Devoir Classification et Apprentissage session 1

Benoit Gachet, Guillaume Mulier

29/11/2020

Table of Contents

[I. Exercice classification non supervisée 1](#_Toc57549154)

[1. Introduction 1](#_Toc57549155)

[2. Mise en oeuvre des modèles 2](#_Toc57549156)

[a. Description du jeu de données 2](#_Toc57549157)

[b. Classification ascendante hiérarchique 3](#_Toc57549158)

[c. Les K-means 4](#_Toc57549159)

[d. Modèle de mélanges Gaussiens 7](#_Toc57549160)

[3. Discussion et comparaisons des modèles 9](#_Toc57549161)

[II. Exercice classification supervisée 11](#_Toc57549162)

[1. Introduction 11](#_Toc57549163)

[2. Description de la base 12](#_Toc57549164)

[3. Mise en oeuvre des modèles 14](#_Toc57549165)

[a. Analyse factorielle discriminante 14](#_Toc57549166)

[b. K plus proches voisins 15](#_Toc57549167)

[c. Random forest 17](#_Toc57549168)

[4. Comparaison des 3 modèles 18](#_Toc57549169)

# 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.

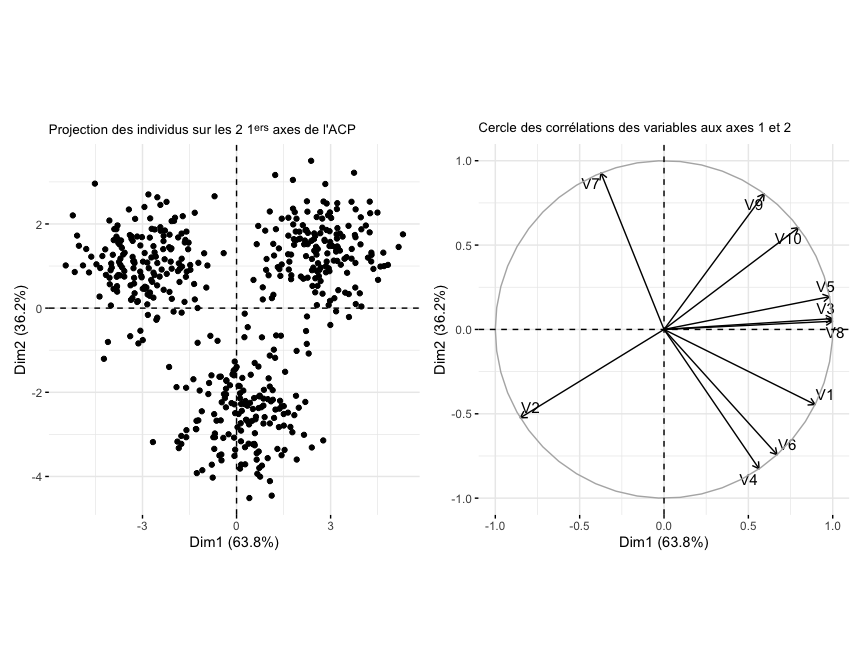
Il existe de nombreuses méthodes de classification non supervisées. Nous nous sommes intéressés à 3 méthodes : la classification hiérarchique ascendante, les k-means et les modèles de mélange Gaussiens. Nous les avons ensuite comparés et avons discuté les avantages et les inconvénients de chacune dans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

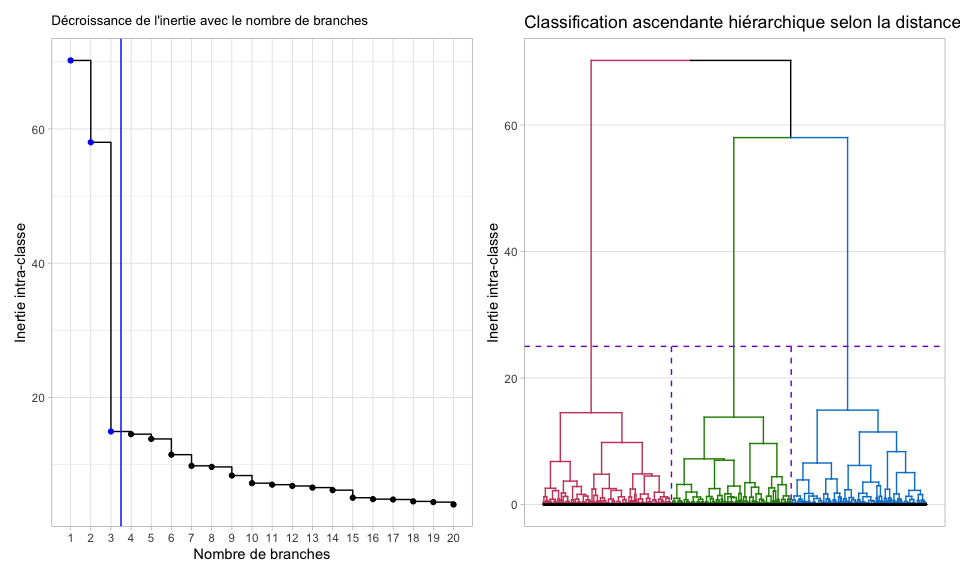


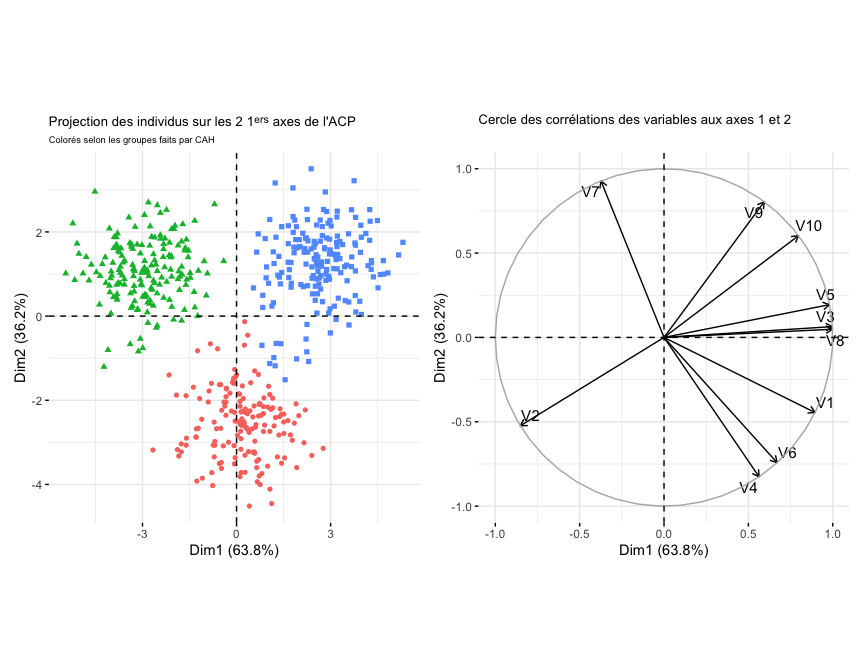
### Classification ascendante hiérarchique

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.

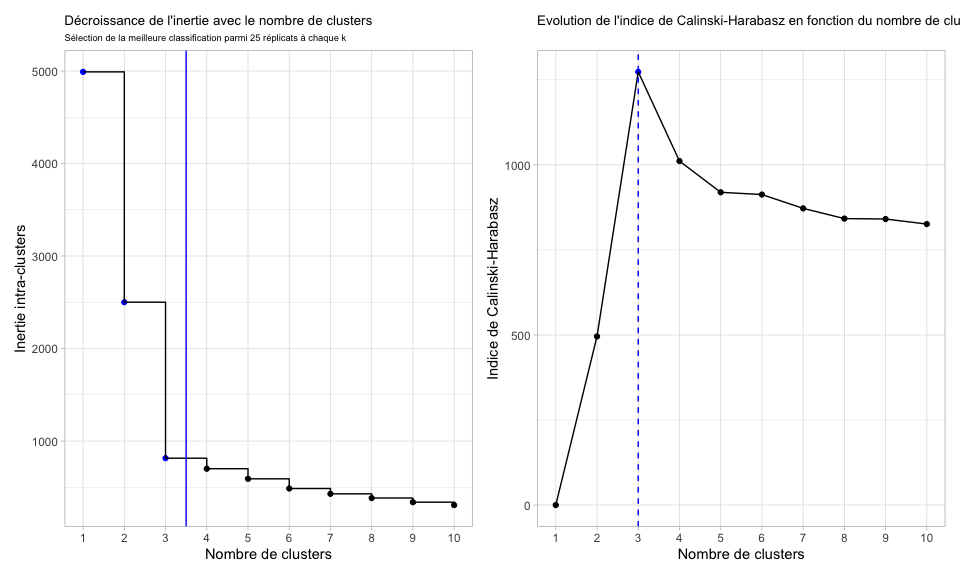
Pour déterminer le nombre de clusters, nous avons calculé l’inertie intra-classe en fonction du nombre de clusters. Puis, notre nombre de clusters était le nombre pour lequel il y a une rupture dans la diminution d’inertie intra-classe (stratégie, parfois dénommée ”critère du coude”). Cette classification 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.



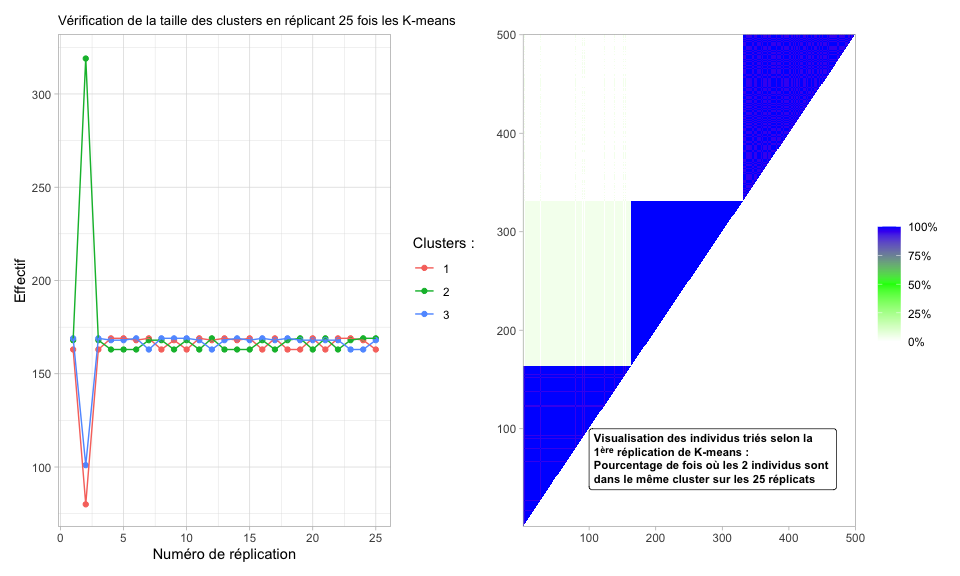
### Les K-means

Pour sélectionner le nombre de groupes optimal par la méthode 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.



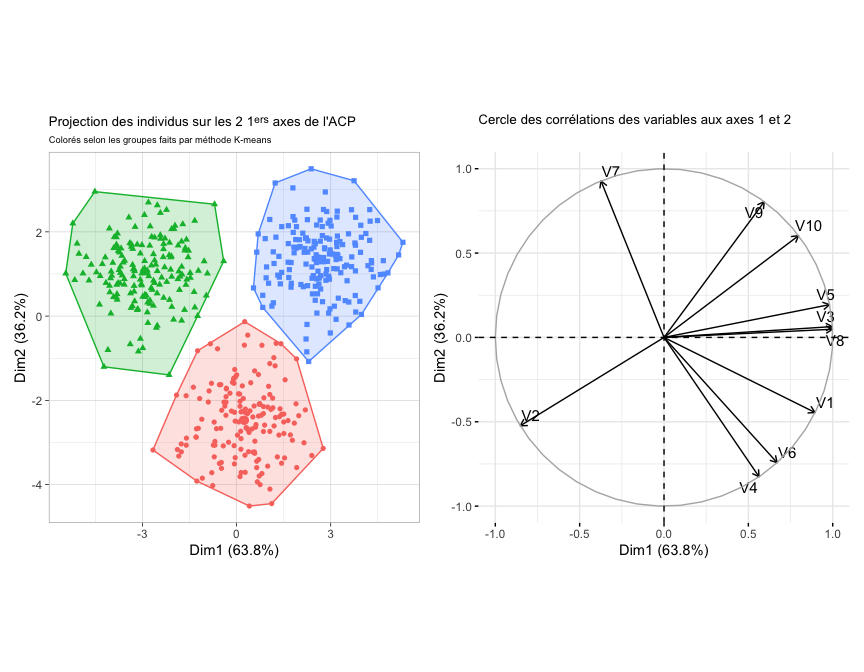
Pour valider la stabilité de la méthode à l’initialisation aléatoire, nous avons généré 25 réplicats d’une sélection de 3 groupes par les K-means. Les 3 clusters créés ont des effectifs similaires autour de 160-170 dans 24 des réplicats. Mais le 2ème réplicat montre un cluster qui a plus de patients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

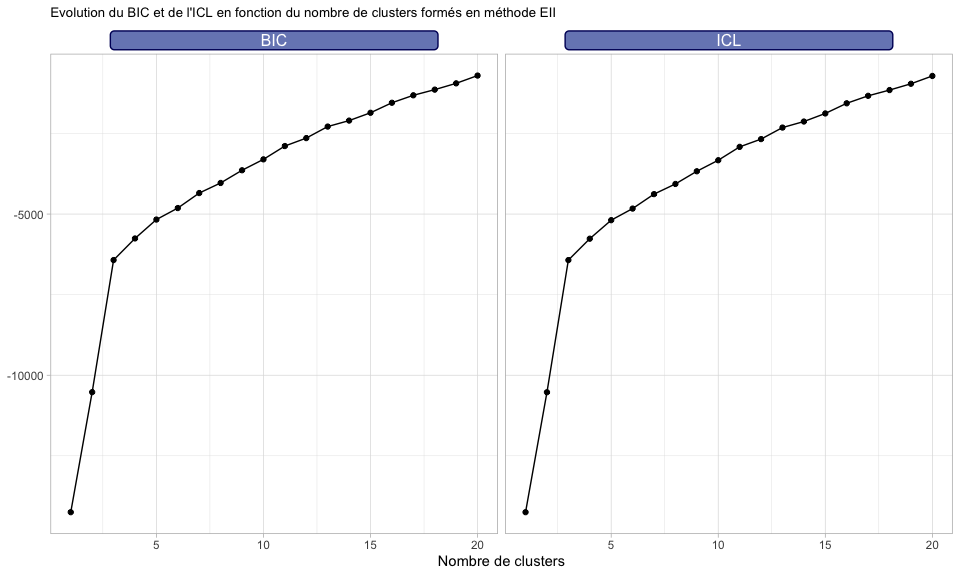
Pour sélectionner notre classification finale, nous avons réalisé là encore 25 réplicats de la classification, puis nous avons sélectionné la classification qui minimisait le plus l’inertie intra-clusters. Nous avons représenté cette classification sur la figure suivante.



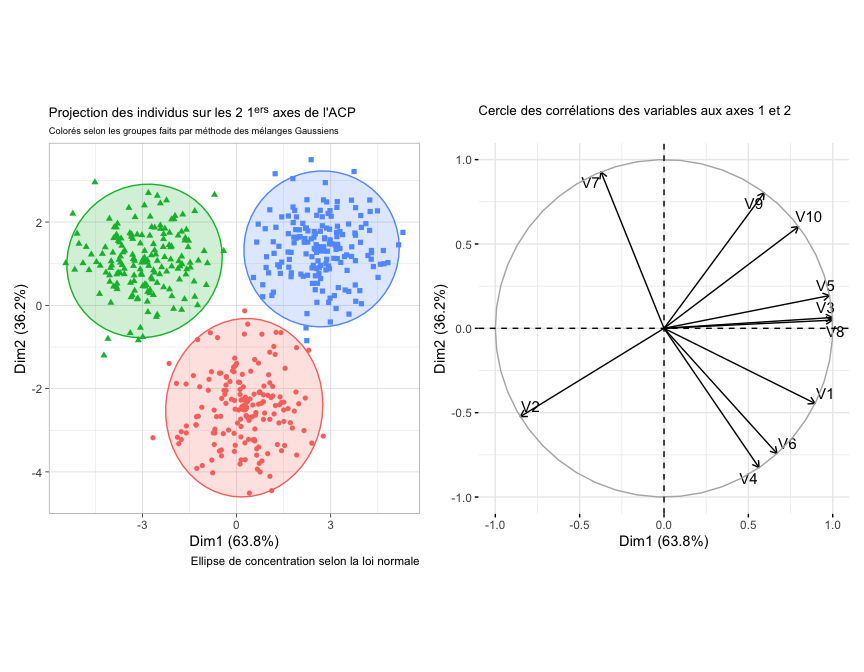
### Modèle de mélanges Gaussiens

A la vue des résultats de l’ACP, nous nous sommes concentrés 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 les scores 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 orientés vers 3 classes. De plus ça a l’air d’être le moment où 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 de façon analogue à notre analyse principale.



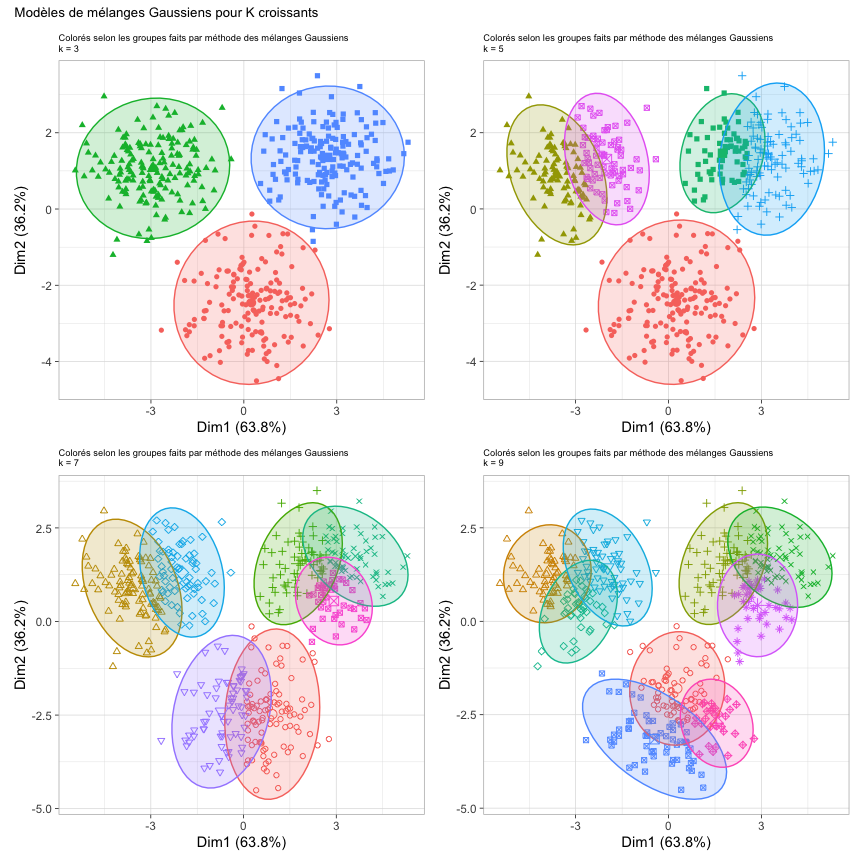
Nous avons représenté notre classification à 3 classes sur les 2 1ers axes principaux et pouvont constater que les groupes formés sont similaires à ceux obtenus avec les méthodes précédentes avec les ellipses de concentrations selon la loi normale.



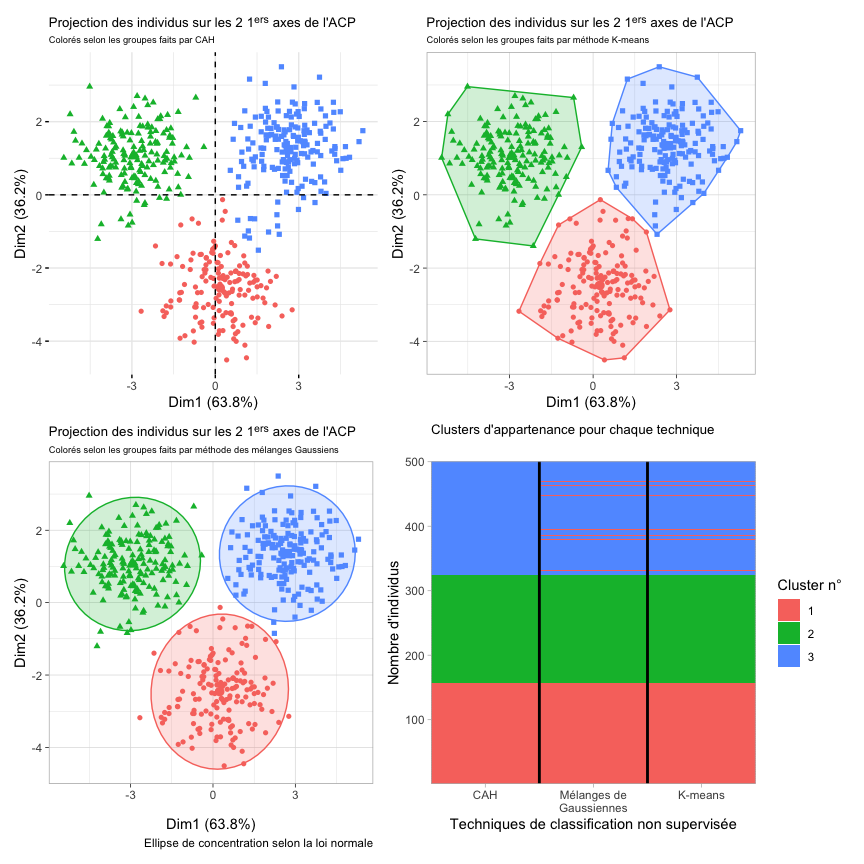
## Discussion et comparaisons des modèles

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.



# Exercice classification supervisée

## Introduction

L’objectif de notre devoir était de réaliser une classification supervisée en utilisant les données de la base simulée ElecGrid qui s’intéresse à la stabilité du réseau électrique. Nous devions réaliser plusieurs méthodes de classification, valider nos résultats et comparer les différentes méthodes entre elles.

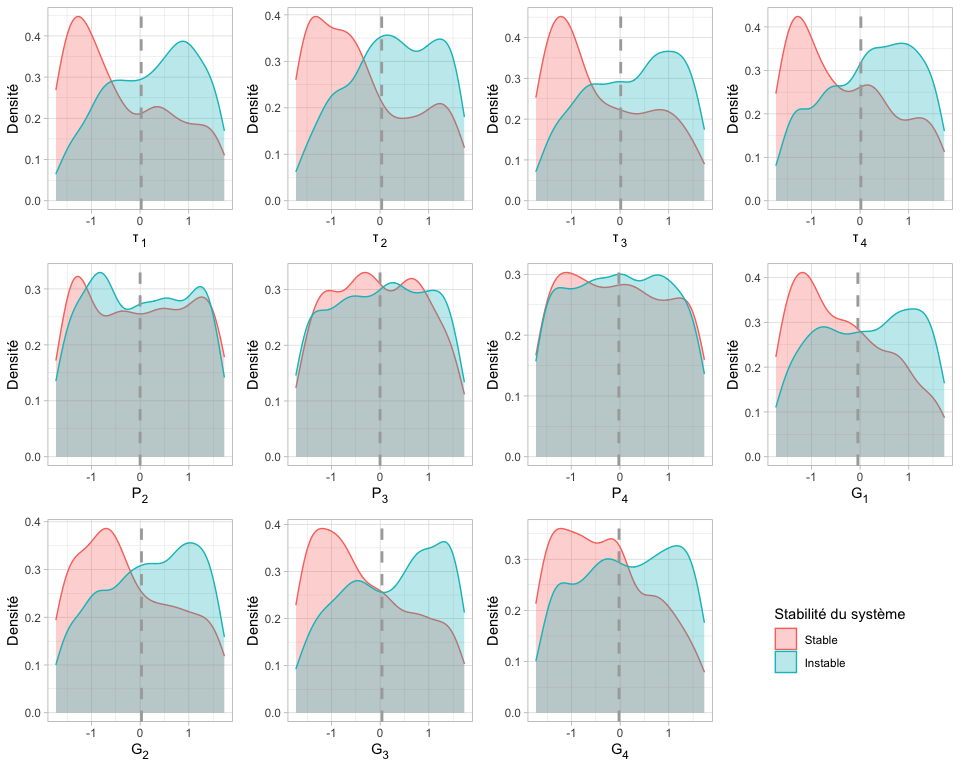
## Description de la base

Chaque observation représentait l’évaluation du système électrique avec 11 variables explicatives et 1 variable à expliquer :

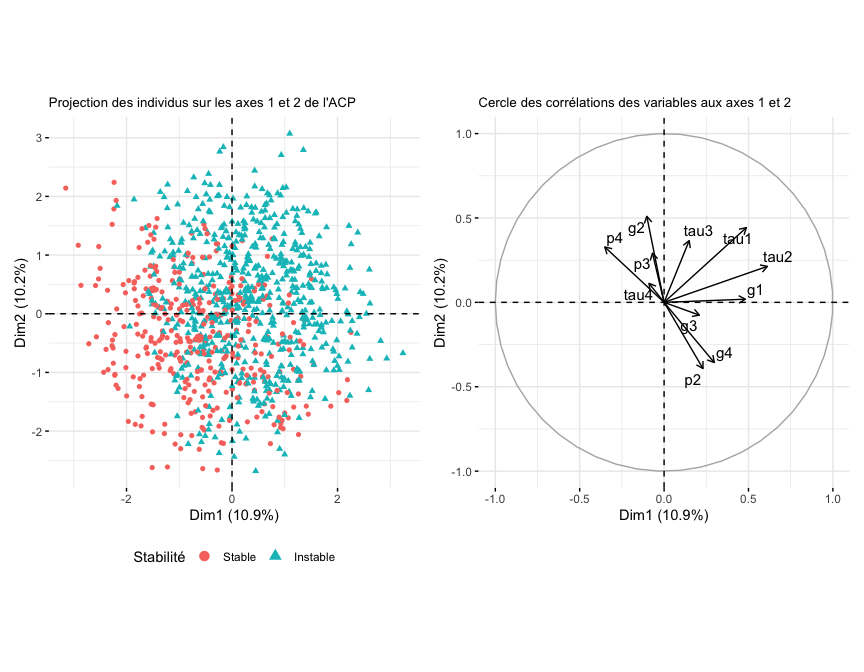
* La variable à expliquer est stabf et nous informait sur la stabilité du réseau électrique : “stable” ou “instable” ;
* Les variables explicatives sont :
  + , , , (électricité produite respectivement à partir du nœud 1, 2, 3 et 4) ;
  + , et (énergie consommée/produite respectivement sur le 2, 3 et 4ème nœuds). quant à lui, réprésentait la valeur absolue de la somme de , et . Cette variable étant une combinaison linéaire de 3 autres variables, elle n’apportait aucun information et n’a pas été utilisée ;
  + , , et (prix de l’électricité sur les noeuds 1 à 4 à chaque observation).

Nous avons travaillé sur un échantillon de 1000 observations sélectionnées au hasard parmi les 10000 observations de la base initiale. 361 observations étaient stables et 639 instables. Il n’y avait aucune donnée manquante. L’ordre de grandeur entre les variables était différent, nous avons donc centré et réduit les valeurs.

Nous avons analysé les courbes de densité pour chaque variable en fonction du statut “stable” et “instable”. Les courbes étaient superposables pour les variables p contrairement aux autres variabless. Nous pouvions donc faire l’hypothèse que les variables et étaient plus discriminantes pour classer les observations.



L’analyse en composante principale à mis en avant 11 composantes principales qui expliquaient toutes environ la même part de variance (7 à 11%). La représentation des 2 groupes sur les 6 1ers axes principaux, mettait en évidence un nuage de point difficile à individualiser, mais se démarquant quand-même. Globalement, les groupes avaient une forme sphérique et un volume assez équivalent (sur la projection en tout cas).



## Mise en oeuvre des modèles

Nous avons mis en oeuvre 3 modèles de classification supervisée : une analyse factorielle discriminante, une analyse des k plus proches voisins et une approche par random forest.

Nous allons essayer de mettre en oeuvre ces classifications et les comparer en estimant les risques de classification pour chacune des 3 techniques étudiées. Pour l’approche par les random forests, le risque sera estimé par les erreurs out-of-the-bag. Pour l’analyse factorielle discriminante et les k plus proches voisins, nous estimerons les risques par cross-validation. Pour ce faire, nous avons aléatoirement découpé notre jeu de données en 10 groupes qui seront les échantillons tests tour à tour afin d’évaluer le risque. Le risque global calculé sera la moyenne des 10 risques obtenus par cette méthode.

De plus, pour l’analyse factorielle discriminante et les k plus proches voisins, nous avons découpé l’échantillon aléatoirement en un échantillon d’apprentissage contenant 70% des observations et un jeu test contenant les 30% restants.

### Analyse factorielle discriminante

Selon la description de l’ACP, nous avons pris les modèles sphériques. Nous n’étions pas sûr du volume, donc nous avons comparé EII et VII sur le BIC sur l’échantillon d’apprentissage. Pour les 2 cas de figures testés (EDDA = matrices de covariances identiques dans toutes les modalités de la variable à prédire ; et MclustDA = matrices de covariances différentes dans chaque modalité de la variable à prédire), le modèle EII avait été choisi. Pour EDDA, le BIC était plus élevé, et pour MclustDA, les BIC étaient équivalents, mais on minimise le nombre de paramètres avec EII donc nous avons choisi l’approche EII.

BIC selon l’estimation de la matrice de covariance

| Méthode d'estimation de la matrice de covariance | EII | VII |
| --- | --- | --- |
| EDDA | **-21978.73** | -21989.43 |
| MclustDA | **-21601.48** | -21601.48 |

Pour retenir ensuite la meilleure méthode concernant l’estimation des matrices de covariance, nous avons estimé le risque avec 2 méthodes : en estimant sur l’échantillon test, et par cross-validation. Sur l’échantillon test et la cross-validation, la méthode MclustDA donne une meilleure minimisation du risque de classification. C’est donc l’analyse factorielle discriminante avec modèle EII et des matrices de covariances différentes selon les modalités stable et instable qui sera retenu. Le risque de classification pour ce modèle est de 17% en échantillon test et 19.5% en cross-validation. Il est à noter que nous avons aussi comparé avec la fonction cvMclustDA du package mclust qui donne un risque un peu différent du notre pour la cross-validation de 18%. Mais nous avons préféré utiliser notre méthode pour pouvoir comparer avec les k plus proches voisins.

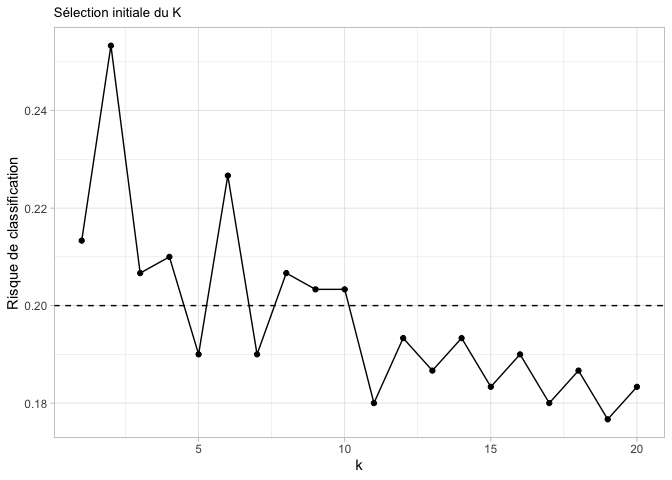
Risque calculé sur le modèle EII selon les 2 méthodes d’estimation de la matrice de covariance en %

| Méthode | EDDA | MclustDA |
| --- | --- | --- |
| Sur échantillon test | 17.3 | **17** |
| Cross-validation | 22.9 | **19.5** |

### K plus proches voisins

Nous avons réalisé une deuxième analyse en utilisant la méthode des k plus proches voisins avec la même population d’entrainement et de test que pour l’analyse factorielle discriminante afin de pouvoir au mieux comparer les 2 techniques.

Une 1ère sélection du nombre de K s’est faite en screenant les k de 1 à 20 et en évaluant le risque de classification sur l’échantillon test. Nous avons retenu les k pour les k pour lesquels on avait un risque 20%. Ensuite, nous avons réalisé une cross-validation sur le même découpage en 10 groupes que celle effectuée pour l’analyse factorielle discriminante.



Sans surprise, les k impairs ont des risques moins élevés. En effet, on s’imagine que comme nous avons 2 groupes à différencier, rechercher la majorité par rapport à un nombre impair de voisins permet de se soustraire aux ex aequo. Avec un K à 5, 7 et de 11 à 20, nous avons un risque 20%. Nous avons donc regardé les risques en cross-validation pour ces k.

Détermination du nombre de voisins optimal par cross-validation

| k | Risque de classification |
| --- | --- |
| **5** | **18.1** |
| 7 | 18.7 |
| 11 | 18.8 |
| 12 | 20.2 |
| 13 | 18.3 |
| 14 | 20.5 |
| 15 | 18.4 |
| 16 | 20.1 |
| 17 | 18.3 |
| 18 | 20.2 |
| 19 | 18.3 |
| 20 | 19.7 |

Finalement, le risque de classification minimal est trouvé pour k = 5 et nous retenons donc ce modèle pour la classification par les k plus proches voisins. Les risque de classification est alors de 18.3%.

### Random forest

Pour déterminer le nombre d’arbres optimal nous avons recherché parmi 5 nombres d’arbres (1000, 2000, 3000, 4000 et 5000) et parmi 5 nombres de prédicteurs (1, 2, 3, 4 et 5). Pour chaque couple, nous allons regarder le risque out-of-the-bag afin de déterminer le couple prédicteur/arbre.

Optimisation du nombre d’arbres et de prédicteur pour le modèle random forest

| Nombre d'arbres | Nombre de prédicteurs | Risque de classification |
| --- | --- | --- |
| 1000 | 1 | 14.3 |
| 1000 | 2 | 12.4 |
| 1000 | 3 | 12.4 |
| 1000 | 4 | 12.8 |
| 1000 | 5 | 13.5 |
| 2000 | 1 | 14.0 |
| 2000 | 2 | 12.2 |
| **2000** | **3** | **11.8** |
| 2000 | 4 | 12.5 |
| 2000 | 5 | 13.2 |
| 3000 | 1 | 14.1 |
| 3000 | 2 | 12.1 |
| 3000 | 3 | 12.3 |
| 3000 | 4 | 13.0 |
| 3000 | 5 | 13.6 |
| 4000 | 1 | 14.1 |
| 4000 | 2 | 12.6 |
| 4000 | 3 | 12.4 |
| 4000 | 4 | 12.5 |
| 4000 | 5 | 13.3 |
| 5000 | 1 | 14.0 |
| 5000 | 2 | 12.3 |
| 5000 | 3 | 12.3 |
| 5000 | 4 | 12.7 |
| 5000 | 5 | 13.2 |

On a donc un risque de classification minimal pour 2000 arbres et 3 prédicteurs. Ce sera donc notre modèle.

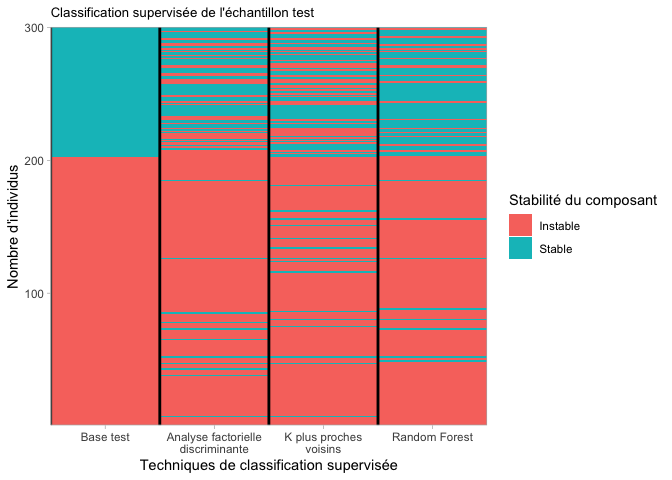
Pour ce modèle, le risque de classification est de 11.8%. En inspectant les importances des variables dans la construction de la forêt, les 3 variables ayant le moins d’importance sont , et , ainsi que nous l’avions suspecté avec la description des densités des variables.

Nous avons essayé de retirer ces variables pour voir si la qualité de la prédiction s’améliorait, mais le risque de classification restait dans le même ordre de grandeur (11.9%).

## Comparaison des 3 modèles

L’analyse factorielle discriminante et les k plus proches voisins donnent des résultats similaires en terme de risque et sont une bonne approche dans un 1er temps. La méthode des random forest permet d’affiner le résultat et d’obtenir un meilleur risque de classification.

Nous avons représenté les résultats des différentes classifications pour l’échantillon test. Il y a quelques erreurs de classification du groupe instable mais c’est assez similaire pour les 3 méthodes. Par contre, les random forests font moins d’erreurs que les 2 autres méthodes sur la classification des stables. Ainsi, les random forests ont l’air de moins surestimer l’instabilité. Nous pensons donc que des 3 méthodes utilisées, les random forests sont la méthode la plus adaptée à utiliser pour classifier notre jeu de données.



Il serait sûrement possible de gagner encore en précision en utilisant des modèles plus compliqués. Tout dépend du risque d’erreur qu’on est prêt à faire dans notre classification. Ici, c’est pour un outil de contrôle des flux et des réseaux de transport d’électricité. Nous n’avons aucune idée de l’erreur qu’il est au maximum toléré dans ces situations, mais nous nous doutons que pour des ingénieurs en électricité soucieux d’être les plus précis possible, notre résultat est peut-être encore à améliorer.