TP Classification non supervisée

Analyse des données

Master ISEFAR - M1

1 Packages R utiles

```
library("tidyverse") #pour avoir de 'beaux' graphiques
library("FactoMineR") #pour effectuer l'ACP
library("factoextra") #pour extraire et visualiser les résultats issus de FactoMineR
```

Dans ce TP, on reprend les études 1 (criminalité aux USA) et 2 (fertilité/indicateurs socio-économiques) du TP sur l'ACP.

2 Criminalités aux USA

On souhaite établir une typologie des 50 états en fonction du type d'arrestations et du taux d'urbanisation. On effectue une Classification Hierarchique Ascendante (CAH).

1. Charger les données et les packages utiles.
2. Normaliser les données et calculer la distance euclidienne entre individus.

SCRIPTS UTILES

```
USArrests.cr <-  USArrests %>% scale(.,scale=TRUE, center=TRUE)
USArrests.dist <- USArrests.cr %>% dist(., method = "euclidean")
```

3. Effectuer la CAH et tracer le dendrogramme pour différentes distances entre groupe ("single"= saut minimum, "average"= saut moyen, "ward.D"= saut de Ward). Commenter.
    a. Combien y-a-t'il de groupes à l'étape initiale? à l'étape finale?
    b. A quel type d'information correspond l'inertie intra-groupe? l'inertie inter-groupe?
    c. Que vaut l'inertie intra-groupe à l'étape initiale? à l'étape finale? Comment se comporte-t-elle en fonction du nombre de groupes?
    d. Représenter l'augmentation de l'inertie intra-groupe en fonction du nombre d'itérations.
    e. A l'aide du dendrogramme et du graphe précédent, combien de groupes choisisseriez-vous? Représenter le dendrogramme avec ce nombre de groupes.
    f. Interpréter les groupes obtenues.

SCRIPTS UTILES

**CAH avec différentes distances entre

```r
# Lien simple
USArrests.single<-USArrests.dist %>% hclust(., method = "single")
fviz_dend(USArrests.single, cex = 0.5)
# Lien complet
USArrests.average<-USArrests.dist %>% hclust(., method = "average")
fviz_dend(USArrests.average, cex = 0.5)
# distance de ward
USArrests.ward<-USArrests.dist^2 %>% hclust(., method = "ward.D")
fviz_dend(USArrests.ward, cex = 0.5)
```

**Evolution de l'inertie-intra en fonction du nombre d'itérations

```r
plot(USArrests.ward$height,type="s",xlab="nb itérations",ylab="heigth")
```

Remarque: si la distance de Ward est utilisée, la sortie de la fonction `data.ward$height` retourne un vecteur dont les valeurs sont égales aux distances de Ward, c'est-à-dire à l'augmentation de l'inertie intra-groupe, à chaque étape d'aggrégation.

**Dendrogramme avec nombre de groupes spécifié

```r
fviz_dend(USArrests.ward,
          k=4,
          cex = 0.8,
          palette="jco",
          rect = TRUE, rect_fill = TRUE, # Rectangle autour des groupes
          rect_border = "jco",
          labels_track_height = 70
          )
```

**Représentation des groupes sur le plan principal de l'ACP et moyennes par groupe

```r
# On récupère les k groupes
cluster.CAH <- USArrests.ward %>%  cutree(., k =4)
# ACP
res.pca=PCA(USArrests,scale.unit = TRUE,ncp = 4,graph=FALSE)
# visualiser les classes sur le premier plan factoriel de l'ACP
fviz_pca_ind(res.pca,axes=c(1,2),habillage=as.factor(cluster.CAH))
fviz_pca_biplot(res.pca,axes=c(1,2),habillage=as.factor(cluster.CAH))
# moyennes des variables par groupe
kable(aggregate(USArrests, by=list(as.factor(cluster.CAH)),mean),digits=1)
```

4. Réalisation les k-means.

   a. Faire tourner l'algorithme avec nstart= 1 et le nombre de groupes choisi précédemment. Représenter les groupes sur l'ACP.

   b. Recommencer a. Comparer les résultats entre les deux classifications obtenues et commenter.

   c. Refaire avec nstart= 10. Comparer avec la CAH.

SCRIPTS UTILES

```r
# 1 fois
res.kmeans <- USArrests %>% kmeans(.,centers =4,nstart = 1)
cluster.kmeans <- res.kmeans$cluster
fviz_pca_biplot(res.pca,axes=c(1,2),habillage=as.factor(cluster.kmeans))

# 1 fois
res.kmeans <- USArrests %>% kmeans(.,centers =4,nstart = 1)
cluster.kmeans <- res.kmeans$cluster
fviz_pca_biplot(res.pca,axes=c(1,2),habillage=as.factor(cluster.kmeans))

# 10 fois
res.kmeans <- USArrests %>% kmeans(.,centers =4,nstart = 10)
cluster.kmeans <- res.kmeans$cluster
fviz_pca_biplot(res.pca,axes=c(1,2),habillage=as.factor(cluster.kmeans))
```

3 Fertilité et indicateurs socio-économiques en Suisse

Proposez une classification des 47 provinces francophones de la Suisse par la méthode CAH. Représenter et interpréter les résultats de la classifcation à l'aide de l'ACP.

4 Décathlon

On s'intéresse à la performance de 28 athlètes du décathlon lors des JO d'athènes Août 2004. Le décathlon est constitué de 10 épreuves: quatre courses (100m, 400m, 110m haies et 1500m), de trois sauts (longueur, hauteur et perche) et de trois lancers (poids, disque et javelot). Pour chaque athlète, on a

- ⋆ les 10 scores des 10 épreuves
- ⋆ le classement des athlètes et le total des points
- ⋆ le nom de la compétition (il y a deux évenements)

Les données sont issues du package `FactoMineR` et peuvent être récupérées via la commande `data(decathlon)`. Comme précisé au début, on ne s'intéresse qu'aux résultats de la compétition "OlympicG". Faire une ACP et une classification des athlètes.

```r
data(decathlon)
decathlon <- decathlon %>% filter(Competition=="OlympicG") %>% dplyr::select(-Competition)
```