Rapport SY09 TP 2

Analyse factorielle d'un tableau de distances, Classification automatique

SY09p019
Shuhan LIN
Grégory Mayemba

Ex1. Analyse factorielle d'un tableau de distance

Cet exercice a pour but de nous faire comprendre la méthode de l'AFTD en l'utilisant sur des données suivantes:

$$X = \begin{pmatrix} 8.5 & 1.5 \\ 3.5 & 5.0 \\ 2.0 & 6.5 \\ 9.5 & 1.5 \\ 8.5 & 2.5 \\ 3.0 & 6.5 \\ 9.0 & 2.5 \\ 2.0 & 5.5 \end{pmatrix}$$

Question 01

On centre en colonne la matrice X puis on utilise la fonction "dist ()" qui calcule le tableau D des distances euclidiennes associé aux données.

On utilise la fonction "as.matrix ()" sur le tableau de distances pour le mettre sous forme de matrice.

On multiplie cette matrice D pour avoir la matrice des distances euclidiennes D^2

Question 02

On calcul W de deux manières différentes, la première, à partir de la formule:
$W = X \times X^t$

Cette manière suppose d'avoir les données sous forme de matrice comme c'est le cas dans cet exercice. En générale on réalise l'AFTD car on ne dispose qu'un tableau de distances, dans ce cas, on utilise la formule:

$W = -0.5 \times Qn \times D^2 \times Qn$
Avec $Qn = In - U_n/n$

Question 03

On diagonalise la matrice W avec la fonction "Eigen()" ,les valeurs propres sont:

1.39e+01	2.21e-01	1.75e-15	1.054e-15
7.51e-17	6.16e-18	-3.46e-17	-4.35e-17

Les valeurs propres 7 et 8 sont négatives, mais elle sont proches de zéro (10^(-17)) et l'on peut donc les considérer comme nulle. Les valeurs propres deviennent donc:

1	2	3	4	5	6	7	8
1.39 e+01	2.21 e-01	1.75 e-15	1.054 e-15	7.51 e-17	6.16 e-18	0	0

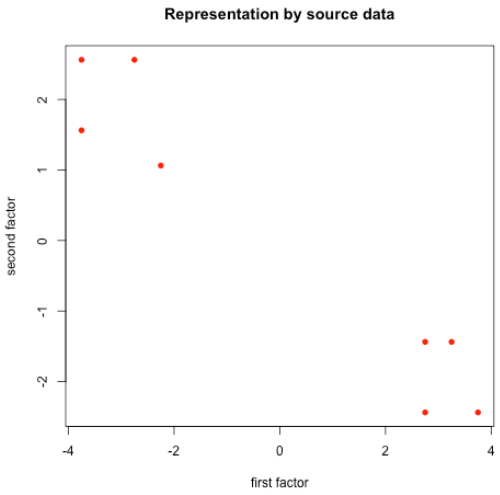
On considère donc la matrice $\frac{1}{n}W$ comme semi défini positive.

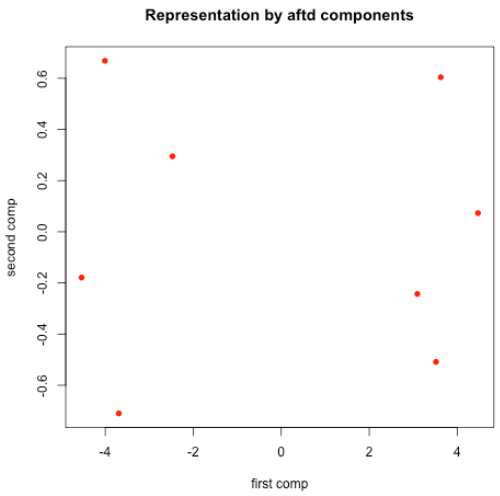
Question 04

La matrice des vecteur propres et la matrice diagonale des valeurs propres sont donnée par la fonction "eigen()"

avec les 2 dernières valeurs propres à 0.

Question 05&06

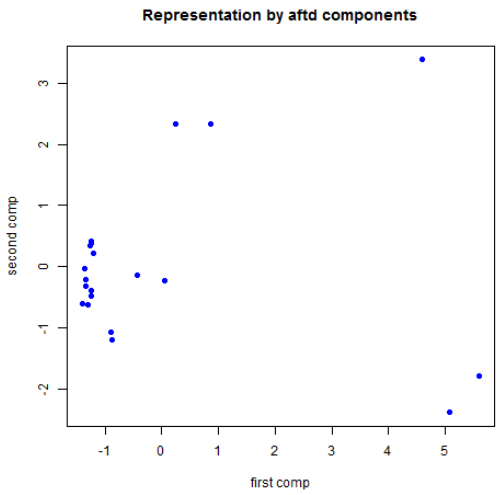


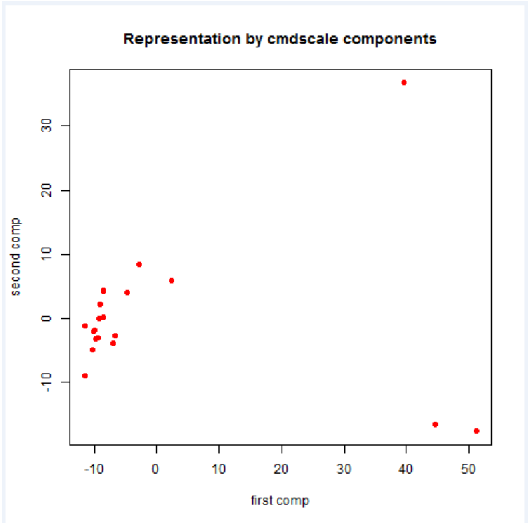


On distingue 2 groupe sur les 2 graphiques. Sur La représentation associé au tableau initiale, les points appartenant a un même groupe sont beaucoup plus proche, ce qui donne l'aspect d'avoir deux "paquet" sur le graphique. Il semble que les axes factorielles ne soient pas les même sur les 2 graphiques, c'est pour cela, que les représentation sont différentes.

Données Mutation

Nous appliquons la méthode AFTD sur le jeu de données de la mutation des espèces





Question 01

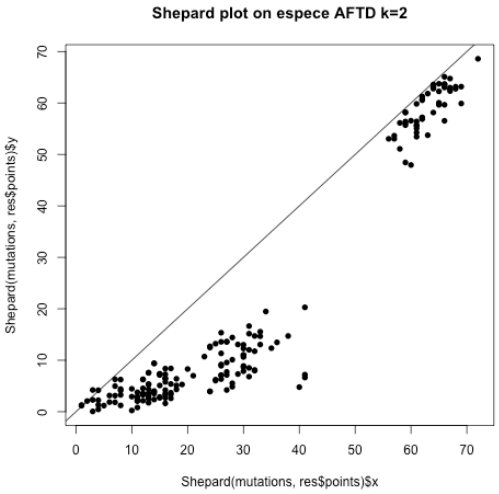
Les résultats sont quasiment similaires à un facteur d'échelle près. Certains points ne sont pas aux mêmes endroits mais on retrouve les mêmes groupes de points sur les 2 graphiques.

Question 02

Pour étudier la qualité de représentation, nous utilisons le graphe Shepard sur notre résultat obtenu par AFTD.

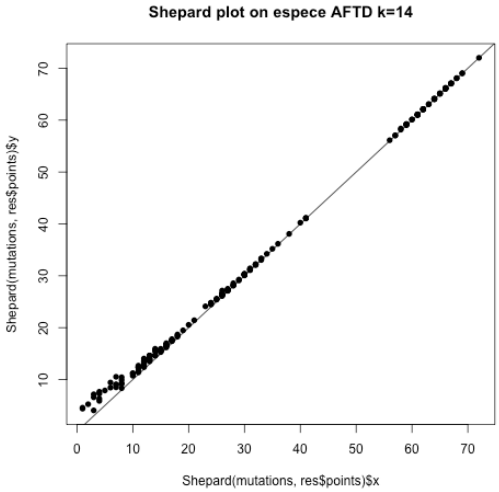
Le digramme shepard se servi de mesurer la cohérence entre les distances originales des données et les distances obtenues par AFTD.

Nous obtenons le résultat pour un AFTD de K = 2.



Nous trouvons que le résultat n'est pas très fidèle, car une partie de points n'est pas très proches du droit.

Si nous avons effectué le shepard avec un AFTD de toutes les dimensions, nous obtenons un graphe comme ci-dessous :



Nous trouvons que maintenant presque tous les pointes sont proches du droit. Une représentation de K = 14 est fidèle et nous assurons qu'il ne perd pas des informations

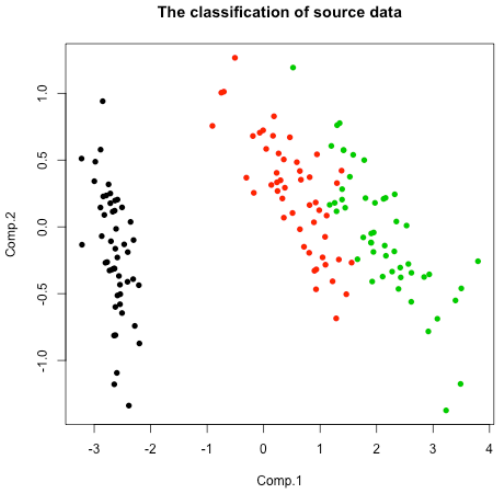
Ex2 Méthode de Centre mobile

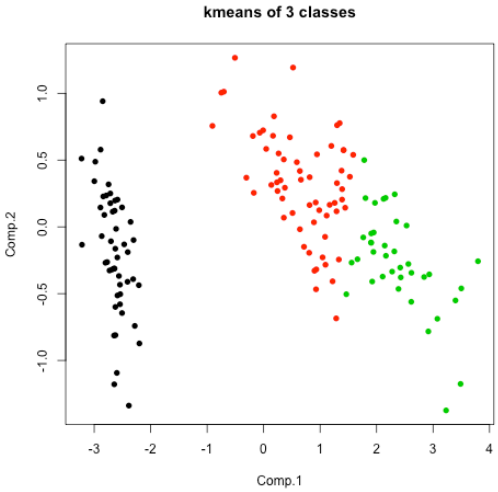
Données Iris

Question 01

En regardant les données de source, nous savons qu'il y a trois espèces de l'iris : setosa, versicolor et virginica.

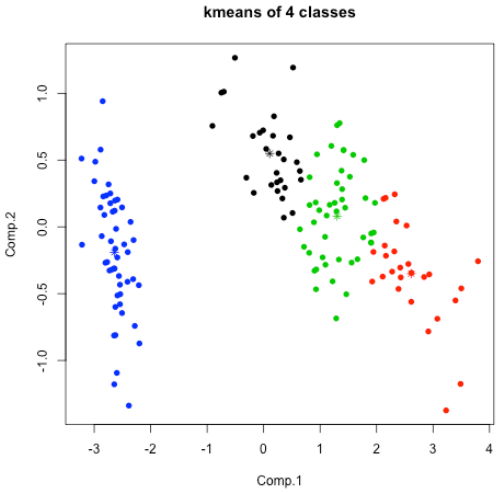
Nous avons d'abord désigné un graphe sur le jeu de données source.

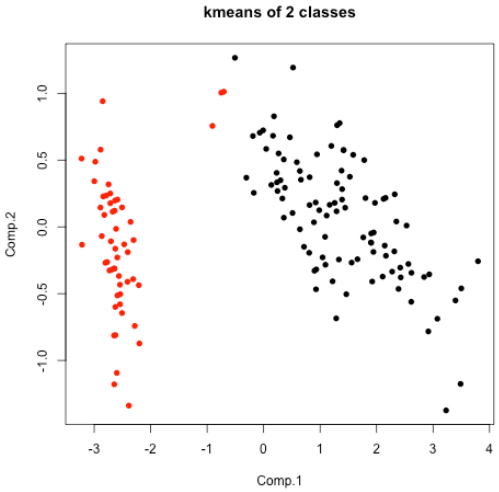




En appliquant la fonction k-means, nous avons calculé la partition des données Iris en K classes. $K = \{2,3,4\}$

Nous avons obtenu les graphes comme ci-dessous:





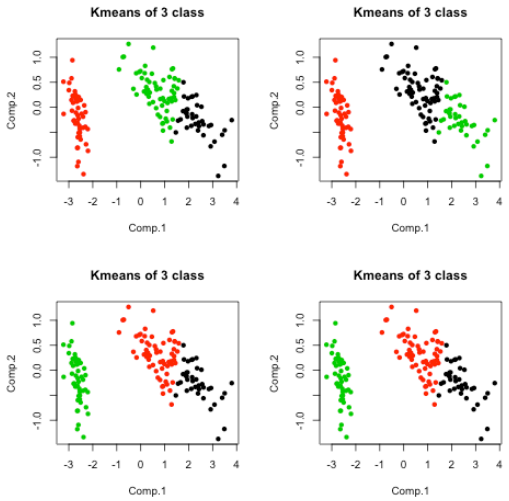
Pour le résultat de K-means en 2 classes, il mélange 2 classes des espèces comme un espèce.

Pour le résultat de K-means en 3 classes, nous avons obtenu une partition plus similaire que celle en réel.

Pour le résultat de K-means en 4 classes, il prend une partie de classe2 et une partie de classe3 pour générer une 4eme classes.

Question 02

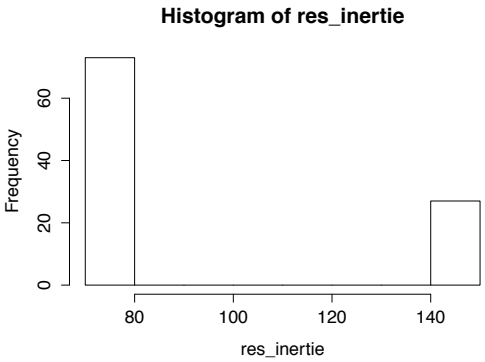
Pour K=3, nous avons effectué 4 fois le K-means pour étudier sa stabilité.



D'après les graphes, il paraît que le résultat est stable. Les partitions des classes sont les mêmes.

Ensuite, nous lançons le k-means en 3 classes 100 fois pour regarder l'inertie intra-classes totale.

Nous avons effectué un histogramme pour illustrer le résultat.



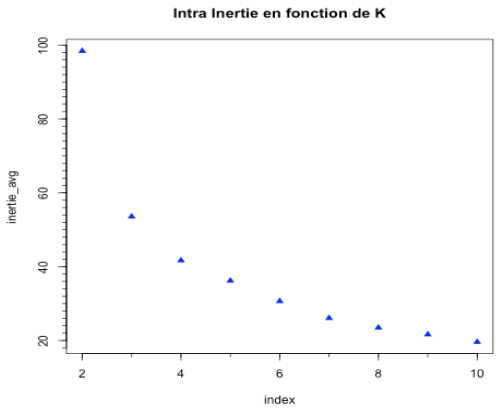
Nous constatons que l'inertie intra classes totale a deux valeurs, par conséquent, cette partition n'est pas stable.

Question 03

D'abord, nous utilisons *Sample* (), une fonction in R pour faire l'échantillon. Elle nous rend un échantillon iid de 100 individus.

Ensuite, à partir de cet échantillon, nous avons appliqué le k-means pour K de 2 à 10. Pour chaque valeur de K, nous avons lancé 100 fois et nous avons calculé la moyenne de l'inertie intra-classes totale pour chaque K.

En finale, nous avons désigné un graphe à représenter le résultat.



Pour le choix du nombre de classes, théoriquement, nous devons choisir la partition dont l'inertie intra-classe totale est le plus faible. Par conséquent, nous devons choisir K=1, dont l'inertie vaut 0.

Cependant, si nous choisissons K = 1, il n'y a aucun sens, parce que nous n'avons pas fait la partition. Par résultat, nous déduisons que l'inertie intra classe n'est pas le seul standard à estimer la qualité d'une partition.

Le standard est que nous devons

choisir une partition qui n'a pas trop de classes et l'inertie est assez faible.
A niveau du graphe, nous devons choisir le 'Coude' de la courbe.

Donc, nous pouvons choisir une partition de 4 classes ou 5 classes.

Question 04

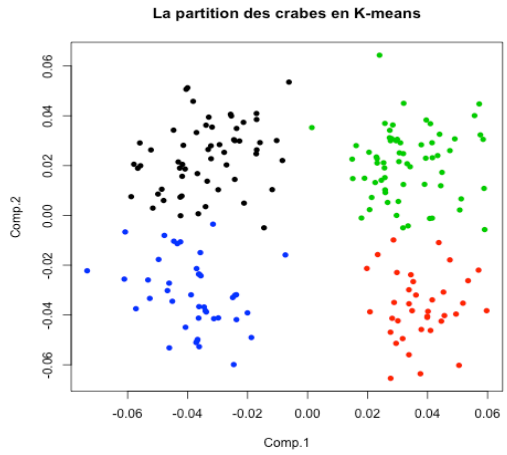
Dans la question 01, nous avons déjà désigné le graphe en réel et le graphe de partition en 3 classes.

D'après ces deux graphes, nous avons observé que la première classe (la classe à gauche) est bien partitionné.
Cependant, pour les deux classes à droite, il y a quelques pointes mal classées.

Données Crabes

Question 01

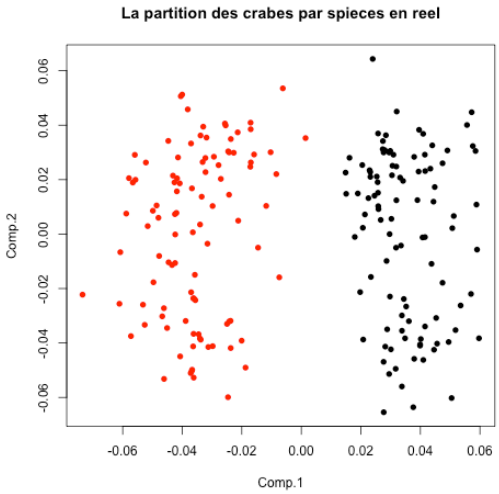
Selon l'exercice précédent, nous savons bien que les données de crabes sont partionnées en 4 classes. Nous avons donc effectué un kmeans de 4 classes.



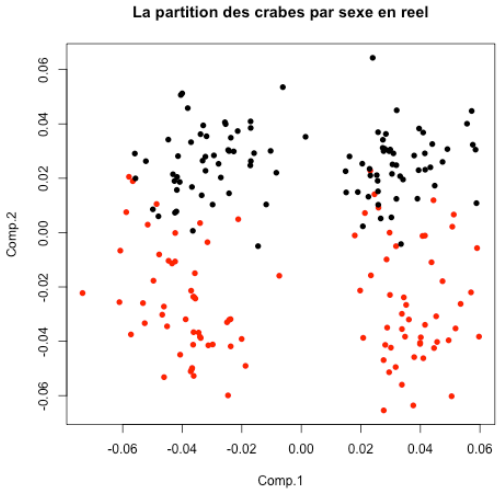
Question 02

Nous avons ensuite observé la partition en réel.

Graphe séparé par espèce



Graphe séparé par sexe



D'après ces trois graphes, nous remarquons que le résultat de kmeans est la même avec celle en réel. Il y a juste quelques points mal classés.

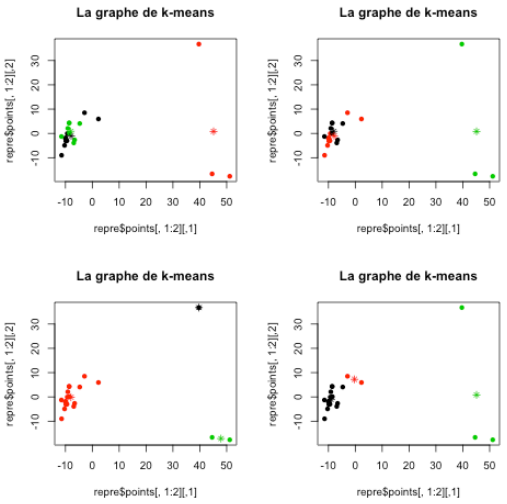
Donnée Mutation

Comme une préparation, nous avons effectué AFTD à notre jeu de données et nous avons gardons 5 premières

dimensions.

Question 01

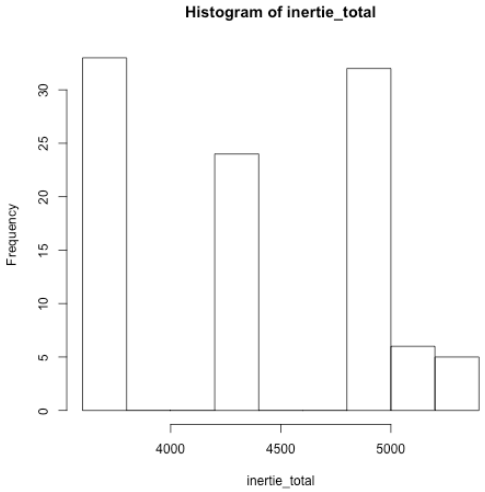
Nous avons effectué le kmeans 4 fois et le représenté dans le premier plan.



Le résultat des ces partitions n'est pas les mêmes. Cela veut dire que ce n'est pas une partition stable.

Question 02

Nous avons encore effectué le k-means 100 fois et nous calculé des inertie intra-classes totale. Ensuite, nous avons désigné le histogramme.



Nous constatons qu'il y a plusieurs valeur de l'inertie intra classes, donc, cette classes n'est pas stable.

Conclusion

Premièrement, ce TP nous permet de comprendre le mécanisme de l'AFTD et l'appliquer sur un jeu de données en réel. (Les mutations des espèces).

Deuxièmes, nous avons appliqué la méthode de la classification automatique, K-means, dans trois jeux de données : Iris, Crabes et également mutation. Nous avons étudié la stabilité d'une classification et le choix du nombre de classes. D'ailleurs, nous avons comparé la classification de la méthode centre mobile avec la vraie classification, pour voir la qualité de classification.