Bayesian Inference for Cox Proportional Hazard Models with Partial Likelihoods, Semi-Parametric Covariate Effects and Correlated Observations

07/02/2020

Abstract

We develop an approximate Bayesian inference methodology for the Cox Proportional Hazards model for survival data with partial likelihood, semi-parametric covariate effects and correlated survival times. Bayesian inference provides model-based uncertainty quantification of the smoothness parameters and between-group standard deviations. The use of partial likelihood avoids smoothness assumptions on the baseline hazard, leading to improved inferences over current methods for approximate Bayesian inference for this model (INLA). A simulation study demonstrates the superior accuracy of our approximations over existing methods when the baseline hazard is not smooth. Analysis of two benchmark datasets demonstrates the improved accuracy of our method over INLA when compared to the frequentist generalized additive models (GAMs), but in contrast to GAMs, we yield full posterior uncertainty for the smoothness of the semi-parametric effect and the between-subject standard deviation. An R package implementing our method will be released publicly.

1 Introduction

Survival data consists of times to an event of interest such as mortality or morbidity. When analysing survival data, the Cox proportional hazards (Cox PH) model is a common choice. The Cox PH model assumes that any two subjects' event hazards are proportional as a function of time, with the ratio depending on covariate effects which are modelled as unknown linear or smooth functions and inferred from the observed data. Event times may be correlated within the sample, for example when the response is time to kidney failure for the left and right kidneys from the same subject. In order to avoid having to make assumptions about the shape of the unknown common baseline event hazard, a partial likelihood is often used, which does not depend on this quantity. The use of Bayesian inference with the Cox PH model is desirable as this yields model-based

estimation and uncertainty quantification for the smoothness of the covariate effects, and in the case of correlated survival times, the between-group standard deviations. However, existing methods for approximate Bayesian inference based on Integrated Nested Laplace Approximations (INLA) Rue et al. (2009) cannot be applied to the Cox PH model with partial likelihood, and hence require restrictive smoothness assumptions to be made about the baseline hazard.

Recently, Stringer et al. (2020) developed an approximate Bayesian inference methodology for a model involving a partial likelihood. Their methodology includes smooth covariate effects and yields full posterior uncertainty for the smoothness parameters, an improvement over existing frequentist methods based on Generalized Additive Models (GAMs), and they demonstrate up to an order of magnitude improvement in computation time when compared to sampling-based approaches to Bayesian inference. They note in their discussion that the partial likelihood involved in the Cox PH model has a similar form to the partial likelihood from their model, and that the Cox PH model with partial likelihood would be a feasible extension of their work.

In this paper we extend the approximate Bayesian inference methodology of Stringer et al. (2020) to the Cox proportional hazard models with partial likelihood. Our methodology accommodates semi-parametric smoothing effects and correlation between observed survival times. We demonstrate improved accuracy over INLA in simulations and two data analysis examples, and provide model-based estimation and uncertainty quantification for the smoothness of effects and between-subject standard deviations.

The remainder of this paper is organized as follows. In §2, we describe the Cox proportional hazard model and the partial likelihood function, and review the approximate Bayesian inference methodology of Stringer et al. (2020). In §3, we describe our proposed methodology. In §4 we illustrate our methodology in a simulation study and through the analysis of two benchmark datasets. We conclude in §5 with a discussion.

2 Preliminaries

2.1 Cox Proportional Hazard Model

Let T denote a random variable representing the time to some event, supported on the interval $[0, \infty)$. For $t \in [0, \infty)$ the *hazard function* $h(t)$ of T is defined as:

$$h(t) = \lim_{s \to 0} \frac{P(t \leq T \leq t + s | T \geq t)}{s} \tag{1}$$

Suppose we observe $i = 1, \ldots, n$ groups each with $j = 1, \ldots, n_i$ survival times. For example, we may observe n subjects with n_i measurements per subject. Denote the random variable representing the j^{th} survival time in the i^{th} group by Y_{ij}, and denote the survival times by $y = \{y_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$. When there are some survival data in a data-set that are not observed exactly, we call these observations *censored*. Right-censoring is a common case of censoring when some survival times are only known to be longer than some specific times. For example, if the observation y_{ij} is right-censored, then the only information available for Y_{ij} is that $Y_{ij} > y_{ij}$. We will focus on this type of censoring in this paper. Define $h_{ij}(t)$ to be the hazard function for random variable Y_{ij}. The Cox PH model assumes (Cox, 1972)

$$h_{ij}(t) = h_0(t)\exp(\eta_{ij}) \tag{2}$$

where $h_0(t)$ is an unknown baseline hazard function that does not depend on the covariates. The additive predictor $\eta = \{\eta_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$ links the covariates for observation y_{ij} to the observed survival time:

$$\eta_{ij} = x_{ij}^T \beta + \sum_{q=1}^{r} \gamma_q(u_{q_{ij}}) + \xi_i \tag{3}$$

Here x_{ij} is a p-dimensional vector of covariates that are modelled as having linear associations with the log-hazard, and $\beta = (\beta_1, \ldots, \beta_p)$ are regression coefficients. The $u_q = \{u_{qij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}, q = 1, \ldots, r$ are covariate vectors whose association with the log-hazard is modelled non-parametrically through unknown smooth functions $\gamma_1, \ldots, \gamma_r$. The vector of group intercepts $\xi = \{\xi_i : i = 1, \ldots, n\}$, referred to as "frailty" coefficients in the context of survival analysis (Vaupel et al., 1979), are included to model correlation between survival times coming from the same group i.

Inference is carried out via a partial likelihood function. Define the *risk set* $R_{ij} = \{k, l : y_{kl} \geq y_{ij}\}$, the set of indices of survival times which occur at or after y_{ij}. For simplicity, assumes for each group i observations $1, 2, ..., r_i$ are actually *observed* distinct survival times and the rest $n_i - r_i$ observations are right-censored, then partial likelihood can be written as follows:

$$\begin{aligned} \pi(y|\eta) &= \prod_{i=1}^{n} \prod_{j=1}^{r_i} \left\{ \frac{\exp[\eta_{ij}]}{\sum_{l,k \in R_{ij}} \exp[\eta_{lk}]} \right\} \\ &= \prod_{i=1}^{n} \prod_{j=1}^{r_i} \left\{ \frac{1}{1 + \sum_{l,k \in R_{ij}, (l,k) \neq (i,j)} \exp[\Delta_{lk,ij}]} \right\} \end{aligned} \tag{4}$$

where $\Delta_{lk,ij} = \eta_{lk} - \eta_{ij}$. Note that $h_0(t)$ does not appear in the partial likelihood, and hence inference may be carried out in the absence of assumptions about $h_0(t)$. Also note that this partial likelihood can be written in the following form:

$$\pi(y|\eta) = \prod_{i=1}^{n} \prod_{j=1}^{r_i} \pi(y_{ij}|\eta) \tag{5}$$

while in order for a model to be compatible with INLA, its likelihood must have the form:

$$\pi(y|\eta) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \pi(y_{ji}|\eta_{ij}) \tag{6}$$

2.2 Approximate Bayesian Inference

To perform Bayesian inference for this model, we specify prior distributions for the parameters. A joint Gaussian prior distribution with fixed covariance matrix is used for $\beta \sim \mathrm{N}(0, \Sigma_\beta)$; we follow Stringer et al. (2020) and use $\Sigma_\beta = \sigma_\beta^2 I_p$, with $\sigma_\beta^2 = 1000$. The group intercepts are given independent Gaussian priors $\xi_i|\theta \overset{iid}{\sim} \mathrm{N}(0, \sigma_\xi), i = 1, \ldots, n$ where σ_ξ is the between-groups standard deviation. Let $U_q = \{U_{ql}; l = 1, ..., m_q\}$ be the ordered vector of *unique* values of covariate $u_q, q = 1, \ldots, r$; often these values are set by the user by discretizing the covariate u_q into m_q pre-specifed bins. To infer the infinite-dimensional parameters $\gamma_q, q = 1, \ldots, r$, we approximate each by a piecewise constant function with jumps at the U_{ql}, which we denote as $\gamma(U_{ql}) = \Gamma_{ql}$. We define the vectors of function values $\Gamma_q = \left\{\Gamma_{q1}, \ldots, \Gamma_{qm_q}\right\}$ and these are given a joint Gaussian distribution $\Gamma_q|\theta \sim \mathrm{N}\left[0, \Sigma_q(\sigma_q)\right]$ which is parametrized through its precision matrix $\Sigma_q(\sigma_q)$ depending on a variance parameter σ_q. A popular choice which we adopt in our analysis is the second-order random walk model (Lindgren and Rue, 2008), but different choices including longitudinal or spatial correlation structures are possible under our approach. Let $\Gamma = (\Gamma_1, \ldots, \Gamma_r)$; we have that $\Gamma|\sigma_1, \ldots, \sigma_q \sim \mathrm{N}\left(0, \Sigma_\Gamma^{-1}\right)$ with $\Sigma_\Gamma^{-1} = \mathrm{diag}\left[\Sigma_1^{-1}(\sigma_1), \ldots, \Sigma_r^{-1}(\sigma_r)\right]$. Finally, define the variance parameter vector $\theta = (\theta_0, \ldots, \theta_r)$ where $\theta_q = -2\log\sigma_q, q = 1, \ldots, r$, and $\theta_0 = -2\log\sigma_\xi$. The variance parameters are given prior distribution $\theta \sim \pi(\theta)$.

For computational purposes, we follow Rue et al. (2009) and Stringer et al. (2020) to add a small random noise on the linear predictor, redefining:

$$\eta_{ij} = x_{ij}^T\beta + \sum_{q=1}^{r} \gamma_q(u_{q_{ij}}) + \xi_i + \epsilon_{ij} \tag{7}$$

where $\epsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0, \tau^{-1})$ for some large, fixed τ. We follow the established default used by Rue et al. (2009) and Stringer et al. (2020) and set $\tau = \exp(12)$ so the addition of the ϵ noise does not significantly change the inferential result. In particular, Stringer et al. (2020) demonstrate in their Web Appendix E that choices of τ in the broad range of $\exp(2), \ldots, \exp(14)$ yield virtually identical inferences and similar running times. Further redefine $\Delta_{lk,ij} = \eta_{lk} - \eta_{ij}$ in terms of the augmented additive predictors (7), and note that since $\Delta_{lk,ij} = \Delta_{11,ij} - \Delta_{11,lk}$ for every (i, j, l, k), the entire partial likelihood (4) depends on η only through the vector $\Delta = \{\Delta_{11,ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$. For the remainder of the paper we reflect this in our notation, writing $\pi(y|\Delta) \equiv \pi(y|\eta)$ and defining the log-likelihood $\ell(\Delta; y) = \log\pi(y|\Delta)$.

Define $W = (\Delta, \Gamma, \beta, \xi)$ which we refer to as the *mean parameters* and let $\dim(W) = m$. Our model specifies $W|\theta \sim \mathrm{N}\left[0, Q_\theta^{-1}\right]$. An expression and derivation for Q_θ is given by Stringer et al. (2020) (Web Appendix C), and in §3 we discuss the differences in our Q_θ compared to theirs. Our main inferential interest is to obtain the joint posterior distribution of the mean parameters,

$$\pi(W|y) = \int \pi(W|y,\theta)\pi(\theta|y)d\theta \tag{8}$$

the marginal posterior distributions of the mean parameters,

$$\pi(W_k|y) = \int \pi(W_i|y,\theta)\pi(\theta|y)d\theta, k = 1,\ldots,m \tag{9}$$

and of the joint posterior distributions of the variance parameters:

$$\pi(\theta|y) = \frac{\int \pi(W,y,\theta)dW}{\int\int \pi(W,y,\theta)dW\,d\theta} \tag{10}$$

All of the quantities of interest (8) – (10) depend on intractable high-dimensional integrals. Stringer et al. (2020) utilize Gaussian and Laplace approximations combined with numerical quadrature to approximate each of these integrals accurately and efficiently. Their approximations take the form

$$\begin{aligned} \tilde{\pi}(W_j|y) &= \sum_{k=1}^{K} \tilde{\pi}_G(W_j|y,\theta^k)\tilde{\pi}_{LA}(\theta^k|y)\delta_k \\ \tilde{\pi}(W|y) &= \sum_{k=1}^{K} \tilde{\pi}_G(W|y,\theta^k)\tilde{\pi}_{LA}(\theta^k|y)\delta_k \end{aligned} \tag{11}$$

For any fixed θ, define

$$\begin{aligned} \widehat{W}_\theta = \left(\widehat{\Delta}_\theta, \widehat{\Gamma}_\theta, \widehat{\beta}, \widehat{\xi}_\theta\right) &= \mathrm{argmax}_W \log \pi(W|\theta, Y) \\ H_\theta(W) &= -\frac{\partial^2}{\partial W \partial W^T} \log \pi(W|\theta, Y) \\ v(\theta)_j^2 &= \left[H_\theta\left(\widehat{W}_\theta\right)^{-1}\right]_{jj} \end{aligned} \tag{12}$$

The approximations (11) depend on Gaussian approximations,

$$\begin{aligned} \pi(W|\theta,Y) \approx \tilde{\pi}_G(W|y,\theta) &\propto \exp\left\{-\frac{1}{2}\left[W - \widehat{W}_\theta\right]^T H_\theta[\widehat{W}_\theta]\left[W - \widehat{W}_\theta\right]\right\} \\ \pi(W_k|\theta,Y) \approx \tilde{\pi}_G(W_k|y,\theta) &\propto \exp\left\{-\frac{1}{2v(\theta)_j^2}\left[W_j - \widehat{W}_{\theta j}\right]^2\right\}, k = 1,\ldots,m \end{aligned} \tag{13}$$

and Laplace approximations Tierney and Kadane (1986):

$$\pi(\theta|Y) \approx \tilde{\pi}_{LA}(\theta|y) \propto \pi(\theta)\frac{|Q_\theta|^{1/2}}{|H_\theta[\widehat{W}_\theta]|^{1/2}} \exp\left\{-\frac{1}{2}\widehat{W}_\theta^T Q_\theta \widehat{W}_\theta + \ell\left(\widehat{\Delta}_\theta; y\right)\right\} \tag{14}$$

The Hessian matrix $H_\theta(W)$ has the form $H_\theta(W) = Q_\theta + C(W)$ where

$$C(W) = -\frac{\partial^2}{\partial W \partial W^T}\ell(\Delta) = -\begin{pmatrix} \frac{\partial^2 \ell(\Delta;y)}{\partial\Delta\partial\Delta^T} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Because the partial likelihood takes the form (5), $C(W)$ has a dense structure. In contrast, Rue et al. (2009) assume that the likelihood takes the form (6) which has $C(W) \equiv \text{diag}(c)$, and hence cannot fit the Cox PH model with partial likelihood. Stringer et al. (2020) relax this assumption to allow $C(W)$ to have a block-diagonal structure; our work extends this to permit a fully dense $C(W)$.

3 Methodology

In §2, it is shown that the partial likelihood function only depends on η_{ij} through $\Delta_{11,ij}$. The relationship between Δ and η can be written as:

$$\Delta = D\eta \tag{15}$$

Where D is the differencing matrix defined as:

$$D = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ & & \ddots & & \\ 1 & & & 0 & -1 \end{pmatrix} \tag{16}$$

The dimension of this differencing matrix is $(\sum_{i=1}^{n} n_i - 1) \times (\sum_{i=1}^{n} n_i)$ with rank $(\sum_{i=1}^{n} n_i - 1)$, and DD^T is an non-singular matrix. The D matrix in this case is differencing every observations with the first observation, which is different from the differencing matrix in Stringer et al. (2020). The differencing for Cox PH model is across all subjects in the data set, while the differencing for case crossover model is only within subjects. Because this D matrix will affect the precision matrix of W directly through the transformation on η, the precision matrix for Cox PH model will be different from the one Stringer et al. (2020) obtained in their paper. Notice since that the partial likelihood of Cox PH model is a marginal likelihood of the ranks of all the observed survival times, if there are more than one subject in the data set, the between-subjects frailty will be identifiable from the partial likelihood which makes its inference feasible in the model. While in case-crossover model, the between-subject frailty will not be identifiable from its partial likelihood function so we cannot carry out inference for it.

To compute the conditional mode $\hat{W}(\theta)$, we utilized trust region optimization with quasi-Newton updates method called Symmetric Rank 1 method (SR1). Such methods approximates the true Hessian matrix at each iteration with a rank-1 update from the previous iteration (Braun, 2014). This quasi-Newton type of method is feasible to use for the optimization in the presence of a dense Hessian matrix, because it does not require the actual evaluation of the Hessian matrix at each iteration. In Stringer et al. (2020)'s work, the computation

of the conditional mode $\hat{W}(\theta)$ was done using trust region methods with conjugate-gradient updates. Because the Hessian matrix in case-crossover model is highly sparse and block-diagonal, the computation of this matrix at each iteration will be fast. This type of method will not be feasible for our problem here as it will require the evaluation of a dense matrix at each iteration, which has computation time scales quadratically with the sample size.

In our previous model specification, we assume that there are no ties of survival times in the data set. When there exists ties between some of the survival times, we correct our partial likelihood and its corresponding Hessian matrix using Breslow's method(Breslow, 1974). The computational advantage of Breslow's approximation is that it does not need to average over the possibilities of possible ordering of these tied survival times, and it preserves the basic structure of the partial likelihood and its corresponding Hessian matrix. This correction method tends to be accurate as long as the number of tied observations is not significantly large compared to the sample size. In Stringer et al. (2020)'s work on case-crossover model, the tied observations are directly excluded from the analysis, but here we take in account the presence of tied observations without ignoring any data point.

If a semi-parametric smoothing component is included in the model, we will model it using second order random walk, i.e. $\Gamma_q \sim \mathrm{RW}_2(\sigma_q^2)$, where the standard deviation σ_q controls its smoothness. However, because here we are using partial likelihood for inference, a regular RW_2 component will not be identifiable from the model. Since shifting the Γ_q up or down by any constant will result in the same partial likelihood value. To make this model identifiable, a linear constraint is required to be applied on $\Gamma_q = \left\{\Gamma_{q1}, \ldots, \Gamma_{qm_q}\right\}$. Though other choices are possible, we will choose to constraint the semi-parametric effects Γ_q by setting $\Gamma_{ql_q} = 0$ for the reason of interpretability, where the index l_q represents the mean observed value of the covariate U_q. In this case, the estimated effect for each component in Γ_q can be interpreted as relative to the effect at Γ_q.

4 Examples

We will illustrate the accuracy of our method over INLA, and the ability of our method to yield full posterior uncertainty for the smoothness of the semi-parametric effect and between subject standard deviation, through a simulation study and two real data analysis examples.

4.1 Simulation study

To illustrate the accuracy of our method over INLA when the smoothness assumption for baseline hazard function is violated, we performed this simulation study. In this simulation study, we generated 400 data points from a oscillating baseline hazard function. The plot of that baseline hazard function is shown at below. The risk function for this simulated dataset is $\gamma(u) = 1.5[\sin(0.8x) + 1]$. There are 20 percent of the data points are randomly selected to be censored in this dataset.

For both INLA and our algorithm, the values of x are discretized into 50 bins with equal width, and model the risk function semi-parametrically using a second order random walk, i.e. $\Gamma_1 \sim \text{RW}_2(\sigma^2)$. The prior-distribution of σ is set to be a Penalized Complexity prior such that $\text{P}(\sigma > 2.5) = 0.5$ (Simpson et al., 2017). For INLA, the baseline hazard function also needs to be modelled semi-parametrically, and its default RW_1 method was implemented.

Based on the Figures 1 and at above, it can be seen that our method captures the true risk function very well and outperforms INLA, which has to assume the baseline hazard function in 1 is smooth. The posterior distribution for σ from our method is shown in the figure above as well.

4.2 Leukaemia Data

In this example, we will use our proposed methodology to analyse the Leukaemia data set Martino et al. (2011) did using INLA, and compare these results with the frequentist method GAM. The data set contains 1043 of adult leukaemia patients' information with around 16 percent of them being right-censored.

For our implementation, we are interested in quantifying the relationship between survival rate of leukaemia patients with the age of the patient (age), the count of white blood cells at diagnosis (wbc), the Townsend deprivation index (tpi) and patient's sex (sex). The effects of sex, age and wbc were modelled linearly, and the tpi was modelled as a semi-parametric smoothing effect. The smoothing variable tpi was discretized into 50 equally spaced bins. Our methodology does not require the smoothness assumption on the baseline hazard function nor a specification of it.

We set the prior distributions for all the linear effects β as $\beta \overset{iid}{\sim} \text{N}(0, 0.05^{-1})$, and for the second order random walk of $\Gamma_1 = \{\Gamma_{1,1}, ..., \Gamma_{1,50}\}$ as $\Gamma \sim \text{RW}_2(\sigma^2)$ where a PC prior is put on σ such that $\text{P}(\sigma > 3) = 0.5$ (Simpson et al., 2017). Therefore, the final dimension of W in this example will be 1094. We fit this model using both INLA and our proposed methodology, and then their smoothing results are compared with the result of GAM. For INLA's implementation, it has to assume the smoothness of the baseline hazard and then

model it semi-parametrically using random walk model.

Figure 2 shows the comparison result of these three methods. It can be seen that the smoothing result given by our method is very similar to the one given by GAM, while the result given by INLA seems comparatively more wiggly. This lack of smoothness may be caused by the inappropriate assumption that INLA put on the form of this baseline hazard function, and in this case it becomes a reason to choose our method or GAM which use the partial likelihood function and hence do not require the specification of the baseline hazard function.

Figure 2 shows the posterior distribution of variance parameter given by our algorithm. Based on this plot, it seems like σ's value is likely to be very close to zero, which indicates high smoothness on the smoothing function Γ. This type of model-based quantification of smoothness is a reason to choose Bayesian method over frequentist method such as GAM.

4.3 Kidney Catheter Data

In this example, we implement our proposed methodology to analyse the kidney catheter data set that McGilchrist and Aisbett (1991) analysed using Cox proportional hazard model. This data set contains 76 recurrence times to infection, at the point of insertion of the catheter, for 38 kidney patients. In this data set, each kidney patient has exactly two observations, each observation from one kidney. When the catheter is removed for other reason than infections, the observation is right censored.

In our analysis, we mostly followed the procedures McGilchrist and Aisbett (1991) did in their work, and aimed to quantify the relationship between recurrence time of bladder infection with patient's age, sex and disease types. The variable sex is coded as 1 for male and 2 for female. The four diseases types are represented by three dummy variables GN, AN and PKD, with the reference group being *Other*. A patient level frailty is also added to the model, such that observations within the same patient are correlated.

We specified the prior distributions for all the linear effect β as $\beta \overset{iid}{\sim} \mathrm{N}(0, 0.05^{-1})$, and the prior distribution for the standard deviation of the between-subjects frailty as a PC prior such that $\mathrm{P}(\sigma > 2) = 0.5$. As a comparison, we also implemented INLA and frequentist maximum partial likelihood method for this model. The result is summarized at the figures below.

From figure 3, it can be seen that for the inference of linear effects, the posterior means given by our proposed method are very similar to the frequentist's maximum partial likelihood estimates. While the posterior means given by INLA tends to be less similar to the results of the above two methods. Besides that,

the posterior standard deviations of these linear effects given by our proposed methods are similar to the estimated standard errors given by maximum partial likelihood methods, but the posterior deviations given by INLA tend to be smaller.

As contrast to maximum partial likelihood method, our proposed method is able to give a model-based quantification of the between-subject standard deviation σ. The figure 4 above shows the posterior distribution for the between-subject standard deviation.

5 Discussion

The novel methodology we proposed in this paper provides a flexible way to do approximate Bayesian inference on Cox proportional hazard model with linear effects, semi-parametric smoothing effects and between-groups frailty. This methodology uses partial likelihood hence does not require the smoothness assumption on the baseline hazard function, which is assumed by INLA as it uses the full likelihood instead. It provides model-based uncertainty quantification of the smoothness parameter and between-groups standard deviation as compared to the bootstrapping method used by frequentist method such as GAM. We have demonstrated its accuracy over alternative approaches through the simulation study, and illustrated its model-based uncertainty quantification through the simulation study and the two real data analysis. As long as the inference on baseline hazard function is of secondary interest, our proposed method will be an appealing option to adopt for the analysis of small to median-size data set.

One limitation of our proposed methodology would be its unscalability to data set with massive size. Since the Hessian matrix in our methodology is fully dense and its number of entries increases quadratically with the sample size, the memory cost will become too heavy for our proposed method to be feasible if the sample size is very large. We avoid the computation of this Hessian matrix during the optimization step by implementing a quasi-Newton method that approximates the true Hessian matrix using update of rank 1, but the true Hessian matrix is still required to be evaluated at the maximum to obtain the posterior inferential result.

The framework of this proposed methodology can be easily extended to fit more complex model, by modifying the covariance structure of the covariate with semi-parametric effect. For example, adding a covariate with spatially correlated covariance structure such as simultaneously autoregressive model (SAR) can allow the inclusion of spatial effect into the Cox PH model (Wall, 2004). We will leave these possible extensions to future works.

References

Braun, M. (2014). trustOptim: An R package for trust region optimization with sparse hessians. *Journal of Statistical Software* **60,** 1–16.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* **30,** 89–99.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* **34,** 187–220.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics* **35,** 691–700.

Martino, S., Akerkar, R., and Rue, H. (2011). Approximate bayesian inference for survival models. *Scandinavian Journal of Statistics* **38,** 514–528.

McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics* **47,** 461–466.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **71,** 319 – 392.

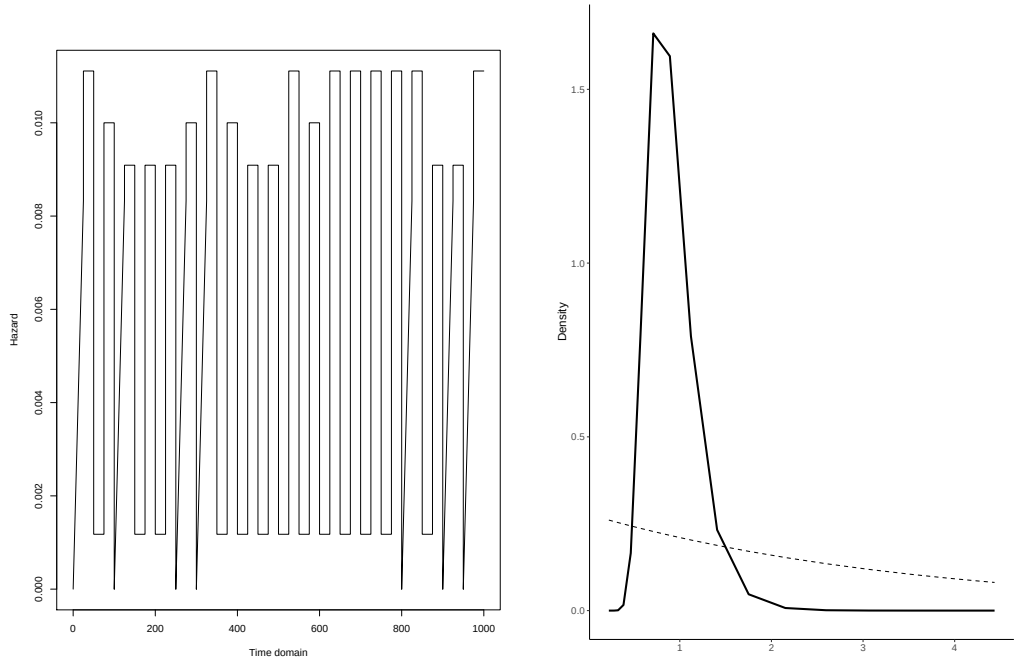
Simpson, D., Rue, H., Martins, T. G., Riebler, A., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science* **32,**.

Stringer, A., Brown, P., and Stafford, J. (2020). Approximate bayesian inference for case crossover models. *Biometrics* .

Tierney, L. and Kadane, J. B. (1986). Accurate approximations to posterior moments and marginal densities. *Journal of the American Statistical Association* **81,**.

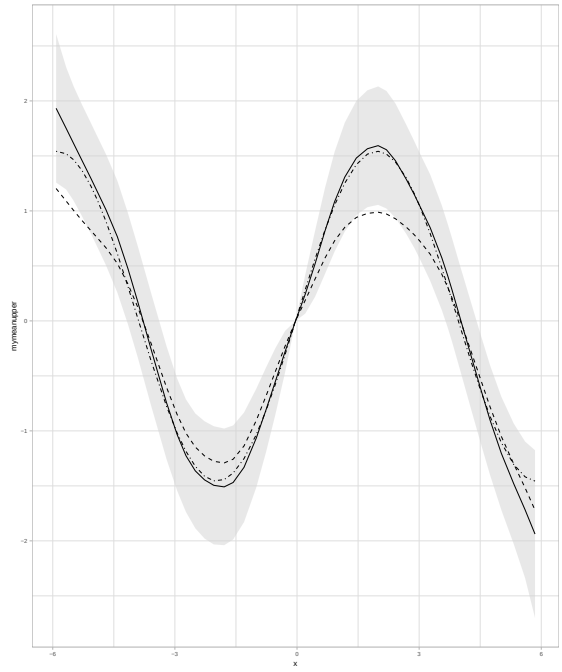
Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16,** 439–454.

Wall, M. M. (2004). A close look at the spatial structure implied by the car and sar models. *Journal of Statistical Planning and Inference* **121,** 311 – 324.



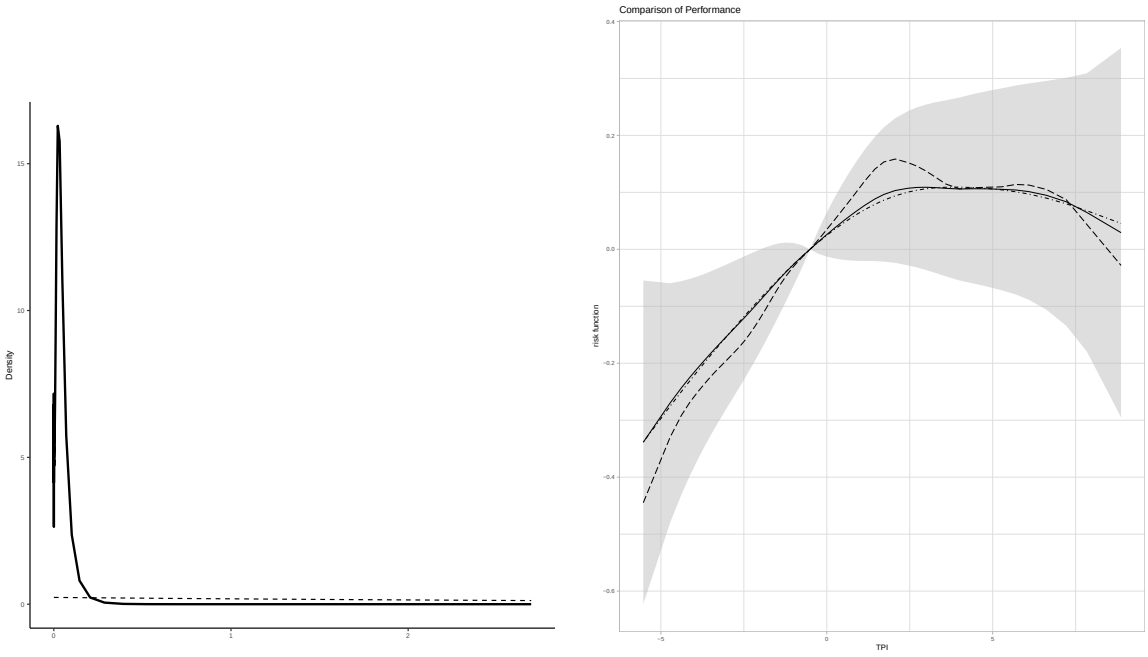
(a) baseline hazard function

(b) Posterior for variance parameter σ



(c) Smoothing result

Figure 1: Baseline hazard function in this simulation (top left panel) Posterior Estimation for variance parameter (—) and its prior (- - -) (top right panels). Bottom panel shows the true risk function (- · -), posterior mean (—) and 95% credible interval using proposed method, posterior mean using INLA (- - -).



(a) Posterior for variance parameter σ

(b) Smoothing result

Figure 2: Posterior Estimation for variance parameter (—) and its prior (- - -) (left panel). Right panel shows the posterior mean (—) and 95% credible interval using proposed method, posterior mean using INLA (- - -) and the smoothing result of GAM (- · -).

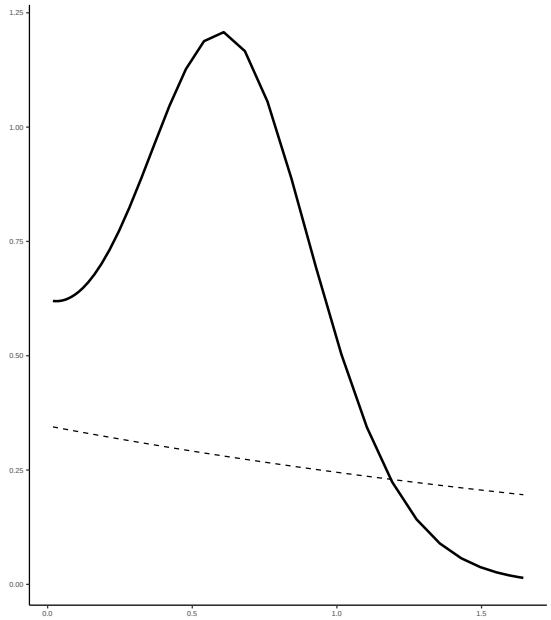
	fixed_Mean	Ours	Coxph	INLA
1	age	0.004632599	0.005180556	0.002291203
2	sex	−1.620617237	−1.678981921	−1.608088118
3	GN	0.170976991	0.180739851	0.117956592
4	AN	0.391823233	0.393639827	0.521933207
5	PKD	−1.167069525	−1.140011592	−1.029157891

(a) Estimated means of linear effects

	fixed_SD	Ours	Coxph	INLA
1	age	0.01444006	0.01472696	0.01295325
2	sex	0.45024348	0.45819377	0.38269245
3	GN	0.52010322	0.53545868	0.46890836
4	AN	0.52152608	0.53683292	0.46273471
5	PKD	0.77747235	0.80987521	0.69859291

(b) Posterior standard deviations of linear effects

Figure 3: Posterior Estimation and Maximum Partial likelihood Estimation for linear effects



(a) Posterior for the between-subject standard deviation

σ

Figure 4: Posterior Estimation for the between-subject standard deviation (—) and its prior (- - -) (left panel)