



Algèbre Linéaire et Analyse de Données

Licence 2 MIASHS (2021-2022)

Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Résumé

Cette deuxième séance a pour objectif de vous travailler sur l'Analyse en Composantes Principales (ACP) Dans un premier temps, nous allons, sur un exemple simple, effectuer manuellement les différentes transformations à effectuer sur les données pour mettre en oeuvre cette méthode. Nous regarderons ensuite comment effectuer cela rapidement sur  en utilisant les fonctions appropriées. On pourra ainsi comparer les résultats obtenus manuellement avec ceux obtenus avec . Enfin, on se propose également d'interpréter les résultats de cette ACP sur le jeu de données étudiée, l'objectif étant de synthétiser l'information et d'en avoir uen représentation visuelle "simple" et facile à interpréter.

1 Analyse en Composantes Principales

1.1 Introduction

Pour mettre en oeuvre l'ACP sur  ainsi que l'interprétation de ces résultats, nous allons travailler sur un jeu de données qui donne des informations sur *le passage des jeunes du système éducatif au travail*, i.e. notre étude portera sur des données Sociologiques.

Le cas qui nous intéresse aujourd'hui est extrait de la référence suivante : D. Busca et S. Toutain, *Analyse factorielle simple en sociologie. Méthodes d'interprétation et étude de cas*, 2009, De Coeck.

Les données consistent en l'étude de 5 critères sur un de 9 pays.

Les 9 pays étudiés sont l'Autriche, la Belgique, l'Espagne, la Finlande, la France, la Grèce, la Hongrie, l'Italie et la Suède.

Enfin les 5 critères/variables étudiées sont les suivants :

- **V1_1** : l'âge moyen lors de la sortie du système éducatif de la population active (15-64 ans)
- **V1_2** : l'âge moyen d'obtention d'un niveau de formation primaire ou secondaire (collège) lors de la sortie du système éducatif
- **V1_4** : l'âge moyen d'obtention d'un niveau de formation premier ou deuxième cycle de l'enseignement supérieur (licence ou master) lors de la sortie du système éducatif
- **V3_1** : pourcentage de parents ayant terminé un niveau de formation primaire ou secondaire¹
- **V3_3** : pourcentage de parents ayant terminé un niveau de formation premier ou deuxième cycle de l'enseignement supérieur.

Nous serons, par la suite, amenés à renommer ces variables comme suit : *age_moy*, *age_moy_ps*, *age_moy_sup*, *pc_par_ps*, *pc_par_sup* afin d'identifier clairement les variables étudiées mais aussi pour nous aider à interpréter les résultats dans la suite.

Les données se trouvent ci-dessous, vous avez simplement à copier-coller le code

```
#### Création du jeu de données ####  
  
# Variables  
  
V1_1=c(19.9,20.6,19.1,21.6,20.8,19.4,18.3,18.4,23.9)
```

1. Sous-entendu, sans avoir terminé un niveau de formation supérieur (universitaire). Il s'agit donc de parents ne disposant de "hauts" diplômes.

```

V1_2=c(18,18.3,15.2,16.9,17.9,14.5,14.9,14.6,22.6)
V1_4=c(24.7,22.8,22.5,25.3,23.2,23.3,23.1,25.0,26.4)
V3_1=c(27,45,80,21,51,66,26,68,26)
V3_3=c(19,26,10,36,15,9,13,6,36)

# Nom des individus

liste_obs=c("Autriche","Belgique","Espagne","Finlande",
            "France","Grèce","Hongrie","Italie","Suède")
liste_var=c("age_moy","age_moy_ps","age_moy_sup",
            "pc_par_ps","pc_par_sup")

# Enregistrement des données dans une base

# save(V1_1,V1_2,V1_4,V3_1,V3_3,liste_obs,liste_var,file="europe.RData")

```

1.2 Préparation des données

Cette première partie se concentre sur la préparation et la transformation des données. Cette étape là est essentielle lorsque l'on souhaite réaliser notre ACP manuellement.

1. Stockez votre jeu de données, *i.e.* les 5 vecteurs qui contiennent les informations relatives aux 5 variables dans une variable que l'on notera D .

```

D = cbind(V1_1,V1_2,V1_4,V3_1,V3_3)
colnames(D) = liste_var

```

2. Entrer et exécuter les commandes suivantes :

```

mean(D[1,])
## [1] 21.72

apply(D,1,mean)
## [1] 21.72 26.54 29.36 24.16 25.58 26.44 19.06 26.40 26.98

mean(D[,1])
## [1] 20.22222

m=apply(D,2,mean)

```

Que représentent ces différentes quantités et notamment m ?

m est le vecteur du barycentre du nuage de points.

- Déterminer à partir de D le nombre d'individus et le nombre de variables à l'aide des commandes `nrow` et `ncol`. Ces informations là seront stockées dans les variables n et p .

```
# nombre de lignes
n = nrow(D)
# nombre de colonnes
p = ncol(D)
```

- Entrer et exécuter les commandes suivantes :

```
# vecteur des écart-types des variables
apply(D,2,sd)

##      age_moy  age_moy_ps age_moy_sup   pc_par_ps  pc_par_sup
##  1.766195   2.610768   1.349074   21.938044   11.340684

# estimation non biaisée
s = apply(D,2,sd)
# estimation biaisée
s = s*sqrt((n-1)/n)
```

Que représente s ?

s est un vecteur contenant l'écart-type **biaisé** des différentes variables.

Remarque : la fonction `sd` (standard deviation) est l'estimateur sans biais de l'écart-type d'une variable. Dans le cas l'ACP, on utilise plutôt l'estimateur biaisé. C'est la raison pour laquelle nous effectuons l'opération $\sigma \sqrt{\frac{n-1}{n}}$.

- Nous allons à présent centrer, réduire et diviser par \sqrt{n} le terme général de la matrice D et nous allons stocker la nouvelle matrice X dans une variable X . Ce que l'on peut faire avec la commande suivante

```
X = scale(D,center = m, scale = s)/sqrt(n)
```

Vérifier que le barycentre des individus de X est le vecteur nul. Vérifier également la norme des vecteurs colonnes de X vaut 1. Pour cela, utiliser la commande `apply`.

```

# Barycentre des individus dans X
apply(X,2,mean)

##      age_moy  age_moy_ps  age_moy_sup  pc_par_ps  pc_par_sup
## 1.572756e-16 2.197316e-16 -4.009139e-16 -2.023242e-17 -2.164188e-17

# Norme des colonnes
apply(X^2,2,sum)

##      age_moy  age_moy_ps  age_moy_sup  pc_par_ps  pc_par_sup
##           1           1           1           1           1

```

6. Pour que la table de données soit riche en information, nous pouvons donner des noms aux lignes et colonnes d'une matrice. Entrer et exécuter les commandes suivantes :

```

rownames(X) = liste_obs
colnames(X) = liste_var

```

1.3 Analyse du nuage des individus

1. A partir de X , stocker dans la variable C , la matrice des corrélations des variables.

```
C=t(X)%*%X
```

2. Procéder à la décomposition en éléments propres de C et stocker le résultat de cette décomposition dans la variable $C.eigen$.

```
C.eigen=eigen(C)
```

3. Entrer et exécuter la commande suivante :

```

C.eigen$values
## [1] 3.66779589 0.60428535 0.51885166 0.17559116 0.03347594
sort(C.eigen$values)
## [1] 0.03347594 0.17559116 0.51885166 0.60428535 3.66779589

```

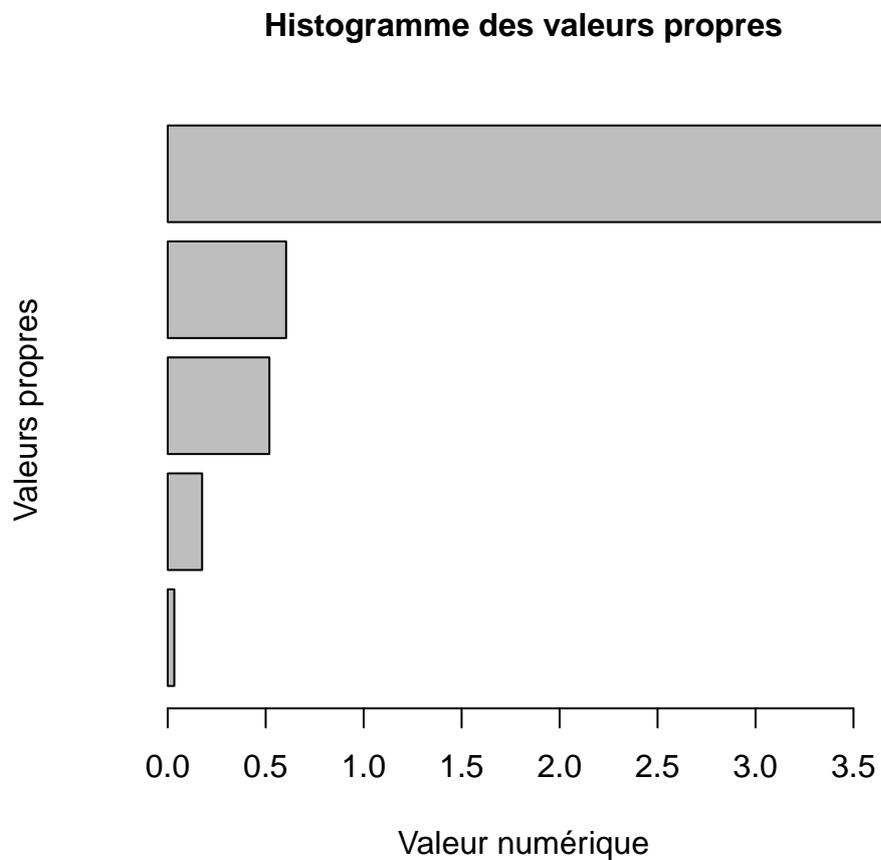
Que fait la commande *sort* ?

Cette commande va trier les valeurs propres par ordre croissant. Noter que vous pouvez également préciser que vous souhaitez trier les valeurs propres par ordre décroissant de la façon suivante :

```
sort(C.eigen$values, decreasing = TRUE)
## [1] 3.66779589 0.60428535 0.51885166 0.17559116 0.03347594
```

4. Entrer et exécuter la commande suivante :

```
# Histogramme des valeurs propres
barplot(sort(C.eigen$values),horiz=TRUE,
        main="Histogramme des valeurs propres",
        xlab="Valeur numérique", ylab="Valeurs propres ",
        cex.lab=1, cex.axis=1, cex.main=1)
```



Ceci est une commande graphique qui comporte plusieurs paramètres. A l'aide des noms de ces paramètres et en modifiant les valeurs de ces derniers, essayer de comprendre leur rôle. Vous pourrez également consulter l'aide pour vous aider.

5. Combien d'axes principaux proposez-vous de garder, au regard du graphique précédent.

Ici on se propose de garder les deux premiers axes principaux qui permettent de conserver plus de 90% de la variance initialement présente dans notre jeu de données.

On aurait presque pu être tenté de ne conserver que le premier axe principal mais cela aurait grandement limité le reste de l'étude.

6. Stocker dans deux variables $u1$ et $u2$, les deux premiers axes principaux.

```
#Axes principaux  
u1=C.eigen$vector[,1]  
u2=C.eigen$vector[,2]
```

7. Calculer les composantes principales associées aux deux premiers axes principaux. Il s'agit des coordonnées des individus sur ces deux axes. Vous stockerez ces deux vecteurs dans les variables $f1$ et $f2$.

```
#Composantes principales  
f1=X%*%u1  
f2=X%*%u2
```

8. Calculer la somme des valeurs propres de C que vous stockerez dans une variable $sum.eig$.

```
# Inertie totale  
sum.eig=sum(C.eigen$values)  
sum.eig  
  
## [1] 5
```

A quoi correspond cette valeur ?

On rappelle que la somme des valeurs propres est égale à la somme des variances des différentes variables. Ici toutes les variables sont centrées et réduites, donc de variance 1, ce qui explique que cette somme soit égale à 5 soit le nombre de variables.

9. Calculer le pourcentage de l'inertie associée à l'axe $u1$. Il s'agit de la valeur propre associée à cet axe divisé par l'inertie totale. Faites de même pour l'axe principal $u2$. Le premier plan principal est l'espace engendré par $u1$ et $u2$. Le pourcentage de l'inertie expliquée par ce plan factoriel (ou plan principal) est la somme des inerties de chaque axe le constituant. Quel est le pourcentage d'information que contient ce premier plan factoriel ?

```

#Pourcentage de l'inertie sur le premier axe
u1.int=C.eigen$values[1]/sum.eig
u1.int

## [1] 0.7335592

# Pourcentage d'inertie sur le deuxième axe
u2.int=C.eigen$values[2]/sum.eig
u2.int

## [1] 0.1208571

# Pourcentage d'inertie sur le plan factoriel
u1.int+u2.int

## [1] 0.8544162

```

10. Stocker dans la variable F , la matrice F qui comporte dans ses deux colonnes, les coordonnées des individus sur les axes $u1$ et $u2$. Donner des noms aux lignes de F . Les noms des colonnes de F seront $u1$ et $u2$.

```

# Plan factoriel
F=cbind(f1,f2)
rownames(F)=liste_obs
colnames(F)=c("u1","u2")

```

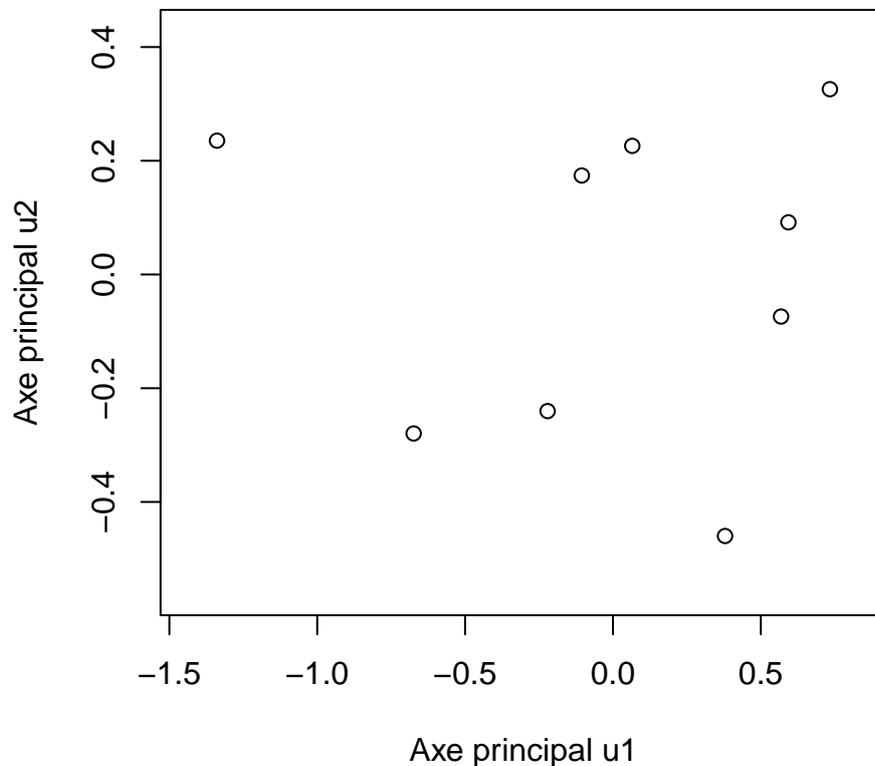
11. Entrer et exécuter la commande suivante :

```

#Représentation graphique
plot(F,xlab = "Axe principal u1",ylab = "Axe principal u2",
     main = "Plan principal (u1,u2)",
     xlim = c(min(f1)-0.1,max(f1)+0.1),
     ylim=c(min(f2)-0.1,max(f2)+0.1),
     cex.lab=1,cex.axis=1,cex.main =1)

```

Plan principal (u1,u2)



Remarque : *plot* est une commande graphique utilisée pour représenter un nuage de points dans un repère orthonormé. Dans la commande précédente, les points sont des lignes de F qui sont des vecteurs dont les composantes sont données par les colonnes de F (qui sont donc les éléments de la base qui est ici de dimension 2).

Que font les paramètres *xlim* et *ylim* ?

Ils permettent de modifier les valeurs min et max des axes des abscisses et ordonnées.

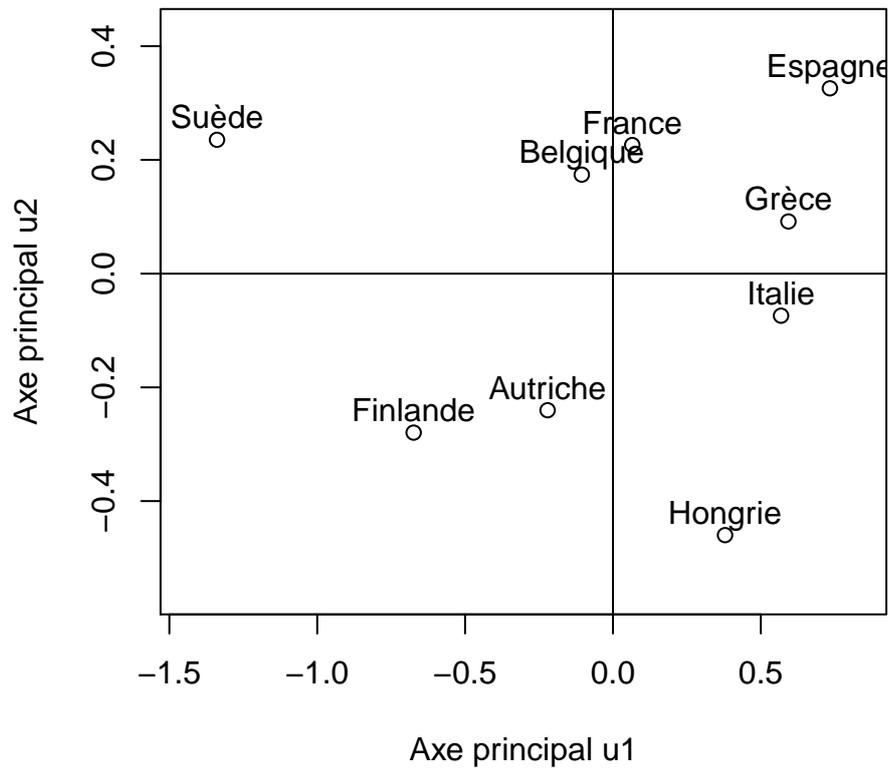
12. Entrer et exécuter l'une après l'autre les commandes suivantes :

```
#Représentation graphique  
plot(F,xlab = "Axe principal u1",ylab = "Axe principal u2",  
      main = "Plan principal (u1,u2)",  
      xlim = c(min(f1)-0.1,max(f1)+0.1),
```

```
ylim=c(min(f2)-0.1,max(f2)+0.1),
cex.lab=1,cex.axis=1,cex.main =1)

text(F,labels=rownames(F),pos=3,cex=1,offset=0.3)
abline(h=0)
abline(v=0)
```

Plan principal (u1,u2)



Expliquer ce que font ces différentes commandes.

Les fonctions *abline* permettent de tracer une droite horizontale qui représentera l'axe des abscisses ($h=0$) ou une droite verticale qui représentera l'axe des ordonnées ($v=0$). Enfin, la fonction *text* permet d'attribuer les noms données aux lignes de notre *F* aux différents points qui forment notre nuage.

1.4 Analyse du nuage des variables

1. Calculer la matrice des produits scalaires entre individus. Cette matrice sera stockée dans la variable K .

```
K = X%*%t(X)
```

2. Procéder à la décomposition en éléments propres de K et stocker les résultats dans la variable $K.eigen$. Observer les valeurs propres de K et commenter.

```
# Décomposition en valeurs/vecteurs propres
K.eigen=eigen(K)
K.eigen$values

## [1] 3.667796e+00 6.042854e-01 5.188517e-01 1.755912e-01 3.347594e-02
## [6] 1.723936e-16 6.746031e-17 1.995615e-17 1.054433e-18

#Formules de transition
K.eigen$vectors[,1]

## [1] 0.11542607 0.05505589 -0.38304884 0.35191927 -0.03409321 -0.30975071
## [7] -0.19788931 -0.29672970 0.69911054

f1/sqrt(C.eigen$values[1])

##           [,1]
## Autriche -0.11542607
## Belgique -0.05505589
## Espagne  0.38304884
## Finlande -0.35191927
## France   0.03409321
## Grèce    0.30975071
## Hongrie  0.19788931
## Italie   0.29672970
## Suède    -0.69911054
```

3. Stocker dans deux variables $v1$ et $v2$ les deux premiers axes factoriels $v1$ et $v2$.

```
# Axes factoriels
v1=K.eigen$vectors[,1]
v2=K.eigen$vectors[,2]
```

Le spectre des valeurs propres de K est le même que le spectre des valeurs propres de F .

4. Calculer puis stocker dans les variables $g1$ et $g2$ les coordonnées des vecteurs variables sur les deux premiers axes factoriels

```

#Composantes des variables sur les axes factoriels v
g1=t(X)%*%v1
g1=ifelse(sum(g1*u1)<0,-1,1)*g1
g2=t(X)%*%v2
g2=ifelse(sum(g2*u2)<0,-1,1)*g2

# Vérifications de propriétés
mean(g1)

## [1] -0.5495554

sum(g1^2)

## [1] 3.667796

# Plan factoriel
G=cbind(g1,g2)
rownames(G)=liste_var
colnames(G)=c("v1", "v2")

```

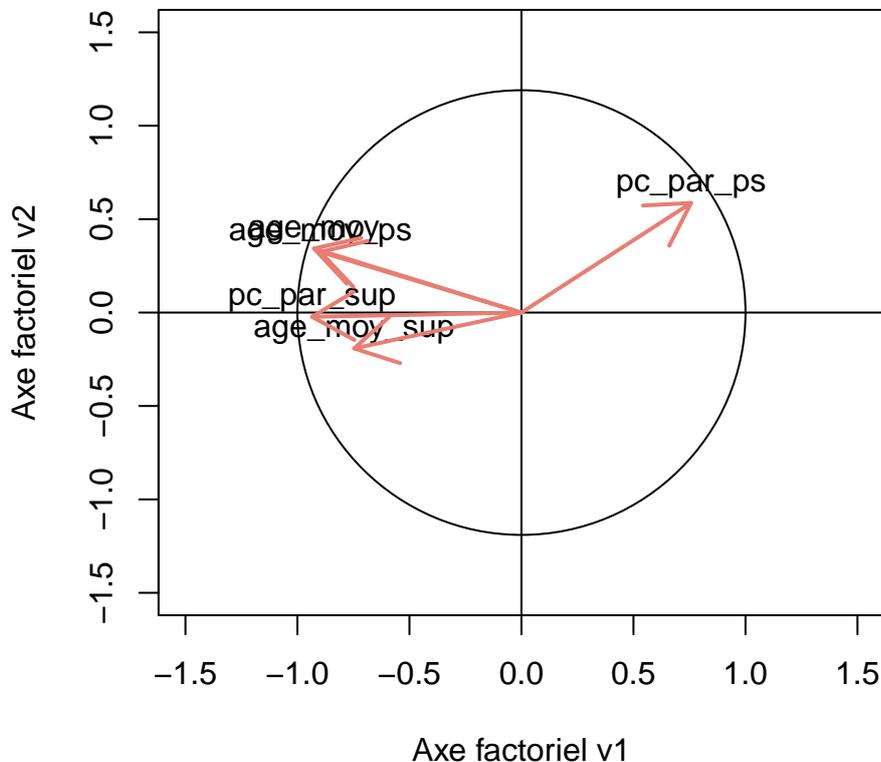
5. Faire un graphique représentant les vecteurs variables dans le premier plan factoriel. Vous donnerez des titre et légendes adéquats aux axes.

```

plot(NA,xlab = "Axe factoriel v1",ylab = "Axe factoriel v2",
main = "Cercle des corrélations (v1,v2)",
xlim = c(-1.5,1.5),ylim=c(-1.5,1.5),
cex.lab=1,cex.axis=1,cex.main=1)
text(G,labels=rownames(G),pos=3,cex=1,offset=0.3)
abline(h=0)
abline(v=0)
for (i in 1:nrow(G)){
  arrows(0,0,G[i,1],G[i,2],col="#e97d72", lwd = 2)
}
symbols(0, 0, circles=1, inches=FALSE, add=TRUE)

```

Cercle des corrélations (v1,v2)



6. Ajouter les noms des variables au graphique précédent ainsi que des droites représentant l'axe des abscisses et celui des ordonnées.

Voir code précédent

1.5 Interprétation des corrélations variables-variables et variables-axes factoriels

A partir de cette section, nous allons mettre en oeuvre des calculs permettant d'interpréter plus robustement les résultats de l'ACP que nous pouvons appréhender visuellement à l'aide des 2 graphiques précédents qui doivent être analysés conjointement. Dans cette section en particulier, il s'agit d'interpréter les axes factoriels du nuage des variables. S'agissant d'une ACP qui est normée, nous avons dans ce cas un cercle de corrélations : les coordonnées des variables sur les axes sont des corrélations linéaires, elles sont donc dans un cercle de rayon 1 ! In fine, nous souhaitons associer des groupes de variables de part et d'autre de chaque axe afin de leur donner une certaine sémantique.

Nous allons restreindre notre étude au premier plan factoriel, *i.e.* à l'espace de dimension 2 engendré par $v1$ et $v2$. Néanmoins, ce qui suit pourra être également appliqué à d'autres plans factoriels tel que celui engendré par $(v1, v3)$ par exemple.

1. Etudier la matrice des coefficients de corrélation (variable C déjà calculée) et identifier les variables qui sont corrélées positivement et celles qui sont corrélées négativement.

Les variables age_moy et pc_par_sup sont fortement linéairement corrélées. Les pays (individus de notre étude) où la part des parents ayant terminés un niveau de formation 1^{er} ou 2^{me} cycle universitaire est importante ont tendance à avoir un âge moyen d'obtention d'un niveau de formation primaire ou secondaire élevé (et inversement). Typiquement, il s'agit de pays où le système éducatif est long (on sort âgé du système éducatif) et où la proportion des parents qui ont un diplôme universitaire est plus grande...

2. Quelle est la mesure de corrélation entre age_moy et pc_par_sup ? Comment expliqueriez vous la corrélation entre ces deux variables?

Voir question précédente.

Mêmes question pour le coupe de variables age_moy_ps et pc_par_ps .

3. Que représentent les coordonnées des différentes variables sur les axes factoriels $v1$ et $v2$?

Dans une ACP normée les coordonnées des variables sur les axes principaux sont égales aux coefficients de corrélation linéaire entre les variables et les axes.

4. Quelles sont les variables fortement corrélées à l'axe $v1$? A l'axe $v2$?

Ainsi, les variables les plus fortement corrélées à l'axe 1 sont celles ayant, en valeur absolue, les coordonnées les plus grandes sur le *s.e.v.* engendré par $v1$ (1^{er} axe factoriel). Ici s'il fallait prendre les 3 variables les plus fortement corrélées, il s'agirait de pc_par_sup , age_moy et age_moy_ps . Ces variables sont corrélées négativement à l'axe 1. Pour l'axe $v2$ on a essentiellement la variable pc_par_ps qui est corrélée positivement : lorsque les valeurs de l'axe 2 augmentent, les valeurs de pc_par_ps ont tendance à augmenter également ET vice-versa.

5. Au travers du positionnement des variables pc_par_ps et pc_par_sup vis à vis de l'axe $v1$, que pouvez-vous dire de la corrélation entre ces deux variables?

pc_par_ps est fortement corrélée positivement à l'axe 1 tandis que pc_par_sup lui est fortement corrélée négativement. Par transitivité, on peut supposer que ces 2 variables sont corrélées négativement. C'est effectivement ce que l'on peut constater en regardant la valeur de leur coefficient de corrélation linéaire. Du point de vue de l'ADD, cela laisse sous-entendre que l'axe 1 va notamment opposer les pays ayant un pc_par_ps fort (à droite de l'axe 1 - côté positif de l'axe) et ceux qui ont un pc_par_sup fort (à gauche de l'axe 1 - côté négatif de l'axe).

6. Que pourriez-vous dire à propos d'une variable hypothétique y dont les composantes sur les axes $v1$ et $v2$ seraient $(0.2, -0.1)$?

Une telle variable hypothétique, y , est proche de l'origine du repère. Elle est donc faiblement corrélée linéairement aux axes donc on se gardera de l'interpréter dans l'analyse.

7. Si vous deviez associer à chaque axe et à chacune de ses parties positives et négatives un groupe de variables, que proposeriez vous ?

Clairement pour l'axe 1 on a du côté positif la variable pc_par_ps et du côté négatif toutes les autres variables. Pour l'axe 2, l'unique variable que l'on peut lui associer est pc_par_ps vis à vis de son côté positif (en haut). Aucune variable ne peut vraiment être associée au côté négatif de l'axe 2. Il est cependant important de mentionner que la corrélation forte d'une variable à un axe à une double interprétation : dans le cas de pc_par_ps et l'axe 2, la partie positive est corrélée positivement à pc_par_ps **mais** la partie négative est également corrélée négativement à pc_par_ps . Autrement dit, et en gardant à l'esprit les relations de dualités entre $v2$ et $u2$ (et donc $f2$), les pays qui ont une composante positive sur l'axe 2 ont un pc_par_ps qui ont tendance à être grand TANDIS que les pays ayant une composante négative sur l'axe 2 ont un pc_par_sup qui ont tendance à être faible.

1.6 Interprétation de la position des individus et des distances individus-individus dans le premier plan principal

Dans cette section, nous nous intéressons au nuage des individus projetés sur le premier plan principal. Dans ce cas, nous pouvons robustifier l'interprétation par le calcul de mesures de qualité et de contributions. Comme précédemment, nous cherchons à caractériser les axes principaux en associant des groupes d'individus à chaque partie positive et négative d'un axe. Ces oppositions traduit de façon synthétique cette notion d'axes le long desquels le nuage des individus s'étend le plus. Il est important de garder à l'esprit que les interprétations que nous faisons sont relatives au barycentre dans le

mesure ou lors du centrage nous positionnons l'individu moyen au centre du repère. Autrement dit, lorsque nous disons qu'un groupe d'individus a tendance à avoir des valeurs élevées (ou faibles) pour un groupe de variables, c'est par rapport à l'individu moyen.

1. Dans la fenêtre des graphiques, revenez sur la projection des individus sur le premier plan principal. A première vue, quels groupes de pays l'axe principal $u1$ oppose-t-il ?

A première vue, l'axe 1 oppose d'un côté l'Espagne, la Grèce et l'Italie et de l'autre la Suède et la Finlande.

Même question pour l'axe principal $u2$.

L'axe 2 oppose d'un côté l'Espagne et de l'autre la Hongrie.

2. Il faut compléter la visualisation des groupes par le calcul de la qualité des individus afin de privilégier les éléments les plus pertinents. Pour ce faire, calculer la qualité de la représentation de chaque individu sur l'axe $u1$. Vous stockerez le résultat dans la variable $q1t.u1$.

```
# Qualité de la représentation des
# individus sur les axes principaux
q1t.u1 = f1^2/apply(X^2,1,sum)
q1t.u1

##           [,1]
## Autriche 0.34203160
## Belgique 0.05822471
## Espagne  0.82091252
## Finlande 0.72380191
## France   0.04319693
## Grèce    0.92308683
## Hongrie  0.34121604
## Italie   0.54350476
## Suède    0.94950834
```

Quels sont les deux pays les moins bien représentés sur ce premier axe principal ?

Les 2 pays les moins bien représentés sur $u1$ sont la France et la Belgique tandis que les 2 pays les mieux représentés sont la Grèce et la Suède.

3. Calculer la qualité de la représentation de chaque individu sur l'axe $u2$. Vous stockerez le résultat dans la variable $q2t.u2$.

```

# Qualité de la représentation sur le deuxième axe
qlt.u2 = f2^2/apply(X^2,1,sum)
qlt.u2

##           [,1]
## Autriche 0.403716264
## Belgique 0.158750420
## Espagne  0.162056003
## Finlande 0.124518527
## France   0.518650254
## Grèce    0.022084394
## Hongrie  0.502263031
## Italie   0.009199026
## Suède    0.029350797

```

Quels sont les deux pays les moins bien représentés sur ce deuxième axe principal ?

Les deux pays les moins bien représentés sont l'Italie et la Grèce et les deux pays les mieux représentés sont la France et la Hongrie.

En pratique on peut décider que les individus dont les contributions sont supérieures à la moyenne ou à la médiane sont significativement représentés sur un axe.

```

# Médiane des contributions
median(qlt.u2)

## [1] 0.1587504

```

4. Dans la section précédente, nous avons cherché à associer des variables aux axes factoriels. Nous pouvons faire de même en ce qui concerne les individus en regardant la mesure de contribution de chacun d'entre eux pour la construction des axes principaux. Calculer les contributions des individus aux axes $u1$ et $u2$. Vous stockerez ces résultats dans les variables $ctr.u1$ et $ctr.u2$ respectivement.

```

# Contribution des individus à chaque axe
ctr.u1 = f1^2/C.eigen$values[1]
ctr.u2 = f2^2/C.eigen$values[2]

# Qualité de la représentation du nuage des
# individus sur les axes principaux
qlt.u1.NO = C.eigen$values[1]/p
qlt.u2.NO = C.eigen$values[2]/p

```

5. Comme précédemment, la significativité d'un élément peut être appréciée en comparant la valeur de sa contribution vis à vis d'une tendance centrale telle que la moyenne ou la médiane. Déterminer les individus qui contribuent le plus à l'axe u_1 puis ceux qui contribuent le plus à l'axe u_2 en vous basant sur la médiane.

```
# Contribution des individus à chaque axe
```

```
ctr.u1 = f1^2/C.eigen$values[1]
```

```
ctr.u1
```

```
##           [,1]
```

```
## Autriche 0.013323177
```

```
## Belgique 0.003031151
```

```
## Espagne  0.146726412
```

```
## Finlande 0.123847172
```

```
## France   0.001162347
```

```
## Grèce    0.095945505
```

```
## Hongrie  0.039160180
```

```
## Italie   0.088048512
```

```
## Suède    0.488755546
```

```
ctr.u2 = f2^2/C.eigen$values[2]
```

```
ctr.u2
```

```
##           [,1]
```

```
## Autriche 0.095451090
```

```
## Belgique 0.050162397
```

```
## Espagne  0.175808405
```

```
## Finlande 0.129319326
```

```
## France   0.084707240
```

```
## Grèce    0.013932551
```

```
## Hongrie  0.349872289
```

```
## Italie   0.009045313
```

```
## Suède    0.091701389
```

```
median(ctr.u1)
```

```
## [1] 0.08804851
```

```
median(ctr.u2)
```

```
## [1] 0.09170139
```

Il suffit de regarder les individus dont la contribution est supérieure à la valeur médiane.

Pour l'axe u_1 nous avons : l'Espagne, la Finlande, La Grèce et la Suède et l'Italie

Pour l'axe u_2 nous avons : la Suède, l'Autriche, l'Espagne, la Finlande et la Hongrie.

```
# Contribution des individus au plan
ctr.u1u2 = ctr.u1+ctr.u2
ctr.u1u2

##           [,1]
## Autriche 0.10877427
## Belgique 0.05319355
## Espagne  0.32253482
## Finlande 0.25316650
## France   0.08586959
## Grèce    0.10987806
## Hongrie  0.38903247
## Italie   0.09709383
## Suède    0.58045693
```

6. Si vous deviez proposer à chaque axe et à chacune de ses parties positives et négatives un groupe d'individus, que proposeriez-vous ?

A la suite des premières observations et des calculs complémentaires précédents, nous pouvons dire que l'axe 1 oppose d'un côté l'Espagne, la Grèce et de l'autre côté la Suède et la Finlande. En ce qui concerne l'axe 2, il met en opposition d'une part, l'Espagne et la France du côté positif et d'autre part la Hongrie et l'Autriche du côté négatif.

7. Interpréter les axes en utilisant conjointement les analyses deux deux nuages de points.

En complétant l'analyse faite à la question précédente avec celle effectuée sur le nuage des variables, nous pouvons dire de plus que l'Espagne et la Grèce sont des pays où la part des parents ayant terminé un niveau de formation primaire ou secondaire élevé et une part des parents ayant terminé un niveau de formation de 1^{er} ou 2^{me} cycle universitaire faible. Il s'agit également de pays où l'âge de sortie du système éducatif (tout niveau confondu) est faible. A l'opposé de l'Espagne et de la Grèce on retrouve en particulier des pays nordiques, la Suède et la Finlande où l'âge moyen de sortie du système éducatif (tout niveau confondu) est plus élevé et où la part des parents ayant terminé un niveau de formation dans le supérieur est particulièrement grand. Concernant le deuxième axe, nous pouvons dire que la Hongrie et l'Autriche sont des pays dont la part des parents ayant terminé un niveau de formation primaire ou secondaire est faible contrairement à l'Espagne et la France qui, par conséquent, ont une part de parents avec un "faible" niveau de diplôme plus grande. Nous

pouvons également observer des individus qui sont bien représentés à la fois pour le premier et le deuxième axe. Il s'agit de l'Espagne et de la Finlande. Le premier a une part de parents ayant terminé un niveau de formation primaire ou secondaire particulièrement fort et un âge moyen de sortie du système éducatif (tout niveau confondu) jeune. A l'opposé la Finlande se caractérise par une faible part de parents ayant un "faible" niveau de formation. Pour ce pays, il y a un fort taux de parents diplômés de l'enseignement supérieur et l'âge de sortie du système éducatif est grand pour tout type de diplôme.

1.7 Ajouts d'individus fictifs supplémentaires sur le premier plan principal

1. Stocker dans une variable Tp la matrice T_+ de taille 4×5 dont les lignes sont les individus fictifs suivants :

$$t_1 = \begin{pmatrix} 24 \\ 20 \\ 26 \\ 10 \\ 70 \end{pmatrix}, \quad t_2 = \begin{pmatrix} 17 \\ 15 \\ 20 \\ 50 \\ 30 \end{pmatrix}, \quad t_3 = \begin{pmatrix} 21 \\ 19 \\ 25 \\ 70 \\ 10 \end{pmatrix} \quad \text{et} \quad t_4 = \begin{pmatrix} 19 \\ 17 \\ 24 \\ 20 \\ 20 \end{pmatrix}$$

On remarquera par exemple que pour le pays fictif t_1 , l'âge moyen de sortie du système éducatif (24 ans), l'âge moyen d'obtention d'un diplôme du secondaire (20 ans) lors de la sortie du système éducatif et l'âge moyen d'obtention d'un diplôme du supérieur (26 ans) lors de la sortie du système éducatif, sont élevés signifiant que ces individus sortent "âgés" du système éducatif peu importe le niveau d'étude. Par ailleurs, pour ce même pays fictif, le pourcentage des parents ayant un diplôme du primaire ou secondaire (10%) est très bas alors que le pourcentage des parents ayant un diplôme du supérieur (70%) est très élevé ce qui indique que ces individus ont des parents ayant eux même fait des études longues.

```
t1 = c(24, 20, 26, 10, 70)
t2 = c(17, 15, 20, 50, 30)
t3 = c(21, 19, 25, 70, 10)
t4 = c(19, 17, 24, 20, 20)
```

```
Tp = rbind(t1, t2, t3, t4)
```

2. Transformer la matrice T_+ précédente en centrant, réduisant selon les statistiques estimées sur la matrice D et en divisant par n les différentes valeurs afin d'obtenir une matrice X_+ que vous stockerez dans Xp . Pour cela, vous pourrez utiliser la commande *scale* vue précédemment mais avec les mêmes variables m et s estimées sur la population originale.

```
# Centrage et réduction des individus
Xp=scale(Tp,center = m,scale = s)/sqrt(n)
```

3. Déterminer la projection des 4 nouveaux individus sur le premier plan principal. Vous stockerez les composantes principales de ces individus dans les variables *fp1* et *fp2*.

```
# Projection des individus sur les axes u1 et u2
fp1=Xp%*%u1
fp2=Xp%*%u2

Fp=cbind(fp1,fp2)
```

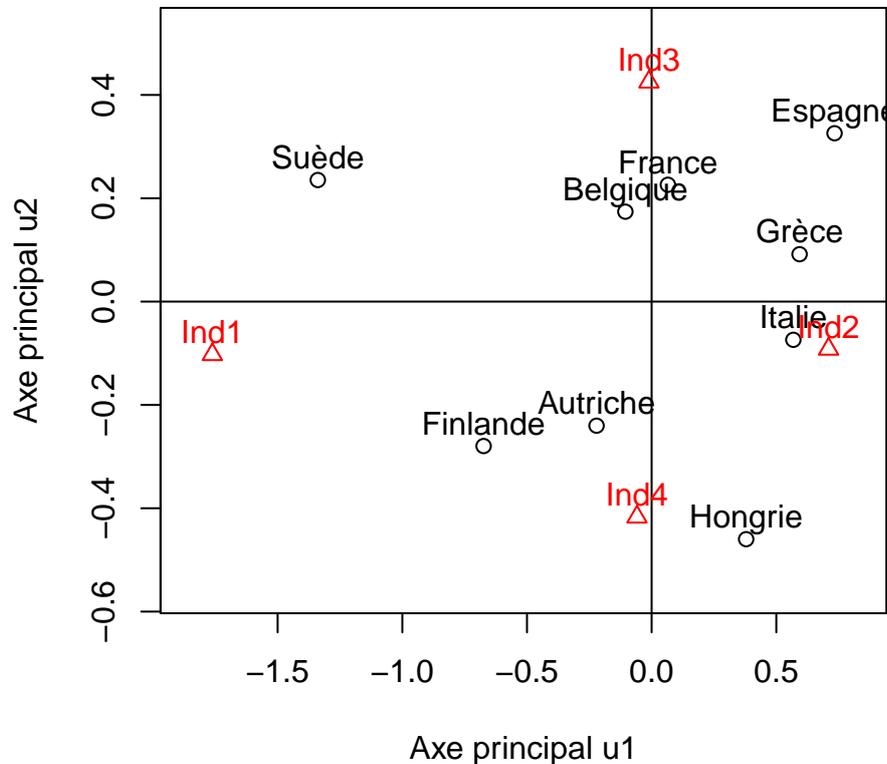
4. Représentez ces individus supplémentaires sur le premier plan principal en exécutant les commandes suivantes :

```
# Représentation graphique
plot(F,xlab = "Axe principal u1",ylab = "Axe principal u2",
     main = "Plan principal (u1,u2)",
     xlim = c(min(c(f1,fp1))-0.1,max(c(f1,fp1))+0.1),
     ylim=c(min(c(f2,fp2))-0.1,max(c(f2,fp2))+0.1),
     cex.lab=1, cex.axis=1, cex.main=1)

text(F,labels=rownames(F),pos=3,cex=1,offset=0.3)
abline(h=0)
abline(v=0)

# Ajout des nouveaux points
points(fp1,fp2, pch = 2, col="red")
text(Fp,labels=c("Ind1","Ind2","Ind3","Ind4"),
     pos=3,cex=1,offset =0.3,col="red")
```

Plan principal (u1,u2)



5. Interprétez à nouveau les axes principaux à l'aide de ces nouveaux éléments.

A vous de jouer !

1.8 Exemple d'interprétation

Les interprétations issues de la référence² d'où est extraite cette étude sont présentées ci-après :

Par rapport à l'axe 2 : Cet axe décrit la population des pays au regard de la proportion de parents peu diplômés. Il identifie 3) des pays comme la Hongrie, l'Autriche,

2. D. Busca et S. Toutain, *Analyse factorielle simple en sociologie. Méthodes d'interprétation et étude de cas*, 2009, De Coeck

et la Finlande caractérisés par une faible proportion de parents peu diplômés.

Par rapport au plan 1-2 : Le plan P1-2 identifie un pays, l'Espagne, caractérisé par des actifs ayant un âge précoce de sortie du système éducatif (tous niveaux de diplômes confondus), une faible part de parents diplômés du premier et du deuxième cycle de l'enseignement supérieur, et une proportion importante de parents ayant un niveau de formation primaire ou de premier cycle du secondaire. En parallèle, la Finlande est caractérisée par des actifs ayant un âge élevé de sortie du système éducatif (quel que soit le niveau de formation), une part élevée de parents diplômés de l'enseignement supérieur et une moindre proportion de parents peu diplômés. *Par rapport à l'axe 1 : L'analyse du cercle des corrélations souligne que l'axe 1 synthétise la relation entre l'âge moyen de sortie du système éducatif des niveaux de formation les plus faibles aux plus élevées, et le pourcentage de parents avec un niveau de formation élevé. e.*

Il oppose (i) les pays comme la Suède ou la Finlande caractérisés par des actifs ayant un âge élevé de sortie du système éducatif (quelque soit le niveau de formation), une part élevée de parents diplômés de l'enseignement supérieur et une moindre proportion de parents avec un faible niveau de diplôme, (ii) aux pays comme la Grèce, l'Italie ou l'Espagne marqués par une proportion élevée de parents peu diplômés, un âge plus précoce de sortie du système éducatif et une part plus faible de parents diplômés de l'enseignement supérieur.

Par rapport à l'axe 2 : Cet axe décrit la population des pays au regard de la proportion de parents peu diplômés. Il identifie 3) des pays comme la Hongrie, l'Autriche, et la Finlande caractérisés par une faible proportion de parents peu diplômés.

Par rapport au plan 1-2 : Le plan P1-2 identifie un pays, l'Espagne, caractérisé par des actifs ayant un âge précoce de sortie du système éducatif (tous niveaux de diplômes confondus), une faible part de parents diplômés du premier et du deuxième cycle de l'enseignement supérieur, et une proportion importante de parents ayant un niveau de formation primaire ou de premier cycle du secondaire. En parallèle, la Finlande est caractérisée par des actifs ayant un âge élevé de sortie du système éducatif (quel que soit le niveau de formation), une part élevée de parents diplômés de l'enseignement supérieur et une moindre proportion de parents peu diplômés.

2 Utilisation de *FactoMineR*

*Cette deuxième section ne requiert que très peu de manipulations. On va simplement reprendre les questions précédentes mais à l'aide d'une fonction de **R** qui vous permettra de faire l'ACP de façon automatique.*

*Bien que tous les calculs soient effectués par **R**, il vous restera le travail d'interprétation à effectuer.*

Pour cela, exécuter les commandes suivantes et analyser les différentes sorties de cette fonction :

```
# Installation des packages
install.packages("FactoMineR")
install.packages("factoextra")

# Chargement des bibliothèques
library("FactoMineR")
library("factoextra")

# Réalisation de l'ACP sur les deux premiers axes

res <- PCA(D, scale.unit = TRUE, ncp = 2, graph = TRUE)
res
summary(res)
```

3 Une analyse *from scratch*

Dans les sections précédentes, vous étudiez guidés sur le processus d'Analyse de Données. On se place maintenant dans une situation plus concrète ou vous allez vous-même étudier le jeu de données sans que plus aucune étape ne vous soit donnée. Il faudra donc réaliser vous-même l'ACP (manuellement ou à l'aide d'une fonction sur ) et extraire l'information présente dans le jeu de données : aussi bien sous forme de tableaux que de graphes.

Pour cela, nous considérons un jeu de données relatif à un individu qui suit des formations en ligne et les différentes caractéristiques sur les cours suivis par cet individu. Les caractéristiques étudiées sont les suivantes :

- **Inscription** : nombre de jours écoulés depuis l'inscription au cours
- **Progression** : progression dans le cours (il s'agit d'un pourcentage)
- **MoyenneDeClasse** : moyenne de l'ensemble des étudiants qui ont terminé le cours (en pourcentage)
- **Duree** : durée estimée du cours (en heures)
- **Difficulté** : difficulté estimée du cours (1 : facile, 2 : moyenne, 3 difficile)
- **nbChapitres** : nombre de chapitres composant le cours
- **ratioQuizEvaluation** : proportion de quiz par rapport au nombre total d'évaluations (nombre d'évaluations : nombre de quiz + nombre d'activités)
- **nbEvaluations** : nombre d'évaluations qui compose ce cours
- **derniereMiseAJour** : temps écoulé depuis la dernière consultation du cours

- *idCours* : identifiant du cours sur le site de formation en ligne.

Les variables *idCours* et *derniereMiseAJour* ne sont pas nécessaires à la suite de cette étude, on pourra donc les supprimer de notre jeu de données. Les données se trouvent dans le fichier *data_open.csv*.

Objectif *Etudier le jeu de données à l'aide en effectuant une ACP et essayer d'extraire un maximum d'informations sur les variables et les individus de notre jeu de données.*

Pour cela, vous pourrez vous aider de la démarche effectuée dans l'exercice précédent et l'appliquer à ces données là.

Finalement, à la fin, vous pourrez comparer les résultats obtenus à ceux de la fonction PCA du package FactoMineR