Bayesian Analysis (0000) **00**, Number 0, pp. 1

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

Ziang Zhang, Alex Stringer Patrick Brown and James Stafford

Abstract. In this paper, we introduce a novel approximate Bayesian inference methodology for the Cox Proportional Hazard model with partial likelihood that allows the inclusion of linear covariate effects, semi-parametric covariate effects and correlated survival times. The existing method for approximate Bayesian inference for Cox Proportional Hazard model (Integrated Nested Laplace Approximations, INLA), uses full likelihood instead of partial likelihood, and hence requires parametric or a smooth semi-parametric baseline hazard function. Combing the use of a generalization of the Laplace approximation method used in the existing methods with the use of quasi-Newton type of optimization, we propose a way to do approximate Bayesian inference for Cox Proportional Hazard model using partial likelihood, that does not require any assumption on the form of the base hazard function. From a few simulation studies with different types of underlying baseline hazard function, We show that our proposed method provides accurate inference even when the sample size is relatively small. We then utilize our proposed method to re-analyse the dataset of Leukaemia survival times and the dataset of Kidney infection times. An R package implementing our method will be released publicly.

Keywords: Cox Proportional Hazard Model, Partial Likelihood, Approximate 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 (Cox, 1972). The Cox PH model assumes that any two subjects' event hazards are proportional as a function of time, with the 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) (Rue

[*]

DOI: 0000

et al., 2009) 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 (Martino et al., 2011), but this requires restrictive smoothness assumptions to be made about the baseline hazard.

Recently, Stringer et al. (2020) developed an approximate Bayesian inference methodology for case-crossover models, which applies the approximation strategy of INLA to a log-partial 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. (2020) does not apply to this model.

In this paper we extend the approximate Bayesian inference methodology of Stringer et al. (2020) to the Cox proportional hazard model with partial likelihood. Our methodology accommodates semi-parametric smoothing effects and correlation between observed survival times. The remainder of this paper is organized as follows. In §2, we describe our proposed methodology and how the quasi-Newton method is used to mitigate the computational problems brought by the dense Hessian matrix. In §3 we illustrate our methodology in a simulation study and through the analysis of Leukaemia survival data analysed by Martino et al. (2011) and the Kidney catheter data analysed by McGilchrist and Aisbett (1991). We conclude in §4 with a discussion.

2 Methods

2.1 A latent Gaussian Cox PH Model

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 for observation ij, if $c_{ij} < T_{ij}$, then T_{ij} 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 $y = \{(y_{ij}, d_{ij}) : i = 1, \dots, n; j = 1, \dots, n_i\}$. 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}:

$$\begin{aligned} \eta_{ij} &= x_{ij}^T\beta + \sum_{q=1}^{r} \gamma_q(u_{qij}) + \xi_i \\ \xi_i &\overset{iid}{\sim} \mathcal{N}(0, \sigma_\xi) \\ \gamma_q(\cdot) &\overset{ind}{\sim} \mathcal{GP}(\sigma_q) \end{aligned} \tag{1}$$

Let $\eta = \{\eta_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$ 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 $\beta = (\beta_1, \ldots, \beta_p)$ 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. There is no global intercept β_0 as this would be absorbed by $h_0(t)$.

2.2 Modelling Semi-parametric covariate effect

The semi-parametric covariate effect γ_q are modelled as Gaussian processes, each with a standard deviation parameter σ_q controlling its covariance function. A typical choice for the Gaussian process is the random walk of order 2 (Lindgren and Rue, 2008), which has similarities with cubic spline functions. The model above is a Latent Gaussian Model (LGM), as ξ_i and γ_q are normally distributed and a Normal prior is specified for the regression coefficients β. The Gaussian assumption for ξ_i and γ_q is essential for the inference methodology in the next section.

Let $U_q = \{U_{ql}; l = 1, ..., m_q\}$ be the ordered vector of distinct values of covariate $u_q, q = 1, \ldots, r$; often these values are set by the user by discretizing the covariate u_q into m_q pre-specifed bins. To infer the infinite-dimensional parameters $\gamma_q, q = 1, \ldots, r$, we approximate each by a piecewise constant function with jumps at the U_{ql}, which we denote as $\gamma_q(U_{ql}) = \Gamma_{ql}$. We define the vectors of function values $\Gamma_q = \left\{\Gamma_{q1}, \ldots, \Gamma_{qm_q}\right\}$ and these are have Gaussian distributions $\Gamma_q|\sigma_q \sim \mathcal{N}\left[0, \Sigma_q(\sigma_q)\right]$ for each $q = 1, \ldots, m_q$. These distributions are parametrized through their precision matrices $\Sigma_q(\sigma_q)$ corresponding to the specific Gaussian processes chosen, which depend on parameters σ_q. We define $\Gamma = (\Gamma_1, \ldots, \Gamma_r)$ and write $\Gamma|\sigma_1, \ldots, \sigma_q \sim \mathcal{N}\left(0, \Sigma_\Gamma^{-1}\right)$ with $\Sigma_\Gamma^{-1} = \text{diag}\left[\Sigma_1^{-1}(\sigma_1), \ldots, \Sigma_r^{-1}(\sigma_r)\right]$.

These models usually contain an intercept β_0 and a *sum-to-zero* constraint $\sum_{l=1}^{m_q} \Gamma_{ql} = 0$, for identifiability of parameters. However, β_0 is not identifiable when using the partial likelihood for inference, and hence the sum-to-zero constraint is difficult to interpret in this setting. We fit the following modified RW2 model for each $q = 1, \ldots, r$:

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

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 $u_{q,a}$.

Finally, define the variance parameter vector $\theta = (\theta_0, \ldots, \theta_r)$ where $\theta_q = -2 \log \sigma_q, q = 1, \ldots, r$, and $\theta_0 = -2 \log \sigma_\xi$. In other words, the variance parameter θ are defined to be the log precision of the RW2 increment for convenience. The variance parameters are given prior distribution $\theta \sim \pi(\theta)$.

2.3 Approximate Bayesian Inference

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:

$$\begin{aligned} \pi(y|\eta) &= \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left\{ \frac{\exp[\eta_{ij}]}{\sum_{l,k \in R_{ij}} \exp[\eta_{lk}]} \right\}^{d_{ij}} \\ &= \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left\{ \frac{1}{1 + \sum_{l,k \in R_{ij}, (l,k) \neq (i,j)} \exp[\Delta_{lk,ij}]} \right\}^{d_{ij}} \end{aligned} \tag{3}$$

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

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

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

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

Martino et al. (2011) are able to write the likelihood for their Cox PH model in the form (5) using the full, not partial likelihood (3). Because of this, they require assumptions to be made about the baseline hazard.

Further redefine $\Delta_{lk,ij} = \eta_{lk} - \eta_{ij}$ in terms of the additive predictors (1). 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 (3) depends on η only through $\Delta = \{\Delta_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$. 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 *latent parameters* and let $\dim(W) = m$. Since Δ is a function of Γ, β, ξ, the joint distribution of W has a singular precision matrix. Furthermore, the use of improper prior such as RW2 for γ also causes the precision matrix Σ_Γ to be singular. However, the Laplace approximation for θ is not applicable when the precision matrix of prior is singular. Therefore, in order to carry out appropriate Laplace approximation for the posterior of θ, we follow Rue et al. (2009) and Stringer et al. (2020) to add a small random noise ϵ_{ij} on Δ_{ij}, where $\epsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0, \tau^{-1})$ for some large, fixed τ. The addition of these ϵ_{ij} gives the joint distribution of $(\Delta, \Gamma, \beta, \xi)$ a non-singular precision matrix, and enables the use of improper prior in the model specification. We set $\tau = \exp(7)$ which is well within the broad range of $\exp(2), \ldots, \exp(14)$ which Stringer et al. (2020) found to yield very similar inferences and running times.

Our model specifies $W|\theta \sim \mathrm{N}\left[0, Q_\theta^{-1}\right]$. An expression for Q_θ is given in §2 and a derivation is given in Appendix A. Our main inferential interest is to obtain the marginal posterior distributions of the latent parameters:

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

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

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

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 latent 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. (2020) utilize Gaussian and Laplace approximations combined with numerical quadrature to approximate each of these integrals accurately and efficiently. Their approximations take the form:

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

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 computed as follows. For any fixed θ, define

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

For the conditional posterior

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

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:

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

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

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

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 (4), $C(W)$ has a dense structure. In contrast, Rue et al. (2009) assume that the likelihood takes the form (5) 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. (2020) relax this assumption to allow $C(W)$ to have a block-diagonal structure. Our work extends this to permit a fully dense $C(W)$.

For fixed design matrices A, B and X, we may write the additive predictor η and the difference additive Δ as:

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

where $\epsilon \sim \mathrm{N}\left(0, \tau^{-1} I_N\right)$ and D is an $(N-1) \times N$-dimensional matrix of rank $N-1$. The partial likelihood (3) depends on η only through Δ. Note that because of this, a global intercept β_0 is not estimable when using partial likelihood. The precision matrix is given by

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

Expressions for D and the derivation of this precision matrix are given in Appendix A.

2.4 Optimization method

To compute the conditional mode $\hat{W}(\theta)$, we use trust region optimization (Braun, 2014). 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 the dense upper-left bock from $C(W)$. We utilize quasi-Newton updates within each iteration of the trust region optimization, which uses 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 increase the number of iterations, 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 Examples

In this section we present a simulation study and two data analysis examples. Code is available in the online supplementary materials.

3.1 Simulation study

To illustrate the accuracy of our proposed approach, we simulate six different datasets, under three baseline hazard functions, and two true covariate effect functions. The three baseline hazard functions are respectively a hazard function constant at 0.03, an regular hazard function that is periodic and continuous, and a complicated hazard function that has three different types of oscillating behaviours and two discontinuities. The two true covariate effect functions are a smooth function $\gamma(u) = \log((u+1)^2) - 1$ with domain $u \in (0, 6)$, and a complicated function $\gamma(u) = 1.5\sin(0.8u)$ with domain $u \in (-5, 5)$.

In all the six simulation settings, we generated 500 data points from the corresponding model $h(t) = h_0(t)\exp(\gamma(u))$, and randomly censored 10 percent of the data points for each setting. The covariate u is generated uniformly from its corresponding domain, and discretized into 50 disjoint, evenly-spaced bins. For the single variance parameter θ that controls the smoothness of the inferred γ, we use an $Exponential(\lambda)$ prior with λ chosen such that $\mathbb{P}(\theta > 2.5) = 0.5$ for the setting when true γ is complicated, and $\mathbb{P}(\theta > 1) = 0.5$ for the setting when true γ is smooth. This is a penalized complexity prior of Simpson et al. (2017). We fit a RW2 model using our proposed approach to infer the true smoothing function γ, with inferred γ constrained to be zero at the median of the observed covariate values of each dataset. The resulting plots are shown in Figure 1, where the covariate effect functions are exponentiated to represent the relative risk function, and the variance parameter θ is transformed back to standard deviation σ.

According to Figure 1, it can be seen that our proposed method gives relative accurate point estimation for the true risk functions regardless of the true baseline hazard function, even with a sample size that is comparatively small. In all of these six simulation examples, our 95 percent credible interval contains the true risk function throughout the domain, with an reasonable interval width at most of the regions in the domain.

3.2 Leukaemia Data

We implemented our proposed procedure to fit a semi-parametric Cox PH model to the Leukaemia data set analysed by Martino et al. (2011). 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, sex and white blood cell count were modelled linearly. Martino et al. (2011) also included a spatial effect in the model, which we ignore for computational reason. The deprivation index was discretized into 50 equally spaced bins and modelled as a semi-parametric effect. Prior distributions $\beta \overset{iid}{\sim} \mathcal{N}(0, 0.001^{-1})$, were used for the regression coefficients. The semi-parametric effects $\Gamma = \{\Gamma_1, \cdots, \Gamma_{50}\}$ were modelled using the RW2 model of §2.2 with the reference constraint $\gamma(0) = 0$. The single variance parameter σ_1 was given an Exponential prior with a prior median of 2.

Figure 3 shows the posterior result of the exponentiated covariate effect of tpi and the corresponding standard deviation σ. Our inferred covariate effect of tpi is similar to the result gotten by Martino et al. (2011), where the risk of death initially grows with the value of tpi with diminishing rate and eventually begins to decrease after tpi reaches around 5. However, the approach utilized by Martino et al. (2011)l relies on the use of full likelihood function with baseline hazard function modelled semi-parametrically. However, our approach implements the approximate Bayesian inference using the partial likelihood, hence has no assumption on the form of the baseline hazard function.

3.3 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 uncertainty 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 prior distribution for σ_ξ with median 2.

Table 1 shows the results of our procedure compared to that obtained using frequentist maximum partial likelihood methods. Our posterior means and standard deviations for the linear covariate effects are comparable to the frequentist estimates. However, as shown in Figure **??**, our method provides full posterior uncertainty for σ_ξ while the frequentist approach does not.

4 Discussion

The methodology we proposed in this paper provides a flexible way to carry out Bayesian inference for Cox proportional hazard models with partial likelihood, that accommodates the inference on linear covariate effects, semi-parametric covariate effects and frailties. The use of partial likelihood does not require any assumption on the baseline hazard function, an advantage over INLA. We have demonstrated the accuracy of our new approach through some simulation studies, even when the sample size is small, and we have implemented our proposed method to analyse two classical datasets in survival analysis. 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. Therefore, our methodology may not be computationally advantageous when the sample size is too large.

Another limitation of our proposed method is that the number of dense Hessian matrices required to be evaluated and stored equals to the number of quadrature points used for the variance parameter θ. With finer grids of quadrature points, the inference will be more accurate especially for the posterior of variance parameter. However, when the dimension of θ becomes large, the number of quadrature points required in the grids to achieve a specified resolution for the inference grows exponentially to the dimension of θ. This makes our methodology infeasible when the number of semi-parametric effects in the model is too large.

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 analysed 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.

Data Availability Statement

The simulated data of example 3.1 are available in the code available with this paper at the Statistics in Medicine website on Wiley Online Library. Data for example 3.2 were obtained from R package "INLA" (Rue et al., 2009) and are freely available. Data for example 3.3 were obtained from R package "survival" (Therneau, 2015) and are freely available.

Appendix A: Derivation of Precision Matrix

In this section we give a brief derivation of the precision matrix Q_θ from Equation (16). The derivation is similar to that of Stringer et al. (2020) (Web Appendix C), with a different differencing matrix. The differencing matrix D is:

$$D = \begin{pmatrix} 1 & -1 & 0 & & 0 \\ 1 & 0 & -1 & & 0 \\ & & & \ddots & \\ 1 & & & 0 & -1 \end{pmatrix} \tag{17}$$

As described in §2 ,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 unless otherwise specified. The vector of additive linear predictors can be written as $\eta = A\Gamma + B\xi + X\beta + \epsilon$ and $\Delta = D\eta$ where D is defined through Equation (17). This gives a joint Gaussian distribution for $W|\theta$ as:

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

where

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

The precision matrix $Q(\theta) = \Sigma^{-1}$ is obtained through direct inversion.

References

Braun, M. (2014). "trustOptim: An R package for trust region optimization with sparse hessians." *Journal of Statistical Software*, 60(4): 1–16. 7

Breslow, N. (1974). "Covariance analysis of censored survival data." *Biometrics*, 30(1): 89–99.
URL http://www.jstor.org/stable/2529620 4

Cox, D. R. (1972). "Regression models and life-tables." *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2): 187–220. 1

Lindgren, F. and Rue, H. (2008). "On the second-order random walk model for irregular locations." *Scandinavian Journal of Statistics*, 35(4): 691–700. 3

Martino, S., Akerkar, R., and Rue, H. (2011). "Approximate Bayesian inference for survival models." *Scandinavian Journal of Statistics*, 38(3): 514–528. 2, 4, 8

McGilchrist, C. A. and Aisbett, C. W. (1991). "Regression with frailty in survival analysis." *Biometrics*, 47(2): 461–466. 2

Rue, H. and Martino, S. (2007). "Approximate Bayesian inference for hierarchical Gaussian Markov random field models." *Journal of Statistical Planning and Inference*, 137: 3177 – 3192.

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(2): 319 – 392. 1, 5, 6, 10

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(1). 8

Stringer, A., Brown, P., and Stafford, J. (2020). "Approximate Bayesian inference for case-crossover models." *Biometrics*, In press. 2, 5, 6, 10

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. Version 2.38.
URL https://CRAN.R-project.org/package=survival 10

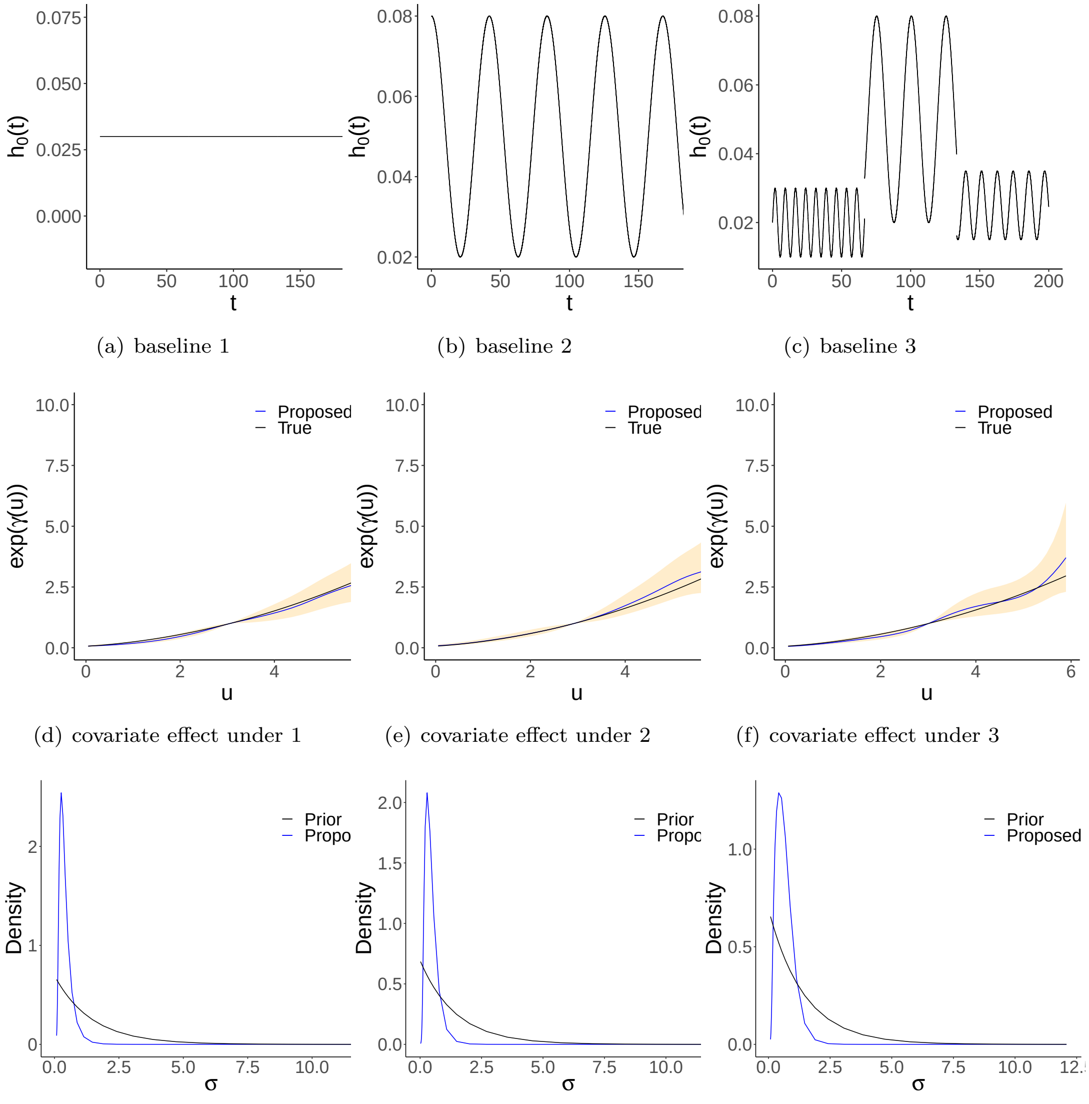
Tierney, L. and Kadane, J. B. (1986). "Accurate approximations to posterior moments and marginal densities." *Journal of the American Statistical Association*, 81(393). 6

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). "The impact of heterogeneity in individual frailty on the dynamics of mortality." *Demography*, 16(3): 439–454. 3

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

		Proposed		Freq PL	
Variables/Reference	Levels	Mean	SD	Mean	SD
Age		0.0048	0.015	0.0052	0.015
Sex		-1.7	0.46	-1.7	0.46
Disease Type/Other	GN	0.17	0.53	0.18	0.54
	AN	0.39	0.53	0.39	0.54
	PKD	-1.2	0.80	-1.1	0.81

Table 1: Estimated means and standard deviations of linear effects by proposed method, frequentist partial likelihood maximization method (Freq PL) for the kidney data in section 3.3.



(a) baseline 1

(b) baseline 2

(c) baseline 3

(d) covariate effect under 1

(e) covariate effect under 2

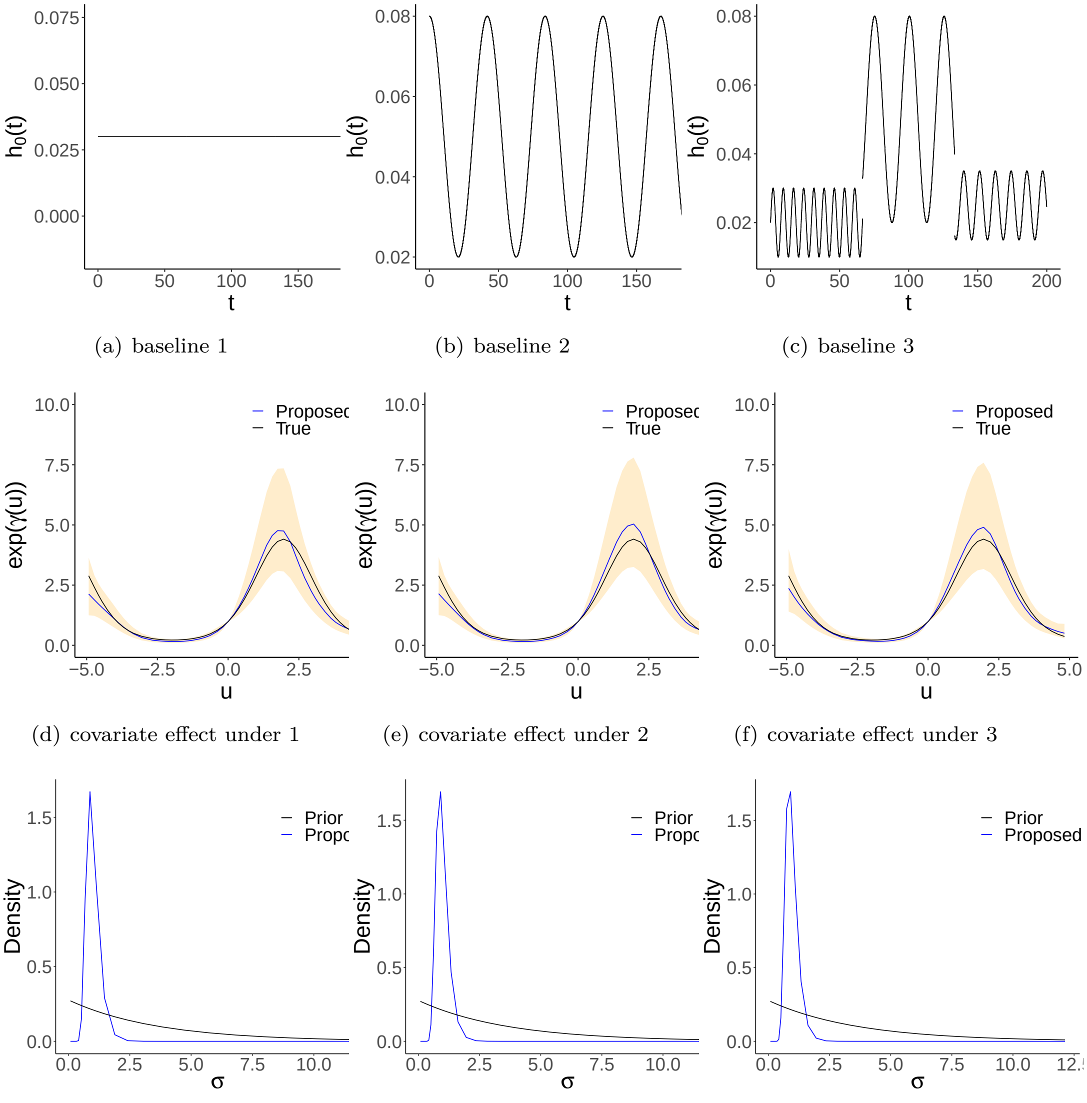
(f) covariate effect under 3

(g) variance parameter under 1

(h) variance parameter under 2

(i) variance parameter under 3

Figure 1: Inferred covariate effects and variance parameters under different baseline hazard functions



(a) baseline 1

(b) baseline 2

(c) baseline 3

(d) covariate effect under 1

(e) covariate effect under 2

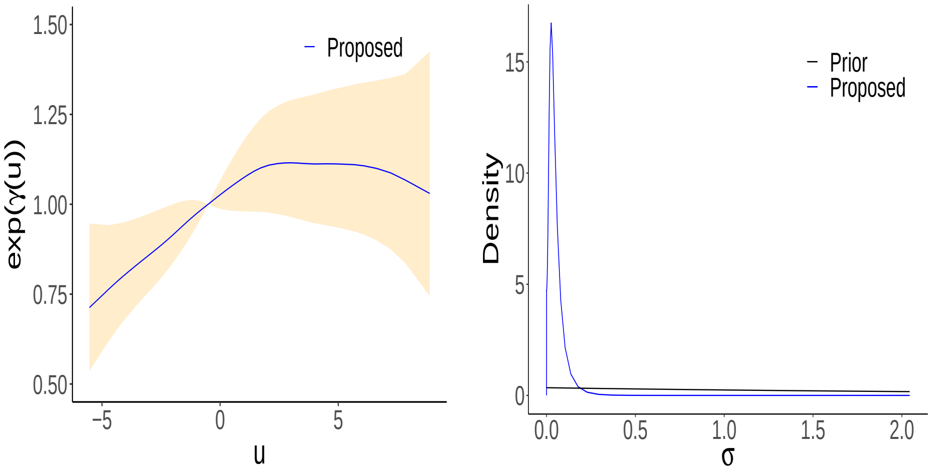
(f) covariate effect under 3

(g) variance parameter under 1

(h) variance parameter under 2

(i) variance parameter under 3

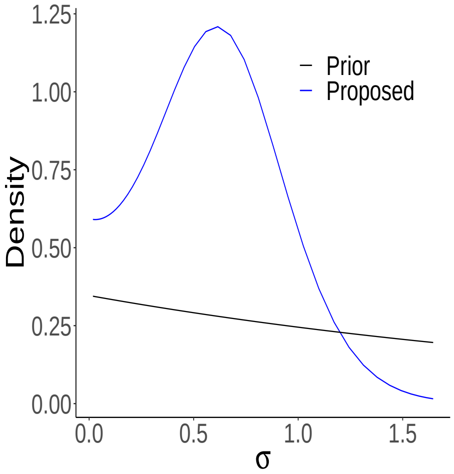
Figure 2: Inferred covariate effects and variance parameters under different baseline hazard functions



(a) Posterior effect of `tpi`

(b) Posterior distribution of σ

Figure 3: Results for the Leukaemia data in section 3.2. (a): posterior mean (blue) and 95% credible interval (shaded) using our method (b): prior (black) and approximate posterior distribution for σ using our method (blue).



(a) Posterior distribution of σ

Figure 4: Posterior distribution for the between-subject standard deviation by our method (blue) and its prior (black), for the kidney data in section 3.3.