Partitionnement (clusterization)

Christophe Genolini

10 novembre 2010

Table des matières

1 Introduction

Objectif. Partitionner, c'est découper des données en sous-groupes homogènes. Cela permet, à l'intérieur d'une population, de définir des groupes. Par exemple, on peut identifier le type de maladie d'un patient parce que différents type de symptômes ont été identifiés comme apparaissant souvent ensemble. On les a donc regroupé sous un même nom, celui d'une maladie.
Du point de vue statistique, partitionner permet de résumer un grand nombre de variables en une seule (qui sera [groupe]) que l'on peut ensuite inclure dans une régression : au lieu de rentrer [symptome1], [symptome2], ..., [symptomeN] dans la régression (et que rien ne soit significatif pour cause de colinéarité), il suffit de rentrer [groupe].

Exemple. Un patient a :

- Fièvre, douleur, maux de gorges : On ne sait pas
- Fièvre, douleur, maux de gorges, maux de tête, grande fatigue, courbature, toux sèche. On sait que c'est la grippe (et on s'empresse de ne rien faire !)
- Fièvre, douleur, maux de gorges, écoulement nasal. On sait que c'est un rhume (et on ne peut rien faire non plus...)

Méthode sous-jacente : on note les symptômes de nombreux patients puis on cherche des "grandes familles" de symptômes. Chaque grande famille est probablement une maladie.
Pour regrouper les individus qui se ressemblent, plusieurs méthodes sont possibles.
– Certaines se basent sur la proximité entre les individus.
– D'autres utilisent la vraisemblance.

2 Distances et similarités

Pour mesurer la proximité entre deux individus, on utilise deux types de fonctions : les distances et les similarités.

– Une **distance** est une fonction qui prend pour argument deux individus et qui renvoie une valeur élevée si les deux individus sont dissemblables, petite si les individus se ressemblent.
– Une **similarité** est une fonction qui prend pour argument deux individus et qui renvoie une valeur élevée si les deux individus sont semblables, petite si les individus sont dissemblables. En général, les similarités sont à valeur dans [0,1], 0 signifiant une disconcordance parfaite et 1 une concordance parfaite.

Mathématiquement, on peut assez facilement passer de l'une à l'autre. Dans ce cours, on utilisera donc soit l'une, soit l'autre, selon ce qui est le plus intuitif à définir.

2.1 Variables continues

On considère deux individus I_1 et I_2 que l'on mesure sur trois variables U, V et W.

– Distance **Euclidienne** : $Dist_{Eucl}(I_1, I_2) = \sqrt{(U_1 - U_2)^2 + (V_1 - V_2)^2 + (W_1 - W_2)^2}$
– Distance **Manhattan** : $Dist_{Manh}(I_1, I_2) = |U_1 - U_2| + |V_1 - V_2| + |W_1 - W_2|$
– Distance **Maximum** : $Dist_{Max}(I_1, I_2) = max\{|U_1 - U_2|, |V_1 - V_2|, |W_1 - W_2|\}$
– Distance **Minkowski** : $Dist_{M(k)}(I_1, I_2) = \sqrt[k]{(U_1 - U_2)^k + (V_1 - V_2)^k + (W_1 - W_2)^k}$

Beaucoup d'autres sont possibles comme $\frac{1-correlation(I_1,I_2)}{2}$. À noter, la Minkowski est une généralisation des trois autres ; $k = 2$ donne l'Euclidienne, $k = 1$ donne la Manhattan et $k = +\inf$ donne le maximum.

Exemple

Id	math	francais	histoire
I1	5	12	14
I2	6	15	12

On obtient :

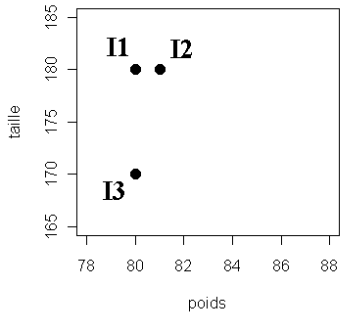
– $Dist_{Eucl}(I1, I2) = \sqrt{(5-6)^2 + (12-15)^2 + (14-12)^2} = 3,74$
– $Dist_{Manh}(I1, I2) = |5-6| + |12-15| + |14-12| = 6$
– $Dist_{Max}(I1, I2) = max\{|5-6|, |12-15|, |14-12|\} = 3$

Problème : dans l'exemple précédent, toutes les variables sont mesurées dans la même unité. Quand ça n'est pas le cas, certaines variables peuvent avoir un poids plus important que d'autres. Pire, un simple changement d'unité peut tout changer :

Exemple :

Id	poids	taille (cm)
I1	80	180
I2	81	180
I3	80	170

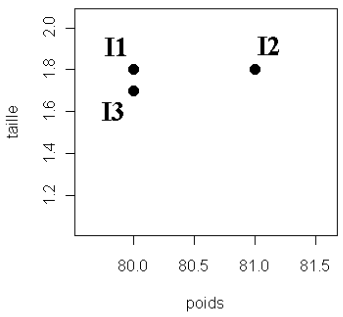
$Dist_{Eucl}(I1, I2) = \sqrt{1^2 + 0^2} = 1$
$Dist_{Eucl}(I1, I3) = \sqrt{0^2 + 10^2} = 10$



La taille a plus d'importance que le poids ;
$I1$ est plus proche de $I2$ que de $I3$.

Id	poids	taille (m)
I1	80	1,80
I2	81	1,80
I3	80	1,70

$Dist_{Eucl}(I1, I2) = \sqrt{1^2 + 0^2} = 1$
$Dist_{Eucl}(I1, I3) = \sqrt{0^2 + 0.1^2} = 0.1$



La taille a peu d'importance ;
$I1$ est plus proche de $I3$ que de $I2$.

Pour éviter cela, il est possible de calculer la distance sur des variables normalisées. Différentes méthodes de normalisation sont possibles : on peut utiliser la classique division par l'écart type mais la division par l'étendue a un intérêt supplémentaire : c'est celui de borner la contribution de chaque composante de la distance par [0,1] ce qui permet ensuite une transformation facile en similitude. C'est particulièrement utile dans le cas de combinaison de distances comme nous le verrons au paragraphe 2.4...

Dans notre exemple, on obtient $Dist_{Eucl-Norm}(I1, I2) = 1$ et $Dist_{Eucl-Norm}(I1, I3) = 1$.

2.2 Variables binaires

On considère deux individus I_1 et I_2 que l'on mesure sur six variables B_1 à B_6. On dresse le tableau croisé des deux individus. On note :

- a le nombre de variables B_i pour lesquelles $B_i(I_1) = 1$ et $B_i(I_2) = 1$
- b le nombre de variables B_i pour lesquelles $B_i(I_1) = 1$ et $B_i(I_2) = 0$
- c le nombre de variables B_i pour lesquelles $B_i(I_1) = 0$ et $B_i(I_2) = 1$
- d le nombre de variables B_i pour lesquelles $B_i(I_1) = 0$ et $B_i(I_2) = 0$

Au final, une similitude possible entre I_1 et I_2 est la **concordance simple** : $S_{CS}(I_1, I_2) = \frac{a+d}{a+b+c+d}$

Exemple

Id	B1	B2	B3	B4	B5	B6
I1	1	1	1	1	0	1
I2	1	0	1	0	0	1

On a $a = 3$ (pour B_1, B_3 et B_5), $b = 2$ (pour B_2 et B_4), $c = 0$ et $d = 1$ (pour B_5), donc $S_{CS}(I_1, I_2) = \frac{3+1}{3+2+0+1}$. Cette similitude peut s'interpréter comme le pourcentage de points communs entre I_1 et I_2.

Problème : dans le calcul précédent, a et d on un rôle symétrique. Or, ça n'est pas toujours justifié. Plus précisément, avoir le même caractère peut rendre deux individus proches alors que ne pas avoir le même caractère n'a pas de signification. Par exemple, si deux individus sont tous les deux *membres* du rotary club, ils sont proches. Si deux individus sont tous les deux *non membres* du rotary club, ils ne sont pas spécialement proches. Ce genre de caractéristiques (être membre d'un club, parler l'espéranto, faire de la plongée sous-marine, être séropositif) est dit **dissymétrique**. Par opposition, les caractéristiques comme le sexe, avoir le bac, être propriétaire de son logement

sont **symétriques**.

En utilisant la formule ci-dessus sur des variables dissymétriques, on trouve des résultats étranges :

Id	6 pates	Ailes	Poils	Antennes
I1	0	0	0	0
I2	0	1	0	0
I3	1	0	1	1
I4	1	1	0	1

$S_{CS}(I_1, I_2) = \frac{0+3}{0+1+0+3} = 0,75$
$S_{CS}(I_3, I_4) = \frac{2+0}{2+1+1+0} = 0,50$

Or, après enquête, I_1 est un ver de terre, I_2 est un ptérodactyle, I_3 est une fourmi et I_4 est une abeille... Dire qu'un ver de terre et un ptérodactyle sont plus proches qu'une abeille et une fourmi est assez étrange.

Solution : on ne compte plus le fait de ne pas avoir un critère en commun. On obtient le **coefficient de Jaccard** ou **similitude de Jaccard** : $S_J(I_1, I_2) = \frac{a}{a+b+c}$. Appliqué à nos individus :

- $S_J(I_1, I_2) = \frac{0}{0+1+0} = 0$
- $S_J(I_3, I_4) = \frac{2}{2+1+1} = 0,50$

Cela qui semble un peu plus raisonnable.

2.3 Variables nominales non binaires

Pour une nominale à plusieurs niveaux, on compte le nombre de paires concordantes et on le divise par le nombre total de paires.

Exemple :

Id	Activité à 8 ans	Activité à 9 ans	Activité à 10 ans
I1	Judo	Judo	Tennis
I2	Gym	Tennis	Tennis

On obtient : $S(I_1, I_2) = \frac{1}{3} = 0,33$.

2.4 Combinaison de plusieurs types de variable

Si on a des variables de natures différentes, on utilise la **similarité de Gower**. Elle a pour avantage de mixer différents types de similarités et de prendre en compte les valeurs manquantes. Pour la calculer, on définit des similarités S_i pour chaque variable V_i. Ensuite, on combine les similarités pertinentes grâce à un facteur d'inclusion w_i. Plusieurs cas de figure :

- Pour une variable nominale, $S_i(I_1, I_2)$ vaut 1 si $V_i(I_1)$ et $V_i(I_2)$ sont égaux, 0 sinon.
- Pour une variable continue, on considère la distance de Manhattan normalisée par l'étendue de la variable (normalisation qui permet d'éviter les problèmes d'échelle vus au paragraphe 2.1) : $\frac{|V_i(I_1)-V_i(I_2)|}{Etendue(V_i)}$. Puis on transforme cette distance en similarité : $S_i(I_1, I_2) = 1 - \frac{|V_i(I_1)-V_i(I_2)|}{Etendue(V_i)}$

Le coefficient de Gower est ensuite la moyenne des différentes similarités S_i obtenues. Toutefois, certaines similarités sont à exclure. C'est le rôle de w_i :

- Si $V_i(I_1)$ ou $V_i(I_2)$ est manquante, alors w_i vaut zéro.
- Si V_i est une variable binaire non symétrique et que $V_i(I_1)$ et $V_i(I_2)$ valent toutes les deux zéro (on est dans le cas où *ne pas* avoir le même caractère est sans importance), alors w_i vaut zéro.
- Dans tous les autres cas, w_i vaut 1.

Au final, la similarité de Gower est $S_G(I_1, I_2) = \frac{\sum_i w_i.S_i(I_1,I_2)}{\sum_i w_i}$

Exemple :

Id	Age	$Taille$	$Sexe$	$Rougeole$	$Oreillon$
I_1	8	$< NA >$	H	Oui	Non
I_2	9	$1,12$	F	Oui	Non
I_3	13	$1,30$	F	Non	Oui

On cherche la similarité de Gower entre I_1 et I_2 :

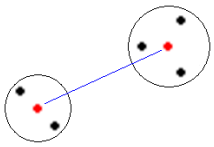
	Age	$Taille$	$Sexe$	$Rougeole$	$Oreillon$
S_i	$1 - \frac{\|8-9\|}{13-8}$	$1 - \frac{<NA>-1.12}{1.12-1.30}$	0	1	1
w_i	1	0	1	1	0

Au final, $S_G = \frac{1\times0,8+0\times<NA>+1\times0+1\times1+0\times1}{1+0+1+1+0} = 0,6$

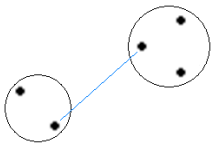
2.5 Distance entre groupes

Pour finir avec les distances, certaines méthodes utilisent la notion de **distance entre groupes** d'individus. Là encore, plusieurs possibilités :

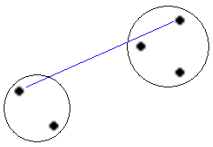
Distance des centres : distance entre les centres de gravité des groupes.



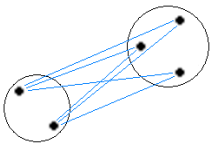
Distance minimum : distance la plus courte entre les deux groupes.



Distance maximum : distance la plus longue entre les deux groupes.

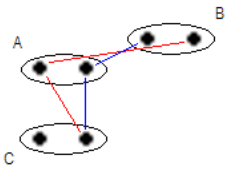


Distance moyenne : moyenne des distances entre toutes les paires d'individus.



De manière générale, les distances min et max sont sensibles aux valeurs aberrantes. La distance moyenne est coûteuse à calculer. Au final, laquelle doit-on utiliser ? Ca dépend des domaines et des problèmes...

À noter que le choix d'une distance n'est pas anodin :



Le groupe A est plus proche du groupe B si on considère la distance min (en bleu), mais plus proche de C si on considère la distance max (en rouge)...

3 Partitionnement

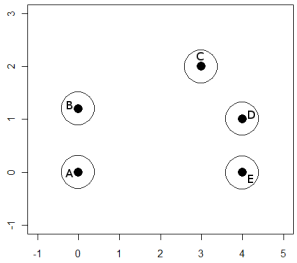
Dans la suite, pour plus de simplicité et afin de pouvoir représenter graphiquement la construction des clusters, nous considérons que le partitionnement se fait sur deux variables continues.

3.1 Classification hiérarchique ascendante

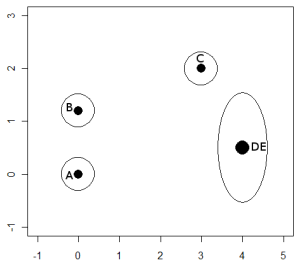
Initialement, chaque individu est un cluster. Puis on fusionne les deux clusters les plus proches. Et on itère jusqu'à ce qu'il ne reste qu'un seul cluster.

Exemple : on considère les points $A = \begin{pmatrix}0\\0\end{pmatrix}$, $B = \begin{pmatrix}0\\1,2\end{pmatrix}$, $C = \begin{pmatrix}3\\2\end{pmatrix}$, $D = \begin{pmatrix}4\\1\end{pmatrix}$ et $E = \begin{pmatrix}4\\0\end{pmatrix}$; on utilise la distance euclidienne entre les points, la distance des centres entre les groupes.

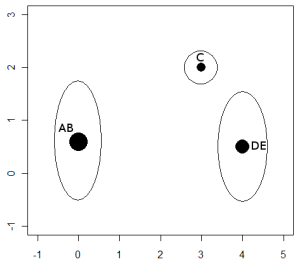
Initialement, chaque point “est” un cluster.



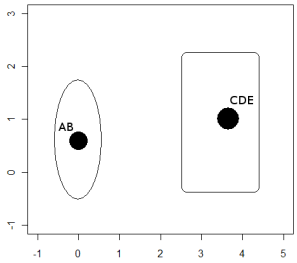
Les clusters D et E étant les plus proches, on les fusionne. Le nouveau cluster obtenu est nommé DE. Dans la suite, on considère son centre de gravité.



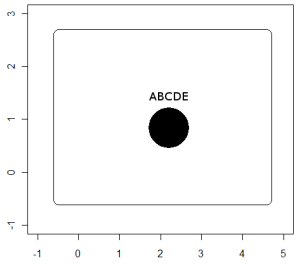
Les clusters A et B étant les plus proches, on les fusionne en AB et on considère le centre de gravité de AB.



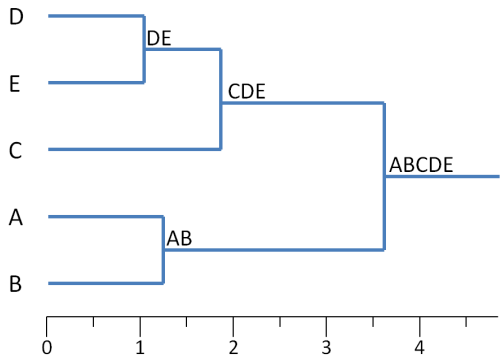
Les clusters C et DE étant les plus proches, on les fusionne en CDE et on considère le centre de gravité de CDE.



Les clusters AB et CDE étant les deux derniers (!), on les fusionne.



Une fois que la classification est terminée, on peut représenter graphiquement les différentes étapes : pour cela, on écrit les noms des deux premiers clusters qui fusionnent. On trace sur leur droite (ou à leur verticale) un trait ayant pour longueur la distance qui sépare les deux clusters. Puis on joint les deux traits. Dans notre exemple, on a initialement fusionné D et E, et ils étaient à une distance de 1. Puis A et B étaient à distance 1.2 ; C et DE étaient à distance 1,8 ; AB et CDE étaient à distance 3,69. L'arbre obtenu est :



Cet arbre permet ensuite de choisir le nombre de clusters.

Avantages et inconvénients
+ Pas besoin de choisir a priori le nombre de clusters ;
- Deux points séparés ne pourront plus être rassemblés ;
- Il faudra quand même choisir un nombre de clusters.

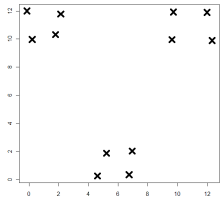
Hiérarchique descendante : le processus de la hiérarchique ascendante peut s'inverser. On part d'un seul clusters que l'on coupe en deux. Puis on coupe encore jusqu'à obtenir des individus. Cette méthode est toutefois plus complexe à mettre en œuvre puisque choisir la meilleure manière de couper un clusters en deux est quelque chose de complexe.

3.2 k-means

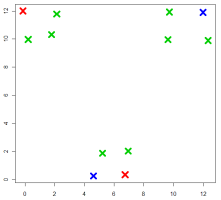
On commence par choisir le nombre de groupe. Puis on répartit aléatoirement tous les individus dans l'un des groupes. Commence alors une alternance de 2 phases : espérance et maximisation.
– Espérance : le centre de gravité de chaque groupe est calculé
– Maximisation : chaque individu est affecté au groupe dont il est le plus proche
L'algorithme s'arrête quand plus aucun individu ne change de cluster.

Exemple :

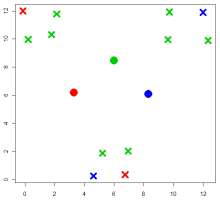
1. : on veut partitionner ces 12 points en trois clusters.



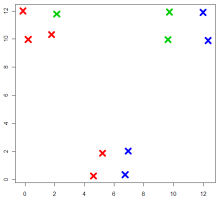
2. Randomisation : chaque individu est aléatoirement affecté à un des groupes.



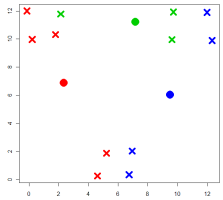
3. Espérance : Les centres de gravité de chaque groupe sont calculés.



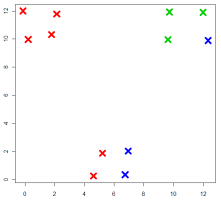
4. Maximisation : Chaque individu est affecté au groupe dont il est le plus proche.



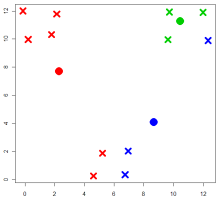
5. Espérance : Les nouveaux centres de gravité sont calculés.



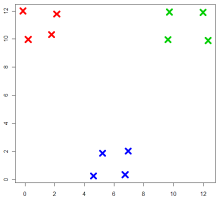
6. Maximisation : Les individus sont ré-affectés.



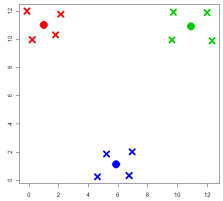
7. Espérance : Et encore.



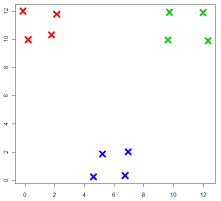
8. Maximisation : Et toujours.



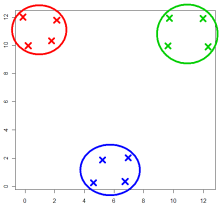
9. Espérance : Et encore.



10. Maximisation : Et toujours.



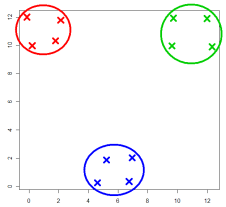
11. Fin : Comme rien ne change entre 8 et 10, l'algorithme s'arrête. Les clusters sont formés.



Problème : comment trouver le "bon" nombre de clusters ? Premier élément de réponse, le "bon" nombre de clusters est une notion subjective, il n'est pas vraiment défini. Néanmoins, comme on a quand même besoin de choisir, on utilise un critère. Le meilleur du moment est **Calinski & Harabatz** : $C(k) = \frac{Trace(V_{Inter})}{Trace(V_{Intra})} \cdot \frac{n-g}{g-1}$.
Il se base sur la division de $Trace(V_{Inter})$ par $Trace(V_{Intra})$. V_{Inter} représente la matrice des séparation entre clusters. Si les clusters sont bien distincts les uns des autres, V_{Inter} prendra des valeurs élevées. Si les clusters sont proches, V_{Inter} prendra des valeurs faibles. À rebourg, V_{Intra} mesure la compacité des clusters. Si les clusters sont compacts, V_{Intra} prendra des valeurs élevées. Si les clusters sont dispersés, V_{Inter} prendra des valeurs faibles.
Au final, $C(k)$ est grand quand les clusters sont compacts et bien séparés ; $C(k)$ est petit quand les clusters sont dispersés et mal séparés.

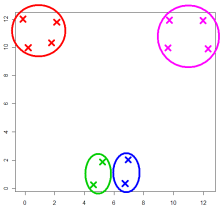
1. Compacts et bien séparés

V_{Inter} est grand
V_{Intra} est petit
Donc C est grand



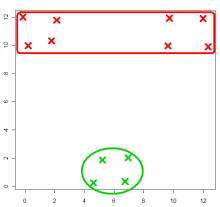
2. Compacts et mal séparés

V_{Inter} est petit
V_{Intra} est petit
Donc C est moyen



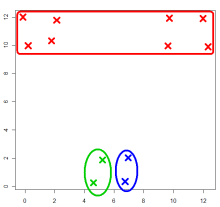
3. Non compacts et bien séparés

V_{Inter} est grand
V_{Intra} est grand
Donc C est moyen



4. Non compacts et mal séparés

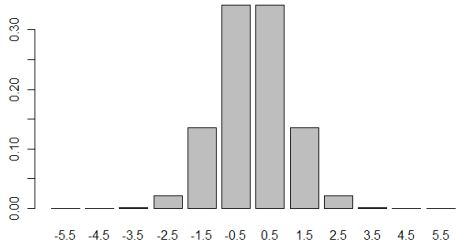
V_{Inter} est petit
V_{Intra} est grand
Donc C est petit



Au final, $C(k)$ est grand quand le découpage est compact et bien séparé. Il permet donc de trouver le "bon" nombre de clusters (ici, la préférence va au premier découpage plutôt qu'au deuxième ou au troisième) mais aussi de choisir entre deux découpages ayant le même nombre de clusters (le premier plutôt que le quatrième).

3.3 Modèles de mélange

Le partitionnement basé sur les modèles de mélange fait l'hypothèse que la population est en fait le mélange de plusieurs sous populations, sous-populations que la méthode se propose d'identifier.
Dans cette section, nous poussons un peu plus loin la simplification du problème puisque nous considérons que le partitionnement se fait sur une unique variable continue. Naturellement, tout se généralise à un nombre quelconque de variables. Nous allons également considérer les distributions dans leur forme discrète. Ainsi, la loi normale centré réduite discrète est :

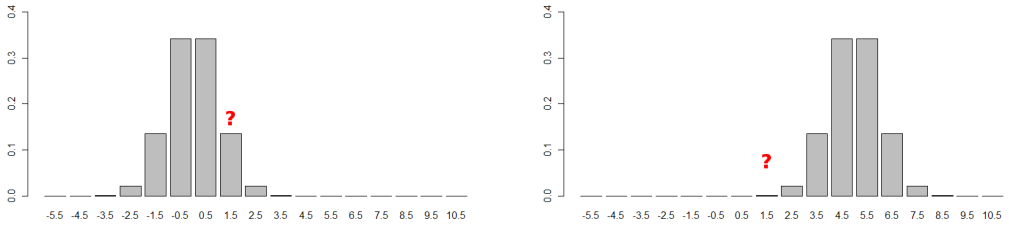


Intervalle	Probabilité
[0; 1[	0.341
[1; 2[	0.136
[2; 3[	0.021
[3; 4[	0.001
[4; 5[	0.0003

Vraisemblance d'un individu : La classification par modèle de mélange se base sur la notion de **vraisemblance**. La vraisemblance est une fonction qui prend deux arguments : un individu et une population (dont on connaît la distribution). Étant donné une population et un individu, la vraisemblance prend une valeur élevée si l'individu appartient vraisemblablement à la population, une valeur faible sinon. Ainsi, si on a un individu et DEUX populations, la vraisemblance peut nous permettre de "deviner" la population dont l'individu est probablement issu.

Exemples "de bon sens" : un individu aux cheveux long provient plus probablement de la population "femme" que de la population "homme". Un patient atteint d'un cancer des poumons provient plus probablement de la population "fumeur" que "non-fumeur". Un patient avec fièvre, douleurs et maux de gorges provient plus probablement de la population "rhume" que "appendicite".

Exemple "mathématique" : considérons deux populations et un individu :
La première suit la loi normale discrétisée $\mathcal{N}_D(0,1)$; la deuxième suit la normale discrétisée $\mathcal{N}_D(5,1)$. Considérons un individu dont la variable vaut 1,5. La première population a 13,6% de chance de produire un tel individu. La deuxième population a 0,1% de chance de produire un tel individu.



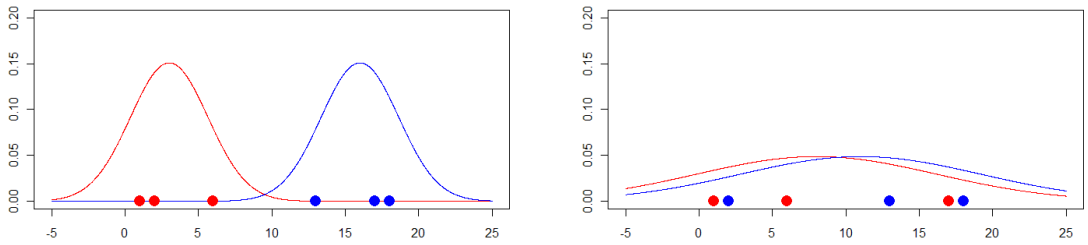
On peut donc raisonnablement penser que notre individu est issu de la première population.

Vraisemblance d'un groupe d'individu : Le processus est généralisable à un groupe d'individus en multipliant les vraisemblances individuelles afin d'obtenir la vraisemblance globale : ainsi, la vraisemblance que le groupe d'individus $\{0.5,1.5,3.5\}$ appartienne à la première population est de $0.34 \times 0.13 \times 0.001 = 6e^{-5}$ alors que la vraisemblance qu'il appartienne à la deuxième population est de $0.00003 \times 0.001 \times 0.13 = 5e^{-9}$. Le groupe est donc plus vraisemblablement issu de la première population.

Maximiser la vraisemblance : maintenant que la vraisemblance d'un groupe est définie, on peut **maximiser la vraisemblance**. On dispose d'un groupe d'individus. On suppose qu'ils sont issus d'une distribution que l'on ne connaît pas, mais dont on connaît néanmoins la forme générale. Par exemple, on sait que le groupe est issu d'une loi normale dont on ignore la moyenne et l'écart type. **Maximiser la vraisemblance**, c'est trouver les paramètres de la loi (moyenne et écart type dans notre cas) tels que la vraisemblance soit maximale. On trouve ainsi la loi dont, le plus vraisemblablement, est issu notre groupe.

Partitionner en utilisant le maximum de vraisemblance : Retour à notre problème initial, le partitionnement. Nous l'avons vu dans l'introduction, l'hypothèse des modèles de mélange est que la population que nous souhaitons partitionner est en fait le mélange de plusieurs sous-populations. La méthode du maximum de vraisemblance consiste à considérer plusieurs découpages possibles, puis à évaluer leur vraisemblance. La partition finale sera celle qui aura la plus grande vraisemblance globale. La logique interne de cette méthode est qu'un cluster homogène aura une grande vraisemblance alors qu'un cluster hétérogène aura une petite vraisemblance.

Exemple : Considérons les points 1,2,6,13,17 et 18. Considérons deux découpages : le premier en deux partitions {1,2,6} et {13,17,18}. Le deuxième en {1,6,17} et {2,13,18}. Dans le premier cas, les groupes sont probablement issus des lois $\mathcal{N}(3, 2.6)$ et $\mathcal{N}(16, 2.6)$. Dans le deuxième cas, les groupes sont probablement issus des lois $\mathcal{N}(8, 8.2)$ et $\mathcal{N}(11, 8.2)$:



Clairement, sur cet exemple, il est bien plus probable que le "bon" découpage soit le premier : sur le graphe de gauche, les trois points rouges sont regroupés et la loi normale $\mathcal{N}(3, 2.6)$ est une population dont on peut raisonnablement penser qu'ils sont extraits. La vraisemblance est haute.
Sur le graphe de droite, les trois points rouges sont dispersés. La loi normale $\mathcal{N}(8, 8.2)$ n'est pas une très bonne candidate pour extraire les points rouges, mais dans cette configuration, c'est la meilleurs possible. La vraisemblance est basse.
Au final, on peut donc raisonnablement penser que le découpage de gauche est meilleur que celui de droite.

Ainsi fonctionnent les modèles de mélange basés sur le maximum de vraisemblance.

A Codes R

Codes utilisés pour tracer les graphes de ce document.

A.1 Classification hiérarchique ascendante

```
###############
### Exemple de classification hierarchique

### Définition des points
a <- c(0,0)
b <- c(0,1.2)
c <- c(3,2)
d <- c(4,1)
e <- c(4,0)
```

```
### Etape 1
M1 <- rbind(a,b,c,d,e)
plot(M1,cex=2,pch=16,xlim=c(-1,5),ylim=c(-1,3),xlab="",ylab="")
dist(M1)
de <- apply(M[4:5,],2,mean)

### Etape 2
M2 <- rbind(a,b,c,de)
plot(M2,cex=c(2,2,2,3),pch=16,xlim=c(-1,5),ylim=c(-1,3),xlab="",ylab="")
dist(M2)
ab <- apply(M[1:2,],2,mean)

### Etape 3
M3 <- rbind(ab,c,de)
plot(M3,cex=c(4,2,3),pch=16,xlim=c(-1,5),ylim=c(-1,3),xlab="",ylab="")
dist(M3)
cde <- apply(M[3:5,],2,mean)

### Etape 4
M4 <- rbind(ab,cde)
plot(M4,cex=c(4,5),pch=16,xlim=c(-1,5),ylim=c(-1,3),xlab="",ylab="")
dist(M4)
abcde <- apply(M,2,mean)

### Etape 5
M5 <- rbind(abcde)
plot(M5,cex=c(9),pch=16,xlim=c(-1,5),ylim=c(-1,3),xlab="",ylab="")
dist(M4)
```

A.2 k means

```
#########################
### K means

### 1/ Définition de la population
a <- c(0,0,0,2,2,0,2,2)
M <- jitter(matrix(c(a+c(0,10),a+c(5,0),a+c(10,10)),2))
plot(t(M),xlab="",ylab="",col=1,pch=4,lwd=5,cex=2)

### Initialisation du vecteur couleur, utile pour les représentation graphiques
color <- rep(0,12)

### 2/ Répartition aléatoire des individus dans les groupes
alea <- floor(runif(12,1,4))
G1 <- (1:12)[alea==1]
G2 <- (1:12)[alea==2]
G3 <- (1:12)[alea==3]

### 2b/ Représentation graphique des individus
color[G1] <- 2
color[G2] <- 3
color[G3] <- 4
plot(t(M),xlab="",ylab="",col=color,cex=2,pch=4,lwd=5)

### 3/ Calcul des centres de gravité
C1 <- apply(M[,G1,drop=F],1,mean)
C2 <- apply(M[,G2,drop=F],1,mean)
C3 <- apply(M[,G3,drop=F],1,mean)
```

```
### 3b/ Représentation graphique des centres de gravité
points(t(C1),col=2,cex=3,pch=16)
points(t(C2),col=3,cex=3,pch=16)
points(t(C3),col=4,cex=3,pch=16)

### 4/ Calcul de la distance entre les individus et chacun des centres
distG1 <- apply(M,2,function(x)dist(rbind(C1,x)))
distG2 <- apply(M,2,function(x)dist(rbind(C2,x)))
distG3 <- apply(M,2,function(x)dist(rbind(C3,x)))

### 4b/ Ré affectation de tous les individus
G1 <- (1:12)[distG1 <= pmin(distG2,distG3)]
G2 <- (1:12)[distG2 <= pmin(distG1,distG3)]
G3 <- (1:12)[distG3 <= pmin(distG1,distG2)]

### Puis retour au 2b
```

A.3 Modèle de mélange

```
###############
### Exemple de densité selon les clusters

### Définition des points
a <- c(1,2,6,13,17,18)
x <- (-500:2500)/100


### Densités du découpage correcte
plot(0,ylim=c(0,0.2),xlim=c(-5,25),ylab="",xlab="",type="n")
m1a <- mean(a[1:3])
s1a <- sd(a[1:3])
lines(x,dnorm(x,m1a,s1a),type="l",col=2)

m1b <- mean(a[4:6])
s1b <- sd(a[4:6])
lines(x,dnorm(x,m1b,s1b),type="l",col=4)
points(a,rep(0,6),col=c(2,2,2,4,4,4),pch=16,cex=2)


### Densités du découpage incorrect
plot(0,ylim=c(0,0.2),xlim=c(-5,25),ylab="",xlab="",type="n")
m2a <- mean(a[c(1,3,5)])
s2a <- sd(a[c(1,3,5)])
lines(x,dnorm(x,m2a,s2a),type="l",col=2)

m2b <- mean(a[c(2,4,6)])
s2b <- sd(a[c(2,4,6)])
lines(x,dnorm(x,m2b,s2b),type="l",col=4)
points(a,rep(0,6),col=c(2,4,2,4,2,4),pch=16,cex=2)
```