Université de Strasbourg
M2 - Statistique
Outils pour la statistique avancée

Ségolen Geffray
geffray@math.unistra.fr
Année 2018/2019

TP2: estimation des paramètres d'un mélange de deux gaussiennes au moyen de l'algorithme EM

Soit $(X_1, ..., X_n)$ un échantillon i.i.d. issu d'une loi admettant la densité suivante par rapport à la mesure de Lebesgue sur $\mathbb{R}$:

$$f(x) = p\varphi_{m_1,\sigma_1^2}(x) + (1-p)\varphi_{m_2,\sigma_2^2}(x), \quad x \in \mathbb{R}$$

où $0 < p < 1$ et où φ_{m_j,σ_j^2} représente la densité de la loi $\mathcal{N}(m_j, \sigma_j^2)$ avec $m_j \in \mathbb{R}$ et $\sigma_j^2 > 0$, pour $j = 1, 2$.
Le but du TP est de programmer l'algorithme EM pour déterminer l'estimateur du maximum de vraisemblance du paramètre à estimer $\theta = (m_1, m_2, \sigma_1^2, \sigma_2^2, p)$ et d'apprécier graphiquement la qualité de l'estimation ainsi que la vitesse de convergence.

La démarche adoptée sera la suivante:

- simuler un échantillon i.i.d. de taille n issu de la loi de mélange

  $$f(x) = p\varphi_{m_1,\sigma_1^2}(x) + (1-p)\varphi_{m_2,\sigma_2^2}(x), \quad x \in \mathbb{R}$$

  en fixant n à 200 et en choisissant des valeurs telles que $0 < p < 1$, $m_1 \neq m_2$ et $\sigma_j^2 > 0$, pour $j = 1, 2$.

- vérifier que les données simulées suivent bien la loi demandée (pour cela, il peut être utile d'augmenter momentanément la taille de l'échantillon).

- estimer les paramètres du mélange grâce à l'algorithme EM en initialisant l'algorithme sur une grille de valeurs initiales pour explorer la sensibilité de l'algorithme aux choix de ces valeurs initiales

- tracer la densité estimée

  $$\widehat{f}(x) = \widehat{p}\varphi_{\widehat{m}_1,\widehat{\sigma}_1^2}(x) + (1-\widehat{p})\varphi_{\widehat{m}_2,\widehat{\sigma}_2^2}(x), \quad x \in \mathbb{R}$$

- superposer la courbe de la densité calculée avec les vrais valeurs utilisées pour simuler les observations.

- tracer la courbe de la densité estimée aux itérations $t = 0, 1, 2, 3, 4, 5$ dans des couleurs différentes.

Méthode pour simuler un échantillon i.i.d. issu d'une loi de mélange:

Indépendamment pour $i = 1, ..., n$, on tire $\mathbf{Z}_i = \begin{pmatrix} Z_i^{(1)} \\ \vdots \\ Z_i^{(K)} \end{pmatrix}$ de sorte que $\sum_{k=1}^{K} Z_i^{(k)} = 1$ et $Z_i^{(k)}(\Omega) = \{0, 1\}$ avec $\mathbb{P}(Z_i^{(k)} = 1) = p_k$ et $\mathbb{P}(Z_i^{(k)} = 0) = 1 - p_k$ pour $i = 1, ..., n$ et $k = 1, .., K$. Cela revient à tirer une variable discrète J_i à valeurs dans $\{1, ..., K\}$ avec $\mathbb{P}(J = k) = p_k$ pour $k = 1, ..., K$. Conditionnellement à $\{J = j\}$, on tire X_i selon une loi $\mathcal{N}(m_j, \sigma_j^2)$ de densité notée $\varphi_{m_j, \sigma_j^2}$.

NB: dans le cas particulier où $K = 2$, indépendamment pour $i = 1, .., n$,

- on tire J_i selon une loi $\mathcal{B}(p_1)$
- si $J_i = 1$, on tire X_i selon une loi $\mathcal{N}(m_1, \sigma_1^2)$,
  si $J_i = 0$, on tire X_i selon une loi $\mathcal{N}(m_2, \sigma_2^2)$.

NB2: lorsqu'on ne travaille pas avec un logiciel dédié à la statistique, il peut arriver que le logiciel en question ne comporte pas de fonction permettant de tirer J selon une loi discrète, mais seulement un générateur de loi uniforme sur $[0, 1]$. Dans ce cas, on peut procéder comme suit. On partitionne l'intervalle $[0, 1]$ en

$$[0,1] = \cup_{k=1}^{K-1} \underbrace{[p_0 + ...p_{k-1}; p_0 + ... + p_k[}_{:=I_k} \cup \underbrace{[p_0 + ...p_{K-1}; p_0 + ... + p_K[}_{:=I_K}$$

en notant $p_0 = 0$ et $p_K = 1$. Indépendamment pour $i = 1, ..., n$, on tire U_i selon une loi $\mathcal{U}(0, 1)$. Lorsque $U_i \in I_j$, on pose $J = j$ ou de manière équivalente $Z_i^{(j)} = 1$ et $Z_i^{(k)} = 0$ pour tout $k \neq j$.