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 improved accuracy of the proposed partial likelihood method over the existing full likelihood method is this what the simulations show?. 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

*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

et al., 2009) cannot be applied to the Cox PH model with partial likelihood because the Hessian matrix of the log partial-likelihood is fully dense while INLA requires this matrix to be diagonal.

added paragraph break

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 assigned modelled with a extended this detail necessary? gamma process prior and the posterior distribution is derived to be another extended gamma process Did these authors show that the partial likelihood is obtained as the limit as the gamma process prior on the hazard becomes more diffuse? That would be relevant here I think. 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 with step function this isn't quite true, and isn't necessary, which implicitly requires smoothness assumptions to be made about the baseline hazard function. The literature review still only includes 3 papers. I think you read much more than this, can you expand this part? We want to be thorough here, for the benefit of the "Bayesian reader"

Stringer et al. (2020) developed an approximate Bayesian inference methodology for case-crossover models, which applies the approximation strategy of INLA to a log-partial likelihood with a non-diagonal Hessian matrix. Their methodology includes semi-parametric covariate effects and yields full posterior uncertainty for the corresponding smoothness parameters, an improvement over existing frequentist methods. Though related, t The partial likelihood they consider is simpler than that of the Cox PH model, and the Hessian matrix of their log-partial likelihood is block-diagonal and sparse. In contrast, the Hessian matrix of log-partial likelihood of Cox PH model is fully dense, so the method of Stringer et al. (2020) does not apply to this model this isn't entirely true, you could still apply it (we did, before!) it would just be horribly slow. Perhaps something like: "...is 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, a tedious operation this is true and it's not going to offend anybody since all the authors of that paper are also on this paper and agree with this. But generally when critiquing others' work it's best to avoid editorializing and to just stick to facts. which requires specialist knowledge to do properly exactly, this is the relevant objective point to make. 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. This is a bit repetative, and also we don't implement our approach on their method so it's not substantiated by the current paper. Perhaps something like: "In order to make approximate Bayesian inferences for the Cox PH model with partial likelihood, an alternative computational strategy is needed."

Recently, Stringer et al. (2021) 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 Stringer et al. (2021) 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. probably just remove this whole paragraph.

In this paper, we utilized improve upon the posterior approximation methodology of Stringer et al. (2021) Stringer et al. (2020) on and apply it to fitting Cox PH models with partial likelihood, to propose an approximate Bayesian inference method that allows the inclusion with fixed and semi-parametric covariate effects, semi-parametric smoothing effect and frailties for modelling correlations between survival times. Through two simulation studies, we illustrate that under the certain circumstances under which the proposed method based on partial likelihood would yields more reliable 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, 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 §2 we describe the semi-parametric Cox PH model and different the method of semi-parametric smoothing that will be used in this paper. In §3, we describe existing methods for approximate Bayesian inference on Cox PH model, and why the proposed inference method based on ELGM type posterior approximation of Stringer et al. (2021) should be preferred. where is the novel content? 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. Can we add several references here which use a linear predictor, $\eta_{ij} = x_{ij}^T\beta$? This would help address the referee's comment about the model construction, without having to define the model twice like they seem to suggest.

To accomodate 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}:

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

I've taken to using the $i \in [n]$ notation to represent $i = 1,\ldots,n$, I think it's cleaner. If you agree, can you make the change elsewhere, or if you disagree you can change it back Let $\eta = \{\eta_{ij} : i = 1,\ldots,n; j = 1,\ldots,n_i\}$ be the vector of all the additive linear predictors. Here x_{ij} is a p-dimensional vector of covariates that are modeled as having linear associations with the log-hazard, and $\beta = (\beta_1,\ldots,\beta_p)$ perhaps should use boldface for vectors, this notation is a bit confusing. Also, have to be consistent with notation for vectors, here it's $(\beta_1,\ldots,\beta_p)$ but everywhere else we're using the "implicit" notation are regression coefficients. The $u_q = \{u_{qij} : i = 1,\ldots,n; j = 1,\ldots,n_i\}, q = 1,\ldots,r$ are covariate vectors whose association with the log-hazard is modeled semi-parametrically through unknown smooth functions $\gamma_1,\ldots,\gamma_r$. The vector of group intercepts $\xi = \{\xi_i : i = 1,\ldots,n\}$—referred to as *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 $\{\gamma_q\}_{q=1}^r$ use consistent vector notation. You don't have to choose the one I suggested, but you do have to choose one. 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 2-fold Integrated Wiener process(Shepp, 1966), which has a connection to cubic smoothing splines (Wahba, 1978). To infer the infinite-dimensional parameters $\{\gamma_q\}_{q=1}^r$, Lindgren and Rue (2008) 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 this is what we wrote in the case crossover paper, but it's not quite true. 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 Lindgren and Rue (2008) can be understood as a special case of Bayesian penalizing regression splines, with basis being linear B splines, and penalty matrix being $\Sigma_q^{-1}(\sigma_q)$ (Miller et al., 2020). 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 function values vector $\Gamma_q()$, and hence will be incompatible with the type of Laplace approximation in Tierney and Kadane (1986), which we utilized in the proposed approach.

here is my suggested rewrite of these two paragraphs, starting from "To infer the infinite-dimensional parameters...":

The unknown processes $\gamma_q, q \in [r]$ are infinite-dimensional, complicating inference. To proceed, each process $\gamma(\cdot)$ (dropping the subscript q) is modelled using a finite-dimensional basis function expansion of the form $\gamma(u) = \sum_{j=1}^d \phi_j(u)\Gamma_j$ where $\Gamma_j, j \in [d]$ are parameters to be inferred and $\phi_j(\cdot), j \in [d]$ are fixed, known basis functions which must be chosen. Lindgren and Rue (2008) show how a choice of linear B-spline basis functions, combined with the Gaussian process prior, lead to a generalization of the second-order Bayesian P-Spline (Lang and Brezger, 2004) to unequally-spaced covariates, and provide an explicit construction of the precision matrix $\Sigma^{-1}(\sigma)$. 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. Here, we use cubic B-splines for the $\phi_j(\cdot)$ and choose a covariance function whose precision matrix is that obtained by using an integrated second derivative penalty of Wood (2017). Note that owing to the established connection between penalized smoothing and Gaussian processes (Lindgren et al., 2011; Miller et al., 2020), the Gaussian process itself does not need to be constructed, only its precision matrix.

To fix the 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 (Stringer et al., 2020; Rue et al., 2009). 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 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 $\Sigma_q^{-1}(\sigma_q)$ by a very small constant, hence will not change any conditional independence structure in the original prior. I suggest putting this in section 3, see below.

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 $\theta \sim \pi(\theta)$. I suggest putting this in section 4.

3 Methods

3.1 Approximate Bayesian Inference

partial likelihood in section 2, and then start with the Δ stuff here?

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.

start section 3 here? Everything before this point is review, everything after it is new.

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 $\Delta = \{\Delta_{ij} : i = 1, \ldots, n; j = 1, \ldots, n_i\}$. 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 $\ell(\Delta; y) = \log \pi(y|\Delta)$.

This section has to be modified to incorporate the fact that we don't add the noise term

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 (Rue et al., 2009; Martino et al., 2011; Stringer et al., 2020), 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 (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 feasible for the inference on Cox PH model with partial likelihood.

The recent posterior approximation method of Stringer et al. (2021) 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 size 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:

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

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

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

This appears primarily as an intermediate step in the calculation of the marginal posteriors (6) but is also useful in calculating complicated posterior summary statistics such as joint credible bands.

All of the quantities of interest (6) – (8) depend on intractable high-dimensional integrals. Stringer et al. (2020) utilize Gaussian and Laplace approximations combined with numerical quadrature to approximate each of these integrals accurately and efficiently. Their approximations take the form:

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

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. In the present work we replace these with the nodes and weights from an adaptive Gauss-Hermite rule, which requires less user input. 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.

How much of this is necessary? A lot of it is straight out of the case crossover paper, and some of it is even verbatim from that paper which is not ideal. Stuff from CC that's necessary should go in section 2.

In the proposed method, the approximations (9) are computed as follows. For any fixed θ, define

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

For the conditional posterior

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

a second-order Taylor expansion of $\log\pi(W|\theta,y)$ 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:

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

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

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

With these approximations available, inference for W makes use of the approximation (9).

3.2 Adaptive quadrature and automatic differentiation

Regarding the following, I think there are several novelties that can be addressed:

- The lack of noise term,
- The use of adaptive quadrature (this is well addressed),
- The use of TMB's automatic differentiation and Laplace approximation.

All three of these are methodological innovations over the case crossover paper. The noise term results in smaller matrices; the adaptive quadrature reuslts in the procedure being more automatic/requiring less specialist input; and the automatic differentiation and Laplace approximation results in a very efficient procedure (see comments below). A clear narrative of these three challenges with CC and then our solutions should be established.

Computing the approximations (9) requires choosing a quadrature rule consisting of nodes $\left\{\theta^k\right\}_{k=1}^K$ and weights $\left\{\delta_k\right\}_{k=1}^K$ 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 (12) and Laplace (14) 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 (9), 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:

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

as well as the (low-dimensional) Hessian matrix of $\log \widetilde{\pi}_{LA}(\theta|y)$ and its Cholesky decomposition. These matrix quantities are straightforward to obtain as θ is low-dimensional. For the optimization, we follow Rue et al. (2009) and use numerical derivatives and a BFGS algorithm This is not true anymore. We use TMB, which implements automatic computation of the Laplace approximation *and its gradient*, using the "inverse subset algorithm" which avoids repeated "inner" optimizations; see the TMB software paper. Evaluating the Laplace approx and its gradient together only use a single inner optimization, and further, TMB is very clever about the way that it assigns starting values, to make these converge fast. which limits the number of derivatives which must be computed. Computations of the numerical integrations make use of the `aghq` package (Stringer, 2020), and the required first two derivatives of $\log \pi(W|\theta, y)$ are computed using the automatic differentiation method of the TMB package (Kristensen et al., 2016).

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 and 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 $n = 100$ groups, each group with $n_i \equiv n_1 m$ we already defined group sizes to be n_i, not sure the best way to notate them when they're all equal observations. The group-level frailties $\{\xi_1, ..., \xi_n\}$ be consistent with vector notation were simulated independently from $\mathcal{N}(0, \sigma_\xi^2)$, with $\sigma_\xi = 0.8$ why was this σ chosen? This seems like something that could be varied and that would be likely to change the results. Besides the independent frailties, we also assumed there is a covariate x generated from $\mathcal{N}(0, 9)$, 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 Isn't the use of a non-smooth baseline likely to contribute to partial likelihood doing better than INLA?. The baseline hazard function in this simulation study is shown in figure 1. We consider five different levels of frailty sparsity in this simulation study by respectively setting m to 2, 4, 6, 8 and 10.

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. 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 $K = 15$ Do you need this many? That's like $O(n^{-5})$ error or something. 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 replications in order to make the comparison accurate.

The comparison metrics are shown in table 1 and 2. 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 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 2.

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 in the full likelihood (Cox, 1972a). 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. Is this paragraph necessary? What is it trying to say?

These simulation results are nice. Can I suggest using $m = 2, 3, 4, 5, 10$ perhaps, since for $m = 2$ and $m = 4$ the results are quite different? I also think you can be a bit more enthusiastic, even when $m = 2$ ours does pretty well, especially for β, while INLA does bad for both.

I think unless it's a crazy amount of work, results for different combinations of m and σ_ξ should be added. I think σ_ξ should have a big effect on how hard it is to estimate ξ_i in general, and choosing one value only obscures this. Also, you might want to prepare yourself for the reviewers to ask the whole thing to be re-run with different n as well. I would suggest pre-emptively writing the code in a way such that it is really easy to change values and collate the results, if you haven't already.

Simulation with non-smooth baseline

To illustrate the accuracy of our method over INLA when the smoothness assumption for baseline hazard function is violated I don't think this simulation convincingly shows that we're more accurate than INLA under any reasonble baseline hazard. I think it shows that the smoothness assumption has to be really strongly violated for this to become a problem for INLA. I would suggest either coming up with more reasonable "bad" baselines, or changing the narrative to say that having sparse frailties/other difficult to estimate models is more of a reason to use partial likelihood than if you think the baseline hazard is non-smooth, 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. The three baseline hazards $h_0(t)$ are shown in Figure 2. 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 $\sum_{i=1}^n \gamma(u_i) = 0$. should this point instead be mentioned in section 2?

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 do you need 50 knots? The splines should work with many less knots, like 5 or 10. I'm not trying to be too picky here but I wonder if this contributes to some of the instability you've been seeing. 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 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 $\mathbb{P}(\sigma > 2) = 0.5$, which is a *penalized complexity* prior of Simpson et al. (2017) The level of detail here should match that from the other simulation description, either less here or more there. 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 3 and table 3. 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% whereas the proposed method yielded coverage rate of 96.8%. 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. Your description here makes sense but remember the comments from the reviewer from the stats in medicine version. The complicated baseline hazard has no reasonable physical justification. So we haven't shown some "aha! gotcha!" on INLA here, what we've actually demonstrated is that INLA with full likelihood is actually pretty robust with respect to the form of the baseline hazard.

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 $n = 1043$ 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 2.2, with fifty again, need 50? 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 4(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 4(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 full-likelihood using INLA, 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 I think this is probably going to have a big impact on inference. Have you tried different values? The referees might ask and it would be good to have an understanding of whether this matters. The setting for baseline hazard is set to default in INLA's implementation. As a comparison, why is INLA the main story and we're the 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 why?. Ties are handled using the method of Breslow (1974).

Table 4 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 5(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 5 shows the comparison between the results from proposed approach with the results from MCMC. Figure 5(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

We can do the discussion later once we're done the rest fo the paper. Comments about the figures and tables:

- Some of the figures are unreadable. I know it's tedious but you have to make sure all the figures look nice, this is extremely important. Labels the proper size and not stretched, all text sizes and other aesthetics exactly the same across all figures, no colour.
- Tables 4 and 5 should be one table, the leftmost 3/4 of them are identical.
- Figures 1 and 2 can be combined into one four-panel figure.

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 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 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. 15

Cox, D. R. (1972a). “Discussion on Professor Cox’s Paper.” *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2): 202–220. 11

— (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

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. 13

Jin, S. and Andersson, B. (2020). “A note on the accuracy of adaptive Gauss–Hermite quadrature.” *Biometrika*. 10

Kim, Y. and Kim, D. (2009). “Bayesian partial likelihood approach for tied observations.” *Journal of Statistical Planning and Inference*, 139(2): 469–477. 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. 10

Lang, S. and Brezger, A. (2004). "Bayesian P-Splines." *Journal of Computational and Graphical Statistics*, 13(1): 183–212. 5

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, 5, 12

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. 5, 13

Liu, Q. and Pierce, D. A. (1994). "A note on Gauss-Hermite quadrature." *Biometrika*, 81(3): 624–629. 10

Martino, S., Akerkar, R., and Rue, H. (2011). "Approximate Bayesian inference for survival models." *Scandinavian Journal of Statistics*, 38(3): 514–528. 2, 3, 6, 7, 9, 13, 14, 15

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. 5

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). 14

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. 10

Ogden, H. (2013). "A sequential reduction method for inference in generalized linear mixed models." *arXiv: Computation*. 11, 16

O'Sullivan, F. (1986). "A Statistical Perspective on Ill-Posed Inverse Problems." *Statistical Science*, 1(4): 502–527. 5

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

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(2): 319 – 392. 1, 5, 7, 10, 16

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). 12

Stringer, A. (2020). "Implementing Approximate Bayesian Inference using Adaptive Quadrature: the aghq Package." *In preparation*. 10

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, 9, 10

— (2021). "Fast, Scalable Approximations to Posterior Distributions in Extended Latent Gaussian Models." 3, 7

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. Version 2.38. 16

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. 14

Tierney, L. and Kadane, J. B. (1986). "Accurate approximations to posterior moments and marginal densities." *Journal of the American Statistical Association*, 81(393). 5, 9

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. 5

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. 5

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 4.1.

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 4.1.

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 three baselines in the second simulation study in section 4.1.

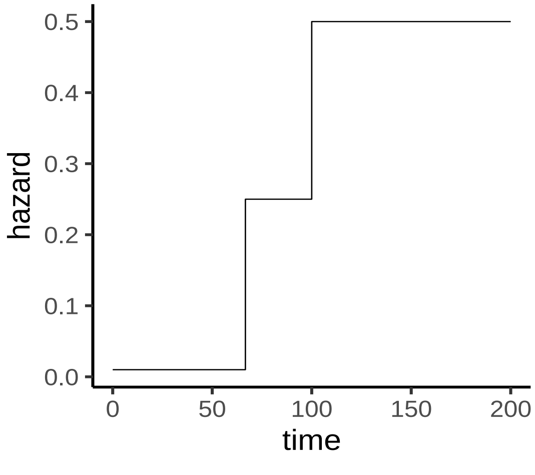
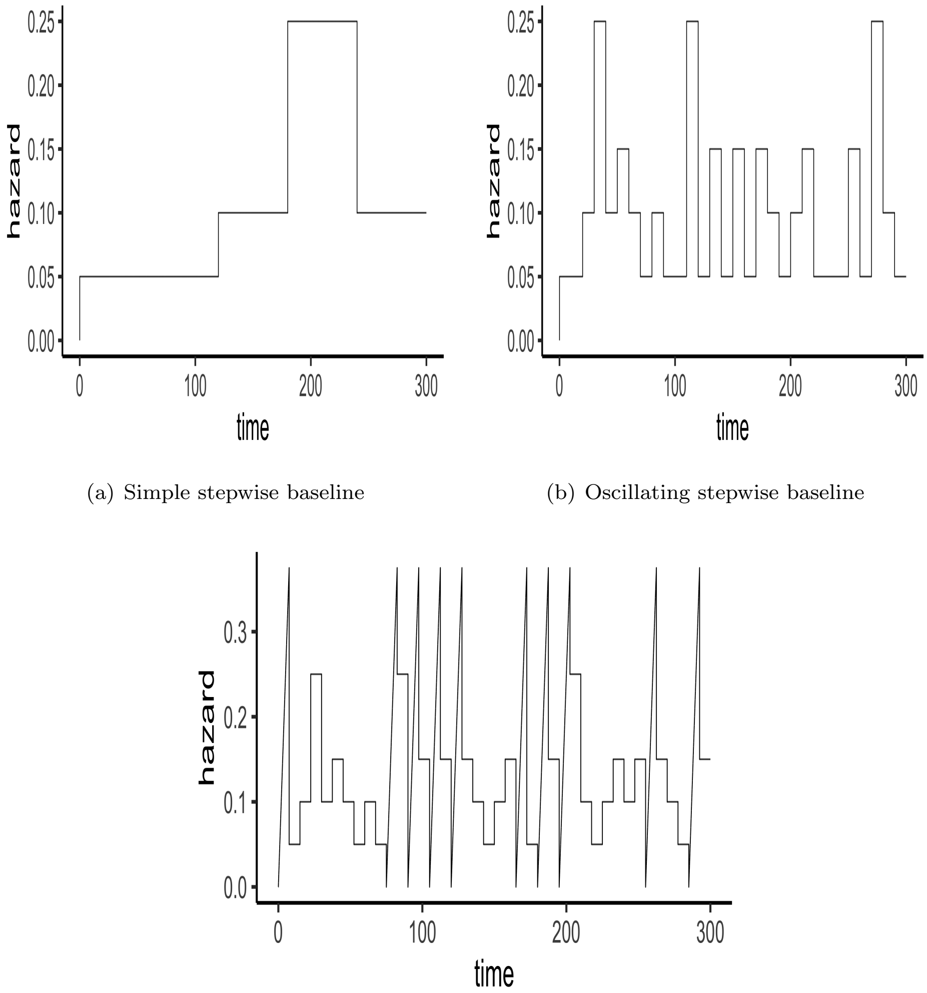


Figure 1: True Baseline Hazard in the first example in 4.1.

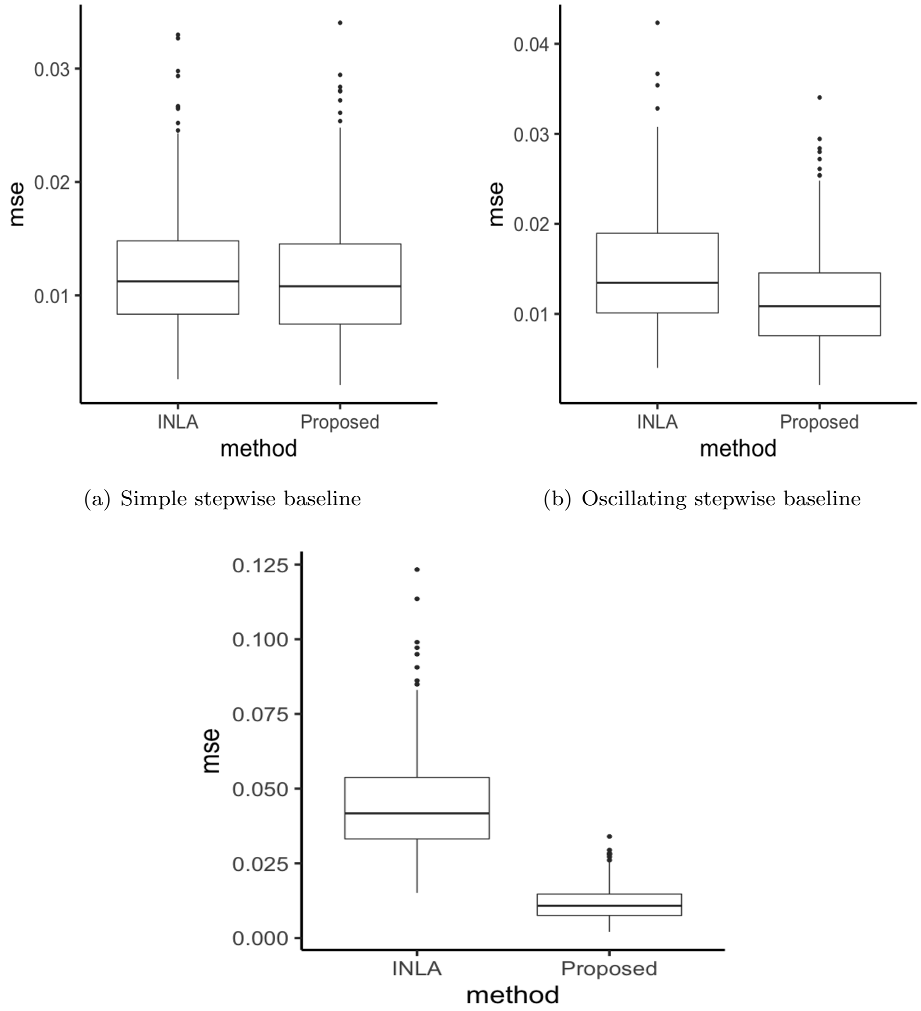


(a) Simple stepwise baseline

(b) Oscillating stepwise baseline

(c) Complicated baseline

Figure 2: True Baseline Hazards in the second example in 4.1.

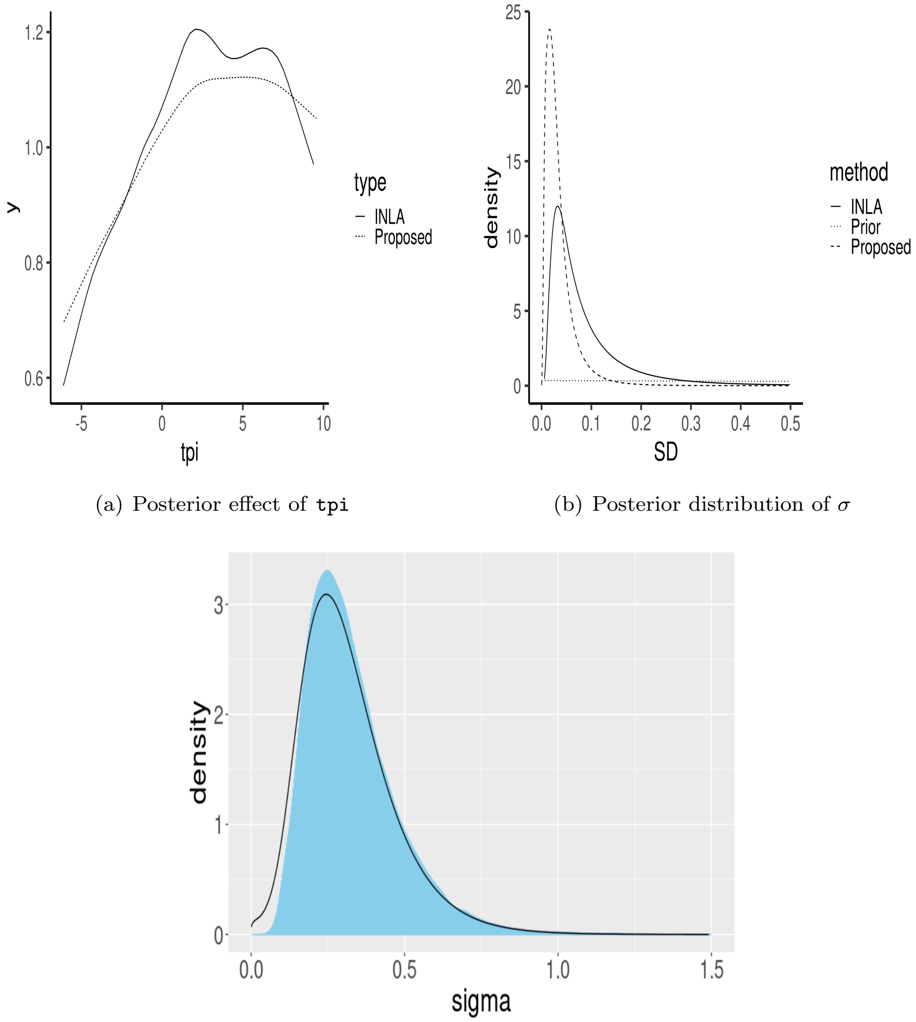


(a) Simple stepwise baseline

(b) Oscillating stepwise baseline

(c) Complicated baseline

Figure 3: 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

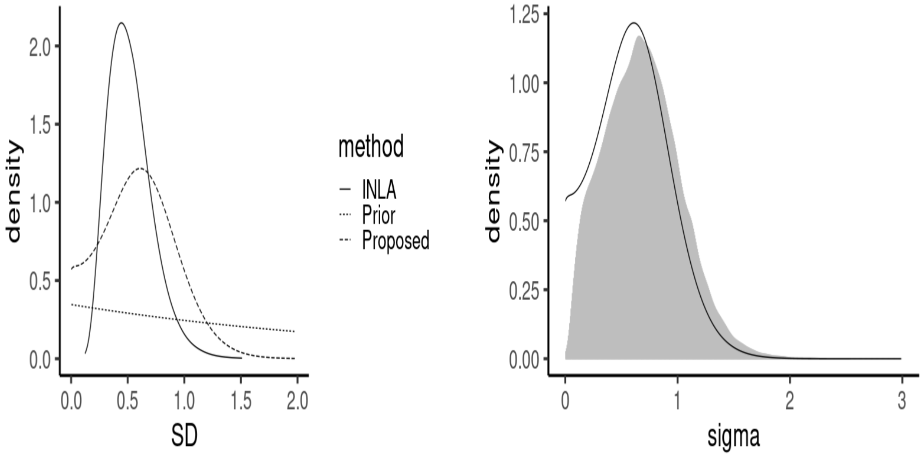
Figure 4: Results for the Leukemia data in section 4.2. (a): (Exponentiated) posterior mean for the semi-parametric tpi effect using our proposed method(dashed) and INLA(solid). (b): Prior(dotted) and posterior distributions for σ using our proposed method(dashed) and INLA(solid). (c): Posterior distribution for σ obtained using MCMC(gray histogram), and using the proposed method(black line).

		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 4.3.

		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 4.3.



(a) Posterior distributions of σ

(b) Posterior distributions of σ

Figure 5: 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).