Homework 4 - STAT 540
Amal Agarwal

Answer 1

Given that $A_1, A_2, ..., A_m \sim$ Exponential(θ). Assume we observe $Y_1, Y_2, ..., Y_m$, where $Y_i :=$ min(A_i, τ) where $\tau = 100$ days. The observed log-likelihood function which is needed to be maximized is w.r.t. θ is

$$l_{\text{obs}}(\theta) = \sum_{i:Y_i<\tau} \left(-\log(\theta) - \frac{Y_i}{\theta} \right) + \sum_{i:Y_i=\tau} \left(\frac{-Y_i}{\tau} \right)$$

The complete log-likelihood function can be written as

$$l_{\text{comp}}(\theta) = \sum_{i=1}^{m} \left(-\log(\theta) - \frac{A_i}{\theta} \right)$$

The pseudocode for the algorithm is given as follows:

(a) Start off with some initial guess for $\theta^{(1)}$.

(b) At the k^{th} iterate of EM algorithm, the E-step involves writing down surrogate function which is the expectation of complete log likelihood function as defined above given the observed data $\underset{\sim}{Y}$ and $\theta^{(k)}$. This can be evaluated as follows:

$$\begin{aligned} E_{\theta^{(k)}}(l_{\text{comp}}(\theta)|\underset{\sim}{Y}) &= \sum_{i:Y_i<\tau} \left(-\log(\theta) - \frac{Y_i}{\theta} \right) + \sum_{i:Y_i=\tau} \left(-\log(\theta) - \frac{E_{\theta^{(k)}}(X_i|\underset{\sim}{Y})}{\theta} \right) \\ &= -m\log(\theta) - \frac{\sum\limits_{i:Y_i<\tau} Y_i}{\theta} - \frac{(\tau + \theta^{(k)})(m-u)}{\theta} \qquad (1) \\ &= -m\log(\theta) - \frac{\sum\limits_{i=1}^{m} Y_i}{\theta} - \frac{(\theta^{(k)})(m-u)}{\theta} \end{aligned}$$

Note here u is the number of observations that are uncensored i.e. strictly less than τ.

(c) The M-step involves maximizing the function in (1) w.r.t. θ. Differentiating it and putting the derivative equal to zero, we obtain the maximizer $\theta^{(k+1)}$ as

$$\theta^{(k+1)} = \frac{\sum\limits_{i=1}^{m} Y_i + \theta^{(k)}(m-u)}{m} \qquad (2)$$

Since the E step and M steps leads to (2), repeat this until convergence.

Answer 2

The following plot gives the estimate of θ versus the number of iterations:

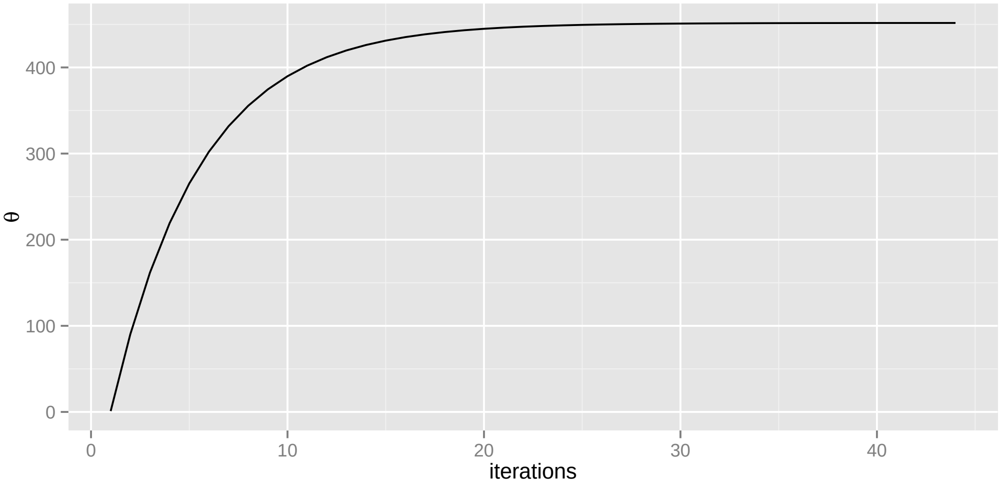

Figure 1: Plot of θ vs. iterations

Now we can choose a suitable starting value by calculating the MLE for the uncensored data (< 100) assuming that follows a truncated exponential distribution. The starting value was found to be **244.9**. Running the algorithm for this starting value leads us to similar convergence as before.

For convergence criteria in this EM algorithm, at k^{th} iterate, the absolute value of the difference of last two estimates i.e. $|\theta^{(k)} - \theta|$ is checked against some arbitrary pre-defined tolerance of 10^{-6}. The algorithm is terminated when the difference becomes less than this tolerance.

Answer 3

The final estimate of θ was obtained as **451.66**.

Answer 4

Non-parametric bootstrap sampling was used to estimate the standard error estimate for $\hat{\theta}_{\text{EM}}$. The following steps describe the bootstrap approach:

(a) Observed data $\underset{\sim}{Y}$ was resampled with replacement to obtain a Bootstrap sample $\underset{\sim}{Y}^*$.

(b) For number of EM iterations defined as 50 and starting value 451, the EM algorithm was used to obtain MLE estimate $\hat{\theta}^*_{\text{EM}}$ for the bootstrap sample $\underset{\sim}{Y}^*$.

(c) The above two steps were repeated B times, where B is the Bootstrap sample size, to obtain $\hat{\theta}_{\text{EM}}^{*(1)}, \hat{\theta}_{\text{EM}}^{*(2)}, ..., \hat{\theta}_{\text{EM}}^{*(B)}$. The sample standard deviation of $\hat{\theta}_{\text{EM}}^{*(1)}, \hat{\theta}_{\text{EM}}^{*(2)}, ..., \hat{\theta}_{\text{EM}}^{*(B)}$ gives an estimate of standard error for $\hat{\theta}_{\text{EM}}$. Note that, here B was chosen such that the absolute value of the difference of estimate of the standard error for B and $(B - 100)$ samples becomes less than an arbitrary pre-defined tolerance of 10^{-6}.

The estimate of the standard error obtained was **32.98**. The corresponding bootstrap sample size is **24208**.

The convergence of the standard error estimate with increasing number of bootstrap samples can be seen in the following plot.

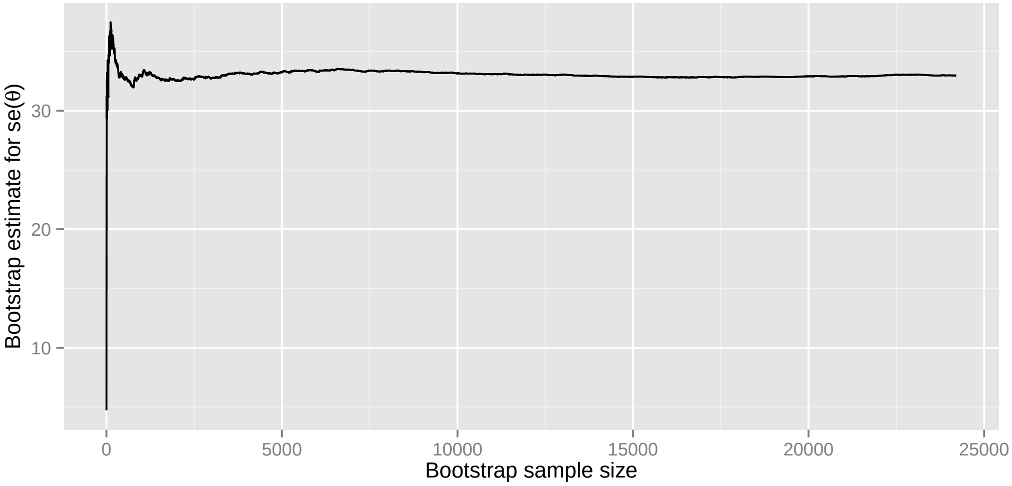

Figure 2: Plot of se(θ) vs. Bootstrap sample size

Answer 5

The bootstrap 95% t confidence interval for θ was obtained as (**378.28**, **506.38**).

Answer 6

An alternative approach to obtain standard error estimates would be to compute the square root of inverse of the fisher information. This asymptotic approach is much easier to implement computationally atleast in this case since the fisher information $E\left(\frac{\partial^2}{\partial\theta^2}(l_{\text{comp}}(\theta))\right)$ can be computed analytically. However for cases cases where this analytical calculation is not possible, the numerical approximation to the Fisher information might be too computationally expensive compared to Bootstrap approach.