Gaussian Mixture Model

何浩勋
16212799

EM algorithm for the ML fitting of the parameter metric mixture model

The mixture model is expressed as,

$$f(y_j;\Psi)=\sum_{i=1}^{g}\pi_i f_i(y_j;\theta_i)$$

where $\theta_i=(\mu_i,\sigma_i)$,

Let $\pi=(\pi_1,...,\pi_g)^T$, $\mu=(\mu_1,...,\mu_g)^T$, $\sigma=(\sigma_1,...\sigma_g)^T$, $\xi=(\mu^T,\sigma^T)^T$, $\Psi=(\pi^T,\xi^T)^T$, $y=(y_1^T,...,y_n^T)$.
The log likelihood for Ψ is given by,

$$\log L(\Psi)=\sum_{j=1}^{n}\log\left[\sum_{i=1}^{g}\pi_i f_i(y_j;\theta_i)\right]$$

Solving the likelihood equation,

$$\partial\log L(\Psi)/\partial\Psi=0$$

We have $\hat{\Psi}$, satisfies

$$\hat{\pi}_i = \sum_{j=1}^{n} \tau_i(y_j; \hat{\Psi})/n \quad (i = 1, ..., g)$$

and

$$\sum_{i=1}^{g} \sum_{j=1}^{n} \tau_i(y_j; \hat{\Psi}) \partial \log f_i(y_j; \hat{\theta}_i)/\partial \xi$$

where

$$\tau_i(y_j; \hat{\Psi}) = \pi_i f_i(y_j; \theta_i) / \sum_{h=1}^{g} \pi_h f_h(y_j; \theta_h)$$

If y is viewed as being incomplete, as the associated component vectors, $z = (z_1, ..., z_n)$ are not available, z_j is a g-dimensional vector with $z_{ij} = (z_j)_i = 1 \; or \; 0$, according to whether y_j is did or did not arise from the ith component of the mixture $(i = 1, ..., g \;, j = 1, ..., n)$. The complete vector is therefore declared to be,

$$y_c = (y^T, z^T)$$

The component-label vector $z_1, ..., z_n$ are taken to be the realized values of the random vectors $Z_1, ...Z_n$, where, for independent feature data, it is appropriate to assume that they are distributed unconditionally as,

$$Z_1, ..., Z_n \sim \mathrm{Mult}_g(1, \pi)$$

This assumption means that the sitribution of the complete-data vector Y_C implies the appropriate distribution of the incomplete-data vector Y. The complete-data log likelihood for $L_c(\Psi)$, is given by

$$\log L_c(\Psi) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\{\log\pi_i + \log f_i(y_j;\theta_i)\}$$

E-step

$$w_{ij} = E_{\Psi^{(k)}}(Z_{ij}|y) = \pi_i^{(k)} f_i(y_j;\theta_i^{(k)}) / \sum_{h=1}^{g} \pi_h f_h(y_j;\theta_h^{(k)})$$

M-step

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} w_{ij}/n \qquad (i = 1, ..., g)$$

$$\mu_i^{(k+1)} = \sum_{j=1}^{n} w_{ij} y_j / \sum_{j=1}^{n} w_{ij} \qquad (i = 1, ..., g)$$

$$\sigma_i^{(k+1)} = \sqrt{\sum_{j=1}^{n} w_{ij}(y_j - \mu_i^{(k+1)})^2 / \sum_{j=1}^{n} w_{ij}} \qquad (i = 1, ..., g)$$

Random starting values

Specigying a random start,

$$\mu_1^{(0)}, ..., \mu_g^{(0)} \sim N(\bar{y}, V)$$

$$\Sigma_0^{(0)} = V$$

$$\pi_1 = ... = \pi_g = 1/g$$

where $V = \sum_{j=1}^{n}(y_j - \bar{y})(y_j - \bar{y})^T/n$

Stopping criterion

$$l^{(k)} = logL_c(\Psi^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\{\log\pi_i + \log f_i(y_j;\theta_i)\}$$

$$a^{(k)} = (l^{(k+1)} - l^{(k)})/(l^{(k)} - l^{(k-1)}) \quad ,k > 1$$

$$l_A^{(k+1)} = l^{(k)} + \tfrac{1}{1-a^{(k)}}(l^{(k+1)} - l^{(k)})$$

The EM algorithm can be stopped if

$$|l_A^{(k+1)} - l_A^{(k)}| < tol$$

Simulation

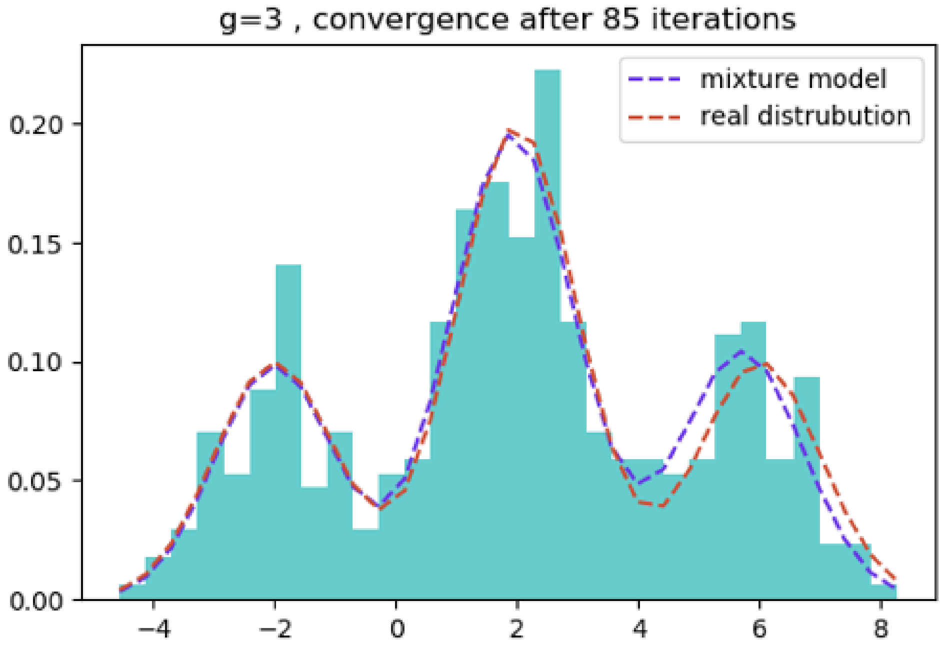
Example 1

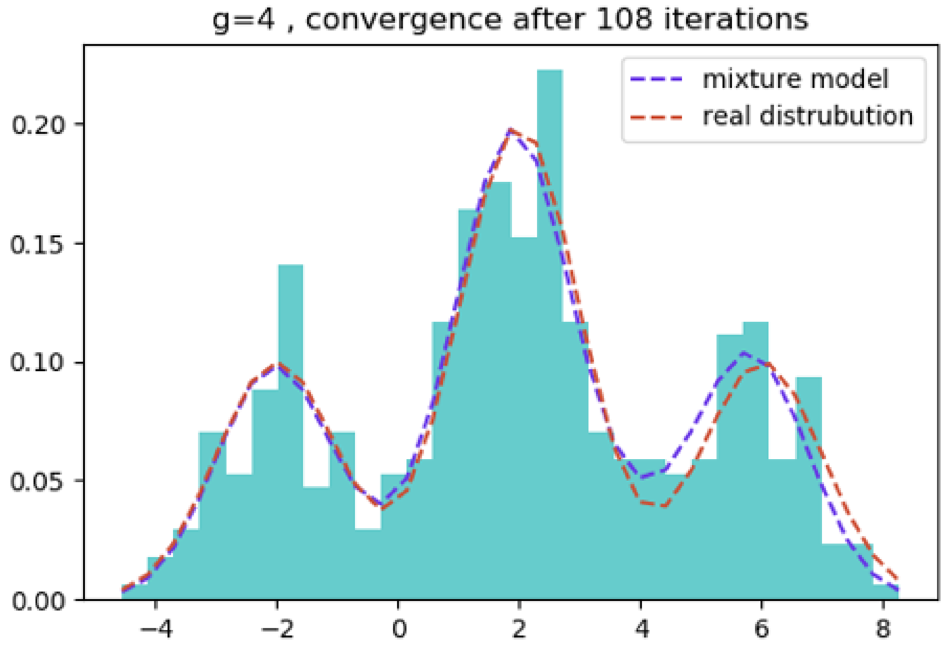
I randomly generated Y_1, Y_2, Y_3,

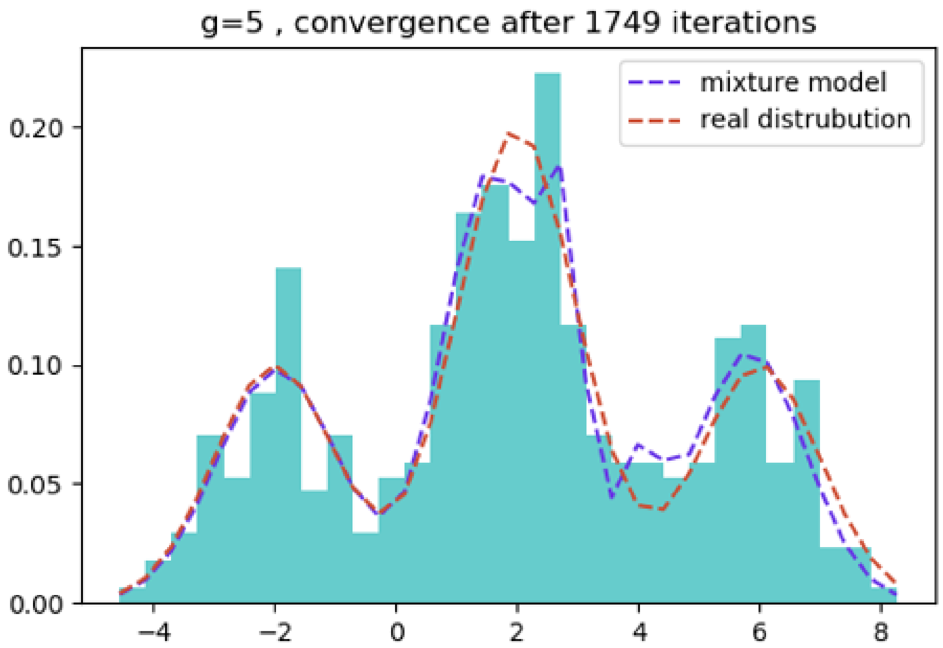
$$\begin{aligned} Y_1 &\sim N(-2, 1) \\ Y_2 &\sim N(2, 1) \\ Y_3 &\sim N(6, 1) \end{aligned}$$

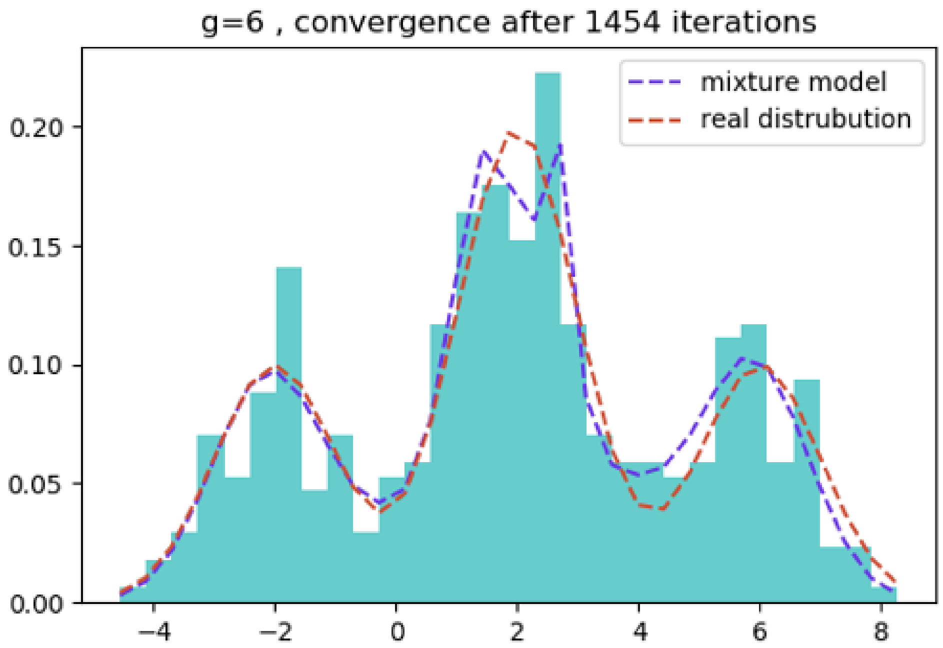
Y_1, Y_3 each have 100 observations, Y_2 have 200 observations, that is, 400 observations total. $Y = (Y_1^T, Y_2^T, Y_3^T)$

The mixture model fits the real distribution well when $g \leq 4$, we can see that there are signs of overfit when $g > 4$.


g=3 , convergence after 85 iterations
mixture model
real distrubution
0.00
0.05
0.10
0.15
0.20
−4
−2
0
2
4
6
8


g=4 , convergence after 108 iterations
mixture model
real distrubution
0.00
0.05
0.10
0.15
0.20
−4
−2
0
2
4
6
8


g=5 , convergence after 1749 iterations
mixture model
real distrubution
0.00
0.05
0.10
0.15
0.20
−4
−2
0
2
4
6
8


g=6 , convergence after 1454 iterations
mixture model
real distrubution
0.00
0.05
0.10
0.15
0.20
−4
−2
0
2
4
6
8

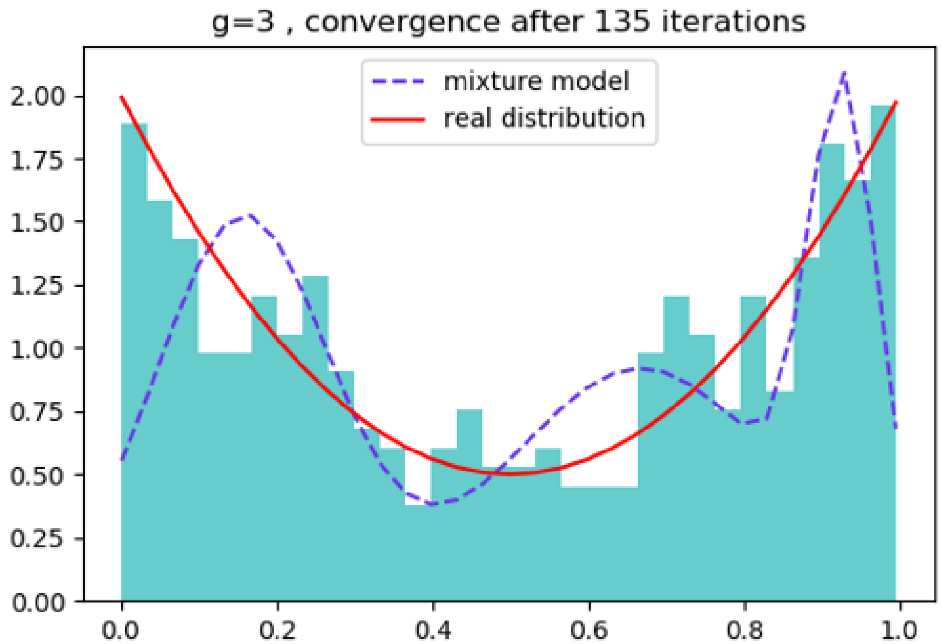
Example 2

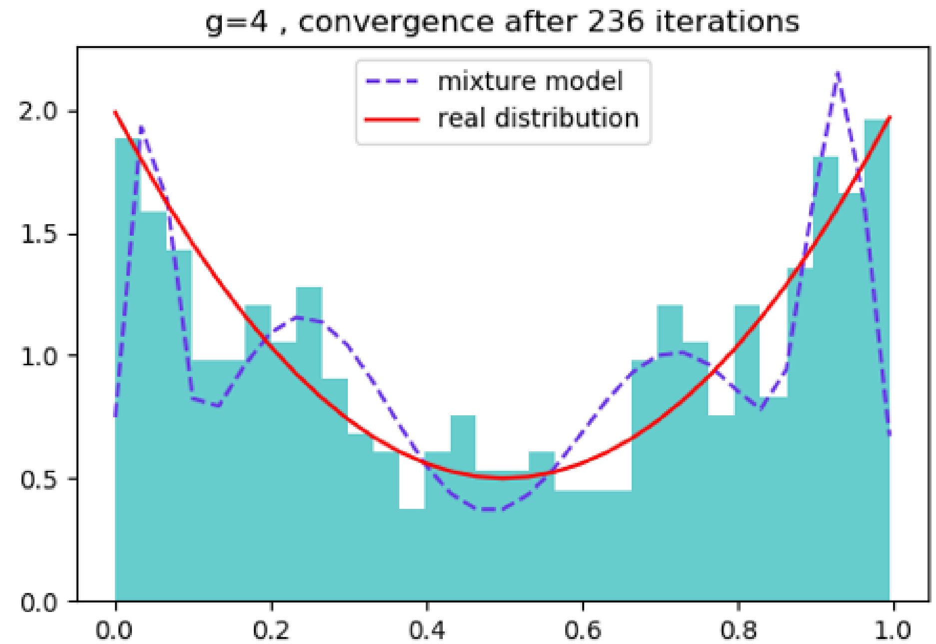
I randomly generated Y_1, Y_2,

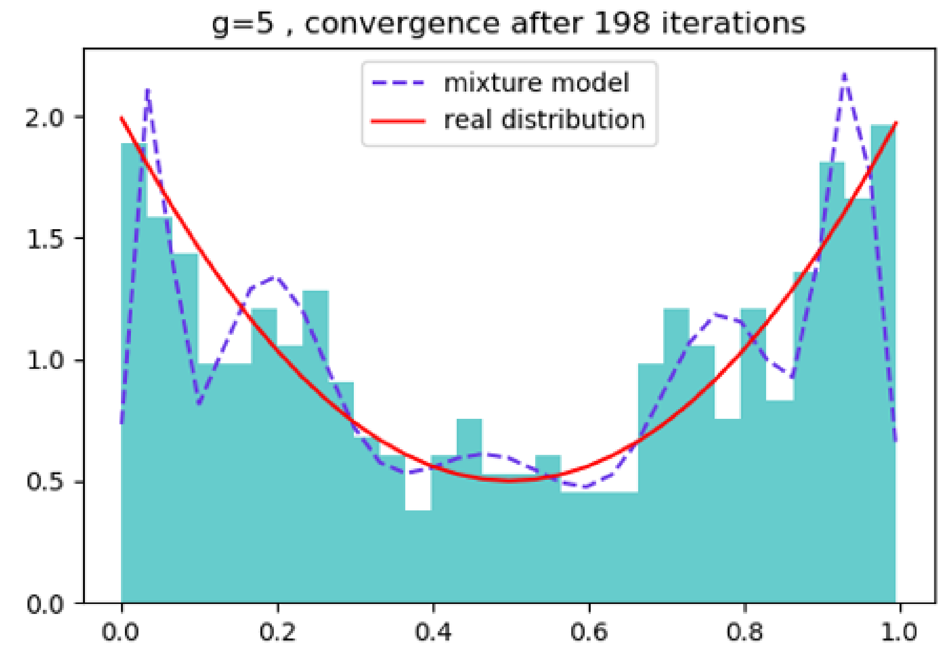
$$Y_1 \sim Beta(1,4)$$
$$Y_2 \sim Beta(4,1)$$

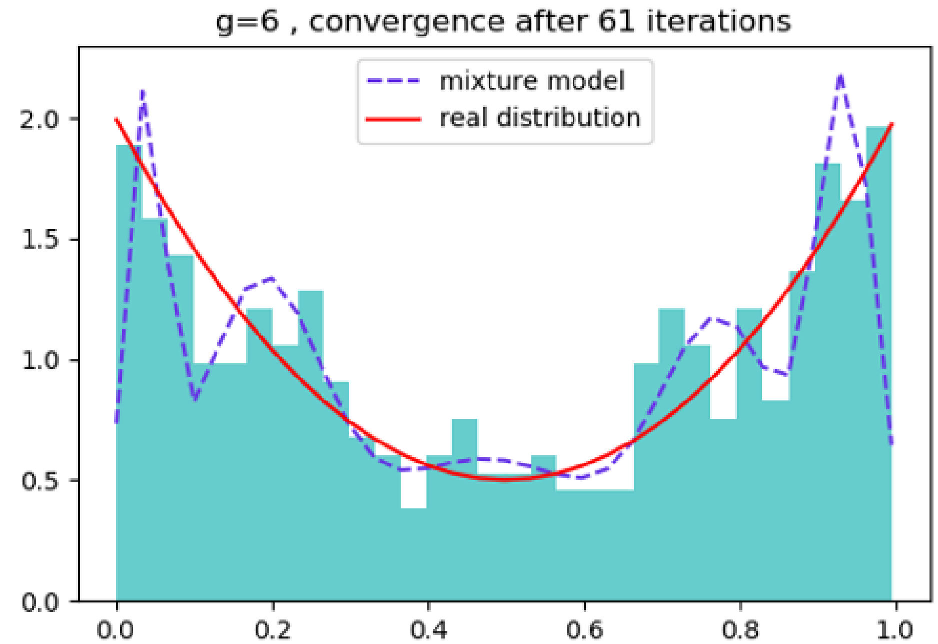
Y_1, Y_2 each have 200 observations, that is, 400 observations total. $Y = (Y_1^T, Y_2^T)$

When $g > 4$, I encountered a precision problem which causes 'Log(0) error', I had to stop EM algorithm and use the current result. However, the parameters seems converged.


g=3 , convergence after 135 iterations
mixture model
real distribution
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0


g=4 , convergence after 236 iterations
mixture model
real distribution
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0


g=5 , convergence after 198 iterations
mixture model
real distribution
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0


g=6 , convergence after 61 iterations
mixture model
real distribution
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0