RAPPORT DE TER

présenté à

L'UNIVERSITÉ LILLE 1

par

Salim Yataghane et Chouaib Brakni

Classification ascendante hiérarchique et clustering.

Sous la direction de Azzouz Dermoune

Table des matières

Chapitre 1

Introduction

La classification est utilisée dans des domaines très divers comme les statistiques, la biologie, la psychologie ou d'autres sciences sociales, dans le but de former des groupes ou des classes à partir des masses de données. Elle joue un rôle important, notamment dans l'extraction de l'information cachée dans des grandes bases de données.

On distingue deux types de classifications. Le premier, que l'on appelle partitionnement, nécessite de connaître au préalable le nombre de classes. A titre d'exemple, en médecine, les patients cardiaques peuvent être classés selon trois grandes catégories par rapport à leur capacités physiques. Pour réaliser ce type de classification nous disposons de nombreuses méthodes, comme par exemple la méthode k-means. Le second consiste, lui, à classifier les individus selon leurs liens hiérarchiques. On peut alors, après classification, en déduire le nombre de classes. Donnons un exemple : si nous souhaitons classer les animaux, notre premier choix est de distinguer les vertébrés et les invertébrés. Ensuite, dans la catégorie des vertébrés, on peut distinguer les animaux avec poils, plumes ou à peau nue, ce qui nous donnera la classe des mammifères, celle des oiseaux et celle des animaux marins. On peut alors répéter le processus jusqu'à ce que tous les animaux soit classifiés. Nous disposons de la méthode CAH pour réaliser ce type de classification.

Dans ce mémoire nous avons d'une part étudié la classification ascendante hiérarchique (CAH) en utilisant différentes mesures de proximité (single linkage, complete linkage, Ward linkage, ...) et d'autre part nous avons étudié l'affectation des individus supplémentaires dans des classes déjà obtenues par la classification. L'une des méthode usuelle utilisée pour résoudre ce type de problème est basée sur la distance géométrique de l'individu au centre de la classe. Brigitte Roux et Frédérik Cassor [**1**] ont montré que cette méthode a ses limites. En effet elle peut créer des individus aberrants dans les classes, ce qui fausse considérablement le résultat. Pour éviter ce genre de problème, ils ont proposé une autre méthode de classification de ces individus supplémentaires, qui repose essentiellement sur la distance de Mahalanobis et la recherche d'un seuil qui minimise le nombre des individus mal classés. Nous avons reprogrammé leur algorithme et nous l'avons testé sur plusieurs exemples.

Chapitre 2

Classification et Affectation

DÉFINITION 2.1 (Classification). *La classification est une action qui consiste à partager un groupe d'individus en classes, de sorte que tout individu appartienne à une et une seule classe (i.e l'ensemble des classes forme une partition du groupe).*

DÉFINITION 2.2 (Classe). *Une classe est un ensemble d'individus possédant des traits de caractères communs.*

1. Quelques domaines d'applications

Domaine médical :
On classifie les individus dans le but de déterminer des groupes de patients, susceptibles ou non d'être soumis à des protocoles thérapeutiques. Chaque groupe contient tous les patients qui réagissent identiquement.

Marketing :
Dans ce domaine, la classification est appelée plus fréquemment la segmentation.
On recherche des différents profils de clients constituant la clientèle. Après avoir détecté les classes de la clientèle, l'entreprise peut adapter sa stratégie marketing à chaque profil.

Commentaire
Les classes sont considérées comme des groupes contenant des objets ou individus qui sont similaires les uns les autres. Ceux qui n'appartiennent pas à la même classe ne le sont pas (dissimilaires). D'où la problématique suivante : quand est-ce que deux objets (ou individus) sont similaires, autrement dit, sur quelle mesure de proximité peut-on se baser pour conclure.

2. Mesures de proximités

2.1. Définition des mesures de proximité.

Une mesure de proximité est une généralisation de la similarité et de la dissimilarité.

Une dissimilarité ou une distance D sur un ensemble de données $X = (x_1, ..., x_n)$ tel que $x_i \in R^d$ satisfait les conditions suivantes :

• Symétrie,

$$D(x_i, x_j) = D(x_j, x_i),\ (i, j) \in \{1, ..., n\}^2$$

• Positivité,

$$D(x_i, x_j) \geq 0,\ \textit{pour tout } x_i \textit{ et } x_j$$

Si D vérifie aussi les conditions suivantes :
• Inégalité triangulaire,

$$D(x_i, x_j) \leq D(x_i, x_k) + D(x_k, x_j)\ \textit{pour tout } x_i \textit{ et } x_j$$

• Réflexivité,

$$D(x_i, x_j) = 0\ \textit{si et seulement si } x_i = x_j$$

Alors D est une métrique.
Si seulement l'inégalité triangulaire qui n'est pas vérifiée , alors D est une semi-métrique.

Egalement une mesure de similarité S satisfait les conditions suivantes :
• Symétrie,

$$S(x_i, x_j) = S(x_j, x_i)$$

• Positivité,

$$0 \leq S(x_i, x_j) \leq 1\ \textit{pour tout } x_i \textit{ et } x_j$$

Si S satisfait aussi les conditions suivantes :

- $S(x_i, x_j)S(x_j, x_k) \leq [S(x_i, x_j) + S(x_i, x_k)]S(x_i, x_k)$
- $S(x_i, x_j) = 1\ \textit{si et seulement si } x_i = x_j$

Alors S est appelée une métrique de similarité.

Pour un ensemble de données $X \in \mathbb{R}^{N*d}$ de N individus (objets), on peut définir alors une matrice symétrique de taille $N * N$, qui contient les mesures de proximité des individus i et j, pour $(i, j = 1, ..., N)$.

2.2. Mesures de proximité pour des variables continues.

• La distance la plus utilisée est la distance euclidienne définie comme suit :

$$D(x_i, x_j) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^2\right)^{1/2}$$

avec $x_i \in \mathbb{R}^d$. Une étude plus approfondie montre que la distance euclidienne a tendance à former des classes hypersphériques (Duda et al., 2001).

Mais si les caractéristiques sont mesurées avec des unités assez différentes, les caractéristiques qui ont des variances très grande dominent les autres caractéristiques.
Il existe plusieurs méthodes pour résoudre ce problème qui procèdent par la normalisation des données. A titre d'exemple :

$$x^*_{il} = \frac{x_{il} - m_l}{s_l}$$

avec

$$m_l = \frac{1}{N}\sum_{i=1}^{N} x_{il} \quad et\ s_l = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_{il} - m_l)^2}$$

Une autre méthode de normalisation basée sur le maximum et le minimum des données.

$$x^*_{il} = \frac{x_{il} - min(x_{il})}{max(x_{il}) - min(x_{il})} \quad tel\ que\ x^*_{il} \in [0,1]$$

La distance euclidienne est un cas particulier d'une famille de distances appelée la distance de Minkowski ou la norm L_p définie comme suit :

$$D(x_i, x_j) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^p\right)^{1/p}$$

Dans le cas où $p = 2$, on obtient la distance euclidienne. Si $p = 1$ ou $p = \infty$, on obtient aussi des cas particuliers de Minkowski, appelées :
La norm L_1

$$D(x_i, x_j) = \sum_{l=1}^{d} |x_{il} - x_{jl}|$$

Et la norm L_∞

$$D(x_i, x_j) = \max_{1 \le l \le d} |x_{il} - x_{jl}|.$$

•La distance de Mahalanobis
Le carré de la distance de Mahalanobis est aussi une métrique et définie comme suit :

$$D(x_i, x_j) = (x_i - x_j)^\top S^{-1}(x_i - x_j)$$

Où S est la matrice de covariance définie comme $S = E[(X - \mu)(X - \mu)^\top]$, avec μ est le vecteur moyen. La distance de Mahalanobis a tendance à former

des classes hyperellipsoides (Jain and Dubes, 1988 ; Mao and Jain, 1996).

• Le coefficient de corrélation
Une mesure de distance peut être dérivée du coefficient de corrélation ρ tel que :

$$\rho_{ij} = \frac{\sum_{l=1}^{d}(x_{il} - \bar{x_i})(x_{jl} - \bar{x_j})}{\sqrt{\sum_{l=1}^{d}(x_{il} - \bar{x_i})^2 \sum_{l=1}^{d}(x_{jl} - \bar{x_j})^2}}$$

Où $\bar{x_i} = \frac{1}{d}\sum_{l=1}^{d} x_{il}$.
Notons que $\rho_{ij} \in [-1, 1]$, avec -1 et 1 indique une forte corrélation négative et positive respectivement.
Et la mesure de distance est définie comme :

$$D(x_i, x_j) = (1 - \rho_{ij})/2$$

Dans ce cas, $D \in [0, 1]$.

• Similarité cosinus
Enfin, nous considérons l'application directe de mesure de similarité pour comparer deux individus basée sur le cosinus, définie comme suit :

$$S(x_i, x_j) = cos(\alpha) = \frac{x_i^\top x_j}{\|x_i\|\|x_j\|}$$

Comme la valeur $\cos(\alpha) \in [-1, 1]$, la valeur -1 indique que les individus sont opposés, 0 non similaires, 1 similaires et la valeurs intermédiaires permettent d'évaluer le degré de similarité.

2.3. Mesures de proximité et leurs applications.

Mesure	métrique	Exemple et/ou applications
Distance de Minkowski	oui	fuzzy c-means, réseaux de neurones (Hathaway et al,2000)
Distance euclidienne	oui	k-means
Distance max	oui	fuzzy c-means avec la norme sup
Mahalanobis	oui	algorithme de classification ellipsoïdal (Mao and jain, 1996)
Coefficient de corrélation	non	Expression des gènes
Similarité cosinus	non	Classification des documents

A partir des mesures de proximités que nous avons définie ci-dessus, nous pouvons désormais procéder à l'étape de formation des classes basée sur celles-ci.
Généralement, il y a deux méthodes de classification, la méthode de partitionnement et la méthode hiérarchique, que nous allons étudier dans la section suivante.

3. Méthodes de classification

•**Méthode de partitionnement (Non hiérarchique) :** Les classes sont deux à deux disjointes et le nombre de classes est connu à priori.
L'algorithme le plus connu pour réaliser ce genre de classification est : K-means.

Avantage : Elles permettent de classifier des ensembles volumineux.
Inconvénient : Elles imposent le nombre de classes au départ.

•**Méthode hiérarchique :** Deux classes sont disjointes ou l'une contient l'autre.
Exemple : Classification ascendante (descendante) hiérarchique (CAH).

Avantage : Permettent de déterminer le nombre optimal de classes.
Inconvénient : Le temps de calcul élevé.

4. La méthode de partitionnement

Cette méthode consiste à partager un groupe d'individus en k classes homogènes. Soit $x_i \in R^d, i = 1...N$ un ensemble d'individus qu'on voudrais classer dans k classes $\{c_1, ..., c_k\}$ telle que chaque classe soit la plus homogène possible. La meilleure partition pour ce problème est celle qui réalise le minimum du critère suivant :

$$J(\delta, M) = \sum_{j=1}^{k} \sum_{i=1}^{N} \delta_{ij} \|x_i - m_j\|^2$$

Avec $\delta = \{\delta_{ij}\}$ tel que $\delta_{ij} = \begin{cases} 1 & si \quad x_i \in c_j \\ 0 & sinon \end{cases}$

$M = [m_1, \ldots, m_k]$ la matrice des moyennes des classes.

$m_i = \frac{1}{N_i} \sum_{j=1}^{N} \delta_{ij} x_j$ la moyenne de la classe c_i avec N_i individus.

Pour trouver le minimum de ce critère, on doit essayer toutes les partitions possibles. Le nombre de partitions est donné par la formule suivante (Liu, 1968) :

$$p(N, k) = \frac{1}{k!} \sum_{m=1}^{k} (-1)^{k-m} C_k^m m^N$$

Exemple
Pour partitionner 30 individus en 3 classes, le nombre de partitions possible est $2 * 10^{14}$.
D'où l'inconvénient de cette méthode (trop coûteuse en temps de calcule).

Il existe des algorithmes qui recherchent des solutions approximatives, à savoir k-means (Forgy, 1965 ; MacQueen, 1967) que nous allons étudier ci-après.

Algorithme des k-means
Le principe de l'algorithme consiste à chercher à regrouper autant que possible les individus les plus semblables (du point de vue des mesures de proximités que l'on possède) tout en séparant les classes le mieux possible les unes des autres.
Il est adapté au nuage sphérique.

• On choisit k centres aléatoirement, où k est le nombre de classes connu au préalable.

$$M = [m_1, ..., m_k]$$

• On calcule la matrice de proximité entre tous les individus et les k centres.

• On forme alors k classes 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 la première partition p_1 de $I = \{1, ..., N\}$.

$$x_i \in c_l, \quad si \quad \|x_j - m_l\| < \|x_j - m_i\| \quad pour \quad j = 1, ..., N; \quad i \neq l; \quad i = 1, ..., k.$$

• On calcule d'abord les centres des classes obtenus précédemment, ensuite la matrice de proximité entre tous les individus et les nouveaux k centres.

$$m_i = \frac{1}{N_i} \sum_{x_j \in c_i} x_j \quad i = 1, .., k; \quad j = 1, ..., N$$

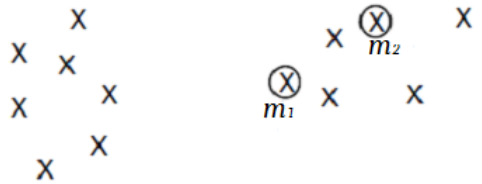
• On forme alors k nouvelles classes, telle que chaque classe étant constitué d'un centre et des individus les plus proches de ce centre que d'un autre. On a une nouvelle partition p_2 de I.

• On itère la procédure jusqu'à ce que deux itérations conduisent à la même partition.

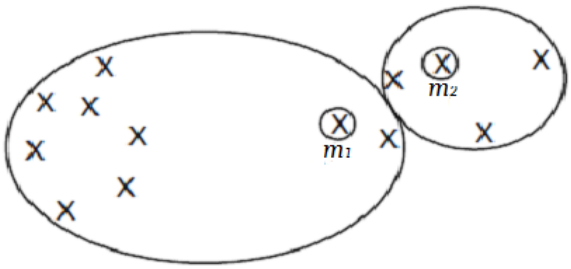
REMARQUE 2.1. *Cette méthode dépend du choix des centres initiaux. Plusieurs méthodes existent pour choisir judicieusement ces centres.*
L'inertie intra-classe diminue à chaque étape.

REMARQUE 2.2. *Après l'exécution de cet algorithme, on obtient un minimum locale. alors pour essayer d'avoir un minimum globale, on ré-exécute l'algorithme plusieurs fois avec différentes initialisations.*

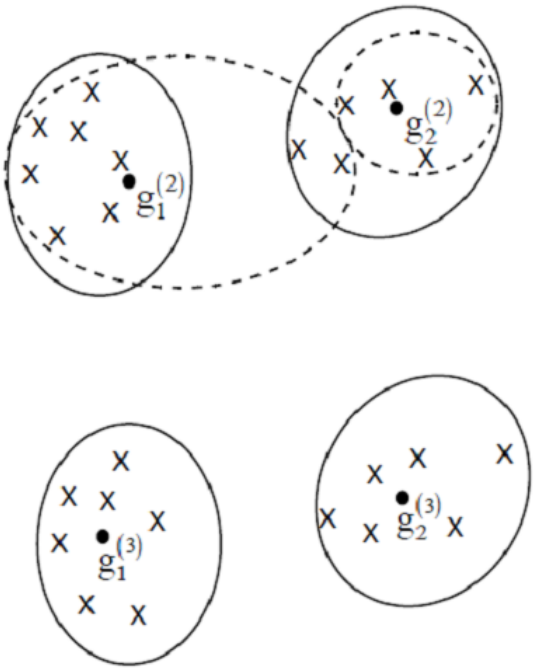
Exemple :
Nous avons un nuage de points, et nous voudrions former 2 classes de telle sorte que chaque classe soit la plus homogène possible.
Étape 0 : Choix des deux centres m_1, m_2 aléatoirement.



Étape 1 : Construction des classes autour des centres m_1, et m_2.
• Classe1 : points plus proches de m_1 que de m_2.
• Classe2 : points plus proches de m_2 que de m_1.



Étape 2 :
Calcul des centres de gravité g_1, g_2 des deux classes formées à l'étape 1. Et la définition de nouvelles classes autour de ces nouveaux centres de gravité.



Etape 3 :
Calcul des centres de gravité des classes formées à l'étape 2.
Nouvelle définition des classes autour de ces centres.

5. Classification ascendante hiérarchique

Le principe de la méthode est le suivant :
production d'une structure (arborescence) qui permet :
•La mise en évidence de lien hiérarchique entre individus ou groupe d'individus.
• La détection d'un nombre de classes au sein de notre population.

Pour réaliser une classification, il faut d'une part une mesure de proximité D entre les individus et d'autre part un indice d'agrégation, c'est-à-dire définir une mesure de proximité entre deux classes c_i et c_j.

• Distance euclidienne
Soit μ le centre de gravité de c_i, la distance euclidienne est :

$$D(x, c_i) = d(x, \mu) = \sum_{i=1}^{p} (x_i - \mu_i)^2$$

• **La plus petite distance (single linkage)**
La distance entre un individu x et une classe d'individus c_l est définie comme suit :

$$D(x, c_l) = \delta_{min} = \min_{y \in c_l} d(x, y)$$

Avec cet indice nous allons obtenir à la fin de la classification un groupe démesurément gros et plusieurs petits groupes satellites. De plus, cette méthode permet de repérer les agrégats filiformes.

• **La plus grande distance (Complete linkage)**

$$D(x, c_l) = \delta_{max} = \max_{y \in c_l} d(x, y)$$

Cet indice forme des groupes de taille égale. Cependant, la méthode est très sensible aux points aberrants et est peu utilisée en pratique.

• **Distance de Ward** La distance entre deux classes d'individus c_1 et c_2 est :

$$D(c_1, c_2) = \delta_w = \frac{n_1 n_2}{n_1 + n_2} d(\mu_1, \mu_2)$$

avec μ_i (resp. n_i) centre de gravité (resp. effectif) du groupe $c_i, i \in \{1, 2\}$.
Cet indice consiste à regrouper les classes en minimisant l'inertie intra-classe (maximisant l'inertie inter-classe).

•**La variance**

$$D(c_1, c_2) = \delta_{var} = \frac{n_{12}}{n} d(\mu_1, \mu_2)$$

avec $n_{12} = 1/(\frac{1}{n_1} + \frac{1}{n_2})$
Cet indice est utilisé dans le cas où les objets à classer sont représentés par des points d'un nuage euclidien. Il est défini par la contribution intra du dipôle des points moyens à regrouper.
Cet indice tend à agréger d'abord des classes composées d'un petit nombre d'observations.

6. Algorithme de la Classification ascendante hiérarchique

Principe de l'algorithme

Le principe de l'algorithme consiste à créer à chaque étape une partition obtenue en agrégeant deux à deux les classes d'objets les plus proches au sens de l'indice d'agrégation utilisé .
L'objectif de l'algorithme est de construire une suite de partitions emboîtées des données en n classes, $n-1$ classes, ... , 1 classe.

6.1. Algorithme. Soit $I = \{i_1 ... i_n\}$ l'ensemble des n individus.

Étape 1 : On part de la partition la plus fine de I constitué des n classes élémentaires $\{i_1\}, \{i_2\} ... \{i_n\}$. On calcule les distances entre individus deux à deux, et on agrège la paire (i_k, i_l) qui réalise le minimum pour crée le noeud l_1. On a maintenant une partition en $n-1$ classes constituée d'une classe à deux éléments correspondant au regroupement de i_k et i_l et $n-2$ classes élémentaires.

Étape 2 : On calcule les distances en prenant le centre de gravité de la classe $\{i_k, i_l\}$ obtenue précédemment comme un individu et les $n-2$ autres classes élémentaires deux à deux, et on agrège les classes $\{i_{k'}\}$ et $\{i_{l'}\}$ qui réalise le minimum (distance euclidienne), d'où le noeud l_2 et une nouvelle partition de I constituée des classes correspondant aux noeuds l_1, l_2 et des $n-4$ classes $\{i_1\} ... \{i_{n-2}\}$.

Étape 3 : Comme précédemment, en prenant les centres de gravité des classes obtenue comme individus, et on calcule les distances entre $n-2$ individus.

Et ainsi de suite jusqu'à ce que, soit créé le $n-1$ noeud qui n'est que l'ensemble I tout entier qui prend le numéro l_{n-1}

6.2. Exemple. avec $n = 10$, :
Dans cet exemple on remplace l'algorithme défini précédemment par un autre algorithme qui utilise les voisins réciproques, dont le principe remonte à McQuity (1966), qui introduit un principe d'accélération en agrégeant, à chaque étape, les paires de noeuds qui sont voisins réciproques.

DÉFINITION 2.3 (Voisins réciproques). *Deux point-individus i', i'' sont dits voisins réciproques si :*

$$\forall i \neq i', \min_{i \in I} \delta(i, i') \text{ est atteint pour } i = i''$$

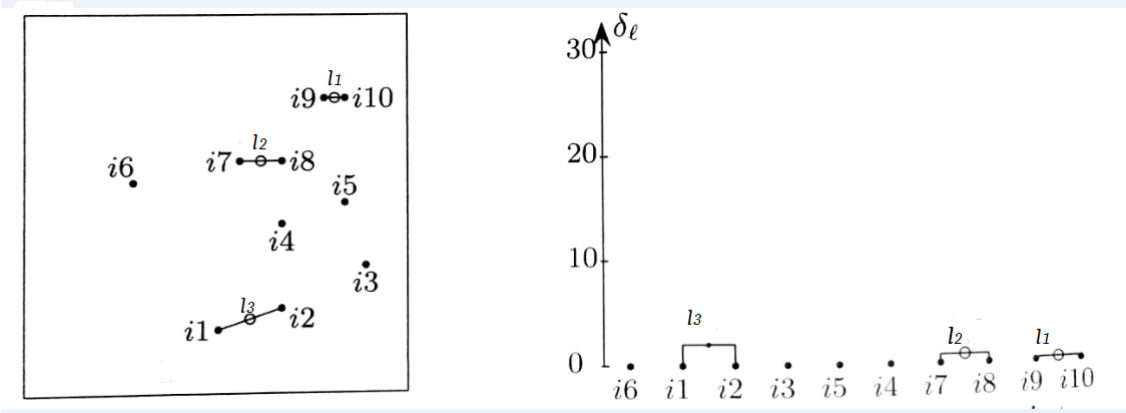
et

$$\forall i \neq i'', \min_{i \in I} \delta(i, i'') \text{ est atteint pour } i = i'$$

Étape 1 : L'indice d'agrégation utilisé dans cet exemple est δ_{var} où le poids est égale à $1/(\frac{1}{1/10} + \frac{1}{1/10}) = 0.05$, c'est le même pour les 45 dipôles.

δ	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9
i_1									
i_2	**2**								
i_3	11.6	4							
i_4	6.8	3.2	4						
i_5	14.4	6.8	2	2					
i_6	13	17	27.4	10.6	20.2				
i_7	13	10.6	12.2	2.6	5.8	5.2			
i_8	14.6	9.8	8.2	1.8	2.6	10	**0.8**		
i_9	29.2	20.8	13.6	8	5.2	19.4	5	2.6	
i_{10}	31.4	21.8	13	9	5	23.2	6.8	3.6	**0.2**

A partir des la matrice de proximité ci-dessus, on constate qu'il y a 3 paires de voisins réciproques : $(i_9, i_{10}), (i_7, i_8), (i_1, i_2)$, d'où les noeuds l_1, l_2 et l_3 et les indices de niveau $\delta_{l1} = 0.2$, $\delta_{l2} = 0.8$ et $\delta_{l3} = 2$.

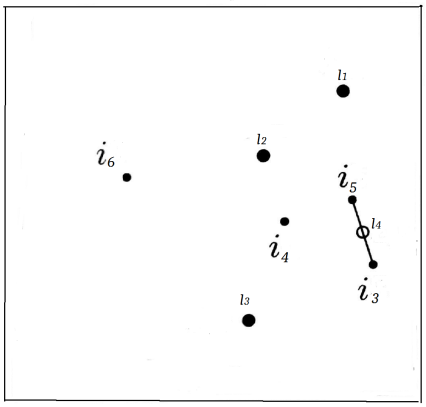


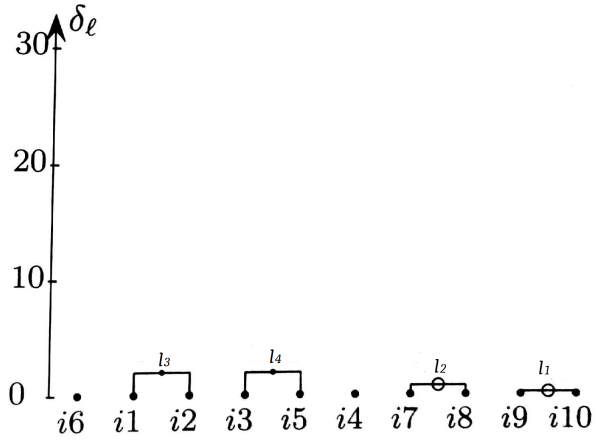
A ce niveau, nous avons 7 classes dont 3 comportent 2 éléments (classes élémentaires) et 4 classes élémentaires.

Étape 2 : Dans cette étape, on prend les centres de gravité des trois classes et les quatre autres individus, et on calcule les distances deux à deux.
Aprés le calcule nous avons obtenu la matrice suivante.

δ	i_3	i_4	i_5	i_6	l_1	l_2
i_3						
i_4	4					
i_5	**2**	**2**				
i_6	27.4	10.6	20.2			
l_1	17.7	11.3	6.7	28.3		
l_2	13.3	2.7	5.3	9.9	8.5	
l_3	9.7	6	13.5	19.3	50.5	22.6

Dans cette étape, on remarque que les couple qui réalisent les minimum sont : (i_3, i_5) et (i_4, i_5), puisque les distances sont égale, on a le choix d'agréger i_5 avec i_4 ou i_3, dans notre cas, on choisit i_3, d'où le noeud l_4. Et l'indice de niveau $\delta_{l4} = 2$

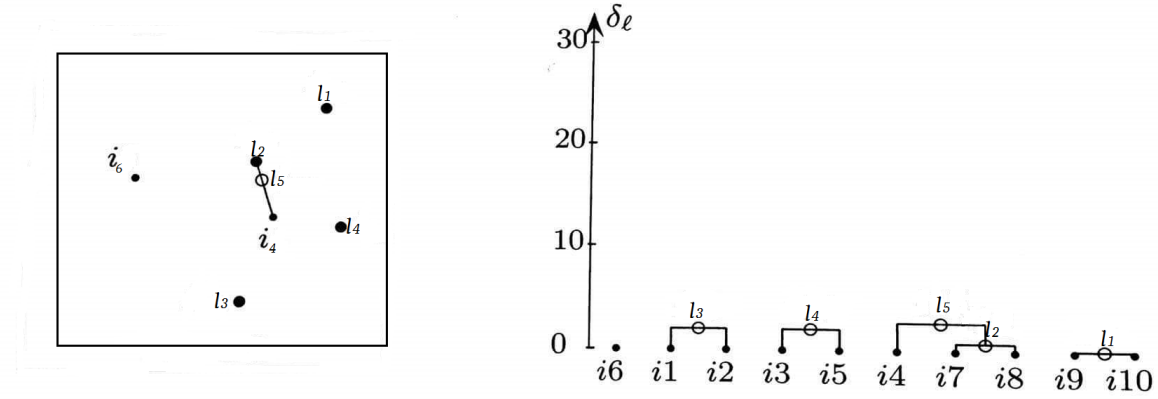




Étape 3 : On calcule les distances entre les 4 noeuds que nous avons obtenu et les 2 autres individus. on obtient la matrice suivante :

δ	i_4	i_6	l_1	l_2	l_3
i_4					
i_6	10.6				
l_1	11.3	28.3			
l_2	**2.7**	9.9	8.5		
l_3	6	19.3	50.5	22.6	
l_4	3.3	31.1	17.3	13	16.4

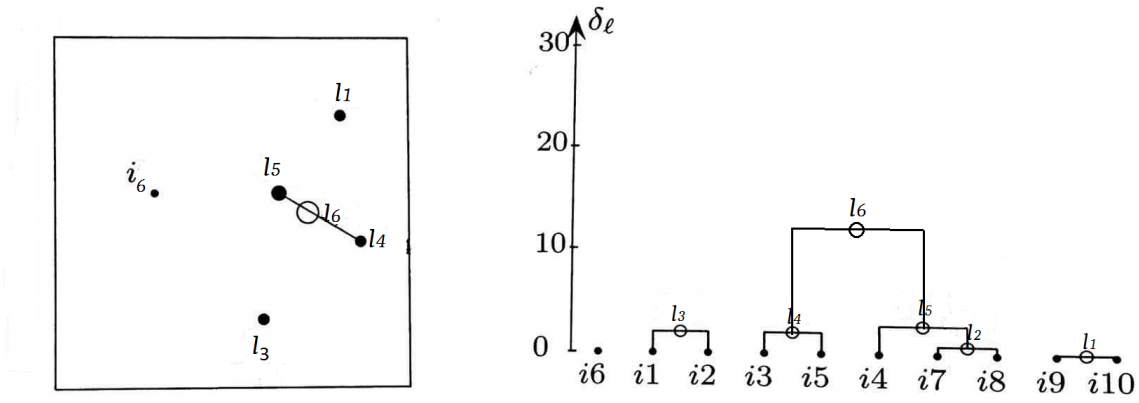
Ici le minimum est réalisé par les couple (i_4, l_2) et nous avons $\delta_{l5} = 2.7$.



Étape 4 : A ce niveau on prend l_5 comme un individu, et on calcule les indices d'agrégation entre l_5 et les 4 autres classes.

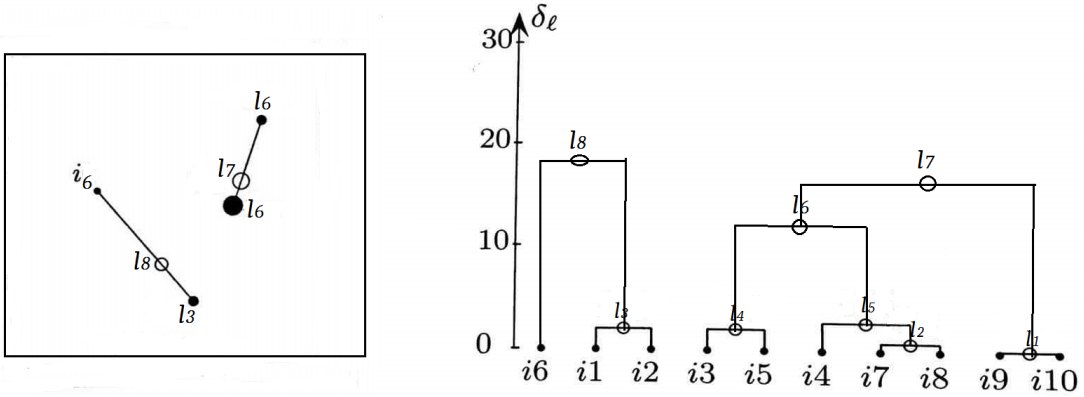
δ	i_6	l_1	l_3	l_4
i_6				
l_1	28.3			
l_3	19.3	50.5		
l_4	31.1	17.3	16.4	
l_5	12	12.5	20.6	**11.3**

On agrège l_4 et l_5, d'où le noeud l_6 avec l'indice de niveau $\delta_{l6} = 11.3$.

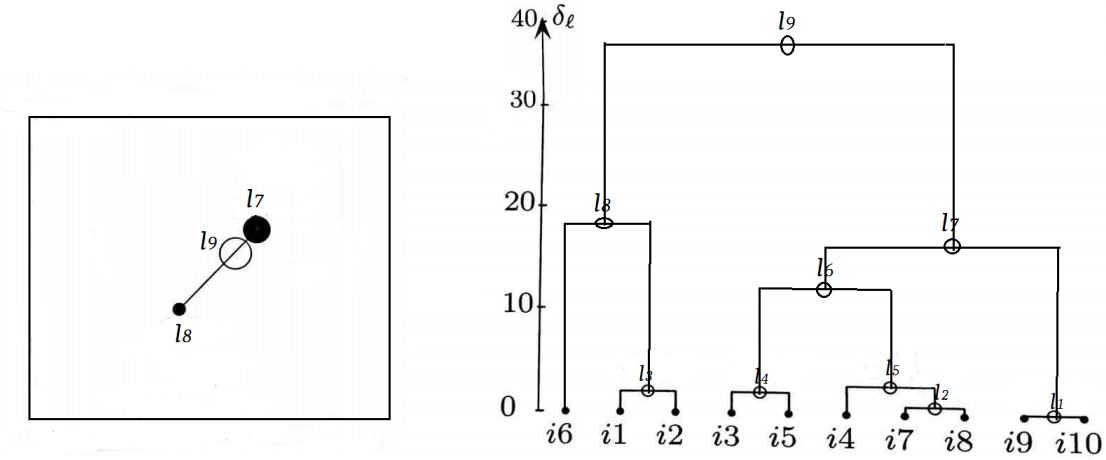


Étape 5 : On calcule les indices d'agrégation entre l_6 et les 3 autres classes. Et cette fois, il y a deux voisins réciproques (l_1, l_6) et (i_6, l_3), d'où les noeuds l_7 et l_8 avec les indices de niveau $\delta_{l7} = 15.6$ et $\delta_{l8} = 19.3$.

δ	i_6	l_1	l_3
i_6			
l_1	28.3		
l_3	**19.3**	50.5	
l_6	20.7	**15.6**	20.9

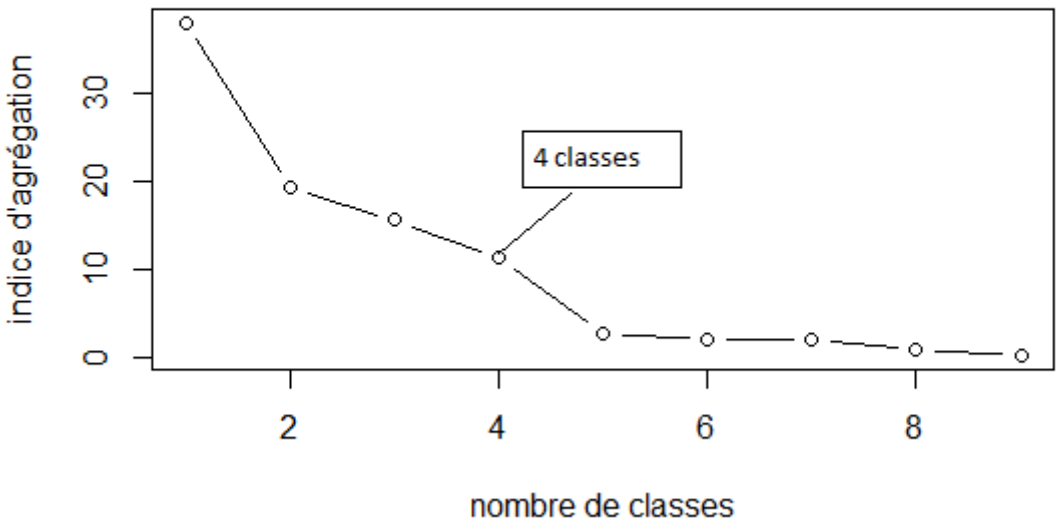


Étape 6 : C'est la dernière étape, il nous reste à agréger les deux noeuds l_7 et l_8, d'où le dernier noeud l_9 qui termine la classification avec $\delta_{l9} = 38.1$.



6.3. Choix du nombre des classes. Le choix de la coupure et du nombre de classes peut se faire à partir du diagramme décroissant des indices de niveau. On coupera entre δ_l et δ_{l-1} si l'écart entre ces deux indices est important relativement aux écarts des autre niveau de l'arbre, en s'aidant de l'inspection visuelle.

Application à l'exemple précédent :

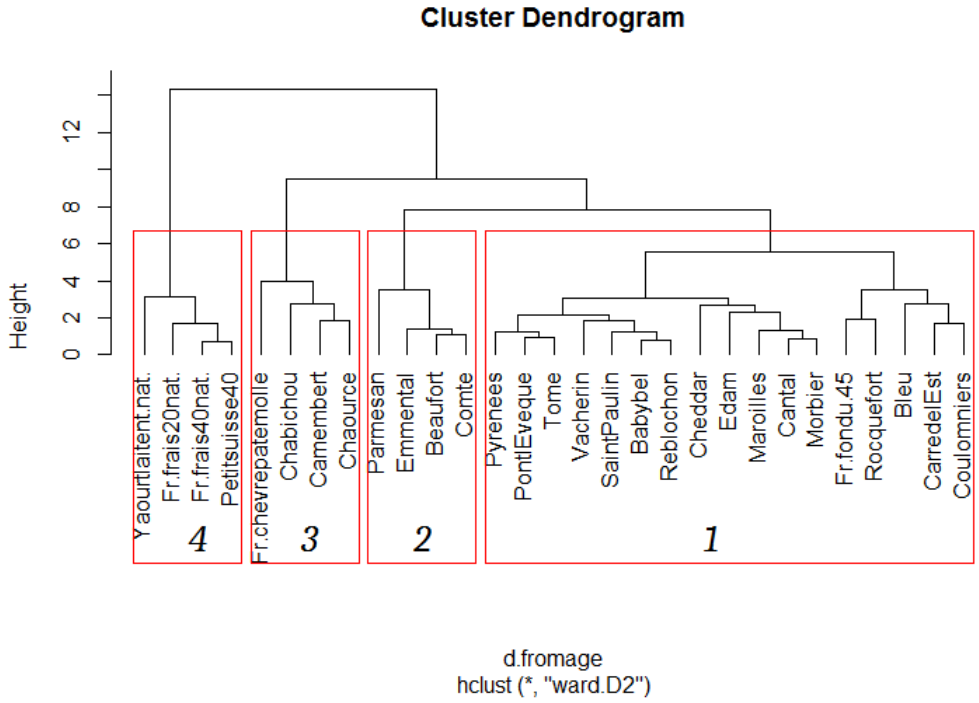


REMARQUE 2.3. *La structure de l'arbre obtenue après l'application de la classification ascendante hiérarchique dépend de l'indice d'agrégation choisi.*

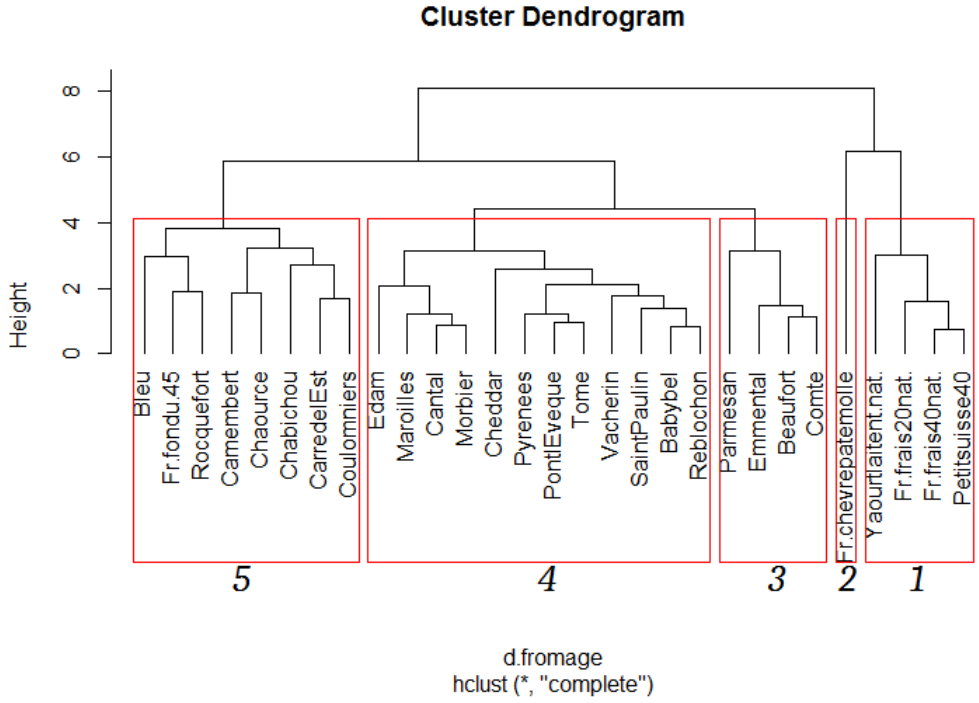
Illustration avec des exemples

Dans cet exemple, Nous allons appliqué la CAH sur un ensemble de fromages afin de découper en sous ensembles homogènes qui partagent des caractéristiques similaires.

Avec la distance de Ward



Avec complete linckage



Commentaire

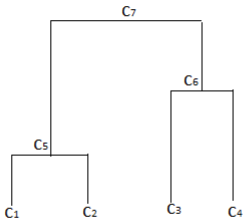
Nous constatons que la structure de l'arbre et le nombre de classes changent en fonction de l'indice d'agrégation.

Le "complete linckage" est sensible aux données aberrantes, comme le montre l'exemple ci-dessus, en donnant une classe qui contient qu'un seul individu.

7. Affectation d'individus supplémentaires en utilisant la distance de Mahalanobis

Classer un individu supplémentaire est une étape très importante dans la classification.
Soit C un ensemble des classes obtenus en appliquant la méthodes de classification ascendante hiérarchique euclidienne sur les individus de références tel que $C = \{c_1...c_m\}$. Donc nous avons m classes primaires d'effectifs $n_{c_i}, i \in \{1...m\}$, et de $m-1$ classes associées aux noeuds de l'arbre.

Le classement d'un individus supplémentaires se fera par voie descendante en procédant comme suit :
Prenant un exemple avec 4 classes primaires et 3 classes associées aux noeuds.
$C = \{c_1, c_2, c_3, c_4\}$



• On décide d'abord auquel des deux successeurs (ici c_5 où c_6) du sommet c_7 qu'on va affecter l'individu.

• Après l'affectation (supposant au noeud c_6), on déicide maintenant au quel des deux successeurs de celui-ci on va l'affecter c_3 où c_4, disons c_4.

• On procède de la même manière jusqu'à parvenir à une classe primaires.

Distance d'un individu à une classe.
Pour mesurer la distance d'un individu à une classe, on utilise la distance de Mahalanobis. Soit V_l la matrice de covariance de la classe c_l, et m_l son vecteur moyen, et x_i le vecteur des coordonnées de l'individu i. La distance de Mahalanobis de l'individu i et classe c_l est :

$$D^2_{c_l}(i) = (x_i - m_l)^\top V_l^{-1}(x_i - m_l)$$

Critère d'affectation.
Affecter l'individu i à la classe c_l où à la classe $c_{l'}$ est équivalent à tester

$$H_0 : i \in c_l$$

contre

$$H_1 : i \in c_{l'}$$

Mathématiquement ceci revient à tester

(2.9) $H_0 :$ est tirée selon la loi l_0

contre

$H_1 :$ est tirée selon la loi l_1

On propose une région d'acceptation de H_0 basée sur la distance de Mahalanobis, c'est-à-dire

$$T = \frac{(x - m_0)^\top V_0^{-1}(x - m_0)}{(x - m_1)^\top V_1^{-1}(x - m_1)} < \alpha$$

On mesure l'erreur de se tromper à l'aide de la somme de l'erreur de première espèce plus l'erreur de deuxième espèce :

$$P(T > \alpha \,|\, l_0) + P(T < \alpha \,|\, l_1).$$

Afin de minimiser cette erreur on cherche un seuil $\alpha > 0$ qui minimise

$$e(\alpha) = P(T > \alpha \,|\, l_0) + P(T < \alpha \,|\, l_1).$$

Pour calculer cette erreur il faut qu'on soit capable de calculer les probabilités

$$P(Q(Y) > c)$$

où Q est une forme quadratique, c est une constante et Y est un vecteur centré réduit.

Le calcul exact est impossible en général, de plus dans la pratique nous avons uniquement deux populations finies I_0 et I_1. Dans ce cas

$$P(T > \alpha \,|\, l_0) = \frac{\sum_{i \in I_0} \mathbf{1}_{[T_i > \alpha]}}{n_0},$$

$$P(T > \alpha \,|\, l_1) = \frac{\sum_{i \in I_1} \mathbf{1}_{[T_i < \alpha]}}{n_1},$$

où $n_0 = card(I_0)$, $n_1 = card(I_1)$. L'entier $N_0(\alpha) := \sum_{i \in I_0} \mathbf{1}_{[T_i > \alpha]}$ est le nombre d'individus mal classés dans I_0. L'entier $N_1(\alpha) := \sum_{i \in I_1} \mathbf{1}_{[T_i < \alpha]}$ est le nombre d'individus mal classés dans I_1.

Nous retenons comme critère de choix du seuil le minimum de la fonction

$$\alpha \in (0, +\infty) \to N_0(\alpha) + N_1(\alpha) = N(\alpha)$$

Remarque. Si $\max(T_i : i \in I_0) < \alpha$, alors $N_0(\alpha) = 0$. Si $\min(T_i : i \in I_1) > \alpha$, alors $N_1(\alpha) = 0$. Ainsi $N_0(\alpha) + N_1(\alpha) = 0$ si et seulement si

$$\max(T_i : i \in I_0) < \min(T_i : i \in I_1),$$

avec

$$\max(T_i : i \in I_0) < \alpha < \min(T_i : i \in I_1).$$

Si $\max(T_i : i \in I_0) > \min(T_i : i \in I_1)$, alors nécessairement $N_0(\alpha) + N_1(\alpha) \geq 1$. S'il y a un seul couple $(i_0, i_1) \in I_0 \times I_1$ tel que

$$T_{i_1} < T_{i_0},$$

alors

$$N_0(\alpha) + N_1(\alpha) = 1,$$

pour

$$T_{i_0} < \alpha < \min(T_i : i \in I_1 \setminus \{i_1\}),$$

où bien pour

$$\max(T_i : i \in I_0 \setminus \{i_0\}) < \alpha < T_{i_1}.$$

En général il existe une unique suite $i_0^{(1)}, \ldots, i_0^{(k_0)} \in I_0$, $i_1^{(1)}, \ldots, i_1^{(k_1)} \in I_1$ telle que

$$\max(T_i : i \in I_0 \setminus \{i_0^{(1)}, \ldots, i_0^{(k_0)}\}) < T_{i_1^{(1)}} < \ldots < T_{i_1^{(k_1)}} < T_{i_0^{(1)}} < \ldots < T_{i_0^{(k_0)}} < \min(T_i : i \in I_1 \setminus \{i_1^{(1)}, \ldots, i_1^{(k_1)}\}).$$

Dans ce cas

$$\min\{N_0(\alpha) + N_1(\alpha) : \alpha > 0\} = \min(k_1, k_0).$$

En effet si $\min(k_1, k_0) = k_1$, alors il suffit de prendre

$$T_{i_0^{(k_0)}} < \alpha < \min(T_i : i \in I_1 \setminus \{i_1^{(1)}, \ldots, i_1^{(k_1)}\}).$$

Si $\min(k_1, k_0) = k_0$, alors il suffit de prendre

$$\max(T_i : i \in I_0 \setminus \{i_0^{(1)}, \ldots, i_0^{(k_0)}\}) < \alpha < T_{i_1^{(1)}}.$$

Calcul du seuil. On ordonne la suite $(T_i : i \in I_0 \cup I_1)$. On obtient

$$T_{(1)} < \ldots < T_{(n)}$$

où $n = n_0 + n_1$. On calcule le milieu $\tau_i = \frac{T_{(i)}+T_{(i+1)}}{2}$ de chaque intervalle $[T_{(i)}, T_{(i+1)}]$, puis on calcule

$$\tau_{opt} := \arg\min\{N(\tau_i) : \quad i = 1, \ldots, n-1\}.$$

Algorithme implémenté sous R :

Fonction pour calculer les T_i.

#Une fonction qui permet de calculer les rapports des distances $T_(j)$ que nous utiliserons dans l'algorithme principale pour calculer le seuil.

```
#calculer les T_i on utilisant la distance de Mahalanobis
dist_Mahal =function(classe1,classe2){
donnee ← rbind(cls,cls1)
vect←vector(length = dim(donnee)[1])
d1 ← mahalanobis(donnee,moy(cls),cov(classe1))
d2 ← mahalanobis(donnee,moy(cls1),cov(classe2))
d ←(d1/d2)
for(s in 1 :(dim(donnee)[1])) { if(s<=(dim(classe1)[1])) { vect[s] ← 1}
if(s>(dim(classe1)[1])) { vect[s] ← 2}
}
return(matrix(c(d,vect), nrow=length(d) , ncol= 2 )) }
```

Fonction pour le seuil α

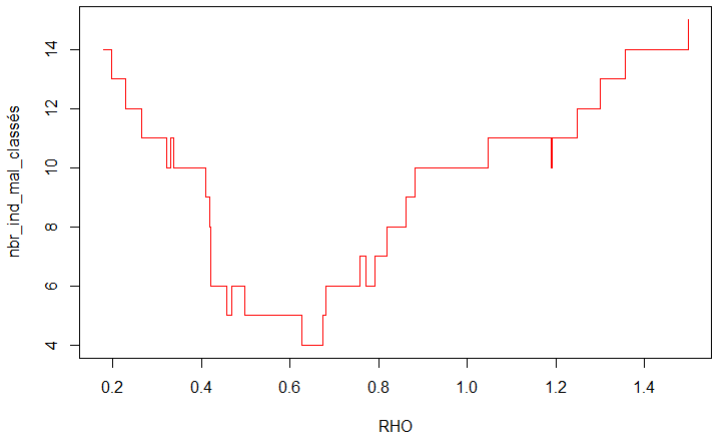
```
#calculer le seuil d'affectation en minimisant le nombre d'individus mal_classés
seuil=function(classe1,classe2)
{
    #calcul des rapports des distances de Mahalanobis
    res ← dist_Mahal(classe1,classe2)

#ordonner les rapports (orde croissant)
res2 ← res[order(res[,1],decreasing=FALSE),]
    nbr=vector(length = dim(res2)[1] )
    for(i in 1 :(dim(res2)[1]))
{
Ncls1= dim(classe1)[1]   #initialisation du nombre de mal classés dans la classe1
Ncls2=0                  #initialisation du nombre de mal classés dans la classe 2
for(j in 1 :i)
{
if(res2[j,2]==1){Ncls1 ← Ncls1-1 }
if(res2[j,2]==2){Ncls2← Ncls2+1 }
```

```
}
nbr[i] ← Ncls+Ncls1 #nombre total de mal classés
}
    #recuperer l'indice du minimum dans le vecteur des rapports

    indice ← which.min(nbr)
    alpha ← ((res2[indice,1]+res2[indice+1,1])/2)
    retour ← list(seuil=alpha,tableau=res2,nbr_mal_classé=nbr)
return (retour)
}
```

Commentaire :
Cet algorithme nous permet de calculer le seuil α associé aux deux classes et d'avoir un tableau qui contient le nombre d'individus mal classés associé à chaque $T_(j)$ tel que $j \in \{1...n_{c_l} + n_{c_{l'}}\}$.

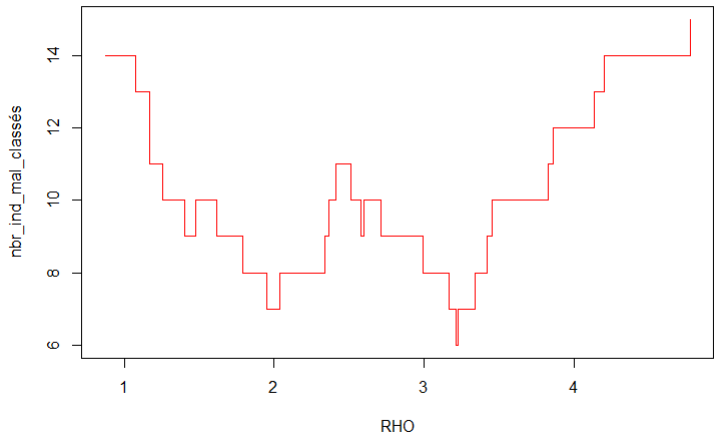
8. Exemples d'applications

Exemple1 : Résultats obtenus à partir de deux échantillons gaussiens. Ici $x = (x_1, x_2)$ tel que $x_i \sim N(6, 3.5)$ et $y = (y_1, y_2)$ tel que $y_i \sim N(6, 3)$ avec $i \in \{1, 2\}$.



$\tau_{opt} = \alpha = 0.65$

Exemple2 : Dans ce cas on a : $x = (x_1, x_2)$ tel que $x_i \sim N(0, 2)$ et $y = (y_1, y_2)$ tel que $y_i \sim N(2, 2)$ avec $i \in \{1, 2\}$.



$\tau_{opt} = \alpha = 3.245$.

9. Algorithme d'affectation d'individus supplémentaires sous R

Algorithme pour l'affectation d'un individu supplémentaire

```
affectation = function(index,cls,cls1,alpha)
{
k←0
    #calcul du rapport de la distance de Mahalanobis aux deux classes
distcls ← mahalanobis(index,colMeans(cls),cov(cls))
distcls1 ← mahalanobis(index,colMeans(cls1),cov(cls1))
τ← (distcls/distcls1)
    #l'affectation de l'individu
    if(+ < alpha) {cls ← rbind(cls,index) ; k←1}
    if(rho >= alpha) {cls1 ← rbind(cls1,index) ; k←2}
    resu ← list(classe1=cls,classe2=cls1,numclasse=k)
return (resu)
}
```

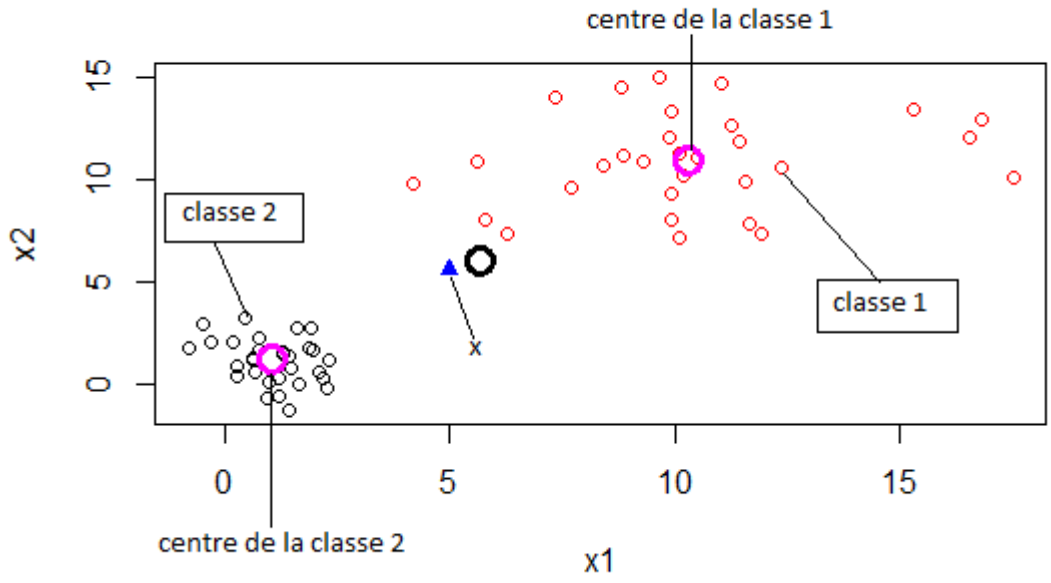
Algorithme principal

```
#l'affectation d'un ensemble d'individus supplémentaires
affect_Mahal=function(data,cls,cls1)
{
#Appel a la fonction "seuil"
alpha ← seuil(cls,cls1)$seuil
nbr_cls1=0
nbr_cls2=0
last_result ← list()
vect.Mahal ← vector(length = dim(data)[1])
for (eter in 1 :(dim(data)[1]))
{
```

```
#appel à la fonction "affectation"
last_result[[eter]] ← affectation(data[eter,],cls,cls1,alpha)
vect.Mahal[eter]← last_result[[eter]]$numclasse
cls ← last_result[[eter]]$classe1
cls1 ← last_result[[eter]]$classe2
if(last_result[[eter]]$numclasse == 1){nbr_cls1 ← nbr_cls1+1}
if(last_result[[eter]]$numclasse == 2){nbr_cls2 ← nbr_cls2+1}
}
    last ← list(les_classes=last_result[[(dim(data)[1])]],z.Mahalanobis=vect.Mahal)
return(last)
}
```

9.1. Quelques résultats de l'algorithme.

Exemple 1 : affectation d'un individu supplémentaire
Soit les vecteurs $X = (x_1, ..., x_{30})^t$ tel que chaque x_i est tiré selon la loi $N(\mu_1, \sum_1)$, et $Y = (y_1, ..., y_{30})^t$ tel que chaque y_i est tiré selon la loi $N(\mu_2, \sum_2)$, avec $\mu_1 = (1, 1)^t, \mu_2 = (11, 11)^t, \sum_1 = I_2$ et $\sum_2 = 2I_2$. nous avons une nouvelle observation $x = (5, 5.6)$ représentée par le petit carré sur la figure ci-dessous, et nous voudrions savoir, si elle est tirée selon la loi de X ou bien celle de Y.

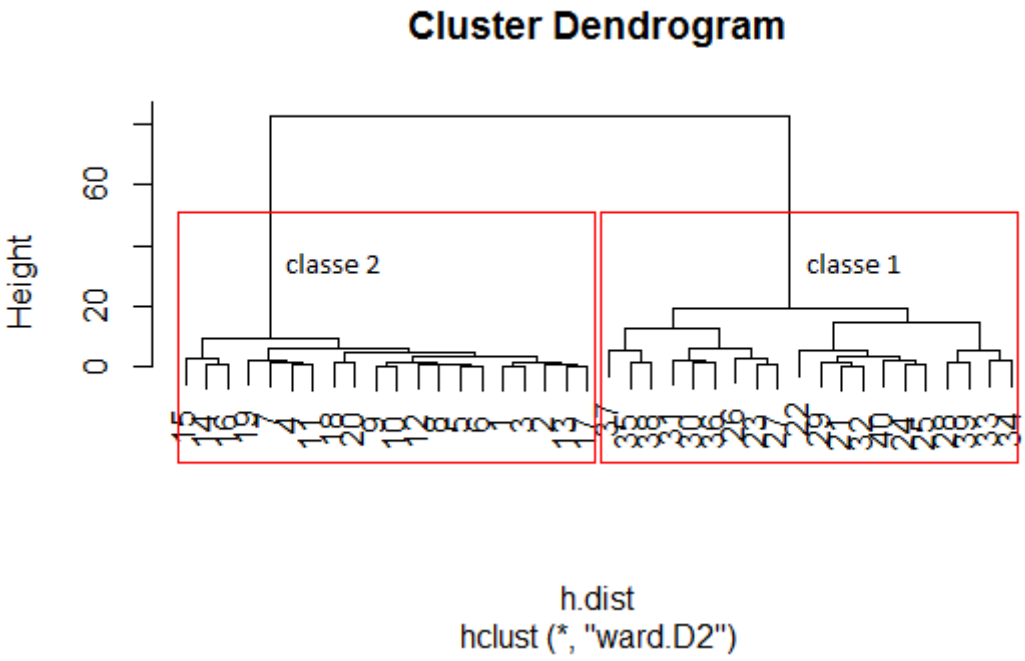


Le résultat de notre algorithme est le suivant :

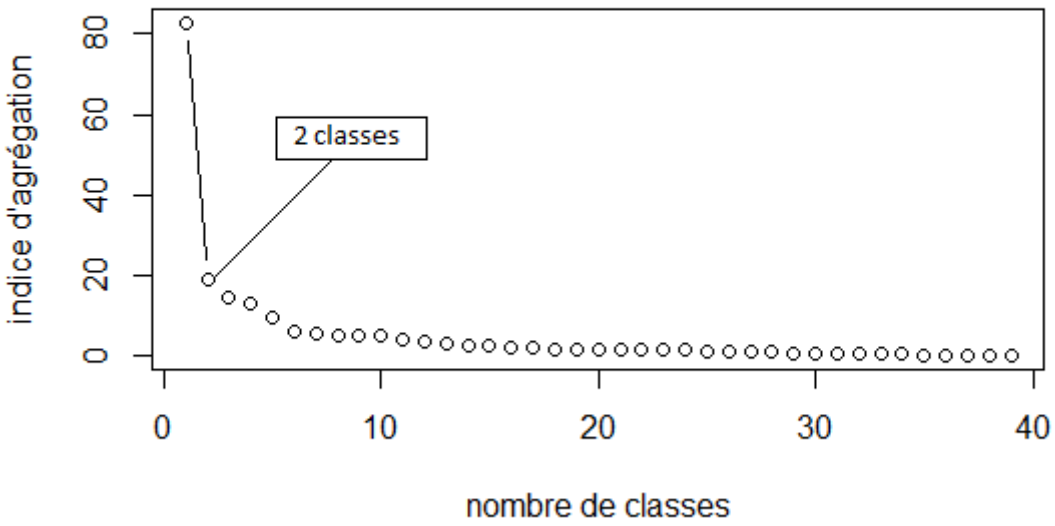
```
$z.Mahalanobis
[1] 1

$z.Eucli
[1] 2
```

Nous constatons que, si on utilise la distance de Mahalanobis, on affecte l'observation x à la classe 1, par contre si on utilise la distance euclidienne l'observation sera affectée à la classe 2.

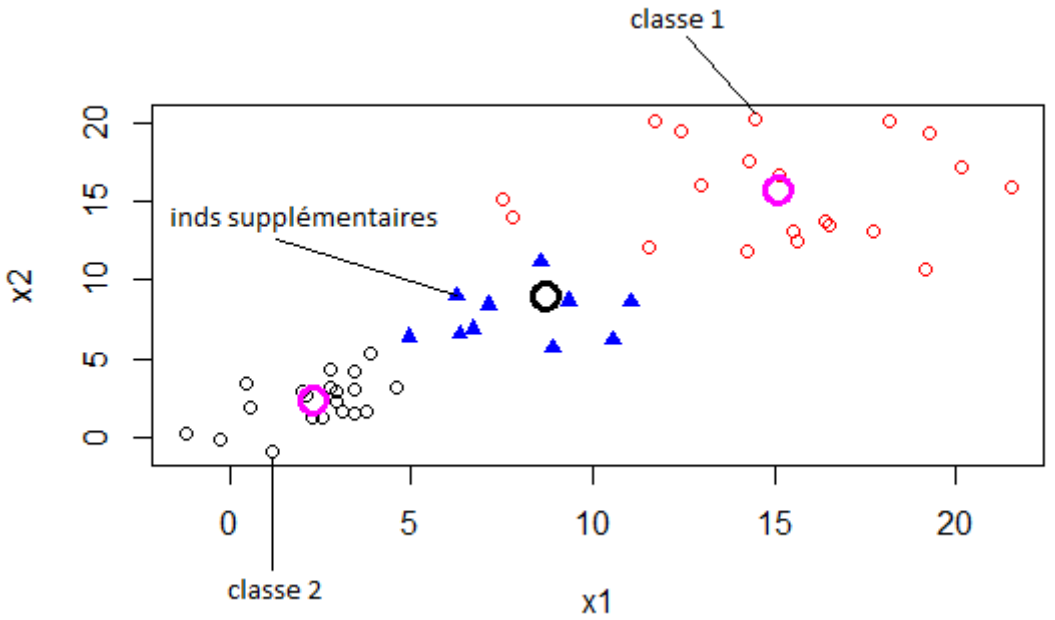
Exemple 2 : affectation de plusieurs individus supplémentaires Dans cette exemple, nous avons 40 observations qu'on va classer en utilisant la CAH, comme le montre la figure ci-dessous.



L'arbre ci-dessus met en évidence deux classes, comme nous le confirme le diagramme suivant.



Maintenant nous voudrions affecter dix observations supplémentaires aux classes obtenues, en utilisant la distance de Mahalanobis.



les deux petits cercles dans les classes représentent leurs centres, et celui du milieu, est le centre de tout le nuage.

Le résultat de notre algorithme est le suivant :

```
$z.Mahalanobis
 [1] 1 1 1 1 1 1 1 1 1 1

$z.Eucli
 [1] 2 2 1 2 2 1 1 2 2 1
```

Le premier vecteur correspond aux résultats de l'affectation en utilisant la distance de Mahalanobis, et le deuxième en utilisant la distance euclidienne.

Chapitre 3

Appendix : Distance de Mahalanobis

1. La distance de Mahalanobis

Soit V une matrice de covariance d'ordre $p \times p$, c'est-à-dire symétrique et définie positive. La distance $d(x,y) = \sqrt{(x-y)^\top V^{-1}(x-y)}$ est appelée la distance de Mahalanobis entre les deux vecteurs x, y. Dans le cas où X est un vecteur aléatoire Gaussien $N(m,V)$, le carrée de la distance $d^2(X,m)$ de X à sa moyenne m est une variable χ^2_p.

Démonstration: Soit X un vecteur aléatoire Gaussien suivant la loi $N(m,V)$. Il existe une variable aléatoire $Z \sim N(0,I)$, tel que $X = AZ + m \sim N(m,V)$ avec $V = AA^\top$
On a V une matrice symétrique définie positive, alors elle admet une décomposition comme suit : $V = PDP^\top$ avec D une matrice diagonale et P une matrice orthogonale.
D'après l'unicité de la décomposition, on peut écrire $V = PD^{1/2}P^\top(PD^{1/2}P^\top)^\top$ avec $A = PD^{1/2}P^\top$.

$$\begin{aligned}
d^2(X,m) &= (X-m)^\top V^{-1}(X-m)p \\
&= (AZ)^\top V^{-1}(AZ) \\
&= Z^\top A^\top V^{-1} AZ \\
&= Z^\top (PD^{1/2}P)^\top V^{-1} PD^{1/2}P^T Z \\
&= Z^\top (PD^{1/2}P)^\top (PD^{1/2}P^\top (PD^{1/2}P^\top)^\top)^{-1} PD^{1/2}P^T Z \\
&= Z^\top Z \\
&= \sum_{i=1}^{p} Z_i^2 \sim \chi^2_p
\end{aligned}$$

□

La matrice V^{-1} est appelée matrice de précision. Elle contient plus d'information que la matrice de covariance V car elle nous permet de voir la covariance entre deux variables conditionnellement aux autres, comme nous le montre la formule suivante :

$$cov(X_i, X_j / X_k, k \neq i,j) = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{ij}}}$$

Où les σ^{ij} sont les coefficients de la matrice de précision.

REMARQUE 3.1. *Si $\sigma^{ij} = 0$ les variables X_i et X_j sont conditionnellement indépendantes.*

REMARQUE 3.2. *L'écriture analytique de la distance de Mahalanobis*

$$d^2(x, m) = \sum_{i,j} \sigma^{ij}(x_i - m_i)(x_j - m_j)$$

où $V^{-1} = [\sigma^{ij}]$ désigne l'inverse de la matrice de covariance X.

2. L'interprétation géométrique de la distance de Mahalanobis

Soit $X = (X_1...X_n)$ un vecteur aléatoire de moyenne m, et soit $x \in \mathbb{R}^n$ tel que $r^2 = d(x, m)^2$ où d est la distance de Mahalanobis. alors l'ensemble $\{y \in \mathbb{R}^n \ d(y, m)^2 \leq r^2\}$ forme un hyper-ellipsoïde centré en m.

Démonstration:
Soit V la matrice de covariance de X, elle est symétrique et définie positive, alors on peut la décomposer de la façon suivante $V = PDP^\top$ avec $D = diag(\lambda_1, ..., \lambda_p)$ est une matrice diagonal qui contient toutes les valeurs propres de V, et $P = (e_1, ..., e_p)$ est une matrice orthogonale où les e_i sont les vecteurs propres de la matrice V.

$$V = PDP^\top$$

Alors

$$V^{-1} = PD^{-1}P^\top$$

d'où

$$\begin{aligned} d^2(y, m) &= (y - m)^\top V^{-1}(y - m) \\ &= (y - m)^\top PD^{-1}P^\top(y - m) \\ &= U^\top D^{-1}U \quad avec \quad U = P^\top(y - m) \\ &= \sum_{i=1}^{p} \frac{1}{\lambda_i} U_i^2 \end{aligned}$$

Enfin on a :

$$\begin{aligned} &\forall y \in \mathbb{R}^p : d^2(y, m) \leq r^2 \\ &\Leftrightarrow \sum_{i=1}^{p} \frac{1}{\lambda_i} U_i^2 \leq r^2 \end{aligned}$$

□

2.1. Indépendance conditionnelle et régression linéaire. Soit $X = (X_1, .., X_p) \in \mathbb{R}^p$ un vecteur aléatoire gaussien centré, V sa matrice de covariance inversible d'inverse $V^{-1} = [\sigma^{ij}]$. La variable X_i est indépendante de X_j sachant les autres variables $(X_l : l \neq i, j)$ si et seulement si :

$$\sigma^{ij} = 0.$$

Le graphe d'indépendance conditionnelle gaussienne est défini par les noeuds $\{1, ..., p\}$ reliés par des arrêtes non orientées : il y a une arrête entre les noeuds i et j si et seulement si : $\sigma^{ij} \neq 0$.

Dans le contexte des réseaux de régulation des gènes, les variables X_i représentent les P gènes.
L'explication du gène i par les autres gènes est donnée par la régression linéaire suivante :

$$X_i = \sum_{j \neq i} \beta_j^i X_j + E_i, \quad i = 1, ..., p.$$

où les E_i représentent les résidus de la régression.

Calcul de la corrélation partielle
La corrélation partielle entre deux gènes est définie comme suit :

$$\rho_{ij} = cor(E_i, E_j), \quad i \neq j.$$

On dit qu'il y a une arrête entre les gènes i et j si :

$$\rho_{ij} \neq 0.$$

Il est difficile de calculer les résidus d'estimation d'une variable en fonction d'un groupe de variables, principalement quand le nombre de variables à considérer est important.
la corrélation partielle peut être calculée par deux manières équivalentes suivantes : **[3]**, **[5]**, **[6]**.

1) Inversion de la matrice de covariance V tel que, $V^{-1} = [\sigma^{ij}]$ et on aura :

$$\rho_{ij} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}.$$

2) Régression des moindres carrés

$$\begin{aligned}\beta_j^i &= \arg\min_{\beta\in\mathbb{R}^{p-1}} \mathbf{E}(\|X_i - \sum_{j\neq i} \beta_j X_j\|_2^2) \\ &= -\frac{\sigma^{ij}}{\sigma^{ii}}\end{aligned}$$

La corrélation partielle entre les gènes i et j est donnée par :

$$\rho_{ij} = sign(\beta_j^i)\sqrt{\beta_j^i \beta_i^j}.$$

Finalement les propositions suivantes sont équivalentes :
1) $\sigma^{ij} \neq 0$,

2) $\rho_{ij} \neq 0$,

3) $\beta_j^i \neq 0$.

Chapitre 4

Conclusion

En guise de conclusion, nous aimerons tout d'abord souligner que nous avons de la chance de faire partie d'un travail qui nous a permis de développer une méthode très importante de l'analyse de données qui est la classification.

Premièrement, dans ce mémoire, nous avons étudié les deux grandes familles de méthodes classification : les méthodes de partitionnement et les méthodes hiérarchique.
Dans les méthodes hiérarchique, nous avons étudié plus particulièrement la méthode CAH. Nous avons présenté l'algorithme de la méthode CAH, et l'importance du choix de la mesure de proximité entre les individus et entre les groupes d'individus sur le résultats final de celle-ci.
le partitionnement, nous avons vu l'une des méthodes la plus utilisée, à savoir la méthode des k-means. Cette méthode dépend aussi du choix de la mesure de proximité choisi, comme elle dépend aussi de l'initialisation.

Finalement, nous avons aussi étudié le problème d'affectation d'un individu supplémentaire en utilisant la distance de Mahalanobis dans des classes obtenue par une CAH.

Bibliographie

[1] Le Roux Brigitte.(2014) : *Analyse géométrique des données multidimensionnelles*, pages 327-339.

[2] Frédérik Cassor, Brigitte Le Roux.(2014) : *Un critère basé sur la distance de Mahalanobis pour l'affectation d'objets supplémentaires aux classes d'une CAH euclidienne*

[3] J. Whittaker *Graphical Models in Applied Multivariate Statistics, Wiley New York, 1990.*

[4] P.J. Bickel, Y. Ritov, A.B. Tsybakov (2009), *Simultaneous analysis of Lasso and Dantzig selector, The Annals of Statistics 37, 1705-1732.*

[5] N. Krämer, J. Schäfer, A.L. Boulesteix, *Regularized estimation of large scale gene association networks using graphical Gaussian models, Technical Report no ; 057, 2009.*

[6] Meinshausen N, Bühlmann P (2006) : *High dimensional graphs and variable selection with Lasso, Annals of Statistics, 34, 1436-1462.C*