

ACP d'un jeu de données très simple

P^rJean R. Lobry

Contents

1	Analyse univariée	1
1.1	l'objet <code>result</code>	1
1.2	l'objet <code>tab</code> et l'objet <code>cent</code>	2
1.3	Analyse univarié	3
1.4	Analyse bivariée	4
2	Une première ACP	5
3	Une deuxième ACP	8
4	Exercice	11
	Références	11

1 Analyse univariée

On ne se lance jamais dans une analyse multivariée avant d'avoir soigneusement épuisé le point de vue univarié. Les données utilisées ici sont disponibles dans le paquet `ade4` de  :

```
library(ade4)
data(deug)
```

Lire la documentation du jeu de données avec `?deug`. Pour information, DEUG était l'acronyme pour « Diplôme d'Études Universitaires Générales » et correspondait aux deux premières années de la licence avant la réforme dite LMD, encore un acronyme pour « Licence Master Doctorat ». Cet objet `deug` est donc une liste à trois composantes :

```
class(deug)
[1] "list"
names(deug)
[1] "tab"      "result"   "cent"
```

1.1 l'objet result

Cet objet contient ce qui intéresse le plus les étudiants : le résultat final global à l'examen. Voyons en quoi il consiste.

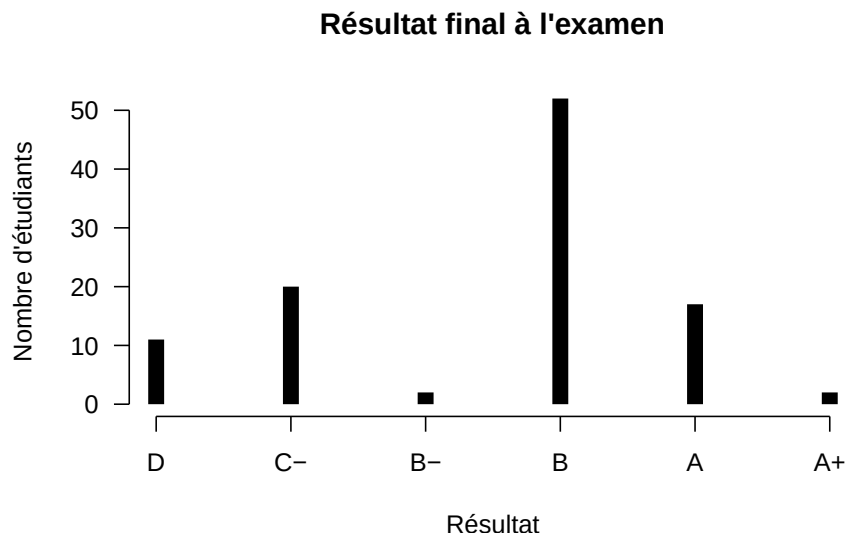
```
class(deug$result)
[1] "factor"
deug$result
[1] C- B B A B C- B B D A B B A C- B B A D B A+ A B B B B B
[27] C- B B B B B B B B C- D D D B B B A B C- B B B B B B D
[53] A+ A A C- B A C- A B B B C- C- D B- C- C- A B D A B B C- C- C-
[79] A C- C- C- B B B A B A A D D B- C- C- D A B B B B B B B
Levels: D A A+ C- B- B
```

C'est donc une variable qualitative non ordonnée. On décide de la transformer en une variable qualitative ordonnée pour mettre un peu d'ordre.

```
deug$result <- ordered(deug$result, levels = c("D","C-","B-","B","A","A+"))
deug$result
[1] C- B B A B C- B B D A B B A C- B B A D B A+ A B B B B B
[27] C- B B B B B B B B C- D D D B B B A B C- B B B B B B D
[53] A+ A A C- B A C- A B B B C- C- D B- C- C- A B D A B B C- C- C-
[79] A C- C- C- B B B A B A A D D B- C- C- D A B B B B B B B
Levels: D < C- < B- < B < A < A+
```

C'est bien plus clair ainsi, les plus mauvais résultats correspondent à la modalité D et les meilleurs à la modalité A+. Regardons comment se répartissent nos 104 étudiants.

```
table(deug$result)
D C- B- B A A+
11 20 2 52 17 2
plot(table(deug$result), lwd = 10, lend = "butt", las = 1,
main = "Résultat final à l'examen", ylab = "Nombre d'étudiants",
xlab = "Résultat")
```



Commentez cette distribution. Nous utiliserons plus tard cette variable qualitative comme variable illustrative pour aider à l'interprétation des plans factoriels.

1.2 l'objet tab et l'objet cent

C'est ici que l'on trouve les résultats des 104 étudiants aux 9 matières :

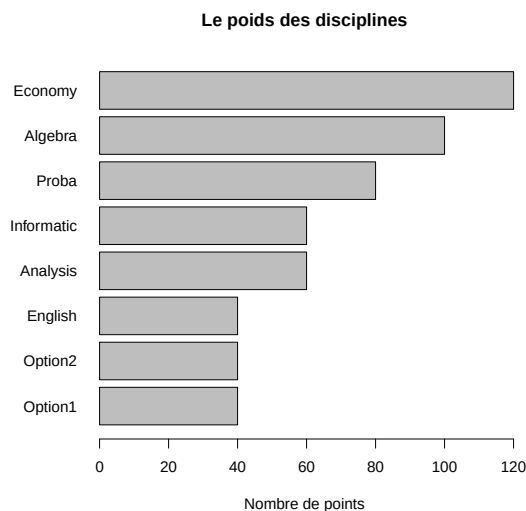
```
class(deug$tab)
[1] "data.frame"
dim(deug$tab)
[1] 104 9
head(deug$tab)
  Algebra Analysis Proba Informatc Economy Option1 Option2 English Sport
1      40      26.0    26      26.0     51.9      17      24      19.0    11.5
2      37      34.5    37       32.0     72.0      24      22      26.0    11.5
3      37      41.0    29       34.5     72.0      24      27      19.6    11.5
4      63      37.5    57       35.5     77.4      23      23      21.0    14.0
5      55      31.5    34       36.0     57.9      19      24      24.0    11.5
6      50      38.0    32       20.0     66.9      20      15      22.2     0.0
```

Une valeur faible correspond à une mauvaise note, une valeur élevée correspond à une bonne note. Mais c'est noté sur combien m'sieur ? Très bonne question, c'est l'objet **cent** qui nous dit combien de points il faut pour avoir la moyenne.

```
deug$cent
  Algebra Analysis    Proba Informatc    Economy    Option1    Option2
English      50         30      40         30         60         20         20
Sport       20         0
```

Notez le cas particulier du sport : à votre avis à quoi correspond la valeur 0 ici ? Représenter le poids des disciplines.

```
coefs <- 2*deug$cent[1:8]
par(mar = c(5,5,4,2)+0.1)
barplot(sort(coefs), horiz=T, las = 1, xlab = "Nombre de points", main = "Le poids des disciplines")
```

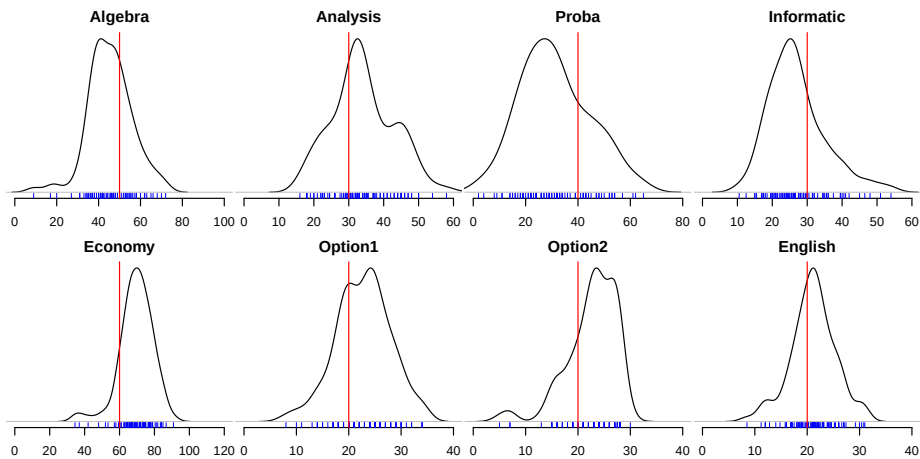


À votre avis dans quel type de filière universitaire étaient inscrits ces étudiants ?

1.3 Analyse univarié

Nous nous intéressons donc ici à la distribution des notes pour chaque discipline.

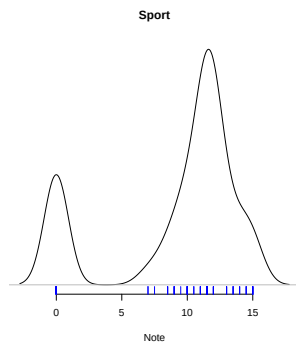
```
par(mfrow = c(2,4), mar = c(2,0,2,0)+0.1)
for(i in 1:8){
  plot(density(deug$tab[,i]), xlim = c(0,2*deug$cent[i]),
       main = colnames(deug$tab)[i], yaxt = "n", bty = "n")
  rug(deug$tab[,i], col = "blue")
  abline(v = deug$cent[i], col = "red")
}
```



Quel est la signification de la ligne verticale rouge ? Quelles sont les matières difficiles ? Si vous n'aviez qu'une discipline à réviser, laquelle choisiriez-vous ?

Le sport est un peu à part ici, commentez la distribution des notes de sport :

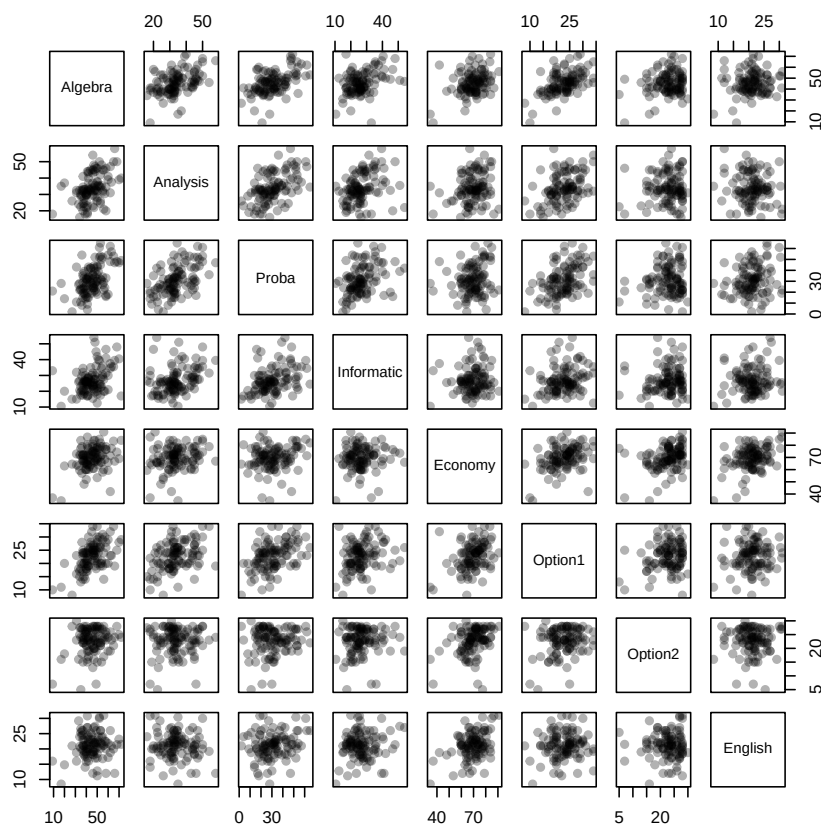
```
x <- deug$tab[,"Sport"]
plot(density(x), xlab = "Note", bty = "n", yaxt = "n",
     ylab = "", main = "Sport")
rug(x, col = "blue")
```



1.4 Analyse bivariée

On croise toutes les notes deux à deux pour se faire une première idée des corrélations entre disciplines.

```
plot(deug$tab[,1:8], pch = 19, col=rgb(0,0,0,0.3))
```

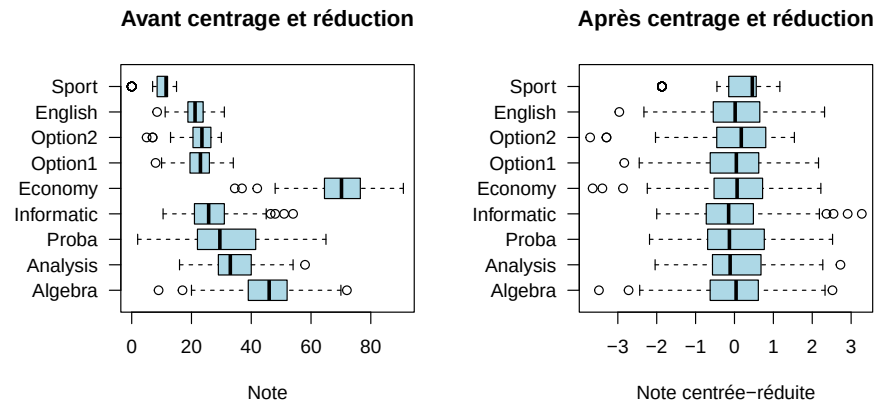


Les corrélations sont elles plutôt positives ou négatives ? À votre avis pourquoi ?

2 Une première ACP

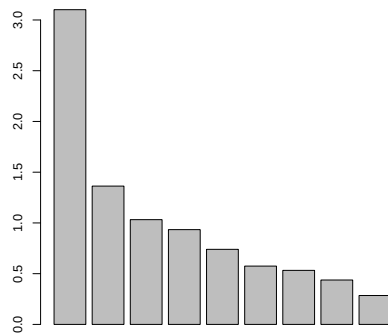
Nous allons commencer par faire une ACP centrée-réduite. Dans ce cas le tableau analysé est centré et réduit par colonne, ce qui correspond à la transformation suivante :

```
par(mfrow=c(1,2), mar = c(5,5,4,2)+0.1)
boxplot(deug$tab, horizontal=T, las = 1, main = "Avant centrage et réduction",
xlab = "Note", col = "lightblue")
boxplot(as.data.frame(scale(deug$tab)), horizontal=T, las = 1,
main = "Après centrage et réduction",
xlab = "Note centrée-réduite", col = "lightblue")
```



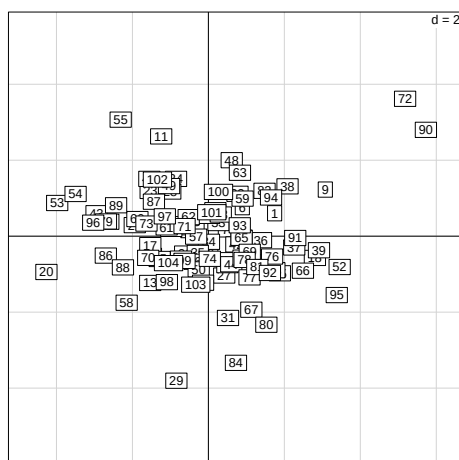
Examinez l'allure du graphe des valeurs propres :

```
acp1 <- dudi.pca(deug$tab, scan = F)
barplot(acp1$eig)
```



Représentez les individus dans le premier plan factoriel :

```
s.label(acp1$li)
```

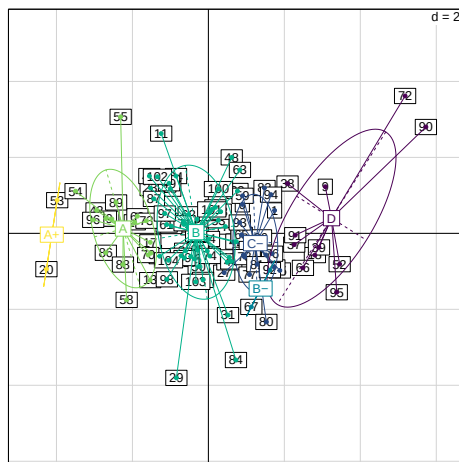


Ce n'est pas évident à interpréter parce que l'on ne connaît pas personnellement les individus. Examinons les résultats de ceux qui ressortent sur le premier axe :

```
deug$result[c(72,90)]
[1] D D
Levels: D < C- < B- < B < A < A+
deug$result[c(20,53,54)]
[1] A+ A+ A
Levels: D < C- < B- < B < A < A+
```

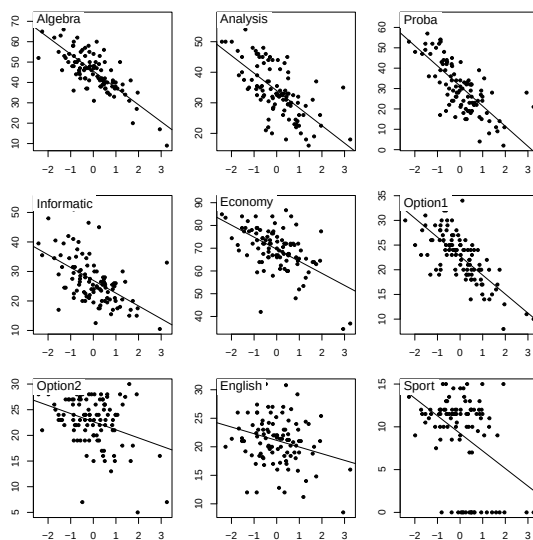
Le premier axe oppose donc les individus qui ont eu un très bon résultat à ceux qui ont eu un très mauvais résultat. On peut généraliser cette approche en utilisant le résultat comme variable illustrative sur le premier plan factoriel :

```
s.label(acp1$li)
s.class(acp1$li,deug$result, add.plot = TRUE, col = hcl.colors(6))
```



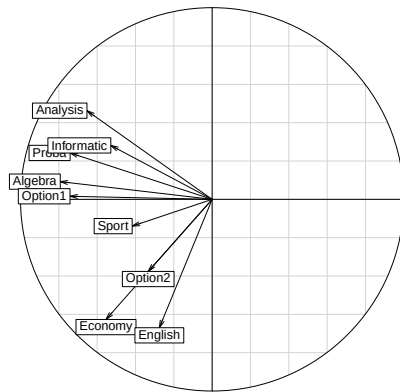
L'interprétation du premier facteur devient alors assez évidente. Une autre aide à l'interprétation consiste à comparer les coordonnées des individus sur le premier axe avec les données originelles :

```
score(acp1)
```



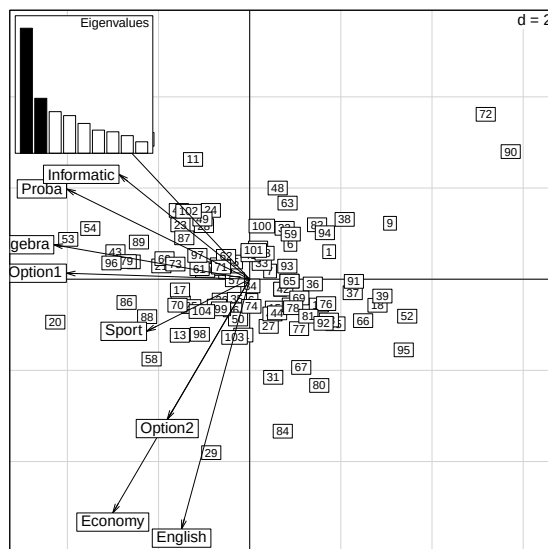
Enfin, le plan des variables donne souvent une aide précieuse à l'interprétation :

```
s.corcircle(acp1$co)
```



Une vue de synthèse est donnée par la fonction `scatter` :

```
scatter(acp1)
```

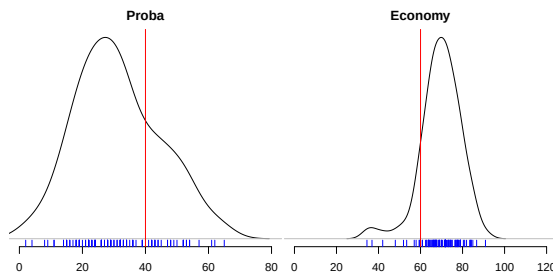


3 Une deuxième ACP

L'ACP centrée-réduite nous a donc montré qu'il y avait un gradient entre les très bons étudiants qui ont de bonnes notes dans toutes les matières et les très mauvais étudiants qui ont de mauvaises notes dans toutes les matières. C'est une parfaite trivialité qui ne nous aide en rien pour préparer l'examen de façon astucieuse. Tout ce que l'ACP centrée-réduite nous apprend ici c'est qu'il y a intérêt à bosser pour avoir ses examens, si c'est pour apprendre ça, on ne voit pas trop l'intérêt de la méthode.

Le gros problème de l'ACP centrée-réduite est qu'elle a complètement gommé un point essentiel qui est celui d'avoir la moyenne ou pas dans une matière. Nous avons centré les notes par rapport à la moyenne des étudiants dans une matière, et ce n'est pas très pertinent :

```
par(mfrow = c(1,2), mar = c(2,0,2,0)+0.1)
for(i in c(3,5)){
  plot(density(deug$tab[,i]), xlim = c(0,2*deug$cent[i]),
       main = colnames(deug$tab)[i], yaxt = "n", bty = "n")
  rug(deug$tab[,i], col = "blue")
  abline(v = deug$cent[i], col = "red")
}
```



Que préférez vous ? Avoir 39/80 en Probabilité ?

```
(ncr <- (39 - mean(deug$tab$Proba))/sd(deug$tab$Proba))
[1] 0.5798552
```

vous êtes à +0.6 écart-types au dessus de la moyenne de la promo, super, sauf que vous avez planté votre examen puisque vous n'avez pas la moyenne. Ou bien préférez-vous avoir 61/120 en économie ?

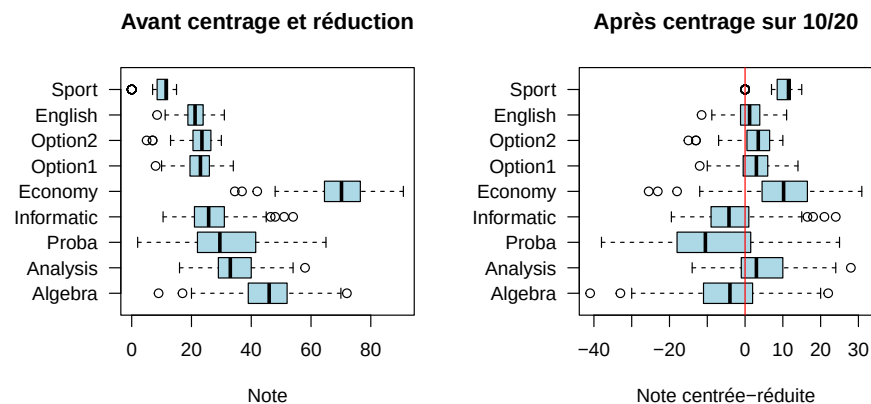
```
(ncr <- (61 - mean(deug$tab$Economy))/sd(deug$tab$Economy))
[1] -0.889822
```

vous êtes à -0.9 écart-type au dessous de la moyenne de la promo, ce n'est pas terrible, il n'empêche que vous avez la moyenne à votre examen. Le centrage fait autour de la moyenne de la promotion dans l'ACP centrée-réduite n'est pas du tout ce que l'on a envie de faire. Mais il y a bien pire, c'est l'opération de réduction :

```
apply(deug$tab,2,sd)
      Algebra      Analysis      Proba Informatic      Economy      Option1      Option2
10.489722    8.811802    13.381987    8.247383    9.609744    5.204087    4.759564
      English      Sport
4.261956    4.944313
```

Un écart-type en probabilités vaut 13.3 points au final contre 4.2 points en anglais. Mais ce qui nous intéresse ce sont les points au dessus de la moyenne dans l'absolu, surtout s'il existe un mécanisme de compensation entre les matières. Voici donc la transformation que l'on veut faire :

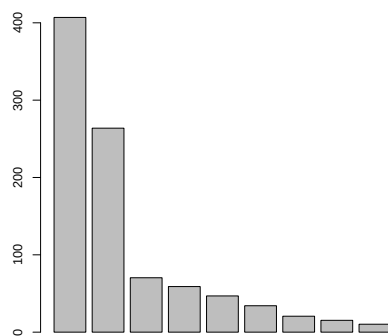
```
par(mfrow=c(1,2), mar = c(5,5,4,2)+0.1)
boxplot(deug$tab, horizontal=T, las = 1, main = "Avant centrage et réduction",
       xlab = "Note", col = "lightblue")
boxplot(as.data.frame(scale(deug$tab, center = deug$cent, scale = FALSE)), horizontal=T, las = 1,
       main = "Après centrage sur 10/20",
       xlab = "Note centrée-réduite", col = "lightblue")
abline(v=0,col="red")
```



Le graphique de droite est très intéressant, c'est lui qui nous permet d'optimiser l'investissement à effectuer. Supposons que je sois un étudiant très moyen, en travaillant normalement je peux espérer avoir la note médiane dans chaque matière. En anglais et dans les deux options je peux espérer grappiller quelques points, mettons +8 en tout, en économie +10, en analyse +3 donc je peux espérer +21 de ce côté. En informatique je vais perdre 4 points, en probabilité 10 points et en algèbre 4, donc -18 points. Je pense que si je travaille une matière à fond, je peux passer de la médiane au premier quartile, quelle matière choisir ? Les probabilités sans hésiter, je peux espérer gagner ainsi 12 points, bien plus qu'ailleurs.

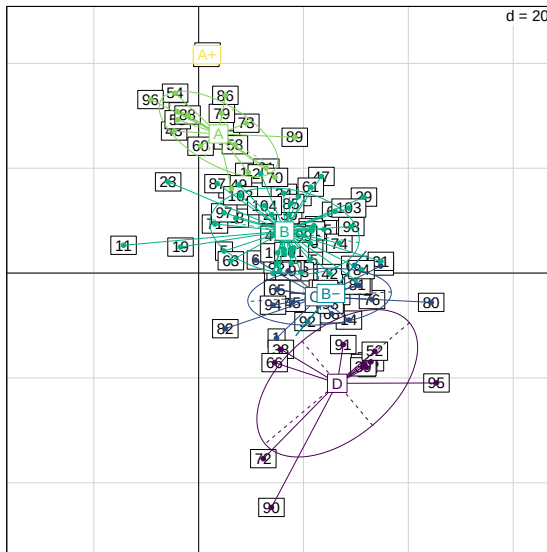
Commenter le graphe des valeurs propres :

```
acp2 <- dudi.pca(deug$stab, center = deug$cent, scale = FALSE, scann=F)
barplot(acp2$eig)
```



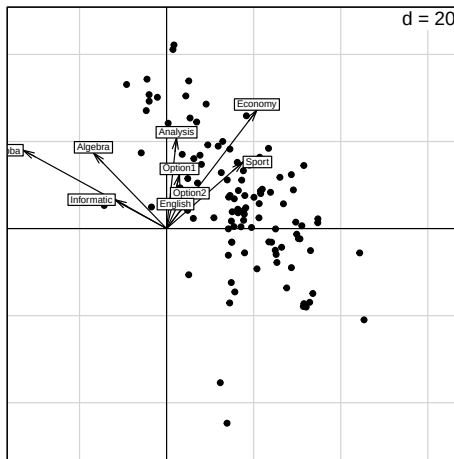
Représenter les individus dans le premier plan factoriel en utilisant le résultat comme variable illustrative :

```
s.label(acp2$li)
s.class(acp2$li, deug$result, add.plot=T, col=hcl.colors(6))
```

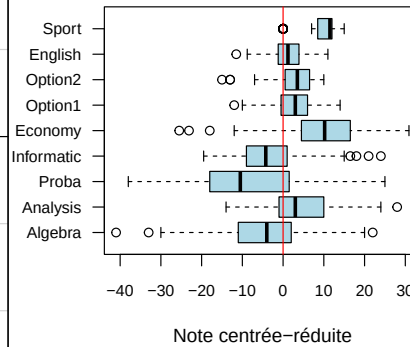


On retrouve sur le deuxième axe le gradient de résultat final, mais que signifie le premier axe ? La vue de synthèse devrait aider :

```
par(mfrow=c(1,2))
scatter(acp2, posieig = "none", clab.col = 0.5, clab.row=0)
NULL
boxplot(as.data.frame(scale(deug$tab, center = deug$cent, scale = FALSE)), horizontal=T, las = 1,
main = "Après centrage sur 10/20",
xlab = "Note centrée-réduite", col = "lightblue", cex.axis = 0.8)
abline(v=0,col="red")
```



Après centrage sur 10/20



4 Exercice

Le jeu de données `seconde` contient des variables de nature très similaires. Analysez ce jeu de données.

```
data(seconde)
names(seconde)
[1] "HGE0" "FRAN" "PHYS" "MATH" "BIOL" "ECON" "ANGL" "ESPA"
```