A Bayesian Model for LTRC Data using Splines

1 References

The ideas of this work are generated from following references:

- Chun Pan and Bo Cai (2020), A Bayesian model for spatial partly interval censored data. *Communication in Statistics – Simulation and Computations*, 51, 7513–7525.
- Chun Pan, Bo Cai, and Xuemei Sui (2024), A Bayesian proportional hazard mixture cure rate model for interval censored data. *Lifetime Data Analysis*, 30, 327–344.
- Chun Pan, Bo Cai, and Lianming Wang (2020), A Bayesian approach for analyzing partly interval censored data under proportional hazard model. *Statistical Methods in Medical Research*, 29, 3192–3204.

2 Problem Formulation

Observed LTRC data can be represented as

$$\{l_i, t_i, \nu_i, \delta_i, \boldsymbol{x}_i\}_{i=1,2,\ldots,n},$$

where

what is ti, is it ti to infinite vaala, i.e. RC?

this is RC na?

and for DT, we will be needing two Indi?

$$\begin{aligned}
l_i &= \text{ left truncation time,}\\
t_i &= \text{ failure or censoring time,}\\
\delta_i &= \text{censoring indicator} = \begin{cases} 1 & \text{if } i\text{-th observation is not censored} \\ 0 & \text{otherwise,} \end{cases}\\
\nu_i &= \text{ truncation indicator} = \begin{cases} 1 & \text{if } i\text{-th observation is truncated} \\ 0 & \text{otherwise,} \end{cases}\\
\boldsymbol{x}_i &= \text{ vector of covariates.}
\end{aligned}$$

In the following model formulation, a categorical covariate with c levels is represented using $c-1$ dummy variables. Assume that there are total Q covariates in the model including numerical covariates and dummy variable. To incorporate covariates, we use proportional hazard model. Therefore, the cumulative hazard function (CHF) of lifetime is given by

$$\Lambda(t|\boldsymbol{x}) = \Lambda_0(t)\exp\{\boldsymbol{\beta}'\boldsymbol{x}\},$$

where $\Lambda_0(\cdot)$ is baseline CHF, which is independent of covariates, and $\beta = (\beta_1, \ldots, \beta_Q)$ is the vector of coefficients.

We further assume that the baseline CHF can be approximated by

$$H_0(t) = \sum_{l=1}^{K} \gamma_l I_l(t), \qquad (1)$$

H is Pi
h is lambda

where γ_l are non-negative parameters, $\{I_l(\cdot)\}_{l=1}^{K}$ is the set of basis I-splines, the number of basis I-splines, K, equals to the degree of each I-spline plus number of interior knots. This is tantamount to approximate baseline hazard function (HF), $\lambda_0(\cdot)$, using

$$h_0(t) = \sum_{l=1}^{K} \gamma_l M_l(t),$$

where $\{M_l(\cdot)\}_{l=1}^{K}$ is the set of basis M-splines. This is true as

$$I_l(t) = \int_0^t M_l(s)ds \quad \text{for all } l = 1, 2 \ldots, K.$$

See Ramsay (1988) for more details.

Now, using the approximation of baseline CHF as given in (1), the likelihood contribution of different observations can be written as follows:

- For $\nu_i = 1$ and $\delta_i = 1$, the likelihood contribution is

$$L_{1i}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = h_0(t_i) \exp\left\{\boldsymbol{\beta}'\boldsymbol{x}_i - (H_0(t_i) - H_0(l_i))\, e^{\boldsymbol{\beta}'\boldsymbol{x}_i}\right\}.$$

- For $\nu_i = 1$ and $\delta_i = 0$, the likelihood contribution is

$$L_{2i}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \exp\left\{-(H_0(t_i) - H_0(l_i))\, e^{\boldsymbol{\beta}'\boldsymbol{x}_i}\right\}.$$

- For $\nu_i = 0$ and $\delta_i = 1$, the likelihood contribution is

$$L_{3i}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = h_0(t_i) \exp\left\{\boldsymbol{\beta}'\boldsymbol{x}_i - H_0(t_i) e^{\boldsymbol{\beta}'\boldsymbol{x}_i}\right\}.$$

- For $\nu_i = 0$ and $\delta_i = 0$, the likelihood contribution is

$$L_{4i}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \exp\left\{-H_0(t_i) e^{\boldsymbol{\beta}'\boldsymbol{x}_i}\right\}.$$

Define the sets

$$A_1 = \{i : \nu_i = 1, \delta_i = 1\}, \qquad A_2 = \{i : \nu_i = 1, \delta_i = 0\},$$
$$A_3 = \{i : \nu_i = 0, \delta_i = 1\}, \qquad A_4 = \{i : \nu_i = 0, \delta_i = 0\}.$$

Then the likelihood function is given by

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{j=1}^{4} \left\{ \prod_{i \in A_j} L_{ji}(\boldsymbol{\gamma}, \boldsymbol{\beta}) \right\}.$$

The Metropolis-Hastings algorithm is a Markov Chain Monte Carlo (MCMC)

For any prior specification, sampling from the posterior may be performed using Metropolis-Hastings algorithm as a general sampler. However, the full conditionals are, in general, not in known form and quite complicated to sample from. As a result, the chains either fail to move and explore the parametric place properly or fail to converge. By augmenting appropriate data, we try to overcome this issue.

Note that presence of sum in

$$h_0(t) = \sum_{l=1}^{K} \gamma_l M_l(t)$$

posts challenges in direct sampling of coefficients, γ_l, of M-splines from the posterior. To overcome it, we introduce latent variable

$$\boldsymbol{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iK}) \sim Multinomial\left(K, \frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}\right)$$

for $i \in A_{13} = A_1 \cup A_3$. Then, the augmented data likelihood contributions for an observation with index $i \in A_{13}$ is given by

- For $i \in A_1$, the augmented data likelihood contribution is

$$\widetilde{L}_{1i}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = K \left\{ \prod_{l=1}^{K} (\gamma_l M_l(t_i))^{u_{il}} \right\} \exp\left\{ \boldsymbol{\beta}' \boldsymbol{x}_i - (H_0(t_i) - H_0(l_i))\, e^{\boldsymbol{\beta}' \boldsymbol{x}_i} \right\}.$$

- For $i \in A_3$, the augmented data likelihood contribution is

$$\widetilde{L}_{3i}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = K \left\{ \prod_{l=1}^{K} (\gamma_l M_l(t_i))^{u_{il}} \right\} \exp\left\{ \boldsymbol{\beta}' \boldsymbol{x}_i - H_0(t_i) e^{\boldsymbol{\beta}' \boldsymbol{x}_i} \right\}.$$

We do not augment any data for an observation with index $i \in A_{24} = A_2 \cup A_4$. Therefore, the likelihood contribution for such an observation does not change. The likelihood function based on the augmented data can be expressed as

$$\widetilde{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \left\{ \prod_{i \in A_1} \widetilde{L}_{1i}(\boldsymbol{\gamma}, \boldsymbol{\beta}) \right\} \left\{ \prod_{i \in A_2} L_{2i}(\boldsymbol{\gamma}, \boldsymbol{\beta}) \right\} \left\{ \prod_{i \in A_3} \widetilde{L}_{3i}(\boldsymbol{\gamma}, \boldsymbol{\beta}) \right\} \left\{ \prod_{i \in A_4} L_{4i}(\boldsymbol{\gamma}, \boldsymbol{\beta}) \right\}.$$

The prior specifications are as follows: $\gamma_l \sim Exp(\eta)$ for all l, $\eta \sim Gamma(a_\eta, b_\eta)$. Assume that the number of numerical covariates is $Q_1 \leq Q$. Also assume that, without loss of generality, the first Q_1 covariates, $x_1, x_2, \ldots, x_{Q_1}$, are numerical. For a regression coefficient corresponding to a numerical covariate, the prior is assumed to be $N(0, \sigma^2)$. Thus, for $r = 1, 2, \ldots, Q_1$, $\beta_r \sim N(0, \sigma^2)$. We treat the coefficient of a categorical covariate differently. The reason is that by specifying a Gamma prior $Gamma(a_\phi, b_\phi)$ for $\phi_r = \exp(\beta_r)$, $r > Q_1$, the resulting posterior happens to be Gamma which can be directly sampled from and renders better MCMC chains. Then we transform ϕ_r back to β_r. Therefore, the posterior

is proportional to

$$\begin{aligned}
&\pi(\boldsymbol{\gamma},\, \boldsymbol{\beta}_{num},\, \boldsymbol{\phi}, \boldsymbol{u}|\text{Data}) \\
&\quad= \prod_{l=1}^{K}\Bigg[\left\{\prod_{i\in A_{13}}\left(\gamma_l M_l(t_i)\right)^{u_{il}}\exp\left(\boldsymbol{\beta}'_{num}\boldsymbol{x}_{num,i}\right)\prod_{r=Q_1+1}^{Q}\phi_r^{x_{ir}}\right\} \\
&\qquad\times\exp\Bigg\{-\gamma_l\Bigg(\sum_{i\in A_{12}}\left(I_l(t_i)-I_l(l_i)\right)e^{\boldsymbol{\beta}'_{num}\boldsymbol{x}_{num,i}}\prod_{r=Q_1+1}^{Q}\phi_r^{x_{ir}} \\
&\qquad\qquad+\sum_{i\in A_{34}}I_l(t_i)e^{\boldsymbol{\beta}'_{num}\boldsymbol{x}_{num,i}}\prod_{r=Q_1+1}^{Q}\phi_r^{x_{ir}}\Bigg)\Bigg\}\Bigg] \\
&\qquad\times\left(\prod_{l=1}^{K}e^{-\eta\gamma_l}\right)\times\eta^{a_\eta-1}e^{-b_\eta\eta}\times\exp\left\{\frac{1}{2}\sum_{r=1}^{Q_1}\frac{\beta_r}{\sigma^2}\right\}\times\left(\prod_{r=Q_1+1}^{Q}\phi_r^{a_\phi-1}e^{-b_\phi\phi_r}\right),
\end{aligned}$$

where $\boldsymbol{\beta}_{num} = (\beta_1, \beta_2, \ldots, \beta_{Q_1})$, $\boldsymbol{x}_{num,i} = (x_{i1}, x_{i2}, \ldots, x_{iQ_1})$, $\boldsymbol{\phi} = (\phi_{Q_1+1}, \ldots, \phi_Q)$, $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ and $A_{ij} = A_i \cup A_j$. Now, the sampling scheme based on the full conditionals are given by

- For $l = 1, 2, \ldots, K$, sample γ_l from

$$Gamma\left(\sum_{i\in A_{13}}u_{il}+1,\, S_{1l}(\boldsymbol{\beta},\,\eta)\right),$$

  where $S_{1l}(\boldsymbol{\beta},\,\eta) = \sum_{i\in A_{12}}\left(I_l(t_i)-I_l(l_i)\right)e^{\boldsymbol{\beta}'\boldsymbol{x}_i}+\sum_{i\in A_{34}}I_l(t_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i}+\eta$.

- For $r = 1, 2, \ldots, Q_1$, sample β_r from a PDF proportional to

$$\exp\left\{K\beta_r\sum_{i\in A_{13}}x_{ir}-\sum_{l=1}^{K}\gamma_l\left(\sum_{i\in A_{12}}\left(I_l(t_i)-I_l(l_i)\right)e^{\boldsymbol{\beta}'\boldsymbol{x}_i}+\sum_{i\in A_{34}}I_l(t_i)e^{\boldsymbol{\beta}'\boldsymbol{x}_i}\right)-\frac{\beta_r^2}{2\sigma^2}\right\}.$$

- For $r = Q_1+1, \ldots, Q$, sample ϕ_r from

$$Gamma\left(K\sum_{i\in A_{13}}x_{ir}+a_\phi,\, S_{2r}(\boldsymbol{\beta}_{(-r)},\, b_\phi)\right),$$

  where

$$\begin{aligned}
S_{2r}(\boldsymbol{\beta}_{(-r)},\, b_\phi) = b_\phi+\sum_{l=1}^{K}\gamma_l\Bigg\{&\sum_{i\in A_{12}}x_{ir}\left(I_l(t_i)-I_l(l_i)\right)\exp\left\{\boldsymbol{\beta}'_{(-r)}\boldsymbol{x}_{i(-r)}\right\} \\
&+\sum_{i\in A_{34}}x_{ir}I_l(t_i)\exp\left\{\boldsymbol{\beta}'_{(-r)}\boldsymbol{x}_{i(-r)}\right\}\Bigg\},
\end{aligned}$$

  then convert ϕ_r to $\beta_r = \ln\phi_r$.

- For $i \in A_{13}$, sample $\boldsymbol{u}_i$ from

$$Multinomial\left(K,\, p_{i1},\, \ldots,\, p_{iK}\right),$$

  where $p_{il} = \dfrac{\gamma_l M_l(t_i)}{\sum_{j=1}^{K}\gamma_j M_j(t_i)}$.