
SF2955 Computer Intensive Methods in Mathematical Statistics

# Home Assignment 2: Statistical inference from coal mine disaster and mixture model data using Markov chain Monte Carlo and the EM-algorithm

*Zi Ying Tan, 961026-T167*
*Xiaohan Hu 970616-T206*

May 20, 2018

# Contents

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

In this problem, we analyzed the data of British coal mining disasters. Assume that there exist $(d-1)$ breakpoints which are denoted by $t_i, i = 2, ..., d$. Set $t_1 = 1851$ and $t_{d+1} = 1963$. The disaster intensity in each interval $[t_i, t_{i+1})$ is $\lambda_i$ and we let $\boldsymbol{\lambda} = (\lambda_1, .., \lambda_d)$.
The time points of the $n = 191$ disasters are denoted by $\boldsymbol{\tau} = (\tau_1, ..., \tau_n)$. We modeled:

$$\lambda(t) = \sum_{i=1}^{d} \lambda_i \mathbb{1}_{[t_i, t_{i+1})}(t) \tag{1.1}$$

$$n_i(\tau) = \text{number of disasters in the sub-interval}[t_i, t_{i+1}) = \sum_{j=1}^{d} \mathbb{1}_{[t_i, t_{i+1})}(\tau_j) \tag{1.2}$$

A $\Gamma(2, \theta)$ prior is put on the intensities with a $\Gamma(2, \alpha)$ hyperprior on $\theta$, where $\alpha$ is a fixed hyperparameter that needs to be specified. In addition, we put a prior

$$f(\boldsymbol{t}) \propto \begin{cases} \prod_{i=1}^{d}(t_{i+1} - t_i) & \text{for } t_1 < t_2 < ... < t_d < t_{d+1} \\ 0 & \text{otherwise} \end{cases} \tag{1.3}$$

on the breakpoints.Then, we can know that

$$f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t}) \propto \exp(-\sum_{i=1}^{d} \lambda_i (t_{i+1} - t_i)) \prod_{i=1}^{d} \lambda_i^{n_i(\boldsymbol{\tau})} \tag{1.4}$$

## 1.1 the marginal posteriors

### 1.1.1 $f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})$

$$\begin{aligned} f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau}) &\propto f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t}, \theta) \cdot f(\boldsymbol{\lambda}, \boldsymbol{t}, \theta) \\ &\propto f(\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{t}) \cdot f(\boldsymbol{\lambda}|\theta) \cdot f(\theta) \cdot f(\boldsymbol{t}) \\ &\propto (\prod_{i=1}^{d} \lambda_i exp(-\lambda_i \theta)\theta^2)(\theta exp(-\theta\alpha)) \\ &\propto exp(-(\alpha + \sum_{i=1}^{d} \lambda_i)\theta)\theta^{2d+1} \end{aligned}$$

i.e. $\Gamma(2d + 2, \alpha + \sum_{i=1}^{d} \lambda_i)$

### 1.1.2 $f(\boldsymbol{\lambda}|\boldsymbol{\tau}, \boldsymbol{t}, \theta)$

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

Thus,

$$f(\lambda_i|\boldsymbol{\tau},\boldsymbol{t},\theta) \propto exp(-(t_{i+1}-t_i+\theta)\lambda_i)\lambda_i^{n_i(\boldsymbol{\tau})+1}$$

i.e. $\Gamma(n_i(\boldsymbol{\tau})+2, t_{i+1}-t_i+\theta)$

#### 1.1.3 $f(\boldsymbol{t}|\boldsymbol{\tau},\boldsymbol{\lambda},\theta)$

$$\begin{aligned} f(\boldsymbol{t}|\boldsymbol{\tau},\boldsymbol{\lambda},\theta) &\propto f(\boldsymbol{\tau}|\boldsymbol{\lambda},\boldsymbol{t},\theta)\cdot f(\boldsymbol{\lambda},\boldsymbol{t},\theta) \\ &\propto f(\boldsymbol{\tau}|\boldsymbol{\lambda},\boldsymbol{t})\cdot f(\boldsymbol{\lambda}|\theta)\cdot f(\theta)\cdot f(\boldsymbol{t}) \\ &\propto exp(-\sum_{i=1}^{d}\lambda_i(t_{i+1}-t_i))\prod_{i=1}^{d}(t_{i+1}-t_i) \qquad \text{for } t_1<t_2<...<t_d<t_{d+1} \end{aligned}$$

### 1.2 Construct a hybird MCMC algorithm that samples from the posterior $f(\theta|\boldsymbol{\lambda},\boldsymbol{t},\boldsymbol{\tau})$

**Algorithm 1** hybird MCMC

**Initialize** $\theta^0, \boldsymbol{\lambda}^0 = (\lambda_1^0,...,\lambda_d^0), \boldsymbol{t}^0 = (t_1^0,...,t_{d+1}^0)$
**for** $k = 1 \to N$ **do**
  $\theta^{k+1} \sim f(\theta|\boldsymbol{\lambda}^k,\boldsymbol{t}^k,\tau)$
  **for** $i = 1 \to d$ **do**
    $\lambda_i^{k+1} \sim f(\lambda_i|\boldsymbol{\tau},\boldsymbol{t}^k,\theta^k)$
  **end for**
  **for** $i = 2 \to d$ **do**
    $R = \rho(t_{i+1}^k - t_i^{k+1})$
    $\epsilon \sim \mathrm{U}(-R,R)$
    $t_i^* = t_i^k + \epsilon$
    set $\delta \leftarrow 1 \bigwedge \frac{f(t_i^*|t_1^{k+1},...,t_{i-1}^{k+1},t_{i+1}^k,...,t_{d+1}^k)p(t_i^k|t_i^*)}{f(t_i^k|t_1^{k+1},...,t_{i-1}^{k+1},t_{i+1}^k,...,t_{d+1}^k)p(t_i^*|t_i^k)}$
    $u \sim U(0,1)$
    **if** $u \le \epsilon$ **then**
      $t_i^{k+1} \leftarrow t_i^*$
    **else**
      $t_i^{k+1} \leftarrow t_i^k$
    **end if**
  **end for**
**end for**

From what we have gotten above,

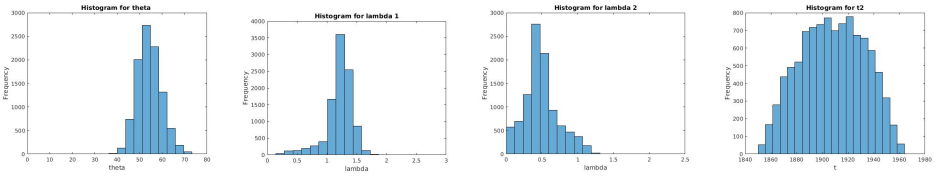
$$f(\theta|\boldsymbol{\lambda},\boldsymbol{t},\boldsymbol{\tau}) \propto \Gamma(2d+2, \alpha+\sum_{i=1}^{d}\lambda_i), \quad f(\lambda_i|\boldsymbol{\tau},\boldsymbol{t},\theta) \propto \Gamma(n_i(\boldsymbol{\tau})+2, t_{i+1}-t_i+\theta)$$

### 1.3 Investigate

We would now investigate the behavior of the MCMC chain for 1,2,3,4 breakpoints with the use of histogram. We have set hyperparameter $\vartheta$ as 2 and $\rho$ as 0.5 for our plots.

### 1.3.1 one breakpoint

We got the plot for $\theta, \boldsymbol{\lambda}, \boldsymbol{t}$:

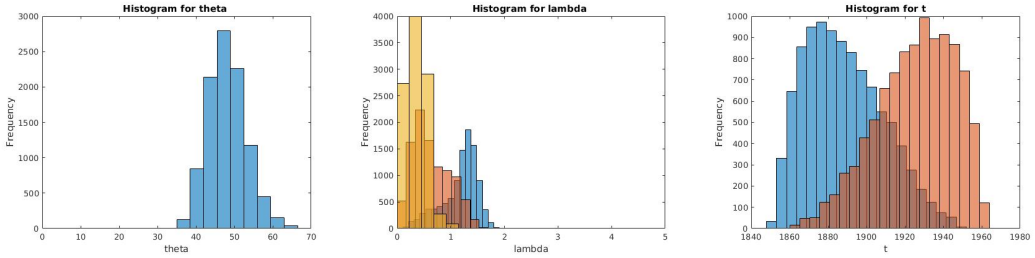



(a) $\theta$ (b) $\lambda_1$ (c) $\lambda_2$ (d) $t_2$

We can see that theta intensity focuses around circa 55 disasters per year and the lambda intensities focuses around circa 1.3 and 0.8 disasters per year respectively, breakpoints focus around 1890 to 1920.

### 1.3.2 two breakpoints

We got the plot for $\theta, \boldsymbol{\lambda}, \boldsymbol{t}$:

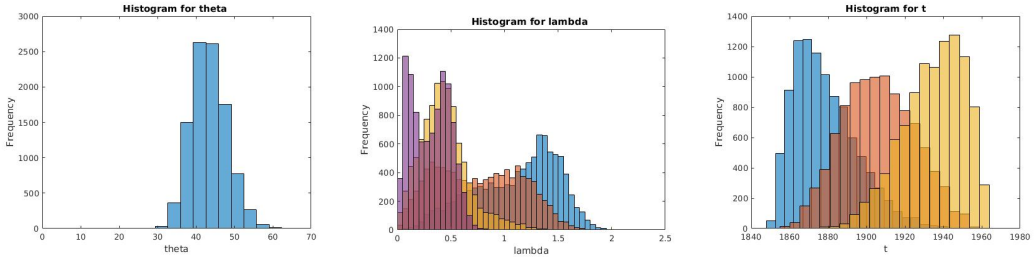



(a) $\theta$ (b) $\lambda$ (c) $t$

We can see that theta intensity focuses around circa 43 disasters per year and the lambda intensities focuses around circa 1.5, 0.4 and 0.3 disasters per year respectively, breakpoints focus around 1875 and 1930.

### 1.3.3 three breakpoints

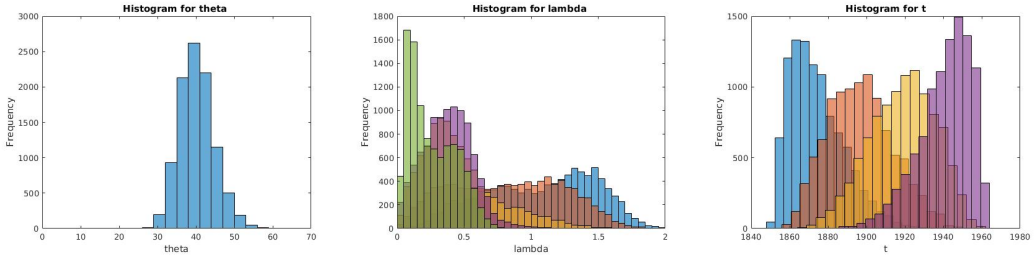
We got the plot for $\theta, \boldsymbol{\lambda}, \boldsymbol{t}$:




(a) $\theta$ (b) $\lambda$ (c) $t$

### 1.3.4 four breakpoints

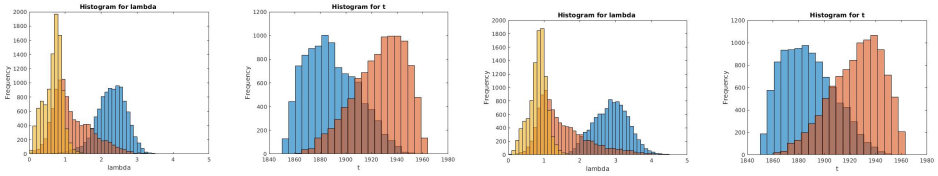
We got the plot for $\theta, \boldsymbol{\lambda}, \boldsymbol{t}$:




(a) $\theta$ (b) $\lambda$ (c) $t$

## 1.4 Sensitivity to the choice of the hyperparameter $\vartheta$

There are many ways to test how sensitive are the posteriors to the choice of the hyperparameter $\vartheta$. The test we decided to utilize would be investigating the differences in the behavior of the MCMC chain for $\boldsymbol{\lambda}, \boldsymbol{t}$ with the varying hyperparmater values($\boldsymbol{\vartheta} = 2, 20, 200$) and $d = 2$. We have set our $\boldsymbol{\rho}$ as 0.5 for our plots as shown below:



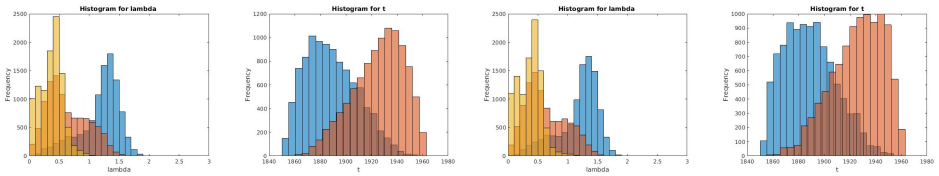

(a) $\lambda$ for $\boldsymbol{\vartheta} = 20$ (b) $t$ for $\boldsymbol{\vartheta} = 20$ (c) $\lambda$ for $\boldsymbol{\vartheta} = 200$ (d) $t$ for $\boldsymbol{\vartheta} = 200$

Figure 5: Histogram for $\lambda$ and $t$ with varying $\boldsymbol{\vartheta}$

From the plots, we are able to observe that changes in $\boldsymbol{\vartheta}$ does not result in large changes in the different histogram plots. Hence, we can conclude that the posteriors are not sensitive to the choice of hyperparameter $\boldsymbol{\vartheta}$.

## 1.5 Sensitivity to the choice of $\rho$

There are many ways to test how sensitive is the mixing and the posteriors to the choice of $\rho$. The test we decided to utilize would be investigating the differences in the behavior of the MCMC chain for $\boldsymbol{\lambda}, \boldsymbol{t}$ with the varying values for $\boldsymbol{\rho} = 0.2, 0.5, 0.8$. We have set our $\boldsymbol{\vartheta}$ as 2 and $d = 2$ for our plots as shown below:




(a) $\lambda$ for $\boldsymbol{\vartheta} = 20$ (b) $t$ for $\boldsymbol{\vartheta} = 20$ (c) $\lambda$ for $\boldsymbol{\vartheta} = 200$ (d) $t$ for $\boldsymbol{\vartheta} = 200$

Figure 6: Histogram for $\lambda$ and $t$ with varying $\boldsymbol{\rho}$

From the plots, we are able to observe that changes in $\boldsymbol{\rho}$ does not result in large changes in the different histogram plots. Hence, we can conclude that the mixing and the posteriors are not sensitive to the choice of $\boldsymbol{\rho}$.

# 2 EM-based inference in mixture models

$X$ is a $\{0,1\}$-valued random variable and $\mathbb{P}(X=1) = 1 - \mathbb{P}(X=0) = \theta$.($\theta$ is unknown)
$Y$ is an observable random variable and:

$$Y|X = 0 \sim g_0(y)dy$$
$$Y|X = 1 \sim g_1(y)dy$$

where $g_0$ and $g_1$ are Gaussian with known means $(\mu_0, \mu_1) = (0, 1)$ and standard deviations $(\sigma_0, \sigma_1) = (1, 2)$.

## 2.1 The complete data log-likelihood function

Set $x_0$ denotes $X = 0$ and $x_1$ denotes $X = 1$

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

for each $y^i (i = 1, ..., n)$ :

$$f_\theta(x_j, y^i) = f_\theta(y^i|x_j) \cdot p_\theta(x_j) \quad for \quad j = 0, 1$$

Thus,

$$f_\theta(x_0, y^i) = g_0(y^i)(1-\theta)$$
$$f_\theta(x_1, y^i) = g_1(y^i)\theta$$

The complete data log-likelihood function:

$$log f_\theta(x_0, y^i) = log(1-\theta) + log(g_0(y^i))$$
$$log f_\theta(x_1, y^i) = log\theta + log(g_1(y^i))$$

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

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

for each $y^i (i = 1, ..., n)$ :

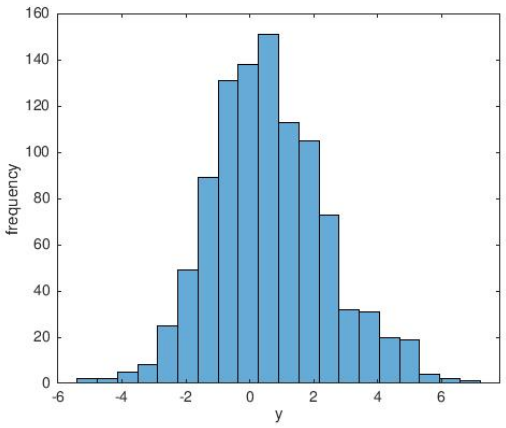
$$\begin{aligned} f_\theta(y^i) &= f_\theta(y^i|x_0)p(x_0) + f_\theta(y^i|x_1)p(x_1) \\ &= g_0(y^i)(1-\theta) + g_1(y^i)\theta \end{aligned}$$

Thus,

$$p_\theta(x_0|y^i) = \frac{f_\theta(x_0, y^i)}{f_\theta(y^i)} = \frac{g_0(y^i)(1-\theta)}{g_0(y^i)(1-\theta) + g_1(y^i)\theta}$$
$$p_\theta(x_1|y^i) = \frac{f_\theta(x_1, y^i)}{f_\theta(y^i)} = \frac{g_1(y^i)\theta}{g_0(y^i)(1-\theta) + g_1(y^i)\theta}$$

## 2.3 Histogram for y

It is paramount to inspect the data by plotting a histogram for y observations. We would obtain the following plot:



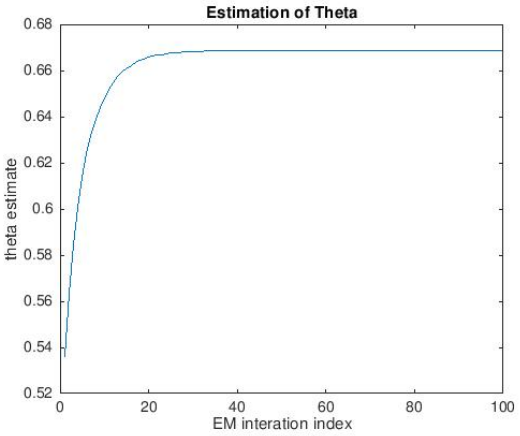
## 2.4 EM algorithm for $\theta$

$$
\begin{aligned}
l(\theta) &= \sum_{j=0}^{1} log \sum_{y^i} p_\theta(x_j, y^i) \\
&= \sum_{j=0}^{1} log \sum_{y^i} p_\theta(x_j|y^i) \frac{p_\theta(x_j, y^i)}{p_\theta(x_j|y^i)} \\
&\geq \sum_{j=0}^{1} \sum_{y^i} p_\theta(x_j|y^i) log \frac{p_\theta(x_j, y^i)}{p_\theta(x_j|y^i)} = \mathscr{Q}(\theta) \\
\theta &= argmax \mathscr{Q}(\theta) = \frac{1}{N} \sum_{i=1}^{N} p_\theta(x_j, y^i)
\end{aligned}
$$

**Algorithm 2** EM algorithm

**Data**: Initial value $\theta_0$
**Result** :$\{\theta_l; l \in \mathbb{N}\}$
**for** $l \to 0, 1, 2, ...$ **do**
  **for** $i \to 0, 1, ..., N$ **do**
    $p_{\theta_l}(x_0|y^i) = \frac{g_0(y^i)(1-\theta_l)}{g_0(y^i)(1-\theta_l)+g_1(y^i)\theta_l}$
    $p_{\theta_l}(x_1|y^i) = \frac{g_1(y^i)\theta_l}{g_0(y^i)(1-\theta_l)+g_1(y^i)\theta_l}$
  **end for**
  set $\theta_{l+1} \to \frac{1}{N} \sum_{i=1}^{N} p_{\theta_l}(x_j, y^i)$
**end for**

After running the EM algorithm, we would obtain the following EM learning curve:




Based on the graph and some calculations, we are able to observe that the final estimate of $\theta$ is 0.6687