Examen de clustering

02/02/2023

Durée : 2h
Les calculatrices et les téléphones portables ne sont pas autorisés. Une feuille A4 recto de notes manuscrites est autorisée et sera rendue dans la copie.
Vous prendrez soin à la rédaction de vos réponses.

Exercice 1

On considère un jeu de données $X = (X_{ij})_{1 \leq i \leq n, 1 \leq j \leq 6}$ où n individus sont décrits par $p = 6$ variables quantitatives. On souhaite ici mettre en place une classification non supervisée **des variables** à l'aide d'une classification hiérarchique ascendante avec la dissimilarité suivante

$$d(X^{(j)}, X^{(\ell)}) = 1 - |\rho(X^{(j)}, X^{(\ell)})|$$

où $\rho(X^{(j)}, X^{(\ell)})$ dénote la corrélation empirique entre les variables j et ℓ, et la mesure d'agrégation complète entre deux classes $\mathcal{C}_k$ et $\mathcal{C}_{k'}$

$$D(\mathcal{C}_k, \mathcal{C}_{k'}) = \max_{j \in \mathcal{C}_k, \ell \in \mathcal{C}_{k'}} d(X^{(j)}, X^{(\ell)})$$

La matrice des dissimilarités entre les 6 variables vaut

	V1	V2	V3	V4	V5	V6
V1	0.0	0.7	0.9	0.8	0.9	0.9
V2	0.7	0.0	0.6	0.6	0.7	0.9
V3	0.9	0.6	0.0	0.3	0.5	0.6
V4	0.8	0.6	0.3	0.0	0.4	0.6
V5	0.9	0.7	0.5	0.4	0.0	0.5
V6	0.9	0.9	0.6	0.6	0.5	0.0

Q1 : Montrez que la formule de Lance et Williams pour mettre à jour le calcul de la mesure d'agrégation est donnée par

$$D(\mathcal{C}_u, \mathcal{C}_{k \cup k'}) = \frac{1}{2} D(\mathcal{C}_u, \mathcal{C}_k) + \frac{1}{2} D(\mathcal{C}_u, \mathcal{C}_{k'}) + \frac{1}{2} |D(\mathcal{C}_u, \mathcal{C}_k) - D(\mathcal{C}_u, \mathcal{C}_{k'})|$$

Q2 : Construisez le dendrogramme associé à cette procédure de clustering. Vous préciserez les étapes de calculs effectuées.

Exercice 2

On considère dans cet exercice des données $\underline{\mathbf{y}} = (y_i)_{1\leq i\leq n}$ où $y_i \in \mathbb{R}^p$. On désire obtenir une classification de ces n individus. Soit $\underline{\mathbf{z}} = (z_1, \ldots, z_n)$ le vecteur des labels inconnus tel que $z_i = (z_{i1}, z_{i2}, z_{i3})$ avec $z_{ik} = 1$ si i est dans la classe k et 0 sinon. On modélise la densité des observations par le mélange suivant :

$$y_i \in \mathbb{R}^p \mapsto \pi_1\phi(y_i| - \alpha, I_p) + \pi_2\phi(y_i|0, I_p) + \pi_3\phi(y_i|\alpha, I_p),$$

où

- I_p est la matrice identité, $\alpha \in \mathbb{R}^p$,
- $\phi(x|\mu, \Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right]$ est la densité de la loi gaussienne $\mathcal{N}_p(\mu, \Sigma)$
- $(\pi_1, \pi_2, \pi_3) \in [0, 1]^3,\ \pi_1 + \pi_2 + \pi_3 = 1$

Q1 : Montrez que la logvraisemblance complétée est donnée par

$$\mathcal{L}(\underline{\mathbf{y}}, \underline{\mathbf{z}}|\theta) = \sum_{i=1}^{n}\sum_{k=1}^{3} z_{ik}\ln(\pi_k) - \frac{1}{2}\sum_{i=1}^{n}\left[z_{i1}\|y_i + \alpha\|^2 + z_{i3}\|y_i - \alpha\|^2\right] + \blacksquare$$

où $\blacksquare$ ne dépend pas des paramètres à estimer $\theta = (\pi_1, \pi_2, \pi_3, \alpha)$.

On note par la suite $Q(\theta|\theta^{(r)}) = \mathbb{E}[\mathcal{L}(\underline{\mathbf{y}}, \underline{\mathbf{z}}|\theta)|\underline{\mathbf{y}}, \theta^{(r)}]$.

Q2 : On met en place un algorithme EM pour estimer les paramètres $\theta = (\pi_1, \pi_2, \pi_3, \alpha)$. On initialise l'algorithme par $\theta^{(0)} = (\pi_1^{(0)}, \pi_2^{(0)}, \pi_3^{(0)}, \alpha^{(0)})$ et on se place à l'itération r.

- **Q2.a.** Montrez que l'étape E de cet algorithme consiste à évaluer

$$t_{ik}^{(r)} = \mathbb{E}[z_{ik}|y_i, \theta^{(r)}],\ \forall i \in \{1, \ldots, n\},\ \forall k \in \{1, 2, 3\}$$

et donnez une expression en fonction des $\pi_k^{(r)}$ et $\alpha^{(r)}$.

- **Q2.b.** Montrez que la mise à jour des paramètres dans l'étape M consiste en

$$\begin{cases} \pi_k^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}^{(r)},\ \forall k \in \{1, 2, 3\} \\ \\ \alpha^{(r+1)} = \dfrac{\sum_{i=1}^{n}(t_{i3}^{(r)} - t_{i1}^{(r)})y_i}{n - \sum_{i=1}^{n} t_{i2}^{(r)}} \end{cases}$$

Exercice 3 :

Dans cet exercice, on s'intéresse au fonctionnement de la thyroide de $n = 215$ patients. Pour ces patients, on a mesuré les 4 variables suivantes :

- T4 : Thyroxine sérique totale mesurée par la méthode du déplacement isotopique
- T3 : Triiodothyronine sérique totale mesurée par dosage radio-immunologique
- TSH : Hormone thyréostimulante basale, mesurée par dosage radio-immunologique
- DTSH : Différence absolue maximale de la valeur de la TSH après injection de 200 microgrammes d'hormone de libération de la thyrotropine par rapport à la valeur basale

L'objectif est de déterminer une classification des patients à partir de ces 4 variables.

```
##        T4              T3              TSH             DTSH
##  Min.   : 0.500   Min.   : 0.20   Min.   : 0.10   Min.   :-0.700
##  1st Qu.: 7.100   1st Qu.: 1.35   1st Qu.: 1.00   1st Qu.: 0.550
##  Median : 9.200   Median : 1.70   Median : 1.30   Median : 2.000
##  Mean   : 9.805   Mean   : 2.05   Mean   : 2.88   Mean   : 4.199
##  3rd Qu.:11.300   3rd Qu.: 2.20   3rd Qu.: 1.70   3rd Qu.: 4.100
##  Max.   :25.300   Max.   :10.00   Max.   :56.40   Max.   :56.300
```

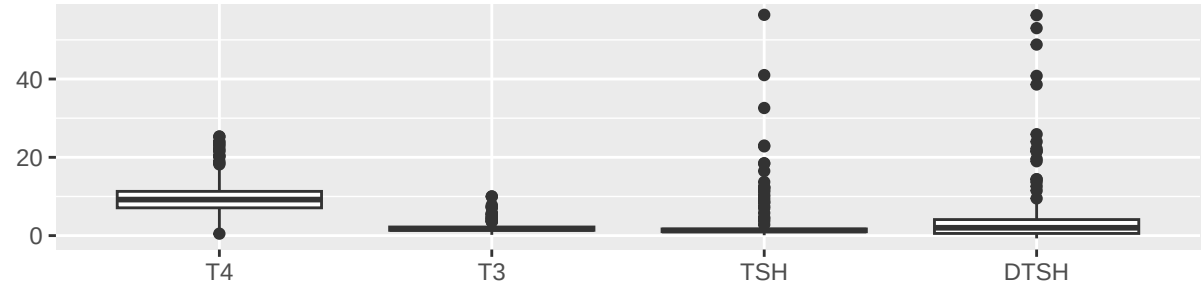


Figure 1: Boxplot des 4 variables

On a également à notre disposition une variable `Diag` qui est le diagnostique de la thyroide suite à une opération pour chaque patient avec 3 états possibles : Hypo, Normal, et Hyper. Elle servira seulement pour interpréter les résultats.

```
table(Diag)
```

```
## Diag
##   Hypo Normal  Hyper
##     30    150     35
```

Q1 : On trace l'inertie interclasse obtenue par classification des patients par méthode des Kmeans en fonction du nombre de classes K (Figure 2).

- **Q1.a.** Exprimez l'inertie interclasse et l'inertie intraclasse associées à une classification en K classes.
  Démontrez une relation reliant ces deux types d'inertie.
- **Q1.b.** Quel nombre de classes proposez-vous de retenir à partir de la Figure 2 ?

Q2 : Toujours avec l'algorithme des Kmeans, on a obtenu les graphiques présentés en Figure 4 à l'aide de la méthode Silhouette. Qu'en déduisez-vous ? Commentez les résultats sur la classification retenue.

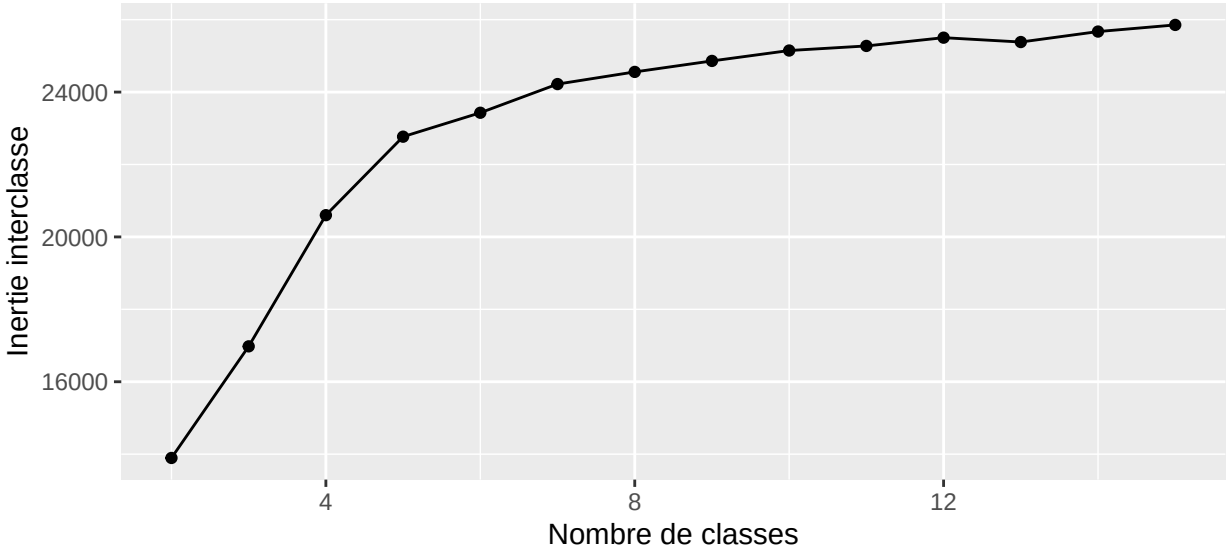


Figure 2: Inertie interclasse en fonction du nombre de classes

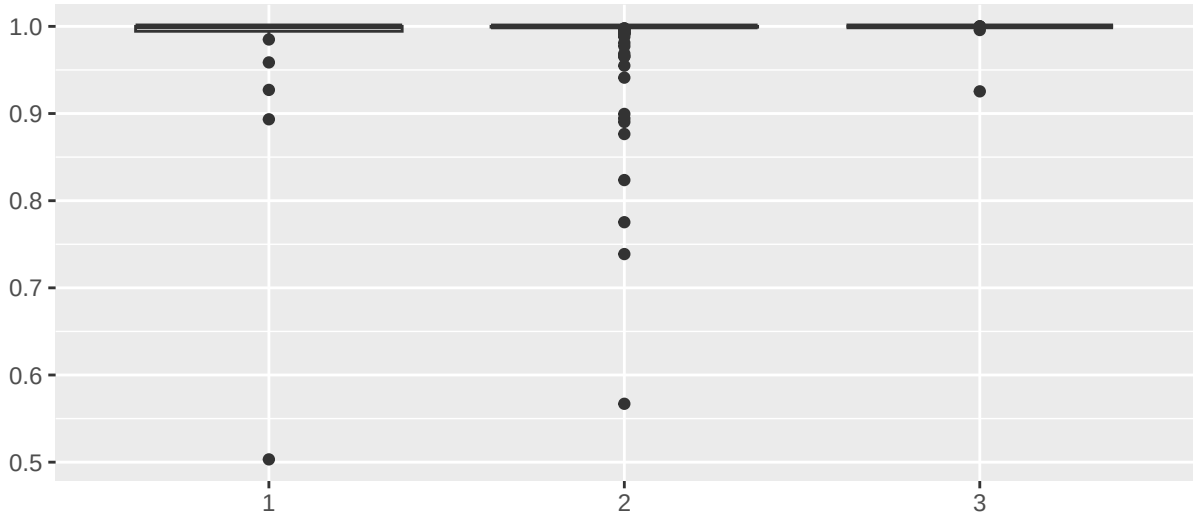


Figure 3: .....

Q3 : Dans cette question on souhaite déterminer une classification des patients à l'aide de mélanges gaussiens.

- **Q3.a. :** On execute les commandes suivantes

```
res<-Mclust(Data,G=2:7,modelNames=c("VII","VVI","EEE","VVV"))
```

Définissez la collection de modèles considérée.

- **Q3.b. :** A quoi sert la Figure 5. Qu'en concluez-vous ?
- **Q3.c. :** Notons $f(.|\hat{\theta}) = \sum_{k=1}^{3} \hat{\pi}_k \phi(.|\hat{\mu}_k, \hat{\Sigma}_k)$ la densité estimée du mélange gaussien retenu dans `res` (cf Q3.a.). Expliquez mathématiquement comment est obtenue la classification des patients à partir de $f(.|\hat{\theta})$.
- **Q3.d. :** Que représente la Figure 3 obtenue avec le code suivant ? Commentez cette figure.

```
df<-data.frame(x=as.factor(res$classification),y=apply(res$z,1,max))
ggplot(df,aes(x=x,y=y))+geom_boxplot()
```

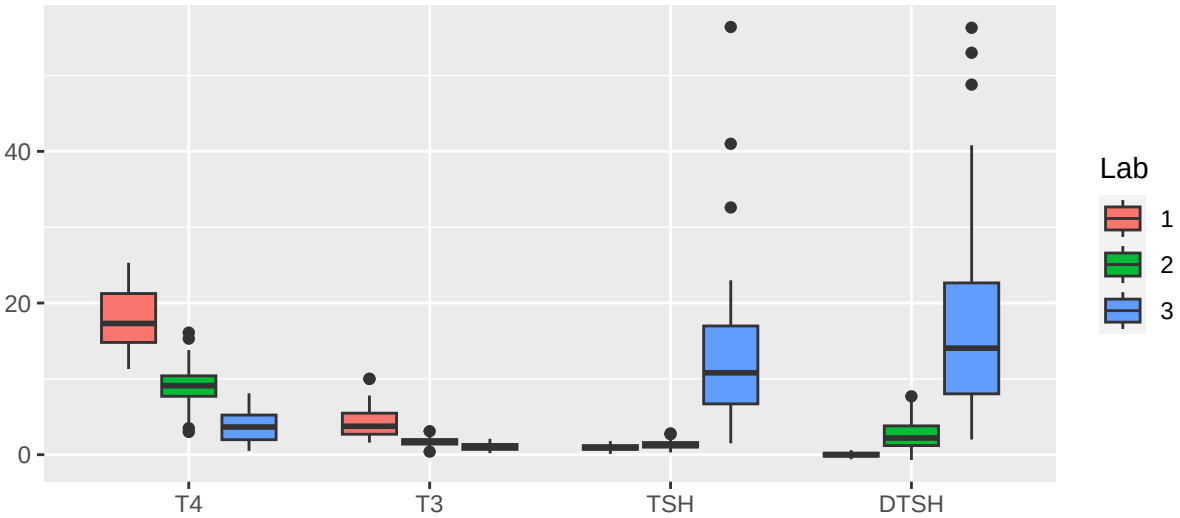
- **Q3.e. :** Avec les sorties suivantes, commentez la classification obtenue dans `res`.

```
table(Diag,res$classification)
```

```
##
## Diag       1   2   3
##   Hypo     0   4  26
##   Normal   1 147   2
##   Hyper   33   2   0
```

```
adjustedRandIndex(Diag,res$classification)
```

```
## [1] 0.8611583
```



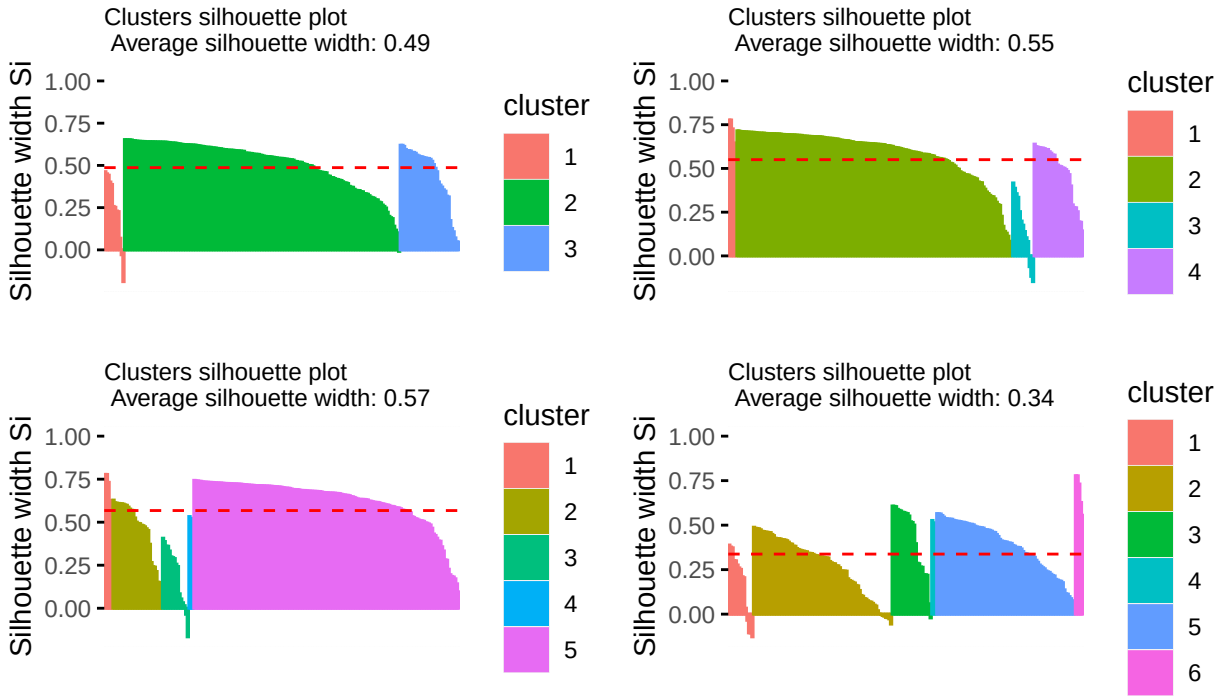


Figure 4: Graphiques obtenus avec Silhouette pour la méthode des Kmeans.

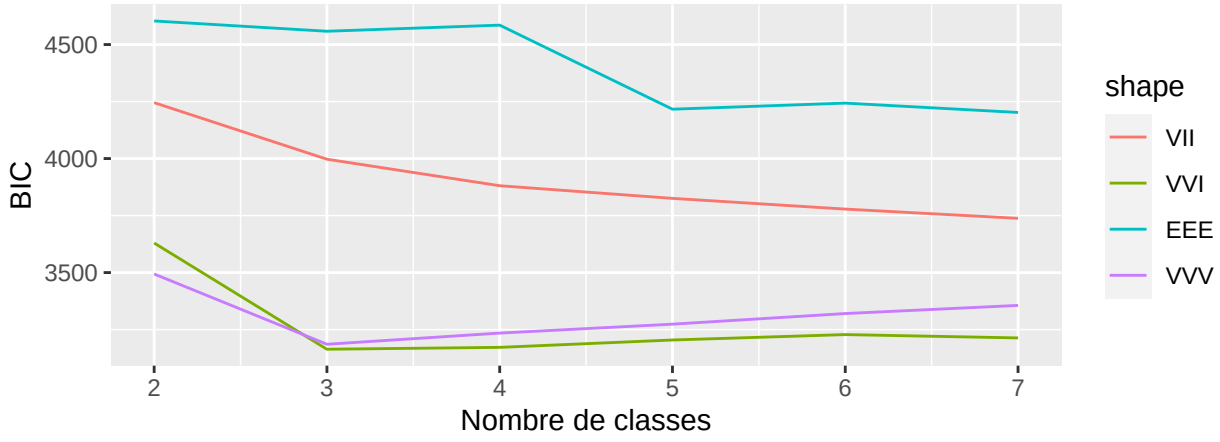


Figure 5: Valeurs du critère BIC en fonction du nombre de classes.