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. The model we consider includes semi-parametric covariate effects and correlated survival times. The proposed method is based on nested approximations and adaptive quadrature, and the computational burden of working with the log-partial likelihood is mitigated through automatic differentiation and Laplace approximation. We provide two simulation studies to show the accuracy of the proposed approach, compared with the existing methods. We demonstrate the practical utility of our method and its computational advantages over MCMC methods 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 (Cox, 1972b). 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) (Rue et al., 2009) cannot be applied to the Cox PH model with partial likelihood because

*Department of Statistical Science, University of Toronto, aguero.zhang@mail.utoronto.ca alex.stringer@mail.utoronto.ca patrick.brown@utoronto.ca jamie.stafford@utoronto.ca

†Centre for Global Health Research, St Michael's Hospital, alex.stringer@mail.utoronto.ca patrick.brown@utoronto.ca

 DOI: 0000

the Hessian matrix of the log partial-likelihood is fully dense while INLA requires this matrix to be diagonal.

Alternative methods of doing Bayesian inference on this kind of survival model have been considered in the literature. Dykstra and Laud (1981) considered a fully non-parametric approach for Bayesian survival analysis, where the entire hazard function is modelled with an extended gamma process prior and the posterior distribution is derived to be another extended gamma process. Kim and Kim (2009) considered Bayesian analysis on Cox PH model on partial likelihood and on full likelihood with a extended gamma process prior for the baseline hazard, and carried out inferences based on Markov Chain Monte Carlo (MCMC). Martino et al. (2011) considered application of the INLA methodology to the Cox PH model, using the full likelihood with baseline hazard modeled semi-parametrically. Kalbfleisch (1978) derived the partial likelihood to be the limiting posterior when baseline hazards are modelled with noninformative priors, and Sinha et al. (2003) later extended the result to allow the inclusion of grouped survival data, and implemented partial likelihood based Bayesian inference with a Gibbs sampling algorithm. Henschel et al. (2009) proposed a Bayesian inference method using MCMC on the full likelihood, with baseline hazard function modelled either as piecewise constant or as linear combination of B splines, and they accommodated the inclusion of different types of frailties in their method. Hennerfeind et al. (2006) developed a general geo-additive Cox PH model that allows the inclusion of components such as non-linear covariate effect, spatial effect and group level frailties, with inference carried out using MCMC on full likelihood and baseline hazards modelled using P splines. Kneib (2005) generalized the method of Hennerfeind et al. (2006) to acommodate left truncation, left censoring, and interval censoring of the survival times. Most of the existing methods for Bayesian inference of Cox PH model have been utlizing MCMC method to obtain the posterior, and are based on the full likelihood with an explicit form to model the unknown baseline hazard.

Stringer et al. (2020) developed an approximate Bayesian inference methodology for case-crossover model, which can be viewed as a special case of Cox PH model, by applying 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. The partial likelihood they considered correponds to one of the simplest special case of the general 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 generally fully dense, leading to increased computational burden when compared to the model considered by Stringer et al. (2020). Further, they use a manual integration strategy which requires the user to supply their own quadrature points, which requires specialist knowledge to do properly. This limits the practical utility of their method. In order to make approximate Bayesian inferences for the Cox PH model with partial likelihood, an alternative computational strategy is needed.

In this paper, we improve upon the posterior approximation methodology of Stringer et al. (2020) and apply it to fitting Cox PH models with partial likelihood, with fixed

and semi-parametric covariate effects, and frailties for modelling correlations between survival times. Through two simulation studies, we illustrate the circumstances under which the proposed method yields improved results compared to existing methods based on full likelihood. Through the analysis of two datasets with semi-parametric effects and correlated survival times, respectively, we demonstrate the accuracy of the posterior approximation and the computational advantages compared to partial likelihood method fit with MCMC.

The remainder of this paper is organized as follows. In §2 we describe the semi-parametric Cox PH model and the method of semi-parametric smoothing that will be used in this paper. In §3, we describe our proposed methodology and the introduced improvements to solve the computational challenges presented by the complicated partial likelihood. In §4 we illustrate advantages of the proposed methodology in two simulation studies and through the analysis of Leukemia survival data analyzed by Martino et al. (2011) and the Kidney catheter data analyzed by McGilchrist and Aisbett (1991). We conclude in §5 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 $y = \{(y_{ij}, d_{ij}) : i \in [n]; j \in [n_i]\}$. 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. Kim and Kim (2009) only considered the inference on linear fixed effects with linear predictor defined as $\eta_{ij} = x_{ij}^T\boldsymbol{\beta}$, and they briefly discussed the possibility of generalizing their method to accommodate group frailty terms. Sinha et al. (2003) proposed a MCMC method for inference with both fixed effects and group level frailties using partial likelihood, but their method is not flexible enough to accommodate nonlinear covariate effect. Dykstra and Laud (1981) on the other hand considered to model the entire hazard function $h_{ij}(t)$ nonparametrically, but their method cannot be used to quantify the effect of a particular covariate on the survival times.

Stringer et al. (2020) considered a general linear predictor that accommodates both linear fixed and nonlinear semi-parametric covariate effects, but the type of likelihood they considered is one of the simplest special cases of the general partial likelihood of Cox PH model, and does not allow the estimation of group level frailty terms that account for the correlation between survival times within the same group.

To accommodate nonlinear covariate effects and correlated survival times, we define an additive predictor η_{ij} which links the covariates for the ijth observation to the survival time T_{ij}:

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

Let $\boldsymbol{\eta} = \{\eta_{ij} : i \in [n]; j \in [n_i]\}$ 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 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are regression coefficients. The $u_q = \{u_{qij} : i \in [n]; j \in [n_i]\}, q \in [r]$ are covariates whose associations with the log-hazard are modeled semi-parametrically through unknown smooth functions $\{\gamma_i, i \in [r]\}$. The vector of group intercepts $\xi = \{\xi_i : i \in [n]\}$—referred to as *frailties* 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 $h_0(t)$. However, in contrast to the model considered by Stringer et al. (2020), the group-specific intercepts ξ_i are estimable in this model.

2.2 Modeling Semi-parametric covariate effect

The semi-parametric covariate effects γ_q, $q \in [r]$ are modeled as $r \in \mathbb{N}$ independent zero-mean Gaussian processes, each defined by its covariance function C_{σ_q}. The covariance functions are each parametrized by a single parameter $\sigma_q > 0$. A typical choice of covariance function is the covariance function of second fold Integrated Wiener process(Shepp, 1966), which has a connection to cubic smoothing splines (Wahba, 1978).

To infer the infinite-dimensional parameters γ_q, $q \in [r]$, Lindgren and Rue (2008) proposed the use of second order random walk model (RW2) to discretize the Integrated Wiener process prior. Miller et al. (2020) showed that the RW2 model proposed by Lindgren and Rue (2008) can be understood as a penalizing basis expansion of the form $\gamma(u) = \sum_{j=1}^{d} \phi_j(u)\Gamma_j$ for each $\gamma(\cdot)$ (dropping the subscript q), where the random weights $\boldsymbol{\Gamma} = \{\Gamma_j, j \in [d]\}$ are parameters to be inferred and $\phi_j(\cdot), j \in [d]$ are fixed, known basis functions which must be chosen. Yue et al. (2014) note that a similar discretization technique to the one used by Lindgren and Rue (2008) yields the B-spline smoothing with integrated derivative penalty of general order proposed by O'Sullivan (1986), and Wood (2017) provide an explicit construction of the corresponding precision matrix. In the method of Lindgren and Rue (2008), the basis function $\phi_j(\cdot)$ is chosen to be the linear B spline function, and the random weights are given zero-mean Gaussian prior with a banded precision matrix $\Sigma^{-1}(\sigma)$ controlled by a single variance parameter σ. In the proposed approach, we use cubic B-splines for the $\phi_j(\cdot)$ and choose the precision matrix $\Sigma^{-1}(\sigma)$ that is obtained by using an integrated second derivative penalty of Wood (2017).

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 $\theta \sim \pi(\theta)$. Since the overall intercept parameter cannot be identified in partial likelihood, we put a sum-to-zero constraint such that $\sum_{i=1}^{n}\gamma(u_i) = 0$ in all the following example as default.

2.3 Partial Likelihood

Our 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 (2) 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:

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

Stringer et al. (2020) 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 (2). Martino et al. (2011) are able to write the likelihood for their Cox PH model in the form (4) using the full, not partial likelihood (2). 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 (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 (2) depends on η only through $\Delta = \{\Delta_{ij} : i \in [n]; j \in [n_i]\}$. 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 (Rue et al., 2009; Martino et al., 2011; Stringer et al., 2020). Approximate Bayesian inference of this type requires the precision matrix of W to be non-singular (Tierney and Kadane, 1986), 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. Define the noised linear predictors as $\tilde{\Delta}_{ij} = \eta_{11} - \eta_{ij} + \epsilon_{ij}$, then the resulting precision matrix of (noised) latent parameters $\tilde{W} := (\tilde{\Delta}, \Gamma, \beta, \xi)$ 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 (4), the resulting log likelihood Hessian matrix will be diagonal and hence efficient to be computed and stored (Rue et al., 2009). Alternatively, if the likelihood is in the form of (5), the Hessian matrix will still be sparse even it is no longer diagonal (Stringer et al., 2020). 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 Rue et al. (2009); Martino et al. (2011); Stringer et al. (2020) are not computationally feasible for the inference on Cox PH model with partial likelihood.

3 Methods

3.1 Approximate Bayesian Inference

To deal with the problem of dense Hessian matrix, we proposed a new way to utilize the Laplace approximation, by defining the latent parameter vector to only include the parameters of interest, $W = (\Gamma, \beta, \xi)$. Note that in our definition, the dimension of this latent parameter vector will be constant, and hence the size of the dense Hessian matrix will be small regardless of the sample size N. This will significantly reduces the memory requirement and computational challenge introduced by computing, storing and factorizing the Hessian matrix, which are necessary for the inferential procedures.

When semi-parametric covariate effect is included in the model, W will have a singular precision matrix as the precision matrix of Γ is rank deficient, and hence direct application of the Laplace approximation of Tierney and Kadane (1986) will be problematic. This problem of singular precision matrix is typically fixed by introducing a small Gaussian noise into the additive linear predictors, which makes the precision matrix full rank (Stringer et al., 2020; Rue et al., 2009). In our proposed approach, no noises will be added into the linear predictors; instead we fix this problem by adding a small constant term (i.e. 0.0001) into the diagonal terms of the precision matrix of Γ, Σ_Γ^{-1}, which will also result in a full rank precision matrix for W. Our approach is essentially similar to the method of Stringer et al. (2020); Rue et al. (2009), with main difference being that our approach only adds noise into the precision matrix corresponding to Γ, but the method of Stringer et al. (2020); Rue et al. (2009) adds noise to the precision matrix of the full latent parameter vector. Furthermore, our modification only shifts the diagonal terms of Σ_Γ^{-1} by a very small constant, hence will not change any conditional independence structure in the original prior.

Define $W|\theta \sim \mathrm{N}\left[0, Q_\theta^{-1}\right]$, where Q_θ is the covariance matrix for W. We are interested in estimating and sampling from the joint posterior distribution of the latent parameters:

$$\pi(W|y) = \int \pi(W|y,\theta)\pi(\theta|y)d\theta. \tag{6}$$

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).

For the posterior of variance parameter (7), we follow the procedure of Stringer et al. (2020) to approximate it with its corresponding Laplace approximation $\tilde{\pi}_{LA}(\theta|y)$. The posterior of the latent parameter vector (6) is approximated by $\tilde{\pi}(W|y)$ defined as:

$$\tilde{\pi}(W|y) = \sum_{k=1}^{K} \tilde{\pi}_G(W|y,\theta^k)\tilde{\pi}_{LA}(\theta^k|y)\delta_k, \tag{8}$$

where $\left\{\theta^k, \delta_k\right\}_{k=1}^K$ is a set of nodes and weights corresponding to a Gauss-Hermite quadrature rule. The $\tilde{\pi}_G(W|y,\theta^k)$ is a Gaussian approximation for $\pi(W|y,\theta^k)$ and the $\tilde{\pi}_{LA}(\theta^k|y)$ is a Laplace approximation for $\pi(\theta^k|y)$, which we describe at below.

For any fixed θ, define

$$\begin{aligned} \widehat{W}_\theta = \left(\widehat{\Gamma}_\theta, \widehat{\beta}, \widehat{\xi}_\theta\right) &= \mathrm{argmax}_W \log \pi(W|\theta,y), \\ H_\theta(W) &= -\frac{\partial^2}{\partial W \partial W^T} \log \pi(W|\theta,y). \end{aligned} \tag{9}$$

For the conditional posterior

$$\pi(W|\theta,y) \propto \exp\left\{-\frac{1}{2}W^T Q_\theta W + \ell\left(\Delta;Y\right)\right\}, \tag{10}$$

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

$$\pi(W|\theta,y) \approx \tilde{\pi}_G(W|y,\theta) \propto \exp\left\{-\frac{1}{2}\left(W-\widehat{W}_\theta\right)^T H_\theta\left(\widehat{W}_\theta\right)\left(W-\widehat{W}_\theta\right)\right\}. \tag{11}$$

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

$$\pi(\theta|y) \approx \tilde{\pi}_{LA}(\theta|y) \propto \pi(\theta)\left\{\frac{|Q_\theta|}{\left|H_\theta\left(\widehat{W}_\theta\right)\right|}\right\}^{1/2} \exp\left\{-\frac{1}{2}\widehat{W}_\theta^T Q_\theta \widehat{W}_\theta + \ell\left(\widehat{\Delta}_\theta;y\right)\right\}. \tag{12}$$

With these approximations available, inference for θ can be directly obtained using the analytical form of $\tilde{\pi}_{LA}(\theta|y)$ as in equation (12). Inference for W or its marginal component can be easily obtained if it is possible to get independent samples from $\tilde{\pi}(W|y)$.

To sample from $\tilde{\pi}(W|y)$, note that by equation(6), $\tilde{\pi}(W|y)$ is Guassian mixture distribution with K mixture components each with mixture weight being

$$\phi_k = \tilde{\pi}_{LA}(\theta^k|y)\delta_k, \quad k \in [K].$$

For a large integer B, we sample independent $\{Z_i\}_{i=1}^B$ from Multinomial$(\phi_1, ..., \phi_K)$, and then for each $Z_i \in [K]$, sample W_i from $\tilde{\pi}_G(W|y, \theta^{Z_i})$. The resulting sample $\{W_i\}_{i=1}^B$ then contains B independent draws from $\tilde{\pi}(W|y)$, and all the posterior summaries for W can be obtained using this independent sample.

3.2 Adaptive quadrature and automatic differentiation

Computing the approximations (8) 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 $K \in \mathbb{N}$. Stringer et al. (2020) lay a user-chosen grid over a range of θ that is thought to be plausible, and then compute the Gaussian (11) and Laplace (12) 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 since K has to be chosen large enough such that the quadrature points densely cover the range where the majority of mass in $\pi(\theta|Y)$ lies. In our case, this problem is made more severe by the presence of a dense Hessian. Martino et al. (2011) 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 for this same reason.

To mitigate the computational challenges associated with applying a manual quadrature rule for (8), we implement Adaptive Gauss-Hermite Quadrature (AGHQ). This technique has been motivated as a useful tool for Bayesian inference (Naylor and Smith, 1982) and work has been done to show that it is very accurate when using only a very small number of quadrature points (Liu and Pierce, 1994; Jin and Andersson, 2020), 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 a small number of dense Hessian matrices need to be stored in memory, an improvement over Stringer et al. (2020) 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:

$$\widehat{\theta} = \operatorname{argmax} \log \widetilde{\pi}_{LA}(\theta|y), \tag{13}$$

as well as the (low-dimensional) Hessian matrix of $\log \widetilde{\pi}_{LA}(\theta|y)$ and its Cholesky decomposition. For the optimization, we use the TMB package, which implements automatic computation of the Laplace approximation *and its gradient*, using the inverse subset algorithm which avoids repeated inner optimizations to find $\widehat{W}_\theta$ at different values of θ (Kristensen et al., 2016). Evaluating the Laplace approximation and its gradient together only use a single inner optimization of $\widehat{W}_\theta$, and further, TMB is very clever about the way that it assigns starting values, to make these converge fast.

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 and under which situations the accuracy is improved over the existing full likelihood method INLA.

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 $n = 60$ groups, each group with $n_i \equiv m$ observations. The group-level frailties $\{\xi_i, i \in [n]\}$ were simulated independently from $\mathcal{N}(0, \sigma_\xi^2)$, with $\sigma_\xi = 1$. Besides the independent frailties, we also assumed there is a covariate x generated from $\mathcal{N}(0, 1)$, 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. This choice of piecewise constant baseline function corresponds to the piecewise Exponential model, which is a type of Cox PH model that is frequently used in the literature (Friedman, 1982). The baseline hazard function in this simulation study is shown in figure 1(a). We consider six different levels of frailty sparsity in this simulation study by respectively setting m to 1, 2, 3, 4, 5 and 10. The parameters of primary interest in this study are the group level frailties.

The fixed effect β was given a prior $\mathcal{N}(0, 1000)$. The variance parameter σ_ξ was given an Exponential prior with median of 1, which is a *penalized complexity* prior of Simpson et al. (2017). 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 500 independent replications.

The comparison metrics are shown in table 1. 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 $m = 1$, both methods didn't achieve very accurate inferential results for the frailties in terms of coverage rates, because both methods involved using Laplace approximation for $\pi(\sigma_\xi|y)$, which is known to be inaccurate when the frailties are extremely sparse (Ogden, 2013). However, the result from our proposed method is still significantly more accurate than INLA as shown in table 1.

For completeness, we also implemented the same simulation setting for $\sigma_\xi = 0.4, 0.8$ and 1.3. We found that as σ_ξ gets larger, inference from INLA suffers more from the sparse frailties, especially in terms of coverage probability for ξ. On the other hand, the proposed approach tends to provide consistent coverage rate across different σ_ξ. The detailed comparison metrics for each setting of σ_ξ can be found in the appendix. These results show that unless the group level frailties have trivial variation, the proposed method yields more reliable result than INLA, especially when frailties are sparse.

The simulation result above seems to be related to the type of full likelihood INLA utlized in its inference (Martino et al., 2011). In Cox (1972a), Cox pointed out one problem associated with the use of the type of full likelihood considered by Martino et al. (2011), that is the large number of parameters introduced in order to model the unknown baseline hazard function. In this simulation study, the sample size is only $60m$, but the latent parameter contains more than 60 parameters not counting the additional parameters INLA introduced for modeling the baseline hazard, which is likely to become a problem if m is small.

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. Specifically baseline hazard function is respectively set to simple step function, oscillating step function and an extremely complicated function that switches between linear and constant. Again, the first two piecewise constant baseline hazards correspond to piecewise Exponential models with different complexity (Friedman, 1982). All of the three baseline hazards $h_0(t)$ are shown in Figure 1. 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 $u_i \overset{iid}{\sim} \text{Unif}(-6, 6)$, $i \in [n]$ and randomly censored 10% of all the survival times.

To infer the unknown risk function γ, we used the Bayesian cubic B-spline smoothing method mentioned in section 2.2 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 its second-order random walk model for γ (Lindgren and Rue, 2008). As shown in Stringer et al. (2020), the Bayesian semi-parametric smoothing methods we considered here are not sensitive to the choice of number and placement of knots. 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 modeled with an Exponential(λ) prior with λ chosen such that $\mathbb{P}(\sigma > 2) = 0.5$. For the adaptive quadrature we used in our inference, the number of grid points for variance parameter to is set to be $K = 7$.

As in the first 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 replications.

The comparison metrics under the three settings of baseline hazard are shown in figure 2 and table 2. Based on the boxplots in figure 2, the proposed method provides inference results that are at least as accurate as INLA in terms of MSE, in all the settings. The same observation can be made from the table 2, where the proposed method provides consistent coverage rate that is close to the nominal level 95% in all the settings.

While the performance of the proposed method is not affected by the choice of true baseline hazard function, the performance of INLA is sensitive to the true baseline hazard function, as shown in the correponding boxplots and table. 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 setting such as 1(c) 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 form of the baseline hazard, and hence unaffected by the wiggliness of 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 Martino et al. (2011) as well as previously by Lindgren et al. (2011); Henderson et al. (2002). The dataset contains information from $n = 1043$ 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 2.2, with fifty equally spaced knots. Prior distributions $\beta \overset{iid}{\sim} \mathcal{N}(0, 1000)$, 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 3(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 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) 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) (Monnahan and Kristensen, 2018). The runtimes respectively for the proposed method and MCMC are 1.88 minutes and 8.63 hours. Figure 3(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. 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

Therneau et al. (2003) analyzed 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. An observation for the survival time of a kidney is censored if the catheter is removed for reasons other than an infection. Each patient $i = 1, \ldots, n$ forms a group, and the survival times are the time to infection of each patient's $n_i = 2$ kidneys. This is therefore a practical example of the type of sparse frailty model on which our partial likelihood approach performed better than INLA in the simulations of section 4.1.

We first analyzed this dataset on partial likelihood using the propose method, with $\mathcal{N}(0, 1000)$ 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. For the adaptive quadrature we used in our inference, the number of grid points for variance parameter to is set to be 18 why?. Ties are handled using the method of Breslow (1974). As a comparison, we also used INLA to fit a Cox PH model to these grouped data on its full likelihood, with the same set of priors as above. The setting for baseline hazard is set to default in INLA's implementation.

Table 3 shows the results of our procedure compared to that obtained using the full likelihood method of Martino et al. (2011). 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 4(a), where our posterior distribution for σ_ξ is wider than that of INLA. As we have shown in 4.1, 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 3 shows the comparison between the results from proposed approach with the results from MCMC, in which results from the proposed method are shown to be much closer to MCMC than results from INLA. Figure 4(b) compares the MCMC samples from posterior of σ_ξ with the approximate posterior obtained using our approach. The maximal absolute difference between the cumulative posterior distribution obtained from the proposed method and the one obtained from MCMC is 0.09. The runtimes are respectively 0.53 seconds for our approach and 1.98 minutes for MCMC with 35000 iterations. The number of MCMC iterations that can be obtained with the runtime of the proposed method is only 157.

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 and the computational efficiency 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 (Ogden, 2013). 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 ef-

fect. 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 4.1.1 and 4.1.2 are available in the supplementary material with this paper. Data for example 4.2 were obtained from R package "INLA" (Rue et al., 2009) and are freely available. Data for example 4.3 were obtained from R package "survival" (Therneau, 2015) and are freely available.

References

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

Breslow, N. (1974). "Covariance analysis of censored survival data." *Biometrics*, 30(1): 89–99. 13

Cox, D. R. (1972a). "Discussion on Professor Cox's Paper." *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2): 202–220. 10

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

Dykstra, R. L. and Laud, P. (1981). "A Bayesian Nonparametric Approach to Reliability." *The Annals of Statistics*, 9(2): 356–367. 2, 3

Friedman, M. (1982). "Piecewise Exponential Models for Survival Data with Covariates." *The Annals of Statistics*, 10(1): 101 – 113. 9, 10

Geyer., C. J. (2020). *trust: Trust Region Optimization*. R package version 0.1-8.

Gray, R. J. (1992). "Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis." *Journal of the American Statistical Association*, 87(420): 942–951.

Henderson, R., Shimakura, S., and Gorst, D. (2002). "Modelling spatial variation in Leukaemia survival." *Journal of the American Statistical Association*, 97(460): 965–972. 11

Hennerfeind, A., Brezger, A., and Fahrmeir, L. (2006). "Geoadditive survival models." *Journal of the American Statistical Association*, 101(475): 1065–1075. 2

Henschel, V., Engel, J., Hölzel, D., and Mansmann, U. (2009). "A semiparametric Bayesian proportional hazards model for interval censored data with frailty effects." *BMC medical research methodology*, 9(1): 1–15. 2

Jin, S. and Andersson, B. (2020). "A note on the accuracy of adaptive Gauss–Hermite quadrature." *Biometrika*. 8

Kalbfleisch, J. D. (1978). "Non-Parametric Bayesian Analysis of Survival Time Data." *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2): 214–221. 2

Kim, Y. and Kim, D. (2009). "Bayesian partial likelihood approach for tied observations." *Journal of Statistical Planning and Inference*, 139(2): 469–477. 2, 3

Kneib, T. (2005). "Geoadditive hazard regression for interval censored survival times." 2

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). "TMB: Automatic Differentiation and Laplace Approximation." *Journal of Statistical Software*, 70(5): 1–21. 9

Lang, S. and Brezger, A. (2004). "Bayesian P-Splines." *Journal of Computational and Graphical Statistics*, 13(1): 183–212.

Lindgren, F. and Rue, H. (2008). "On the second-order random walk model for irregular locations." *Scandinavian Journal of Statistics*, 35(4): 691–700. 4, 11

Lindgren, F., Rue, H., and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 73(4): 423–498. 11

Liu, Q. and Pierce, D. A. (1994). "A note on Gauss-Hermite quadrature." *Biometrika*, 81(3): 624–629. 8

Martino, S., Akerkar, R., and Rue, H. (2011). "Approximate Bayesian inference for survival models." *Scandinavian Journal of Statistics*, 38(3): 514–528. 2, 3, 5, 6, 8, 10, 11, 12, 13

McGilchrist, C. A. and Aisbett, C. W. (1991). "Regression with frailty in survival analysis." *Biometrics*, 47(2): 461–466. 3

Miller, D. L., Glennie, R., and Seaton, A. E. (2020). "Understanding the stochastic partial differential equation approach to smoothing." *Journal of Agricultural, Biological and Environmental Statistics*, 25(1): 1–16. 4

Monnahan, C. and Kristensen, K. (2018). "No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages." *PloS one*, 13(5). 12

Naylor, J. and Smith, A. F. M. (1982). "Applications of a Method for the Efficient Computation of Posterior Distributions." *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 31(3): 214–225. 8

Ogden, H. (2013). "A sequential reduction method for inference in generalized linear mixed models." *arXiv: Computation*. 10, 13

O'Sullivan, F. (1986). "A Statistical Perspective on Ill-Posed Inverse Problems." *Statistical Science*, 1(4): 502–527. 4

Rue, H. and Martino, S. (2007). "Approximate Bayesian inference for hierarchical Gaus-

sian 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, 6, 14

Shepp, L. A. (1966). "Radon-Nikodym Derivatives of Gaussian Measures." *The Annals of Mathematical Statistics*, 37(2): 321 – 354. 4

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). 9

Sinha, D., Ibrahim, J. G., and Chen, M. (2003). "A Bayesian justification of Cox's partial likelihood." *Biometrika*, 90(3): 629–641. 2, 3

Stringer, A., Brown, P., and Stafford, J. (2020). "Approximate Bayesian inference for case-crossover models." *Biometrics*, In press. 2, 3, 4, 5, 6, 7, 8, 11

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. Version 2.38. 14

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). "Penalized Survival Models and Frailty." *Journal of Computational and Graphical Statistics*, 12(1): 156–175. 12

Tierney, L. and Kadane, J. B. (1986). "Accurate approximations to posterior moments and marginal densities." *Journal of the American Statistical Association*, 81(393). 6, 7

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. 4

Wahba, G. (1978). "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression." *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(3): 364–372. 4

Wood, S. N. (2017). "P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data." *Statistics and Computing*, 27: 985–989. 4

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.

Yue, Y. R., Simpson, D., Lindgren, F., and Rue, H. (2014). "Bayesian Adaptive Smoothing Splines Using Stochastic Differential Equations." *Bayesian Analysis*, 9(2): 397–424. 4

Number of Measurements	ξ Coverage Rate (Proposed/INLA)	ξ MSE (Proposed/INLA)	β Coverage Rate (Proposed/INLA)	β MSE (Proposed/INLA)
m = 1	0.934/0.549	0.659/0.903	0.946/0.934	0.0421/0.0266
m = 2	0.918/0.848	0.491/0.567	0.944/0.942	0.0174/0.0156
m = 3	0.934/0.919	0.350/0.370	0.950/0.952	0.0103/0.0098
m = 4	0.937/0.930	0.278/0.288	0.954/0.956	0.0072/0.0069
m = 5	0.940/0.934	0.227/0.233	0.940/0.940	0.0058/0.0056
m = 10	0.944/0.944	0.126/0.127	0.944/0.942	0.0024/0.0023

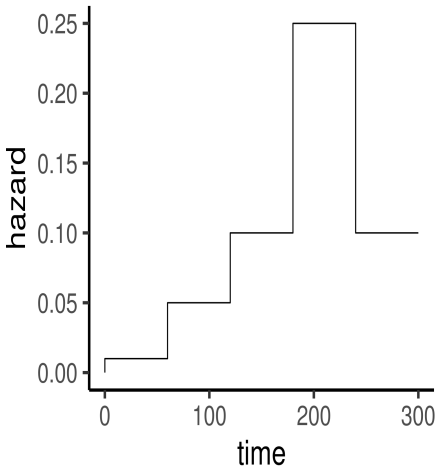
Table 1: Comparison metrics in terms of MSE and posterior coverage rate from 500 independent replications, for the 60 frailty effects and the fixed effect in the first simulation study in section 4.1.

Baseline Hazards	γ Coverage Rate (Proposed)	γ Coverage Rate (INLA)	γ MSE (Proposed)	γ 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

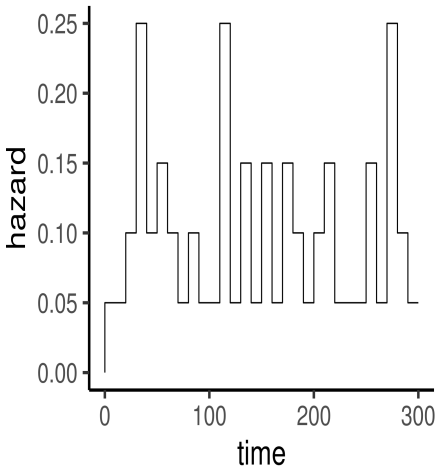
Table 2: Comparison metrics in terms of MSE and posterior coverage rate from 300 independent replications, for the three baselines in the second simulation study in section 4.1.

		Proposed		INLA		MCMC	
Variables/Reference	Levels	Mean	SD	Mean	SD	Mean	SD
Age		0.00467	0.0149	0.00235	0.0130	0.00516	0.0158
Sex/Male	Female	-1.65	0.463	-1.64	0.385	-1.72	0.507
Disease	GN	0.178	0.532	0.111	0.474	0.172	0.576
Type/Other							
	AN	0.420	0.528	0.519	0.467	0.415	0.573
	PKD	-1.15	0.817	-1.06	0.708	-1.26	0.859

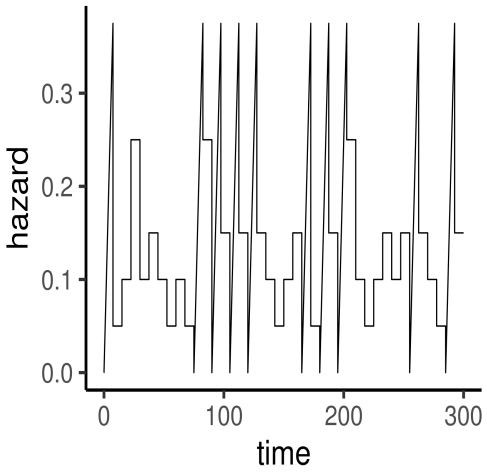
Table 3: Estimated means and standard deviations of linear effects by proposed method, INLA and MCMC for the kidney data in section 4.3.



(a) Simple stepwise baseline

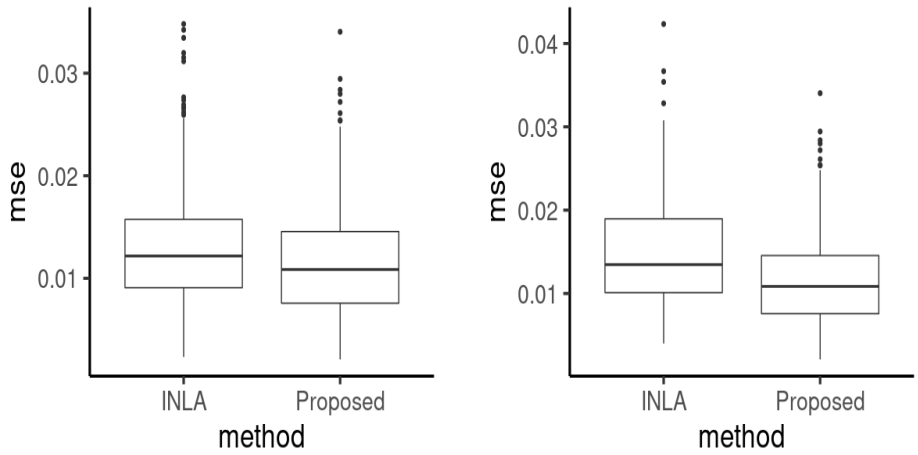


(b) Oscillating stepwise baseline



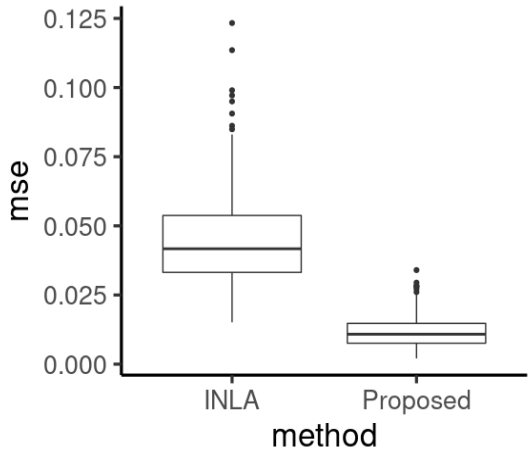
(c) Complicated baseline

Figure 1: True Baseline Hazards in the two examples in 4.1.



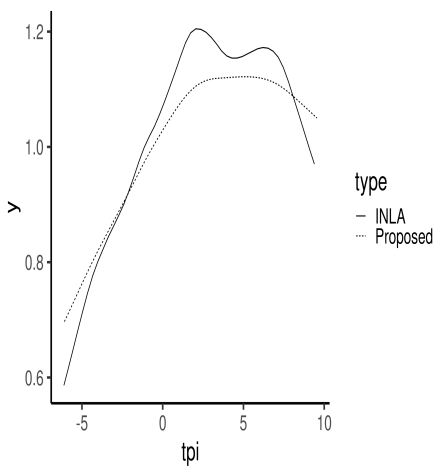
(a) Simple stepwise baseline

(b) Oscillating stepwise baseline

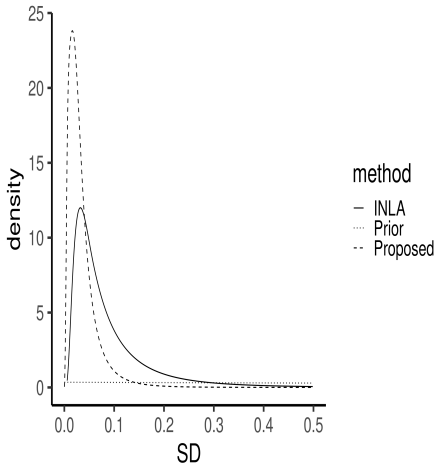


(c) Complicated baseline

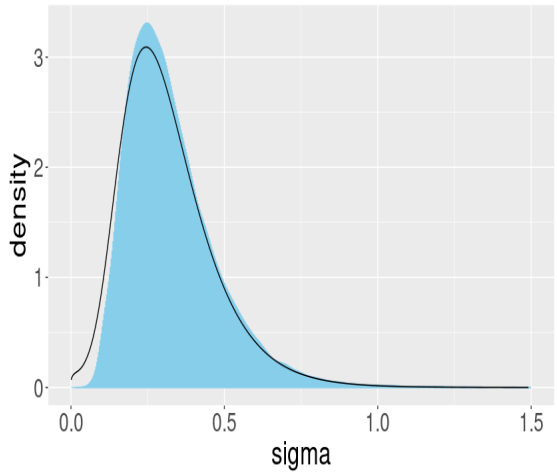
Figure 2: Results for the second simulation in section 4.1. (a): Box-plot of MSE from 300 replications with simple baseline, using INLA and the proposed method. (b): Box-plot of MSE from 300 replications with oscillating baseline, using INLA and the proposed method. (c): Box-plot of MSE from 300 replications with complicated baseline, using INLA and the proposed method.



(a) Posterior effect of `tpi`

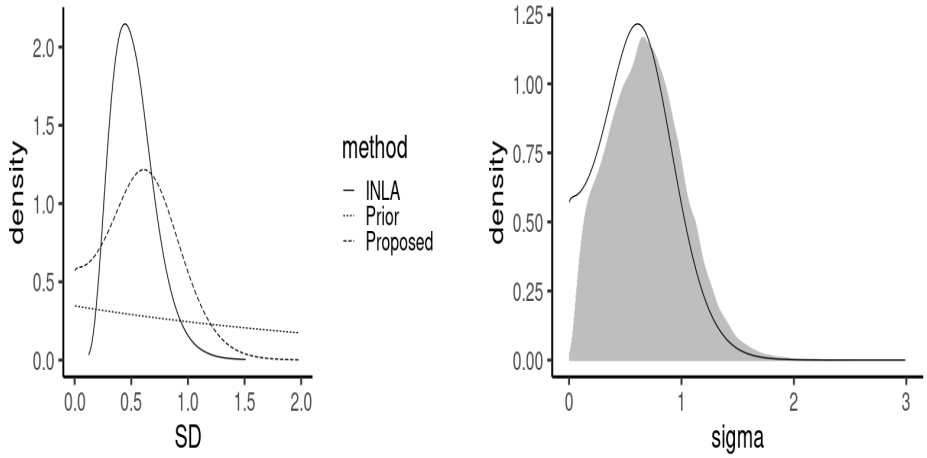


(b) Posterior distribution of σ



(c) Posterior distributions of σ

Figure 3: Results for the Leukemia data in section 4.2. (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).



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

Figure 4: Results for the kidney data in section 4.3. (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).