TISD – FICHE 7

Algorithme EM et applications (avec R)

Adrien Hardy, adrien.hardy@math.univ-lille1.fr

1 Données censurées

Problème typique : À la fin d'une expérience qui a duré T heures, on reporte la durée de vie de n composants électroniques : $x_1,\dots,x_n$. On modélise les vraies durées de vie $z_1,\dots,z_n$ (qui peuvent donc être supérieures à T) comme des réalisations indépendantes d'une même variable de loi exponentielle de paramètre $\theta > 0$ inconnu. Le problème est alors d'estimer θ avec pour seule information la donnée des x_i et la durée de l'expérience T.

Cadre théorique (cf. cours) : Étant donné un paramètre d'initialisation $\theta_0 > 0$, la suite (θ_k) obtenue par itération de l'algorithme EM est caractérisée par la relation de récurrence :

$$\theta_{k+1} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{m_0}{n}\frac{1}{\theta_k}} \ ,$$

où m_0 est le nombre de x_i supérieurs ou égaux à T.

Le but de cet exercice est de tester l'efficacité de cet algorithme sur un jeu de données construit artificiellement.

1. Créer une fonction m_0 qui, étant donné $T \in \mathbb{R}$ et un vecteur x, renvoie le nombre d'entrées de x qui sont supérieures ou égales à T.

2. Créer une fonction qui, étant donné un vecteur x, un entier $k \geq 1$, $T \in \mathbb{R}$, et un paramètre initial $\theta_0 > 0$, renvoie la suite $\theta_0,\dots,\theta_k$ obtenue par itération de l'algorithme EM.

3. Générer une réalisation $z_1,\dots,z_{200}$ d'un échantillon de loi $\mathcal{E}(0.1)$, puis construire $x_1,\dots,x_{200}$ avec $T = 15$ à l'aide de la commande `pmin`. Tester l'algorithme EM sur ce jeu de données : Après avoir choisi aléatoirement θ_0, représenter graphiquement θ_k comme une fonction de k

4. Même question qu'en **3.** mais avec $T = 2$. Qu'en pensez-vous ?

5. Dessiner θ_{15} en fonction de T, où T varie de 0.1 à 10 par pas de 0.01.

2 Mélanges gaussiens

2.1 Simulation de mélanges gaussiens

Un mélange gaussien X est défini de la façon suivante : Étant donné $r \geq 1$ et $\pi_1, \ldots, \pi_r > 0$ tels que $\pi_1 + \cdots + \pi_r = 1$, on tire une variable Z à valeurs dans $\{1, \ldots, r\}$ avec $\mathbb{P}(Z = j) = \pi_j$. Si $Z = j$, alors on tire une variable X de loi normale $\mathcal{N}(\mu_j, \sigma_j^2)$, où $\mu_j \in \mathbb{R}$ et $\sigma_j > 0$ sont donnés.

1. (Théorique) Calculer la fonction de répartition de X ainsi que sa densité.

2. Ecrire une fonction qui, étant donné les paramètres $\pi \in]0,1[$, $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$ et un entier $n \geq 1$, renvoie une réalisation d'un échantillon de taille n du mélange gaussien X associé à ces paramètres où $\pi_1 := \pi$ et $\pi_2 := 1 - \pi$.

3. On prend pour paramètres $\pi = 0.25$, $\mu_1 = 1$, $\mu_2 = 7$, $\sigma_1 = \sigma_2 = 1$ et $n = 1000$. Donner l'histogramme de l'échantillon généré. Ensuite, dessiner la densité approchée de l'échantillon associé à l'aide de la commande `density` et superposer la courbe théorique obtenue en **1.**

4. Même questions qu'en **3.** mais avec $\mu_2 = 2$. Quelle différence avec le mélange précédent ?

2.2 Retrouver les paramètres avec l'algorithme EM

Cadre théorique (cf. cours) : On considère l'espace des paramètres

$$\Theta = \left\{ \theta = (\pi_j, \mu_j, \sigma_j)_{j=1\ldots r} \ : \quad \pi_j, \sigma_j > 0, \quad \mu_j \in \mathbb{R}, \quad \pi_1 + \cdots + \pi_r = 1 \right\}.$$

Étant donné une réalisation $x_1, \ldots, x_n$, on définit pour chaque $\theta \in \Theta$ la matrice

$$H_{ij}^{\theta} = g_{\theta}(j|x_i) = \frac{\pi_j \gamma_{\mu_j,\sigma_j}(x_i)}{\sum_{j=1}^{r} \pi_j \gamma_{\mu_j,\sigma_j}(x_i)}, \qquad \gamma_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(\frac{x-\mu}{\sigma})^2/2}. \tag{1}$$

Le paramètre $\theta_k = (\pi_j^{(k)}, \mu_j^{(k)}, \sigma_j^{(k)})_{j=1\ldots r} \in \Theta$, obtenu après k itérations de l'algorithme EM, est donné par les formules de récurrence :

$$\pi_j^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} H_{ij}^{\theta_k}, \qquad \mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} x_i H_{ij}^{\theta_k}}{\sum_{i=1}^{n} H_{ij}^{\theta_k}}, \qquad \sigma_j^{(k+1)} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu_j^{(k+1)})^2 H_{ij}^{\theta_k}}{\sum_{i=1}^{n} H_{ij}^{\theta_k}}} \tag{2}$$

A partir de maintenant, on considère le cas où $r = 2$.

5. Créer une fonction qui, étant donné un vecteur $(x_1, \ldots, x_n)$ et un vecteur de paramètres $\theta = (\pi, \mu_1, \mu_2, \sigma_1, \sigma_2)$ renvoie la matrice $[H_{ij}^{\theta}]$ définie en (1).

6. Créer une fonction qui étant donné un entier $k \geq 1$, un vecteur d'observations $(x_1, \ldots, x_n)$ et un vecteur de paramètres θ_0 initial, renvoie la suite des vecteurs $\theta_0, \ldots, \theta_k$ obtenus par itérations de l'algorithme EM, décrits en (2).

7. Tester cet algorithme sur les deux mélanges gaussiens obtenus en **3.** et **4.** Commenter vos résultats.