UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2009:17

On Simulation Methods for Two Component Normal Mixture Models under Bayesian Approach

Liwen Liang

Examensarbete i matematisk statistik, 30 hp
Handledare och examinator: Silvelyn Zwanzig

September 2009

Department of Mathematics
Uppsala University

Abstract

EM-Algorithm and Gibbs sampler are two useful Bayesian simulation methods for parameter estimation of finite normal mixture model. The EM-Algorithm is an iterative estimate of maximum likelihood for incomplete data problem. Gibbs sampler is an approach of generating random sample from a multivariate distribution. We introduce and derive Dempster EM-Algorithm for the two-component normal mixture models to get the iterative computation estimates, also use data augmentation and general Gibbs sampler to get the sample from posterior distribution under conjugate prior. The estimate results from both simulation methods under two-component normal mixture model with unknown mean parameters are compared and the connections and differences between both methods are represented. Data set from astronomy is used for comparison.

Acknowledgement

I would like to thank my supervisor Silvelyn Zwanzig for the patience, guidance and encouragement that she always gave to me, not only in the thesis, but also in the whole procedure of my statistics studying.

I would also like to thank my friend Han Jun for the the assistances of LATEX, thank Alena for the data source, and thank my parents for the spiritual and substantial support and wholesouled love they gave me all my life.

At last I would like to thank the department of mathematics of Uppsala University for giving me the opportunity to study.

Contents

1 Introduction

Bayesian approach has been widely used in social sciences and continuously developed for statistical analysis after 1990s when Markov chain Monte Carlo was discovered. There are a lot of simulation method for parameter estimation of finite mixture models, especially finite mixture of exponential family distribution. We focus on the EM-Algorithm and Gibbs sampler. Both of them are useful tools for solving difficult computation problem of finite mixture models.

The EM-Algorithm is an iterative estimate of maximum likelihood for incomplete data[1] problem. Gibbs sampler is an approach of generating random sample from a multivariate distribution. Finite normal mixture model is a classical example of how EM-Algorithm fits with incomplete data situation and we focus on the two-component normal mixture model int this thesis.

In Section 1, we first introduce the two-component normal mixture model with some basic properties. Then in Section 2 we propose an iterative plug-in procedure for solving nonlinear likelihood function to get the parameter estimators and give the EM-Algorithm. The monotonicity and convergency of the EM-Algorithm are discussed. After that, in Section 3 we present an indicator latent normal mixture model to get the Bayesian posterior distribution and introduce the Gibbs sampler. At last, compare the parameter estimation results got from both algorithm procedures in Section 4. We find that both simulation methods get similar estimation results, the EM-Algorithm is stabler but contains less information, and Gibbs sampler is just opposite.

1.1 Data source and notation

Two data sets are used for the parameter estimation comparison. One is the fictitious data set from Hastie, Tibshirani and Friedman(2001), which contains 20 data. The other is a real astronomic data set which observed by FLAMES-ARGUS, a spectrograph on the Very Large Telescope (VLT)[2] in 2006. It contains a group of 81 datum range from -1250.6222 to 502910.81. Detailed background of the astronomic data see Appendix A.

[1]The incomplete data situation means where there are missing data, truncated distributions or censored or grouped observations. In fact, the EM-Algorithm can also be applied to the whole variety of situations where the incompleteness of the data is not so obvious, such as latent variable structure which will be introduced later for our model. Details see [10].

[2]The VLT is a system which built and operated by the European Southern Observatory (ESO). The VLT is constituted of four separate optical telescopes: the Antu telescope, the Kueyen telescope, the Melipal telescope and the Yepun telescope. The VLT provides the total light collecting power of a 16 meter single telescope, making it the largest optical telescope in the world.

Figure 1: Picture of Planetary Nebulae (PNe)

We follow the symbols and procedure forms in [1] primarily to take the basic expression, but also give forms from other books and authors for comparison. Now introduce the two-component normal mixture model.

1.2 Two-component normal mixture model

Suppose Y is a mixture model of two normal distributions: $Y_1 \sim N(\mu_1, \sigma_1^2)$, $Y_2 \sim N(\mu_2, \sigma_2^2)$, Y_1 and Y_2 are independent. Then we have the two-component normal mixture model:

$$Y = (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2, \tag{1.1}$$

where $\Delta = 0$ or 1 with $Pr(\Delta = 1) = \pi$, which means $\Delta \sim Ber(\pi)$. Δ is called the *unobservable* or *latent* data.

Let $\theta = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ denote the unknown parameter vector, where $\theta_1 = (\mu_1, \sigma_1^2)$ and $\theta_2 = (\mu_2, \sigma_2^2)$. Denote the parameter space by $\mathbf{\Omega}$, where

$$\mathbf{\Omega} = [0, 1] \times \mathcal{R}^2 \times \mathcal{R}_+^2$$

Let $\varphi_{\theta_j}(y)$, j=1 or 2 denote the normal density of each component. Then the density of two-component normal mixture model Y under parameter θ is:

$$g_Y(y) = (1 - \pi)\varphi_{\theta_1}(y) + \pi\varphi_{\theta_2}(y) \tag{1.2}$$

Since Y_1, Y_2 are independent and Δ is i.i.d, we have the following basic properties of Y:

$$E[Y|\theta] = (1 - \pi)\mu_1 + \pi\mu_2$$
$$E[Y^2|\theta] = (1 - \pi)(\mu_1 + \sigma_1^2) + \pi(\mu_2 + \sigma_2^2)$$

2 Iterative estimation and the EM-Algorithm

We are interested in the parameter estimation. First introduce the training data $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$, where Z_i i.i.d $\sim Y$, $Y_{1i} \sim N(\mu_1, \sigma_1^2)$, $Y_{2i} \sim N(\mu_2, \sigma_2^2)$, $\Delta_i \sim Ber(\pi)$, i.i.d., $i = 1, \ldots, N$, which has form:

$$Z_i = (1 - \Delta_i)Y_{1i} + \Delta_i Y_{2i},$$

and

$$Z_i|\Delta_i = 1 \sim N(\mu_2, \sigma_2^2)$$

$$Z_i|\Delta_i = 0 \sim N(\mu_1, \sigma_1^2)$$

$$p(Z_i|\theta) = (1 - \pi)\varphi_{\theta_1}(y_i) + \pi\varphi_{\theta_2}(y_i)$$

From equation (1.1) and (1.2), the log-likelihood of model (1.1) based on the N training cases $\mathbf{Z}$ is:

$$\ell(\theta; \mathbf{Z}) = \sum_{i=1}^{N} log[(1 - \pi)\varphi_{\theta_1}(y_i) + \pi\varphi_{\theta_2}(y_i)] \tag{2.1}$$

Then we have the local derivations under each parameter:

$$(*) \begin{cases} \dfrac{\partial \ell}{\partial \pi} = \displaystyle\sum_{i=1}^{N} \frac{\varphi_{\theta_2}(y_i) - \varphi_{\theta_1}(y_i)}{(1 - \pi)\varphi_{\theta_1}(y_i) + \pi\varphi_{\theta_2}(y_i)} \\ \dfrac{\partial \ell}{\partial \mu_1} = \displaystyle\sum_{i=1}^{N} \frac{(1 - \pi)\varphi_{\theta_1}(y_i)}{(1 - \pi)\varphi_{\theta_1}(y_i) + \pi\varphi_{\theta_2}(y_i)} \cdot \frac{(y_i - \mu_1)}{\sigma_1^2} \\ \dfrac{\partial \ell}{\partial \sigma_1^2} = \displaystyle\sum_{i=1}^{N} \frac{(1 - \pi)\varphi_{\theta_1}(y_i)}{(1 - \pi)\varphi_{\theta_1}(y_i) + \pi\varphi_{\theta_2}(y_i)} \cdot \frac{1}{2\sigma_1^2} \left(\frac{(y_i - \mu_1)^2}{\sigma_1^2} - 1 \right) \\ \dfrac{\partial \ell}{\partial \mu_2} = \displaystyle\sum_{i=1}^{N} \frac{\pi\varphi_{\theta_2}(y_i)}{(1 - \pi)\varphi_{\theta_1}(y_i) + \pi\varphi_{\theta_2}(y_i)} \cdot \frac{(y_i - \mu_2)}{\sigma_2^2} \\ \dfrac{\partial \ell}{\partial \sigma_2^2} = \displaystyle\sum_{i=1}^{N} \frac{\pi\varphi_{\theta_2}(y_i)}{(1 - \pi)\varphi_{\theta_1}(y_i) + \pi\varphi_{\theta_2}(y_i)} \cdot \frac{1}{2\sigma_2^2} \left(\frac{(y_i - \mu_2)^2}{\sigma_2^2} - 1 \right) \end{cases}$$

We want to solve $\partial \ell / \partial \theta = 0$ to get the MLEs. Trying to maximize $\ell(\theta; \mathbf{Z})$ directly for the estimation of parameters is quite difficult since those equations are nonlinear and no analytic solutions can be found. So numerical procedure like iterative optimization methods often be used to get successive approximation of the solution. In that case we focus on the EM-algorithm in this section. First consider the idea of iterative plug-in procedure.

2.1 Iterative estimation for nonlinear likelihood function

From the local derivation equation system $(*)$, we notice that expression

$$\frac{(1-\pi)\varphi_{\theta_1}(y_i)}{(1-\pi)\varphi_{\theta_1}(y_i)+\pi\varphi_{\theta_2}(y_i)}$$

or

$$\frac{\pi\varphi_{\theta_2}(y_i)}{(1-\pi)\varphi_{\theta_1}(y_i)+\pi\varphi_{\theta_2}(y_i)}$$

exists in the last four equations, and for the first equation, we also have:

$$\begin{aligned}\frac{\partial \ell}{\partial \pi} &= \sum_{i=1}^{N}\frac{\varphi_{\theta_2}(y_i)-\varphi_{\theta_1}(y_i)}{(1-\pi)\varphi_{\theta_1}(y_i)+\pi\varphi_{\theta_2}(y_i)} \\ &= \sum_{i=1}^{N}\left(\frac{\pi\varphi_{\theta_2}(y_i)}{(1-\pi)\varphi_{\theta_1}(y_i)+\pi\varphi_{\theta_2}(y_i)}\cdot\frac{1}{\pi}-\frac{(1-\pi)\varphi_{\theta_1}(y_i)}{(1-\pi)\varphi_{\theta_1}(y_i)+\pi\varphi_{\theta_2}(y_i)}\cdot\frac{1}{(1-\pi)}\right)\end{aligned}$$

So we denote

$$\gamma_i(\theta)=\frac{\pi\varphi_{\theta_2}(y_i)}{(1-\pi)\varphi_{\theta_1}(y_i)+\pi\varphi_{\theta_2}(y_i)} \tag{2.2}$$

and introduce $\gamma_i(\theta)$ into $(*)$, then we get the plug-in derivation equation system as followed:

$$(**)\begin{cases}\dfrac{\partial \ell}{\partial \pi}=\displaystyle\sum_{i=1}^{N}\left(\frac{\gamma_i(\theta)}{\pi}-\frac{1-\gamma_i(\theta)}{1-\pi}\right)\\ \dfrac{\partial \ell}{\partial \mu_1}=\displaystyle\sum_{i=1}^{N}\frac{(1-\gamma_i(\theta))(y_i-\mu_1)}{\sigma_1^2}\\ \dfrac{\partial \ell}{\partial \sigma_1^2}=\displaystyle\sum_{i=1}^{N}\frac{1-\gamma_i(\theta)}{2\sigma_1^2}\left(\frac{(y_i-\mu_1)^2}{\sigma_1^2}-1\right)\\ \dfrac{\partial \ell}{\partial \mu_2}=\displaystyle\sum_{i=1}^{N}\frac{\gamma_i(\theta)(y_i-\mu_2)}{\sigma_2^2}\\ \dfrac{\partial \ell}{\partial \sigma_2^2}=\displaystyle\sum_{i=1}^{N}\frac{\gamma_i(\theta)}{2\sigma_2^2}\left(\frac{(y_i-\mu_2)^2}{\sigma_2^2}-1\right)\end{cases}$$

Let $\partial\ell/\partial\theta=0$, then rewrite these equation systems $(*)$ and $(**)$ as matrix forms respectively:

$$\begin{aligned}L'(\hat{\theta}_{MLE}) &= \mathbf{0}\\ \mathbf{H}(\gamma(\hat{\theta}_{MLE}),\hat{\theta}_{MLE}) &= \mathbf{0}\end{aligned} \tag{2.3}$$

We want to solve (2.3) and get the parameter estimator as:

$$\hat{\theta}_{MLE} = \mathbf{h}(\gamma(\hat{\theta}_{MLE}))$$

Notice that $\hat{\theta}$ is also contained in $\gamma(\hat{\theta})$, which is called the *plug − in estimatior*, so we propose an iterative procedure to solve problem (2.3):

1. Let $\hat{\theta}^{(j)}$ denote the current jth parameter statement, so $\hat{\gamma}_i^{(j)} = \gamma(\hat{\theta}^{(j)})$ is the current plug-in estimator, where:

$$\hat{\gamma}_i^{(j)} = \frac{\hat{\pi}^{(j)}\varphi_{\hat{\theta}_2^{(j)}}(y_i)}{(1-\hat{\pi}^{(j)})\varphi_{\hat{\theta}_1^{(j)}}(y_i) + \hat{\pi}^{(j)}\varphi_{\hat{\theta}_2^{(j)}}(y_i)}$$

2. Introduce $\hat{\gamma}_i^{(j)}$ into equation (2.3) and get the iterative plug-in parameter estimator $\hat{\theta}_r^{(j+1)}$ as:

$$\hat{\theta}_r^{(j+1)} = (\hat{\pi}^{(j+1)}, \hat{\mu}_1^{(j+1)}, (\hat{\sigma}_1^2)^{(j+1)}, \hat{\mu}_2^{(j+1)}, (\hat{\sigma}_2^2)^{(j+1)})$$

where

$$\begin{aligned}
\hat{\pi}^{(j+1)} &= \frac{1}{N}\sum_{i=1}^{N}\hat{\gamma}_i^{(j)} \\
\hat{\mu}_1^{(j+1)} &= \frac{\sum_{i=1}^{N}(1-\hat{\gamma}_i^{(j)})y_i}{\sum_{i=1}^{N}(1-\hat{\gamma}_i^{(j)})} \\
(\hat{\sigma}_1^2)^{(j+1)} &= \frac{\sum_{i=1}^{N}(1-\hat{\gamma}_i^{(j)})(y_i-\hat{\mu}_1^{(j+1)})^2}{\sum_{i=1}^{N}(1-\hat{\gamma}_i^{(j)})} \\
\hat{\mu}_2^{(j+1)} &= \frac{\sum_{i=1}^{N}\hat{\gamma}_i^{(j)}y_i}{\sum_{i=1}^{N}\hat{\gamma}_i^{(j)}} \\
(\hat{\sigma}_2^2)^{(j+1)} &= \frac{\sum_{i=1}^{N}\hat{\gamma}_i^{(j+1)}(y_i-\hat{\mu}_2^{(j+1)})^2}{\sum_{i=1}^{N}\hat{\gamma}_i^{(j+1)}}
\end{aligned} \tag{2.4}$$

By this iterative procedure, we get the parameter estimator as:

$$\hat{\theta}^{(j+1)} = \mathbf{h}(\gamma(\hat{\theta}^{(j)}))$$

If we have some initial value of θ as $\hat{\theta}^{(0)}$, we can continue this iterative procedure for $j = 1, 2, \ldots, n, \ldots$ and consider the iterative parameter estimator $\hat{\theta}^{(n)}$, $n \to \infty$ as the approximation of the solution for (2.3) if it is convergent. This iterative procedure is similar as the iteratively reweighted least squares. The EM-Algorithm has equal form with iteratively reweighted least squares for normal mixture model and we will discuss the general form of the EM-Algorithm in the following section.

2.2 The EM-Algorithm

The earliest literature related to an EM-type algorithm appears in Newcomb(1886) with estimation of parameters of a mixture of two univariate normal models. McKendrick(1926) also gives a method with basic EM-Algorithm spirit.

The formulation of EM algorithm is first introduced by Dempster, Laird and Rubin in 1977. The convergency and other basic properties of the EM-Algorithm under general conditions was established in their literature.

Before recommending the EM-Algorithm procedure, we introduce the *latent variable normal mixture model*(we called it *latent model* for short in this thesis) first to give a further comprehension for the latent variable. The latent model is another form for the two-component normal mixture model which consider the latent data Δ as part of the model (also see [1]).

2.2.1 Latent variable model

Take a review of the original model (1.1)

$$Y = (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2,$$

where $\Delta \sim Ber(\pi)$. For training data $\mathbf{Z} = \{Z_1, \ldots, Z_N\}$, where Z_i i.i.d $\sim Y$, $i = 1, \ldots, N$, we have

$$Z_i = (1 - \Delta_i)Y_{1i} + \Delta_i Y_{2i},$$

and the log-likelihood function:

$$\ell(\theta; \mathbf{Z}) = \sum_{i=1}^{N} log[(1 - \pi)\varphi_{\theta_1}(y_i) + \pi\varphi_{\theta_2}(y_i)].$$

Now we consider unobserved latent variables Δ_i: when $\Delta_i = 0$, Y_i comes from $N(\mu_1, \sigma_1^2)$, when $\Delta_i = 1$, Y_i comes from $N(\mu_2, \sigma_2^2)$. If we knew the value of Δ_i's, then get the *latent variable model* with log-likelihood as followed:

$$\ell_0(\theta; \mathbf{Z}, \Delta) = \sum_{i=1}^{N} [(1 - \Delta_i) log\varphi_{\theta_1}(y_i) + \Delta_i log\varphi_{\theta_2}(y_i)] \tag{2.5}$$

Actually the values of Δ_i's is unknown, so we want to use the expected value of $\Delta_i|\theta, \mathbf{Z}$ to substitute each Δ_i in (2.5). We have

$$\begin{aligned} E(\Delta_i|\theta, \mathbf{Z}) &= p(\Delta_i = 1|\theta, Z_i) \\ &= \frac{p(\Delta_i = 1, Z_i|\theta)}{p(Z_i|\theta)} = \frac{p(Z_i|\theta, \Delta_i = 1)p(\Delta_i = 1|\theta)}{P(Z_i|\theta)} \\ &= \frac{\varphi_{\theta_2}(y_i)\pi}{(1 - \pi)\varphi_{\theta_1}(y_i) + \pi\varphi_{\theta_2}(y_i)} \\ &= \gamma_i(\theta) \end{aligned} \tag{2.6}$$

In fact, $\gamma_i(\theta)$ was defined as the expected value of $\Delta_i|\theta, \mathbf{Z}$, which called *responsibility* in [1]. Now give the definition as followed:

Definition 1. *The expected value of* $\Delta_i|\theta, \mathbf{Z}$ *is called the* ***responsibility*** *of model for observation i, denoted as* $\gamma_i(\theta)$:

$$\gamma_i(\theta) = E(\Delta_i|\theta, \mathbf{Z})$$

Then we have the expected latent log-likelihood function as:

$$\begin{aligned}
\ell_0(\theta; \mathbf{Z}) &= E_{\Delta_i}(\ell_0(\theta; \mathbf{Z}, \Delta)|\theta, Z_i) \\
&= E_{\Delta_i}\left(\sum_{i=1}^{N}[(1-\Delta_i)log\varphi_{\theta_1}(y_i) + \Delta_i log\varphi_{\theta_2}(y_i)]\right) \\
&= \sum_{i=1}^{N}[E_{\Delta_i}((1-\Delta_i)log\varphi_{\theta_1}(y_i)) + E_{\Delta_i}(\Delta_i log\varphi_{\theta_2}(y_i))] \\
&= \sum_{i=1}^{N}[(1-\gamma_i)log\varphi_{\theta_1}(y_i) + \gamma_i log\varphi_{\theta_2}(y_i)] \\
&= -\frac{1}{2}Nlog(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\big[(1-\gamma_i)(log\sigma_1^2 + (y_i-\mu_1)^2/\sigma_1^2) \\
&\quad +\gamma_i(log\sigma_2^2 + (y_i-\mu_2)^2/\sigma_2^2)\big] \qquad (2.7)
\end{aligned}$$

Obviously $\hat{\pi} = \frac{1}{N}\sum_{i=1}^{N}\hat{\gamma}_i$ since $\sum_{i=1}^{N}\hat{\gamma}_i$ is the estimation of expected number that Y_i comes from $N(\mu_2, \sigma_2^2)$. Then set $\partial\ell_0/\partial\mu_j = \partial\ell_0/\partial\sigma_j^2 = 0$, and follow the iterative procedure in Section 2.1, we can get the same iterative parameter estimators as (2.4).

The latent data Δ is part of the model in our latent normal mixture situation. In general problems the latent data could also be actual data that should been observed but missing. Now come to the EM-Algorithm procedures.

2.2.2 General Dempster EM-Algorithm

Denote $\mathbf{Z}$ as the observed incomplete data, $\mathbf{Z}^m$ as the latent data (or missing data), and $\mathbf{T} = (\mathbf{Z}, \mathbf{Z}^m)$ as the unobserved complete data, where $t(\mathbf{T}) = \mathbf{Z}$ collapses $\mathbf{T}$ to $\mathbf{Z}$. First we give the general form for Dempster EM-Algorithm (also see [2] and [12]):

Algorithm 1. *Dempster EM algorithm*

1. *Choose a start value* $\hat{\theta}^{(0)}$.

2. Expectation Step: at the jth step, calculate

$$Q(\theta|\hat{\theta}^{(j)}) = E[\ln(f(\mathbf{T})|\theta), \mathbf{Z}, \hat{\theta}^{(j)}]$$

3. Maximization Step: deterimine the new estimate $\hat{\theta}^{(j+1)}$ *as*

$$\hat{\theta}^{(j+1)} = \arg\max Q(\theta|\hat{\theta}^{(j)})$$

4. Iterate step 2 and 3 until convergence.

The essence of the EM algorithm is that maximizing $Q(\theta', \theta)$ leads to an increase in the log-likelihood of the observed data. Remind the Jensen's inequality and information inequality which prove this property:

Proposition 2.1. *(Jensen' Inequality). Assume that the values of the random variable* W *are confined to the possibly infinite interval* (a, b)*. If* $h(w)$ *is convex on* (a, b)*, then* $E[h(W)] \geq h[E(W)]$*, provided both expectations exist. For a strictly convex function* $h(w)$*, equality holds in Jensen's inequality if and only if* $W = E(W)$ *almost surely.*

Proof. See [12]. □

Proposition 2.2. *(Information Inequality). Let* f *and* g *be probability densities with respect to a measure* μ*. Suppose* $f > 0$ *and* $g > 0$ *almost everywhere relative to* μ*. If* E_f *denotes expectation with respect to the probability measure* $f d\mu$*, then* $E_f(\ln f) \geq E_f(\ln g)$*, with equality only if* $f = g$ *almost everywhere relative to* μ*.*

Proof. See [12]. □

Now give the fundamental property for the EM-Algorithm in general (also see [12]).

Proposition 2.3. *Suppose that* $g(Z|\theta)$ *and* $f(T|\theta)$ *are the probability densities of the observed and complete data, respectively. Then the EM iterates obey*

$$\ln g(Z|\hat{\theta}^{(j+1)}) \quad \geq \quad \ln g(Z|\hat{\theta}^{(j}),$$

with strict inequality when $f(T|\hat{\theta}^{(j+1)})/g(Z|\hat{\theta}^{(j+1)})$ *and* $f(T|\hat{\theta}^{(j)})/g(Z|\hat{\theta}^{(j)})$ *are different conditional densities or when the surrogate function* $Q(\theta|\hat{\theta}^{(j)}))$ *satisfies*

$$Q(\hat{\theta}^{(j+1)}|\hat{\theta}^{(j)}) \quad > \quad Q(\hat{\theta}^{(j)}|\hat{\theta}^{(j)})$$

Proof. (See [12]) Since both $f(T|\theta)/g(Z|\theta)$ and $f(T|\hat{\theta}^{(j)})/g(Z|\hat{\theta}^{(j)})$ are conditional densities of $\mathbf{T}$ on $\{T : t(T) = Z\}$ with respect to some measure μ_Z, then by the information inequality, for $Q(\theta|\hat{\theta}^{(j)}) = E[\ln(f(\mathbf{T})|\theta), \mathbf{Z} = Z, \hat{\theta}^{(j)}]$, we have:

$$\begin{aligned} Q(\theta|\hat{\theta}^{(j)}) - \ln g(Z|\theta) &= E[\ln(\frac{f(\mathbf{T}|\theta)}{g(\mathbf{Z}|\theta)})|\mathbf{Z} = Z, \hat{\theta}^{(j)}] \\ &\leq E[\ln(\frac{f(\mathbf{T}|\hat{\theta}^{(j)})}{g(\mathbf{Z}|\hat{\theta}^{(j)})})|\mathbf{Z} = Z, \hat{\theta}^{(j)}] \\ &= Q(\hat{\theta}^{(j)}|\hat{\theta}^{(j)}) - \ln g(Z|\hat{\theta}^{(j)}) \end{aligned}$$

Hence the difference $Q(\theta|\hat{\theta}^{(j)}) - \ln g(Z|\theta)$ attains its maximum when $\theta = \hat{\theta}^{(j)}$.

If we choose $\hat{\theta}^{(j+1)}$ to maximize $Q(\theta|\hat{\theta}^{(j)})$, then we have:

$$\begin{aligned} \ln g(Z|\hat{\theta}^{(j+1)}) &= Q(\hat{\theta}^{(j+1)}|\hat{\theta}^{(j)}) - [Q(\hat{\theta}^{(j+1)}|\hat{\theta}^{(j)}) - \ln g(Z|\hat{\theta}^{(j+1)})] \\ &\geq Q(\hat{\theta}^{(j)}|\hat{\theta}^{(j)}) - [Q(\hat{\theta}^{(j)}|\hat{\theta}^{(j)}) - \ln g(Z|\hat{\theta}^{(j)})] \\ &= \ln g(Z|\hat{\theta}^{(j)}) \end{aligned}$$

□

Proposition 2.3 provide that the EM iteration never decreases the log-likelihood of observed data, hence the EM algorithm works in general and also be called *generalized* EM algorithm(GEM).

2.2.3 Dempster EM-Algorithm for latent model

Now apply Dempster EM-Algorithm to latent model. We have $\mathbf{T} = (\mathbf{Z}, \mathbf{\Delta})$ and log-likelihood for complete data as (2.5):

$$\ln f(\mathbf{T}|\Delta, \theta) = \ell_0(\theta; \mathbf{Z}, \Delta) = \sum_{i=1}^{N} [(1 - \Delta_i) log\varphi_{\theta_1}(y_i) + \Delta_i log\varphi_{\theta_2}(y_i)]$$

Then

$$Q(\theta|\hat{\theta}^{(j)}) = E_{\Delta|\hat{\theta}^{(j)}}[\ell_0(\theta; \mathbf{Z}, \Delta)] = \ell_0(\theta; \mathbf{Z}, \gamma(\hat{\theta}^{(j)})),$$

according to (2.7). Hence the Dempster EM-Algorithm for latent model is as followed:

Algorithm 2. *Dempster EM algorithm for latent model*

1. *Choose a start value* $\hat{\theta}^{(0)}$.
2. *Expectation Step: at the jth step, calculate*

$$Q(\theta|\hat{\theta}^{(j)}) = \ell_0(\theta; \mathbf{Z}, \gamma(\hat{\theta}^{(j)})),$$

where $\ell_0(\theta; \mathbf{Z}, \gamma(\hat{\theta}^{(j)}))$ *defined as (2.7)*

3. Maximization Step: deterimine the new estimate $\hat{\theta}^{(j+1)}$ *as*

$$\hat{\theta}^{(j+1)} = \arg\max Q(\theta|\hat{\theta}^{(j)})$$

4. Iterate step 2 and 3 until convergence.

2.2.4 HTF EM-Algorithm

We consider the iterative plug-in parameter estimator given in Section 2.1 as the root of $\arg\max \ell_0(\theta; \mathbf{Z}, \Delta)$ by Iteration Theorem.

Since the result of iterative estimation for both original normal mixture model (1.2) and latent model (2.5) are the same, we get a concrete form of the EM-Algorithm procedure for two-component normal mixture model, called HTF EM-Algorithm [3], as followed:

Algorithm 3 (also see [1])**.** *HTF EM-Algorithm*

1. Take initial guesses for the parameters $\hat{\pi}, \hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2$.

2. Expectation Step: compute the responsibilities

$$\hat{\gamma}_i = \frac{\hat{\pi}\varphi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\varphi_{\hat{\theta}_1}(y_i) + \hat{\pi}\varphi_{\hat{\theta}_2}(y_i)}, i = 1, 2, \ldots, N.$$

3. Maximization Step: compute the weighted means and variances:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)},$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i},$$

and $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$.

4. Iterate steps 2 and 3 until convergence.

HTF EM-Algorithm procedure is substantively the same as iterative procedure introduced in Section 2.2.1. For the initial guess for the parameters, usually take $\hat{\pi} = 0.5$, take two of y_i randomly as the initial guesses for $\hat{\mu}_1$ and $\hat{\mu}_2$, and take $\hat{\sigma}_1^2 = \hat{\sigma}_2^2 = \Sigma_{i=1}^N (y_i - \bar{y})^2 / N$. Details see [1].

[3] We call this algorithm as HTE EM-Algorithm since Hastie, Tibshirani and Friedman (first) introduce this algorithm procedure in [1].

2.3 Convergency of the EM-Algorithm

We just prove the monotonicity of the EM-Algorithm in Section 2.2.2 and now we are interested in the relationship between the iterative plug-in parameter estimator sequence $\{\hat{\theta}^{(n)}\}$, $n \to \infty$ and the maximum likelihood estimator $\hat{\theta}_{MLE}$. Wu (1983) have given the convergency properties of the EM-Algorithm and we take a statement of the main convergency theorem here without a proof.

Theorem 2.1 (Also see [24] and [26]). *Let $\{\hat{\theta}^{(j)}\}$ be an instance of a GEM algorithm generated by $\hat{\theta}^{(j+1)} = \arg\max Q(\theta|\hat{\theta}^{(j)})$. Suppose that*

1. *$\hat{\theta}^{(j+1)} = \arg\max Q(\theta|\hat{\theta}^{(j)})$ is closed over the complement of $\mathcal{S}$, the set of stationary points in the interior of $\mathbf{\Omega}$ and,*

2. *Have*
$$\ell_0(\theta; \mathbf{Z}, \gamma(\hat{\theta}^{(j+1)})) > \ell_0(\theta; \mathbf{Z}, \gamma(\hat{\theta}^{(j)}))$$
for all $\{\hat{\theta}^{(j)}\}$ do not belong to $\mathcal{S}$.

Then all the limit points of $\{\hat{\theta}^{(j)}\}$ are stationary points and $\ell_0(\theta; \mathbf{Z}, \gamma(\hat{\theta}^{(j)}))$ converges monotonically to $\ell_0^* = \ell_0^*(\theta; \mathbf{Z}, \gamma(\hat{\theta}^*))$ for some stationary point $\hat{\theta}^*$.

Wu(1983) and Hartley & Hocking(1971) have given some very useful theorems about the convergency properties for an EM sequence(details see [24], pp.82-84) and we focus on the situation when there is only when stationary point in the interior of $\mathbf{\Omega}$. Now we have the convergency theorem for the iterative plug-in parameter estimator sequence to a unique maximum likelihood estimate.

Theorem 2.2 (Also see [24]). *Suppose that $\ell_0(\theta; \mathbf{Z})$ is unimodal in $\mathbf{\Omega}$ with $\hat{\theta}^*$ being the only stationary point and that $\partial Q(\theta, \varphi)\partial\theta$ is continuous in θ and φ. Then any EM sequence $\{\theta^{(j)}\}$ converges to the unique maximizer θ^* of $\ell_0(\theta; \mathbf{Z})$. That is, it converges to the unique MLE of θ.*

Proof. See [24]. □

There are many other methods for computing MLE, like the Newton-Raphson and Fisher's scoring method. EM-Algorithm has several advantages relative to those iterative algorithms, such as the numerically stability, reliable global convergence, easy implementation, small storage space requirement, and so on. It also has a few disadvantages, like the slowly convergency. But the disadvantages do not prevent the EM-Algorithm from being one of the most appealing iterative simulation methods for finding MLE. Details see [24].

3 Posterior distribution and Gibbs sampler

In this section, we give the posterior distribution for two-component normal mixture model under a conjugate prior and present the Gibbs sampler for both general case and normal mixture application. Gibbs sampler is a useful simulation method which generates sample from the posterior distribution.

3.1 Indicator normal mixture model

Calculating the posterior under conjugate prior for normal mixture model (1.2)

$$g_Y(y) = (1-\pi)\varphi_{\theta_1}(y) + \pi\varphi_{\theta_2}(y)$$

is very complicated (see [3]), so we introduce the zero-one component indicator vector variable $z = \{z_1, z_2\}$ [4] into the model to make the calculation easier and the result more straightforward. Consider

$$z_{ij} = \begin{cases} 1, & \text{if } y_i \sim \varphi_{\theta_j}(y_i) \\ 0, & \text{otherwise} \end{cases}$$

Since y_i can only belongs to one normal model for each i, we have $z_{1i} + z_{2i} = 1$. Also z_{ij} is unobservable and with conditional expectation as

$$\tau_{ij} = E[z_{ij}|\mathbf{y}] = \frac{\varphi_{\theta_j}(y_i)\pi_j}{\pi_1\varphi_{\theta_1}(y_i) + \pi_2\varphi_{\theta_2}(y_i)}$$

where $\pi_1 = 1-\pi$, $\pi_2 = \pi$.

Now see the joint density of the observed data $\mathbf{y}$ and unobserved data $\mathbf{z}$:

$$f(\mathbf{y}, \mathbf{z}; \theta) = f(\mathbf{z}, \theta) \cdot f(\mathbf{y}|\mathbf{z}, \theta)$$

From the definition of z_j, we have

$$\begin{aligned} f(z, \theta) &= (1-\pi)^{z_1}\pi^{z_2} \\ f(y|z_1 = 1, \theta) &= \varphi_{\theta_2}(y) = (\varphi_{\theta_2}(y))^{z_1} \\ f(y|z_2 = 1, \theta) &= \varphi_{\theta_2}(y) = (\varphi_{\theta_2}(y))^{z_2} \end{aligned}$$

So we get the indicator normal mixture model with joint density

$$f(\mathbf{y}, \mathbf{z}; \theta) = \prod_{i=1}^{N} [(1-\pi)\varphi_{\theta_1}(y_i)]^{z_{1i}} \cdot [\pi\varphi_{\theta_2}(y_i)]^{z_{2i}} \tag{3.1}$$

[4] Here the indicator vector variable z is the same as Δ in Section 2, in fact $z_1 = 1 - \Delta$ and $z_2 = \Delta$. We change the symbol just for the conveniency of later calculation.

The log-likelihood function under indicator normal mixture model is:

$$\ell(\theta) = \sum_{i=1}^{N}[z_{1i}log(1-\pi) + z_{2i}log\pi] + \sum_{i=1}^{N}[z_{1i}log\varphi_{\theta_1}(y_i) + z_{2i}log\varphi_{\theta_2}(y_i)]$$

If we take the iterative procedure as Section 2.2.1 into model (3.1), by using τ_{ij} instead of z_{ij} and letting

$$\begin{aligned} \partial\ell(\theta)/\partial\pi_j &= 0 \\ \partial\ell(\theta)/\partial\mu_j &= 0 \\ \partial\ell(\theta)/\partial\sigma_j^2 &= 0, \end{aligned}$$

we get the same iterative plug-in-estimators as:

$$\begin{aligned} \hat{\pi}_j^{(k+1)} &= \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_{ij}^{(k)} \\ \hat{\mu}_j^{(k+1)} &= \frac{\sum_{i=1}^{N}\hat{\tau}_{ij}^{(k)}y_i}{\sum_{i=1}^{N}\hat{\tau}_{ij}^{(k)}} \\ (\hat{\sigma}_j^2)^{(k+1)} &= \frac{\sum_{i=1}^{N}\hat{\tau}_{ij}^{(k)}(y_i-\hat{\mu}_i^{(k+1)})^2}{\sum_{i=1}^{N}\hat{\tau}_{ij}^{(k)}} \end{aligned}$$

So there is no difference in parameter estimation result among original normal mixture model, latent model and indicator normal mixture through the EM-Algorithm. More details of g-component normal mixture model see [10] pp. 16-20 and pp. 68-71.

3.2 Conjugate prior and posterior

Choosing a proper prior distribution is very important for Bayesian method, since the improper prior may not lead to an analytical tractable form of posterior. Specify a conjugate prior can guarantee the easily calculable form of the posterior. First give the definition of *conjugate*.

Definition 2. *A family $\mathcal{F}$ of probability distributions on Θ is said to be* ***conjugate*** *for a likelihood function $f(x|\Theta)$ if, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.*

We already know the inverse gamma and normal priors are conjugate for the variance and mean parameter in normal distribution model (see [3], [22] and [9]). Now we give the following theorem of expounding the conjugacy property for the normal, inverse gamma and beta prior for two-component normal mixture model.

Theorem 3.1. *(Conjugacy of the normal, inverse gamma and beta prior for two-component normal mixture model) In the normal mixture model, a normal prior along with a mixture normal joint likelihood function produced a normal posterior for the mean parameter; an inverse gamma prior with the same mixture normal likelihood produced an inverse gamma posterior for the variance parameter; a beta prior with the same mixture normal likelihood produced a beta posterior for the proportion parameter.*

Proof. Suppose the prior distributions are followed:

$$\begin{aligned} \mu_j|\sigma_j^2 &\sim \mathcal{N}\left(\xi_j, \frac{\sigma_j^2}{n_j}\right) \\ \sigma_j^2 &\sim \mathcal{IG}\left(\frac{\nu_j}{2}, \frac{s_j^2}{2}\right) \\ \pi &\sim \mathcal{B}e\big(\alpha, \beta\big) \end{aligned} \tag{3.2}$$

with density functions

$$\begin{aligned} p(\mu_j|\xi_j, \frac{\sigma_j^2}{n_j}) &\propto (\sigma_j^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_j^2/n_j}(\mu_j - \xi_j)^2\right] \\ p(\sigma_j^2|\frac{\nu_j}{2}, \frac{s_j^2}{2}) &\propto (\sigma_j^2)^{-(\frac{\nu_j}{2}+1)} \exp\left[-\frac{s_j^2}{2\sigma_j^2}\right] \\ p(\pi|\alpha, \beta) &\propto \pi^{\alpha-1}(1-\pi)^{\beta-1} \end{aligned}$$

For indicator normal mixture model (3.1), we have the following joint distribution under all parameters

$$p(\theta|\mathbf{y}, \mathbf{z}) \;=\; p(\mathbf{y}, \mathbf{z}|\theta)p(\pi|\alpha, \beta)\prod_{j=1}^{2}\left[p(\sigma_j^2|\frac{\nu_j}{2}, \frac{s_j^2}{2})p(\mu_j|\xi_j, \frac{\sigma_j^2}{n_j})\right]$$

$$\begin{aligned} = \;& (1-\pi)^{\sum_{i=1}^N z_{1i}+\beta-1}\pi^{\sum_{i=1}^N z_{2i}+\alpha-1} \times \prod_{i=1}^{N}\big(\phi_{\theta_1}(y_i)\big)^{z_{1i}}\prod_{i=1}^{N}\big(\phi_{\theta_2}(y_i)\big)^{z_{2i}} \\ & \times \prod_{j=1}^{2}\left[(\sigma_j^2)^{-\frac{\nu_j}{2}-\frac{3}{2}}\exp\left(-\frac{s_j^2}{2\sigma_j^2} - \frac{1}{2\sigma_j^2/n_j}(\mu_j-\xi_j)^2\right)\right] \end{aligned} \tag{3.3}$$

Denote $\bar{z}_j = \sum_{i=1}^N z_{ij}$ and $\bar{y}_j(z) = \frac{1}{\bar{z}_j}\sum_{i=1}^N z_{ij}y_i$. From (3.3) we get the

posterior distribution for $\theta = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ as followed:

$$\begin{aligned}
p(\theta|\mathbf{y},\mathbf{z}) &\propto (1-\pi)^{\bar{z}_1+\beta-1}\pi^{\bar{z}_2+\alpha-1}\prod_{j=1}^{2}(\sigma_j^2)^{-\frac{\bar{z}_j}{2}-\frac{\nu_j}{2}-\frac{3}{2}} \\
&\quad \times\prod_{j=1}^{2}\exp\left[-\frac{s_j^2}{2\sigma_j^2}-\frac{1}{2\sigma_j^2}\sum_{i=1}^{N}z_{ij}(y_i-\mu_j)^2-\frac{1}{2\sigma_j^2/n_j}(\mu_j-\xi_j)^2\right] \\
&= (1-\pi)^{\bar{z}_1+\beta-1}\pi^{\bar{z}_2+\alpha-1}\prod_{j=1}^{2}(\sigma_j^2)^{-\frac{\bar{z}_j}{2}-\frac{\nu_j}{2}-\frac{3}{2}}\prod_{j=1}^{2}\exp\left[-\frac{s_j^2}{2\sigma_j^2}\right. \\
&\quad \left.-\frac{1}{2\sigma_j^2}\left(\sum_{i=1}^{N}z_{ij}y_i^2-2\bar{z}_j\bar{y}_j(z)\mu_j+\bar{z}_j\mu_j^2\right)-\frac{1}{2\sigma_j^2/n_j}(\mu_j^2-2\mu\xi_j+\xi_j^2)\right] \\
&= (1-\pi)^{\bar{z}_1+\beta-1}\pi^{\bar{z}_2+\alpha-1}\prod_{j=1}^{2}(\sigma_j^2)^{-\frac{\bar{z}_j}{2}-\frac{\nu_j}{2}-\frac{1}{2}}\cdot\exp\left[-\frac{s_j^2}{2\sigma_j^2}-\frac{1}{2\sigma_j^2}\left(\sum_{i=1}^{N}z_{ij}y_i^2-k_jy^2\right)\right] \\
&\quad \times\prod_{j=1}^{2}(\sigma_j^2)^{-1}\cdot\exp\left[-\frac{1}{2\sigma_j^2}\left((\bar{z}_j+n_j)\mu_j^2-2(\bar{z}_j\bar{y}_j(z)+\xi_jn_j)\mu_j+(k_jy^2+n_j\xi_j^2)\right)\right]
\end{aligned} \tag{3.4}$$

where

$$k_jy^2 = \frac{\left(\bar{z}_j\bar{y}_j(z)+\xi_jn_j\right)^2-(\bar{z}_j+n_j)n_j\xi_j^2}{\bar{z}_j+n_j}$$

The second line of (3.4) is a product of two normal kernels for μ_1 and μ_2, the first line only contain π and σ_j^2. Parameters π, (μ_1, σ_1^2) and (μ_2, σ_2^2) are all independent to each other. So we take the operation over π, μ_1, μ_2 and σ_2^2 to get the posterior for σ_1^2:

$$\begin{aligned}
p(\sigma_1^2|\mathbf{y},\mathbf{z}) &= \int_0^\infty\int_0^\infty\int_0^\infty\int_0^1 p(\mu_1,\sigma_1^2,\pi|\mathbf{y},\mathbf{z})d\pi d\mu_1 d\mu_2 d\sigma_2 \\
&\propto ((\sigma_1^2)^{-\frac{\bar{z}_1}{2}-\frac{\nu_1}{2}-\frac{1}{2}}\exp\left[-\frac{1}{\sigma_1^2}\left(\frac{s_1^2}{2}+\frac{1}{2}\left(\sum_{i=1}^{N}z_{1i}y_i^2-k_1y^2\right)\right)\right]
\end{aligned}$$

Similarly we can get the posterior for σ_2^2:

$$p(\sigma_2^2|\mathbf{y},\mathbf{z}) \propto ((\sigma_2^2)^{-\frac{\bar{z}_2}{2}-\frac{\nu_2}{2}-\frac{1}{2}}\exp\left[-\frac{1}{\sigma_2^2}\left(\frac{s_2^2}{2}+\frac{1}{2}\left(\sum_{i=1}^{N}z_{2i}y_i^2-k_2y^2\right)\right)\right]$$

Hence the posterior distribution of σ_j^2 is also inverse gamma:

$$\sigma_j^2|\mathbf{y},\mathbf{z} \sim \mathcal{IG}\left(\frac{\nu_j+\bar{z}_j}{2},\frac{1}{2}\left[s_j^2+\sum_{i=1}^{N}z_{ij}\big(y_i-\bar{y}_j(z)\big)^2+\frac{n_j\bar{z}_j(\bar{y}_j(z)-\xi_j)^2}{n_j+\bar{z}_j}\right]\right) \tag{3.5}$$

Then the posterior distribution of (π, μ_1, μ_2) is:

$$\begin{aligned} p(\pi, \mu_1, \mu_2|\mathbf{y}, \mathbf{z}) &= \frac{p(\theta|\mathbf{y}, \mathbf{z})}{\prod_{j=1}^{2} p(\sigma_j^2|\mathbf{y}, \mathbf{z})} \\ &\propto (1-\pi)^{\bar{z}_1+\beta-1}\pi^{\bar{z}_2+\alpha-1}\prod_{j=1}^{2}(\sigma_j^2)^{-1}\cdot\exp\left[-\frac{1}{2\sigma_j^2}\Big((\bar{z}_j+n_j)\mu_j^2\right. \\ &\quad \left.-2(\bar{z}_j\bar{y_j}(z)+\xi_j n_j)\mu_j+(k_j y^2+n_j\xi_j^2)\Big)\right] \end{aligned}$$

Take operation over π and μ_2 to get the posterior for μ_1:

$$p(\mu_1|\mathbf{y}, \mathbf{z}) \propto \sigma_1^{-2}\exp\left[-\frac{1}{2\sigma_1^2/(\bar{z}_1+n_1)}\left(\mu_1^2-2\frac{\bar{z}_1\bar{y_1}(z)+\xi_1 n_1}{\bar{z}_1+n_1}\mu_1+\frac{k_1y^2+n_1\xi_1^2}{\bar{z}_1+n_1}\right)\right]$$

Similarly the posterior distribution of μ_2 is:

$$p(\mu_2|\mathbf{y}, \mathbf{z}) \propto \sigma_2^{-2}\exp\left[-\frac{1}{2\sigma_2^2/(\bar{z}_2+n_2)}\left(\mu_2^2-2\frac{\bar{z}_2\bar{y_2}(z)+\xi_2 n_2}{\bar{z}_2+n_2}\mu_2+\frac{k_2y^2+n_2\xi_2^2}{\bar{z}_2+n_2}\right)\right]$$

Hence the posterior distribution of μ_j is also normal:

$$\mu_j|\sigma_j^2, \mathbf{y}, \mathbf{z} \sim \mathcal{N}\left(\frac{n_j\xi_j+\bar{z}_j\bar{y}_j(z)}{n_j+\bar{z}_j}, \frac{\sigma_j^2}{n_j+\bar{z}_j}\right) \tag{3.6}$$

Last the posterior distribution of (π) is:

$$\begin{aligned} p(\pi|\mathbf{y}, \mathbf{z}) &= \int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty p(\mu, \sigma^2, \pi|\mathbf{y}, \mathbf{z})d\mu_1 d\mu_2 d\sigma_1^2 d\sigma_2^2 \\ &\propto (1-\pi)^{\bar{z}_1+\beta-1}\pi^{\bar{z}_2+\alpha-1} \end{aligned}$$

Hence the posterior distribution of π is also Beta:

$$\pi|\mathbf{y}, \mathbf{z} \sim \mathcal{B}e\left(\bar{z}_2+\alpha, \bar{z}_1+\beta_2\right) \tag{3.7}$$

Conclude from previous calculation, we get that: for two-component normal mixture model, if

$$\begin{aligned} \mu_j|\sigma_j^2 &\sim \mathcal{N}\left(\xi_j, \frac{\sigma_j^2}{n_j}\right) \\ \sigma_j^2 &\sim \mathcal{IG}\left(\frac{\nu_j}{2}, \frac{s_j^2}{2}\right) \\ \pi &\sim \mathcal{B}e\big(\alpha, \beta\big) \end{aligned}$$

Then

$$\mu_j|\sigma_j^2, \mathbf{y}, \mathbf{z} \sim \mathcal{N}\left(\frac{n_j\xi_j+\bar{z}_j\bar{y}_j(z)}{n_j+\bar{z}_j}, \frac{\sigma_j^2}{n_j+\bar{z}_j}\right)$$

$$\sigma_j^2|\mathbf{y},\mathbf{z} \sim \mathcal{IG}\left(\frac{\nu_j+\bar{z}_j}{2}, \frac{1}{2}\left[s_j^2+\sum_{i=1}^{N} z_{ij}\big(y_i-\bar{y}_j(z)\big)^2+\frac{n_j\bar{z}_j(\bar{y}_j(z)-\xi_j)^2}{n_j+\bar{z}_j}\right]\right)$$

$$\pi|\mathbf{y},\mathbf{z} \sim \mathcal{B}e\left(\bar{z}_2+\alpha, \bar{z}_1+\beta_2\right)$$

□

This property is also true for g-component normal mixture model if we use the g-category Dirichlet prior $\mathcal{D}(\pi|\alpha_1,\ldots,\alpha_g)$ for the proportion parameter, where Dirichlet distribution is the multicategory generalization of Beta distribution, details see [22] and [9].

3.3 Gibbs sampler

Gibbs sampler which originating by the Geman's(1984) is a useful approach to draw sample from the joint posterior when the joint distribution is complicated and difficult to handle. It is a Markov Chain Monte Carlo (MCMC) procedure which sampling from conditional distribution of each parameter when the other parameters and observed data $\mathbf{Z}$ are given.

In this section, we fist give the the general Gibbs sampler procedure and a discussion for stationary. Then apply Gibbs sampler to two-component normal mixture model in both situation with unknown and known variance parameter.

We start with a review of the basic Bayesian framework for prior $p(\theta)$ and posterior $p(\theta|\mathbf{X})$:

$$\begin{aligned} p(\theta|\mathbf{X}) &= \frac{p(\theta)\ell(\theta|\mathbf{X})}{p(\mathbf{X})} \\ &= \frac{p(\theta)\ell(\theta|\mathbf{X})}{\int_\Theta p(\theta)\ell(\theta|\mathbf{X})d\theta)} \\ E(\theta|\mathbf{X}) &= \int_\Theta \theta p(\theta|\mathbf{X})d\theta \end{aligned}$$

Since $\int_\Theta p(\theta)\ell(\theta|\mathbf{X})d\theta)$ is the expression only for $p(\mathbf{X})$, we have:

$$\begin{aligned} p(\theta|\mathbf{X}) &\propto p(\theta) \times L(\theta|\mathbf{X}) \\ Posterior Probability &\propto Prior Probability \times Likelihood Function \end{aligned}$$

Then posterior distribution can be considered as the conditional distributions for unknown parameter when the other parameter and observed data are given. So we generate parameter sample from the posterior distribution.

3.3.1 General Gibbs sampler and its properties

Consider random variables $U_1, U_2, \ldots, U_K$, we simulate sample from the conditional distributions $P(U_j|U_1, U_2, \ldots, U_{j-1}, U_{j+1}, \ldots, U_K)$, $j = 1, 2, \ldots, K$ instead of the joint distribution. Then give general Gibbs sampler as followed (also see [1]):

Algorithm 4. *General Gibbs sampler*

1. *Take some initial values* $U_k^{(0)}$, $k = 1, 2, \ldots, K$.
2. *Repeat for* $t = 1, 2, \ldots, .$:
   For $k = 1, 2, \ldots, K$ *generate* $U_k^{(t)}$ *from*
   $$P(U_k^{(t)}|U_1^{(t)}, \ldots, U_{k-1}^{(t)}, U_{k+1}^{(t-1)}, \ldots, U_K^{(t-1)})$$
3. *Continue step 2 until the joint distribution of* $(U_1^{(t)}, U_2^{(t)}, \ldots, U_K^{(t)})$ *does not change.*

If the explicit form of the conditional density $P(U_k|U_l, l \neq k)$ is available, we can estimate the marginal density of U_k by:

$$\hat{P}_{U_k}(u) = \frac{1}{M - m + 1} \sum_{t=m}^{M} P(u|U_l^{(t)}, l \neq k).$$

This equation can be obtained from following formula:

$$P(A) = \int P(A|B) d(P(B))$$

Now give a discussion for the convergency of Gibbs sampler procedure. Gibbs sampler produces a Markov chain with stationary distribution to be the true joint distribution. We will take the *data augmentation* for example to exposit the stationary. Data augmentation is the first Gibbs sampling brought forth by Tanner and Wong(1987), it is a simple form of general Gibbs sampler when $K = 2$. The algorithm is followed (also see [3]):

Algorithm 5. *Data augmentation*

1. *Initialization: Start with an arbitrary value* $\lambda^{(0)}$.
2. *Iteration* t*: Given* $\lambda^{(t-1)}$, *generate*
   (a) $\theta^{(t)}$ *according to* $p_1(\theta|x, \lambda^{(t-1)})$
   (b) $\lambda^{(t)}$ *according to* $p_2(\lambda|x, \theta^{(t-1)})$

This data augmentation algorithm leads to good convergence properties:

Proposition 3.1. *If $p_1(\theta|x,\lambda) > 0$ on Θ($p_2(\lambda|x,\theta > 0$ on Λ, resp.), both sequences $\left(\theta^{(m)}\right)$ and $\left(\lambda^{(m)}\right)$ are ergodic Markov chains with invariant distributions $p(\theta|x)$ and $p(\lambda|x)$.*

Proof. See [15]. □

Proposition 3.2. *(Duality Principle) If the convergence is uniformly geometric for one of the two chains, e.g., if it takes values in a finite space, the convergence to the stationary distribution is also uniformly geometric for the other chain.*

Proof. See [15]. □

Extending these properties to the general Gibbs sampler, the joint distribution $P(U_1, U_2, \ldots, U_K)$ is stationary at each step since the $P_k's$ are the full condition of $P(U_1, U_2, \ldots, U_K)$. Then the whole procedure is stationary.

Now see the convergency property of Gibbs sampler. Chao (1970) have given following proposition:

Proposition 3.3. *If $\hat{\theta}$ is the MLE and $\tilde{\theta}$ is the posterior mean from a Bayesian model using the same likelihood, but any proper prior(and most improper priors), then:*

$$\sqrt{n}(\tilde{\theta} - \hat{\theta}) \to 0 \ as \ n \to \infty$$

almost assuredly for reasonable starting values of θ.

Details see [9] and [23].

3.3.2 Gibbs sampler for normal mixture model

Now we apply Gibbs sampler to the indicator normal mixture model, We need to know the conditional distributions for all parameters.

As calculation in the pervious section, choose conjugate prior for indicator normal mixture model to get a posterior distribution. Besides, also consider $z = \{z_1, z_2\}$ as additional parameter. So we get the following joint distributions:

$$\begin{aligned} z_i|\theta &\sim \mathcal{B}er(\pi), \\ y_i|z_i,\theta &\sim \mathcal{N}\left(\prod_{j=1}^{2}\mu_i^{z_{ij}}, \prod_{j=1}^{2}\sigma_i^{2z_{ij}}\right). \end{aligned}$$

According to the computation in Section 3.2, we have the following posterior distributions:

For $i = 1, \ldots, N$ and $j = 1, 2$,

$$\begin{aligned} z_i|\mathbf{y}, \theta &\sim \mathcal{B}er\left(\frac{\pi\varphi_{\theta_2}(y_i)}{(1-\pi)\varphi_{\theta_1}(y_i) + \pi\varphi_{\theta_2}(y_i)}\right) \\ \mu_j|\sigma_j^2, \mathbf{y}, \mathbf{z} &\sim \mathcal{N}\left(\frac{n_j\xi_j + \bar{z}_j\bar{y}_j(z)}{n_j + \bar{z}_j}, \frac{\sigma_j^2}{n_j + \bar{z}_j}\right), \\ \pi|\mathbf{y}, \mathbf{z} &\sim \mathcal{B}e\left(\bar{z}_2 + \alpha, \bar{z}_1 + \beta_2\right), \end{aligned}$$

and

$$\sigma_j^2|\mathbf{y}, \mathbf{z} \sim \mathcal{IG}\left(\frac{\nu_j + \bar{z}_j}{2}, \frac{1}{2}\left[s_j^2 + \sum_{i=1}^{N} z_{ij}\big(y_i - \bar{y}_j(z)\big)^2 + \frac{n_j\bar{z}_j(\bar{y}_j(z) - \xi_j)^2}{n_j + \bar{z}_j}\right]\right). \tag{3.8}$$

Then we give the Gibbs sampler for two-component normal mixture model.

Algorithm 6. *Gibbs sampler for two-component normal mixtures.*

1. *Take some initial values* $\theta^{(0)} = (\pi^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}, (\sigma_1^2)^{(0)}, (\sigma_2^2)^{(0)})$*, where those parameters come from the prior distributions (3.2).*

2. *Repeat for t=1,2,. . . ,.*

   (a) *For* $i = 1, 2, \ldots, N$*, generate* $z_i^{(t)} \in \{0, 1\}$ *with*

$$z_i^{(t)} \sim \mathcal{B}er\left(\frac{\pi^{(t)}\varphi_{\theta_2}(y_i)}{(1-\pi^{(t)})\varphi_{\theta_1}(y_i) + \pi^{(t)}\varphi_{\theta_2}(y_i)}\right)$$

(b) For $j = 1, 2$, generate parameters as followed [5]*:*

$$\begin{aligned}
\pi^{(t+1)} &\sim \mathcal{B}e\left(\bar{z}_2^{(t)} + \alpha, \bar{z}_1^{(t)} + \beta_2\right), \\
(\sigma_j^2)^{(t+1)} &\sim \mathcal{IG}\left(\frac{\nu_j + \bar{z}_j^{(t)}}{2}, \frac{1}{2}\left[s_j^2 + \sum_{i=1}^{N} z_{ij}^{(t)}\big(y_i - \bar{y}_j(z)^{(t)}\big)^2 \right.\right. \\
&\qquad \left.\left. + \frac{n_j \bar{z}_j^{(t)} (\bar{y}_j(z)^{(t)} - \xi_j)^2}{n_j + \bar{z}_j^{(t)}}\right]\right), \\
\mu_j^{(t+1)} &\sim \mathcal{N}\left(\frac{n_j \xi_j + \bar{z}_j^{(t)} \bar{y}_j(z)^{(t)}}{n_j + \bar{z}_j^{(t)}}, \frac{(\sigma_j^2)^{(t+1)}}{n_j + \bar{z}_j^{(t)}}\right).
\end{aligned}$$

3. *Continute step 2 until the joint distribution of $(z^{(t)}, \theta^{(t)})$ does not change.*

3.3.3 Gibbs sampler for normal mixture with known σ_j^2 and π

At last give a specific application to normal mixture model with variances and mixture proportion known:

Algorithm 7. *Gibbs sampler for two-component normal mixtures with known variance and mixture proportion.*

1. *Take some initial values $\theta^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)})$.*
2. *Repeat for t=1,2,. . . ,.*
   (a) *For $i = 1, 2, \ldots, N$, generate $\Delta_i^{(t)} \in \{0, 1\}$ with $Pr(\Delta_i^{(t)} = 1) = \hat{\gamma}_i(\theta^{(t)})$, by Equation (2.6).*
   (b) *Set*

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N}(1 - \Delta_i^{(t)}) \cdot y_i}{\sum_{i=1}^{N}(1 - \Delta_i^{(t)})}, \qquad \hat{\mu}_2 = \frac{\sum_{i=1}^{N} \Delta_i^{(t)} \cdot y_i}{\sum_{i=1}^{N} \Delta_i^{(t)}} \tag{3.9}$$

and generate $\mu_1^{(t+1)} \sim N(\hat{\mu}_1, \hat{\sigma}_1^2)$ and $\mu_2^{(t+2)} \sim N(\hat{\mu}_2, \hat{\sigma}_2^2)$.

[5] The priors and posterior under other parameters must be proper otherwise the geometric convergence will not be established. [1] give an example with following simulation

$$\sigma_j^2 | \mathbf{y}, \mathbf{z}, \mu_i \sim \mathcal{IG}\left(\frac{\nu_j + n_j + 1}{2}, \frac{1}{2}\left[s_j^2 + \sum_{i=1}^{N} z_{ij}(y_j - \xi_j)^2\right]\right),$$

The convergence will be much more difficult to established since geometric convergence must be established under imposing restrictions on σ_i^2. So μ_i should not be involved in the conditional posteriors distribution of σ_i^2. Details see [3].

3. *Continute step 2 until the joint distribution of* $(\Delta^{(t)}, \mu_1^{(t)}, \mu_2^{(t)})$ *does not change.*

Compare this Gibbs sampler with the simulation result in Algorithm 3:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N}(1-\hat{\gamma}_i)\cdot y_i}{\sum_{i=1}^{N}(1-\hat{\gamma}_i)}, \qquad \hat{\mu}_2 = \frac{\sum_{i=1}^{N}\hat{\gamma}_i \cdot y_i}{\sum_{i=1}^{N}\hat{\gamma}_i}$$

Notice that the iterative prior $\hat{\mu}_j$ in (3.9) has the similar form as the plug-in-estimators of the EM-Algorithm. But these two approaches have essential differences in both parameter and procedure.

In Gibbs sampler, we consider the latent data $\mathbf{\Delta}$ to be another parameter. Compared with the "E" step of the EM-Algorithm, Gibbs sampler use generated latent data Δ_i from distributions $P(\Delta_i|\theta, \mathbf{Z})$ instead of using the responsibilities $\gamma_i(\theta) = E(\Delta_i|\theta, \mathbf{Z})$. Also compared with the "M" step of the EM-Algorithm, Gibbs sampler generate the iterative items from the conditional distribution $P(\mu_1, \mu_2|\mathbf{\Delta}, \mathbf{Z})$ instead of maximizing the posterior $P(\mu_1, \mu_2, \mathbf{\Delta}|\mathbf{Z})$.

4 Application

In the last part of Section 3.3.3, we compared the EM-Algorithm and Gibbs sampler in procedure through a simplified normal mixture model. Now take a further comparison practically for the parameter estimators received from both algorithm procedures.

4.1 Simulation results

We run the algorithm procedures under a simplified two-component normal mixture model with variances σ_1^2, σ_2^2 and the mixture proportion π known. We focus on the estimation of unknown parameters μ_1 and μ_2.

4.1.1 Simulation results for the fictitious data set

From Figure 4, we have the density plot of the fictitious data set and it indicates a two-component normal mixture model.

In Section 2.2.4, we mentioned the common way of initial guesses for unknown parameters. In fact, different initial guesses lead to different iterative estimate results and those get the highest maximized likelihood are the best. Hastie, Tibshirani and Friedman run the EM-Algorithm for 20 fictitious data and received a best group of estimate values in [1]. So we just choose the best estimate values given in [1] to evaluate the known parameters σ_1^2, σ_2^2 and π, also to be the initial guesses for μ_1 and μ_2. Then we have σ_1^2=0.87, σ_2^2=0.77, and π=0.546 (See [1] pp. 237-240).

Run Algorithm 3 and Algorithm 7 through R programme with initial guesses $\mu_1^{(0)}$=4.62 and $\mu_2^{(0)}$=1.06, we get the simulation results of the fictitious data set. From Figure 5 and Figure 6 we find these two algorithms get similar iterative estimation results. The EM-Algorithm estimator, the mean value of the Gibbs sampler estimator and the Gibbs sampler estimator with the highest density are very close in Figure 6. But The EM-Algorithm is much stabler and faster for this data sample. From Figure 5, we see it get converged in less than 10 steps, when Gibbs sampler is still fluctuating widely after 200 iterations.

4.1.2 Simulation results for the astronomy data set

Now come to the real astronomy data. We choose a data set with 81 datum which is the numbers of photons over wavelength in [nm](we call it *photon data*). From the density plot(Figure 4), we consider the photon data as following a two-component normal mixture model. Use the common way to choose the initial values of parameters and run the EM-Algorithm and Gibbs sampler, we get the estimation results of this photon data (Figure 8 and Figure 9).

Both algorithms also get similar estimate results for the real photon data, and the EM-Algorithm still seems faster and stabler. But we notice that the density plot of Gibbs sampler estimators has more than one local maximization(Figure 9), which means there might be some information that out of our model. Come back to the initial data set information, a group of noise at each point was mentioned. Fixing of this kind of noise requires more astronomy background which is out of this thesis, so we just use the original photon data.

Here we could find the advantages of Gibbs sampler: it contains more information. The EM-Algorithm can only get the stable convergency estimator value, but Gibbs sampler shows more information of the data set itself besides the estimator, which is more important for statistical analysis, especially with large number of data set.

The estimator generated from Gibbs sampler should follow a normal distribution, but Figure 9 are obviously not normal. We wonder whether it is the problem of the data set and the model or the problem of the method of Gibbs sampler, so we generate a two-component normal mixture model with parameters as followed:

$$\pi = 0.5,\ \mu_1 = 18,\ \sigma_1^2 = 0.8,\ \mu_2 = 25,\ \sigma_2^2 = 4.$$

Then suppose the mean parameters are unknown and generate the mean parameter estimators by Gibbs sampler. We get the results as Figure 11. From Figure 11 we have that the parameter estimators generated from Gibbs

sampler for generated two-component normal mixture data are typical normal distribution and the estimation is not bad compared with the original parameter value. So Gibbs sampler might not be responsible for the bad results of the photon data.

Come back to Figure 4. We suppose the photon data follow a two-component normal mixture model according to the shape of density plot. But it may also be a finite normal mixture with higher dimension than two. Figure 12 shows two density plots of a three-component and a four-component normal mixture model with parameters as followed:

$$\begin{aligned} \pi_1 = 0.2, \quad \mu_1 &= 18, \quad \sigma_1^2 = 0.8, \\ \pi_2 = 0.3, \quad \mu_2 &= 25, \quad \sigma_2^2 = 3.4, \\ \pi_3 = 0.5, \quad \mu_3 &= 60, \quad \sigma_3^2 = 11.2; \end{aligned}$$

and

$$\begin{aligned} \pi_1 = 0.1, \quad \mu_1 &= 18, \quad \sigma_1^2 = 0.8, \\ \pi_2 = 0.2, \quad \mu_2 &= 25, \quad \sigma_2^2 = 3.4, \\ \pi_3 = 0.3, \quad \mu_3 &= 40, \quad \sigma_3^2 = 11.2, \\ \pi_4 = 0.4, \quad \mu_4 &= 110, \quad \sigma_4^2 = 34. \end{aligned}$$

Both density plots are similar to the density plot of the photon data, and it is difficult to judge the exactly dimension of the photon data. So the bad results from previous Gibbs sampler for the photon data are probably caused by the data and model, but not the simulation method. And Gibbs sampler might be a good method for simulation under two-component normal mixture model. The situation of the higher dimension of finite normal mixtures is worth for further study.

4.2 Confidence interval

Now give the definition of Bayesian Credible Interval:

Definition 3. *Credible set(also see [9]) Define $\mathcal{C}$ as a subset of the parameter space Θ such that a $100(1-\alpha)$ credible interval meets the condition:*

$$1 - \alpha = \int_C p(\theta|\mathbf{X}) d\theta.$$

Denote T as the estimate of parameter θ, a_p as the quantiles of $\hat{\theta} - \theta$. Then we have:

$$P(T - \theta \leq a_\alpha) = P(T - \theta \geq a_{1-\alpha}) = \alpha.$$

Suppose T is continuous. We want an equal tail interval with tail errors equal to α. The limits of $1-2\alpha$ equal tail interval is

$$\hat{\theta}_\alpha = t - a_{1-\alpha}, \quad \hat{\theta}_{1-\alpha} = t - a_\alpha.$$

Using T-test, we get the two-side confidence intervals for both data sets. The 95% confidence intervals for μ_1 and μ_2 under fictitious data are (3.819986, 4.193414) and (1.555847, 1.922822). The 95% confidence intervals for μ_1 and μ_2 under photon data are (189317.6, 191720.1) and (-380.9429, -356.9297). Box plots also show the similar results. See Figure 7 and Figure 10.

4.3 Discussion

For general case of Gibbs sampler, iterative sample need to be generated from the posterior distribution. One important point is the parameter selection for the prior distribution. Since we already have some initial guesses for the unknown parameter vector θ, the prior parameter need to be chosen carefully to make the initial guesses fit the prior distribution. Plenty values of prior parameters are tested and the final choice for our model are:

$$\alpha = 2, \qquad \beta = 2.48,$$

$$\frac{\nu_1}{2} = 5.46, \qquad \frac{s_1^2}{2} = 1.032 * 10^{10},$$

$$\frac{\nu_2}{2} = 5.46, \qquad \frac{s_2^2}{2} = 10^6,$$

$$\xi_1 = 191942.2, \quad \sigma_1^2 = 2847896799, \quad n_1 = 5 * 10^8,$$

$$\xi_2 = -379.9798, \quad \sigma_2^2 = 206657.3, \quad n_2 = 42000,$$

The selection of the prior parameter is important but complicate work and the further study of more efficient way to choose proper prior parameter might be attractive.

The EM-Algorithm and Gibbs sampler are both good simulation methods for parameter estimation, and they usually get similar results. The EM-Algorithm needs no prior information and it is stabler. Gibbs sampler is more complicated in computing but also contains more information, which is better for further statistical analysis. Summarizing the simulation results, we suggest to use the EM-Algorithm when have a small number of data set, and choose Gibbs sampler when more data and information need to be dealed with.

A Appendix: Astronomic background of the data set

Planetary nebulae (PNe) has a strong emission-line spectrum and it can be used for studying the expansion properties of PNe. When the PNe expands, the Doppler-shifting can course a broadening of the emission-line.

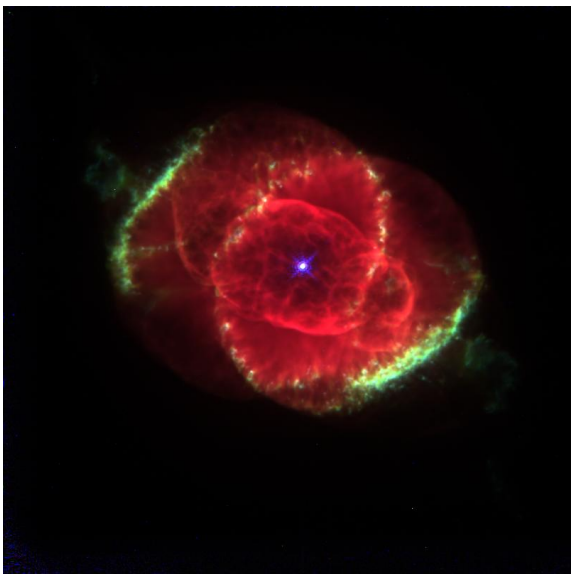

Figure 2: Picture of Planetary Nebulae (PNe)

Two different velocities v1 and v2 (Figure 3) are obtained ,which result in two symmetric peaks and the Doppler-shift can be assumed to causes a mixture of two normal distributions.

The spectrograph detects the number of photons for different wavelength-bins. The outcome spectra(numbers of photons over wavelength in [nm]) shows a line broadening, or in some cases, a double line profile. It is possible to translate the wavelength in radial velocity (RV). So the number of photons can be used to set the two-component normal mixture model. The number of photons in the "$StrWr2\_Hr8\_O2\_sigmaright$" data set is chosen for the comparison in this thesis.

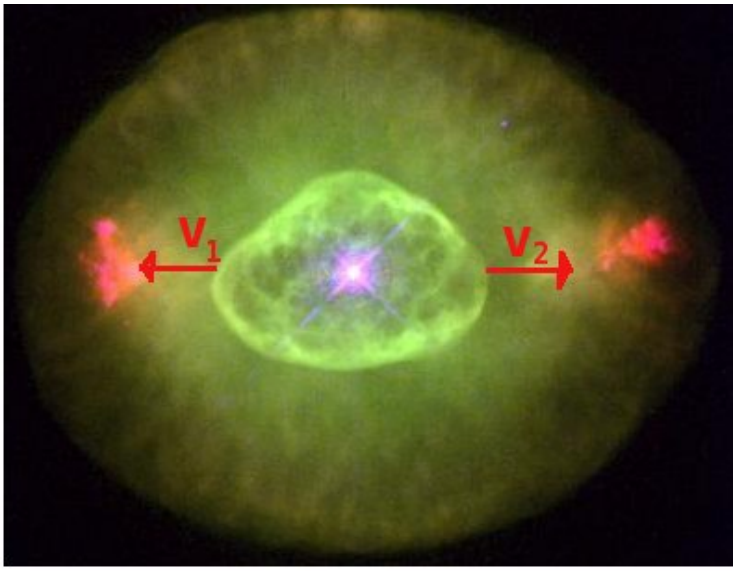

Figure 3: Picture of Velocities

B Appendix: Programme results

B.1 Simulation results

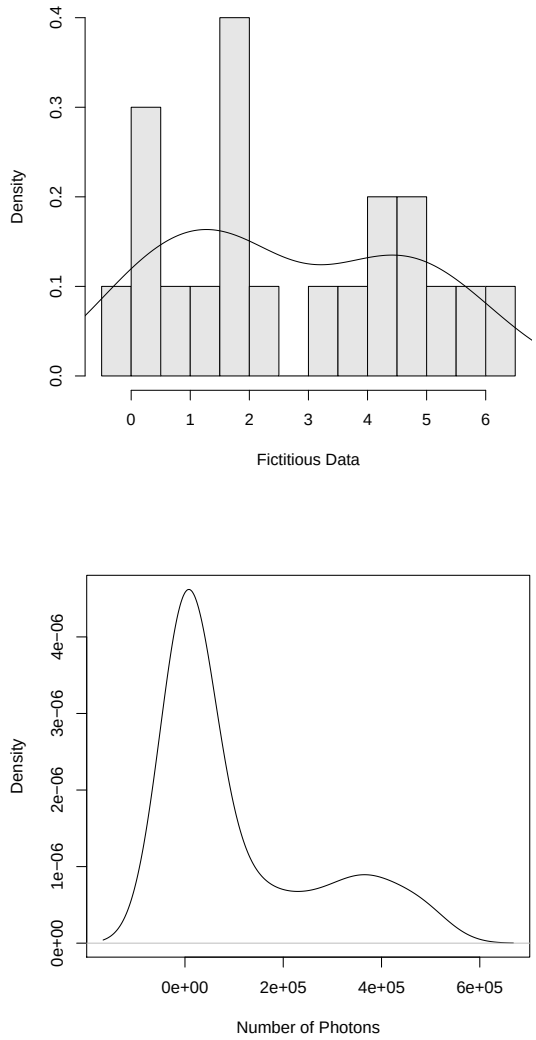

Figure 4: Density plots for both data sets

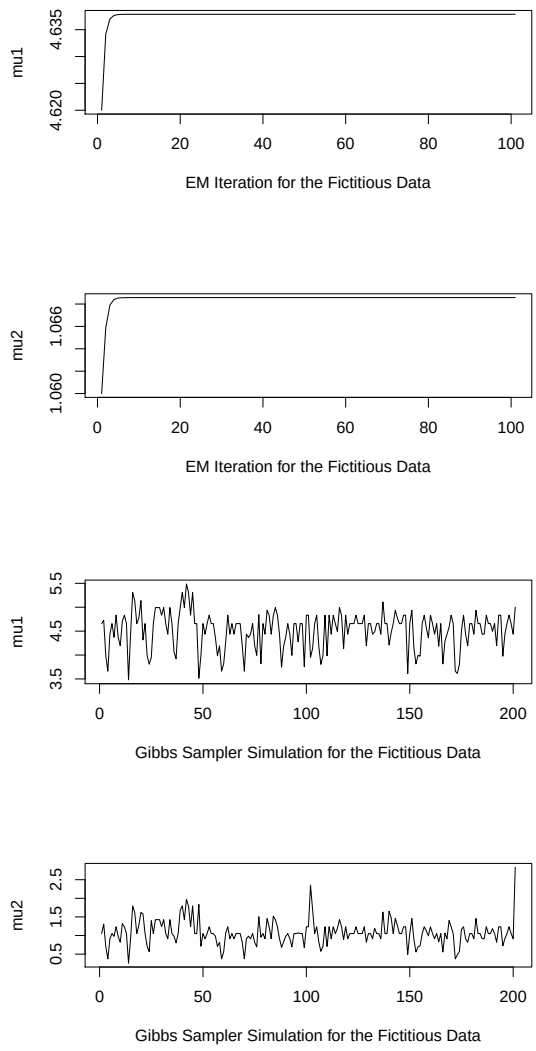

Figure 5: Mean value estimators for the Fictitious data

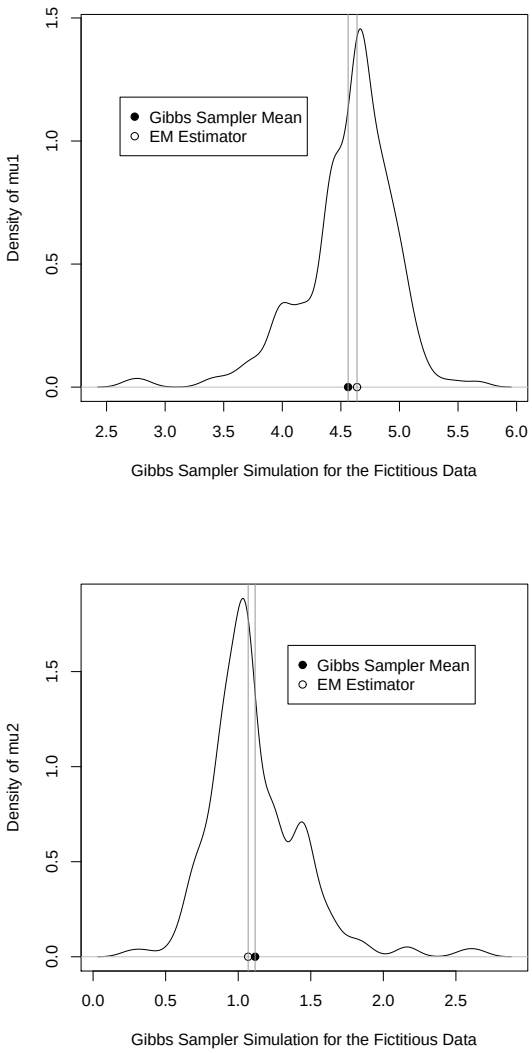

Figure 6: Density plots from Gibbs sampler for the Fictitious data

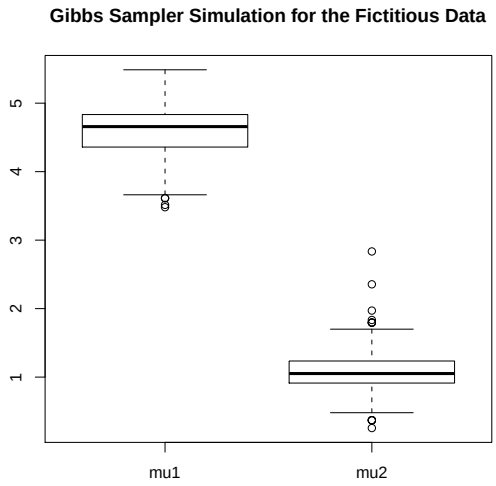

Figure 7: Box plot from Gibbs sampler for the Fictitious data

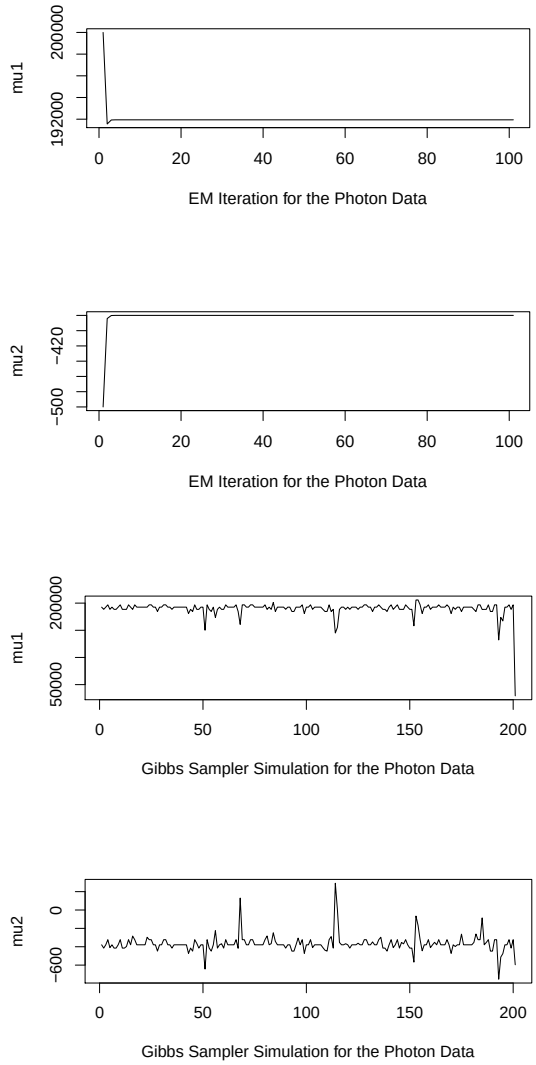

Figure 8: Mean value estimators for the Photon data

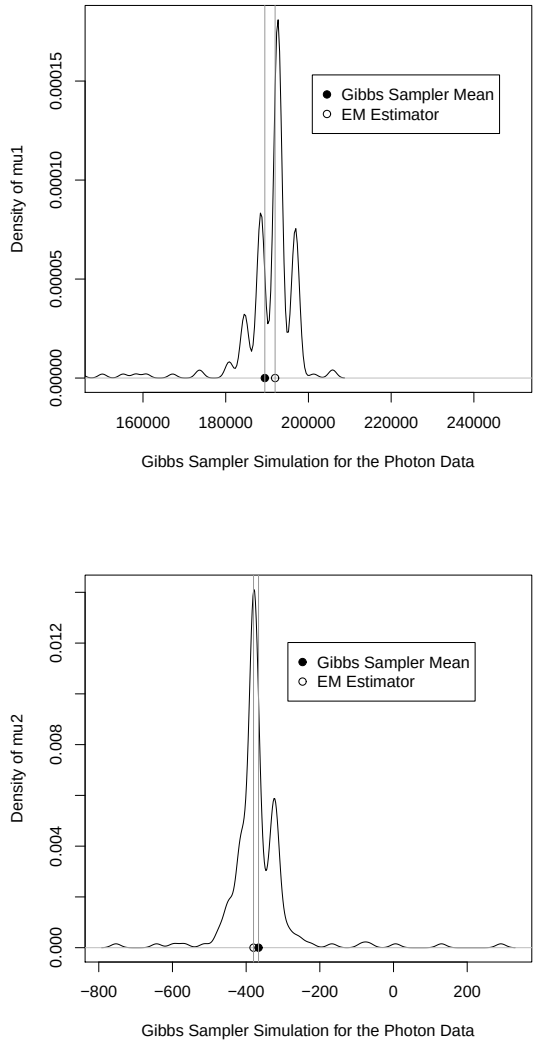

Figure 9: Density plots from Gibbs sampler for the Photon data

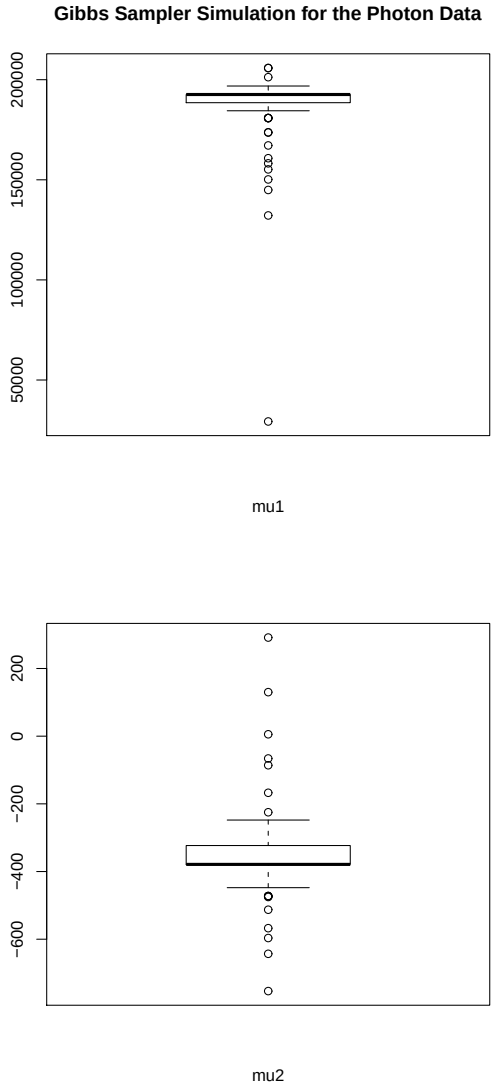

Figure 10: Box plots from Gibbs sampler for the Photon data

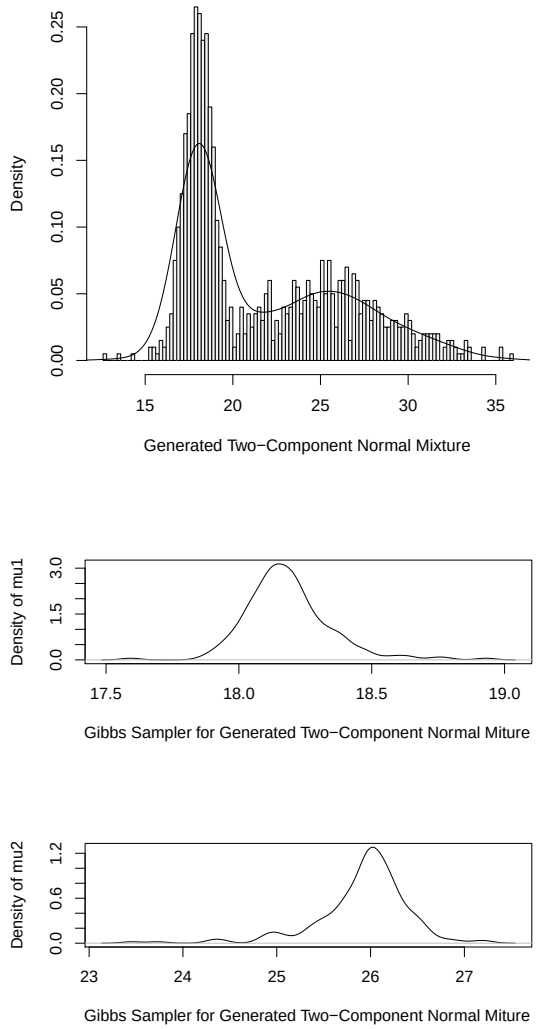

Figure 11: Gibbs sampler for generated two-component normal mixture

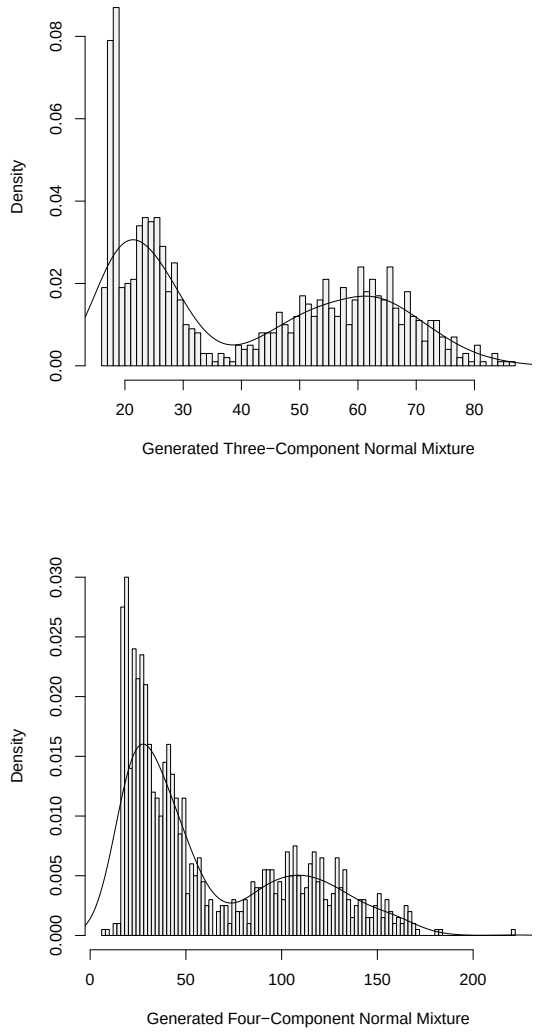

Figure 12: Density plots for high dimension-component normal mixture model

C Appendix: R programme code

C.1 R code for Algorithm 3 with σ_1^2, σ_2^2 and π known

```
data<-c(-0.39,0.12,0.94,1.67,1.76,2.44,3.72,4.28,4.92,5.53,
        0.06,0.48,1.01,1.68,1.80,3.25,4.12,4.60,5.28,6.22)
pi<-0.546
sigmas1<-0.87
sigmas2<-0.77
mu1<-numeric(0)
mu2<-numeric(0)
r<-numeric(0)
R1<-matrix(0,20,100)
mu1[1]<-4.62
mu2[1]<-1.06
for (j in 1:100){
for (i in 1:20){
r[i]<-pi*dnorm(data[i],mu2[j],sigmas2^(1/2))/((1-pi)*dnorm(data[i],
     mu1[j],sigmas1^(1/2))+ pi*dnorm(data[i],mu2[j],sigmas2^(1/2)))
R1[i,j]<-r[i]
}
mu1[j+1]<-sum((1-r)*data)/sum(1-r)
mu2[j+1]<-sum(r*data)/sum(r)
Muu1<-mu1[j+1]
Muu2<-mu2[j+1]
}
Muu1
Muu2
x11()
layout(matrix(c(1,2)))
plot(mu1,type="l",main="",xlab="EM Iteration for the
Fictitious Data")
plot(mu2,type="l",main="",xlab="EM Iteration for the
Fictitious Data")
```

C.2 R code for Algorithm 7

```
delta<-numeric(0)
r<-numeric(0)
R2<-matrix(0,20,200)
for (j in
1:200){
for (i in 1:20){
r[i]<-pi*dnorm(data[i],mu2[j],sigmas2^(1/2))/((1-pi)*dnorm(data[i],
```

```
      mu1[j],sigmas1^(1/2))+pi*dnorm(data[i],mu2[j],sigmas2^(1/2)))
delta[i]<-rbinom(1,1,r[i])
R2[i,j]<-r[i]
}
delta
mu1[j+1]<-sum((1-delta)*data)/sum(1-delta)
mu2[j+1]<-sum(delta*data)/sum(delta)

mu1[j+1]=rnorm(1,mu1[j],0.87^(1/2))
mu2[j+1]=rnorm(1,mu2[j],0.77^(1/2))

MU1<-mu1[j+1]
MU2<-mu2[j+1]
}
MU1
MU2
x11()
layout(matrix(c(1,2)))
x11()
layout(matrix(c(1,2)))
plot(mu1,type="l",main="",xlab="Gibbs Sampler Simulation for
the Fictitious Data")
plot(mu2,type="l",main="",xlab="Gibbs Sampler Simulation for
the Fictitious Data")
```

C.3 R code for plots

```
x11()
plot(density(mu1),type="l",main="",ylab="Density of mu1",
xlab="Gibbs Sampler Simulation for the Fictitious Data")
points(mean(mu1),0,pch=19)
abline(v=mean(mu1),col="gray60")
points(4.637871,0,pch=21)
abline(v=4.637871,col="gray60")
legend(locator(n=1),legend=c("Gibbs Sampler Mean","EM
Estimator"),pch=c(19,21))

t.test(mu1,conf.level=0.95)
t.test(mu2,conf.level=0.95)

x11()
boxplot(mu1,mu2,names=c("mu1","mu2"),main="Gibbs Sampler
Simulation for the Fictitious Data")
```

References

[1] Hastie, T., Tibshirani, R., Friedman, J. (2001). *The elements of statistical learning; data mining, inference, and prediction*, Springer, ISBN 0-387-95284-5

[2] Liero, H., Zwanzig, S. (2008). *Script for "Computer Intensive Methods in Statistics—a short course with R"*, Home page of Silvelyn Zwanzig, Uppsala University, http://www.math.uu.se/ zwanzig/index.html

[3] Robert, Christian P. (2001). *The Bayesian choice, 2nd edition*, Springer, ISBN 0-387-95231-4

[4] Marin, J-M., Robert, Christian P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, Springer, ISBN 978-0-387-38979-0

[5] Efron, B. (1979). *Bootstrap Method: Another Look at the Jackknife*, The Annals of Statistics, Vol. 7, No. 1, pp. 1-26

[6] Davison, A.C., Hinkley, D.V. (1997). *Bootstrap methods and their applications*, Cambridge University Press, ISBN 978-0-521-57391-7

[7] Efron, B., Tibshirani, R. (1993). *An introduction to the bootstrap*, Chapman & Hall, Inc., ISBN 0-412-04231-2

[8] Woodward, M. (1999). *Epidemiology; study design and data analysis*, Chapman & Hall/CRC, ISBN 1-58488-009-0

[9] Gill, J. (2008). *Bayesian Methods; a social and behavioral sciences approach, 2nd edition*, Chapman & Hall/CRC, ISBN 1-58488-562-9

[10] McLachlan, G.J.,Krishnan T. (1997). *The EM Alogorithm and Extensions*, Jhon Wiley & Sons, Inc., ISBN 0-471-12358-7

[11] Ghosh, J.K., Ramamoorthi R.V. (2003). *Bayesian Nonparnmetrics*, Springer, ISBN 0-387-95537-2

[12] Lange K. (1998). *Numerical Analysis for Statisticians*, Springer, ISBN 0-387-94979-8

[13] Titterington D.M. (1985). *Statistical Analysis of Finite Mixture Distributions*, Jhon Wiley & Sons, Inc., ISBN 0-471-90763-4

[14] Dempster A.D., Laird N.M., Rubin D.B. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Staitistical Society, Series B, Vol.39, No.1, pp.1-38

[15] Diebolt J., Robert C.P. (1994). *Estimation of Finite Mixture Distributions through Bayesian Sampling*, Journal of the Royal Staitistical Society, Series B, Vol.56, No.2, pp.363-375

[16] Verzani J. (2004). *Using R for Introductory Statistics*, Chapman & Hall/CRC, ISBN 1-58488-4509

[17] Albert J. (2007). *Bayesian Computation with R*, Springer, ISBN 978-0-387-71384-7

[18] Givens G.H., Hoeting J.A. (2005). *Computational Statistics*, John Wiley & Sons, Inc., ISBN 0-471-46124-5

[19] Garthwaite P.H., Jolliffe I.T., Jones B. (2002). *Statistical Inference, 2nd edition*,Oxford, ISBN 0-19-857226-3

[20] Figueiredo M.A.T., (2004). *Lecture Notes on the EM Algorithm*, Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001 Lisboa, Portugal, mtf@lx.it.pt

[21] Dellaert F., Hoeting J.A. (2002). *The Expectation Maximization Algorithm*, College of Computing, Georgia Institute of Technology Technical, Report number GIT-GVU-02-20

[22] Marin J.M., Mengersen K., Robert C.P. (2005). *Bayesian Modelling and Inference on Mixtures of Distributions*, Handbook of Statistics, D. Dey & C. Rao, eds., Vol. 25. Elsevier-Sciences, ISBN: 9780444515391

[23] Chao M.T.(1970). *The Asymptotic Behavior of Bayes's Estimators*, Annals of Mathematical Statistics 41, 601-608

[24] McLachlan G.J., Krishnan T.(2008). *The EM Algorithm and Extensions, 2nd edition*, John Wiley & Sons, ISBN 978-0-471-20170-0

[25] Ishwaran H., James L.F.(2002). *Approximate Dirichlet Process Computing in Finite Normal Mixtrues: Smoothing and Prior Information*, Journal of Computational and Graphical Statistics, Vol. 11, No. 3, Page 1-26

[26] Wu C.F.J.(1983). *On the Convergence Properties of the EM Algorithm*, Annals of Statistics 11, 95-103