TD 2 : Algorithme EM pour le mélange de lois de Gauss

M2 Statistique et économétrie
Séries temporelles

2010-11

L'objectif de ce TD est d'aider à mieux comprendre le modèle de mélange de lois de Gauss. On s'intéressera notamment à la simulation de réalisations de ce modèle et au problème de l'inférence dans les modèles de mélange.

Considérons un vecteur aléatoire $X \in \mathbb{R}^p$ distribué comme un mélange de M lois de Gauss de paramètres (μ_1, Σ_1), (μ_2, Σ_2), $\cdots$, (μ_M, Σ_M) avec les proportions $\pi_1, \cdots, \pi_M$ $(\pi_1 + \cdots + \pi_M = 1)$.

Question 1 - Modèle

1. Notons S la variable de classe. Donner la loi de S.
2. Ecrire la densité du vecteur X. On notera $\phi(.; \mu_k, \Sigma_k)$ la densité de la loi de Gauss de paramètres (μ_1, Σ_1).
3. Comment s'écrit cette densité si $p = 1$? Comment s'écrit cette densité si $M = 2$?

Question 2 - Inférence

Soient $p = 1$ et $M = 2$. Soient $x_1, \cdots, x_n$, n réalisations du vecteur X. On va utiliser l'algorithme EM pour estimer les paramètres du modèle.

Etape E

1. Donner le vecteur θ des paramètres à estimer. Quelle est sa dimension?
2. Montrer que la log-vraisemblance des données observées s'écrit

$$\log \mathcal{L}(\theta; x_1, \cdots, x_n) = \sum_{i=1}^{n} \sum_{i=1}^{M} \log(\pi_k \phi(x_i; \mu_k, \Sigma_k))$$

3. En déduire l'expression de la fonction $Q(\theta, \theta_0)$. Se reporter au cours pour la définition de la fonction Q.

Etape M

1. Ecrire les dérivées de $Q(\theta, \theta_0)$ en fonction de θ.

Question 3 - Mise en pratique de l'algorithme EM

1. Générer $n = 1000$ réalisations $\{x_1, \cdots, x_n\}$ d'une variable aléatoire X ayant pour densité de probabilité
$$p(x) = \frac{1}{3}\phi(x; 2, 1) + \frac{2}{3}\phi\left(x; -1, \frac{1}{2}\right)$$

2. Utiliser l'algorithme EM pour estimer les paramètres de X sachant $\{x_1, \cdots, x_n\}$. Vous pouvez utiliser les fonctions du package `mclust` ou la fonction `em_gauss_mixture.R` à télécharger (`http://www-labsticc.univ-ubs.fr/~monbet/doc/COURS/ST/em_gauss_mixture.R`).

3. Estimer le biais et la covariance des estimateurs par simulation. Ces propriétés dépendent-elles de la valeur initiale choisie pour l'estimation par EM?

4. Comparer les résulats à ceux l'estimation de la covariance des estimateurs obtenus dans l'algorithme EM.

Question 4 - Application

On considère les données du nombre de naissances et de morts par 1000 habitants pour 70 pays du monde. On souhaite modéliser la distribution de ces données par un modèle de mélange de 2 ou 3 lois de Gauss.

1. Télécharger les données
(`http://www-labsticc.univ-ubs.fr/~monbet/doc/COURS/ST/birth_death.txt`).
Tracer un nuage de points pour représenter les données.

2. Commenter le choix des modèles proposés.

3. Ajuster les 2 modèles, donner l'estimation des paramètres et interpréter les valeurs obtenues. On pourra utiliser une classification par `kmeans` pour initialiser l'optimisation par EM.

4. Superposer au nuage de points une représentation des densités estimées en utilisant la fonction `contour`.

5. Calculer le critère BIC pour les 2 modèles et commenter les résultats. Vous pourrez aussi comparer les résultats obtenus à ceux d'une classification par kmeans (après avoir rappelé en quoi consiste cette méthode de classification).