SY19 –A 18
TP 5: algorithme EM

Exercice 1

Soit $\boldsymbol{X} = (X_1, \ldots, X_n)$ un échantillon iid issu du mélange d'une loi normale unidimensionnelle $\mathcal{N}(\mu, \sigma)$ et d'une loi uniforme $\mathcal{U}([-a, a])$, de densité

$$p(x; \theta) = \pi\phi(x; \mu, \sigma) + (1 - \pi)c, \tag{1}$$

où $\phi(\cdot; \mu, \sigma)$ est la fonction de densité de la loi normale, $c = (2a)^{-1}$, π est la proportion de la loi normale dans le mélange et $\theta = (\mu, \sigma, \pi)^T$ est un vecteur de paramètres. La constante c est fixée. Typiquement, la loi uniforme représente la distribution des observations aberrantes. La proportion d'observations aberrantes est donc $1 - \pi$. On cherche à estimer le paramètre θ par l'algorithme EM.

1. A l'aide des fonctions `sample`, `rnorm` et `runif`, générer un échantillon de taille $n = 100$, pour une certaine valeur de θ que l'on fixera. Faire un diagramme en boîte.
2. Coder un algorithme EM pour ce problème.
3. Appliquer cet algorithme sur les données, avec différentes initialisations. Tracer les probabilités estimées $1 - y_i^{(t)}$ que l'observation i soit aberrante, en fonction des x_i. Interpréter le résultat.
4. Utiliser la fonction `optim` pour maximiser la vraisemblance et comparer les résultats.

Exercice 2

En utiliser le package `mclust`, rechercher le meilleur modèle de mélange gaussien pour les données `wine`, `seeds` et `e.coli`. (Commencer par décrire les données. Comparer la vraie partition et la partition obtenue par l'algorithme EM.)