SY09 - Analyse des données et Data-Mining

Printemps 2017 (P17)

TP2

Classification automatique

Maxime Churin GI04 - Louis Denoyelle GI04

Introduction

La classification automatique consiste à regrouper les individus d'un jeu de données avec ses semblables dans différentes classes. Les méthodes utilisées sont des algorithmes se basant sur l'apprentissage automatique. Certaines de ces données comme `Iris` et `Crabs` ont déjà été étudiées dans les TP précédents, nous nous appuyerons donc sur les connaissances et analyses de ces jeux de données. Au contraire de `Mutations` qui contient des données nouvelles. Nous effectuerons tout d'abord une visualisation des données à travers des ACP et AFTD. Ensuite, les différents types de classifications hiérarchiques seront étudiés. Enfin, les résultats de la méthode des centres mobiles, effectuée sur ces données, seront analysés.

1 Visualisation des données

Dans cette première partie, nous allons réutiliser les méthodes de visualisation vues au premier TP et les compléter par l'analyse factorielle d'un tableau de distances ou AFTD sur différents jeux de données.

1.1 Données Iris

Le jeu de données des Iris de Fisher contient 150 observations décrites par 4 variables quantitatives : la longueur et la largeur du sépale et du pétale plus une variable qualitative qui indique l'espèce. On effectue un léger nettoyage des données en ne gardant qu'en un seul exemplaire les lignes strictement identiques et utiliser les outils de la plateforme R.

On réalise donc une ACP et une AFTD sur ce jeu de données. Les représentations des résultats de ces analyses étant similaires à une rotation près, on montre uniquement les résultats de l'ACP.

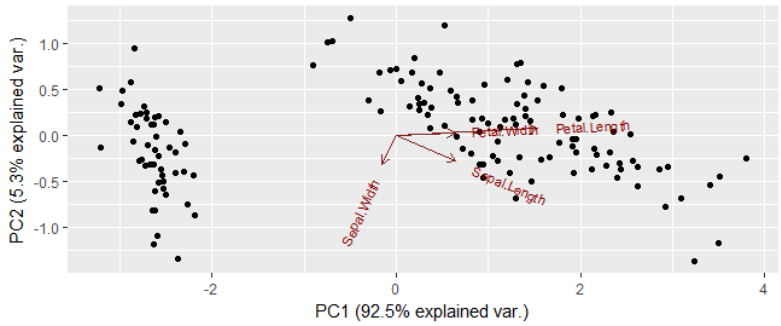


Figure 1.1 – ACP sur les individus (iris) dans le premier plan.

Une première approche consiste à ne pas tenir compte de l'espèce. On constate que l'inertie est essentiellement concentrée sur le premier axe avec près de 93%. Il y a donc une corrélation très importante entre toutes les variables. On pourrait, comme dans le TP1, atténuer cette corrélation mais la représentation nous permet tout de même de conclure notre étude avec l'observation de deux groupes de points. Le groupe le plus à gauche étant très compacte alors que celui de droite est plus dispersé.

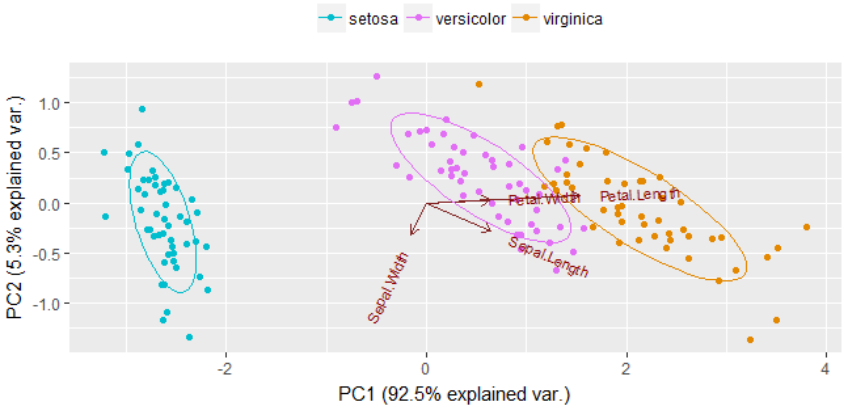


Figure 1.2 – ACP sur les individus (iris) dans le premier plan.

Grâce à l'information *espèce* et aux couleurs, on observe maintenant trois groupes. On constate que les espèces *Versicolor* et *Virginica* ont des caractéristiques moyennes assez proches, il était donc très difficile de les distinguer dans la représentation précédente. La recherche d'une partition optimale des données va donc être différente en fonction des méthodes utilisées, on pourra s'attendre à deux ou trois classes avec des répartitions de points différentes.

1.2 Données Crabs

Le jeu de données `Crabs` à déjà été largement présenté au TP1. Nous nous intéresserons donc ici qu'aux résultats de l'AFTD, une ACP ayant déjà été réalisée.

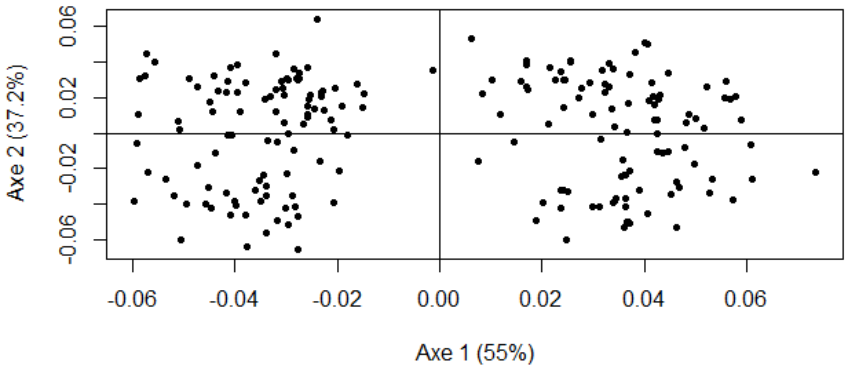


Figure 1.3 – AFTD sur les individus (crabs) dans le premier plan.

Sur cette première représentation de l'AFTD, on ne tient pas compte ni de l'espèce *B/O* ou du sexe *M/F* des individus. La qualité de la représentation est bonne car l'inertie cumulée des deux axes est haute et bien répartie (cf. traitement effectué sur l'ACP du TP1). Pour obtenir l'inertie on a gardé uniquement les valeurs propres positives issues de l'AFTD. On observe deux groupes distincts mais très dispersés, l'un à droite et l'autre à gauche.

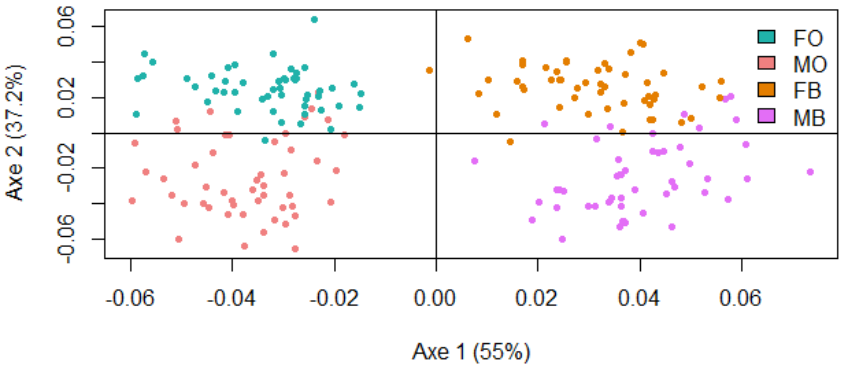


Figure 1.4 – AFTD sur les individus (crabs) dans le premier plan.

Grâce aux informations *espèce* et *sexe* et aux couleurs, on observe maintenant quatre groupes. On constate que les femelles se trouvent sur la partie horizontale haute du graphique alors que les mâles sont regroupés dans la partie horizontale basse. On peut tirer des conclusions similaires sur la répartition via les espèces : les individus de l'espèce *B* se trouvant sur la partie verticale droite et les individus de l'espèce *O* sur la partie verticale gauche. Enfin, on constate que sur ce graphique la partition selon l'espèce parait plus pertinente.

1.3 Données Mutations

Le jeu de données `Mutations` représente un tableau de dissimilarités entre espèces construit à partir de la distance de *Manhattan*. Une ACP nécessitant un tableau individus/variables, on va nécessairement analyser ce jeu de données par une AFTD. On y retrouve le nombre de positions de la molécule qui ont un acide aminé différent de la chaîne de référence.

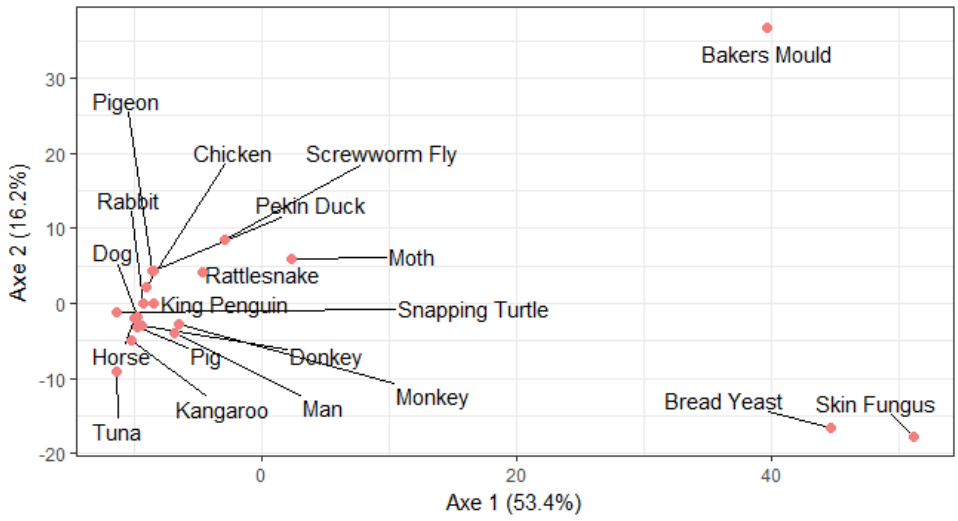
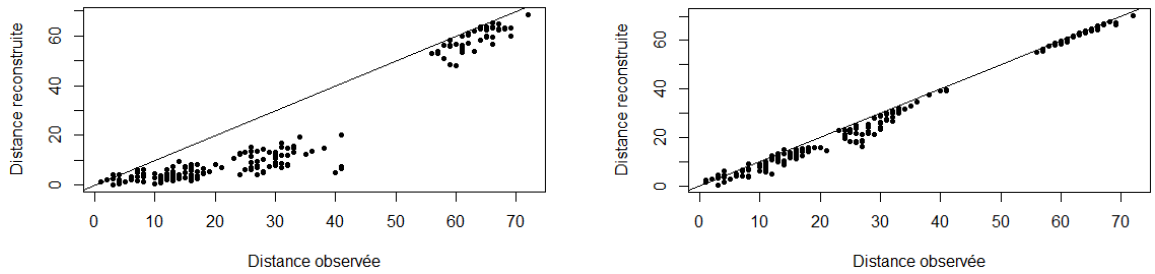


Figure 1.5 – Représentation euclidienne des données par l'AFTD (d=2).

Sur cette représentation, on constate un groupement de points sur la gauche et 3 espèces qui se démarquent par leur grand nombre d'acides aminés différents. Concernant la qualité de cette représentation, on l'obtient en gardant uniquement les valeurs propres positives issues de l'AFTD. Le pourcentage d'inertie pris en compte par le premier plan est donc d'environ 70%, ce qui est une qualité de représentation correcte.



(a) 2 variables et $R^2 = 0.92$.

(b) 5 variables et $R^2 = 0.99$.

Figure 1.6 – Diagrammes de Shepard.

Un diagramme de Shepard permet de vérifier la corrélation entre les distances observées et les distances reconstruites grâce au coefficient de corrélation R^2. On observe que plus on ajoute de variables plus on augmente ce coefficient de corrélation. Les points tendent alors à ce rapprocher de la droite ce qui est tout à fait logique car on reconstruit nos données. La matrice initiale étant obtenue à partir d'une distance L_1, on ne peut pas, en se limitant aux seuls vecteurs propres positifs, retrouver exactement la droite théorique linéaire croissante.

2 Classification hiérarchique

Dans cet exercice, les classifications hiérarchiques ascendante et descendante des jeux de données `Iris` et `Mutations` seront analysées. Ces classifications permettent de mettre en évidence des classes dans les jeux de données à travers l'analyse des dissimilarités entre les individus.

2.1 Données Mutations

La réalisation d'une classification hiérarchique ascendante sur le jeu de données `Mutations` permet de distinguer deux groupes évidents, l'un composé de 17 individus et l'autre de 3. On retrouve également des similarités évidentes entre les espèces, comme l'Homme qui est très proche du singe.

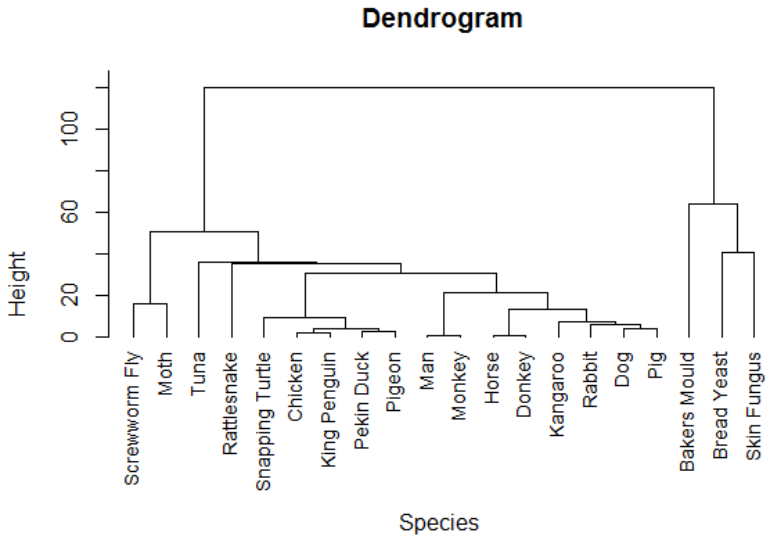


Figure 2.1 – Classification hiérarchique ascendante des données Mutations.

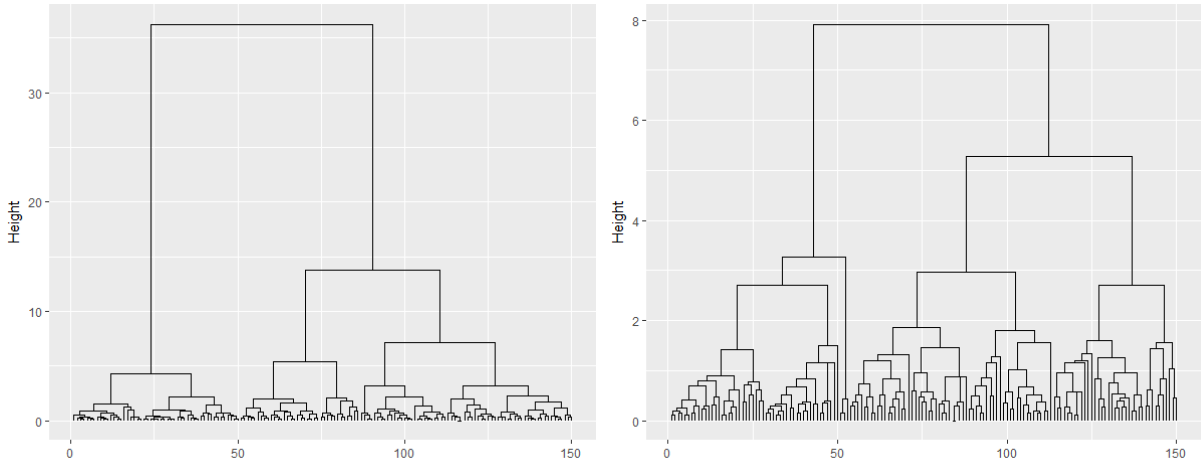
On retrouve dans l'essentiel les regroupements qui avaient pu être mis en évidence par l'AFTD, avec deux groupes bien distincts. Cependant, les trois individus à l'écart, *Bakers Mould*, *Bread Yeast* et *Skin Fungus* semblent bien plus proches sur cette représentation que sur celle de l'AFTD, où *Bakers Mould* était beaucoup plus éloigné des deux autres.

2.2 Données Iris

La classification hiérarchique ascendante des `Iris` (voir Figure 2.2a) nous donne une séparation très nette en deux groupes. Le premier groupe correspond à l'espèce *Setosa*, et le deuxième à la réunion des espèces *Versicolor* et *Virginica*. En effet, nous savons que ces deux espèces sont très proches de par leurs caractéristiques grâce aux études précédentes réalisées sur ce jeu de données. La distinction entre les espèces *Versicolor* et *Virginica* est en revanche moins nette, mais on la retrouve tout de même sur la partie droite du dendrogramme. De plus, on peut retrouver une autre séparation qui pourrait laisser présager l'identification de quatre groupes. Notre connaissance du jeu de données nous permet de savoir qu'il n'existe que trois espèces,

mais une personne devant analyser les données sur la base de ce dendrogramme seul pourrait se tromper.

La classification hiérarchique descendante des `Iris` (voir figure 2.2b) permet de distinguer plus nettement qu'il existe trois groupes qui correspondent aux trois espèces. On note toutefois que les hauteurs des regroupements sont ici plus élevées que précédemment, donc que les distances sont plus élevées qu'elles ne devraient l'être. On remarque également que trois individus se dégagent du groupe correspondant à l'espèce *Setosa*, ce qui n'apparaissait pas dans l'analyse de la classification hiérarchique ascendante et ne devrait pas arriver étant donné que nous savons que les individus de cette espèce sont très proches. C'est pour ces raisons que l'on peut considérer la classification hiérarchique descendante moins optimale que l'ascendante.



(a) Classification hiérarchique ascendante. **(b)** Classification hiérarchique descendante.

Figure 2.2 – Classification hiérarchiques des données Iris.

3 Méthode des centres mobiles

L'objectif de la méthode des centres mobiles est de partitionner en différentes classes des individus d'un jeu de données. Les individus les plus semblables seront regroupés dans un nombre K de classes détérminé au début de l'exécution de l'algorithme.

3.1 Données Iris

On effectue une partition en 2, 3 et 4 classes du jeu de données `Iris` :

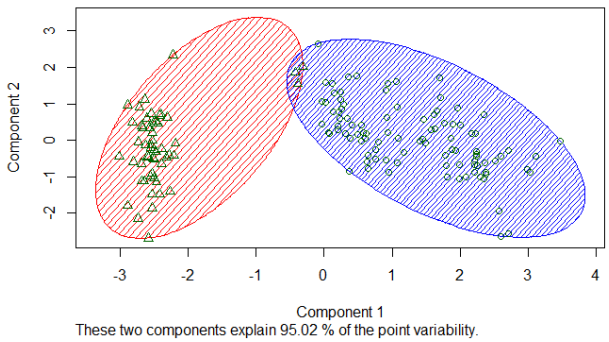


Figure 3.1 – Partition des données Iris en 2 classes.

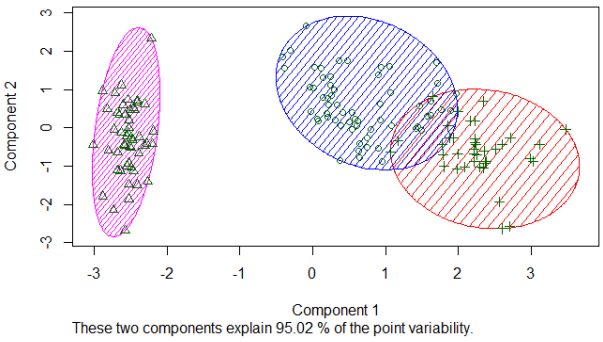


Figure 3.2 – Partition des données Iris en 3 classes.

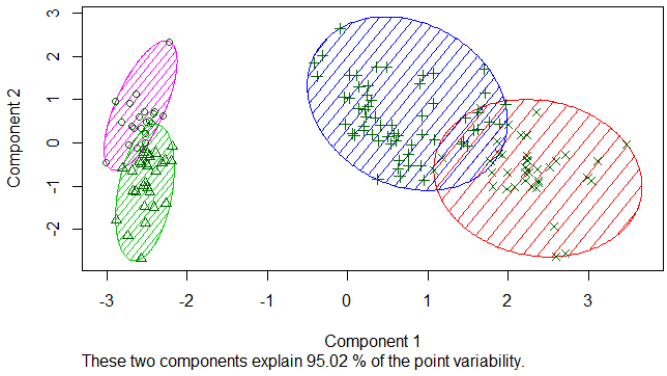


Figure 3.3 – Partition des données Iris en 4 classes.

La séparation en 2 classes amène aux classes suivantes : l'espèce *Setosa* et la réunion des espèces *Versicolor* et *Virginica*. Ce résultat est cohérent avec les observations faites précédemment sur le jeu de données.

La séparation en 3 classes donne une représentation qui correspond à la séparation du jeu de données en 3 espèces. On remarque cependant que cette représentation est légèrement différente. La séparation des espèces *Versicolor* et *Virginica* devrait ressembler à deux ellipses qui ne se croisent pas (voir l'ACP sur la Figure 1.2).

La séparation en 4 classes fait apparaître une nouvelle classe en séparant un des groupes déjà existants. En sachant que le jeu de données ne comporte que 3 espèces, on peut conclure que l'algorithme fait apparaître un nouveau groupe qui n'a pas lieu d'être.

En testant plusieurs classifications en 3 classes, on trouve 2 configurations possibles. Il est possible de trouver des partitions différentes car les premiers points choisis par l'algorithme au départ le sont de manière aléatoire, ce qui peut entraîner des résultats différents pour deux exécutions de cette méthode sur les mêmes données. La première partition correspond bien à la répartition des individus selon les 3 espèces. La deuxième, en revanche, est très différente et s'éloigne assez largement de la répartition selon les espèces. On obtient les résultats suivants pour les deux répartitions :

Taille des classes	Inertie intra-classe	Inertie totale
33 - 96 - 21	6.432121 - 118.651875 - 17.669524	142.7535
38 - 62 - 50	23.87947 - 39.82097 - 15.15100	78.85144

Table 3.1 – Table d'analyse des répartitions en 3 classes.

On remarque que les inerties totales des deux partitions sont très éloignées. En comparant les deux partitions, il est possible de retrouver que la répartition la plus cohérente est la 1^{e}, même en ne connaissant pas le jeu de données car son inertie intra-classe totale est beaucoup plus faible que celle de la 2^{e} partition.

On calcule l'inertie intra-classe minimale d'une partition pour un nombre de classes K allant de 1 à 10 :

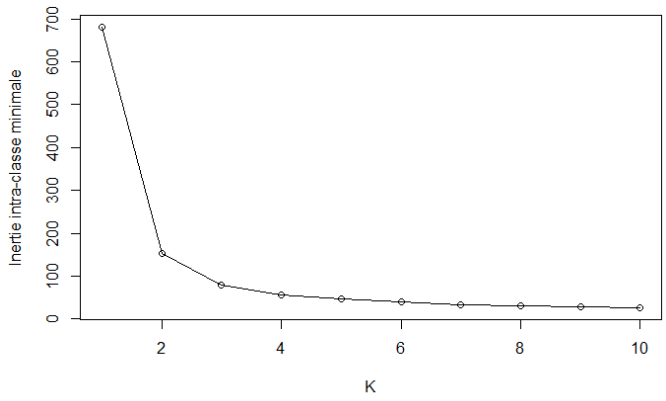


Figure 3.4 – Variation de l'inertie intra-classe minimale en fonction du nombre de classes K.

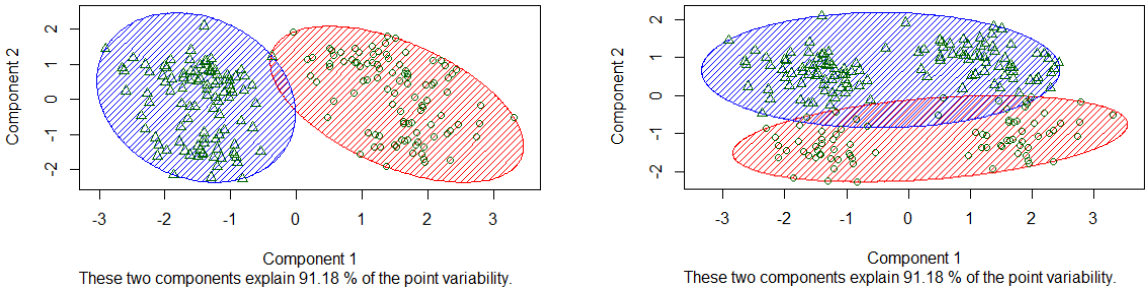
On remarque que l'inertie totale diminue très largement avec $K = 2$ classes. L'inertie diminue encore pour $K = 3$ classes, mais beaucoup moins fortement qu'auparavant. Ensuite, de 4 à 10 classes, l'inertie intra-classe diminue très faiblement et de façon linéaire. En utilisant la méthode du coude, il est possible de décider un nombre de classes $K = 2$ ou $K = 3$.

La partition obtenue avec la méthode des centres mobiles nous donne une partition en 2 ou 3 groupes. Ce résultat est cohérent, car même si nous savons qu'il existe 3 espèces, 2 de ces espèces ont des caractéristiques très proches. Ces deux espèces pourraient être confondues en un seul groupe, ce qui explique qu'une partition en deux groupes est possible. Nous obtenons également une répartition possible en 3 groupes. La partition possédant l'inertie minimale parmi

les partitions en 3 classes correspond globalement aux 3 espèces d'`Iris` même si on peut voir que cette partition n'est pas parfaite. En effet, les effectifs de cette partition nous renseignent déjà sur la différence avec les 3 espèces. L'algorithme nous renvoie 38, 62 et 50 individus pour les 3 classes alors que les individus sont répartis équitablement entre les espèces avec 50 individus par espèce.

3.2 Données Crabs

En effectuant une partition en 2 classes du jeu de données Crabs, on obtient les résultats suivants :



(a) Partition selon l'espèce. **(b)** Partition selon le sexe.

Figure 3.5 – Partitions des données Crabs en 2 classes.

On obtient essentiellement deux types de cluster. Ils correspondent aux partition par sexe et par espèce des individus du jeu de données `Crabs` que nous avions abordé pour la Figure 1.3. En générant ces deux partitions on a observé que la partition selon l'espèce est plus fréquente que la partition selon le sexe, elle doit donc être plus pertinente.

On réalise maintenant une partition en 4 classes :

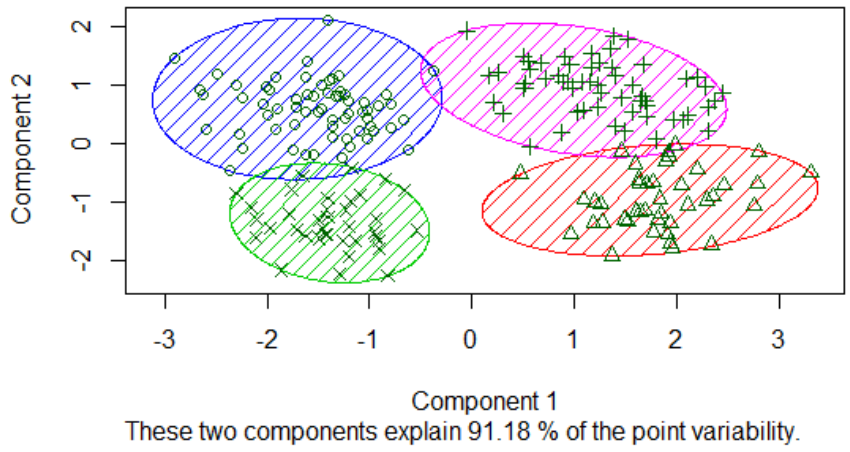


Figure 3.6 – Partition des données Crabs en 4 classes.

On constate la présence des 4 partitions en accord avec la Figure 1.4 qui classait les individus selon les différentes combinaisons sexe/espèce possibles. On voit donc qu'à travers ces deux partitions l'algorithme des centres mobiles est plutôt fidèle à nos prédictions même si elles étaient plutôt évidentes.

3.3 Données Mutations

Avant d'appliquer l'algorithme des centres mobiles au jeu de données on se place dans un espace à 5 dimensions qui reconstruit fiablement nos données (voir Figure 1.6). On a ensuite effectué plus de 200 classifications en $K = 3$ classes et observé six classifications différentes.

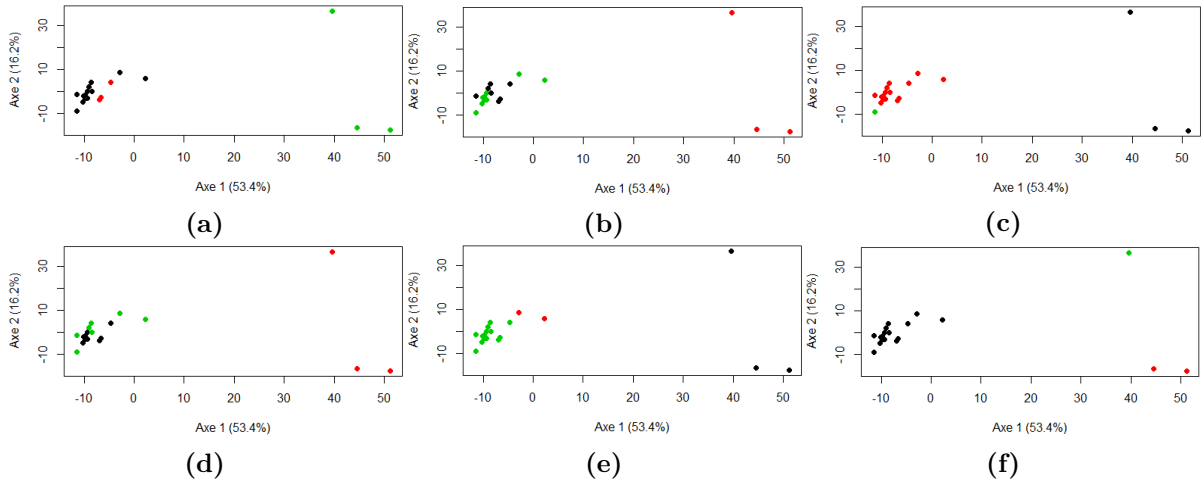


Figure 3.7 – Représentations euclidiennes des données classifiées par l'AFTD ($d = 5$).

On constate que les individus appartiennent à des classes très différentes d'une itération à l'autre de l'algorithme des centres mobiles. Les résultats sont donc très instables car à chaque itération on observe une partition différente. Cependant il est possible de déterminer la meilleure partition en trouvant la représentation minimisant l'inertie intra-classe totale :

Figure	Inertie intra-classe totale
a	4918.1
b	4975
c	5237.2
d	5092.5
e	4398.2
f	3621.9

Table 3.2 – Récapitulatif des inerties intra-classe totales des représentations de la Figure 3.8

C'est donc la partition de la Figure 3.8f qui est la meilleure même si les valeurs sont toutes très proches et qu'aucune des inerties intra-classe totales ne se détache vraiment. On constate qu'elle ne correspond pas à la partition obtenue via la classification hiérarchique ascendante (voir Figure 2.1).

Conclusion

L'étude des jeux de données de ce TP ont permis d'expérimenter la classification à travers des méthodes non supervisées. Si les partitions des données semblent parfois évidentes visuellement, ce n'est pas toujours le cas en pratique on obtient alors des partitions en classes qui sont différentes des groupements d'origine. Cependant, des méthodes comme la méthode du coude et l'étude de l'inertie intra-classe minimale permettent d'aider à retrouver les classifications les plus optimales.