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. We propose a flexible and scalable approximate Bayesian inference methodology for the Cox Proportional Hazards model with partial likelihood that allows the inclusion of semi-parametric covariate effects and correlated survival times. We significantly reduce the computational burden introduced by the dense log Hessian matrix, through the use of a posterior approximation method for Extended Latent Gaussian Models. We further improve on existing methods by using an adaptive quadrature technique to reduce the amount of specialist user input required to fit the model, and to minimize the number of dense Hessian matrices required to be stored. We provide two simulation studies to show the improved accuracy of our proposed partial likelihood method over the existing full likelihood method. We demonstrate the practical utility of our method and its computational advantages over existing method through the analysis of Leukemia survival times, with a semi-parametric covariate effect, and Kidney infection times, which are paired. An R package implementing our method will be released publicly.

Keywords: Cox Proportional Hazard Model, Partial Likelihood, Approximate Bayesian inference, Hierarchical Modeling.

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 (**?**). The Cox PH model assumes that any two subjects' event hazards are proportional as a function of time, with the ratio depending on unknown 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 in the presence of complex models for the hazard, which would be difficult to achieve otherwise. However, existing methods for approximate Bayesian inference based on Integrated Nested Laplace Approximations (INLA) (**?**) cannot be applied to the Cox PH model with partial likelihood because the Hessian matrix of

[??]Department of Statistical Science, University of Toronto, **?? ?? ?? ??**
[??]Centre for Global Health Research, St Michael's Hospital, **?? ??**

 DOI: 0000

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 (**?**), but this requires smoothness assumptions to be made about the baseline hazard. Although Bayesian inference on partial likelihood could be carried out using Markov Chain Monte Carlo (MCMC) method, the computational burden of MCMC is much heavier than method based on Laplace Approximation, especially when some parameters are strongly correlated.

? 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. Though related, 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 **?** does not apply to this model. Further, they use a manual integration strategy which requires the user to supply their own grid, a tedious operation which requires specialist knowledge to do properly. This limits the practical utility of their method. In terms of the scalability to large sample, since the size of Hessian matrix grows quadratically with the sample size, direct generalization of their methodology to the partial likelihood of Cox PH model will introduce prohibited computational load when sample size is large.

Recently, **?** proposed a fast and scalable methodology for posterior approximation for Extended Latent Gaussian Models (ELGM), a broad class of models that includes the Cox PH model with partial likelihood. In their paper, they demonstrated the possibility of approximate Bayesian inference on partial likelihood through their ELGM type method with an example that included fixed covariate effects and spatial random effects. Since the method of **?** does not involve additional noised linear predictors in the latent parameter vector, their Hessian matrix of their log-partial likelihood will have fixed dimension independent of the sample size, and hence is scalable for the analysis of large dataset.

In this paper, we utilized the posterior approximation methodology of **?** on Cox PH models with partial likelihood, to propose an approximate Bayesian inference method that allows the inclusion fixed covariate effects, semi-parametric smoothing effect and frailties for correlations between survival times. Through two simulation studies, we illustrated that under certain circumstances, this proposed method based on partial likelihood would yield more reliable result than existing method based on full likelihood. To demonstrate the accuracy of the posterior approximation and the computational advantages over existing partial likelihood method based on MCMC, we applied the proposed ELGM type method to re-analyze two datasets of survival times.

The remainder of this paper is organized as follows. In §**??** we describe the semi-parametric Cox PH model with different types of semi-parametric smoothing that will be used in this paper. In §**??**, we describe existing methods for approximate Bayesian inference on Cox PH model, and why inference method based on ELGM type posterior approximation of **?** should be preferred. In §**??** we illustrate advantages of the proposed methodology in two simulation studies and through the analysis of Leukemia survival data analyzed by **?** and the Kidney catheter data analyzed by **?**. We conclude in §**??** with a discussion.

2 Model

2.1 A General 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 T_{ij} such that T_{ij} is not directly observable when $c_{ij} < T_{ij}$. The observed survival 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 is 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}:

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

Let $\eta = \{\eta_{ij} : i = 1, \dots, n; j = 1, \dots, n_i\}$ be the vector of all the additive linear predictors. Here x_{ij} is a p-dimensional vector of covariates that are modeled as having linear associations with the log-hazard, and $\beta = (\beta_1, \dots, \beta_p)$ are regression coefficients. The $u_q = \{u_{qij} : i = 1, \dots, n; j = 1, \dots, n_i\}, q = 1, \dots, r$ are covariate vectors whose association with the log-hazard is modeled semi-parametrically through unknown smooth functions $\gamma_1, \dots, \gamma_r$. The vector of group intercepts $\xi = \{\xi_i : i = 1, \dots, n\}$, referred to as *frailties* coefficients in the context of survival analysis (**?**), 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 Modeling Semi-parametric covariate effect

The semi-parametric covariate effects $\{\gamma_q\}_{q=1}^{r}$ are modeled as r independent zero-mean Gaussian processes, each defined by its covariance functions C_{σ_q} which in turn parametrized by $\sigma_q > 0$. A typical choice of covariance function is the covariance function of 2-fold Integrated Wiener process(**?**), which has a connection to cubic smoothing splines(**?**). To infer the infinite-dimensional parameters $\{\gamma_q\}_{q=1}^r$, **?** proposed the use of second order random walk model (RW2) to approximate the Integrated Wiener process prior, which includes discretizing the covariate u_q into m_q pre-specified bins and approximate each γ_q by a piecewise constant function at each bin. The m_q dimensional vector of function values are defined as $\Gamma_q = (\Gamma_{q1}, ..., \Gamma_{qm_q})$, with prior distribution being $\Gamma_q|\sigma_q \sim \mathcal{N}(0, \Sigma_q^{-1}(\sigma_q))$ for each $q = 1, ..., m_q$. Each precision $\Sigma_q^{-1}(\sigma_q)$ is sparse and available in closed form. Define $\Gamma = \{\Gamma_q\}_{q=1}^r$, then $\Gamma|\sigma_1, ..., \sigma_q \sim \mathcal{N}(0, \Sigma_\Gamma^{-1})$, with $\Sigma_\Gamma^{-1} = \text{diag}[\Sigma_1^{-1}(\sigma_1), ..., \Sigma_q^{-1}(\sigma_q)]$.

This RW2 model proposed in **?** can be understood as a special case of Bayesian penalizing regression splines, with basis being linear B splines, and penalty matrix being $\Sigma_q^{-1}(\sigma_q)$ (**?**). In our proposed method, we consider the use of cubic B splines instead of linear B splines as the basis function, in order to achieve higher order smoothness in the inferred function. Since the precision matrix $\Sigma_q^{-1}(\sigma_q)$ will have random deficiency of order 2, both RW2 method and the cubic B splines method will correspond to an improper prior for the unknown function $\gamma_q()$, and hence will be incompatible with the type of Laplace approximation in **?**, which we utilized in the proposed approach.

To mitigate this problem of singular precision matrix, a small Gaussian noise will normally be introduced into the additive linear predictor which makes the precision matrix of the latent parameter vector full rank (**??**). In our proposed method, we fix this problem by adding a small constant term (i.e. 0.0001) into the diagonal terms of $\Sigma_q^{-1}(\sigma_q)$, which will also result in a full rank proper precision matrix. Our approach is essentially similar to the method of **??**, with main difference being that our approach only adds noise into the precision matrix corresponding to Γ, but the method of **??** adds noise to the precision matrix of the full latent parameter vector. Furthermore, the modification only shifts the diagonal terms of $\Sigma_q^{-1}(\sigma_q)$ by a very small constant, hence will not change any conditional independence structure in the original prior.

Finally, define the variance parameter vector $\theta = (\theta_0, \ldots, \theta_r)$ where $\theta_q = -2\log\sigma_q$, $q = 1, \ldots, r$, and $\theta_0 = -2\log\sigma_\xi$. The variance parameters are given prior distribution $\theta \sim \pi(\theta)$.

3 Methods

3.1 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:

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

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

The 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:

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

? extend this to permit partial likelihoods of the form:

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

which still does not include (**??**). **?** are able to write the likelihood for their Cox PH model in the form (**??**) using the full, not partial likelihood (**??**). Because of this, they require assumptions to be made about the baseline hazard.

Further define $\Delta_{lk,ij} = \eta_{lk} - \eta_{ij}$ in terms of the additive predictors (**??**). 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 (**??**) 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 $\ell(\Delta; y) = \log \pi(y|\Delta)$.

In typical Laplace approximation for posterior distributions, the *latent parameters* W will be defined as $W = (\Delta, \Gamma, \beta, \xi)$, where the (differenced) linear predictors Δ are included as part of the latent parameter vector. Approximate Bayesian inference of this type requires the precision matrix of W to be non-singular (**???**), and hence a small noise term $\epsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0, \tau^{-1})$ (for some large, fixed τ) is added into the model to make the required matrices non-singular. Redefine the (differenced) linear predictors as $\Delta_{ij} = \eta_{11} - \eta_{ij} + \epsilon_{ij}$, then the resulting precision matrix of W will be non-singular even if improper prior such as the RW2 prior is used.

Such posterior approximation methods have the advantage that, when the likelihood can be factored out in the form of (**??**), the resulting log likelihood Hessian matrix will be diagonal and hence efficient to be computed and stored (**?**). Alternatively, if the likelihood is in the form of (**??**), the Hessian matrix will still be sparse even it is no longer diagonal (**?**). However, if one considers applying such approximate Bayesian inference on Cox PH model with partial likelihood, the resulting Hessian matrix will be completely dense and with number of elements growing quadratically with sample size N. Therefore, the methods of **???** are not feasible for the inference on Cox PH model with partial likelihood.

The recent posterior approximation method of **?** on the other hand, considers the latent parameter vector W to only contain the parameters of interest (i.e. $W = (\Gamma, \beta, \xi)$).

In this way, the dimension of latent parameter vector will be constant, and hence the dimension of the dense Hessian matrix will be small regardless of the sample size N. This enables the approximate Bayesian inference to be carried out on partial likelihood, as long as all the elements in W have proper priors. The detailed approximate Bayesian inference method will be described at below.

Define $W|\theta \sim \mathrm{N}\left[0, Q_\theta^{-1}\right]$, where Q_θ is the covariance matrix for W. 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 (**??**). 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 (**??**).

All of the quantities of interest (**??**) – (**??**) depend on intractable high-dimensional integrals. **?** 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 a manually-rescaled Gauss-Hermite 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 (**??**) are computed as follows. For any fixed θ, define

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

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 $\log \pi(W|\theta, y)$ 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 **?** yields a Laplace approximation:

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

With these approximations available, inference for W makes use of the approximation (**??**).

Computing the approximations (**??**) requires choosing a quadrature rule consisting of nodes $\left\{\theta^k\right\}_{k=1}^K$ and weights $\{\delta_k\}_{k=1}^K$ for some chosen $K \in \mathbb{N}$. **?** lay a user-chosen grid over a range of θ that is thought to be plausible, and then compute the Gaussian (**??**) and Laplace (**??**) approximations at each point on this grid. This requires the user to choose the location and spread of the grid points, as well as a number K of points that is large enough such that the structure of the resulting posterior approximations is captured. The function $\pi(W|Y, \theta)$ must be optimized, and the Hessian matrix stored, for each of these K points. In addition to this strategy requiring the user to have specialist knowledge to implement, it is potentially computationally wasteful. In our case, this problem is made more severe by the presence of a dense Hessian. **?** use the INLA software which uses a custom adaptive quadrature rule which avoids the need for the user to choose points, however may still result in a large number of points being used.

To mitigate the computational challenges associated with applying a manual quadrature rule for (**??**), we implement Adaptive Gauss-Hermite Quadrature (AGHQ). This technique has been motivated as a useful tool for Bayesian inference (**?**) and work has been done to show that it is very accurate when using only a very small number of quadrature points (**??**), for example attaining $O(N^{-1})$ asymptotic accuracy with $K = 3$ and $O(N^{-2})$ with $K = 5$. The use of a small number of quadrature points means only this number of dense Hessian need to be stored in memory, a marked improvement over **?** that is necessary to extend their method to work with the partial likelihood of the Cox PH model.

Computing the AGHQ rule requires computation of the mode of the Laplace approximation:

$$\widehat{\theta} = \operatorname{argmax} \log \widetilde{\pi}_{LA}(\theta|y), \tag{15}$$

as well as its Hessian matrix and its Cholesky. These matrix quantities are straightforward to obtain as θ is low-dimensional. For the optimization, we follow **?** and use numerical derivatives and a BFGS algorithm which limits the number of derivatives which must be computed. Computations of the numerical integrations make use of the `aghq` package (**?**), and the required first two derivatives of $\log \pi(W|\theta, y)$ are computed using the automatic differentiation method of the TMB package (**?**).

4 Examples

In this section we present two simulation studies and two data analysis examples. All the codes are available in the online supplementary materials.

4.1 Simulation studies

In this section, we will provide two simulation studies to demonstrate the accuracy of our proposed method over the existing full likelihood method INLA under certain settings.

Simulation with sparse frailties

In the first simulation study, we considered the Bayesian inference problem for models with sparse frailties. In other words, survival times were correlated within groups while the number of observations in each group is small. We randomly generated $n = 100$ groups, each group with m observations. The group-level frailties $\{\xi_1, ..., \xi_{100}\}$ were simulated independently from $\mathcal{N}(0, \sigma_{\xi}^{-2})$, with $\sigma_{\xi} = 0.8$. Besides the independent frailties, we also assumed there is a covariate x generated from $\mathcal{N}(0, 9^{-1})$, with covariate effect $\beta = 0.2$. Among all the survival times generated in this study, 10% of observations were randomly selected to be right-censored. In this simulation study, we consider the baseline function to be a simple step function. The baseline hazard function in this simulation study is shown in **??**. We consider five different levels of frailty sparsity in this simulation study by setting m to 2, 4, 6, 8 and 10.

The fixed effect β was given a prior $\mathcal{N}(0, 1000^{-1})$. The variance parameter σ_ξ was given an Exponential prior with median of 1. The same priors were used for implementations of both our proposed method and of INLA. For the adaptive quadrature we used in our inference, the number of grid points for variance parameter was set to be $K = 15$. For the implementation of INLA, we used its first order random walk model for the baseline hazard run under its default settings. To compare the accuracies between the two methods, we used the metrics of posterior mean square error (MSE) and coverage rates of the 95% posterior credible intervals, for both the fixed effect parameter and the frailties. All the metrics were computed by averaging through 300 independent repetitions in order to make the comparison accurate.

The comparison metrics are shown in table **??** and **??**. Based on the table, it can be noticed that our proposed method in general gives more accurate inferential results than INLA, both in terms of smaller MSE and coverage rates closer to the nominal level (i.e. 95%), and these differences get larger as the frailties get sparser (smaller m). When $m = 2$, both methods didn't achieve very accurate inferential results for the frailties in terms of MSE, because both methods involved using Laplace approximation for $\pi(\sigma_\xi|y)$, which is known to be inaccurate when the frailties are sparse (**?**). However, the result from our proposed method is still significantly more accurate than INLA.

In this study, the quantities of interest that need to be inferred include 100 frailties $\{\xi_1, ..., \xi_{100}\}$, one fixed effect parameter β and one variance parameter σ_ξ. However, the number of data points available is only $100m$. The problem is even severer when the inference is carried out on the type of full likelihood used by INLA, due to the more parameters it introduced in order to approximate the baseline hazard (**?**). Hence, as we have illustrated through this simulation study that for inferences on sparse frailties, our proposed method based on partial likelihood will yield more accurate result than INLA under such setting.

Simulation with non-smooth baseline

To illustrate the accuracy of our method over INLA when the smoothness assumption for baseline hazard function is violated, we performed our second simulation study. We generated $n = 1000$ uncorrelated data points from a distribution with known hazard function. For the baseline hazard functions, we consider three different settings corresponding to three different levels of wiggliness. More specifically, we consider the cases when baseline hazard function is simple step function, oscillating step function and an extremely complicated function that switches between linear and constant. The three baseline hazards $h_0(t)$ are shown in Figure **??**. The additive predictor is $\eta_i = \gamma(u_i)$ with $\gamma(u) = 1.5[\sin(0.8u) + 1]$ in all the three simulation settings. We generated the covariates u as $u_1, \ldots, u_n \overset{iid}{\sim} \text{Unif}(-6, 6)$, and randomly censored 10% of all the survival times. Since the overall intercept parameter cannot be identified in partial likelihood, we put a sum-to-zero constraint such that $\sum_{i=1}^{n} \gamma(u_i) = 0$.

To infer the unknown risk function γ, we used the Bayesian B-spline smoothing method mentioned in section **??** in our proposed method, with fifty equally spaced knots. For the smoothing method in INLA, we placed the values of u into 50 discrete bins, and fitted a second-order random walk model for γ (**?**). As before, we implemented INLA under its default setting, with a first-order random walk model for the baseline hazard. 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. In both of the smoothing methods, the single variance parameter σ that controls the smoothness of γ, was modelled with an Exponential(λ) prior with λ chosen such that $\mathbb{P}(\sigma > 2) = 0.5$, which is a *penalized complexity* prior of **?**.

As in the second simulation study, we compared the accuracy of our proposed method with INLA, using the metrics of MSE computed using posterior mean and coverage rate computed using the 95% posterior credible interval. The metrics were still computed by averaging over 300 independent repetitions.

The comparison metrics under the three settings of baseline hazard are shown in figure **??** and table **??**. Based on the table, it can be noticed that as the baseline hazard function gets more complicated, our proposed method tends to yield more accurate results than INLA, both in terms of smaller MSE and coverage rates closer to the nominal level (i.e. 95%). In particular, under the most extreme setting for the baseline hazard, INLA's 95% posterior credible interval only yielded coverage rate of 65.8%. This is not unexpected as the full-likelihood used in INLA's inference implicitly requires that the baseline hazard is smooth enough to be approximated well by its first-order random walk, which will not hold under such setting where the baseline hazard is varying rapidly as time changes. On the other hand, the inference of our proposed method relies on the partial likelihood, which makes no assumption on the smoothness of the baseline hazard, and hence unaffected by the non-smooth baseline hazard in this study.

4.2 Leukemia Data

We implemented our proposed procedure to fit a semi-parametric Cox PH model to the Leukemia data set analyzed by **?**. The dataset contains information from $n = 1043$ independent adult leukemia patients, with 16% of observations right-censored. We are interested in quantifying the relationship between survival rate of leukemia patients with the Townsend deprivation index (tpi) corresponding to the patient's location, controlling effect of the age of the patient, the count of white blood cells at diagnosis (wbc), and sex of the patient.

The effects of age, sex and white blood cell count were modeled linearly. The deprivation index (tpi) was modeled as a semi-parametric effect using method described in section **??**, with fifty equally spaced knots. Prior distributions $\beta \overset{iid}{\sim} \mathcal{N}(0, 0.001^{-1})$, were used for the linear regression coefficients. The semi-parametric effects $\{\gamma(\text{tpi}_1), \cdots, \gamma(\text{tpi}_n)\}$ were modeled with the reference constraint $\sum_{i=1}^{n} \gamma(\text{tpi}_i) = 0$. The single variance parameter σ was given an Exponential prior with a prior median of 2. For the adaptive quadrature we used in our inference, the number of grid points for variance parameter to is set to be $K = 15$. As a comparison, we also implemented INLA for this problem to do inference on the full likelihood, with baseline hazard modeled under its default setting. For the prior on semi-parametric effect $\gamma(\text{tpi})$ in INLA, we utilized its second-order random walk model, with values of tpi placed into 50 equally space bins. The standard deviation parameter σ that controls the smoothness of γ, and the fixed effect parameters were given same priors as before.

Figure **??**(a) shows the posterior results of the exponentiated covariate effect of tpi. Our inferred covariate effect of tpi is more volatile compared to the result reported by **?**, 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 **?** relies on the use of full likelihood function with baseline hazard function modeled semi-parametrically. Our approach implements the approximate Bayesian inference using the partial likelihood, hence requires no assumption on the form of the baseline hazard function.

To demonstrate the accuracy of our proposed approximation and the computational advantage compared to existing method, we also fitted the same partial likelihood model using MCMC method, through STAN's No U-turn Sampler (NUTS) (**?**). The runtimes respectively for the proposed method and MCMC are 1.88 minutes and 8.63 hours. Figure **??**(b) shows the posterior results on the corresponding standard deviation σ. The difference between posterior distributions of σ yielded by the proposed method and that yielded by MCMC method can be quantified using Kolmogorv-Smirov (KS) statistic, which measures the maximal absolute difference between the two cumulative posterior distributions. As shown in figure **??**(c), the KS statistics for this example was computed to be 0.05. This demonstrates that our proposed approach yields accurate approximations to the posterior distributions yielded by MCMC method, with a much faster runtime.

4.3 Kidney Catheter Data

? analysed a Kidney Catheter dataset using their proposed penalized partial likelihood method. The Kidney Catheter dataset contains 76 times to infection at the point of insertion of a catheter, for $n = 38$ patients. Each patient $i = 1, \dots, n$ forms a group, and the survival times are the time to infection of each patient's $n_i = 2$ kidneys. An observation for the survival time of a kidney is censored if the catheter is removed for reasons other than an infection.

We first analyzed this dataset on full-likelihood using INLA, with $\mathcal{N}(0, 0.001^{-1})$ priors on the linear covariate effects for age, sex and pre-existing disease types. Subject-specific intercepts $\xi_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\xi^2)$ were included to account for correlation between kidneys from the same subject. We used an Exponential prior distribution for σ_ξ with median 2. The setting for baseline hazard is set to default in INLA's implementation. As a comparison, we also used our procedure to fit a Cox PH model to these grouped data on partial likelihood, with the same set of priors as above. For the adaptive quadrature we used in our inference, the number of grid points for variance parameter to is set to be 18. Ties are handled using the method of **?**.

Table ?? shows the results of our procedure compared to that obtained using the full likelihood method of ?. Based on the table, our procedure gave different posterior means and reported larger posterior standard deviations compared to INLA, especially for the effects of different disease types. This is also reflected in ??(a), where our posterior distribution for σ_ξ is wider than that of INLA. As we have shown in ??, when sparse frailties exist, Bayesian inference on partial likelihood tends to be more stable than on full-likelihood.

Again, to assess the accuracy of our approximation to the posterior distribution, we implemented the same partial likelihood model using MCMC method. Table ?? shows the comparison between the results from proposed approach with the results from MCMC. Figure ??(b) compares the MCMC samples from posterior of σ_ξ with the approximate posterior obtained using our approach. The posterior cumulative distributions for σ_ξ are compared in ??(c), with maximal absolute difference being 0.09. The runtimes are respectively 0.53 seconds for our approach and 1.98 minutes for MCMC with 35000 iterations.

5 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 for semi-parametric covariate effects and correlated survival times. The use of partial likelihood does not require any assumption on the baseline hazard function, which is an advantage over existing approaches for Bayesian inference in this model. We have demonstrated the accuracy of our new approach through simulation studies and analysis of two classical datasets in survival analysis. Our proposed method is an appealing option to adopt for the analysis of time-to-event data, when the inference of baseline hazard is not of primary interest.

One limitation of our proposed methodology is that for analyses of sparse frailties, the type of posterior approximation will tend to be less accurate, due to the nature of the Laplace approximation (?). For such application, sampling based method such as MCMC might be preferred for higher inference quality, at the cost of longer runtime.

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.

Data Availability Statement

The simulated data of example 3.1 are available in the supplementary material with this paper. Data for example 3.2 were obtained from R package "INLA" (?) and are freely available. Data for example 3.3 were obtained from R package ”survival” (**?**) and are freely available.

Number of Measurements	β Coverage Rate (Proposed)	β Coverage Rate (INLA)	β MSE (Proposed)	β MSE (INLA)
m = 2	0.960	0.880	0.00116	0.00168
m = 4	0.907	0.833	0.000672	0.000942
m = 6	0.977	0.883	0.000294	0.000492
m = 8	0.943	0.893	0.000206	0.000335
m = 10	0.943	0.833	0.000159	0.000316

Table 1: Comparison metrics in terms of MSE and posterior coverage rate from 300 independent replications, for the fixed effect parameter in the first simulation study in section **??**.

Number of Measurements	ξ Coverage Rate (Proposed)	ξ Coverage Rate (INLA)	ξ MSE (Proposed)	ξ MSE (INLA)
m = 2	0.916	0.859	0.371	0.404
m = 4	0.942	0.921	0.224	0.241
m = 6	0.942	0.927	0.162	0.175
m = 8	0.944	0.936	0.130	0.137
m = 10	0.946	0.938	0.106	0.113

Table 2: Comparison metrics in terms of MSE and posterior coverage rate from 300 independent replications, for the 100 frailty effects in the first simulation study in section **??**.

Baseline Hazards	ξ Coverage Rate (Proposed)	ξ Coverage Rate (INLA)	ξ MSE (Proposed)	ξ MSE (INLA)
Simple Baseline	0.969	0.974	0.0116	0.0122
Oscillating Baseline	0.968	0.949	0.0117	0.0148
Complicated Baseline	0.968	0.659	0.0117	0.0448

Table 3: Comparison metrics in terms of MSE and posterior coverage rate from 300 independent replications, for the 100 frailty effects in the first simulation study in section **??**.

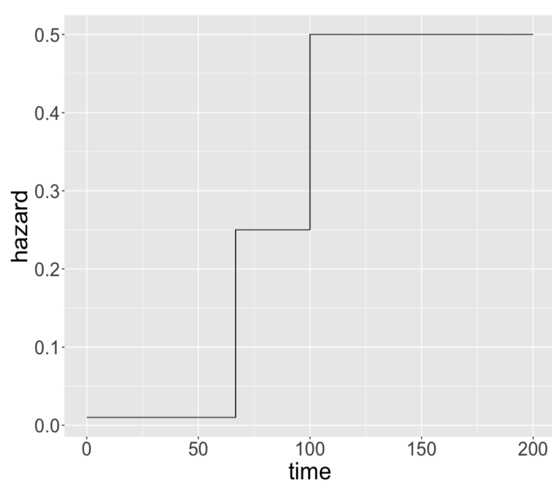
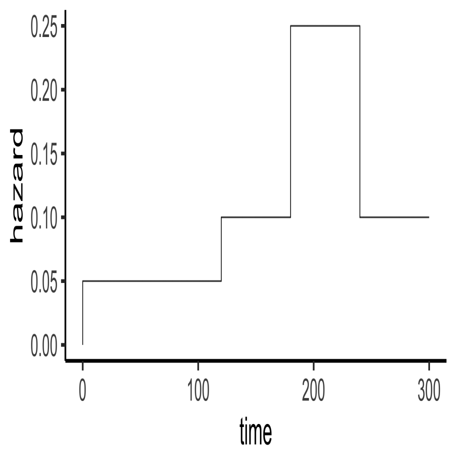
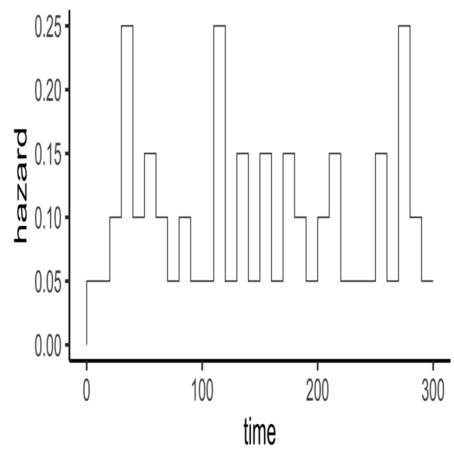


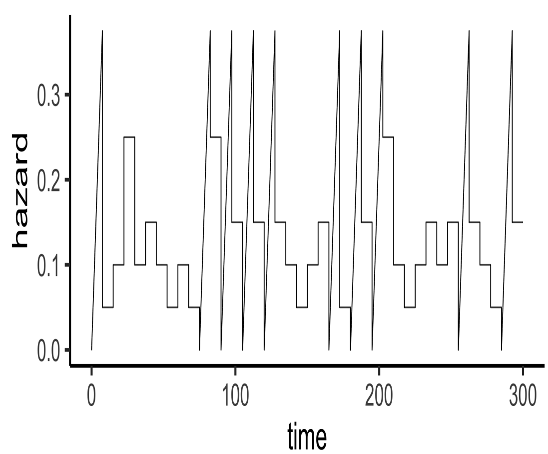
Figure 1: True Baseline Hazard in the first example in **??**.



(a) Simple stepwise baseline

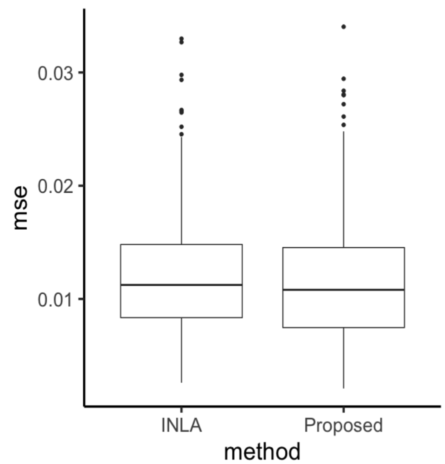


(b) Oscillating stepwise baseline

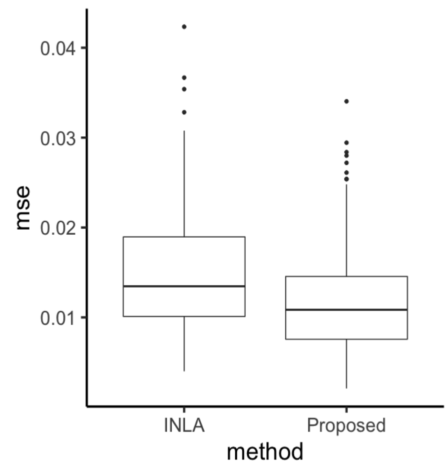


(c) Complicated baseline

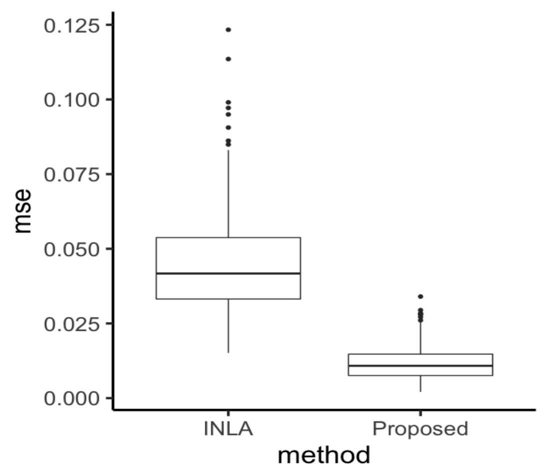
Figure 2: True Baseline Hazards in the second example in **??**.



(a) Simple stepwise baseline

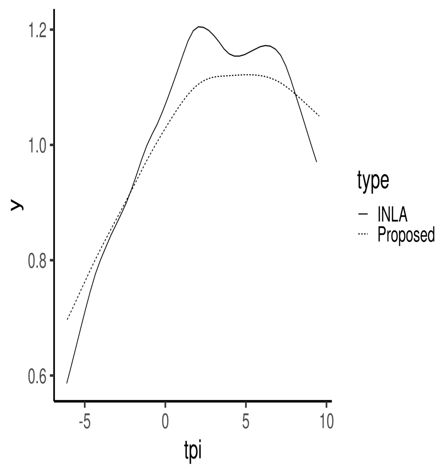


(b) Oscillating stepwise baseline

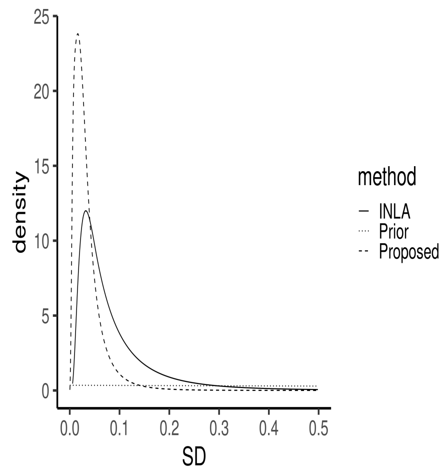


(c) Complicated baseline

Figure 3: Results for the second simulation in section **??**. (a): Box-plot of MSE from 300 replications with simple baseline, using INLA and the proposed method. (b): Box-plot of MSE from 300 replications with oscillating baseline, using INLA and the proposed method. (c): Box-plot of MSE from 300 replications with complicated baseline, using INLA and the proposed method.



(a) Posterior effect of `tpi`



(b) Posterior distribution of σ

leukHyper2.png

(c) Posterior distributions of σ

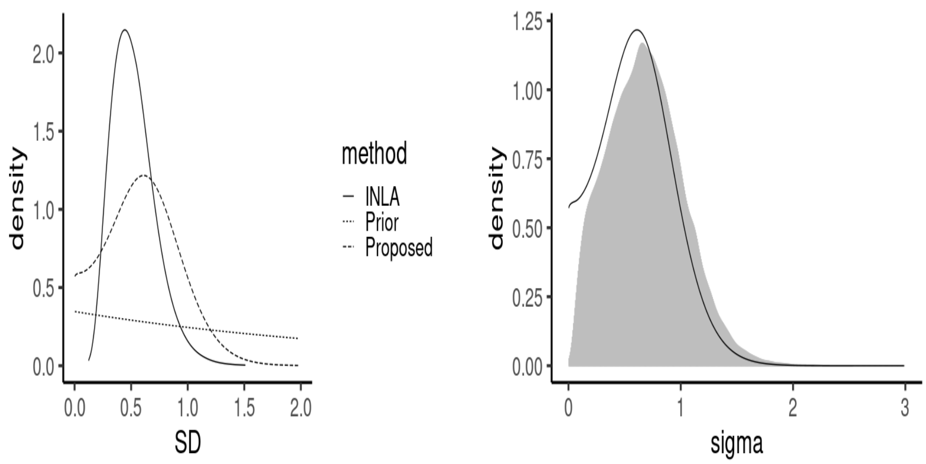
Figure 4: Results for the Leukemia data in section **??**. (a): (Exponentiated) posterior mean for the semi-parametric tpi effect using our proposed method(dashed) and INLA(solid). (b): Prior(dotted) and posterior distributions for σ using our proposed method(dashed) and INLA(solid). (c): Posterior distribution for σ obtained using MCMC(gray histogram), and using the proposed method(black line).

		Proposed		INLA	
Variables/Reference	Levels	Mean	SD	Mean	SD
Age		0.00467	0.0149	0.00235	0.0130
Sex/Male	Female	-1.65	0.463	-1.64	0.385
Disease Type/Other	GN	0.178	0.532	0.111	0.474
	AN	0.420	0.528	0.519	0.467
	PKD	-1.15	0.817	-1.06	0.708

Table 4: Estimated means and standard deviations of linear effects by proposed method and INLA's full likelihood method for the kidney data in section **??**.

		Proposed		MCMC	
Variables/Reference	Levels	Mean	SD	Mean	SD
Age		0.00467	0.0149	0.00516	0.0158
Sex/Male	Female	-1.65	0.463	-1.72	0.507
Disease Type/Other	GN	0.178	0.532	0.172	0.576
	AN	0.420	0.528	0.415	0.573
	PKD	-1.15	0.817	-1.26	0.859

Table 5: Estimated means and standard deviations of linear effects by proposed method and MCMC method for the kidney data in section **??**.



(a) Posterior distributions of σ (b) Posterior distributions of σ

Figure 5: Results for the kidney data in section **??**. (a): Prior(dotted) and posterior distributions for σ using our proposed method(dashed) and INLA(solid) (b): Posterior distribution for σ obtained using MCMC(gray histogram), and using the proposed method(black line).