomposition spectrale donne les valeurs propres et le pourcentage d'inertie associé.

On choisit deux composantes pour le reste de l'analyse.

La représentation simultanée est alors la suivante :

Comme dans l'analyse canonique et l'analyse en composantes principales, il faut porter

Tableau 7 – Profil associés aux colonnes

| | EXPA | SALA | PATR | LICS | CADM | EMPL | OUVR | SERV | AUTR |
|---|---|---|---|---|---|---|---|---|---|
| DROI | 0.0503 | 0.0038 | 0.1055 | 0.2952 | 0.1482 | 0.0911 | 0.1043 | 0.0101 | 0.1916 |
| SEC | 0.0563 | 0.0031 | 0.1158 | 0.2989 | 0.1549 | 0.0814 | 0.1002 | 0.0094 | 0.1800 |
| LETT | 0.0433 | 0.0048 | 0.1009 | 0.2606 | 0.1594 | 0.0909 | 0.1296 | 0.0087 | 0.2017 |
| SCIE | 0.0664 | 0.0040 | 0.0919 | 0.2685 | 0.1772 | 0.0893 | 0.1295 | 0.0074 | 0.1658 |
| MEDD | 0.0324 | 0.0020 | 0.1037 | 0.4369 | 0.1401 | 0.0673 | 0.0633 | 0.0040 | 0.1501 |
| PHAR | 0.0693 | 0.0025 | 0.1312 | 0.4059 | 0.1386 | 0.0743 | 0.0569 | 0.0050 | 0.1163 |
| PLUR | 0.0534 | 0.0049 | 0.1019 | 0.2184 | 0.1748 | 0.0971 | 0.1359 | 0.0097 | 0.2039 |
| IUT | 0.1016 | 0.0070 | 0.1086 | 0.1384 | 0.1524 | 0.0946 | 0.2259 | 0.0140 | 0.1576 |

| | Valeurs propres | % d'inertie | Cumul |
|---|---|---|---|
| 1 | 0.040 | 0.837 | 0.837 |
| 2 | 0.005 | 0.115 | 0.952 |
| 3 | 0.001 | 0.024 | 0.976 |
| 4 | 0.001 | 0.020 | 0.996 |
| 5 | 0.000 | 0.003 | 0.999 |

attention à surinterpréter la distance entre les axes et se concentrer plutôt sur les angles entre les points et les axes.

La dépendance entre le type d'études et l'origine sociale peut être résumée par deux facteurs

- le premier oppose les études de médecine, caractéristiques des fils de professions libérales et cadres supérieurs, aux études en IUT, caractéristiques des fils d'ouvriers ;
- le deuxième oppose les fils d'exploitants agricoles à ceux de la CSP « autres » et les études de pharmacie et d'IUT aux études de lettres.

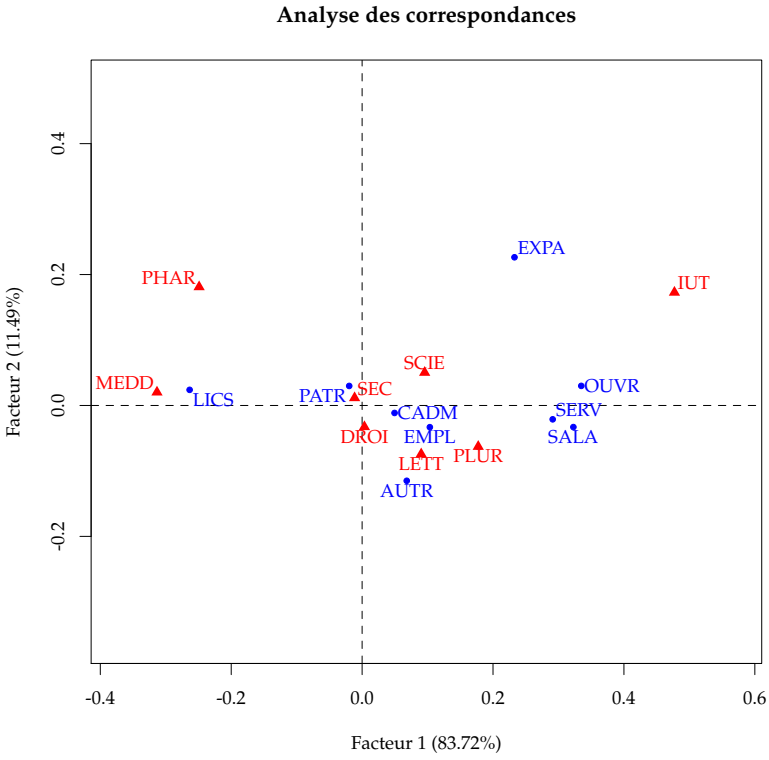


Tableau 8 – Cosinus carré des angles des vecteurs avec les sous-espaces

(a) Vecteurs lignes

| | Facteur 1 | Facteur 2 |
|---|---|---|
| EXPA | 0.50 | 0.47 |
| SALA | 0.93 | 0.01 |
| PATR | 0.06 | 0.15 |
| LICS | 0.99 | 0.01 |
| CADM | 0.39 | 0.02 |
| EMPL | 0.81 | 0.08 |
| OUVR | 0.95 | 0.01 |
| SERV | 0.81 | 0.00 |
| AUTR | 0.25 | 0.73 |

(b) Vecteurs colonnes

| | Facteur 1 | Facteur 2 |
|---|---|---|
| DROI | 0.00 | 0.31 |
| SEC | 0.03 | 0.03 |
| LETT | 0.59 | 0.40 |
| SCIE | 0.56 | 0.15 |
| MEDD | 0.98 | 0.00 |
| PHAR | 0.61 | 0.32 |
| PLUR | 0.85 | 0.11 |
| IUT | 0.87 | 0.11 |

Tableau 9 – Contribution aux inerties associées aux axes factoriels

(a) Lignes

| | Facteur 1 | Facteur 2 |
|---|---|---|
| EXPA | 0.07 | 0.48 |
| SALA | 0.01 | 0.00 |
| PATR | 0.00 | 0.02 |
| LICS | 0.53 | 0.03 |
| CADM | 0.01 | 0.00 |
| EMPL | 0.02 | 0.02 |
| OUVR | 0.32 | 0.02 |
| SERV | 0.02 | 0.00 |
| AUTR | 0.02 | 0.43 |

(b) Colonnes

| | Facteur 1 | Facteur 2 |
|---|---|---|
| DROI | 0.00 | 0.03 |
| SEC | 0.00 | 0.00 |
| LETT | 0.06 | 0.31 |
| SCIE | 0.03 | 0.07 |
| MEDD | 0.50 | 0.02 |
| PHAR | 0.06 | 0.24 |
| PLUR | 0.02 | 0.01 |
| IUT | 0.33 | 0.31 |

# Chapitre 6
# Partitionnement de données

Dans les deux prochains chapitres, nous aborderons quelques techniques de partitionnement de données et de discrimination. La partitionnement de données ("clustering") est une technique d'apprentissage non-supervisée qui consiste à grouper des observations. Les groupes sont constitués basés sur des mesures de dissimilarité (ou de distance) entre les individus. Les groupements sont inconnus, contrairement à l'analyse discriminante que nous verrons dans le prochain chapitre.

À la base de la partitionnement de données, on trouve le concept de dissimilarité . Soit $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k \in \mathbb{R}^p$. Une mesure de dissimilarité satisfait (1) $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$, (2) $d(\boldsymbol{x}_i, \boldsymbol{x}_i) = 0$ et (3) $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = d(\boldsymbol{x}_j, \boldsymbol{x}_i)$. Si en plus $d$ respecte l'inégalité du triangle, (4) $d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq d(\boldsymbol{x}_i, \boldsymbol{x}_k) + d(\boldsymbol{x}_k, \boldsymbol{x}_j)$, alors $d$ définit une distance.

On considère quelques distances, notamment

- Euclidienne : $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^\top \mathbf{A} (\boldsymbol{x} - \boldsymbol{y})}$ avec $\mathbf{A} = \mathbf{I}$ ou plus généralement. On prend souvent $\mathbf{A} = \boldsymbol{S}^{-1}$
- Minkowski :

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left( \sum_{i=1}^{p} |x_i - y_i|^m \right)^{\frac{1}{m}},$$

  appelée Manhattan (ou "city block") quand $m = 1$ et euclidienne si $m = 2$.
- Canberra $d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{p} |x_i - y_i| / (|x_i| + |y_i|)$.

À partir de la distance, on peut construire facilement une similarité ,

$$s_{ik} = \frac{1}{1 + d_{ik}},$$

où $d_{ik}$ représente la distance entre l'objet $i$ et l'objet $k$. Si on dispose d'une matrice de similarité définie non-négative, on peut construire $d_{ik} = \sqrt{2(1 - s_{ik})}$.

Un exemple de dissimilarité entre observations est $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \rho_{ij}$, où $\rho_{ij}$ est la corrélation linéaire de Pearson.

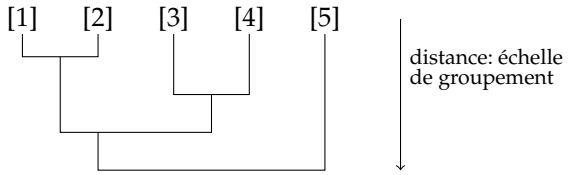
## 6.1 Méthodes de regroupement hiérarchiques

Elles consistent à

- agglomérer des individus ou des groupes d'individus,
- partitionner de plus en plus.

Le résultat est présenté sous la forme d'un dendrogramme. Dans le dendrogramme ci-dessus, on commence par obtenir la distance minimale entre les observations, ici

FIGURE 5 – Exemple de dendrogramme




entre [1] et [2]. On continue itérativement, en groupant [3] et [4], puis le groupe [12] et [34] et finalement en incluant [5] à [1234]. On dénote ci-dessous par $d_{k(ij)}$ la distance entre l'observation $k$ et les groupes $i$ et $j$ (formés respectivement de $n_i \geq 1$ et $n_j \geq 1$ observations).

**Proposition 6.1 (Méthodes hiérarchiques)**

1. distance unique (*single linkage*) : distance minimale $d_{k(ij)} = \min(d_{ki}, d_{kj})$, correspond au plus proche voisin
2. distance complète (*complete linkage*) : distance maximale $d_{k(ij)} = \max(d_{ki}, d_{kj})$, voisins les plus éloignés

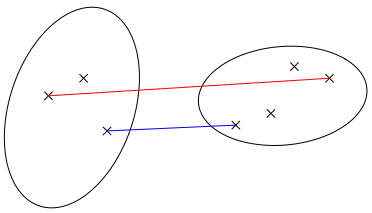


FIGURE 6 – Illustration de la distance unique (bleu) et de la distance complète (rouge)

3. distance moyenne (*average linkage*) : moyenne de toutes les distances,

$$d_{k(ij)} = \frac{1}{n_i n_j} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d_{ij}$$

4. barycentre (*centroid*) : distance entre les deux moyennes,

$$d_{k(ij)} = \left\| \bar{\boldsymbol{X}}_k - \left( \frac{n_i}{n_i + n_j} \bar{\boldsymbol{X}}_i + \frac{n_j}{n_i + n_j} \bar{\boldsymbol{X}}_j \right) \right\|^2$$

5. médiane

$$d_{k(ij)} = \left\| \bar{\boldsymbol{X}}_k - \frac{1}{2} \left( \bar{\boldsymbol{X}}_i + \bar{\boldsymbol{X}}_j \right) \right\|^2,$$

correspondant à la distance entre $\bar{\boldsymbol{X}}_k$ et $\bar{\boldsymbol{X}} = (\bar{\boldsymbol{X}}_i + \bar{\boldsymbol{X}}_j)/2$.

6. méthode de Ward : basée sur la somme des carrés (SC). La distance entre deux groupements est la somme des carrés inter-groupes,

$$d_{ij} = \mathsf{SC}_{\boldsymbol{B}} = n_i\|\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}}\|^2 + n_j\|\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}}\|^2 = \frac{n_i n_j}{n_i + n_j}\|\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}}_j\|^2$$

avec

$$\bar{\boldsymbol{X}} = \frac{n_i\bar{\boldsymbol{X}}_i + n_j\bar{\boldsymbol{X}}_j}{n_i + n_j}.$$

Pour recalculer les distance entre les groupes, on utilise la formule de Lance et Williams, soit

$$d_{k(i,j)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma|d_{ki} - d_{kj}|.$$

Le tableau suivant résumé les relations pour les différents modèles.

| | $\alpha_i$ | $\alpha_j$ | $\beta$ | $\gamma$ |
|---|---|---|---|---|
| Simple | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $-\frac{1}{2}$ |
| Complet | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{2}$ |
| Moyenne | $\frac{n_i}{n_i+n_j}$ | $\frac{n_j}{n_i+n_j}$ | 0 | 0 |
| Centroïde | $\frac{n_i}{n_i+n_j}$ | $\frac{n_j}{n_i+n_j}$ | $-\alpha_i\alpha_j$ | 0 |
| Médiane | $\frac{1}{2}$ | $\frac{1}{2}$ | $-\frac{1}{4}$ | 0 |
| Ward | $\frac{n_k+n_i}{n_k+n_i+n_j}$ | $\frac{n_k+n_j}{n_k+n_i+n_j}$ | $-\frac{n_k}{n_k+n_i+n_j}$ | 0 |

Dans ces méthodes hiérarchiques, le choix du nombre de groupes est souvent fait en maximisant l'expression

$$\frac{\mathrm{tr}(\widehat{\boldsymbol{B}})/(g-1)}{\mathrm{tr}(\widehat{\boldsymbol{W}})/(n-g)}$$

ou en examinant le dendrogramme et en choisissant un niveau adéquat. Cela permet éventuellement de déterminer les aberrances si une observation rejoint le reste seulement à la fin.

## 6.2 Méthodes non hiérarchiques

La problème général de la partitionnement de données optimale en un nombre prédéfini de $g$ groupements est en général NP complet, puisque le nombre de partitions de $n$ individus en $g$ groupes est de l'ordre $\mathrm{O}(g^n/g!)$.

Une alternative consiste à procéder à une affectation initiale des observations en $g$ groupements et réassigner les individus de façon à optimiser un critère à chaque étape jusqu'à convergence du processus de réassignation. Le résultat dépend de la répartition

initiale des groupes, lequel n'est donc pas anodin, puisqu'il peut conduire à des solutions finales différentes si l'algorithme se termine.

Les $K$-moyennes consiste à minimiser $\text{tr}(\widehat{\boldsymbol{W}})$ ou maximiser $\text{tr}(\widehat{\boldsymbol{B}})$.

**Algorithme 6.1 (Partitionnement de données par $K$-moyennes (Lloyd))**
Soit $\mathbf{X}$ un tableau d'observations $n \times p$ et $K$ un nombre pré-spécifié de groupements.

1. Assignation initiale des observations aux $K$ groupements et calcul des centroïdes $\bar{\mathbf{x}}_k$ ou pré-spécification des centroïdes $\bar{\mathbf{x}}_k$ pour $k = 1, \ldots, K$.
2. Calculer la distance Euclidienne carrée entre chaque observation et le centroïde $\bar{\mathbf{x}}_k$ de son groupe d'appartenance $c(\mathbf{x}_i) \in \{c_1, \ldots, c_K\}$

$$\text{ESS} = \sum_{k=1}^{K} \sum_{i:c(\mathbf{x}_i)=c_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)^\top (\mathbf{x}_i - \bar{\mathbf{x}}_k).$$

3. Assigner : $c_k \leftarrow \{\mathbf{x}_j : \|\mathbf{x}_j, \bar{\mathbf{x}}_k\| \leq \|\mathbf{x}_j, \bar{\mathbf{x}}_i\|, i = 1, \ldots, K\}, k = 1, \ldots, K$
4. Mise à jour des centroïdes : $\bar{\mathbf{x}}_\mathbf{k} \leftarrow \text{card}^{-1}(c_k) \sum_{i:c(\mathbf{x}_i)=c_k} \mathbf{x}_i$.
5. Itérer les étapes 3. et 4. tant que des observations sont réassignés à des groupes

L'algorithme de $K$-moyennes minimise le critère ESS, qui décroît avec chaque itération et donc la convergence en temps fini est garantie. Comme indiqué précédemment, l'obtention d'une solution optimale n'est pas garantie avec cet algorithme glouton et dépend de l'allocation initiale des centroïdes ; en pratique, on essaiera plusieurs valeurs initiales. L'algorithme de $K$-moyenne tend à créer des groupes sphériques de taille équivalente, puisque la structure de variance est la même pour chaque groupement.

On peut aussi procéder en utilisant d'autres critères de réallocation.

- minimiser $|\widehat{\boldsymbol{W}}|$, maximiser $|\widehat{\boldsymbol{T}}|/|\widehat{\boldsymbol{W}}|$
- maximiser $\text{tr}(\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{B}})$

Une variante des $K$-moyennes, les $K$-médianes, minimise plutôt la distance de Manhattan ($\ell_1$) entre les observations et la médiane de chaque groupe, à savoir $|\mathbf{x}_k - \text{med}_k(\mathbf{x})|$. Une autre variante utilise comme représentant de groupe une des observations, appelé médoïde, en lieu et place des centroïdes. Le critère de minimisation est basé sur une dissimilarité ou une distance. Ce critère est plus robuste face aux données aberrantes. Deux variantes existent : le partitionnement autour de médoïdes (PAM) est le plus utilisé, puisqu'il utilise un algorithme glouton.

**Algorithme 6.2 ($K$-médoïdes et partitionnement autour de médoïdes)**
Soit $\mathbf{D} = (d_{ij})$ une matrice $n \times n$ de proximité et $K$ un nombre pré-spécifié de groupements.

1. Assignation initiale des items en $K$ groupements et localisation du médoïde pour chaque groupe. Le médoïde du groupe $k$ est l'observation $m_k$ qui minimise la dissimilarité totale $\sum_{i \in c_k} d_{im_k}$.

$K$-médoïdes :

(a) Pour le $k^{\text{e}}$groupe, réassigner l'observation $i_k$ au groupe la plus près telle que la fonction objective

$$\text{ESS}_{\text{med}} = \sum_{k=1}^{K} \sum_{i:c(i)=c_k} d_{ii_k}$$

est réduite.

(b) Répéter l'étape 3 et la réassignation jusqu'à ce que plus aucune réassignation d'observations n'ait lieu.

Partitionnement autour de médoïdes :

(a) Pour chaque groupe, interchanger le médoïde avec l'observation qui réduit le plus la fonction objective $\text{ESS}_{\text{med}}$

(b) Répéter les permutations jusqu'à ce que plus aucune réduction de $\text{ESS}_{\text{med}}$ n'ait lieu.

Un diagnostic graphique utile pour les méthodes de partitionnement de données est la silhouette, qui est basée la matrice de proximité $\mathbf{D}$. Pour chaque observation, la moyenne des distances intra-classe est calculée pour l'observation $i$, dénotée $a_i$, ainsi que la plus petite distance moyenne inter-classe pour l'observation, dénotée $b_i$. Plus précisément, on a

$$\begin{aligned} a_i &= \frac{1}{\text{card}(c(i)) - 1} \sum_{j \in c(i)} d(i,j) \\ b_i &= \min_{\substack{k \in \{1,\ldots,K\} \\ k \neq c(i)}} \frac{1}{\text{card}(c_k)} \sum_{j \in c_k} d(i,j). \end{aligned}$$

La largeur de la silhouette pour $i$ pour une méthode de partitionnement de données avec $K$ groupements est donnée par

$$s_{i,K} = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

et $-1 \leq s_{i,K} \leq 1$. Des grandes valeurs positives ($a_i \approx 0$) indiquent que la $i^{\text{e}}$observation est bien représentée par le groupe, des valeurs négatives ($b_i \approx 0$) indiquent une mauvaise allocation et $s_{i,K}$ presque nulle indique que l'observation se situe entre deux groupements. Une règle subjectives veut que si $\max_i s_{i,K} < 0.25$ , aucune structure substantielle n'a été détectée et qu'entre 0.26 et 0.50, la structure soit potentiellement artificielle.

Graphiquement, on recherche des groupes plutôt homogènes et une moyenne des silhouettes, $n^{-1} \sum_{i=1}^{n} s_{i,K}$, élevée.

**Exemple 6.1 (Interprétation de silhouettes)**

La figure 7 montre le résultat de la partitionnement de données à l'aide de l'algorithme

des $K$-moyennes sur un mélange de 4 lois binormales. Le diagnostic visuel indique dans le cas $K = 3$ que certaines observations sont probablement plus susceptibles d'appartenir à un autre groupe (en l'occurrence rouge). Le groupe vert est trop dispersé. Avec $K = 5$, la séparation des groupements bleu-rouge est artificielle.

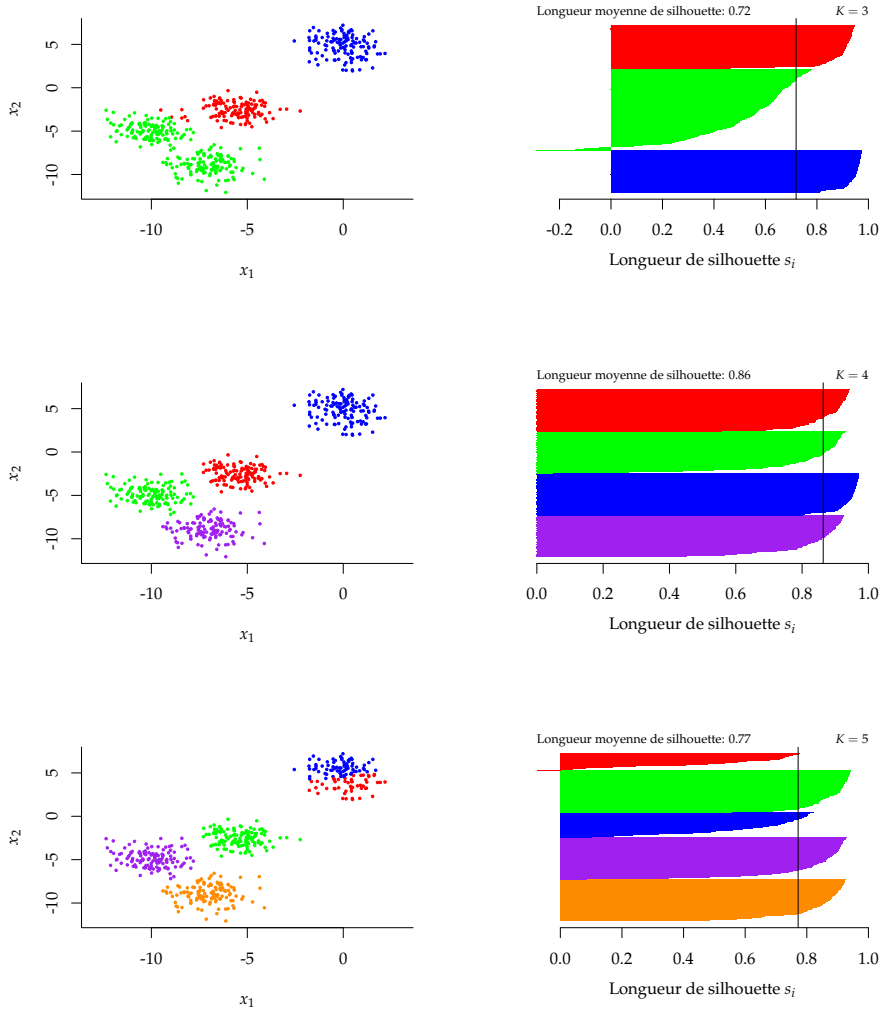



FIGURE 7 – Nuages de points et silhouettes pour la partitionnement de données de données avec l'algorithme des $K$-moyennes

## 6.3 Méthodes basées sur un modèle

Soit un échantillon $\mathbf{X} = (\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)^\top$ où $\boldsymbol{X}_i$ est un $p$-vecteur colonne. On suppose qu'on a $K$ groupes, chacun caractérisé par une densité de dimension $p$, soit $f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$ si

$\boldsymbol{X}_i$ provient du groupe $k$ pour $k = 1, \ldots, K$. On réécrit la vraisemblance en fonction de $\pi_k$, la probabilité qu'une observation tombe dans le groupe $k$,

$$L(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K; \pi_1, \ldots, \pi_K, \mathbf{X}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i, \boldsymbol{\theta}_k).$$

Dans le cas de données manquantes (comme la classe d'une observation), on peut éliminer ces données, remplacer par la moyenne des colonnes (ce qui réduit la variance, alors que les données manquantes devraient augmenter l'incertitude), faire une régression et prédire les données manquantes, voire faire de l'imputation multiple. La dernière solution consiste à utiliser l'algorithme de Dempster, Laird & Rubin (1977). Nous utilisons l'algorithme d'espérance-maximisation puisque dans le cas présent l'information est manquante.

On prend pour la densité

$$f_k(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left( -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right).$$

Nous allons prendre pour la matrice de covariance

$$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{P}_k \boldsymbol{A}_k \boldsymbol{P}_k^\top \tag{6.5}$$

avec $\boldsymbol{P}_k$ une matrice orthogonale de vecteurs propres, $\boldsymbol{A}_k$ une matrice diagonale dont les éléments sont proportionnels aux valeurs propres et $\lambda_k$ un scalaire. $\boldsymbol{P}_k$ gouverne l'orientation de l'ellipsoïde, $\boldsymbol{A}_k$ sa forme et $\lambda_k$ son volume, lequel est proportionnel à $\lambda_k^p \det(A_k)$.

| $\boldsymbol{\Sigma}_k$ | **distribution** | **volume** | **forme** | **orientation** |
|---|---|---|---|---|
| $\lambda \mathbf{I}$ | sphérique | constant | constant | non-définie |
| $\lambda_k \mathbf{I}$ | sphérique | variable | constant | non-définie |
| $\lambda \boldsymbol{P} \boldsymbol{A} \boldsymbol{P}^\top$ | elliptique | constant | constant | constant |
| $\lambda_k \boldsymbol{P}_k \boldsymbol{A}_k \boldsymbol{P}_k^\top$ | elliptique | variable | variable | variable |
| $\lambda \boldsymbol{P}_k \boldsymbol{A} \boldsymbol{P}_k^\top$ | elliptique | constant | constant | variable |
| $\lambda_k \boldsymbol{P}_k \boldsymbol{A} \boldsymbol{P}_k^\top$ | elliptique | variable | constant | variable |

Le choix du nombre de composantes, de même que la forme du modèle, est basée sur la sélection bayésienne de modèles et l'utilisation du facteur de Bayes. Spécifiquement, si plusieurs modèles $M_1, \ldots, M_O$ sont considérés avec chacun probabilité a priori $p(M_o)$ (souvent $1/O$) pour un ensemble de données $\mathbf{X}$, on a par le théorème de Bayes la loi postérieure du modèle $o \in \{1, \ldots, O\}$

$$p(M_o \mid \mathbf{X}) \propto p(\mathbf{X} \mid M_o) p(M_o)$$

où la vraisemblance intégrée $p(\mathbf{X} \mid M_o)$ est obtenue en intégrant les paramètres inconnus du modèle et est donnée par la loi totale de probabilité

$$p(\mathbf{X} \mid M_o) = \int p(\mathbf{X} \mid \boldsymbol{\theta}_o, M_o) p(\boldsymbol{\theta}_o \mid M_o) \, \mathrm{d}\boldsymbol{\theta}_o$$

Le facteur de Bayes pour le modèle $i$ versus $j$ est $B_{ij} = p(\mathbf{X} \mid M_i)/p(\mathbf{X} \mid M_j)$ et la comparaison est favorable au modèle $i$ si $B_{ij} > 1$ et il y a de très fortes preuves en faveur de $M_i$ si $B_{ij} > 100$.

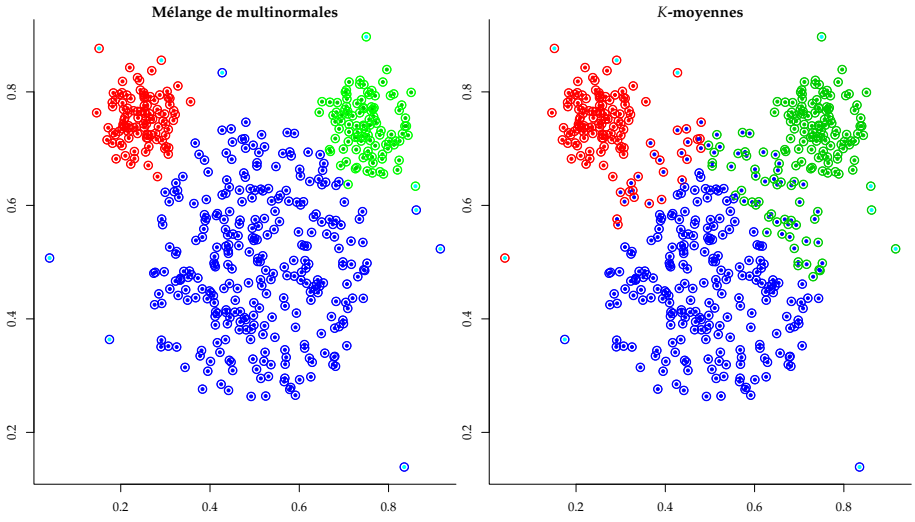
La vraisemblance intégrée $p(\mathbf{X} \mid M_o)$ est approximée par le critère d'information bayésien, ou BIC, donné par

$$\mathsf{BIC}_o = -2 \log p(\mathbf{X} \mid M_o) \approx -2 \log p(\mathbf{X} \mid \widehat{\boldsymbol{\theta}}_o, M_o) + \nu_o \log(n)$$

où $\nu_o = |\boldsymbol{\theta}|$ est le nombre de paramètres inconnus estimés dans le modèle $o$. On cherche à minimiser le BIC. On peut aussi prendre comme critère de comparaisons des modèles l'AIC ou le $C_p$ de Mallows.

Dans le cas de groupements ellipsoïdales, les mélanges de multinormales peuvent mieux capturer certaines partitions inhomogènes, comme l'atteste cet exemple avec ce mélange de normales en forme de souris. Le résultat de la partitionnement de données est indiqué par des cercles, et la couleur des points indique l'appartenance aux différentes classes.

FIGURE 8 – Regroupemen pour le jeu de données `mickey-mouse`



## 6.4 Algorithme d'espérance-maximisation

L'algorithme d'espérance-maximisation est un algorithme itératif d'optimisation numérique, introduit par Dempster, Laird et Rubin (1977), qui permet d'obtenir les estimés du maximum de vraisemblance dans des problèmes avec données manquantes, c'est-à-dire pour lesquelles les données du modèles sont partiellement observées. L'algorithme produisant une séquence d'estimés qui converge vers l'estimé du maximum de vraisemblance des données incomplètes à un rythme linéaire. Dans certains modèles, on peut aussi augmenter les données pour obtenir une vraisemblance complète plus tractable et ainsi faciliter l'analyse.

Soit le couple $(\boldsymbol{Y}, \boldsymbol{U})$ de données complètes. Les données $\boldsymbol{U}$ sont les données manquantes ou variables latentes, tandis que les données observées $\boldsymbol{Y}$, aussi appelées données incomplètes. La vraisemblance pour les données complètes est

$$L_{\boldsymbol{Y},\boldsymbol{U}}(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) = f_{\boldsymbol{Y},\boldsymbol{U}|\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{u}; \theta),$$

tandis que la vraisemblance pour les données incomplètes est

$$L_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) = f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y}; \theta) = \int f_{\boldsymbol{Y},\boldsymbol{U}|\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{u}; \theta) \, \mathrm{d}\boldsymbol{u}.$$

L'algorithme EM facilite la maximisation de la vraisemblance pour les données incomplètes $L(\boldsymbol{\theta}; \boldsymbol{Y})$ en travaillant avec la vraisemblance des données complètes et la distribution conditionnelle

$$f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}) = \frac{f_{\boldsymbol{U},\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{u}, \boldsymbol{y} \mid \boldsymbol{\theta})}{f_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{\theta})}$$

On peut donc exprimer la log-vraisemblance des données incomplètes comme

$$\ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) = \ell_{\boldsymbol{U},\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{u}, \boldsymbol{y}) - \ell_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{\theta}; \boldsymbol{u}, \boldsymbol{y}). \tag{6.6}$$

Si on prend l'espérance de (6.6) par rapport à la densité conditionnelle $f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{u} \mid \boldsymbol{y}, \boldsymbol{\theta}')$ pour $\boldsymbol{\theta}' \subset \Theta$, on obtient

$$\begin{aligned}
\ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) &= \mathsf{E}_{f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}}\left(\ell_{\boldsymbol{U},\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{U}, \boldsymbol{Y}) \mid \boldsymbol{y}, \boldsymbol{\theta}'\right) - \mathsf{E}_{f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}}\left(\ell_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}'}(\boldsymbol{\theta}; \boldsymbol{U}, \boldsymbol{Y}) \mid \boldsymbol{y}, \boldsymbol{\theta}'\right) \\
&= Q(\boldsymbol{\theta}; \boldsymbol{\theta}') - \mathsf{E}_{f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}}\left(\ell_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}'}(\boldsymbol{\theta}; \boldsymbol{U}, \boldsymbol{Y}) \mid \boldsymbol{y}, \boldsymbol{\theta}'\right)
\end{aligned}$$

attendu que la log-vraisemblance des données observées ne dépend pas de $\boldsymbol{U}$. On conditionne sur une valeur spécifique $\boldsymbol{\theta}'$ lors du calcul de l'espérance conditionnelle.

L'algorithme d'espérance maximisation est un algorithme itératif qui produit une séquence d'estimés qui convergent vers l'estimé du maximum de vraisemblance des don-

nées incomplètes. À partir d'une valeur initiale $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, la $r+1^{\text{e}}$ valeur de la séquence, est construite selon

$$\boldsymbol{\theta}^{(r+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathsf{E}_{f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}} \left( \ell_{\boldsymbol{U},\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{U}, \boldsymbol{Y}) \mid \boldsymbol{y}, \boldsymbol{\theta}^{(r)} \right) = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}).$$

L'algorithme comporte deux étapes :

- étape E : calcul de l'espérance conditionelle de la log-vraisemblance pour les données complètes
- étape M : maximisation de ladite espérance.

**Preuve** On veut montrer que $\ell_{\boldsymbol{Y}}(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{y}) \geq \ell_{\boldsymbol{Y}}(\boldsymbol{\theta}^{(r)}; \boldsymbol{y})$. Puisque l'on choisit $\boldsymbol{\theta}^{(r+1)}$ qui maximise la fonction $Q$, on a $Q(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{\theta}^{(r)}) \geq Q(\boldsymbol{\theta}^{(r)}; \boldsymbol{\theta}^{(r)})$ et il suffit de démontrer que

$$\mathsf{E}_{f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}^{(r)}}} \left( \ell_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}^{(r)}}(\boldsymbol{\theta}^{(r)}; \boldsymbol{U}, \boldsymbol{Y}) \right) \geq \mathsf{E}_{f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}^{(r)}}} \left( \ell_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}^{(r)}}(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{U}, \boldsymbol{Y}) \right)$$

Or, pour toutes fonctions de densité $f_1, f_2$ pour une variable aléatoire $\boldsymbol{U}$ donnée, on a

$$\begin{aligned}
\mathsf{E}_{f_1}(\log f_1(\boldsymbol{U})) - \mathsf{E}_{f_1}(\log f_2(\boldsymbol{U})) &= -\mathsf{E}_{f_1}\left( \log\left( \frac{f_2(\boldsymbol{U})}{f_1(\boldsymbol{U})} \right) \right) \\
&\geq -\log \mathsf{E}_{f_1}\left( \frac{f_2(\boldsymbol{U})}{f_1(\boldsymbol{U})} \right) \\
&= -\log \int \frac{f_2(\boldsymbol{u})}{f_1(\boldsymbol{u})} f_1(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \\
&= -\log \int f_2(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} = 0
\end{aligned}$$

par l'inégalité de Jensen. En appliquant ce résultat avec $f_1 = f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}^{(r)}}$ et $f_2 = f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}^{(r+1)}}$, on obtient que la séquence $\{\ell_{\boldsymbol{Y}}(\boldsymbol{\theta}^{(r)})\}_{r=1}^{\infty}$ converge vers un maximum local de la log-vraisemblance $\ell_{\boldsymbol{Y}}(\boldsymbol{\theta})$ si la fonction est bornée. ■

Des méthodes pour accélérer la convergence d'EM existent. Puisque l'algorithme EM converge vers un maximum local, il peut être nécessaire de démarrer l'algorithme à plusieurs valeurs de départ pour traiter les cas où la fonction objective est multimodale. L'algorithme est très facile à coder, bien que les calculs pour l'espérance puissent être difficiles en premier lieu. On peut déterminer la convergence en calculant (si disponible) $|\ell_{\boldsymbol{Y}}(\boldsymbol{\theta}^{(r+1)}; \boldsymbol{y}) - \ell_{\boldsymbol{Y}}(\boldsymbol{\theta}^{(r)}; \boldsymbol{y})| < \varepsilon$ ou $|\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}| < \delta$ pour des paramètres de tolérance numérique $\delta$ et $\varepsilon$.

**Exemple 6.2 (Données manquantes pour une variable binormale)**
Soit $\boldsymbol{W} = (W_1, W_2)^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ des données tirées d'une loi bivariée normale avec moyenne $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ et matrice de covariance $\boldsymbol{\Sigma} = \left(\begin{smallmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{smallmatrix}\right)$. Les paramètres du modèle sont $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})^\top$. Supposons que l'échantillon obtenu contient $m$ données complètes $\boldsymbol{w}_j = (w_{1j}, w_{2j})^\top$ pour $j = 1, \ldots, m$ ainsi que $\boldsymbol{w}_{2j} = (?, w_{2j})^\top$, $j = m+1, \ldots, m+m_1$ paire de données partiellement observées auxquelles manquent

la première coordonnée et finalement $\boldsymbol{w}_{1j} = (w_{1j}, ?)^\top$, $j = m + m_1 + 1, \ldots, n$ les $m_2 = n - m - m_1$ paires auxquelles manquent la deuxième coordonnée.

La log-vraisemblance des données observées est donc

$$\begin{aligned}\ell(\boldsymbol{\theta}; \boldsymbol{W}) = &-\frac{n+m}{2}\log(2\pi) - \frac{1}{2}m\log(\det(\boldsymbol{\Sigma})) - \frac{1}{2}\sum_{j=1}^{m}(\boldsymbol{w}_j - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{w}_j - \boldsymbol{\mu}) \\ &- \frac{1}{2}m_1\log(\sigma_{22}) - \frac{1}{2}m_2\log(\sigma_{11}) - \sum_{j=m+m_1+1}^{n}\frac{(w_{1j}-\mu_1)^2}{2\sigma_{11}} - \sum_{j=m+1}^{m+m_1}\frac{(w_{2j}-\mu_2)^2}{2\sigma_{22}}\end{aligned}$$

Les données complètes consistent ici des $n$ paires d'observations et la log-vraisemblance complète est donnée par

$$\begin{aligned}\ell_{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{W}, \boldsymbol{U}) = &-n\log(2\pi) - \frac{n\log(\xi)}{2} - \frac{1}{2\xi}\left[\sigma_{22}T_{11} + \sigma_{11}T_{22}\right. \\ &- 2\sigma_{12}T_{12} - 2\{T_1(\mu_1\sigma_{22} - \mu_2\sigma_{12}) + T_2(\mu_2\sigma_{11} - \mu_1\sigma_{12})\} \\ &\left.+ n(\mu_1^2\sigma_{22} + \mu_2^2\sigma_{11} - 2\mu_1\mu_2\sigma_{12})\right]\end{aligned}$$

où

$$T_i = \sum_{j=1}^{n} w_{ij}, \qquad T_{hi} = \sum_{j=1}^{n} w_{hj}w_{ij}, \qquad \xi = \sigma_{11}\sigma_{22}(1-\rho^2), \qquad \rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}.$$

où $h, i = 1, 2$. La distribution normale est une famille exponentielle régulière avec statistique exhaustive $(T_1, T_2, T_{12}, T_{11}, T_{22})^\top$. Les estimateurs du maximum de vraisemblance pour les données complètes sont

$$\widehat{\mu}_i = \frac{T_i}{n}, \qquad \widehat{\sigma}_{hi} = \frac{T_{hi}}{n} - \frac{T_h T_i}{n^2}, \qquad h, i = 1, 2.$$

Pour calculer $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = \mathsf{E}\left(\log f(\boldsymbol{Y}, \boldsymbol{U}; \boldsymbol{\theta}) \mid \boldsymbol{y}, \boldsymbol{\theta}^{(r)}\right)$, on doit évaluer

$$\begin{aligned}&\mathsf{E}_{\boldsymbol{\theta}^{(r)}}\left(W_{1j} \mid w_{2j}\right), \quad \mathsf{E}_{\boldsymbol{\theta}^{(r)}}\left(W_{1j}^2 \mid w_{2j}\right), \quad j = m+1, \ldots, m+m_1 \\ &\mathsf{E}_{\boldsymbol{\theta}^{(r)}}\left(W_{2j} \mid w_{1j}\right), \quad \mathsf{E}_{\boldsymbol{\theta}^{(r)}}\left(W_{2j}^2 \mid w_{1j}\right), \quad j = m+m_1+1, \ldots, n\end{aligned}$$

Or on sait par les propriétés de la loi multinormale que $W_2 \mid w_1$ suit une loi normale

$$W_2 \mid W_1 = w_1 \sim \mathcal{N}\left(\mu_2 + \sigma_{12}\sigma_{11}^{-1}(w_1 - \mu_1), \sigma_{22}(1-\rho^2)\right).$$

Ainsi, l'étape E à l'itération $r$ de l'algorithme EM prend la forme

$$\mathsf{E}_{\boldsymbol{\theta}^{(r)}}\left(W_{2j} \mid w_{1j}\right) = w_{2j}^{(r)} = \mu_2^{(r)} + \frac{\sigma_{12}^{(r)}}{\sigma_{11}^{(r)}}(w_{1j} - \mu_1^{(r)})$$

$$\mathsf{E}_{\boldsymbol{\theta}^{(r)}}\left(W_{2j}^2 \mid w_{1j}\right) = \left(w_{2j}^{(r)}\right)^2 + \sigma_{22.1}^{(r)},$$

pour $j = m + m_1 + 1, \ldots, n$. Le même raisonnement s'applique pour $w_{1j}$ pour $j = m + 1, \ldots, m + m_1$.

Dans l'étape M de l'algorithme, on remplace les valeurs manquantes $w_{ij}$ et $w_{ij}^2$ pour $i = 1, 2$ par leurs espérances conditionnelles, substituées dans les expressions pour $T_i, T_{hi}$, soit

$$\mu_i^{(r+1)} = \frac{T_i^{(r)}}{n}, \qquad \sigma_{hi}^{(r+1)} = \frac{T_{hi}^{(r)}}{n} - \frac{T_h^{(r)} T_i^{(r)}}{n^2}, \qquad h, i = 1, 2.$$

On considère un cas simplifié où les observations manquent à un seul groupe.

Tableau 10 – Données manquantes pour une variable binormale)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\boldsymbol{w}_1$ | 8 | 11 | 16 | 18 | 6 | 4 | 20 | 25 | 9 | 13 |
| $\boldsymbol{w}_2$ | 10 | 14 | 16 | 15 | 20 | 4 | 18 | 22 | ? | ? |

Tableau 11 – Premières 10 itérations pour l'algorithme EM de l'exemple 6.2

| $r$ | $\mu_1$ | $\mu_2$ | $\sigma_{11}$ | $\sigma_{22}$ | $\sigma_{12}$ | $-2\ell(\boldsymbol{\theta}^{(r)})$ |
|---|---|---|---|---|---|---|
| 1 | 13 | 14.05537 | 40.2 | 130.53371 | 23.22854 | 125.27 |
| 2 | 13 | 14.47994 | 40.2 | 47.40122 | 21.20237 | 116.26 |
| 3 | 13 | 14.58502 | 40.2 | 30.89012 | 20.95190 | 113.49 |
| 4 | 13 | 14.60853 | 40.2 | 27.58288 | 20.89990 | 113.09 |
| 5 | 13 | 14.61375 | 40.2 | 26.92016 | 20.88843 | 113.03 |
| 6 | 13 | 14.61490 | 40.2 | 26.78735 | 20.88588 | 113.02 |
| 7 | 13 | 14.61516 | 40.2 | 26.76073 | 20.88532 | 113.01 |
| 8 | 13 | 14.61522 | 40.2 | 26.75539 | 20.88519 | 113.01 |
| 9 | 13 | 14.61523 | 40.2 | 26.75432 | 20.88516 | 113.01 |
| 10 | 13 | 14.61523 | 40.2 | 26.75411 | 20.88516 | 113.01 |

**Exemple 6.3 (Mélanges finis)**
Soit un échantillon $\mathbf{X}$ tiré d'un mélange fini de $K$ composantes. La vraisemblance d'une observation $\boldsymbol{X}_i \in \mathbb{R}^p$ est alors

$$f_{\boldsymbol{X}_i|\boldsymbol{\theta}}(\boldsymbol{x}_i \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k), \qquad \sum_{k=1}^{K} \pi_k = 1$$

pour $0 < \pi_k < 1$. On peut augmenter la vraisemblance en considérant l'indicateur de composante $C_k$, dénoté $Z_{ik} = \mathsf{I}(C_i = k)$ pour chaque observation $i$ et ainsi se débarrasser de la somme. Les données complètes sont donc constituées des tuples $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ avec $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})$. Les $z_{ik}$ correspond dans le cas présent aux variables latentes et

suivent une distribution multinomiale avec 1 essai. La log-vraisemblance peut s'écrire

$$\ell(\boldsymbol{\theta}_k, \pi_k; \mathbf{z}, \mathbf{x}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left( \log\left(\pi_k\right) + \log\left(f_k(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k)\right)\right).$$

À l'étape $r$ de l'algorithme, on impute $\boldsymbol{z}_i$ par son espérance conditionnelle $\boldsymbol{\gamma}_i^{(r)}$ sachant $\boldsymbol{\pi}^{(r)}, \boldsymbol{\theta}_1^{(r)}, \ldots, \boldsymbol{\theta}_K^{(r)}$, qui découle de la formule de Bayes. La distribution conditionnelle pour $C_k$ est une distribution discrète sur $\{1, \ldots, K\}$ et puisque $\{C_i = k\} \equiv \{Z_{ik} = 1\}$

$$\mathsf{E}_{f_{\boldsymbol{Z}_i \mid \boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\pi}}} \left(\mathsf{I}\left(C_i = k\right) \mid \boldsymbol{x}_i, \boldsymbol{\pi}, \boldsymbol{\theta}\right) = \mathsf{P}_{f_{\boldsymbol{Z}_i \mid \boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\pi}}} \left(C_i = k \mid \boldsymbol{x}_i, \boldsymbol{\pi}, \boldsymbol{\theta}\right)$$

que l'on peut écrire comme

$$\mathsf{P}_{f_{\boldsymbol{Z}_i \mid \boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\pi}}} \left(Z_{ik} = 1 \mid \boldsymbol{x}_i, \boldsymbol{\pi}, \boldsymbol{\theta}\right) = \frac{\pi_k f_k(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k)}{\sum_{j=1}^{K} \pi_j f_j(\boldsymbol{x}_i \mid \boldsymbol{\theta}_j)} = \gamma_k(\boldsymbol{x}_i) \tag{6.7}$$

où les variables latentes $Z_{ik}$ sont conditionnellement indépendantes.

Dès lors,

$$\begin{aligned} Q(\boldsymbol{\pi}, \boldsymbol{\theta} \mid \boldsymbol{\pi}^{(r)}, \boldsymbol{\theta}^{(r)}) &= \mathsf{E}_{f_{\boldsymbol{Z}_i \mid \boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\pi}}} \left(\ell(\boldsymbol{\theta}_k, \pi_k; \boldsymbol{z}_{ik}, \mathbf{x})\right) \\ &= \sum_{k=1}^{K} \sum_{i=1}^{n} \gamma_k^{(r)}(\boldsymbol{x}_i) \log(\pi_k) + \sum_{k=1}^{K} \sum_{i=1}^{n} \gamma_k^{(r)}(\boldsymbol{x}_i) \log(f_k(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k)) \end{aligned} \tag{6.8}$$

On optimise alors chacune des termes séparément de façon à obtenir $\boldsymbol{\theta}_1^{(r+1)}, \ldots, \boldsymbol{\theta}_K^{(r+1)}$ et $\boldsymbol{\pi}^{(r+1)}$. Le premier terme de l'équation (6.8) correspond à une vraisemblance multinomiale, et donc comme précédemment, on obtient l'estimateur du maximum de vraisemblance

$$\pi_k^{(r+1)} = \frac{\sum_{i=1}^{n} \gamma_k^{(r)}(\boldsymbol{x}_i)}{\sum_{j=1}^{K} \sum_{i=1}^{n} \gamma_j^{(r)}(\boldsymbol{x}_i)}.$$

Le deuxième terme peut être maximisé séparément, numériquement ou analytiquement selon la nature de $f_1, \ldots, f_K$. On pose

$$\boldsymbol{\theta}_k^{(r+1)} = \arg\max_{\boldsymbol{\theta}_k} \sum_{i=1}^{n} \gamma_k^{(r)}(\boldsymbol{x}_i) \log(f_k(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k))$$

Dans le cas particulier où $f_k$ est une loi multinormale, les expressions pour le maximum de vraisemblance de $\boldsymbol{\theta}_k \equiv (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ pour $k = 1, \ldots, K$ sont explicites.

**Algorithme 6.3 (Espérance-maximisation pour mélange fini de lois multinormales)**

- Initialiser $\gamma_k^{(0)}(\boldsymbol{x}_i)$ pour $i = 1, \ldots, n$ et $k = 1, \ldots, K$
- Pour itération $r = 1, \ldots$ jusqu'à convergence
  - Étape M : calculer les paramètres qui maximisent la vraisemblance étant donné

$\gamma_k^{(r-1)}(\boldsymbol{x}_i)$, soit

$$n_k^{(r)} = \sum_{i=1}^{n} \gamma_k^{(r-1)}(\boldsymbol{x}_i), \qquad \pi_k^{(r)} = \frac{n_k^{(r)}}{n}, \qquad \boldsymbol{\mu}_k^{(r)} = \frac{1}{n_k^{(r)}} \sum_{i=1}^{n} \gamma_k^{(r-1)}(\boldsymbol{x}_i)\boldsymbol{x}_i$$

et $\boldsymbol{\Sigma}_k^{(r)}$ dépend de la structure de la matrice de covariance (voir équation (6.5)).

– Étape E : imputer les valeurs $z_{ik}$ par $\gamma_k^{(r)}(\boldsymbol{x}_i)$ étant donnés les paramètres estimés à l'étape M,

$$\gamma_k^{(r)}(\boldsymbol{x}_i) = \frac{\pi_k^{(r)} f_k\big(\boldsymbol{\mu}_k^{(r)}, \boldsymbol{\Sigma}_k^{(r)} \mid \boldsymbol{x}_i\big)}{\sum_{j=1}^{K} \pi_j^{(r)} f_j\big(\boldsymbol{\mu}_j^{(r)}, \boldsymbol{\Sigma}_j^{(r)} \mid \boldsymbol{x}_i\big)}$$

Dans notre problème, il peut arriver que la matrice $\widehat{\boldsymbol{\Sigma}}_k$ soit mal conditionnée.[8] La convergence de l'algorithme EM peut être lente si le nombre de groupe est élevé.

## Estimation de la variance avec l'algorithme EM - cas régulier

Si les algorithmes de type quasi-Newton offrent un estimé de la matrice de variance par le biais de la matrice hessienne, les écarts-types ne sont pas automatiquement disponibles avec l'algorithme EM. Nous assumerons subséquemment que des conditions de régularité permettent d'interchanger dérivée et intégrale.

La fonction de score des données incomplètes est donnée par

$$\begin{aligned} S(\boldsymbol{y};\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\ell_{\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{y}) &= \mathsf{E}_{f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ell_{\boldsymbol{U},\boldsymbol{Y}|\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{U},\boldsymbol{Y}) \mid \boldsymbol{y},\boldsymbol{\theta}\right) \\ &= \mathsf{E}_{f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}}\left(S_{\mathsf{c}}(\boldsymbol{u},\boldsymbol{y};\boldsymbol{\theta})\right) \\ &= \left[\frac{\partial}{\partial\boldsymbol{\theta}}Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')\right]_{\boldsymbol{\theta}'=\boldsymbol{\theta}}. \end{aligned} \tag{6.9}$$

Cette autoconsistance de l'algorithme justifie le choix particulier de l'étape M ; au maximum de vraisemblance, l'estimé $\widehat{\boldsymbol{\theta}}$ résout l'équation de score $S(\boldsymbol{y},\widehat{\boldsymbol{\theta}}) = 0$.

On peut décomposer la variance en fonction de l'information pour les données complètes et l'information manquante, données respectivement par

$$\begin{aligned} \mathcal{I}_{\boldsymbol{U}|\boldsymbol{Y}}(\boldsymbol{\theta};\boldsymbol{y}) &= -\mathsf{E}_{f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_{\boldsymbol{U}|\boldsymbol{Y}}(\boldsymbol{\theta};\boldsymbol{U},\boldsymbol{Y}) \mid \boldsymbol{y}\right) \\ \mathcal{I}_{\boldsymbol{U},\boldsymbol{Y}}(\boldsymbol{\theta}) &= -\mathsf{E}_{f_{\boldsymbol{U},\boldsymbol{Y}|\boldsymbol{\theta}}}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_{\boldsymbol{U},\boldsymbol{Y}}(\boldsymbol{\theta};\boldsymbol{U},\boldsymbol{Y})\right). \end{aligned}$$

8. S'il y a un groupe avec un seul individu $\boldsymbol{x}_l$, on aura alors $\widehat{\boldsymbol{\mu}}_k = \boldsymbol{x}_l$, $\widehat{\boldsymbol{\Sigma}}_k \to \mathbf{O}_d$ et la log-vraisemblance tendra vers l'infini.

En dérivant équation (6.6) par rapport à $\boldsymbol{\theta}$ deux fois, on obtient

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_{\boldsymbol{Y}}(\boldsymbol{\theta};\boldsymbol{y}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_{\boldsymbol{U},\boldsymbol{Y}}(\boldsymbol{\theta};\boldsymbol{u},\boldsymbol{y}) - \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{u},\boldsymbol{y}).$$

En prenant l'espérance par rapport à $f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}$ à $\boldsymbol{Y} = \boldsymbol{y}$, on obtient

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\ell_{\boldsymbol{Y}}(\boldsymbol{\theta};\boldsymbol{y}) = \mathcal{I}_{\boldsymbol{U},\boldsymbol{Y}}(\boldsymbol{\theta}) - \mathcal{I}_{\boldsymbol{U}|\boldsymbol{Y}}(\boldsymbol{\theta};\boldsymbol{y}).$$

La première contribution est l'information maximale que l'on peut obtenir en n'ayant pas observé $\boldsymbol{U}$ en prenant l'espérance sur tous les $\boldsymbol{U}$ possibles ; le deuxième terme est la quantité d'information manquante, conséquence des données manquantes, que l'on doit soustraire. En prenant l'espérance cette fois par rapport à $f_{\boldsymbol{Y}|\boldsymbol{\theta}}$, on obtient

$$\mathcal{I}_{\boldsymbol{Y}}(\boldsymbol{\theta}) = \mathcal{I}_{\boldsymbol{U},\boldsymbol{Y}}(\boldsymbol{\theta}) - \mathsf{E}_{f_{\boldsymbol{Y}|\boldsymbol{\theta}}}\left(\mathcal{I}_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{Y})\right)$$

Un article de Louis (1982) montre que l'on peut exprimer l'information manquante

$$\mathcal{I}_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{Y}) = \mathsf{E}_{f_{\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\theta}}}\left(S_{\mathsf{c}}(\boldsymbol{u},\boldsymbol{y};\boldsymbol{\theta})S_{\mathsf{c}}^\top(\boldsymbol{u},\boldsymbol{y};\boldsymbol{\theta})\right) - S(\boldsymbol{y};\boldsymbol{\theta})S(\boldsymbol{y};\boldsymbol{\theta})^\top$$

en utilisant équation (6.9). Or, la fonction de score au maximum de vraisemblance est égale à 0, donc le terme de droite disparaît. On peut donc obtenir la matrice d'information observée à $\widehat{\boldsymbol{\theta}}$ comme

$$\begin{aligned}\mathcal{I}_{\boldsymbol{Y}}(\widehat{\boldsymbol{\theta}}) &= \mathcal{I}_{\boldsymbol{U},\boldsymbol{Y}}(\widehat{\boldsymbol{\theta}}) - \mathsf{E}_{f_{\boldsymbol{U}|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}}}\left(S_{\mathsf{c}}(\boldsymbol{u},\boldsymbol{y};\widehat{\boldsymbol{\theta}})S_{\mathsf{c}}^\top(\boldsymbol{u},\boldsymbol{y};\widehat{\boldsymbol{\theta}})\right)\\ &= \mathcal{I}_{\boldsymbol{U},\boldsymbol{Y}}(\widehat{\boldsymbol{\theta}}) - \mathsf{Var}_{f_{\boldsymbol{U}|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}}}\left(S_{\mathsf{c}}(\boldsymbol{u},\boldsymbol{y};\widehat{\boldsymbol{\theta}})\right).\end{aligned}$$

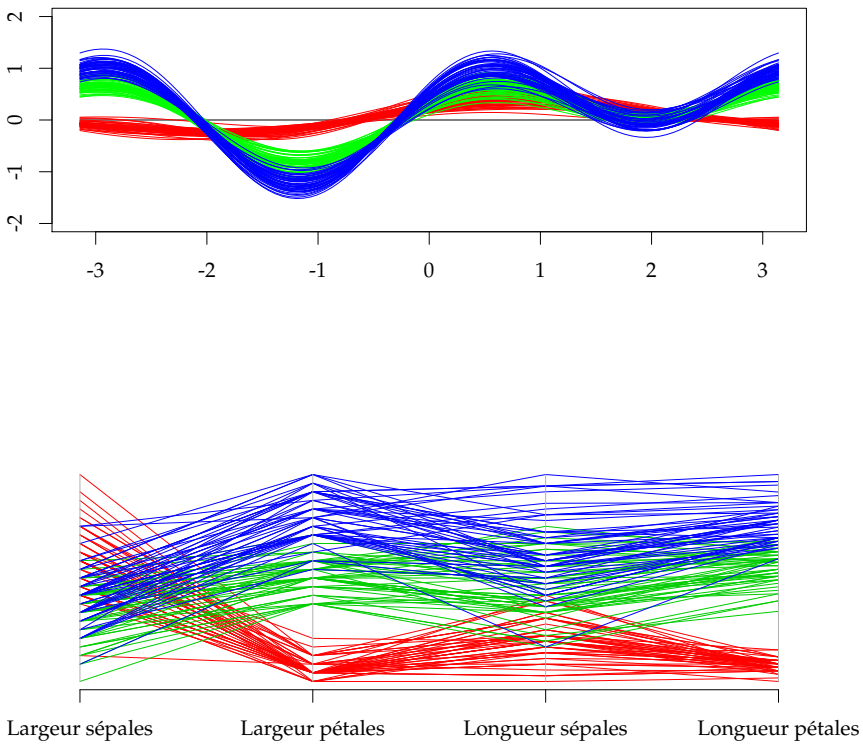
## Détection d'aberrances et diagnostics visuels

Une analyse exploratoire des données (à l'aide de nuages de points) peut être utile pour diagnostiquer les groupements et détecter les aberrances. D'autres possibilités incluent les profils (coordonnées parallèles), les diagrammes en étoiles, les faces de Chernoff, les ACP et les diagrammes Andrews, correspondant à

$$f_x(t) = \frac{x_1}{\sqrt{2}} + x_2\sin(t) + x_3\cos(t) + \cdots$$

Les couleurs correspondent aux étiquettes des groupements. Si les courbes s'empilent, alors les individus sont ressemblants.

FIGURE 9 – Diagrammes d'Andrews et en coordonnées parallèles pour le jeu de données `iris`



# Chapitre 7
# Analyse discriminante et classification

L'analyse discriminante vise deux buts :

1. Séparation : décrire les caractères différentiels d'observations provenant de populations connues. Essayer de trouver les caractères discriminants qui séparent le plus une collection d'individus.
2. Classification (ou allocation) : trier les observations en deux ou plusieurs groupes. Obtenir une règle optimale d'allocation qui permet d'assigner un individu nouveau à l'un des groupes.

En pratique, les deux buts se mélangent.

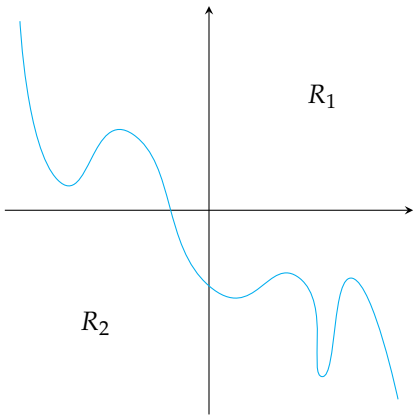
## 7.1 Le problème général de la classification

Soit deux populations $\Pi_1, \Pi_2$ de densités $f_1(x), f_2(x)$, respectivement et $R_i$ la région d'allocation à la population $i$. On dénotera par

$$\mathsf{P}\left(R_2 \mid \Pi_1\right) = \mathsf{P}\left(X \in R_2 \mid X \sim \Pi_1\right) = \int_{R_2} f_1(x)\,\mathrm{d}x$$
$$\mathsf{P}\left(R_1 \mid \Pi_2\right) = \mathsf{P}\left(X \in R_1 \mid X \sim \Pi_2\right) = \int_{R_1} f_2(x)\,\mathrm{d}x$$

où $R_2 = \Omega \setminus R_1$ pour $\Omega \coloneqq R_1 \sqcup R_2$. Cela revient à classifier un individu de la population $\Pi_2$ en 1 et réciproquement. On dénote également $p_i$ la probabilité a priori de classer un individu dans $\Pi_i$ et $\mathsf{C}\left(R_i \mid \Pi_j\right)$ le coût de mauvaise classification d'une observation de $\Pi_j$ dans $\Pi_i$. Le coût espéré de mauvaise classification, dénoté CEMC est

FIGURE 10 – Deux régions d'allocation en 2D




donné par

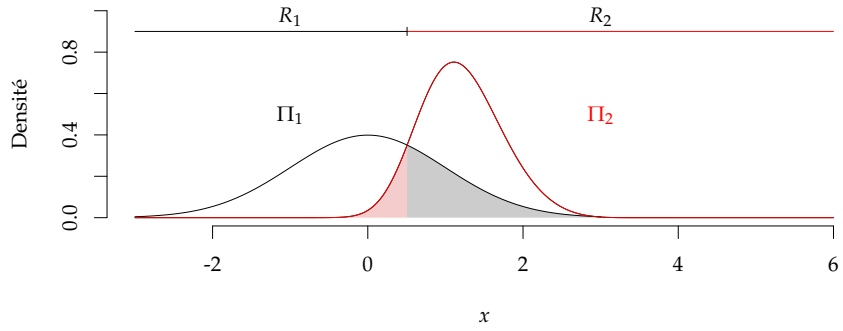
$$\mathsf{CEMC} = \mathsf{C}\left(R_2 \mid \Pi_1\right)\mathsf{P}\left(R_2 \mid \Pi_1\right)p_1 + \mathsf{C}\left(R_1 \mid \Pi_2\right)\mathsf{P}\left(R_1 \mid \Pi_2\right)p_2$$

**Proposition 7.1**
Les régions $R_1$ et $R_2$ qui minimisent ce coût CEMC sont

$$R_1 := \left\{ \frac{f_1(x)}{f_2(x)} \geq \frac{\mathsf{C}\left(R_1 \mid \Pi_2\right) p_2}{\mathsf{C}\left(R_2 \mid \Pi_1\right) p_1} \right\}, \qquad R_2 := \left\{ \frac{f_1(x)}{f_2(x)} < \frac{\mathsf{C}\left(R_1 \mid \Pi_2\right) p_2}{\mathsf{C}\left(R_2 \mid \Pi_1\right) p_1} \right\}.$$

En pratique, on sélectionne souvent des probabilités à priori égales, de sorte qu'on a le ratio des densités à comparer avec 1.

FIGURE 11 – Discrimination de deux populations et erreur de classification




**Preuve** Par définition,

$$\begin{aligned}
\mathsf{CEMC} &= \mathsf{C}\left(R_2 \mid \Pi_1\right) p_1 \int_{R_2} f_1(x)\,\mathrm{d}x + \mathsf{C}\left(R_1 \mid \Pi_2\right) p_2 \int_{R_1} f_2(x)\,\mathrm{d}x \\
&= \mathsf{C}\left(R_2 \mid \Pi_1\right) p_1 \left(1 - \int_{R_1} f_1(x)\,\mathrm{d}x\right) + \mathsf{C}\left(R_1 \mid \Pi_2\right) p_2 \int_{R_1} f_2(x)\,\mathrm{d}x \\
&= \int_{R_1} \left(\mathsf{C}\left(R_1 \mid \Pi_2\right) p_2 f_2(x) - \mathsf{C}\left(R_2 \mid \Pi_1\right) p_1 f_1(x)\right)\,\mathrm{d}x + \mathsf{C}\left(R_2 \mid \Pi_1\right) p_1
\end{aligned}$$

en notant que tous les termes sont positifs. Le minimum sera donc atteint lorsque l'intégrant est plus petite ou égale à zéro. Ainsi, $R_1$ sera telle que $\mathsf{C}\left(R_1 \mid \Pi_2\right) p_2 f_2(x) - \mathsf{C}\left(R_2 \mid \Pi_1\right) p_1 f_1(x) \leq 0$, c'est-à-dire tel que

$$\frac{f_1(x)}{f_2(x)} \geq \frac{\mathsf{C}\left(R_1 \mid \Pi_2\right) p_2}{\mathsf{C}\left(R_2 \mid \Pi_1\right) p_1}.$$

Pour $R_2$, c'est évidemment le contraire. ■

On peut adopter d'autres stratégies

1. Minimiser la probabilité totale de mauvaise classification (revient au cas précédent avec $\mathsf{C}(R_1 \mid \Pi_2) = \mathsf{C}(R_2 \mid \Pi_1)$.
2. Allouer une nouvelle observation $X_0$ à la population avec la plus grande probabilité a posteriori

$$\mathsf{P}(\Pi_1 \mid X_0) = \frac{\mathsf{P}(X_0 \mid \Pi_1)\, p_1}{\mathsf{P}(X_0 \mid \Pi_1)\, p_1 + \mathsf{P}(X_0 \mid \Pi_2)\, p_2} \dot{\sim} \frac{p_1 f_1(x_0)}{p_1 f_1(x_0) + p_2 f_2(x_0)}$$
$$\mathsf{P}(\Pi_2 \mid X_0) = \frac{\mathsf{P}(X_0 \mid \Pi_2)\, p_1}{\mathsf{P}(X_0 \mid \Pi_1)\, p_1 + \mathsf{P}(X_0 \mid \Pi_2)\, p_2} \dot{\sim} \frac{p_2 f_2(x_0)}{p_1 f_1(x_0) + p_2 f_2(x_0)}$$

Cela revient au cas (1).

La généralisation à plus de deux populations (disons $g$) du CEMC allouera l'observation $X_0 = x_0$ à la population $\Pi_k$ pour laquelle $\sum_{i=1, i\neq k}^{g} f_i(x) p_i \mathsf{C}(R_k \mid \Pi_i)$ est la plus petite. Si $\mathsf{C}(R_k \mid \Pi_i) = 1$ (toutes sont égales), cette expression sera la plus petite si $p_k f_k(x) > p_i f_i(x)$ pour $i \neq k$.

**Exemple 7.1 (Application du CEMC à deux lois normales)**
Soit $\Pi_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ et $\Pi_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$

1. Cas $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ : la densité est donnée par

$$f_i(\boldsymbol{x}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)\right)$$

et donc

$$R_1 : \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \log\left(\frac{\mathsf{C}(R_1 \mid \Pi_2)\, p_2}{\mathsf{C}(R_2 \mid \Pi_1)\, p_1}\right)$$
$$R_2 : \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < \log\left(\frac{\mathsf{C}(R_1 \mid \Pi_2)\, p_2}{\mathsf{C}(R_2 \mid \Pi_1)\, p_1}\right)$$

Dans ce premier cas, la discrimination est une droite formée par les points d'intersection des ellipses. En pratique, on utilisera les estimateurs usuels et on remplacera $\boldsymbol{\mu}_1$ par $\bar{\boldsymbol{x}}_1$, $\boldsymbol{\mu}_2$ par $\bar{\boldsymbol{x}}_2$ et $\boldsymbol{\Sigma}$ par $\boldsymbol{S}_p$. La fonction de classification/discriminant d'Anderson sera donnée par

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_p^{-1} \boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_p^{-1} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \geq \log\left(\frac{\mathsf{C}(R_1 \mid \Pi_2)\, p_2}{\mathsf{C}(R_2 \mid \Pi_1)\, p_1}\right)$$

2. Le cas où $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ mène à

$$R_1 : \quad -\frac{1}{2}\boldsymbol{x}^\top (\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}) \boldsymbol{x} + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1}) \boldsymbol{x} - C \geq \log\left(\frac{\mathsf{C}(R_1 \mid \Pi_2)\, p_2}{\mathsf{C}(R_2 \mid \Pi_1)\, p_1}\right)$$

où $C$ est donné par

$$C = \frac{1}{2}\log\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} + \frac{1}{2}\left(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right)$$

La structure de la courbe dépend de $\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}$, qui peut être une parabole ou une ellipse, mais peut ne pas être définie positive.

**Exemple 7.2 (Cas de plusieurs lois normales)**
Supposons que $\mathsf{C}\,(R_k \mid \Pi_i) = 1$ ; allouer $\boldsymbol{x}$ à la population $\Pi_k$ si $p_k f_k(x) > p_i f_i(\boldsymbol{x})$ pour $i \neq k$, c'est-à-dire allouer $\boldsymbol{x}$ à $\Pi_k$ pour laquelle

$$-\frac{1}{2}\log\left(|\boldsymbol{S}_i|\right) - \frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{x}}_i)^\top \boldsymbol{S}_i^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_i)^\top + \log(p_i).$$

est le plus grand. Si on fait en plus l'hypothèse que les $\boldsymbol{\Sigma}_i$ sont égaux, $\boldsymbol{x}$ est alloué à la population $\Pi_k$ pour laquelle

$$-\frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{x}}_i)^\top \boldsymbol{S}_p^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_i)^\top + \log(p_i)$$

est le plus grand. La quantité qui apparaît plus haut n'est rien d'autre que la distance de Mahalanobis $D_i^2$ vue précédemment.

## 7.2 Analyse discriminante et classification par la méthode de Fisher

Nous entamerons l'étude de la méthode pour deux populations. Soit $\boldsymbol{X} = \boldsymbol{x}$ un vecteur d'observations de dimension $p$ appartenant à une population, soit $\Pi_1$ ou $\Pi_2$, tel que $\mathsf{E}\,(\boldsymbol{X} \mid \Pi_1) = \boldsymbol{\mu}_1$ et $\mathsf{E}\,(\boldsymbol{X} \mid \Pi_2) = \boldsymbol{\mu}_2$ et $\boldsymbol{\Sigma} = \mathsf{Var}\,(\boldsymbol{X})$. L'idée est d'effectuer la transformation $\boldsymbol{Y} = \mathbf{a}^\top\boldsymbol{X}$ tel que $\mu_{1Y} = \mathbf{a}^\top\boldsymbol{\mu}_1$, $\mu_{2Y} = \mathbf{a}^\top\boldsymbol{\mu}_2$, $\sigma_Y^2 = \mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}$ qui séparera le mieux les points projetés, c'est-à-dire celle qui maximisera

$$\frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2} = \frac{\mathbf{a}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\mathbf{a}}{\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}} = \frac{(\mathbf{a}^\top\boldsymbol{\delta})^2}{\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}}$$

où $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. On cherche donc à maximiser

$$\max_{\mathbf{a}} \frac{(\mathbf{a}^\top\boldsymbol{\delta})^2}{\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}} = \boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}.$$

En utilisant la décomposition de Choleski pour la matrice de covariance, $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$ et $\boldsymbol{\Sigma}^{-1} = \mathbf{L}^{-\top}\mathbf{L}^{-1}$ et en posant $\mathbf{z} = \mathbf{L}^\top\mathbf{a}$, $\mathbf{y} = \mathbf{L}^{-1}\boldsymbol{\delta}$, on peut appliquer l'inégalité de Cauchy–Schwartz (eq. 3.1) et ainsi obtenir

$$(\mathbf{a}^\top\boldsymbol{\delta})^2 \leq \mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}\boldsymbol{\delta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$$

avec égalité si $\mathbf{a} = k\boldsymbol{\Sigma}^{-1}\delta$. Le maximum vaut donc $\delta^\top\boldsymbol{\Sigma}^{-1}\delta$ (d'où l'égalité) et il est atteint en prenant $\mathbf{a} = \boldsymbol{\Sigma}^{-1}\delta$. La combinaison linéaire $\mathbf{y} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ est appelée fonction discriminante de Fisher.

La règle d'allocation est la suivante : posons

$$m = \frac{1}{2}(\mu_{1Y} + \mu_{2Y}) = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

et soit $\boldsymbol{x}$ à allouer à

$$\begin{array}{ll} \Pi_1 & \text{si } (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \geq m \\ \Pi_2 & \text{si } (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} < m \end{array}$$

avec en pratique $\bar{x}_1, \bar{x}_2$ et $S_p = ((n_1 - 1)S_1 + (n_2 - 1)S_2)/(n_1 + n_2 - 2)$.

Les graphiques suivants illustrent la classification de 75 femmes en deux populations (30 individus sains et 45 hémophiles) en fonction de l'activité et de la quantité de facteur anti-hémophilique A (FAH) présent dans le sang, ainsi que la classification pour un mélange de lois binormales.

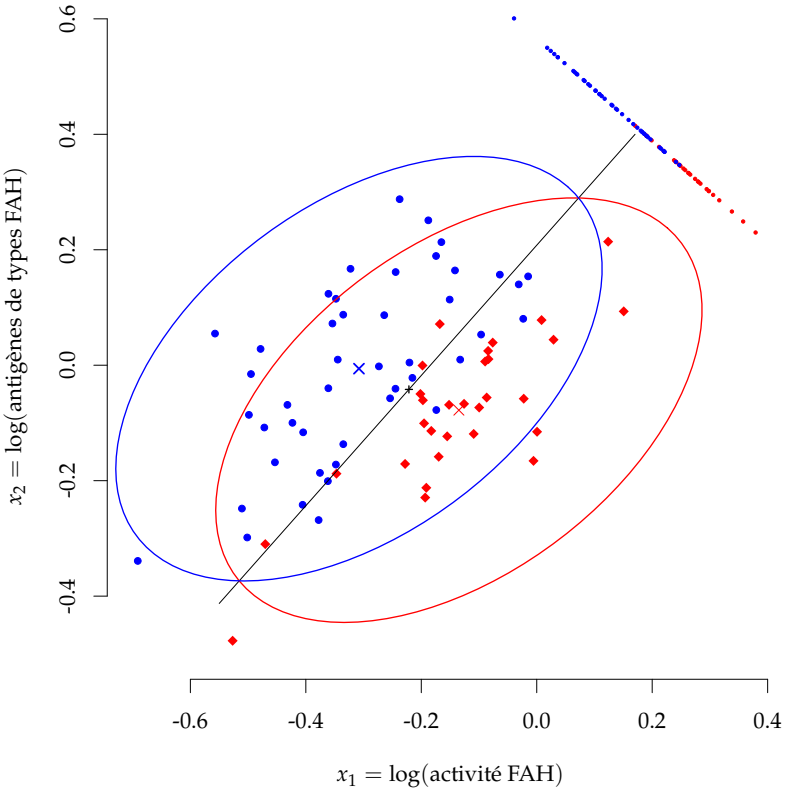



FIGURE 12 – Discrimination linéaire et méthode de Fisher pour données d'hémophilie

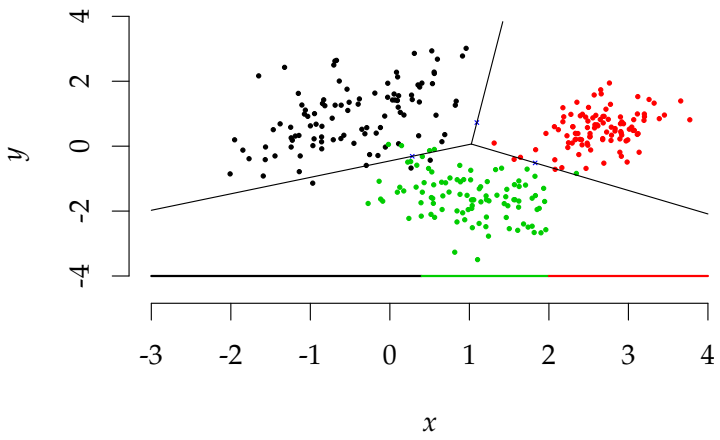



FIGURE 13 – Discrimination linéaire et méthode de Fisher pour trois populations binormales

En projetant dans une direction, on obtient le même critère dans le cas de lois multinormales en dimension deux. Le résultat diffère dans le cas général, comme l'illustre la Figure 13.

**Proposition 7.2 (Méthode de Fisher pour plusieurs populations)**
L'idée est de déterminer un nombre restreint de combinaisons linéaires $\mathbf{a}_1^\top \boldsymbol{X}, \mathbf{a}_2^\top \boldsymbol{X}, \ldots$ utiles pour séparer les populations $\Pi_1, \ldots, \Pi_g$ pour $g \geq 2$. On va supposer l'égalité des matrices de variance-covariance $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$. On dénotera

$$\bar{\boldsymbol{\mu}} := \frac{1}{g} \sum_{j=1}^{g} \boldsymbol{\mu}_j$$

et

$$\boldsymbol{B} = \sum_{j=1}^{g} (\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})^\top,$$

la matrice de variance inter-groupe. On considère $Y = \mathbf{a}^\top \boldsymbol{X}$, avec $\mathbf{a}^\top \boldsymbol{\mu}_j = \mu_{jY}$ ainsi que $\mathsf{Var}\left(Y \mid \Pi_j\right) = \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}$, $\bar{\mu}_Y = \mathbf{a}^\top \bar{\boldsymbol{\mu}}$. Formons le quotient

$$\frac{\sum_{j=1}^{g} \left(\mu_{jY} - \bar{\mu}_Y\right)^2}{\sigma_Y^2} = \frac{\mathbf{a}^\top \sum_{j=1}^{g} \left(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}}\right)\left(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}}\right)^\top \mathbf{a}}{\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}} = \frac{\mathbf{a}^\top \boldsymbol{B} \mathbf{a}}{\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}}.$$

Maximisons $\mathbf{a}^\top \boldsymbol{B}\mathbf{a}$ sujet à la contrainte $\mathbf{a}^\top \boldsymbol{\Sigma}\mathbf{a} = 1$, avec Lagrange, soit

$$\mathcal{L} = \mathbf{a}^\top \boldsymbol{B}\mathbf{a} - \lambda\left(\mathbf{a}^\top \boldsymbol{\Sigma}\mathbf{a} - 1\right)$$

et donc

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}} = 2\boldsymbol{B}\mathbf{a} - 2\lambda\boldsymbol{\Sigma}\mathbf{a} = 0,$$

qui donne $\boldsymbol{B}\mathbf{a} = \lambda\boldsymbol{\Sigma}\mathbf{a}$ ou $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\mathbf{a} = \lambda\mathbf{a}$. On retombe sur un problème de valeurs propres–vecteurs propres. La maximisation successive (avec orthonormalité par rapport à $\boldsymbol{\Sigma}$) de cette expression est donnée par les vecteurs propres de $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$ et est atteinte pour les $\min(p, g-1)$ vecteurs propres correspondants (car $\text{rang}(\boldsymbol{B}) = g-1$). On vérifie que $\text{Var}\left(\mathbf{a}_i^\top \boldsymbol{X}\right) = 1$ et $\text{Cov}\left(\mathbf{a}_i^\top \boldsymbol{X}, \mathbf{a}_j^\top \boldsymbol{X}\right) = \delta_{ij}$.

En pratique, on calcule les valeurs propres et les vecteurs propres de $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$, puisque comme lors de l'analyse canonique, la matrice de départ n'est pas symétrique. Si $\lambda_i$ est valeur propre de $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$ pour $i = 1, \ldots, k = \min(g-1, p)$ et $\mathbf{e}_i$ est vecteur propre de $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$, alors $\mathbf{a}_i = \boldsymbol{\Sigma}^{-1/2}\mathbf{e}_i$ et de ce fait $\mathbf{a}_i^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{a}_j = \delta_{ij}$.

Dans le cas d'un échantillon, on remplace $\boldsymbol{\mu}_i$ par $\bar{\boldsymbol{X}}_i$ et $\boldsymbol{\Sigma}$ par

$$\boldsymbol{S}_p = \frac{\sum_i (n_i - 1)\boldsymbol{S}_i}{\sum_i (n_i - 1)} = \frac{\widehat{\boldsymbol{W}}}{\sum_i n_i - g}$$

avec comme à l'accoutumée

$$\bar{\boldsymbol{X}} = \frac{\sum_{j=1}^g n_j \bar{\boldsymbol{X}}_j}{\sum_{j=1}^g n_j}$$

$$\widehat{\boldsymbol{B}} = \sum_j n_j \left(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}}\right)\left(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}}\right)^\top$$

$$\widehat{\boldsymbol{W}} = \sum_{j=1}^g (n_j - 1)\boldsymbol{S}_j.$$

La fonction $\boldsymbol{Y}_i = \widehat{\mathbf{a}}_i^\top \boldsymbol{x}$, où $\widehat{\mathbf{a}}_i$ est le vecteur propre correspondant à la valeur propre $\widehat{\lambda}_i$ de $\boldsymbol{S}_p^{-1}\widehat{\boldsymbol{B}} \propto \widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{B}}$ avec $\widehat{\mathbf{a}}_i^\top \boldsymbol{S}_p \widehat{\mathbf{a}}_j = \delta_{ij}$, est appelé $i^{\text{e}}$ fonction discriminante.

## Allocation

Considérons une observation hypothétique $\boldsymbol{X}$ et $Y_j = \mathbf{a}_j^\top \boldsymbol{X}$ pour $j = 1, \ldots k = \min(g-1, p)$, la transformation correspondant à chacun des vecteurs propres $\mathbf{a}_j$. Alors $\boldsymbol{Y} = (Y_1, \ldots, Y_k)^\top$ est, par construction, tel que $\text{Var}(\boldsymbol{Y}) = \mathbf{I}_k$. Le vecteur $\boldsymbol{Y}$ a pour moyenne $\boldsymbol{\mu}_{l\boldsymbol{Y}} = (\mathbf{a}_1^\top \boldsymbol{\mu}_l, \mathbf{a}_2^\top \boldsymbol{\mu}_l, \ldots, \mathbf{a}_k^\top \boldsymbol{\mu}_l)^\top$ si l'observation provient du groupe $l$. Il est alors approprié de calculer l'éloignement de $\boldsymbol{Y} = \boldsymbol{y}$ par rapport au centre de chacune des populations $\boldsymbol{\mu}_{l\boldsymbol{Y}}$, soit $(\boldsymbol{y} - \boldsymbol{\mu}_{l\boldsymbol{Y}})^\top (\boldsymbol{y} - \boldsymbol{\mu}_{l\boldsymbol{Y}})$. On attribuera $\boldsymbol{Y} = \boldsymbol{y}$ à la population pour laquelle cette distance est minimisée.

On peut aussi se résoudre à n'utiliser que $r < k$ fonctions discriminantes, comme lors d'une ACP.

## 7.3 Autres considérations sur la discrimination

1. L'analyse discriminante peut s'effectuer comme une analyse canonique particulière. On écrit la matrice de données $n \times p$ comme $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_g)^\top$ où $\mathbf{X}_i$ est $m_i \times p$ avec la matrice $\mathbf{Y}$ de dimension $n \times (g-1)$ donnée par

$$\mathbf{Y} = \begin{pmatrix} \mathbf{1}_{m_1} & \mathbf{0}_{m_1} & \cdots & \mathbf{0}_{m_1} \\ \mathbf{0}_{m_2} & \mathbf{1}_{m_2} & \cdots \mathbf{0}_{m_2} & \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0}_{m_g} & \mathbf{0}_{m_g} & \cdots & \mathbf{1}_{m_g} \end{pmatrix}$$

tel que $y_{ij} = 1$ si $x_i$ appartient au groupe $j$, et $n\boldsymbol{S}_{11}^* = \widehat{\boldsymbol{T}}$ et $n\boldsymbol{S}_{12}^*\boldsymbol{S}_{22}^{*\,-1}\boldsymbol{S}_{21}^* = \widehat{\boldsymbol{B}}$. Aussi $\widehat{\boldsymbol{T}}^{-1}\widehat{\boldsymbol{B}}$ a valeur propre $\widehat{\gamma}_i$, $\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{B}}$ a valeur propre $\widehat{\lambda}_i$ et ces dernières sont reliées par la relation $\widehat{\lambda}_i = \widehat{\gamma}_i/(1-\widehat{\gamma}_i)$

2. L'analyse discriminante est très fortement liée au test MANOVA (moyennes différentes pour faire une discrimination).

3. Il existe une technique de régression pas à pas permettant de faire une analyse discriminante.

## 7.4 Évaluation de la qualité de la discrimination

La validation croisée ou le calcul empirique du taux de classification sont des moyens de déterminer la qualité de la règle.

1. Cas normal : On considère le cas de deux populations, avec $\Pi_1 : \boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ et $\Pi_2 : \boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. On a

$$h(\boldsymbol{x}) = \mathbf{a}^\top \left( \boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right)$$

avec

$$\mathsf{P}(R_1 \mid \Pi_2) = \mathsf{P}(R_2 \mid \Pi_1) = \Phi\left(-\frac{1}{2}\Delta\right)$$

où

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

avec les valeurs empiriques $\widehat{\mathsf{P}}(R_1 \mid \Pi_2) = \widehat{\mathsf{P}}(R_2 \mid \Pi_1) = \Phi(-0.5\widehat{\Delta})$ avec $\widehat{\Delta}^2 = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^\top \boldsymbol{S}_p^{-1}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)$. En utilisant deux fois les données, on obtiendra un résultat (trop) optimiste.

2. **Resubstitution :**

Soit $n_j$ la taille de $\Pi_j$. On fait la discrimination, menant à l'allocation des données initiales. Si $n_{ij}$ dénote le nombre d'individus de $\Pi_j$ classés dans $R_i(\Pi_i)$, avec comme

estimé empirique $\widehat{\mathsf{P}} = n_{ij}/n_j$. Même réserves pour cette méthode.

3. **Validation croisée :**
Effectuer la discrimination sur $\boldsymbol{X}_{-r}$, en ôtant l'observation numéro $r$, et calculer les régions $R_1^{(r)}, \dots, R_g^{(r)}$. Observer si l'observation $r \in R_1^{(r)}$. Soit $n_{i1}$ le nombre d'individus de $\Pi_1$ pour lesquels $x_r \in R_i^{(r)}, i \neq 1$, autrement dit sont mal classés. Alors $p_{i1}^* = n_{i1}^*/n_1$, pour $i \neq 1$. Faire la même chose pour les échantillons $2, \dots, g$. Cette méthode est meilleure au niveau de l'évaluation, mais est intensive en calculs.

## 7.5 Discrimination et régression logistique

Considérons la classification sur la base de variables explicatives $\mathbf{X}$ d'observations appartenant à 2 groupes. On suppose que $Y \in \{0, 1\}$ suit une loi Bernoulli, dont la densité en caractérisation exponentielle est

$$\begin{aligned} f_Y(y; p) &= \exp\left(y \log\left(\frac{p}{1-p}\right) + \log(1-p)\right) \\ &= \exp\left(y\theta + \log(1+e^\theta)\right). \end{aligned}$$

avec comme paramètre naturel $\theta = \operatorname{logit}(p) := \log(p) - \log(1-p)$. La moyenne de la distribution est $p = \operatorname{expit}(\theta) := e^\theta/(1+e^\theta) \in (0, 1)$.

Un modèle généralisé linéaire est un modèle de régression de la forme

$$g(\mathsf{E}(Y \mid \boldsymbol{X})) = \eta = \alpha + \boldsymbol{X}\boldsymbol{\beta}, \quad \eta \in \mathbb{R}$$

pour une fonction lien $g$ adéquate. On peut choisir par exemple la fonction lien naturelle $g(\cdot) = \operatorname{logit}(\cdot)$, qui correspond à $p = \operatorname{expit}(\eta)$. Le modèle binomial résultant satisfait les contraintes inhérente à l'hypothèse distributionnelle, dans le cas présent $\mathsf{E}(Y \mid \boldsymbol{X}) \in (0, 1)$, tout en relâchant l'hypothèse de symétrie des résidus inhérente au modèle linéaire.

La **discrimination logistique** est antérieure aux modèles généralisés linéaires et est analogue à la discrimination linéaire. Soit la probabilité d'appartenance conditionnelle à une population $k \in \{1, 2\}$, $\mathsf{P}(G_k \mid \boldsymbol{X})$. Par le théorème de Bayes, on peut écrire

$$\begin{aligned} \mathsf{P}(G_1 \mid \boldsymbol{X}) &= \frac{\mathsf{P}(\boldsymbol{X} \mid G_1)\,\mathsf{P}(G_1)}{\mathsf{P}(\boldsymbol{X} \mid G_1)\,\mathsf{P}(G_1) + \mathsf{P}(\boldsymbol{X} \mid G_2)\,\mathsf{P}(G_2)} \\ &= \left(1 + \frac{\mathsf{P}(\boldsymbol{X} \mid G_2)\,\mathsf{P}(G_2)}{\mathsf{P}(\boldsymbol{X} \mid G_1)\,\mathsf{P}(G_1)}\right)^{-1} \end{aligned}$$

où $\mathsf{P}(G_k)$ dénote la probabilité a priori d'appartenir à la population $k$. On travaille alors avec la vraisemblance des $\boldsymbol{X}$. En supposant une discrimination logistique linéaire, en assumant que les lois sont normales de même variance, soit $\boldsymbol{X} \mid G_1 \sim \mathcal{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ et

$\boldsymbol{X} \mid G_2 \sim \mathcal{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, on peut exprimer en forme close le maximum a posteriori, soit

$$\begin{aligned}
\mathsf{P}\left(G_1 \mid \boldsymbol{X}\right) &= \exp\left(\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}\right) \mathsf{P}\left(G_2 \mid \boldsymbol{X}\right) \\
\mathsf{P}\left(G_2 \mid \boldsymbol{X}\right) &= \left(1 + \exp\left(\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}\right)\right)^{-1}
\end{aligned}$$

avec

$$\begin{aligned}
\boldsymbol{\beta} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
\alpha &= -\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \log\left(\frac{\mathsf{P}\left(G_1\right)}{1 - \mathsf{P}\left(G_1\right)}\right)
\end{aligned}$$

Cette formulation nous force à imposer une hypothèse distributionnelle sur les $\boldsymbol{X}$. La régression logistique (avec $\eta = \text{logit}(p)$ est plus parcimonieuse puisque l'on modélise directement $\mathsf{P}\left(G_j \mid \boldsymbol{X}_i\right)$ et que seuls les coefficients de régressions sont spécifiés (la variance d'une famille exponentielle étant fonction du paramètre naturel).

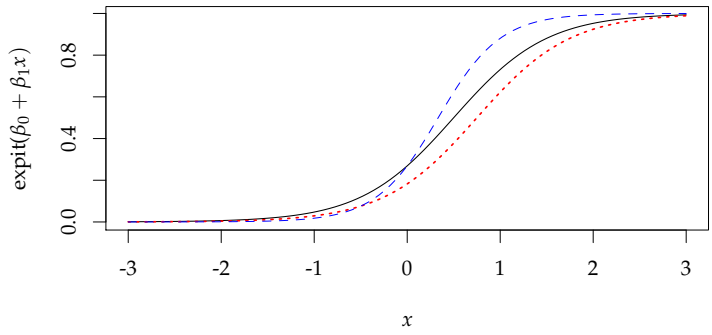



FIGURE 14 – Illustration d'un changement dans la probabilité lors d'un décalage du taux de base (changement de $\alpha$ (rouge)) et de l'effet $\beta$ (bleu) en fonction de la covariable $x$.

On peut généraliser ce résultat au cas de $K$ populations (données multinomiales) avec une série de modèles

$$\log\left(\frac{\mathsf{P}\left(G_k \mid \boldsymbol{X}\right)}{\mathsf{P}\left(G_K \mid \boldsymbol{X}\right)}\right) = \alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{X}, \qquad k = 1, \ldots, K-1$$

avec ici $K$ comme classe de référence. Cela revient à imposer $\alpha_K = 0, \boldsymbol{\beta}_K = \boldsymbol{0}$ pour satisfaire les contraintes d'identifiabilité $\sum_{k=1}^{K} \mathsf{P}\left(G_k \mid \boldsymbol{X}\right) = 1$. La probabilité que l'ob-

servation $i$ appartienne au groupement $j$ est alors donnée par

$$\mathsf{P}\left(G_j \mid \boldsymbol{X}_i\right) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^\top \boldsymbol{X}_i)}{\sum_{k=1}^{K} \exp(\alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{X}_i)} = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j^\top \boldsymbol{X}_i)}{1 + \sum_{k=1}^{K-1} \exp(\alpha_k + \boldsymbol{\beta}_k^\top \boldsymbol{X}_i)}.$$

On attribue les nouvelles données $\boldsymbol{X}_i$ au groupe $\max_{k \in \{1,\ldots,K\}} \mathsf{P}\left(G_k \mid \boldsymbol{X}_i\right)$.

## 7.6 Arbres de décisions

Un arbre de décision est une séquence de partitions binaires récursives qui mène au partitionnement des données. On parle d'arbre de régression si la variable dépendante est continue, et d'arbre de classification dans le cas de données catégoriques.

### Arbres de classification

Soit $Y \in \{1, \ldots, K\}$ une variable catégorique et $\boldsymbol{X}$ un $p$-vecteur de variables explicatives. On considère une séquence de $M$ règles de décisions binaires de la forme $X_{j_m} \leq c_m$ pour $m = 1, \ldots, M$. Ces règles définissent des régions (des hyperrectangles) dans l'espace des $\boldsymbol{X}$ dans lesquelles nous prédirons $Y$ par un indicateur d'appartenance à une catégorie $k$. Graphiquement, cette séquence peut être représentée par un arbre ; chaque règle de décision crée un noeud $\tau$ scindant une branche en deux.

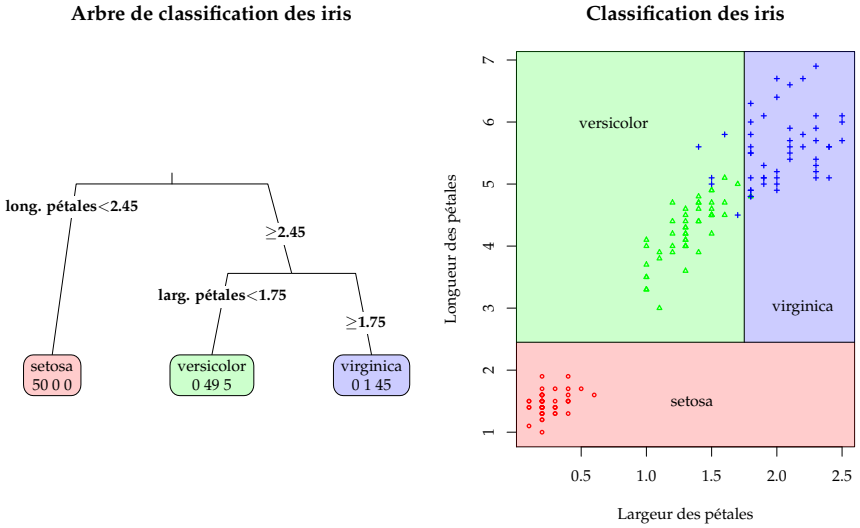



FIGURE 15 – Arbre de classification pour les données `iris` (gauche) et régions correspondantes (droite).

Comment choisir les conditions booléennes pour séparer un noeud $\tau_m$ ? En sélectionnant la variable $X_{j_m}$ et une valeur $c_m$ de séparation, on partitionne la région $R_m$ définie

à $\tau_m$ en deux hyperplans définis par

$$R_m^{(g)}(j_m, c_m) = \{\boldsymbol{X} \mid X_{j_m} \leq c_m, \boldsymbol{X} \in R_m\}, \quad R_m^{(d)}(j_m, c_m) = \{\boldsymbol{X} \mid X_{j_m} > c_m, \boldsymbol{X} \in R_m\},$$

respectivement les branches droites et gauches du noeuds $\tau_m$. La région $R_m$ dépend des ancêtres de $\tau_m$.

Le choix optimal de la variable explicative et du point de séparation prend la forme

$$\min_{j_m, c_m} \left[ \min_{k_g} \sum_{x_i \in R_m^{(g)}(j_m, c_m)} \left| y_i - \mathsf{I}\left(y_i = k_g\right) \right| + \min_{k_d} \sum_{x_i \in R_m^{(d)}(j_m, c_m)} \left| y_i - \mathsf{I}\left(y_i = k_d\right) \right| \right],$$

un problème d'optimisation avec ici une fonction de coût binaire.

Si la région $R_m$ compte $N_m$ observations, la proportion d'observations du groupe $k$ dans ce noeud est donnée par

$$\widehat{p}_{mk} = \frac{1}{N_m} \sum_{\boldsymbol{x}_i \in R_m} \mathsf{I}\left(y_i = k\right)$$

et représente la probabilité conditionnelle qu'une observation tombe dans le groupe $k$ dans la région $R_m$. Si le coût de mauvaise classification est binaire, le choix optimal est d'assigner les observations au groupe $k$ représentant la classe majoritaire des observations dans la feuille $m$, soit au groupe $k(m) = \max_k \widehat{p}_{mk} = \min_k \sum_{x_i \in R_m} |y_i - \mathbf{1}_{y_i = k}|$. On peut ensuite énumérer toutes les séparations possibles (au plus $nd$ choix) afin de trouver les paramètres $j_m, c_m$ optimaux et ainsi la règle de décision binaire optimale à une étape donnée.

La qualité de la séparation $c_m$ au noeud $\tau_m$ dépend de la proportions d'observations bien classées. Plusieurs mesures mesurant la qualité de la séparation existent : elles sont appelées fonction d'impureté de noeud et dénotées $Q_m$. Les trois principales sont

l'erreur de classification : $1 - \widehat{p}_{mk}$ ;
l'index de Gini : $\sum_{k=1}^{K} \widehat{p}_{mk}(1 - \widehat{p}_{mk})$ ;
l'entropie croisée : $-\sum_{k=1}^{K} \widehat{p}_{mk} \log(\widehat{p}_{mk})$ (ou la déviance $-2\sum_{k=1}^{K} \widehat{p}_{mk} \log(\widehat{p}_{mk})$).

L'erreur de classification est moins sensible aux changements dans les probabilités d'assignation aux noeuds et est moins utilisée pour guider la croissance de l'arbre.

L'algorithme procède comme suit : à partir d'une première division, on traverse l'arbre en séparant chaque noeud terminal intermédiaire jusqu'à saturer l'arbre, en divisant une branche seulement si l'effectif de cette dernière est supérieure à un certain seuil $n_{\min}$.

On continue cette procédure jusqu'à saturer l'arbre, surajustant ce faisant le modèle. La taille de l'arbre saturé $T_0$ correspond au nombre de feuilles (noeuds terminaux) de $T_0$,

dénotée $|T_0|$. Le choix optimal du sous-arbre de classification $T \subset T_0$ est déterminé à l'aide d'une critère prenant en compte la complexité du modèle et la qualité des prédictions.

En indexant les noeuds terminaux de $T$ par $l = 1, \ldots, |T|$ et les régions associées par $R_l$, on définit le critère de coût complexité

$$C_\alpha(T) := \sum_{l=1}^{|T|} N_l Q_l(T) + \alpha|T|,$$

où $Q_l(T)$ est une fonction d'impureté de noeud et $\alpha$ un paramètre de réglage qui dicte le compromis entre taille de l'arbre et adéquation du modèle. Ce dernier est typiquement choisi par validation croisée.

Pour chaque valeur de $\alpha$, il existe un sous-arbre minimal unique, $T_\alpha$, tel que $T_\alpha = \arg\min_{T \subset T_0} C_\alpha(T)$. Ce dernier peut être trouvé par une séquence d'itérations en fusionnant les noeuds intermédiaires et en choisissant celui qui produit la plus faible hausse de $\sum_l N_l Q_l(T)$, jusqu'à obtenir un arbre avec un seul noeud. Il est possible de démontrer que cette séquence de sous-arbres contient $T_\alpha$. Les observations dans une feuille de $T_\alpha$ seront attribuées au groupe présentant la plus forte proportion d'observations.

## Forces et faiblesses des arbres de décision

1. Les solutions sont simples et faciles à implémenter. Il y a peu d'hypothèses probabilistes. Par exemple la discrimination linéaire fait l'hypothèse d'égalité des matrices de variance-covariance.
2. L'ajustement se fait par le biais d'un algorithme glouton (les décisions préalables ne sont pas remises en cause et aucune recherche exhaustive de l'espace n'est effectuée).
3. La classification se fait variable par variable, ce qui est un avantage s'il y a des données manquantes. Si les variables sont fortement corrélées, les arbres de classifications sont peu performants.
4. Les arbres de classification sont très instables et peu robustes et l'arbre n'a pas de caractère inférentiel. On peut se protéger en le construisant avec un échantillon d'entraînement et en utilisant le reste des données pour estimer l'erreur.

Les arbres de décisions sont simplistes, mais sont à la base de plusieurs techniques modernes d'apprentissage statistique, comme les forêts aléatoires, les réseaux de neurones et les machines à vecteurs de support.

# Chapitre 8
# Copules

Dans ce dernier chapitre, nous aborderons la notion de copule, qui permet de regrouper sous un même étendard les distributions multivariées.

## 8.1 Motivation et définition

Il existe peu de distributions multivariées canoniques, et la plupart de ces modèles sont des généralisation de lois unidimensionnelles, dont le domaine et les marges sont prescrits. Les exemples les plus connus sont les lois multinomiale (marge binomiale, support $\mathbb{Z}_+^d$), Dirichlet (marges Beta, avec pour support le simplexe unitaire $\mathbb{S}_{d-1}$), Wishart (marges $\chi^2$, matrices positives semidéfinies) et elliptiques.

Deux questions motivent notre approche, soit

1. comment obtenir des fonctions de répartition multivariées valides avec des marges arbitraires ?
2. comment comparer des données dont les marges diffèrent ?

**Lemme 8.1**
Soit $F$ une fonction de répartition univariée et sa fonction de quantile,

$$F^{\leftarrow}(u) := \inf\{x : F(x) \geq u\}, \quad u \in [0,1].$$

Alors

1. $F^{\leftarrow}$ est non-décroissante et continue à gauche. Si $F$ est continue, alors $F^{\leftarrow}$ est strictement croissante (mais pas nécessairement continue).
2. $F^{\leftarrow} \circ F(x) \leq x$, pour tout $x \in \mathbb{R}$. Si $F$ est strictement croissante, $F^{\leftarrow} \circ F(x) = x$.
3. $F \circ F^{\leftarrow}(u) \geq u$ pour tout $u \in [0,1]$. Si $F$ est continu, alors $F \circ F^{\leftarrow}(u) = u$

**Preuve**

2. $F^{\leftarrow} \circ F(x) = \inf\{y \in \mathbb{R} : F(y) \geq F(x)\} =: A$. Alors, $x \in A$ et $\inf(A) \leq x$, donc $\inf(A) = F^{\leftarrow} \circ F(x)$. Supposer $F$ est strictement croissant, soit $F^{\leftarrow} \circ F(x) < x$ pour $x \in \mathbb{R}$. Par contradiction ; en supposant qu'il existe $a \in A$ tel que $a < x$. Mais $F(a) < F(x)$ puisque $F$ est strictement croissant, ce qui contredit le fait que $a \in A$ et donc $F(a) \geq F(x)$.
3. $F$ est par définition continue à droite. Nous avons égalité à 0, alors assumons sans perte de généralité que $y \in (0,1]$ et $F^{\leftarrow}(1) < \infty$. Alors $F^{\leftarrow}(u) \in \mathbb{R}$ pour $u \in (0,1)$ et

$$F(F^{\leftarrow}(u)) = \lim_{\substack{x_n \to F^{\leftarrow}(u) \\ x_n > F^{\leftarrow}(u)}} F(x_n).$$

Maintenant $x_n > F^{\leftarrow}(u)$, donc il existe $x_n^* \in A := \{x : F(x) \geq u\}$ tel que $x_n^* < x_n$.

$F$ est non-décroissante, donc $u \leq F(x_n^*) \leq F(x_n)$ et de ce fait $x_n \in A$, c'est-à-dire $F(x_n) \geq u$. Ainsi $F \circ F^{\leftarrow}(u) = \lim_{n\to\infty} F(x_n) \geq u$. Si $F$ est continue,

$$F(F^{\leftarrow}(u)) = \lim_{\substack{n\to\infty \\ y_n \to F^{\leftarrow}(u) \\ y_n < F^{\leftarrow}(u)}} F(y_n) \quad \Leftrightarrow \qquad y_n < \inf(A)$$

et donc $y_n \notin A$ implique que $F(y_n) < u$.

■

**Proposition 8.2 (transformée intégrale de probabilité)**
Si $Y$ est une variable aléatoire absolument continue de fonction de répartition $G$, alors

$$G(Y) \sim \mathcal{U}(0,1)$$

**Proposition 8.3 (transformation des quantiles)**
Si $U \sim \mathcal{U}(0,1)$ a une loi uniforme standard, alors

$$\mathsf{P}\left(G^{\leftarrow}(U) \leq x\right) = G(x)$$

où l'inverse généralisée est définie comme $F^{\leftarrow}(p) := \min\{x : F(x) \geq p\}$

**Preuve de la proposition 8.2** Observer que pour tout $u \in (0,1), x \in \mathbb{R}$,

$$F^{\leftarrow}(u) \leq x \quad \Leftrightarrow \quad u \leq F(x)$$

En effet, $F$ est non-décroissante et donc si $F^{\leftarrow}(u) \leq x$, donc $F(F^{\leftarrow}(u)) \leq F(x)$ et par notre lemme 8.1, $u \leq F(F^{\leftarrow}(u))$, ce qui prouve la nécessité de la condition.

Pour la suffisance, si $u \leq F(x)$, alors $x \in \{y : F(y) \geq u\}$ et $\inf\{y : F(y) \geq u\} \leq x \Rightarrow F^{\leftarrow}(u) \leq x$. Soit $U \sim \mathcal{U}(0,1)$ et fixons $X = F^{\leftarrow}(U)$;

$$F(x) = \mathsf{P}\left(X \leq x\right) = \mathsf{P}\left(F^{\leftarrow}(U) \leq x\right) = \mathsf{P}\left(U \leq F(x)\right) = F(x)$$

puisque $F(x)$ a pour domaine $[0,1]$. ■

**Preuve de la proposition 8.3** Nous avons

$$\mathsf{P}\left(F^{\leftarrow}(U) \leq x\right) = \mathsf{P}\left(U \leq F(x)\right) = F(x), \quad x \in \mathbb{R}.$$

en utilisant le point 3 du lemme 8.1 ■

La transformation des quantiles sert à générer des variables pseudo-aléatoires.

**Exemple 8.1 (Génération de variables exponentielles)**
Si $U \sim \mathcal{U}(0,1)$, alors $-\log(U) \sim \mathcal{E}(1)$.

Si les distributions marginales sont inconnues, elles devront être estimées. On peut utiliser un estimateur nonparamétrique de $F$ : la fonction de répartition empirique, $\widetilde{F}_n$. Ce choix est motivé par le résultat suivant :

**Théorème 8.4 (Glivenko–Cantelli)**
Soit $X \sim F$ et $\widetilde{F}_n(x) := n^{-1} \sum_{i=1}^{n} \mathsf{I}\,(X_i \leq x)$

$$\|\widetilde{F}_n(x) - F(x)\|_\infty \xrightarrow{\text{a.s.}} 0$$

Il n'est pas difficile de dériver la normalité asymptotique de $\widetilde{F}_n$ puisque les indicateurs sont des variables Bernoulli et que donc $\sqrt{n}(\widetilde{F}_n(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, F(x)(1-F(x)))$ par application du théorème central limite.

Pour transformer non-paramétriquement les données à l'échelle (pseudo)-uniforme, on calcule les rangs des observations, dénotés $R_i$ que l'on renormalise par un facteur $(n+1)^{-1}$. Les données ainsi obtenues sont dénommées pseudo-observations. Le choix de $(n+1)^{-1}$ plutôt que $n^{-1}$ évite les problèmes à la bordure et n'a pas d'effet asymptotiquement. [9] On aura ainsi

$$\widetilde{U}_i = \frac{\mathsf{rang}(X_i)}{n+1} = \frac{R_i}{n+1}.$$

La transformée intégrale de probabilité sert dans les diagrammes P–P (voir définition 2.19) pour vérifier l'adéquation d'une hypothèse paramétrique. La fonction de quantile est utilisée dans les diagrammes Q–Q.

## Copule

Une copule lie une fonction de répartition et des lois marginales. Formellement, c'est une fonction de répartition multivariée dont les marges sont uniformes.

**Définition 8.5 (Copule)**
Une copule est une application $C : [0,1]^d \to [0,1]$ satisfaisant

1. $C$ a des marges uniformes : $C(\boldsymbol{u}) = u_i$ si $u_j = 1 \; \forall \; j \neq i$ et $u_i \in [0,1]$.
2. $C(\boldsymbol{u}) = 0$ si $u_i = 0$ pour au moins un $i$ ;
3. $C$ est $\Delta$-monotone, c'est-à-dire est non-décroissante en toute composante,

$$\mathsf{P}\left(\boldsymbol{U} \in (\boldsymbol{a}, \boldsymbol{b}]\right) = \Delta_{(\boldsymbol{a},\boldsymbol{b}]} C := \sum_{i_1=0}^{1} \cdots \sum_{i_d=0}^{1} (-1)^{\sum_{j=1}^{d} i_j} C(v_{1i_1}, \ldots, v_{di_d}) \geq 0$$

pour des vecteurs $\boldsymbol{a}, \boldsymbol{b} \in [0,1]^d$ avec $\boldsymbol{a} \leq \boldsymbol{b}$ et $v_{ji_j} = \mathsf{I}\,(i_j = 0) a_j + \mathsf{I}\,(i_j = 1) b_j$.

9. En effet, si le support d'une variable aléatoire est $\mathbb{R}$, alors $\widetilde{F}^{\leftarrow}(1) = \infty$.

La condition 3. équivaut à nécessiter que le volume entre deux points de $\mathbb{R}^d$ soit non-négatif (voir définition 1.2). Si la densité de $C(\boldsymbol{u})$ existe, cette condition revient à requérir que $c(\boldsymbol{u}) \geq 0$ pour tout $u \in (0,1)^d$.

**Exemple 8.2 (Copules fondamentales)**
La copule antimonotone $\mathrm{W}(u_1, u_2) = \max\{u_1 + u_2 - 1, 0\}$ est une fonction de répartition valide décrivant la dépendance négative parfaite (c'est-à-dire la fonction de répartition de $U, 1-U$). Inversement, la copule comonotone $\mathrm{M} \coloneqq \bigwedge_{i=1}^d u_i$ décrit la dépendance parfaite, soit la fonction de répartition de $(U, \ldots, U)$. Les deux copules sont singulières. Un autre cas intéressant est celui de l'indépendance : les composantes d'un vecteur aléatoire $\boldsymbol{X}$ sont mutuellement indépendantes ($X_i \perp\!\!\!\perp X_j$ pour tout $i \neq j$) si et seulement si

$$\Pi(\boldsymbol{u}) \coloneqq C(u_1, \ldots, u_d) = \prod_{i=1}^d u_i, \qquad \boldsymbol{u} \in [0,1]^d$$

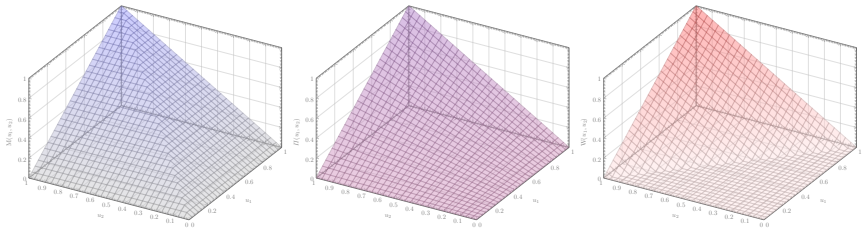
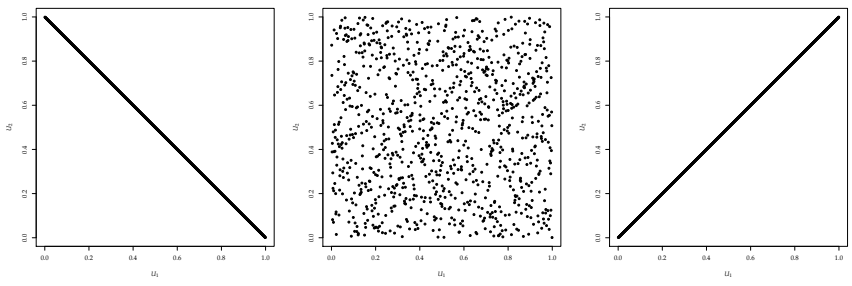


FIGURE 16 – Copules antimonotone (gauche), d'indépendance (centre) et comonotone (droite)



**Exemple 8.3 (Copule de Marshall–Olkin)**
La famille de copules bivariées de Marshall–Olkin est

$$C(u_1, u_2; \alpha_1, \alpha_2) = \min(u_1^{1-\alpha_1} u_2, u_1 u_2^{1-\alpha_2}) = \begin{cases} u_1^{1-\alpha_1} u_2, & \text{si } u_1^{\alpha_1} \geq u_2^{\alpha_2}, \\ u_1 u_2^{1-\alpha_2}, & \text{si } u_1^{\alpha_1} \leq u_2^{\alpha_2}. \end{cases}$$

Ces copules sont à la fois absolument continues et dotées d'une singularité. Puisque

$$\frac{\partial^2}{\partial u_1 \partial u_2} C(u_1, u_2; \alpha_1, \alpha_2) = \begin{cases} u_1^{-\alpha_1}, & \text{si } u_1^{\alpha_1} > u_2^{\alpha_2} \\ u_2^{-\alpha_2}, & \text{si } u_1^{\alpha_1} < u_2^{\alpha_2}; \end{cases}$$

la masse de la composante singulière est concentrée sur la courbe $u_1^{\alpha_1} = u_2^{\alpha_2} \in [0,1]^2$.

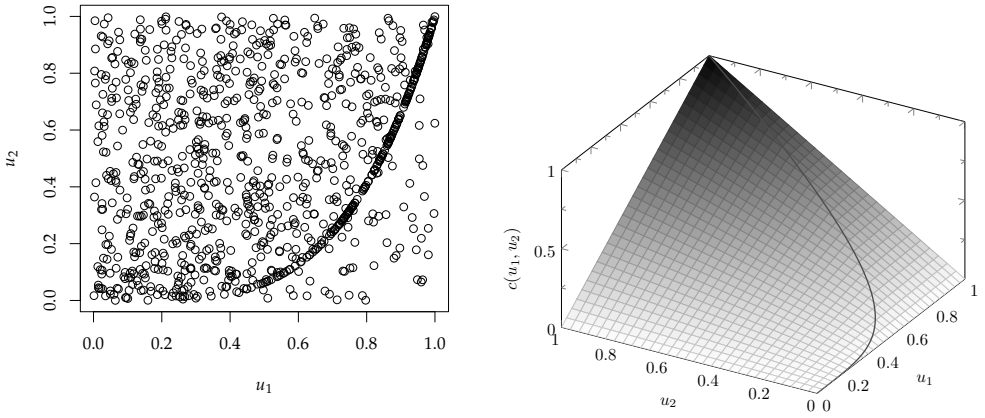



FIGURE 17 – Échantillon et graphique d'une copule de Marshall–Olkin avec $\alpha_1 = 0.8, \alpha_2 = 0.2, n = 1000$

## 8.2 Théorème de Sklar

Pourquoi restreindre son attention aux distributions multivariées dont les marges sont uniformes ? Le résultat suivant démontre qu'il existe une relation entre les fonctions de répartitions $F$ de marges $F_1, \ldots, F_d$ et les copules.

**Théorème 8.6 (Sklar (1959))**
Soit $F$ une distribution conjointe avec des distributions marginales $F_1, \ldots, F_d$. Alors, il existe une copule $C : [0,1]^d \to [0,1]$ telle que, pour tout $x_1, \ldots, x_d$ dans $\overline{\mathbb{R}} = [-\infty, \infty]$,

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \tag{8.10}$$

Si les distributions marginales sont continues, alors $C$ est unique. Autrement, $C$ est uniquement définie sur $\mathsf{Im}(F_1) \times \cdots \times \mathsf{Im}(F_d)$.

Inversement, si $C$ est une copule et $F_1, \ldots, F_d$ sont des fonctions de répartition univariées, alors la fonction définie par eq. (8.10) est une distribution conjointe de distribution marginales $F_1, \ldots, F_d$.

**Preuve** Nous supposerons que les fonctions $F_1, \ldots, F_d$ sont continues. Le cas général est traité dans Nelsen (2006).

Soit $\boldsymbol{X} \sim F$ et $U_j \coloneqq F_j(X_j)$ pour $j \in \{1, \ldots, d\}$. Par la transformée intégrale de probabilité, $U_j \sim \mathcal{U}(0,1)$ et de ce fait la fonction de répartition $C$ est une copule. Puisque $F_j$ est croissante sur $\mathsf{Im}(X_j)$, $X_j = F_j^{\leftarrow}(F_j(X_j)) = F^{\leftarrow}(U_j)$ pour $j = 1, \ldots, d$. Ainsi, pour $\boldsymbol{x} \in \mathbb{R}^d$.

$$\begin{aligned} F(\boldsymbol{x}) = \mathsf{P}\left(X_j \leq x_j \; \forall \, j\right) &= \mathsf{P}\left(F_j^{\leftarrow}(U_j) \leq x_j \; \forall \, j\right) \\ &= \mathsf{P}\left(U_j \leq F_j(x_j) \; \forall \, j\right) \\ &= C(F_1(x_1), \ldots, F_d(x_d)) \end{aligned}$$

$C$ est donc une copule qui satisfait l'équation (8.10).

Soit $\boldsymbol{U} \sim C$ et le vecteur $\boldsymbol{X} \coloneqq (F_1^{\leftarrow}(U_1), \ldots, F_d^{\leftarrow}(U_d))$. Alors, pour $\boldsymbol{x} \in \mathbb{R}^d$,

$$\begin{aligned} \mathsf{P}\left(\boldsymbol{X} \leq \boldsymbol{x}\right) &= \mathsf{P}\left(F_j^{\leftarrow}(U_j) \leq x_j \; \forall \, j\right) \\ &= \mathsf{P}\left(U_j \leq F_j(x_j) \; \forall \, j\right) \\ &= C(F_1(x_1), \ldots, F_d(x_d)). \end{aligned}$$

$F$ telle que définie par équation (8.10) est donc la fonction de répartition de $\boldsymbol{X}$ et ses marges sont $F_1, \ldots, F_d$ par la transformation des quantiles. ■

La continuité absolue est cruciale pour l'unicité de la copule.

**Exemple 8.4 (Distribution bivariée avec marges Bernoulli)**
Soit $(X_1, X_2)$ suivant une loi bivariée de Bernoulli avec $\mathsf{P}\left(X_1 = k, X_2 = l\right) = 1/4$ pour $k, l \in \{0, 1\}$. On déduit que $\mathsf{P}\left(X_j = k\right) = 1/2$ pour $k \in \{0, 1\}$ et l'image $\mathsf{Im}(F_j) = \{0, 1/2, 1\}$ pour $j = 1, 2$. Toute copule telle que $C(1/2, 1/2) = 1/4$ satisfait ce critère (par exemple la copule d'indépendance $C(u_1, u_2) = u_1 u_2$ ou bien encore la copule diagonale $C(u_1, u_2) = \min\{u_1, u_2, (u_1^2 + u_2^2)/2\}$.

Une propriété fondamentale des copules est leur invariance à des transformations strictement croissantes des distributions marginales.

**Proposition 8.7**
Soit $(X_1, \ldots, X_d)$ un vecteur aléatoire dont les marges sont continues et une collection $\{\phi_i\}_{i=1}^d$ de fonctions **strictement croissantes** sur $\mathsf{Im}(X_i)$, respectivement. Alors $(\phi_1(X_1), \ldots, \phi_d(X_d))$ et $(X_1, \ldots, X_d)$ ont la même copule $C$.

**Preuve** Soit $\boldsymbol{X} \sim F$ de distributions marginales $F_1, \ldots, F_d$. Sans perte de généralité, il est possible d'assumer que $\phi_j$ est continue à droite aux points dénombrables de continuité

(puisque $X_j$ a par hypothèse une fonction de répartition continue, $\phi_j(X_j)$ change sur un ensemble de mesure nulle). Puisque $\phi_j$ est croissant sur $\mathsf{Im}(X_j)$ et que $X_j$ a une distribution continue, $\phi_j(X_j)$ a également une distribution continue et

$$\begin{aligned} F_{\phi_j(X_j)}(x) &= \mathsf{P}\left(\phi_j(X_j) \leq x\right) \\ &= \mathsf{P}\left(T_j(X_j) < x\right) = \mathsf{P}\left(X_j < T_j^{\leftarrow}(x)\right) \\ &= \mathsf{P}\left(X_j \leq \phi_j^{\leftarrow}(x)\right) \\ &= F_j(\phi_j^{\leftarrow}(x)) \end{aligned}$$

pour $x \in \mathbb{R}$. Cette dernière égalité implique que

$$\mathsf{P}\left(F_j(\phi_j^{\leftarrow}(\phi_j(X_j))) \leq u_j \; \forall \; j\right) = \mathsf{P}\left(F_j(X_j) \leq u_j \; \forall \; j\right) = C(\boldsymbol{u})$$

où la dernière égalité découle de l'unicité de la copule associée à une fonction de répartition $F$. ■

Les implications du théorème de Sklar sont nombreuses. D'une part, il permet d'étudier la dépendance en séparant les effets marginaux de la structure conjointe. En effet,

- les copules sont invariantes aux transformations marginales strictement monotones croissantes sur l'image, puisqu'elles sont basées sur les rangs. Cela veut dire que l'on peut bâtir un modèle avec des marges (continues) arbitraires.
- il pose le principe que toutes les distributions multivariées de dimension $d$ dont les marges sont absolument continues correspondent à une copule, la recherche de modèles adéquats peut se réduire à celui d'une copule correspondante.
- les formules dérivées pour les copules sont valides pour tous les modèles peu importe leurs marges, et le support borné facilite les dérivations analytiques.

D'autre part, le résultat bien que général n'est pas constructif. Mikosch (2006) articule plusieurs critiques, dont nous retenons un échantillon sélectif :

- la classe des copules est trop large pour être utile en pratique.
- le choix d'une copule en est souvent un de convenance mathématique.
- les copules ne s'insèrent pas dans le cadre existant des processus stochastiques : ce sont des modèles strictement statiques.

**Proposition 8.8 (Propriétés des copules)**
Si $\boldsymbol{U} \sim C$ alors $\mathbf{1} - \boldsymbol{U} \sim \overline{C}$, la copule de survie de $C$. Pour $d = 2$,

$$\begin{aligned} \overline{C}(u_1, u_2) &= 1 - (1 - u_1) - (1 - u_d) + C(1 - u_1, 1 - u_2) \\ &= C(1 - u_1, 1 - u_2) + u_1 + u_2 - 1. \end{aligned}$$

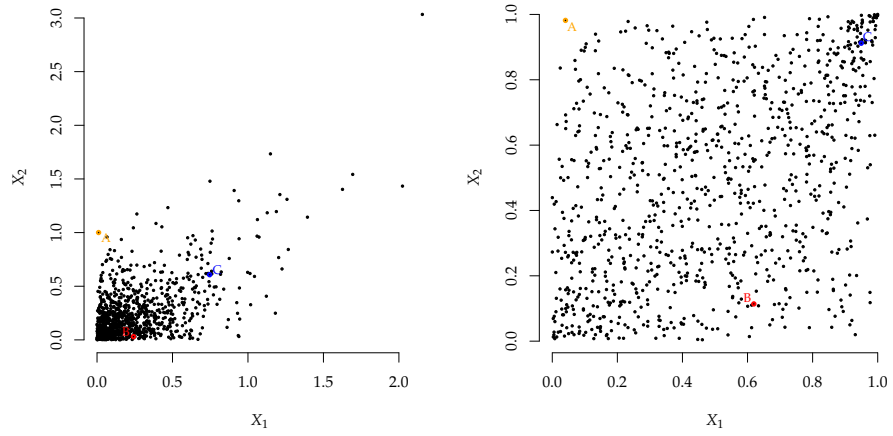



FIGURE 18 – Illustration de l'invariance des rangs à des changements de marges

Si $\overline{C} = C$, $C$ est une fonction symétrique radiale.

Le vecteur aléatoire $\boldsymbol{X}$ est dit échangeable si

$$(X_1, \ldots, X_d) \overset{\mathrm{d}}{=} (X_{\pi(1)}, \ldots X_{\pi(d)})$$

pour toute permutation $(\pi(1), \ldots, \pi(d))$ de $(1, \ldots, d)$.

Par extension, une copule $C$ est échangeable si c'est la fonction de répartition d'un vecteur $\boldsymbol{U}$ échangeable dont les marges sont uniformes, c'est à dire si $C$ est symétrique.

**Exemple 8.5 (Distribution de Marshall–Olkin)**
Soit $X_1, X_2$ deux variables aléatoires représentant la durée de vie de deux composantes. Si des chocs surviennent suivant trois processus ponctuels de Poisson indépendants de paramètres $\lambda_1, \lambda_2, \lambda_{12} \geq 0$ (selon qu'ils affectent la composante 1, 2 ou les deux simultanément). Le temps de ces chocs sont des variables exponentielles indépendantes de paramètre $\lambda_1, \lambda_2, \lambda_{12}$ respectivement et la fonction de survie de la paire $(X_1, X_2)$ est donnée par

$$\bar{F}(x_1, x_2) = \mathsf{P}\left(X_1 > x_1, X_2 > x_2\right) = \mathsf{P}\left(Z_1 > x_1\right) \mathsf{P}\left(Z_2 > x_2\right) \mathsf{P}\left(Z_{12} > \max\{x_1, x_2\}\right).$$

Les fonctions de survie univariées de $X_1$ et $X_2$ sont $\bar{F}_1(x_1) = \exp(-(\lambda_1 + \lambda_{12})x_1)$ et $\bar{F}_2(x_2) = \exp(-(\lambda_2 + \lambda_{12})x_2)$.

En prenant $\alpha_1 = \lambda_{12}/(\lambda_1 + \lambda_{12})$ et $\alpha_2 = \lambda_{12}/(\lambda_2 + \lambda_{12})$, on trouve que la copule de survie $(Z_1, Z_2)$ est $\overline{C}(u_1, u_2) = \min(u_1^{1-\alpha_1} u_2, u_1 u_2^{1-\alpha_2})$. Ce modèle correspond à la copule de Marshall–Olkin de l'exemple 8.3.

L'intervalle borné $[0,1]^d$ est utile pour la visualisation et la distribution uniforme facilite l'obtention des formules, puisque la forme de la densité est extrêmement simple. En

utilisant la transformation intégrale de probabilité et la fonction de quantiles, il est en revanche possible d'obtenir des distributions avec des marges autres que uniformes. Une telle construction est appelée méta copule..

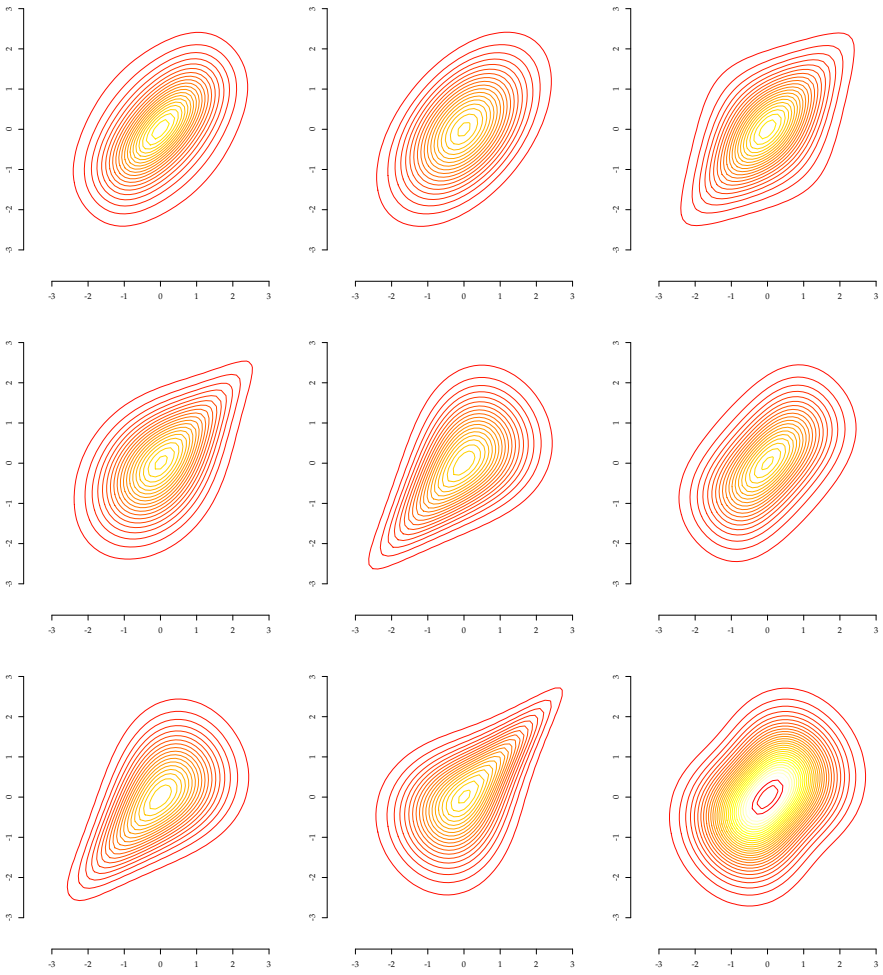


FIGURE 19 – Méta copules (via `Mvdc`) avec lois marginales $\mathcal{N}(0,1)$. De gauche à droite : haut en bas : Plackett, Gaussienne, Student–$t$ ; Gumbel–Hougaard, Clayton, Frank ; Ali-Mikhail-Haq, Joe, Farlie-Gumbel-Morgenstern. Modèles (sauf FGM) calibrés pour que $\tau \approx 1/3$.

En résumé : en variant $F_i$ dans l'équation (8.10) on peut construire des distributions dont les marges sont arbitraires. En variant $C$, on peut altérer la structure de dépendance entre $X_1, \ldots, X_d$.

## 8.3 Mesure de la dépendance

Le théorème de Sklar n'est pas un résultat constructif, et en pratique le choix d'une copule est souvent question de jugement. Avant de présenter quelques modèles classiques, nous révisons les mesures de dépendance qui puissent être utiles pour la calibration d'un modèle aux estimés empiriques issus d'un échantillon.

### Bornes de Fréchet–Hoeffding

Les distributions multivariées ont des bornes naturels en terme de dépendance. Dans le cas bivarié, ces deux bornes correspondent à des copules et sont strictes. Dans le cas $d \geq 3$, la borne inférieure ne décrit plus un modèle valide. Les bornes de Fréchet–Hoeffding, de la série 12, peuvent s'écrire pour toute copule $C(u_1, \ldots, u_d)$

$$\max\left\{\sum_{i=1}^{d} u_i + 1 - d, 0\right\} \leq C(\boldsymbol{u}) \leq \min\{u_1, \ldots, u_d\} =: \mathrm{M}(\boldsymbol{u})$$

et sont valides ponctuellement.

### Mesures d'association

Comment mesurer le degré de dépendance entre deux variables $X, Y$, au vu des bornes de Fréchet ? Clairement, la dépendance négative versus positive entraîne un choix de modèle potentiellement différent. Avant d'aborder quelques mesures d'association, révisons les propriétés désirables que ce dernier devrait satisfaire.

**Propriétés désirables d'un coefficient d'association** $r$ selon Embrechts, McNeil, Straumann (1999) :

P1. symétrique : $r(X, Y) = r(Y, X)$
P2. normalisée : $-1 \leq r \leq 1$
P3. $r = 1 \Leftrightarrow$ comonotonicité, $r = -1 \Leftrightarrow$ antimonotonicité
P4. Pour $T : \mathbb{R} \to \mathbb{R}$ strictement monotone sur $\mathsf{Im}(X)$,

$$r(T(X), Y) = \begin{cases} r(X, Y) & \text{si } T \text{ est croissante} \\ -r(X, Y) & \text{si } T \text{ est décroissante} \end{cases}$$

P5. $r = 0$ si et seulement si $X \perp\!\!\!\perp Y$

Notez qu'il n'existe pas de mesure satisfaisant P4 et P5 simultanément. Cela est dû à l'existence de distributions sphériques. On peut en revanche exiger que $X \perp\!\!\!\perp Y$ implique que $r = 0$. Ces conditions, ré-énoncées dans un ouvrage non-technique, suivent les grandes lignes de Schweizer & Wolff (1981), qui proposent de baser une mesure d'association sur la copule. Cela permet de respecter P4 (contrairement au $\rho$ de Pearson) et l'estimé résultant est toujours bien défini puisque l'ensemble $[0, 1]$ est borné (par

contraste, $\rho$ nécessite des variances finies pour les lois marginales). Une ultime condition souhaitable, hormis celle qui ont déjà été énoncées, provient de Scarsini (1984) et veut que

P6. si $(X_n, Y_n) \xrightarrow{d} (X, Y)$ alors $r(X_n, Y_n) \to r(X, Y)$

## Corrélation linéaire ($\rho$ de Pearson)

Mesure de dépendance linéaire la plus connue, elle est adaptée aux lois elliptiques. Facile à calculer, elle est directement interprétable pour la loi multinormale, pour laquelle $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$. Pour une distribution arbitraire, les propriétés suivantes sont valides :

- les corrélations linéaires atteignables forment un ensemble fermé $[\rho_{\min}, \rho_{\max}]$ incluant 0.
- la corrélation minimale (maximale) est atteinte dans le cas de variables antimonotones (comonotones).
- $\mathrm{Cor}\,(X, Y) = \pm 1$ si et seulement si $Y \overset{\mathrm{d}}{=} \pm aX + b$ pour $a > 0, b \in \mathbb{R}$.
- $X \perp\!\!\!\perp Y$ implique que $\mathrm{Cor}\,(X, Y) = 0$. L'inverse est vrai si la loi est multinormale (faux en général)

La corrélation linéaire n'est pas invariante aux transformations marginales strictement monotones. Un dernier point notable est le fait que $\rho$ n'est pas toujours bien défini : l'existence de $\rho(X, Y)$ est acquise si $\mathrm{Var}\,(X) < \infty$ et $\mathrm{Var}\,(Y) < \infty$.

Les quelques exemples suivants servent à démolir quelques mythes entourant la corrélation linéaire.

**Exemple 8.6 (Lois de Pareto)**
Concernant les derniers points, si $X, Y$ proviennent d'une loi Pareto de paramètre 3 et sont indépendantes, alors $\rho(X, Y) = 0$, mais $\rho(X^2, Y)$ n'existe pas.

**Exemple 8.7 (Lois lognormales)**
Reprenons l'exemple des bornes de Fréchet de la série 12, cette fois en fonction de lois marginales $X_i \sim \mathcal{LN}(0, \sigma_i^2), i \in \{1, 2\}$. Nous avons déjà démontré que les bornes $\pm 1$ ne sont pas toujours atteignables, comme les graphiques suivants l'illustrent. Si $\sigma_1^2 = 1, \sigma_2^2 = 25$ alors $\rho \in [-0.000002824, 0.000419091]$ !

**Exemple 8.8 (Corrélation nulle)**
Cet exemple montre que la corrélation de 0 n'implique pas l'indépendance. Soit la copule

$$C(u_1, u_2) = u_1 u_2 \left(1 - 2\theta \left(u_1 - \frac{1}{2}(u_1 - 1)(u_2 - 1)\right)\right)$$

On peut montrer aisément que les marges sont uniformes et que $\rho = 0$ pour tout $\theta \in [-1, 1]$, bien que la copule ne soit pas l'indépendance. Plus généralement : toute

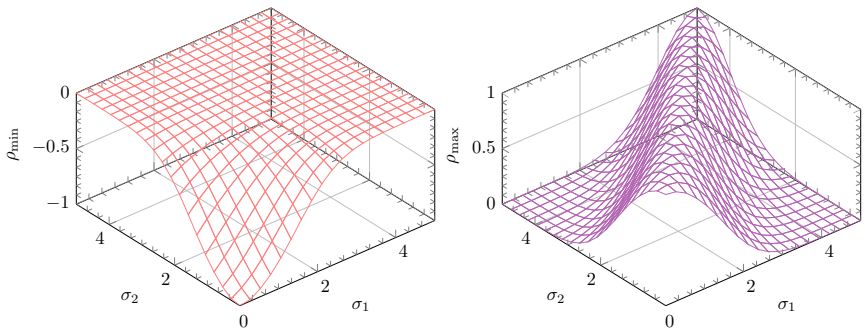



FIGURE 20 – Bornes de Fréchet–Hoeffding pour le $\rho$ de Pearson de l'exemple 8.7

distribution symétrique autour de 0 avec 2^e moments finis fournit un exemple de ce point. Si $X \sim \mathcal{N}(0,1)$, $Y = X^2$, alors $\mathsf{Cor}\,(X,Y) = \mathsf{E}\left(X^3\right) - \mathsf{E}\,(X)\,\mathsf{E}\left(X^2\right) = 0$.

## $\varrho$ de Spearman

Cette mesure d'association est équivalente à la corrélation linéaire de Pearson, mais à l'échelle uniforme. Pour deux variables aléatoires $X_1, X_2$, il est défini comme

$$\varrho(X_1, X_2) = \rho(F_1(X_1), F_2(X_2)).$$

En fonction de la copule, le $\varrho$ de Spearman est égal à

$$\begin{aligned}\varrho(X_1, X_2) &= \frac{\mathsf{E}\,(U_1U_2) - \mathsf{E}\,(U_1)\,\mathsf{E}\,(U_2)}{\sqrt{\mathsf{Var}\,(U_1)\,\mathsf{Var}\,(U_2)}} \\ &= 12\int_0^1\int_0^1 u_1u_2\,\mathrm{d}C(u_1,u_2) - 3 \\ &= 12\int_0^1\int_0^1 C(u_1,u_2)\,\mathrm{d}u_1\,\mathrm{d}u_2 - 3\end{aligned}$$

qui découle de l'identité de Hoeffding, c'est-à-dire

$$\mathsf{Cov}\,(X_1,X_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(F(x_1,x_2) - F_1(x_1)F_2(x_2))\,\mathrm{d}x_1\,\mathrm{d}x_2$$

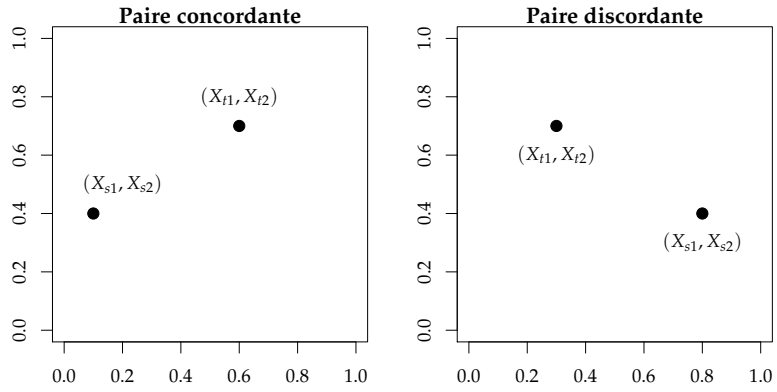
et du fait que $\mathsf{E}\,(U) = 1/2, \mathsf{Var}\,(U) = 1/12$ pour $U \sim \mathcal{U}(0,1)$.

Un estimateur empirique du coefficient de rang de Spearman est donné comme suit : pour $\{\mathrm{rang}(X_{t,i}), \mathrm{rang}(X_{t,j})\}$, les colonnes $i, j$ de $\mathbf{X}$ et $t = 1, \ldots, n$,

$$\varrho(\mathbf{X}_i, \mathbf{X}_j) = \frac{12}{n(n^2-1)}\sum_{t=1}^{n}\left(\mathrm{rang}\left(X_{t,i} - \frac{1}{2}(n+1)\right)\mathrm{rang}\left(X_{t,j} - \frac{1}{2}(n+1)\right)\right).$$

## $\tau$ de Kendall

C'est une mesure nonparamétrique basée sur les rangs, qui mesure la concordance. Les paires $(X_{t1}, X_{t2})$ et $(X_{s1}, X_{s2})$ sont concordantes si $(X_{t1} - X_{s1})(X_{t2} - X_{s2}) > 0$.




**Définition 8.9 ($\tau$ de Kendall)**
Le $\tau$ de Kendall est définie comme étant $\tau := \mathsf{P}\,(\text{concordance}) - \mathsf{P}\,(\text{discordance})$, ou

$$\tau := 4 \iint C(u_1, u_2)\, \mathrm{d}C(u_1, u_2) - 1$$

Un estimateur empirique du coefficient de Kendall pour les colonnes $i, j$ de $\mathbf{X}$ est

$$\begin{aligned}\tau_n(\mathbf{X}_i, \mathbf{X}_j) &= \binom{n}{2}^{-1} \sum_{1 \leq t < s \leq n} \operatorname{sign}\left((X_{t,i} - X_{s,i})(X_{t,j} - X_{s,j})\right) \\ &= \frac{P_n - Q_n}{\binom{n}{2}} = \frac{4P_n}{n(n-1)} - 1\end{aligned}$$

où $P_n$ dénote le nombre de paires concordantes et $Q_n = n - P_n$ le nombre de paires discordantes. La collection résulte en des matrices positives semi-définies.

Le $\tau$ de Kendall et le $\varrho$ de Spearman sont basés sur les rangs et satisfont P1, P2 et P4. En revanche, leur dérivation peut être ardue.

**Note**
Il est impossible de résumer la dépendance et de condenser toute l'information dans un scalaire : les méta copules de la Figure 19 ont toutes le même $\tau$ de Kendall (à l'exception de la copule de FGM dont la borne supérieure est plus faible que $\tau = 1/3$). La plupart des mesures ne capturent pas les relations nonlinéaires complexes, comme un nuage de point en étoile tel qu'illustré dans la figure 21. C'est pourquoi la représentation graphique des données demeure cruciale.

### Coefficient de dépendance de queue

Il est possible de calculer la probabilité conditionnelle d'être dans le coin inférieur gauche (supérieur droit) du carré unitaire, cette probabilité mesure la dépendance entre les valeurs extrêmes. Dans le cas bivarié, le coefficient de dépendance de queue inférieure est défini comme la limite

$$\begin{aligned}\lambda_l &= \lim_{q\to 0^+} \mathsf{P}\left(X_2 \leq F_2^{\leftarrow}(q) \mid X_1 \leq F_1^{\leftarrow}(q)\right)\\ &= \lim_{q\to 0^+} \frac{\mathsf{P}\left(X_2 \leq F_2^{\leftarrow}(q), X_1 \leq F_1^{\leftarrow}(q)\right)}{\mathsf{P}\left(X_1 \leq F_1^{\leftarrow}(q)\right)}\\ &= \lim_{q\to 0^+} \frac{C(q,q)}{q}\end{aligned}$$

si la limite $\lambda_l \in [0,1]$ existe. Similairement, le coefficient de dépendance de queue supérieure est

$$\begin{aligned}\lambda_u &= \lim_{q\to 1^-} \mathsf{P}\left(X_2 > F_2^{\leftarrow}(q) \mid X_1 > F_1^{\leftarrow}(q)\right)\\ &= \lim_{q\to 1^-} \frac{1-2q+C(q,q)}{1-q}.\end{aligned}$$

Ce dernier est non-nul pour les modèles de valeurs extrêmes.

## 8.4 Familles et modèles

Les classes de modèles de copules les plus communes sont

1. induites ou implicites (copule découlant du théorème de Sklar, comme par exemple les copules elliptiques ou de valeurs extrêmes)
2. explicites (construction directe de $C$ ; notamment les copules archimédiennes).

Nous revoyons à tour de rôle les différentes classes de modèles et les propriétés définissant ces dernières.

### 8.4.1. Copules elliptiques

Bien connues, ces copules sont dites induites, c'est-à-dire que leur construction découle du théorème de Sklar et des modèles définis précédemment (définition 2.17). Quelques métacopules gaussiennes (modèles dont les marginales sont transformées d'uniformes à normales) sont présentées en figure 21.

Les copules elliptiques sont populaires pour plusieurs raisons. D'abord, leur densité est explicite, bien que la copule soit implicite (sous forme d'intégrale, comme la fonction de répartition de la loi normale). Par exemple, la fonction de répartition de la copule

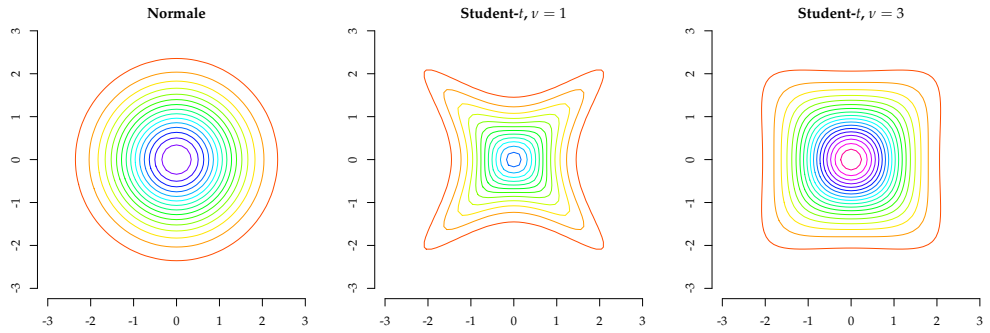



FIGURE 21 – Contour de la densité des métacopules pour des lois elliptiques (marges $\mathcal{N}(0,1)$).

gaussienne pour $d = 2$

$$C_{\mathcal{N}}(\boldsymbol{u};\rho) = \frac{1}{\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \exp\left(-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}\right) \mathrm{d}y\,\mathrm{d}x$$

tandis que la fonction de densité est

$$c_{\mathcal{N}}(\boldsymbol{u};\rho) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(\frac{2\rho\Phi^{-1}(u_1)\Phi^{-1}(u_2) - \rho^2(\Phi^{-1}(u_1)^2 + \Phi^{-1}(u_2)^2)}{2(1-\rho^2)}\right).$$

Les copules elliptiques sont faciles à échantillonner et sont utilisées couramment en pratique de ce fait. Une restriction est la symétrie radiale qui associe la fonction de répartition centrée à sa fonction de survie, $\bar{C}(\boldsymbol{u}) = C(\boldsymbol{u})$ ; cette propriété a des implications sur les coefficients de queue, qui doivent être égaux ($\lambda_l = \lambda_u$). Des exemples connus sont les copules gaussienne, $t_\nu$, normale inverse gaussienne (NIG), hyperbolique symétrique généralisée, Laplace, etc.

**Exemple 8.9 (Copules elliptiques)**
On peut considérer des distributions où $\boldsymbol{\mu} = \mathbf{0}_d$ et $\boldsymbol{\Sigma} = \wp(\boldsymbol{\Sigma})$ où $\wp(\Sigma)$ est la matrice de corrélation. La densité est alors

$$f_{\wp(\boldsymbol{\Sigma})} = \frac{1}{|\wp(\boldsymbol{\Sigma})|^{1/2}} g\left(-\frac{1}{2}\boldsymbol{x}^\top(\wp(\boldsymbol{\Sigma}))^{-1}\boldsymbol{x}\right).$$

Les dérivations peuvent découler de la forme quadratique $R^2 = \boldsymbol{X}^\top(\wp(\boldsymbol{\Sigma}))^{-1}\boldsymbol{X}$ ; voir le tableau 1. Le coefficient de dépendance de queue de la copule gaussienne est 0 (ce qui en fait un très mauvais choix pour la modélisation de retours financiers), tandis que

pour la copule de Student-$t$,

$$\lambda_l = \lambda_u = 2t_{\nu+1}\sqrt{\nu+1}\sqrt{\frac{1-\rho_{ij}}{1+\rho_{ij}}}$$

où $\rho_{ij}$ est l'élément hors-diagonale de $\wp(\mathbf{\Sigma})$ (voir Demarta & McNeil (2005)).

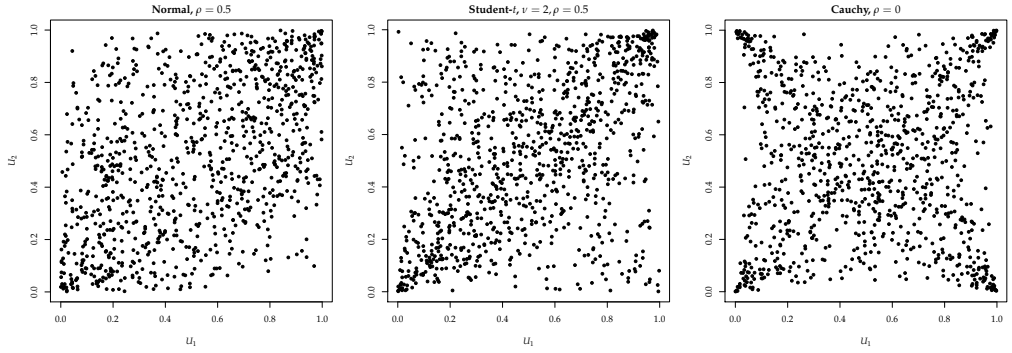



FIGURE 22 – Échantillons de copules elliptiques

**Proposition 8.10 (Relation entre mesures de concordance)**
Nous notons au passage quelques identités fort utiles : une relation entre $\rho$ et $\tau$, valide pour toute copule elliptique, de même qu'une relation entre $\rho$ et $\varrho$ qui elle n'est valide que pour la copule gaussienne. Elles sont données respectivement par

$$\tau = \frac{2}{\pi}\arcsin(\rho)$$
$$\varrho = \frac{6}{\pi}\arcsin\left(\frac{\rho}{2}\right)$$

## 8.4.2. Copules archimédiennes

**Définition 8.11 (copule archimédienne)**
Une copule copule archimédienne admet la représentation

$$C_\psi(\boldsymbol{u}) = \psi(\psi^{\leftarrow}(u_1) + \cdots + \psi^{\leftarrow}(u_d)) \tag{8.11}$$

où $\psi : [0,\infty) \to [0,1]$, le générateur archimédien de la copule, est une application $d$-monotone avec $\psi(0) = 1$ et $\lim_{x\to\infty}\psi(x) = 0$ (si strict). Voir McNeil & Nešlehová (2009) pour plus à ce sujet et pour une représentation stochastique.

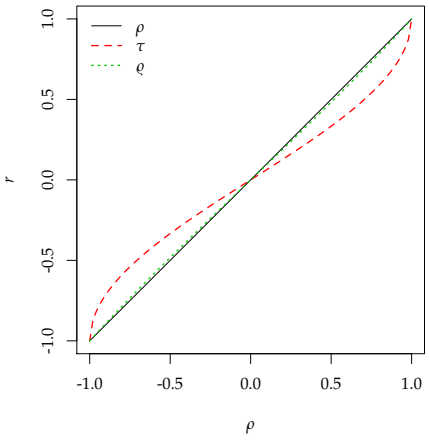



FIGURE 23 – Relation entre $\rho$, $\varrho$ et $\tau$ pour la copule gaussienne

Le (pseudo)-inverse $\psi^{\leftarrow} : [0,1] \to [0,\infty]$ est donné par l'inverse de $\psi$ sur $(0,1]$ et par

$$\psi^{\leftarrow}(0) = \inf\{x : \psi(x) = 0\}.$$

Dans le cas bivarié, on a

**Définition 8.12 (copule archimédienne bivariée)**

$$C_\psi(u_1, u_2) = \psi(\psi^{\leftarrow}(u_1) + \psi^{\leftarrow}(u_2))$$

pour une fonction convexe $\psi$ telle que $\psi(0) = 1$ et $\lim_{x\to\infty} \psi(x) = 0$.

Les copules archimédiennes sont très populaires, notamment parce qu'elles sont apparues très tôt dans la littérature. Elles ont souvent un ou deux paramètres de dépendance, ce qui facilite l'estimation mais limite leur applicabilité. De nombreuses propriétés extrémales et de dépendance sont disponibles à partir du générateur $\psi$. Elles sont échangeables (symétriques). Certaines familles couramment utilisées incluent les copules de Gumbel–Hougaard, Joe, Clayton, Ali–Mikhail–Haq et Frank. Pour les cinq familles, les densités sont disponibles sous forme fermée et l'erreur moyenne quadratique est $\propto 1/nd$ si les marges sont connues, voir l'article de Höfert, Machler et McNeil (2012) pour des indications à ce sujet et les difficultés numériques sous-jacentes. Le modèle est en revanche difficilement réaliste en dimension élevée, mais des modèles de copules imbriquées ont commencé à émerger dans la littérature.

Une formule pour le $\tau$ de Kendall est disponible, dérivée par Genest & Rivest (1993),

est

$$\tau = 1 + 4 \int_0^1 \frac{\psi^{\leftarrow}(t)}{(\psi^{\leftarrow}(t))'} \, \mathrm{d}t$$

tandis que les coefficient de dépendance de queue sont donnés par

$$\lambda_u = 2 - 2 \lim_{q \to 0^+} \frac{\psi'(2q)}{\psi'(q)}$$
$$\lambda_l = 2 \lim_{q \to \infty} \frac{\psi'(2q)}{\psi'(q)}$$

Tableau 12 – Générateurs et propriétés de quelques copules archimédiennes

| **Famille** | $\Theta_C$ | $\psi(t)$ |
|---|---|---|
| A | $[0,1)$ | $(1-\theta)/(e^t-\theta)$ |
| C | $(0,\infty)$ | $(1+\theta t)^{-1/\theta}$ |
| F | $(0,\infty)$ | $-\log(1-(1-e^{-\theta})e^{-t})/\theta$ |
| G | $[1,\infty)$ | $\exp\left(-t^{-1/\theta}\right)$ |
| J | $[1,\infty)$ | $1-(1-e^{-t})^{1/\theta}$ |

| **Famille** | $\tau$ | $\mathsf{Im}(\tau)$ | $\lambda_L$ | $\lambda_U$ |
|---|---|---|---|---|
| A | $1-[2(\theta+(1-\theta)^2\log(1-\theta)]/3\theta^2$ | $(0,\frac{1}{3})$ | 0 | 0 |
| C | $\theta/(\theta+2)$ | (0,1) | $2^{-1/\theta}$ | 0 |
| F | $1-4\left(\int_0^\theta t/[\theta(e^t-1)]\,\mathrm{d}t-1\right)/\theta$ | (0,1) | 0 | 0 |
| G | $(\theta-1)/\theta$ | [0,1] | 0 | $2-2^{1/\theta}$ |
| J | $1-4\sum_{k=1}^{\infty}1/[k(\theta k+2)(\theta(k-1)+2)]$ | [0,1] | 0 | $2-2^{1/\theta}$ |

Les familles sont celles de Ali-Mikhail-Haq (A), Clayton (C), Frank (F), Gumbel-Hougaard (G) et Joe (J).

Nous nous cantonnerons dorénavant à l'étude de fonctions dites complètement monotone et sont valide pour tout $d \in 2, 3, \ldots$, comme celles présentées dans le tableau 12.

**Définition 8.13 (fonction complètement monotone)**
Une fonction est complètement monotone si $(-1)^k \psi^{(k)}(x) \geq 0$ pour tout $k \in \mathbb{Z}^+$ et $\psi$ est continue sur $[0,\infty]$.

Par le théorème de Bernstein, la transformée de Laplace d'une variable aléatoire positive est complètement monotone, et donc $\psi = \mathcal{L}(F)$ est un candidat pour un générateur archimédien. Pour l'échantillonnage, on procède généralement en utilisant l'algorithme de Marshall & Olkin (1988).

**Algorithme 8.1 (Algorithme de Marshall-Olkin (1988))**
Si $\psi$ est complètement monotone :

1. Échantillonner $X_0 \sim F_0 = \mathcal{L}^{-1}(\psi)$

2. Échantillonner $\boldsymbol{X}$, où $X_j \sim \mathcal{U}(0,1)$ pour $j = 1, \ldots, d$.
3. Retourner $U_i = \psi(-\log(X_j)/X_0)$ pour $j = 1, \ldots, d$.

### 8.4.3. Copules de valeurs extrêmes

En dimension $d = 2$, ces copules sont de la forme

$$C_A(u_1, u_2) = \exp\left(\log(u_1 u_2) \mathrm{A}\left(\frac{\log(u_2)}{\log(u_1 u_2)}\right)\right)$$

où la fonction de Pickands $\mathrm{A} : [0,1] \to [1/2, 1]$ est une application convexe telle que

$$\max(t, 1-t) \leq \mathrm{A}(t) \leq 1, \qquad t \in [0,1]$$

Comme leur nom l'indique, les copules de valeurs extrêmes ont des connections avec les distributions max-stables, auxquelles elles sont complètement équivalentes.

### 8.4.4. Exemples de modèles archimédiens

**Définition 8.14 (Copule de Gumbel–Hougaard)**
C'est une copule archimédienne et de valeur extrême (correspondant à la distribution de valeur extrêmes logistique), dont le générateur est donné par

$$\psi(t) = \exp(-t^{1/\theta}), \qquad \psi^{\leftarrow}(t) = (-\log(t))^{\theta}$$

pour $\theta \in [1, \infty)$ et la copule s'écrit comme

$$C_\theta(\boldsymbol{u}) = \exp\left(-\left(\sum_{i=d}^{d}(-\log(u_i))^{\theta}\right)^{1/\theta}\right).$$

Cas limites :

- $C_{\mathrm{Gu}}(\boldsymbol{u} \mid \theta = 1) = \Pi(\boldsymbol{u})$, la copule d'indépendance et
- $\lim_{\theta \to \infty} C_{\mathrm{Gu}}(\boldsymbol{u} \mid \theta) = \mathrm{M}(\boldsymbol{u})$.

$\tau_{\mathrm{Gu}} = (\theta - 1)/\theta$ et seule la dépendance positive est modélisée : $\tau \in (0,1)$

**Définition 8.15 (Copule de Clayton)**
Pour $\theta \geq -1/(d-1)$, le générateur

$$\psi(t) = (1 + \theta t)_+^{-1/\theta}, \qquad \psi^{\leftarrow}(t) = \frac{t^{-\theta} - 1}{\theta}$$

strict si $\theta > 0$, peut être vu comme transformée de Laplace d'une loi $\Gamma(1/\theta, 1/\theta)$. Le $\tau$ de Kendall est $\tau_{\mathrm{Cl}} = \theta/(\theta + 2)$ et est borné entre $-1/(2d-3)$ et 1. D'autre propriétés sont dérivées dans la série 13.

**Note**

La paramétrisation de la copule de Clayton présenté ci-dessus diffère de celle de R, qui utilise plutôt $\psi(t) = (1+t)^{-1/\theta}$ comme générateur.

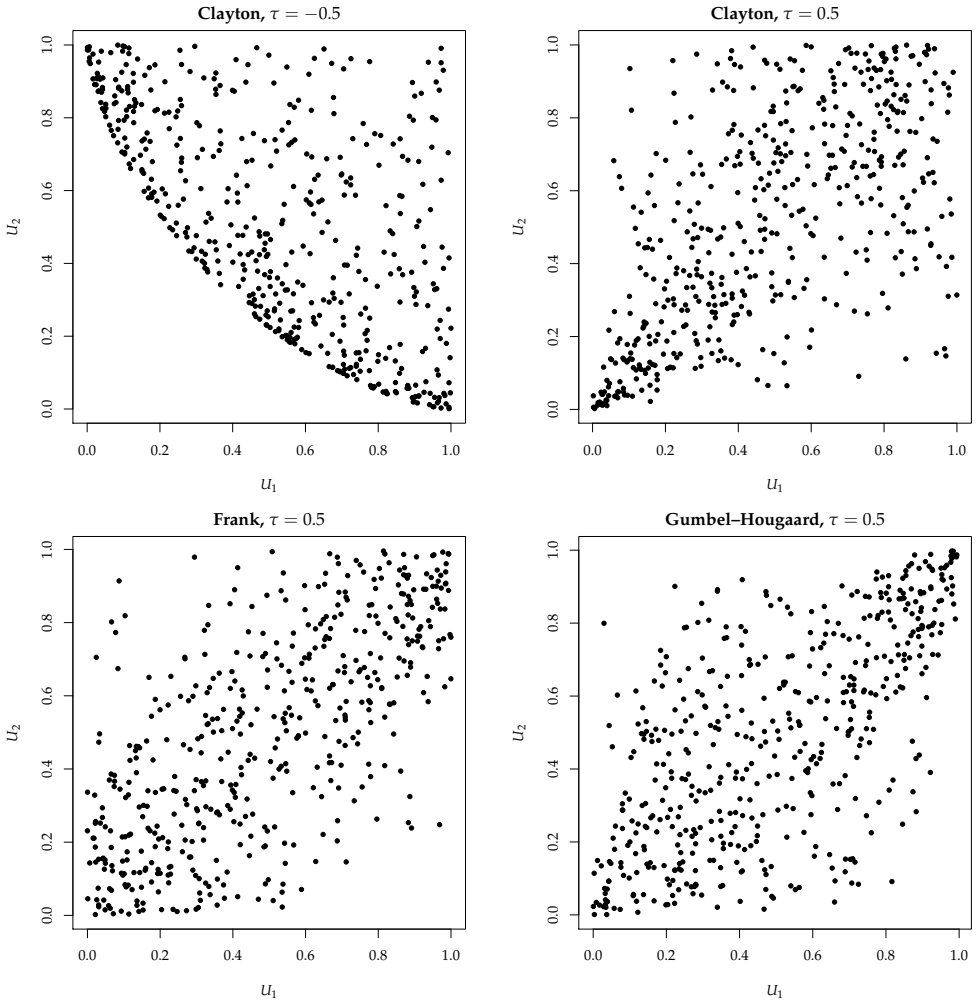



FIGURE 24 – Échantillons de copules archimédiennes

## 8.5 Estimation et inférence

### Inférence nonparamétrique

Même l'estimation nonparamétrique copule en utilisant une extension multivariée de la fonction de répartition empirique, il est préférable de spécifier un modèle paramétrique pour la copule. L'approche non-paramétrique repose sur le processus empirique $\sqrt{n}(C_n - C)$ ; ce dernier converge en loi vers un champ gaussien sous quelques hypo-

thèses de régularité, mais nécessite $n$ grand. La copule empirique est néanmoins au coeur de la plupart des tests d'indépendance et d'adéquation.

**Définition 8.16 (copule empirique)**
La copule empirique, proposée par Deheuvels (1979), est

$$C_n(\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{I}\left(\bigcup_{j=1}^{d} \frac{R_{i,j}}{n+1} \leq u_j\right)$$

Cette copule est l'équivalent de la fonction empirique en dimension $d$ avec des marges uniformes.

## Inférence paramétrique

Soit $\mathbf{X}$ un échantillon $n \times d$ de vecteurs aléatoires $\boldsymbol{X}$ de fonctions de répartition $F$, ou de façon équivalente de lois marginales continues $F_1, \ldots, F_d$ et de copule $C$. On suppose que les données sont générées d'un modèle paramétrique $F(\boldsymbol{\theta}_0)$, où le vecteur de paramètres peut être partitionné en $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,1}, \ldots, \boldsymbol{\theta}_{0,d}, \boldsymbol{\theta}_{0,C}) \in \boldsymbol{\theta}$, avec $F_j = F_j(\cdot; \boldsymbol{\theta}_{0,j})$ pour $j = 1, \ldots, d$ et $C = C(\cdot; \boldsymbol{\theta}_{0,C})$ et tel que $F_j(; \boldsymbol{\theta}_j)$ est continue pour tout $\boldsymbol{\theta}_j \in \Theta_j$.

**Définition 8.17 (Densité d'une copule)**
En vertu du théorème de Sklar, la densité de $F$ (si elle existe) prend la forme

$$f(\boldsymbol{x}; \boldsymbol{\theta}_0) = c(F_1(x_1; \boldsymbol{\theta}_{0,1}, \ldots F_d(x_d; \boldsymbol{\theta}_{0,d}); \boldsymbol{\theta}_{0,C}) \prod_{j=1}^{d} f_j(x_j; \boldsymbol{\theta}_{0,j}),$$

où

$$c(\boldsymbol{u}) = \frac{\partial^d}{\partial u_1 \cdots \partial u_d} C(u_1, \ldots, u_d), \qquad \boldsymbol{u} \in [0,1]^d$$

La log-vraisemblance basée sur $\mathbf{X}$ est

$$\begin{aligned}\ell(\boldsymbol{\theta}; \mathbf{X}) &= \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; \boldsymbol{X}_i) \\ &= \sum_{i=1}^{n} \ell_C(\boldsymbol{\theta}_C; F_1(X_{i1}; \boldsymbol{\theta}_1), \ldots, F_d(X_{id}; \boldsymbol{\theta}_d)) + \sum_{i=1}^{n}\sum_{j=1}^{d} \ell_j(\boldsymbol{\theta}_j; X_{ij}),\end{aligned}$$

où

$$\begin{aligned}\ell_C(\boldsymbol{\theta}_C; u_1, \ldots, u_d) &= \log c(u_1, \ldots, u_d; \boldsymbol{\theta}_C) \\ \ell_j(\boldsymbol{\theta}_j; x) &= \log f_j(x; \boldsymbol{\theta}_j), \qquad j = 1, \ldots, d.\end{aligned}$$

L'estimateur du maximum de vraisemblance est donné par

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}; \mathbf{X})$$

Les modèles multivariés peuvent impliquer un grand nombre de paramètres si on optimise simultanément les distributions conjointes et marginales. L'optimisation multivariée peut être également numériquement instable.

Pour cette raison, la plupart des praticiens procèdent en deux étapes : estimation des lois marginales d'abord, puis des paramètres de la copule.

Joe et Xu (1996) ont proposé une approche pour l'inférence en deux étapes.

**Proposition 8.18 (Inférence pour les marges)**

1. Estimer $\boldsymbol{\theta}_{0,j}$ par son maximum de vraisemblance $\widehat{\boldsymbol{\theta}}_j$ à partir de la log-vraisemblance marginale $\sum_{i=1}^{n} \log f_j(\boldsymbol{\theta}_j; \boldsymbol{x}_{ij})$
2. Estimer $\boldsymbol{\theta}_{0,C}$ par

$$\widehat{\boldsymbol{\theta}}_C^{\mathsf{IPM}} = \arg\max_{\boldsymbol{\theta}_C\in\Theta_C} \ell(\boldsymbol{\theta}_C, \widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_d; \mathbf{X}).$$

Les fonctions d'inférence pour les estimateurs sont donc $(\widehat{\boldsymbol{\theta}}_C^{\mathsf{IPM}}, \widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_d)$.

Cette méthode est facile à implémenter puisqu'elle implique typiquement $d$ optimisations de dimension respective $|\boldsymbol{\theta}_j|$ pour $j = 1, \ldots, d$ ainsi qu'une optimisation pour $\boldsymbol{\theta}_{0,C}$. Les estimés sont prouvables asymptotiquement normaux. Ils sont en revanche sujets à propagation de l'erreur en cas de mis-spécification des distributions marginales.

## Inférence semiparamétrique

Une autre approche, la maximisation de la pseudo-vraisemblance, est suggérée par Oakes (1994), Genest & Rivest (1995) et adaptée pour les données censurées par Shih & Louis (1995) consiste à estimer les marginales **nonparamétriquement**.

**Proposition 8.19 (Estimateur du maximum de pseudo-vraisemblance)**
À partir d'un échantillon $\mathbf{X}$,

1. Calculer les pseudos-observations basées sur les rangs $\widetilde{\boldsymbol{U}}_1, \ldots, \widetilde{\boldsymbol{U}}_n$ où

$$\widetilde{U}_{ij} = \frac{n}{n+1} \widetilde{F}_{n,j}(X_{ij}) = \frac{R_{ij}}{n+1}$$

et $R_{ij}$ dénote le rang de $X_{ij}$ parmi $X_{1j}, \ldots, X_{nj}$.

2. Estimer $\widehat{\boldsymbol{\theta}}_{0,C}$ par

$$\widehat{\boldsymbol{\theta}}_C^{\mathsf{EMPV}} = \underset{\boldsymbol{\theta}_C \in \Theta_C}{\arg\max} \sum_{i=1}^{n} \ell_C(\boldsymbol{\theta}_C; \widetilde{U}_{i1}, \ldots, \widetilde{U}_{id}) = \underset{\boldsymbol{\theta}_C \in \Theta_C}{\arg\max} \sum_{i=1}^{n} \log c(\widetilde{\boldsymbol{U}}_i; \boldsymbol{\theta}_C).$$

L'estimateur du maximum de pseudo-vraisemblance n'est pas asymptotiquement efficace, mais est plus robuste et adéquat pour $n$ grand. Cette méthode est également facile à implémenter, les estimés sont asymptotiquement normaux et crucialement invariants à des transformations monotones croissantes des données puisque basés sur des rangs. Il ne faut cependant pas que les marges dépendent de covariables.

La dernière méthode semi-paramétrique est basée sur le méthode des moments. Dans le cas où $\theta_{0,C}$ est unidimensionnel, il est possible d'obtenir pour plusieurs familles de copules des formules explicites pour le $\tau$ de Kendall ou le $\rho$ de Spearman. On peut alors procéder à l'inversion de $\varrho$ ou de $\tau$ : si la copule est paramétrisée par $\theta_{0,C}$, cela revient à estimer $\widehat{\theta}_n := \tau^{-1}(\tau_n)$. $\tau_n$ est une $U$-statistique et de ce fait il est possible de prouver la normalité asymptotique de cette méthode basée sur la méthode $\delta$.

## Validation de modèles

Une fois l'estimation complétée, une question cruciale demeure la validation du modèle. À ce sujet, quelques tests d'adéquations, c'est-à-dire le test de l'hypothèse $\mathscr{H}_0$ : $C \in \mathcal{C}_\theta$, sont disponibles : voir Genest & al (2009) et la revue de littérature de l'article, de même que Kojadinovic & Yan (2011). Ces tests sont basés sur une technique d'auto-amorçage (*bootstrap*) paramétrique ou multiplicatif, et parmi les hypothèses nécessaires à leur application figurent notamment l'hypothèse d'unicité des rangs (absence de doublons).

### Remarque

L'adéquation ne doit pas être confondue avec la sélection de modèles. Le praticien doit utiliser un modèle dont les propriétés empiriques ressemblent à celle du modèle choisi. Peu d'articles se penchent sur la sélection, mais Huard, Évin & Favre (2006) proposent une alternative bayésienne.

# Index