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 :

```
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 Méthode AMM

On décide tout d'abord de faire une classification par Agrégation autour des Moyennes Mobiles (AMM ou k-means en anglais). On utilisera la fonction `kmeans()`.

3.1 Question

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 `$withinss` et l'inertie totale par `$totss` ce qui permet de calculer le R^2.

Solution

3.2 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

3.3 Question

Créer un vecteur contenant les numéros des classes obtenues avec la fonction `kmeans()` et donner les effectifs des classes obtenues.

Conseils

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

Solution

3.4 Question

Réaliser les boîtes à moustaches des variables pour chaque groupe, puis calculer les moyennes des variables à l'intérieur de chaque groupe.

Conseils

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

Solution

3.5 Question

Calculer le rapport de corrélation pour chaque variable et les classer par ordre d'importance.

Conseils

Pour calculer le rapport de coorélation R^2 pour chaque variable en fonction de votre cluster, vous pouvez utiliser une régression linéaire avec la fonction `lm()` et extraire la valeur de R^2 à partir de `$r.squared`. N'oubliez pas de répéter cette opération pour chaque variable de manière adéquate.

Solution

3.6 Question

Représenter le nuage de points des deux variables les plus importantes, en distinguant les observations par des couleurs différentes selon leur groupe d'appartenance. Vous ajouterez également les centres des classes sur le graphique.

Conseils

Pour ajouter les centres des classes, vous pouvez utiliser la fonction `points()`. Pour inclure une légende et identifier chaque groupe, vous pouvez utiliser la fonction `legend()`.

Solution

3.7 Question

Définir une typologie des pays en prenant les précautions nécessaires. Ensuite, représentez les observations sur les premières composantes principales de l'ACP, que vous choisirez après avoir interprété les résultats de l'ACP.

Conseils

Pour définir une typologie des pays, commencez par utiliser les informations obtenues précédemment. Pour effectuer l'ACP, vous pouvez utiliser la fonction `PCA()`. Sélectionnez ensuite les axes principaux en justifiant votre choix (précisez le critère utilisé). Interprétez les résultats en examinant les corrélations des variables avec les axes principaux pour en comprendre la signification. Enfin, représentez les observations sur les premières composantes principales en utilisant `$ind$coord`.

Solution

4 Méthode CAH

On réalise ensuite une Classification Ascendante Hiérarchique (CAH) sur ces mêmes données. On utilisera la fonction `hclust()` du package **cluster**.

4.1 Question

En effectuant une CAH, commencer par calculer la matrice des distances entre individus à partir des données centrées et réduites. Ensuite, utiliser la fonction `hclust()` pour réaliser la CAH. Déterminer le nombre de groupes à retenir et justifier votre choix en vous appuyant sur des critères appropriés.

Conseils

Pour centrer et réduire les données, utilisez la fonction `scale()`. Ensuite, pour calculer la matrice des distances, utilisez la fonction `dist()`.
Pour déterminer le nombre de groupes, vous pouvez examiner le graphique des hauteurs et le dendrogramme (référez-vous aux notes de cours pour plus de détails).

Solution

4.2 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

4.3 Question

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

Solution