ST740 HW1

Chenyin Gao

2023-09-13

1

(a)

Choose the prior distribution $\pi(\beta)$ as $\beta \sim N_q(\mu_\beta, \Sigma_\beta)$ and derive its posterior distribution:

$$\begin{aligned}
p(\beta \mid \boldsymbol{Y}) &\propto \pi(\beta) f(\boldsymbol{Y} \mid \beta) \\
&\propto \exp\left\{-\frac{1}{2}(\beta-\mu_\beta)^\intercal \Sigma_\beta^{-1}(\beta-\mu_\beta)\right\} \times \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}(Y_i-\boldsymbol{X}_i\beta)^\intercal \Sigma_i^{-1}(Y_i-\boldsymbol{X}_i\beta)\right\} \\
&\propto \exp\left[-\frac{1}{2}\left\{\beta^\intercal(\Sigma_\beta^{-1}+\sum_i \boldsymbol{X}_i^\intercal W_i \boldsymbol{X}_i)\beta - 2\beta^\intercal(\Sigma_\beta^{-1}\mu_\beta + \sum_i \boldsymbol{X}_i^\intercal W_i Y_i)\right\}\right],
\end{aligned}$$

where $Y_i = (Y_{i1}, \cdots, Y_{im_i})$ and $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \cdots, \boldsymbol{X}_{im_i})^\intercal \in \mathbb{R}^{m_i \times q}$ which concatenate m_i the outcomes and covariates for i-th subject.

Denote $\boldsymbol{M} = \Sigma_\beta^{-1} + \sum_i \boldsymbol{X}_i^\intercal W_i \boldsymbol{X}_i$ and $\boldsymbol{b} = \Sigma_\beta^{-1}\mu_\beta + \sum_i \boldsymbol{X}_i^\intercal W_i y_i$, we can complete the square and obtain the posterior distribution for β as $p(\beta \mid \boldsymbol{Y}) = N_q(\boldsymbol{M}^{-1}\boldsymbol{b}, \boldsymbol{M}^{-1})$.

(b)

Let $m_i = 1$ for all $i \in \{1, \cdots, n\}$ and $W_i = \Sigma_i = 1$ for simplicity, we can verify that $Y_{ij} \sim N(\boldsymbol{X}_{ij}^\intercal\beta, 1)$:

$$\begin{aligned}
p(\beta \mid \boldsymbol{Y}) &\propto \exp\left\{-\frac{1}{2}(\beta-\mu_\beta)^\intercal \Sigma_\beta^{-1}(\beta-\mu_\beta)\right\} \times \prod_{i=1}^{n}\prod_{j=1}^{m_i} \exp\left\{-\frac{1}{2}(Y_{ij}-\boldsymbol{X}_{ij}\beta)^\intercal \Sigma_i^{-1}(Y_{ij}-\boldsymbol{X}_{ij}\beta)\right\} \\
&\propto \exp\left[-\frac{1}{2}\left\{\beta^\intercal(\Sigma_\beta^{-1}+\sum_{i,j} \boldsymbol{X}_{ij}^\intercal W_i \boldsymbol{X}_{ij})\beta - 2\beta^\intercal(\Sigma_\beta^{-1}\mu_\beta + \sum_{i,j} \boldsymbol{X}_{ij}^\intercal W_i Y_{ij})\right\}\right],
\end{aligned}$$

which matches our claims in (a).

(c)

To find the subjects who provide the most information about β in terms of posterior distribution, we could compute the standardized mean contribution of each subject by $\left(\sum_i \boldsymbol{X}_i^\intercal W_i \boldsymbol{X}_i\right)^{-1/2}\left(\boldsymbol{X}_i^\intercal W_i Y_i\right)$ and find the largest one.

2

```r
library(survival)
Y <- lung$time
```

```
delta <- lung$status == 1
sex <- lung$sex == 2
df <- data.frame(Y = Y, delta = delta, sex = sex)
```

(a)

The likelihood function in terms of parameter λ and data $(Y_1, T_1, \delta_1), \cdots, (Y_n, T_n, \delta_n)$ is

$$
\begin{aligned}
L(\lambda \mid Y, T, \delta) &= \prod_{i=1}^{n} \{\lambda \exp(-\lambda Y_i)\}^{1-\delta_i} \times \{\exp(-\lambda T_i)\}^{\delta_i} \\
&= \lambda^{\sum_i (1-\delta_i)} \exp\left[-\lambda \sum_{i=1}^{n} \{(1-\delta_i) Y_i + \delta_i T_i\}\right].
\end{aligned}
$$

(b)

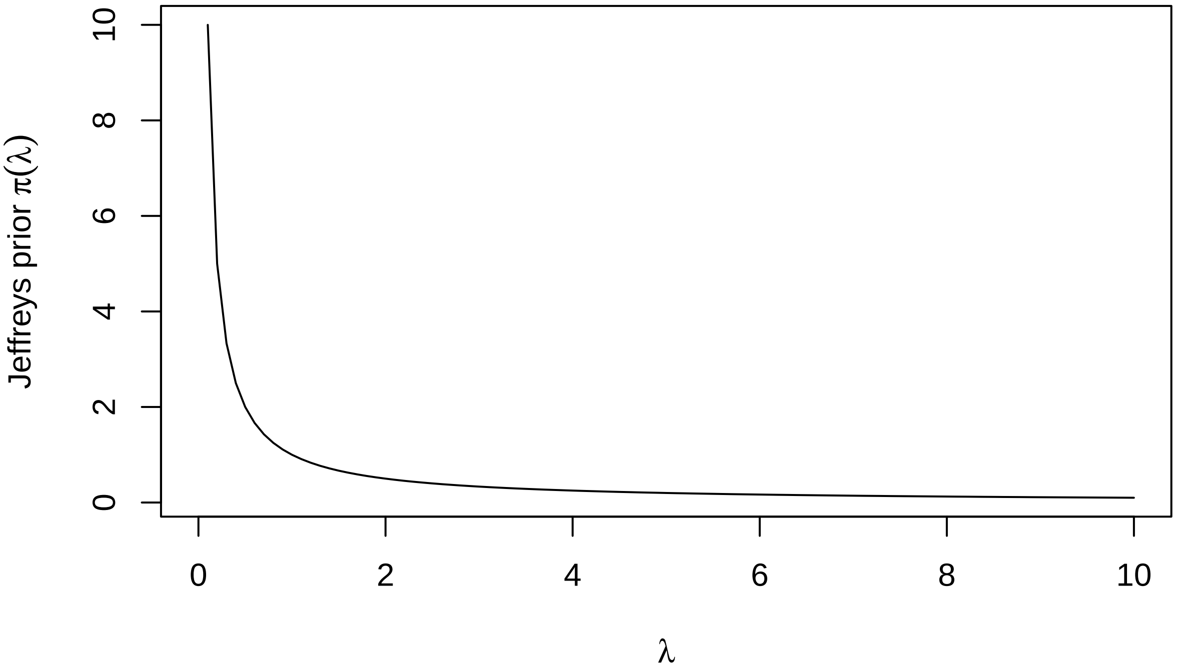
To derive the Jeffreys prior, we need to compute the Fisher information of the model at any λ:

$$
I(\lambda) = -E\left\{\frac{\partial^2}{\partial \lambda^2} \log L(\lambda \mid Y, T, \delta)\right\} = E\left\{\frac{\sum_{i=1}^{n}(1-\delta_i)}{\lambda^2}\right\} = \frac{n - \sum_i \exp(\lambda T_i)}{\lambda^2}.
$$

Therefore, the Jeffreys prior is a non-informative prior:

$$
\pi(\lambda) \propto \sqrt{I(\lambda)}.
$$

To ensure that the Jeffreys prior gives a proper posterior, it is necessary that for any data $(Y_1, T_1, \delta_1), \cdots, (Y_n, T_n, \delta_n)$, the marginal distribution is finite, which only holds if and only if there is at least one uncensored observation in the sample (page 10, *Jeffreys priors for survival models with censored data*). The plot of this prior is presented below and it seems quite uninformative (although it puts more weights near zero).



(c)

Let the prior for λ be $\pi(\lambda) \propto \text{Gamma}(\alpha_0, \beta_0)$, the posterior distribution $\pi(\lambda \mid T, Y, \delta)$

$$\pi(\lambda \mid T, Y, \delta) \propto L(\lambda \mid T, Y, \delta)\pi(\lambda) = \lambda^{\alpha_0 + \sum_i (1-\delta_i)} \exp\left(-\lambda\left[\beta_0 + \sum_{i=1}^{n}\{(1-\delta_i)Y_i + \delta_i T_i\}\right]\right).,$$

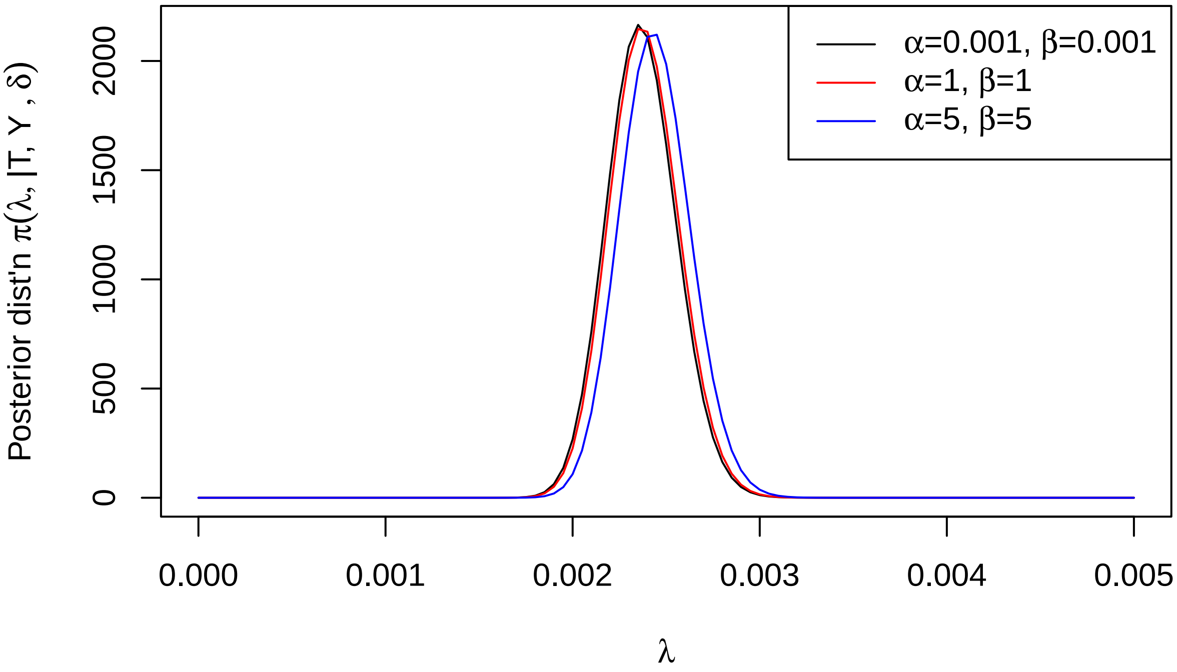
which is $\text{Gamma}(\alpha_0 + \sum_i(1-\delta_i), \beta_0 + \sum_{i=1}^{n}\{(1-\delta_i)Y_i + \delta_i T_i\})$

(d)

From (c), we know the posterior distribution of λ under the conjugate prior is

$$p(\lambda \mid T, Y, \delta) \propto L(\lambda \mid T, Y, \delta)\pi(\lambda) = \text{Gamma}(\alpha_0 + \sum_i(1-\delta_i), \beta_0 + \sum_i\{(1-\delta_i)Y_i + \delta_i T_i\}).$$

Plot the posterior distribution of λ under a few conjugate priors below and observe that the results are not senstive to the choice of priors even if α_0 and β_0 are large.



(e)

Choose an uninformative conjugate prior $\pi(\lambda_F), \pi(0\lambda_M) \propto \text{Gamma}(\alpha_0, \beta_0)$ with $\alpha_0 = \beta_0 = 0.001$ for women and men, separately. Let $S_i = 1$ if i-th patient is female and 0 otherwise. The joint posterior distribution for (λ_F, λ_M):

$$
\begin{aligned}
&p(\lambda_F, \lambda_M \mid Y, T, \delta) \propto f(Y, T, \delta \mid \lambda_F) f(Y, T, \delta \mid \lambda_M) \pi(\lambda_F) \pi(\lambda_M) \\
&\propto \lambda_F^{\sum_i (1-\delta_i) S_i} \exp\left\{ -\lambda_F \sum_i S_i y_i \right\} \lambda_M^{\sum_i (1-\delta_i)(1-S_i)} \exp\left\{ -\lambda_M \sum_i (1 - S_i) y_i \right\} \cdot \pi(\lambda_F) \pi(\lambda_M),
\end{aligned}
$$

which implies $\lambda_F \mid Y, T, \delta \sim \text{Gamma}(\alpha_0 + \sum_i (1-\delta_i) S_i, \beta_0 + \sum_i S_i y_i)$ and $\lambda_M \mid Y, T, \delta \sim \text{Gamma}(\alpha_0 + \sum_i (1-\delta_i)(1 - S_i), \beta_0 + \sum_i (1 - S_i) y_i)$. It can be shown the these two posteriors are independent and we can use Monte Carlo sampling to compare the posterior distributions.

```
## setup for the hyper-parameters
alpha0 <- beta0 <- 0.001
## size of the posterior sample
N <- 100000
## female
set.seed(123456)
lambda0 <- rgamma(N, shape = alpha0 + sum((1-df$delta) * df$sex),
                  rate = beta0 + sum(df$sex * df$Y))
## male
lambda1 <- rgamma(N, shape = alpha0 + sum((1-df$delta) * (1-df$sex)),
                  rate = beta0 + sum((1 - df$sex) * df$Y))
## construct the credible interval based on the posterior samples
```

```
### male - female
quantile(1/lambda1 - 1/lambda0, c(0.025, 0.975))
```

```
##       2.5%      97.5%
## -424.78223  -77.83035
```

Since the 95% credible interval of the difference of mean survival times for women and men does not contain zero, we claim that there is enough evidence to support that the survival distribution varies by sex.