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

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. On s'intéressera notamment à la simulation de réalisations de ce modèle et au problème de l'inférence dans les modèles de mélange.

Nous considérons les données des nombres de tremblements de terre par an entre 1900 et 2006. Vous pouvez les trouver sur la page web du cours. Charger les données sous `R` et tracer la série temporelle.

1. Statistiques descriptives

1. Estimer les premiers moments (centrés réduits) du nombre de tremblements de terre par an.
2. Tracer un histogramme (ou diagramme en batons) normalisé illustrant la distribution des données.
3. Une des lois les plus usuelles pour un nombre d'occurences est la loi de Poisson. Que pensez-vous de ce modèle dans le cas présent? Pourquoi? Ajouter sur votre diagramme en batons la représentation de la loi de Poisson adhoc (utiliser par exemple la fonction `dpois`).

2. Modèle de mélange

Considérons une variable aléatoire $X \in \mathbb{N}$ distribuée comme un mélange de M lois de Poisson de paramètres $\lambda_1, \cdots, \lambda_M$ 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 loi de X sous la forme $P(X = x) = \cdots$.

Question 2 - Inférence

Soient $p = 1$ et $M = 2$. Soient $x_1, \cdots, x_n$, n réalisations i.i.d. 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 complètes s'écrit

$$\log \mathcal{L}(\theta; x_1, \cdots, x_n) = \sum_{i=1}^{n} \sum_{m=1}^{M} \delta_m(s_i) \log(\pi_m P_{\lambda_m}(X = x_i))$$

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

Etape M

1. Ecrire les dérivées de $Q(\theta, \theta_0)$ en fonction de θ et en déduire la formulation de l'étape M.

Question 3 - Mise en pratique de l'algorithme EM

1. Ecrire une fonction pour l'algorithme EM[1] dans le cas d'un mélange univarié de M lois de Poisson.
2. Nous allons valider le programme et étudier les propriétés des estimateurs. Générer un échantillon de taille $n = 107$ suivant un modèle de mélange de deux lois de Poisson de paramètres $\lambda_1 = 5$ et $\lambda_2 = 15$.
3. Estimer le biais et la covariance des estimateurs par simulation : simuler $B = 500$ échantillons ayant les mêmes propriétés que dans la questions précédente, pour chacun d'eux estimer les paramètre et en déduire une estimation du biais et de la covariance des estimateurs. Ces propriétés dépendent-elles de la valeur initiale choisie pour l'estimation par EM?
4. Comparer les résultats à ceux de l'approximation de la covariance des estimateurs obtenus par l'information de Fisher.

Question 4 - Application

On revient maintenant aux données.

1. Proposer une (ou plusieurs) modélisation(s) de la loi du nombre de tremblements de terre par an.
2. Estimer les paramètres du (ou des) modèles. Si vous avez considéré plusieurs modèles, calculer les critères BIC associés pour vous aider à choisir le "meilleur".
3. Valider votre modèle en comparant les moments théoriques et empiriques ainsi que les distributions.

[1] Dans R, le package `mclust= propose des algorithmes EM.`