UBS/Faculté des Sciences et Sciences pour L'Ingénierie

Cancer du Colon

Analyse exploratoire et statistique

Rédigés par:
par :

DOUMBALY CAMARA SY

Tuteurs: Olivier Sire & Salim Larjane

2018-2019

Université Bretagne sud de Vannes

Faculté des sciences et sciences pour l'ingénierie

(Data Science Modélisation Statistique)

Président :	**Pr. Jean PEETERS**
Doyen / Faculté SSI :	**Mr. Frederic BEDEL**
Responsable :	**Mr. ion GRAMA**
Maitre de conférences :	**Mr. Salim LARJANE**
Professeur :	**Pr. Olivier SIRE**
Secrétaire Pédagogique :	**Sandrine STEINMANN**

Table des matières

Liste des graphiques

Remerciements

Nos remercîments vont à toutes personnes de l'université Bretagne Sud de Vannes qui ont contribué chacune à leur manière à la réalisation de ce travail. Nous les adressons tout particulièrement :

A Mr Gramma qui a fait preuve de beaucoup d'engagement pour la réalisation du programme Data Science.

A Mr Olivier Sire et Mr Salim Larjane, pour leurs explications lors des cours et leurs esprits d'écoute.

A tous nos camarades de programme avec qui nous avons eu des échanges toujours constructifs

Avant-propos

L'objectif de ce projet est de fournir un environnement interactif d'analyse exploratoire de données, statistiques et prévisions, doté d'outils graphiques performants et permettant une adaptation aisée aux besoins des utilisateurs (médecin).

Le principe est de faire un diagnostic des patients jugés atteints du cancer du côlon. Afin de pouvoir les classes par groupes après injection d'un médicament anti-cancéreux a la premier visite.

Toute l'analyse s'est faite à l'aide du logiciel statistique R.

À cet égard, nous remercions ici toute la communauté des chercheurs qui travaillent continuellement à améliorer cet outil de travail de qualité, accessible gratuitement et d'une grande puissance.

R est un langage de programmation pour l'analyse et la modélisation des données. R peut être utilise comme un langage orienté objet tout comme un environnement statistique dans lequel des listes d'instructions peuvent être exécutées en séquence sans l'intervention de l'utilisateur.

Étant conscient que la science est basée sur la critique, nous sommes toujours intéressés à recevoir toutes les remarques ou suggestions de correction ou d'amélioration de ce qui a été présenté dans ce mémoire.

Les codes des résultats seront en annexes. Ils seront générés automatiquement à l'aide du package Rmarkdown.

Introduction

La statistique est l'étude de la collecte de données, leur analyse, leur traitement, l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques. Parmi ces méthodes on peut la biostatistique qui est un ensemble de méthodes qui a pour objet : la collecte des données, le traitement des données et l'interprétation des données tout cela au service de sciences biomédicales.

La statistique constitue, en médecine, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence en médecine :

Quelle est la valeur normale d'une grandeur biologique, taille, poids, glycémie ?

Quel est le risque de complication d'un état pathologique, et quel est le risque d'un traitement ?

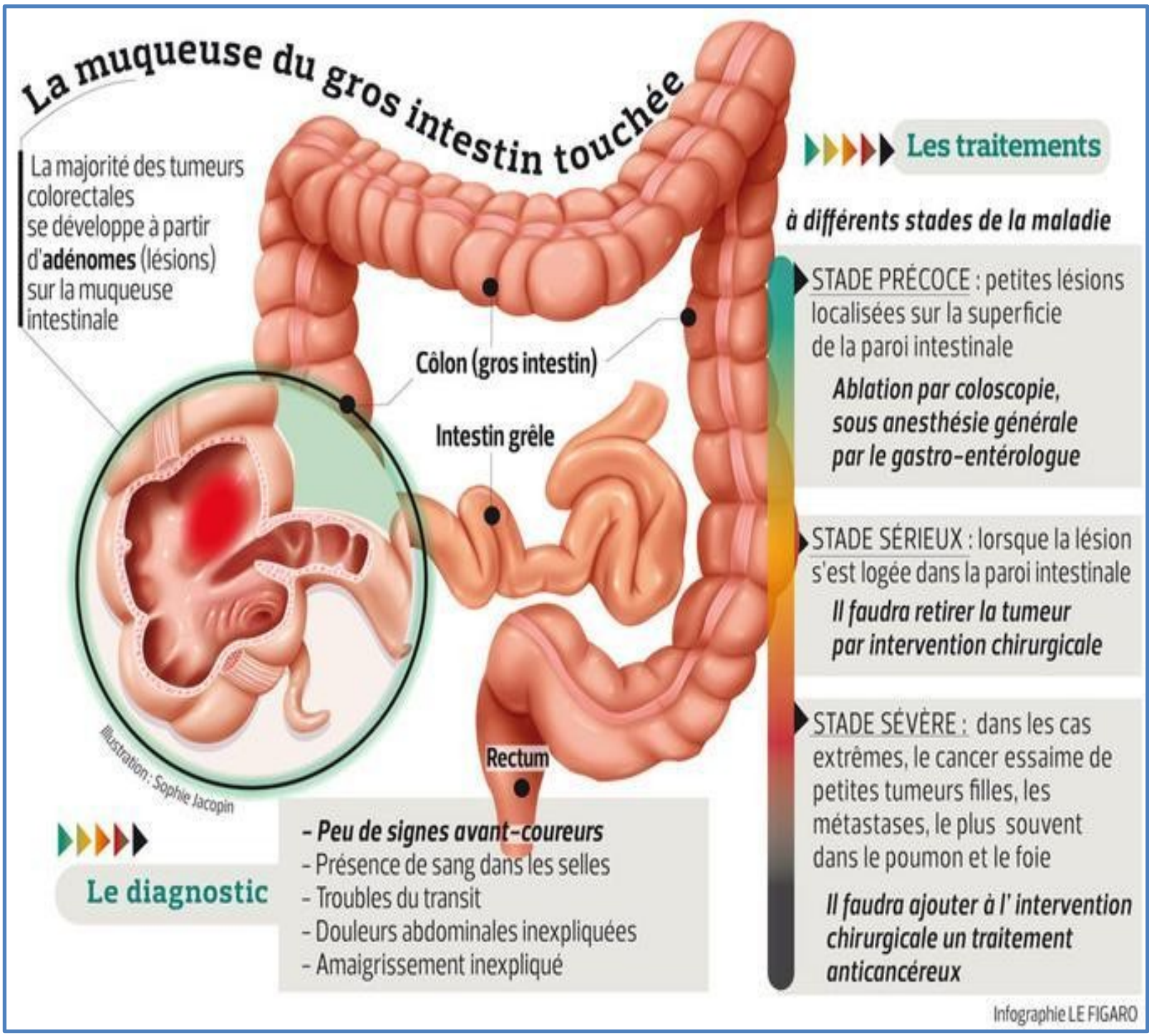
Le traitement A est-il plus efficace que le traitement B ?

Nous allons dans un premier temps expliquer l'objectif de l'étude puis importer le jeu de données dans l'environnement de R, ensuite l'analyser, constituer des groupes de patients et enfin faire des prévisions d'appartenance de groupes.

Dans ce présent rapport nous disposons d'un jeu de données contenant 1041 individus et 1600 variables (Longueur d'onde).

Nous tenons aussi à préciser que le jeu de données est obtenu grâce à un prélèvement effectue sur des dépôts de plasma après 24h de séchage, ne dispose pas de données manquantes et que l'ensemble du traitement de données se fera avec le logiciel R.

Cancer du Colon



I-Présentation :

I-1-Définition :

2Le cancer du côlon (colorectal) (CCR) est une tumeur maligne du côlon ou du rectum. Il fait suite dans 60 % à 80 % des cas à une tumeur bénigne, appelée communément polype, qui peut correspondre histologiquement à un polype adénomateux ou à un polype festonné. La durée de transformation d'un polype en cancer est estimée de 5 à 10 ans. Le CCR évolue fréquemment sans symptôme avant-coureur. Il peut être diagnostiqué devant l'apparition de signes fonctionnels digestifs (rectorragies, melæna, syndrome rectal, douleurs abdominales, modifications du transit abdominal, etc.), de signes généraux (amaigrissement inexpliqué, asthénie, fièvre, etc.), de signes physiques (masse abdominale, etc.) ou de signes biologiques (anémie, syndrome inflammatoire, etc.).

Le CCR est le 3e cancer le plus fréquent en France et le 2e en termes de mortalité1. La France est l'un des pays d'Europe où l'incidence du CCR est la plus élevée pour les deux sexes. Les cancers colorectaux sont sporadiques dans 80 % des cas, surviennent dans un contexte familial dans 15 % des cas et sont liés à une prédisposition génétique dans 5 % des cas. Plus le CCR est diagnostiqué tôt, meilleur est le taux de survie (90 % de survie à 5 ans, pour les stades localisés). L'objectif de cette fiche est de faire le point sur les modalités de dépistage du CCR et de prévention chez le sujet à risque élevé et très élevé.

I-2 Motivation (Contexte) :

On administre parfois un traitement ciblé pour traiter le cancer colorectal. On a alors recours à des médicaments pour cibler des molécules spécifiques, comme les protéines, présentes à la surface des cellules cancéreuses. Ces molécules contribuent à l'envoi de signaux qui indiquent aux cellules de croître ou de se diviser. En ciblant ces molécules, les médicaments interrompent la croissance et la propagation des cellules cancéreuses tout en limitant les dommages aux cellules normales.

C'est dans ce sens que le médecin effectue des traitements en faisant des injections de **bevacizumab** en t_0 (avant traitement), t_{15} (15 jours après) et en t_{30} (30 jours après).

Cependant, nous allons effectuer des classification afin de prévoir l'appartenance d'un patient a un groupe et prévoir pour un nouveau cas atteint du cancer du côlon après seulement une première visite pour optimiser son traitement.

➔ [2] Les cancers en France, Les Données, INCa, édition 2015, avril 2017

I-3-Objectifs :

Notre objectif comme cité précédemment est de pourvoir classer les individus atteints par le cancer du côlon dans des groupes homogènes en utilisant l'algorithme de kmeans. La segmentation des patients atteints, permettra aux médecins de réduire le temps de traitement et de diminuer le cout des médicaments.

II- Analyses et explorations :

L'analyse exploratoire des données (AED) permet d'identifier les relations systématiques entre des variables, lorsqu'il n'existe aucune hypothèse a priori (ou des hypothèses incomplètes) quant à la nature de ces relations. Lors d'une analyse exploratoire typique, le chercheur prend en compte et compare, à l'aide de diverses techniques, de nombreuses variables pour mettre en évidence des structures systématiques.

II-1. Importation du jeu de donnée :

Cette partie sera beaucoup plus explicitée au niveau du livre des codes en annexes avec une explication détaillée de la démarche suivie.

Dans ce jeu de données les patients sont les individus et les variables sont des longueurs d'ondes.

Il faut tout de même affirmer qu'il n'y pas de variable qualitative.

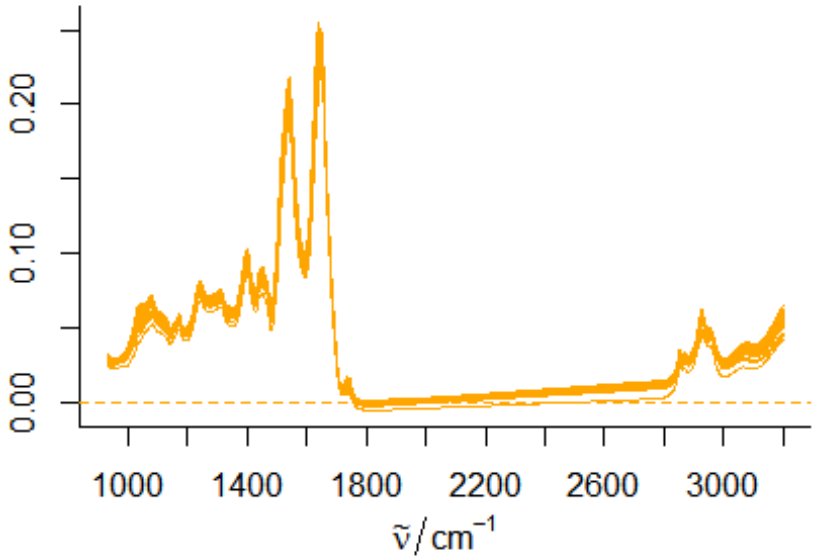
II.1 Présentations des données :

La base utilisée dans notre analyse est constituée de 1041 observations et 1600 variable, les observations représentent des individus qui ont le cancer du côlon et les variables représentant des longueurs d'ondes.

Notons que pour chaque patient le médecin a effectué trois traitements, ce qui fait que les 1041 observations ont été réduis à 113 en sélectionnant le spectre qui était proche de la moyenne pour les trois spectres d'une seul personne.

II-1.1 Représentation graphique de la base :

Figure 1 Representation graphique de la base colon



II-2. La méthode utilisées :

II.2.1. classification

La classification est une technique d'apprentissage automatique qui implique le regroupement de points de données. Avec un ensemble de points de données, nous pouvons utiliser un algorithme de classification pour classifier chaque point de données dans un groupe spécifique. En théorie, les points de données appartenant au même groupe devraient avoir des propriétés et / ou des caractéristiques similaires, tandis que les points de données de différents groupes devraient avoir des propriétés et / ou des caractéristiques très différentes.

Il existe beaucoup de méthodes de classification comme :

A. L'algorithme du kmeans :

La méthode des k-means est un outil de classification classique qui permet de répartir un ensemble de données en k classes homogènes. La plupart des images (photos, dessins vectoriels 2D, synthèses 3D, …) vérifient localement des propriétés d'homogénéité, notamment en termes d'intensité lumineuse. L'algorithme des k-means permet donc d'apporter une solution à la segmentation d'images.

•Avantages de l'algorithme :

L'algorithme de k-means est très populaire du fait qu'il est très facile à comprendre et à mettre en œuvre.
Sa simplicité conceptuelle et sa rapidité,
Applicable à des données de grandes tailles, et aussi à tout type de données (mêmes textuelles), en choisissant une bonne notion de distance.

•Inconvénients de l'algorithme :

Le nombre de classe doit être fixé au départ,
Le résultat dépend de tirage initial des centres des classes ,
Les clusters sont construits par rapports à des objets inexistants (les milieux).

B- Mise en œuvre :

Choisir K éléments initiaux "centres" des K groupes Placer les objets dans le groupe de centre le plus proche. Recalculer le centre de gravité de chaque groupe Itérer l'algorithme jusqu'à ce que les objets ne changent plus de groupe

C- Problèmes de l'algorithme :

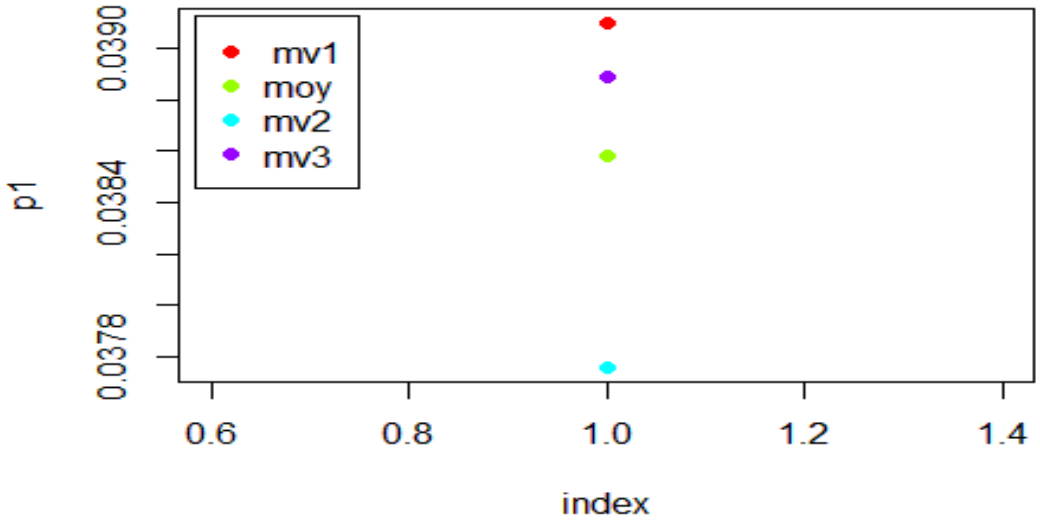
-> Obligation de fixer K.

-> le résultat dépend fortement du choix des centres initiaux.

-> ne fournit pas nécessairement le résultat optimum fournit un optimum local qui dépend des centres initiaux.

II-3. Calcul des moyennes spectrale :

Figure 2 *Representation graphique du choix des spectres*



Les 113 patients ont été sélectionnés en regardant la longueur d'onde pour chaque patients, que l'on compare à la moyenne des3 triplicate pour chaque spectres et nous sélectionnons la longueur d'onde qui sera la plus proche de la moyenne des triplicates.

L'étape qui suit sera la programmation de l'algoritme kmeans , mais avant nous transformons les trois bases obtenues en object hyperSpectrale en utilisant le package "hyperSpec".

[3] Triplicate c'est par exemple pour le traitement 1 nous avons 1.1, 1.2, 1.3 qui représente des spectres pour un patient.

III-Détermination des groupe pour le V1 :

Rappelons que pour le traitement des patients nous avons trois temps de visites t_0, t_{15} et t_{30}.

V1 représente le jeu de donne concernant que les traitements en t0.

III-1. Importation de la base :

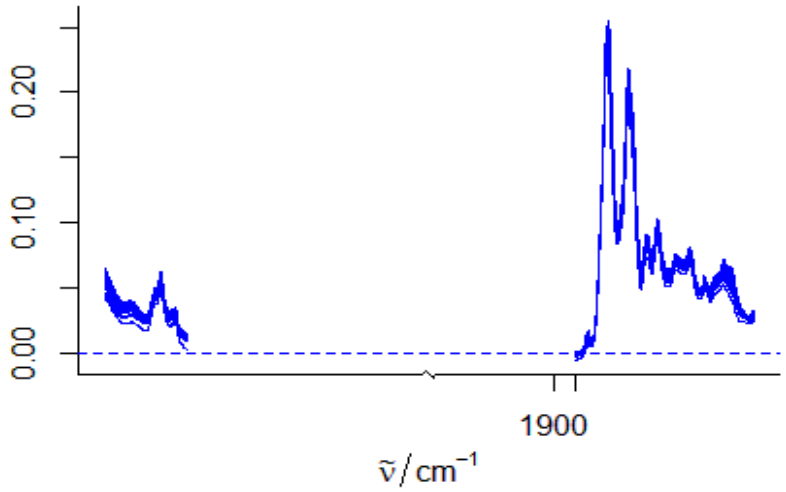
Et après nous avons effectué la troncature pour éliminer une partie qui ne comporte pas de spectres valides.

III-1.1 Représentation des spectres bruts après troncature :

La troncature est faite entre 3200-2800 et 1800-600, les valeurs ou longueur d'onde λ qui se trouvent entre 2800 et 1800

La troncature est faite entre 3200-2800 et 1800-600, les valeurs ou longueur d'onde λ qui se trouvent entre 2800 et 1800 sont presque nulles d'où la suppression.

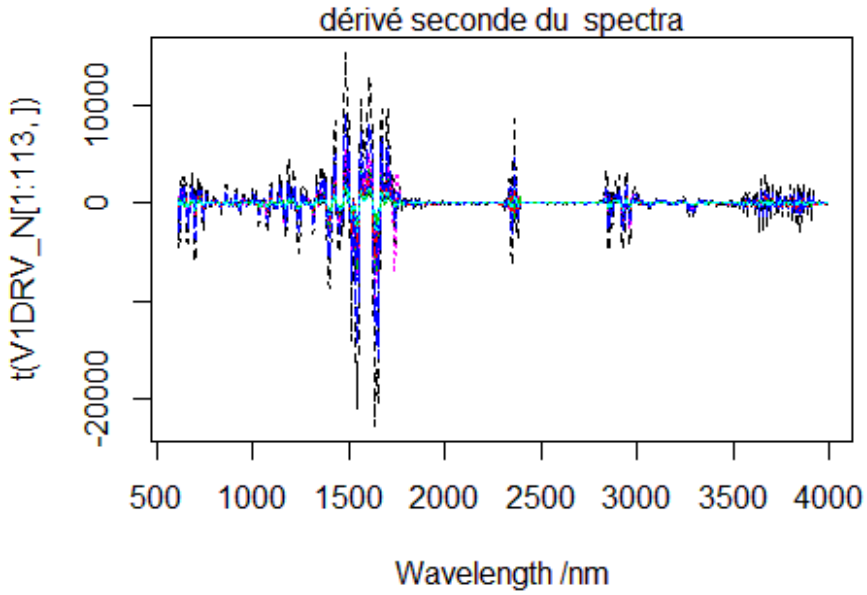
Figure 3 *Representation graphique de V1*



III-1.2. Dérivé du spectre et Normalisation de la base :

D'abord nous allons calculer la dérivé seconde la base afin d'éliminer les différences d'amplitudes et ensuite faire la normalisation de la base avant d'appliquer Kmeans notre algorithme qui va nous aider à faire le groupement de patients dans des groupes homogènes.

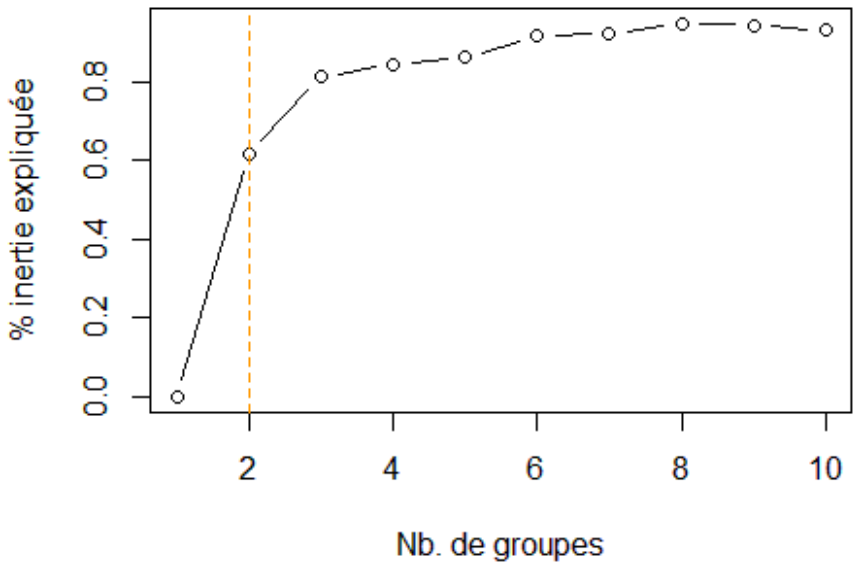
Figure 4 *Representation graphique de la derivé seconde*



Nous avons aussi effectué la troncature après le calcul de la dérivé seconde pour V1.

III-2 Recherche du nombre de groupes :

Figure 5 *Representation du choix des groupes*

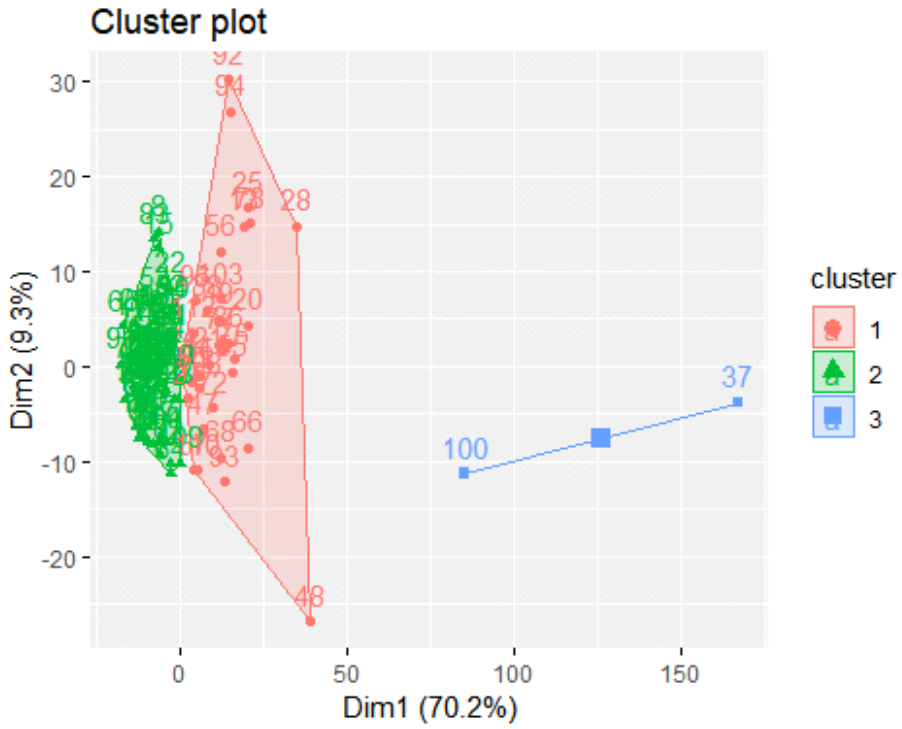


Le graphe ci-dessus nous permet de déterminer le nombre de clusters. On cherche le coude du graphe.

Le choix correct de k est souvent ambigu, mais à partir du graphique ci-dessus, nous allons essayer notre analyse par le choix de 3 clusters.

III-2.1 Mise en oeuvre de l'algoritme kmeans :

Figure 6 *Representation des groupes pour V1*



Nous avons la constitution des groupes : G1=32, G2= 79 et G3=2

On remarque que le troisième groupe est moins représenté, ce qui nous pousse à penser que c'est des « outliers » valeurs aberrantes.Il s'agit du patient 100 et du patient 37.

Que nous avons décidé de supprimer afin d'avoir une bonne représentation inter-groupes et intra-groupes.

- **Note** : variance inter et variance intra

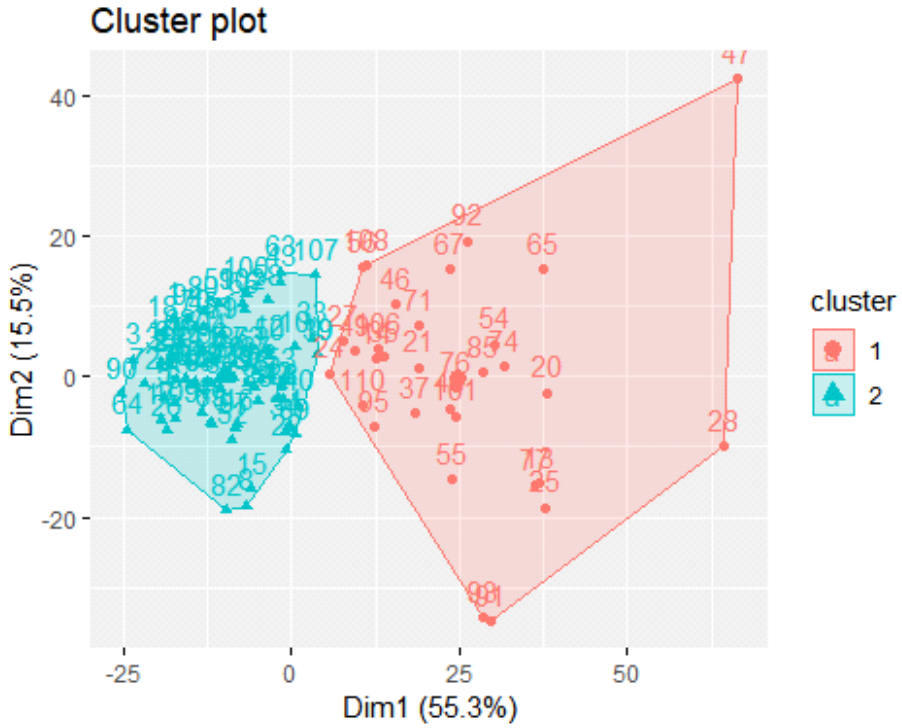
Si on considère une partition en K groupes ayant pour centres de gravité (g1,…,gK), nous obtenons l'inertie totale (IT) telle que

IT représente la variabilité totale du nuage de point : elle est constante pour un jeu de données fixé IW (within variance) représente la dispersion des points autour de leur centre. IB (between variance) représente la séparabilité des groupes : à maximiser

Stratégie d'optimisation

On cherche à trouver une partition "optimale". Une partition optimale sera définie par une IW minimale et une IB maximale Etant donné que IT = IW + IB , si on maximise IB on minimise IW

Figure 7 *Representation graphique des groupes sans les 2 patients*



Dans le premier groupe nous avons 68 patients et dans le deuxieme groupe 45 patients.

III-2.2. Détermination de groupe pour le V3 :

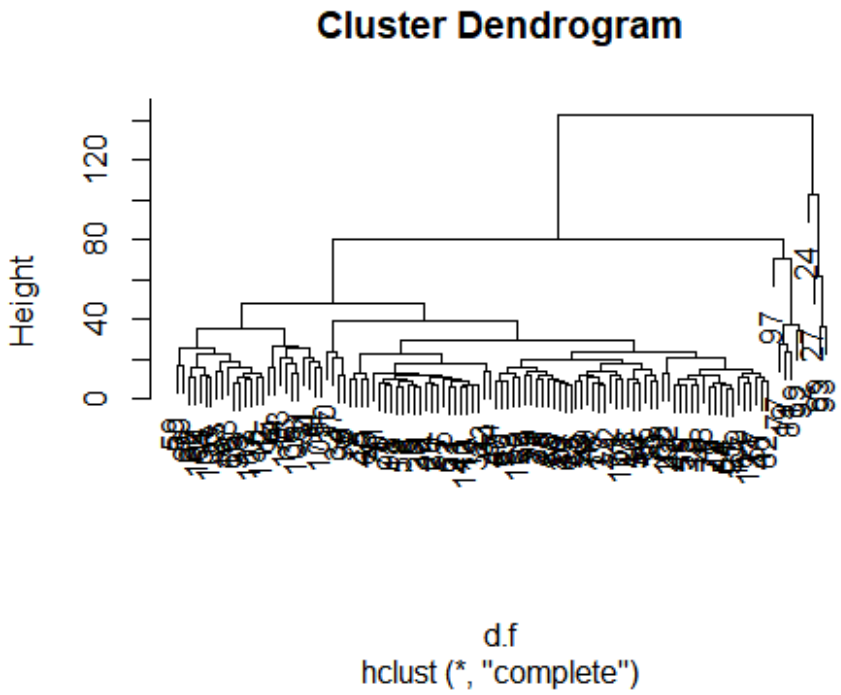
V3 correspond à la visite t_{15} c'est à dire 15 jour après la première visite.

Ici vu que pour représenter les groupes pour V2, il nous faudra refaire la même procédure que pour V1.

C'est dans ce sens que nous nous sommes mis d'accord de ne pas publier les même procédures. Mais cela a été traite dans le script.r.

III-3. Evaluation du nombre de groupe :

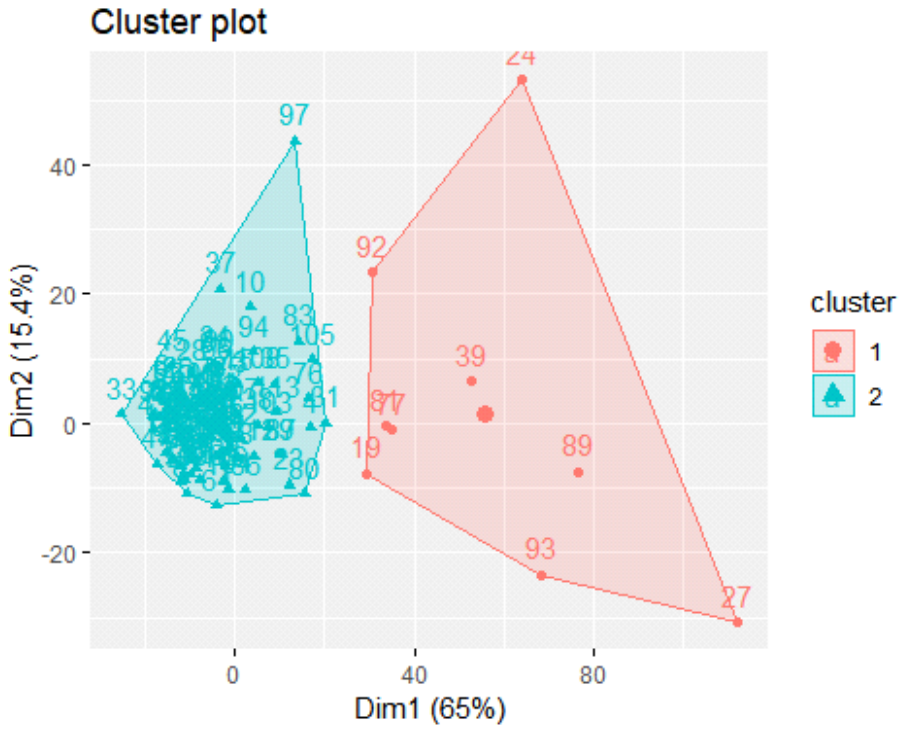
Figure 8 *La représentation du dendrogramme*



Le dendrogramme nous suggère un découpage 3 groupes.

III-3.1. Representation des groupes :

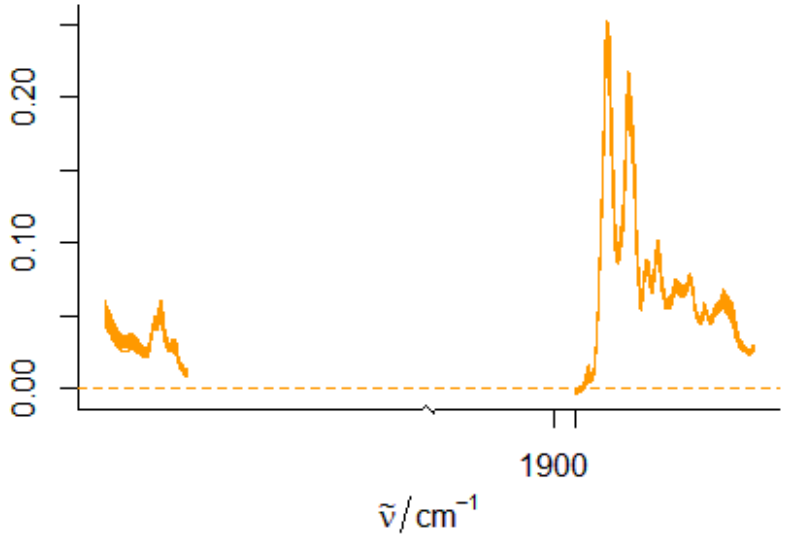
Figure 9 *Représentation graphique des groupes*



Dans le premier groupe nous avons 9 patients et les 104 autre patients sont tous dans le groupe 2.

III-3.2 Détermination de groupes pour V4 :

Figure 10 *Représentation graphique du spectre pour V4*



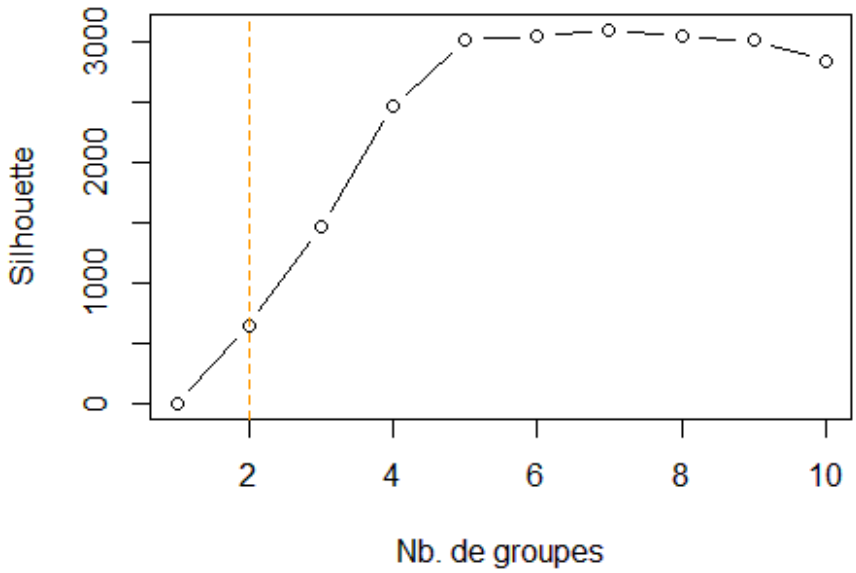
III-4 Représentation graphique des groupes :

Figure 11 *Representation graphique des groupe pour V4*

La variance intergroupe qu'on cherche à optimiser est grande soit 266.145 ,on conclut qu'on quatrième visite presque tous les patients se sont rétablis et se sont retrouvés dans un seul groupe.

III-4.1. Validation de groupe par le critère de che :

Figure 12 *Graphique de validation de notre choix*



IV -La regression LDA :

L'analyse discriminante linéaire (LDA) est une machine bien établie d'apprentissage technique pour prédire les catégories. Ses principaux avantages, par rapport à d'autres algorithmes de classification tels que les réseaux de neurones et forêts aléatoires, sont que le modèle est interprétable et que la prédiction est facile.

L'objectif ici est de prévoir les deux autres visites à partir de V1.

IV -Mise en oeuvre :

Le principe est de modéliser la distribution de chaque variable prédictive par une loi de probabilité gaussienne qui dépend de la classe à prédire et de calculer les paramètres de ces lois de probabilité. Puis, lors de la prédiction, on applique la loi de Bayes pour en déduire la probabilité de chaque classe connaissant les valeurs des variables prédictives. Dans la LDA, les frontières entre classes prédites sont en fait linéaires (d'où le nom).

- Verification des groupes avec la LDA :

```
   1  2
1 67  1
2  0 45
```

D'après les résultats de lda, dans le groupe 1 il ya 67 individus qui sont bien classés et un individu qui est mal classé par contre dans le groupe 2 tous les individus sont bien classés c'est - à-dire 45 individus. Ce qui vient conforter notre méthode précédente.

IV.1- Prédiction de V3 et V4

Figure 13 Tableau de classification des groupes

```
[1] 2 2 1 1 1 2 2 2 1 2 2 1 1 2 1 2 1 1 1 1 1 1 1 2 2 2 1 2 1 2 2 2 2 1 2 2 2
  2 2 2 1 1 2 1 1 1 2 2 2 2 1 2 1 2
[55] 1 2 1 2 1 2 2 2 1 1 1 1 2 2 2 2 2 1 2 1 1 2 2 2 2 1 1 1 2 2 2 1 1 1 1 2 2
  2 1 1 1 1 2 1 1 2 2 2 2 1 2 1 2 2
[109] 1 2 1 1 2
Levels: 1 2
```

Nous remarquons sur le graphique ci-dessus que la prevision du LDA nous donne aussi deux classes.

Le tableau ci dessous nous montre la matrice avec une colonne par classe indiquant la probabilité de chaque classe pour chaque individu prédit.

Table 1 Representation de la probabilite classe/ind_predit

	1	2
1	2.702741e-01	7.297259e-01
2	1.236912e-03	9.987631e-01
3	9.994587e-01	5.413382e-04
4	7.712160e-01	2.287840e-01
5	9.597852e-01	4.021482e-02
6	3.222696e-22	1.000000e+00
7	3.376067e-02	9.662393e-01
8	4.941190e-12	1.000000e+00
9	5.505753e-01	4.494247e-01
10	2.507385e-15	1.000000e+00
11	3.003711e-10	1.000000e+00
12	8.072995e-01	1.927005e-01
13	1.000000e+00	1.559001e-09
14	2.229065e-10	1.000000e+00
15	5.096335e-01	4.903665e-01
16	2.526841e-13	1.000000e+00

V-Conclusion :

Dans ce rapport nous avons utilisé une base de données qui correspond à des prélèvements effectués sur des patients atteint du cancer du côlon. Elle nous a permis de nous familiariser avec l'analyse des données médicales, la procédure pour obtenir la base et a l'importance de la statistique dans ce domaine.

Premièrement, nous avons représenté les groupes pour chaque visites grâce a la méthode de classification de k-means.

Deuxièmes nous avons effectué une régression en LDA afin des prédire l'appartenance d'un ou de plusieurs patients grâce à la première visite (V1).

En sommes cette étude permettra aux médecins de pouvoir savoir la réaction d'un patient juste après une première visite, afin de distinguer le dosage du médicament en minimisant le cout et en maximisant le taux de survie des cas.

Références bibliographiques

Pierre Lafaye de Micheaux, Remy Drouillhet, Benoit Liquet (2011). Le logiciel R- Maitriser le langage Aragon, Y. (2011).

Biostatistique : Alain-Jacques VALLERON

IHEAL, Guide : Comment rédiger son mémoire en M1 et en M2 ? Documentation interne, 2016, 10p

ANNEXE

DOUMBALY CAMARA SY

Vous présentes les codes du projet génères automatiquement à l'aide du package Rmarkdown.

Livre de codes

I -Importation de la base

```
PV <- read_excel("C:/Users/mirei/OneDrive/V1_V3_V4 2/V1.xlsx")

## New names:
## * `` -> ...1

PV1 <- as.data.frame(PV)
PV1=PV1[,-1]
```

➔ Transformation en object spectral :

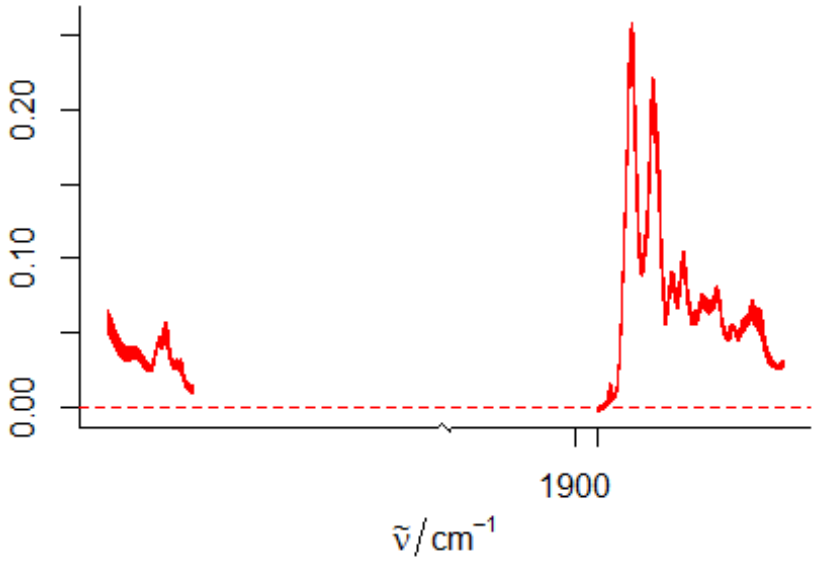
```
{
new ("hyperSpec")
V1 <-PV1
V1<- new ("hyperSpec", spc =V1,label = list (sp = "I / a.u.",.wavelength = ex
pression (tilde (nu) / cm^-1) ))
}
V1

## hyperSpec object
##    113 spectra
##    1 data columns
##    1650 data points / spectrum
## wavelength: tilde(nu)/cm^-1 [numeric] 3996.047 3993.987 ... 599.4072
## data:  (113 rows x 1 columns)
##    1. spc:  [matrix1650] -0.002102177 -0.003283256 ... 0.1601786
```

*Troncature

```
V1 =V1[,, c (3200 ~ 2800, 1800 ~ 928)]
V1

## hyperSpec object
##    113 spectra
##    1 data columns
##    620 data points / spectrum
## wavelength: tilde(nu)/cm^-1 [numeric] 3200.957 3198.897 ... 928.978
## data:  (113 rows x 1 columns)
##    1. spc:  [matrix620] 0.06480237 0.05613868 ... 0.02803944
```

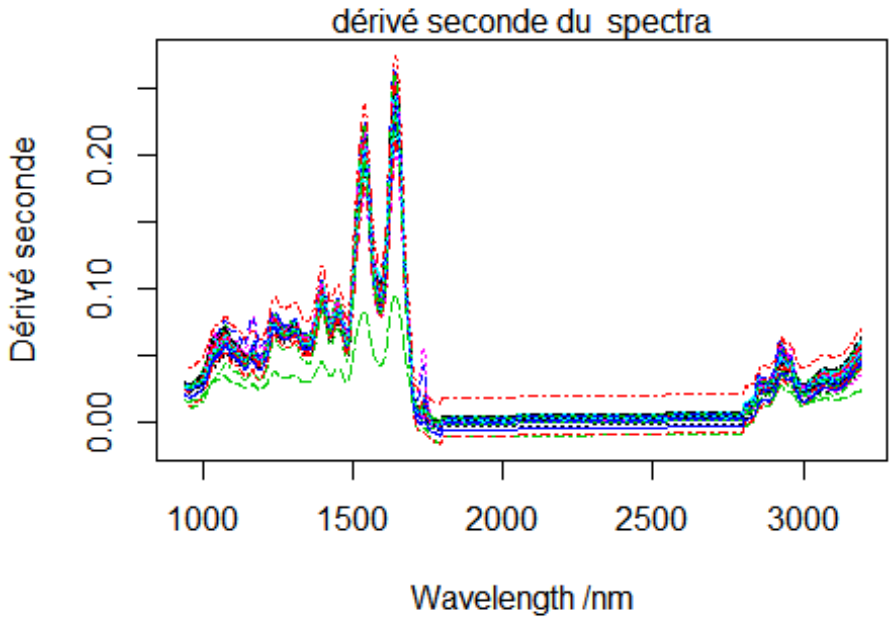
4.1.3.3 Représentation des spectres bruts

```
plot(V1,col="red")
```



I.1 Derivee seconde

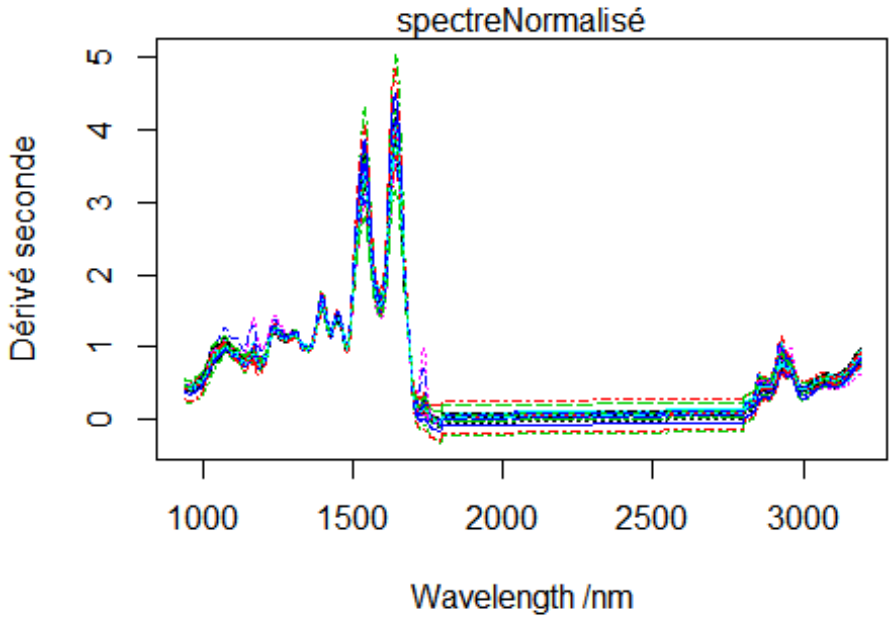
```
V1DRV <- savitzkyGolay(V1$spc, p = 3, w = 11, m = 0)
matplot(as.numeric(colnames(V1DRV)),t(V1DRV[1:113,]),type='l',xlab='Wavelengt
h /nm',ylab='Dérivé seconde')
      mtext( 'dérivé seconde du  spectra')
```



4.1.3.5 Normalisation de la base dérivée

Enfin nous allons faire la normalisation de la base avant d'appliquer Kmeans notre algorithme qui va nous aider ? faire le groupement depatients dans des groupes hommogènes.

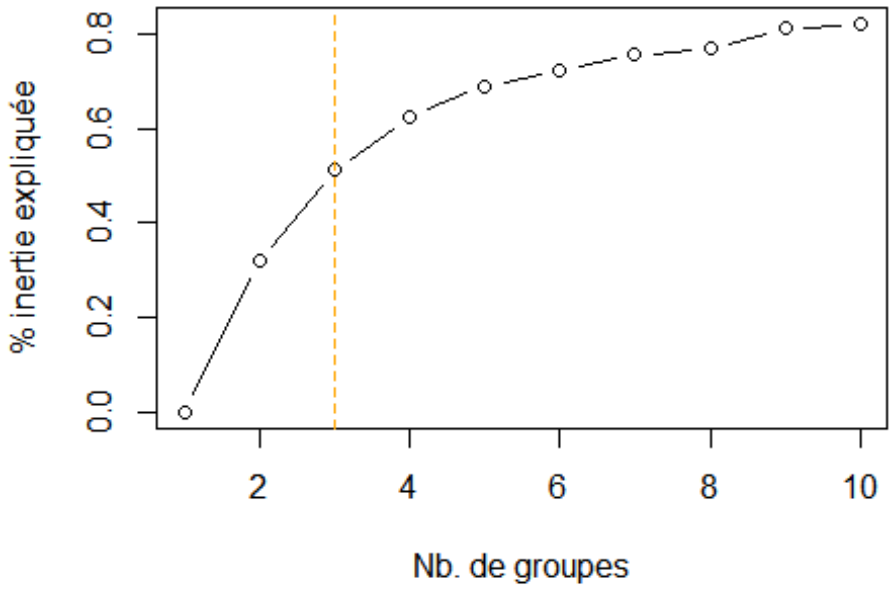
```
V1DRV_N<- V1DRV/ rowMeans(V1DRV)
matplot(as.numeric(colnames(V1DRV_N)),t(V1DRV_N[1:113,]),type='l',xlab='Wavel
ength /nm',ylab='Dérivé seconde')
      mtext( 'spectreNormalisé')
```



4.1.3.6 Evaluation du nombre de groupe

```
library(tidyverse)

## Registered S3 method overwritten by 'rvest':
##   method            from
##   read_xml.response xml2

## -- Attaching packages ---------------------------------------------------
-------------------- tidyverse 1.2.1 --

## v tibble  2.1.1     v purrr   0.3.2
## v tidyr   0.8.3     v dplyr   0.8.1
## v readr   1.3.1     v stringr 1.4.0
## v tibble  2.1.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------
-------------- tidyverse_conflicts() --
## x dplyr::collapse() masks hyperSpec::collapse()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()

library(cluster)
library(factoextra)

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at http
s://goo.gl/13EFCZ
```

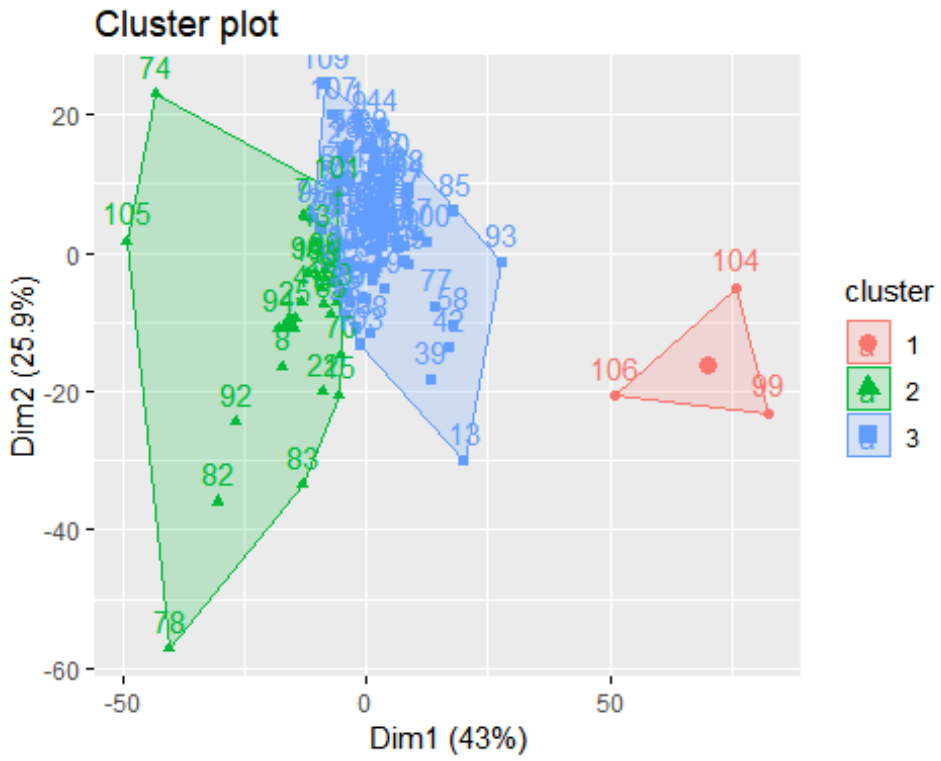
```
library(stats)
inertie.expl1<-rep(0,times=10)
for (k in 2:10) {
clus<-kmeans(V1DRV_N,centers=k,nstart=5)
inertie.expl1[k] <-clus$betweenss/clus$totss}
#graphique
plot(1:10,inertie.expl1,type="b",xlab="Nb. de groupes",ylab="% inertie expliq
uée")
abline(v=3, lty = 'dashed',col = 'orange')
```



Le graphe donne l'idée de trois groupe ,de ce faire nous allons commencer en inutialisan à trois groupes et nous changeons le nombre de k au fur à mesure

4.1.3.7 Mise en oeuvre de l'algoritme kmeans

```
set.seed(123)
km<- kmeans(V1DRV_N,center=3)
fviz_cluster(km, data =V1DRV_N)
```




```
km$size
```

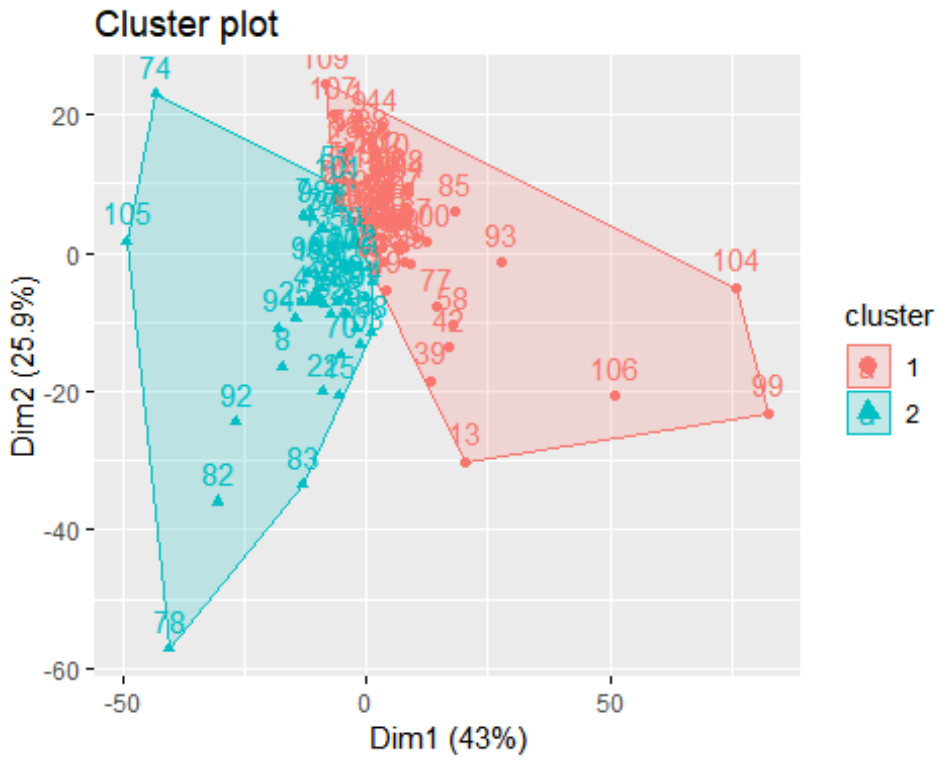
```
## [1]  3 24 86
```

```
#km$centers
km$cluster
```

```
##   [1] 3 3 3 2 3 3 2 2 3 3 3 3 3 3 2 3 2 3 3 3 3 2 3 3 2 2 3 3 3 3 3 3 3 3
## [36] 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 2 3 2 3 3 3 3 3 3 3 3 3 2 3 3 3 3
## [71] 3 3 3 2 3 3 3 2 3 3 3 2 2 3 3 3 3 3 3 2 3 2 3 2 3 2 3 3 1 3 2 3 3 1
## [106] 1 3 3 3 3 2 3 3
```

On remarque que le troisième groupe est moins représenté, nous essayons alors avec 2

```
set.seed(123)

km1<- kmeans(V1DRV_N,center=2)
fviz_cluster(km1, data =V1DRV_N)
```




```
km1$size
```

```
## [1] 68 45
```

```
#km$centers
```

```
km1$size
```

```
## [1] 68 45
```

```
km1$cluster
```

```
##   [1] 1 1 1 2 2 2 2 2 1 1 1 1 1 1 2 2 2 1 1 2 1 2 1 1 2 2 1 2 1 1 1 2 1 2
1
##  [36] 1 2 2 1 1 1 1 2 1 1 1 1 2 1 2 2 1 2 2 2 2 1 1 1 1 1 1 1 2 1 1 1 1
2
##  [71] 1 1 1 2 1 2 1 2 2 1 1 2 2 2 1 1 1 1 2 2 1 2 1 2 1 2 1 2 1 1 2 1 2 1
2
## [106] 1 1 1 1 1 2 2 1
```

```
km1$totss
```

```
## [1] 279.013
```

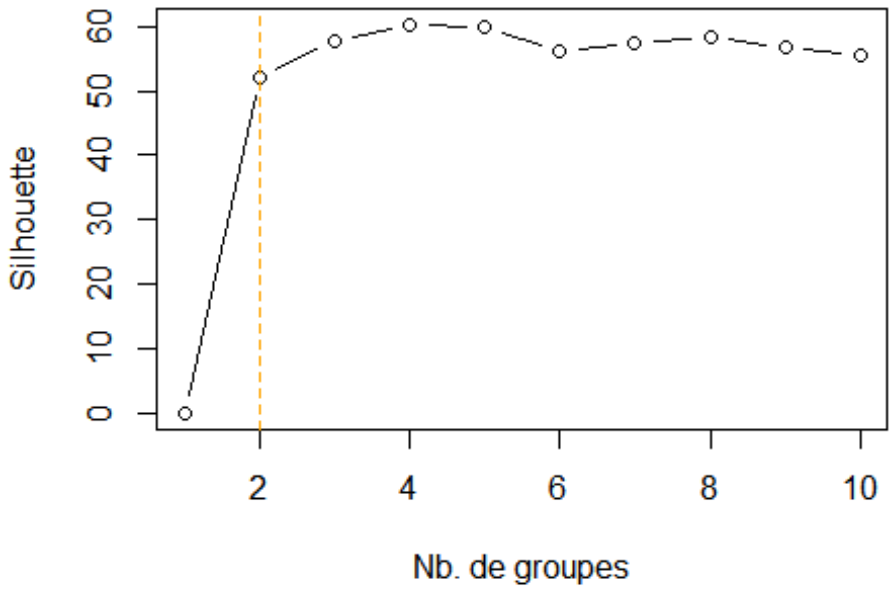
```
km1$withinss
```

```
## [1] 131.62455  82.01462
```

```
km1$tot.withinss

## [1] 213.6392

km1$betweenss

## [1] 65.37388
```

4.1.3.8 validation de groupe par le critère de che

```
 library(fpc)

sol.kmeans1<-kmeansruns(V1DRV_N,krange=2:10,criterion="ch")
#graphique
plot(1:10,sol.kmeans1$crit,type="b",xlab="Nb. de groupes",ylab="Silhouette")
abline(v=2, lty = 'dashed',col = 'orange')
```



IV- Régression LDA :

```{r}

library(MASS)

ldaA=lda(BLDA$GROUPE~.,data=BLDA)

predit=predict(ldaA,BLDA)
mat1=table(BLDA[,"GROUPE"],predit$class)
mat
pred <- predict(ldaA, newdata = PV3)

```

IV-Prediction de V3 :

```{r}
Gp_v3=km3$cluster
Gp_v3=data.frame(Gp_v3)#->Groupe constitué de 1 et 2 groupes
Gp3=Gp_v3[Gp_v3==1]
Gp2=Gp_v3[Gp_v3==2]
vv3=data.frame(V3DRV_N) #->v3
BLDA3=cbind(Gp_v3,vv3) #base concatenée
colnames(BLDA)[1] = "GROUPE"
 pred <- predict(ldaA, newdata = vv3)     #Prevision de V3
attributes(pred)
```

```
pred$class

pred$posterior

mat=table(vv3,predit$class)
```