TP 2 - ACP : mise en oeuvre des calculs avec R et interprétations

L2 MIASHS - Université Lyon 2 - 2020/2021

Responsables : Julien Ah-Pine, Stéphane Chrétien et Antoine Gourru

1 Objectif du TP et étude de cas

Le premier objectif de ce TP est de mettre en pratique, à l'aide du langage R, la méthode ACP telle que vue en cours. A l'aide d'un exemple numérique, nous allons effectuer pas à pas les différents calculs permettant de réaliser l'ACP d'une table de données. Il s'agit à la fois de mieux maîtriser le langage R et de connaître les commandes permettant de mettre en oeuvre les formules du cours.

Au-delà des calculs et code R, le deuxième objectif du TP est d'aborder plus précisément les interprétations que nous pouvons faire une fois les différents axes principaux déterminés et les nuages de points projetés dans les espaces réduits. Pour nous aider dans cette analyse, nous pouvons calculer plusieurs mesures permettant d'apprécier la qualité de représentations des variables et individus dans leurs plans factoriels respectifs. De plus, il est possible de projeter des éléments supplémentaires dits illustratifs dans l'un ou l'autre espace réduit. Ces éléments permettent d'ajouter de nouvelles informations ce qui conduit à une interprétation des résultats de l'ACP plus riche.

L'ensemble des calculs consistent à coder des équations que nous avons vues en CM. Il est donc important de faire le TP accompagné du CM.

Le cas auquel nous nous intéressons est extrait de la référence suivante : *D. Busca et S. Toutain. (2009). "Analyse factorielle simple en sociologie, Méthodes d'interprétation et études de cas". De Boeck.*

Cet exemple présente pour 9 pays européens 5 variables sur "le passage des jeunes du système éducatif au monde du travail" dont les données proviennent de : *Eurostat. (2000). Enquête sur les forces de travail, Résultats 1999.*

Les 9 pays sont : Autriche, Belgique, Espagne, Finlande, France, Grèce, Hongrie, Italie, Suède.

Les 5 variables sont :

- `V1_1` : Age moyen lors de la sortie du système éducatif de la population active (15-64 ans).
- `V1_2` : Age moyen d'obtention d'un niveau de formation primaire ou secondaire (collège) lors de la sortie du système éducatif.
- `V1_4` : Age 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 renommons respectivement ces variables par : `age_moy`, `age_moy_1_2`, `age_moy_1_3`, `pc_par_1_2` et `pc_par_1_3`.

2 Chargement d'un fichier de données RData

1. Téléchargez le fichier `acp.RData` sur votre disque dur et chargez celui-ci dans votre espace de travail à l'aide de la commande suivante (Attention ! il faut que le fichier soit dans le répertoire de travail ou alors que vous rajoutiez le chemin ou se situe le fichier -pour changer le répertoire

1. Sous-entendu sans avoir terminé un niveau de formation supérieur (universitaire). Il s'agit donc de parents avec un "faible" niveau de diplôme.

de travail dans Rstudio allez dans l'onglet `Session` puis `Set Working Directory`, ou alors utiliser la commande `setwd`-) :

```
#Chargement d'un fichier de variables RData
load("acp.RData")
```

Que constatez-vous ? Remarque : vous pouvez vous-même sauvegarder des variables instanciées lors d'une session R à l'aide de la commande `save` (si besoin consulter l'aide pour le bon fonctionnement de la commande). Dans ce cas, la convention pour l'extension du fichier de données est `.RData` (avec ou sans majuscules...).

3 Préparation et transformation des données

2. Stockez dans une variable `T` (représentant la table de données $\mathbf{T}$) les 5 vecteurs représentants les variables de l'étude.
3. Entrez et exécutez les commandes suivantes :

```
mean(T[1,])
apply(T,1,mean)
mean(T[,1])
#vecteur des moyennes des variables (barycentre individus)
m=apply(T,2,mean)
```

Que représente `m` ?

4. Déterminez à partir de `T` le nombre d'individus et le nombre de variables en utilisant les commandes `nrow` et `ncol`. Vous stockerez ces valeurs respectivement dans les variables `n` et `p`.
5. Entrez et exécutez les commandes suivantes :

```
#vecteur des écart-types des variables
apply(T,2,sd)
s=apply(T,2,sd)#estimation non-biaisée
s=s*sqrt((n-1)/n)#estimation biaisée
```

Que représente `s` ? Remarque : la fonction `sd` ("standard deviation") est par défaut l'estimation sans-biais de l'écart-type d'une variable. Dans le cas de l'ACP, il est d'usage de prendre l'estimation biaisée. C'est la raison pour laquelle nous prenons $\sigma\frac{\sqrt{n-1}}{\sqrt{n}}$.

6. Nous allons à présent centrer, réduire et diviser par $\sqrt{n}$ la terme général de la matrice $\mathbf{T}$ et nous allons stocker la nouvelle matrice $\mathbf{X}$ dans la variable `X`. Entrez et exécutez les commandes suivantes :

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

Vérifiez que le barycentre des individus de $\mathbf{X}$ est le vecteur nul. Vérifiez également que la norme des vecteurs colonnes de $\mathbf{X}$ vaut 1. Pour cela utilisez la commande `apply`.

7. Pour que la table de données soit riche en information, nous pouvons donner des noms au lignes et colonnes d'une matrice. Entrez et exécutez les commandes suivantes :

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

4 Analyse du nuage des individus

8. A partir de `X`, stockez dans la variable `C`, la matrice des corrélations des variables $\mathbf{C}$.
9. Procédez à la décomposition en éléments propres de $\mathbf{C}$ et vous stockerez le résultat de cette décomposition dans la variable `C.eigen`.

10. Entrez et exécutez la commande suivante :

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

Que fait la commande `sort` ?

11. Entrez et exécutez 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 comportant plusieurs paramètres. Amusez-vous à changer quelques paramètres pour comprendre leur signification. Consultez l'aide également.

12. Combien d'axes principaux proposez-vous de garder ?
13. Stockez dans deux variables `u1` et `u2`, les deux premiers axes principaux.
14. Calculez 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`.
15. Calculez la somme des valeurs propres de $\mathbf{C}$ que vous stockerez dans la variable `O.int`. A quoi correspond cette valeur ?
16. Calculez le pourcentage de l'inertie associée à l'axe $\mathbf{u}_1$. Il s'agit de la valeur propre associée à cet axe divisé par l'inertie totale. Faites de même pour l'axe principal $\mathbf{u}_2$. Le premier plan principal est l'espace engendré par $\mathbf{u}_1$ et $\mathbf{u}_2$. Le pourcentage de l'inertie expliquée par ce plan factoriel est la somme des inerties de chaque axe le constituant. Quel est le pourcentage d'information que contient le premier plan factoriel ?
17. Stockez dans la variable `F` la matrice $\mathbf{F}$ qui comporte dans ses 2 colonnes les coordonnées des individus sur $\mathbf{u}_1$ et $\mathbf{u}_2$. Donnez des noms aux lignes de `F`. Les noms des colonnes de `F` seront `u1` et `u2`.
18. Entrez et exécutez la commande graphique 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 très utilisée qui sert à représenter des points dans un repère orthonormé. Dans la commande précédente, les points sont les 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` ?

19. Entrez et exécutez l'une après l'autre les commandes graphiques suivantes :

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

Expliquez ce que fait chacune de ces commandes.

5 Analyse du nuage des variables

20. Calculez la matrice des produits scalaires entre individus $\mathbf{K}$. Stockez la dans la variable `K`.

21. Procédez à la décomposition en éléments propres de $\mathbf{K}$ et vous stockerez les résultats dans la variable `K.eigen`. Observez les valeurs propres de $\mathbf{K}$ et commentez.
22. Stockez dans deux variables `v1` et `v2` les deux premiers axes factoriels $\mathbf{v}_1$ et $\mathbf{v}_2$.
23. Calculez puis stockez dans les variables `g1` et `g2` les coordonnées des vecteurs variables sur les deux premiers axes factoriels.
24. Faites un graphique représentant les vecteurs variables dans le premier plan factoriel. Vous donnerez des titres et légendes des axes adéquats.
25. Ajoutez les noms des variables au graphique précédent ainsi que des droites représentant l'axe des abscisses et celui des ordonnées.

6 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 ie à l'espace de dimension 2 engendré par $\mathbf{v}^1$ et $\mathbf{v}^2$. Néanmoins, ce qui suit pourra être également appliqué à d'autres plans factoriels tel que celui engendré par $(\mathbf{v}^1, \mathbf{v}^3)$ par exemple.

26. Etudiez la matrice des coefficients de corrélation (variable `C` déjà calculée), et identifiez les variables qui sont corrélées positivement et celles qui sont corrélées négativement.
27. Quelle est la mesure de corrélation entre `age_moy` et `pc_par_1_3` ? Comment expliqueriez vous la corrélation entre ces deux variables ? Mêmes questions pour le couple `age_moy_1_2` et `pc_par_1_2`.
28. Que représentent les coordonnées des différentes variables sur les axes factoriels $\mathbf{v}^1$ et $\mathbf{v}^2$?
29. Quelles sont les variables fortement corrélées à l'axe $\mathbf{v}^1$? à l'axe $\mathbf{v}^2$?
30. Au travers du positionnement des variables `pc_par_1_2` et `pc_par_1_3` vis à vis de l'axe $\mathbf{v}_1$, que pouvez vous dire de la corrélation entre ces deux variables ?
31. Que pourriez vous dire à propos d'une variable hypothétique $\mathbf{y}$ dont les composantes sur les axes $\mathbf{v}_1$ et $\mathbf{v}_2$ seraient $(0.2, -0.1)$?
32. Si vous deviez associer à chaque axe et à chacune de ses parties positives et négatives un groupe de variables que proposeriez vous ?

7 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 calul 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 où 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.

33. 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 $\mathbf{u}_1$ oppose t-il ? Même question pour l'axe principal $\mathbf{u}_2$.
34. 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, calculez la qualité de la représentation de chaque individu sur l'axe $\mathbf{u}_1$. Vous stockerez le résultat dans la variable `qlt.u1`. Quels sont les deux pays les moins bien représentés sur ce premier axe principal ?
35. Calculez la qualité de la représentation de chaque individu sur l'axe $\mathbf{u}_2$. Vous stockerez le résultat dans la variable `qlt.u2`. 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.
36. 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. Calculez les contributions des individus aux axes $\mathbf{u}_1$ et $\mathbf{u}_2$. Vous stockerez ces résultats dans les variables `ctr.u1` et `ctr.u2` respectivement.
37. 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. Déterminez les individus qui contribuent le plus à l'axe $\mathbf{u}_1$ puis ceux contribuant le plus à l'axe $\mathbf{u}_2$ en vous basant sur la médiane.
38. Si vous deviez associer à chaque axe et à chacune de ses parties positives et négatives un groupe d'individus que proposeriez vous ?
39. Interprétez les axes en utilisant conjointement les analyses des deux nuages de points.

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

40. Stockez dans une variable `Tp` la matrice $\mathbf{T}_+$ de taille (4×5) dont les lignes sont les individus fictifs suivants :

$$\mathbf{t}_{+,1} = \begin{pmatrix} 24 \\ 20 \\ 26 \\ 10 \\ 70 \end{pmatrix} \quad ; \quad \mathbf{t}_{+,2} = \begin{pmatrix} 17 \\ 15 \\ 20 \\ 50 \\ 30 \end{pmatrix} \quad ; \quad \mathbf{t}_{+,3} = \begin{pmatrix} 21 \\ 19 \\ 25 \\ 70 \\ 10 \end{pmatrix} \quad ; \quad \mathbf{t}_{+,4} = \begin{pmatrix} 19 \\ 17 \\ 24 \\ 20 \\ 20 \end{pmatrix} \tag{1}$$

    On remarquera par exemple que pour le pays fictif $\mathbf{t}_{+,1}$, 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.
41. Transformez la matrice $\mathbf{T}_+$ en centrant, réduisant selon les statistiques estimées sur $\mathbf{T}$ (matrice des individus actifs) et en divisant par n les différentes valeurs afin d'obtenir la matrice $\mathbf{X}_+$ que vous stockerez dans la variable `Xp`. Pour cela, vous pourrez utiliser la commande `scale` vue précédement mais avec les mêmes variables `m` et `s` estimées sur la population originale.
42. Déterminez 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`.
43. 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=1.5,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")
```

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

9 Exemple d'interprétations selon les axes

45. Voici ci-dessous les interprétations issues de la référence [2] d'où est extrait cette étude de cas. Faites le lien entre ces observations et les résultats obtenus précédemment en identifiant notamment les indicateurs permettant de justifier ces interprétations.

Par rapport à l'axe 1 : L'analyse du cercle des corrélations souligne que l'axe 1 syntétise 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. Il oppose 1) 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 élévée de parents diplômés de l'enseignement supérieur et une moindre proportion de parents avec un faible niveau de diplôme, 2) 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 diplomés. 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. *D. Busca et S. Toutain. (2009). "Analyse factorielle simple en sociologie, Méthodes d'interprétation et études de cas". De Boeck.*