TP 3 : Classification non supervisée (Clustering)

Zaineb Smida

On reprend ici le jeu de données sur les 25 pays de l'Union Européenne (source : Eurostat 2002) qu'on importe de la façon suivante :

```
pays <- read.table("pays_eu.txt", header = T)
row.names(pays) <- pays$PAYS
```

En revanche, on va considérer uniquement les variables suivantes :

- espérance de vie à la naissance pour un homme (en années)
- espérance de vie à la naissance pour une femme (en années)
- taux d'activité (en pourcentage) : population active/population en âge de travailler
- taux d'inflation (en pourcentage)
- taux d'emploi (en pourcentage) : actifs occupés/population en âge de travailler
- taux de chômage (en pourcentage): chômeurs/population active

On extrait donc les colonnes qui nous intéressent :

```
paysred <- pays[, c("EVH", "EVF", "TEMP",  "TINF", "TACT", "TCHOM")]
```

Packages nécessaires :

```
library("FactoMineR")
library("cluster")
```

```
Warning: le package 'cluster' a été compilé avec la version R 4.4.2
```

1 Question

Pourquoi faut-il centrer et réduire les variables ?

Conseils

Pour regarder les différences de magnitude entre les variables, on peut utiliser des indicateurs statistiques de base.

Solution

2 Question

Centrer et réduire les variables

Conseils

On pourra utiliser la fonction `scale()`

Solution

3 Question

On décide tout d'abord de faire une classification par Agrégation autour des Moyennes Mobiles (AMM ou k-means en anglais). Représenter le R^2 global en fonction du nombre de groupes et justifier ainsi le choix du nombre de groupes.

Conseils

On utilisera la fonction `kmeans()` en faisant varier le paramètre `centers` pour modifier le nombre de classe. On pourra fixer `nstart` à 100. L'objet retourné contient un certain nombre d'informations sur la classification : par exemple l'inertie inter-classes est donnée dans `$withinss` et l'inertie totale par `$tot.withinss` ce qui permet de calculer le R^2

Solution

4 Question

Confirmer votre choix avec le calcul du coefficient silhouette moyen que vous représenterez graphiquement en fonction du nombre de groupes.

Conseils

On utilisera la fonction `avg_sil()` suivante qui fait appel au package **cluster**

```
library(cluster) #pour avoir la fonction silhouette
avg_sil <- function(k) {
  km.res = kmeans(paysred2, centers = k, nstart = 100)
  ss = silhouette(km.res$cluster, dist(paysred2))
  mean(ss[, 3])
}
```

Solution

5 Question

On choisit ainsi de garder 4 classes. Ajouter une variable dans le jeu de données qui donne le numéro de la classe obtenue avec la fonction `kmeans()`. On utilisera la classe `factor` pour caractériser les classes. Donner les effectifs des classes trouvées.

Conseils

On peut récupérer les numéros des classes à partir de la syntaxe `$cluster`; on peut créer un `factor` avec la fonction `factor()`.

Solution

6 Question

Faire les boîtes à moustaches des variables par groupe. Calculer les moyennes des variables à l'intérieur de chaque groupe. Calculer le rapport de corrélation pour chaque variable et les classer par ordre d'importance. Représenter le nuage de points des deux variables les plus

importantes dans lesquel vous représenterez les observations avec des couleurs différentes selon leur groupe d'appartenance. Vous représenterez aussi les barycentres des classes.

Conseils

Pour calculer les moyennes par classe, on pourra utiliser la fonction `tapply()` (voir notes de cours).

Solution

7 Question

Définir une typologie des pays en prenant les précautions nécessaires. Pour cela, on pourra d'abord utiliser les informations trouvées précédemment. Dans un second temps, on pourra représenter les observations sur les deux premières composantes principales de l'ACP. Pour cela, il faudra interpréter les résultats de l'ACP.

Solution

8 Question

On réalise ensuite une classification ascendante hiérarchique (CAH) sur ces mêmes données. On calculera dans un premier temps la matrice des distances entre individus sur les données centrées et réduites, avec la fonction `dist()`. Dans un second temps, on utilisera la fonction `hclust()` du package **cluster**. Combien de groupes décidez-vous de choisir pour cette CAH ? Justifier votre réponse.

Conseils

On pourra utiliser le graphique des hauteurs et le dendrogramme (voir notes de cours).

Solution

9 Question

Comparer le résultat de cette classification avec les classes obtenues par la méthode des K-means.

Conseils

On pourra utiliser la fonction `cutree()` pour récupérer les classes de la CAH et la fonction `table()` pour comparer les résultats des deux classifications.

Solution

10 Question

Réaliser la CAH sur les composantes principales données par l'ACP. Commenter.

Solution