INSA-GMM
Simulation Stochastique
2005-2006

TP de MATLAB du mercredi 14 juin 2006
Introduction à l'algorithme EM et SEM

Dans ce TP on va mettre en place deux algorithmes permettant la classification et l'estimation dans un modèle de mélange à deux composantes. Le modèle est le suivant. On considère les trois suites indépendantes de variables aléatoires i.i.d. (X_n), (Y_n) (T_n). X_n a pour densité f_θ et Y_n a pour densité g_η où θ et η sont des paramètres réels. On suppose que lorsque l'on observe la suite (X_n) (resp. (Y_n)) jusqu'à l'instant N on dispose de l'estimateur du maximum de vraisemblance $\widehat{\theta}_N$ de θ (resp. $\widehat{\eta}_N$ de η). T_n est une variable de loi de Bernoulli de paramètre $0 < p < 1$. On observe la suite $Z_n = T_n X_n + (1 - T_n) Y_n$ jusqu'à l'instant N. Il s'agit alors d'estimer les paramètres p, θ, η et de prédire les valeurs de $T_1, \dots, T_N$.

1. Algorithme EM

Initialisation
On choisit arbitrairement $\widehat{p}^0$ (par exemple 1/2). Soit $n_0 = [N\widehat{p}^0]$ On pose $X_i^0 = Z_i, i = 1, \dots, n_0$ et $Y_i^0 = Z_{n_0+i}, i = 1, \dots, N - n_0$.

Etape j
Estimation des paramètres
On calcule $\widehat{\theta}_n^j$ (resp. $\widehat{\eta}_n^j$) à partir de l'échantillon $X_i^j, i = 1, \dots, n_j$ (resp. $Y_i^j, i = 1, \dots, N - n_j$). On pose $\widehat{p}^j = n_j/N$.

Reclassement des variables
Pour chacune des observation Z_i on calcule

$$u_i^{j+1} = \frac{f_{\widehat{\theta}_n^j}(Z_i)}{f_{\widehat{\theta}_n^j}(Z_i) + g_{\widehat{\eta}_n^j}(Z_i)}.$$

n_{j+1} est alors le nombre de Z_i tels que $u_i^{j+1} > 0.5$. Ces Z_i permettent alors de construire l'échantillon $X_i^{j+1}, i = 1, \dots, n_{j+1}$, alors que les Z_i qui ne vérifient pas la condition précédente conduisent à l'échantillon $Y_i^{j+1}, i = 1, \dots, N - n_{j+1}$.
On arrête l'algorithme lorsque les estimateurs ont convergé.

2. Algorithme SEM

On procède comme dans l'algorithme EM sauf que la phase de ré-assignation des variables devient aléatoire on effectue un tirage au sort : à l'étape $j + 1$, on affecte Z_i à la population des X avec probabilité u_i^{j+1}.

3. Le TP

Expérimenter les deux algorithmes dans le cas de deux gaussiennes de même variance mais de moyennes différentes, pour des lois exponentielles de paramètres différents, puis pour des lois de Pareto.