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 R en utilisant les fonctions appropriées. On pourra ainsi comparer les résultats obtenus manuellement avec ceux obtenus avec R. 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 R 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 secondaire [1]
- **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 concentration 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.
2. Entrer et exécuter les commandes suivantes :

   ```
   mean(D[1,])
   apply(D,1,mean)
   mean(D[,1])
   m=apply(D,2,mean)
   ```

   Que représentent ces différentes quantités et notamment m ?
3. 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.
4. Entrer et exécuter les commandes suivantes :

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

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

Que représente s ?
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* $\sigma\sqrt{\frac{n-1}{n}}$.

5. Nous allons à présent centrer, réduire eet 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*.

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.
2. Procéder à la décomposition en éléments propres de C et stocker le résultat de cette décomposition dans la variable *C.eigen*.
3. Entrer et exécuter la commande suivante :

```
C.eigen$values
sort(C.eigen$values)
```

Que fait la commande *sort* ?

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=1.5, cex.axis=1.5, cex.main=1.5)
```

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.
6. Stocker dans deux variables *u1* et *u2*, les deux premiers axes principaux.
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*.
8. Calculer la somme des valeurs propres de C que vous stockerez dans une variable *sum.eig*.

   A quoi correspond cette valeur ?
9. Calculer le pourcentage de l'inertie associée à l'axe *u1*. Il s'agit de la valeur propre associé 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 le 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 ?
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*.
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=1.5,cex.axis=1.5,cex.main =1.5)
```

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* ?

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

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

Expliquer ce que font ces différentes commandes.

1.4 Analyse du nuage des variables

1. Calculer la matrice des produits scalaires entre individus. Cette matrice sera stockée dans la variable K.
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.
3. Stocker dans deux variables *v1* et *v2* les deux premiers axes factoriels *v1* et *v2*.
4. Calculer puis stocker dans les variables *g1* et *g2* les coordonnées des vecteurs variables sur les deux premiers axes factoriels
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=1.5,cex.axis=1.5,cex.main=1.5)
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)
```

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

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

A partir ce 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.
2. Quelle est la mesure de corrélation entre *age_moy* et *pc_par_sup* ? Comment expliqueriez vous la corrélation entre ces deux variables ? 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* ?
4. Quelles sont les variables fortement corrélées à l'axe *v1* ? A l'axe *v2* ?
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 ?
6. Que pourriez-vous dire à propos d'une variable hypothétique y dont les composantes sur les axes *v1* et *v2* seraient $(0.2, -0.1)$?
7. Si vous deviez associer à chaque axe et à chacune de ses parties positives et négatives un groupe de variables, que proposeriez vous ?

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 ‘a 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 ? Même question pour l'axe principal *u2*.

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 *qlt.u1*. Quels sont les deux pays les moins bien représentés sur ce premier axe principal ?
3. Calculer la qualité de la représentation de chaque individu sur l'axe *u2*. Vous stockerez le résultat dans la variable *qlt.u2*. Quels sont les deux pays les moins bien représentés sur ce deuxième axe principal ?
   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.
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.
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 *u1* puis ceux qui contribuent le plus à l'axe *u2* en vous basant sur la médiane.
6. Si vous deviez proposer à chaque axe et à chacune de ses parties positives et négatives un groupe d'individus, que proposeriez-vous ?
7. Interpréter les axes en utilisant conjointement les analyses deux deux nuages de points.

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.

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

```
# Représentation graphique
plot(G,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=1.5,cex.axis=1.5,cex.main=1.5)

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")
```

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

1.8 Exemple d'interprétation

Les interprétation issues de la référence [2] d'où est extraite cette étude sont présentées ci-après :

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

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.*Par rapport à l'axe 1 : L'analyse du cercle des corrélations souligne que l'axe 1 synthétise la relation entre l'^age 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 librairies
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 étudiez 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 R) 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)*
- ***Difficultée :*** *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