Devoir Classification et Apprentissage session 1

Benoit Gachet, Guillaume Mulier

28/11/2020

Table of Contents

# Exercice classification non supervisée

## Introduction

L’objectif de notre devoir était de réaliser une classification non supervisée afin de caractériser l’appartenance des observations de notre base de données à k classes, en définissant le nombre de classes.

Notre base de données comportait un ensemble de 500 observations et 10 variables mesurées pour chacune des observations. Nous n’avions aucune donnée manquante pour l’ensemble de la base. Les ordres de grandeur des variables étaient similaires ; néanmoins, afin d’éviter qu’une variable ne prenne trop d’importance du simple fait de son unité de mesure, nous avons centré et réduit l’ensemble des variables.

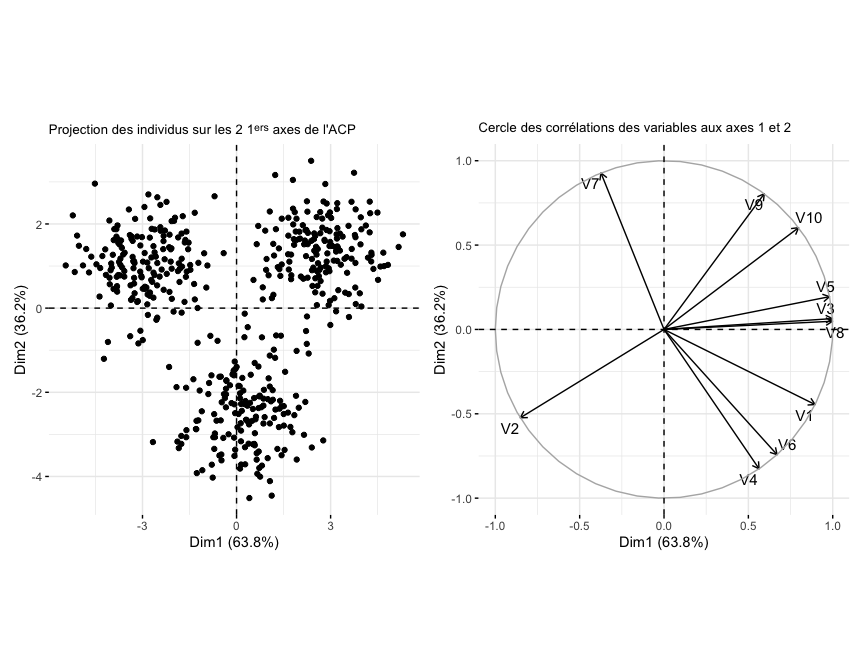
Il existe de nombreuses méthodes de classification non supervisées. Nous nous sommes intéressés à 3 méthodes : la classification hiérarchique ascendante, les k-means et les modèles de mélange Gaussiens. Nous les avons ensuite comparé et avons discuté les avantages et les inconvénients de chacune dans notre classification.

## Mise en oeuvre des modèles

### Description du jeu de données

Pour visualiser les données, nous avons mis en oeuvre une analyse en composantes principales. Les 2 1ères composantes principales expliquent la quasi-totalité de l’inertie du jeu de données. Nous utiliserons donc exclusivement ces 2 dimensions pour explorer le jeu de données.

En représentant les projections des individus sur ces 2 composantes, nous pouvons remarquer un nuage de points bien étallé. Et on a l’impression de distinguer 3 groupes de points dans ce nuage. Ces 3 groupes ont l’air d’être de forme sphérique, ce qui guidera nos choix de classifications par la suite.

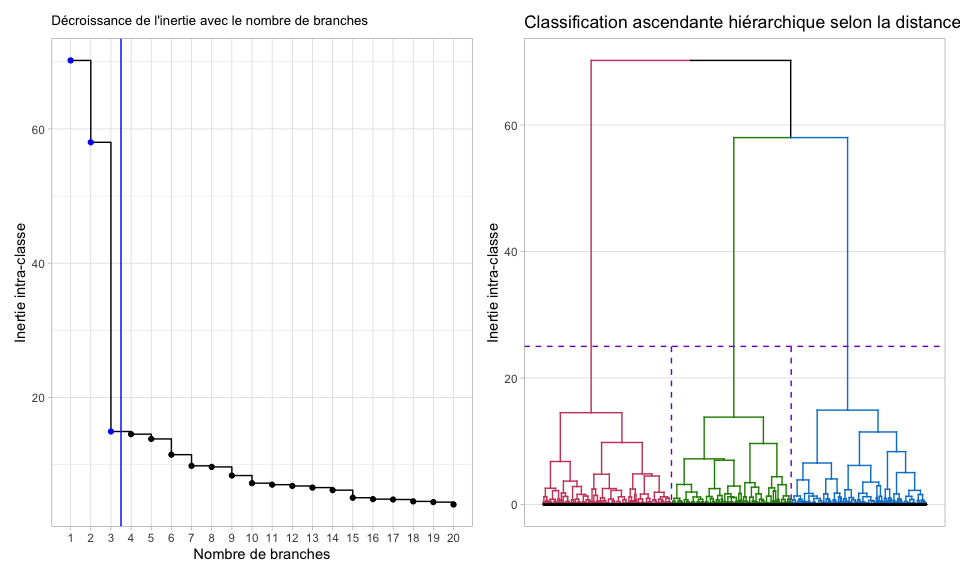


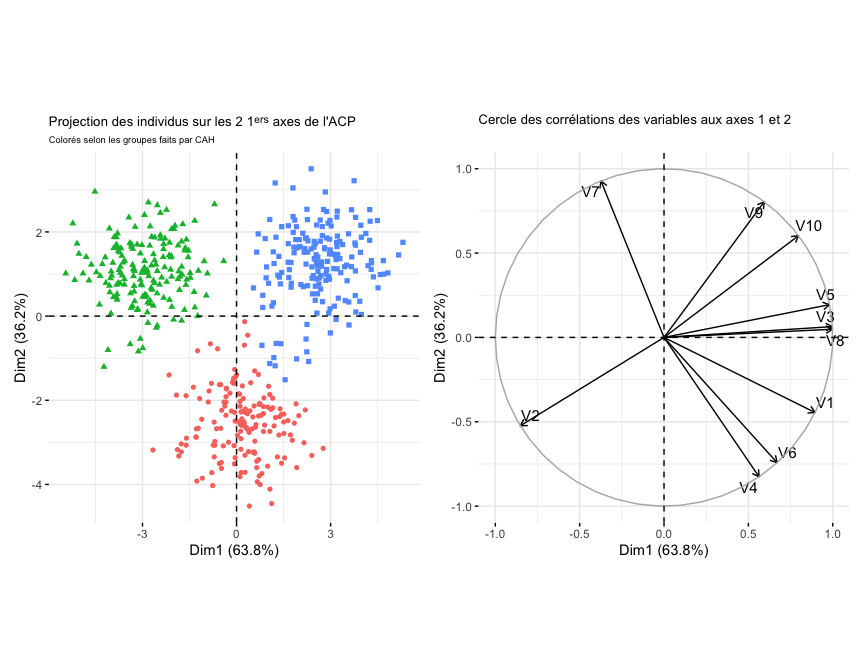
### Classification ascendante hiérarchique

Nous avons dans un premier temps réalisé une classification hiérarchique ascendante et présenté ses résultats sous la forme d’un dendrogramme.

Etant donné le résultat de l’ACP avec la projection des variables sur un plan qui explique la quasi-totalité de l’inertie, ainsi que le caractère numérique de toutes nos variables prédictrices, nous pouvons bien nous mettre dans la situation d’un espace métrique. De ce fait, nous avons choisi d’utiliser la distance euclidienne entre les différents individus statistiques, et la méthode pour aggréger les individus est la méthode de Ward, adaptée aux clusters sphériques.

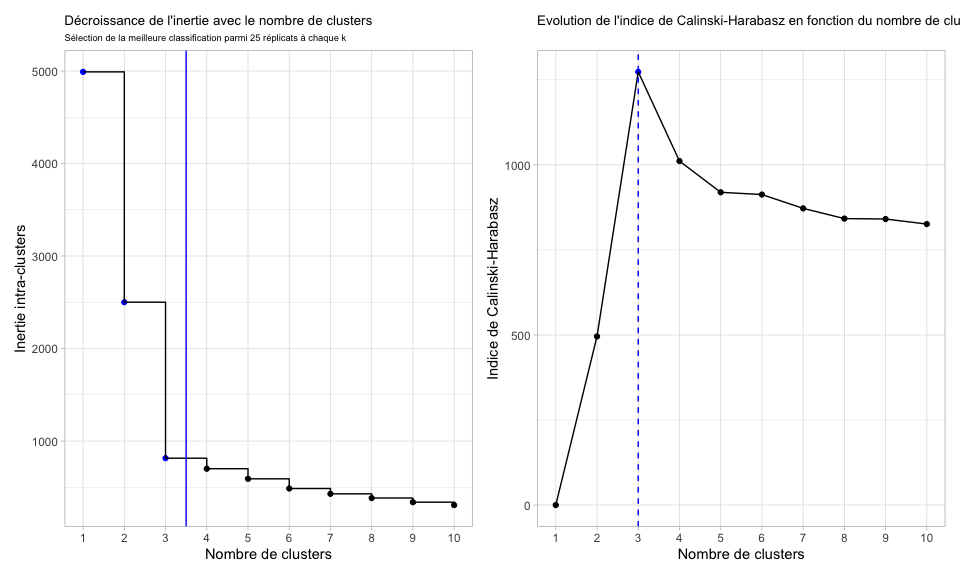
Pour déterminer le nombre de clusters, nous avons calculé l’inertie intra-classe en fonction du nombre de clusters. Puis, notre nombre de clusters était le nombre pour lequel il y a une rupture dans la diminution d’inertie intra-classe (stratégie, parfois dénommée ”critère du coude”). Cette classification permet de créer 3 classes d’effectifs semblables puisqu’ils sont respectivement de 156, 168, 176.

 En réprésentant les individus sur les 2 1ères composantes principales, nous pouvons voir les 3 clusters qui sont dans 3 directions du plan représenté.



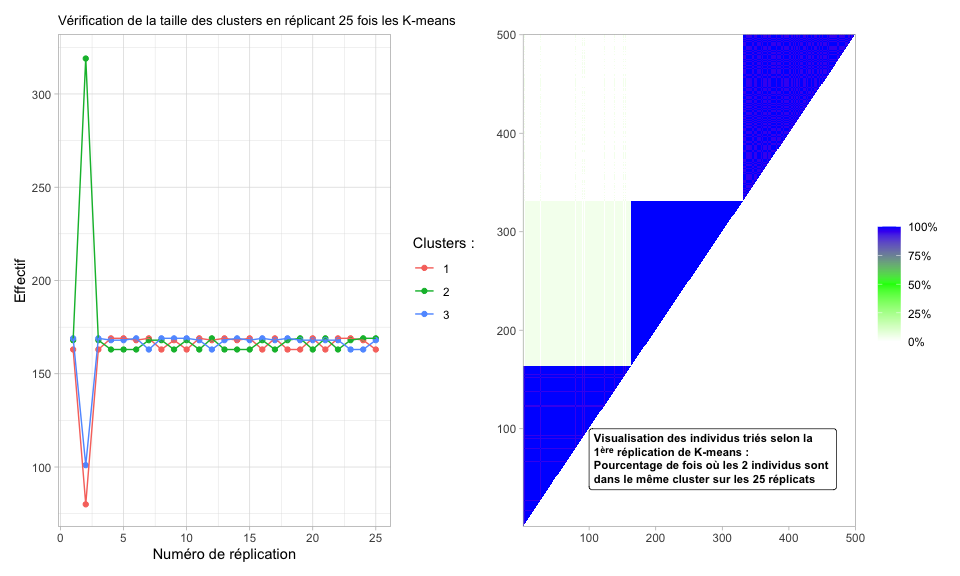
### Les K-means

Pour sélectionner le nombre de groupes optimal par les méthodes des K-means, nous avons eu recours à 2 méthodes. La 1ère est analogue à celle utilisée pour la classification ascendante hiérarchique : nous avons repéré le saut dans la diminution de l’inertie intra-classe avec le nombre de classes. Nous avons aussi représenté l’indice de Calinski-Harabasz qui doit être maximisé. Ces 2 méthodes nous indiquent un nombre optimal de clusters de 3. Nous réaliserons donc des clusters selon la méthode des K-means avec k = 3.



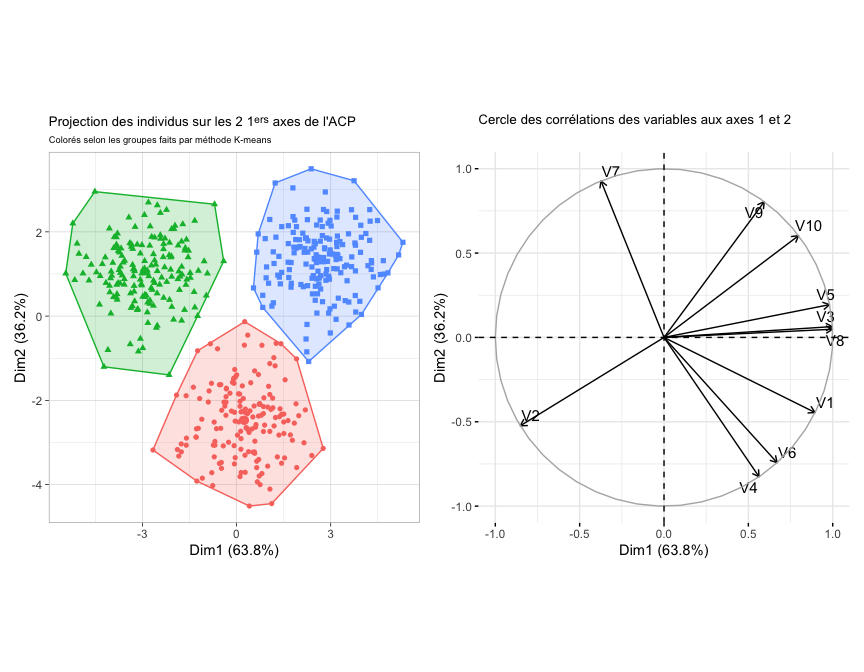
Pour valider la stabilité de la méthode à l’initialisation aléatoire, nous avons généré 25 réplicats d’une sélection de 3 groupes par les K-means. Les 3 clusters créés ont des effectifs similaires autour de 160-170 dans 24 des réplicats. Mais le 2ème réplicat montre un cluster qui a plus de patients que les 2 autres.

Nous avons donc regardé si les individus étaient répartis différemment selon le réplicat. Pour ce faire, pour les réplicats 2 à 25, nous avons attribué le numéro de cluster du réplicat 1 pour lequel l’effectif dans le réplicat (2 à 25) est majoritaire. Nous avons constaté que dans 70% des individus statistiques, le cluster est constant avec les 25 réplicats. En revanche, pour 30% des individus, on a des clusters différents avec les réplicats. En analysant les séquences de numéros de clusters, on a un déséquilibre pour le réplicat n°2. On peut remarquer que dans ce réplicat, en aggrégeant avec notre méthode, le cluster 1 du réplicat disparaît car il est minoritaire dans tous les clusters du réplicat 1. Ainsi, notre méthode n’est pas totalement satisfaisante dans ce cas de figure précis. Néanmoins, elle nous a permis de montrer la non constance de la classification pour le 2ème réplicat. De plus, nous avons représenté la matrice de similarité des patients qui réprésente les mêmes patients en abscisse et en ordonnée, et les intersections entre 2 individus sont de plus en plus bleues en fonction du nombre de fois où ces 2 individus sont dans le même cluster parmi chaque réplicat. On voit là encore qu’il y a un léger déséquilibre avec un cluster qui n’est pas homogène avec les clusters des autres réplicats et qui est vraisemblablement le cas du réplicat 2.



Ainsi, la méthode des K-means apparaît robuste pour classifier notre jeu de données puisque les clusters sont presque constants tout au long des 25 réplicats.

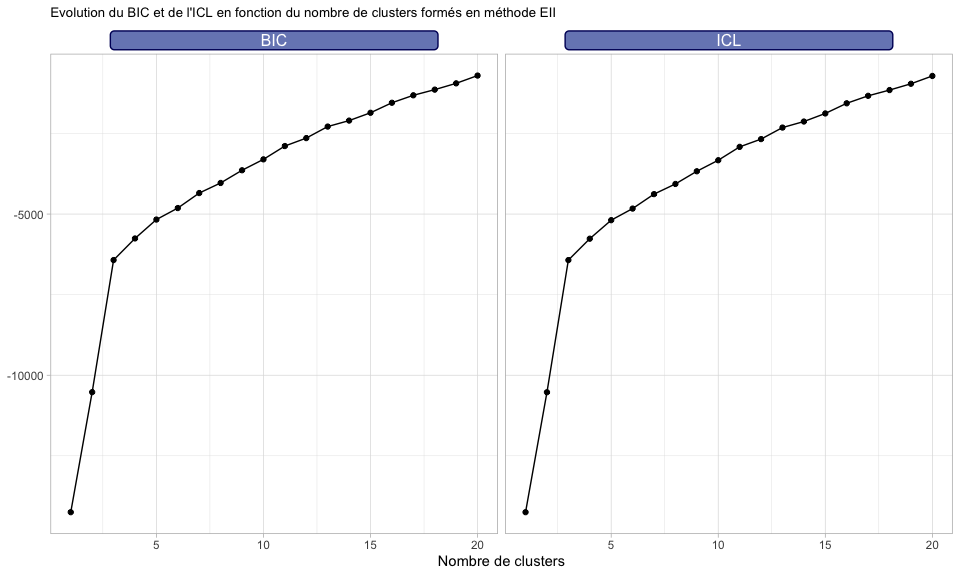
Pour sélectionner notre classification finale, nous avons réalisé là encore 25 réplicats de la classification, puis nous avons sélectionné la classification qui minimisait le plus l’inertie intra-clusters. Nous avons représenté cette classification sur la figure suivante.



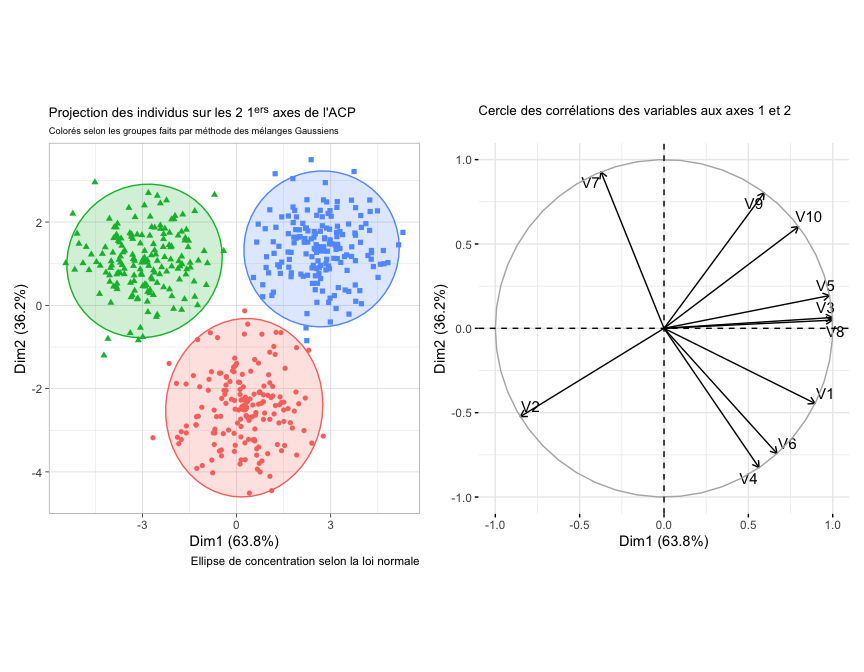
### Modèle de mélanges Gaussiens

A la vue des résultats de l’ACP, nous nous sommes concentré sur les modèles sphériques (EII et VII). Nous avons écarté la méthode EII de notre analyse principale car c’était redondant avec les K-means.

Afin de choisir le nombre de classes, nous nous sommes appuyés sur l’analyse des maximums de vraissemblance, représentés par le score BIC et ICL. Néanmoins, la représentation suivante montre que le BIC et l’ICL n’atteignent pas de valeur maximale pour un nombre de classes de 20 ou moins. Etant donné nos résultats précédents et la représentation de l’ACP, nous nous sommes orienté vers 3 classes. De plus ça a l’air d’être le moment ou il y a une inflexion de l’augmentation du BIC/ICL. Nous avons regardé ce qui se passait pour les modèles non sphériques de matrice de covariance. Pour les modèles non orientés sur les axes, la fonction Mclust n’arrivait pas à calculer de classes (dans l’article de Scrucca, il est dit que cela demande une correction de la matrice de covariance), et pour les autres, on n’atteignait pas de maximum ni pour le BIC ni pour l’ICL.



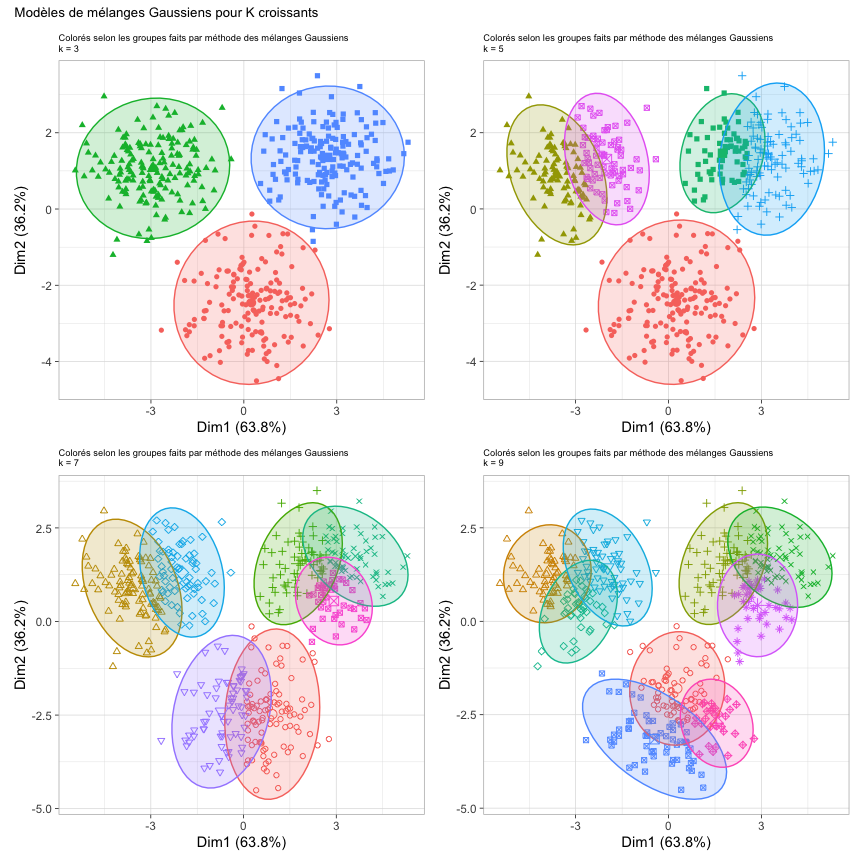
Nous avons représenté notre classification à 3 classes sur les 2 1ers axes principaux et pouvont constater que les groupes formés sont similaires à ceux obtenus avec les méthodes précédentes avec les ellipses de concentrations selon la loi normale.



## Discussion et comparaisons des modèles

L’objectif étant in fine la mise en évidence de la structure sous-jacente des données, une méthode pragmatique pour le choix du nombre de classes est de choisir une partition dont il sera possible d’interpréter les classes.

Le modèle Gaussien minimise la variabilité intra-groupe, mais la sélection du nombre de groupes n’a pas l’air de s’arrêter (sur les critères du BIC et de l’ICL). Nous avons représenté des nombres de clusters croissants sur la figure suivante. En augmentant le nombre de clusters, on remarque que cela ne fait que segmenter les 3 groupes que nous avions définis. Il y a donc peut-être des groupes d’individus à l’intérieur de nos 3 clusters qui se ressemblent et peuvent se regrouper, mais le découpage en 3 groupes apparaît pertinent. La méthode des mélanges Gaussiens n’est peut-être ici pas adaptée pour déterminer le nombre de groupes.



Les groupes formés par les 3 méthodes sont très similaires, et on voit que peu d’individus sont dans des clusters différents entre les méthodes. La différence se fait surtout sur les individus à la frontière entre les groupes 1 et 3 (aux coordonnées x = 1 et y = -0.5 environ sur les 2 1ers axes principaux). Nous pensons donc avoir une classification assez proche de la réalité. La méthode des mélanges Gaussiens donne les 3 mêmes groupes environ, mais a plus de mal pour déterminer le nombre de groupes.

