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 introduce 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. 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 ratio depending on unknown linear or smooth covariate effects which are 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 conducted using a partial likelihood which does not require assumptions to be made about the form of the baseline hazard. Further, the use of Bayesian inference with the Cox PH model 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 that our method yields improved inferences over INLA when compared to existing frequentist methods, but retains all the advantages of a fully Bayesian approach.

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:

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

Suppose we observe $i = 1, \ldots, n$ groups each with $j = 1, \ldots, n_i$ observations. For example, we may observe n subjects with n_i measurements per subject. Denote the random variable rep-

resenting the j^{th} survival time in the i^{th} group by Y_{ij}, and denote observations by the pairs $(y, d) = \{(y_{ij}, d_{ij}) : i = 1, \ldots, n; j = 1, \ldots, n_i\}$. Here y_{ij} is the observed time and d_{ij} is an indicator of whether an observation is right-censored. Specifically, $d_{ij} = 1$ if $y_{ij} = Y_{ij}$ and $d_{ij} = 0$ if $Y_{ij} > y_{ij}$. The total number of rows in the data set will be denoted by $N = \sum_{i=1}^{n} n_i$.

Define $h_{ij}(t)$ to be the hazard function for the 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 semi-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* $R_{ij} = \{k, l : y_{kl} \geq y_{ij}\}$. Assuming $y_{ij} \neq y_{kl}$ when $(i, j) \neq (k, l)$, the partial likelihood can be written as follows:

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

where $\Delta_{lk,ij} = \eta_{lk} - \eta_{ij}$. Ties in survival times are handled according to the method of Breslow (Breslow, 1974). 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:

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

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 (4), and hence require restrictive smoothness assumptions to be made about 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 $\tau = \exp(7)$ 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). Note that $\Delta_{lk,ij} = \Delta_{11,ij} - \Delta_{11,lk}$ for every (i, j, l, k). To simplify notation, define $\Delta_{ij} = \Delta_{11,ij}, \Delta_{11} = 0$, and note that the entire partial likelihood (4) depends on η only through $\Delta = \{\Delta_{ij} : i = 1, \dots, n; j = 1, \dots, n_i\}$. 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:

$$\pi(W_s|y) = \int \pi(W_s|y,\theta)\pi(\theta|y)d\theta, s = 1, \ldots, m \tag{8}$$

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 smoothness of effects and between-subject standard deviations, 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:

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

where $\left\{\theta^k, \delta_k\right\}_{k=1}^K$ is a set of nodes and weights corresponding to an appropriate numerical quadrature rule.

The approximations (11) are obtained as follows. For any fixed θ, define

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

For the conditional posterior

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

a second-order Taylor expansion of $\log(W|\theta, Y)$ about $W = \widehat{W}_\theta$ yields 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 corresponding marginal density:

$$\tilde{\pi}_G(W_s|y,\theta) = \int \tilde{\pi}_G(W|y,\theta)dW_{-s} \propto \exp\left\{-\frac{1}{2v^2_{\theta,s}}\left(W_s - \widehat{W}_{\theta s}\right)^2\right\}, s = 1,\ldots,m \tag{15}$$

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

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

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 $\epsilon \sim \text{N}\left(0, \tau^{-1}I_N\right)$. In §2, it is shown that the partial likelihood function only depends on η through Δ. Hence the partial likelihood depends on $\Delta = D\eta$ where D is an $(N-1)\times N$-dimensional matrix of rank

$N-1$. The precision matrix is given by

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

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 and posterior summaries

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. Stringer et al. (2020) also use trust region optimization, and they exploit the sparsity of their Hessian matrix to ensure their procedure is fast and scalable. The Hessian of the objective function is $H_\theta(W) = Q_\theta + C(W)$, and hence inherits its sparsity pattern from that of Q_θ and $C(W)$. In our model, these matrices are dense.

The presence of a dense block in $H_\theta(W)$ presents potential memory challenges when implementing our procedure. To mitigate this we utilize quasi-Newton updates within each iteration of the trust region optimization. Quasi-Newton updates use a low-rank approximation to $H_\theta(W)$ at each iteration and hence do not require evaluation or storage 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 perform the optimization (12) when $H_\theta(W)$ is dense.

The use of quasi-Newton method avoids the need to compute H_θ in iterations of the optimization, but $H_\theta(\widehat{W}_\theta)$ still needs to be evaluated and stored after each optimization. We use $\widehat{W}_\theta$ and $H_\theta(\widehat{W}_\theta)$ to compute the approximations (11), with marginal moments and quantiles computed in an analogous manner.

3.3 Models for latent variables

Our method allows for any jointly-Gaussian model for Γ. In our experiments we implement a second-order random walk (RW2) model for each $\Gamma_q, q = 1 \ldots r$ (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), the intercept itself is not identifiable when using the partial likelihood for inference,

and the sum-to-zero constraint is difficult to interpret in this setting. We instead fit the following modified RW2 model for each $q = 1, \ldots, r$:

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

where $a \in \{1, \ldots, m_q\}$ 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 value $u_{q,l}$ compared to an individual with covariate value $u_{q,a}$.

4 Examples

In this section we present a simulation study and two real data analysis 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.

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 $u_1, \ldots, u_n \overset{iid}{\sim} \mathrm{Unif}(-6, 6)$, and then discretized these values into 50 disjoint, evenly-spaced intervals. Further, we randomly censored 84 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 $\mathbb{P}(\sigma > 2.5) = 0.5$, which is a *penalized complexity* prior of (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, and there are short periods of time where there is no possibility of an event ocurring. The inferred baseline hazard from INLA does not accurately capture the true baseline hazard, and

the inferred covariate effect is too smooth to capture the truth. Our procedure inferrs only the covariate effect, and captures the truth accurately.

4.2 Leukaemia Data

We demonstrate the advantages of our procedure by fitting a semi-parametric Cox PH model to the Leukaemia data set analyzed by Martino et al. (2011). The dataset contains information from 1043 independent adult leukaemia patients, with 16 percent of observations right-censored. We compare our results with INLA and frequentist GAMs. Specifically, we are interested in quantifying the relationship between survival rate of leukaemia patients with the `age` of the patient, the count of white blood cells at diagnosis (`wbc`), the Townsend deprivation index (`tpi`) corresponding to the patient's location, and `sex` of the patient.

The effects of `age`, `wbc` and `sex` were modelled linearly. The `tpi` was discretized into 50 equally spaced bins and modelled as a semi-parametric effect. Prior distributions $\beta \overset{iid}{\sim} \mathrm{N}(0, 0.001^{-1})$, were used for the linear effects. The semi-parametric effects $\Gamma_1 = \{\Gamma_{1,1}, \cdots, \Gamma_{1,50}\}$ were modelled using the RW2 model of §3.3 with the reference constraint $\gamma(0) = 0$. The single variance parameter σ was given an Exponential(λ) prior with λ chosen such that $\mathrm{P}(\sigma > 2) = 0.5$.

Figure 2 shows the results of our procedure compared to INLA and GAM. Our inferred covariate effect closely matches the estimate returned by the GAM, however we provide full model-based posterior uncertainty for σ where the GAM does not. The covariate effect inferred by INLA is less smooth than ours, as a result of assuming that the baseline hazard is smooth. This is reflected both in the shape of the posterior median for Γ_1 and in the posterior for σ.

4.3 Kidney Catheter Data

We use our procedure to fit a Cox PH model to grouped data, providing full posterior undertainty over the between-subject standard deviation. We compare our results to frequentist maximum partial likelihood methods and INLA. The Kidney Catheter dataset contains 76 times to infection, at the point of insertion of the catheter, for $n = 38$ kidney patients. Each patient $i = 1, \ldots, n$ forms a group, and the time to infection of each patient's $n_i = 2$ kidneys represent a survival time. An observation for the survival time of a kidney is censored if the catheter is removed for reasons other than an infection.

We associate survival times with covariates `sex`, `age`, and indicator of one of four types of disease each

patient may have. Subject-specific intercepts $\xi \overset{iid}{\sim} \mathrm{N}(0, \sigma_\xi^2)$ are included to account for correlation between kidneys from the same subject. We use an Exponential(λ) prior distribution for σ_ξ with λ chosen such that $\mathrm{P}(\sigma_{xi} > 2) = 0.5$.

Variables/Reference	Levels	Proposed:Mean/SD	Coxph:Mean/SD	INLA:Mean/SD
age		0.0048/0.015	0.0052/0.015	0.0024/0.013
sex		-1.7/0.46	-1.7/0.46	-1.6/0.38
disease type/Other	GN	0.17/0.53	0.18/0.54	0.11/0.47
	AN	0.39/0.53	0.39/0.54	0.52/0.47
	PKD	−1.2/0.80	−1.1/0.81	−1.1/0.71

Table 1: Estimated means and standard deviations of linear effects by proposed method, Coxph and INLA

Table 1 shows the results of our procedure compared to that obtained using frequentist maximum partial likelihood methods and INLA. Our posterior means and standard deviations for the linear covariate effects are comparable to the frequentist estimates. However, as shown in 3, our method provides full posterior uncertainty for σ_ξ while the frequentist approach does not. INLA gives different estimates for the linear effects and reports lower posterior standard deviations, indicating both potential bias and underestimation of posterior uncertainy. This is also reflected in Figure 3, where our posterior for σ_ξ is wider than that of INLA.

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 Derivation of Precision Matrix

In this section 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:

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

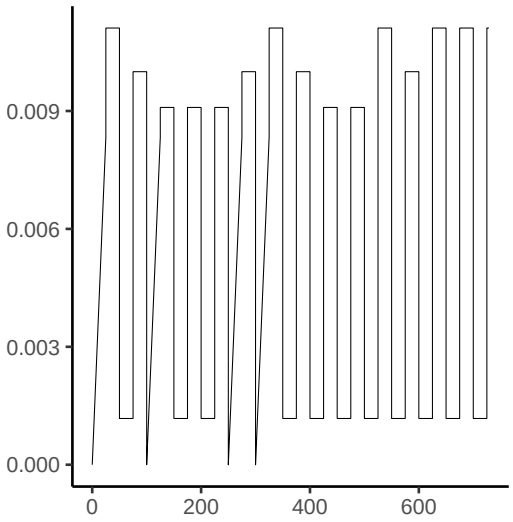
all independent of each other, and of θ unless otherwise specified. The additive predictor is $\eta = A\Gamma + B\xi + X\beta + \epsilon$ and $\Delta = D\eta$ where D is defined through Equation .... This gives a joint distribution for $W|\theta$:

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

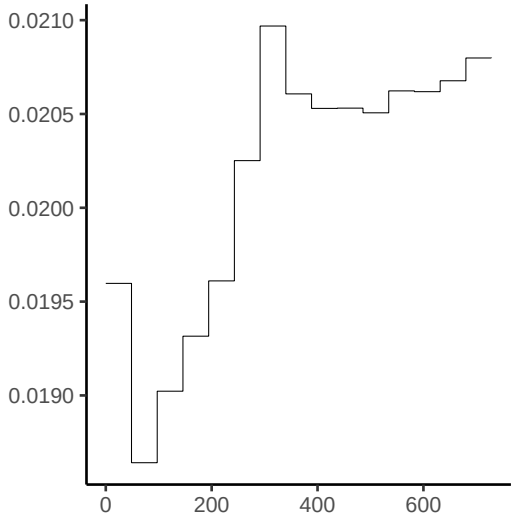
where

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

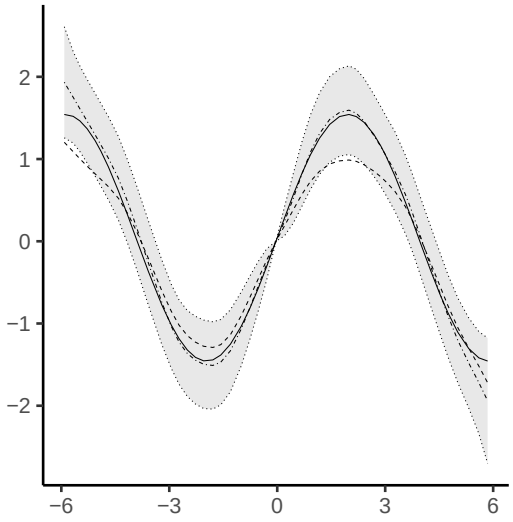
Direct calculation using formulas for block matrix inversion yields $Q(\theta) = \Sigma^{-1}$.



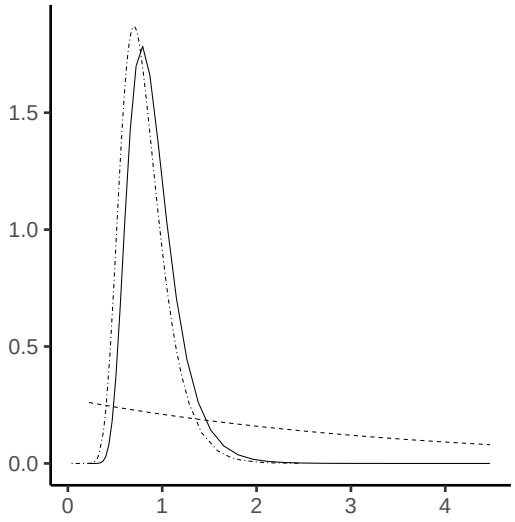
(a) True baseline hazard function



(b) Estimated baseline hazard function (INLA)

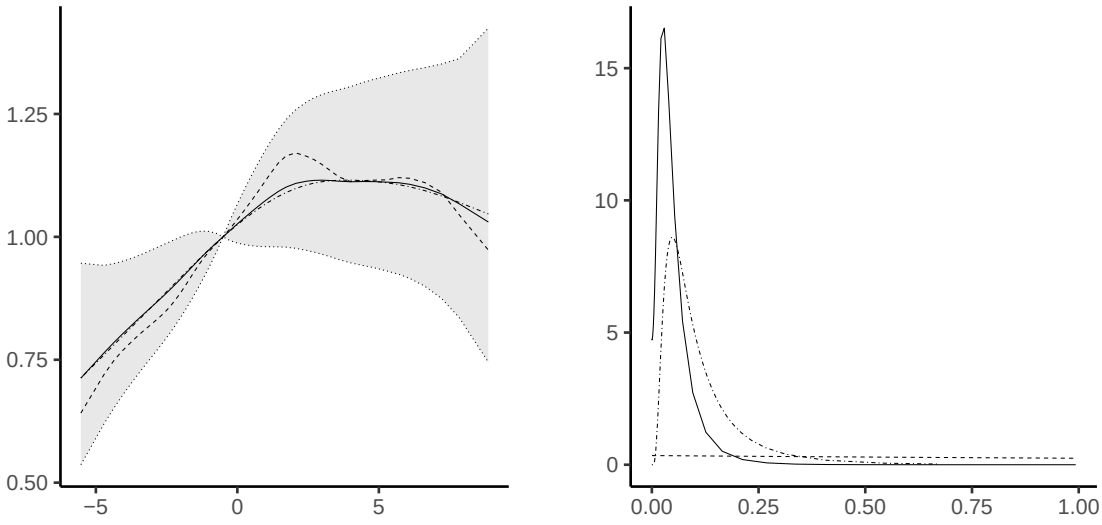


(c) Posterior estimate of covariate effect



(d) Posterior distribution of σ

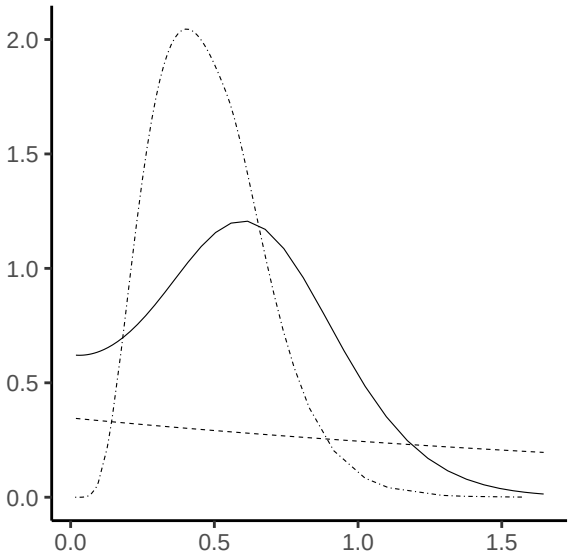
Figure 1: (a): true baseline hazard function of this simulation. (b): estimated baseline hazard function from INLA. (c): true risk function (—); posterior median (- · -) and 95% credible interval ($\cdots$) using proposed method; posterior mean using INLA (- - -). (d): prior (- - -) and approximate posterior distribution for variance parameter by our method (—) and by INLA (- · -).



(a) Posterior effect of `tpi`

(b) Posterior distribution of σ

Figure 2: (a): posterior mean (—) and 95% credible interval ($\cdots$) using our method, posterior mean using INLA (- - -), and the result of fitting a GAM (- · -). (b): prior (- - -) and approximate posterior distribution for σ using our method (—) and INLA (- · -).



(a) Posterior distribution of σ

Figure 3: Posterior Estimation for the between-subject standard deviation by our method (—) and by INLA (- · -), and its prior (- - -)