TP ACP

Analyse des données

Master ISEFAR - M1

1 Packages R utiles

```
library("knitr") #pour avoir un format table dans les sorties
library("tidyverse")
library("FactoMineR") #pour effectuer l'ACP
library("factoextra") #pour extraire et visualiser les résultats issus de FactoMineR
library("corrplot") #pour avoir une représentation des corrélations
library("MASS") #pour disposer du jeu de données crabes
library("ppcor") #pour calculer les corrélations partielles
```

2 Criminalités aux USA

On s'intéresse à différents crimes violents commis au cours de l'année 1973 dans 50 états des Etats-Unis. Pour ces 50 états, on a noté le pourcentage de population urbaine (variable `UrbanPop`) et le nombre, par 100 000 habitants, d'arrestations pour meurtres (variable `Murder`), agressions (variable `Assault`) et viols (variable `Rape`). Les données sont issues du package `FactoMineR` et peuvent être récupérées via la commande `data(USArrests)`.

1. Charger les données et les packages utiles.
2. Calculer quelques statistiques simples sur les variables (moyennes, écart-types, corrélations) et les représenter graphiquement (boxplot, corrplot). Commenter.

Remarque: la commande suivante retourne les boxplots de chaque variable (colonne du jeu de données)

```
USArrests %>% boxplot()
```

3. Réaliser cette ACP normée.

```
res.pca=PCA(USArrests,scale.unit = TRUE,ncp = 4,graph=FALSE)
# scale.unit: TRUE pour réalisée une ACP normée et FALSE pour une ACP simple
# graph : FALSE ne donne pas les graphes
# ncp : nombres variables sur lesquelles est réalisée l'ACP

#Remarque: si on veut faire une ACP sans la première variable mais
# qu'elle soit projetée quand même
res.pca=PCA(USArrests,scale.unit = TRUE,ncp = 4,graph=FALSE,quanti.sup = 1)
```

4. Représenter la distribution des nouvelles variables (boxplot). Commenter.

```
res.pca$ind$coord %>% boxplot()
```

5. Quelle est le pourcentage d'inertie portée par les deux premiers axes? par le plan principal? Combien vaut la somme des valeurs propres? Ce résultat est-il attendu?

```
kable(res.pca$eig)
```

6. Représenter le graphe des valeurs propres. Combien d'axes retiendriez-vous? Justifier.

```
fviz_eig(res.pca, addlabels = TRUE)
```

7. Tracer le cercle des corrélations.

    a. En quoi la représentation des variables sur le cercle des corrélations illustre-t-elle la matrice des corrélations?

    b. Comment interprétez-vous les composantes principales retenues?

```
fviz_pca_var(res.pca,axes=c(1,2))
# fviz_pca_var(res.pca,axes=c(3,4))

# Coordonnées des nouvelles variables sur les axes (vecteurs propres)
kable(cor(res.pca$var$cor),digits=2)
```

8. Faire le graphe des individus sur les composantes choisies. Commenter.

    a. Interprétez la position des états sur le plan principal.

    b. En quoi s'opposent les états "Florida" et "North-Dakota"?

```
fviz_pca_ind(res.pca, col.ind = "cos2",axes=c(1,2), repel = TRUE,pointsize = 0.7, labelsize = 3)
# repel=TRUE: évite la superposition des noms des individus (si noms et pas numéros)
# pointsize: taille des points
# labelsize: taille des noms des individus

# Si on veut le nuage des individus projeté sur les axes 3 et 4
# fviz_pca_ind(res.pca,axes=c(3,4))

# Si on veut colorier les poinds selon uen variable qualitative, par exemple sex
#fviz_pca_ind (res.pca, habillage = "sex")


#Qualité de la représentation des individus sur le plan principal
cos2 = rowSums(res.pca$ind$cos2[, 1:2])
cos2
#Contribution des individus sur le plan principal
contrib = rowSums(res.pca$ind$contrib[, 1:2])
contrib
```

```r
fviz_pca_biplot(res.pca,axes=c(1,2))
```

9. Refaire l'étude avec une ACP simple cette fois.

   a. Quelles sont les variables qui contribuent le plus au premier axe? au deuxième axe?

   b. Comparer le cercle de corrélation avec celui de l'ACP normée. Commenter en vous appuyant sur les résultats de la question 1.

```r
res.pca.simple=PCA(USArrests,scale.unit = FALSE,ncp = 4,graph=FALSE)
fviz_pca_var(res.pca.simple,axes=c(1,2))   #sur les axes 1 et 2
fviz_pca_ind(res.pca.simple,axes=c(1,2))   #sur les axes 1 et 2


#Représentation du nuage des individus sur le plan principal
fviz_pca_ind (res.pca, col.ind = "cos2",axes=c(1,2)) #dégradé de couleur selon la représentativité
```

3 Fertilité et indicateurs socio-économiques en Suisse

On étudie les données `swiss` disponibles dans `R` donnant une mesure de la fécondité et des indicateurs socio-économiques pour 47 provinces francophones de la Suisse en 1888. On a

- variable `Fertility`: mesure de la fertilité
- variable `Agriculture`: % d'hommes employés dans l'agriculture
- variable `Examination`: % de reçus avec une bonne note à l'examen militaire
- variable `Education`: & % d'études effectuées au-delà de l'école primaire (et réussies)
- variable `Catholic`: % de catholiques (par opposition aux protestants)
- variable `Infant.Mortality`: mortalité infantile avant un an

L'objectif de l'étude associée à ces données est d'expliquer la fertilité en fonction des variables socio-économiques. On souhaite donc effectuer une analyse descriptive de ces 5 dernières variables.

1. Charger les données. Calculer quelques statistiques simples sur les variables (moyennes, écart-types, corrélations) et les représenter graphiquement (boxplot, corrplot). Commenter.

On effectue une ACP normée.

2. Les variables sont-elles toutes bien représentées?
3. Quelles sont les variables qui expliquent le premier axe?
4. Combien d'axes retiendriez-vous? Interprétez les deux premiers axes.
5. Quel lien la variable fertility a avec les variables? avec les axes principaux ?
6. Que peut-on dire de la qualité de représentation de l'individu `Rive Droite`?
7. Etudier la contribution des individus aux deux premiers axes.
8. Caractériser les provinces `Porrentruy`, `Broye`, `Herens`, `V. De Geneve?` en fonction de toutes les variables.

4 Crabes

On s'intéresse à différentes variables mesurées sur 200 crabes de 2 espèces différentes (100 crabes par espèce): la taille du lobe frontal (variable `FL`), la largeur arrière (variable `RW`), la longeur de la carapace (variable `CL`) et l'épaisseur de la carapace (variable `BD`). L'espèce est donnée par la variable `sp` et est codée en B (bleu) et O (orange). On a aussi noté le sexe des crabes (variable `sex`). Les données sont issues du package `MASS` et peuvent être récupérées via la commande `data(crabs)`. On se demande si il est possible d'identifier le sexe et l'espèce à partir des données morphologiques.

1. Charger les données et identifier la nature des différentes variables. Retirer la variable ìndex'.
2. Calculer quelques statistiques simples sur les variables (moyennes, écart-types, corrélations) et les représenter graphiquement (boxplot, corrplot). Commenter.
3. Représenter la distribution des différentes variables en fonction de l'espèce et du sexe.

Pour cela, on réorgansie le jeu de données en faisant

```
crabs.G <- crabs  %>% group_by(sp,sex) %>% gather(key='att',value='mesure',-sp,-sex)
```

4. Effectuer une ACP normée. Combien d'axes garder?
5. Que peut-on dire des crabes qui ont une coordonnée positive sur le premier axe et ceux qui ont une coordonnée négative? Comment peut-on interpréter cet axe?
6. Représenter les individus dans le plan principal en fonction du sexe, puis de l'espèce. Cette représentation permet-elle de distinguer l'espèce et/ou le sexe et selon quelles variables morphologiques? Faire la même chose sur les premier et troisième axes principaux. Qu'obtient-on? (on pourra utiliser la représentation par biplot).
7. Le lien entre deux variables, exprimé par la corrélation, n'est pas toujours un indicateur pertinent. En effet, le lien entre deux variables peut venir du fait qu'elles sont liées à une troisième. La corrélation partielle permet de mesurer la corrélation entre deux variables X et Y corrigée de leurs liens avec une troisème variable Z. Elle est donnée par

$$\rho(X,Y|Z) = \frac{\rho(X,Y) - \rho(X,Z)\rho(Y,Z)}{\sqrt{(1-\rho(X,Z)^2)(1-\rho(Y,Z)^2)}}$$

   où $\rho(X,Y)$ est la corrélation entre X et Y.

Calculer les corrélation partielles entre les variables morphologiques à l'aide de la fonction `pcor` du package `ppcor`. Comparer ces corrélations partielles aux corrélations obtenues précédemment.

```
corrp <- crabs %>% select_if(is.numeric)%>% pcor()
corrp$estimate
corrplot(corrp$estimate)
```