Éléments de classification

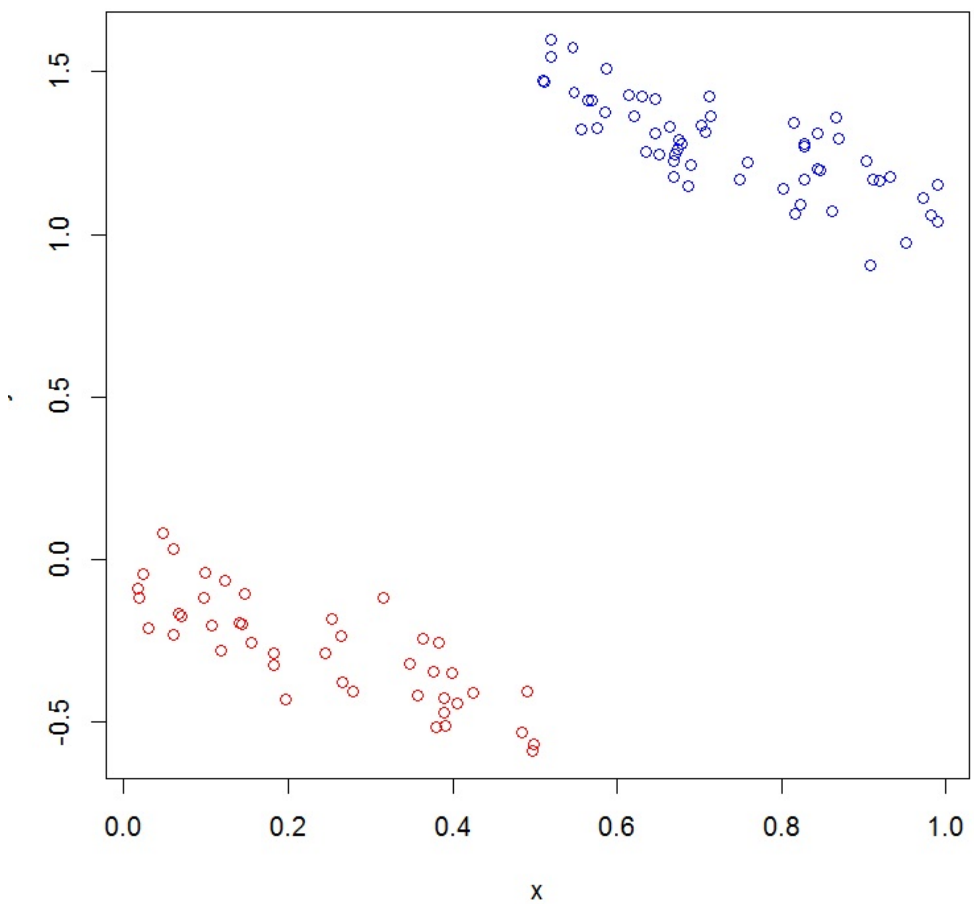
Christophe Chesneau

Université de Caen

Éléments de classification

Christophe Chesneau

https://chesneau.users.lmno.cnrs.fr/



Caen, le 20 Février 2020

Table des matières

∼ Note ∼

Ce document résume les notions abordées dans le cours *Éléments de classification* du Master 1 MIASHS de l'université de Caen.

Un des objectifs est de donner des pistes de réflexion au regroupement/classification des individus à partir de données.

Les méthodes statistiques y sont décrites de manière concise, avec les commandes R associées.

N'hésitez pas à me contacter pour tout commentaire :

`christophe.chesneau@gmail.com`

Bonne lecture !

1 Introduction

On présente ici les enjeux de la classification non-supervisée et de la classification supervisée.

1.1 Classification non-supervisée

Contexte : Pour n individus d'une population, on dispose des valeurs de p caractères $X_1, \ldots, X_p$. Ces valeurs constituent les données.

Objectif : Partant des données, l'objectif est de regrouper/classer les individus qui se ressemblent le plus/qui ont des caractéristiques semblables.

Ce regroupement peut avoir des buts divers : tenter de séparer des individus appartenant à des sous-populations distinctes, décrire les données en procédant à une réduction du nombre d'individus pour communiquer, simplifier, exposer les résultats...

Exemple : Dans une classe, un professeur souhaite faire des binômes constitués d'élèves ayant des compétences semblables. Parmi ceux-ci, 6 élèves ont obtenu les notes suivantes :

	Maths	Physique	Ed Mus	Art Plas
Boris	20	20	0	0
Mohammad	8	8	12	12
Stéphanie	20	20	0	0
Jean	0	0	20	20
Lilly	10	10	10	10
Annabelle	2	2	18	18

Tous les élèves ont une moyenne de 10/20 mais, vu les notes,

- Boris et Stéphanie ont un profil similaire,
- Mohammad et Lilly ont un profil similaire,
- Jean et Annabelle ont un profil similaire.

Finalement, le professeur décide de faire 2 groupes cohérents de 3 élèves avec ces 6 élèves. Lesquels proposez-vous ?

En comparant les notes par matière, on propose :

- *Groupe 1 :* Boris, Stéphanie et Lilly,
- *Groupe 2 :* Mohammad, Jean et Annabelle.

De plus, par exemple, le profil de Jean est plus proche de celui de Lilly, que celui de Stéphanie.

Bien entendu, cette analyse intuitive n'est pas possible si, par exemple, on a 30 élèves à classer par groupes de 3 et on considère 12 matières. C'est pourquoi des méthodes mathématiques ont été mises en place.

Applications :

- *Application 1 :* En biologie, on veut regrouper les espèces suivant leurs caractéristiques et donc leurs origines communes.
- *Application 2 :* En psychologie, on veut classer les individus selon leur type de personnalités.
- *Application 3 :* En chimie, on veut classer des composés selon leurs propriétés.
- *Application 4 :* Dans l'industrie, on souhaite
  - analyser des résultats d'enquêtes,
  - identifier les clients potentiels d'une entreprise,
  - identifier les clients susceptibles de partir à la concurrence,
  - déterminer des lieux de ventes (pose de distributeurs de billets. . .),
  - analyser, identifier les risques (dégâts des eaux. . .),
  - analyser des données textuelles.

1.2 Classification supervisée

Contexte : On considère une population divisée en q groupes d'individus différents. Ces groupes sont distinguables suivant les valeurs de p caractères $X_1, \ldots, X_p$, sans que l'on ait connaissance des valeurs de $X_1, \ldots, X_p$ les caractérisant.

On dispose

- de n individus avec, pour chacun d'entre eux, les valeurs de $X_1, \ldots, X_p$ et son groupe d'appartenance,

- d'un individu ω_* de la population avec ses valeurs de $X_1, \ldots, X_p$, mais sans connaissance de son groupe d'appartenance.

Objectif : Partant des données, l'objectif est de déterminer à quel groupe l'individu ω_* a le plus de chances d'appartenir.

Exemple : Dans une classe, un professeur considère deux groupes d'élèves, $G1$ et $G2$, en fonction de leur compétence. On dispose uniquement des notes et de l'affectation de 6 élèves :

	Maths	Physique	Ed Mus	Art Plas	Groupe
Boris	20	20	0	0	G1
Mohammad	8	8	12	12	G2
Stéphanie	20	20	0	0	G1
Jean	0	0	20	20	G2
Lilly	10	10	10	10	G1
Annabelle	2	2	18	18	G2

D'autre part, un étudiant de la classe, Bob, a les résultats suivants :

	Maths	Physique	Ed Mus	Art Plas	Groupe
Bob	9	15	13	11	inconnu

À partir de ses notes, à quel groupe Bob a le plus de chances d'appartenir ?

Autrement écrit, quelle est la probabilité que Bob appartienne au groupe G1 sachant qu'il a obtenu les notes (Maths, Physique, Ed Mus, Art Plas)$= (9, 15, 13, 11)$?

Soit encore, avec une modélisation probabiliste adaptée, que vaut la probabilité :

"$\mathbb{P}(\{\text{Bob} \in G1\}/\{(\text{Maths, Physique, Ed Mus, Art Plas}) = (9, 15, 13, 11)\})$" ?

La réponse n'est pas immédiate ; c'est pourquoi des méthodes mathématiques ont été mises en place.

Applications :

- *Application 1 :* Un archéologue cherche à déterminer si des restes humains sont ceux d'un homme ou d'une femme.
- *Application 2 :* Dans une banque, une commission de crédit doit décider, à partir de paramètres financiers, si on accorde ou non un prêt à un particulier.

- *Application 3 :* Étant donné un ensemble de symptômes, un médecin doit poser un diagnostic.
- *Application 4 :* Dans l'industrie, on veut
  - identifier des visages, des empreintes digitales,
  - identifier des objets dans des séquences vidéos,
  - rechercher des clients potentiels dans des bases de données,
  - rapprocher un ou plusieurs mots de manière pertinente au texte le plus pertinent.

1.3 Les métiers

Il y a de nombreux métiers où la classification est utilisée dont

- responsable logistique du traitement et de l'analyse des études,
- chargé d'études junior : prise en charge de la documentation, codage des questionnaires, traitement statistiques simples,
- chargé d'études senior, assistant du chargé d'étude : prise en main d'une étude de marché,
- analyste statisticien, études quantitatives, aide à la décision, expert en statistiques,
- chef de projet.

2 Enjeu de la classification non-supervisée

Contexte : Pour n individus $\omega_1, \ldots, \omega_n$ d'une population, on dispose des valeurs de p caractères quantitatifs $X_1, \ldots, X_p$. Pour tout $i \in \{1, \ldots, n\}$, celles associées à ω_i sont notées $x_{1,i}, \ldots, x_{p,i}$. Elles sont généralement présentées sous la forme d'un tableau :

	X_1	$\ldots$	X_p
ω_1	$x_{1,1}$	$\ldots$	$x_{p,1}$
$\vdots$	$\vdots$	$\ldots$	$\vdots$
ω_n	$x_{1,n}$	$\ldots$	$x_{p,n}$

Ces valeurs constituent les données.

Objectif : Partant des données, l'objectif est de regrouper/classer les individus qui se ressemblent le plus/qui ont des caractéristiques semblables.

Règle n^{o}1 : Qui se ressemble s'assemble.

Méthodes : Pour atteindre l'objectif, plusieurs méthodes sont possibles. Parmi elles, il y a

- l'algorithme de classification ascendante hiérarchique (CAH),
- l'algorithme des centres mobiles,
- l'algorithme de classification descendante hiérarchique (CDH),
- la méthode des nuées dynamiques (partitionnement autour d'un noyau),
- la méthode de classification floue,
- la méthode de classification par voisinage dense.

Ce document aborde quelques aspects des deux premiers points.

3 Étude de la ressemblance

3.1 Nuage de points

Ensemble des n individus : L'ensemble des n individus considérés est noté $\Gamma = \{\omega_1, \ldots, \omega_n\}$.

Matrice de données : On appelle matrice de données associée à Γ la matrice $\mathbf{X}$ a n lignes et p colonnes définie par

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \ldots & x_{p,1} \\ \vdots & \ldots & \vdots \\ x_{1,n} & \ldots & x_{p,n} \end{pmatrix}.$$

Nuage de points : Pour tout $i \in \{1, \ldots, n\}$, l'individu ω_i peut être représenté dans $\mathbb{R}^p$ par un point m_i de coordonnées $(x_{1,i}, \ldots, x_{p,i})$. On appelle nuage de points la représentation graphique de l'ensemble de ces points. Il est noté $\mathcal{N} = \{m_1, \ldots, m_n\}$.

Ressemblance : On dira que des individus se ressemblent si les points associés sont proches les uns des autres/si les distance qui les séparent sont petites.

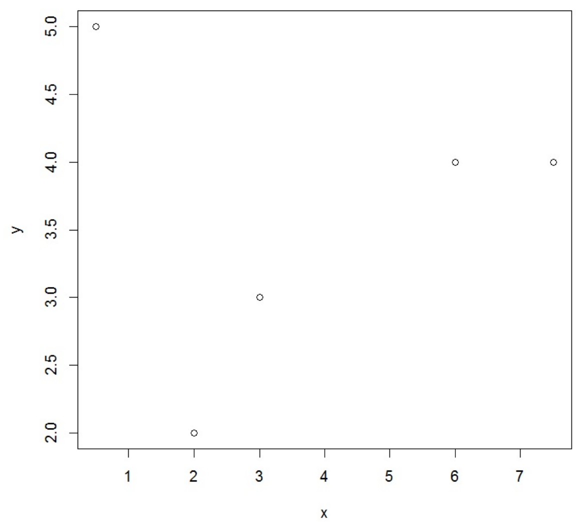
Ainsi, on souhaite rechercher dans $\mathcal{N}$ les zones denses pouvant correspondre à des groupes d'individus qu'il s'agira d'interpréter par la suite.

Exemple 1 : On considère la matrice de données $\mathbf{X}$ associée à 5 individus, $\Gamma = \{\omega_1, \ldots, \omega_5\}$, définie par

$$\mathbf{X} = \begin{pmatrix} 2 & 2 \\ 7.5 & 4 \\ 3 & 3 \\ 0.5 & 5 \\ 6 & 4 \end{pmatrix}.$$

Implicitement, on considère donc 2 caractères X_1 et X_2. Par exemple, l'individu ω_2 a pour caractéristiques $X_1 = 7.5$ et $X_2 = 4$.

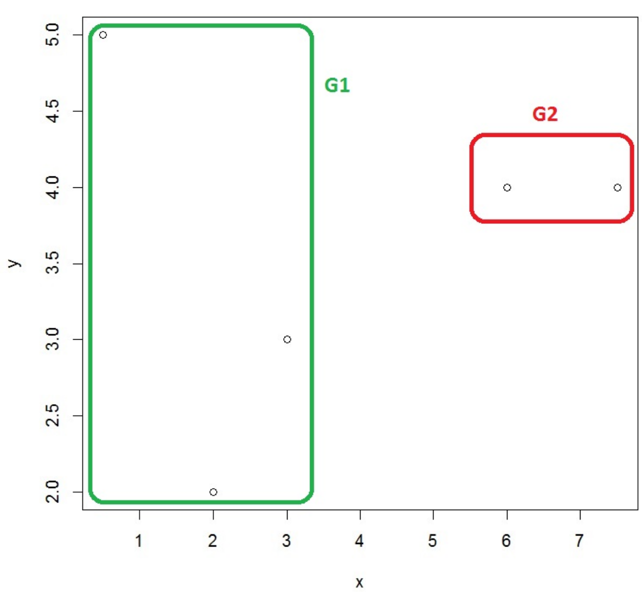
Le nuage de point associé est :



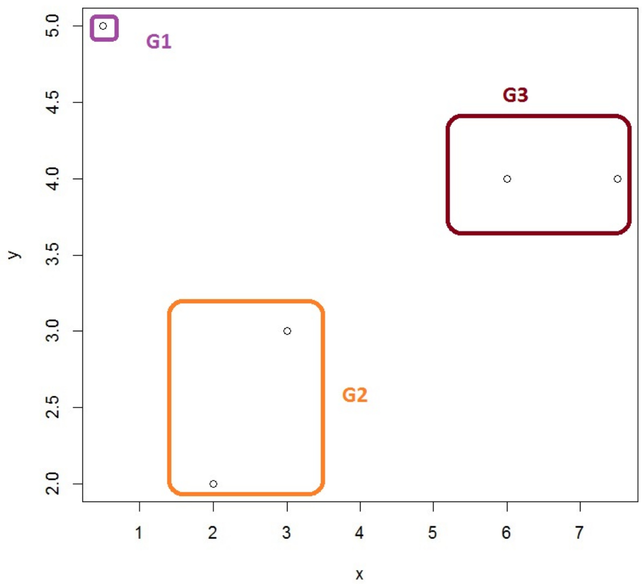
La problématique est la suivante : comment regrouper ces individus en 2 ou 3 groupes, par exemple, en fonction de leur position dans $\mathbb{R}^2$?

Visuellement, en fonction des zones denses, on peut envisager

∘ les 2 groupes suivants :



◦ les 3 groupes suivants :



Exemple 2 : On considère un tableau de notes de 6 élèves :

	Maths	Physique	Ed Mus	Art Plas
Boris	19	17	2	8
Mohammad	7	8	12	12
Stéphanie	20	19	9	9
Jean	1	6	18	17
Lilly	10	11	12	12
Annabelle	2	12	18	18

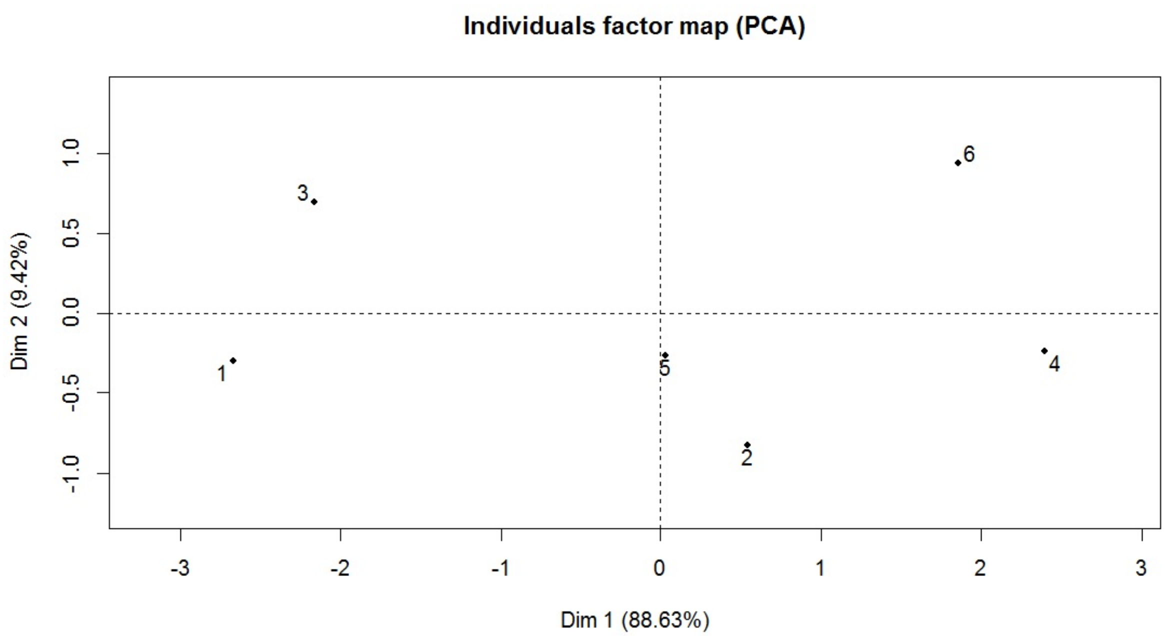
La matrice de données $\mathbf{X}$ associée est

$$\mathbf{X} = \begin{pmatrix} 19 & 17 & 2 & 8 \\ 7 & 8 & 12 & 12 \\ 20 & 19 & 9 & 9 \\ 1 & 6 & 18 & 17 \\ 10 & 11 & 12 & 12 \\ 2 & 12 & 18 & 18 \end{pmatrix}$$

La problématique est la suivante : comment regrouper ces individus en 2 ou 3 groupes, par exemple, en fonction de leur position dans $\mathbb{R}^4$?

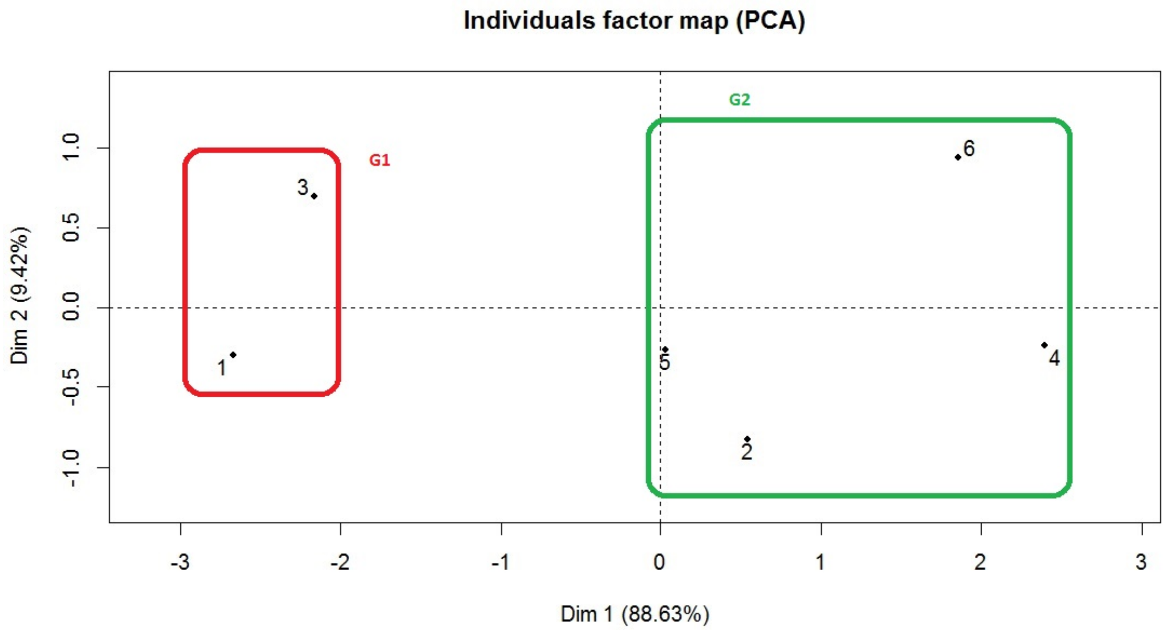
La représentation graphique dans $\mathbb{R}^2$ n'est donc pas possible.

Une solution est de considérer le plan principal d'une analyse en composante principale (ACP), méthode statistique qui ne sera pas développée ici. Cela donne la représentation graphique suivante :

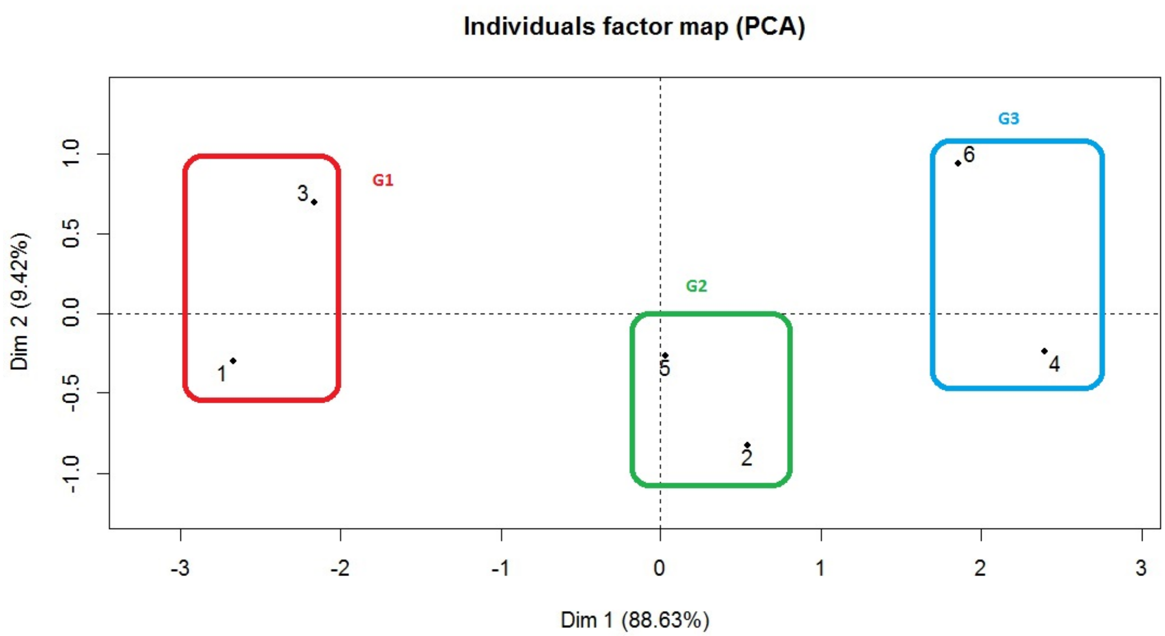


Visuellement, en fonction des zones denses, on peut envisager

◦ les 2 groupes suivants :



◦ les 3 groupes suivants :



3.2 Distances

Distances : On peut donc aborder le problème de la ressemblance entre individus par le biais de la notion de distance. On appelle distance sur un ensemble M toute application $d : M^2 \to [0, \infty[$ telle que

- pour tout $(x, y) \in M^2$, on a $d(x, y) = 0$ si, et seulement si, $x = y$,
- pour tout $(x, y) \in M^2$, on a $d(x, y) = d(y, x)$,
- pour tout $(x, y, z) \in M^3$, on a $d(x, y) \leq d(x, z) + d(z, y)$.

Exemple 1 : distance euclidienne : Soient $m \in \mathbb{N}^*$, $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ et $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$. On appelle distance euclidienne entre x et y la distance :

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}.$$

Exemple 2 : distance de Manhattan : Soient $m \in \mathbb{N}^*$, $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ et $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$. On appelle distance de Manhattan entre x et y la distance :

$$d(x, y) = \sum_{i=1}^{m} |x_i - y_i|.$$

Exemple 3 : distance de Minkowski : Soient $m \in \mathbb{N}^*$, $p \geq 1$, $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ et $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$. On appelle distance de Minkowski entre x et y la distance :

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

Exemple 4 : distance de Canberra : Soient $m \in \mathbb{N}^*$, $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ et $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$. On appelle distance de Canberra entre x et y la distance :

$$d(x, y) = \sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i + y_i|}.$$

Exemple 5 : distance maximum : Soient $m \in \mathbb{N}^*$, $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$ et $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$. On appelle distance maximum entre x et y la distance :

$$d(x, y) = \sup_{i \in \{1, \ldots, m\}} |x_i - y_i|.$$

Quelques commandes R : Quelques commandes R associées à ces distances sont :

```
x = c(1, 16, 2, 9, 10, 16, 1)
y = c(14, 9, 9, 12, 4, 3, 13)
z = rbind(x, y)
dist(z, method = "euclidean")
dist(z, method = "manhattan")
dist(z, method = "minkowski", p = 6)
dist(z, method = "maximum")
```

Dorénavant, pour raison de simplicité et de popularité, seule la distance euclidienne sera considérée.

Distance entre 2 individus : Pour tout $(u, v) \in \{1, \ldots, n\}^2$ avec $u \neq v$, la distance euclidienne entre les individus ω_u et ω_v est

$$d(\omega_u, \omega_v) = \sqrt{\sum_{j=1}^{p} (x_{j,u} - x_{j,v})^2}.$$

Tableau des distances : Soit d une distance. On appelle tableau des distances associées aux individus $(\omega_1, \ldots, \omega_n)$ le tableau :

$\mathbf{D} =$

	ω_1	ω_2	$\ldots$	ω_{n-1}	ω_n
ω_1	0	$d_{1,2}$	$\ldots$	$d_{1,n-1}$	$d_{1,n}$
ω_2	$d_{2,1}$	0	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
ω_{n-1}	$d_{n-1,1}$	$\ldots$	$\ldots$	0	$d_{n-1,n}$
ω_n	$d_{n,1}$	$\ldots$	$\ldots$	$d_{n,n-1}$	0

Pour tout $(u,v) \in \{1,\ldots,n\}^2$ avec $u \neq v$, on a posé

$$d_{u,v} = d(\omega_u,\omega_v) = \sqrt{\sum_{j=1}^{p}(x_{j,u} - x_{j,v})^2}.$$

Exemple : On considère la matrice de données $\mathbf{X}$ définie par

$$\mathbf{X} = \begin{pmatrix} 2 & 2 \\ 7.5 & 4 \\ 3 & 3 \\ 0.5 & 5 \\ 6 & 4 \end{pmatrix}$$

En prenant 2 chiffres après la virgule, on a, par exemple,

$$d(\omega_1,\omega_2) = \sqrt{(2-7.5)^2 + (2-4)^2} = 5.85.$$

En procédant de même, on obtient le tableau des distances :

$\mathbf{D} =$

	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0	5.85	1.41	3.35	4.47
ω_2	5.85	0	4.60	7.07	1.50
ω_3	1.41	4.60	0	3.20	3.16
ω_4	3.35	7.07	3.20	0	5.59
ω_5	4.47	1.50	3.16	5.59	0

3.3 Écarts

Écarts : Soit $\mathcal{P}(\Gamma)$ l'ensemble des parties de Γ. On appelle écart toute application $e : \mathcal{P}(\Gamma)^2 \to [0, \infty[$ définie à partir d'une distance et évaluant la ressemblance entre deux groupes d'individus.

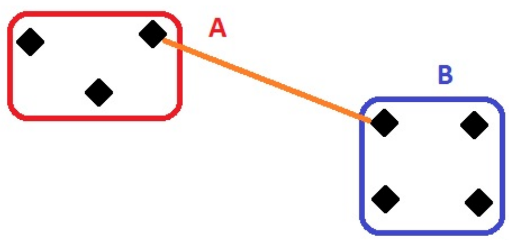
Règle centrale : Plus l'écart entre deux éléments est petit, plus ils se ressemblent.

Écarts usuels : Parmi les écarts usuels entre deux groupes A et B/méthodes usuelles mesurant la ressemblance entre deux groupes A et B, il y a :

- **Écart simple (single linkage)/Méthode du plus proche voisin :**

$$e(A, B) = \min_{(\omega, \omega_*) \in A \times B} d(\omega, \omega_*).$$

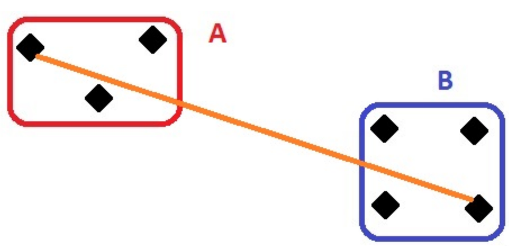
L'écart entre deux groupes A et B est caractérisé par la distance la plus faible entre un point de A et un point de B :



- **Écart complet (complete linkage)/Méthode du voisin le plus éloigné :**

$$e(A, B) = \max_{(\omega, \omega_*) \in A \times B} d(\omega, \omega_*).$$

L'écart entre deux groupes A et B est caractérisé par la distance la plus forte entre un point de A et un point de B :

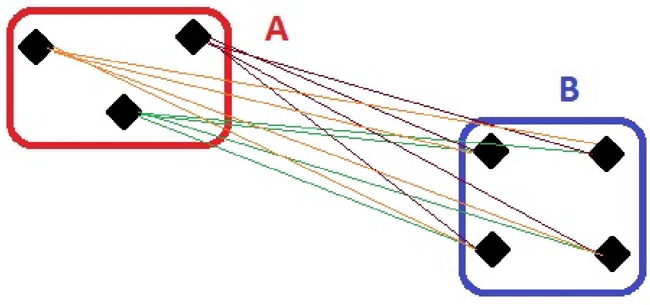


∘ **Écart moyen (average linkage)/Méthode de la distance moyenne :**

$$e(A,B) = \frac{1}{n_A n_B} \sum_{\omega \in A} \sum_{\omega_* \in B} d(\omega, \omega_*),$$

où n_A est le nombre d'individus dans A, et n_B le nombre d'individus dans B.

L'écart entre deux groupes A et B est caractérisé par la distance moyenne entre les points de A et B :



∘ **Écart de Ward :** Soit d la distance euclidienne. La méthode de Ward considère l'écart :

$$e(A,B) = \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B),$$

où g_A est le centre de gravité de A, et g_B celui de B. On rappelle que g_A est le point de coordonnées $(\overline{x}_{1,A}, \ldots, \overline{x}_{p,A})$, où, pour tout $j \in \{1, \ldots, p\}$, $\overline{x}_{j,A}$ désigne la moyenne des valeurs observées du caractère X_j sur les n_A individus du groupe A. De même pour g_B.

Cette méthode prend en compte à la fois la dispersion à l'intérieur d'un groupe et la dispersion entre les groupes. Elle est utilisée par défaut dans la plupart des programmes informatiques. Elle fera l'objet d'un chapitre à venir.

Tableau des écarts : Soit e un écart défini par une des méthodes précédentes. On appelle tableau des écarts associé aux groupes d'individus $(A_1, \ldots, A_n)$ le tableau :

$\mathbf{E} =$

	A_1	A_2	$\ldots$	A_{n-1}	A_n
A_1	0	$e_{1,2}$	$\ldots$	$e_{1,n-1}$	$e_{1,n}$
A_2	$e_{2,1}$	0	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
A_{n-1}	$e_{n-1,1}$	$\ldots$	$\ldots$	0	$e_{n-1,n}$
A_n	$e_{n,1}$	$\ldots$	$\ldots$	$e_{n,n-1}$	0

où, pour tout $(u, v) \in \{1, \ldots, n\}^2$ avec $u \neq v$,

$$e_{u,v} = e(A_u, A_v).$$

Exemple : On considère la matrice de données $\mathbf{X}$ dans $\mathbb{R}^2$ définie par

$$\mathbf{X} = \begin{pmatrix} 2 & 2 \\ 7.5 & 4 \\ 3 & 3 \\ 0.5 & 5 \\ 6 & 4 \end{pmatrix}$$

On considère la méthode du voisin le plus éloigné munie de la distance euclidienne.

Le tableau des écarts associé à $(\{\omega_1\}, \ldots, \{\omega_5\})$ est en fait le tableau des distances :

$\mathbf{E} =$

	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0	5.85	1.41	3.35	4.47
ω_2	5.85	0	4.60	7.07	1.50
ω_3	1.41	4.60	0	3.20	3.16
ω_4	3.35	7.07	3.20	0	5.59
ω_5	4.47	1.50	3.16	5.59	0

Soit A le couple d'individus : $A = \{\omega_1, \omega_3\}$. Par la même méthode, on obtient

$$e(\omega_2, A) = \max(e(\omega_2, \omega_1), e(\omega_2, \omega_3)) = \max(5.85, 4.60) = 5.85,$$

$$e(\omega_4, A) = \max(e(\omega_4, \omega_1), e(\omega_4, \omega_3)) = \max(3.35, 3.20) = 3.35$$

et

$$e(\omega_5, A) = \max(e(\omega_5, \omega_1), e(\omega_5, \omega_3)) = \max(4.47, 3.16) = 4.47.$$

Le tableau des écarts associé à $(\{\omega_2\}, \{\omega_4\}, \{\omega_5\}, A)$ est

$\mathbf{E} =$

	ω_2	ω_4	ω_5	A
ω_2	0	7.07	1.50	5.85
ω_4	7.07	0	5.59	3.35
ω_5	1.50	5.59	0	4.47
A	5.85	3.35	4.47	0

4 Algorithme de classification ascendante hiérarchique (CAH)

4.1 Introduction

CAH : L'idée de l'algorithme de classification ascendante hiérarchique (CAH) est de créer, à chaque étape, une partition de $\Gamma = \{\omega_1, \ldots, \omega_n\}$ en regroupant les deux éléments les plus proches. Le terme "élément" désigne aussi bien un individu qu'un groupe d'individus.

Objectif : On veut

- mettre en relief les liens hiérarchiques entre les individus ou groupe d'individus,
- détecter les groupes d'individus qui se démarquent le plus.

4.2 Description de l'algorithme

Algorithme CAH : L'algorithme de CAH est décrit ci-dessous :

- On choisit un écart. On construit le tableau des écarts pour la partition initiale des n individus de Γ :

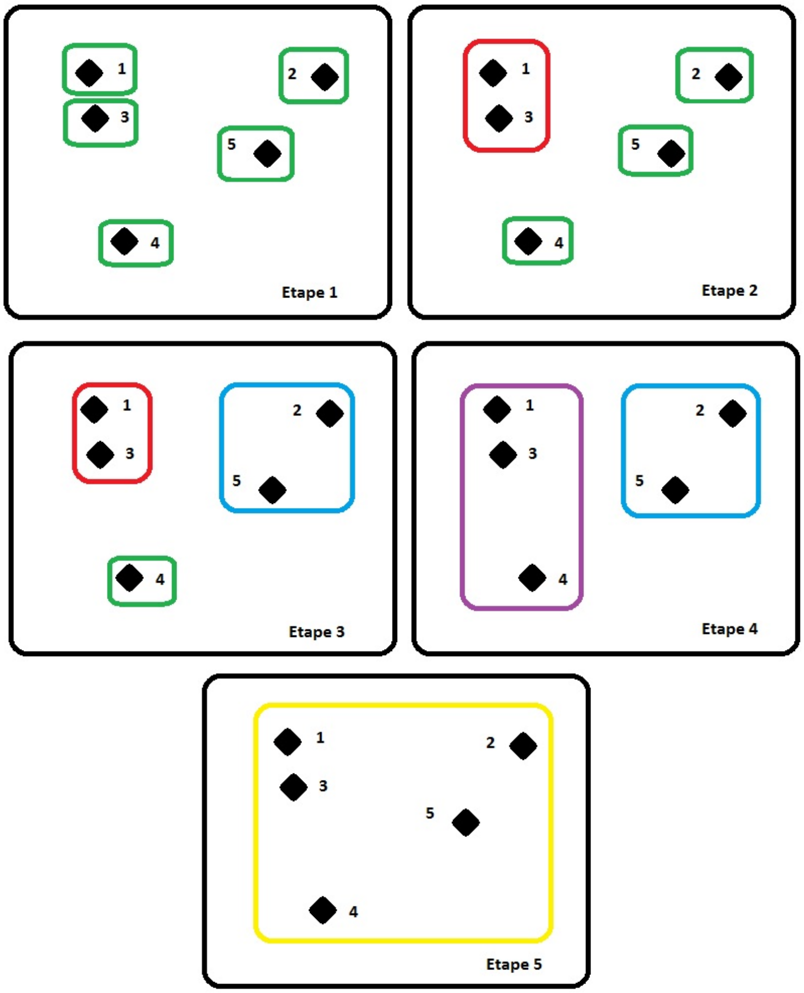
$$\mathcal{P}_0 = (\{\omega_1\}, \ldots, \{\omega_n\}).$$

  Chaque individu constitue un élément.

- On parcours le tableau des écarts pour identifier le couple d'individus ayant l'écart le plus petit. Le regroupement de ces deux individus forme un groupe A. On a donc une partition de Γ de $n-1$ éléments : A et les $n-2$ individus restants.
- On calcule le tableau des écarts entre les $n-1$ éléments obtenus à l'étape précédente et on regroupe les deux éléments ayant l'écart le plus petit (cela peut être deux des $n-2$ individus, ou un individu des $n-2$ individus restants avec A). On a donc une partition de Γ de $n-2$ éléments.
- On itère la procédure précédente jusqu'à ce qu'il ne reste que deux éléments.

- On regroupe les deux éléments restants. Il ne reste alors qu'un seul élément contenant tous les individus de Γ.

Exemple graphique : Ci-dessous, un exemple graphique des étapes de l'algorithme CAH :



Quelques commandes R : hclust et agnes : On peut aussi utiliser la commande hclust (pour Hierarchical CLUSTering) :

```
x = c(2.4, 7.1, 3.8, 1.2, 6.5, 2.1, 4.3, 3, 5.1, 4.3)
m = matrix(x, ncol = 2, nrow = 5)
d = dist(m, method = "euclidean")
cah = hclust(d, "complete")
cah$merge
```

Si on a directement affaire au tableau des distances, la commande est `as.dist` :

```
M = matrix(c(0, 23, 15, 22, 30, 26, 20, 23, 0, 26, 25, 16, 25, 33, 15, 26,
0, 28, 37, 28, 20, 22, 25, 28, 0, 22, 7, 28, 30, 16, 37, 22, 0, 20, 22, 26,
25, 28, 7, 20, 0, 18, 20, 33, 20, 28, 22, 18, 0), byrow = T, ncol = 7)
rownames(M) = c("A","B","C","D","E","F","G")
colnames(M) = c("A","B","C","D","E","F","G")
d = as.dist(M)
cah = hclust(d, "single")
cah$merge
```

Alternativement à `hclust`, on peut utiliser la commande `agnes` (pour AGglomerative NESting) qui offre plus de possibilités :

```
x = c(1, 16, 2, 9, 10, 16, 1, 17, 15, 2, 1, 37, 0, 14, 9, 9, 12, 4, 3, 13)
m = matrix(x, ncol = 5, nrow = 4)
library(cluster)
ag = agnes(m, method = "average")
ag$merge
```

4.3 Dendrogramme

Dendrogramme : Les partitions de Γ faites à chaque étape de l'algorithme de la CAH peuvent se visualiser via un arbre appelé dendrogramme. Sur un axe apparait les individus à regrouper et sur l'autre axe sont indiqués les écarts correspondants aux différents niveaux de regroupement. Cela se fait graphiquement par le biais de branches et de nœuds.

Une partition naturelle se fait en coupant l'arbre au niveau du plus grand saut de nœuds.

Exemple : On considère la matrice de données $\mathbf{X}$ dans $\mathbb{R}^2$ définie par

$$\mathbf{X} = \begin{pmatrix} 2 & 2 \\ 7.5 & 4 \\ 3 & 3 \\ 0.5 & 5 \\ 6 & 4 \end{pmatrix}$$

On va regrouper les individus avec l'algorithme CAH et la méthode du voisin le plus éloigné munie de la distance euclidienne.

∘ Le tableau des écarts associé à $\mathcal{P}_0 = (\{\omega_1\}, \ldots, \{\omega_5\})$ est

	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0	5.85	**1.41**	3.35	4.47
ω_2	5.85	0	4.60	7.07	1.50
ω_3	**1.41**	4.60	0	3.20	3,16
ω_4	3.35	7.07	3.20	0	5.59
ω_5	4.47	1.50	3.16	5.59	0

Les éléments (individus) ω_1 et ω_3 ont l'écart le plus petit : ce sont les éléments les plus proches. On les rassemble pour former le groupe : $A = \{\omega_1, \omega_3\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_1 = (\{\omega_2\}, \{\omega_4\}, \{\omega_5\}, A).$$

∘ Le tableau des écarts associé à $\mathcal{P}_1$ est

	ω_2	ω_4	ω_5	A
ω_2	0	7.07	**1.50**	5.85
ω_4	7.07	0	5.59	3.35
ω_5	**1.50**	5.59	0	4.47
A	5.85	3.35	4,47	0

On a

$$e(\omega_2, A) = \max(e(\omega_2, \omega_1), e(\omega_2, \omega_3)) = \max(5.85, 4.60) = 5.85,$$

$$e(\omega_4, A) = \max(e(\omega_4, \omega_1), e(\omega_4, \omega_3)) = \max(3.35, 3.20) = 3.35$$

et

$$e(\omega_5, A) = \max(e(\omega_5, \omega_1), e(\omega_5, \omega_3)) = \max(4.47, 3.16) = 4.47.$$

Les éléments (individus) ω_2 et ω_5 sont les plus proches. On les rassemble pour former le groupe : $B = \{\omega_2, \omega_5\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_2 = (\{\omega_4\}, A, B).$$

- Le tableau des écarts associé à $\mathcal{P}_2$ est

	ω_4	A	B
ω_4	0	**3.35**	7.07
A	**3.35**	0	5.85
B	7.07	5.85	0

On a

$$e(B, \omega_4) = \max(e(\omega_2, \omega_4), e(\omega_5, \omega_4)) = \max(7.07, 5.59) = 7.07$$

et

$$e(B, A) = \max(e(\omega_2, A), e(\omega_5, A)) = \max(5.85, 4.47) = 5.85.$$

Les éléments ω_4 et A sont les plus proches. On les rassemble pour former le groupe : $C = \{\omega_4, A\} = \{\omega_1, \omega_3, \omega_4\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_3 = (B, C).$$

- Le tableau des écarts associé à $\mathcal{P}_3$ est

	B	C
B	0	**7.07**
C	**7.07**	0

On a

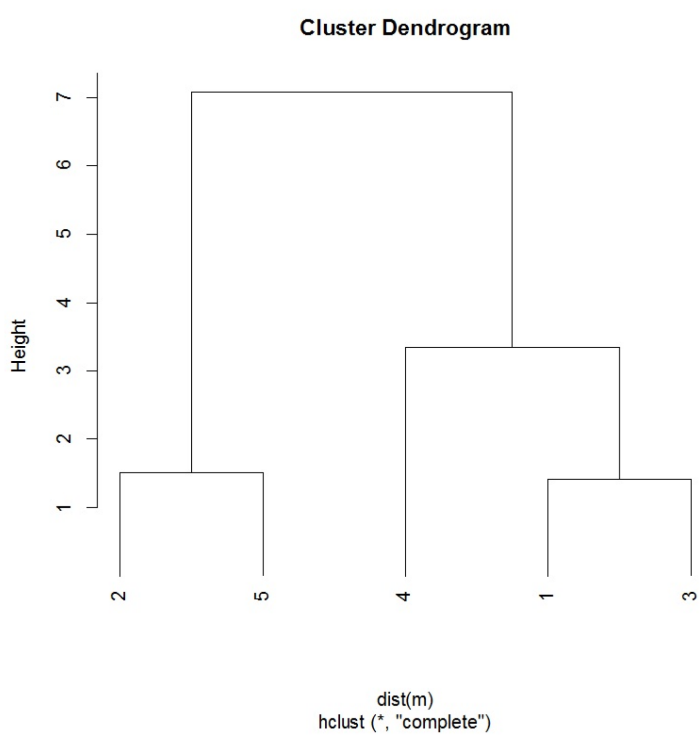
$$e(C,B) = \max(e(\omega_4, B), e(A,B)) = \max(7.07, 5.85) = 7.07.$$

Il ne reste plus que 2 éléments, B et C ; on les regroupe. On obtient la partition $\mathcal{P}_4 = \{\omega_1, \ldots, \omega_5\} = \Gamma$. Cela termine l'algorithme de CAH.

Au final,

- les éléments $\{\omega_1\}$ et $\{\omega_3\}$ ont été regroupés avec un écart de 1.41,
- les éléments $\{\omega_2\}$ et $\{\omega_5\}$ ont été regroupés avec un écart de 1.50,
- les éléments $A = \{\omega_1, \omega_3\}$ et $\{\omega_4\}$ ont été regroupés avec un écart de 3.35,
- les éléments $C = \{\omega_4, A\}$ et $B = \{\omega_2, \omega_5\}$ ont été regroupés avec un écart de 7.07.

On peut donc construire le dendrogramme associé :



Comme le plus grand saut se situe entre les éléments B et C (on a $7.07 - 3.35 = 3.72$), on propose les deux groupes : B et C.

4.4 Quelques commandes R

Avec la commande hclust :

◦ D'abord, on met les données dans une matrice et on trace le nuage de points :

```
x = c(2, 7.5, 3, 0.5, 6, 2, 4, 3, 5, 4)

m = matrix(x, ncol = 2, nrow = 5)

plot(m)
```

◦ On calcule les distances euclidiennes :

```
dist(m)
```

◦ On met en œvre l'algorithme CAH avec la méthode du voisin le plus éloigné (complete linkage) :

```
hc = hclust(dist(m), "complete")
```

On affiche les regroupements :

```
hc$merge
```

Cela renvoie :

```
      [,1]  [,2]
[1,]   -1    -3
[2,]   -2    -5
[3,]   -4     1
[4,]    2     3
```

Ainsi, à la première étape, les individus ω_1 et ω_3 ont été regroupés, formant ainsi le groupe 1, à la deuxième étape, ω_2 et ω_5 ont été regroupés, formant ainsi le groupe 2, à la troisième étape ω_4 et le groupe 1, ont été regroupés, formant ainsi le groupe 3, et pour finir, les groupes 2 et 3 ont été regroupés.

∘ On affiche les écarts de regroupements :

```
hc$height
```

Cela renvoie : 1.414214 1.500000 3.354102 7.071068, rejoignant ainsi la conclusion de l'exercice, à savoir :

- les éléments $\{\omega_1\}$ et $\{\omega_3\}$ ont été regroupés avec un écart de 1.41,
- les éléments $\{\omega_2\}$ et $\{\omega_5\}$ ont été regroupés avec un écart de 1.50,
- les éléments $A = \{\omega_1, \omega_3\}$ et $\{\omega_4\}$ ont été regroupés avec un écart de 3.35,
- les éléments $C = \{\omega_4, A\}$ et $B = \{\omega_2, \omega_5\}$ ont été regroupés avec un écart de 7.07.

∘ On trace le dendrogramme :

```
plot(hc, hang = -1)
```

∘ On peut demander à quel groupe chaque individu appartient suivant la hauteur des sauts avec la commande `cutree`. Avec 2 groupes, on a :

```
b = cutree(hc, k = 2)
b
```

∘ Les effectifs dans chaque groupe s'obtiennent en faisant :

```
table(b)
```

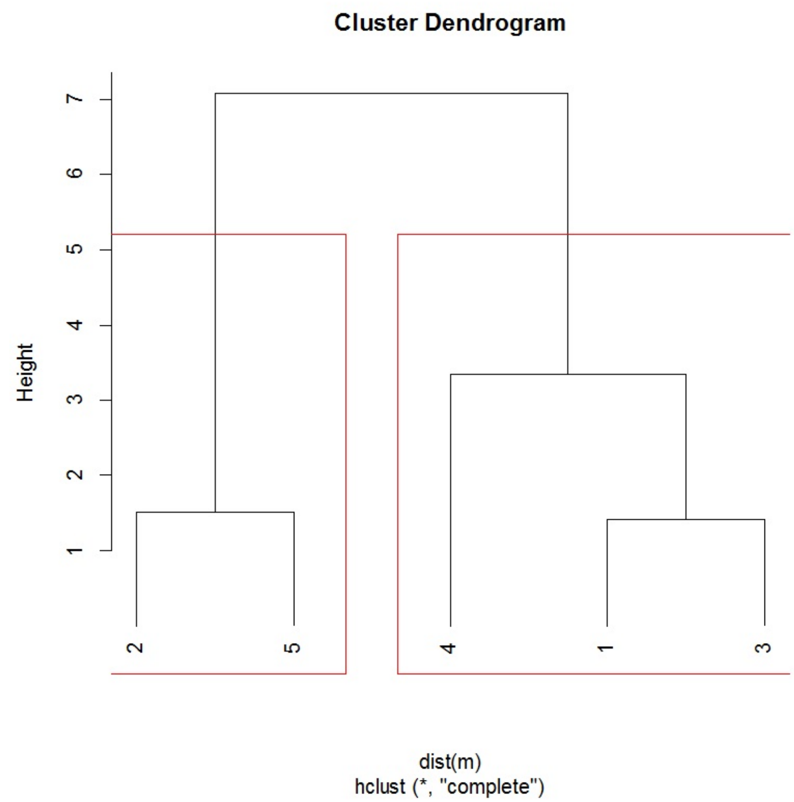
∘ Les indices des individus dans le groupe 1 (par exemple) peuvent s'obtenir en faisant :

```
(1:5)[b == 1]
```

(ou `rownames(m)[b == 1]` si des noms aux lignes de la matrice ou de la `data.frame` existent)

∘ On peut alors afficher clairement les groupes sur le dendrogramme :

```
rect.hclust(hc, 2)
```

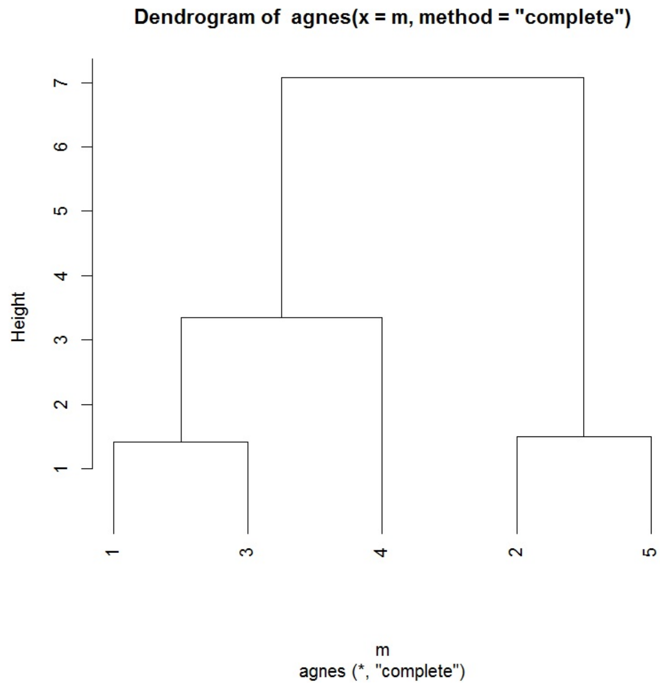


Avec la commande agnes : Avec la commande agnes, on propose :

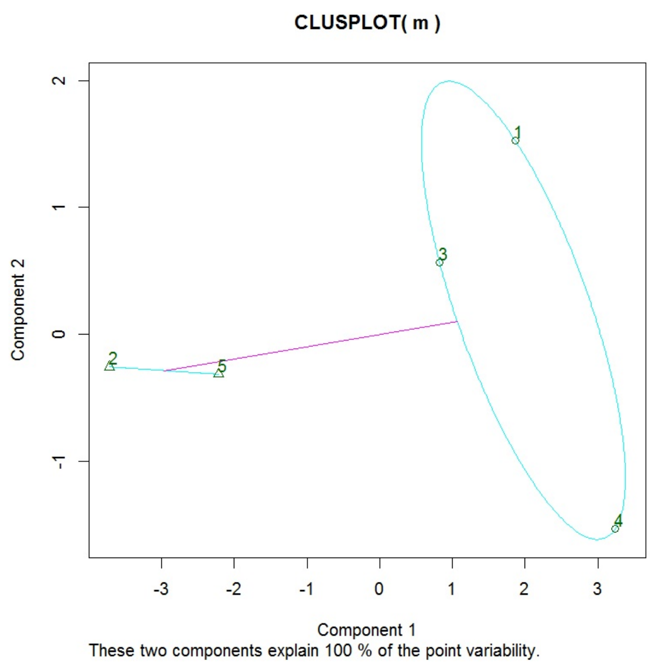
```
x = c(2, 7.5, 3, 0.5, 6, 2, 4, 3, 5, 4)
m = matrix(x, ncol = 2, nrow = 5)
library(cluster)
ag = agnes(m, method = "complete")
ag$merge
ag$height
b = cutree(ag, k = 2)
b
table(b)
(1:5)[b == 1]
(1:5)[b == 2]
pltree(ag, hang = -1)
clusplot(m, cutree(ag, k = 2), labels = 3)
```

On obtient, entre autre,

∘ le dendrogramme :



∘ une représentation graphique des regroupements obtenus (utilisant l'ACP) :



5 CAH et méthode de Ward ; compléments

Centre de gravité : On appelle centre de gravité du nuage de points $\mathcal{N} = \{m_1, \ldots, m_n\}$ le point g de coordonnées $(\overline{x}_1, \ldots, \overline{x}_p)$, où, pour tout $j \in \{1, \ldots, p\}$,

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{j,i}.$$

Pour raison de simplicité, on dira que g est le centre de gravité associé à $\Gamma_n = \{\omega_1, \ldots, \omega_n\}$; on ne se ramènera pas toujours au nuage de point associé.

Inertie totale : On appelle inertie totale de $\mathcal{N}$ autour de son centre de gravité g le réel :

$$\mathcal{I}_{tot} = \frac{1}{n} \sum_{i=1}^{n} d^2(\omega_i, g).$$

On peut remarquer que

$$\mathcal{I}_{tot} = \sum_{j=1}^{p} \sigma_j^2, \qquad \sigma_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{j,i} - \overline{x}_j)^2.$$

L'inertie de $\mathcal{N}$ est une mesure de l'homogénéité de $\mathcal{N}$.

Inertie d'un sous-nuage de points : Soient $h \in \{1, \ldots, n\}$ et $\mathcal{P} = (\mathcal{N}_\ell)_{\ell \in \{1,\ldots,h\}}$ une partition de $\mathcal{N}$. Ainsi, pour tout $\ell \in \{1, \ldots, h\}$, $\mathcal{N}_\ell$ est un sous-nuage de points de $\mathcal{N}$. On note

- n_ℓ le nombre d'individus représentés par $\mathcal{N}_\ell$,
- g_ℓ le centre de gravité de $\mathcal{N}_\ell$, donc le point de coordonnées $(\overline{x}_{1,\ell}, \ldots, \overline{x}_{p,\ell})$, où, pour tout $j \in \{1, \ldots, p\}$, $\overline{x}_{j,\ell}$ désigne la moyenne des valeurs observées du caractère X_j sur les n_ℓ individus du sous-nuage $\mathcal{N}_\ell$.

- Inertie totale : On appelle inertie totale de $\mathcal{N}_\ell$ autour de son centre de gravité g_ℓ le réel :

$$\mathcal{I}(\mathcal{N}_\ell) = \frac{1}{n_\ell} \sum_{i \in \mathcal{N}_\ell} d^2(\omega_i, g_\ell).$$

◦ Inertie intra-classes : On appelle inertie intra-classes le réel :

$$\mathcal{I}_{intra}(\mathcal{P}) = \sum_{\ell=1}^{h} \frac{n_\ell}{n} \mathcal{I}(\mathcal{N}_\ell) \quad \left(= \frac{1}{n} \sum_{j=1}^{p} \sum_{\ell=1}^{h} \sum_{i \in \mathcal{N}_\ell} (x_{j,i} - \overline{x}_{j,\ell})^2 \right).$$

L'inertie intra-classes mesure l'homogénéité de l'ensemble des sous-nuages de la partition.

◦ Inertie inter-classes : On appelle inertie inter-classes le réel :

$$\mathcal{I}_{inter}(\mathcal{P}) = \sum_{\ell=1}^{h} \frac{n_\ell}{n} d^2(g_\ell, g) \quad \left(= \frac{1}{n} \sum_{j=1}^{p} \sum_{\ell=1}^{h} \sum_{i \in \mathcal{N}_\ell} (\overline{x}_{j,\ell} - \overline{x}_j)^2 \right).$$

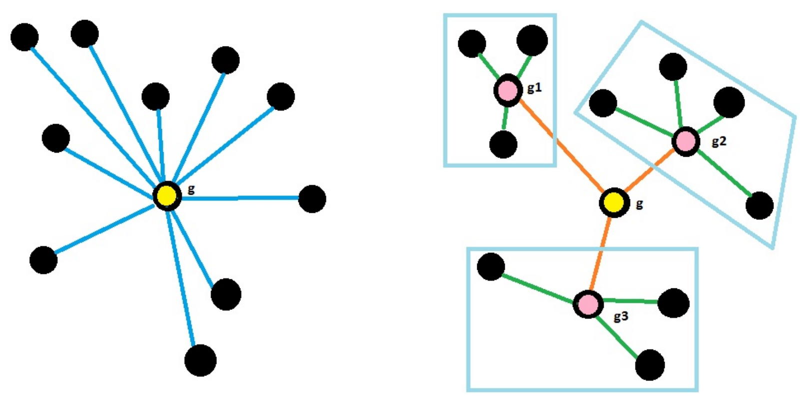
L'inertie inter-classes mesure la séparation entre les sous-nuages de la partition.

Décomposition de Huygens : Pour toute partition $\mathcal{P}$ de $\mathcal{N}$, on a

$$\mathcal{I}_{tot} = \mathcal{I}_{intra}(\mathcal{P}) + \mathcal{I}_{inter}(\mathcal{P}).$$

On constate que minimiser l'inertie intra-classes est équivalent à maximiser l'inertie inter-classes.

Cette décomposition est illustrée par les schémas ci-dessous :



Le point g est le centre de gravité du nuage de points, g_1 est celui du sous-nuage de points à gauche, g_2 est celui du sous-nuage de points à droite et g_3 est celui du sous-nuage de points en bas. Les traits de couleurs représentent les distances entre les points et les centres de gravité.

Alors la somme des distances des traits bleus au carré est égale à la somme des distances des traits verts au carré plus la somme des traits orange au carré.

Sur l'écart de Ward : L'utilisation de l'algorithme de CAH avec la méthode de Ward est justifiée par le résultat suivant :

Soient $\Gamma_n = \{\omega_1, \ldots, \omega_n\}$ n individus et g le centre de gravité associé. Soient A et B deux groupes d'individus

- d'effectifs respectifs n_A et n_B,
- de centres de gravité associés respectifs g_A et g_B.

Le regroupement de A et B, noté $A \cup B$, a pour centre de gravité :

$$g_{A\cup B} = \frac{n_A g_A + n_B g_B}{n_A + n_B}.$$

La perte d'inertie inter-classes lors du regroupement de A et B est égale à $1/n$ multiplié par

$$n_A d^2(g_A, g) + n_B d^2(g_B, g) - (n_A + n_B) d^2(g_{A\cup B}, g) = \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B).$$

On reconnait alors l'écart de Ward qui est donc une mesure de la perte d'inertie inter-classes lors du regroupement de A et B. Ainsi, à chaque étape de l'algorithme de CAH, on veut regrouper des éléments dont le regroupement provoque une perte minimale de l'inertie inter-classes.

Dendrogramme associé à l'écart de Ward : Pour la hauteur des branches, on peut soit prendre les écarts, soit prendre les inerties intra-classe correspondants aux différents niveaux de regroupement.

Commande agnes et écart de Ward : La commande `agnes` avec `method = "ward"` considère un écart défini comme une transformation de l'écart de Ward original :

$$e(A, B) = \sqrt{2 \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B)} = \sqrt{2 \times \text{écart de Ward}}.$$

Cela ne change rien quant à la hiérarchie de la classification.

Exemple : On considère la matrice de données $\mathbf{X}$ dans $\mathbb{R}^2$ définie par

$$\mathbf{X} = \begin{pmatrix} 2 & 2 \\ 7.5 & 4 \\ 3 & 3 \\ 0.5 & 5 \\ 6 & 4 \end{pmatrix}$$

On fait l'algorithme de CAH avec la méthode de Ward.

- Le tableau des écarts associé à $\mathcal{P}_0 = (\{\omega_1\}, \ldots, \{\omega_5\})$ est

	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0	17.12	**1**	5.62	10
ω_2	17.12	0	10.62	25	1,12
ω_3	**1**	10.62	0	5.12	5
ω_4	5.62	25	5.12	0	15.62
ω_5	10	1.12	5	15.62	0

Par exemple, on a

$$e(\omega_1, \omega_2) = \frac{1 \times 1}{1+1} \left((2-7.5)^2 + (2-4)^2 \right) = 17.12.$$

Les éléments (individus) ω_1 et ω_3 ont l'écart le plus petit : ce sont les éléments les plus proches.

On les rassemble pour former le groupe : $A = \{\omega_1, \omega_3\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_1 = (\{\omega_2\}, \{\omega_4\}, \{\omega_5\}, A).$$

L'inertie intra-classes de $\mathcal{P}_1$ est

$$\mathcal{I}_{intra}(\mathcal{P}_1) = \frac{1}{5} \times 1 = 0.2.$$

- Le centre de gravité associé à A est le point g_A de coordonnées :

$$\left(\frac{2+3}{2}, \frac{2+3}{2}\right) = (2.5, 2.5).$$

Le tableau des écarts associé à $\mathcal{P}_1$ est

	ω_2	ω_4	ω_5	A
ω_2	0	25	**1.12**	18.16
ω_4	25	0	15.62	6.83
ω_5	**1.12**	15.62	0	9.66
A	18.16	6.83	9.66	0

Par exemple, on a

$$e(\omega_2, A) = \frac{1 \times 2}{1+2}\left((7.5-2.5)^2 + (4-2.5)^2\right) = 18.16.$$

Les éléments (individus) ω_2 et ω_5 sont les plus proches. On les rassemble pour former le groupe : $B = \{\omega_2, \omega_5\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_2 = (\{\omega_4\}, A, B).$$

L'inertie intra-classes de $\mathcal{P}_2$ est

$$\mathcal{I}_{intra}(\mathcal{P}_2) = 0.2 + \frac{1}{5} \times 1.12 = 0.424.$$

- Le centre de gravité associé à B est le point g_B de coordonnées $(6.75, 4)$.

Le tableau des écarts associé à $\mathcal{P}_2$ est

	ω_4	A	B
ω_4	0	**6.83**	26.7
A	**6.83**	0	20.31
B	26.7	20.31	0

On a, par exemple,

$$e(B, A) = \frac{2 \times 2}{2+2}\left((6.75 - 2.5)^2 + (4 - 2.5)^2\right) = 20.31.$$

Les éléments ω_4 et A sont les plus proches. On les rassemble pour former le groupe : $C = \{\omega_4, A\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_3 = (B, C).$$

L'inertie intra-classes de $\mathcal{P}_3$ est

$$\mathcal{I}_{intra}(\mathcal{P}_3) = 0.424 + \frac{1}{5} \times 6.83 = 1.79.$$

- Le centre de gravité associé à C est le point g_C de coordonnées :

$$\left(\frac{2+3+0.5}{3}, \frac{2+3+5}{3}\right) = (1.83, 3.33).$$

Le tableau des écarts associé à $\mathcal{P}_3$ est

	B	C
B	0	**29.58**
C	**29.58**	0

On a

$$e(B, C) = \frac{2 \times 3}{2+3}\left((6.75 - 1.83)^2 + (4 - 3.33)^2\right) = 29.58.$$

Il ne reste plus que 2 éléments, B et C ; on les regroupe. Cela donne la partition

$\mathcal{P}_4 = \{\omega_1, \ldots, \omega_5\} = \Gamma$.

L'inertie intra-classes de $\mathcal{P}_4$ est

$$\mathcal{I}_{intra}(\mathcal{P}_4) = 1.79 + \frac{1}{5} \times 29.58 = 7.706.$$

Cela termine l'algorithme de CAH.

Un indicateur qu'aucune erreur n'a été commise est l'égalité : $\mathcal{I}_{intra}(\mathcal{P}_4) = \mathcal{I}_{tot}$. On vérifie alors cela : on a

$$\mathcal{I}(\mathcal{N}) = \sigma_1^2 + \sigma_2^2,$$

avec

$$\sigma_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)^2} = 2.5807, \qquad \sigma_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{2,i} - \overline{x}_2)^2} = 1.0198.$$

Donc

$$\mathcal{I}(\mathcal{N}) = 2.5807^2 + 1.0198^2 = 7.701.$$

On admet alors l'égalité (en prenant en compte les arrondis de décimales).

Au final,

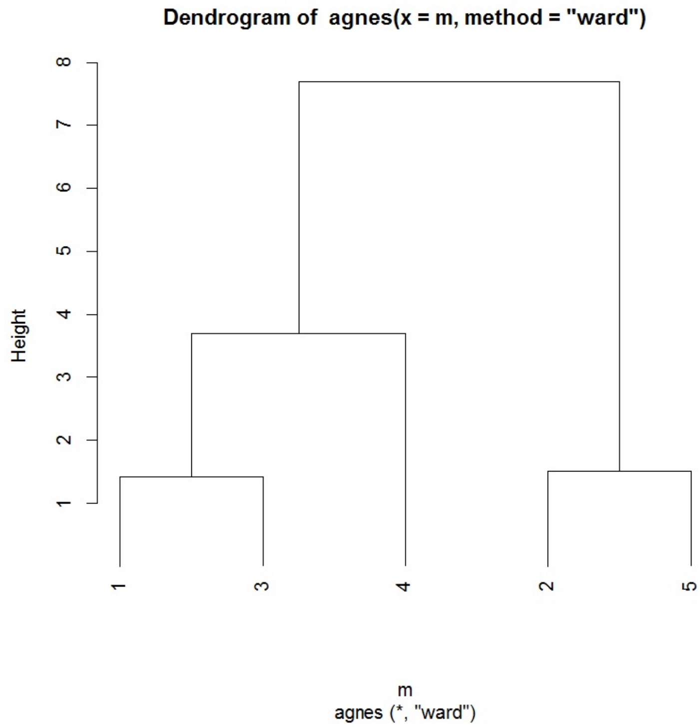
- les éléments $\{\omega_1\}$ et $\{\omega_3\}$ ont été regroupés avec un écart de 1,
- les éléments $\{\omega_2\}$ et $\{\omega_5\}$ ont été regroupés avec un écart de 1.12,
- les éléments $A = \{\omega_1, \omega_3\}$ et $\{\omega_4\}$ ont été regroupés avec un écart de 6.83,
- les éléments $B = \{\omega_2, \omega_5\}$ et $C = \{\omega_4, A\}$ ont été regroupés avec un écart de 29.58.

On peut donc construire le dendrogramme associé.

Quelques commandes R : Avec la commande `agnes`, on propose :

```
x = c(2, 7.5, 3, 0.5, 6, 2, 4, 3, 5, 4)

m = matrix(x, ncol = 2, nrow = 5)

library(cluster)

ag = agnes(m, method = "ward")

pltree(ag, hang = -1)
```

On obtient :



Comme le plus grand écart se situe entre les éléments B et C, on envisage de considérer ces deux groupes.

Avec la commande `agnes`, notons que la formule : $\sqrt{2 \times \text{écart de Ward}}$ a été utilisée pour les hauteurs des branches du dendrogramme : on a $\sqrt{2 \times 1} = 1.41$, $\sqrt{2 \times 1.12} = 1.49$, $\sqrt{2 \times 6.83} = 3.69$ et $\sqrt{2 \times 29.58} = 7.69$.

6 Qualité d'une partition

Coefficient d'agglomération : On appelle coefficient d'agglomération le réel :

$$AC = \frac{1}{n} \sum_{i=1}^{n} \left(1 - \frac{e(\omega_i, A_i)}{e(Q, R)} \right),$$

où

- pour tout $i \in \{1, \ldots, n\}$, A_i désigne le premier élément avec lequel ω_i a été regroupé,
- Q et R désignent les deux éléments rassemblés à l'étape finale de l'algorithme.

On a $AC \in]0, 1[$.

Plus AC est proche de 1, plus les individus sont fortement structurés en plusieurs groupes. Une valeur proche de 0 signifie que les individus appartiennent tous à un même groupe.

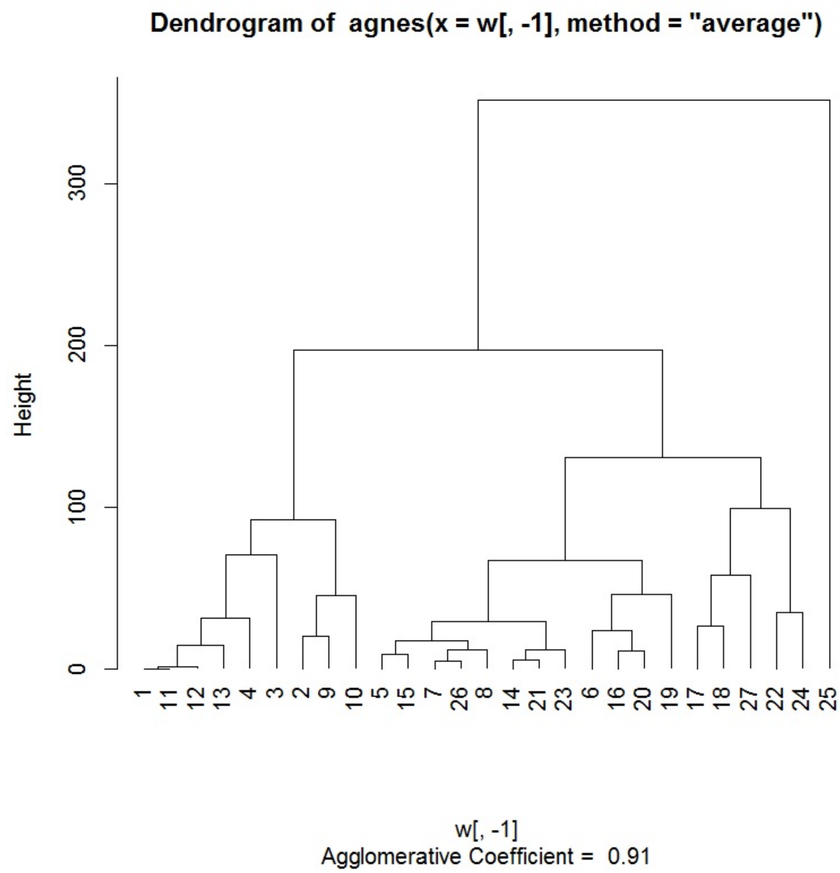
Quelques commandes R : On considère le jeu de données "aliments" dont voici l'entête :

	Individus	X1	X2	X3	X4	X5
1	BB	340	20	28	9	2.60
2	HR	245	21	17	9	2.70
3	BR	420	15	39	7	2.00
4	BS	375	19	32	9	2.50
5	BC	180	22	10	17	3.70
6	CB	115	20	3	8	1.40

Un exemple de CAH et AC avec la commande agnes est :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/aliments.txt",
header = T)
w
attach(w)
library(cluster)
ag = agnes(w[ ,-1], method = "average")
ag$ac
plot(ag, which = 2, hang = -1)
```

Cela renvoie le coefficient d'agglomération $AC = 0.9054413$ et le dendrogramme :



On constate alors une bonne structure de groupes, confirmée par le coefficient d'agglomération proche de 1.

Indice de silhouette : Pour tout $i \in \{1, \ldots, n\}$, on appelle indice de silhouette associé à l'individu ω_i le réel :

$$S(i) = \frac{b_i - a_i}{\max(a_i, b_i)},$$

où

- a_i est la moyenne des distances entre ω_i et les individus de son groupe,
- b_i est la moyenne des distances entre ω_i et les individus du groupe le plus proche de celui auquel il appartient.

On a $S(i) \in]-1, 1[$.

Plus $S(i)$ est proche de 1, plus l'appartenance de ω_i a son groupe est justifiée.

Ainsi, les individus ayant des grands indices de silhouette sont bien regroupés.

Si l'indice de silhouette d'un individu est négatif, l'individu n'est pas dans le bon groupe et pourrait être déplacé dans le groupe le plus proche.

Largeur de silhouette : On appelle largeur de silhouette de la partition le réel :

$$S = \frac{1}{n}\sum_{i=1}^{n} S(i).$$

On a alors l'interprétation suivante :

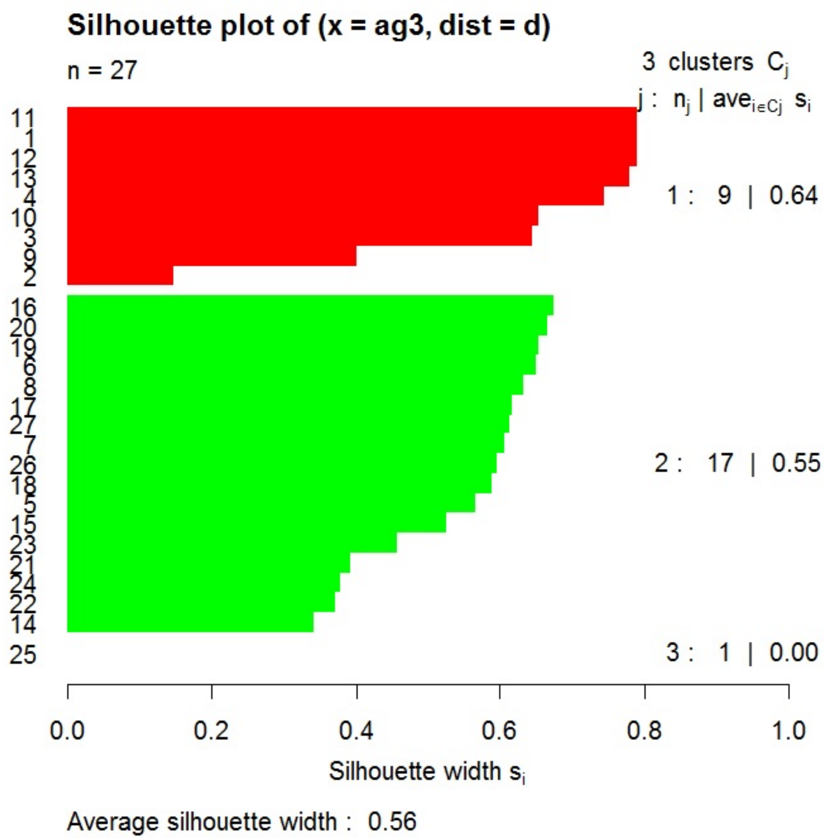
Valeur de S	Nature de la structure
$\in]0.51, 1]$	Forte
$\in]0.31, 0.50]$	Raisonnable
$\in [0, 0.30[$	Faible
$\in [-1, 0[$	Inexistante

On peut également calculer S pour les individus d'un groupe.

Quelques commandes R : Ci-dessous un exemple utilisant les indices de silhouette :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/aliments.txt",
header = T)
w
attach(w)
d = dist(w[ ,-1], method = "euclidean")
library(cluster)
ag = agnes(d, method = "average")
ag3 = cutree(ag, 3)
si = silhouette(ag3, d)
plot(si, col = c("red", "green", "blue"))
```

Cela renvoie le graphique :



Cela renvoie une largeur de silhouette de 0.56, soit une structure forte de la partition, un individu isolé dans le troisième groupe (ω_{25}) et pas d'individu mal regroupé ; aucun indice de silhouette n'est négatif.

Remarques : D'autres indices de qualité existent. Il y a notamment :

- l'indice d'inertie,
- l'indice de connectivité,
- l'indice de Dunn,
- le Cubic Clustering Criterion (CCC).

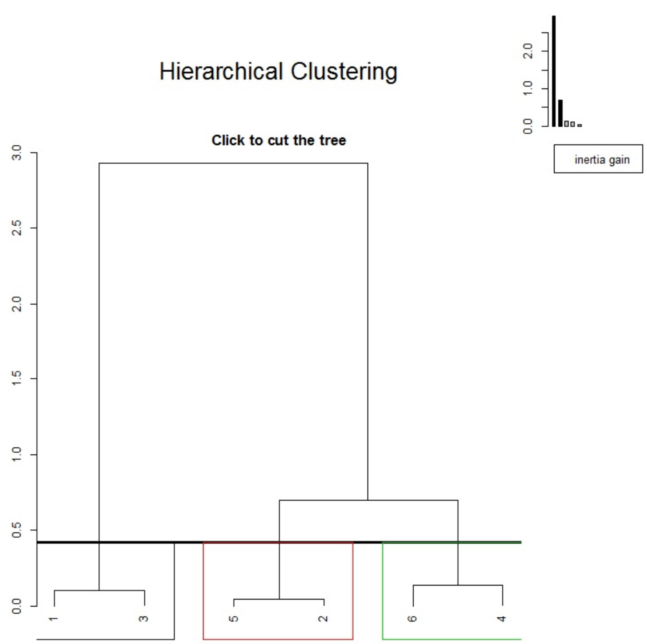
7 ACP et CAH

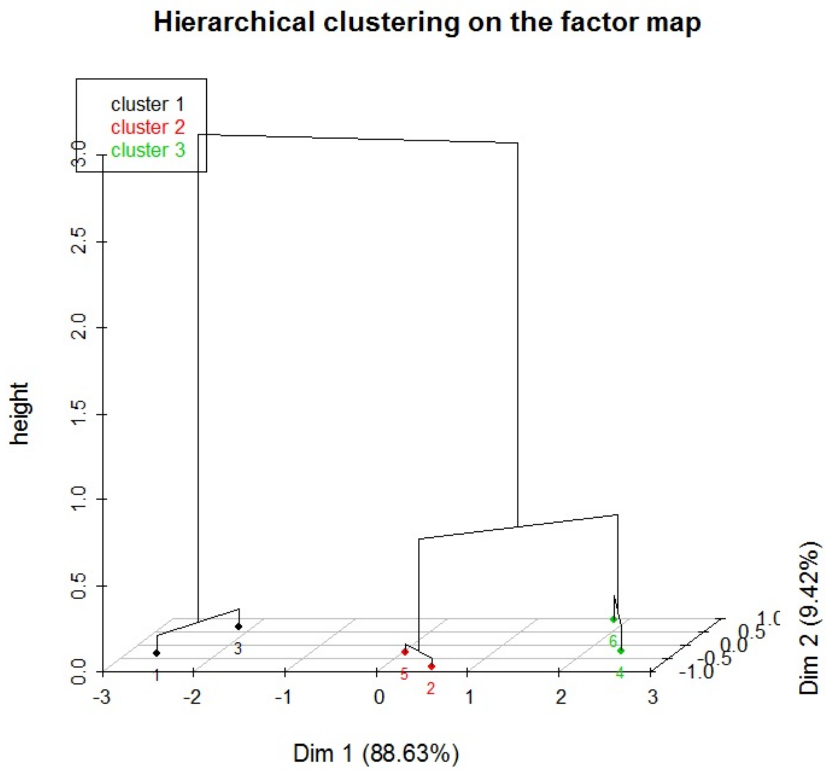
Idée : Lorsque l'on travaille avec plus de 3 variables quantitatives, donc $p \geq 3$, on peut faire une analyse en composantes principales (ACP) et considérer les coordonnées des individus sur le plan principal.

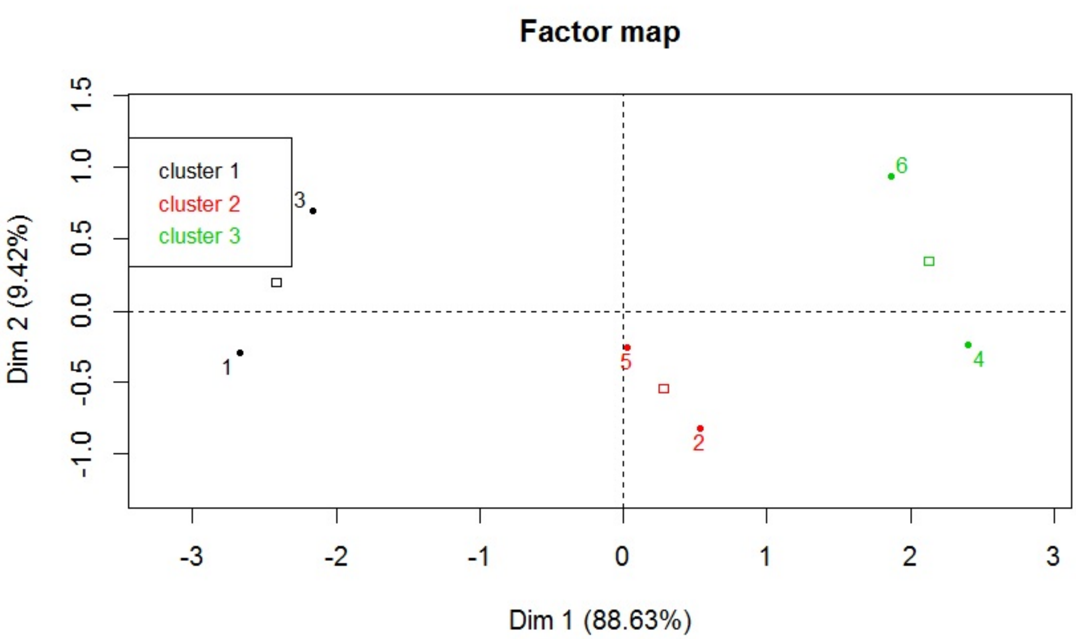
Quelques commandes R : Un exemple de commandes utilisant le package FactoMineR et l'écart de Ward est présenté ci-dessous :

```
library(FactoMineR)
x = c(19, 7, 20, 1, 10, 2, 17, 8, 19, 6, 11, 12, 2, 12, 9, 18, 12, 18, 8,
12, 9, 17, 12, 18)
m = matrix(x, ncol = 4, nrow = 6)
acp = PCA(m, ncp = 2, graph = F)
res = HCPC(acp)
```

On décide de faire 3 groupes (la coupure se fait interactivement sur le dendrogramme affiché). On obtient les graphiques :




Hierarchical clustering on the factor map
cluster 1
cluster 2
cluster 3
height
Dim 1 (88.63%)
Dim 2 (9.42%)


Factor map
cluster 1
cluster 2
cluster 3
Dim 1 (88.63%)
Dim 2 (9.42%)

8 Caractérisation des groupes

Parangons : Pour chaque groupe formé, on appelle parangon l'individu dont le point est le plus proche du centre de gravité du groupe. Le profil de cet individu caractérise alors le groupe auquel il appartient.

Caractères dominants dans la classification : Pour connaître les caractères qui jouent un rôle important dans la classification, on peut faire p ANOVA à 1 facteur avec, pour tout $j \in \{1, \ldots, p\}$,

- le facteur G ayant pour modalités les q groupes formés : $G_1, \ldots, G_q$,
- le caractère X_j (variable quantitative).

Pour chacun des p tests d'hypothèses, le test de Fisher renvoie alors une p-valeur évaluant l'influence du facteur sur le caractère considéré. Ainsi, les caractères associés aux p-valeurs les plus petites sont ceux qui importent le plus dans la classification obtenue.

Caractères dominants d'un groupe : On peut déterminer les caractères dominants pour chacun des groupes formés. Pour se faire, pour chacun des caractères, on peut faire un test d'hypothèses reposant sur loi normale. Soient $G_1, \ldots, G_q$ les q groupes formés. Pour tout $g \in \{G_1, \ldots, G_q\}$ et tout $j \in \{1, \ldots, p\}$, on calcule :

- $\overline{x}_{j,g}$: la moyenne des valeurs du caractère X_j pour les individus du groupe g,
- $\overline{x}_j$: la moyenne des valeurs du caractère X_j,
- n_g : le nombre d'individus dans le groupe g,
- s_j : l'écart-type corrigé des valeurs du caractère X_j,
- le $z_{obs,(j,g)}$:

$$z_{obs,(j,g)} = \frac{\overline{x}_{j,g} - \overline{x}_j}{\sqrt{\frac{s_j^2}{n_g}\left(\frac{n-n_g}{n-1}\right)}}.$$

Pour chaque groupe g, plus $|z_{obs,(j,g)}|$ est grand, plus le caractère X_j importe dans la constitution du groupe ; le caractère le plus important est X_{j_*} avec $j_* = \underset{j\in\{1,\ldots,p\}}{\operatorname{argmax}} |z_{obs,(j,g)}|$.

On peut évaluer le degré d'importance de X_j dans g avec la p-valeur :

$$\text{p-valeur}_{(j,g)} = \mathbb{P}(|Z| \geq |z_{obs,(j,g)}|), \qquad Z \sim \mathcal{N}(0,1).$$

Si p-valeur$_{(j,g)}$ > 0.5, on admet que l'importance de X_j dans g n'est pas significative, si p-valeur$_{(j,g)}$ $\in]0.01, 0.05]$, l'importance de X_j dans g est significative ($\star$), si p-valeur$_{(j,g)}$ $\in]0.001, 0.01]$, elle est très significative ($\star\star$) et si p-valeur$_{(j,g)}$ < 0.001, elle est hautement significative ($\star\star\star$).

Quelques commandes R : On considère le jeu de données "zebu" dont voici l'entête :

	vif	carc	qsup	tota	gras	os
1	395	224	35.10	79.10	6.00	14.90
2	410	232	31.90	73.40	9.70	16.40
3	405	233	30.70	76.50	7.50	16.50
4	405	240	30.40	75.30	8.70	16.00
5	390	217	31.90	76.50	7.80	15.70
6	405	243	32.10	77.40	7.10	15.50

On décrit ci-dessous des exemples de commandes R avec FactoMineR :

```
library(FactoMineR)

w = read.table("https://chesneau.users.lmno.cnrs.fr/zebu.txt", header = T)

w

attach(w)

acp = PCA(w, ncp = 5, graph = F)

res = HCPC(acp, consol = F)
```

On décide de faire 2 groupes.

∘ Classification :

```
res$data.clust
```

∘ Parangons (donnés par les premiers noms de chaque liste) :

```
res$desc.ind
```

L'individu ω_{13} est un parangon pour le premier groupe et ω_3 est un parangon pour le deuxième.

∘ Étude des caractères dominants dans la classification et des caractères dominants d'un groupe :

```
res$desc.var
```

On remarque que *gras* (caractère X_3), *tota* (caractère X_4) et *qsup* (caractère X_5) caractérisent le mieux la partition et ce sont les caractères dominants des groupes.

9 Algorithme des centres mobiles (k means)

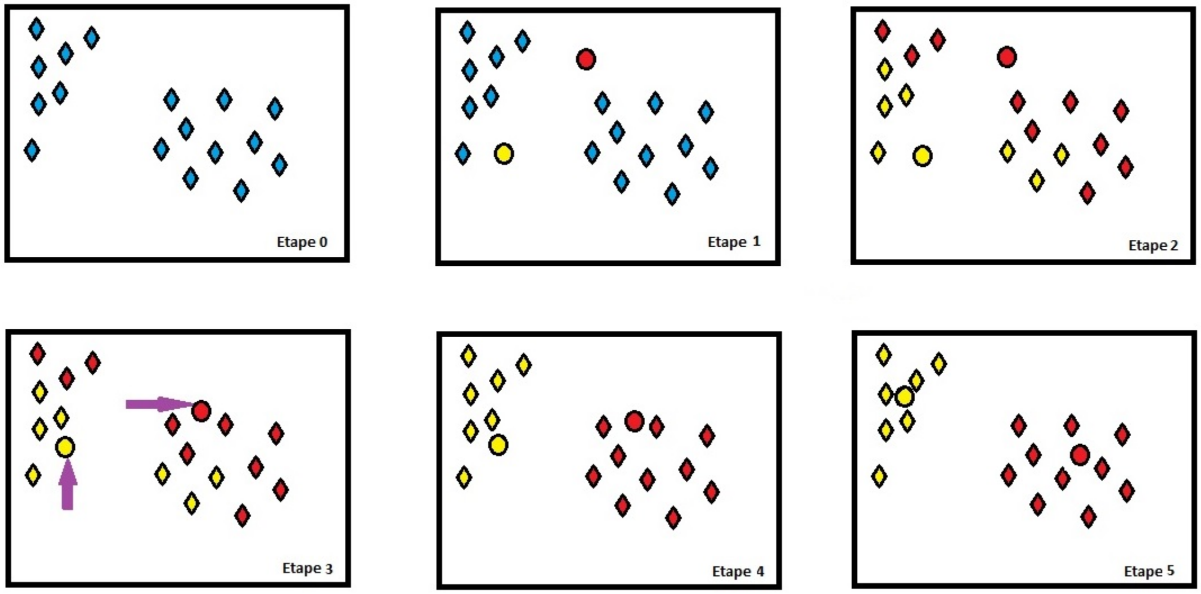
Algorithme des centres mobiles (k means) : L'algorithme des centres mobiles vise à classer une population Γ en q classes. Cela se fait de manière automatique ; il n'y a pas de lien hiérarchique dans les regroupements contrairement à l'algorithme CAH. Il est le mieux adapté aux très grands tableaux de données.

L'algorithme des centres mobiles avec la méthode de Lloyd (la plus standard) est décrit ci-dessous :

- On choisit q points au hasard dans $\mathbb{R}^p$. Ces points sont appelés centres.
- On calcule le tableau de distances entre tous les individus et les q centres.
- On forme alors q groupes de la manière suivante : chaque groupe est constitué d'un centre et des individus les plus proches de ce centre que d'un autre. On obtient une partition $\mathcal{P}_1$ de Γ.
- On calcule le centre de gravité de chacun des q sous-nuages de points formés par les q groupes. Ces q centres de gravité sont nos nouveaux q centres.
- On calcule le tableau de distances entre tous les individus et les nouveaux q centres.
- On forme alors q groupes, chaque groupe étant constitué d'un centre et des individus les plus proches de ce centre que d'un autre. On a une nouvelle partition $\mathcal{P}_2$ de Γ.
- On itère la procédure précédente jusqu'à ce que deux itérations conduisent à la même partition.

Remarque importante : La classification des individus dépend du choix des centres initiaux. Plusieurs méthodes existent pour choisir judicieusement ces centres.

Illustration : Une illustration de l'algorithme des centres mobiles est présentée ci-dessous :



Exemple : Dans une étude industrielle, on a étudié 2 caractères : X_1 et X_2, sur 6 individus $\omega_1, \ldots, \omega_6$.

Les données recueillies sont :

	X_1	X_2
ω_1	-2	2
ω_2	-2	-1
ω_3	0	-1
ω_4	2	2
ω_5	-2	3
ω_6	3	0

1. Dans un premier temps, on fait une classification par l'algorithme des centres mobiles avec, pour centres initiaux, c_1^0 de coordonnées $(-1, -1)$ et c_2^0 de coordonnées (2.3).
2. Dans un deuxième temps, on fait de même avec, pour centres initiaux, c_1^0 de coordonnées $(-1, 2)$ et c_2^0 de coordonnées $(1, 1)$.

1. ∘ On considère les centres initiaux c_1^0 de coordonnées $(-1, -1)$ et c_2^0 de coordonnées $(2, 3)$.

Le tableau des distances entre les individus et ces centres est

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
c_1^0	3.16	1	1	4.24	4.12	4.12
c_2^0	4.12	5.66	4.47	1	4	3.16

Exemple de calcul : $d(\omega_1, c_1^0) = \sqrt{(-2-(-1))^2 + (2-(-1))^2} = 3.16$.

D'où les deux groupes :

$$A = \{\omega_1, \omega_2, \omega_3\}, \qquad B = \{\omega_4, \omega_5, \omega_6\}.$$

- On considère deux nouveaux centres, c_1^1 et c_2^1, lesquels sont les centres de gravité des deux groupes A et B. Donc c_1^1 a pour coordonnées $\left(\frac{-2-2+0}{3}, \frac{2-1-1}{3}\right) = (-1.33, 0)$ et c_2^1 a pour coordonnées $\left(\frac{2-2+3}{3}, \frac{2+3+0}{3}\right) = (1, 1.67)$.

  Le tableau des distances entre les individus et ces centres est

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
c_1^1	2.11	1.20	1.66	3.88	3.07	4.33
c_2^1	3.02	4.02	2.85	1.05	3.28	2.61

  D'où les deux groupes :

$$A = \{\omega_1, \omega_2, \omega_3, \omega_5\}, \qquad B = \{\omega_4, \omega_6\}.$$

- On considère deux nouveaux centres, c_1^2 et c_2^2, lesquels sont les centres de gravité des deux groupes A et B. Donc c_1^2 a pour coordonnées $\left(\frac{-2-2+0-2}{4}, \frac{2-1-1+3}{4}\right) = (-1.5, 0.75)$ et c_2^2 a pour coordonnées $\left(\frac{2+3}{2}, \frac{2+0}{2}\right) = (2.5, 1)$.

Le tableau des distances entre les individus et ces centres est

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
c_1^2	1.35	1.82	2.30	3.72	2.30	4.56
c_1^2	4.61	4.92	3.20	1.12	4.92	1.12

D'où les deux groupes :

$$A = \{\omega_1, \omega_2, \omega_3, \omega_5\}, \qquad B = \{\omega_4, \omega_6\}.$$

On retrouve la même classification que l'étape précédente, on arrête l'algorithme.

2. Considérons maintenant les centres initiaux c_1^0 de coordonnées $(-1, 2)$ et c_2^0 de coordonnées $(1, 1)$.

   - On considère les centres initiaux c_1^0 de coordonnées $(-1, 2)$ et c_2^0 de coordonnées $(1, 1)$.

     Le tableau des distances entre les individus et ces centres est

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
c_1^0	1	3.16	3.16	3	1.41	4.47
c_2^0	3.16	3.60	2.24	1.41	3.60	2.24

     D'où les deux groupes :

     $$A = \{\omega_1, \omega_2, \omega_5\}, \qquad B = \{\omega_3, \omega_4, \omega_6\}.$$

   - On considère deux nouveaux centres, c_1^1 et c_2^1, lesquels sont les centres de gravité des deux groupes A et B. Donc c_1^1 a pour coordonnées $\left(\frac{-2-2-2}{3}, \frac{2-1+3}{3}\right) = (-2, 1.33)$ et c_2^1 a pour coordonnées $\left(\frac{0+2+3}{3}, \frac{-1+2+0}{3}\right) = (1.67, 0.33)$.

Le tableau des distances entre les individus et ces centres est

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
c_1^1	0.67	2.33	3.07	4.06	1.67	5.17
c_2^1	4.03	3.90	2.13	1.70	4.54	1.37

D'où les deux groupes :

$$A = \{\omega_1, \omega_2, \omega_5\}, \qquad B = \{\omega_3, \omega_4, \omega_6\}.$$

On retrouve la même classification que l'étape précédente, on arrête l'algorithme.

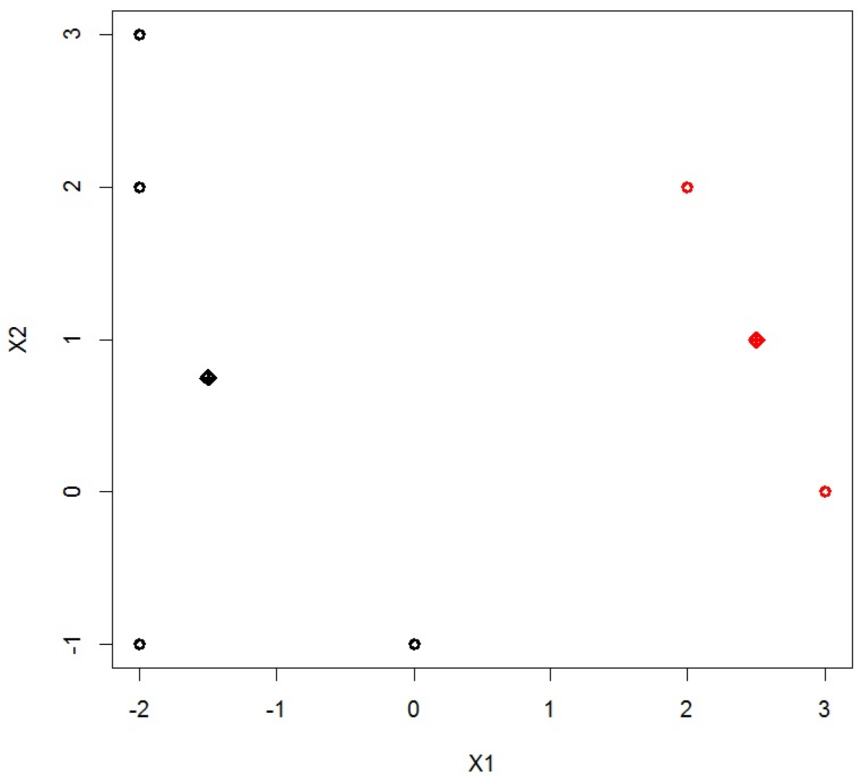
Remarque : On obtient deux classifications différentes suivant les choix des centres initiaux.

Commandes R de l'exemple :

Pour le 1., les commandes R associées sont :

```
x = c(-2, -2, 0, 2, -2, 3, 2, -1, -1, 2, 3, 0)
m = matrix(x, ncol = 2, nrow = 6)
clus = kmeans(m, centers = rbind(c(-1, -1), c(2, 3)), algorithm = "Lloyd")
clus$cluster
clus$centers
plot(m, col = clus$cluster, pch = 1, lwd = 3, xlab = "X1", ylab = "X2")
points(clus$centers, col = 1:2, pch = 9, lwd = 3)
```

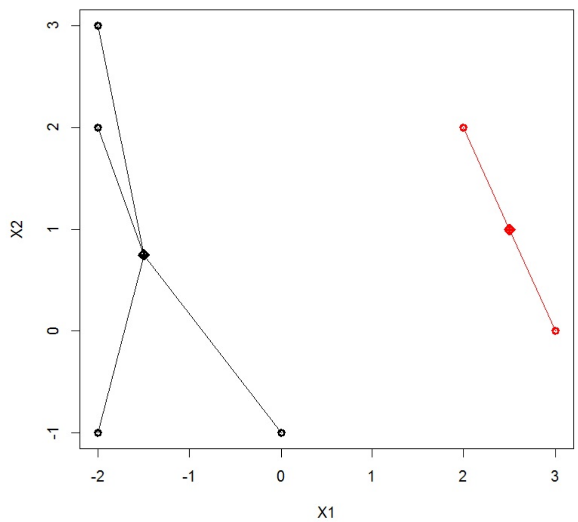
Cela renvoie les groupes d'affectation de chaque individu (`clus$cluster`), les coordonnées des centres de gravité de chaque groupe (`clus$centers`) et le graphique :



On peut rejoindre les individus au centre de gravité dans chaque groupe avec la commande `segments` :

```
segments(m[clus$cluster == 1, ][ ,1], m[clus$cluster == 1, ][ ,2],
clus$centers[1, 1], clus$centers[1, 2])
segments(m[clus$cluster == 2, ][ ,1], m[clus$cluster == 2, ][ ,2],
clus$centers[2, 1], clus$centers[2, 2], col = 2)
```

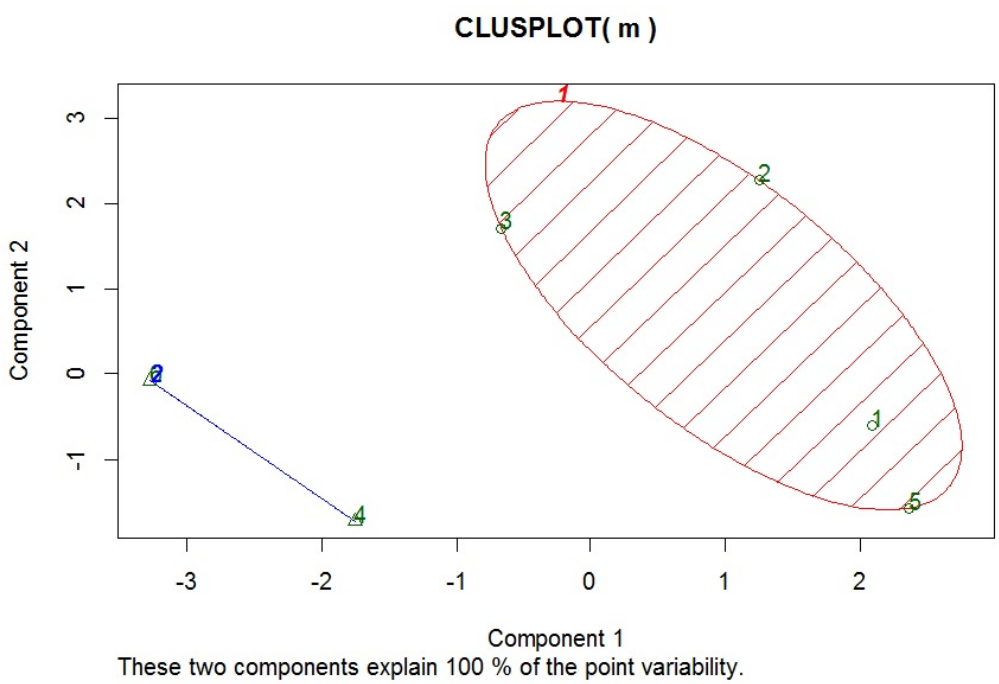
Cela renvoie :



On peut aussi utiliser clusplot pour la visualisation des groupes :

```
library(cluster)

clusplot(m, clus$cluster, color = T, shade = T, labels = 2, lines = 0)
```

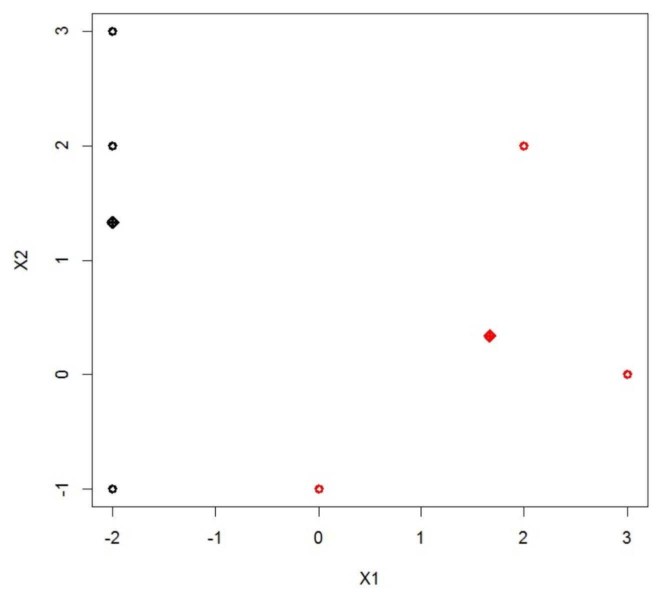
Cela renvoie :



Pour le 2., les commandes R associées sont :

```
x = c(-2, -2, 0, 2, -2, 3, 2, -1, -1, 2, 3, 0)
m = matrix(x, ncol = 2, nrow = 6)
clus = kmeans(m, centers = rbind(c(-1, 2), c(1, 1)), algorithm = "Lloyd")
clus$cluster
clus$centers
plot(m, col = clus$cluster, pch = 1, lwd = 3, xlab = "X1", ylab = "X2")
points(clus$centers, col = 1:2, pch = 9, lwd = 3)
```

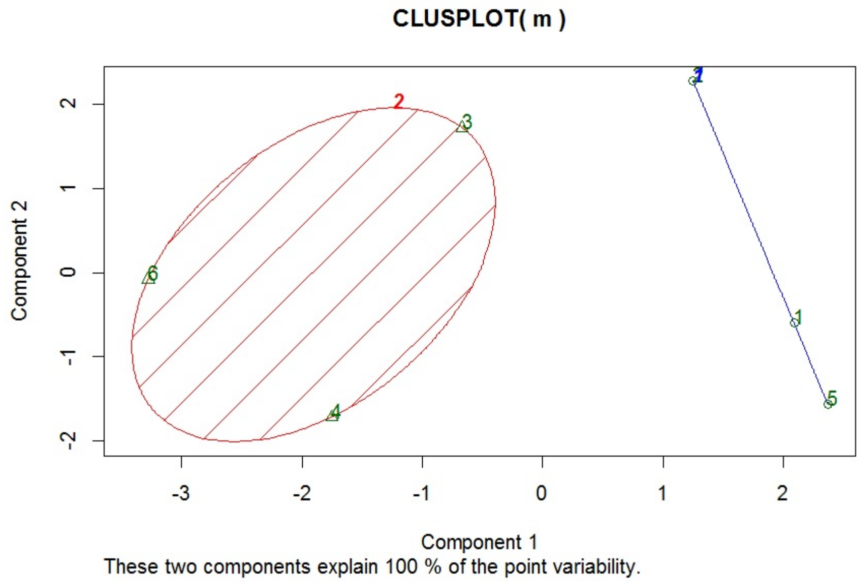
Cela renvoie le graphique :



On peut aussi utiliser clusplot pour la visualisation des groupes :

```
library(cluster)
clusplot(m, clus$cluster, color = T, shade = T, labels = 2, lines = 0)
```

Cela renvoie :



Présentation du jeu de données iris : Une célèbre jeu de données étudié par le statisticien Fisher en 1936 est "les iris de Fisher". Pour 3 variétés d'iris : Setosa, Versicolor, Virginica, et pour 150 iris par variété, on considère 4 caractères quantitatifs :

- X_1 la longueur en cm d'un pétale,
- X_2 la largeur en cm d'un pétale,
- X_3 la longueur en cm d'un sépale,
- X_4 la largeur en cm d'un sépale.

Ce sont les variables explicatives X_1, X_2, X_3 et X_4. La variable à expliquer Y est une variable qualitative dont les modalités sont les espèces d'iris $\{setosa, versicolor, virginica\}$.

Voici l'entête du jeu de données "iris" :

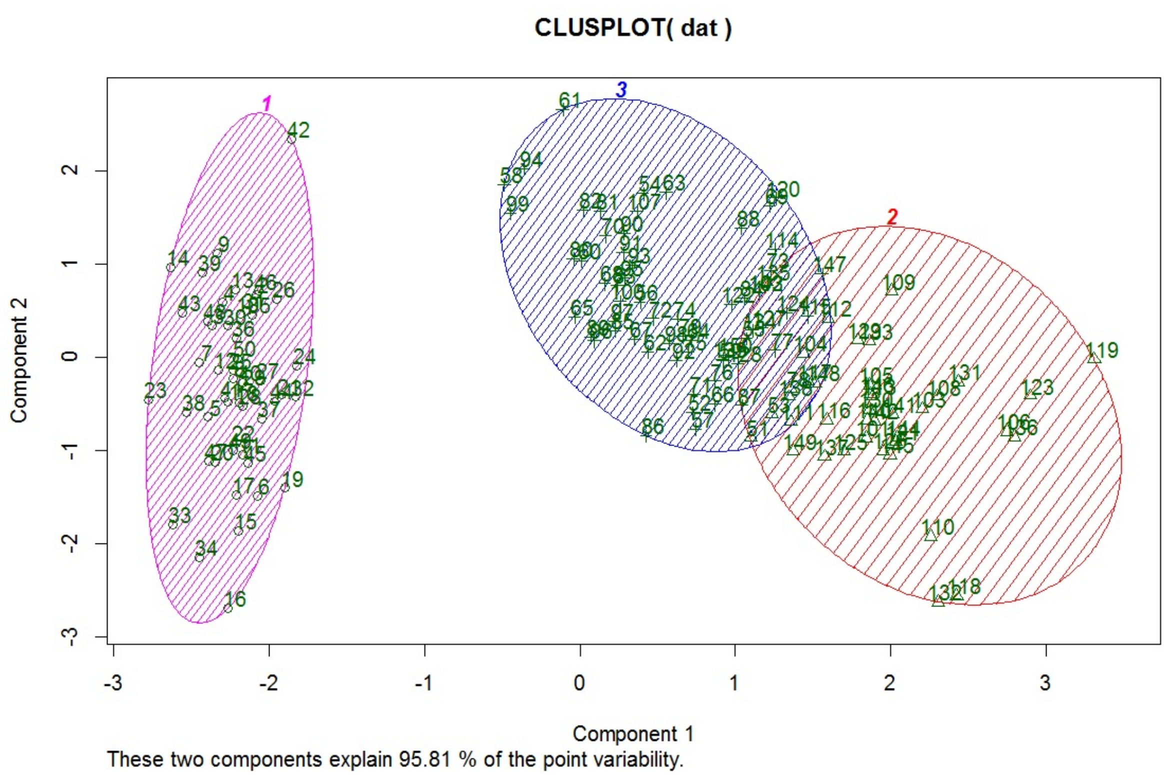
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.10	3.50	1.40	0.20	setosa
2	4.90	3.00	1.40	0.20	setosa
3	4.70	3.20	1.30	0.20	setosa
4	4.60	3.10	1.50	0.20	setosa
5	5.00	3.60	1.40	0.20	setosa
6	5.40	3.90	1.70	0.40	setosa

Quelques commandes R : Un exemple de commandes R utilisant le jeu de données `iris` et l'algorithme des centres mobiles est présenté ci-dessous :

```
dat = iris[ ,1:4]

library(stats)

clus = kmeans(dat, centers = dat[c(15, 135, 65), ], algorithm = "Lloyd")

library(cluster)

clusplot(dat, clus$cluster, color = T, shade = T, labels = 2, lines = 0)
```

Dans cet exemple, on a donc considéré l'algorithme des centres mobiles avec 3 centres initiaux qui sont les individus correspondants aux lignes 15, 135 et 65 du jeu de données `iris`.

Le graphique obtenu est :



Méthodes alternatives : Il existe de nombreuses méthodes autres que celle de Lloyd. Il y a notamment :

- la méthode de Forgy : les centres initiaux sont tirés au hasard parmi ceux associés aux individus de Γ.

```
clus = kmeans(dat, 3, algorithm = "Forgy")
```

∘ la méthode de MacQueen : les centres sont recalculés à chaque réaffectation d'un seul individu.

```
clus = kmeans(dat, 3, algorithm = "MacQueen")
```

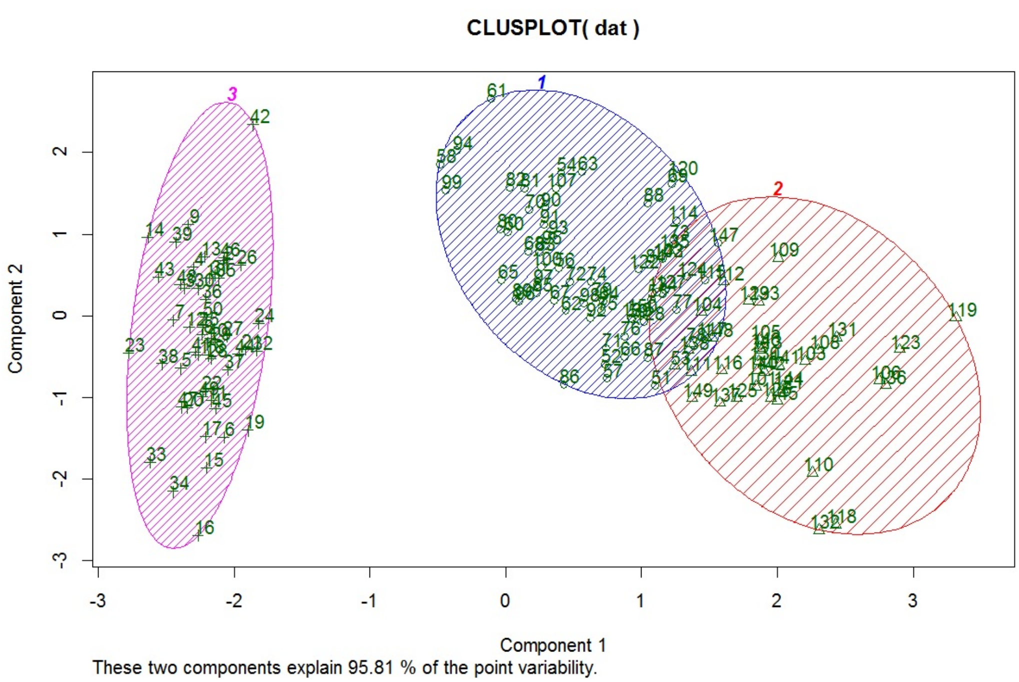
∘ la méthode de Hartigan-Wong : c'est la méthode par défaut de la commande kmeans. Elle est considérée comme la plus robuste de toutes.

```
clus = kmeans(dat, 3)
```

Quelques commandes R : Un exemple de commandes R pour utiliser l'algorithme des centres mobiles avec la méthode de Hartigan-Wong :

```
dat = iris[ ,1:4]
clus = kmeans(dat, 3)
library(cluster)
clusplot(dat, clus$cluster, color = T, shade = T, labels = 2, lines = 0)
```

Le graphique obtenu est :



Le résultat est identique à celui obtenu avec la méthode de Lloyd ; cela est un hasard.

Alternatives à l'algorithme des centres mobiles : Il existe de nombreuses alternatives à l'algorithme des centres mobiles. Il y a notamment :

- la méthode PAM (Partition Around Medoids) : cette méthode a la particularité de marcher aussi avec un tableau des distances et d'être moins sensible que l'algorithme des centres mobiles aux individus atypiques.

```
library(cluster)
clus = pam(dat, 3)
plot(clus)
```

- la méthode CLARA (Clustering LARge Application) :

```
library(cluster)
clus = clara(dat, 3)
plot(clus)
```

- la méthode FANNY :

```
library(cluster)
clus = fanny(dat, 3)
plot(clus)
```

10 Consolidation de l'algorithme de CAH

Idée : On peut consolider/améliorer les regroupements obtenus via l'algorithme de CAH en utilisant l'algorithme des centres mobiles. On prend alors pour centres initiaux les parangons obtenus lors de la CAH. Il est donc possible que des individus changent de groupes.

Quelques commandes R : Un exemple de code R utilisant la librairie FactoMineR et la commande `consol = T` est :

```
library(FactoMineR)
w = read.table("https://chesneau.users.lmno.cnrs.fr/zebu.txt", header = T)
w
attach(w)
acp = PCA(w, ncp = 5, graph = F)
res = HCPC(acp, consol = T)
res$data.clust
```

En fait, à la base, on dispose du jeu de données zebu avec une classification des individus en deux groupes. Celle-ci est visible par :

```
w2 = read.table("https://chesneau.users.lmno.cnrs.fr/zebu-g.txt",
header = T)
w2
```

On constate alors que la classification obtenue avec la CAH consolidée est exacte. Ce n'est pas le cas sans consolidation :

```
res2 = HCPC(acp, consol = F)
res2$data.clust
```

11 Complément : CAH avec des caractères qualitatifs

Indice de similarité : Soit $\Gamma = \{\omega_1, \ldots, \omega_n\}$. On appelle indice de similarité toute application $s : \Gamma^2 \to [0, \infty[$ telle que, pour tous individus ω et ω_* dans Γ, on a

- $s(\omega, \omega_*) = s(\omega_*, \omega)$,
- $s(\omega, \omega_*) \leq s(\omega, \omega)$,
- on a $s(\omega, \omega_*) = s(\omega, \omega)$ si, et seulement si, $\omega = \omega_*$.

Règle centrale : Plus l'indice de similarité entre deux individus est élevé, plus ils se ressemblent.

Cas des caractères quantitatifs : Quand les caractères $X_1, \ldots, X_p$ sont quantitatifs, on peut choisir comme indice de similarité la fonction s telle que

$$s(\omega, \omega_*) = d_{max} - d(\omega, \omega_*),$$

où d désigne une distance et $d_{max} = \max\limits_{\omega \times \omega_* \in \Gamma^2} d(\omega, \omega_*)$.

Tableau disjonctif complet : Dans le cas où les caractères $X_1, \ldots, X_p$ sont qualitatifs, on peut présenter les données sous la forme d'un tableau disjonctif complet (TDC) de dimension $n \times r$, où r est le nombre total de modalités des p caractères considérés. Pour tout $i \in \{1, \ldots, n\}$, la i-ème ligne du tableau est constituée du vecteur $(n_{1,i}, \ldots, n_{k,i}, \ldots n_{r,i})$, avec $n_{k,i} = 1$ si i possède la modalité k, et 0 sinon. Il n'y a donc que des 0 et 1 dans le tableau.

Valeurs intermédiaires : Pour tout $(u, v) \in \{1, \ldots, n\}^2$, on pose

- $a_{u,v}$ le nombre de $(1, 1)$ aux (u, v)-ème lignes du TDC,
- $b_{u,v}$ le nombre de $(1, 0)$ aux (u, v)-ème lignes du TDC,
- $c_{u,v}$ le nombre de $(0, 1)$ aux (u, v)-ème lignes du TDC,
- $d_{u,v}$ le nombre de $(0, 0)$ aux (u, v)-ème lignes du TDC.

Notons que

$$a_{u,v} + b_{u,v} + c_{u,v} + d_{u,v} = r.$$

Indices de similarité usuels : Les indices de similarité les plus utilisés sont les suivants :

- Indice de Russel et Rao :

$$s(\omega_u, \omega_v) = \frac{a_{u,v}}{r}.$$

- Indice de Jaccard :

$$s(\omega_u, \omega_v) = \frac{a_{u,v}}{a_{u,v} + b_{u,v} + c_{u,v}} = \frac{a_{u,v}}{r - d_{u,v}}.$$

- Indice de Dice :

$$s(\omega_u, \omega_v) = \frac{2a_{u,v}}{2a_{u,v} + b_{u,v} + c_{u,v}}.$$

- Indice de d'Anderberg :

$$s(\omega_u, \omega_v) = \frac{a_{u,v}}{a_{u,v} + 2(b_{u,v} + c_{u,v})}.$$

- Indice de Rogers et Tanimoto :

$$s(\omega_u, \omega_v) = \frac{a_{u,v} + d_{u,v}}{a_{u,v} + d_{u,v} + 2(b_{u,v} + c_{u,v})}.$$

- Indice de Pearson :

$$s(\omega_u, \omega_v) = \frac{a_{u,v}d_{u,v} - b_{u,v}c_{u,v}}{\sqrt{(a_{u,v} + b_{u,v})(a_{u,v} + c_{u,v})(d_{u,v} + b_{u,v})(d_{u,v} + c_{u,v})}}.$$

- Indice de Yule :

$$s(\omega_u, \omega_v) = \frac{a_{u,v}d_{u,v} - b_{u,v}c_{u,v}}{a_{u,v}d_{u,v} + b_{u,v}c_{u,v}}.$$

Distances à partir d'un indice de similarité et CAH : À partir d'un indice de similarité s, on définit une application $d_* : \Gamma^2 \to [0, \infty[$ par

$$d_*(\omega_u, \omega_v) = s_{max} - s(\omega_u, \omega_v),$$

où $s_{max} = s(\omega_u, \omega_u) \ \ (= s(\omega_v, \omega_v))$.

Cette application est appelée dissimilarité.

On peut alors faire de la CAH avec cette dissimilarité d_* au lieu de d et l'écart de son choix.

Sur l'indice de Jaccard : Si s est l'indice de Jaccard, alors $s_{max} = 1$ et on peut prendre la dissimilarité :

$$d_*(\omega_u, \omega_v) = 1 - s(\omega_u, \omega_v) = 1 - \frac{a_{u,v}}{r - d_{u,v}}.$$

Quelques commandes R : Ci-dessous, des exemples de commandes R illustrant plusieurs dissimilarités :

```
m = matrix(sample(c(0, 1), 100, replace = T), ncol = 10)
m
library(arules)
d = dissimilarity(m, method = "jaccard")
d
d = dissimilarity(m, method = "pearson")
d
```

Exemple : On interroge 6 individus en leur demandant leur sexe X_1 (F : femme, H : homme), leur type de logement X_2 (R : rural, U : urbain) et leur état civil X_3 (C : célibataire, M : marié, A : autre). On obtient :

	X_1	X_2	X_3
ω_1	H	U	C
ω_2	F	U	C
ω_3	F	R	M
ω_4	F	U	A
ω_5	H	R	M
ω_6	H	R	A

1. En considérant l'indice de Jaccard, calculer $s(\omega_1, \omega_2)$ et $s(\omega_3, \omega_6)$.
2. Est-ce que ω_1 est plus proche de ω_2, que ω_3 de ω_6 ?

Solution :

1. Le TDC associé est

	F	H	R	U	C	M	A
ω_1	0	1	0	1	1	0	0
ω_2	1	0	0	1	1	0	0
ω_3	1	0	1	0	0	1	0
ω_4	1	0	0	1	0	0	1
ω_5	0	1	1	0	0	1	0
ω_6	0	1	1	0	0	0	1

On a $a_{1,2} = 2$, $b_{1,2} = 1$ et $c_{1,2} = 1$ (et $d_{1,2} = 3$). Donc

$$s(\omega_1, \omega_2) = \frac{a_{1,2}}{a_{1,2} + b_{1,2} + c_{1,2}} = \frac{2}{2+1+1} = 0.5.$$

On a $a_{3,6} = 1$, $b_{3,6} = 2$ et $c_{3,6} = 2$ (et $d_{3,6} = 2$). Donc

$$s(\omega_3, \omega_6) = \frac{a_{3,6}}{a_{3,6} + b_{3,6} + c_{3,6}} = \frac{1}{1+2+2} = 0.2.$$

2. Comme

$$s(\omega_1, \omega_2) > s(\omega_3, \omega_6),$$

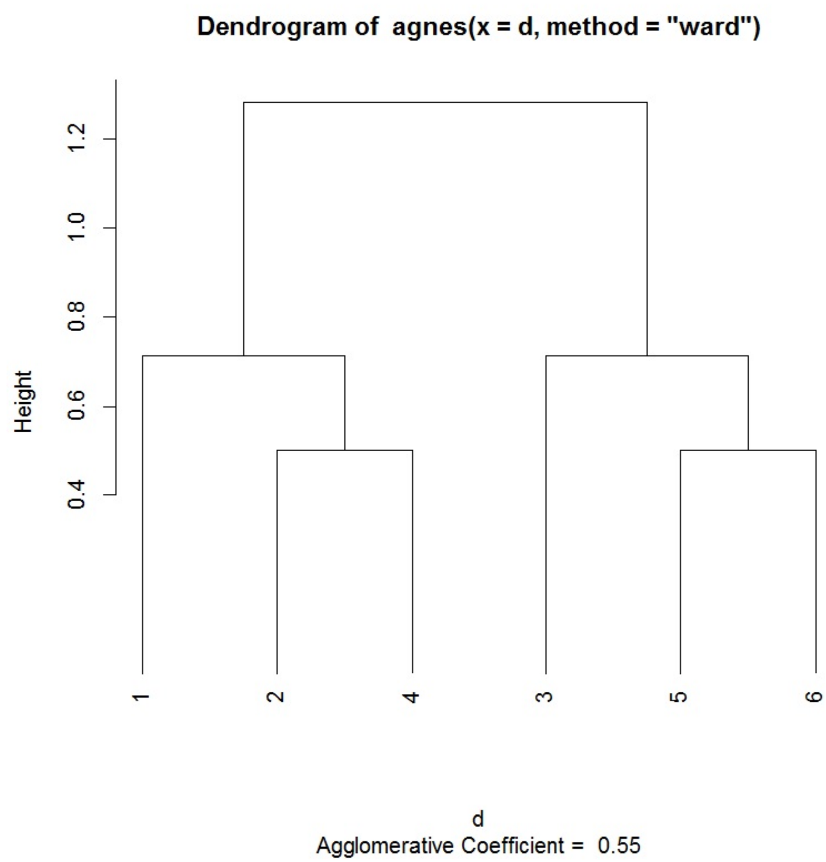
ω_1 est plus proche de ω_2 que ω_3 de ω_6.

On peut aller plus loin en calculant les distances entre tous les individus et faire une CAH.

Commandes R de l'exemple : Les commandes R ci-dessous renvoient les dissimilarités entre tous les individus avec l'indice de Jaccard et l'algorithme de CAH est mis en œuvre.

```
x = c(0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0,
0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1)
m = matrix(x, nrow = 6)
library(arules)
d = dissimilarity(m, method = "jaccard")
d
library(cluster)
ag = agnes(d, method = "ward")
cutree(ag, k = 2)
plot(ag, which = 2, hang = -1)
```

Cela renvoie le graphique :



"Distance" du Chi-deux : À partir du tableau disjonctif complet, on appelle "distance" du Chi-deux entre ω_u et ω_v la distance :

$$d(\omega_u, \omega_v) = \sqrt{\sum_{k=1}^{r} \frac{1}{\rho_k}(f_{u,k} - f_{v,k})^2},$$

où

$$f_{u,k} = \frac{n_{u,k}}{n_{u,\bullet}}, \qquad n_{u,\bullet} = \sum_{k=1}^{r} n_{u,k}, \qquad \rho_k = \frac{n_{\bullet,k}}{n}, \qquad n_{\bullet,k} = \sum_{i=1}^{n} n_{i,k}.$$

On peut aussi utiliser cette "distance" pour mettre en œuvre l'algorithme de CAH.

ACM et CAH : Une autre stratégie est de se ramener à des caractères quantitatifs en considérant les premières dimensions d'une analyse des correspondances multiples (ACM). Dès lors, on peut faire l'algorithme de CAH à partir des composantes principales de l'ACM. Une information résiduelle n'est alors volontairement pas prise en compte.

Modalités dominantes d'un groupe : On peut déterminer les modalités dominantes pour chacun des groupes formés. Pour se faire, pour chacune des r modalités, on peut faire un test d'hypothèses reposant sur loi normale. Soient $G_1, \ldots, G_q$ les q groupes formés. Pour tout $g \in \{G_1, \ldots, G_q\}$ et tout $k \in \{1, \ldots, r\}$, on calcule :

- $m_{k,g}$: le nombre d'individus du groupe g possèdant la k-ème modalité,
- $m_{k,\bullet}$: le nombre d'individus dans Γ possédant la k-ème modalité,
- $m_{\bullet,g}$: le nombre d'individus dans le groupe g,
- le $z_{obs,(k,g)}$:

$$z_{obs,(k,g)} = \frac{m_{k,g} - \frac{m_{k,\bullet} m_{\bullet,g}}{n}}{\sqrt{\left(\frac{n-m_{\bullet,g}}{n-1}\right) \frac{m_{k,\bullet} m_{\bullet,g}}{n} \left(1 - \frac{m_{k,\bullet}}{n}\right)}}.$$

Pour chaque groupe g, plus $|z_{obs,(k,g)}|$ est grand, plus la k-ème modalité importe dans la constitution du groupe. On peut évaluer le degré de significativité de cette importance avec la p-valeur :

$$\text{p-valeur}_{(k,g)} = \mathbb{P}(|Z| \geq |z_{obs,(k,g)}|), \qquad Z \sim \mathcal{N}(0,1).$$

Caractères de natures différentes : Si certains caractères sont qualitatifs et d'autres quantitatifs, on peut toujours transformer les caractères quantitatifs en qualitatifs en introduisant des classes de valeurs et en les considérant comme des modalités.

12 Enjeu de la classification supervisée

Contexte : On considère une population divisée en q groupes d'individus différents $\{G_1, \ldots, G_q\}$. Ces groupes sont distinguables suivant les valeurs de p caractères $X_1, \ldots, X_p$, sans que l'on ait connaissance des valeurs de $X_1, \ldots, X_p$ les distinguant. Soit Y un caractère qualitatif nominal égal au groupe dans lequel un individu appartient.

Pour n individus $\omega_1, \ldots, \omega_n$ de cette population, on dispose des valeurs de $Y, X_1, \ldots, X_p$. Elles sont généralement présentées sous la forme d'un tableau :

	Y	X_1	$\ldots$	X_p
ω_1	y_1	$x_{1,1}$	$\ldots$	$x_{p,1}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
ω_n	y_n	$x_{1,n}$	$\ldots$	$x_{p,n}$

Ces valeurs constituent les données.

Objectif : On s'intéresse à un nouvel individu ω_* de la population dont on connaît les valeurs $x_1, \ldots, x_p$ de $X_1, \ldots, X_p$, sans avoir connaissance de son groupe d'appartenance. Partant des données, l'objectif est de déterminer à quel groupe l'individu ω_* vérifiant

$(X_1, \ldots, X_p) = (x_1, \ldots, x_p) = x$ a le plus de chances d'appartenir.

Avec une modélisation probabiliste adaptée, ce groupe inconnu G vérifie : pour tout

$g \in \{G_1, \ldots, G_q\}$,

$$\mathbb{P}\left(\{Y = g\}/\{(X_1, \ldots, X_p) = x\}\right) \leq \mathbb{P}\left(\{Y = G\}/\{(X_1, \ldots, X_p) = x\}\right).$$

On peut alors l'écrire sous la forme :

$$G = \underset{g \in \{G_1, \ldots, G_q\}}{\operatorname{argmax}} \mathbb{P}\left(\{Y = g\}/\{(X_1, \ldots, X_p) = x\}\right).$$

Règle n^{o}1 : Qui se ressemble s'assemble.

Règle n^{o}2 : L'ami de mon ami est mon ami.

Méthodes : Pour déterminer G, plusieurs méthodes sont possibles. Parmi elles, il y a

- la méthode des k plus proches voisins (kNN pour K Nearest Neighbors),
- l'analyse discriminante,
- le modèle de régression logistique (pour $q = 2$),
- les arbres de décision,
- les réseaux de neurone,
- le support vector machine (SVM),
- les forêts aléatoires.

Ce document aborde quelques aspects des trois premiers points.

13 Méthode des k plus proches voisins

Méthode des k plus proches voisins (kNN pour K Nearest Neighbors) : Soient d une distance et $k \in \{1, \ldots, n\}$. Les k plus proches voisins de ω_* sont des k individus de $\Gamma = \{\omega_1, \ldots, \omega_n\}$ qui ressemblent le plus à ω_*. Ainsi, en notant $\mathcal{U}_k$ l'ensemble de ces k voisins, pour tout $\omega_i \in \mathcal{U}_k$ et tout $\omega_j \in \Gamma - \mathcal{U}_k$, on a

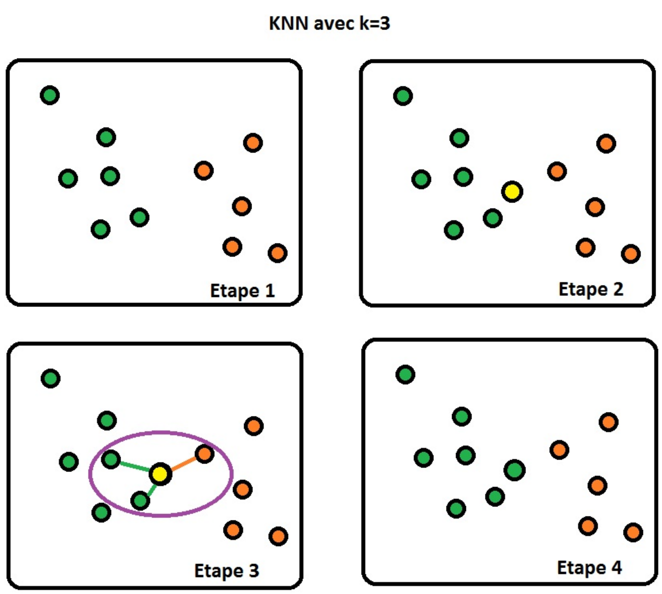
$$d(\omega_*, \omega_i) < d(\omega_*, \omega_j).$$

La méthode des k plus proches voisins propose d'affecter ω_* au groupe auquel la majorité de ses k voisins appartiennent. Mathématiquement, ce groupe peut s'écrire comme :

$$G_* = \underset{g \in \{G_1, \ldots, G_q\}}{\operatorname{argmax}} \sum_{i \in \mathcal{I}_k} \mathbb{I}_{\{y_i = g\}},$$

où $\mathcal{I}_k = \{i \in \{1, \ldots, n\};\ \omega_i \in \mathcal{U}_k\}$, $\mathbb{I}_{\{y_i = g\}} = 1$ si $y_i = g$ (donc ω_i appartient au groupe g), et 0 sinon.

Illustration : Une illustration de la méthode des k plus proches voisins avec $k = 3$ est présentée ci-dessous :



Quelques commandes R : Un exemple simple de commandes R est décrit ci-dessous.

On introduits 3 individus $A1$, $A2$ et $A3$ qui vont former un groupe A :

```
A1 = c(0.1, 0.4)

A2 = c(0.8, 0.9)

A3 = c(3, 3.5)
```

On introduits 3 individus $B1$, $B2$ et $B3$ qui vont former un groupe B :

```
B1 = c(5.7, 6.1)

B2 = c(5.5, 6.8)

B3 = c(6.5, 4.9)
```

On considère la matrice de données correspondante et on spécifie l'appartenance des individus aux groupes A et B :

```
train = rbind(A1, A2, A3, B1, B2, B3)

cl = factor(c(rep("A", 3), rep("B", 3)))
```

On s'intéresse à un nouvel individu ω_* de caractéristiques $X_1 = 4.1$ et $X_2 = 3.8$ et on trace le nuage de points :

```
point = c(4.1, 3.8)

plot(rbind(train, point))
```

On évalue le groupe dans lequel ω_* a le plus de chance d'appartenir avec la méthode des k plus proche voisins avec $k = 1$:

```
library(class)

knn(train, point, cl, k = 1)
```

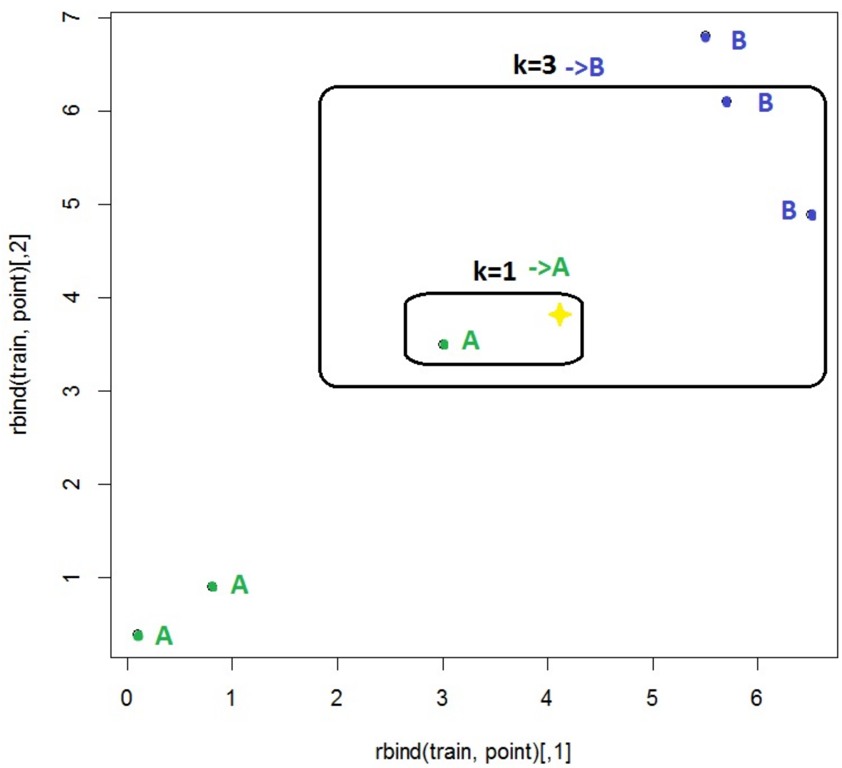
Cela renvoie A.

On fait la même chose avec $k = 3$:

```
knn(train, point, cl, k = 3)
```

Cela renvoie B.

On se rend compte de ce qui se passe par un graphique :



On considère un autre exemple portant sur des mesures du crane associées aux chiens et aux loups. L'entête du jeu de données "loups-g.txt" associé est :

	LCB	LMS	LPM	LP	LM	LAM	RACE
1	129	64	95	17.50	11.20	13.80	CHIEN
2	154	74	76	20.00	14.20	16.50	CHIEN
3	170	87	71	17.90	12.30	15.90	CHIEN
4	188	94	73	19.50	13.30	14.80	CHIEN
5	161	81	55	17.10	12.10	13.00	CHIEN
6	164	90	58	17.50	12.70	14.70	CHIEN

On propose les commandes R suivantes :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/loups-g.txt",
header = T)
attach(w)
w
cl = factor(w[ ,7])
point = c(210, 200, 76, 22, 12, 15)
library(class)
knn(w[ ,1:6], point, cl, k = 3)
```

Cela renvoie CHIEN. Ainsi, l'individu ω_* de caractéristiques $X_1 = 210$, $X_2 = 200$, $X_3 = 76$, $X_4 = 22$, $X_5 = 12$ et $X_6 = 15$ appartient au groupe des chiens.

kNN avec validation croisée : On peut aussi évaluer la qualité de l'algorithme des k plus proches voisins à l'aide d'une validation croisée. Cela consiste à extraire un petit groupe d'individus du jeu de données dont on connait parfaitement leur groupe d'affectation et de faire l'algorithme des k plus proches voisins sur ceux-ci. On peut ainsi voir le nombre de fois où l'algorithme se trompe.

Un exemple avec le jeu de données `iris` est présenté ci-dessous :

```
data(iris)
u = iris[ ,-5]
library(class)
class = as.factor(iris[ ,5])
results = knn.cv(u, class, 1:length(class))
levels(results) = levels(class)
table(results, class)
```

Cela renvoie :

```
setosa     versicolor    virginica
setosa      50     0      0
versicolor  0     47    3
virginica    0     3      47
```

On voit alors qu'il y a eut 6 iris qui ont été mal affectés.

Taux d'erreur de classification : Partant de l'algorithme des k plus proches voisins avec validation croisée, le taux moyen d'erreur de classification, noté t, est donné par le nombre d'individus mal affectés sur le nombre total d'individus.

Plus t est proche de 0, meilleur est la qualité prédictive du modèle.

On convient que la qualité de la classification est mauvaise lorsque $t > 0.5$.

Exemple : Sur l'exemple du jeu de données `iris`, ce taux est bon :

$$t = \frac{3+3}{50+47+47+3+3} = 0.04.$$

14 Analyse discriminante

Modélisation : On adopte le contexte de la classification supervisée. On modélise les caractères $Y, X_1, \ldots, X_p$ par des variables aléatoires réelles, en gardant les mêmes notations par convention.

Enjeu : Partant des données, pour tout $k \in \{1, \ldots, q\}$, on souhaite estimer la probabilité inconnue qu'un individu ω_* vérifiant $(X_1, \ldots, X_p) = (x_1, \ldots, x_p) = x$ appartienne au groupe G_k :

$$p_{G_k}(x) = \mathbb{P}(\{Y = G_k\}/\{(X_1, \ldots, X_p) = x\}).$$

Hypothèse de normalité : On suppose que, pour tout $k \in \{1, \ldots, q\}$,

$$\mathbb{P}(\{(X_1, \ldots, X_p) = x\}/\{Y = G_k\}) = \phi(x, \mu_k, \Sigma_k),$$

où $\phi(x, \mu_k, \Sigma_k)$ est la densité associée à la loi normale multivariée $\mathcal{N}_p(\mu_k, \Sigma_k)$:

$$\phi(x, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{p/2}\sqrt{\det(\Sigma_k)}} \exp\left(-\frac{1}{2}(x - \mu_k)^t \Sigma_k^{-1}(x - \mu_k)\right).$$

Les paramètres μ_k et Σ_k sont inconnus.

Règle de Bayes : Par la règle de Bayes et l'hypothèse de normalité, on peut montrer que

$$p_{G_k}(x) = \frac{r_k \phi(x, \mu_k, \Sigma_k)}{f(x, \mu, \Sigma, r)},$$

où $r_k = \mathbb{P}(Y = G_k)$ est inconnu et $f(x, \mu, \Sigma, r)$ désigne la densité inconnue de $(X_1, \ldots, X_p)$:

$$f(x, \mu, \Sigma, r) = \sum_{k=1}^{q} r_k \phi(x, \mu_k, \Sigma_k).$$

Par conséquent, en procédant par substitution, des estimations ponctuelles de $\mu = (\mu_1, \ldots, \mu_p)$, $\Sigma = (\Sigma_1, \ldots, \Sigma_p)$ et $r = (r_1, \ldots, r_q)$ amèneront une estimation ponctuelle de $p_{G_k}(x)$.

Estimateurs du maximum de vraisemblance : Pour estimer ponctuellement les paramètres μ, Σ et r, on utilise la méthode du maximum de vraisemblance. On note μ^*, Σ^* et r^* ces estimateurs. En pratique, ils sont approchés avec l'algorithme de Newton-Raphson.

Estimation : Une estimation ponctuelle de $p_{G_k}(x)$ est

$$p^*_{G_k}(x) = \frac{r^*_k \phi(x, \mu^*_k, \Sigma^*_k)}{f(x, \mu^*, \Sigma^*, r^*)}.$$

Prédiction du groupe : Le groupe auquel ω_* a le plus de chances d'appartenir est $G_* = G_{k_*}$ avec

$$k_* = \underset{k\in\{1,\ldots,q\}}{\operatorname{argmax}}\, p^*_{G_k}(x).$$

On peut montrer que

$$k_* = \underset{k\in\{1,\ldots,q\}}{\operatorname{argmax}} \left(2\ln(r^*_k) - \ln(\det(\Sigma^*_k)) - (x-\mu^*_k)^t(\Sigma^*_k)^{-1}(x-\mu^*_k)\right).$$

Quelques commandes R : Rappel du jeu de données `iris` : pour 3 variétés d'iris : Setosa, Versicolor, Virginica, et pour 150 iris par variété, on considère 4 caractères quantitatifs :

- X_1 la longueur en cm d'un pétale,
- X_2 la largeur en cm d'un pétale,
- X_3 la longueur en cm d'un sépale,
- X_4 la largeur en cm d'un sépale.

Ce sont les variables explicatives X_1, X_2, X_3 et X_4. La variable à expliquer Y est une variable qualitative dont les modalités sont les espèces d'iris $\{setosa, versicolor, virginica\}$.

Voici un problème de classification supervisée possible : on dispose d'un iris vérifiant :

$$X_1 = 2.1, \qquad X_2 = 3, \qquad X_3 = 2.3, \qquad X_4 = 4.3.$$

À l'aide des mesures effectuées, à quelle variété a t'il le plus de chances d'appartenir ?

Les commandes ci-dessous, dont `lda` signifiant Linear Discriminant Analysis, apportent une réponse :

```
data(iris)
library(MASS)
results = lda(Species ~ ., iris, prior = c(1, 1, 1) / 3)
library(MASS)
newiris = data.frame(Sepal.Length = 2.3, Sepal.Width = 4.3,
Petal.Length = 2.1, Petal.Width = 3)
plda = predict(results, newiris)
plda
```

Cela renvoie Virginica avec une probabilité de 0.9980522. On est donc presque sûr que l'iris observé est de l'espèce Virginica.

Validation croisée : On peut aussi évaluer la qualité de l'analyse discriminante à l'aide d'une validation croisée. Cela consiste à extraire un petit groupe d'individus du jeu de données dont on connait parfaitement leur groupe d'affectation et les tester avec l'analyse discriminante. On peut ainsi voir le nombre de fois où l'analyse se trompe.

Un exemple avec le jeu de données `iris` et la commande lda est présenté ci-dessous :

```
data(iris)
library(MASS)
results = lda(Species ~ ., iris, prior = c(1, 1, 1) / 3, CV = T)
table(iris$Species, results$class)
```

Cela renvoie :

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	1	49

On voit alors qu'il y a eut 3 iris qui ont été mal affectés.

Taux d'erreur de classification : Partant de l'analyse discriminante, le taux moyen d'erreur de classification, noté t, est donné par le nombre d'individus mal affectés sur le nombre total d'individus.

Plus t est proche de 0, meilleur est la qualité prédictive du modèle.

On convient que la qualité de la classification est mauvaise lorsque $t > 0.5$.

Exemple : Sur l'exemple du jeu de données `iris`, ce taux est :

$$t = \frac{1+2}{50+48+49+1+2} = 0.02.$$

Cela est très correct.

15 Modèle de régression logistique

Modélisation : On adopte le contexte de la classification supervisée. On modélise les caractères $Y, X_1, \ldots, X_p$ par des variables aléatoires réelles, en gardant les mêmes notations par convention.

Enjeu : On suppose que la population est divisée en $q = 2$ groupes G_1 et G_2. Partant des données, on souhaite estimer la probabilité inconnue qu'un individu ω_* vérifiant $(X_1, \ldots, X_p) = (x_1, \ldots, x_p) = x$ appartienne au groupe G_1 :

$$p(x) = \mathbb{P}(\{Y = G_1\}|\{(X_1, \ldots, X_p) = x\}).$$

Si l'estimation de $p(x)$ est supérieure à 0.5, alors ω_* a plus de chances d'appartenir à G_1.

Transformation logit : On appelle transformation logit la fonction :

$$\text{logit}(y) = \log\left(\frac{y}{1-y}\right) \in \mathbb{R}, \qquad y \in]0, 1[.$$

Son inverse est la fonction : $\text{logit}^{-1}(y) = \dfrac{\exp(y)}{1 + \exp(y)} \in]0, 1[$, $y \in \mathbb{R}$.

Modèle de régression logistique : On appelle modèle de régression logistique la modélisation :

$$\text{logit}(p(x)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

où $\beta_0, \ldots, \beta_p$ désignent $p + 1$ réels inconnus.

On en déduit l'expression de $p(x)$:

$$p(x) = \text{logit}^{-1}(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}.$$

Par conséquent, en procédant par substitution, des estimations ponctuelles de $\beta_0, \ldots, \beta_p$ amèneront une estimation ponctuelle de $p(x)$.

Estimateurs du maximum de vraisemblance : Pour estimer ponctuellement $\beta_0, \ldots, \beta_p$, on utilise la méthode du maximum de vraisemblance. On note $b_0, \ldots, b_p$ ces estimateurs.

En pratique, ils sont approchés avec l'algorithme de Newton-Raphson.

Estimation : Une estimation ponctuelle de $p(x)$ est

$$p^*(x) = \text{logit}^{-1}(b_0 + b_1 x_1 + \ldots + b_p x_p) = \frac{\exp(b_0 + b_1 x_1 + \ldots + b_p x_p)}{1 + \exp(b_0 + b_1 x_1 + \ldots + b_p x_p)}.$$

Prédiction du groupe : Le groupe auquel ω_* a le plus de chances d'appartenir est

$$G_* = \begin{cases} G_1 & \text{si } p^*(x) \geq 0.5, \\ G_2 & \text{sinon.} \end{cases}$$

Quelques commandes R : On considère le jeu de données "puits" dont voici l'entête :

Y	X1	X2
1	2.36	16.83
1	0.71	47.32
0	2.07	20.97

Un exemple de commandes R associées est donné ci-dessous :

```
w = read.table("https://chesneau.users.lmno.cnrs.fr/puits.txt", header = T)
attach(w)
w
library(stats)
reg = glm(Y ~ X1 + X2, family = binomial)
pred.prob = predict.glm(reg, data.frame(X1 = 1, X2 = 60), type = "response")
pred.mod = factor(ifelse(pred.prob > 0.5, "G1", "G2"))
pred.mod
```

Cela renvoie $G2$.

Taux d'erreur de classification : Partant du modèle de régression logistique, le taux moyen d'erreur de classification noté t est donné par le nombre d'individus mal affectés sur le nombre total d'individus.

Plus t est proche de 0, meilleur est la qualité prédictive du modèle.

On convient que la qualité de la classification est mauvaise lorsque $t > 0.5$.

Quelques commandes R : Un exemple de commandes R associées est donné ci-dessous :

```
pred.prob = predict(reg, type = "response")
pred.mod = factor(ifelse(pred.prob > 0.5, "G1", "G2"))
mc = table(Y, pred.mod)
t = (mc[1, 2] + mc[2, 1]) / sum(mc)
t
```

Cela renvoie $t = 0.6182119$. La classification est donc mauvaise.

Plus d'éléments seront donnés en Master 2. Voir, par exemple, les documents :

`https://chesneau.users.lmno.cnrs.fr/Reg-M2.pdf`

`https://chesneau.users.lmno.cnrs.fr/etudes-reg.pdf`

16 Exercices

Exercice 1. Soit $\mathbf{X}$ la matrice de données associée à 4 individus $\omega_1, \ldots, \omega_4$ définie par

$$\mathbf{X} = \begin{pmatrix} 2.0 & 3.0 \\ 7.0 & 4.0 \\ 3.5 & 3.0 \\ 0.5 & 5.0 \end{pmatrix}$$

1. Tracer le nuage de points $\mathcal{N}$ formé par $\{\omega_1, \ldots, \omega_4\}$ et donner son inertie totale.
2. Déterminer le tableau des écarts associé à $\mathcal{P}_0 = (\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\})$ obtenu par la méthode du plus proche voisin.
3. Au hasard, on forme le groupe d'individus : $\{\omega_3, \omega_4\}$. Déterminer le tableau des écarts associé à $\mathcal{P}_1 = (\{\omega_1\}, \{\omega_2\}, \{\omega_3, \omega_4\})$ obtenu par la méthode du plus proche voisin.

Exercice 2. Dans une étude en sciences sociale, on a étudié 2 caractères X_1 et X_2 sur 5 individus $\omega_1, \ldots, \omega_5$. Les données recueillies sont :

	X_1	X_2
ω_1	2.4	2.1
ω_2	7.1	4.3
ω_3	3.8	3.0
ω_4	1.2	5.1
ω_5	6.5	4.3

1. Déterminer la matrice des données.
2. Tracer le nuage de points $\mathcal{N}$ formé par $\Gamma = \{\omega_1, \ldots, \omega_5\}$.
3. Déterminer le centre de gravité g de $\mathcal{N}$, le tracer, et calculer les distances entre les individus et celui-ci.

4. Déterminer l'inertie totale de $\mathcal{N}$.

5. (a) Faire une classification avec l'algorithme CAH muni de la méthode du voisin le plus éloigné.

   (b) Tracer le dendrogramme associé. Proposer 2 groupes d'individus semblables.

6. (a) Faire une classification avec l'algorithme CAH muni de la méthode de Ward. À chaque nouvelle partition, indiquer l'inertie intra-classes associée.

   (b) Tracer le dendrogramme associé. Proposer 2 groupes d'individus semblables.

   (c) Déterminer le parangon du groupe $\{\omega_1, \omega_3, \omega_4\}$.

7. Comparer les classifications obtenues par les deux méthodes précédentes.

Exercice 3. Décrire brièvement l'enjeu des commandes R suivantes :

```
x = matrix(c(1, 16, 2, 9, 10, 16, 1, 17, 15, 2, 1, 37, 0, 14, 9, 9, 12, 4, 3, 13), ncol = 5, nrow =
4)


library(cluster)
ag = agnes(x, method = "average")
ag$merge
```

Exercice 4. Afin de comprendre l'évolution du singe, des chercheurs souhaitent faire une classification des primates. Pour ce faire, l'ADN d'un individu de chaque espèce est analysé : ω_1 est un humain, ω_2 est un chimpanzé, ω_3 est un bonobo, ω_4 est un gorille, ω_5 est un orang-outan et ω_6 est un gibbon. Les distances euclidiennes entre chacun de ces individus, caractérisant leur ressemblance quant à l'ADN, sont données dans le tableau incomplet suivant :

$\mathbf{E} =$

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
ω_1	0		0.95	1.49	1.85	2.56
ω_2	0.84	0	0.7	1.11	1.87	2.38
ω_3	0.95	0.7		1.35		2.15
ω_4	1.49	1.11	1.35	0	1.96	2.32
ω_5	1.85	1.87	2.05	1.96	0	2.30
ω_6	2.56		2.15	2.32	2.30	0

1. Recopier et compléter $\mathbf{E}$.
2. Faire une classification par l'algorithme CAH muni de la méthode du plus proche voisin. Tracer le dendrogramme associé.
3. Faire une classification par l'algorithme CAH muni de la méthode du voisin le plus éloigné. Tracer le dendrogramme associé.
4. Faire une classification par l'algorithme CAH muni de la méthode de la distance moyenne. Tracer le dendrogramme associé.
5. Proposer des commandes R pour répondre aux questions précédentes.

Exercice 5. Le tableau des distances associé à la partition initiale $\mathcal{P}_0 = (\{\omega_1\}, \ldots, \{\omega_4\})$ est

	ω_1	ω_2	ω_3	ω_4
ω_1	0	2	4	7
ω_2	2	0	4	5
ω_3	4	4	0	3
ω_4	7	5	3	0

1. Faire une classification avec l'algorithme CAH muni de la méthode de la distance moyenne.
2. Calculer le coefficient d'agglomération. Est-ce que la classification obtenue est fortement structurée en groupes ?

Exercice 6. Comprendre pourquoi la sortie finale des commandes R suivantes est `TRUE` :

```
x = matrix(c(2, 7.5, 3, 0.5, 6, 2, 4, 3, 5, 4), ncol = 2)

library(cluster)
ag = agnes(x, method = "complete")
pltree(ag, hang = -1)

a = (1 / 5) * ((1 - 1.41 / 7.07) + (1 - 1.5 / 7.07) + (1 - 1.41 / 7.07) + (1 - 1.5 / 7.07) + (1
- 3.35 / 7.07))

round(ag$ac, 2) == round(a, 2)
```

Exercice 7. Thomas, Fatiha, Emilie, Robert, Igor et Mathias sont malades. Ci-dessous quelques données relatives à leur maladie : Sexe $\in$ {Homme, Femme}, Fièvre $\in$ {Oui, Non},

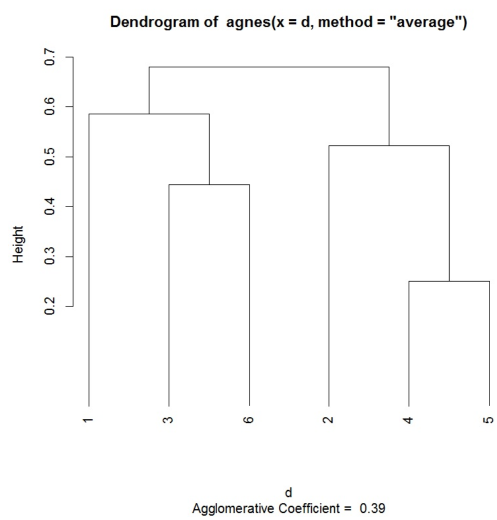
Toux $\in$ {Positif, Négatif} et pour les 4 tests : Test $\in$ {Positif, Négatif} :

Nom	Sexe	Fièvre	Toux	Test 1	Test 2	Test 3	Test 4
Thomas	H	O	P	P	P	N	P
Fatiha	F	O	N	P	N	P	P
Emilie	F	O	P	N	N	N	N
Robert	H	O	N	P	N	P	N
Igor	H	O	N	P	P	P	N
Mathias	H	O	P	P	N	N	N

On désire classer ces individus en 2 groupes suivant la ressemblance de leur diagnostique.

1. Dresser le tableau disjonctif complet.

2. Proposer des commandes R pour faire une classification avec l'algorithme CAH muni de la méthode de la distance moyenne avec la dissimilarité de Jaccard permettant d'obtenir le dendrogramme suivant :



3. Au vu du dendrogamme, quels sont les 2 groupes que vous suggéreriez ?

Exercice 8. La ressemblance génétique de cinq lignées d'une céréale (ℓ_1, ℓ_2, ℓ_3, ℓ_4 et ℓ_5) peut se mesurer par le nombre des bandes à la même hauteur révélées sur des profils obtenu par une certaine technique biologique (électrophorèse). Après analyse, on obtient les profils suivants :

ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_5
	—			—
—	—	—	—	—
—			—	
—		—		
—	—	—	—	—
—	—	—		
	—		—	—
—	—	—	—	—
		—		

Dans de contexte, on considère l'indice de similarité de Dice défini par :

$$s_{\ell,\ell_*} = \frac{2N_{\ell,\ell_*}}{N_\ell + N_{\ell_*}},$$

où N_ℓ est le nombre de bandes de la lignée ℓ, N_{ℓ_*} est le nombre de bandes de la lignée ℓ_* et N_{ℓ,ℓ_*} est le nombre de bandes à la même hauteur entre les lignées ℓ et ℓ_*. De même, on considère la dissimilarité de Dice défini par :

$$d_{\ell,\ell_*} = 1 - s_{\ell,\ell_*}.$$

1. Recopier et compléter le tableau des indices de similarités :

	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_5
ℓ_1		$\frac{8}{12}$			$\frac{6}{11}$
ℓ_2					
ℓ_3			1		
ℓ_4					$\frac{8}{10}$
ℓ_5					

.

En déduire le tableau des dissimilarités, avec des valeurs décimales.

2. Faire une classification par l'algorithme CAH avec la méthode du voisin le plus éloigné et la dissimilarité proposée. Tracer le dendrogramme associé.
3. Calculer le coefficient d'agglomération.

Exercice 9. On a relevé les valeurs de 2 caractères X_1 et X_2 sur 6 individus $\omega_1, \ldots, \omega_6$. Les données recueillies sont :

	X_1	X_2
ω_1	-2	3
ω_2	-2	1
ω_3	-2	-1
ω_4	2	-1
ω_5	2	1
ω_6	1	0

1. Tracer le nuage de points $\mathcal{N}$ formé par $\{\omega_1, \ldots, \omega_6\}$.
2. Faire une classification avec l'algorithme des centres mobiles avec, pour centres initiaux, ω_1 et ω_2.
3. Reprendre la question précédente avec, pour centres initiaux, ω_4 et ω_6.

Exercice 10. On a relevé les valeurs de 2 caractères : X_1 et X_2, sur 7 individus $\omega_1, \ldots, \omega_6$. Les données recueillies sont :

	X_1	X_2
ω_1	1.42	3.58
ω_2	4.23	4.65
ω_3	0.34	1.04
ω_4	1.30	1.95
ω_5	2.56	3.46
ω_6	1.33	4.67
ω_7	3.17	4.59

1. Tracer le nuage de points $\mathcal{N}$ formé par $\{\omega_1, \ldots, \omega_7\}$.
2. Faire une classification par l'algorithme des centres mobiles avec, pour centres initiaux, m_1 de coordonnées $(1, 1)$ et m_2 de coordonnées $(3, 3)$.
3. Proposer des commandes R pour répondre à la question précédente.

Exercice 11. Dans une étude industrielle, on a étudié 2 caractères X_1 et X_2 sur 6 individus $\omega_1, \ldots, \omega_6$. Les données recueillies sont :

Groupe A			Groupe B		
ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
$(0, 0)$	$(1, 1)$	$(2, 2)$	$(6, 6)$	$(5.5, 7)$	$(6.5, 5)$

Déterminer le groupe d'appartenance d'un individu ω_* vérifiant $X_1 = 4$ et $X_2 = 4$ avec la méthode des k plus proches voisins (kNN) pour $k = 3$. Faire un dessin.

Exercice 12. Soit X le caractère donnant la taille en mètres d'un client qui entre dans un magasin de vêtements masculins au centre ville de Caen. On suppose que :

– si le client est une femme, X suit la loi $\mathcal{N}(1.65, 0.16^2)$,

– si le client est un homme, X suit la loi $\mathcal{N}(1.75, 0.15^2)$,

– la probabilité qu'un client homme rentre est 0.7.

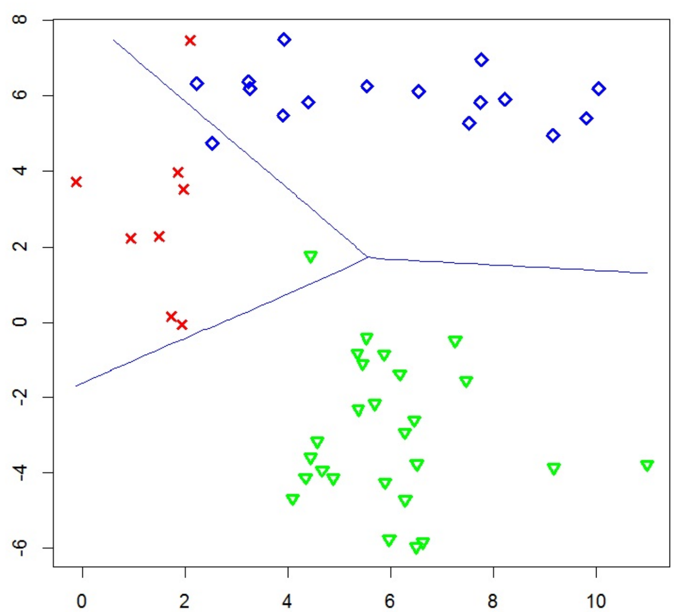
Un client rentre dans le magasin. On sait qu'il mesure 1.60 mètres. Qu'elle est la probabilité que ce soit un homme ?

Exercice 13. Voici un exemple de commandes R illustrant la classification supervisée par l'analyse discriminante :

```
p1 = 0.2 ; p2 = 0.3 ; p3 = 0.5
n = 50
s1 = c(1, 2) ; s2 = c(3, 1) ; s3 = c(1.5, 2)
s = rbind(s1, s2, s3)

m1 = c(1, 2) ; m2 = c(6, 6) ; m3 = c(6, -2)
m = rbind(m1, m2, m3)

c = sample(c(1, 2, 3), size = n, prob = c(p1, p2, p3), replace = TRUE)

x = cbind(rnorm(n, m[c, 1], s[c, 1]), rnorm(n, m[c, 2], s[c, 2]))

couleur = rep("red", n)
couleur[c == 2] = "blue"
couleur[c == 3] = "green"
x11()

plot(x,col = couleur, pch = 3 + c, lwd = 3, , xlab = " ", ylab = " ")
library(MASS)
T = as.factor(couleur)
x.lda = lda(x, T)
len = 50
xp = seq(min(x[ ,1]), max(x[ ,1]), length = len)
yp = seq(min(x[,2]), max(x[,2]), length = len)
grille = expand.grid(z1 = xp, z2 = yp)
```

```
Z = predict(x.lda, grille)
T.lda = predict(x.lda)$class


zp = Z$post[ ,3] - pmax(Z$post[ ,2], Z$post[ ,1])
contour(xp, yp, matrix(zp, len), add = TRUE, levels = 0, drawlabels = FALSE, col = "blue")


zp = Z$post[ ,1] - pmax(Z$post[ ,2], Z$post[ ,3])
contour(xp, yp, matrix(zp, len), add = TRUE, levels = 0, drawlabels = FALSE, col = "blue")
```

On obtient :



Quelle est la proportion de points mal classés ?

Exercice 14. On cherche à modéliser la probabilité de survenue du cancer du poumon chez l'homme en fonction de trois caractères : l'âge, le poids et le tabagisme. Pour ce faire, 10000 hommes dans la population sont considérés. On mesure les valeurs de (Y, X_1, X_2, X_3), où Y est un caractère qui vaut 1 si l'homme présente un tel cancer et 0 sinon, X_1 est l'âge en années, X_2 est le poids de l'individu en kilogrammes et X_3 la quantité de cigarettes consommées ces 10 dernières années.

1. Proposer brièvement une modélisation adaptée au problème.
2. Expliquer par quelle méthode on obtient les estimations des paramètres inconnus du modèle proposé au résultat de la question 1.

Exercice 15. On considère le jeu de données "anesthésie" :

```
w = read.table("https ://chesneau.users.lmno.cnrs.fr/anesthésie.txt", header = T)
```

Trente patients ont reçu un certain niveau de dosage d'agent anesthésique pendant 15 minutes. Puis une incision leur est faite. Il est ensuite noté si le patient a bougé ou pas lors de l'incision. Ainsi, pour chaque patient, on dispose :

– du dosage de l'agent anesthésique pendant 15 minutes (variable $X1$),

– du fait qu'il ait bougé ou pas (variable Y, avec $Y = 1$ pour bougé).

On souhaite expliquer Y en fonction de $X1$.

1. On exécute les commandes R suivantes :

```
attach(w)
library(stats)
reg = glm(Y ~ X1, family = binomial)
summary(reg)
predict.glm(reg, data.frame(X1 = 1.25), type = "response")
confint.default(reg, level = 0.95)
```

Cela renvoie :

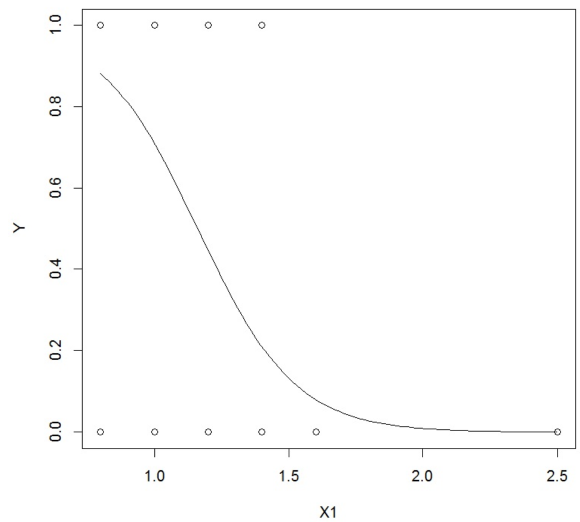
```
> summary(reg)

Call:
glm(formula = Y ~ X1, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.06900  -0.68666  -0.03413   0.74407   1.76666

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    6.469      2.418   2.675  0.00748 **
X1            -5.567      2.044  -2.724  0.00645 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.455  on 29  degrees of freedom
Residual deviance: 27.754  on 28  degrees of freedom
AIC: 31.754

Number of Fisher Scoring iterations: 5

> predict.glm(reg, data.frame(X1 = 1.25), type="response")
       1
0.379946
> confint.default(reg, level=0.95)
                2.5 %    97.5 %
(Intercept)  1.728560 11.208790
X1          -9.572126 -1.561398
```

Quel est le modèle considéré ? Est-ce qu'un patient ayant eut pour dosage X1 = 1.25 a plus de chance de bouger que de ne pas bouger ?

2. On exécute les commandes R suivantes :

```
plot(X1, Y)
curve(predict(reg, data.frame(X1 = x), type = "response"), add = T)
```

Cela renvoie :



Que représente ce graphique ?

3. On exécute les commandes R suivantes :

```
pred.prob = predict(reg, type = "response")
pred.mod = factor(ifelse(pred.prob >0.5, "1", "0"))
mc = table(Y, pred.mod)
t = (mc[1, 2] + mc[2, 1]) / sum(mc)
t
```

Cela renvoie $t = 0.2$.

Que représente cette quantité ? Est-ce que le résultat est satisfaisant ?

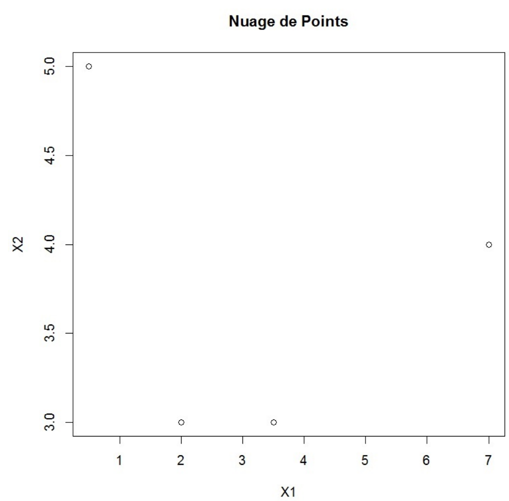
17 Solutions

Solution 1.

1. On peut utiliser les commandes R :

```
x = c(2, 7, 3.5, 0.5, 3, 4, 3, 5)
m = matrix(x, ncol = 2, nrow = 4)
plot(m, main = "Nuage de Points", xlab = "X1", ylab = "X2")
```

Cela renvoie :



L'inertie totale du nuage de points est

$$\mathcal{I}(\mathcal{N}) = \sigma_1^2 + \sigma_2^2,$$

avec

$$\sigma_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)^2} = 2,41, \qquad \sigma_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{2,i} - \overline{x}_2)^2} = 0.82.$$

Donc

$$\mathcal{I}(\mathcal{N}) = 2.41^2 + 0.82^2 = 6.50.$$

La commande R associée est :

```
Itot = (3/4) * (sd(m[ ,1])^2 + sd(m[ ,2])^2)
Itot
```

2. Le tableau des écarts associé à $\mathcal{P}_0 = (\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\})$ obtenu par la méthode du plus proche voisin est

	ω_1	ω_2	ω_3	ω_4
ω_1	0	5.09	1.50	2.50
ω_2	5.09	0	3.64	6.57
ω_3	1.50	3.64	0	3.60
ω_4	2.50	6.57	3.60	0

Cela est donné par les commandes R :

```
dist(m)
d
```

3. Le tableau des écarts associé à $\mathcal{P}_1 = (\{\omega_1\}, \{\omega_2\}, \{\omega_3, \omega_4\})$ obtenu par la méthode du plus proche voisin est

	ω_1	ω_2	$\{\omega_3, \omega_4\}$
ω_1	0	5.09	1.50
ω_2	5.09	0	3.64
$\{\omega_3, \omega_4\}$	1.50	3.64	0

Solution 2.

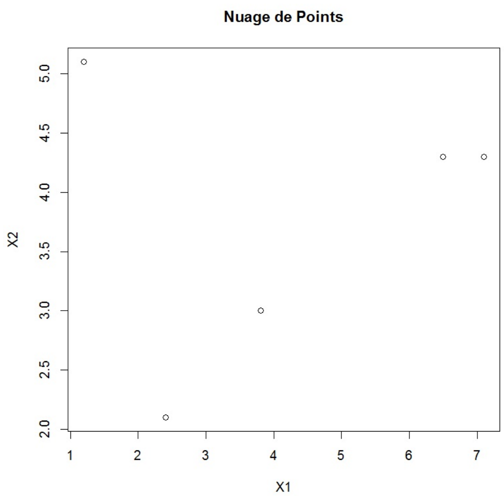
1. La matrice des données $\mathbf{X}$ associée à $\{\omega_1, \ldots, \omega_5\}$ est

$$\mathbf{X} = \begin{pmatrix} 2.4 & 2.1 \\ 7.1 & 4.3 \\ 3.8 & 3.0 \\ 1.2 & 5.1 \\ 6.5 & 4.3 \end{pmatrix}$$

2. On peut utiliser les commandes R :

```
x = c(2.4, 7.1, 3.8, 1.2, 6.5, 2.1, 4.3, 3, 5.1, 4.3)

m = matrix(x, ncol = 2, nrow = 5)

plot(m, main="Nuage de Points", xlab = "X1", ylab = "X2")
```

Cela renvoie :



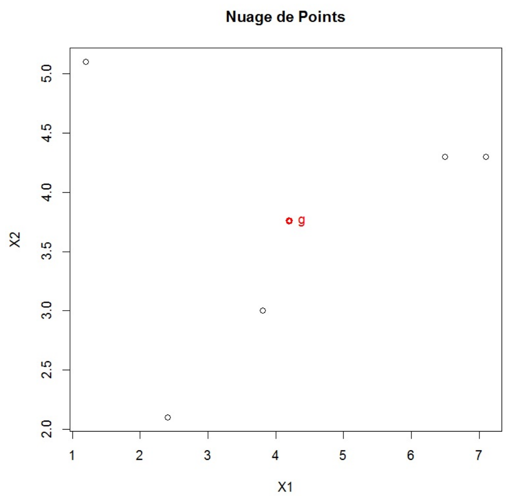
3. Le centre de gravité g de $\mathcal{N}$ est le point de coordonnées

$$\left(\frac{2.4 + 7.1 + 3.8 + 1.2 + 6.5}{5}, \ \frac{2.1 + 4.3 + 3.0 + 5.1 + 4.3}{5} \right) = (4.20, \ 3.76).$$

On peut utiliser les commandes R pour le visualiser :

```
g = colMeans(m)

g

points(g[1], g[2], col = "red", lwd = 3)

text(4.20, 3.76, "g", col = "red")
```

Cela renvoie :



Les distances euclidiennes entre les individus et g sont confectionnées dans le tableau suivant :

	ω_1	ω_2	ω_3	ω_4	ω_5
g	2.44	2.94	0.85	3.28	2.36

On peut utiliser la boucle :

```
dg = c(0, 0, 0, 0, 0)
for (k in 1 :5){
dg[k] = dist(rbind(m[k, ], g))
}
dg
```

4. L'inertie totale du nuage de points est

$$\mathcal{I}(\mathcal{N}) = \sigma_1^2 + \sigma_2^2,$$

avec

$$\sigma_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)^2} = 2.284, \qquad \sigma_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{2,i} - \overline{x}_2)^2} = 1.068.$$

Donc

$$\mathcal{I}(\mathcal{N}) = 2.284^2 + 1.068^2 = 6.35.$$

Les commandes R associées sont :

```
Itot = (4 / 5) * (sd(m[ ,1]) ^2+sd(m[ ,2]) ^2)
Itot
```

5. On rappelle que la méthode du voisin le plus éloigné est caractérisée par l'écart :

$$e(A, B) = \max_{(\omega,\omega_*)\in A\times B} d(\omega, \omega_*).$$

◦ Le tableau des écarts associé à $\mathcal{P}_0 = (\{\omega_1\}, \ldots, \{\omega_5\})$ est

	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0	5.18	1.66	3.23	4.65
ω_2	5.18	0	3.54	5.95	**0.60**
ω_3	1.66	3.54	0	3.34	2.99
ω_4	3.23	5.95	3.34	0	5.36
ω_5	4.65	**0.60**	2.99	5.36	0

On fait :

```
d = dist(m)
d
```

Les éléments (individus) ω_2 et ω_5 ont l'écart le plus petit : ce sont les éléments les plus proche. On les regroupe pour former $A = \{\omega_2, \omega_5\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_1 = (\{\omega_1\}, \{\omega_3\}, \{\omega_4\}, A).$$

◦ On a

$$e(\omega_1, A) = \max(e(\omega_1, \omega_2), e(\omega_1, \omega_5)) = \max(5.18, 4.65) = 5.18,$$

$$e(\omega_3, A) = \max(e(\omega_3, \omega_2), e(\omega_3, \omega_5)) = \max(3.54, 2.99) = 3.54$$

et

$$e(\omega_4, A) = \max(e(\omega_4, \omega_2), e(\omega_4, \omega_5)) = \max(5.95, 5.36) = 5.95.$$

Le tableau des écarts associé à $\mathcal{P}_1$ est

	ω_1	ω_3	ω_4	A
ω_1	0	**1.66**	3.23	5.18
ω_3	**1.66**	0	3.34	3.54
ω_4	3.23	3.34	0	5.95
A	5.18	3.54	5.95	0

Les éléments (individus) ω_1 et ω_3 sont les plus proche. On les regroupe pour former $B = \{\omega_1, \omega_3\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_2 = (\{\omega_4\}, A, B).$$

∘ On a

$$e(\omega_4, B) = \max(e(\omega_4, \omega_1), e(\omega_4, \omega_3)) = \max(3.23, 3.34) = 3.34$$

et

$$e(A, B) = \max(e(\omega_1, A), e(\omega_3, A)) = \max(5.18, 3.54) = 5.18.$$

Le tableau des écarts associé à $\mathcal{P}_2$ est

	ω_4	A	B
ω_4	0	5.95	**3.34**
A	5.95	0	5.18
B	**3.34**	5.18	0

Les éléments ω_4 et B sont les plus proche.

On les regroupe pour former $C = \{\omega_4, B\} = \{\omega_1, \omega_3, \omega_4\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_3 = (A, C).$$

∘ On a

$$e(A, C) = \max(e(\omega_4, A), e(A, B)) = \max(5.95, 5.18) = 5.95.$$

Le tableau des écarts associé à $\mathcal{P}_3$ est

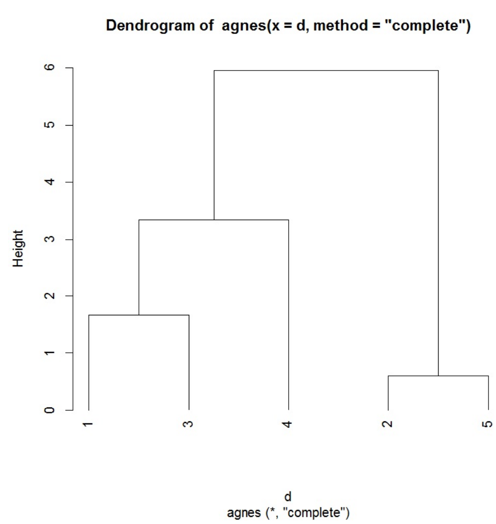
	A	C
A	0	**5.95**
C	**5.95**	0

Il ne reste plus que 2 éléments, A et C, on les regroupe. Il vient la partition $\mathcal{P}_4 = \{\omega_1, \ldots, \omega_5\} = \Gamma$. Cela termine l'algorithme de CAH.

Au final,

- les éléments $\{\omega_2\}$ et $\{\omega_5\}$ ont été regroupés avec un écart de 0.60,
- les éléments $\{\omega_1\}$ et $\{\omega_3\}$ ont été regroupés avec un écart de 1.66,
- les éléments $B = \{\omega_1, \omega_3\}$ et $\{\omega_4\}$ ont été regroupés avec un écart de 3.34,
- les éléments $C = \{\omega_4, B\}$ et $A = \{\omega_2, \omega_5\}$ ont été regroupés avec un écart de 5.95.

On peut donc construire le dendrogramme associé :



Cela est obtenu avec les commandes :

```
library(cluster)
ag = agnes(d, method = "complete")
pltree(ag, hang = -1)
```

Comme le plus grand saut se situe entre les éléments C et A, on envisage de considérer les deux groupes :

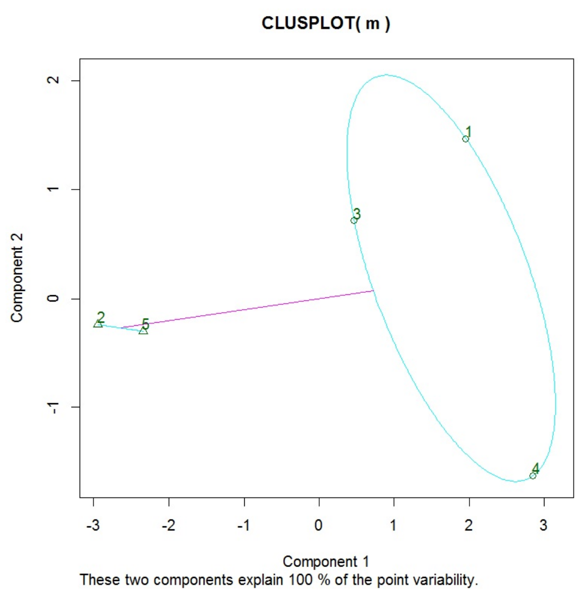
$$C = \{\omega_1, \omega_3, \omega_4\}, \qquad A = \{\omega_2, \omega_5\}.$$

Pour compléter ce résultat, on peut

○ afficher le regroupement obtenu :

```
ag2 = cutree(ag, 2)
clusplot(m,ag2,labels=3)
```

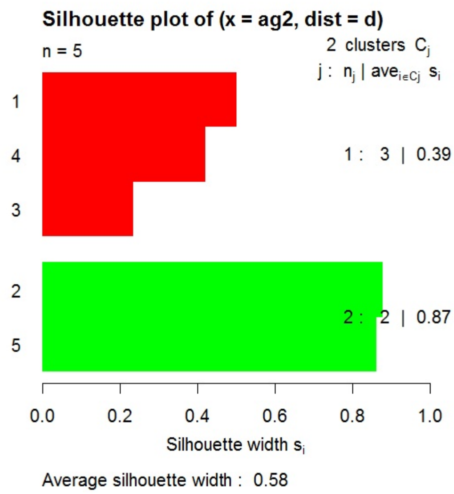
Cela renvoie :



○ s'intéresser à la qualité de partition en étudiant les indices de silhouette et la largeur de silhouette :

```
ag2 = cutree(ag, 2)
si = silhouette(ag2, d)
plot(si, col = c("red", "green"))
```

Cela renvoie :



On constate alors une structure forte de la partition car $S \in]0.51, 1]$.

6. (a) On rappelle que la méthode de Ward est caractérisée par l'écart :

$$e(A, B) = \frac{n_A n_B}{n_A + n_B} d^2(g_A, g_B).$$

∘ Le tableau des écarts associé à $\mathcal{P}_0 = (\{\omega_1\}, \ldots, \{\omega_5\})$ est

	ω_1	ω_2	ω_3	ω_4	ω_5
ω_1	0	13.46	1.38	5.22	10.82
ω_2	13.46	0	6.29	17.72	**0.18**
ω_3	1.38	6.29	0	5.58	4.49
ω_4	5.22	17.72	5.58	0	14.36
ω_5	10.82	**0.18**	4.49	14.36	0

Les éléments étant tous des individus, on a utilisé la formule :

$$e(\omega_u, \omega_v) = \frac{1 \times 1}{1 + 1} d^2(\omega_u, \omega_v).$$

Les éléments (individus) ω_2 et ω_5 ont l'écart le plus petit : ce sont les éléments les plus proche. On les regroupe pour former $A = \{\omega_2, \omega_5\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_1 = (\{\omega_1\}, \{\omega_3\}, \{\omega_4\}, A).$$

L'inertie intra de $\mathcal{P}_1$ est

$$\mathcal{I}_{intra}(\mathcal{P}_1) = \frac{1}{5} \times 0.18 = 0.036.$$

○ Le centre de gravité associé à A est le point g_A de coordonnées $\left(\frac{7.1+6.5}{2}, \frac{4.3+4.3}{2}\right) = (6.8, 4.3)$. On a

$$e(\omega_1, A) = \frac{1 \times 2}{1+2}((2.4 - 6.8)^2 + (2.1 - 4.3)^2) = 16.13,$$

$$e(\omega_3, A) = \frac{1 \times 2}{1+2}((3.8 - 6.8)^2 + (3 - 4.3)^2) = 7.12$$

et

$$e(\omega_4, A) = \frac{1 \times 2}{1+2}((1.2 - 6.8)^2 + (5.1 - 4.3)^2) = 21.33.$$

Le tableau des écarts associé à $\mathcal{P}_1$ est

	ω_1	ω_3	ω_4	A
ω_1	0	**1.38**	5.22	16.13
ω_3	**1.38**	0	5.58	7.12
ω_4	5.22	5.58	0	21.33
A	16.13	7.12	21.33	0

Les éléments (individus) ω_1 et ω_3 sont les plus proche. On les regroupe pour former $B = \{\omega_1, \omega_3\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_2 = (\{\omega_4\}, A, B).$$

L'inertie intra de $\mathcal{P}_2$ est

$$\mathcal{I}_{intra}(\mathcal{P}_2) = 0.036 + \frac{1}{5} \times 1.38 = 0.312.$$

∘ Le centre de gravité associé à B est le point g_B de coordonnées $\left(\frac{2.4+3.8}{2}, \frac{2.1+3.0}{2}\right) = (3.1, 2.55)$. On a

$$e(\omega_4, B) = \frac{1 \times 2}{1+2}((1.2 - 3.1)^2 + (5.1 - 2.55)^2) = 6.74$$

et

$$e(A, B) = \frac{2 \times 2}{2+2}((6.8 - 3.1)^2 + (4.3 - 2.55)^2) = 16.75.$$

Le tableau des écarts associé à $\mathcal{P}_2$ est

	ω_4	A	B
ω_4	0	21.33	**6.74**
A	21.33	0	16.75
B	**6.74**	16.75	0

Les éléments ω_4 et B sont les plus proche. On les regroupe pour former $C = \{\omega_4, B\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_3 = (A, C).$$

L'inertie intra de $\mathcal{P}_3$ est

$$\mathcal{I}_{intra}(\mathcal{P}_3) = 0.312 + \frac{1}{5} \times 6.74 = 1.66.$$

∘ Le centre de gravité associé à C est le point g_C de coordonnées

$$\left(\frac{2.4 + 3.8 + 1.2}{3}, \frac{2.1 + 3.0 + 5.1}{3}\right) = (2.46, 3.4).$$

On a

$$e(A, C) = \frac{2 \times 3}{2+3}((6.8 - 2.46)^2 + (4.3 - 3.4)^2) = 23.57.$$

Le tableau des écarts associé à $\mathcal{P}_3$ est

	A	C
A	0	**23.57**
C	**23.57**	0

Il ne reste plus que 2 éléments, A et C, on les regroupe, d'où la partition $\mathcal{P}_4 = \{\omega_1, \ldots, \omega_5\} = \Gamma$. On a

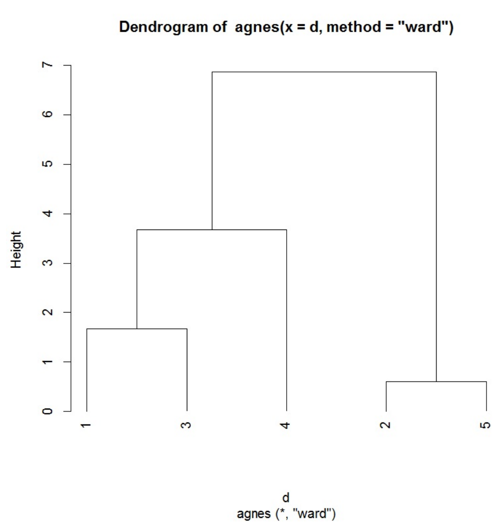
$$\mathcal{I}_{intra}(\mathcal{P}_4) = 1.66 + \frac{1}{5} \times 23.57 = 6.34.$$

On retrouve (avec les approximation) l'inertie totale $\mathcal{I}(\mathcal{N})$, ce qui rassurant sur l'aspect numérique.

Cela termine l'algorithme de CAH avec l'écart de Ward. Au final,

- les éléments $\{\omega_2\}$ et $\{\omega_5\}$ ont été regroupés avec un écart de 0.18,
- les éléments $\{\omega_1\}$ et $\{\omega_3\}$ ont été regroupés avec un écart de 1.38,
- les éléments $B = \{\omega_1, \omega_3\}$ et $\{\omega_4\}$ ont été regroupés avec un écart de 6.74,
- les éléments $C = \{\omega_4, B\}$ et $A = \{\omega_2, \omega_5\}$ ont été regroupés avec un écart de 23.57.

On peut donc construire le dendrogramme associé :



Cela est obtenu avec les commandes :

```
library(cluster)
ag = agnes(d, method = "ward")
pltree(ag, hang = -1)
```

Précisons que pour 2 éléments Q et R qui se regroupent, avec la commande agnes, la hauteur de la branche correspondante est donnée par la formule : $\sqrt{2e(Q,R)}$. On a $\sqrt{2 \times 0.18} = 0.6$, $\sqrt{2 \times 1.38} = 1.66$, $\sqrt{2 \times 6.74} = 3.67$ et $\sqrt{2 \times 23.57} = 6.86$.

Comme le plus grand saut se situe entre les éléments C et A, on envisage de considérer les deux groupes :

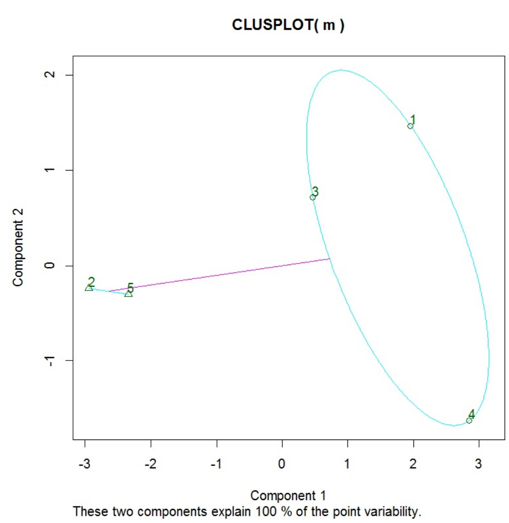
$$C = \{\omega_1, \omega_3, \omega_4\}, \qquad A = \{\omega_2, \omega_5\}.$$

Pour compléter ce résultat, on peut

○ afficher le regroupement obtenu :

```
ag2 = cutree(ag, 2)
clusplot(m, ag2, labels = 3)
```

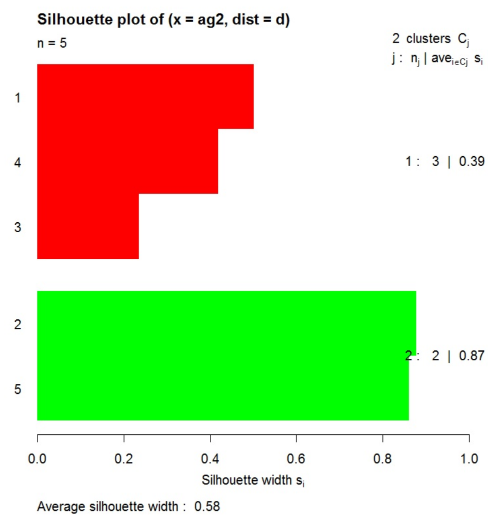
Cela renvoie :



○ s'intéresser à la qualité de partition en étudiant les indices de silhouette et la largeur de silhouette :

```
ag2 = cutree(ag, 2)
si = silhouette(ag2, d)
plot(si, col = c("red", "green"))
```

Cela renvoie :



On constate alors une structure forte de la partition car $S \in]0.51, 1]$.

(b) Le centre de gravité de $C = \{\omega_1, \omega_3, \omega_4\}$ est de coordonnées :

$$\left(\frac{2.4 + 3.8 + 1.2}{3}, \frac{2.1 + 3.0 + 5.1}{3} \right) = (2.46, 3.4).$$

Les distances euclidiennes entre les individus du groupe C et g_C sont confectionnées dans le tableau suivant :

	ω_1	ω_3	ω_4
g_C	1.30	1.39	2.11

La plus petite distance correspondant à $d(\omega_1, g_C)$, le parangon de C est ω_1.

(c) Les classifications obtenues dans les questions 5 et 6 sont identiques.

Solution 3. L'enjeu des commandes R présentées est de mettre en œuvre l'algorithme CAH avec la méthode de la distance moyenne (average linkage) sur 4 individus et 5 caractères. Pour ce faire, on utilise la commande agnes du package cluster. Il en ressort les regroupements obtenus avec la commande ag$merge.

Solution 4.

1. On a

$\mathbf{E} =$

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
ω_1	0	**0.84**	0.95	1.49	1.85	2.56
ω_2	0.84	0	0.70	1.11	1.87	2.38
ω_3	0.95	0.70	**0**	1.35	**2.05**	2.15
ω_4	1.49	1.11	1.35	0	1.96	2.32
ω_5	1.85	1.87	2.05	1.96	0	2.30
ω_6	2.56	**2.38**	2.15	2.32	2.30	0

2. CAH avec la méthode du plus proche voisin.

   ◦ Par le résultat de la question 1, le tableau des écarts associé à $\mathcal{P}_0 = (\{\omega_1\}, \ldots, \{\omega_6\})$ est

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
ω_1	0	0.84	0.95	1.49	1.85	2.56
ω_2	0.84	0	**0.70**	1.11	1.87	2.38
ω_3	0.95	0.70	0	1.35	2.05	2.15
ω_4	1.49	1.11	1.35	0	1.96	2.32
ω_5	1.85	1.87	2.05	1.96	0	2.30
ω_6	2.56	2.38	2.15	2.32	2.30	0

   Les éléments (individus) ω_2 et ω_3 ont l'écart le plus petit : ce sont les éléments les plus proches. On les regroupe pour former le groupe : $A = \{\omega_2, \omega_3\}$. On a une nouvelle partition de Ω : $\mathcal{P}_1 = (\{\omega_1\}, \{\omega_4\}, \{\omega_5\}, \{\omega_6\}, A)$.

- On a

$$e(\omega_1, A) = \min(e(\omega_1, \omega_2), e(\omega_1, \omega_3)) = \min(0.84, 0.95) = 0.84.$$

$$e(\omega_4, A) = \min(e(\omega_4, \omega_2), e(\omega_4, \omega_3)) = \min(1.11, 1.35) = 1.11.$$

$$e(\omega_5, A) = \min(e(\omega_5, \omega_2), e(\omega_5, \omega_3)) = \min(1.87, 2.05) = 1.87$$

et

$$e(\omega_6, A) = \min(e(\omega_6, \omega_2), e(\omega_6, \omega_3)) = \min(2.38, 2.15) = 2.15.$$

Le tableau des écarts associé à $\mathcal{P}_1$ est

	ω_1	ω_4	ω_5	ω_6	A
ω_1	0	1.49	1.85	2.56	**0.84**
ω_4	1.49	0	1.96	2.32	1.11
ω_5	1.85	1.96	0	2.30	1.87
ω_6	2.56	2.32	2.30	0	2.15
A	0.84	1.11	1.87	2.15	0

Les éléments ω_1 et A sont les plus proches. On les regroupe pour former le groupe : $B = \{\omega_1, A\}$. On a une nouvelle partition de Ω : $\mathcal{P}_2 = (\{\omega_4\}, \{\omega_5\}, \{\omega_6\}, B)$.

- On a

$$e(\omega_4, B) = \min(e(\omega_4, \omega_1), e(\omega_4, A)) = \min(1.49, 1.11) = 1.11.$$

$$e(\omega_5, B) = \min(e(\omega_5, \omega_1), e(\omega_5, A)) = \min(1.85, 1.87) = 1.85.$$

et

$$e(\omega_6, B) = \min(e(\omega_6, \omega_1), e(\omega_6, A)) = \min(2.56, 2.15) = 2.15.$$

Le tableau des écarts associé à $\mathcal{P}_2$ est

	ω_4	ω_5	ω_6	B
ω_4	0	1.96	2.32	**1.11**
ω_5	1.96	0	2.30	1.85
ω_6	2.32	2.30	0	2.15
B	1.11	1.85	2.15	0

Les éléments ω_4 et B sont les plus proches. On les regroupe pour former le groupe : $C = \{\omega_4, B\}$. On a une nouvelle partition de Ω : $\mathcal{P}_3 = (\{\omega_5\}, \{\omega_6\}, C)$.

- On a

$$e(\omega_5, C) = \min(e(\omega_5, \omega_4), e(\omega_5, B)) = \min(1.96, 1.85) = 1.85$$

et

$$e(\omega_6, C) = \min(e(\omega_6, \omega_4), e(\omega_6, B)) = \min(2.32, 2.15) = 2.15.$$

Le tableau des écarts associé à $\mathcal{P}_3$ est

	ω_5	ω_6	C
ω_5	0	2.30	**1.85**
ω_6	2.30	0	2.15
C	1.85	2.15	0

Les éléments ω_5 et C sont les plus proches. On les regroupe pour former le groupe : $D = \{\omega_5, C\}$. On a une nouvelle partition de Ω : $\mathcal{P}_4 = (\{\omega_6\}, D)$.

- On a

$$e(\omega_6, D) = \min(e(\omega_6, \omega_5), e(\omega_6, C)) = \min(2.30, 2.15) = 2.15.$$

Le tableau des écarts associé à $\mathcal{P}_4$ est

	ω_6	D
ω_6	0	**2.15**
D	2.15	0

Il ne nous reste plus que 2 éléments, ω_6 et D, nous les regroupons. Ce qui nous donne la partition $\mathcal{P}_5 = \{\omega_1, \ldots, \omega_6\} = \Omega$. Cela termine l'algorithme de CAH.

Au final,

– les éléments $\{\omega_2\}$ et $\{\omega_3\}$ ont été regroupés avec un écart de 0.70,

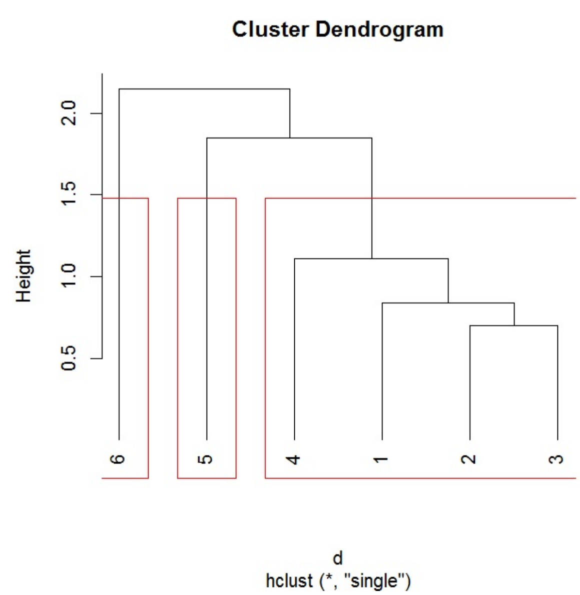
– les éléments $\{\omega_1\}$ et $A = \{\omega_2, \omega_3\}$ ont été regroupés avec un écart de 0.84,

– les éléments $\{\omega_4\}$ et $B = \{\omega_1, A\}$ont été regroupés avec un écart de 1.11,

– les éléments $\{\omega_5\}$ et $C = \{\omega_4, B\}$ ont été regroupés avec un écart de 1.85,

– les éléments $\{\omega_6\}$ et $D = \{\omega_5, C\}$ ont été regroupés avec un écart de 2.15.

Comme le plus grand saut se situe entre les éléments $\{\omega_5\}$ et $C = \{\omega_4, B\}$, on envisage de considérer les trois groupes : $\{\omega_1, \omega_2, \omega_3, \omega_4\}$, $\{\omega_5\}$ et $\{\omega_6\}$. Le dendrogramme associé est :



3. CAH avec la méthode du voisin le plus éloigné.

- Par le résultat de la question 1, le tableau des écarts associé à $\mathcal{P}_0 = (\{\omega_1\}, \ldots, \{\omega_6\})$ est

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
ω_1	0	0.84	0.95	1.49	1.85	2.56
ω_2	0.84	0	**0.70**	1.11	1.87	2.38
ω_3	0.95	0.70	0	1.35	2.05	2.15
ω_4	1.49	1.11	1.35	0	1.96	2.32
ω_5	1.85	1.87	2.05	1.96	0	2.30
ω_6	2.56	2.38	2.15	2.32	2.30	0

Les éléments (individus) ω_2 et ω_3 ont l'écart le plus petit ; ce sont les éléments les plus proches. On les regroupe pour former le groupe : $A = \{\omega_2, \omega_3\}$. On a une nouvelle partition de Ω : $\mathcal{P}_1 = (\{\omega_1\}, \{\omega_4\}, \{\omega_5\}, \{\omega_6\}, A)$.

- On a

$$e(\omega_1, A) = \max(e(\omega_1, \omega_2), e(\omega_1, \omega_3)) = \max(0.84, 0.95) = 0.95.$$

$$e(\omega_4, A) = \max(e(\omega_4, \omega_2), e(\omega_4, \omega_3)) = \max(1.11, 1.35) = 1.35.$$

$$e(\omega_5, A) = \max(e(\omega_5, \omega_2), e(\omega_5, \omega_3)) = \max(1.87, 2.05) = 2.05$$

et

$$e(\omega_6, A) = \max(e(\omega_6, \omega_2), e(\omega_6, \omega_3)) = \max(2.38, 2.15) = 2.38.$$

Le tableau des écarts associé à $\mathcal{P}_1$ est

	ω_1	ω_4	ω_5	ω_6	A
ω_1	0	1.49	1.85	2.56	**0.95**
ω_4	1.49	0	1.96	2.32	1.35
ω_5	1.85	1.96	0	2.30	2.05
ω_6	2.56	2.32	2.30	0	2.38
A	0.95	1.35	2.05	2.38	0

Les éléments ω_1 et A sont les plus proches. On les regroupe pour former le groupe : $B = \{\omega_1, A\}$. On a une nouvelle partition de Ω : $\mathcal{P}_2 = (\{\omega_4\}, \{\omega_5\}, \{\omega_6\}, B)$.

∘ On a

$$e(\omega_4, B) = \max(e(\omega_4, \omega_1), e(\omega_4, A)) = \max(1.49, 1.35) = 1.49.$$

$$e(\omega_5, B) = \max(e(\omega_5, \omega_1), e(\omega_5, A)) = \max(1.85, 2.05) = 2.05$$

et

$$e(\omega_6, B) = \max(e(\omega_6, \omega_1), e(\omega_6, A)) = \max(2.56, 2.38) = 2.56.$$

Le tableau des écarts associé à $\mathcal{P}_2$ est

	ω_4	ω_5	ω_6	B
ω_4	0	1.96	2.32	**1.49**
ω_5	1.96	0	2.30	2.05
ω_6	2.32	2.30	0	2.56
B	1.49	2.05	2.56	0

Les éléments ω_4 et B sont les plus proches. On les regroupe pour former le groupe : $C = \{\omega_4, B\}$. On a une nouvelle partition de Ω : $\mathcal{P}_3 = (\{\omega_5\}, \{\omega_6\}, C)$.

∘ On a

$$e(\omega_5, C) = \max(e(\omega_5, \omega_4), e(\omega_5, B)) = \max(1.96, 2.05) = 2.05$$

et

$$e(\omega_6, C) = \max(e(\omega_6, \omega_4), e(\omega_6, B)) = \max(2.32, 2.56) = 2.56.$$

Le tableau des écarts associé à $\mathcal{P}_3$ est

	ω_5	ω_6	C
ω_5	0	2.30	**2.05**
ω_6	2.30	0	2.56
C	2.05	2.56	0

Les éléments ω_5 et C sont les plus proches. On les regroupe pour former le groupe : $D = \{\omega_5, C\}$. On a une nouvelle partition de Ω : $\mathcal{P}_4 = (\{\omega_6\}, D)$.

◦ On a

$$e(\omega_6, D) = \max(e(\omega_6, \omega_5), e(\omega_6, C)) = \max(2.30, 2.56) = 2.56.$$

Le tableau des écarts associé à $\mathcal{P}_4$ est

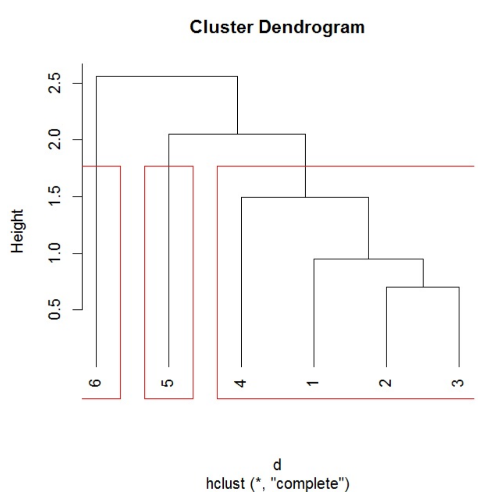
	ω_6	D
ω_6	0	**2.56**
D	2.56	0

Il ne nous reste plus que 2 éléments, ω_6 et D, nous les regroupons. Ce qui nous donne la partition $\mathcal{P}_5 = \{\omega_1, \ldots, \omega_6\} = \Omega$. Cela termine l'algorithme de CAH.

Au final,

- les éléments $\{\omega_2\}$ et $\{\omega_3\}$ ont été regroupés avec un écart de 0.70,
- les éléments $\{\omega_1\}$ et $A = \{\omega_2, \omega_3\}$ ont été regroupés avec un écart de 0.95,
- les éléments $\{\omega_4\}$ et $B = \{\omega_1, A\}$ont été regroupés avec un écart de 1.49,
- les éléments $\{\omega_5\}$ et $C = \{\omega_4, B\}$ ont été regroupés avec un écart de 2.05,
- les éléments $\{\omega_6\}$ et $D = \{\omega_5, C\}$ ont été regroupés avec un écart de 2.56.

Comme le plus grand saut se situe entre les éléments $\{\omega_5\}$ et $C = \{\omega_4, B\}$, on envisage de considérer les trois groupes : $\{\omega_1, \omega_2, \omega_3, \omega_4\}$, $\{\omega_5\}$ et $\{\omega_6\}$. Le dendrogramme associé est :



4. CAH avec la méthode de la distance moyenne.

  ◦ Par le résultat de la question 1, le tableau des écarts associé à $\mathcal{P}_0 = (\{\omega_1\}, \ldots, \{\omega_6\})$ est

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
ω_1	0	0.84	0.95	1.49	1.85	2.56
ω_2	0.84	0	**0.70**	1.11	1.87	2.38
ω_3	0.95	0.70	0	1.35	2.05	2.15
ω_4	1.49	1.11	1.35	0	1.96	2.32
ω_5	1.85	1.87	2.05	1.96	0	2.30
ω_6	2.56	2.38	2.15	2.32	2.30	0

  Les éléments (individus) ω_2 et ω_3 ont l'écart le plus petit : ce sont les éléments les plus proches. On les regroupe pour former le groupe : $A = \{\omega_2, \omega_3\}$. On a une nouvelle partition de Ω : $\mathcal{P}_1 = (\{\omega_1\}, \{\omega_4\}, \{\omega_5\}, \{\omega_6\}, A)$.

  ◦ On a

$$e(\omega_1, A) = \frac{1}{2}(e(\omega_1, \omega_2) + e(\omega_1, \omega_3)) = \frac{1}{2}(0.84 + 0.95) = 0.895.$$

$$e(\omega_4, A) = \frac{1}{2}(e(\omega_4, \omega_2) + e(\omega_4, \omega_3)) = \frac{1}{2}(1.11 + 1.35) = 1.23.$$

$$e(\omega_5, A) = \frac{1}{2}(e(\omega_5, \omega_2) + e(\omega_5, \omega_3)) = \frac{1}{2}(1.87 + 2.05) = 1.96$$

  et

$$e(\omega_6, A) = \frac{1}{2}(e(\omega_6, \omega_2) + e(\omega_6, \omega_3)) = \frac{1}{2}(2.38 + 2.15) = 2.265.$$

Le tableau des écarts associé à $\mathcal{P}_1$ est

	ω_1	ω_4	ω_5	ω_6	A
ω_1	0	1.49	1.85	2.56	**0.89**
ω_4	1.49	0	1.96	2.32	1.23
ω_5	1.85	1.96	0	2.30	1.96
ω_6	2.56	2.32	2.30	0	2.26
A	0.89	1.23	1.96	2.26	0

Les éléments ω_1 et A sont les plus proches. On les regroupe pour former le groupe : $B = \{\omega_1, A\}$. On a une nouvelle partition de Ω : $\mathcal{P}_2 = (\{\omega_4\}, \{\omega_5\}, \{\omega_6\}, B)$.

- On a

$$e(\omega_4, B) = \frac{1}{3}(e(\omega_4, \omega_1) + e(\omega_4, \omega_2) + e(\omega_4, \omega_3)) = \frac{1}{3}(1.49 + 1.11 + 1.35) = 1.32.$$

$$e(\omega_5, B) = \frac{1}{3}(e(\omega_5, \omega_1) + e(\omega_5, \omega_2) + e(\omega_5, \omega_3)) = \frac{1}{3}(1.85 + 1.87 + 2.05) = 1.92$$

et

$$e(\omega_6, B) = \frac{1}{3}(e(\omega_6, \omega_1) + e(\omega_6, \omega_2) + e(\omega_6, \omega_3))\frac{1}{3}(2.56 + 2.38 + 2.15) = 2.36.$$

Le tableau des écarts associé à $\mathcal{P}_2$ est

	ω_4	ω_5	ω_6	B
ω_4	0	1.96	2.32	**1.32**
ω_5	1.96	0	2.30	1.92
ω_6	2.32	2.30	0	2.36
B	1.32	1.92	2.36	0

Les éléments ω_4 et B sont les plus proches. On les regroupe pour former le groupe : $C = \{\omega_4, B\}$. On a une nouvelle partition de Ω : $\mathcal{P}_3 = (\{\omega_5\}, \{\omega_6\}, C)$.

- On a

$$\begin{aligned} e(\omega_5, C) &= \frac{1}{4}(e(\omega_5,\omega_1) + e(\omega_5,\omega_2) + e(\omega_5,\omega_3) + e(\omega_5,\omega_4)) \\ &= \frac{1}{4}(1.85 + 1.87 + 2.05 + 1.96) = 1.93 \end{aligned}$$

et

$$\begin{aligned} e(\omega_6, C) &= \frac{1}{4}(e(\omega_6,\omega_1) + e(\omega_6,\omega_2) + e(\omega_6,\omega_3) + e(\omega_6,\omega_4)) \\ &= \frac{1}{4}(2.56 + 2.38 + 2.15 + 2.32) = 2.35. \end{aligned}$$

Le tableau des écarts associé à $\mathcal{P}_3$ est

	ω_5	ω_6	C
ω_5	0	2.30	**1.93**
ω_6	2.30	0	2.35
C	1.93	2.35	0

Les éléments ω_5 et C sont les plus proches. On les regroupe pour former le groupe : $D = \{\omega_5, C\}$. On a une nouvelle partition de Ω : $\mathcal{P}_4 = (\{\omega_6\}, D)$.

- On a

$$\begin{aligned} e(\omega_6, D) &= \frac{1}{5}(e(\omega_6,\omega_1) + e(\omega_6,\omega_2) + e(\omega_6,\omega_3) + e(\omega_6,\omega_4) + e(\omega_6,\omega_5)) \\ &= \frac{1}{5}(2.56 + 2.38 + 2.15 + 2.32 + 2.3) = 2.34. \end{aligned}$$

Le tableau des écarts associé à $\mathcal{P}_4$ est

	ω_6	D
ω_6	0	**2.34**
D	2.34	0

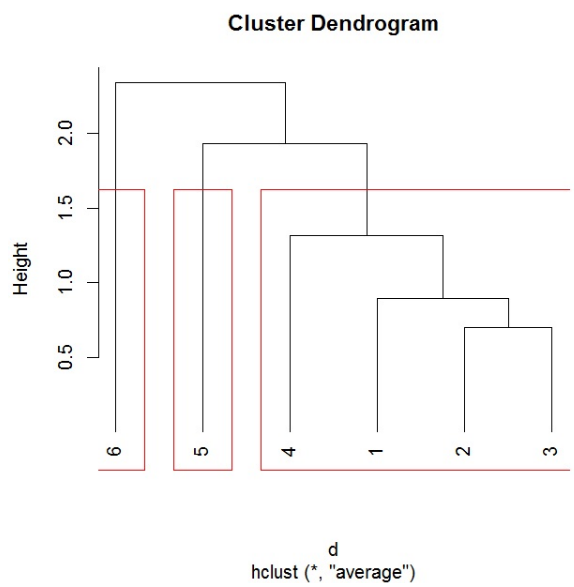
Il ne nous reste plus que 2 éléments, ω_6 et D, nous les regroupons. Ce qui nous donne la partition $\mathcal{P}_5 = \{\omega_1, \ldots, \omega_6\} = \Omega$. Cela termine l'algorithme de CAH.

Au final,

- les éléments $\{\omega_2\}$ et $\{\omega_3\}$ ont été regroupés avec un écart de 0.70,
- les éléments $\{\omega_1\}$ et $A = \{\omega_2, \omega_3\}$ ont été regroupés avec un écart de 0.89,
- les éléments $\{\omega_4\}$ et $B = \{\omega_1, A\}$ont été regroupés avec un écart de 1.32,
- les éléments $\{\omega_5\}$ et $C = \{\omega_4, B\}$ ont été regroupés avec un écart de 1.93,
- les éléments $\{\omega_6\}$ et $D = \{\omega_5, C\}$ ont été regroupés avec un écart de 2.34.

Comme le plus grand saut se situe entre les éléments $\{\omega_5\}$ et $C = \{\omega_4, B\}$, on envisage de considérer les trois groupes : $\{\omega_1, \omega_2, \omega_3, \omega_4\}$, $\{\omega_5\}$ et $\{\omega_6\}$.

Le dendrogramme associé est :



5. On propose les commandes R :

```
x = c(0 , 0.84, 0.95, 1.49, 1.85 , 2.56, 0.84, 0, 0.7 , 1.11, 1.87, 2.38,
0.95, 0.7, 0, 1.35, 2.05, 2.15, 1.49, 1.11, 1.35, 0, 1.96, 2.32, 1.85, 1.87,
2.05, 1.96, 0, 2.30, 2.56, 2.38, 2.15, 2.32, 2.30, 0)
A = matrix(x, ncol = 6)
d = as.dist(A)
cah1 = hclust(d, "single")
cah1$merge
cah1$height
plot(cah1, hang = -1)
rect.hclust(cah1, 3)
cah2 = hclust(d, "complete")
cah2$merge
cah2$height
plot(cah2, hang = -1)
rect.hclust(cah2, 3)
cah3 = hclust(d, "average")
cah3$merge
cah3$height
plot(cah3, hang = -1)
rect.hclust(cah3, 3)
```

Solution 5.

1. ∘ Le tableau des écarts associé à $\mathcal{P}_0 = (\{\omega_1\}, \ldots, \{\omega_4\})$ est

	ω_1	ω_2	ω_3	ω_4
ω_1	0	**2**	4	7
ω_2	**2**	0	4	5
ω_3	4	4	0	3
ω_4	7	5	3	0

Les éléments (individus) ω_1 et ω_2 ont l'écart le plus petit. On les regroupe pour former $A = \{\omega_1, \omega_2\}$. On a une nouvelle partition de $\Gamma = \{\omega_1, \ldots, \omega_4\}$:

$$\mathcal{P}_1 = (\{\omega_3\}, \{\omega_4\}, A).$$

∘ On a

$$e(\omega_3, A) = \frac{1}{2}(4 + 4) = 4$$

et

$$e(\omega_4, A) = \frac{1}{2}(7 + 5) = 6.$$

Le tableau des écarts associé à $\mathcal{P}_1$ est

	ω_3	ω_4	A
ω_3	0	**3**	4
ω_4	**3**	0	6
A	4	6	0

Les éléments (individus) ω_3 et ω_4 sont les plus proche. On les regroupe pour former $B = \{\omega_3, \omega_4\}$. On a une nouvelle partition de Γ :

$$\mathcal{P}_2 = (A, B).$$

- On a

$$e(B,A) = \frac{1}{2\times 2}(4+4+7+5) = 5.$$

Le tableau des écarts associé à $\mathcal{P}_2$ est

	A	B
A	0	**5**
B	**5**	0

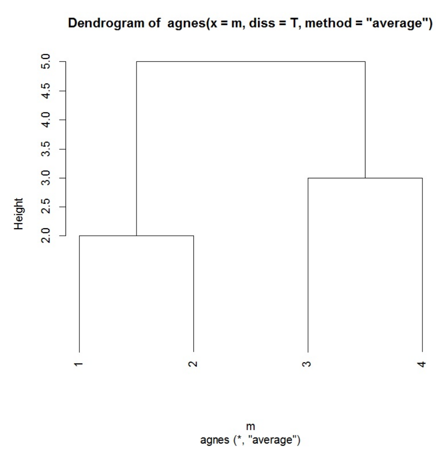
Il ne reste plus que 2 éléments, A et B, on les regroupe. Ce qui donne la partition $\mathcal{P}_3 = \{\omega_1,\ldots,\omega_4\} = \Gamma$. Cela termine l'algorithme de CAH. Au final,

– les éléments $\{\omega_1\}$ et $\{\omega_2\}$ ont été regroupés avec un écart de 2,

– les éléments $\{\omega_3\}$ et $\{\omega_4\}$ ont été regroupés avec un écart de 3,

– les éléments $A = \{\omega_1,\omega_2\}$ et $B = \{\omega_3,\omega_4\}$ ont été regroupés avec un écart de 5.

On peut tracer le dendrogramme :



Les commandes R associées sont :

```
x = c(0, 2, 4, 7, 2, 0, 4, 5, 4, 4, 0, 3, 7, 5, 3, 0)
m = matrix(x, ncol = 4, nrow = 4)
m = as.dist(m)
library(cluster)
ag = agnes(m, method = "average"
pltree(ag, hang = -1)
```

2. *Rappel* : on appelle coefficient d'agglomération le réel :

$$AC = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{e(\omega_i, A_i)}{e(Q, R)}\right),$$

où

- pour tout $i \in \{1, \ldots, n\}$, A_i désigne le premier élément avec lequel ω_i a été regroupé,
- Q et R désignent les deux derniers groupes rassemblés à l'étape finale de l'algorithme.

Ainsi, par le résultat de la question 1, on a

$$\begin{aligned} AC &= \frac{1}{4}\left(\left(1 - \frac{e(\omega_1, \omega_2)}{e(A, B)}\right) + \left(1 - \frac{e(\omega_2, \omega_1)}{e(A, B)}\right) + \left(1 - \frac{e(\omega_3, \omega_4)}{e(A, B)}\right) + \left(1 - \frac{e(\omega_4, \omega_3)}{e(A, B)}\right)\right) \\ &= \frac{1}{4}\left(\left(1 - \frac{2}{5}\right) + \left(1 - \frac{2}{5}\right) + \left(1 - \frac{3}{5}\right) + \left(1 - \frac{3}{5}\right)\right) = 0.5. \end{aligned}$$

Par conséquent, pour la classification obtenue, on a donc une structure de groupes moyenne. Les commandes R associées sont :

```
ag$ac
```

Solution 6. On calcule le coefficient d'agglomération associée à la main et avec la commande `agnes`. On obtient alors le même résultat à l'arrondi près.

Solution 7.

1. Le tableau disjonctif complet est :

	H	F	O	N	$P0$	$N0$	$P1$	$N1$	$P2$	$N2$	$P3$	$N3$	$P4$	$N4$
ω_1	1	0	1	0	1	0	1	0	1	0	0	1	1	0
ω_2	0	1	1	0	0	1	1	0	0	1	1	0	1	0
ω_3	0	1	1	0	1	0	0	1	0	1	0	1	0	1
ω_4	1	0	1	0	0	1	1	0	0	1	1	0	0	1
ω_5	1	0	1	0	0	1	1	0	1	0	1	0	0	1
ω_6	1	0	1	0	1	0	1	0	0	1	0	1	0	1

2. On propose les commandes R :

```
m = matrix(c(1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0,
1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1,
0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1), ncol = 14, nrow = 6)

library(arules)
d = dissimilarity(m, method = "jaccard")

library(cluster)
ag = agnes(d, method = "average")
plot(ag, which = 2, hang = -1)
```

L'enjeu des commandes R présentées est de mettre en œuvre l'algorithme CAH avec la dissimilarité de Jaccard et la méthode de la distance moyenne sur 6 individus et 7 caractères qualitatifs qui totalisent 14 modalités. Pour ce faire, on utilise :

◦ la commande dissimilarity du package arules,

◦ la commande agnes du package cluster.

3. Vu le dendrogramme, les 2 groupes qui se distinguent le plus sont : $A = \{\omega_1, \omega_3, \omega_6\}$ et $B = \{\omega_2, \omega_4, \omega_5\}$, soit :

$$A = \{\text{Thomas, Emilie, Mathias}\} \qquad B = \{\text{Fatiha, Robert, Igor}\}.$$

Solution 8.

1. Après comptage, par la définition de l'indice de similarité de Dice, le tableau des indices de similarités est :

	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_5
ℓ_1	1	$\frac{8}{12}$	$\frac{10}{12}$	$\frac{8}{11}$	$\frac{6}{11}$
ℓ_2	$\frac{8}{12}$	1	$\frac{8}{12}$	$\frac{8}{11}$	$\frac{10}{11}$
ℓ_3	$\frac{10}{12}$	$\frac{8}{12}$	1	$\frac{6}{11}$	$\frac{6}{11}$
ℓ_4	$\frac{8}{11}$	$\frac{8}{11}$	$\frac{6}{11}$	1	$\frac{8}{10}$
ℓ_5	$\frac{6}{11}$	$\frac{10}{11}$	$\frac{6}{11}$	$\frac{8}{10}$	1

Soit encore, avec des valeurs décimales (en gardant 3 décimales) :

	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_5
ℓ_1	1.000	0.667	0.833	0.727	0.545
ℓ_2	0.667	1.000	0.667	0.727	0.909
ℓ_3	0.833	0.667	1.000	0.545	0.545
ℓ_4	0.727	0.727	0.545	1.000	0.800
ℓ_5	0.545	0.909	0.545	0.800	1.000

Le tableau des dissimilarités, avec des valeurs décimales, est :

	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_5
ℓ_1	0.000	0.333	0.167	0.273	0.455
ℓ_2	0.333	0.000	0.333	0.273	0.091
ℓ_3	0.167	0.333	0.000	0.455	0.455
ℓ_4	0.273	0.273	0.455	0.000	0.200
ℓ_5	0.455	0.091	0.455	0.200	0.000

2. CAH avec la méthode du voisin le plus éloigné.

- Par le résultat de la question 1, le tableau des écarts associé à $\mathcal{P}_0 = (\{\ell_1\}, \ldots, \{\ell_5\})$ est

	ℓ_1	ℓ_2	ℓ_3	ℓ_4	ℓ_5
ℓ_1	0.000	0.333	0.167	0.273	0.455
ℓ_2	0.333	0.000	0.333	0.273	**0.091**
ℓ_3	0.167	0.333	0.000	0.455	0.455
ℓ_4	0.273	0.273	0.455	0.000	0.200
ℓ_5	0.455	0.091	0.455	0.200	0.000

Les éléments ℓ_2 et ℓ_5 ont l'écart le plus petit : ce sont les éléments les plus proches. On les regroupe pour former le groupe : $A = \{\ell_2, \ell_5\}$. On a une nouvelle partition de Ω : $\mathcal{P}_1 = (\{\ell_1\}, \{\ell_3\}, \{\ell_4\}, A)$.

- On a

$$e(\ell_1, A) = \max(e(\ell_1, \ell_2), e(\ell_1, \ell_5)) = \max(0.333, 0.455) = 0.455,$$

$$e(\ell_3, A) = \max(e(\ell_3, \ell_2), e(\ell_3, \ell_5)) = \max(0.333, 0.455) = 0.455$$

et

$$e(\ell_4, A) = \max(e(\ell_4, \ell_2), e(\ell_4, \ell_5)) = \max(0.273, 0.200) = 0.273.$$

Le tableau des écarts associé à $\mathcal{P}_1$ est

	ℓ_1	ℓ_3	ℓ_4	A
ℓ_1	0.000	**0.167**	0.273	0.455
ℓ_3	0.167	0.000	0.455	0.455
ℓ_4	0.273	0.455	0.000	0.273
ℓ_5	0.455	0.455	0.273	0.000

Les éléments ℓ_1 et ℓ_3 sont les plus proches. On les regroupe pour former le groupe : $B = \{\ell_1, \ell_3\}$. On a une nouvelle partition de Ω : $\mathcal{P}_2 = (\{\ell_4\}, A, B)$.

∘ On a

$$e(\ell_4, B) = \max(e(\ell_4, \ell_1), e(\ell_4, \ell_3)) = \max(0.273, 0.455) = 0.455$$

et

$$e(A, B) = \max(e(\ell_1, A), e(\ell_3, A)) = \max(0.455, 0.455) = 0.455.$$

Le tableau des écarts associé à $\mathcal{P}_2$ est

	ℓ_4	A	B
ℓ_4	0.000	**0.273**	0.455
A	0.273	0.000	0.455
B	0.455	0.455	0.000

Les éléments ℓ_4 et A ont l'écart le plus petit : ce sont les éléments les plus proches. On les regroupe pour former le groupe : $C = \{\ell_4, A\}$. On a une nouvelle partition de Ω : $\mathcal{P}_3 = (B, C)$.

∘ On a

$$e(B, C) = \max(e(\ell_4, B), e(A, B)) = \max(0.455, 0.455) = 0.455.$$

Le tableau des écarts associé à $\mathcal{P}_3$ est

	B	C
B	0	**0.455**
C	0.455	0

Il ne nous reste plus que 2 éléments, B et C, nous les regroupons. Ce qui nous donne la partition $\mathcal{P}_4 = \{\ell_1, \ldots, \ell_5\} = \Omega$. Cela termine l'algorithme de CAH.

Au final,

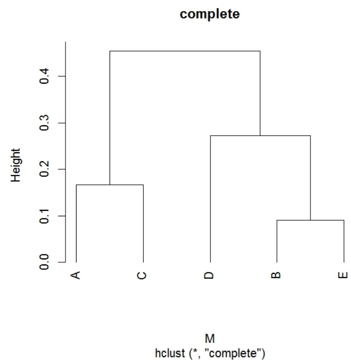
– les éléments $\{\ell_2\}$ et $\{\ell_5\}$ ont été regroupés avec un écart de 0.091,

– les éléments $\{\ell_1\}$ et $\{\ell_3\}$ ont été regroupés avec un écart de 0.167,

– les éléments $\{\ell_4\}$ et $A = \{\ell_2, \ell_5\}$ ont été regroupés avec un écart de 0.273,

– les éléments $B = \{\omega_1, \ell_3\}$ et $C = \{\ell_4, A\}$ ont été regroupés avec un écart de 0.455.

Comme le plus grand saut se situe entre les éléments B et C, on envisage de considérer les 2 groupes : $\{\omega_1, \ell_3\}$ et $\{\ell_2, \ell_4, \ell_5\}$. Le dendrogramme associé est :



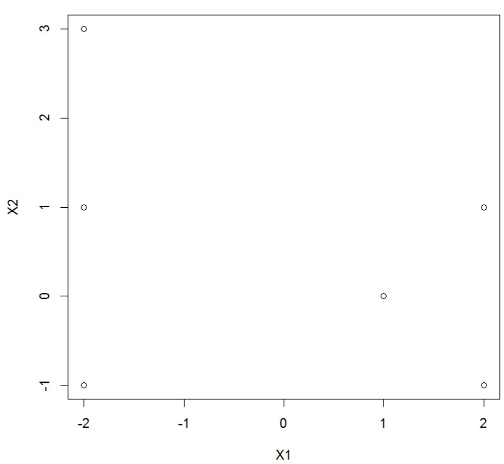
3. Par le résultat de la question 2, on a

$$\begin{aligned} AC &= \frac{1}{5}\left(\left(1-\frac{0.167}{0.455}\right)+\left(1-\frac{0.091}{0.455}\right)+\left(1-\frac{0.167}{0.455}\right)+\left(1-\frac{0.273}{0.455}\right)+\left(1-\frac{0.091}{0.455}\right)\right) \\ &= 0.6531868. \end{aligned}$$

Par conséquent, pour la classification obtenue, on a donc une structure de groupes moyenne, voire bonne.

Solution 9.

1. On affiche le nuage de points :



Les commandes R associées sont :

```
x = c(-2, -2, -2, 2, 2, 1, 3, 1, -1, -1, 1, 0)
m = matrix(x, ncol = 2, nrow = 6)
plot(m, xlab = "X1", ylab = "X2")
```

2. ∘ On considère les centres initiaux $c_1^0 = \omega_1$ de coordonnées $(-2, 3)$ et $c_2^0 = \omega_2$ de coordonnées $(-2, 1)$. Le tableau des distances entre les individus et ces centres est :

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
c_1^0	0	2	4	5.65	4.47	4.24
c_2^0	2	0	2	4.47	4	3.16

D'où les deux groupes :

$$A = \{\omega_1\}, \qquad B = \{\omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}.$$

∘ On considère deux nouveaux centres, c_1^1 et c_2^1, lesquels sont les centres de gravité des deux groupes A et B. Donc c_1^1 a pour coordonnées $(-2, 3)$ et c_2^1 a pour coordonnées :

$$\left(\frac{-2-2+2+2+1}{5}, \frac{1-1-1+1}{5} \right) = (0.2, 0).$$

Le tableau des distances entre les individus et ces centres est :

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
c_1^1	0	2	4	5.65	4.47	4.24
c_2^1	3.72	2.41	2.41	2.05	2.05	0.80

D'où les deux groupes :

$$A = \{\omega_1, \omega_2\}, \qquad B = \{\omega_3, \omega_4, \omega_5, \omega_6\}.$$

∘ On considère deux nouveaux centres, c_1^2 et c_2^2, lesquels sont les centres de gravité des deux groupes A et B. Donc c_1^2 a pour coordonnées $\left(\frac{-2-2}{2}, \frac{3+1}{2}\right) = (-2, 2)$ et c_2^2 a pour coordonnées $\left(\frac{-2+2+2+1}{4}, \frac{-1-1+1+0}{4}\right) = (0.75, -0.25)$.

Le tableau des distances entre les individus et ces centres est :

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
c_1^2	1	1	3	5	4.12	3.60
c_2^2	4.25	3.02	2.85	1.45	1.76	0.35

D'où les deux groupes :

$$A = \{\omega_1, \omega_2\}, \qquad B = \{\omega_3, \omega_4, \omega_5, \omega_6\}.$$

On retrouve les mêmes regroupements qu'à l'étape précédente, l'algorithme s'arrête.

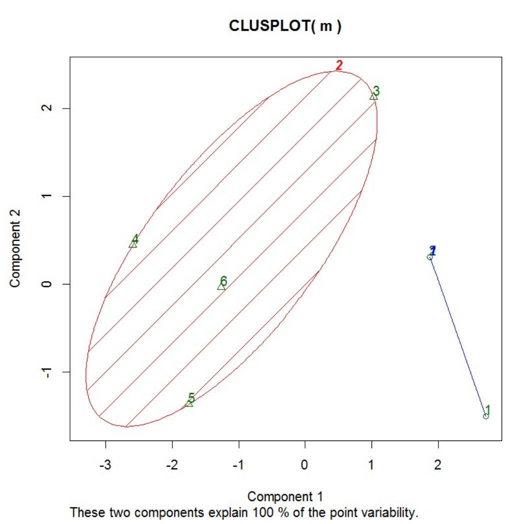
On propose donc la partition finale $\mathcal{P}(\Gamma) = (A, B)$, avec

$$A = \{\omega_1, \omega_2\}, \qquad B = \{\omega_3, \omega_4, \omega_5, \omega_6\}.$$

Avec R, on propose :

```
library(stats)
clus = kmeans(m, centers = m[c(1, 2), ], algorithm = "Lloyd")
clus$cluster
clus$centers
clusplot(m, clus$cluster, color = T, shade = T, labels = 2, lines = 0)
```

Cela renvoie les groupes formés, les centres de gravité associés et le graphique :



3. ∘ On considère les centres initiaux $c_1^0 = \omega_4$ de coordonnées $(2, -1)$ et $c_2^0 = \omega_6$ de coordonnées $(1, 0)$. Le tableau des distances entre les individus et ces centres est :

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
c_1^0	5.65	4.47	4	0	2	1.41
c_2^0	4.24	3.16	3.16	1.41	1.41	0

D'où les deux groupes :

$$A = \{\omega_4\}, \qquad B = \{\omega_1, \omega_2, \omega_3, \omega_5, \omega_6\}.$$

∘ On considère deux nouveaux centres, c_1^1 et c_2^1, lesquels sont les centres de gravité des deux groupes A et B. Donc c_1^1 a pour coordonnées $(2, -1)$ et c_2^1 a pour coordonnées :

$$\left(\frac{-2-2-2+2+1}{5}, \frac{3+1-1+1+0}{5} \right) = (-0.6, 0.8).$$

Le tableau des distances entre les individus et ces centres est :

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
c_1^1	5.65	4.47	4	0	2	1.41
c_2^1	2.6	1.41	2.28	3.16	2.60	1.78

D'où les deux groupes :

$$A = \{\omega_4, \omega_5, \omega_6\}, \qquad B = \{\omega_1, \omega_2, \omega_3\}.$$

∘ On considère deux nouveaux centres, c_1^2 et c_2^2, lesquels sont les centres de gravité des deux groupes A et B. Donc c_1^2 a pour coordonnées $\left(\frac{2+2+1}{3}, \frac{-1+1+0}{3}\right) = (1.66, 0)$ et c_2^2 a pour coordonnées $\left(\frac{-2-2-2}{3}, \frac{3+1-1}{3}\right) = (-2, 1)$. Le tableau des distances entre les individus et ces centres est :

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
c_1^2	4.73	3.79	3.79	1.05	1.05	0.66
c_2^2	2	0	2	4.47	4	3.16

D'où les deux groupes :

$$A = \{\omega_4, \omega_5, \omega_6\}, \qquad B = \{\omega_1, \omega_2, \omega_3\}.$$

On retrouve les mêmes regroupements qu'à l'étape précédente, l'algorithme s'arrête.

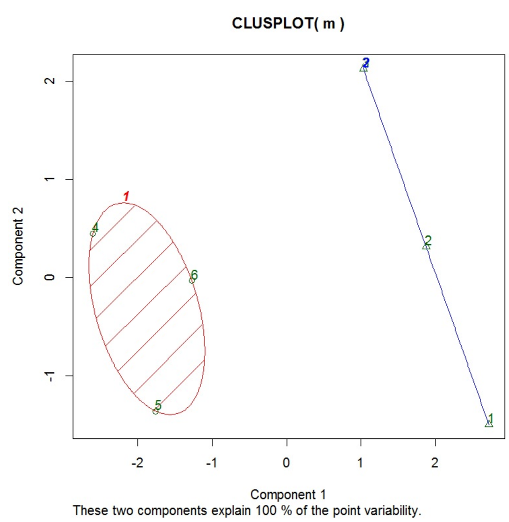
On propose donc la partition finale $\mathcal{P}(\Gamma) = (A, B)$, avec

$$A = \{\omega_4, \omega_5, \omega_6\}, \qquad B = \{\omega_1, \omega_2, \omega_3\}.$$

Avec R, on propose :

```
library(stats)
clus = kmeans(m, centers = m[c(4, 6), ], algorithm = "Lloyd")
clus$cluster
clus$centers
clusplot(m, clus$cluster, color = T, shade = T, labels = 2, lines = 0)
```

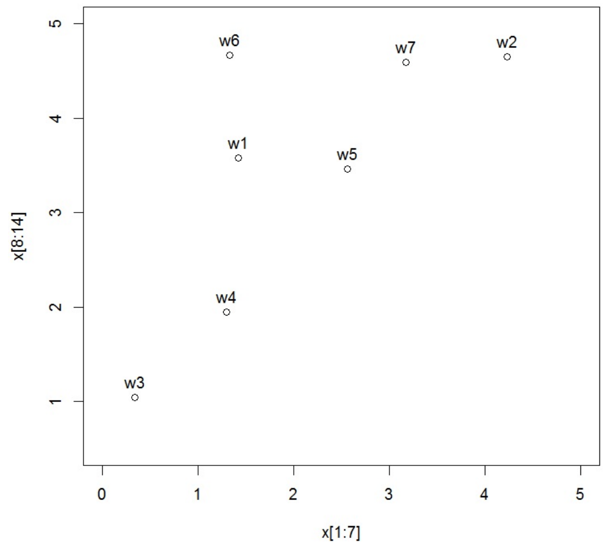
Cela renvoie les groupes formés, les centres de gravité associés et le graphique :



On obtient une classification différente de la question précédente. Cela illustre la sensibilité de l'algorithme des centres mobiles aux choix des centres initiaux.

Solution 10.

1. On trace le nuage de points :



- On considère les centres initiaux c_1^0 de coordonnées $(1, 1)$ et c_2^0 de coordonnées $(3, 3)$. Le tableau des distances entre les individus et ces centres est

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7
c_1^0	2.61	4.87	0.66	1.00	2.91	3.68	4.19
c_2^0	1.68	2.06	3.30	2.00	0.64	2.36	1.60

Exemple de calcul : $d(\omega_1, c_1^0) = \sqrt{(1.42 - 1)^2 + (3.58 - 1)^2} = 2.61$.

D'où les deux groupes :

$$A = \{\omega_1, \omega_2, \omega_5, \omega_6, \omega_7\}, \qquad B = \{\omega_3, \omega_4\}.$$

- On considère deux nouveaux centres, c_1^1 et c_2^1, lesquels sont les centres de gravité des deux groupes A et B. Donc c_1^1 a pour coordonnées

$$\left(\frac{1.42 + 4.23 + 2.56 + 1.33 + 3.17}{5}, \frac{3.58 + 4.65 + 3.46 + 4.67 + 4.59}{5}\right) = (2.54, 4.19)$$

et c_2^1 a pour coordonnées

$$\left(\frac{0.34 + 1.30}{2}, \frac{1.04 + 1.95}{2}\right) = (0.82, 1.49).$$

Le tableau des distances entre les individus et ces centres est

	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7
c_1^1	1.28	1.75	3.84	2.56	0.73	1.30	0.75
c_2^1	2.17	4.65	0.66	0.66	2.63	3.22	3.89

D'où les deux groupes :

$$A = \{\omega_1, \omega_2, \omega_3, \omega_5\}, \qquad B = \{\omega_4, \omega_6\}.$$

On retrouve la même classification que l'étape précédente, on arrête l'algorithme.

2. On propose les commandes R :

```
x = c(1.42, 4.23, 0.34, 1.30, 2.56, 1.33, 3.17, 3.58, 4.65, 1.04, 1.95,
3.46, 4.67, 4.59)
individus = c("w1", "w2", "w3", "w4", "w5", "w6", "w7")
plot(x[1:7], x[8:14], xlim = c(1, 5), ylim = c(1.8, 5))
text(x[1:7], x[8:14], individus, pos = 3)
m = matrix(x, ncol = 2, nrow = 7)
clus = kmeans(m, centers = rbind(c(-1, -1), c(2, 3)), algorithm = "Lloyd")
clus$cluster
clus$centers
plot(m, col = clus$cluster, pch = 1, lwd = 3, xlab = "X1", ylab = "X2")
points(clus$centers, col = 1:2, pch = 9, lwd = 3)
```

Solution 11. Soit ω_* l'individu dont les coordonnées sont $(4, 4)$. Les distances entre cet individu et les autres sont :

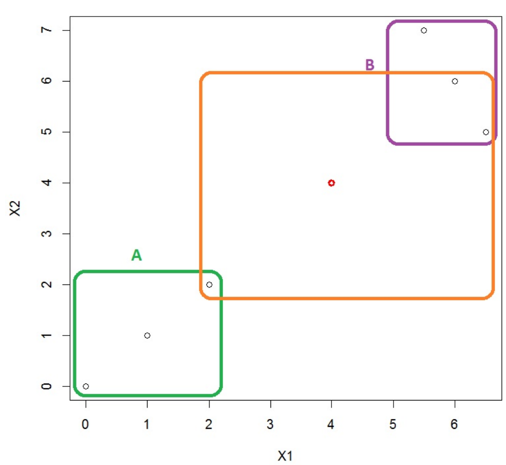
	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6
Groupe	A	A	A	B	B	B
ω_*	5.65	4.24	2.82	2.82	3.35	2.69

Les $k = 3$ individus les plus de proche de ω_* sont : ω_3, ω_4 et ω_6. Le premier appartient au groupe A et les deux autres au groupe B. Par conséquent, la méthode des k plus proche voisin nous dit que ω_* a plus de chance d'appartenir à B.

On peut vérifier cela avec des commandes R :

```
x = c(0, 1, 2, 6, 5.5, 6.5, 0, 1, 2, 6, 7, 5)
m = matrix(x, ncol = 2, nrow = 6)
cl = factor(c(rep("A", 3), rep("B", 3)))
library(class)
knn(m, c(4, 4), cl, k = 3)
```

Cela renvoie B. On s'en rend compte graphiquement :



Solution 12. On divise la population en 2 groupes : $G1$ ={hommes} et $G2$ ={femmes}. Soit Y le caractère indiquant le sexe d'un individu qui rentre dans le magasin avec $Y = G1$ si c'est un homme et $Y = G2$ si c'est une femme. On veut calculer : $\mathbb{P}(\{Y = G1\}|\{X = 1.60\})$. Par l'énoncé, on sait que

– la densité de X sachant que $Y = G2$ est

$$f_2(x) = \frac{1}{\sqrt{2\pi \times 0.16^2}} \exp\left(-\frac{1}{2 \times 0.16^2}(x-1.65)^2\right), \qquad x \in \mathbb{R},$$

– la densité de X sachant que $Y = G1$ est

$$f_1(x) = \frac{1}{\sqrt{2\pi \times 0.15^2}} \exp\left(-\frac{1}{2 \times 0.15^2}(x-1.75)^2\right), \qquad x \in \mathbb{R}.$$

De plus, on a $\mathbb{P}(Y = G1) = 0.7$ et $\mathbb{P}(Y = G2) = 0.3$. La règle de Bayes donne :

$$\begin{aligned}\mathbb{P}(\{Y = G1\}|\{X = 1.60\}) &= \frac{\mathbb{P}(Y = G1)f_1(1.60)}{\mathbb{P}(Y = G1)f_1(1.60) + \mathbb{P}(Y = G2)f_2(1.60)} \\ &= \frac{0.7 \times 1.6131}{0.7 \times 1.6131 + 0.3 \times 2.3745} = 0.6131.\end{aligned}$$

Il y a donc plus de chances que l'individu soit un homme qu'une femme.

Solution 13. Il y a 50 points et on en compte 3 mal classés sur le graphique obtenu (cela peut différer d'une expérience à l'autre, les données étant générées aléatoirement). Par conséquent, la proportion de points mal classés est $3/50 = 0.06$.

Solution 14.

1. Comme on veut expliquer un caractère binaire $Y \in \{0, 1\}$ en fonction de 3 caractère qualitatifs, le modèle de régression logistique est approprié. Ainsi, on considère

$$p(x) = \mathbb{P}(\{Y = 1\}|\{(X_1, X_2, X_3) = x\}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)},$$

   où $x = (x_1, x_2, x_3)$ et $\beta_0, \beta_1, \beta_2, \beta_3$ désignent des coefficients inconnus.

2. On obtient les estimations des paramètres inconnus $\beta_0, \beta_1, \beta_2, \beta_3$ avec la méthode du maximum de vraisemblance.

Solution 15.

1. Le modèle considéré est le modèle de régression logistique :

$$p(x) = \mathbb{P}(\{Y = 1\}|\{X_1 = x\}) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)},$$

   où β_0 et β_1 désignent des coefficients inconnus.

   Les estimations sont données par la méthode du maximum de vraisemblance. La sortie R nous informe que la probabilité qu'un patient ayant eu pour dosage X1 = 1.25 bouge est de 0.37994 et, par conséquent, la probabilité qu'il ne bouge pas est $1 - 0.37994 = 0.62$. Il a donc plus de chance de ne pas bouger que de ne pas bouger.

2. Ce graphique représente la réalisation de l'estimateur $\hat{p}(x)$ de $p(x)$ pour tout $x \in [0, 2.5]$.

3. Cette quantité représente le taux d'erreur. Celui-ci étant proche de 0, le modèle a une bonne qualité prédictive.

Index