1. (Bayesian)

(a)

$$p(\boldsymbol{T}|\theta, k) = \prod_{i=1}^{n} \theta k T_i^{k-1} \exp\left(-\theta T_i^k\right) = (\theta k)^n \left(\prod_{i=1}^{n} T_i\right)^{k-1} \exp\left(-\theta \sum_{i=1}^{n} T_i^k\right)$$

(b)

$$\begin{aligned} p(\theta|\boldsymbol{T}, k) \propto p(\boldsymbol{T}|\theta, k)\pi(\theta) \propto \theta^n \exp\left(-\theta \sum_{i=1}^{n} T_i^k\right) \theta^{a-1} \exp(-b\theta) \\ \sim \text{Gamma}\left(a+n, b+\sum_{i=1}^{n} T_i^k\right) \end{aligned}$$

(c)

$$\begin{aligned} p(\boldsymbol{T}) &= \int p(\boldsymbol{T}|\theta)\pi(\theta)d\theta = \int (\theta k)^n \left(\prod_{i=1}^{n} T_i\right)^{k-1} \exp\left(-\theta \sum_{i=1}^{n} T_i^k\right) \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp(-b\theta) d\theta \\ &= k^n \left(\prod_{i=1}^{n} T_i\right)^{k-1} \frac{b^a}{\Gamma(a)} \int \theta^{n+a-1} \exp\left\{-\theta\left(\sum_{i=1}^{n} T_i^k + b\right)\right\} d\theta \\ &= k^n \left(\prod_{i=1}^{n} T_i\right)^{k-1} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+n)}{\left(\sum_{i=1}^{n} T_i^k + b\right)^{a+n}} \end{aligned}$$

(d) With $a_p = a+n, b_p = b + \sum_{i=1}^{n} T_i^k$,

$$\begin{aligned} p(\tilde{T}|\boldsymbol{T}) &= \int p(\tilde{T}|\theta)p(\theta|\boldsymbol{T})d\theta = \int \theta k \tilde{T}^{k-1} \exp(-\theta \tilde{T}^k) \frac{b_p^{a_p}}{\Gamma(a_p)} \theta^{a_p-1} \exp(-\theta b_p) d\theta \\ &= k\tilde{T}^{k-1} \frac{b_p^{a_p}}{\Gamma(a_p)} \int \theta^{a_p+1-1} \exp\left\{-\theta(b_p + \tilde{T}^k)\right\} d\theta \\ &= k\tilde{T}^{k-1} \frac{b_p^{a_p}}{\Gamma(a_p)} \frac{\Gamma(a_p+1)}{(b_p+\tilde{T}^k)^{a_p+1}} \end{aligned}$$

To obtain the predictive posterior samples, for the j-th iteration,

- draw $\theta^{(j)} \sim \text{Gamma}\left(a+n, b+\sum_{i=1}^{n} T_i^k\right)$;
- sample $\tilde{T}^{(j)} \sim \text{Weibull}(\theta^{(j)}, k)$.

Thus, we can obtain $\left\{\tilde{T}^{(j)}\right\}_{j=1}^{M}$ as M predictive posterior samples.

(e) By taking $\tilde{u} = \tilde{T}^k$,

$$p(\tilde{u}|\boldsymbol{T}) = k\tilde{u}^{\frac{k-1}{k}} \frac{b_p^{a_p}}{\Gamma(a_p)} \frac{\Gamma(a_p+1)}{(b_p+\tilde{u})^{a_p+1}} \frac{1}{k} \tilde{u}^{\frac{1}{k}-1} = \frac{b_p^{a_p} a_p}{(b_p+\tilde{u})^{a_p+1}}.$$

Therefore, $\tilde{u} + b_p = \tilde{T}^k + b_p$ follows a Pareto distribution with parameters (b_p, a_p).

(f)

$$\begin{aligned}
P(\boldsymbol{T}|M_j) &= \int \pi(\theta_j|M_j) p(\boldsymbol{T}|\theta_j, M_j) d\theta_j \\
&= k_j^n \left(\prod_{i=1}^n T_i\right)^{k_j-1} \frac{b_j^{a_j}}{\Gamma(a_j)} \frac{\Gamma(a_j+n)}{\left(\sum_{i=1}^n T_i^{k_j} + b_j\right)^{a_j+n}} \\
P(M_j|\boldsymbol{T}) &= \frac{P(\boldsymbol{T}|M_j)P(M_j)}{\sum_{l=1}^3 P(\boldsymbol{T}|M_l)P(M_l)} = \frac{P(\boldsymbol{T}|M_j)}{\sum_{l=1}^3 P(\boldsymbol{T}|M_l)} \\
&= \frac{k_j^n \left(\prod_{i=1}^n T_i\right)^{k_j-1} \frac{b_j^{a_j}}{\Gamma(a_j)} \frac{\Gamma(a_j+n)}{\left(\sum_{i=1}^n T_i^{k_j} + b_j\right)^{a_j+n}}}{\sum_{l=1}^3 k_l^n \left(\prod_{i=1}^n T_i\right)^{k_l-1} \frac{b_l^{a_l}}{\Gamma(a_l)} \frac{\Gamma(a_l+n)}{\left(\sum_{i=1}^n T_i^{k_l} + b_l\right)^{a_l+n}}}
\end{aligned}$$

2. (EM algorithm)

(a) We should first divide it to two cases: exactly observed ($\Delta_i = 1$, using $f(L)$) and right-censored ($\Delta_i = 0$, using $F(R) - F(L)$). Observed likelihood (only contain observed data X_i, Δ_i):

$$\begin{aligned}
p(\{X_i, \Delta_i\}_{i=1}^n|\lambda) &= \prod_{i=1}^n f(X_i|\lambda)^{\Delta_i} S(X_i|\lambda)^{1-\Delta_i} = \prod_{i=1}^n \{\lambda \exp(-\lambda X_i)\}^{\Delta_i} \{\exp(-\lambda X_i)\}^{1-\Delta_i} \\
&= \lambda^{\sum_{i=1}^n \Delta_i} \exp\left(-\lambda \sum_{i=1}^n X_i\right)
\end{aligned}$$

(b)

$$\frac{\partial \log p(\{X_i, \Delta_i\}_{i=1}^n|\lambda)}{\partial \lambda} = \frac{\sum_{i=1}^n \Delta_i}{\lambda} - \sum_{i=1}^n X_i = 0,$$

we can derive the MLE

$$\hat{\lambda} = \frac{\sum_{i=1}^n \Delta_i}{\sum_{i=1}^n X_i}.$$

The Fisher information matrix of λ has the form,

$$I(\lambda) = -E\left[\frac{\partial^2 \log p(\{X_i, \Delta_i\}_{i=1}^n|\lambda)}{\partial \lambda^2}\right] = E\left[\frac{\sum_{i=1}^n \Delta_i}{\lambda^2}\right].$$

Thus, $\widehat{\text{Var}}(\hat{\lambda}) = \frac{1}{I(\lambda)} = \left(\frac{\sum_{i=1}^n \Delta_i}{\hat{\lambda}^2}\right)^{-1} = \frac{\hat{\lambda}^2}{\sum_{i=1}^n \Delta_i}$

(c) Complete data likelihood:

$$p(\{T_i, X_i, \Delta_i\}_{i=1}^n|\lambda) = \lambda^n \exp\left(-\lambda \sum_{i=1}^n T_i\right)$$

(d) Do expectation/integration for missing data T_i on θ_{old}, we get the Q-function. For exactly observation $\Delta_i = 1$, we have $E(T_i) = X_i$, for right-censored with censoring time X_i, $\Delta_i = 0$, it's a truncexpon distribution, easy to compute $E(T_i) = X_i + \frac{1}{\lambda_{old}}$.

$$\begin{aligned}
Q(\lambda|\lambda^{old}) &= E_{T_i \sim \lambda^{old}}\left[\log p(\{T_i, X_i, \Delta_i\}_{i=1}^n|\lambda)\right] \\
&= n\log(\lambda) - \lambda \sum_{i=1}^n E(T_i) \\
&= n\log(\lambda) - \lambda \sum_{i=1}^n \left\{\Delta_i X_i + (1-\Delta_i)\left(X_i + \frac{1}{\lambda^{\text{old}}}\right)\right\} \\
&= n\log(\lambda) - \lambda \sum_{i=1}^n (X_i + \frac{1-\Delta_i}{\lambda^{\text{old}}})
\end{aligned}$$

(e) – M-step: let $\frac{\partial Q}{\partial \lambda} = 0$, Update

$$\lambda^{\text{new}} = \max_\lambda Q(\lambda|\lambda^{\text{old}}) = \frac{n}{\sum_{i=1}^n T_i'} = \frac{n}{\sum_{i=1}^n \left(X_i + \frac{1-\Delta_i}{\lambda^{\text{old}}}\right)}$$

Repeat the iteration, λ will converge.

(f) The MLE $\hat{\lambda} = \frac{\sum_{i=1}^n \Delta_i}{\sum_{i=1}^n X_i} = \frac{4}{5.9} = 0.6780$.

By using the EM algorithm with $\lambda^{(0)} = 0.8$,

$$\lambda^{\text{new}} = \frac{n}{\sum_{i=1}^n \left(X_i + \frac{1-\Delta_i}{\lambda^{\text{old}}}\right)} = \frac{6}{5.9 + \frac{2}{\lambda^{\text{old}}}}$$

$$\lambda^{(1)} = 0.7143;$$
$$\lambda^{(2)} = 0.6897;$$
$$\lambda^{(3)} = 0.6818.$$

The estimation of $\hat{\lambda}$ obtained by the EM algorithm is close to the MLE.

*** **Difference with Gibbs sampling in Ass2**

In Gibbs sampling, we draw samples for latent variables T_i, then **draw samples** for λ, do it repeatedly and λ will converge.
In EM, we integrate out latent variables T_i to get Q, then maximize Q-func to obtain $\lambda = f(\lambda_{old})$ expression directly, there is **no sampling**. But when we can't integrate out T_i, we still need to do sampling, like MCEM.

So Gibbs sampling seems to be more general.