Approximate Bayesian Inference for Semi-parametric Proportional Hazard Models

Ziang Zhang

11/09/2019

1 Survival Analysis Model:

1.1 Introduction to Survival Analysis:

Survival analysis refers to situations in which the response variable of interest is the time until the occurrence of a particular event. Examples include time to death of patients with a specific kind of disease, time to failure of a lightbulb, add one more example.. Models for analyzing time to event data quantify the association between treatments or risk factors and the time to an event. For example, we may quantify any relationship between the lifetimes of patients with the types of medicine they are using, to conclude whether a certain type of medicine is associated with a change in the overall survival times of patients.

Let T be a continuous non-negative random variable representing the time to some event, defined over the interval $[0, \infty)$. Let the probability density function of T be denoted as $f(t)$ and its cumulative distribution function be $F(t)$. The survivor function $S(t)$ of T can be defined as:

$$S(t) = P(T > t) = \int_t^\infty f(x)dx = 1 - F(t) \tag{1}$$

Notice that $S(t)$ is the probability of an observation to survive to time t, and therefore it is a monotone decreasing function with $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(x) = 0$.

The hazard function, denoted h(t), is defined as instantaneous rate of occurrence at a specific time t given that the event does not occur before t, which can be written as:

$$h(t) = \lim_{s\to 0} \frac{P(t \leq T \leq t + s|T \geq t)}{s} = \frac{f(t)}{S(t)} = -\frac{\partial}{\partial t}\log[S(t)] \tag{2}$$

The corresponding cumulative hazard function $H(t)$ will be defined by:

$$H(t) = \int_0^t h(u)du = -\log[S(t)] \tag{3}$$

Cumulative hazard function is an alternative measure of risk of occurrence, and will be more convenient to use in some cases.

Event times are often only partially observed. Right censoring occurs when it is only known that an event occurred after some fixed timepoint, such as patients in a medical study who are lost to follow up. Left truncation occurs when event times are only included in the dataset conditional on having surpassed some threshold, such as patients joining a medical study only after having a disease for several weeks or months. Interval censoring refers to the presence of both right censoring and left truncation in the same observed event times. Partially observed observations present challenges in the analysis of survival data, and are too common to be ignored. We expand on these challenges in Section 1.2.

1.2 Types of Censoring and Truncation:

In survival analysis, we are mainly dealing with the problem of right-censoring, interval-censoring and left truncation. Right-censoring is when an individual's lifetime T_i is not exactly known because the individual is still alive when the study terminates at C_i, so we are only sure about that $T_i > C_i$ but not sure what exactly T_i is. Interval-censoring on the other hand, rises when the survival time is only known to be in an interval (L_i, U_i), and the left truncation problem happens when some survival times are not recorded unless they are bigger than a specified start time t^{tr}, so all the data with survival times less than t^{tr} are missed.

In general, we use the term "censoring" to refer to the scenarios where some lifetimes are only known to exceed their cutting times C_i, but we do not know how long do they last exactly. On the other hand, the term "truncation" mostly refer to a data collected problem where only lifetimes greater than the start time t^{tr} are collected and observed. So these terms should be used in different situations depending on what kind of survival data are we dealing with.

There are two types of right-censoring that appear most frequently in the context of survival analysis, which are Type-I and Type-II right-censoring.

Type-I right-censoring occurs when each individual's censoring time C_i is fixed and known beforehand. That means when we collect a bunch of survival times, we know whether each survival time is right-censored and when is it censored exactly. For example, in a medical study, we may give each individual 30 minutes of observation time after each of them took a specific type of medicine, and the C_i in this case is fixed to be 30 minutes for all i.

For this type of censoring mechanism, we will be able to write our original data-set $\{T_i, C_i : i = 1, ..., n\}$ as $\{t_i, \delta_i : i = 1, ..., n\}$ where:

$$t_i = \min\{T_i, C_i\}, \qquad \delta_i = I(T_i \leq C_i) \tag{4}$$

Many statistical procedures dealing with right censoring are assuming this is the underlying censoring mechanism (Kwan, 1997), so we will focus on this type of censoring for the rest of the paper.

Type-II right-censoring occurs when we only observed the r smallest survival times in our sample. So the survival times that we can observed will be like $t_{(1)} < t_{(2)} < ... < t_{(r)}$, and the other survival times will be censored so we don't know the exact numbers. In this scenario, we have a censoring time $t_{(r)}$ that is itself random. For example, in an engineering study, we can measure the length of time for each component to become defective, and we will end the study after collecting ten defectives in our sample. In this case, the censoring time will be t_{10}, and all the times after this will not be observable.

Lastly, independent random censoring happens when both the ith survival time T_i and the ith censoring time C_i are random variable that are independent.For example, if we want to measure the lifetimes of patients in a hospital, then it is possible that some patients are going to switch to anther hospital before the study ends. In this case, if we assume that switching hospital does not have relationship with the severity of the deterioration of diseases, then the censoring times will be independent with the lifetimes of patients.

1.3 Cox Proportional Hazard Model:

In most survival analysis study, we are interested in incorporating some covariates $\tilde{X} = \{X_1, X_2, ..., X_p\}$ into the distribution of survival time T, and studying their associations with the survival time T.Define u_{qi} be the covariates that are modelled semi-parametrically using unknown smooth functions γ_q. Therefore, to specify the dependence of T on $\tilde{X}$,the proportional hazard model introduced by Cox(1972) is the a popular choice.

Let $h(t|\tilde{x})$ denote the hazard function of T at time t for a subject with covariates $\tilde{x} = (x_1, x_2, ..., x_p)$. The Cox Proportional Hazard Model can be specified as follows:

$$h(t|\tilde{x}) = h_0(t)\text{exp}(\beta_1 x_1 + ... + \beta_p x_p + \sum_{q=1}^{R} \gamma_q(u_{qi})) \tag{5}$$

where $h_0(t)$ is an arbitrary baseline hazard function that does only depend on time, and β_i's are the unknown parameters that we are interested in estimating. The reason that it is called a "proportional" hazard model is because for any two subjects, the ratio of their hazard function will be constant over time. This is a model assumption, and it should be checked before adopting the Cox proportional hazard model.

Notice that the baseline hazard function is left to be arbitrary, which implies that the Cox Proportional Hazard Model will be a semi-parametric model. To specify the baseline hazard functions, we are going to use the piece-wise constant baseline hazard model, and I will introduce it in details in the next section.

1.4 Proportional Hazard Model with Piece-wise Constant Baseline Hazard:

Firstly, we break the time axis into K intervals with endpoints $0 = s_0 < s_1 < ... < s_K < \max\{t_i : i = 1, ..., n\}$, and assumes that the baseline hazard function is constant in each interval, i.e: $h_0(t) = b_k$ for $t \in (s_{k-1}, s_k),\ k = 1, 2, ..., K$.

Let $\eta_{ik} = \log(b_k) + \tilde{x_i}^T\tilde{\beta} + \sum_{q=1}^{R} \gamma_q(u_{qi})$, the model that we will be focusing on will be the semi-parametric proportional hazard model, specified at below:

$$\begin{aligned} h(t_i) &= h_0(t_i)\text{exp}[\tilde{x_i}^T\tilde{\beta} + \sum_{q=1}^{R} \gamma_q(u_{qi})] \\ &= \text{exp}[\log(b_k) + \tilde{x_i}^T\tilde{\beta} + \sum_{q=1}^{R} \gamma_q(u_{qi})] \qquad t_i \in (s_{k-1}, s_k] \\ &= \text{exp}(\eta_{ik}) \end{aligned} \tag{6}$$

Using this information, we can derive the likelihood for that single observation to be:

$$\begin{aligned} \pi\big[(t_i, \delta_i)|\eta_{i1}...\eta_{ik}\big] &= f(t_i)^{\delta_i} S(t_i)^{(1-\delta_i)} \\ &= h(t_i)^{\delta_i} S(t_i) \\ &= \text{exp}(\delta_i \eta_{ik})\Big\{\text{exp}\big[-\int_0^{t_i} h(u)du\big]\Big\} \\ &= \text{exp}(\delta_i \eta_{ik})\Big\{\text{exp}\big[-\sum_{j=1}^{k-1}(s_j - s_{j-1})\text{exp}(\eta_{ij}) - (t_i - s_{k-1})\text{exp}(\eta_{ik})\big]\Big\} \end{aligned} \tag{7}$$

Therefore, the full-likelihood of the data-set will be:

$$\begin{aligned} L &= \prod_{i=1}^{n} \exp(\delta_i \eta_{ik_{(i)}}) \exp\Bigg\{ - \sum_{j=1}^{k_{(i)}-1} (s_j - s_{j-1}) \exp(\eta_{ij}) - (t_i - s_{k_{(i)}-1}) \exp(\eta_{ik_{(i)}}) \Bigg\} \\ &= \prod_{i=1}^{n} \exp\big\{ \delta_i \eta_{ik_{(i)}} - \sum_{j=1}^{k_{(i)}-1} (s_j - s_{j-1}) \exp(\eta_{ij}) - (t_i - s_{k_{(i)}-1}) \exp(\eta_{ik_{(i)}}) \big\} \end{aligned} \tag{8}$$

I emphasize the subscript for $k_{(i)}$ because each survival time will correspond to a different value of $k_{(i)}$, depending on which interval the survival time lies in. More specifically, $k_{(i)}$ denotes which one of the k sub-intervals the observation is in.

By taking the logarithm, the log-likelihood function for the i^{th} observation $t_i \in (s_{k-1}, s_k]$ can be written as :

$$\begin{aligned} l &= \log[f(t_i)^{\delta_i} S(t_i)^{(1-\delta_i)}] \\ &= \log[h(t_i)^{\delta_i} S(t_i)] \\ &= \delta_i \eta_{ik} - (t_i - s_{k-1}) \exp(\eta_{ik}) - \sum_{j=1}^{k-1} [(s_j - s_{j-1}) \exp(\eta_{ij})] \end{aligned} \tag{9}$$

Similarly, the full log-likelihood can be derived as:

$$l = \sum_{i=1}^{n} \Bigg\{ \delta_i \eta_{ik_{(i)}} - (t_i - s_{k_{(i)}-1}) \exp(\eta_{ik_{(i)}}) - \sum_{j=1}^{k_{(i)}-1} [(s_j - s_{j-1}) \exp(\eta_{ij})] \Bigg\} \tag{10}$$

It can see from the above expression that by considering a piece-wise constant baseline hazard, we make the corresponding log-likelihood much easier to work with, since the integral $\int_0^{t_i} h(u) du$ can be replaced by a sum.

2 INLA's Inference Methodology:

In this section, I will briefly introduce what INLA actually does to do the approximate Bayesian inference for Cox proportional hazard model, and show what are the current limitations of INLA for this type of problem.

2.1 Data Augmentation Using Poisson Likelihood:

For Cox proportional hazard model, the INLA algorithm cannot directly be applied, because if we look at the log-likelihood of a single survival time $\{t_i, \delta_i\}$, we can find that it depends on more than one η. To use INLA, we required a conditional independent latent field together with a sparse Hessian matrix for the log-likelihood. That means we need to make sure that for a single data point, the log-likelihood should be free of terms from latent field once we condition on one of the term from the latent field.

To solve this puzzle, INLA utilizes a data "augmentation" trick to transform the log-likelihood of a single data point into the form that INLA likes. Notice that if we are looking at a random variable X_i that follows a Poisson distribution with mean $(t_i - s_{k-1}) \exp(\eta_{ik})$, then the log-likelihood corresponding to a single data point $\{X_i = 0\}$ will be:

$$\begin{aligned} l &= \log\Bigg\{ P\big[X_i = 0 | \lambda = (t_i - s_{k-1}) \exp(\eta_{ik})\big] \Bigg\} \\ &= 0 \times \ln[(t_i - s_{k-1}) \exp(\eta_{ik})] - (t_i - s_{k-1}) \exp(\eta_{ik}) - \ln(0!) \\ &= -(t_i - s_{k-1}) \exp(\eta_{ik}) \end{aligned} \tag{11}$$

Similarly, when $X_i = 1$, the log-likelihood of this single data point is:

$$\begin{aligned} l &= \log\bigg(P(X_i = 1|\lambda = (t_i - s_{k-1})\exp(\eta_{ik}))\bigg) \\ &= 1 \times \ln((t_i - s_{k-1})\exp(\eta_{ik})) - (t_i - s_{k-1})\exp(\eta_{ik}) - \ln(1!) \\ &= \ln(t_i - s_{k-1}) + \eta_{ik} - (t_i - s_{k-1})\exp(\eta_{ik}) \\ &\propto \eta_{ik} - (t_i - s_{k-1})\exp(\eta_{ik}) \end{aligned} \tag{12}$$

Here we can basically ignore the term $\ln(t_i - s_{k-1})$ as it does not depend on any term from the latent field. So when we later take derivative, this term will just disappear which means it won't affect our C matrix.

We showed that the first two terms of the log-likelihood of a single data point $\{t_i, \delta_i\}$ can be viewed as the log-likelihood of a single data point $X_i \sim \text{Poisson}\big[\lambda = (t_i - s_{k-1})\exp(\eta_{ik})\big]$ being 0 when $\delta_i = 0$ and being 1 when $\delta_i = 1$.

Next step will be to figure out a similar way to deal with the last term in equation (3). Notice that for a Poisson random variable Y_j with mean $(s_j - s_{j-1})\exp(\eta_{ij})$, the log-likelihood for observing it being 0 will be:

$$\begin{aligned} l &= \log\bigg\{P\big[Y_j = 0|\lambda = (s_j - s_{j-1})\exp(\eta_{ij})\big]\bigg\} \\ &= -(s_j - s_{j-1})\exp(\eta_{ij}) \end{aligned} \tag{13}$$

Similarly, if we gather a sample of $\{Y_{i_1} = 0, Y_{i_2} = 0, ..., Y_{i_k} = 0\}$ where each $Y_{i_j} \sim \text{Poisson}\big[\lambda = (s_j - s_{j-1})\exp(\eta_{ij})\big]$ is independent of others, then the log-likelihood of this sample will simply be the sum of log-likelihood of each term due to independence, which sums to be $\sum_{j=1}^{k-1}(s_j - s_{j-1})\exp(\eta_{ij})$,that is exactly what we want.

Putting these two pieces information together, which means if we have a sample being $\{X_i = \delta_i, Y_{i_1} = 0, Y_{i_2} = 0, ..., Y_{i_k} = 0\}$, and all the terms in this sample being mutually independent, then the log-likelihood of this sample will just be the log-likelihood of the single data point $\{t_i, \delta_i\}$. Doing this for all the data points $\{t_i, \delta_i | i = 1, ..., n\}$. We retrieve the original log-likelihood from the log-likelihood of a sample of $\sum_{i=1}^{n} k_{(i)}$ number of independent, but non-identical Poisson random variables. In other words, we augment our original data-set $\{t_i, \delta_i | i = 1, ..., n\}$ into a huge data-set$\{x_i, y_{i_1}, y_{i_2}, ..., y_{i_{k_{(i)}}} | i = 1, 2, ...n\}$, where all the terms in this new data-set are mutually independent. This is the cure for our problem since the log-likelihood of each term from this new "augmented" data-set, will only depend on the latent field through one η.

2.2 Derivation of the Negated Hessian Matrix:

Here I will present how the Bayesian approximation can be carried out using an INLA-type of algorithm. Firstly, to make the covariance matrix of the joint Gaussian latent field non-singular, and to simplify the Hessian matrix that we are going to derive later, we will assume that for each η_{ij}, a normal random noise ϵ_{ij} is added. We assume that $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ being mutually independent across different i and j. In other words, we will write $\eta_{ik} = \log(b_k) + \tilde{x_i}^T\tilde{\beta} + \sum_{q=1}^{R}\gamma_q(u_{qi}) + \epsilon_{ik}$. Let $U_q = \{U_{ql}, l = 1, ..., M_q\}$ be an ordered vector of all unique values of u_{ql}. Define $\Gamma_q = \{\gamma_q(U_{ql}), l = 1, ..., M_q\}$, and $\Gamma = \{\Gamma_q, q = 1, ..., R\}$.

Then, the latent field can be denoted as:

$$\tilde{W} = \big[\eta_{11}, \eta_{12}, ..., \eta_{1k}, \eta_{2k}, ..., \eta_{nk}, \Gamma_1, ..., \Gamma_R, \beta_1, ..., \beta_p, \log(b_1), ..., \log(b_k)\big]^T \tag{14}$$

which has dimension $nk + \sum_{q=1}^{R} M_q + p + k$.

Besides assume that $\tilde{W}$ is a Gaussian Markov Random Field, we also assume that $\log(b_{k+1}) - \log(b_k)$ follows $N(0, \tau^{-1})$, a random walk with order 1. So we will just use $\tilde{\theta}$ to denote the hyper-parameter vector that determines the precision matrix of our latent field.

Now, let's derive the negated Hessian matrix of the log-likelihood with respect to the latent field. To do that, let's first consider the log-likelihood consider only one survival time $\{t_i, \delta_i\}$ where $t_i \in (s_{k_{(i)}-1}, s_{k_{(i)}}]$. In this case, the log-likelihood for this data point will be:

$$l = \delta_i \eta_{ik_{(i)}} - (t_i - s_{k_{(i)}-1})\exp(\eta_{ik_{(i)}}) - \sum_{j=1}^{k_{(i)}-1} [(s_j - s_{j-1})\exp(\eta_{ij})] \tag{15}$$

The derivative with respect to $\eta_{ik_{(i)}}$ will be

$$\frac{\partial l}{\partial \eta_{ik_{(i)}}} = \delta_i - (t_i - s_{k_{(i)}-1})\exp(\eta_{ik_{(i)}}) \tag{16}$$

That means the negated second derivative will be:

$$-\frac{\partial^2 l}{\partial {\eta_{ik_{(i)}}}^2} = (t_i - s_{k_{(i)}-1})\exp(\eta_{ik_{(i)}}) \tag{17}$$

For first and negated second derivatives with η_{ij} where $j < k_{(i)}$, we have:

$$\begin{aligned} \frac{\partial l}{\partial \eta_{ij}} &= -(s_j - s_{j-1})\exp(\eta_{ij}) \\ -\frac{\partial^2 l}{\partial {\eta_{ij}}^2} &= (s_j - s_{j-1})\exp(\eta_{ij}) \end{aligned} \tag{18}$$

Apparently, for η_{ij} where $j > k_{(i)}$, we have the second derivatives of log-likelihood being 0's. Combine them together, we know that the negated Hessian matrix for the log-likelihood of $\{t_i, \delta_i\}$, C_i will be:

$$\begin{bmatrix} (s_1 - s_0)\exp(\eta_{i1}) & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & (s_2 - s_1)\exp(\eta_{i2}) & 0 & \ddots & & & \\ 0 & 0 & \ddots & & \ddots & & \\ \vdots & \cdots & \cdots & (s_{k_{(i)}-1} - s_{k_{(i)}-2})\exp(\eta_{i(k_{(i)}-1)}) & \ddots & \vdots & \\ \vdots & & \ddots & 0 & (t_i - s_{k_{(i)}-1})\exp(\eta_{ik_{(i)}}) & \cdots & \vdots \\ \vdots & & & \ddots & & \ddots & \vdots \\ \vdots & & & & \cdots & \cdots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix} \tag{19}$$

This is a very sparse matrix with only diagonal terms. It is a $k \times k$ matrix, but only the diagonal terms at the first $k_{(i)}$ rows are non-zero.

Repeating this procedure for the rest data points, using the property of independence, we can get the negated Hessian matrix C for the full log-likelihood will be:

$$C = \begin{pmatrix} C_1 & 0 & 0 & \cdots & & & \\ 0 & C_2 & 0 & \cdots & & & \\ & \cdots & \ddots & & & & \\ & & & & C_n & \cdots & \vdots \\ & & & \ddots & & & \vdots \\ & & & & & & 0 \end{pmatrix} \tag{20}$$

Here we build a block diagonal matrix C using each block C_i obtained from above procedures. The negated Hessian matrix is sparse and diagonal, which is exactly what it has to be to be fitted using INLA. This full matrix has dimension $(nk + \sum_{q=1}^{R} M_q + p + k) \times (nk + \sum_{q=1}^{R} M_q + p + k)$, and each block matrix H_i has different patterns of zeros but same dimension.

However, in a lot of cases, the data augmentation trick that INLA does will not work, and the negated Hessian matrix is not diagonal. For example, if we want to only use the partial likelihood instead of the full likelihood, because the baseline function is not of interest, then the resulting negated Hessian matrix will be sparse but not diagonal. INLA is currently not applicable to this type of problem, and here we are suggesting a novel method that will be general enough to be applicable even when the data augmentation trick cannot work.

3 Proposed Methodology for Approximation:

In the paper "Approximate Bayesian Inference for Case-Crossover Models", Stringer et al.(2019) suggested a new type of algorithm to do the approximation while allowing the log-likelihood of each observation to be dependent on more than one element from the latent field, which means the ad-hoc method using "data augmentation" is no longer needed. Here we will demonstrate how that algorithm can be used to estimate the parameters in Cox Proportional Hazard Model, and when this new algorithm will be preferred than INLA's algorithm.

3.1 Approximation using Partial Likelihood with Right censoring only:

For simplicity, let's assume that our main interests are the components in the linear predictor η, but not the baseline hazard $h_0(t)$, and the only type of censoring present is right-censoring. Let the C-matrix denote the negated Hessian matrix of the log-likelihood with respects to the latent field $\tilde{W}$.

Assume that $\{t_i : i = 1, ..., k\}$ is a set of k distinct lifetimes that we actually *observed*, such that $t_{(1)} < t_{(2)} < ... < t_{(k)}$, and the result n-k lifetimes are the censored lifetimes that are not observed. Let $R_i = R(t_{(i)})$ be the set of individuals who are alive and uncensored prior to time $t_{(i)}$ (including the i-th individual who dies at $t_{(i)}$).

Define the hazard function for the i-th individual to be $h_0(t)\exp(\eta_i)$, and let $\Delta_{i,j} = \eta_i - \eta_j$, then the partial likelihood for Cox Proportional Hazard Model can be written as:

$$\begin{aligned} L^{partial} &= \prod_{i=1}^{k} \left\{ \frac{\exp[\eta_{(i)}]}{\sum_{l \in R_i} \exp[\eta_{(l)}]} \right\} \\ &= \prod_{i=1}^{k} \left\{ \frac{1}{\sum_{l \in R_i} \exp[\eta_{(l)} - \eta_{(i)}]} \right\} \\ &= \prod_{i=1}^{k} \left\{ \frac{1}{\sum_{l \in R_i} \exp[\Delta_{l,i}]} \right\} \\ &= \prod_{i=1}^{k} \left\{ \frac{1}{1 + \sum_{l \in R_i, l \neq i} \exp[\Delta_{l,i}]} \right\} \end{aligned} \tag{21}$$

Which can be written as:

$$= \prod_{i=1}^{k} \left\{ \frac{1}{1 + \sum_{l \in R_i, l \neq i} \exp[(\Delta_{1,i} - \Delta_{1,l})]} \right\} \tag{22}$$

Because:

$$\Delta_{i,j} = \eta_i - \eta_j = \eta_1 - \eta_j - (\eta_1 - \eta_i) = \Delta_{1,j} - \Delta_{1,i} \tag{23}$$

Define $\Delta := (\Delta_{1,2}, \Delta_{1,3}, ..., \Delta_{1,n})^T$, which are all the Δ_{ij}'s that will contribute to the likelihood.

Notice that this partial likelihood does not include any information on the baseline hazard function $h_0(t)$, meaning that all of the information are used to estimate the regression parameters in the model, which should result in a more precise estimation for them. Here it is shown that the partial likelihood only depend on those "differenced linear predictors" $\{\Delta_{1,i} - \Delta_{1,j}\}$, so our latent field in this case will be $\{\Delta, \beta, \Gamma\}$. More importantly, because we are not estimating those baseline hazards, the algorithm's convergence rate is expected to be faster. INLA does not allow this type of approximation because using partial likelihood to ignore the baseline hazard invalidates the "Poisson data-augmentation" trick that INLA does to make the C-matrix diagonal. While non-diagonal C matrix is not feasible in INLA's algorithm, it will be feasible in the new proposed algorithm.

3.1.1 Derivation of Hessian matrix and Precision matrix:

For i-th observation, the partial log-likelihood will be:

$$\begin{aligned} l &= -\ln(1 + \sum_{j \in R_i, j \neq i} \exp[\Delta_{j,i}]) \\ &= -\ln(1 + \sum_{j \in R_i, j \neq i} \exp[(\Delta_{1,i} - \Delta_{1,j})]) \end{aligned} \tag{24}$$

Therefore, taking derivative with respect to $\Delta_{1,i}$, we can get:

$$\frac{\partial l}{\partial \Delta_{1,i}} = -\frac{\sum_{j \in R_i, j \neq i} \exp[(\Delta_{1,i} - \Delta_{1,j})]}{1 + \sum_{j \in R_i, j \neq i} \exp[(\Delta_{1,i} - \Delta_{1,j})]} \tag{25}$$

Similarly, we can see that:

$$\frac{\partial^2 l}{\partial \Delta_{1,i}^2} = -\frac{\sum_{j \in R_i, j \neq i} \exp[(\Delta_{1,i} - \Delta_{1,j})]}{\left\{1 + \sum_{j \in R_i, j \neq i} \exp[(\Delta_{1,i} - \Delta_{1,j})]\right\}^2} \tag{26}$$

Suppose that $w \neq i$, if $w \notin R_i$ then we have:

$$\frac{\partial l}{\partial \Delta_{1,w}} = 0 \tag{27}$$

If $w \in R_i$ then we have:

$$\frac{\partial l}{\partial \Delta_{1,w}} = \frac{\exp[(\Delta_{1,i} - \Delta_{1,w})]}{1 + \sum_{j \in R_i, j \neq i} \exp[(\Delta_{1,i} - \Delta_{1,j})]} \tag{28}$$

Similarly, we get the following result for $w \in R_i$:

$$\frac{\partial^2 l}{\partial \Delta_{1,i} \partial \Delta_{1,w}} = \frac{\exp(\Delta_{1,i} - \Delta_{1,w})}{\left\{1 + \sum_{j \in R_i, j \neq i} \exp[(\Delta_{1,i} - \Delta_{1,j})]\right\}^2} \tag{29}$$

$$\frac{\partial^2 l}{\partial \Delta_{1,w}^2} = -\frac{\left[\sum_{j \in R_i, j \neq i,w} \exp(2\Delta_{1,i} - \Delta_{1,w} - \Delta_{1,j})\right] + \exp(\Delta_{1,i} - \Delta_{1,w})}{\left\{1 + \sum_{j \in R_i, j \neq i} \exp[(\Delta_{1,i} - \Delta_{1,j})]\right\}^2} \tag{30}$$

If $M \in R_i$ but $M \neq i, M \neq w$, then:

$$\frac{\partial^2 l}{\partial \Delta_{1,w} \partial \Delta_{1,M}} = \frac{\exp(2\Delta_{1,i} - \Delta_{1,w} - \Delta_{1,M})}{\left\{1 + \sum_{j \in R_i, j \neq i} \exp[(\Delta_{1,i} - \Delta_{1,j})]\right\}^2} \tag{31}$$

From the information above, we can derive the negated Hessian matrix of log-likelihood of the i-th observation, with respect to Δ being:

$$C_i = - \begin{pmatrix} \frac{\partial^2 l}{\partial \Delta_{1,2}^2} & \frac{\partial^2 l}{\partial \Delta_{1,2} \Delta_{1,3}} & \cdots & \frac{\partial^2 l}{\partial \Delta_{1,2} \Delta_{1,n}} \\ \frac{\partial^2 l}{\partial \Delta_{1,3} \Delta_{1,2}} & \frac{\partial^2 l}{\partial \Delta_{1,3}^2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \Delta_{1,n} \Delta_{1,2}} & \cdots & \cdots & \frac{\partial^2 l}{\partial \Delta_{1,n}^2} \end{pmatrix} \tag{32}$$

For each $i \in \{1, ..., n\}$, the corresponding C_i matrix will be $(n-1) \times (n-1)$. Since the log-likelihood of the sample will be the sum of log-likelihood of each data, the negated Hessian matrix C^Δ of the full log-likelihood with respect to Δ will be the sum of each C_i matrix, i.e.: $C^\Delta = C_1 + ... + C_k$.

Therefore, the negated Hessian matrix C of the full log-likelihood with respect to the whole latent field will be:

$$C = \begin{pmatrix} C^\Delta & & & & & \\ & 0 & & & & \\ & & 0 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \tag{33}$$

To get the vector $\Delta = (\Delta_{1,2}, \Delta_{1,3}, ..., \Delta_{1,n})^T$, we need to apply a linear transformation on the vector $\eta = (\eta_1, ..., \eta_n)^T$, in other words:

$$\Delta = \begin{pmatrix} \eta_1 - \eta_2 \\ \vdots \\ \eta_1 - \eta_n \end{pmatrix} = D\eta \tag{34}$$

Where D is the differencing matrix defined as:

$$D = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ & & \ddots & & \\ 1 & & & 0 & -1 \end{pmatrix} \tag{35}$$

In this case, D will be a matrix with dimension $(n-1) \times n$ and rank $n-1$. Therefore, DD^T will be an invertible matrix.

Using the differenced matrix D defined above, we can proceed to get the precision matrix Q. The rest procedures will be identical to the procedures Stringer did in the paper “Approximate Bayesian Inference for Case-Crossover Models” (Stringer et al,2019).

3.1.2 Adjustments of Likelihood for tied observations

Throughout our discussion on how to use partial likelihood for doing approximate Bayesian Inference, we made an important assumption that the uncensored lifetimes are all unique, in other words $t_{(i)} \neq t_{(j)}$ for all $i \neq j$. However, in real life applications, there will be some cases where this assumption is not met. For example, we might use a "discretized" measure of lifetime when we are recording both 12.11 hours and 12.12 hours as 12.1 hours in the study of sustainability of a certain type of battery. That means, we should be able to deal with "ties" in our observed lifetimes. Here we will focus on Efron and Breslow's methods of approximation.

From now, let's assume that for an observed lifetime t_j, there are $d_{(j)}$ number of other observed lifetimes are "tied" with it. Let $D_{(j)}$ be the set of all the individuals with lifetimes tied at t_j, that means there will be $d_{(j)}$ number of elements in $D_{(j)}$.

Breslow's method of approximation will be the easiest and fastest method to use, when accuracy of estimations is not extremely required. The main idea of Breslow approximation is to include all the events occured at t_j in the risk set $R(t_j)$. Therefore, we can write the partial likelihood at t_j as:

$$L_j^B = \frac{\prod_{i \in D_{(j)}} \exp(\eta_i)}{\left[\sum_{i \in R_j} \exp(\eta_i)\right]^{d_{(j)}}} \tag{36}$$

The use of Breslow method is really convenient, because its partial likelihood has the same form of our ordinary partial likelihood when all the observations are distinct. When the number of ties in the data-set is not a lot, Breslow's method does give a fairly accurate result, and its computation time is exactly the same as no ties in the set. However, despite its appealing efficiency, if too many ties occuring at the same time, then its accuracy will be influenced comparatively heavily. In that case, we would like to consider an alternative method, which is Efron method of approximation.

Efron initially suggested this method as an approximation for the case when these ties are "exactly continuous". However, Efron's method in general gives very accurate estimates, and is computationally fast enough, even when the number of ties is very large. The main idea of this approach is to give a partial approximation to the contribution to $R(t_j)$ of each tied occurence, and it uses a partial likelihood in the form below:

$$L_j^{Ef} = \frac{\prod_{i \in D_{(j)}} \exp(\eta_i)}{\prod_{h=1}^{d_{(j)}} \left\{ \sum_{i \in R_j} \exp(\eta_i) - \frac{h-1}{d_j} \sum_{k \in D_j} \exp(\eta_k) \right\}} \tag{37}$$

Although the partial likelihood of Efron's approximation is not as easy as the partial likelihood of Breslow's method, it is still not bad for our computational efficiency. In the case of a lot of ties occuring at a specific time, it gives a more accurate result than Breslow's method.

There are other exact methods to deal with ties such as D.Cox's "exact discrete" method and Kalbfleisch and Prentice's "exact continuous" method, but they are computationally too heavy to be used for our purpose. So we choose to not use them for now.

3.1.2.1 Adjustment to C-matrix using Breslow's Method:

If Breslow's method is used to adjust for the effect of tied observations at time $t_{(j)}$, then the form of C-matrix will be very similar to the C-matrix under regular setting derived in section 3.1.1. However, there will be some small changes we need to add to the C-matrix.

Firstly, the partial likelihood at time $t_{(j)}$ will be:

$$L_j^B = \prod_{i \in D_{(j)}} \left[\frac{1}{\sum_{k \in R_j} \exp(\Delta_{k,i})} \right] \tag{38}$$

Therefore, its logarithm will be:

$$\begin{aligned} l_j^B &= \sum_{i\in D_{(j)}} \log\Bigg[\frac{1}{1+\sum_{k\in R_j,k\neq i}\exp(\Delta_{k,i})}\Bigg] \\ &= -\sum_{i\in D_{(j)}} \log\Bigg[1+\sum_{k\in R_j,k\neq i}\exp(\Delta_{k,i})\Bigg] \\ &= -\sum_{i\in D_{(j)}} \log\Bigg[1+\sum_{k\in R_j,k\neq i}\exp(\Delta_{1,i}-\Delta_{1,k})\Bigg] \end{aligned} \tag{39}$$

That means, for the subject i such that $i \in D_{(j)}$, we have its log-likelihood being:

$$l_{j(i)}^B = -\text{log}(1+\sum_{k\in R_j,k\neq i}\exp(\Delta_{1,i}-\Delta_{1,k})) \tag{40}$$

The negated Hessian matrix $C_{j_{(i)}}$ of this log-likelihood with respect to Δ can be computed using the same procedures in section 3.1.1.

Consequently, the negated Hessian matrix of the log-likelihood at time $t_{(j)}$ will be:

$$C_j = \sum_{i\in D_{(j)}} C_{j_{(i)}} \tag{41}$$

Similarly, we are able to get negated Hessian matrix C_i at each observed time point $t_{(i)}$, and combing them using the method from section 3.1.1 gives our final C matrix of the sample.

3.1.2.2 Adjustment to C matrix using Efron's Method:

Define the vector $\tilde{h} = (1,2,...,d_{(j)})^T$, where h_i is its ith component, then the partial likelihood at time $t_{(j)}$ can be written as:

$$L_j^{Ef} = \prod_{i\in D_{(j)}} \frac{1}{\sum_{k\in R_j}\exp(\Delta_{k,i}) - \frac{h_i-1}{d_{(j)}}\sum_{m\in D_{(j)}}\exp(\Delta_{m,i})} \tag{42}$$

where for the ith subject such that $i \in D_{(j)}$, its partial likelihood will be:

$$L_{j_{(i)}}^{Ef} = \frac{1}{\sum_{k\in R_j}\exp(\Delta_{k,i}) - \frac{h_i-1}{d_{(j)}}\sum_{m\in D_{(j)}}\exp(\Delta_{m,i})} \tag{43}$$

and its logarithm will be:

$$l_{j_{(i)}}^{Ef} = -\text{log}\Bigg\{1+\sum_{k\in R_j,k\neq i}\exp(\Delta_{1,i}-\Delta_{1,k}) - \frac{h_i-1}{d_{(j)}}\Big[\sum_{m\in D_{(j)},m\neq i}\exp(\Delta_{1,i}-\Delta_{1,m})\Big]\Bigg\} \tag{44}$$

From this log-likelihood, we can derive that:

$$\frac{\partial l_{j_{(i)}}^{Ef}}{\partial \Delta_{1,i}} = -\frac{\sum_{k\in R_j,k\neq i}\exp(\Delta_{1,i}-\Delta_{1,k}) - \frac{h_i-1}{d_{(j)}}\big[\sum_{m\in D_{(j)},m\neq i}\exp(\Delta_{1,i}-\Delta_{1,m})\big]}{\Big\{1+\sum_{k\in R_j,k\neq i}\exp(\Delta_{1,i}-\Delta_{1,k}) - \frac{h_i-1}{d_{(j)}}\big[1+\sum_{m\in D_{(j)},m\neq i}\exp(\Delta_{1,i}-\Delta_{1,m})\big]\Big\}} \tag{45}$$

If $w \notin R_{(j)}$, then the derivative just simplifies to:

$$\frac{\partial l_{j_{(i)}}^{Ef}}{\partial \Delta_{1,w}} = 0 \tag{46}$$

Therefore, assumes $w \in D_{(j)}$, then we have:

$$\frac{\partial l^{Ef}_{j_{(i)}}}{\partial \Delta_{1,w}} = -\frac{-\exp(\Delta_{1,i} - \Delta_{1,w}) + \frac{h_i-1}{d_{(j)}}\exp(\Delta_{1,i} - \Delta_{1,w})}{\left\{1 + \sum_{k\in R_j, k\neq i} \exp(\Delta_{1,i} - \Delta_{1,k}) - \frac{h_i-1}{d_{(j)}}\left[1 + \sum_{m \in D_{(j)}, m\neq i} \exp(\Delta_{1,i} - \Delta_{1,m})\right]\right\}} \tag{47}$$

Define the notation ι as below:

$$\iota = -\exp(\Delta_{1,i} - \Delta_{1,w}) + \frac{h_i - 1}{d_{(j)}}\exp(\Delta_{1,i} - \Delta_{1,w}) \tag{48}$$

Therefore, we know that:

$$\frac{\partial \iota}{\partial \Delta_{1,w}} = \exp(\Delta_{1,i} - \Delta_{1,w}) - \frac{h_i - 1}{d_{(j)}}\exp(\Delta_{1,i} - \Delta_{1,w}) = \iota \tag{49}$$

and

$$\frac{\partial \iota}{\partial \Delta_{1,i}} = \iota \tag{50}$$

Define the notation ζ as below:

$$\zeta = \sum_{k \in R_j, k\neq i} \exp(\Delta_{1,i} - \Delta_{1,k}) - \frac{h_i - 1}{d_{(j)}}\Big[\sum_{m\in D_{(j)}, m\neq i} \exp(\Delta_{1,i} - \Delta_{1,m})\Big] \tag{51}$$

It can be shown that $\frac{\partial \zeta}{\partial \Delta_{1,w}} = \iota$, $\frac{\partial \zeta}{\partial \Delta_{1,i}} = \zeta$ and similarly that:

$$\begin{aligned} \frac{\partial l^{Ef}_{j_{(i)}}}{\partial \Delta_{1,w}} &= -\frac{\iota}{\zeta + \frac{d_{(j)} - h_i + 1}{d_{(j)}}} \\ \frac{\partial l^{Ef}_{j_{(i)}}}{\partial \Delta_{1,i}} &= -\frac{\zeta}{\zeta + \frac{d_{(j)} - h_i + 1}{d_{(j)}}} \end{aligned} \tag{52}$$

Consequently, the second derivatives of log-likelihood can be shown to be:

$$\frac{\partial^2 l^{Ef}_{j_{(i)}}}{\partial \Delta_{1,w} \partial \Delta_{1,i}} = -\frac{\iota[\frac{d_{(j)} - h_i + 1}{d_{(j)}}]}{\left[\zeta + \frac{d_{(j)} - h_i + 1}{d_{(j)}}\right]^2} \tag{53}$$

$$\begin{aligned} \frac{\partial^2 l^{Ef}_{j_{(i)}}}{\partial \Delta^2_{1,w}} &= -\frac{\frac{\partial \iota}{\partial \Delta_{1,w}}[\zeta + \frac{d_{(j)} - h_i + 1}{d_{(j)}}] - \iota^2}{\left[\zeta + \frac{d_{(j)} - h_i + 1}{d_{(j)}}\right]^2} \\ &= \frac{\iota\big[\zeta + \frac{d_{(j)} - h_i + 1}{d_{(j)}}\big] + \iota^2}{\left[\zeta + \frac{d_{(j)} - h_i + 1}{d_{(j)}}\right]^2} \end{aligned} \tag{54}$$

For $M \in D_{(i)}$ and $M \neq i, k$, we have:

$$\frac{\partial^2 l^{Ef}_{j_{(i)}}}{\partial \Delta_{1,w} \partial \Delta_{1,M}} = \frac{\iota\left[-\exp(\Delta_{1,i} - \Delta_{1,M}) + \frac{h_i - 1}{d_{(j)}}\exp(\Delta_{1,i} - \Delta_{1,M})\right]}{\left[\zeta + \frac{d_{(j)} - h_i + 1}{d_{(j)}}\right]^2} \tag{55}$$

Finally, for $M \in R_j$ but $M \notin D_j$, we have:

$$\frac{\partial l^{Ef}_{j_{(i)}}}{\partial \Delta_{1,M}} = -\frac{\exp(\Delta_{1,i} - \Delta_{1,M})}{\zeta + \frac{d_j - h_i + 1}{d_j}} \tag{56}$$

$$\frac{\partial^2 l^{Ef}_{j_{(i)}}}{\partial \Delta^2_{1,M}} = \frac{-\exp(\Delta_{1,i} - \Delta_{1,M})[\zeta + \frac{d_j - h_i + 1}{d_j}] + \exp(2\Delta_{1,i} - 2\Delta_{1,M})}{\left[\zeta + \frac{d_j - h_i + 1}{d_j}\right]^2} \tag{57}$$

$$\frac{\partial^2 l^{Ef}_{j_{(i)}}}{\partial \Delta_{1,M} \partial \Delta_{1,i}} = \frac{\exp(\Delta_{1,i} - \Delta_{1,M})\frac{d_j - h_i + 1}{d_j}}{\left[\zeta + \frac{d_j - h_i + 1}{d_j}\right]^2} \tag{58}$$

$$\frac{\partial^2 l^{Ef}_{j_{(i)}}}{\partial \Delta_{1,M} \partial \Delta_{1,w}} = \frac{-\exp(\Delta_{1,i} - \Delta_{1,M})\iota}{\left[\zeta + \frac{d_j - h_i + 1}{d_j}\right]^2} \tag{59}$$

Now, assumes that $a \in R_j$ but $a \notin D_j$, then:

$$\frac{\partial^2 l^{Ef}_{j_{(i)}}}{\partial \Delta_{1,M} \partial \Delta_{1,a}} = \frac{\exp(2\Delta_{1,i} - \Delta_{1,M} - \Delta_{1,a})\iota}{\left[\zeta + \frac{d_j - h_i + 1}{d_j}\right]^2} \tag{60}$$

Using these information above, we can construct the negated Hessian matrix $C_{j_{(i)}}$ for this ith subject, and therefore construct the negated Hessian matrix C_j of the log-likelihood at time $t_{(j)}$, which is $C_j = \sum_{i \in D_{(j)}} C_{j_{(i)}}$. Consequently, we are able to get the C matrix of the full sample by summing these C_j matrixes at different observed times.

3.1.3 Reduction of dimension of C-matrix:

It is worth noticing that in the above derivations, the C-matrix will have a dimension that is more than $n \times n$. Since the sample size n in modern data-set will be super large, the C-matrix will also have a very large dimension with non-sparse components. That will cause some troubles in storage cost and computation time, which is a major concern during the application. Therefore, the Ph.D student Alex Stringer proposed a method that can greatly reduce the dimension of C-matrix. That can be a solution for the computational problems listed above. The idea of the proposed method will be described at below.

Redefine η as $\eta_i := \tilde{X}_i\beta + \sum_{q=1}^{R} \gamma_q(u_{qi})$, which means $\tilde{\eta} = X\beta + A\Gamma$, and we redefine the latent field $\tilde{W} = (\Gamma, \beta)$.

Notice that, η can be linearly transformed into $\Delta := (\Delta_{1,2}, \Delta_{1,3}, ..., \Delta_{1,n})^T$ by $D\eta$, where D is the differencing matrix defined at previous section. Also, $\eta = \Bigg(A, X\Bigg)W$.Therefore, we know that $\Delta = D\Bigg(A, X\Bigg)W = \Bigg(DA, DX\Bigg)W$. Define $DA = A^*$ and $DX = X^*$, then $\Delta = \Bigg(DA, DX\Bigg)W$.

Now, define matrix C as the Hessian matrix of the log-likelihood with respect to Δ, which is already derived in section 3.1.1, and the Hessian matrix of the log-likelihood with respect to W (i.e. $D_w(\nabla_w l)$ the jocobian matrix of the gradient) is of interest. Since log-likelihood l only depends on W through Δ, we can implement chain rule to simplify the computation:

Note that:

$$\begin{aligned} D_w(\Delta) &= \begin{pmatrix} A^* & X^* \end{pmatrix} \\ D_\Delta(\nabla_\Delta l\ ) &= C \end{aligned} \tag{61}$$

Using chain rule we get:

$$\begin{aligned} \nabla_w l &= D_w(l)^T \\ &= D_w(\Delta)^T \nabla_\Delta l \\ &= \begin{pmatrix} A^{*T} \\ X^{*T} \end{pmatrix} \nabla_\Delta l \end{aligned} \tag{62}$$

Then, if we compute the Jocobian matrix of $\nabla_w l$ with respect to W, we can get the Hessian matrix of l with respect to W as following:

$$\begin{aligned} D_w(\nabla_w l\ ) &= \begin{pmatrix} A^{*T} \\ X^{*T} \end{pmatrix} D_\Delta(\nabla_\Delta l\ ) D_w(\Delta) \\ &= \begin{pmatrix} A^{*T} \\ X^{*T} \end{pmatrix} D_\Delta(\nabla_\Delta l\ ) \left(A^*, X^*\right) \\ &= \begin{pmatrix} A^{*T} C A^* & A^{*T} C X^* \\ X^{*T} C A^* & X^{*T} C X^* \end{pmatrix} \end{aligned} \tag{63}$$

The main problem here is that matrix C is highly non-sparse, and has a very big dimension, so the computation of the four matrix cross-products will be challenging when sample size is large. To solve that, procedures below are suggested.

Let C_{ij} denotes the ij th component of matrix C, and a_i^*, x_i^* denote the i^{th} columns of matrix A and X, then the matrix C only affects the Hessian matrix $D_w(\nabla_w l\ )$ through the relationships below:

$$\begin{aligned} \left[A^{*T} C A^*\right]_{ij} &= a_i^{*T} C a_j^* = \sum_{k=1}^{N} \sum_{l=1}^{N} A_{ik}^* A_{jl}^* C_{kl} \\ \left[X^{*T} C A^*\right]_{ij} &= x_i^{*T} C a_j^* = \sum_{k=1}^{N} \sum_{l=1}^{N} X_{ik}^* A_{jl}^* C_{kl} \\ \left[A^{*T} C X^*\right]_{ij} &= a_i^{*T} C x_j^* = \sum_{k=1}^{N} \sum_{l=1}^{N} A_{ik}^* X_{jl}^* C_{kl} \\ \left[X^{*T} C X^*\right]_{ij} &= x_i^{*T} C x_j^* = \sum_{k=1}^{N} \sum_{l=1}^{N} X_{ik}^* X_{jl}^* C_{kl} \end{aligned} \tag{64}$$

The first three sums only consists of a small number of non-zero terms because of the sparsity of matrix A. The last sum will consist of a lot of non-zero terms because matrix X is not as sparse as matrix A in most settings. But that sum can be estimated using random approximation method as below:

$$\begin{aligned} \sum_{k=1}^{N} \sum_{l=1}^{N} X_{ik}^* X_{jl}^* C_{kl} &= N^2 \times \left[\frac{1}{N^2} \sum_{k=1}^{N} \sum_{l=1}^{N} X_{ik}^* X_{jl}^* C_{kl}\right] \\ &\approx N^2 \times \left[\frac{1}{|S|^2} \sum_{k,l \in S} X_{ik}^* X_{jl}^* C_{kl}\right] \end{aligned} \tag{65}$$

Here S is a random sample, which has sample size $|S| \ll N$. If S is randomly selected with enough sample size, then by weak law of large number it is known that this approximation will be very close to the true value of $\left[X^{*T} C X^*\right]_{ij}$.

3.2 Approximation using full-likelihood with left-truncation:

If there are both right-censoring and left-truncations in our data-set, the "data augmentation trick" that INLA uses actually still works theoretically, but the software does not actually allow the user to run the approximation under this scenario. Fortunately, it can be solved using the package of the new proposed algorithm.

Recall that left-truncation happens when we cannot observed the i-th individual lifetime t_i, unless it is greater than the entry time u_i. Under this setup, all the observed lifetimes t_i's are known to be greater than their corresponding entry times u_i's. In other words, we should use conditional probability given $t_i > u_i$ to form our likelihood. For simplicity, let's still consider the same semi-parametric proportional hazard model with piece-wise constant basline hazard.

Denote the i-th lifetime as t_i, the i-th left truncation time is u_i, and assume that $t_i \in (s_{k_{(i)}-1}, s_{k_{(i)}}]$, $u_i \in (s_{m_{(i)}-1}, s_{m_{(i)}}]$. Therefore, we have the likelihood being:

$$\begin{aligned} L &= \prod_{i=1}^{n} \left[\frac{f(t_i)}{S(u_i)}\right]^{\delta_i} \left[\frac{S(t_i)}{S(u_i)}\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} f(t_i)^{\delta_i} \frac{S(t_i)^{1-\delta_i}}{S(u_i)} \\ &= \prod_{i=1}^{n} h(t_i)^{\delta_i} \frac{S(t_i)}{S(u_i)} \end{aligned} \tag{66}$$

Using the likelihood above, we can easily derive the log-likelihood being:

$$l = \sum_{i=1}^{n} \delta_i \text{log}[h(t_i)] + \sum_{i=1}^{n} \left[\text{log}S(t_i) - \text{log}S(u_i)\right] \tag{67}$$

Recall that if $t_i \in (s_{k_{(i)}-1}, s_{k_{(i)}}]$, and $u_i \in (s_{m_{(i)}-1}, s_{m_{(i)}}]$, we have the followings:

$$\begin{aligned} \text{log}S(t_i) &= -\int_0^{t_i} h(x)dx \\ &= -\sum_{j=1}^{k_{(i)}-1} (S_j - S_{j-1})\text{exp}(\eta_{ij}) - (t_i - S_{k_{(i)}-1})\text{exp}(\eta_{ik_{(i)}}) \end{aligned} \tag{68}$$

Similarly:

$$\begin{aligned} \text{log}S(u_i) &= -\int_0^{u_i} h(x)dx \\ &= -\sum_{j=1}^{m_{(i)}-1} (S_j - S_{j-1})\text{exp}(\eta_{ij}) - (u_i - S_{m_{(i)}-1})\text{exp}(\eta_{im_{(i)}}) \end{aligned} \tag{69}$$

Therefore, the difference between this two terms can be written as:

$$\text{log}S(t_i) - \text{log}S(u_i) = -\int_{u_i}^{t_i} h(x)dx \tag{70}$$

Whereas:

$$-\int_{u_i}^{t_i} h(x)dx = -\sum_{j=m_{(i)}+1}^{k_{(i)}-1} (S_j - S_{j-1})\text{exp}(\eta_{ij}) - (S_{m_{(i)}} - u_i)\text{exp}(\eta_{im_{(i)}}) - (t_i - S_{k_{(i)}-1})\text{exp}(\eta_{n_{ik_{(i)}}}) \tag{71}$$

Then, combine all of the information above together, we can derive an expression for the log-likelihood of the sample:

$$l = \sum_{i=1}^{n} \delta_i \eta_{ik_{(i)}} - \sum_{i=1}^{n} \sum_{j=m_{(i)}+1}^{k_{(i)}-1} (S_j - S_{j-1}) \exp(\eta_{ij}) - \sum_{i=1}^{n} (S_{m_{(i)}} - u_i) \exp(\eta_{im_{(i)}}) - \sum_{i=1}^{n} (t_i - S_{k_{(i)}-1})(\eta_{ik_{(i)}}) \tag{72}$$

If we have $m_{(i)} \leq k_{(i)} - 1$, then the above expression simplify to:

$$l = \sum_{i=1}^{n} \delta_i \eta_{ik_{(i)}} - \sum_{i=1}^{n} (S_{m_{(i)}} - u_i) \exp(\eta_{im_{(i)}}) - \sum_{i=1}^{n} (t_i - S_{k_{(i)}-1})(\eta_{ik_{(i)}}) \tag{73}$$

Next step, I will derive the corresponding C and Q matrix in this case.

3.2.1 Derivation of C-matrix with left-truncation

To make the derivation most general, I will assume that $k_{(i)} - 1 \geq m_{(i)} + 1$, since otherwise the computation will be simplified to trivial. For the i-th observation $t_{(i)}$ with left-truncation time $u_{(i)}$, assume that: $t_i \in (s_{k_{(i)}-1}, s_{k_{(i)}}]$, and $u_i \in (s_{m_{(i)}-1}, s_{m_{(i)}}]$, then the likelihood of this observation will be:

$$l = \delta_i \eta_{ik_{(i)}} - \sum_{m_{(i)}+1}^{k_{(i)}-1} (S_j - S_{j-1}) \exp(\eta_{ij}) - (S_{m_{(i)}} - u_{(i)}) \exp(\eta_{im_{(i)}}) - (t_{(i)} - S_{k_{(i)}-1}) \exp(\eta_{ik_{(i)}}) \tag{74}$$

For $j < m_{(i)}$ or $j > k_{(i)}$, apparently we have $\frac{\partial l}{\partial \eta_{ij}} = \frac{\partial^2 l}{\partial \eta_{ij}^2} = 0$.

For $j = m_{(i)}$, we can compute that $\frac{\partial l}{\partial \eta_{ij}} = \frac{\partial^2 l}{\partial \eta_{ij}^2} = -(S_{m_{(i)}} - u_{(i)}) \exp(\eta_{im_{(i)}})$.

For $m_{(i)} < j < k_{(i)}$, $\frac{\partial l}{\partial \eta_{ij}} = \frac{\partial^2 l}{\partial \eta_{ij}^2} = -(S_j - S_{j-1}) \exp(\eta_{ij})$.

For $j = k_{(i)}$, it can be shown that:

$$\frac{\partial l}{\partial \eta_{ij}} = \delta_i - (t_{(i)} - S_{k_{(i)}-1}) \exp(\eta_{ik_{(i)}}) \tag{75}$$

So,

$$\frac{\partial^2 l}{\partial \eta_{ij}^2} = -(t_{(i)} - S_{k_{(i)}-1}) \exp(\eta_{ik_{(i)}}) \tag{76}$$

From now, let's denote $\exp(\eta_{ij})$ as b_{ij}. Now, we can use the above information, to obtain the C-matrix (negated Hessian) of the i-th observation:

$$\begin{pmatrix}
0 & & & \cdots & \cdots & & \cdots \\
0 & \ddots & & & & & \\
\vdots & \cdots & (S_{m_{(i)}} - u_{(i)})b_{im_{(i)}} & & \cdots & \cdots & \cdots \\
\vdots & & & (S_{m_{(i)}+1} - S_{m_{(i)}})b_{i(m_{(i)}+1)} & & \cdots & \cdots \\
\vdots & & & \ddots & \ddots & \ddots & \cdots \\
\vdots & & & \cdots & (S_{k_{(i)}-1} - S_{k_{(i)}-2})b_{i(k_{(i)}-1)} & \cdots & \\
\vdots & & & & \cdots & (t_i - S_{k_{(i)}-1})b_{ik_{(i)}} & \cdots \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots
\end{pmatrix} \tag{77}$$

This is a matrix with $k \times k$ dimension.

Let's call the C-matrix of observation i as C_i, then the C-matrix of the whole sample will be:

$$C = \begin{pmatrix}
C_1 & 0 & 0 & \cdots & & & \\
0 & C_2 & 0 & \cdots & & & \\
& \cdots & \ddots & & & & \\
& & & & C_n & \cdots & \vdots \\
& & & \ddots & & & \vdots \\
& & & & & & 0
\end{pmatrix} \tag{78}$$

This C matrix has a dimension of $(nk + \sum_{q=1}^{R} M_q + p + k) \times (nk + \sum_{q=1}^{R} M_q + p + k)$, and each block matrix C_i has the same dimension being $k \times k$.

We can see that the present of left-truncation does not change the overall shape of the C-matrix. The only effect of it is to change the diagonal terms of each individual observation's C_i matrix depending on the i-th left-truncation time. Therefore, the computation efficiency and precision will not be affected too much.

3.3 Approximation using full-likelihood with interval-censoring:

Suppose that we are not observing the exact lifetimes of individiuals, but only the set of intervals that contain each lifetime. In other words, our data-set is $\{L_i, R_i; i = 1, ...n\}$, where $L_i \leq t_i \leq R_i$. Using the same way to define the piece-wise constant hazard functions as before, we can assume that for the i-th observation, we have $R_i \in (S_{k_{(i)}-1}, S_{k_{(i)}}]$, and $L_i \in (S_{m_{(i)}-1}, S_{m_{(i)}}]$.

Now, we can write down the log-likelihood of the i-th individual using the above information:

$$\begin{aligned} l_i &= \delta_i \log\big[h_i(R_i)\big] - \int_0^{R_i} h_i(u)du + \log\Big\{1 - \exp\big[-\int_{L_i}^{R_i} h_i(u)du\big]\Big\} \\ &= \delta_i \eta_{ik_i} - \sum_{j=1}^{k_{(i)}-1} \exp(\eta_{ij}) - (R_i - S_{k_{(i)}-1})\exp(\eta_{ik_{(i)}}) + \log\big[1 - \exp(\vartheta_i)\big] \end{aligned} \tag{79}$$

Where ϑ_i is defined as:

$$\vartheta_i = -\int_{L_i}^{R_i} h_i(u)du = -\sum_{j=m_{(i)}+1}^{k_{(i)}-1} (S_j - S_{j-1})\exp(\eta_{ij}) - (S_{m_{(i)}} - L_i)\exp(\eta_{im_{(i)}}) - (R_i - S_{k_{(i)}-1})\exp(\eta_{ik_{(i)}}) \tag{80}$$

Now, we can take derivative of this log-likelihood with respect to the ij-th linear predictor (assume that $m_{(i)} + 1 \leq j \leq k_{(i)} - 1$), and get the following result:

$$\frac{\partial l_i}{\partial \eta_{ij}} = -(S_j - S_{j-1})\exp(\eta_{ij}) - \frac{\exp(\vartheta_i)\frac{\partial \vartheta_i}{\partial \eta_{ij}}}{1 - \exp(\vartheta_i)} \tag{81}$$

Where $\frac{\partial \vartheta_i}{\partial \eta_{ij}}$ is $-(S_j - S_{j-1})\exp(\eta_{ij})$ in this case.

Since ϑ_i depends on more than one linear predictors, so it follows naturally that $\frac{\partial l_i}{\partial \eta_{ij}}$ will be a function of several different linear predictors, which means the C-matrix of each observation's log-likelihood will be non-diagonal. That will be a serious problem for INLA, as its package relies on the diagonality of C-matrix, but it will not cause any probelm for our new proposed algorithm, as the diagonality of C-matrix is no longer required here.

4 Example from Diabetics Data-set

Firstly, I will use the data-set "diabetics" to demonstrate the equivalence between "coxph" approach in survival package, INLA's approach and our proposed approach. This data-set contains the results from a trail of laser coagulation for the treatment of diabetic retinopathy from 197 patients. Each patient had one eye randomized to laser treatment and the other eye received no treatment. (To be continued on "proposed approach")

The variable "id" specifies the subject's ID.

The variable "laser" is a categorical variable with levels xenon or argon.

The variable "age" is the age of the subject at diagnosis.

The variable "eye" is a categorical variable with levels left or right.

The variable "trt" is a categorical variable with levels 0 for no treatment and 1 for treatment using laser.

The variable "risk" classifies the risk levels of the patients.

The response variable in this data-set will be “time”, which are the actual time to blindness in months, minus the minimum possible time to event (6.5 months), and “status” indicates whether the time is censored with 1 for visual loss and 0 for censored. The censoring can be due to death, dropout, or end of the study.

Let use briefly view the structure of “diabetics”:

4.1 Data-set:

```
head(as_tibble(diabetic))
```

```
## # A tibble: 6 x 8
##      id laser   age eye     trt  risk  time status
##   <int> <fct> <int> <fct> <int> <int> <dbl>  <int>
## 1     5 argon    28 left      0     9  46.2      0
## 2     5 argon    28 right     1     9  46.2      0
## 3    14 xenon    12 left      1     8  42.5      0
## 4    14 xenon    12 right     0     6  31.3      1
## 5    16 xenon     9 left      1    11  42.3      0
## 6    16 xenon     9 right     0    11  42.3      0
```

We can see that those survival times are right-censored. We will fit a cox proportional hazard model with piece-wise constant baseline hazard, assuming that each individual will have the same baseline hazard function. The variable ID will be treated as a random effect(which can be added using code “frailty.gaussian(id)”). The variables “age”, “eye”, “trt” and “laser” will be included as fixed effects.

4.2 Survival: coxph

```
diabetic.CoxPh <- coxph(Surv(time, status)~age + eye + trt + laser + frailty.gaussian(id), data = diabe
summary(diabetic.CoxPh)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ age + eye + trt + laser + 
##     frailty.gaussian(id), data = diabetic)
## 
##   n= 394, number of events= 155 
## 
##                      coef      se(coef) se2      Chisq  DF    p      
## age                   0.009548 0.01323  0.009879   0.52  1.00 4.7e-01
## eyeright              0.483005 0.17501  0.168693   7.62  1.00 5.8e-03
## trt                  -1.007507 0.17930  0.174315  31.57  1.00 1.9e-08
## laserargon           -0.182388 0.39471  0.293566   0.21  1.00 6.4e-01
## frailty.gaussian(id)                             131.35 79.63 2.4e-04
## 
##            exp(coef) exp(-coef) lower .95 upper .95
## age           1.0096     0.9905    0.9837    1.0361
## eyeright      1.6209     0.6169    1.1503    2.2842
## trt           0.3651     2.7388    0.2569    0.5189
## laserargon    0.8333     1.2001    0.3844    1.8062
## 
## Iterations: 6 outer, 24 Newton-Raphson
##      Variance of random effect= 0.9447455 
## Degrees of freedom for terms=  0.6  0.9  0.9  0.6 79.6 
## Concordance= 0.867  (se = 0.867 )
## Likelihood ratio test= 228  on 82.62 df,   p=2e-15
```

From the output above, it can be seen that variables “age” and “eyeright” have positive association with the rate of occurrence of visual loss, and variables “trt” and whether using “argon” type of laser are negatively associated with the rate. All the fixed effects that we included in this study have significant effects for the risk of visual loss.

4.3 Bayesian: INLA

Now we fit the same model using INLA:

```
formula = inla.surv(time, status) ~ age + eye + trt + laser + f(id, model = "iid")
diabetic.INLA <- inla(formula, control.compute = list(dic = TRUE), family = "coxph",
                      data = diabetic, control.hazard=list(model="rw2", n.intervals=20))
diabetic.INLA$summary.fixed
```

```
##                     mean         sd  0.025quant     0.5quant  0.975quant
## (Intercept) -4.763724360 0.24995385 -5.27245603 -4.757703164 -4.28882306
## age          0.007159963 0.01085615 -0.01385335  0.007024012  0.02907695
## eyeright     0.385773402 0.17638804  0.04934098  0.382233466  0.74283671
## trt         -0.867270167 0.19242675 -1.26420670 -0.860404353 -0.50839819
## laserargon  -0.118815204 0.32295500 -0.77665993 -0.111806295  0.49861264
##                     mode          kld
## (Intercept) -4.746076640 3.754775e-06
## age          0.006834285 8.264532e-06
## eyeright     0.375260163 6.104375e-06
## trt         -0.846042871 1.464370e-05
## laserargon  -0.099015911 3.861603e-06
```

It seems like these two approaches are similar enough. Though in the classic “coxph” approach, the effects of age and using argon-type laser are significant, but INLA gives insignificant results(the 95% credible interval contains zero). There is an estimate for intercept in the INLA’s method because we used a random walk prior in that.

5 Example from Bladder Data-set:

Next, we will study the two approaches on the data-set “bladder1”. This is the full data-set that contains the result from a study on recurrences of bladder cancer from 118 subjects. In this data-set, the variables that we are interested in are “id”, “number”, “size”, “recur”, “times” and “censored”.

The variable “id” is just the patient ID.

The variable “number” specifies initial number of tumors of each subject.

The variable “size” is the size of largest initial tumor.

The variable “recur” is the number of recurrence of bladder cancer for that subject.

The response variable will be “time” which is computed to be the duration of times until recurrence or death, censored by the variable “censored” with 0 means being censored.

5.1 Data-set:

```
data <- as_tibble(bladder1)
data <- select(data,-c(rsize,rtumor,enum))
data <- data %>% mutate(censored = status==0)
for (i in 1:length(data$censored)) {
  if(data$censored[i]) data$censored[i] <- 0
  else data$censored[i] <- 1
}
data <- data %>% mutate(times = stop-start)
head(data)
```

```
## # A tibble: 6 x 10
##      id treatment number  size recur start  stop status censored times
##   <int> <fct>      <int> <int> <int> <int> <int>  <dbl>    <dbl> <int>
## 1     1 placebo        1     1     0     0     0      3        1     0
## 2     2 placebo        1     3     0     0     1      3        1     1
## 3     3 placebo        2     1     0     0     4      0        0     4
## 4     4 placebo        1     1     0     0     7      0        0     7
## 5     5 placebo        5     1     0     0    10      3        1    10
## 6     6 placebo        4     1     1     0     6      1        1     6
```

```
tail(data)
```

```
## # A tibble: 6 x 10
##      id treatment number  size recur start  stop status censored times
##   <int> <fct>      <int> <int> <int> <int> <int>  <dbl>    <dbl> <int>
## 1   115 thiotepa       4     1     3    24    47      1        1    23
## 2   115 thiotepa       4     1     3    47    50      0        0     3
## 3   116 thiotepa       3     4     0     0    54      0        0    54
## 4   117 thiotepa       2     1     1     0    38      1        1    38
## 5   117 thiotepa       2     1     1    38    54      0        0    16
## 6   118 thiotepa       1     3     0     0    59      3        1    59
```

Here the variable “id” specifies different individuals, and should be treated as a random effect. The variable “time” is computed using the difference between variables “start” and “stop”, which denote the start time and end time of each time interval. It seems like a interval censoring problem but the start time is known before hand, so we can treat it as a regular type-I right censoring.

In this study, a nonzero value of “status” can be death from bladder disease, death from other reason or recurrence. Here we will just view all of these situations as “occurrence” for simplicity. So the variable “censored” is created such that it is 1 if “status” is non-zero, otherwise 0. We will include “number”, “size” and “recur” as fixed effects in this study.

5.2 Survival:coxph

```
bladder.CoxPh <- coxph(Surv(times, censored)~ number + size + recur + frailty.gaussian(id), data = data
summary(bladder.CoxPh)
```

```
## Call:
## coxph(formula = Surv(times, censored) ~ number + size + recur +
##     frailty.gaussian(id), data = data)
##
##   n= 294, number of events= 218
##
##                         coef     se(coef) se2     Chisq DF   p
## number                 0.064848 0.04246  0.04051  2.33 1.00 1.3e-01
## size                   0.008336 0.04797  0.04631  0.03 1.00 8.6e-01
## recur                  0.229848 0.02559  0.02431 80.66 1.00 2.7e-19
## frailty.gaussian(id)                              5.12 4.59 3.5e-01
##
##        exp(coef) exp(-coef) lower .95 upper .95
## number     1.067     0.9372    0.9818     1.160
## size       1.008     0.9917    0.9179     1.108
## recur      1.258     0.7947    1.1968     1.323
##
## Iterations: 8 outer, 43 Newton-Raphson
##      Variance of random effect= 0.0257231
## Degrees of freedom for terms= 0.9 0.9 0.9 4.6
## Concordance= 0.694  (se = 0.694 )
## Likelihood ratio test= 105.4  on 7.33 df,   p=<2e-16
```

Fitting this model using the traditional partial likelihood approach gives insignificant results for all the fixed effects except "recur", which has a strong positive effect. But we will still proceed to check what will happen if we fit it using a Bayesian approach.

5.3 Bayesian: INLA

```
formula = inla.surv(times, censored) ~ number + size + recur + f(id, model = "iid")
bladder.INLA <- inla(formula, control.compute = list(dic = TRUE), family = "coxph",
                     data = data, control.hazard=list(model="rw2", n.intervals=20))
bladder.INLA$summary.fixed
```

```
##                      mean         sd  0.025quant    0.5quant  0.975quant
## (Intercept) -3.945620507 0.26869179 -4.49395885 -3.93815413 -3.43930106
## number       0.061328623 0.04045256 -0.02007891  0.06202239  0.13881964
## size         0.009253765 0.04631404 -0.08484274  0.01037918  0.09704015
## recur        0.215846265 0.02392154  0.16899563  0.21580583  0.26288177
##                    mode          kld
## (Intercept) -3.92341056 5.445038e-06
## number       0.06339941 2.829580e-07
## size         0.01261179 7.226282e-07
## recur        0.21572706 1.743425e-06
```

Indeed, the two results seem pretty similar in general. In both cases, we can see that there are no apparent relationships between all of the fixed effects and the rate of occurrence of bladder cancer's recurrence, or death, except the variable "recur" with a positive effect.