The EM algorithm for multiple object recognition

Shotaro Akaho
Electrotechnical Laboratory
1–1–4 Umezono, Tsukuba-shi, Ibaraki 305 Japan
akaho@etl.go.jp

ABSTRACT

We propose a mixture model that can be applied to the recognition of multiple objects in an image plane. The model consists of any shape of modules; Each module is a probability density function of data points with scale and shift parameters, and the modules are combined with weight probabilities. We present the EM (Expectation-Maximization) algorithm to estimate those parameters. We also modify the algorithm in the case that data points are restricted in an attention window.

1. Introduction

We consider a mixture model that is a statistical model consisting of modules. Each module is a parametric probability distribution and those modules are integrated by taking a weighted sum of them. In order to make our discussion concrete, we consider the application to an image recognition problem. Suppose there are scattered points in an image plane and clusters of those points form objects. Our purpose is to fit a mixture model to those objects, where each module corresponds to the model of a certain object. We assume that objects in a real image are scaled or shifted from the original object model.

Recently, the EM (Expectation and Maximization) algorithm, which is a technique to find a locally maximum likelihood estimation from incomplete data, has attracted much attention because it can be applied the parameter estimation of many kinds of neural networks or related statistical models[4, 3, 5, 2]

In this paper, we consider unsupervised learning, whose statistical model is an unconditional probability distribution, and each module can be any kind of an object model. We adjust the scale and shift parameters to fit the module to an object. Each module is approximated by a normal mixture model and the parameters of the module are estimated in advance.

Another important technique to recognize multiple objects is to focus an "attention" to the target. When there are a lot of objects in a image, it is not easy to fit a complicated model. However, if we restrict the region to estimate parameters, the data out of the region should be treated as missing values.

We present the recurrence formulae of the estimation of the parameters of the mixture model with shift and scale parameters in section 3 and 4.

2. Mixture model

2.1. Mixture model

Suppose there are n probability density functions $g_i(\boldsymbol{x} \mid \boldsymbol{\theta}_i)$ $(i = 1, \ldots, n)$, where $\boldsymbol{x}$ is a d-dimensional random variable and $\boldsymbol{\theta}_i$ is a parameter vector. Each g_i is called "**module**". Mixture model is a probability density function defined as the weighted sum of those modules,

$$p(\boldsymbol{x} \mid \boldsymbol{\lambda}) = \sum_{i=1}^{n} \frac{\xi_i}{\sum_k \xi_k} g_i(\boldsymbol{x} \mid \boldsymbol{\theta}_i), \qquad \xi_i \in \mathcal{R}, \quad \xi_i > 0, \tag{1}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$, $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ and $\boldsymbol{\lambda} = (\boldsymbol{\xi}, \boldsymbol{\theta}^*)$. When we consider the application to object recognition, $\boldsymbol{x}$ is considered to be a 2-dimensional vector.

The goal of the maximum likelihood estimation (MLE) is to estimate $\boldsymbol{\xi}$ and $\boldsymbol{\theta}^*$ that maximizes the likelihood calculated from given samples of $\boldsymbol{x}$,

$$\sum_{\boldsymbol{x}} \log p(\boldsymbol{x} \mid \boldsymbol{\lambda}), \tag{2}$$

where the summation is taken over the whole samples.

Since it is difficult to estimate $\boldsymbol{\xi}$ and $\boldsymbol{\theta}^*$ directly, let us introduce a hidden random variable $z \in \{1, 2, \ldots, n\}$, and observed $\boldsymbol{x}$ is generated from the module $g_z(\boldsymbol{x} \mid \boldsymbol{\theta}_z)$. The joint distribution of $(\boldsymbol{x}, z)$ is written by

$$p(\boldsymbol{x}, z \mid \boldsymbol{\lambda}) = \frac{\xi_z}{\sum_k \xi_k} g_z(\boldsymbol{x} \mid \boldsymbol{\theta}_z). \tag{3}$$

If we regard observed sample x as the incomplete data of $(\boldsymbol{x}, z)$, we can apply the EM algorithm for the estimation.

It is also remarkable that we can combine any different types of modules. The difficulty of the estimation only depends on the form of each module g_i as long as the parameters of the modules are independent of each other. If g_i is a normal distribution, we obtain a simple form of the EM algorithm. In the following section, we introduce a mixture module which can approximate a wider class of distributions instead of a normal module. Shift and scale parameters of each module are successfully estimated by the EM algorithm, in the sense that the estimation is explicitly obtained in each EM step. The explicit form of the estimation is described in section 4. In the following sections, we omit subscripts of $g_i(x \mid \boldsymbol{\theta}_i)$ as $g(x \mid \boldsymbol{\theta})$ unless confusing.

2.2. Mixture module with scale and shift parameters

Let us consider a class of density functions

$$\{Af(\mathrm{diag}[\boldsymbol{a}]\,\boldsymbol{x} + \boldsymbol{b}) \mid a_\alpha > 0, \quad a_\alpha, b_\alpha \in \mathcal{R}\}, \tag{4}$$

where $f(\boldsymbol{x})$ is an arbitrary smooth density function, $\boldsymbol{a}$ and $\boldsymbol{b}$ denote a scale parameter and a shift parameter respectively, A is a normalization parameter $A = \prod_\alpha a_\alpha$, and $\mathrm{diag}[\boldsymbol{a}]$ is a diagonal matrix whose α-th diagonal element is a_α.

Of course, it is difficult to estimate $\boldsymbol{a}$ and $\boldsymbol{b}$ in general. Therefore we approximate $f(\boldsymbol{x})$ in advance by normal mixture model as follows.

$$\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{m} \frac{\zeta_j}{\sum_k \zeta_k} \phi(\boldsymbol{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2), \tag{5}$$

where $\phi(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ is an orthogonal normal distribution

$$\phi(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \prod_\alpha \frac{1}{\sqrt{2\pi {\sigma_\alpha}^2}} \exp\left\{ -\frac{(x_\alpha - \mu_\alpha)^2}{2{\sigma_\alpha}^2} \right\}. \tag{6}$$

According to the result of the density approximation theory, $f(\boldsymbol{x})$ can be approximated by $\hat{f}(\boldsymbol{x})$ as precisely as possible with sufficiently large m, when $f(\boldsymbol{x})$ satisfies some regularity conditions.

We adopt

$$\{A\hat{f}(\mathrm{diag}[\boldsymbol{a}]\,\boldsymbol{x} + \boldsymbol{b}) \mid a_\alpha > 0, \quad a_\alpha, b_\alpha \in \mathcal{R}\} \tag{7}$$

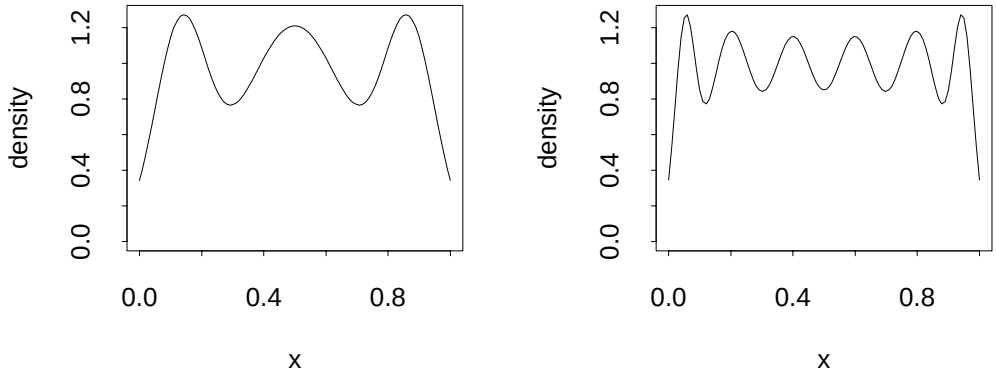
as a module g of mixture model instead of (4). Each Gaussian module included in $\hat{f}$ is called "**submodule**". However, even if we approximate f by $\hat{f}$, the estimation of $\boldsymbol{a}$ and $\boldsymbol{b}$ is not trivial, because $\boldsymbol{a}$ and $\boldsymbol{b}$ are not independent but common parameters in normal submodules included in $\hat{f}$.

Figure 1 shows a simple example of the approximation: f is the uniform distribution on $[0, 1]$. As the number of submodules increases, the accuracy of approximation increases. However, since the time complexity also increases, we should select an appropriate model based on AIC/MDL or other kinds of model selection criteria.

In order to apply the EM algorithm to our model, let us introduce another hidden random variable $w \in \{1, 2, \ldots, m\}$, and observed $\boldsymbol{x}$ is generated from w-th normal submodule $\phi(\boldsymbol{x} \mid \boldsymbol{\mu}_w, \boldsymbol{\sigma}_w^2)$ of z-th module g_z. The joint distribution of $(\boldsymbol{x}, z, w)$ is written by

$$p(\boldsymbol{x}, z, w \mid \boldsymbol{\lambda}) = \frac{\xi_z}{\sum_k \xi_k} \frac{\zeta_{z,w}}{\sum_k \zeta_{z,k}} A_z \phi(\mathrm{diag}[\boldsymbol{a}_z]\,\boldsymbol{x} + \boldsymbol{b}_z \mid \boldsymbol{\mu}_{z,w}, \boldsymbol{\sigma}_{z,w}^2), \tag{8}$$

where $\boldsymbol{\theta}^* = (\boldsymbol{a}_1, \boldsymbol{b}_1; \boldsymbol{a}_2, \boldsymbol{b}_2; \ldots; \boldsymbol{a}_n, \boldsymbol{b}_n)$ and $\boldsymbol{\lambda} = (\boldsymbol{\xi}, \boldsymbol{\theta}^*)$.



(a): $m = 3$ (b): $m = 6$

Fig. 1: Example of mixture module (1-dim). The uniform distribution on $[0,1]$ is approximated by normal mixture model in the sense of MLE (not squared error). Mean log-likelihood values of (a) and (b) are -0.0600 and -0.0269 respectively.

3. Estimation in an attention window

To give an attention is an efficient method for the recognition of objects and the learning of environments. However, when we focus our attention to a certain region, the data out of the region are censored. We can apply the EM algorithm in such a case. In this section, we show a slight modification of the EM algorithm for that problem.

Let C be the region of attention. Suppose random variable $\boldsymbol{x}$ is observed only when $\boldsymbol{x} \in C$, and the number of data $\boldsymbol{x} \notin C$ is unknown.

Since the EM algorithm can be applied when the number of data is known, we consider the following algorithm where the number of missing values are estimated in each EM step. The main loop of the algorithm of this paper is as follows.

The main loop of the algorithm

1. Let $\boldsymbol{\lambda}^{(0)}$ be an initial parameter, and $M^{(0)}$ an initial number of missing values (usually 0).
2. Repeat the following two steps for $t = 0, 1, 2, \ldots$,
   (a) Apply an EM step based on N observed samples and $M^{(t)}$ missing samples. Let $\boldsymbol{\lambda}^{(t+1)}$ be the solution.
   (b) Estimate $M^{(t+1)}$ with fixing $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(t)}$.

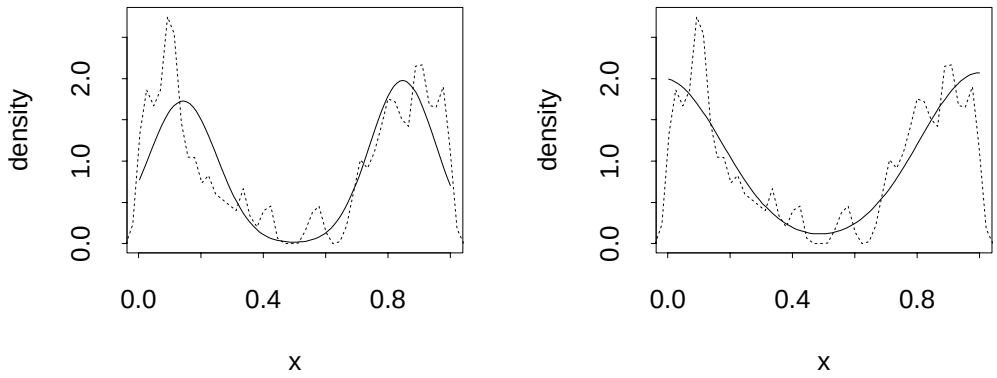
When we fix parameters $\boldsymbol{\lambda}$, observations can be regarded as Bernoulli trials, where each observation is censored in probability $P(\boldsymbol{\lambda}^{(t)})$ and observed in probability $1 - P(\boldsymbol{\lambda}^{(t)})$, where $P(\boldsymbol{\lambda}) \equiv \int_{C^c} p(\boldsymbol{x} \mid \boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{x}$ and C^c is a complement of C. Thus **the number of censored data** can be estimated by

$$M^{(t+1)} = \frac{P(\boldsymbol{\lambda}^{(t)})}{1 - P(\boldsymbol{\lambda}^{(t)})} N. \tag{9}$$

We show a simple simulation results for this algorithm in figure 2.

4. The EM algorithm for the mixture model

In this section, we give an explicit form of the EM algorithm for the mixture model in an attention window C (See [1] about the detail of the derivation).



(a) Usual EM algorithm (b) Proposed algorithm

Fig. 2: The EM algorithm for attention region (1-dim). 100 sample are generated on $C = [0, 1]$ with respect to the mixture $0.5\phi(x, 0, 0.2^2) + 0.5\phi(x, 1, 0.2^2)$. Sample density(dotted) and estimated density(solid) are plotted.

4.1. Weight value estimation

The recurrence formula for the weight value $\boldsymbol{\xi}$ is given by

$$\xi_i^{(t+1)} = \sum_{\boldsymbol{x}} p(i \mid \boldsymbol{x}, \boldsymbol{\lambda}^{(t)}) \ + M^{(t)} Q_i(\boldsymbol{\lambda}^{(t)}), \tag{10}$$

where $p(i \mid \boldsymbol{x}, \boldsymbol{\lambda}) = \dfrac{p(\boldsymbol{x}, i \mid \boldsymbol{\lambda})}{p(\boldsymbol{x} \mid \boldsymbol{\lambda})}$ is the conditional probability of hidden variable, which can be calculated from (1) and (3); and $Q_i(\boldsymbol{\lambda}) \equiv \dfrac{1}{P(\boldsymbol{\lambda})} \displaystyle\int_{C^e} p(\boldsymbol{x}, i \mid \boldsymbol{\lambda}) \, \mathrm{d}\boldsymbol{x}$.

4.2. Scale and shift parameter estimation

Scale and shift parameters of mixture module described in 2.2 can be explicitly estimated in each EM step by optimizing separately, namely, estimating $\boldsymbol{a}$ with fixing $\boldsymbol{b}$ after(or before) estimating $\boldsymbol{b}$ with fixing $\boldsymbol{a}$.

In order to simplify the recurrence formulae, let us consider the conditional distribution of hidden variables,

$$q_{i,j}(\boldsymbol{x}, \boldsymbol{\lambda}) \equiv p(i, j \mid \boldsymbol{x}, \boldsymbol{\lambda}) = \frac{p(\boldsymbol{x}, i, j \mid \boldsymbol{\lambda})}{p(\boldsymbol{x} \mid \boldsymbol{\lambda})}, \tag{11}$$

where i denotes the current module and j denotes the normal submodule of the i-th module, and let us consider the moment-like values defined by

$$R_{i,j,\alpha}^{\langle k \rangle}(\boldsymbol{\lambda}) \equiv \frac{1}{P(\boldsymbol{\lambda})} \int_{C^e} (x_\alpha)^k \, p(\boldsymbol{x}, i, j \mid \boldsymbol{\lambda}) \, \mathrm{d}\boldsymbol{x}. \tag{12}$$

$q_{i,j}$ can be calculated from (1) and (8).

Since those parameters can be estimated independently for each module, we omit the subscript i below.

The recurrence formula for the scale parameter $\boldsymbol{a}$ with fixing shift parameter is given by

$$a_\alpha^{(t)} = \frac{\sqrt{(Y_\alpha)^2 + 4X_\alpha Z_\alpha} - Y_\alpha}{2X_\alpha}, \tag{13}$$

where

$$X_\alpha \ = \ \sum_j \frac{1}{{\sigma_j}^2} \left[\sum_{\boldsymbol{x}} (x_\alpha)^2 \, q_j(\boldsymbol{x}, \boldsymbol{\lambda}^{(t)}) \ + M^{(t)} R_{j,\alpha}^{\langle 2 \rangle}(\boldsymbol{\lambda}^{(t)}) \right], \tag{14}$$

$$Y_\alpha = \sum_j \frac{b_\alpha - \mu_{j,\alpha}}{{\sigma_j}^2} \left[\sum_{\boldsymbol{x}} x_\alpha q_j(\boldsymbol{x}, \boldsymbol{\lambda}^{(t)}) \ + M^{(t)} R^{\langle 1 \rangle}_{j,\alpha}(\boldsymbol{\lambda}^{(t)}) \right], \tag{15}$$

$$Z_\alpha = (N + M^{(t)}) \, \xi_i^{(t+1)} \sum_j \zeta_j . \tag{16}$$

The recurrence formula for the shift parameter $\boldsymbol{b}$ with fixing shift parameter is given by

$$b_\alpha^{(t)} = \frac{V_\alpha}{U}, \tag{17}$$

where

$$U = \sum_j \frac{1}{{\sigma_j}^2} \left[\sum_{\boldsymbol{x}} q_j(\boldsymbol{x}, \boldsymbol{\lambda}^{(t)}) \ + M^{(t)} Q(\boldsymbol{\lambda}^{(t)}) \right], \tag{18}$$

$$V_\alpha = \sum_j \frac{1}{{\sigma_j}^2} \left[\sum_{\boldsymbol{x}} (\mu_{j,\alpha} - a_\alpha x_\alpha) \, q_j(\boldsymbol{x}, \boldsymbol{\lambda}^{(t)}) \ + M^{(t)} \left\{ \mu_{j,\alpha} Q(\boldsymbol{\lambda}^{(t)}) - a_\alpha R^{\langle 1 \rangle}_{j,\alpha}(\boldsymbol{\lambda}^{(t)}) \right\} \right]. \tag{19}$$

It is difficult to calculate Q's and $R^{\langle k \rangle}$'s in general. However, when C is a hyper-rectangle that is parallel to axes, we can calculate those values from the error function.

5. Simple simulation results

Let us define a module "face" that consists of 3 normal submodules: two 'eyes' and one 'mouth'. We show a simulation result in figure 3. There are 100 sample points in a image generated from the mixture of two face modules (fig.3(a)): One face is in the upper right and scaled by 0.8 along x-axis and 0.7 along y-axis from the template face, and another face is in the lower left and scaled by 0.9 and 1.1. Although those faces look very vague, the parameters converged to the correct answer. However, they don't converge in the following cases: (1) there is almost no overlap between samples points and the initial solution, (2) there is a lot of overlap between modules of the initial solution.

6. Concluding remarks

We have presented the EM algorithm for a mixture model with mixture modules that can approximate any kinds of the models of objects. We have also considered the case that data are restricted in an attention region.

We can consider the algorithm to estimate the parameters of mixture modules other than shift and scale parameters. However, it causes unstable behavior in general and the estimation can be done just for minor tuning. M. Revow et al[6] have proposed the algorithm to make the algorithm more stable by a kind of regularization method.

It is difficult to get appropriate initial values of parameters especially when the number of modules is large. It is necessary to give an attention in a smaller area. We have not mentioned how to decide the attention region. That problem is related to the field of active vision or active learning.

It is also a future task to apply our algorithm to real world images.

Acknowledgement

The author would like to thank Dr. Suwa and Dr. Otsu of Electrotechnical Laboratory, for affording an opportunity of this study. He also expresses his thanks to all members of Mathematical Informatics Section for their helpful discussions. A part of this work is supported by Real World Computing Program.

References

[1] S. Akaho, "Mixture model for image understanding and the EM algorithm," Tech. Rep. TR-95-13, Electrotechnical Laboratory, 1995. ftp://etlport.etl.go.jp/pub/akaho/ETL-TR-95-13.ps.Z.

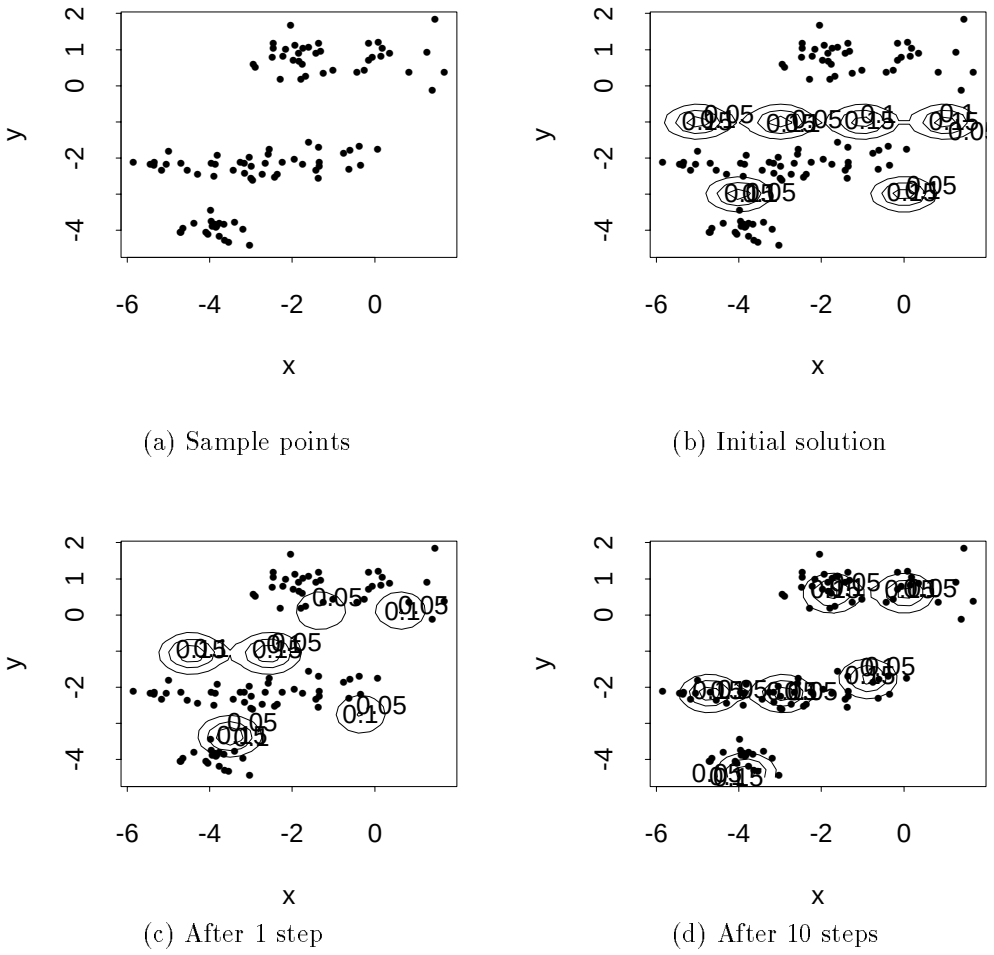
[2] S. Amari, "Information geometry of the EM and em algorithms for neural networks," Tech. Rep. METR 94–04, University of Tokyo, 1994. (to appear in Neural Networks).

[3] S. Amari, K. Kurata, and H. Nagaoka, "Information geometry of Boltzmann machines," *IEEE Trans. Neural Networks*, vol. 3, no. 2, 1992.

[4] L. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[5] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," in *Proc. of IJCNN'93*, (Nagoya), pp. 1339–1344, 1993.

[6] M. Revow, C. Williams, and G. Hinton, "Using generative models for handwritten digit recognition,". to appear IEEE Trans. on PAMI.



(a) Sample points (b) Initial solution

(c) After 1 step (d) After 10 steps

Fig. 3: A simulation result for "face" mixture model (2-dim). Dots: sample points, Solid circles: contour plot of estimated density function. The parameters converged to the correct answer. Mean log-likelihood values of (b), (c) and (d) are -11.8, -5.39 and -3.12 respectively.