Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

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

Ziang Zhang[1] | Alex Stringer[1,2] | Patrick Brown[1,2] | James Stafford[1]

[1]Department of Statistical Sciences, University of Toronto, Ontario, Canada
[2]Centre for Global Health Research, St. Michael's Hospital, Ontario, Canada

Correspondence
*Ziang Zhang Email: aguero.zhang@mail.utoronto.ca

Summary

We introduce a novel approximate Bayesian inference methodology for the Cox Proportional Hazard model with partial likelihood, semi-parametric covariate effects and correlated survival times. Existing methods for approximate Bayesian inference for survival models (INLA) require the Hessian matrix of the log-likelihood to be diagonal and are restricted to models with a parametric or smooth semi-parametric baseline hazard. In contrast, the use of partial likelihood in the Cox proportional hazards model does not require assumptions to be made about the baseline hazard, but the Hessian of the log-partial likelihood is fully dense and hence this model cannot be fit using existing methods. We overcome this limitation through the use of quasi-Newton optimization methods, and fit a Cox proportional hazards model with linear covariate effect, semi-parametric covariate effects and correlated survival times using partial likelihood and approximate Bayesian inference. A simulation study demonstrates the superior accuracy of our approximations over INLA when the baseline hazard is not smooth. Analysis of a dataset of Leukaemia survival times and a dataset of Kidney infection times demonstrates the use of our method to yield full posterior estimation and uncertainty quantification for the parameters of interest without making smoothness assumptions about the baseline hazard. An R package implementing our method will be released publicly.

KEYWORDS:
Cox Proportional Hazard Model, Partial Likelihood, Bayesian Inference, Semi-parametric Smoothing

1 | INTRODUCTION

For problems involving time-to-event data, the combination of Cox proportional hazard (Cox PH) models and inference via partial likelihood has been the dominant methodology following its development by Cox[1]. 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 that is conducted via partial likelihood 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 is desirable as this yields model-based estimation and uncertainty quantification for all parameters of interest. However, existing methods for approximate Bayesian inference based on Integrated Nested Laplace Approximations (INLA)[2] cannot be applied to the Cox PH model with partial likelihood because the Hessian matrix of the log partial-likelihood is fully dense while INLA requires this matrix to be diagonal. Application of the INLA methodology to the Cox PH model without partial likelihood has been considered[3], but this requires restrictive smoothness assumptions to be made about the baseline hazard.

Recently, Stringer et al.[4] developed an approximate Bayesian inference methodology for a particular type of survival model which uses partial likelihood, and hence applied the approximation strategy of INLA to a log-likelihood with a non-diagonal Hessian matrix. Their methodology includes semi-parametric covariate effects and yields full posterior uncertainty for the corresponding smoothness parameters, an improvement over existing frequentist methods. However, the partial likelihood they consider is simpler than that of the Cox PH model, and the Hessian matrix of their log-partial likelihood is block-diagonal and sparse. In contrast, the Hessian matrix of log-partial likelihood of Cox PH model is fully dense, so the method of Stringer et. al.[4] does not apply to this model.

In this paper we extend the approximate Bayesian inference methodology of Stringer et al.[4] to the Cox proportional hazard model 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. Our point estimates compare favourably to those based on frequentist generalized additive models (GAMs) which use partial likelihood, in cases where INLA does not compare favourably. However, we retain all the advantages of a fully Bayesian approach and yield full posterior uncertainty estimates for all parameters including those related to the smoothness of the covariate effects.

The remainder of this paper is organized as follows. In §2, we describe the Cox proportional hazard model and the partial likelihood function. In §3, we describe our proposed methodology and how the quasi-Newton method is used to alleviate the computational problems brought by the dense Hessian matrix. In §4 we illustrate our methodology in a simulation study and through the analysis of Leukaemia survival data analysed by Martino et al in 2011[3] and the Kidney catheter data analysed by McGilchrist[5] in 1991. We conclude in §5 with a discussion.

2 | PRELIMINARIES

Suppose we observe n groups indexed by i, each with n_i observations indexed by j. 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 T_{ij}, and denote its realization by t_{ij}. Let c_{ij} denote the censoring time of the survival time T_{ij} which is not directly observable. The observed time is $y_{ij} = \min\{t_{ij}, c_{ij}\}$. Define $d_{ij} = 1$ if $y_{ij} = t_{ij}$ (a survival time) and $d_{ij} = 0$ if $t_{ij} > y_{ij}$ (a censoring time). The observations for each i, j are hence denoted by pairs $y = \left\{(y_{ij}, d_{ij}) : i = 1, \dots, n; j = 1, \dots, n_i\right\}$. 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 T_{ij}. The Cox PH model assumes $h_{ij}(t) = h_0(t)\exp(\eta_{ij})$ where $h_0(t)$ is an unknown baseline hazard function that does not depend on the covariates. An additive predictor η_{ij} links the covariates for the ijth observation to the survival time T_{ij}:

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

Let $\eta = \left\{\eta_{ij} : i = 1, \dots, n; j = 1, \dots, n_i\right\}$ be the vector of all the additive linear predictors. Here x_{ij} is a p-dimensional vector of covariates that are modelled as having linear associations with the log-hazard, and $\beta = (\beta_1, \dots, \beta_p)$ are regression coefficients. The $u_q = \left\{u_{qij} : i = 1, \dots, n; j = 1, \dots, n_i\right\}, q = 1, \dots, r$ are covariate vectors whose association with the log-hazard is modelled semi-parametrically through unknown smooth functions $\gamma_1, \dots, \gamma_r$. The vector of group intercepts $\xi = \left\{\xi_i : i = 1, \dots, n\right\}$, referred to as "frailty" coefficients in the context of survival analysis[6], are included to model correlation between survival times coming from the same group i. There is no global intercept β_0 as this would be absorbed by $h_0(t)$.

Inference is carried out via a partial likelihood function. Define the *risk set* $R_{ij} = \left\{k, l : y_{kl} \geq y_{ij}\right\}$. Assuming $y_{ij} \neq y_{kl}$ when $(i, j) \neq (k, l)$, the partial likelihood can be written as follows:

$$\begin{aligned}\pi(y|\eta) &= \prod_{i=1}^{n}\prod_{j=1}^{n_i}\left\{\frac{\exp[\eta_{ij}]}{\sum_{l,k\in R_{ij}}\exp[\eta_{lk}]}\right\}^{d_{ij}} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{n_i}\left\{\frac{1}{1+\sum_{l,k\in R_{ij},(l,k)\neq(i,j)}\exp[\Delta_{lk,ij}]}\right\}^{d_{ij}}\end{aligned} \tag{2}$$

where $\Delta_{lk,ij} = \eta_{lk} - \eta_{ij}$. Ties in survival times are handled according to the method of Breslow[7]. 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:

$$\pi(y|\eta) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \pi(y_{ij}|\eta) \tag{3}$$

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

$$\pi(y|\eta) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \pi(y_{ji}|\eta_{ij}), \tag{4}$$

Martino et al.[3] are able to write the likelihood for their Cox PH model in the form (4) because they do not use the partial likelihood (2). Because of this, they require restrictive smoothness assumptions to be made about the baseline hazard.

3 | METHODOLOGY

3.1 | Approximate Bayesian Inference

To perform Bayesian inference we require prior distributions for all unknown quantities. A joint Gaussian prior distribution with fixed covariance matrix is used for the regression coefficients $\beta \sim \mathrm{N}(0, \Sigma_\beta)$. The group intercepts $\xi = \left\{\xi_i, i = 1 \dots n\right\}$ are given independent Gaussian priors $\xi_i|\sigma_\xi \overset{iid}{\sim} \mathrm{N}(0, \sigma_\xi), i = 1, \dots, n$ where the between-group standard deviation σ_ξ is treated as an unknown parameter. Let $U_q = \{U_{ql}; l = 1, ..., m_q\}$ be the ordered vector of distinct 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 $\gamma_q(U_{ql}) = \Gamma_{ql}$. We define the vectors of function values $\Gamma_q = \left\{\Gamma_{q1}, \dots, \Gamma_{qm_q}\right\}$ and these are given independent Gaussian distributions $\Gamma_q|\sigma_q \sim \mathrm{N}\left[0, \Sigma_q(\sigma_q)\right]$ for each $q = 1, \dots, m_q$. These distributions are parametrized through their precision matrices $\Sigma_q(\sigma_q)$ which depend on variance parameter σ_q that controls the smoothness of γ_q. A popular choice which we adopt in our analysis is the second-order random walk model[8]. We define $\Gamma = (\Gamma_1, \dots, \Gamma_r)$ and write $\Gamma|\sigma_1, \dots, \sigma_q \sim \mathrm{N}\left(0, \Sigma_\Gamma^{-1}\right)$ 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.[2] and Stringer et al.[4] to add a small random noise on the additive predictor, redefining:

$$\eta_{ij} = x_{ij}^T\beta + \sum_{q=1}^{r} \gamma_q(u_{q_{ij}}) + \xi_i + \epsilon_{ij} \tag{5}$$

where $\epsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0, \tau^{-1})$ for some large, fixed τ. The addition of these ϵ_{ij} is to make the joint distribution of $(\Delta, \Gamma, \beta, \xi)$ a non-singular Gaussian distribution. We set $\tau = \exp(7)$ which is well within the broad range of $\exp(2), \dots, \exp(14)$ which Stringer et. al.[4] found to yield virtually identical inferences and similar running times. Further redefine $\Delta_{lk,ij} = \eta_{lk} - \eta_{ij}$ in terms of the augmented additive predictors (5). Note that $\Delta_{lk,ij} = \Delta_{11,ij} - \Delta_{11,lk}$ for every (i, j, l, k). To simplify notation, define $\Delta_{ij} = \Delta_{11,ij}$, and note that $\Delta_{11} = 0$. The entire partial likelihood (2) depends on η only through $\Delta = \left\{\Delta_{ij} : i = 1, \dots, n; j = 1, \dots, n_i\right\}$. 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 Appendix A. Our main inferential interest is to obtain the marginal posterior distributions of the mean parameters:

$$\pi(W_s|y) = \int \pi(W_s|y,\theta)\pi(\theta|y)d\theta, s = 1, \dots, m \tag{6}$$

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:

$$\pi(\theta|y) = \frac{\int \pi(W, y, \theta)dW}{\int\int \pi(W, y, \theta)dW d\theta} \tag{7}$$

These are used for point estimates and uncertainty quantification of the variance parameter θ, and appear as integration weights in (6). Of secondary inference is the joint posterior distribution of the mean parameters:

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

This appears primarily as an intermediate step in the calculation of the marginal posteriors (6).

All of the quantities of interest (6) – (8) depend on intractable high-dimensional integrals. Stringer et al.[4] utilize Gaussian and Laplace approximations combined with numerical quadrature to approximate each of these integrals accurately and efficiently. Their approximations take the form:

$$\begin{aligned} \tilde{\pi}(W_s|y) &= \sum_{k=1}^{K} \tilde{\pi}_G(W_s|y,\theta^k)\tilde{\pi}_{LA}(\theta^k|y)\delta_k \\ \tilde{\pi}(W|y) &= \sum_{k=1}^{K} \tilde{\pi}_G(W|y,\theta^k)\tilde{\pi}_{LA}(\theta^k|y)\delta_k \end{aligned} \tag{9}$$

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 $\tilde{\pi}_G(W_s|y, \theta^k)$ is a Gaussian approximation for $\pi(W_s|y, \theta^k)$ and the $\tilde{\pi}_{LA}(\theta^k|y)$ is a Laplace approximation for $\pi(\theta^k|y)$, which we describe at below.

The approximations (9) are obtained as follows. For any fixed θ, define

$$\begin{aligned} \widehat{W}_\theta &= \left(\widehat{\Delta}_\theta, \widehat{\Gamma}_\theta, \widehat{\beta}, \widehat{\xi}_\theta\right) = \mathrm{argmax}_W \log \pi(W|\theta, y) \\ H_\theta(W) &= -\frac{\partial^2}{\partial W \partial W^T} \log \pi(W|\theta, y) \\ v_{\theta,s}^2 &= \left[H_\theta\left(\widehat{W}_\theta\right)^{-1}\right]_{ss}, s = 1, \dots, m \end{aligned} \tag{10}$$

For the conditional posterior

$$\pi(W|\theta, y) \propto \exp\left\{-\frac{1}{2}W^T Q_\theta W + \ell(\Delta; Y)\right\}, \tag{11}$$

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

$$\pi(W|\theta, y) \approx \tilde{\pi}_G(W|y, \theta) \propto \exp\left\{-\frac{1}{2}\left(W - \widehat{W}_\theta\right)^T H_\theta\left(\widehat{W}_\theta\right)\left(W - \widehat{W}_\theta\right)\right\} \tag{12}$$

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

$$\tilde{\pi}_G(W_s|y, \theta) = \int \tilde{\pi}_G(W|y, \theta) dW_{-s} \propto \exp\left\{-\frac{1}{2v_{\theta,s}^2}\left(W_s - \widehat{W}_{\theta s}\right)^2\right\}, s = 1, \dots, m \tag{13}$$

For the joint posterior of the variance parameters, the method of Tierney and Kadane[9] yields a Laplace approximation:

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

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 (3), $C(W)$ has a dense structure. In contrast, Rue et al.[2] assume that the likelihood takes the form (4) which enforces the constraint that $C(W)$ is diagonal and hence their method cannot fit the Cox PH model with partial likelihood. Stringer et al.[4] relax this assumption to allow $C(W)$ to have a block-diagonal structure. Our work extends this to permit a fully dense $C(W)$.

3.2 | Precision Matrix

For fixed design matrices A, B and X, we may write the additive predictor (5) as:

$$\eta = A\Gamma + B\xi + X\beta + \epsilon \tag{15}$$

where $\epsilon \sim \mathrm{N}\left(0, \tau^{-1} I_N\right)$. The partial likelihood (2) depends on η only through $\Delta = D\eta$ where D is an $(N-1) \times N$-dimensional matrix of rank $N - 1$. The precision matrix is given by

$$Q_\theta = \tau \begin{pmatrix} \Lambda^{-1} & -\Lambda^{-1} DA & -\Lambda^{-1} DB & -\Lambda^{-1} DX \\ -A^T D^T \Lambda^{-1} & \frac{1}{\tau}\Sigma_\Gamma^{-1} + A^T D^T \Lambda^{-1} DA & A^T D^T \Lambda^{-1} DB & A^T D^T \Lambda^{-1} DX \\ -B^T D^T \Lambda^{-1} & B^T D^T \Lambda^{-1} DA & \frac{1}{\tau}\Sigma_\xi^{-1} + B^T D^T \Lambda^{-1} DB & B^T D^T \Lambda^{-1} DX \\ -X^T D^T \Lambda^{-1} & X^T D^T \Lambda^{-1} DA & X^T D^T \Lambda^{-1} DB & \frac{1}{\tau}\Sigma_\beta^{-1} + X^T D^T \Lambda^{-1} DX \end{pmatrix} \tag{16}$$

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.[4]. A global intercept β_0 is not estimable in our method.

3.3 | Optimization method

To compute the conditional mode $\hat{W}(\theta)$, we use trust region optimization[10]. The objective function (10) is convex and high-dimensional, and hence trust region methods are well-suited to this problem. The Hessian of the objective function is $H_\theta(W) = Q_\theta + C(W)$, and hence inherits its sparsity pattern from those of Q_θ and $C(W)$. The optimization procedure used by Stringer et. al.[4] relies on the sparsity of the Hessian matrix, whereas in our model, $C(W)$ and hence $H_\theta(W)$ is dense.

To overcome this challenge 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.[4], the computation loads brought by the dense Hessian matrix are greatly reduced and we are able to perform the optimization (10) when $H_\theta(W)$ is dense.

Note that $\widehat{W}_\theta$ and $H_\theta(\widehat{W}_\theta)$ are used to compute the approximations (9) and the associated marginal moments and quantiles and hence the dense $H_\theta(\widehat{W}_\theta)$ needs to be stored after each optimization. The total number of Hessian matrices that needs to be evaluated and stored equals to the number of quadrature points being used for the approximations (9).

3.4 | 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 $\Gamma_q, q = 1 \ldots r$[8]. 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.[4], 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 $q = 1, \ldots, r$:

$$\begin{aligned} \Gamma_{q,l+1} - 2\Gamma_{q,l} + \Gamma_{q,l-1} &\overset{iid}{\sim} \mathrm{N}\left(0, \sigma_q^2\right), \\ \Gamma_{q,a} &= 0, \end{aligned} \tag{17}$$

where $a \in \left\{1, \ldots, m_q\right\}$ 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

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 $n = 400$ uncorrelated data points from a distribution with baseline hazard $h_0(t)$ shown in Figure 1 and additive predictor $\eta_i = \gamma(u_i)$ with $\gamma(u) = 1.5[\sin(0.8u) + 1]$. The covariate is obtained by generating $u_1, \dots, u_n \overset{iid}{\sim} \text{Unif}(-6, 6)$, and discretizing these values into 50 disjoint, evenly-spaced bins. Further, we randomly censored 84 observations. For the single variance parameter σ that controls the smoothness of γ, 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.[11].

We fit a RW2 model using our procedure and the INLA software[2]. For the latter we used the default 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.

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 occurring. 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 infers 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.[3] The dataset contains information from 1043 independent adult leukaemia patients, with 16 percent of observations right-censored. 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 $\beta \overset{iid}{\sim} \text{N}(0, 0.001^{-1})$, 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 §3.4 with the reference constraint $\gamma_1(0) = 0$. The single variance parameter σ_1 was given an Exponential(λ) prior with λ chosen such that $\text{P}(\sigma_1 > 2) = 0.5$.

Figure 2 shows the results of our procedure as well as those of INLA and a frequentist GAM. Our inferred covariate effect closely matches the point 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 more variable than ours, as reflected by both the shape of the posterior mean for Γ_1 and the width of the posterior for σ_1.

4.3 | Kidney Catheter Data

The Kidney Catheter dataset contains 76 times to infection, at the point of insertion of the catheter, for $n = 38$ kidney patients. Each patient $i = 1, \dots, n$ 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 use our procedure to fit a Cox PH model to these grouped data, providing full posterior undertainty over the between-subject standard deviation.

We associate survival times with covariates `sex`, `age`, and indicator of one of four types of pre-existing disease each patient may have. Subject-specific intercepts $\xi_i \overset{iid}{\sim} \mathrm{N}(0, \sigma_\xi^2)$ 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	âĹŠ1.2/0.80	âĹŠ1.1/0.81	âĹŠ1.1/0.71

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 Figure 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. 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 covariate effects and correlated observations. The use of partial likelihood does not require the smoothness assumption on the baseline hazard function, an advantage over INLA. The use of

Bayesian inference provides model-based uncertainty quantification of all parameters of interest, an advantage over the frequentist GAM approach. We have demonstrated the advantages of our new approach over alternative approaches through a simulation study and two data analysess. Our proposed method is an appealing option to adopt for the analysis of time-to-event data.

One limitation of our proposed methodology is the manner in which it scales with the sample size N. Since the Hessian matrix in our methodology is fully dense, its number of non-zero entries increases as $O(N^2)$. The scalability of our procedure is limited by the need to store this matrix in memory. We avoid the computation of this Hessian matrix during the optimization step by using a quasi-Newton method, however the true Hessian matrix is still required to be evaluated and stored at the maximum to compute the posterior approximations that we use.

The framework of this proposed methodology can be extended to fit more complex models, by modifying the covariance structure of the covariate with semi-parametric effect. Temporally- and spatially-correlated survival data may be analyzed through a similar procedure. Because we accommodate the dense Hessian matrix of the log-likelihood, our approach could be extended to approximate Bayesian inference for other models with a dense Hessian matrix. We leave such extensions to future work.

$\square$

APPENDIX

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[4] (Web Appendix C), with a different differencing matrix. The differencing matrix D is:

$$D = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ & & \ddots & & \\ 1 & & & 0 & -1 \end{pmatrix} \tag{A1}$$

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 $W|\theta$:

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

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 $Q(\theta) = \Sigma^{-1}$.

References

1. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 1972; 34(2): 187–220.

2. Rue H, Martino S, Chopin N. Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 2009; 71(2): 319 - 392.

3. Martino S, Akerkar R, Rue H. Approximate Bayesian Inference for Survival Models. *Scandinavian Journal of Statistics* 2011; 38(3): 514-528. doi: 10.1111/j.1467-9469.2010.00715.x

4. Stringer A, Brown P, Stafford J. Approximate Bayesian Inference for Case-Crossover Models. *Biometrics* 2020; To appear.

5. McGilchrist CA, Aisbett CW. Regression with Frailty in Survival Analysis. *Biometrics* 1991; 47(2): 461–466.

6. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 1979; 16(3): 439–454. doi: 10.2307/2061224

7. Breslow N. Covariance Analysis of Censored Survival Data. *Biometrics* 1974; 30(1): 89–99.

8. Lindgren F, Rue H. On the Second-Order Random Walk Model for Irregular Locations. *Scandinavian Journal of Statistics* 2008; 35(4): 691–700.

9. Tierney L, Kadane JB. Accurate approximations to posterior moments and marginal densities. *Journal of the American Statistical Association* 1986; 81(393).

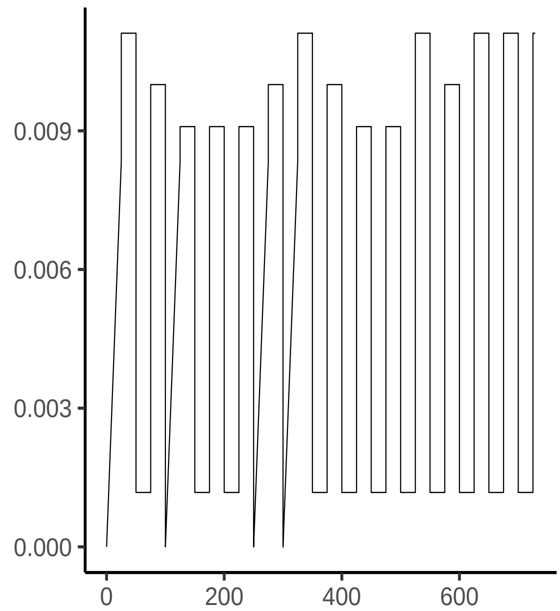
10. Braun M. trustOptim: An R Package for Trust Region Optimization with Sparse Hessians. *Journal of Statistical Software* 2014; 60(4): 1–16.

11. Simpson D, Rue H, Martins TG, Riebler A, SÃÿrbye SH. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science* 2017; 32(1).

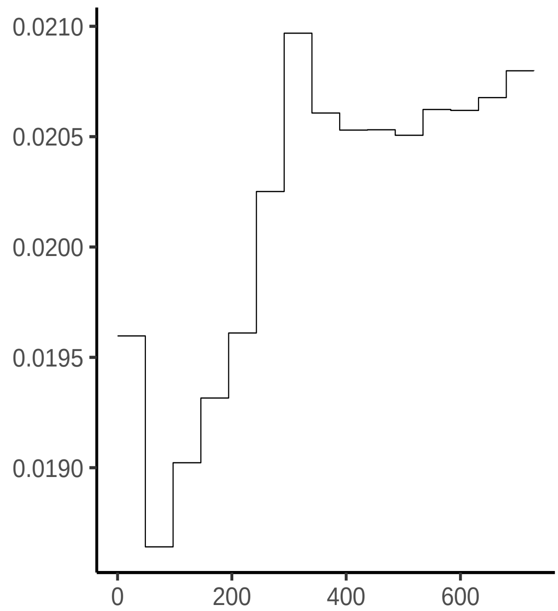
12. Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* 2016; 111(516): 1548–1563.

13. Wall MM. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference* 2004; 121(2): 311 - 324. doi: https://doi.org/10.1016/S0378-3758(03)00111-3

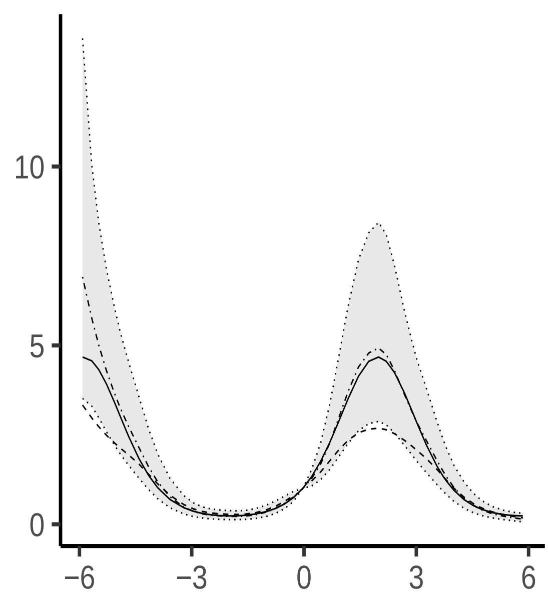
14. Rue H, Martino S. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference* 2007; 137: 3177 - 3192. doi: doi:10.1016/j.jspi.2006.07.016



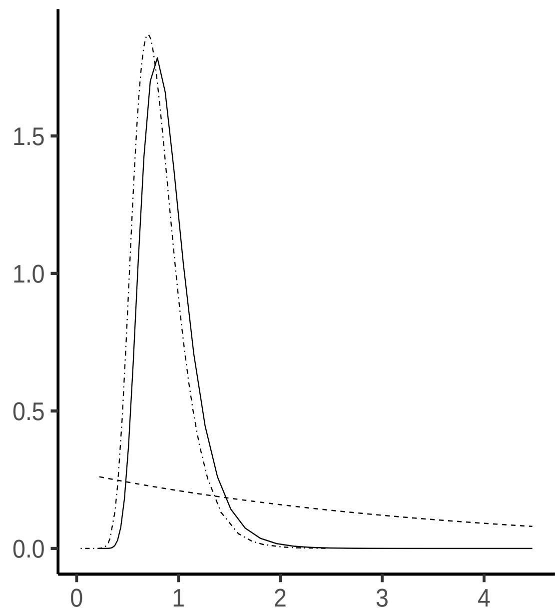
(a) True baseline hazard function



(b) Estimated baseline hazard function (INLA)

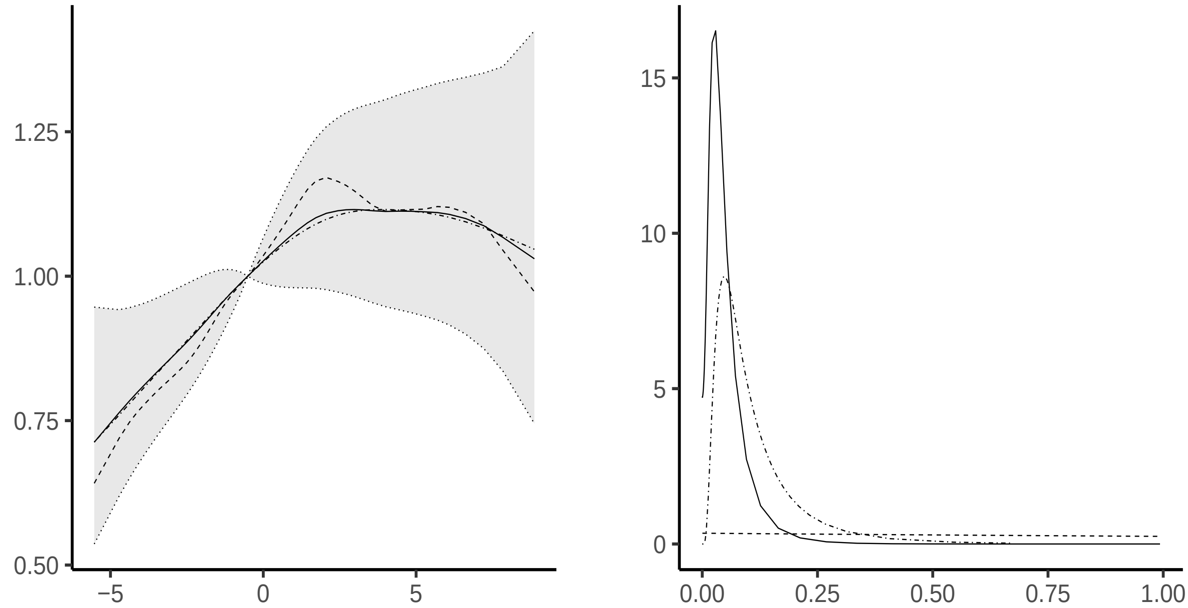


(c) Posterior estimate of covariate effect



(d) Posterior distribution of σ

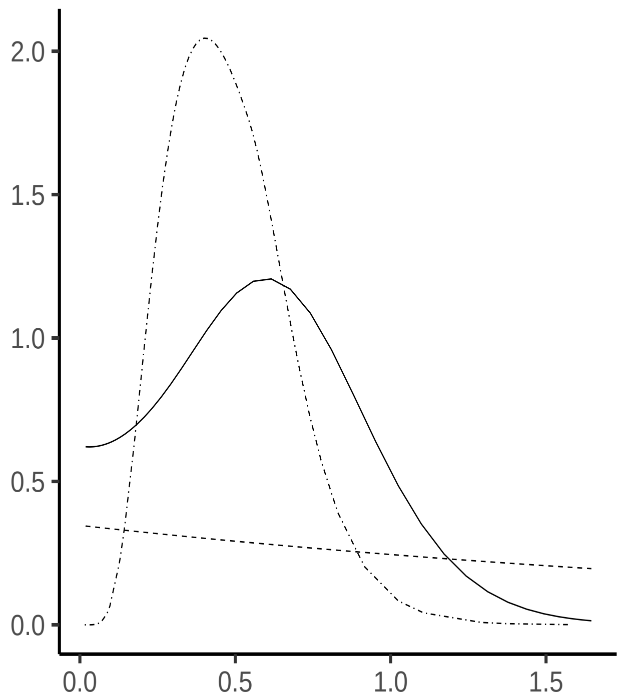
FIGURE 1 (a): true baseline hazard function of this simulation. (b): estimated baseline hazard function from INLA. (c): true risk function (—); posterior median (- · -) and 95% credible interval (···) using proposed method; posterior mean using INLA (- - -). (d): prior (- - -) and approximate posterior 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 (···) 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 (- - -)