Analyse des données: Diagnostiquer les cancers du sein - Tumeur bénigne ou maligne

Atontsa Nguemo Miradain, miradain.atontsan@gmail.com

Contents

1 Introduction

L'objectif de cette analyse est de diagnostiquer les tumeurs du sein sur les patients de l'hôpital universitaire du Wisconsin. Plus précisément, nous voulons, à partir des données d'imagerie médicales prises sur les patients, decider si une tumeur est bénigne ou maligne.

La base de donnée contient 569 patients. Les 11 variables d'étude sont:

1) diagnosis: le diagnostic M pour maligne et B pour bénigne;
2) radius_mean: le rayon moyen de la tumeur;
3) Texture_mean: La texture moyenne. Ceci capture la variation de l'intensité du gris des pixels;
4) Perimeter_mean: le périmètre nucléaire approximatif;
5) Area_mean: la surface nucléaire moyenne. On compte les pixels à l'intérieur de la cellule et sur le périmètre;
6) Smoothness_mean: la régularité moyenne du contour nucléaire;
7) Compactness_mean: La compacité moyenne est obtenue via le ratio périmètre/surface plus une certaine compensation;
8) Concavity_mean:La concavité nucléaire moyenne;
9) Concave_points_mean: le nombre moyen de points de concavité;
10) Symmetry_mean: La symétrie nucléaire moyenne;
11) Fractal_dimension_mean: La dimension fractale moyenne. Cet indicateur capture le defaux d'approximation d'une cellule par un polygone. Il capture donc aussi la régularité du contour.

2 Présentation des données, analyse descriptive

2.1 Données

```
library(bitops)
library(RCurl)
UCI_data_URL <-
  getURL('https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data')
names <- c('id_number', 'diagnosis', 'radius_mean',
         'texture_mean', 'perimeter_mean', 'area_mean',
         'smoothness_mean', 'compactness_mean',
         'concavity_mean','concave_points_mean',
         'symmetry_mean', 'fractal_dimension_mean',
         'radius_se', 'texture_se', 'perimeter_se',
         'area_se', 'smoothness_se', 'compactness_se',
         'concavity_se', 'concave_points_se',
         'symmetry_se', 'fractal_dimension_se',
         'radius_worst', 'texture_worst',
         'perimeter_worst', 'area_worst',
         'smoothness_worst', 'compactness_worst',
         'concavity_worst', 'concave_points_worst',
         'symmetry_worst', 'fractal_dimension_worst')
breast_cancer <- read.table(textConnection(UCI_data_URL), sep = ',', col.names = names)

breast_cancer$id_number <- NULL
breast_cancer<-breast_cancer[1:11]
```

2.2 Description des données

```
dim(breast_cancer)
```

```
[1] 569  11
```

```
str(breast_cancer)
```

```
'data.frame':   569 obs. of  11 variables:
 $ diagnosis             : Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
 $ radius_mean           : num  18 20.6 19.7 11.4 20.3 ...
 $ texture_mean          : num  10.4 17.8 21.2 20.4 14.3 ...
 $ perimeter_mean        : num  122.8 132.9 130 77.6 135.1 ...
 $ area_mean             : num  1001 1326 1203 386 1297 ...
 $ smoothness_mean       : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
 $ compactness_mean      : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
 $ concavity_mean        : num  0.3001 0.0869 0.1974 0.2414 0.198 ...
 $ concave_points_mean   : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
 $ symmetry_mean         : num  0.242 0.181 0.207 0.26 0.181 ...
 $ fractal_dimension_mean: num  0.0787 0.0567 0.06 0.0974 0.0588 ...
```

Nous donnons en annexe la table sommaire des tendances centrales sur les données. Nous allons cependant représenter le tableau de valeurs moyennes des variables par groupe d'interêt: "M" et "B".

```
table<- aggregate(breast_cancer[2:11], by=list(breast_cancer$diagnosis),FUN = mean)
rownames(table)<- table$Group.1
```

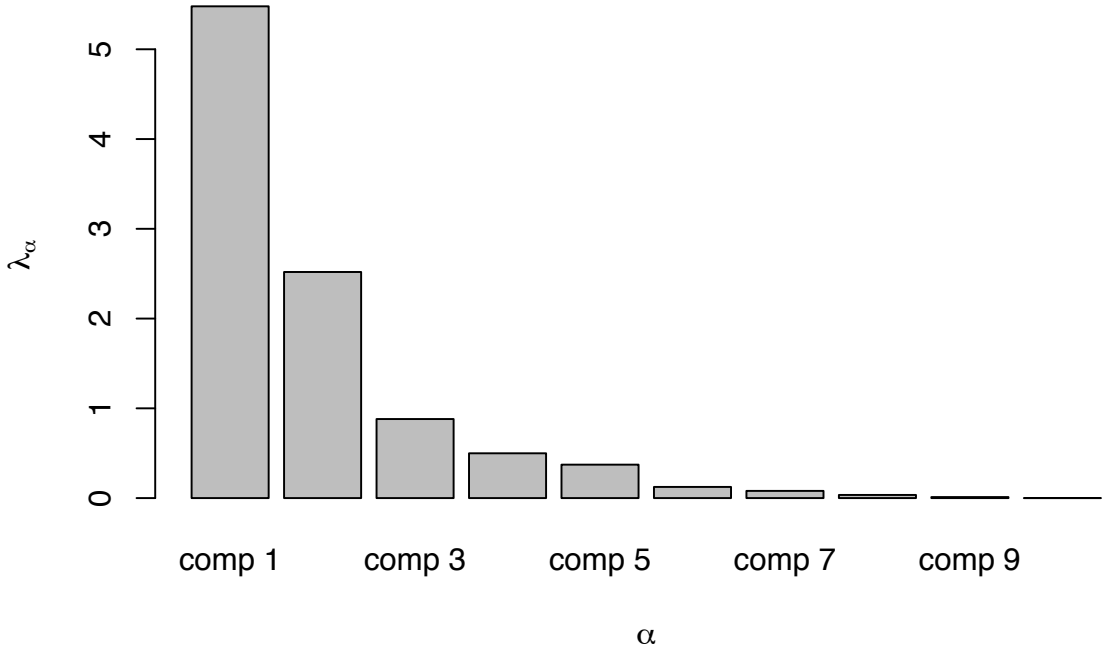
```
table$Group.1<- NULL
table
```

```
  radius_mean texture_mean perimeter_mean area_mean smoothness_mean
B    12.14652     17.91476       78.07541  462.7902      0.09247765
M    17.46283     21.60491      115.36538  978.3764      0.10289849
  compactness_mean concavity_mean concave_points_mean symmetry_mean
B       0.08008462     0.04605762          0.02571741      0.174186
M       0.14518778     0.16077472          0.08799000      0.192909
  fractal_dimension_mean
B             0.06286739
M             0.06268009
```

3 Analyse en composantes principales

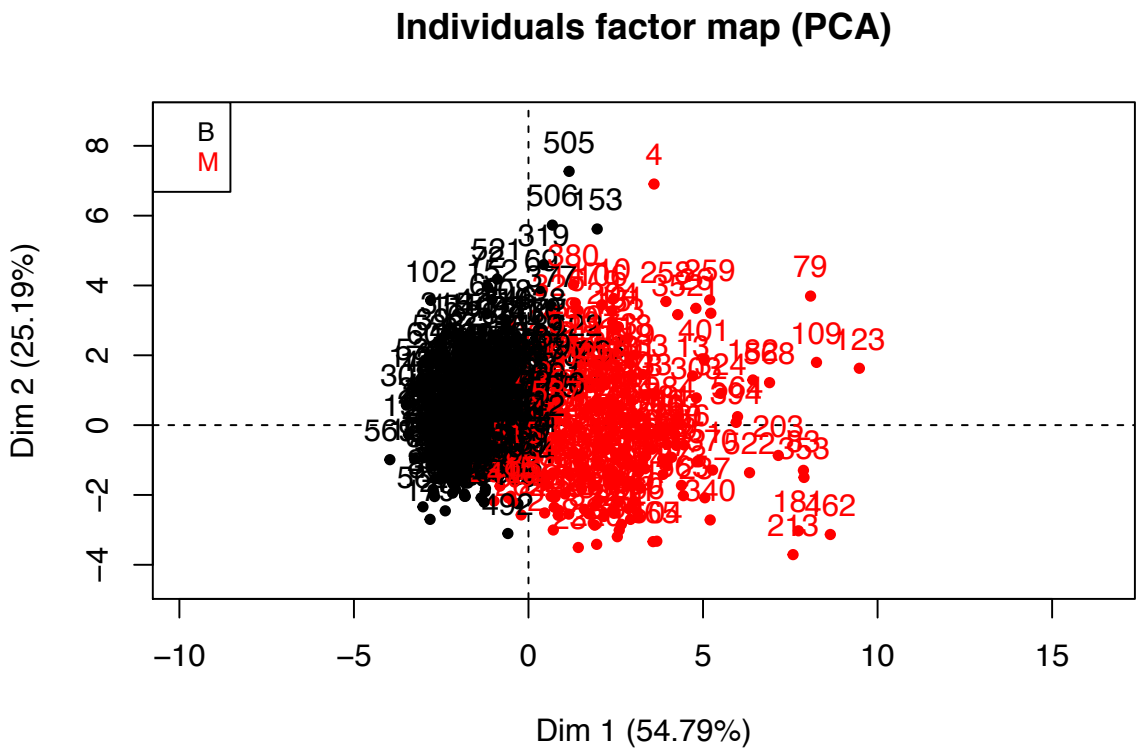
Pour effectuer l'ACP, nous prendrons la variable qualitative à expliquer "diagnosis" comme supplémentaire.

```
library(FactoMineR)
res_pca<- PCA(breast_cancer, scale.unit = TRUE, ncp = 5, quali.sup = 1, graph = F)
barplot(res_pca$eig[,"eigenvalue"], xlab = expression(alpha), ylab = expression(lambda[alpha]))
```



Nous pouvons observer sur cet histogramme que la variabilité est mieux expliqué sur le premier plan factoriel avec une attention particulière sur le premier axe factoriel qui se démarque de tous les autres. Nous allons donc visualiser et interpréter les individus sur ce plan.

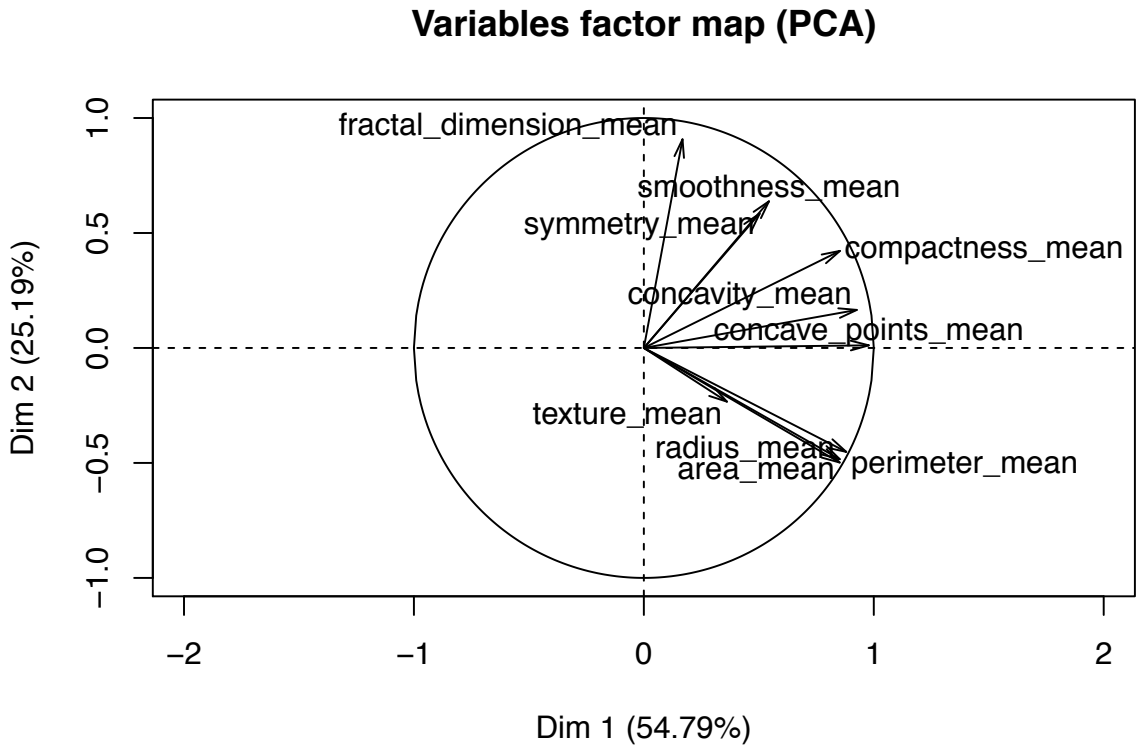
```
plot.PCA(res_pca, axes = c(1,2), choix = "ind", habillage = 1)
```



Nous Déduisons de la visualisation ci-dessus que les deux groupes (“B” et “M”) occupent presque des positions différentes sur le premier axe factoriel. En effet, Les tumeurs malignes sont à droite de l’axe verticale, et les tumeurs bénignes sont à gauche. La dimension 1 caractérise donc au mieux la séparation des deux familles de tumeurs.

Nous allons maintenant nous interesser aux variables sur le premier plan factoriel.

```
plot.PCA(res_pca, axes = c(1,2), choix = "var")
```



Nous pouvons faire les interprétations suivantes:

a) Sur le premier axe: il n'y a pas d'opposition entre les variables. Ce qui veux dire qu'elles contribuent tous à la séparation des groupes "M" et "B" sur cet axe. On peut cependant distinguer que les variables qui contribuent le plus su cet axe, à savoir celles qui ont la plus grande corrélation (voir annexe) sont: "concave_points_mean", "concavity_mean", "perimeter_mean", "compactness_mean", "area_mean", "radius_mean". En lien avec le graphe des individus, on peut aussi dire que les tumeurs malignes ont les plus grande valeurs sur ces variables citées, et que les tumeur bénignes ont des valeurs plus petites.

b) Le deuxième axe est celle qui oppose la régularité des tumeurs à leurs surface ou encore elle oppose la forme de la tumeur à sa grosseur. Elle oppose en effet les variables "fractal_dimension_mean", "smoothness_mean" et "symmetry_mean" aux variables "perimeter_mean", "area_mean" et "radius_mean".

c) La variable "texture_mean" est moins représenté sur ce plan.

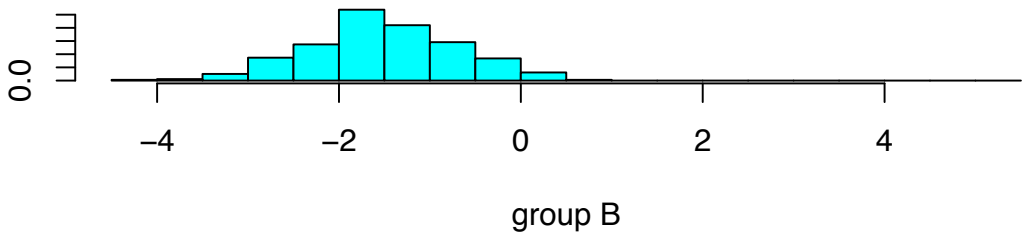
4 Analyse discriminante

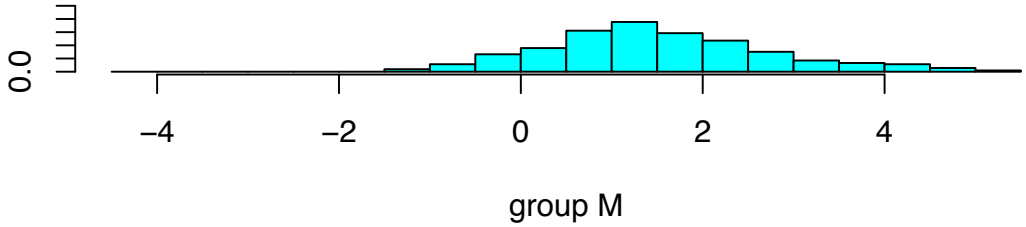
La représentation sur le premier plan factorielle (notamment la première dimension) nous fait remarquer qu'une classification linéaire pourrais bien être utilisée pour classifier les deux groupes "M" et "B". Pour cette raison, nous opérons dans cette section une analyse discriminante linéaire.

```
library(MASS)
res_lda<- lda(breast_cancer$diagnosis ~ ., data=breast_cancer[2:11], method="mle")
```

Les coefficients de l'analyse discriminantes sont repris en annexe. Nous pouvons cependant représenter l'histogramme des differnts groupes sur l'axe discriminante contruite par cette methode.

```
#ldahist(data=z[,1], g=breast_cancer1$diagnosis)
plot(res_lda )
```





4.1 Qualité de la classification

```
# Predicted classes
K.predicted<- predict(object = res_lda, newdata=breast_cancer[,-1])$class
# Table de classes réelles vs classes prédictes
validation<-table(breast_cancer$diagnosis, K.predicted)
validation
```

```
   K.predicted
      B   M
  B 351   6
  M  29 183
```

```
#Qualité de la règle discriminante
proportion<- sum(diag(validation))/nrow(breast_cancer)
proportion
```

```
[1] 0.9384886
```

93.82% des patients sont correctement classés sur les deux classes. Les 6.2% d'erreurs viennent du fait que 29 tumeurs malignes et 6 tumeurs bénignes ont étés mal classés. Ceci se dégage clairement du premier plan factoriel où l'on peut observer qu'il n'y a pas un séparateur linéair parfait entre les deux groupes "M" et "B".

4.2 Sensibilité et spécificité

Pour la classification lda: La spécificité du modèle de classification est de

```
sensibilité<-  validation[2,2]/(validation[2,2]+validation[2,1])
sensibilité
```

```
[1] 0.8632075
```

```
spécificité<- validation[1,1]/(validation[1,1]+validation[1,2])
spécificité
```

```
[1] 0.9831933
```

La sensibilité du modèle de classification (lda) est de l'ordre de 86,32% tandis que la spécificité du modèle de l'ordre de 98,31%. Ce qui voudrais dire en d'autres termes que le modèle est un peu plus précis pour la prédiction des tumeurs benignes que celles malignes. Cette tendance pourrais bien changer en "déplaçant' parallèlement la droite (frontière) qui sépare les deux groupes. En effet sur l'histogramme, on peut observer qu'en déplaçant cette droite de manière parallèle vers la gauche, on augmenterait la sensibilité en réduisant ainsi la spécificité.

5 Analyse des correspondances

5.1 Discretisation des données

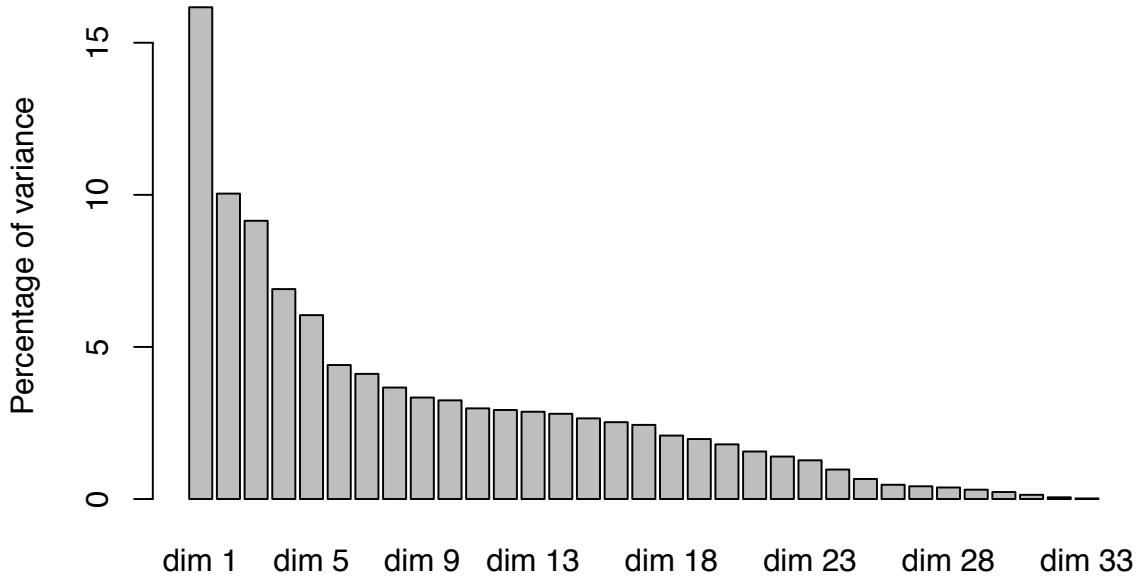
Nous allons utiliser la fonction cut et le kmeans pour discrediter naturellement nos 10 variables quantitatives. Dans cette discrétisation, les indexes des niveaux sont attribués de façon croissante. Par exemple: la variable discrétisée du rayon "radius.disc" aura les niveaux "radius1", "radius2", "radius3" et "radius4" classés ainsi du plus petit au plus grand.

```
Kr<-kmeans(breast_cancer$radius_mean, 4)
radius.disc<- cut(Kr$cluster, breaks =c(0,1.5,2.5,3.5,4), labels=c("radius1", "radius2", "radius3", "ra

Kt<-kmeans(breast_cancer$texture_mean,4)
texture.disc<- cut(Kt$cluster, breaks =c(0,1.5,2.5,3.5,4), labels=c("texture1", "texture2", "texture3",

Kp<-kmeans(breast_cancer$perimeter_mean,4)
perimeter.disc<- cut(Kp$cluster, breaks =c(0,1.5,2.5,3.5,4), labels=c("perimeter1", "perimeter2", "peri

ka<-kmeans(breast_cancer$area_mean, 5)
area.disc<- cut(ka$cluster, breaks =c(0,1.5,2.5,3.5,4.5, 5), labels=c("area1", "area2", "area3", "area4

ks<-kmeans(breast_cancer$smoothness_mean, 6)
smooth.disc<- cut(ks$cluster, breaks =c(0,1.5,2.5,3.5,4.5,5.5,6), labels=c("smooth1", "smooth2", "smooth

kc<-kmeans(breast_cancer$compactness_mean, 4)
compact.disc<- cut(kc$cluster, breaks =c(0,1.5,2.5,3.5,4), labels=c("compact1", "compact2", "compact3",

kcom<-kmeans(breast_cancer$concavity_mean, 4)
concavity.disc<- cut(kcom$cluster, breaks =c(0,1.5,2.5,3.5,4), labels=c("concavity1", "concavity2", "co

kconp<-kmeans(breast_cancer$concave_points_mean, 3)
concavePoint.disc<- cut(kconp$cluster, breaks =c(0,1.5,2.5,3), labels=c("concavePoint1", "concavePoint2

ks<-kmeans(breast_cancer$symmetry_mean, 4)
symmetry.disc<- cut(ks$cluster, breaks =c(0,1.5,2.5,3.5,4), labels=c("symmetry1", "symmetry2", "symmetry

kf<-kmeans(breast_cancer$fractal_dimension_mean, 5)
fractal.disc<- cut(kf$cluster, breaks =c(0,1.5,2.5,3.5,4.5,5), labels=c("fractal1", "fractal2", "fractal

data<-data.frame(radius.disc,texture.disc,perimeter.disc, area.disc,smooth.disc,compact.disc, concavity
```

5.2 Analyse des correspondances

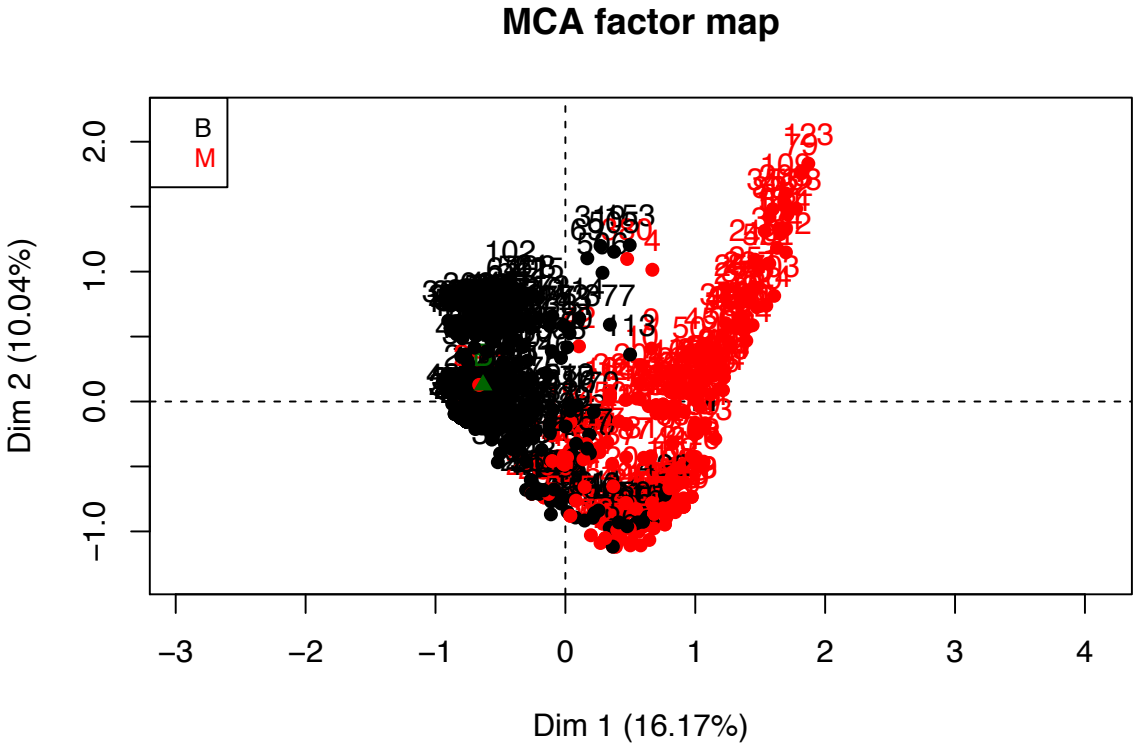
Nous effectuons une analyse des correspondances sur la nouvelle base de données discrétisée en considérant la variable “diagnosis” comme variable qualitative supplémentaire vue que le but ici est d’observer sa reesemblance avec d’autres variables. Ceci nous permettra aussi d’observer les facteurs qualitatifs qui influencent les différents groupes de la variable “diagnosis”.

```
res.mca = MCA(data,  quali.sup=11,  graph = FALSE)
#L'histogramme de variabilité des dimensions
barplot(res.mca$eig[,"percentage of variance"], ylab="Percentage of variance")
```



Sur l’histogramme des pourcentages de variance des différentes dimensions, on vois que le premier axe factoriel se distingue clairement des autres dimensions. Nous apporterons une attention particulière sur cette dimension. Nous explorerons dans la suite en détail pour voir si cette dimension permet d’observer les deux grand groupes de l’étude: “M”, “B”.

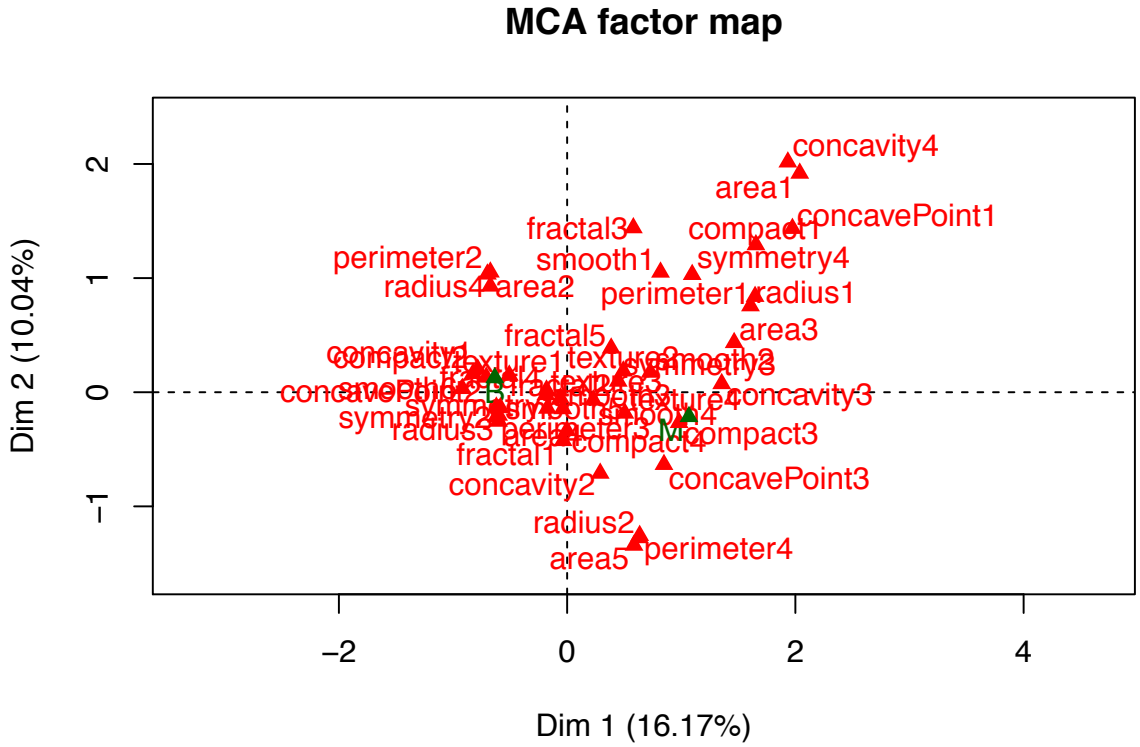
```
#Représentation sur le premier plan factoriel les individus par groupe: "M", "B".
plot.MCA(res.mca, choix = "ind", habillage = 11, invisible=c("var"))
```



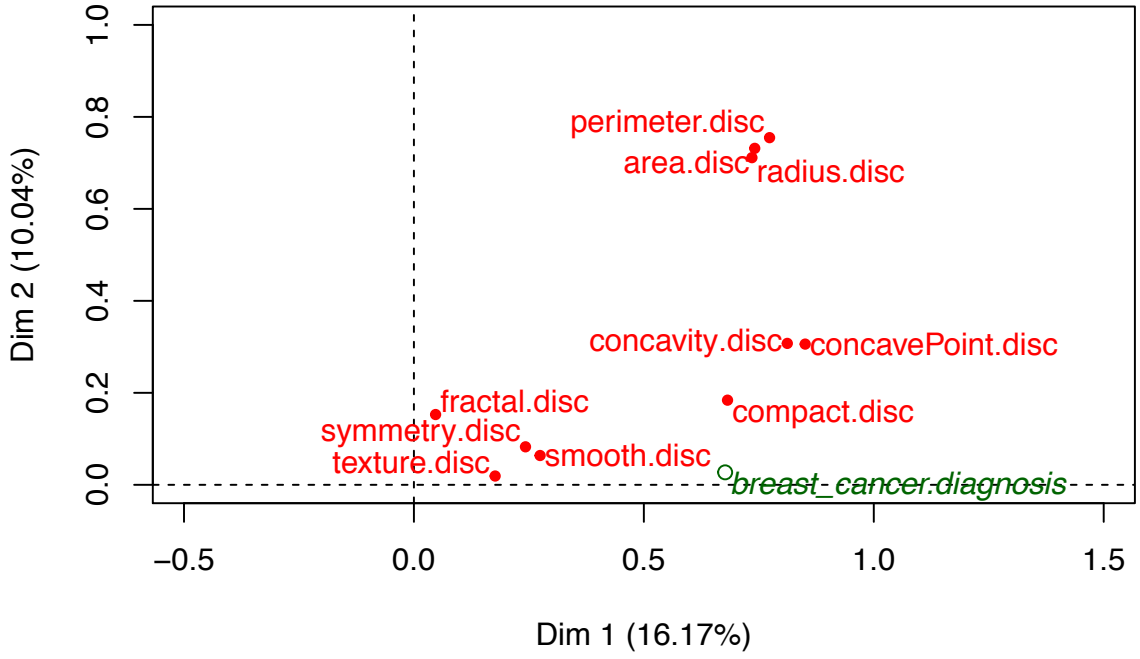
En visualisant nos données sur le premier plan factoriel (voir figure ci-dessus), on peut remarquer que les deux groupes occupent presque des positions différentes sur le premier axe factoriel. En effet, Les tumeurs malignes sont à droite de l'axe verticale, et les tumeurs bénignes sont à gauche. La dimension 1 caractérise donc au mieux la variabilité des deux groupes "M" et "B". Ce résultat est similaire à celui obtenu en ACP sur le premier axe.

Nous allons représenter ci-dessous le graphe conjoint des variables et individus pour interpréter les facteurs (des variables qualitatives) déterminants dans les groupes "B" et "M", ensuite nous représenterons le graphe des variables pour observer celles qui ressemblent à notre variable d'intérêt.

```
#Bonne représentation qui met en avant la variabilité. A inclure dans le rapport
plot.MCA(res.mca, invisible=c("ind"))
```




```
plot.MCA(res.mca, choix = "var")
title(main = "Graphe des variables")
```



Comme pour l'ACP, les diagrammes de l'ACM nous permettent de faire les analyses suivantes:

a) Nous remarquons que les tumeurs malignes ont des plus grosses valeurs sur les variables "concavity.disc",

"compact.disc", "radius.disc", "concavePoint.disc" contrairement au groupe tumeurs bénignes;

b) Les variables "perimeter.disc", "area_disc", "radius_disc", "concavity.disc", concavePoint.disc, "compact.disc" ressemblent à la variable et "breast_cancer.diagnosis" sur le premier axe factoriel. Ce qui veux dire en d'autres termes que ces variables influence le plus dans l'explication de la variable "diagnosis".

6 Conclusions

Les méthodes que nous avons utilisé dans cette analyse de données ACP, LDA et ACM ont été toutes dans le but de diagnostiquer un patient suivant qu'il ait une tumeur maligne ou bénigne.

l'AMC et l'APC ont donnée des résutats concordants, à savoir que le premier axe factoriel permettait au mieux de distinguer des patients appartenant à ces deux groupes. Enfin nous avons appris de ces deux analyses que les variables "perimeter", "area", "radius", "concavity", "concavePoint" et "compact" contribuaient le mieux à la distinction des deux groupes.

La méthode LDA nous a donné un modèle de classification linéaire des deux groupes "M" et "B" avec un taux d'exactitude de 93.82%. Nous avons constaté qu'il classait sur l'échantillon, mieux les tumeurs bénignes que les tumeurs maligne. Nous avons cependant noter que l'on pouvais bien déplacer la frontière en fonction de l'utilisation de ce modèle.

7 Annexes

```
#Statistique descriptive sur les données
summary(breast_cancer)
```

```
 diagnosis  radius_mean      texture_mean   perimeter_mean
 B:357     Min.   : 6.981   Min.   : 9.71   Min.   : 43.79
 M:212     1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17
           Median :13.370   Median :18.84   Median : 86.24
           Mean   :14.127   Mean   :19.29   Mean   : 91.97
           3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10
           Max.   :28.110   Max.   :39.28   Max.   :188.50
   area_mean      smoothness_mean   compactness_mean  concavity_mean
 Min.   : 143.5   Min.   :0.05263   Min.   :0.01938   Min.   :0.00000
 1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492   1st Qu.:0.02956
 Median : 551.1   Median :0.09587   Median :0.09263   Median :0.06154
 Mean   : 654.9   Mean   :0.09636   Mean   :0.10434   Mean   :0.08880
 3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040   3rd Qu.:0.13070
 Max.   :2501.0   Max.   :0.16340   Max.   :0.34540   Max.   :0.42680
 concave_points_mean symmetry_mean    fractal_dimension_mean
 Min.   :0.00000     Min.   :0.1060   Min.   :0.04996
 1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770
 Median :0.03350     Median :0.1792   Median :0.06154
 Mean   :0.04892     Mean   :0.1812   Mean   :0.06280
 3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612
 Max.   :0.20120     Max.   :0.3040   Max.   :0.09744
```

```
#Coefficients de la variable discriminante linéaire
res_lda$scaling
```

```
                LD1
```

```
radius_mean              2.177663123
texture_mean             0.097651089
perimeter_mean          -0.244312909
area_mean               -0.004243099
smoothness_mean          8.625383281
compactness_mean         0.432236655
concavity_mean           3.598687007
concave_points_mean     28.580051340
symmetry_mean            4.496983929
fractal_dimension_mean -0.530147316
```

```
#Description de l'APC sur le premier plan factoriel
dimdesc(res_pca, axes = c(1,2))
```

```
$Dim.1
$Dim.1$quanti
                        correlation       p.value
concave_points_mean       0.9784767  0.000000e+00
concavity_mean            0.9263041 1.576740e-242
perimeter_mean            0.8801838 1.188744e-185
compactness_mean          0.8530271 2.818254e-162
area_mean                 0.8521933 1.234467e-161
radius_mean               0.8518471 2.273559e-161
smoothness_mean           0.5441529  3.600744e-45
symmetry_mean             0.5037943  5.936400e-38
texture_mean              0.3615142  5.218608e-19
fractal_dimension_mean    0.1681455  5.558543e-05

$Dim.1$quali
                  R2       p.value
diagnosis 0.6169068 3.134668e-120

$Dim.1$category
             Estimate       p.value
diagnosis=M  1.901185 3.134668e-120
diagnosis=B -1.901185 3.134668e-120


$Dim.2
$Dim.2$quanti
                        correlation       p.value
fractal_dimension_mean    0.9074214 1.200566e-215
smoothness_mean           0.6379325  2.448070e-66
symmetry_mean             0.5845103  1.944408e-53
compactness_mean          0.4221750  5.339024e-26
concavity_mean            0.1655066  7.288856e-05
texture_mean             -0.2335828  1.722977e-08
perimeter_mean           -0.4517650  5.778245e-30
area_mean                -0.4837977  1.016271e-34
radius_mean              -0.4982197  4.971765e-37

$Dim.2$quali
                   R2      p.value
diagnosis 0.01964015 0.0008018806
```

```
$Dim.2$category
              Estimate      p.value
diagnosis=B  0.2300075 0.0008018806
diagnosis=M -0.2300075 0.0008018806
```