EM Algorithm

Noah Gblonyah, Seth Okyere, Michael Throolin

4/14/2022

The EM Algorithm

The EM Algorithm has its roots in work done in the 1950s but really came into statistical prominence after the seminal work of Dempster, Laird, and Rubin, which detailed the underlying structure of the algorithm and illustrated its use in a wide variety of applications. (Casella and Berger 2002).

Another common tool used for getting a maximum-likelihood estimation of censored data is called the EM Algorithm. Here, the 'E' canonically stands for the 'Expectation' step and 'M' represents the 'Maximization' step. Hence the EM Algorithm takes the expectation of the log-likelihood function, then maximizes that quantity. It repeats that process until the parameter converges to a specified value.

Formally, if we let $\theta^{(p)}$ represent the p^{th} iteration of the algorithm to estimate the parameter θ. these two steps can be written out as follows:

> *Expectation (E-Step):* Compute $Q\left(\theta^{(p)}|\theta^{(p-1)}\right) = E\left[\log\left(f(\mathbf{x}|\theta^{(p)}\right)|\mathbf{y}, \theta^{(p-1)}\right]$ where $\mathbf{x}$ represents the complete data and $\mathbf{y}$ represents the censored, or incomplete data.
>
> *Maximization (M-Step):* Maximize $Q\left(\theta^{(p)}|\theta^{(p-1)}\right)$

The EM algorithm can often lead to functions that are tricky to evaluate. However in special cases, such as the exponential family family case, the algorithm becomes much easier to evaluate. Specifically a function $f(\mathbf{x}|\theta)$ is an exponential family if it can be written as $f(\mathbf{x}|\theta) = h(\mathbf{x})c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(\mathbf{x})\right)$. It has been shown that we can use the complete sufficient statistic $T(\mathbf{X}) = \sum_{i=1}^{k} t_i(\mathbf{x})$ to estimate the parameter θ. This is done as follows:

> *Expectation (E-Step):* Estimate $\mathbf{t}(\mathbf{x})$ by finding $\mathbf{t}^{(p-1)} = E\left(\mathbf{t}(\mathbf{x})|\mathbf{y}, \theta^{(p-1)}\right)$ where $\mathbf{x}$ represents the complete data and $\mathbf{y}$ represents the censored, or incomplete data.
>
> *Maximization (M-Step):* Determine $\theta^{(p)}$ as the solution to $E\left(\mathbf{t}(\mathbf{x})|\theta\right) = \mathbf{t}^{(p-1)}$

Example (Exponential Distribution)

For a definitive example, suppose we have data from an exponential distribution with unknown parameter θ. For each sample, we are giving a vector of values, $(c_{(1,i)}, c_{(2,i)}, x_i)$, where $c_{(1,i)}$ represents a left-censoring point, $c_{(2,i)}$ represents a right-censoring point, and x_i is the value the sample. If the i^{th} sample is present, or rather if $c_{(1,i)} < x_i < c_{(2,i)}$, we will call define $y_i = x_i$. Analagously, if the i^{th} sample is absent, or $c_{(1,i)} \geq x_i$ or $x_i \geq c_{(2,i)}$, we will define $z_i = x_i$. Hence, $\mathbf{x}$ is a vector of our complete data, $\mathbf{y}$ is a vector of our incomplete data, and $\mathbf{z}$ is a vector of our unknown, or missing data. Our object is to use the EM algorithm to estimate θ using only y and the censoring points $c_{(1,i)}$ and $c_{(2,i)}$ corresponding with the missing values known in $\mathbf{z}$.

To begin, note that a random sample for the exponential family has a complete sufficient statistic of $\sum_{i=1}^{n} x_i$. Hence for the *E-Step* we must find the expectation of $E(\sum_{i=1}^{n} x_i|\mathbf{y}, \theta^{(p-1)})$.

Using and indicator variable $d_i = \begin{cases} 1 & x_i \in \mathbf{y} \\ 0 & x_i \in \mathbf{z} \end{cases}$ indicating presence and absence of data, we can expand

$$E\left(\sum_{i=1}^{n} x_i|\mathbf{y}, \theta^{(p-1)}\right) = E\left(\sum_{i=1}^{n} y_i d_i + z_i(1-d_i)\right)$$

Now, we need to estimate or missing data, z_i. To do this, we will use the memoryless property of the exponential distribution to get

$$z_i = \begin{cases} \min\{\theta^{(p-1)}, c_{(1,i)}\} & 0 < z_i \leq c_{(1,i)} \\ c_{(2,i)} + \theta^{(p-1)} & z_i > c_{(2,i)} \end{cases}$$

This means

$$\begin{aligned} E(z_i) &= \frac{\int_0^{c_{(1,i)}} \min\{\theta^{(p-1)}, c_{(1,i)}\} \frac{1}{\theta^{(p-1)}} e^{-x_i/\theta^{(p-1)}}\, dx_i + \int_{c_{(2,i)}}^{\infty} (c_{(2,i)} + \theta^{(p-1)}) \frac{1}{\theta^{(p-1)}} e^{-x_i/\theta^{(p-1)}}\, dx_i}{\int_0^{c_{(1,i)}} \frac{x_i}{\theta^{(p-1)}} e^{-x_i/\theta^{(p-1)}}\, dx_i + \int_{c_{(2,i)}}^{\infty} \frac{x_i}{\theta^{(p-1)}} e^{-x_i/\theta^{(p-1)}}} \\ &= \frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}} \end{aligned}$$

Therefore, we can simplify our expectation step down to

$$E\left(\sum_{i=1}^{n} x_i|\mathbf{y}, \theta^{(p-1)}\right) = E\left(\sum_{i=1}^{n} y_i d_i + z_i(1-d_i)\right)$$

$$= \sum_{i=1}^{n} y_i d_i + (1-d_i)\left(\frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}\right)$$

And our maximization step is the solution to

$$\begin{aligned} E(\mathbf{t}(\mathbf{x})|\theta^{(p)}) &= E(\sum_{i=1}^{n} x_i|\theta^{(p)}) = n\,\theta^{(p)} \\ &= \sum_{i=1}^{n} y_i d_i + (1-d_i)\left(\frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}\right) \end{aligned}$$

$$\implies \theta^{(p)} = \frac{1}{n}\sum_{i=1}^{n} y_i d_i + (1-d_i)\left(\frac{\min\{\theta^{(p-1)}, c_{(1,i)}\} - \min\{\theta^{(p-1)}, c_{(1,i)}\} e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}{\theta^{(p-1)} - (c_{(1,i)} + \theta^{(p-1)}) e^{-c_{(1,i)}/\theta^{(p-1)}} + (c_{(2,i)} + \theta^{(p-1)}) e^{-c_{(2,i)}/\theta^{(p-1)}}}\right)$$

Simulation

In the actuarial context, we could encounter data that includes the deductible, policy limit, and losses for each customer, where the losses would be unreported if they exceed the policy limit or are below the deductible.

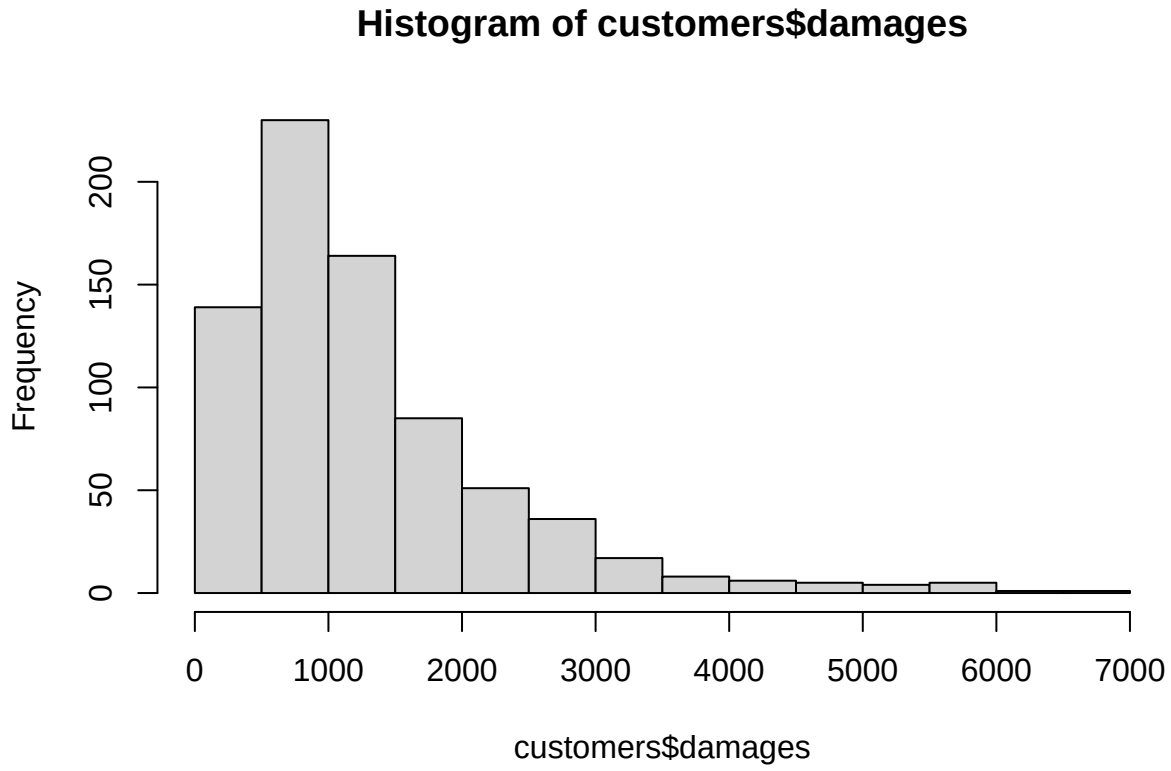
We will use computer simulation to see how well the algorithm holds given the parameter $\theta = 1000$.

```
set.seed(53523)

#Simulate Data
n_customers <- 1000
```

```r
damages <- round(rexp(n_customers, 1/1000),2)
deductible <- round(runif(n_customers, min = 0, max = 600))
limit <- round(runif(n_customers, min = 5000, max = 20000))

#combine data into one data frame
customers <- as.data.frame(cbind(deductible, limit, damages))
#censor data
customers[(damages < deductible),]$damages = NA
customers[(limit < damages),]$damages = NA
hist(customers$damages)
```




```r
#Separate into observed and unobserved data frames
observed_data <- customers[!is.na(customers$damages),]
unobserved_data <- customers[is.na(customers$damages),]


#EM Algorithm
sum_observed <- sum(observed_data$damages)

#Initialize theta
theta_new <- 500
theta <- 0

#iterate until difference between previous theta and new theta is small
while((theta - theta_new)^2 > 0){
```

```
  theta <- theta_new

  #Expectation step
  m <- min(theta, unobserved_data$deductible)

  numerator <- m-m*exp(-unobserved_data$deductible/theta) +
      (unobserved_data$limit +theta)*exp(-unobserved_data$limit/theta)

  denominator <- theta -(unobserved_data$deductible + theta)*
    exp(-unobserved_data$deductible/theta) +
    (unobserved_data$limit + theta)*exp(-unobserved_data$limit/theta)

  expectation <- sum_observed + sum(numerator/denominator)

  #Maximization step
  theta_new <- (expectation)/n_customers
}
theta_new #display outcome
```

```
## [1] 964.4086
```

The output from this simulation estimated the value of θ to be 964.4086115, which is 35.5913885 from the known value of 1000.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference.* Duxbury.