Niveau scolaire d'une classe de 6e

Taniel Rémi - GIS2A4

16/09/2019

Contents

Introduction

Dans cette étude, nous nous intéresserons aux résultats d'une classe de 6e composée de 27 élèves sur les 14 disciplines suivantes :

- ORTH: Orthographe
- GRAM: Grammaire
- EXPR: Expression écrite
- RECI: Récitation
- MATH: Mathématiques
- ANGL: Anglais
- HIST: Histoire
- BIOL: Biologie
- EDMU: Education musicale
- ARTS: Arts plastiques
- TECH: Technologie
- EPS: Education Physique et Sportive
- GEO: Géographie
- EXPO: Exposé

Partie 1 : Importation et visualisation des données

Nos données sont sauvegardées au format CSV dans le fichiers notes, nous commencons donc par importer les données dans la variable `data` grâce à la fonction suivante :

```
data <- read.table("/home/remi/Documents/Cours/AD/data/notes.csv", sep = ";", dec = ",", header = TRUE)
```

On visualise ensuite les types de nos données et la forme de celles-ci grâce à la fonction `str` appliquée à notre variable `data` :

```
str(data)
```

```
## 'data.frame':    27 obs. of  15 variables:
##  $ eleves: Factor w/ 27 levels "EL01","EL02",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ ORTH  : num  13 6.5 14 13 15 5 12 8.5 15.5 20 ...
##  $ GRAM  : num  10 8 6.5 7.5 7.5 8 6.5 2.5 7.5 14.5 ...
##  $ EXPR  : num  2 8.5 8 9 10 5.5 9 9 12.5 16.5 ...
##  $ RECI  : num  4 14 5 5 14 6.5 16 13 16 10 ...
##  $ MATH  : num  9 13 11 10 12 16 18 12 15 18 ...
##  $ ANGL  : num  9 7 8 10.5 11 12 13.5 9.5 13 16.5 ...
##  $ HIST  : num  8 11 9.5 10 9 9 9 12 12 15 ...
##  $ BIOL  : num  7 8.5 8 16 11 7 10 13.5 13.5 10.5 ...
##  $ EDMU  : num  7.5 16 18.5 16 16.5 13.5 15 16.5 17 18 ...
##  $ ARTS  : num  1.5 4 9.5 11.5 13.5 5 11 8 14 13.5 ...
##  $ TECH  : int  14 18 14 0 16 16 16 13 15 12 ...
##  $ EPS   : num  10 18 16.5 11.5 13 12.5 13.5 12 16 14.5 ...
##  $ GEO   : num  10.5 16 14 15 15.5 ...
##  $ EXPO  : num  13 15 13 18 17 17.5 18 17 17 15 ...
```

Notre jeu de données contient 27 observations, représentant nos 27 élèves, pour 15 variables, représentant les matières. On remarque qu'à part la première colonne qui est le numéro de l'élève, toutes les autres variables sont quantitatives, et représentent la moyenne dans chaque matière pour un élève donné.

On se décide alors de visualiser les 6 premières lignes de nos données avec la fonction `head` ;

```
head(data)
```

```
##   eleves ORTH GRAM EXPR RECI MATH ANGL HIST BIOL EDMU ARTS TECH  EPS   GEO
## 1   EL01 13.0 10.0  2.0  4.0    9  9.0  8.0  7.0  7.5  1.5   14 10.0 10.50
## 2   EL02  6.5  8.0  8.5 14.0   13  7.0 11.0  8.5 16.0  4.0   18 18.0 16.00
## 3   EL03 14.0  6.5  8.0  5.0   11  8.0  9.5  8.0 18.5  9.5   14 16.5 14.00
## 4   EL04 13.0  7.5  9.0  5.0   10 10.5 10.0 16.0 16.0 11.5    0 11.5 15.00
## 5   EL05 15.0  7.5 10.0 14.0   12 11.0  9.0 11.0 16.5 13.5   16 13.0 15.51
## 6   EL06  5.0  8.0  5.5  6.5   16 12.0  9.0  7.0 13.5  5.0   16 12.5 13.00
##   EXPO
## 1 13.0
## 2 15.0
## 3 13.0
## 4 18.0
## 5 17.0
## 6 17.5
```

On remarque que nous devons enlever la colonne `eleves` avant de lancer les calculs et faire notre analyse, on décide donc de formater nos données afin de donner un identifiant aux différentes lignes de nos données, dans notre cas, ce sera la variable `eleves` :

```
rownames(data) <- data$eleves
data <- data[,-1]
head(data)
```

```
##      ORTH GRAM EXPR RECI MATH ANGL HIST BIOL EDMU ARTS TECH  EPS   GEO
## EL01 13.0 10.0  2.0  4.0    9  9.0  8.0  7.0  7.5  1.5   14 10.0 10.50
## EL02  6.5  8.0  8.5 14.0   13  7.0 11.0  8.5 16.0  4.0   18 18.0 16.00
## EL03 14.0  6.5  8.0  5.0   11  8.0  9.5  8.0 18.5  9.5   14 16.5 14.00
## EL04 13.0  7.5  9.0  5.0   10 10.5 10.0 16.0 16.0 11.5    0 11.5 15.00
## EL05 15.0  7.5 10.0 14.0   12 11.0  9.0 11.0 16.5 13.5   16 13.0 15.51
## EL06  5.0  8.0  5.5  6.5   16 12.0  9.0  7.0 13.5  5.0   16 12.5 13.00
##      EXPO
## EL01 13.0
## EL02 15.0
## EL03 13.0
## EL04 18.0
## EL05 17.0
## EL06 17.5
```

Partie 2 : Mise en oeuvre de l'ACP

Pour réaliser l'ACP, nous allons avoir besoin des packages suivants :

```
library(FactoMineR)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
```

```
##     format.pval, units
```

Puis on range les résultats de l'ACP (valeurs propres, coordonnées, contribution) dans la variable `data.pca`, dans notre cas, nous ne gardons que les 6 premieres dimensions pour notre analyse et on souhaite que les graphiques ne soient pas générés lors de l'appel de la fonction :

```
data.pca <- FactoMineR::PCA(data, scale.unit = TRUE, ncp = 5, graph = FALSE)
```

Les résultats seront alors stockés dans les variables suivantes :

- `$eig` pour les informations par rapport aux valeurs propres relatives à chaque dimension
- `$var` pour les résultats concernant les différentes variables, pour notre jeu de données, les différentes matières
- `$ind` pour les informations par rapport aux différents individus, élèves dans notre cas

Partie 3 : Analyse des résultats

Question 2 : Tableau des statistiques sommaires

La tableau des statistiques sommaires est obtenu par la fonction `summary` :

```
summary(data)
```

```
##       ORTH            GRAM            EXPR            RECI      
##  Min.   : 0.00   Min.   : 2.000   Min.   : 2.000   Min.   : 4.00  
##  1st Qu.: 5.75   1st Qu.: 6.500   1st Qu.: 7.500   1st Qu.: 6.50  
##  Median :11.00   Median : 7.500   Median : 9.000   Median :10.00  
##  Mean   :10.02   Mean   : 7.556   Mean   : 8.889   Mean   :10.07  
##  3rd Qu.:14.00   3rd Qu.: 8.500   3rd Qu.:10.500   3rd Qu.:12.50  
##  Max.   :20.00   Max.   :14.500   Max.   :16.500   Max.   :16.00  
##       MATH            ANGL            HIST            BIOL      
##  Min.   : 8.00   Min.   : 3.00   Min.   : 5.000   Min.   : 2.000  
##  1st Qu.:10.00   1st Qu.: 8.75   1st Qu.: 7.750   1st Qu.: 7.000  
##  Median :12.00   Median :11.00   Median : 9.000   Median :10.000  
##  Mean   :12.57   Mean   :10.93   Mean   : 9.204   Mean   : 9.593  
##  3rd Qu.:14.50   3rd Qu.:12.75   3rd Qu.:11.000   3rd Qu.:12.000  
##  Max.   :18.00   Max.   :17.00   Max.   :15.000   Max.   :17.000  
##       EDMU            ARTS            TECH            EPS       
##  Min.   : 7.00   Min.   : 1.500   Min.   : 0.00   Min.   : 5.00  
##  1st Qu.:13.75   1st Qu.: 6.750   1st Qu.:11.50   1st Qu.:11.00  
##  Median :16.00   Median : 9.000   Median :14.00   Median :13.50  
##  Mean   :15.20   Mean   : 9.019   Mean   :12.63   Mean   :13.07  
##  3rd Qu.:17.50   3rd Qu.:12.500   3rd Qu.:16.00   3rd Qu.:15.25  
##  Max.   :19.00   Max.   :14.500   Max.   :18.00   Max.   :18.50  
##       GEO             EXPO       
##  Min.   : 0.00   Min.   : 0.00  
##  1st Qu.:13.00   1st Qu.:13.50  
##  Median :14.00   Median :15.00  
##  Mean   :13.80   Mean   :14.48  
##  3rd Qu.:15.25   3rd Qu.:17.00  
##  Max.   :18.00   Max.   :18.00
```

Grâce à ce tableau, nous pouvons déduire les renseignements suivants :

- En grammaire (variable GRAM) 50% des élèves ont en dessous de 7,5, tandis que 50% des élèves ont plus de 15 de moyenne en exposé (variable EXPO).

- En orthographe (ORTH), les notes sont très hétérogènes (minimum 0 et maximum 20 de moyenne), comme en technologie (TECH) ou en géographie (GEO).
- Généralement la moyenne dans chacunes des 14 matières est supérieure à 10, on peut donc déduire que la classe a plutôt un bon niveau.

Question 3 : Coefficient de corrélation

Pour obtenir les différents coefficients de corrélation entre les différentes variables de notre jeu de données, on utilise la fonction `rcorr` se trouvant dans la librairie `Hmisc` :

```
Hmisc::rcorr(as.matrix(data))[1]$r
```

```
##             ORTH        GRAM        EXPR        RECI        MATH
## ORTH  1.00000000  0.53193641 0.512491039 -0.06584006  0.37741318
## GRAM  0.53193641  1.00000000 0.354117977 -0.10403022  0.34116941
## EXPR  0.51249104  0.35411798 1.000000000  0.23620258  0.58934714
## RECI -0.06584006 -0.10403022 0.236202578  1.00000000  0.18054329
## MATH  0.37741318  0.34116941 0.589347137  0.18054329  1.00000000
## ANGL  0.65441947  0.40357186 0.538117448 -0.04141681  0.69692407
## HIST  0.54466302  0.29507903 0.591412881 -0.05760725  0.63763616
## BIOL  0.35035637 -0.02040237 0.401657962  0.13463036  0.11846745
## EDMU  0.18337932  0.20517943 0.634169061  0.41163083  0.27548988
## ARTS  0.30740988  0.14521896 0.724991566  0.32487576  0.37126688
## TECH -0.29362248 -0.16625700 0.003608444  0.34084641  0.20992505
## EPS  -0.03960858 -0.09472332 0.026370611  0.23648194 -0.06024121
## GEO   0.16244818 -0.07176006 0.394955377 -0.11088774  0.32617140
## EXPO  0.14269742  0.05278196 0.314635869  0.32985689  0.41091247
##              ANGL        HIST        BIOL        EDMU        ARTS
## ORTH  0.654419471  0.54466302  0.35035637 0.183379320  0.30740988
## GRAM  0.403571855  0.29507903 -0.02040237 0.205179429  0.14521896
## EXPR  0.538117448  0.59141288  0.40165796 0.634169061  0.72499157
## RECI -0.041416815 -0.05760725  0.13463036 0.411630831  0.32487576
## MATH  0.696924068  0.63763616  0.11846745 0.275489880  0.37126688
## ANGL  1.000000000  0.59218482  0.17154873 0.004124413  0.28688027
## HIST  0.592184818  1.00000000  0.25930812 0.221296794  0.26157530
## BIOL  0.171548732  0.25930812  1.00000000 0.455598901  0.58337642
## EDMU  0.004124413  0.22129679  0.45559890 1.000000000  0.70027355
## ARTS  0.286880270  0.26157530  0.58337642 0.700273555  1.00000000
## TECH -0.226097749 -0.01474255 -0.16504177 0.176027437 -0.05394396
## EPS  -0.273408387  0.06731441  0.09803248 0.292349967  0.06551116
## GEO   0.134354152  0.32889942  0.46281961 0.322071990  0.23962044
## EXPO  0.253589128  0.25840339  0.56177062 0.432835103  0.57057209
##              TECH         EPS         GEO        EXPO
## ORTH -0.293622483 -0.03960858  0.16244818  0.14269742
## GRAM -0.166257002 -0.09472332 -0.07176006  0.05278196
## EXPR  0.003608444  0.02637061  0.39495538  0.31463587
## RECI  0.340846405  0.23648194 -0.11088774  0.32985689
## MATH  0.209925049 -0.06024121  0.32617140  0.41091247
## ANGL -0.226097749 -0.27340839  0.13435415  0.25358913
## HIST -0.014742553  0.06731441  0.32889942  0.25840339
## BIOL -0.165041767  0.09803248  0.46281961  0.56177062
## EDMU  0.176027437  0.29234997  0.32207199  0.43283510
## ARTS -0.053943960  0.06551116  0.23962044  0.57057209
## TECH  1.000000000  0.06968198  0.15458283  0.28872427
## EPS   0.069681976  1.00000000  0.34711802 -0.21845283
```

```
## GEO   0.154582835  0.34711802  1.00000000  0.09439988
## EXPO  0.288724270 -0.21845283  0.09439988  1.00000000
```

On remarque qu'il y a très peu de coefficient négatif entre les variables, et que ces coefficients sont relativement faible, les seuls coefficients de corrélation négatif ayant une valeur "correcte" sont ceux entre :

- L'orthographe et la technologie (-0.29)
- La grammaire et la technologie (-0.17)
- L'anglais et la technologie (-0.22)
- L'anglais et l'éducation physique et sportive (-0.27)
- La biologie et la technologie (-0.17)
- L'éducation physique et sportie et l'exposé (-0.22)

La technologie est souvent négativement corrélé avec les autres disciplines, cela n'est pas étonnant, en effet, cette matière est relativement à part et n'a pas de relation avec les autres, elle est très probablement peu travaillé par rapport aux autres.

Tout comme l'EPS qui n'a aucun rapport avec des matières comme l'anglais ou l'exposé, il n'est donc pas étonnant de voir un coefficient de corrélation négatif entre ces disciplines.

Question 4 : Nombre de composantes principales à retenir

Pour rappel, on obtient les valeurs propres des dimensions grâce à la variable suivante :

```
data.pca$eig
```

```
##         eigenvalue percentage of variance
## comp 1  4.77570045             34.1121461
## comp 2  2.39238450             17.0884607
## comp 3  1.49934217             10.7095870
## comp 4  1.34084173              9.5774409
## comp 5  1.12670192              8.0478709
## comp 6  0.72038480              5.1456057
## comp 7  0.59919071              4.2799337
## comp 8  0.42748490              3.0534636
## comp 9  0.39040787              2.7886277
## comp 10 0.25931678              1.8522627
## comp 11 0.18210022              1.3007158
## comp 12 0.12552548              0.8966105
## comp 13 0.09728799              0.6949142
## comp 14 0.06333049              0.4523606
##         cumulative percentage of variance
## comp 1                            34.11215
## comp 2                            51.20061
## comp 3                            61.91019
## comp 4                            71.48763
## comp 5                            79.53551
## comp 6                            84.68111
## comp 7                            88.96104
## comp 8                            92.01451
## comp 9                            94.80314
## comp 10                           96.65540
## comp 11                           97.95611
## comp 12                           98.85273
## comp 13                           99.54764
## comp 14                          100.00000
```

Pour connaître le nombre d'axe que nous devons retenir, nous pouvons utiliser ces 3 critères :

- Part d'inertie supérieure à la moyenne
- Part d'inertie cumulée supérieure à 80%
- Critère du coude

Nous allons pour chacun de ces critères, déterminer le nombre de composantes principales à retenir.

Part d'inertie supérieure à la moyenne

Pour utiliser ce critère, on commence par calculer la moyenne des pourcentages des valeurs propres, obtenu grâce à :

```
mean(data.pca$eig[,2])
```

```
## [1] 7.142857
```

Puis nous retenons les composantes principales dont le pourcentage d'inertie expliqué est supérieur à cette moyenne, dans notre cas, nous pouvons retenir les 5 premières composantes principales en utilisant ce critère.

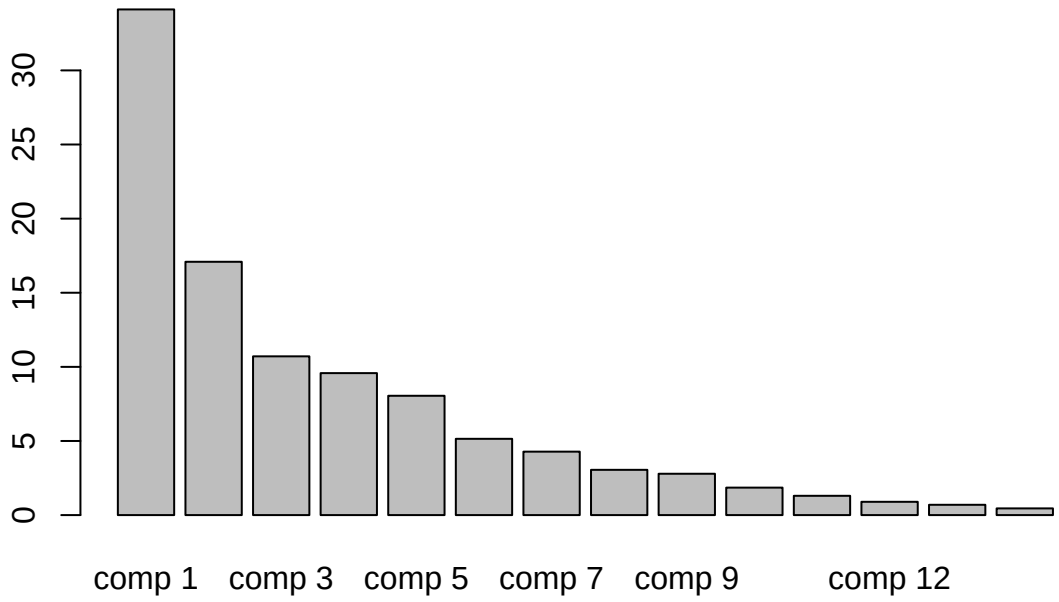
Part d'inertie cumulée supérieure à 80%

Pour ce critère, nous devons additionner les valeurs propres des dimensions jusqu'à obtenir un pourcentage cumulé supérieur à 80%, grâce au tableau précédent, nous pouvons retenir les 6 premières composantes qui à elles seules réprésentent un peu plus 84% de l'information totale.

Critère de coude

Afin d'utiliser ce critère, nous devons dans un premier temps, tracer le graphique suivant :

```
barplot(data.pca$eig[,2])
```



On recherche l'apparition d'un coude, sur notre graphique, le coude apparaît entre le 3e et 4e dimension, donc selon ce critère nous pouvons retenir 3 composantes principales.

Conclusion sur le nombre d'axe à retenir

Selon les critères, précédents, nous décidons de retenir 5 composantes principales pour notre analyse.

Question 5 : Interprétation des 3 premiers axes par rapport aux disciplines

On recupère les données (contribution, qualité de représentation et coordonnées) des disciplines sur les différentes dimensions grâce à la variable suivante :

```
data.pca$var
```

```
## $coord
##            Dim.1       Dim.2        Dim.3       Dim.4       Dim.5
## ORTH 0.650378590 -0.49271901  0.173647652 -0.10480943  0.14895341
## GRAM 0.417388795 -0.48575680 -0.071743415 -0.01051370  0.50581332
## EXPR 0.874166811  0.01540502  0.046797227  0.04371614  0.14884945
## RECI 0.247299110  0.59505913 -0.371423503  0.02016356  0.39197929
## MATH 0.733273680 -0.17286351 -0.306912714  0.42772948 -0.08498142
## ANGL 0.660960956 -0.58211516 -0.178752583  0.06388966 -0.09975141
## HIST 0.698144175 -0.30880357  0.072665513  0.34374243 -0.12739816
## BIOL 0.598226171  0.29312427  0.324794825 -0.44434410 -0.34799317
## EDMU 0.644856795  0.53510937  0.097359762 -0.09289748  0.29956935
## ARTS 0.758076081  0.33115891  0.006050377 -0.36129256  0.08037631
## TECH 0.009501976  0.51511927 -0.466209076  0.56544225 -0.12534130
```

```
## EPS  0.043328615  0.43201931  0.601812160  0.37873765  0.35484197
## GEO  0.447322427  0.22075833  0.540616973  0.38354653 -0.41420069
## EXPO 0.583734627  0.33551118 -0.453503335 -0.27451089 -0.33011273
##
## $cor
##            Dim.1       Dim.2        Dim.3       Dim.4       Dim.5
## ORTH 0.650378590 -0.49271901  0.173647652 -0.10480943  0.14895341
## GRAM 0.417388795 -0.48575680 -0.071743415 -0.01051370  0.50581332
## EXPR 0.874166811  0.01540502  0.046797227  0.04371614  0.14884945
## RECI 0.247299110  0.59505913 -0.371423503  0.02016356  0.39197929
## MATH 0.733273680 -0.17286351 -0.306912714  0.42772948 -0.08498142
## ANGL 0.660960956 -0.58211516 -0.178752583  0.06388966 -0.09975141
## HIST 0.698144175 -0.30880357  0.072665513  0.34374243 -0.12739816
## BIOL 0.598226171  0.29312427  0.324794825 -0.44434410 -0.34799317
## EDMU 0.644856795  0.53510937  0.097359762 -0.09289748  0.29956935
## ARTS 0.758076081  0.33115891  0.006050377 -0.36129256  0.08037631
## TECH 0.009501976  0.51511927 -0.466209076  0.56544225 -0.12534130
## EPS  0.043328615  0.43201931  0.601812160  0.37873765  0.35484197
## GEO  0.447322427  0.22075833  0.540616973  0.38354653 -0.41420069
## EXPO 0.583734627  0.33551118 -0.453503335 -0.27451089 -0.33011273
##
## $cos2
##              Dim.1        Dim.2        Dim.3        Dim.4       Dim.5
## ORTH 4.229923e-01 0.2427720217 3.015351e-02 0.0109850169 0.022187119
## GRAM 1.742134e-01 0.2359596680 5.147118e-03 0.0001105378 0.255847113
## EXPR 7.641676e-01 0.0002373146 2.189980e-03 0.0019111006 0.022156159
## RECI 6.115685e-02 0.3540953710 1.379554e-01 0.0004065693 0.153647766
## MATH 5.376903e-01 0.0298817942 9.419541e-02 0.1829525118 0.007221842
## ANGL 4.368694e-01 0.3388580549 3.195249e-02 0.0040818887 0.009950343
## HIST 4.874053e-01 0.0953596459 5.280277e-03 0.1181588583 0.016230290
## BIOL 3.578746e-01 0.0859218394 1.054917e-01 0.1974416771 0.121099246
## EDMU 4.158403e-01 0.2863420326 9.478923e-03 0.0086299409 0.089741798
## ARTS 5.746793e-01 0.1096662218 3.660706e-05 0.1305323137 0.006460351
## TECH 9.028755e-05 0.2653478631 2.173509e-01 0.3197249347 0.015710442
## EPS  1.877369e-03 0.1866406803 3.621779e-01 0.1434422081 0.125912825
## GEO  2.000974e-01 0.0487342381 2.922667e-01 0.1471079399 0.171562210
## EXPO 3.407461e-01 0.1125677514 2.056653e-01 0.0753562303 0.108974416
##
## $contrib
##              Dim.1        Dim.2        Dim.3        Dim.4      Dim.5
## ORTH  8.857178434 10.147700841  2.011115774  0.819262759  1.9692093
## GRAM  3.647913177  9.862949216  0.343291721  0.008243911 22.7076131
## EXPR 16.001163011  0.009919583  0.146062754  0.142529916  1.9664615
## RECI  1.280583878 14.800939040  9.201063052  0.030321946 13.6369489
## MATH 11.258878062  1.249038115  6.282449446 13.644601594  0.6409718
## ANGL  9.147755175 14.164029878  2.131100334  0.304427329  0.8831389
## HIST 10.205943479  3.985966555  0.352172894  8.812289759  1.4405132
## BIOL  7.493655757  3.591472838  7.035864125 14.725203802 10.7481175
## EDMU  8.707419780 11.968896845  0.632205474  0.643621144  7.9649991
## ARTS 12.033404324  4.583971428  0.002441541  9.735102286  0.5733860
## TECH  0.001890561 11.091355234 14.496417584 23.845091334  1.3943743
## EPS   0.039310859  7.801449997 24.155785266 10.697922436 11.1753449
## GEO   4.189905868  2.037057095 19.492996119 10.971312784 15.2269386
## EXPO  7.134997636  4.705253336 13.717033917  5.620069002  9.6719828
```

Pour chacune des 5 dimensions, nous allons retenir les variables dont la contribution est supérieure à la moyenne (soit 6.25), puis pour chacune des variables retenues, nous allons noter la qualité de leur répresentation sous cet axe, ainsi que le signe de sa projection.

Dim 1. (34,11%)

Discipline	Contribution	Qualité	Signe
EXPR	16.00	0.76	+
MATH	11.26	0.53	+
HIST	10.21	0.48	+
ARTS	12.03	0.57	+
ANGL	9.15	0.44	+
ORTH	8.86	0.43	+
EDMU	8.71	0.41	+
EXPO	7.14	0.34	+
Somme	83.36		

Les disciplines retenus expliquent à elles seules 83.36% de l'information portée par la dimension 1, on remarque que sur cet axe, toutes les coordonnées des variables sont positives, cet axe n'oppose donc aucune variable. Il y a donc un lien entre l'expression écrite (EXPR), les mathématiques (MATH), l'histoire (HIST), les arts plastiques (ARTS), l'anglais (ANGL), l'orthographe (ORTH), l'éducation musicale (EDMU) et l'exposé (EXPO).

Dim. 2 (17.09%)

Discipline	Contribution	Qualité	Signe
ORTH	10.14	0.24	-
GRAM	9.86	0.24	-
RECI	14.80	0.35	+
ANGL	14.16	0.34	-
EDMU	11.96	0.29	+
TECH	11.09	0.27	+
EPS	7.80	0.19	+
Somme	79.81		

Les disciplines retenus expliquent 79.81% de l'information portée par la dimension 2, en regardant le signe des variables sur cet axe, nous pouvons dire que la dimension 2 oppose la récitation (RECI), l'éducation musicale (EDMU), la technologie (TECH) et l'EPS aux matières orthographe (ORTH), grammaire (GRAM) et anglais (ANGL).

On remarque qu'on retrouve cette oposition dans les coefficients de corrélations entre ces disciplines.

Dim. 3 (10.7%)

Disciplines	Contribution	Qualité	Signe
EPS	24.16	0.36	+
GEO	19.5	0.29	+
TECH	14.5	0.22	-
EXPO	13.72	0.21	-
RECI	9.20	0.14	-

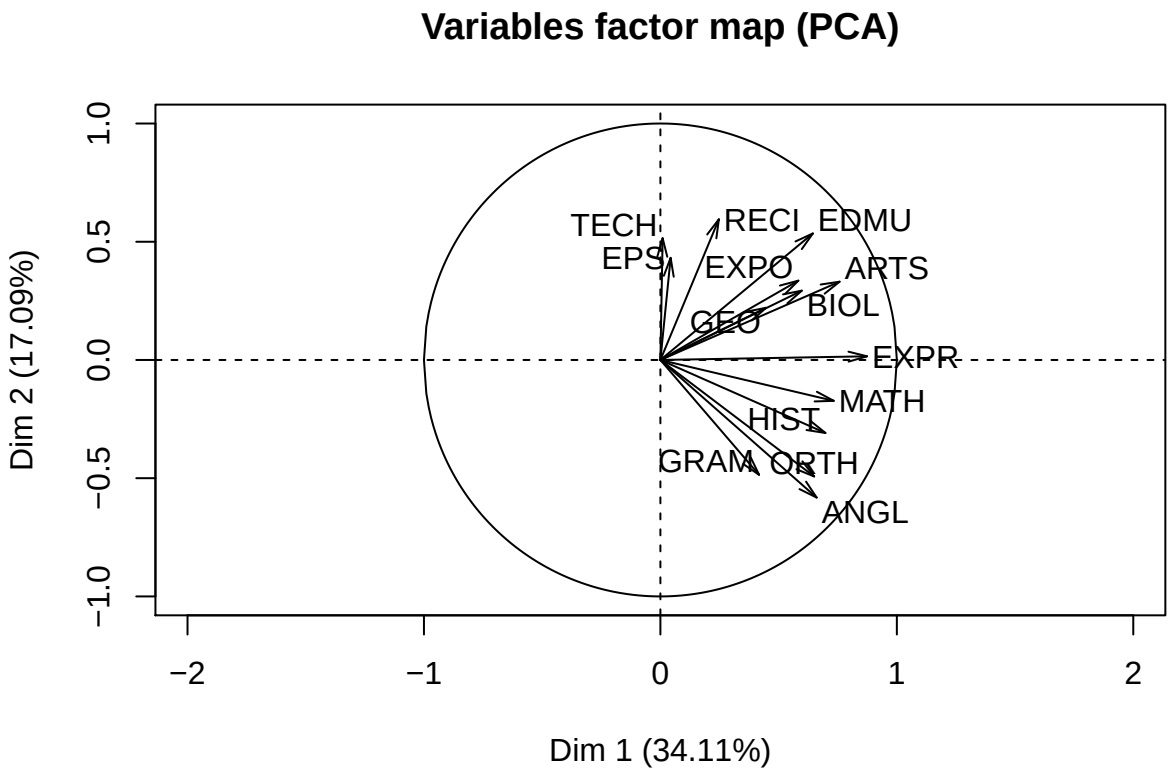
Disciplines	Contribution	Qualité	Signe
Somme	81.08		

Les disciplines retenus expliquent 81.08% de l'information portée par la dimension 3, cet axe met en opposition l'EPS et la géographie (GEO) à la technologie (TECH), l'exposé (EXPO) et la récitation (RECI).

Question 6 : Cercle de corrélation

On obtient le cercle de corrélation avec la fonction `plot.PCA` dans la librairie `FactoMineR` :

```
FactoMineR::plot.PCA(data.pca, axes = c(1,2), choix = "var")
```



On remarque que suivant l'axe 1, toutes les variables sont représentés positivement, cet axe n'oppose donc aucune variable / discipline, elles sont toutes liées entre elles, si la valeur de l'une de ces matières est grande, alors la valeur du reste des matières le sera aussi (mais faiblement voir nul dans le cas de la technologie ou l'EPS étant donné que les flèches sont quasiment orthogonales à l'axe 1, et cela repose encore sur les coefficients de corrélation de ces 2 disciplines avec les autres)

Les disciplines les mieux représentés dans ce plan formé des 2 premiers axes sont les suivantes :

- l'éducation musicale (EDMU)
- l'expression écrite (EXPR)
- l'anglais (ANGL)
- l'orthographe (ORTH)

Question 7 : Interprétation

La majorité des matières étant corrélés positivement, on peut déduire de ce phénomène que si un élève est plutôt bon dans une de ces matières, il sera également bon dans les autres, et inversement, si il a de mauvaises notes dans une de celles-ci, il est fortement probable qu'il en a des mauvaises dans les autres également.

On peut présumer qu'il y a 2 types d'élèves dans cette classe :

- Ceux travaillant dans chacune de ces disciplines et donc ayant une bonne moyenne dans celles-ci
- Ceux ne travaillant pas dans aucunes matières et donc ayant une moyenne faible dans celles-ci

Question 8 : Contribution de certains élèves

La contribution des élèves EL10 et EL12 par rapport aux axes factoriels 1 et 2 est donnée par :

```
data.pca$ind$contrib[c("EL10","EL12"),c("Dim.1","Dim.2")]
```

```
##         Dim.1      Dim.2
## EL10 18.58838 7.22537004
## EL12 20.98589 0.00230761
```

On remarque que les 2 élèves contribuent fortement au premier axe factoriel, tandis que seul l'élève EL10 contribue au deuxième axe factoriel.

Nous pouvons interpréter ces résultats en disant que ces 2 élèves ont une bonne moyenne en expression écrite, en mathématiques, en biologie, en art plastique, en histoire. Tandis que seul l'élève EL10 a de bonne note en technologie, sport et récitation.

Question 9 : Explication des axes 1,2 et 3 par les élèves

La contribution des différents élèves sur les 3 premiers axes est donnée grâce à la variable suivante :

```
data.pca$ind$contrib[,c("Dim.1","Dim.2","Dim.3")]
```

```
##             Dim.1       Dim.2        Dim.3
## EL01  9.328905241  9.59126708  0.616515806
## EL02  0.125275069  3.96118861  0.258445075
## EL03  0.090900704  0.16283004  5.087265764
## EL04  0.565357272  1.30848854  9.031255314
## EL05  1.487781989  1.64284957  0.148851085
## EL06  0.707432497  0.57129277  3.707371169
## EL07  1.853709107  0.76715711  5.398964876
## EL08  0.112046826  2.41932985  0.013041174
## EL09  7.407644214  1.52394402  0.009609208
## EL10 18.588377979  7.22537004  0.003702659
## EL11  0.306502245  1.89876424  1.487091590
## EL12 20.985890664  0.00230761  2.387411364
## EL13  3.231582745 14.07178676 14.446330034
## EL14  0.043589696  0.38069178  0.008154061
## EL15  0.158993355  1.06053122  0.105021975
## EL16  0.008219267  6.63073046  0.008769846
## EL17  1.187170818  5.85814498  1.849005639
## EL18  0.293940859  1.67494396  3.919286109
## EL19  0.002322024  5.62257996  0.012798694
## EL20  2.054270051  0.03005751  4.669204845
## EL21 12.582779600  5.23836937  0.273267018
## EL22  1.781543769  5.83680752  1.427638803
## EL23  4.620411054 16.67349157 22.144458992
```

```
## EL24 11.219745959  0.10732552 15.428296052
## EL25  0.484483948  0.01002931  5.460474394
## EL26  0.057921681  5.00334513  1.027681671
## EL27  0.713201366  0.72637544  1.070086782
```

Pour la dimension 1, les élèves qui l'expliquent sont : 1,9,10,12,21,23,24

Pour la dimension 2, les élèves qui l'expliquent sont : 1,13,16,17,19,21,22,23,26

Pour la dimension 3, les élèves qui l'expliquent sont : 3,4,7,13,20,23,24,25

Partie 3 : Clasification ascendate hiérarchique sur les axes retenus

Pour rappel, nous avons retenus les 5 premiers axes factoriels qui représentent 84% de l'information totale. Notre classification se portera donc seulement sur ces axes.

Réalisation sous R

Pour notre étude, nous aurons besoins des librairies suivantes :

```
library(ggplot2)
library(plyr)
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:Hmisc':
##
##     is.discrete, summarize
```

```
library(philentropy)
library(factoextra)
```

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

On stocke ensuite les résultats de la CAH dans la variable `data.hcpc` et on décide de ne pas afficher les graphiques lors de l'appel à la fonction :

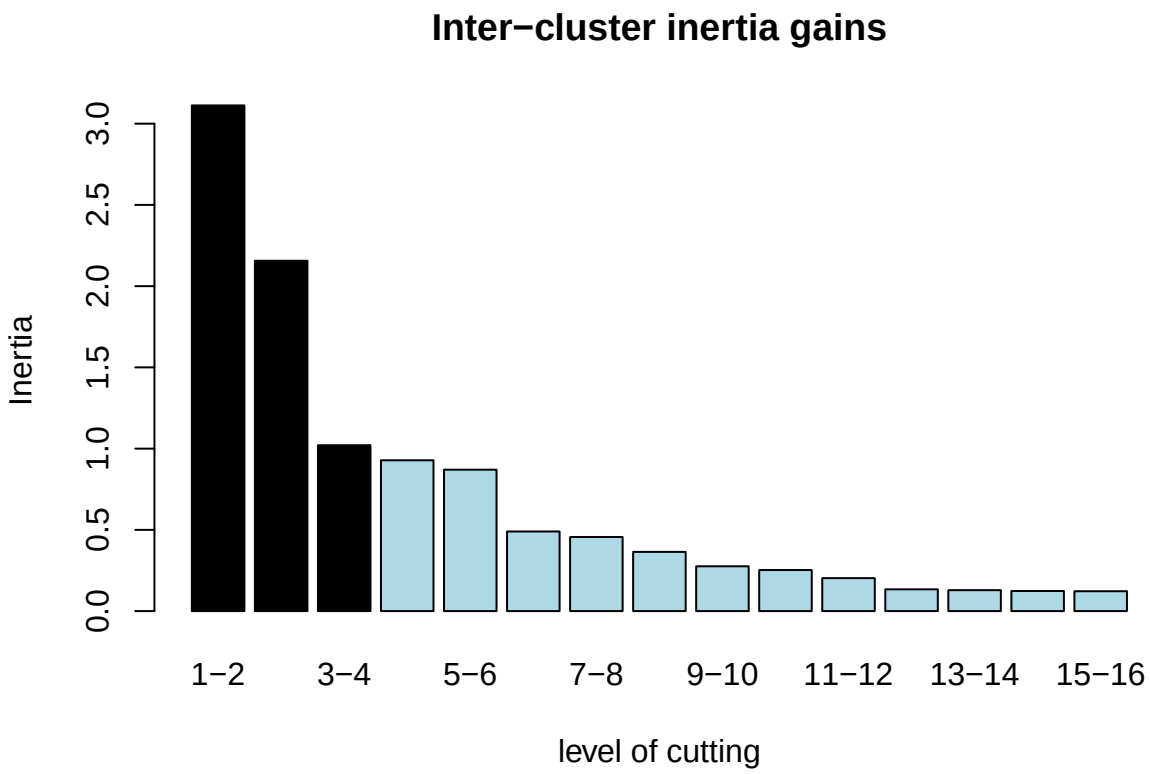
```
data.hcpc <- FactoMineR::HCPC(data.pca, nb.clust = 4, proba = 1, graph = FALSE)
```

Interprétation et analyse des résultats

Interprétation du graphique de gain d'inertie

Avant d'interpréter les résultats de la classification, on décider d'afficher le graphique de gain d'inertie par le nombre de classe retenu, obtenu par le code :

```
plot(data.hcpc, choice = "bar")
```



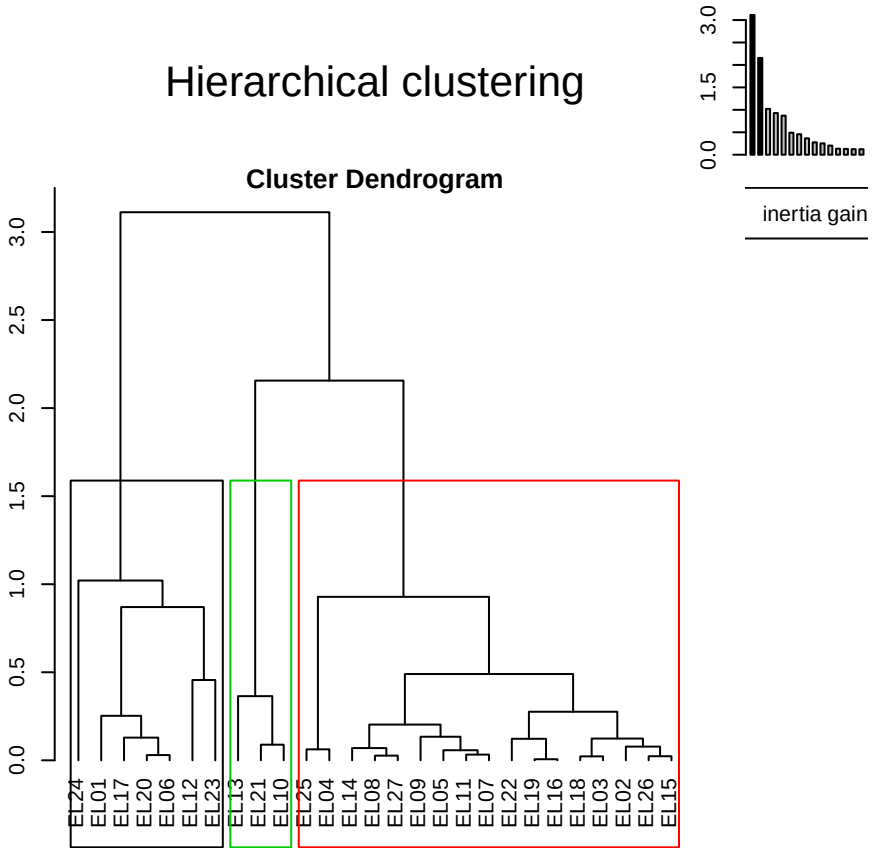
On remarque que le gain d'inertie est faible si on garde la sépération en 4 classes, on relance donc la classification mais cette fois-ci en seulement 3 classes :

```
data.hcpc <- FactoMineR::HCPC(data.pca, nb.clust = 3, proba = 1, graph = FALSE)
```

Typologie en 3 classes

Pour visualiser la coupure de l'arbre en 3 classes, on utilise la fonction suivante :

```
plot(data.hcpc, choice = "tree")
```



On distingue 4 groupes principaux :

- Le premier groupe est formé des élèves suivants : EL24, EL01, EL17, EL20, EL06, EL12 et EL23
- Le second est quant à lui formé de seulement 3 élèves : EL13, EL21 et EL10
- Puis le dernier regroupe les autres élèves de la classe, soit : EL25, EL04, EL14, EL08, EL27, EL09, EL05, EL11, EL07, EL22, EL19, EL16, EL18, EL03, EL02, EL26 et EL15

Analyse des variables les plus explicatives pour les 3 classes

On se décide d'abord de déterminer quelles seront les variables qui seront les plus explicatives pour l'analyse des 3 classes, pour cela on utilise la p.value de nos variables :

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

```
##            Eta2      P-value
## EXPR 0.71539493 2.824304e-07
## EDMU 0.70890932 3.701175e-07
## GRAM 0.53989807 9.000128e-05
## ARTS 0.51153881 1.844849e-04
## BIOL 0.43199002 1.127903e-03
## ANGL 0.40777314 1.861475e-03
## MATH 0.37029314 3.887417e-03
## ORTH 0.34994451 5.693839e-03
## HIST 0.34711756 5.998185e-03
## GEO  0.28050543 1.924555e-02
## EXPO 0.26410391 2.522307e-02
## EPS  0.22104966 4.990192e-02
## RECI 0.16932766 1.079337e-01
```

```
## TECH 0.06459432 4.487454e-01
```

On remarque que les classes seront donc d'abord crée en fonction des notes des élèves en expréssion écrite (EXPR) et éducation musicale (EDMU) puis par la grammaire (GRAM) ou les arts plastiques (ARTS)

Caractérisation des 3 classes en fonction des différentes variables

Pour chaque classe determinée au dessus, on décide de les caractériser en fonction des variables de notre problème, c'est à dire en fonction de leur note dans les différentes matières qui sont enseignés dans ce collège. Pour cela, on regardera la moyenne des matières dans cette échantillon par rapport à la moyenne générale (c'est à dire de l'ensemble des élèves) et de la valeur de la p-value qui doit être inférieur à 5%.

Classe 1

Pour la classe 1, les données sont obtenus par :

```
data.hcpc$desc.var$quanti$`1`
```

```
##          v.test Mean in category Overall mean sd in category Overall sd
## GRAM -0.1356952         7.428571     7.555556       1.699340   2.822966
## ANGL -0.6563349        10.214286    10.925926       3.711537   3.270813
## TECH -0.7595873        11.428571    12.629630       5.900536   4.769876
## EPS  -1.1817674        11.928571    13.074074       2.932924   2.924050
## HIST -1.5331929         8.142857     9.203704       1.726149   2.087260
## MATH -1.7176060        11.000000    12.574074       2.927700   2.764536
## ORTH -1.8344889         7.142857    10.018519       4.509627   4.728717
## RECI -1.9032830         7.857143    10.074074       3.512369   3.513739
## EXPO -2.6204369        10.642857    14.481481       6.890544   4.418986
## GEO  -2.6451029        11.000000    13.796667       4.605897   3.189472
## BIOL -3.2264077         5.714286     9.592593       1.749636   3.626132
## EXPR -3.4205919         5.500000     8.888889       1.732051   2.988662
## ARTS -3.5872961         4.428571     9.018519       2.258770   3.859768
## EDMU -4.2619279        10.714286    15.203704       2.519313   3.177640
##           p.value
## GRAM 8.920622e-01
## ANGL 5.116087e-01
## TECH 4.475013e-01
## EPS  2.372980e-01
## HIST 1.252283e-01
## MATH 8.586850e-02
## ORTH 6.658146e-02
## RECI 5.700363e-02
## EXPO 8.781717e-03
## GEO  8.166608e-03
## BIOL 1.253547e-03
## EXPR 6.248501e-04
## ARTS 3.341248e-04
## EDMU 2.026709e-05
```

On remarque

Question 10 : Résumé

En résumé, nous pouvons dire que la plupart des matières (les matières principales) sont liées entre elles, elles sont donc enseignées équitablement, l'une ne prends pas l'ascendant sur une autre et inversement, les cours sont donc variées et enseignés en période égale.

La seule différence est avec la technologie et l'éducation physique et sportive, ces 2 disciplines sont à part et n'ont aucun lien avec les autres, nous pouvons dire que ces 2 matières possèdent moins d'heure que les autres et que vu qu'elles ne sont pas des matières "logiques" il est normal qu'il n'existe pas de lien notable entre celles-ci et les autres matières.