SF2955
Computer Intensive Methods In Mathematical Statistics
Project 2
Statistical inference from coal mine disaster and mixture model data using MCMC and EM-algorithm

Baixi Jin
20000607-T005
baixi@kth.se

Hanmo Chen
19990904-T072
hanmoc@kth.se

May 2019

Contents

1 Bayesian analysis of coal mine disasters —— constructing a complex MCMC algorithm **2**
1.1 marginal posteriors . . . 2
1.2 Construct a hybrid MCMC algorithm from $f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}|\boldsymbol{\tau})$ . . . 3
1.3 Investigate MCMC chain for different breakpoints . . . 4
1.3.1 one breakpoint . . . 4
1.3.2 two breakpoints . . . 5
1.3.3 three breakpoints . . . 5
1.3.4 four breakpoints . . . 6
1.4 Sensitivity to the choice of ϑ . . . 6
1.4.1 ϑ =1 . . . 6
1.4.2 ϑ =10 . . . 7
1.4.3 ϑ =40 . . . 7

1.4.4 ϑ =100 . . . 7
1.4.5 Summary . . . 8
1.5 Sensitivity to the choice of ρ . . . 8
1.5.1 ρ =0.1 . . . 8
1.5.2 ρ =0.4 . . . 8
1.5.3 ρ =0.7 . . . 9
1.5.4 ρ =1.0 . . . 9
1.5.5 Summary . . . 9

2 EM-based inference in mixture models **10**
2.1 complete data log-likelihood function $f_\theta(\boldsymbol{x}, \boldsymbol{y})$ . . . 10
2.2 conditional distribution $f_\theta(\boldsymbol{x}|\boldsymbol{y})$ . . . 10
2.3 EM algorithm for θ . . . 10

1 Bayesian analysis of coal mine disasters —— constructing a complex MCMC algorithm

1.1 marginal posteriors

Compute the marginal posteriors for $f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})$, $f(\boldsymbol{\lambda}|\theta, \boldsymbol{t}, \boldsymbol{\tau})$ and $f(\boldsymbol{t}|\theta, \boldsymbol{\lambda}, \boldsymbol{\tau})$.

We use *Bayes' Theorem* to rewrite this expression.

$$f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau}) \propto f(\boldsymbol{t}, \boldsymbol{\lambda}, \theta) \cdot f(\boldsymbol{\tau}|\boldsymbol{t}, \boldsymbol{\lambda}, \theta)$$

Since t is independent from $\boldsymbol{\lambda}$ and θ, the first term is rewritten as

$$f(\boldsymbol{t}, \boldsymbol{\lambda}, \theta) = f(\boldsymbol{t}) \cdot f(\boldsymbol{\theta}, \boldsymbol{\lambda}) = f(\boldsymbol{t}) \cdot f(\boldsymbol{\lambda}|\theta) \cdot f(\theta)$$

while the second term

$$f(\boldsymbol{\tau}|\boldsymbol{t}, \boldsymbol{\lambda}, \theta) = f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t})$$

Inserting this expression then yields

$$f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau}) \propto f(t) \cdot f(\theta) \cdot f(\boldsymbol{\lambda}|\theta) \cdot f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t})$$

The expression of the distribution becomes

$$f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau}) \propto \prod_{i=1}^{d} (t_{i+1} - t_i) \cdot \theta \exp\{-\vartheta\theta\} \cdot \theta^{2d} \prod_{i=1}^{d} \lambda_i \exp\left\{-\theta \sum_{i=1}^{d} \lambda_i\right\} \cdot \prod_{i=1}^{d} \lambda_i^{n_i(\boldsymbol{\tau})} \exp\left\{-\sum_{i=1}^{d} \lambda_i (t_{i+1} - t_i)\right\}$$

I. We identify the terms containing θ and finally get that

$$f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau}) \propto \theta^{2d+1} \exp\left\{-\theta \cdot \left(\vartheta + \sum_{i=1}^{d} \lambda_i\right)\right\} \sim \Gamma\left(2(d+1), \vartheta + \sum_{i=1}^{d} \lambda_i\right)$$

II. The procedure of computing is the same as in I.

$$f(\boldsymbol{\lambda}|\theta, \boldsymbol{t}, \boldsymbol{\tau}) \propto f(t) \cdot f(\theta) \cdot f(\boldsymbol{\lambda}|\theta) \cdot f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t})$$

We then identify the terms containing λ and conclude that

$$f(\boldsymbol{\lambda}|\theta, \boldsymbol{t}, \boldsymbol{\tau}) \propto \prod_{i=1}^{d} \lambda_i^{1+n_i(\boldsymbol{\tau})} \cdot \exp\left\{-\sum_{i=1}^{d}(\theta + (t_{i+1} - t_i))\lambda_i\right\}$$

Thus,

$$\lambda_i \sim \Gamma\left(2 + n_i(\boldsymbol{\tau}), \theta + t_{i+1} - t_i\right)$$

III. The procedure of computing is the same as in I.

We then identify the terms containing $\boldsymbol{t}$ and obtain that

$$f(\boldsymbol{t}|\theta, \boldsymbol{\lambda}, \boldsymbol{\tau}) \propto \prod_{i=1}^{d}(t_{i+1} - t_i) \cdot \exp\left\{-\sum_{i=1}^{d} \lambda_i(t_{i+1} - t_i)\right\}$$

for $t_1 < t_2 < ... < t_{d+1}$

1.2 Construct a hybrid MCMC algorithm from $f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}|\boldsymbol{\tau})$

We use RWMH and Gibbs sampling in the algorithm.

Algorithm 1 Hybrid MCMC

Input: $\boldsymbol{\tau}$
Output: $\boldsymbol{\lambda}, \boldsymbol{t}, \theta$
  for $j = 1 : M - 1$ **do**
    Draw $\theta^{j+1} \sim f(\theta|\boldsymbol{\lambda}^j, \boldsymbol{t}^j, \boldsymbol{\tau})$
    compute $n_i(\boldsymbol{\tau})$
    Draw $\boldsymbol{\lambda}^{j+1} \sim f(\boldsymbol{\lambda}|\boldsymbol{\tau}, \boldsymbol{t}^j, \theta^{j+1})$
    for $i = 2 : d$ **do**
      $R = \rho(t_{i+1}^j - t_i^{j+1})$
      $\epsilon \sim U(-R, R)$
      $t_i^* = t_i^j + \epsilon$
      $\alpha = min(1, \frac{f(t_i^j|t_i^*) \cdot f(t_i^*|t^{-i})}{f(t_i^*|t_i^j) \cdot f(t_i^j|t^{-i})})$
      if $U(0, 1) \leq \alpha$ **then**
        $t_i^{j+1} \leftarrow t_i^*$
      else
        $t_i^{j+1} \leftarrow t_i^j$
      end if
    end for
  end for

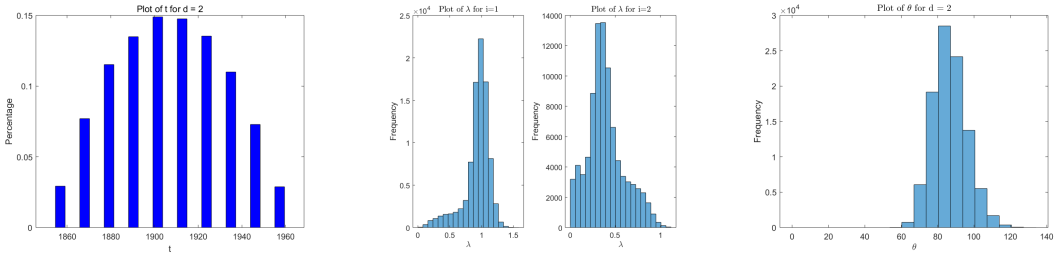
where t^{-i} is given by $(t_1^{j+1}, ..., t_{i-1}^{j+1}, t_{i+1}^{j}, ..., t_{d+1}^{j})$

and $\theta^{j+1} \sim \Gamma(2d+2, \sum \lambda_j + \vartheta)$, $\lambda_i^{j+1} \sim \Gamma(2 + n_i(\tau), t_{i+1} - t_i + \theta^{j+1})$

1.3 Investigate MCMC chain for different breakpoints

In our investigation, we have set $\vartheta = 1, \rho = 1$.

1.3.1 one breakpoint



(a) t (b) λ (c) θ

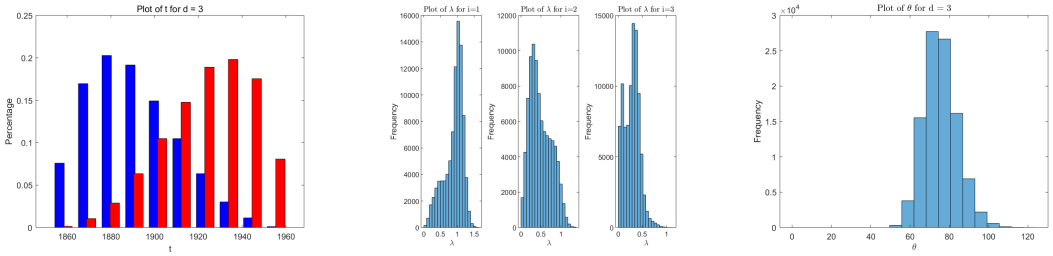
Figure 1: Break point=1

the histogram of t is centered around 1900, but the accuracy of estimation is not quite satisfying.

The histogram of λ_1 and λ_2 suggests that the intensities both have a Γ–distribution where the first intensity is centered around 1.1 whereas the second intensity is centered around 0.4, indicating a decrease of almost 70%.

We can also see that θ intensity is centered around 90.

1.3.2 two breakpoints



(a) t (b) λ (c) θ

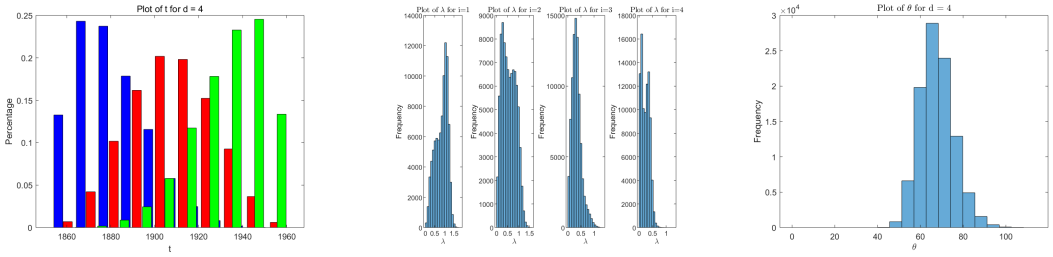
Figure 2: Break point=2

the histogram of t shows that the breakpoints are located at around 1890 and 1920.

The histogram of λ_1, λ_2 and λ_3 suggests that the intensities have a Γ–distribution and focus around 1.0, 0.4 and 0.3.

We can also see that θ intensity is centered around 80.

1.3.3 three breakpoints



(a) t (b) λ (c) θ

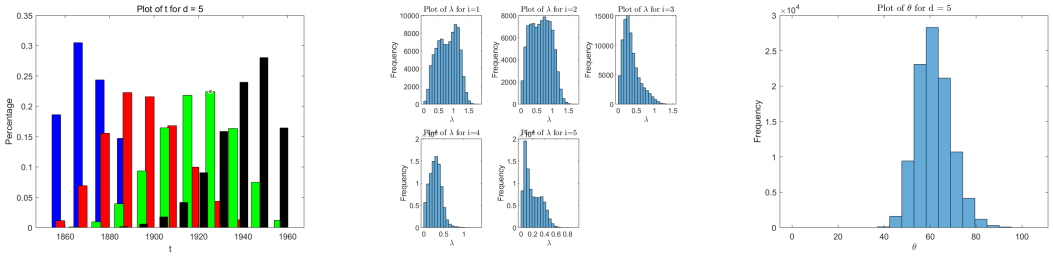
Figure 3: Break point=3

the histogram of t shows that the breakpoints are located at around 1880, 1910 and 1940. As we can easily distinguish the centers of densities in each breakpoint, the result is quite satisfying.

The histogram of $\boldsymbol{\lambda}$ suggests that the intensities have a Γ–distribution and focus around 1.0, 0.7,0.3 and 0.1.

We can also see that θ intensity is centered around 70.

1.3.4 four breakpoints



(a) t (b) λ (c) θ

Figure 4: Break point=4

the histogram of t shows that the breakpoints are located at around 1870,1890,1920 and 1945. As we can easily distinguish the centers of densities in each breakpoint, the estimation is quite satisfying.

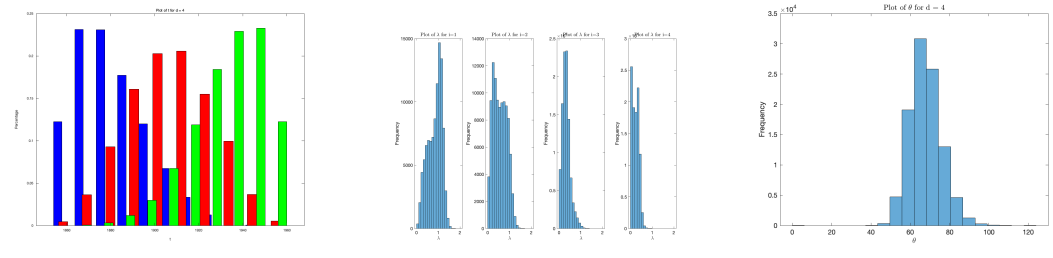
The histogram of $\boldsymbol{\lambda}$ suggests that the intensities have a Γ–distribution and focus around 1.0, 0.7, 0.25, 0.3 and 0.2.

We can also see that θ intensity is centered around 60.

1.4 Sensitivity to the choice of ϑ

The sensitivity of the posteriors to the choice of ϑ is tested under the circumstance of 3 breakpoints and $\rho = 1$.

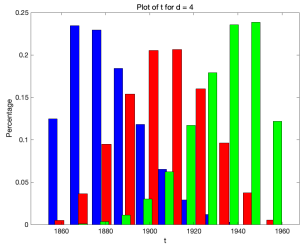
1.4.1 ϑ =1



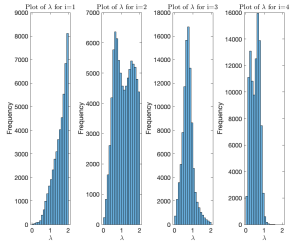
(a) t (b) λ (c) θ

Figure 5: ϑ =1

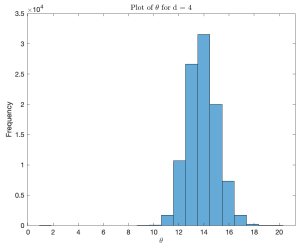
1.4.2 ϑ =10



(a) t



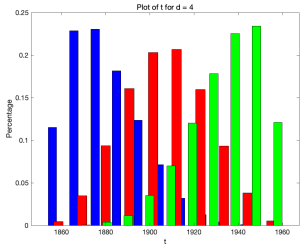
(b) λ



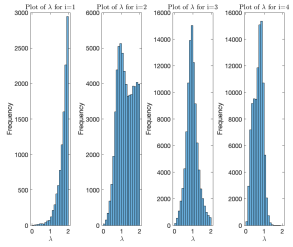
(c) θ

Figure 6: ϑ =10

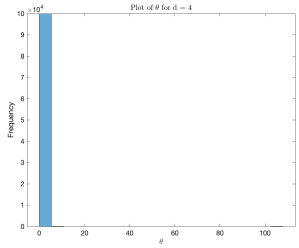
1.4.3 ϑ =40



(a) t



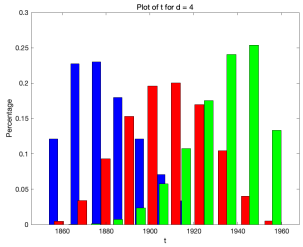
(b) λ



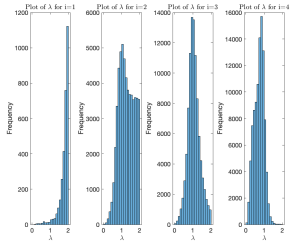
(c) θ

Figure 7: ϑ =40

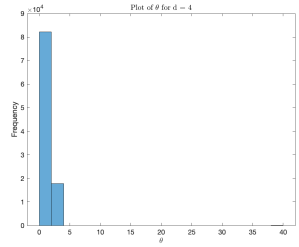
1.4.4 ϑ =100



(a) t



(b) λ



(c) θ

Figure 8: ϑ =100

1.4.5 Summary

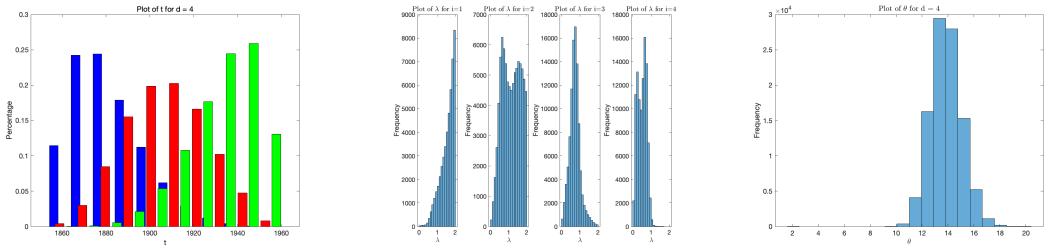
We estimated t, λ, θ with $\vartheta = 1, 10, 40, 100$ and the conclusion is,

- t is not sensitive to the choice of ϑ
- λ is not sensitive to ϑ within a centain range e.g. $[10, 60]$. Even if ϑ was set to some extreme value e.g. 1 or 100, the estimated λ just varies a bit.
- θ is quite sensitive to ϑ, because the estimated θ tends to decrease when ϑ increases.

1.5 Sensitivity to the choice of ρ

The sensitivity of the posteriors to the choice of ρ is tested under the circumstance of 3 breakpoints and $\vartheta = 10$.

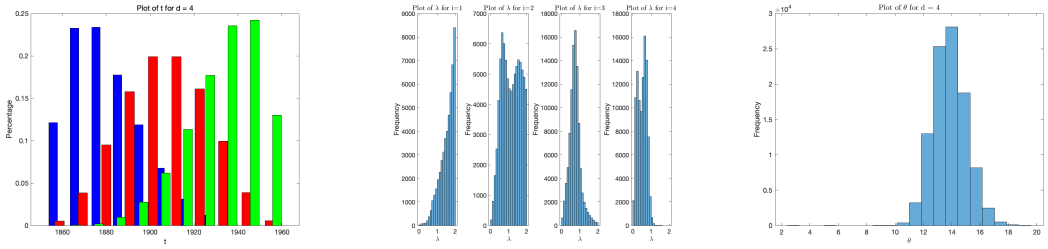
1.5.1 ρ =0.1



(a) t (b) λ (c) θ

Figure 9: ρ =0.1

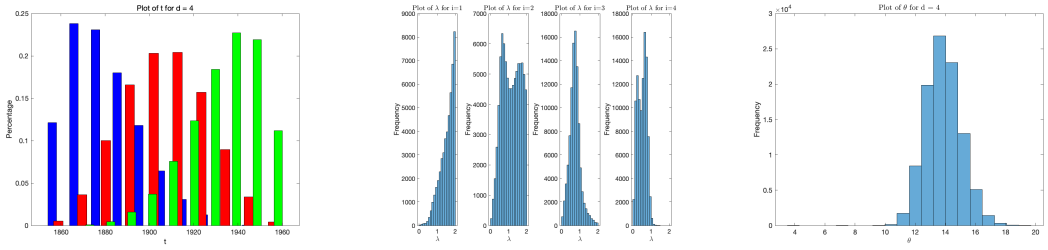
1.5.2 ρ =0.4



(a) t (b) λ (c) θ

Figure 10: ρ =0.4

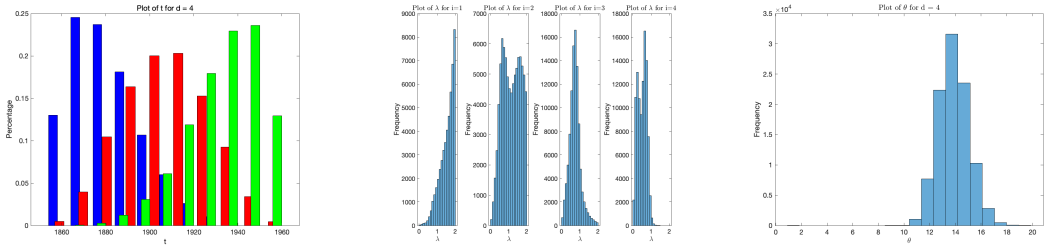
1.5.3 ρ =0.7



(a) t (b) λ (c) θ

Figure 11: ρ =0.7

1.5.4 ρ =1.0



(a) t (b) λ (c) θ

Figure 12: ρ =1.0

1.5.5 Summary

We estimated t, λ, θ with $\rho = 0.1, 0.4, 0.7, 1.0$ and the conclusion is all the posteriors are not sensitive to the choice of ρ.

Also, from the figures of t-posteriors, we can conclude that the mixing for t is also not sensitive to the choice of ρ.

2 EM-based inference in mixture models

2.1 complete data log-likelihood function $f_\theta(\boldsymbol{x}, \boldsymbol{y})$

$$f_\theta(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} f_\theta(x_i, y_i)$$

where

$$f_\theta(x_i, y_i) = f_\theta(y_i|x_i = 0)p_\theta(x_i = 0)(1-x_i)+f_\theta(y_i|x_i = 1)p_\theta(x_i = 1)x_i = g_0(y_i)(1-\theta)(1-x_i)+g_1(y_i)\theta x_i$$

Thus,

$$\log f_\theta(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \log[g_0(y_i)(1-\theta)(1-x_i) + g_1(y_i)\theta x_i]$$

2.2 conditional distribution $f_\theta(\boldsymbol{x}|\boldsymbol{y})$

$$f_\theta(\boldsymbol{x}|\boldsymbol{y}) = \frac{f_\theta(\boldsymbol{x}, \boldsymbol{y})}{f_\theta(\boldsymbol{y})}$$

Since

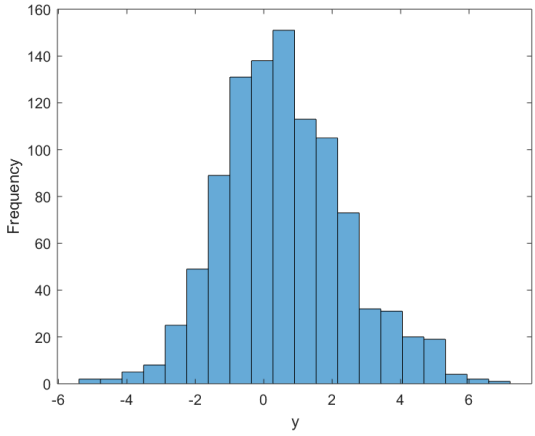
$$f_\theta(y_i) = f_\theta(x_i = 1, y_i)p_\theta(x_i = 1) + f_\theta(x_i = 0, y_i)p_\theta(x_i = 0) = g_0(y_i)(1-\theta) + g_1(y_i)\theta$$

Thus, $X_i|Y_i$ is a Bernoulli distribution, where $P(X_i = 1|Y_i) = \frac{g_1(y_i)\theta}{g_0(y_i)(1-\theta)+g_1(y_i)\theta}$

$$f_\theta(\boldsymbol{x}|\boldsymbol{y}) = \frac{\prod_{i=1}^{n}[g_0(y_i)(1-\theta)(1-x_i) + g_1(y_i)\theta x_i]}{\prod_{i=1}^{n}(g_0(y_i)(1-\theta) + g_1(y_i)\theta)} \tag{1}$$

2.3 EM algorithm for θ

The data is inspected by plotting a histogram.



In the E- step,

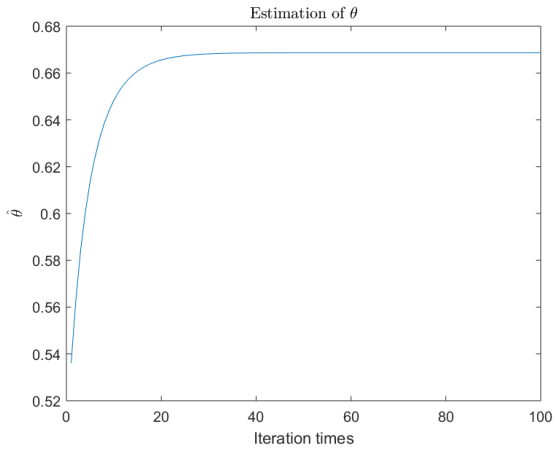
$$\begin{aligned}
Q_{\theta_l}(\theta) = \mathbb{E}_{\theta_l}(\log f_\theta(\mathbf{X},\mathbf{Y})|\mathbf{Y}) &= \mathbb{E}_{\theta_l}(\sum_{i=1}^{n} \log f_\theta(x_i, y_i)|y_i) \\
&= \mathbb{E}_{\theta_l}(\sum_{i=1}^{n} \log[g_0(y_i)(1-\theta)(1-x_i) + g_1(y_i)\theta x_i]) \\
&= \sum_{i=1}^{n} [\log(g_0(y_i)(1-\theta))\ P_{\theta_l}(x_i = 0|y_i) + \log(g_1(y_i)\theta) P_{\theta_l}(x_i = 1|y_i)] \\
&= \sum_{i=1}^{n} \left[\frac{g_0(y_i)(1-\theta_l)\log(1-\theta) + g_1(y_i)\theta_l log(\theta)}{g_0(y_i)(1-\theta_l) + g_1(y_i)\theta_l)}\right] + \text{constant}
\end{aligned}$$

Denote $p_i = \frac{g_1(y_i)\theta_l}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l)}$, so $Q_{\theta_l}(\theta) = \sum_{i=1}^{n}((1-p_i)log(1-\theta) + p_i log(\theta))$

In the M-step

$$Q'_{\theta_l}(\theta) = 0 \Longrightarrow \frac{\sum_{i=1}^{n} p_i}{\theta} = \frac{\sum_{i=1}^{n}(1-p_i)}{1-\theta}$$

$$\theta_{l+1} = \text{argmax}\, Q(\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{\theta_l g_1(y_i)}{\theta_l g_1(y_i) + (1-\theta_l)g_0(y_i)}$$



We set the initial θ_0 to 0.5. As shown in the figure above, the final estimator is 0.66868.