Analyse de données
M1 Statistique et économétrie - 2013-14
V. Monbet
Classification

Étude des caractéristiques d'un ensemble d'hôtels

Le fichier `hotels.CSV` contient des caractéristiques liées au confort, à la qualité et à la situation d'un ensemble d'hôtels. Importer les données sous `R`.

Quelles sont les différentes variables ? Quelle est leur nature ? Qui sont les individus sur qui on va faire porter la classification ? Obtenir les statistiques descriptives, les covariances et les corrélations entre les variables quantitatives du jeu de données.

Classification hiérarchique ascendante

Faire la classification hiérarchique ascendante des observations en utilisant les distances eulidienne et Manhattan et les liaisons simple, complète et de Ward. Dans chacun des cas, représenter le dendogramme, proposer un nombre de classes, lister les hôtels (et les pays) de chaque classe, donner la moyenne de chaque classe et interpréter les résultats.

Par exemple, si `hotels` est la table de données et qu'on choisit la distance euclidienne et le saut de Ward :

```
Pays = hotels[,1]
hotelsnum = hotels[,-1]
d.h = dist(hotelsnum,"euclidean")
cah.h = hclust(d.h,"ward")
str(cah.h)
plot(cah.h,hang =-1)
dev.new()
plot(cah.h$height)
nbcl = 3
split(rownames(hotelsnum),cutree(cah.h,k=nbcl))
(s.pays = split(Pays,cutree(cah.h,k=nbcl))
pca.h = PCA(hotelsnum)
plot(pca.h$ind$coord[,1:2], pch=21,
bg=c("red", "green", "blue")[cutree(cah.h,k=nbcl)])
```

Représenter un graphique de l'évolution du rapport de la variance interclasse sur la variance intraclasse en fonction du nombre de classes.

Obtient-on les mêmes résultats si on réduit les données ?

Quelle est la meilleure classification parmis toutes celles qu'on a testées ? Pourquoi ?

K-moyennes

1. Obtenir la classification des hôtels en trois groupes à l'aide de la méthode des K-moyennes qui portera sur toutes les variables numériques du tableau. Représenter graphiquement les trois groupes sur les premiers plans factoriels de l'ACP. Qu'observe t'on ? Lister les hôtels (et les pays) de chaque classe, donner les centres de chaque classe et interpréter les résultats.

   Pour obtenir la classification, on fait par exemple

   ```
   nbcl = 3
   centers.init = hotelsnum[sample(1:length(hotelsnum[,1]),nbcl),] ;
   K = kmeans(hotelsnum, centers.init)
   ```

   Comment est initialisé l'algorithme ? Obtient-on toujours le même résultat quand on change l'initialisation ?
   Proposer une autre méthode d'initialisation.
   Obtient-on le même résultat si on travaille avec des données réduites ? centrées-réduites ?

2. Obtenir la classification des hôtels en trois groupes à l'aide de la méthode des K-moyennes qui portera sur les coordonnées des hôtels dans le premier plan factoriel. Représenter graphiquement les trois groupes sur les premiers plans factoriels de l'ACP. Qu'observe t'on ? Lister les hôtels (et les pays) de chaque classe, donner les centres de chaque classe et interpréter les résultats.

3. Quelles sont les différences de classement entre les classements des questions 1 et 2 précédentes ? Le premier plan factoriel traduit-il fidèlement l'ensemble des données ? (voir valeurs propres).

4. Sélection de modèle.
   Il est possible d'associer un critère BIC au modèle obtenu quand on fait de la classification par la méthode des moyennes mobiles. On peut en effet supposer qu'on obtient un modèle de mélange gaussien : dans chaque classe le vecteur des variables X suit une loi de Gauss de moyenne μ_k et de variance Σ_k. Ainsi la densité de X est donnée par

$$f_X(x) = \sum_{k=1}^{K} \pi_k \phi_k(x; \mu_k, \Sigma_k)$$

avec π_k la probabilité a priori de la classe k.
Estimer le critère BIC pour des modèles à $K = 2$ jusqu'à $K = 7$ classes. Quel est le meilleur modèle selon ce critère ?
Estimer aussi pour chacun des modèles le rapport de l'inertie inter-classe sur l'inertie totale. Quel est le meilleur modèle selon ce critère ?
Ces résultats sont-ils cohérents ?

5. On décide de vérifier si l'attribution des étoiles est conforme aux critères de constitution des groupes par la méthode des K-moyennes. Puisqu'il existe 6 catégories d'étoiles de 0 à 5, classer les hôtels en 6 groupes à l'aide de la méthode des K-moyennes portant cette fois ci sur toutes les variables à l'exclusion des variables prix et étoiles. Comparer la répartition des hotels dans les groupes et leur nombre d'étoiles. On peut utiliser la fonction `table`. Attention, les groupes ne sont pas nécéssairement numérotés par ordre croissant d'étoiles.