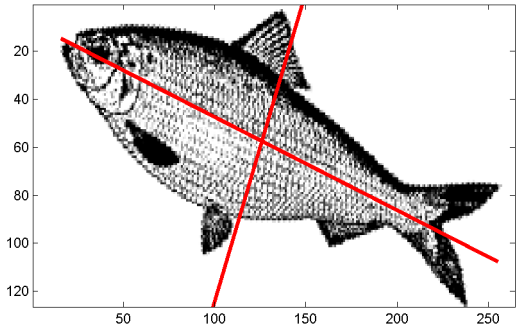
Université de Lille 3, Sciences Humaines et Sociales
UFR Mathématiques, Informatique, Management, Économie

Projet Statistique Multivariée

L'analyse en composantes principales sur un jeux de données : "la note de 22 étudiants sur 8 matières "



Réalisé par :
Ben Yassine Mohamed
Bajti Fatouma

Encadré par : MMe. SOPHIE DABO

19 AVRIL 2019

Table des matières

Première partie

Introduction

Introduction

L'analyse en composantes principales consiste à transformer des variables liées entre elles en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées "composantes principales", ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante.

Mathématiquement, l'analyse en composantes principales est un simple changement de base : passer d'une représentation dans la des facteurs dêfinis par les vecteurs propres de la matrice des corrélations.

Jeux de données

Les données utilisées ici dans mon projet sont disponibles dans le paquet ade4, il s'agit du jeux de données "seconde" qui contient les notes de 8 matières (nos variables) pour 22 étudiants. Les 8 variables sont :

HGEO : Histoire géographie
FRAN : Français
PHYS : Physiques
MATH : Mathématiques
BIOL : Biologie
ECON : Économie
ANGL : Anglais
ESPA : Langue espagnole

Deuxième partie

Statistiques élémentaires

Statistiques élémentaires

.1 Statistiques decriprtives

```
library(ade4) #echo=false
library(FactoMineR)
library("factoextra")
data(seconde)
summary(seconde)
```

```
      HGEO            FRAN            PHYS            MATH            BIOL      
 Min.   : 8.80   Min.   : 8.00   Min.   : 3.700   Min.   : 3.700   Min.   : 8.00  
 1st Qu.:11.25   1st Qu.: 8.80   1st Qu.: 5.125   1st Qu.: 7.625   1st Qu.:10.50  
 Median :12.65   Median : 9.55   Median : 8.600   Median : 8.750   Median :11.50  
 Mean   :12.23   Mean   :10.14   Mean   : 8.532   Mean   : 9.323   Mean   :11.18  
 3rd Qu.:13.55   3rd Qu.:11.50   3rd Qu.:11.000   3rd Qu.: 9.575   3rd Qu.:12.20  
 Max.   :15.20   Max.   :14.00   Max.   :18.000   Max.   :18.800   Max.   :13.60  
      ECON            ANGL            ESPA      
 Min.   : 5.00   Min.   : 5.00   Min.   :11.00  
 1st Qu.:10.00   1st Qu.: 9.00   1st Qu.:12.93  
 Median :11.25   Median :10.00   Median :13.75  
 Mean   :11.18   Mean   :10.23   Mean   :13.88  
 3rd Qu.:13.50   3rd Qu.:11.75   3rd Qu.:14.50  
 Max.   :15.50   Max.   :14.00   Max.   :19.50  
```

Variable	Moyenne	Ecart type
HGEO	12.23182	1.813281
FRAN	10.13636	1.870713
PHYS	8.531818	3.925907
MATH	9.322727	3.351749
BIOL	11.17727	1.519377
ECON	11.18182	2.933978
ANGL	10.22727	2.428635
ESPA	13.87727	1.796539
FRAN	10.13636	1.870713

La moyenne de cet échantillon de 22 étudiants en Biologie et en économie sont presque identiques (11.17727 pour la biologie et 11.18182 pour l'économie) or la valeur de l'écart type en économie est plus élevé que la valeur en biologie donc les notes en économie sont plus distribuées autour de la moyenne que les notes en biologie.

La moyenne des notes des étudiants en mathématiques est 9.322727, l'écart type vaut 3.351749, donc les notes sont largement distribués autour de la moyenne, et donc leurs niveau en mathématiques n'est pas le même

certains étudiants sont beacoup plus bon en mathématiques que les autres (la note maximale est 18.8 et la note minimale est 3.7)

.2 Résultat de la corrélation

```
Mcorrel=cor(seconde[,1:6])
Mcorrel
```

```
          HGEO      FRAN      PHYS      MATH      BIOL      ECON
HGEO 1.0000000 0.6772620 0.6317149 0.5325050 0.2704275 0.6531598
FRAN 0.6772620 1.0000000 0.5566687 0.4386750 0.1577884 0.4581293
PHYS 0.6317149 0.5566687 1.0000000 0.6981611 0.4150919 0.4174341
MATH 0.5325050 0.4386750 0.6981611 1.0000000 0.3150375 0.1792094
BIOL 0.2704275 0.1577884 0.4150919 0.3150375 1.0000000 0.3630957
ECON 0.6531598 0.4581293 0.4174341 0.1792094 0.3630957 1.0000000
```

C'est la matrice de variance covariance des variables centrées réduites.Elle possède p valeurs propres.
Le coefficient de corrélation nous donne deux informations que l'on doit interpréter :
- **le sens de la relation entre les variables :** si le coefficient est négatif, plus la valeur de la première variable est élevé, plus la valeur de la deuxième diminue.
- **la force de la relation :** En examinant la valeur de chaque coefficient, nous pouvons dire que l'effet de la relation entre deux variables est de grande taille et que l'assocciation est très forte, ou bien le contraire.

On remarque que le coefficient de corrélation entre la variable "PHYS" (physique) et la variable "MATH" (Mathematiques) est positive, donc les deux variable sont positivement corrélées, elle varient dans le même sens, de plus le coefficient vaut 0,6981611 donc elles sont fortement corrélées. On peut dire les étudiants qui ont des bonnes notes en mathématiques tendent à avoir des bonnes notes en physique. C'est le cas aussi pour les deux variables "FRAN" et "HGEO".
Les deux variables "PHYS" et "ESPA" sont corrélées négativement donc plus un étudiant est bon en physique, il sera moins bon en espagnole.

Troisième partie

Choix et Nombre d'axes factoriels à retenir

Choix et Nombre d'axes factoriels à retenir

.1 Table des Valeurs propres et vecteurs propres

Les valeurs propres permettent d'effectuer un choix du nombre de composantes principales à retenir pour l'interprétation.

Le choix du nombre d'axes à interpréter se fait sur la base de règles.

- **La règle de Kaiser :** Elle consiste à retenir les axes pour lequels les valeurs propres sont supérieures à 1 (1 étant la moyenne de l'ensemble des valeurs propres). Il est à noter qu'on peut aussi avoir des résultats d'ACP dont la somme des valeurs propres n'est pas égale à p (nombre de variable) (cas de l'ACP non réduite). Dans ce cas, il faut adapter cette règle de Kaiser et retenir les valeurs propres supérieures à la moyenne des valeurs propres, et non plus à 1.

```
pca<- PCA(seconde,scale.unit=TRUE,ncp = 4, graph = FALSE)
names(pca)
```

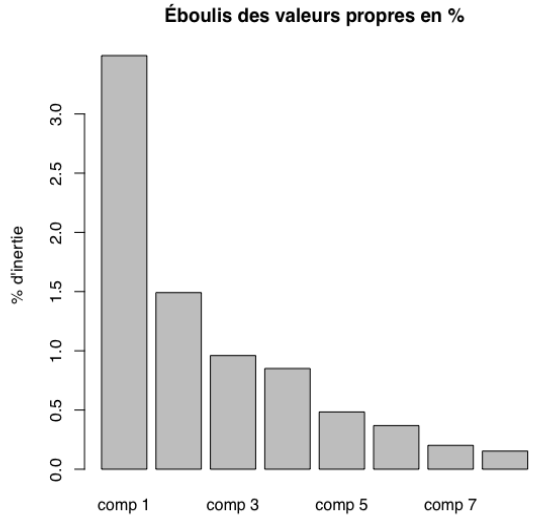
```
[1] "eig"  "var"  "ind"  "svd"  "call"
```

```
pca$eig
```

```
       eigenvalue percentage of variance cumulative percentage of variance
comp 1  3.4926186              43.657732                           43.65773
comp 2  1.4910720              18.638400                           62.29613
comp 3  0.9596309              11.995387                           74.29152
comp 4  0.8502761              10.628452                           84.91997
comp 5  0.4834123               6.042654                           90.96262
comp 6  0.3684670               4.605837                           95.56846
comp 7  0.2015152               2.518940                           98.08740
comp 8  0.1530079               1.912599                          100.00000
```

- **La règle de l'éboulis :** Elle consiste à retenir les 2 premiers axes au moins, puis de "couper" l'éboulis des valeurs propres entre les valeurs propres dont la différence est maximum.

```
#eboulis
inertie<-pca$eig/sum(pca$eig)*100
barplot(pca$eig[,1],ylab="% d'inertie")
title("Éboulis des valeurs propres en %")
```



- La règle de l'éboulis combineée avec celle de Kaiser est une des meilleurs. En effet, on commence par regarder combien de valeurs propres sont supérieures à la moyenne. Puis on regarde si la dernière valeur propre retenue (supérieure à la moyenne) est suffisament éloignée de celle qui la suit (inférieure à la moyenne).Si oui, on reste sur la décision de la règle de Kaiser, si non, on coupera au saut plus important le plus près.

Dans notre, on a appliqué la règle de l'éboulis combineée avec celle de Kaiser Dans notre exmple,le nombre de variables p=8, est bien la somme des valeurs propres. on retiendra donc 2 axes pour l'interprétation. le premier et le deuxième axe comportent respectivement 43.65 et 18.62 de l'inertie totale du nuage, et le plan (1,2) totalise 62.29 % de cette variance totale.

Quatrième partie

Étude des individus : Résultats sous R

Étude des individus : Résultats sous R

.1 Coordonnées des individus, contribution et qualité de la représentation d'un individu

De même, nous stockons le résultat dans une variable, ainsi nous pourrons avoir les coordonnées des individus mais ainsi la qualité de contribution sur chacun des axes.

Résultats des individus :

```
res.ind= pca$ind
res.ind$coord # coordonnées
```

```
        Dim.1      Dim.2      Dim.3       Dim.4
1  -2.2455642  1.0621795 -0.8525192  0.06173170
2   0.4576538 -1.4517620  0.4229862 -1.02171877
3   0.6046177  1.1109252  0.3765567 -0.45343441
4  -2.6389527  0.5072452  0.4653774 -0.06449521
5   0.1184447 -0.4461454  0.3660597 -0.45832461
6  -0.5878830  0.6170288  0.1954045 -0.11910808
7   0.1639646 -1.2939089 -0.5207035 -0.96252733
8   4.1704371  1.1999025 -0.9476905  0.54072761
9   0.8370768  2.4899828  2.3676281  0.82508655
10  2.3464453 -1.0950849 -1.0594219  0.61426336
11 -2.0932682  2.2484318 -0.3043635  0.71634864
12  1.1844011 -1.8376414 -0.8468450  0.97744931
13  0.2772426 -0.7436505  0.5880185 -0.92421513
14 -1.8499631 -0.9414966  1.3689088 -0.18376882
15  2.2585095  1.7432841 -0.8613618 -2.84881399
16  1.8233121 -0.6364264  0.8702628  0.11545198
17 -0.0156162 -0.5256530  0.5107955  0.17127651
18 -1.7716804 -1.4764047  0.9539400  0.15678756
19 -3.0006080  0.4200908 -1.9078163  0.28318286
20  0.5088940 -0.9028870  0.6604060  0.49934300
21 -2.5108827 -0.3545605 -1.3098498  0.05603479
22  1.9634192  0.3065510 -0.5357728  2.01872247
```

On remarque que ce sont les étudiants 8 et 10 et 15 qui sont les plus représentatifs positivement avec le première axe et les étudiants 19 et 21 et 4 et 21 et 1 qui sont les plus représentatifs négativement.

On remarque que ce sont les étudiants 9 et 11 et 15 qui sont les plus représentatifs positivement avec le deuxième axe et les étudiants 12 et 18 et 10 ... qui sont les plus représentatifs négativement.

Elle permet de verifier que tous les individus sont bien représentés par le sous-espace principal choisi ; elle s'exprime comme le carré du cosinus de l'angle entre les individus et sa projection orthogonale.

Pour chaque axe retenu et chaque nuage, on regarde Quels sont les individus qui participent le plus à la formation de l'axe. Il faut s'assurer que les points contribuant le plus a‘ l'axe sont bien représentés sur l'axe (sinon il faut les mettre en éléments supplémentaires.)

Contribution :

```
# Contribution des individus aux axes factoriels
pca$ind$contrib
```

```
          Dim.1      Dim.2      Dim.3       Dim.4
1  6.562618e+00  3.4393352  3.4425590  0.02037201
2  2.725840e-01  6.4249474  0.8474721  5.58058946
3  4.757599e-01  3.7622561  0.6716359  1.09912295
4  9.063363e+00  0.7843584  1.0258495  0.02223680
5  1.825816e-02  0.6067805  0.6347123  1.12295844
6  4.497881e-01  1.1606187  0.1808598  0.07584026
7  3.498858e-02  5.1037115  1.2842632  4.95271652
8  2.263546e+01  4.3890508  4.2540849  1.56305558
9  9.119201e-01 18.9004173 26.5521722  3.63928741
10 7.165508e+00  3.6557313  5.3163179  2.01709594
11 5.702640e+00 15.4112599  0.4387915  2.74325521
12 1.825676e+00 10.2943815  3.3968860  5.10746996
13 1.000337e-01  1.6858406  1.6377781  4.56628991
14 4.454021e+00  2.7021920  8.8760981  0.18053483
15 6.638501e+00  9.2643516  3.5143445 43.38559118
16 4.326614e+00  1.2347388  3.5873514  0.07125578
17 3.173781e-04  0.8423198  1.2358542  0.15682415
18 4.085046e+00  6.6449172  4.3103763  0.13141368
19 1.171776e+01  0.5379786 17.2403539  0.42869779
20 3.370395e-01  2.4851094  2.0658324  1.33295431
21 8.205008e+00  0.3832299  8.1267341  0.01678545
22 5.017094e+00  0.2864733  1.3596725 21.78565240
```

Pour l'axe 1 : Les individus 9, 4, 10 et 21 participent le plus à la création de l'axe du côté positif. En effet les variables contribuent toutes dans le même sens à la formation de l'axe pour l'axe 2 : Les individus 11, 12, 15, et 9 participent le plus à la création de l'axe du coté positif. les variables sont toutes du même côté de l'axe.

Qualité de représentation des individus :

```
# Qualité de représentation des individus

pca$ind$cos2
```

```
          Dim.1      Dim.2       Dim.3        Dim.4
1  6.457180e-01 0.14447333 0.093067964 0.0004879871
2  3.646479e-02 0.36693605 0.031149557 0.1817449121
3  9.417962e-02 0.31795444 0.036530540 0.0529692953
4  7.549852e-01 0.02789402 0.023479336 0.0004509518
5  5.958389e-03 0.08453779 0.056911729 0.0892163238
6  2.718589e-01 0.29948327 0.030035213 0.0111594855
7  7.270419e-03 0.45275866 0.073323015 0.2505448643
8  8.038431e-01 0.06654264 0.041508898 0.0135134181
9  4.908430e-02 0.43431476 0.392680100 0.0476882005
10 4.834766e-01 0.10530528 0.098558133 0.0331332342
11 4.120515e-01 0.47540220 0.008711378 0.0482559243
12 1.899220e-01 0.45719271 0.097092492 0.1293499416
13 2.492346e-02 0.17931886 0.112116666 0.2769710200
14 4.799535e-01 0.12431128 0.262797999 0.0047360545
15 2.995113e-01 0.17844551 0.043565286 0.4765379539
16 6.140415e-01 0.07481214 0.139886741 0.0024619456
17 9.267763e-05 0.10500801 0.099155836 0.0111485843
18 4.081314e-01 0.28342626 0.118323589 0.0031963362
19 6.920276e-01 0.01356411 0.279755097 0.0061636621
20 9.706833e-02 0.30555520 0.163472569 0.0934589289
21 6.858880e-01 0.01367669 0.186656601 0.0003415985
22 4.358063e-01 0.01062363 0.032451045 0.4607026099
```

Pour la qualité, il faut s'assurer que le cosinus carré supérieure à 0.5.

Cinquième partie

Études des variables : Résultats sous R

Études des variables : Résultats sous R

.1 Détermination des variables expliquantle mieux un axe donnée

Lorsque l'on a beaucoup de variables, une description automatique des axes par les variables est possible à l'aide de cette commande pour le plan (1,2)

```
dimdesc(pca, axes=c(1,2))
```

```
$Dim.1
$Dim.1$quanti
     correlation      p.value
HGEO   0.8712886 1.310221e-07
PHYS   0.8650766 2.043435e-07
FRAN   0.7711661 2.654784e-05
MATH   0.7217584 1.495767e-04
ECON   0.6743279 5.781828e-04
BIOL   0.4605160 3.102175e-02


$Dim.2
$Dim.2$quanti
     correlation      p.value
ESPA   0.7855568 1.479555e-05
ANGL   0.7848921 1.521507e-05
BIOL  -0.4383492 4.128818e-02
```

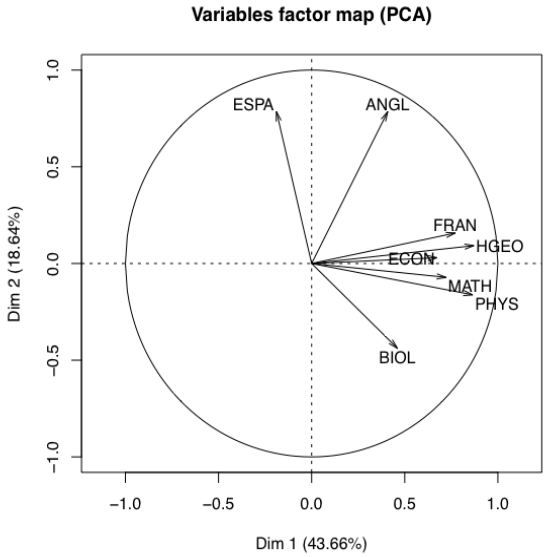
La détermination des variables expliquant chacun des axes est réalisée en examinant leurs coordonnées (table des valeurs propres) qui sont elle-même reliées à leur contribution. -Les variables les plus corrélées à la première dimension sont dans l'ordre : **HGEO, PHYS, FRAN, MATH**. -Les variables les plus corrélées à la deuxième dimension sont dans l'ordre : **ESPA, ANGL, BIOL**

.2 Cercle de corrélation, représentations graphiques

C'est une représentation où, pour deux composantes principales, par exemple c1 et c2, on représente chaque variable z j par un point d'abscisse cor(z j , c1) et d'ordonnée cor(z j , c2).

Les deux premières dimensions contiennent 50% de l'inertie totale (l'inertie est la variance totale du tableau de données, i.e. la trace de la matrice de corrélation).

```
z<- dudi.pca(seconde, center = T, scale = T, scannf = F)
plot.PCA(pca, axes=c(1, 2), choix="var")
```




```
pca$var$cos2  # Qualités de représentation des variables
```

```
        Dim.1       Dim.2       Dim.3       Dim.4
HGEO 0.7591438 0.008528485 0.007629824 0.044123464
FRAN 0.5946972 0.024834261 0.006693173 0.078466745
PHYS 0.7483575 0.026365583 0.059848845 0.020290628
MATH 0.5209352 0.005151395 0.153992651 0.180009110
BIOL 0.2120750 0.192150024 0.268206937 0.266735914
ECON 0.4547182 0.000887104 0.312579568 0.115046882
ANGL 0.1669078 0.616055636 0.026998352 0.005829332
ESPA 0.0357840 0.617099474 0.123681587 0.139774051
```

```
pca$var$contrib #Correlation des variables avec les axes
```

```
        Dim.1       Dim.2      Dim.3     Dim.4
HGEO 21.735662  0.57197006  0.7950790  5.189310
```

```
FRAN 17.027257  1.66553069  0.6974737  9.228384
PHYS 21.426832  1.76823008  6.2366523  2.386358
MATH 14.915318  0.34548267 16.0470703 21.170665
BIOL  6.072091 12.88670355 27.9489673 31.370505
ECON 13.019405  0.05949438 32.5728939 13.530532
ANGL  4.778873 41.31629135  2.8134099  0.685581
ESPA  1.024561 41.38629721 12.8884535 16.438666
```

Sixième partie

Conclusion ACP

Conclusion ACP

.1 Interprétation des axes : synthèse

L'interprétation des nouvelles variables (des axes factoriel) se fera à l'aide des individus et variables contribuant le plus à l'axe avec la règle suivante :

si une variable a une forte contribution positive à l'axe, les individus ayant une forte contribution positive à l'axe sont caractérisés par une valeur élevée de la variable.

On donne un sens à un axe à partir des coordonnées des variables et des individus. Les résultats obtenus dans les chapitres précedents montrent que :

1. Les variables "PHYS" et "MATH" sont les plus corrélées positivement à la première dimension et que sont les variables "ESPA" et "ANGL" les plus corrélées à la deuxième dimension.

donc l'axe DIM1 oppose les matières scientifique(Mathématiques, Physiques, Biologie) aux matières économique et littéraire. l'axe 2 oppose les deux matières Anglais et Espagnole.

2. Les individus (2, 6,..)contribuent à la première dimension et les individus (11,12) contribuent le plus à la deuxième dimension.

3. Ces individus (2, 6,..) sont bien représentés sur les axe.

les deux axes donc opposent les matières scientifique contre les matière littéraire

EX : L'étudiant 12 a des notes en (Mathématiques, physique, biologie) plus élévées que les notes (anglais, espagnole, francais)

On peut aussi diviser le premier plan factoriel en quatre parties : - Des étudiants qui sont plus bon en (Mathématiques, physique, biologie) que en (Histoire géographie, Français, Économie) + ils sont plus bon en Anglais que l'espagnole.

- Des étudiants qui sont plus bon en (Mathématiques, physique, biologie) que en (Histoire géographie, Français, Économie) + ils sont plus bon en Espagnole que en anglais. - Des étudiants qui sont plus bon en (Histoire géographie, Français, Economie) que en (Mathématiques, physique, biologie) + ils sont plus bon en Anglais que l'espagnole.

- Des étudiants qui sont plus bon en (Histoire géographie, Français, Économie) que en (Mathématiques, physique, biologie) + ils sont plus bon en Espagnole que en anglais.

Septième partie

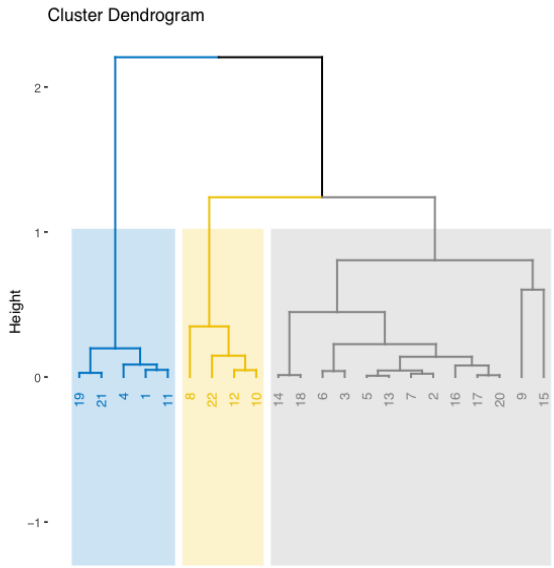
Classification ascendante hiérarchique

Classification ascendante hiérarchique

```
res=HCPC(pca, graph = FALSE)
```

Dendogramme qui suggére 3 classes

```
fviz_dend(res, cex = 0.7, palette = "jco", rect = TRUE, rect_fill = TRUE, # Add rec
rect_border = "jco",           # Rectangle color
labels_track_height = 0.8      # Augment the room for labels
)
```



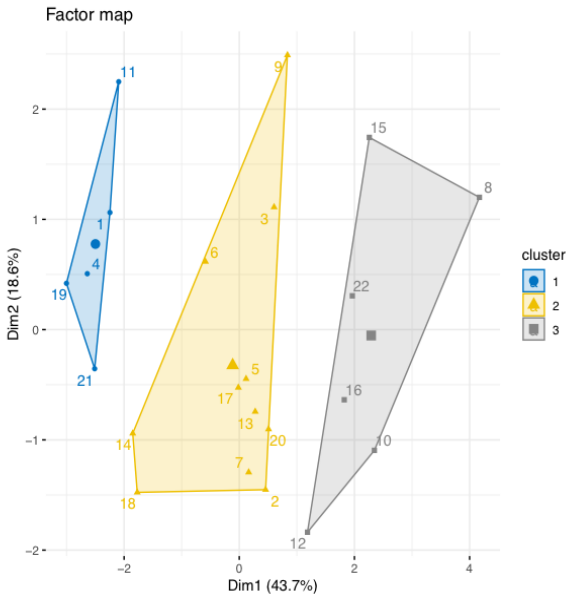
-le dendrogramme fournit une classification des éléments lorsque l'on se donne une " hauteur de coupe " de l'arbre.

-Plus l'arbre est coupé " bas " (proche des éléments initiaux) plus la classification obtenue est fine.

-Une hauteur de coupe est pertinente si elle se trouve entre 2 noeuds dont les hauteurs sont " relativement " éloignées.

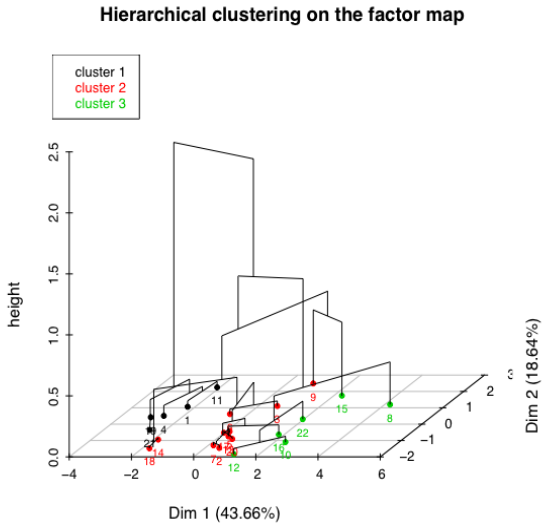
Les classes sur les plans factoriels

```
fviz_cluster(res,repel = TRUE, show.clust.cent = TRUE, # Show cluster centers
palette = "jco",ggtheme = theme_minimal(),main = "Factor map"
)
```



Représentation 3D

```
plot(res, choice = "3D.map")
```



Les variables qui décrivent le plus les classes

```
res$desc.var$quanti
```

```
$`1`
        v.test Mean in category Overall mean sd in category Overall sd      p.value
FRAN -2.170684             8.54     10.136364      0.5782733   1.827702 0.0299550899
PHYS -2.495730             4.68      8.531818      0.6584831   3.835645 0.0125698218
BIOL -2.841567             9.48     11.177273      0.7138627   1.484444 0.0044892359
HGEO -3.215045             9.94     12.231818      1.0423051   1.771591 0.0013042428
ECON -3.538904             7.10     11.181818      1.7435596   2.866521 0.0004017926

$`2`
       v.test Mean in category Overall mean sd in category Overall sd    p.value
ECON 2.034651         12.45455      11.18182       1.356039   2.866521 0.04188602

$`3`
       v.test Mean in category Overall mean sd in category Overall sd      p.value
PHYS 3.512895         13.33333      8.531818      2.4944383   3.835645 0.0004432518
MATH 3.279781         13.15000      9.322727      3.4932077   3.274687 0.0010388752
FRAN 2.784646         11.95000     10.136364      1.9128948   1.827702 0.0053586155
HGEO 2.589640         13.86667     12.231818      0.7086764   1.771591 0.0096076277
```

Les composantes qui sont le plus associées aux classes

```
res$desc.axes$quanti
```

```
$`1`
         v.test Mean in category  Overall mean sd in category Overall sd
Dim.3 -1.983503       -0.7818343 -2.144749e-16      0.8167187  0.9796075
Dim.1 -3.321711       -2.4978552  8.328250e-16      0.3162847  1.8688549
           p.value
Dim.3 0.0473113216
Dim.1 0.0008946726

$`2`
        v.test Mean in category  Overall mean sd in category Overall sd     p.value
Dim.3 3.100219        0.6627273 -2.144749e-16      0.6983835  0.9796075 0.001933775

$`3`
        v.test Mean in category Overall mean sd in category Overall sd      p.value
Dim.1 3.440261         2.291087  8.32825e-16      0.9207801   1.868855 0.0005811541
```

Les individus qui représent le plus les classes

```
res$desc.ind$para
```

```
Cluster: 1
        1        21        19         4        11
0.4151007 1.2579934 1.2856998 1.3128956 1.6773645
------------------------------------------------------------
Cluster: 2
        5        17        13        20         6
0.4601378 0.4797961 0.9079939 1.1169501 1.1572964
------------------------------------------------------------
Cluster: 3
      10       16       22       12        8
1.215457 1.621457 1.847866 2.244652 2.311409
```

K-means classif

```
data("seconde")
dataf <- scale(seconde)
head(dataf)
```

```
         HGEO        FRAN         PHYS         MATH       BIOL        ECON
1 -0.34843925 -0.76781614 -1.026977402 -0.513978617 -1.4330036 -1.59572385
```

```
2  0.75453392  1.15658381 -0.593956491 -0.245462084  0.5414899  0.10844725
3  0.53393929  1.04967270 -0.008104669 -0.901835833  0.2782241 -0.06196986
4 -1.89260169 -1.03509392 -0.899618310 -1.677550263 -0.3799404 -0.40280408
5  0.25819599 -0.82127170  0.373972605 -0.006780721 -0.3799404  0.96053280
6  0.09275002 -0.07289394 -0.772259219 -0.036615891 -0.4457569 -0.06196986
         ANGL       ESPA
1 -0.09358043  0.9032517
2 -1.32884217 -1.0449382
3  1.14168130  0.3466260
4  0.31817348  0.3466260
5 -0.50533435 -0.2099997
6 -0.09358043  0.6249388
```

Distance basée sur les corrélations

```
dist <- get_dist(dataf, method = "pearson")
```

visualisation

```
fviz_dist(dist, lab_size = 8)
```

K-means

```
km=eclust(dataf, "kmeans", nstart = 25)
km
```

```
K-means clustering with 1 clusters of sizes 22

Cluster means:
           HGEO          FRAN         PHYS        MATH          BIOL         ECON
1 -4.087639e-16 -1.709491e-16 1.211152e-16 1.51394e-16 -2.826022e-16 8.074349e-17
          ANGL         ESPA
1 2.624164e-16 1.539173e-16

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1

Within cluster sum of squares by cluster:
[1] 168
 (between_SS / total_SS =  -0.0 %)

Available components:

 [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
 [6] "betweenss"    "size"         "iter"         "ifault"       "clust_plot"
[11] "nbclust"      "data"         "gap_stat"
```