TP1 : Estimation des paramètres d'un mélange de gaussiennes

1) Ecrire une fonction Scilab normale permettant de simuler une loi gaussienne sur $\mathbb{R}$.

2) Selon un article de N. Herpin (*cf.* [2]), la taille moyenne des femmes en 2001 était de $1,619m$ avec un écart-type de $0,065m$ tandis que celle des hommes était de $1,741m$ avec un écart-type de $0,071m$. La proportion française était, à peu près à la même époque, constituée de $51,4\%$ de femmes et $48,6\%$ d'hommes (statistiques disponibles sur le site de l'INSEE). Ecrire une fonction Scilab échantillon pour simuler un tel mélange de n individus en supposant que chaque distribution suive une loi gaussienne indépendante. Représenter les données sous la forme d'un histogramme. Représenter graphiquement la densité du mélange de la question précédente.

3) Comment tester l'adéquation à une loi gaussienne ? Le faire pour l'échantillon simulé à la question précédente. Si votre test précédent conclut qu'on a affaire à une gaussienne, estimez ses paramètres et allez à la question 7 (et reprenez à la question 4 après).

4) On se pose maintenant le problème d'estimer les deux paramètres de chaque gaussienne, ainsi que la proportion du mélange à partir des données non-marquées (c'est–à–dire qu'on ne sait pas si un individu donné est un homme ou une femme). Ecrire le modèle statistique correspondant à la situation de la question précédente et écrire la vraisemblance $\mathcal{L}(x_1, \ldots, x_n)$ d'un échantillon $(x_1, \ldots, x_n)$ dans ce modèle. Quel genre de méthode proposeriez-vous (sans la mettre en oeuvre pour l'instant, ce sera l'objet des questions suivantes) pour estimer les différents paramètres en jeu ? En donner les inconvénients et avantages connus.

5) Fixons quelques notations : dans ce qui suit on note $(X_1, \ldots, X_n)$ la suite de variables aléatoires *i.i.d.* dont l'échantillon précédent est une réalisation. On note $\theta = (\alpha, m_0, \sigma_0, m_1, \sigma_1) \in [0,1] \times (\mathbb{R} \times \mathbb{R}_+)^2$ l'ensemble des paramètres de ce mélange. Par ailleurs, on note Z_i la loi de probabilité sur le sexe de l'individu i c'est–à–dire qu'on a : $\mathbb{P}_\theta[Z_i = 0] = \alpha$, $\mathbb{P}_\theta[Z_i = 1] = 1 - \alpha$ et $X_i|_\theta\{Z_i = z\} \sim \mathcal{N}(m_z, \sigma_z)$. On note aussi $\mathcal{L}_c^{(\theta)}$ la vraisemblance de l'échantillon complété, c'est–à–dire l'échantillon (x_i, z_i) où z_i représente le sexe de l'individu i.

Calculer $\mathcal{L}_c^{(\theta)}$ en tant que fonction des x_i, z_i et de θ, puis $\mathbb{E}_\theta[Z_1|X_1, \ldots, X_n]$ en tant que fonction de l'échantillon et de θ et en déduire une expression de

$$Q(\theta, \theta') = \mathbb{E}_{\theta'}[\log(\mathcal{L}_c^{(\theta)})|X_1, \ldots, X_n]$$

Puis calculer

$$\arg\max_{\theta \in \Theta} Q(\theta, \theta')$$

6) Pour réaliser l'estimation par maximum de vraisemblance dans ce genre de situation Dempster, Laird et Rubin ont mis au point (cf. [1]) la méthode suivante.

Étape 0 : Initialisation

On initialise un $\theta^{(0)}$ choisi au hasard.

Étape E (Espérance)

On calcule $Q(\theta, \theta^{(i)})$.

Étape M (Maximisation)

On calcule $\theta^{(i+1)} = \arg\max_\theta Q(\theta, \theta^{(i)})$. On calcule la différence entre la log-vraisemblance de l'échantillon sous $\theta^{(i+1)}$ et sous $\theta^{(i)}$; tant qu'elle est plus grande qu'une certaine valeur, on retourne à l'étape E.

Implémenter cet algorithme à l'aide de Scilab. Représenter la vraisemblance d'un échantillon tiré suivant la méthode de simulation de 2) au cours du temps. Que constate-t-on ? Quels sont les paramètres estimés au final ? Sont-ils proches de ceux qui ont servi à la simulation ?

7) Comment tester l'adéquation à la loi simulée ?

8) Il nous faut prouver ce qu'on a observé : à chaque itération de l'algorithme EM la vraisemblance croît. A partir de la relation existant entre $\log \mathcal{L}^\theta$, $\log \mathcal{L}_c^\theta$, et $\log p^\theta$ où p^θ désigne la vraisemblance de $Z_1, \ldots, Z_n$ (par rapport à la mesure de comptage) conditionnellement à $X_1, \ldots, X_n$ (pour le modèle θ) , écrire une relation liant $\log \mathcal{L}^\theta$, $Q(\theta, \theta')$ et $H(\theta, \theta') \stackrel{\text{def}}{=} \mathbb{E}_{\theta'}[\log(p^\theta) | X_1, \ldots, X_n]$ valable pour tous les couples θ, θ'. Conclure.

Références

[1] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39 :1–38, 1977.

[2] N. Herpin. La taille des hommes : son incidence sur la vie en couple et la carrière professionnelle. *Economie et Statistique*, 361, 2003.