Home Assignment 2 - SF2955

Statistical inference from coal mine disaster and mixture model data using Markov chain Monte Carlo and the EM-algorithm

17 May 2019

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

In the first part of the assignment a coal mining example is generalised from one breakpoint to $d-1$ breakpoints. The end points of the dataset are fixed to $t_1 = 1851$ and $t_{d+1} = 1963$. The breakpoints are denoted by t_i, $i = 2, ..., t_d$. The end points and the break points are collected in a vector $\boldsymbol{t} = (t_1, ..., t_{d+1})$. The disaster intensity in each interval $[t_i, t_{i+1})$ is λ_i and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_d)$. The time points of the $n = 191$ disasters are denoted by $\boldsymbol{\tau} = (\tau_1, ..., \tau_n)$. The data is modeled on the interval $t_1 \leq t \leq t_{d+1}$ using an inhomogeneous Poisson process with intensity

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

From the time points of the disasters the number of disasters in the sub-interval $[t_i, t_{i+1})$ is computed by

$$n_i(\boldsymbol{\tau}) = \sum_{j=1}^{n} \mathbb{1}_{[t_i, t_{i+1})}(\tau_j)$$

A $\Gamma(2, \theta)$ prior is put on the intensities (λ) with a $\Gamma(2, \vartheta)$ hyperprior on θ, where ϑ is a fixed hyperparameter that needs to be specified. A prior

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

is put on the breakpoints. We obtain

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

Problem (a) - Compute the marginal posteriors

We start by computing the marginal posteriors $f(\theta|\boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})$, $f(\lambda|\theta, \boldsymbol{t}, \boldsymbol{\tau})$ and $f(\boldsymbol{t}|\theta, \boldsymbol{\lambda}, \boldsymbol{\tau})$ up to normalizing constants. The joint distribution $f(\theta, \boldsymbol{\lambda}, \boldsymbol{t}, \boldsymbol{\tau})$ can be expressed by

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

We know that $\pi(\boldsymbol{\lambda}|\theta)$ follows a $\Gamma(2,\theta)$-distribution and $\pi(\theta)$ follows a $\Gamma(2,\vartheta)$-distribution. This gives

$$\pi(\boldsymbol{\lambda}|\theta) = \prod_{i=1}^{d}\left(\frac{\theta^2}{\Gamma(2)\lambda_i}\exp\left(-\sum_{i=1}^{d}(\lambda_i\theta)\right)\right)$$

and

$$\pi(\theta) = \frac{\vartheta^2}{\Gamma(2)}\theta\exp(-\theta\vartheta)$$

We then obtain the joint distribution

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

The marginal posteriors are now derived by

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

which corresponds to the distribution $\Gamma(2d+2,\vartheta+\sum_{i=1}^{d}\lambda_i)$.

$$f(\boldsymbol{\lambda}|\theta,\boldsymbol{t},\boldsymbol{\tau}) \propto f(\boldsymbol{\tau}|\boldsymbol{\lambda},\boldsymbol{t},\theta)\pi(\boldsymbol{\lambda}|\theta) \propto \exp\left(-\sum_{i=1}^{d}\lambda_i(t_{i+1}-t_i)\right)\prod_{i=1}^{d}\lambda_i^{n_i(\boldsymbol{\tau})}\prod_{i=1}^{d}\left(\theta^2\lambda_i\right)\exp\left(-\sum_{i=1}^{d}(\lambda_i\theta)\right)$$
$$\propto \prod_{i=1}^{d}\lambda_i^{n_i(\boldsymbol{\tau})+1}\exp\left(-\sum_{i=0}^{d}\lambda_i(t_{i+1}-t_i+\theta)\right)$$

Thus each $f(\lambda_i|\theta,\boldsymbol{t},\boldsymbol{\tau})$ have the distribution $\Gamma(2+n_i(\boldsymbol{\tau}),t_{i+1}-t_i+\theta)$.

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

Problem (b) - Constructing a hybrid MCMC algorithm that samples from the posterior $f(\theta,\boldsymbol{\lambda},\boldsymbol{t}|\boldsymbol{\tau})$

To sample from the posterior $f(\theta,\boldsymbol{\lambda},\boldsymbol{t}|\boldsymbol{\tau})$ a hybrid MCMC algorithm was constructed. All components except for the breakpoints $\boldsymbol{t}$ was updated by using Gibbs sampling. To update the breakpoints a Metropolis-Hastings step was used with a Random walk proposal. This means that one breakpoint at a time was updated and for each breakpoint t_i a candidate t_i^* was generated according to

$$t_i^* = t_i + \epsilon, \quad \text{with} \quad \epsilon \sim \text{Unif}(-R,R)$$

and $R = \rho(t_{i+1}-t_{i-1})$ where ρ is a tuning parameter. The proposed vector $\boldsymbol{t}^*$ is accepted with probability $\alpha = \min\left(1, \frac{f(\boldsymbol{t}^*|\theta,\boldsymbol{\lambda},\boldsymbol{\tau})r(\boldsymbol{t}^*|\boldsymbol{t})}{f(\boldsymbol{t}|\theta,\boldsymbol{\lambda},\boldsymbol{\tau})r(\boldsymbol{t}|\boldsymbol{t}^*)}\right)$. Since we have a symmetric proposal kernel r we have

$$\frac{f(\boldsymbol{t}^*|\theta,\boldsymbol{\lambda},\boldsymbol{\tau})r(\boldsymbol{t}^*|\boldsymbol{t})}{f(\boldsymbol{t}|\theta,\boldsymbol{\lambda},\boldsymbol{\tau})r(\boldsymbol{t}|\boldsymbol{t}^*)} = \frac{f(\boldsymbol{t}^*|\theta,\boldsymbol{\lambda},\boldsymbol{\tau})}{f(\boldsymbol{t}|\theta,\boldsymbol{\lambda},\boldsymbol{\tau})}$$

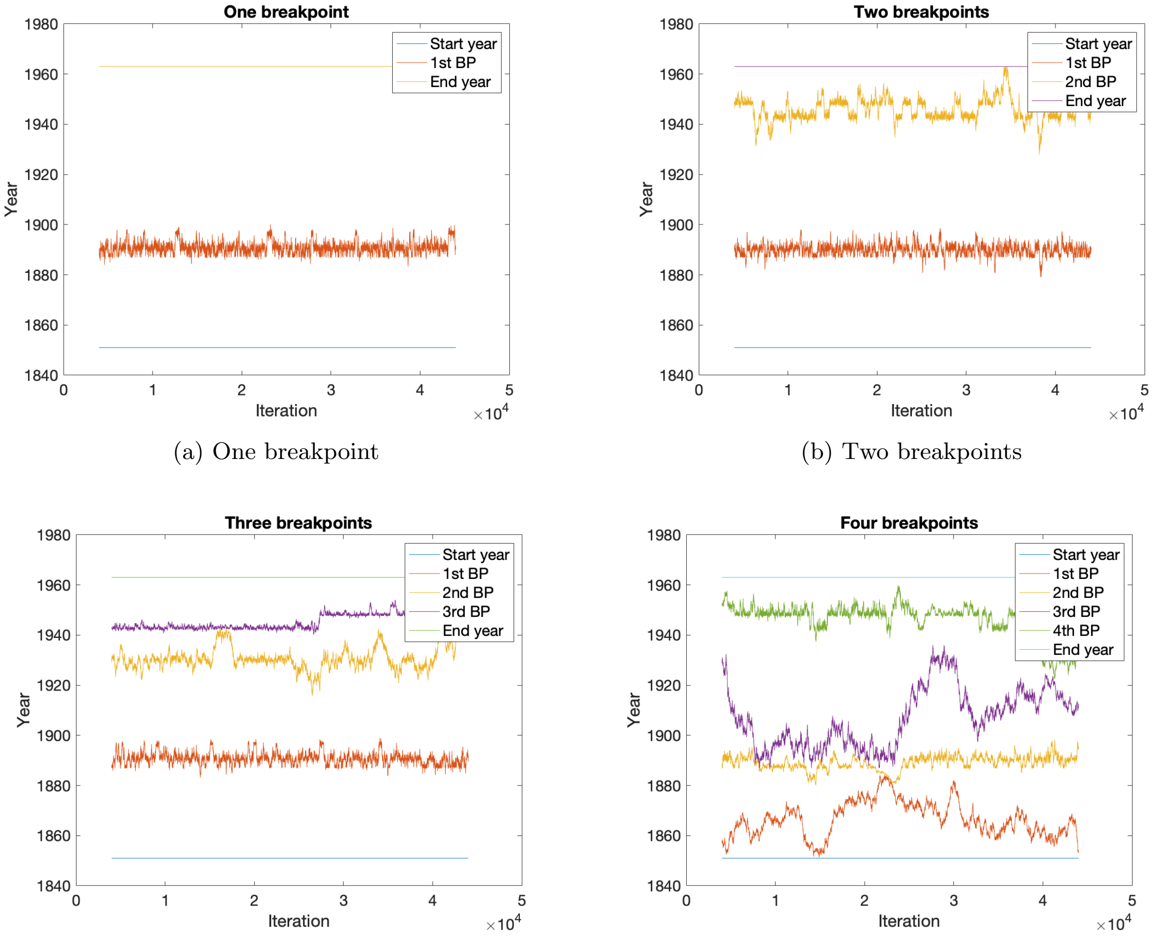
The following hybrid MCMC algorithm is used

```
for $k = 1, ..., N-1$ do
    draw $\theta^{k+1} \sim \Gamma(2d+2, \vartheta + \sum_{i=1}^{d} \lambda_i^k)$
    for $i = 1, ..., d$ draw $\lambda_i^{k+1} \sim \Gamma(2 + n_i(\boldsymbol{\tau}), t_{i+1} - t_i + \theta)$
    for $i = 2, ..., d$ draw $t_i^* = t_i^k + \epsilon$ with $\epsilon \sim \text{Unif}(-R, R)$ and the condition $t_1 < t_2^* < ... < t_d^* < t_{d+1}$
    draw $u$ from $\text{U}(0,1)$
    if $u < \min\left(1, \frac{f(\boldsymbol{t^*}|\theta,\boldsymbol{\lambda},\boldsymbol{\tau})}{f(\boldsymbol{t}|\theta,\boldsymbol{\lambda},\boldsymbol{\tau})}\right)$ then
        $\boldsymbol{t}^{k+1} = \boldsymbol{t^*}$
    else
        $\boldsymbol{t}^{k+1} = \boldsymbol{t^k}$
    end
end
```

Algorithm 1: Hybrid MCMC

Problem (c) - Investigate the behavior of the MCMC chain for 1, 2, 3 and 4 beakpoints

The different number of breakpoints are plotted in figure 1 after a burn-in of 4000 iterations.



(a) One breakpoint

(b) Two breakpoints

(c) Three breakpoints

(d) Four breakpoints

Figure 1: The plots a, b, c and d presents the evolution of the breakpoints after burn-in of 4000 iterations.

For the different breakpoints, we receive the table 1 with parameter values.

# breakpoints	1	2	3	4
t	[1887]	[1893, 1945]	[1882, 1895 ,1947]	[1889.3, 1911.4, 1927.3, 1942.1]
λ	[2.96, 1.09]	[3.14,1.16,0.66]	[3.56, 2.18, 0.83, 0.41]	[3.20, 1.47, 0.69, 1.79, 0.52]
Acceptance rate	38%	30%	34%	29%

Table 1: Parameter values of the posterior for different number of breakpoints

From both the figures in figure 1 and the parameter values in 1, we see that three breakpoints are quite constant. However, when choosing four breakpoints, the values starts to fluctuate. This suggest that we need to increase the burn-in or choose different parameter values.

Problem (d) - How sensitive are the posteriors to the choice of the hyperparameter ν?

We varied ν for three breakpoints and choosing ρ to be $\rho = 0.05$ and generated the table 3.

ν	Acceptance rate	λ	θ	breakpoints
1	29.43%	[3.0740, 0.9217, 1.3419, 0.4457]	[1.1257]	[1892, 1927, 1947]
5	33.17%	[3.2192, 0.7405, 1.4508, 0.2592]	[0.8852]	[1889, 1930, 1948]
10	33.15%	[3.8414, 4.0753, 1.0508, 0.4573]	[0.5733]	[1866, 1889, 1948]
20	35.80%	[3.2748, 1.1103, 1.2115, 0.5393]	[0.1830]	[1893, 1942, 1952]
50	31.15%	[3.3390, 2.7139, 1.1794, 0.6953]	[0.1014]	[1877, 1888, 1952]

Table 2: Table presenting the sensitivity of the parameters t, ρ and λ while varying ν.

Table 2 suggest that λ, the acceptance rate and the breakpoints is quite insensitive to ν, while θ is more sensitive.

Problem (e) - How sensitive is the mixing and the posteriors to the choice of ρ in the proposal distribution?

Letting $\rho = 0.05$ and using three breakpoints, we generate table 3 when varying the value of ν.

ρ	Acceptance rate	λ	θ	breakpoints
1	1.32%	[3.2065 0.7664 1.5779 0.4317]	[0.4612]	[1887, 1944 , 1905]
0.5	3.43%	[4.4419 3.3075 0.8828 1.1383]	[0.3165]	[1871, 1890, 1909]
0.1	18.05 %	[3.5321 3.0844 1.0201 0.7708]	[0.3689]	[1873, 1887, 1947]
0.05	33.02%	[3.1040 1.9799 1.0038 0.9763]	[0.4865]	[1880, 1890, 1952]
0.01	61.80%	[3.5436 1.2377 0.7003 0.3590]	[0.3925]	[1888, 1943, 1955]

Table 3: Table presenting the sensitivity of the mixing and the parameters t, ρ and λ while varying ρ.

From table 3, we see that ρ has a large affect on the acceptance rate, or the mixing, while no significant or systematic implication on the other parameters. The acceptance rate should be around 30%, which means that $\rho = 0.05$ is a good choice.

Part 2: EM-based inference in mixture models

In the second part of the assignment, the aim is to infer the parameter θ with a frequentist approach. Since the mixture model comprises missing data, the maximum likelihood estimator of θ is computed using the EM algorithm.

The mixture model comprises unobservable random variables X, taking on the values {0,1} with the probabilities $\mathbb{P}(X=1)=\theta$ and $\mathbb{P}(X=0)=1-\theta$, and an observable random variable Y, following the distributions

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

Here, $g_0(y) \sim N(0,1)$ and $g_1(y) \sim N(1,2)$. A set of independent observations $\mathbf{y}=(y_1,...,y_n)$ are given and we denote by $\mathbf{x}=(x_1,...,x_n)$ the corresponding unobserved index variables.

Problem (a) - Write, up to additive constants not depending on θ, the complete data log-likelihood function $\theta \mapsto \log f_\theta(\mathbf{x},\mathbf{y})$

We know that

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

where, $f_\theta(\mathbf{y}|\mathbf{x})$ and $f_\theta(\mathbf{x})$ are given in the problem formulation. Thus

$$f_\theta(\mathbf{x},\mathbf{y}) = \prod_{i=1}^{n}\left((g_0(y_i)(1-\theta))^{1-x_i}(g_1(y_i)\theta)^{x_i}\right)$$

The log-likelihood function is then

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

Problem (b) - Determine the conditional distribution $f_\theta(\mathbf{x}|\mathbf{y})$

We have

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

Thus

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

Problem (c) - Implement the EM algorithm and report the results

First, the observed data is inspected in figure 2.

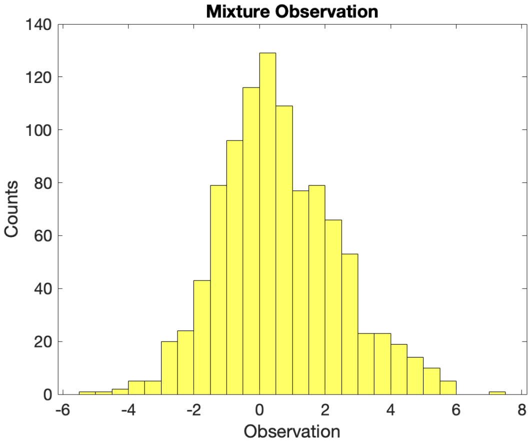


Figure 2: Histogram displaying the distribution of the observed data.

We then want to derive the EM-updating formula for θ

$$\theta_{l+1} = \arg\max Q_{\theta_l}(\theta)$$

where

$$Q_{\theta_l}(\theta) = \mathbb{E}_{\theta_l}[\log f_\theta(\mathbf{x}|\mathbf{y})|\mathbf{y}] = \sum_{i=1}^{n}((1-\mathbb{E}_{\theta_l}[x_i|y_i])\log(1-\theta) + \mathbb{E}_{\theta_l}[x_i|y_i]\log\theta)$$

The expectation value $\mathbb{E}_{\theta_l}[x_i, y_i]$ is

$$\begin{aligned}\mathbb{E}_{\theta_l}[x_i, y_i] &= f(x_i = 0|y_i)\cdot 0 + f(x_i = 1|y_i)\cdot 1 = \frac{g_0(y_i)(1-\theta_l)}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l}\cdot 0 + \frac{g_1(y_i)\theta_l}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l}\cdot 1\\ &= \frac{g_1(y_i)\theta_l}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l}\end{aligned}$$

By inserting this result we get

$$\begin{aligned}Q_{\theta_l}(\theta) &= \sum_{i=1}^{n}\left(\left(1-\frac{g_1(y_i)\theta_l}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l}\right)\log(1-\theta) + \frac{g_1(y_i)\theta_l}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l}\log\theta\right)\\ &= \sum_{i=1}^{n}\left(\frac{g_0(y_i)(1-\theta_l)}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l}\log(1-\theta) + \frac{g_1(y_i)\theta_l}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l}\log\theta\right)\end{aligned}$$

We take the derivative of $Q_{\theta_l}(\theta)$ and set it equal to zero.

$$-\sum_{i=1}^{n}\left(\frac{g_0(y_i)(1-\theta_l)\theta_l}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l}\frac{1}{1-\theta} + \frac{g_1(y_i)\theta_l}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l}\frac{1}{\theta}\right) = 0 \tag{1}$$

From the expression above we get

$$\theta_{l+1} = \frac{1}{n}\sum_{i=1}^{n}\frac{g_1(y_i)\theta_l}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l}$$

The following EM algorithm was implemented

```
**Data:** $g_0(\mathbf{y})$, $g_1(\mathbf{y})$, initial value $\theta_0$
**Result:** $\theta_l, \quad l = 1, ..., N$
**for** $l = 0, ..., N$ **do**
    $\theta_{l+1} = \frac{1}{n}\sum_{i=1}^{n} \frac{g_1(y_i)\theta_l}{g_0(y_i)(1-\theta_l)+g_1(y_i)\theta_l}$
**end**
```

Algorithm 2: EM algorithm

The EM learning curve is displayed in figure 3.

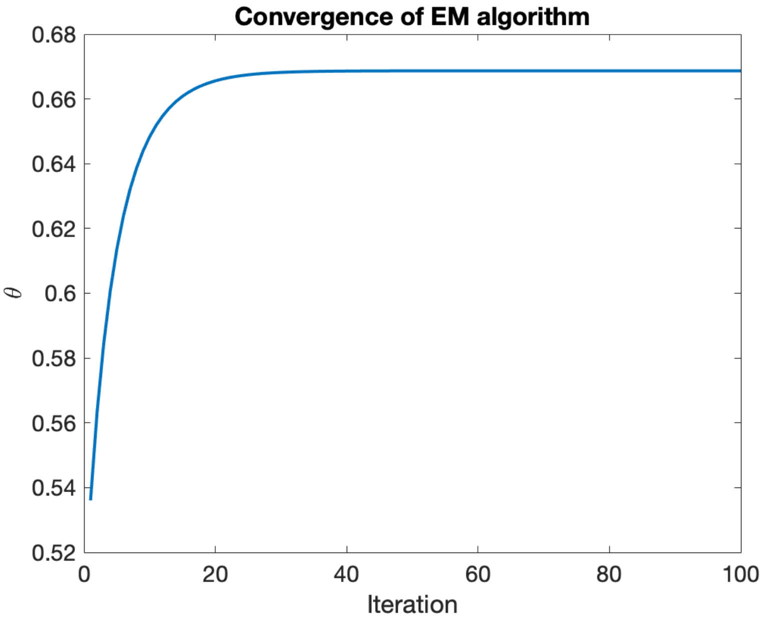


Figure 3: The convergence of the EM algorithm for $N = 100$ iterations and initial guess $\theta_0 = 0.5$.

The EM-algorithm found the optimal value for the parameter to be $\theta = 0.6687$.