Clustering spectral et modèles de mélanges

Exercice 1 (Segmentation d'image)

1. Télécharger le fichier `irm_small.jpeg` et l'importer sous R en utilisant la fonction `readJPEG` du package `jpeg`.
2. Afficher l'image à l'aide de la fonction `image`.
3. Appliquer l'algorithme de partitionnement spectral normalisé afin d'identifier différentes zones dans l'image. On utilisera la fonction de similarité Gaussienne et le graphe du ε-voisinage en choisissant un seuil ε égal au quantile d'ordre 75% des indices de similarité. On justifiera le nombre de classes sélectionnées.
4. Afficher les classes obtenues.

Exercice 2 (Quelques expérimentations numériques)

On considère que $X = [X_1, \ldots, X_d]$ est une variable aléatoire définie par la densité suivante

$$p(x;\theta) = \sum_{k=1}^{2} \pi_k \prod_{j=1}^{d} \phi(x_j; \mu_{kj}, \sigma_{kj}^2)$$

avec $\pi_1 = \pi_2 = 1/2$, $\sigma_{1j} = 1$, $\mu_{1j} = 0$,

$$\mu_{2j} = \begin{cases} 2.5 & \text{si } j < r \\ 0 & \text{sinon} \end{cases} \quad \text{et } \sigma_{2j} = \begin{cases} 2 & \text{si } j < r \\ 0 & \text{sinon} \end{cases}$$

1. Écrire la fonction `generdata` qui prend pour argument la taille de l'échantillon n, le nombre r et le nombre de variables d. Cette fonction retourne l'échantillon généré ainsi que la vraie partition.
2. On souhaite étudier l'importance de la sélection de variables en clustering. Pour cela, on calcule l'ARI entre la vraie partition et les partitions obtenues avec et sans sélection de variables par le package `VarSelLCM`, en considèrant 20 réplicats générés avec $n = 100$, $r = 3$ et $d \in \{3, 6, 20, 50\}$.
3. Modifier la fonction `generdata` pour que les donnée possèdent un taux τ% de valeurs manquantes (MCAR). La fonction retourne maintenant 3 éléments: l'échantillon sans valeurs manquantes, l'échantillon avec valeurs manquantes et la partition.
4. On souhaite étudier l'intérêt des mélanges lorsque les observations présentent des valeurs manquantes. Pour cela on compare les deux approches suivantes:
   - Imputation des valeurs manquantes par la fonction `imputePCA` puis K-means sur les données imputées.
   - Utilisation des mélanges pour l'imputation et le clustering.

   On considère le cas où $n = 100$ et $r = 3$. Montrez l'évolution de la qualité de la partition lorsque τ augmente pour $d \in \{3, 6, 20, 50\}$.