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

Ziang Zhang, Alex Stringer, Patrick Brown, Jamie Stafford

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. The use of Bayesian inference yields 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 use of our method to yield full posterior uncertainty for the smoothness of the semi-parametric effect and the between-subject standard deviation without making assumptions about the baseline hazard. 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. Inference is often conducted using a partial likelihood which does not depend on the baseline hazard, avoiding the need to make assumptions about its form. Further, 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. Application of the INLA methodology to the Cox PH model requires restrictive smoothness assumptions to be made about the smoothness of the baseline hazard (Martino et al., 2011).

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). By applying a strategy similar to INLA they demonstrate up to an order of magnitude improvement in computation time when compared to sampling-based approaches to Bayesian inference. However, the partial likelihood they consider is simpler than that of the Cox PH model.

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 where the assumption of a smooth baseline hazard is violated. Through two data analysis examples we demonstrate the use of our method for model-based estimation and uncertainty quantification for the smoothness of effects and between-subject standard deviations, without making assumptions about the baseline hazard.

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:

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

Add to this section: notation for the observations that accomodates censoring. Redefine an "observation" to be the pair (Y_{ij}, d_{ij}). Add a note that "ties are handled according to the method of Breslow (cite)". Also, I think we should denote the total number of datapoints as $N = \sum_{i=1}^{n} n_i$, and define this here. 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 partial likelihood can be written as follows: I removed all extra text from this paragraph. The censoring notation should already be defined. Make sure to modify the partial likelihood function appropriately, as discussed.

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

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:

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

Martino et al. (2011) use a data-augmentation trick to write their likelihood in the form (6), but do not use the partial likelihood. Their method requires assumptions about the smoothness of the baseline hazard.

2.2 Approximate Bayesian Inference

To perform Bayesian inference for this model, we specify prior distributions for all unknowns. 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 $\xi = \{\xi_i, i = 1 \dots n\}$ are given independent Gaussian priors $\xi_i|\theta \overset{iid}{\sim} \mathrm{N}(0, \sigma_\xi), i = 1, \dots, n$ where σ_ξ is the between-group standard deviation. Let $U_q = \{U_{ql}; l = 1, ..., m_q\}$ be the ordered vector of *unique* values of covariate $u_q, q = 1, \dots, r$; 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 $\gamma_q, q = 1, \dots, r$, 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 $\Gamma_q = \left\{\Gamma_{q1}, \dots, \Gamma_{qm_q}\right\}$ 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). Let $\Gamma = (\Gamma_1, \dots, \Gamma_r)$; we have that $\Gamma|\sigma_1, \dots, \sigma_q \sim \mathrm{N}\left(0, \Sigma_\Gamma^{-1}\right)$ with $\Sigma_\Gamma^{-1} = \mathrm{diag}\left[\Sigma_1^{-1}(\sigma_1), \dots, \Sigma_r^{-1}(\sigma_r)\right]$. Finally, define the variance parameter vector $\theta = (\theta_0, \dots, \theta_r)$ where $\theta_q = -2\log\sigma_q, q = 1, \dots, r$, 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 $\exp(2), \dots, \exp(14)$ 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 for Q_θ is given in §3 and a derivation is given in Web Appendix A. Our main inferential interest is to obtain the marginal posterior distributions of the mean parameters:

$$\pi(W_k|y) = \int \pi(W_i|y, \theta)\pi(\theta|y)d\theta, k = 1, \ldots, m \tag{8}$$

These are used for point estimates and uncertainty quantification of the mean parameters, which often include the effects of primary interest. We are also interested in the joint posterior distributions of the variance parameters:

$$\pi(\theta|y) = \frac{\int \pi(W, y, \theta)dW}{\int\int \pi(W, y, \theta)dW\, d\theta} \tag{9}$$

These are used for point estimates and uncertainty quantification of the variance parameters, and appear as integration weights in (8). Of secondary inference is the joint posterior distribution of the mean parameters:

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

This appears primarily as an intermediate step in the calculation of the marginal posteriors (8).

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

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

To approximate the density,

$$\pi(W|\theta, Y) \propto \exp\left\{-\frac{1}{2}W^T Q_\theta W + \ell\left(\Delta; Y\right)\right\} \tag{13}$$

a second-order Taylor expansion of $\log(W|\theta, Y)$ about $W = \widehat{W}_\theta$ gives a Gaussian approximation:

$$\pi(W|\theta, Y) \approx \tilde{\pi}_G(W|y, \theta) \propto \exp\left\{-\frac{1}{2}\left(W - \widehat{W}_\theta\right)^T H_\theta\left(\widehat{W}_\theta\right)\left(W - \widehat{W}_\theta\right)\right\} \tag{14}$$

Direct integration of this Gaussian approximation yields a Gaussian approximation for the correnspoding marginal density:

$$\pi(W_k|\theta, Y) = \int \pi(W|\theta, Y)dW_{-k} \approx \tilde{\pi}_G(W_k|y, \theta) \propto \exp\left\{-\frac{1}{2v(\theta)_j^2}\left(W_j - \widehat{W}_{\theta j}\right)^2\right\}, k = 1, \ldots, m \tag{15}$$

For the joint posterior of the variance parameters, the method of Tierney and Kadane (1986) yields a Laplace approximation:

$$\pi(\theta|Y) \approx \tilde{\pi}_{LA}(\theta|y) \propto \pi(\theta)\left\{\frac{|Q_\theta|}{\left|H_\theta\left(\widehat{W}_\theta\right)\right|}\right\}^{1/2} \exp\left\{-\frac{1}{2}\widehat{W}_\theta^T Q_\theta \widehat{W}_\theta + \ell\left(\widehat{\Delta}_\theta; y\right)\right\} \tag{16}$$

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 this section we construct the quantities necessary to implement the approximations (11), with a focus on computational considerations. We describe the novel approach taken to the optimization required to compute $\widehat{W}_\theta$, describing how we address the challenge of a dense $C(W)$.

3.1 Precision Matrix

For fixed design matrices A, B and X, we may write the additive predictor (7) as:

$$\eta = A\Gamma + B\xi + X\beta + \epsilon \tag{17}$$

where $\epsilon \sim \text{N}\left(0, \tau^{-1} I_N\right)$. Now, I think you were right to split A and B. It makes our procedure look more novel. Do you have the corresponding formula for the Q_θ matrix? I have only done this for the no-eps

methodology, in which it is trivial to add new terms. If you don't have it, I think the easiest way to derive it would be to take my old formula for Q_θ from the case crossover paper, replace A with $(A : B)$ and Γ with (Γ, ξ), and expand it out. In §2, it is shown that the partial likelihood function only depends on η through Δ note: this is true for the vectors, but not true for the scalars as you put. Hence the partial likelihood depends on $\Delta = D\eta$ where D is an $(N-1) \times N$-dimensional matrix of rank $N-1$ see my note in section 2 about defining N. In general, whenever you find yourself writing a messy quantity more than once, consider defining a term for it. The precision matrix is given by

$$Q_\theta = \tau \begin{pmatrix} \Lambda^{-1} & -\Lambda^{-1}DA & -\Lambda^{-1}DX \\ -A^T D^T \Lambda^{-1} & \frac{1}{\tau}\Sigma_\Gamma^{-1} + A^T D^T \Lambda^{-1} DA & A^T D^T \Lambda^{-1} DX \\ -X^T D^T \Lambda^{-1} & X^T D^T \Lambda^{-1} DA & \frac{1}{\tau}\Sigma_\beta^{-1} + X^T D^T \Lambda^{-1} DX \end{pmatrix} \tag{18}$$

we need to agree on whether or not to include B and then modify this formula where $\Lambda = DD^T$. Expressions for D and Λ^{-1} are given in Appendix A. The specific form of the partial likelihood and this differencing matrix allow estimation of the frailty coefficients $\xi_i, i = 1, \ldots, n$. In contrast, these are not estimable in the model considered by Stringer et al. (2020).

3.2 Optimization

To compute the conditional mode $\hat{W}(\theta)$, we use trust region optimization (Braun, 2014). The objective function (12) is convex and high-dimensional, and hence trust region methods are well-suited to this problem. The Hessian of the objective function is $H_\theta(W) = Q_\theta + C(W)$. The non-zero block of $C(W)$ is fully dense in the Cox PH model with partial likelihood. The prior precision matrix Q_θ is also dense since Λ^{-1} is. Since this density comes from Λ^{-1}, the Hessian $H_\theta(W)$ inherits the same sparsity pattern as Q_θ.

To apply such methods when the Hessian of the objective function is dense, we utilize quasi-Newton updates inside the trust region procedure. Such updates use a low-rank approximation to $H_\theta(W)$ at each iteration and hence do not require evaluation of this matrix during optimization. While this can lead to more iterations than the method used by Stringer et al. (2020), we are able to optimize our objective function efficiently in the presence of a dense $H_\theta(W)$.

note: I removed the paragraph on ties– see our previous conversation. I think this can be reduced to a simple mention in the preliminaries.

3.3 Models for latent variables

We use a second-order random walk (RW2) model for each $\Gamma_q, q = 1 \dots r$ (Lindgren and Rue, 2008). These models usually contain an intercept β_0 and a *sum-to-zero* constraint $\sum_{q=1}^{r} \Gamma_q = 0$, for identifiability of parameters. However, as in Stringer et al. (2020), an intercept itself is not identifiable when using the partial likelihood for inference, and the sum-to-zero constraint may not be applied. We instead fit the following modified RW2 model for each $q = 1, \dots, r$:

$$\begin{aligned} \Gamma_{q,l+1} - 2\Gamma_{q,l} + \Gamma_{q,l-1} &\overset{iid}{\sim} \mathrm{N}\left(0, \sigma_q^2\right), \\ \Gamma_{q,a} &= 0, \end{aligned} \tag{19}$$

where $a \in \mathbb{R}$ is some chosen reference value. This parametrization is identifiable under the partial likelihood and gives a clear interpretation of $\Gamma_{q,l}$ as the change in log-risk for an individual with covariate $u_{q,l}$ compared to an individual with covariate $u_{q,a}$. what is the second constraint you use? Two are needed for identifiability in RW2. Is this the part that you fix by adding the vv^T onto the diagonal of the precision matrix?

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 a simulation study. We generated $N = n = 400$ uncorrelated data points from a distribution with hazard function (1). The baseline hazard $h_0(t)$ shown in Figure 1 and the additive predictor is $\eta_i = \gamma(u_i)$ with $\gamma(u) = 1.5[\sin(0.8u) + 1]$. To generate the covariate u we first generate $u_1, \dots, u_n \overset{iid}{\sim} \mathrm{Unif}(0, 1)$ confirm: what did you actually do here?, and then discretized these values into 50 disjoint, evenly-spaced intervals. Further, we randomly censored 80 observations.

We fit a RW2 model using our procedure and the INLA software (Rue et al., 2009). For INLA, we used their default first-order random walk model for the baseline hazard, run under its default settings. This implicitly assumes that $h_0(t)$ is smooth. In contrast, our procedure does not infer $h_0(t)$, and does not make assumptions about its smoothness. For the single variance parameter σ we use an Exponential(λ) prior with

λ chosen such that $\mathbb{P}(\sigma > 2.5) = 0.5$, corresponding to a penalized complexity prior (Simpson et al., 2017).

Figure 1 demonstrates the superior accuracy of our method over INLA when $h_0(t)$ is not smooth. The oscillating baseline hazard could represent a scenario where mortality or morbidity risk varies from day to night, or across days of the week. can you add the following: 1. INLA's posterior for σ, on the same plot as ours; 2. INLA's estimated baseline hazard function, on a fourth plot? I need these to be able to further comment on the differences between INLA and us.

4.2 Leukaemia Data

I think you should take a stab at writing the two data examples sort of like I wrote the above simulation one, and given our discussion where we changed the abstract and clarified the purpose of these two examples. Also, the changes to the plots for the simulation should be reflected in the Leukemia example (basically add their σ posterior, on the same graph as ours. For the Kidney example, you should create the tables in Latex using "tabular" and "table" (there is lots of documentation on this). And what's going on with the σ posterior in that example? The approximation is convex so the left tail shouldn't inflect like that. Also, INLA's posterior should be shown here too.

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 $\beta \overset{iid}{\sim} \mathrm{N}(0, 0.05^{-1})$, and for the second order random walk of $\Gamma_1 = \{\Gamma_{1,1}, ..., \Gamma_{1,50}\}$ as $\Gamma \sim \mathrm{RW}_2(\sigma^2)$ where a PC prior is put on σ such that $\mathrm{P}(\sigma > 3) = 0.5$ (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.

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 Derivation of Precision Matrix

In this seciton we give a brief derivation of the precision matrix Q_θ from Equation ... The derivation is identical to that of Stringer et al. (2020) (Web Appendix C), with a different differencing matrix. The

differencing matrix D is:

$$D = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ & & \ddots & & \\ 1 & & & 0 & -1 \end{pmatrix} \tag{20}$$

Our model specifies:

$$\Gamma|\theta \sim \text{Normal}\left(0, \Sigma_\Gamma\right);\ \beta \sim \text{Normal}\left(0, \Sigma_\beta\right);\ \epsilon \sim \text{Normal}\left(0, \tau^{-1}I\right)$$

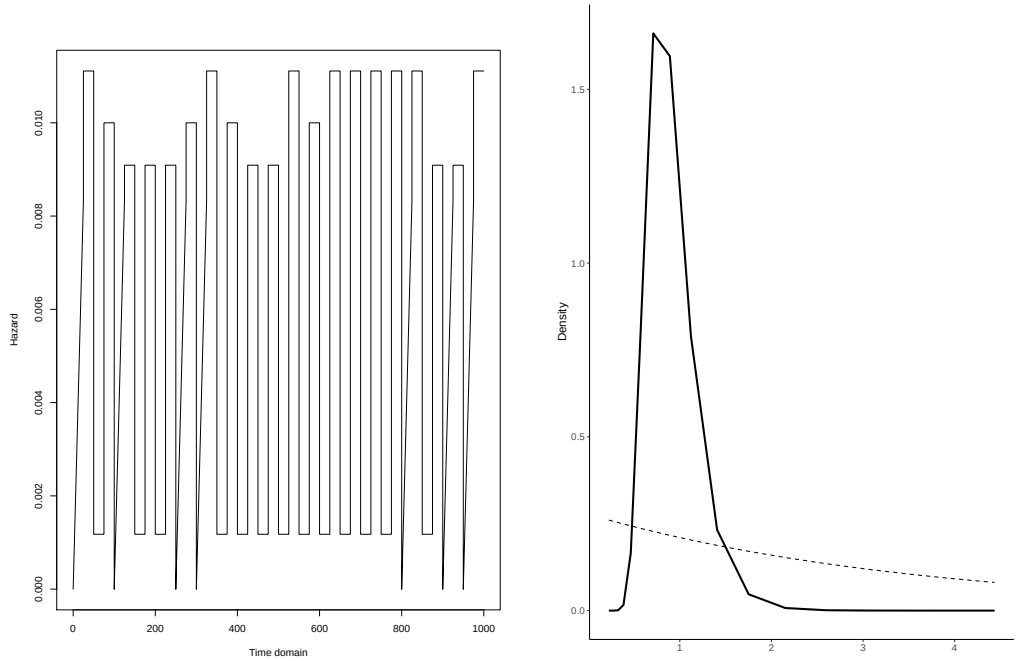
all independent of each other, and of θ unless otherwise specified. The additive predictor is $\eta = A\Gamma + X\beta + \epsilon$ and $\Delta = D\eta$ where D is defined through Equation .... This gives a joint distribution for $W|\theta$:

$$W|\theta = \begin{pmatrix} \Delta \\ \Gamma \\ \beta \end{pmatrix} = \begin{pmatrix} DA & DX & D \\ I & 0 & 0 \\ 0 & I & 0 \end{pmatrix} \begin{pmatrix} \Gamma \\ \beta \\ \epsilon \end{pmatrix} \sim \text{Normal}\left(0, \Sigma\right)$$

where

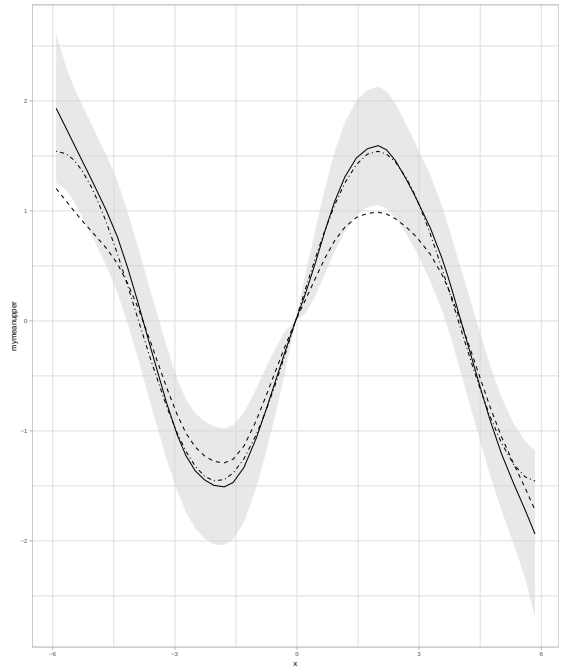
$$\Sigma = \begin{pmatrix} DA\Sigma_\Gamma A^T D^T + DX\Sigma_\beta X^T D^T + \tau^{-1}DD^T & DA\Sigma_\Gamma & DX\Sigma_\beta \\ \Sigma_\Gamma D^T A^T & \Sigma_\Gamma & 0 \\ \Sigma_\beta D^T X^T & 0 & \Sigma_\beta \end{pmatrix}$$

Direct calculation using formuals for block matrix inversion yields $Q(\theta) = \Sigma^{-1}$.



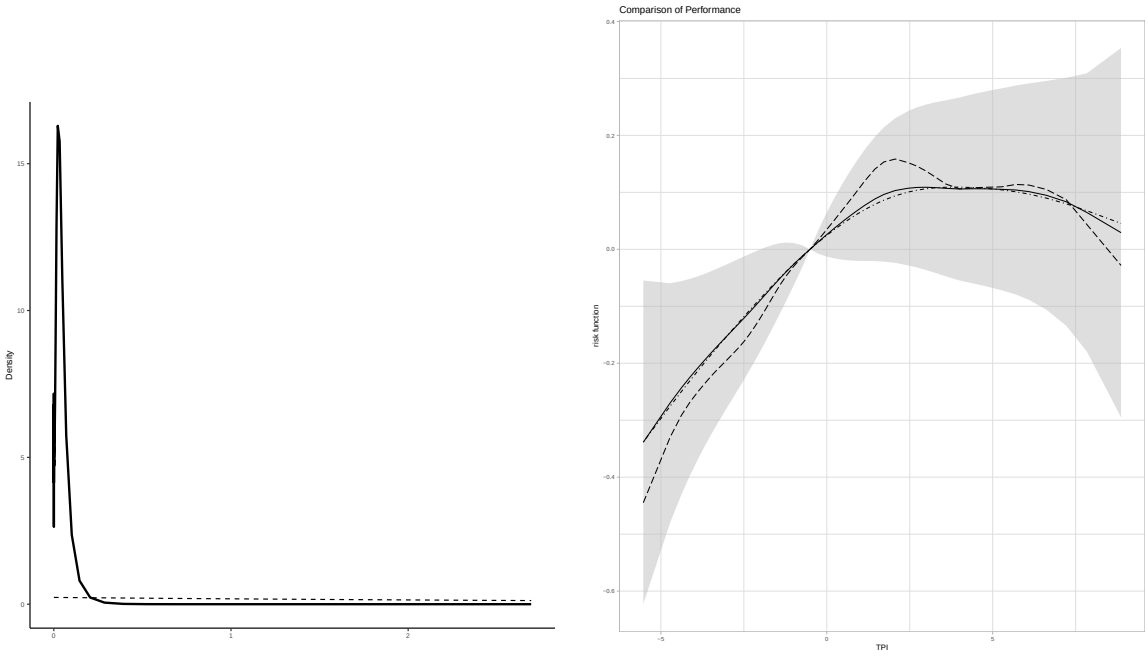
(a) baseline hazard function

(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 (- · -).

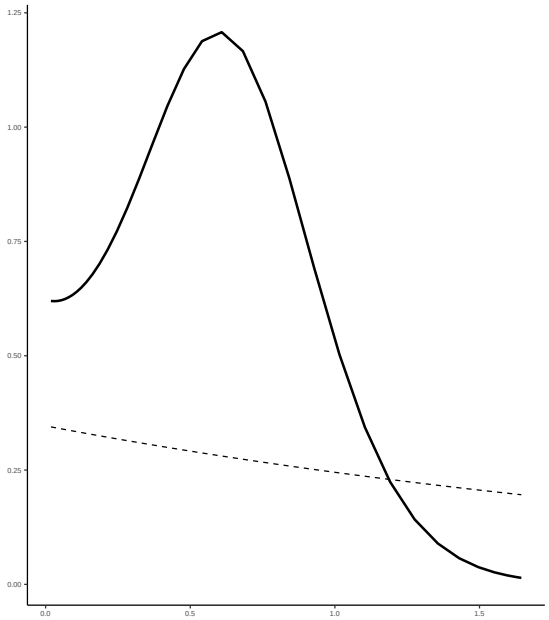
	fixed_Mean	Ours	Coxph	INLA
1	age	0.004632599	0.005180556	0.002291203
2	sex	−1.620617237	−1.678981921	−1.608088118
3	GN	0.170976991	0.180739851	0.117956592
4	AN	0.391823233	0.393639827	0.521933207
5	PKD	−1.167069525	−1.140011592	−1.029157891

(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)