Fondements statistiques - Exercice 1

Céline LY and Hugo LAULLIER

20 Décembre 2020

Chargement des librairies et des données

```
## Loading required package: FactoMineR
## Loading required package: factoextra
## Loading required package: ggplot2
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
## Loading required package: corrplot
## corrplot 0.84 loaded
## Loading required package: cluster
## Loading required package: klaR
## Loading required package: MASS
```

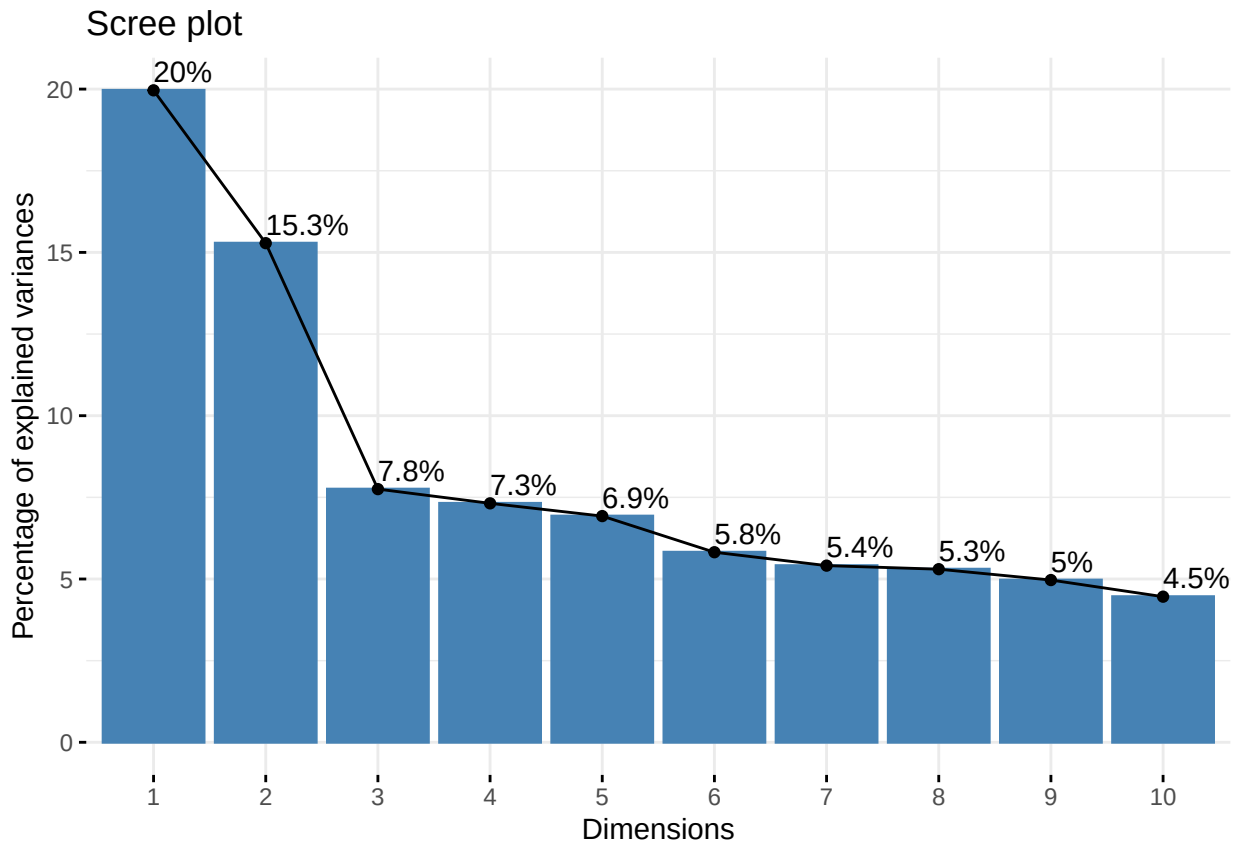
ACP

Afin de mettre en valeur d'éventuelles similitudes entre individus, et entre variables, nous allons réaliser une ACP. Nous préférons une ACP à une AFC, car les variables sont quantitatives.

```
alim.acp <- PCA(alim, scale.unit=TRUE, ncp=5, quali.sup = c(16:18),graph=FALSE)
```

Choix du nombre d'axes factoriel

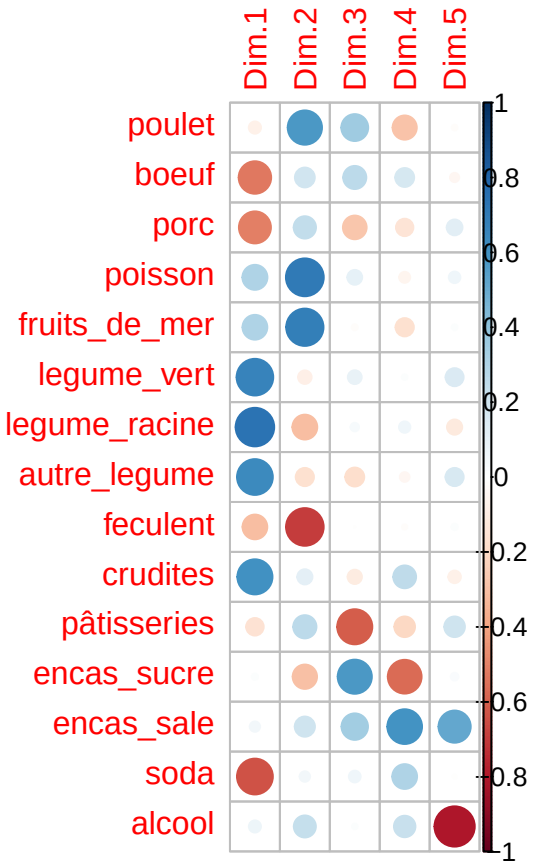
```
fviz_eig(alim.acp, addlabels = TRUE, choice = "variance" )
```



En appliquant le critère du coude, nous choisissons les deux premiers axes factoriels. Ils ne couvriront en effet que 35.3% de l'information, mais il faut prendre beaucoup plus de dimensions pour avoir une augmentation significative de la couverture d'information. Il est donc pertinent de ne choisir que les deux permières, mais il faut rester conscient de l'importante perte d'informations.

Pour se donner une idée, on peut voir, à l'aide de cette matrice, la contribution des aliments aux différents axes factoriels.

```
corrplot(alim.acp$var$cor)
```

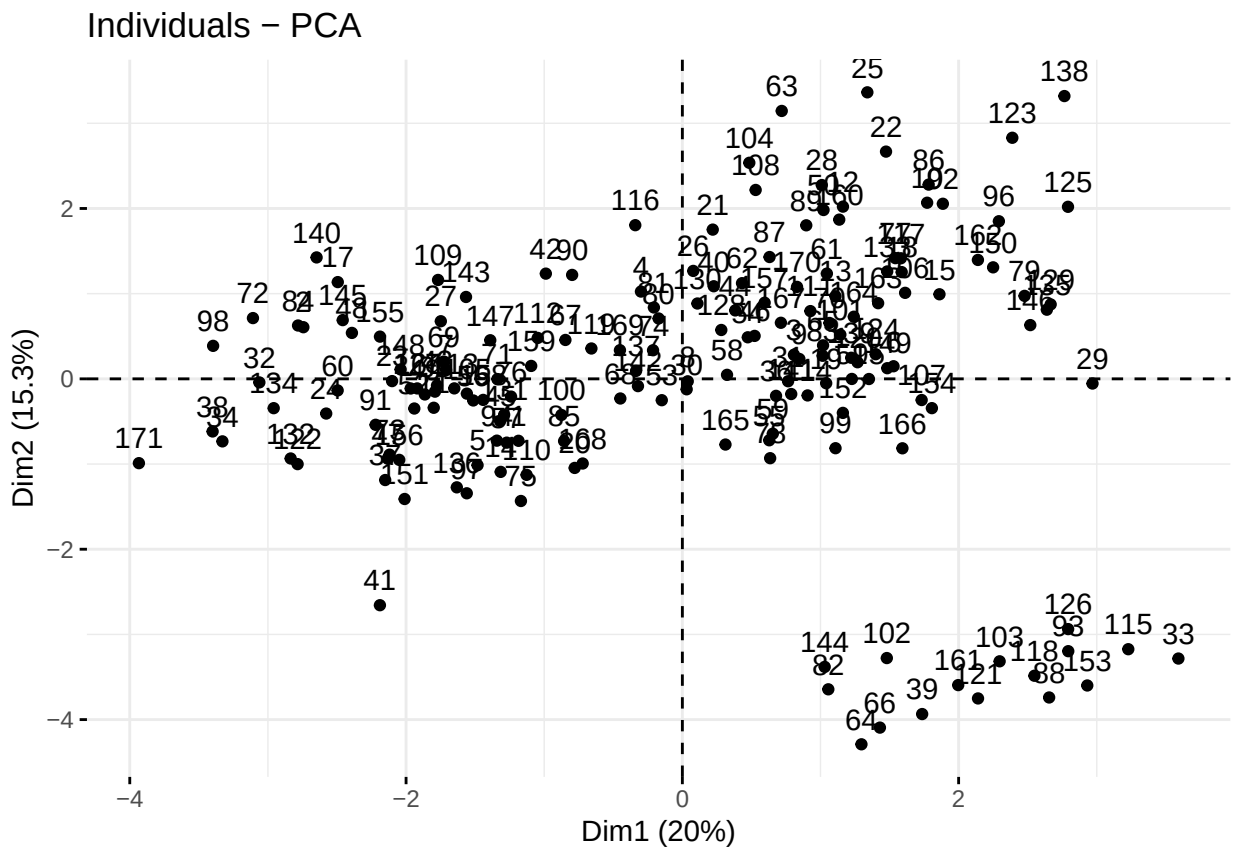


On voit clairement que les aliments d'origine végétale (les légumes) contribuent très fortement au premier axe, et les aliments d'origine animale (poisson, fruits de mer, poulet...) contribuent très fortement au second.

Ressemblances/oppositions entre les individus

Afin de mettre en valeur les ressemblances, nous allons observer la projection des individus sur le plan formé par les deux premiers axes vectoriels.

```
fviz_pca_ind(alim.acp)
```



Les points ne sont pas concentrées, mais au contraire dispersés. Cela signifie qu'il existe bien des oppositions entre les individus. On peut de plus discerner à l'oeil nu quelques semblants de clusters : un en bas à droite, un à gauche, et un en haut à droite. Cela signifie qu'il est possible de réaliser des groupes d'indivus semblables. Ainsi, il existe bien des ressemblances et des oppositions entre les individus. Par exemple, les individus 64 et 66 se ressemblent, mais les individus 98 et 33 s'opposent. Attention, comme dit précédemment, il faut garder en tête que notre ACP nous a conduit à une perte non négligeable d'informations. Nous pourrons ainsi affirmer ces ressemblances/oppositions uniquement sur les 35.3% de l'information que nous possédons, mais les 64.7% peuvent alterner ces ressemblances/oppositions.

Etablissement d'une typologie

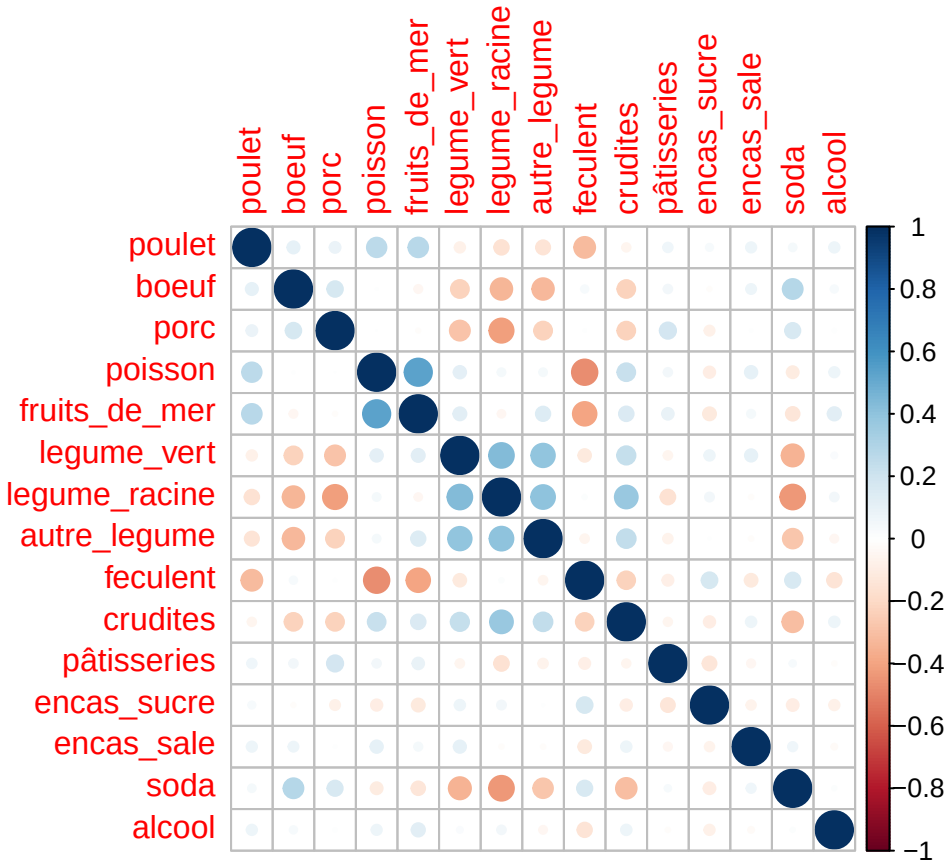
A présent, on ne va plus s'intéresser aux individus, mais aux variables. Afin de mieux comprendre le sens des variables, nous allons les catégoriser. On distingue deux types de variables :

- 15 variables quantitatives représentant, par catégorie d'aliment, la masse (en kg) consommée en moyenne par mois : poulet, poisson, soda, féculent…
- 3 variables qualitatives ordonnées taux fer, taux vitamines, cholesterol relatives à des données biologiques

Corrélation entre les aliments

Nous allons maintenant s'intéresser aux aliments (définis dans la typologie), et plus spécifiquement d'éventuels liens de corrélation entre eux. Pour ce faire, nous allons construire une matrice de corrélation.

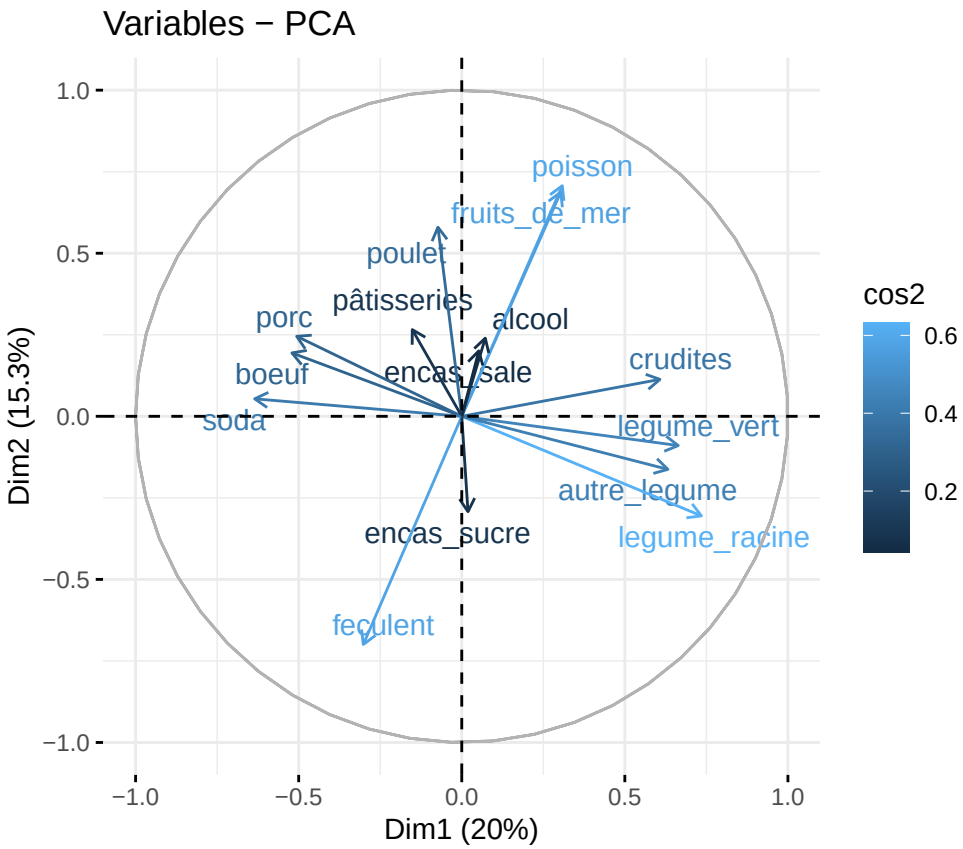
```
corrplot( cor( alim[1:15] ) )
```



Nous pouvons constater que certains aliments sont corrélés entre eux, comme les fruits de mer et le poisson, ou encore tous les types de légumes entre eux. On peut facilement synthétisés ces variables fortement corrélées entre elles en une seule variable. Les fruits de mer et le poisson peuvent être synthétisés en aliments marins. Ou encore, les légumes verts, les légumes racines, les autres légumes et les crudités peuvent tout simplement être synthétisés en légumes.

Nous pouvons aussi retrouver ces résultats sur la projection des variables sur les deux permiers axes factoriels.

```
fviz_pca_var(alim.acp, col.var = "cos2", repel=TRUE)
```

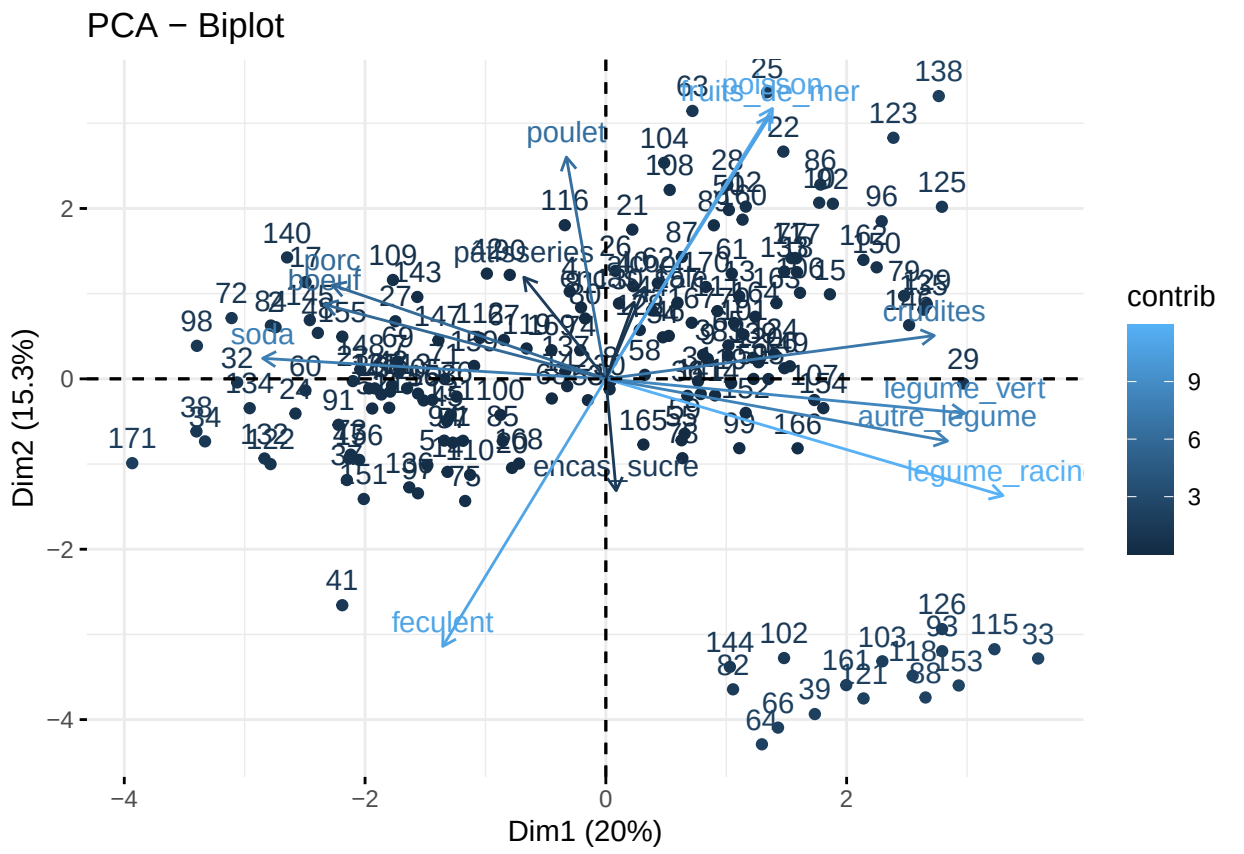


Nous retrouvons en effet deux groupes de variables : les aliments marins et les légumes.

Synthèse des ressemblances et corrélations

Afin de tout synthétiser sur un même schéma, on peut superposer la projection des individus et des variables dans le plan factoriel.

```
fviz_pca_biplot(alim.acp,col.var = "contrib",col.ind = "contrib" )
```



Nous retrouvons tout ce qui a été dit précemment : les ressemblances/oppositions entre les individus et la corrélation entre plusieurs variables, permettant donc de synthétiser de nouvelles variables plus générales. Avec cette représentation, nous pouvons aussi comprendre ce qui contribuent à la ressemblance ou à l'opposition des individus. Par exemple, les repas 22 et 25 se rapprochent car ce sont deux repas marins, mais les repas 33 et 140 s'opposent, car l'un semble être un repas végétarien, alors que l'autre semble être un repas à base de viande.

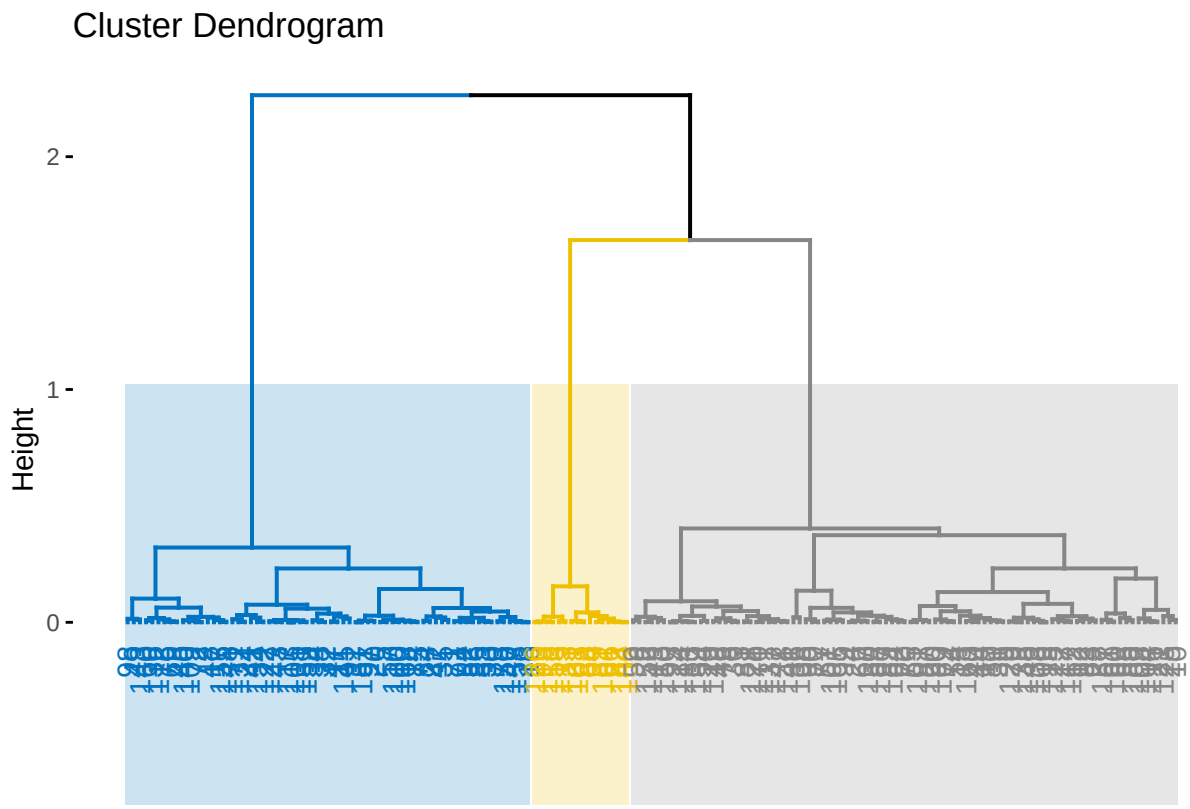
CAH

A présent, nous allons repérer les différents types de comportements alimenataires en identifiant des clusters.

Choix du nombre de clusters

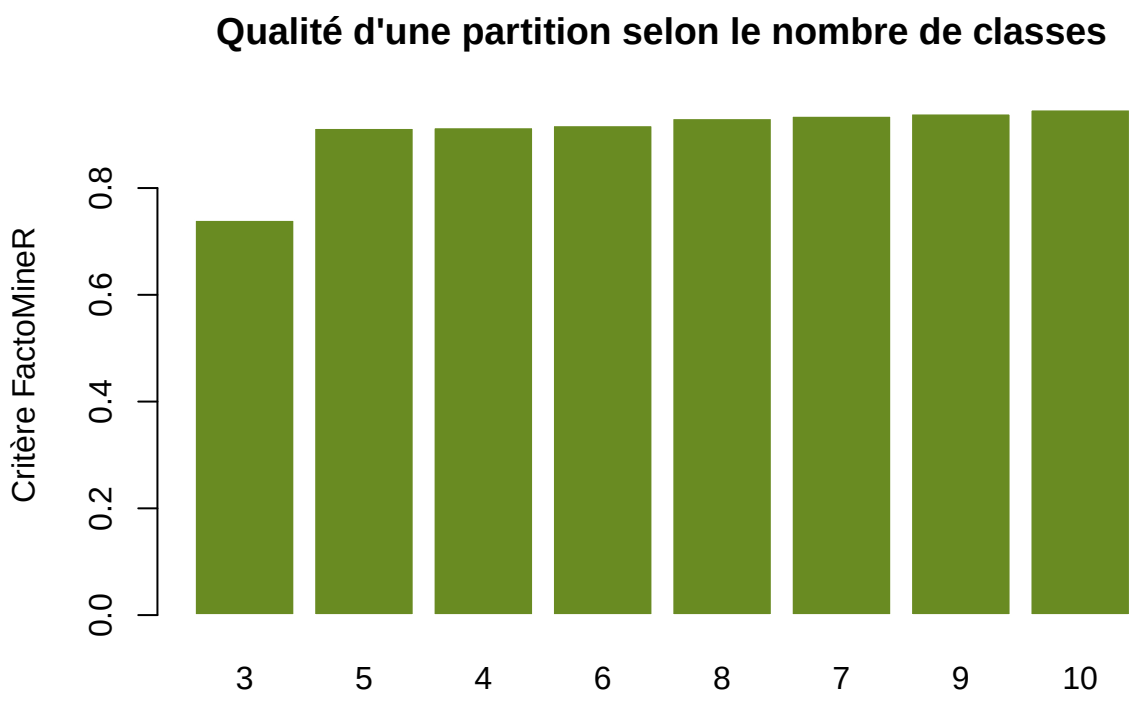
Nous pouvons essayer dans un premier temps d'afficher un dendrogramme faisant un regroupement hiéarchique en déterminant pour quel nombre de clusters le gain d'inertie est maximum.

```
hcpc <- HCPC(alim.acp, graph = FALSE)
fviz_dend(hcpc,
          palette = "jco",
          rect = TRUE, rect_fill = TRUE,
          rect_border = "jco"
          )
```



Bien qu'il soit difficile à lire, nous observons que le nombre de clusters suggérés est 3. Nous allons vérifier que 3 clusters minimise bien un critère de FactoMineR.

```
crit.tri= sort(hcpc$call$t$quot)
coup = order(hcpc$call$t$quot)+2
barplot(crit.tri, names.arg=coup, col="olivedrab4", border="white", ylab= "Critère FactoMineR", main="Q
```

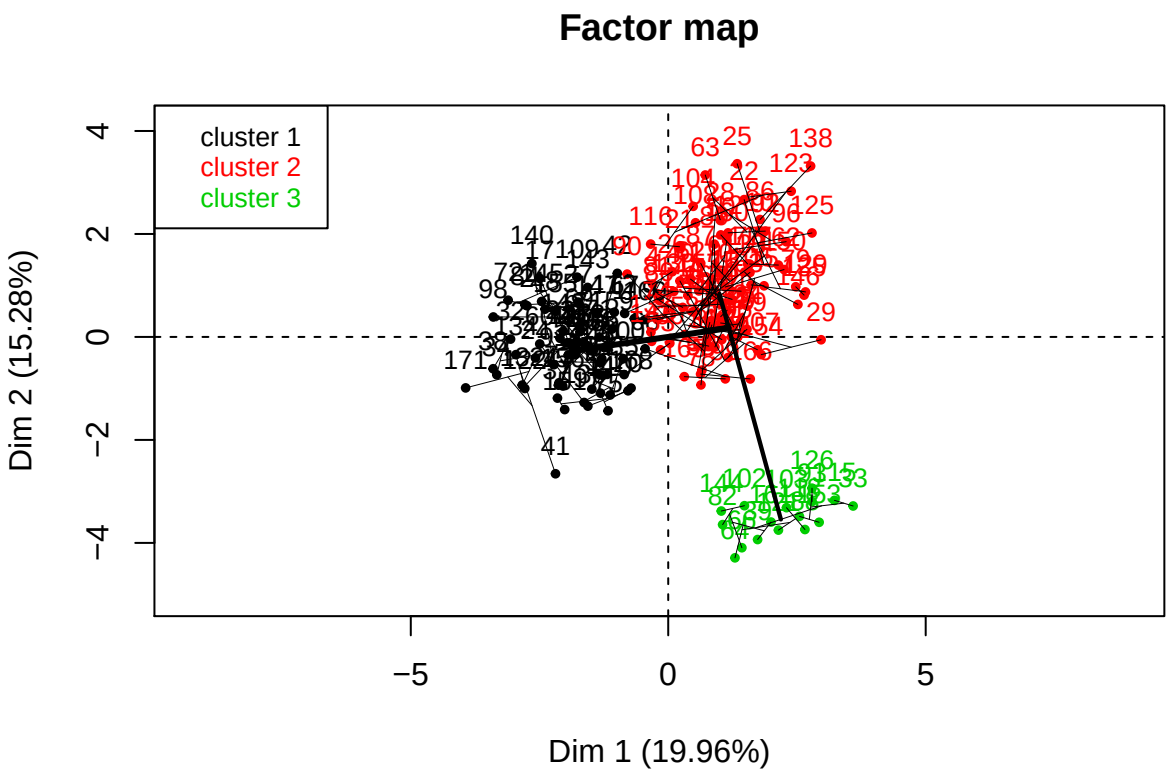


3 minimise bien ce critère, donc il est pertinent de distinguer 3 clusters.

Affichage des clusters identifiés

Voici les clusters que nous pouvons identifiés dans le premier plan factoriel.

```
plot(hcpc, choice = "map")
```



Nous pouvons nous apercevoir, comme suggéré précédemment, de la localisation des clusters : un en bas à droite, un à gauche, et un en haut à droite. Ils sont clairement distincts : le premier plan factoriel est donc suffisant pour les représenter.

Différents types de comportements alimentaires

On peut interprêter ces 3 clusters de la manière suivante : il existe 3 types de comportement alimentaire. Nous allons donc déterminer ce qui les caractrisent.

```
hcpc$desc.var$quanti
```

```
## $`1`
##                   v.test Mean in category Overall mean sd in category Overall sd
## soda            7.405020        14.909265    10.160936       7.644123   6.793207
## feculent        5.362124        18.996912    14.787778       6.936773   8.316032
## porc            5.107019         7.670882     5.470526       5.560782   4.564422
## boeuf           5.056651         6.373235     4.609415       3.962303   3.695316
## poisson        -4.554618         2.722647     4.191404       1.980282   3.416317
## fruits_de_mer -4.631161         1.851176     3.112924       1.217329   2.886310
## crudites       -5.924758         3.874265     6.832047       2.846946   5.288784
## autre_legume   -6.812169         3.821912     8.092807       3.280866   6.641922
## legume_vert    -8.054799         7.502500    12.772982       5.289519   6.931959
## legume_racine -8.080049         9.248971    16.073918       6.510729   8.948405
##                     p.value
## soda           1.311308e-13
## feculent       8.224886e-08
## porc           3.272812e-07
```

```
## boeuf         4.266836e-07
## poisson       5.248091e-06
## fruits_de_mer 3.636217e-06
## crudites      3.127577e-09
## autre_legume  9.613777e-12
## legume_vert   7.960980e-16
## legume_racine 6.474062e-16
##
## $`2`
##                   v.test Mean in category Overall mean sd in category Overall sd
## poisson         7.452843         6.110230     4.191404       3.357976   3.416317
## fruits_de_mer   7.165722         4.671609     3.112924       3.124114   2.886310
## legume_vert     5.370127        15.578391    12.772982       5.188034   6.931959
## crudites        4.769907         8.733218     6.832047       5.377356   5.288784
## autre_legume    3.968394        10.079195     8.092807       6.655896   6.641922
## poulet          3.797884         5.453448     4.538947       3.042139   3.195118
## legume_racine   3.601449        18.502644    16.073918       5.624588   8.948405
## alcool          2.593826         9.488161     8.388421       5.532673   5.625921
## porc           -2.074921         4.756782     5.470526       2.642525   4.564422
## soda           -5.024061         7.588851    10.160936       3.616746   6.793207
## feculent       -8.989674         9.153793    14.787778       4.347940   8.316032
##                     p.value
## poisson       9.135001e-14
## fruits_de_mer 7.737777e-13
## legume_vert   7.868132e-08
## crudites      1.843111e-06
## autre_legume  7.235851e-05
## poulet        1.459365e-04
## legume_racine 3.164485e-04
## alcool        9.491462e-03
## porc          3.799387e-02
## soda          5.059018e-07
## feculent      2.479652e-19
##
## $`3`
##                   v.test Mean in category Overall mean sd in category Overall sd
## legume_racine   7.396486        31.873750    16.073918       4.892886   8.948405
## feculent        6.420597        27.533750    14.787778       4.632291   8.316032
## autre_legume    4.635867        15.443125     8.092807       5.642895   6.641922
## legume_vert     4.317909        19.918125    12.772982       6.032176   6.931959
## encas_sucre     2.604998         7.064375     4.836608       5.111349   3.582457
## pâtisseries    -2.964038         2.063125     4.492398       1.702567   3.433295
## soda           -3.819997         3.966250    10.160936       1.888842   6.793207
## fruits_de_mer  -4.517979         0.000000     3.112924       0.000000   2.886310
## porc           -5.020673         0.000000     5.470526       0.000000   4.564422
## poisson        -5.139489         0.000000     4.191404       0.000000   3.416317
## boeuf          -5.225319         0.000000     4.609415       0.000000   3.695316
## poulet         -5.950958         0.000000     4.538947       0.000000   3.195118
##                     p.value
## legume_racine 1.398353e-13
## feculent      1.357411e-10
## autre_legume  3.554448e-06
## legume_vert   1.575142e-05
## encas_sucre   9.187480e-03
```

```
## pâtisseries   3.036310e-03
## soda          1.334534e-04
## fruits_de_mer 6.243270e-06
## porc          5.149085e-07
## poisson       2.754870e-07
## boeuf         1.738552e-07
## poulet        2.665778e-09
```

Les paragons (i.e. les éléments les plus proches du centre de gravité de chacune des classes) sont les premiers de chaque liste. Cela signifie que l'on peut caractériser les différents types de comportements alimentaires de la manière suivante :

- comportement alimentaire 1 : plats composés de soda, féculent et viande
- comportement alimentaire 2 : plats composés d'aliments marins et de légumes
- comportement alimentaire 3 : plats composés de légumes.

Nous pouvons donc résumer les différents types de comportement alimentaire en 3 plats : les plats avec viande, les plats avec poisson et les plats végétariens.

ADisc

A présent, nous allons chercher à comprendre quels sont les aliments qui permettent de mieux représenter les différentes données biologiques.

Masse de chaque aliment en fonction du taux des différentes données biologiques

Pour commencer, nous allons représenter la masse des aliments composant les plats en fonction du taux des données biologiques pour avoir une idée globale des variables discriminantes.

```
par(mfrow=c(5,3))
for (i in 1:15) {
boxplot(alim[,i]~alim$cholesterol, xlab = "", ylab=colnames(alim)[i])
}
```

En fonction du taux de cholestérol

poulet
0 2 4 6 8 12
bas élevé normal
boeuf
0 5 10 15
bas élevé normal
porc
0 5 10 15 20
bas élevé normal
poisson
0 5 10 15
bas élevé normal
fruits_de_mer
0 2 4 6 8 12
bas élevé normal
legume_vert
0 5 10 20 30
bas élevé normal
legume_racine
0 10 20 30 40
bas élevé normal
autre_legume
0 5 15 25
bas élevé normal
feculent
5 15 25 35
bas élevé normal
crudites
0 5 10 15 20
bas élevé normal
pâtisseries
0 5 10 15 20
bas élevé normal
encas_sucre
0 5 10 15 20
bas élevé normal
encas_sale
0 5 10 15
bas élevé normal
soda
0 10 20 30
bas élevé normal
alcool
0 5 10 20
bas élevé normal

Nous pouvons remarquer à première vue que la masse de fruits de mer, de légumes, de crudités et des sodas permettent de déterminer le taux de cholestérol.

```
par(mfrow=c(5,3))
for (i in 1:15) {
boxplot(alim[,i]~alim$`taux_fer`, xlab = "", ylab=colnames(alim)[i])
}
```

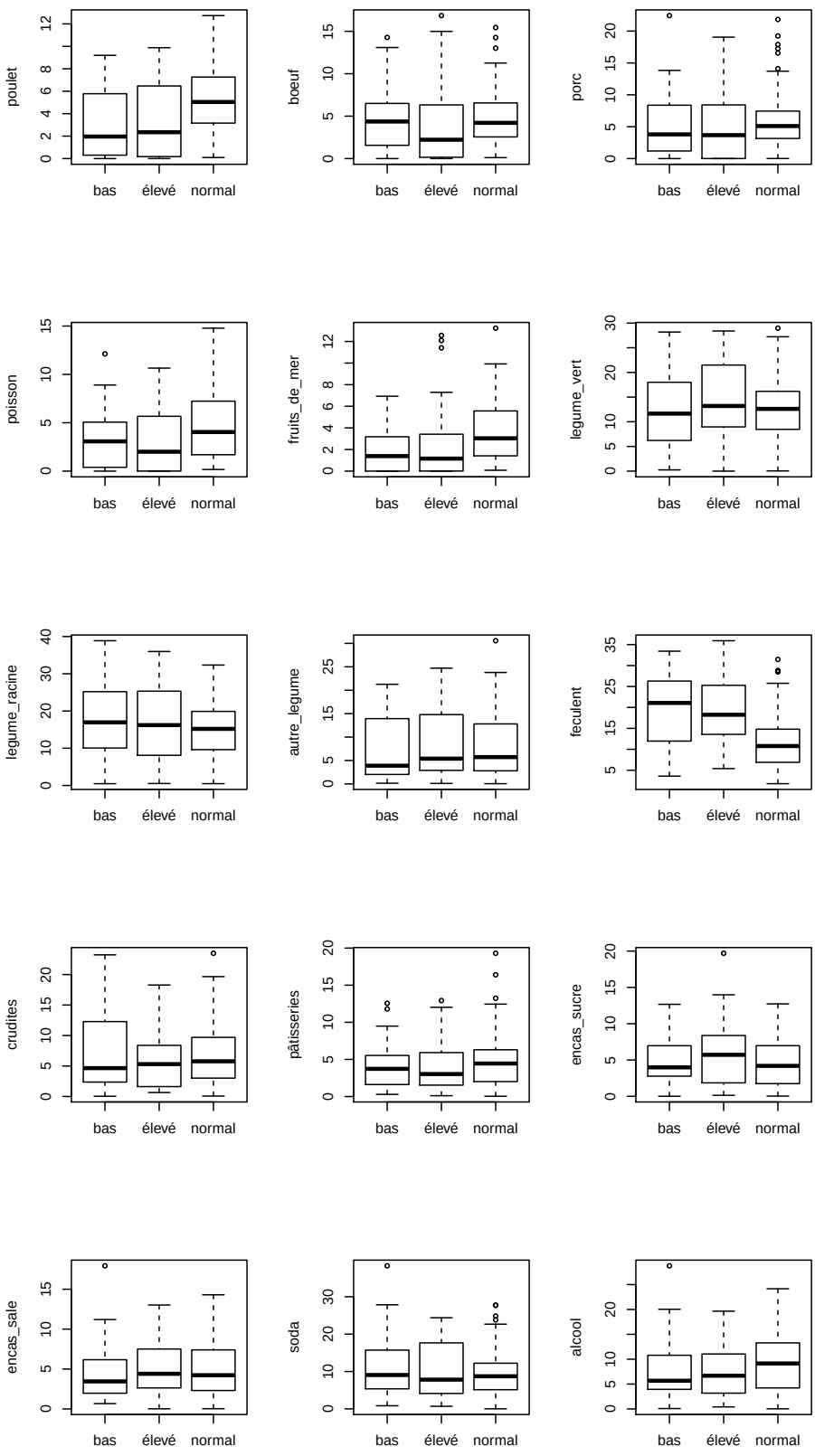
En fonction du taux de fer

poulet
boeuf
porc
poisson
fruits_de_mer
legume_vert
legume_racine
autre_legume
feculent
crudites
pâtisseries
encas_sucre
encas_sale
soda
alcool
bas
élevé
normal

Nous pouvons remarquer à première vue que la masse des fruits de mer, des légumes, des encas sucrés et des sodas permettent de déterminer le taux de fer.

```
par(mfrow=c(5,3))
for (i in 1:15) {
boxplot(alim[,i]~alim$`taux_vitamines`, xlab = "", ylab=colnames(alim)[i])
}
```

En fonction du taux de vitamines


poulet
0 2 4 6 8 12
bas élevé normal
boeuf
0 5 10 15
bas élevé normal
porc
0 5 10 15 20
bas élevé normal
poisson
0 5 10 15
bas élevé normal
fruits_de_mer
0 2 4 6 8 12
bas élevé normal
legume_vert
0 5 10 20 30
bas élevé normal
legume_racine
0 10 20 30 40
bas élevé normal
autre_legume
0 5 15 25
bas élevé normal
feculent
5 15 25 35
bas élevé normal
crudites
0 5 10 15 20
bas élevé normal
pâtisseries
0 5 10 15 20
bas élevé normal
encas_sucre
0 5 10 15 20
bas élevé normal
encas_sale
0 5 10 15
bas élevé normal
soda
0 10 20 30
bas élevé normal
alcool
0 5 10 20
bas élevé normal

Nous pouvons remarquer à première vue que la masse du poulet, du boeuf, du poisson et des féculents permettent de déterminer le taux de vitamines.

Sélection des aliments dicriminants

Afin de déterminer de façon plus précise quels sont les aliments qui permettent de caractériser au mieux le taux des données biologiques dans les plats, nous allons effectuer une sélection progressive en utilisant le lambda de Wilks.

```
greedy.wilks(cholesterol~., data=alim[-c(16,17)], niveau = 0.1)
```

Pour le taux de cholestérol

```
## Formula containing included variables:
##
## cholesterol ~ legume_vert + soda + pâtisseries + alcool + feculent +
##     legume_racine
## <environment: 0x55bdbd616b78>
##
##
## Values calculated in each step of the selection procedure:
##
##            vars Wilks.lambda F.statistics.overall p.value.overall
## 1   legume_vert    0.8298552            17.222478    1.571068e-07
## 2          soda    0.7762368            11.274036    1.361292e-08
## 3   pâtisseries    0.7325053             9.318589    1.882266e-09
## 4        alcool    0.6930778             8.298778    2.924076e-10
## 5      feculent    0.6664440             7.378343    1.438489e-10
## 6 legume_racine    0.6472538             6.600845    1.367339e-10
##   F.statistics.diff p.value.diff
## 1         17.222478 1.571068e-07
## 2          5.767751 3.778775e-03
## 3          4.955200 8.117450e-03
## 4          4.693222 1.040643e-02
## 5          3.277059 4.020970e-02
## 6          2.416363 9.241695e-02
```

Les aliments permettant de caractériser au mieux le taux de cholestérol sont les légumes verts, les sodas, les pâtisserises, l'alcool, les féculents et les légumes racines.

```
greedy.wilks(taux_fer~., data=alim[-c(17,18)], niveau = 0.1)
```

Pour le taux de fer

```
## Formula containing included variables:
##
## taux_fer ~ legume_vert + fruits_de_mer + encas_sucre + porc
## <environment: 0x55bdbc98a4d0>
##
##
## Values calculated in each step of the selection procedure:
##
##            vars Wilks.lambda F.statistics.overall p.value.overall
## 1   legume_vert    0.8812868            11.315169    2.453862e-05
```

```
## 2 fruits_de_mer    0.8034707             9.653986    2.126265e-07
## 3   encas_sucre    0.7638273             7.979134    4.704904e-08
## 4          porc    0.7389294             6.736877    3.472071e-08
##   F.statistics.diff p.value.diff
## 1         11.315169 2.453862e-05
## 2          8.086967 4.434187e-04
## 3          4.307785 1.499041e-02
## 4          2.779807 6.494133e-02
```

Les aliments permettant de caractériser au mieux le taux de fer sont les légumes verts, les encas sucrés et le porc.

```
greedy.wilks(taux_vitamines~., data=alim[-c(16,18)], niveau = 0.1)
```

Pour le taux de vitamines

```
## Formula containing included variables:
##
## taux_vitamines ~ feculent + poulet
## <environment: 0x55bdbcfabd58>
##
##
## Values calculated in each step of the selection procedure:
##
##       vars Wilks.lambda F.statistics.overall p.value.overall F.statistics.diff
## 1 feculent    0.8081674             19.93886    1.698681e-08         19.938861
## 2   poulet    0.7788320             11.11600    1.777658e-08          3.145098
##   p.value.diff
## 1 1.698681e-08
## 2 4.560908e-02
```

Les aliments permettant de caractériser au mieux le taux de vitamines sont les féculents et le poulet.

Nous retrouvons dans les trois cas approximativement les mêmes aliments discriminants intuités par la représentation de la masse des aliments composant les plats en fonction du taux des données biologiques.