Lifetime Data Anal (2015) 21:470–490
DOI 10.1007/s10985-014-9305-9

A Bayesian proportional hazards model for general interval-censored data

Xiaoyan Lin · Bo Cai · Lianming Wang · Zhigang Zhang

Received: 2 August 2013 / Accepted: 23 July 2014 / Published online: 7 August 2014

Abstract The proportional hazards (PH) model is the most widely used semiparametric regression model for analyzing right-censored survival data based on the partial likelihood method. However, the partial likelihood does not exist for interval-censored data due to the complexity of the data structure. In this paper, we focus on general interval-censored data, which is a mixture of left-, right-, and interval-censored observations. We propose an efficient and easy-to-implement Bayesian estimation approach for analyzing such data under the PH model. The proposed approach adopts monotone splines to model the baseline cumulative hazard function and allows to estimate the regression parameters and the baseline survival function simultaneously. A novel two-stage data augmentation with Poisson latent variables is developed for the efficient computation. The developed Gibbs sampler is easy to execute as it does not require imputing any unobserved failure times or contain any complicated Metropolis-Hastings steps. Our approach is evaluated through extensive simulation studies and illustrated with two real-life data sets.

Keywords Interval-censored data · Monotone splines · Nonhomogeneous Poisson process · Proportional hazards model · Semiparametric regression

L. Wang · X. Lin (✉)
Department of Statistics, University of South Carolina, Columbia, SC 29208, USA
e-mail: lin9@mailbox.sc.edu

B. Cai
Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC 29208, USA

Z. Zhang
Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA

1 Introduction

The proportional hazards (PH) model (Cox 1972) is the most widely used semiparametric regression model in the survival literature. It specifies that covariates have a multiplicative effect on the hazard function of the failure time of interest as follows,

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}), \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function, $\boldsymbol{x}$ is a $p \times 1$ covariate vector, and $\boldsymbol{\beta}$ is a vector of regression coefficients. The corresponding survival function takes the form $S(t) = \exp\{-\Lambda_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta})\}$ with $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ denoting the baseline cumulative hazard function. The popularity of the PH model is due to the following facts. First, the PH model has a nice interpretation of the regression parameters in terms of the log hazard ratio. Second, the PH model has great modeling flexibility by allowing the baseline hazard function to be unspecified. Third, for right-censored data, the partial likelihood method (Cox 1975) under the PH model allows one to estimate the regression parameters directly without the need of estimating the hazard function. Most statistical software packages have built-in functions for conducting the partial likelihood method and lead the PH model widely used even by non-statisticians. However, the partial likelihood method does not exist for interval-censored data under the PH model (Sun 2006).

Interval-censored data commonly occur in many fields such as demographical, epidemiological, and medical studies. In such studies, participants usually undergo periodical observations or examinations, and the failure time of interest is not observed exactly but is known to fall within some interval. For example, the onset time of HIV for a participant is usually interval censored and the observed interval is formed by the last examination time with negative status and the first examination time with positive status for that participant. Analyzing interval-censored data is challenging due to the complicated data structure being a mixture of left-, right-, and interval-censored observations. Also, the observation process that results in interval-censored data varies from study to study, making the inference on interval-censored data challenging from both theoretic and computational perspectives.

Many approaches have been developed for the regression analysis of interval-censored data under the PH model. Finkelstein (1986) was the first to investigate this problem and proposed to use a Newton-Raphson algorithm to estimate the regression parameters and the baseline hazard function jointly. Aiming to avoid estimating the baseline hazard function, Satten (1996) proposed a marginal likelihood approach, Goggins et al. (1998) developed a Monte Carlo EM algorithm, and Satten et al. (1998) proposed estimating equations, all based on the possible rankings of failure times that are consistent with the observed data. Pan proposed a generalized gradient projection method (Pan 1999) and a multiple imputation method (Pan 2000). Cai and Betensky (2003) developed a penalized likelihood approach and modeled the logarithm of the baseline hazard function with a linear spline. Zhang et al. (2010) proposed a seive maximum likelihood method by adopting B-splines for the logarithm of the cumulative baseline hazard function. Sinha et al. (1999) used piecewise constant functions to model the baseline hazard func-

tion, and Yavuz and Lambert (2011) modeled the baseline density function with penalized B-splines. Recently, Wang et al. (2013) proposed a dynamic regression model with time-varying coefficients for interval-censored data. A comprehensive review of the PH model for interval-censored data can be found in Zhang and Sun (2010).

Gomez et al. (2009) gave a pedagogical tutorial on methods for interval-censored data and their implementation in R. However, to our knowledge, there are only three existing statistical packages, `intcox` (Henschel et al. 2009a), `survBayes` (Henschel et al. 2009b), and `dynsurv` (Wang et al. 2013), that perform semiparametric regression analysis for interval-censored data under the PH model. The R package `intcox` adopts the generalized gradient projection method of Pan (1999), but is of limited use because it only provides the point estimates of the regression parameters without estimated variances. The R package `survBayes` implements a Bayesian method that models the log baseline hazard function with cubic B-splines and imputed the unknown survival times using a piecewise exponential distribution conditioned on the respective interval. However, `survBayes` is observed computationally demanding and slow from our simulations. The recently developed R package `dynsurv` provides functions to fit time-varying coefficient models. However, they use piecewise constants to model the baseline and coefficient functions; therefore their estimates of the baseline and coefficient functions are discrete.

In this paper, we propose a novel and efficient Bayesian approach for analyzing general interval-censored data under the PH model. The approach can be regarded as a novel member of a series of approaches (Lin and Wang 2010, 2011; Wang and Lin 2011; Cai et al. 2011) as they share two common strategies. First, they all adopt monotone splines (Ramsay 1988) to model certain unknown baseline nondecreasing functions, and therefore to produce smooth estimates of baseline functions. Specifically, Lin and Wang (2010) used the monotone splines to model the transformed baseline cumulative distribution under the probit model; Lin and Wang (2011) and Wang and Lin (2011) modeled the baseline odds and the logarithm of the baseline odds, respectively, under the proportional odds (PO) model. This paper models the baseline cumulative hazard function under the PH model as in Cai et al. (2011). Second, they all adopt certain data augmentation to facilitate Bayesian computation. The probit model (Lin and Wang 2010) introduced normal latent variables; the PO model (Wang and Lin 2011) developed two types of normal mixture data augmentation based on the relationship between the PO model and the logistic distribution. Cai et al. (2011) developed a simple two-stage Poisson data augmentation for current status data under the PH model. This paper extends Cai et al.'s data augmentation to a more general and complicated interval-censored data structure based on the relationship between the PH model and a latent nonhomogeneous Poisson process. The newly developed Poisson latent variable augmentation can be extended to the PO model and the generalized odds-rate hazards models (Banerjee et al. 2007) by introducing a gamma frailty term. Thus, it can serve as a general data augmentation framework to facilitate Bayesian methods for analyzing interval-censored data under such models. It can also be extended to handle clustered or multivariate interval-censored data under the frailty PH and PO models. Unlike many existing

Bayesian methods in the literature, our developed Gibbs sampler is easy to execute with only four steps and efficient because it does not require imputing any unobserved failure times or contain any complicated Metropolis-Hastings steps. Furthermore, our R package `ICBayes` provides practitioners an easy access to the proposed approach.

The remainder of this paper is organized as follows. Section 2 discusses the general data structure and the associated likelihood functions. Section 3 provides the details about the proposed methods, including the use of monotone splines for $\Lambda_0(t)$, the two-stage Poisson data augmentation, the prior specification, and the posterior computation. Section 4 shows the simulation results of the proposed approach and the comparison with several existing approaches. Section 5 presents two real-life data applications. Section 6 concludes with some discussions.

2 Data and the likelihoods

Suppose that there are n independent subjects in the study. For each subject i, let T_i denote the failure time of interest and $\boldsymbol{x}_i$ a covariate vector. Due to the design or some special property of the failure event, T_i is never observed exactly but only known to fall in an observed interval $(L_i, R_i]$. Let $F(\cdot|\boldsymbol{x})$ denote the cumulative distribution function (CDF) of the failure time of interest given covariate $\boldsymbol{x}$. Then $F(t|\boldsymbol{x}) = 1 - \exp\{-\Lambda_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta})\}$ under the PH model. Under noninformative censoring, i.e., the failure time is independent of the observation process given covariates, the likelihood based on the observed data $\{L_i, R_i, \boldsymbol{x}_i\}_{i=1}^n$ is

$$L_{obs} = \prod_{i=1}^{n} \Big\{ F(R_i|\boldsymbol{x}_i) - F(L_i|\boldsymbol{x}_i) \Big\}.$$

Note that the observed interval $(L_i, R_i]$ for the failure time T_i takes $(0, R_i]$ in the case of left-censoring, and (L_i, ∞) in the case of right-censoring. To distinguish the types of censoring, we reformulate the observed likelihood as follows,

$$L_{obs} = \prod_{i=1}^{n} F(R_i|\boldsymbol{x}_i)^{\delta_{i1}} \Big[F(R_i|\boldsymbol{x}_i) - F(L_i|\boldsymbol{x}_i) \Big]^{\delta_{i2}} \Big[1 - F(L_i|\boldsymbol{x}_i) \Big]^{\delta_{i3}}, \qquad (2)$$

where δ_{i1}, δ_{i2}, and δ_{i3} are the censoring indicators for subject i denoting left-, interval-, and right-censoring, respectively, subject to the constraint $\delta_{i1} + \delta_{i2} + \delta_{i3} = 1$. Our proposed approach is based on the observed likelihood (2). The observed likelihood is promising to use because it does not require any specific assumptions on the distribution of the observation process. When the distribution of the observation process does not contain the parameters of interest, $\boldsymbol{\beta}$ and Λ_0, the observed likelihood makes an efficient inference of $\boldsymbol{\beta}$ and Λ_0 (Sun 2006).

3 The proposed approach

3.1 Modeling $\Lambda_0(t)$ with monotone splines

Estimation of the unknown nondecreasing function $\Lambda_0(t)$ is challenging when the observed intervals are overlapping and different, a common feature for many real-life interval-censored data sets. In this case, the number of unknown parameters involved is on the order of sample size, which causes great estimation difficulty from both theoretic and computational perspectives. To reduce the number of parameters while also allow adequate modeling flexibility, we model $\Lambda_0(t)$ by a linear combination of monotone splines of Ramsay (1988) as in Joly et al. (1998) and Cai et al. (2011). This strategy has been effectively used to model unknown nondecreasing functions in other survival models, such as modeling the transformed baseline cumulative distribution in the probit model (Lin and Wang 2010) and the baseline odds (Lin and Wang 2011) or the logarithm of the baseline odds (Wang and Dunson 2011) in the PO model. Specifically, we model the baseline cumulative hazard function

$$\Lambda_0(t) = \sum_{l=1}^{k} \gamma_l I_l(t \,|\, d), \tag{3}$$

where $I_l(\cdot \,|\, d)$s are monotone spline basis functions with degree d, each of them is nondecreasing from 0 to 1, and γ_ls are nonnegative spline coefficients to ensure that $\Lambda_0(t)$ is nondecreasing. The degree d controls the smoothness of the splines, taking 1 for piecewise linear functions, 2 for quadratic functions, and 3 for cubic functions, etc.

Monotone splines are also called I- (or integrated) splines because they are the integrated functions of M splines (Ramsay 1988). To make our paper self-contained, we present below how one can construct monotone spline basis functions within a target interval $[a, b]$. First, one needs to choose the degree d and set up m interior knots $\xi_1 < \cdots < \xi_m$ within $[a, b]$. Then one can construct M splines using the following recursive formulas: let $s_1 = a$, $s_2 = \xi_1$, $\cdots$, $s_{m+1} = \xi_m$, and $s_{m+2} = b$, then for $l = 1, \ldots, m + 1$,

$$M_l(t|1) = \begin{cases} \frac{1}{s_{l+1} - s_l}, & \text{for } s_l \le t < s_{l+1}, \\ 0, & \text{elsewhere;} \end{cases}$$

for $d \ge 2$, let $s_1 = \cdots = s_d = a$, $s_{d+1} = \xi_1$, $\cdots$, $s_{d+m} = \xi_m$, and $s_{m+d+1} = \cdots = s_{m+2d} = b$, then for $l = 1, \ldots, m + d$,

$$M_l(t|d) = \begin{cases} \frac{(t - s_l) M_l(t|d-1) + (s_{l+d} - t) M_{l+1}(t|d-1)}{(s_{l+d} - s_l)} \cdot \frac{d}{d-1}, & \text{for } s_l \le t < s_{l+d}, \\ 0, & \text{elsewhere.} \end{cases}$$

Each $M_l(\cdot|d)$ is a piecewise polynomial with nonzero only within $[s_l, s_{l+d})$ for $l = 1, \ldots, m + d$. Last, one can obtain monotone spline basis functions by using $I_l(t|d) = \int_a^t M_l(u|d)du$ for each l. Specifically, for $t \in [s_j, s_{j+1})$, $I_l(t|d)$ takes the following form,

$$I_l(t|d) = \begin{cases} 0, & l > j, \\ \sum_{h=l}^{j} (s_{h+d+1} - s_h) \frac{M_h(t|d+1)}{d+1}, & j - d + 1 \le l \le j, \\ 1, & l < j - d + 1, \end{cases}$$

for each $l = 1, \ldots, m + d$. As seen from the above construction, the spline basis functions are totally determined once the degree and knots are specified. The number k of basis functions is equal to the number of interior knots m plus the degree d.

In specifying the monotone splines, we recommend to use 2 or 3 for the degree to provide adequate smoothness. In the following, we drop d and use $I_l(\cdot)$ directly for notational convenience. In terms of the placement of knots, one may use random knots placement, i.e., to treat the number and positions of knots as random. However, using a random knot placement leads to a great computation burden for deciding to add or drop a knot, or change a knot position, and for re-calculating and evaluating the basis functions whenever a change occurs. In contrast, using a set of appropriately selected fixed knots requires much less computational efforts while maintaining enough modeling flexibility. In general, using 10–30 knots (equally-spaced or based on quantiles) provides adequate modeling flexibility for data sets containing up to thousands of observations (Cai et al. 2011; Wang and Dunson 2011). It is worth noting that the shrinkage priors for the spline coefficients in our method function to prevent overfitting problems that may be caused by using too many knots. Finally, Bayesian model comparison criteria such as the deviance information criterion (DIC) (Spiegelhalter et al. 2002) and log pseudo marginal likelihood (LPML) (Ibrahim et al. 2001) can help with the selection of the degree and knots. In the simulation, we use the two model comparison criteria to compare the two strategies of specifying knots: taking equally-sapced knots or using quantiles. We also use the two model comparison criteria to choose the best setup of the degree and the number of knots for the real data analyses in Sect. 5.

3.2 A two-stage Poisson data augmentation

Bayesian methods typically require sampling all the unknown parameters and latent variables from their posterior distributions formed by combining the likelihood function and the prior distributions. However, using the complicated likelihood function (2) plus priors does not lead to standard posterior distributions for those unknown parameters, which makes sampling challenging. Existing Bayesian approaches typically adopt many complicated Metropolis-Hastings steps as seen in Yavuz and Lambert (2011) and Henschel et al. (2009b) among others. Such Metropolis-Hastings steps may lead to poor mixing of Markov chains in addition to a high computational cost.

To facilitate the posterior computation, we develop a novel data augmentation by taking advantage of the relationship between the PH model and a latent nonhomogeneous Poisson process. Based on the appropriately selected Poisson latent variables from the Poisson process, the augmented data likelihood expands the observed likelihood (2) and has a nice form for sampling.

Specifically, let $N(t)$ be a latent nonhomogeneous Poisson process with a cumulative intensity function $\Lambda_0(t) \exp(\boldsymbol{x}'\boldsymbol{\beta})$. Define $T = \inf\{t : N(t) > 0\}$, the time of the

first occurrence in the Poisson process. Then it is clear that for any t,

$$P(T > t) = P\{N(t) = 0\} = \exp\{-\Lambda_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta})\},$$

indicating that T indeed follows the PH model. For any two time points t_1 and t_2 with $t_1 < t_2$, let $z = N(t_1)$, the count of occurrences up to time t_1, and $w = N(t_2) - N(t_1)$, the count of occurrences between t_1 and t_2, respectively. Then, by the independent increments property of a nonhomogeneous Poisson process, z and w are independent Poisson random variables with $z \sim \mathcal{P}\{\Lambda_0(t_1)\exp(\boldsymbol{x}'\boldsymbol{\beta})\}$ and $w \sim \mathcal{P}[\{\Lambda_0(t_2) - \Lambda_0(t_1)\}\exp(\boldsymbol{x}'\boldsymbol{\beta})]$, respectively. Here $\mathcal{P}(a)$ denotes the Poisson distribution with mean parameter a.

Considering that there is a nonhomogeneous Poisson process $N_i(t)$ with a cumulative intensity function $\Lambda_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$ associated with each subject i, for $i = 1, \ldots, n$, we formalize the data augmentation as follows. Let $t_{i1} = R_i 1_{(\delta_{i1}=1)} + L_i 1_{(\delta_{i2}=1)} + L_i^- 1_{(\delta_{i3}=1)}$ and $t_{i2} = R_i^+ 1_{(\delta_{i1}=1)} + R_i 1_{(\delta_{i2}=1)} + L_i 1_{(\delta_{i3}=1)}$, for $i = 1, \cdots, n$, based on the observed data. Here L_i^- is an arbitrary time point between 0 and L_i, and R_i^+ is an arbitrary time point greater than R_i. We introduce two independent Poisson latent variables, $z_i \sim \mathcal{P}\{\Lambda_0(t_{i1})\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\}$ and $w_i \sim \mathcal{P}[\{\Lambda_0(t_{i2}) - \Lambda_0(t_{i1})\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})]$ for each subject i. Note that introducing w_i in the case of left-censoring is unnecessary because $N_i(t_{i2})$ will not contribute any information about the failure time T_i given $z_i = N_i(t_{i1}) > 0$. For this reason, we can ignore the unconstrained latent variable w_i and have only one constraint $z_i > 0$ in the case of left-censoring. In the case of interval-censoring, there are constraints $z_i = 0$ and $w_i > 0$. In the case of right-censoring, there are constraints $z_i = 0$ and $w_i = 0$. With the above introduction of z_is and w_is, the augmented likelihood can be written as

$$L_{aug1} = \prod_{i=1}^{n} p\Big(z_i|\Lambda_0(t_{i1})\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\Big)\Big[p\Big(w_i|\{\Lambda_0(t_{i2}) - \Lambda_0(t_{i1})\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\Big)\Big]^{\delta_{i2}+\delta_{i3}} \tag{4}$$

subject to the following constraints $z_i > 0$ if $\delta_{i1} = 1$, $z_i = 0$ and $w_i > 0$ if $\delta_{i2} = 1$, and $z_i = 0$ and $w_i = 0$ if $\delta_{i3} = 1$. Here we use $p(\cdot|\lambda)$ to denote the Poisson probability mass function with the rate parameter λ. Integrating z_is and w_is out of (4) leads back to the observed likelihood (2).

The augmented likelihood (4) takes the form of the production of Poisson probability mass functions, which is much easier to deal with than the observed likelihood (2). However, the likelihood (4) still can not provide closed-form updates for the coefficients of monotone spline basis. We thus take a further data augmentation by taking advantage of the additivity property of Poisson random variables and the linearity form of $\Lambda_0(t)$ in (3). For each i, we decompose both z_i and w_i as a sum of k independent Poisson random variables, $z_i = \sum_{l=1}^{k} z_{il}$ and $w_i = \sum_{l=1}^{k} w_{il}$, where $z_{il} \sim \mathcal{P}\{\gamma_l I_l(t_{i1})\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\}$ and $w_{il} \sim \mathcal{P}[\gamma_l\{I_l(t_{i2}) - I_l(t_{i1})\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})]$ for $l = 1, \ldots, k$. In this case, the new augmented likelihood function is

$$L_{aug2} = \prod_{i=1}^{n} \prod_{l=1}^{k} p\left(z_{il} | \gamma_l I_l(t_{i1}) \exp(\boldsymbol{x}_i' \boldsymbol{\beta})\right) \left[p\left(w_{il} | \gamma_l \{I_l(t_{i2}) - I_l(t_{i1})\} \exp(\boldsymbol{x}_i' \boldsymbol{\beta})\right)\right]^{\delta_{i2}+\delta_{i3}} \tag{5}$$

subject to the following constraints $\sum_l z_{il} > 0$ if $\delta_{i1} = 1$, $\sum_l z_{il} = 0$ and $\sum_l w_{il} > 0$ if $\delta_{i2} = 1$, and $\sum_l z_{il} = \sum_l w_{il} = 0$ if $\delta_{i3} = 1$. The augmented likelihood (5) takes the form of the production of Poisson probability mass functions and directly connects with the regression coefficients and the coefficients of monotone spline basis, and therefore forms the basis of our proposed Gibbs sampler in the next section.

The above Poisson data augmentation uses the properties of the PH model and the structure of the interval-censored data. It can be extended to the proportional odds model and the generalized odds-rate hazards models by introducing a Gamma frailty term in a frailty PH model. It can serve as a general data augmentation framework to facilitate Bayesian computation for analyzing interval-censored data under such models.

3.3 Prior specification and posterior computation

We now specify the priors for the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_k)$ to combine with the augmented likelihood (5) for the posterior computation. We assign independent exponential priors $Exp(\lambda)$ for γ_ls and assign a Gamma hyper prior $\mathcal{G}(a_\lambda, b_\lambda)$ for λ with mean a_λ/b_λ and variance a_λ/b_λ^2. This prior specification is appealing from the computational perspective because it leads to conjugate forms for each of the conditional posterior distributions of γ_ls and λ. Theoretically, such a prior specification is closely related to Bayesian Lasso (Park and Casella 2008) and is equivalent to the penalized likelihood approach with L_1 penalty on those spline coefficients, in which λ serves as a tuning parameter. This prior specification penalizes large values of the coefficients γ_ls and functions to shrink the coefficients of those unnecessary spline bases towards zero. This property allows us to use many knots to provide adequate modeling flexibility but does not cause over-fitting problems. In the penalized likelihood approach, selecting a proper λ value requires much additional work using cross-validation method. In contrast, our approach treats λ as random and assigns the gamma hyper prior for λ to allow for automatic tuning with much less computational efforts. Our simulations show that our approach is very robust to the choice of the hyperparameters in the gamma hyper prior of λ. For β_j, $j = 1, \ldots, p$, we assign conventional independent vague normal priors $\pi(\beta_j) = N(\mu_j, \sigma_j^2)$ by letting σ_j^2 be large. This leads to a log-concave conditional posterior distribution for each β_j, which can be sampled easily using the adaptive rejection sampling (ARS) (Gilks and Wild 1992). For the simulations and real data analyses presented in Sects. 4 and 5, we specifically choose $a_\lambda = b_\lambda = 1$ and $\sigma_j^2 = 100$.

Gibbs sampling (Geman and Geman 1984; Gelfand and Smith 1990) is one of the most popular Markov chain Monte Carlo (MCMC) (Robert and Casella 2004) algorithms for Bayesian computation. It repeatedly and sequentially generates all unknown parameters and latent variables from their full conditional distributions. A

full conditional distribution of a parameter is defined as the conditional distribution of the parameter given everything else. The MCMC theory guarantees that the limiting distribution of the samples from a Gibbs sampler is the same as the joint posterior distribution under certain regularity conditions. We adopt Gibbs sampling for our posterior computation and derive the full conditional distributions of all parameters and latent variables by combining the augmented likelihood (5) and the priors. After specifying the initial values of the unknown parameters, our Gibbs sampler iterates through the following steps.

1. Sample z_is, z_{il}s, w_is, and w_{il}s for $l = 1, \ldots, k$ and $i = 1, \ldots, n$. First set all of them to be 0. Then for each i, if $\delta_{i1} = 1$ (i.e. left-censored), sample

$$\begin{aligned} z_i &\sim \mathcal{P}\{\Lambda_0(R_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\}1_{(z_i>0)}, \\ (z_{i1}, \cdots, z_{ik}|\, z_i) &\sim \mathcal{M}(z_i, \boldsymbol{p}_i), \quad \text{with } \boldsymbol{p}_i = (p_{i1}, \ldots, p_{ik}), \\ p_{il} &= \gamma_l I_l(R_i)\Big\{\sum_{j=1}^{k}\gamma_j I_j(R_i)\Big\}^{-1} \text{ for } l = 1, \ldots, k, \end{aligned}$$

   where $\mathcal{M}(z_i, \boldsymbol{p}_i)$ denotes a multinomial distribution with parameters z_i and $\boldsymbol{p}_i$; if $\delta_{i2} = 1$ (i.e. interval-censored), sample

$$\begin{aligned} w_i &\sim \mathcal{P}[\{\Lambda_0(R_i) - \Lambda_0(L_i)\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})]1_{(w_i>0)}, \\ (w_{i1}, \cdots, w_{ik}|\, w_i) &\sim \mathcal{M}(w_i, \boldsymbol{q}_i), \quad \text{with } \boldsymbol{q}_i = (q_{i1}, \ldots, q_{ik}), \\ q_{il} &= \gamma_l\{I_l(R_i) - I_l(L_i)\}\Big[\sum_{j=1}^{k}\gamma_j\{I_j(R_i) - I_j(L_i)\}\Big]^{-1} \\ &\qquad \text{for } l = 1, \cdots, k. \end{aligned}$$

2. Sample β_j, for $j = 1, \ldots, p$, by using the ARS method (Gilks and Wild 1992). The full conditional distribution of β_j is proportional to

$$\begin{aligned} &\exp\Big[\sum_{i=1}^{n}\boldsymbol{x}_i'\boldsymbol{\beta}(z_i\delta_{i1} + w_i\delta_{i2}) \\ &\quad -\sum_{i=1}^{n}\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\{\Lambda_0(R_i)(\delta_{i1} + \delta_{i2}) + \Lambda_0(L_i)\delta_{i3}\}\Big]\pi(\beta_j), \end{aligned}$$

   which is a log-concave function of β_j as long as the prior $\pi(\beta_j)$ is log-concave.
3. Sample γ_l from Gamma distribution $\mathcal{G}(a_{\gamma_l}, b_{\gamma_l})$ for $l = 1, \ldots, k$, where

$$\begin{aligned} a_{\gamma_l} &= 1 + \sum_{i=1}^{n}\Big(z_{il}\delta_{i1} + w_{il}\delta_{i2}\Big), \\ b_{\gamma_l} &= \lambda + \sum_{i=1}^{n}\exp\Big(\boldsymbol{x}_i'\boldsymbol{\beta}\Big)\Big\{(\delta_{i1} + \delta_{i2})I_l(R_i) + \delta_{i3}I_l(L_i)\Big\}. \end{aligned}$$

4. Sample λ from $\mathcal{G}(a_\lambda + k,\ b_\lambda + \sum_{l=1}^{k} \gamma_l)$.

The above Gibbs sampler is fast since all the parameters can be updated either from a standard distribution or by using the automatic ARS. Our R code for this algorithm is easy to implement and the required input includes the observed intervals, the censoring indicators, and the covariates. Users also have the option to specify their desired values for the degree and the number of knots for the monotone spline specification. The R code has been incorporated in our user-friendly R package `ICBayes` and the main function for our method is called "case2ph". The package is available at http://www.stat.sc.edu/~wangl/.

4 Simulation studies

Extensive simulation studies were conducted to evaluate the proposed approach and to compare it with the several existing approaches. First we generated 500 data sets with sample size $n = 200$ from the following model,

$$F(t|x_1, x_2) = 1 - \exp\Big\{ - \Lambda_0(t) \exp(\beta_1 x_1 + \beta_2 x_2)\Big\},$$

where x_1 is a Bernoulli(0.5) random variable and x_2 is a $N(0, 0.5^2)$ random variable. We took true $\Lambda_0(t) = t + \log(1 + t)$, $\beta_1 = 1$, 0 or -1, and $\beta_2 = 0$ or 1. We allowed each subject to have a random number of observations, determined by 1 plus a Poisson random variable with mean 2. The observation times were produced by generating the gap times between adjacent observation times from independent exponential distributions with mean 1. The observed interval was determined by the two adjacent observation times (including 0 and ∞) that bracket the true failure time. According to this data generating scheme, the right censoring rate varies between 7 and 22 % across all setups.

We implemented the Gibbs sampler in Sect. 3.3 for each simulated data set. Good mixing and fast convergence in the chains of the key parameters were observed. The convergence assessment was carried out using various convergence criteria in the R package `coda` (Plummer et al. 2006). We summarized the results based on 10000 iterations of the Gibbs sampler output after discarding the first 1000 iterations as a burn-in.

4.1 Selection of knots

The deviance information criterion (DIC) (Spiegelhalter et al. 2002) and log pseudo marginal likelihood (LPML) (Ibrahim et al. 2001) are two useful tools for Bayesian model comparison with MCMC samples. In general, models with smaller DICs and larger LPMLs are preferred. Detailed definition and formula to calculate DIC and LPML are referred to Spiegelhalter et al. (2002) and Ibrahim et al. (2001), respectively. Wang et al. (2013) also has the specific formula to calculate LPML for interval-censored data. We adopt the two model comparison criteria to compare two strategies of specifying knots: taking equally-spaced knots or quantile-based knots. For the

Table 1 Estimation of the regression parameters based on 500 simulated data sets from the proposed Bayesian method for the equally-spaced knots (type 1) and the quantile-based knots (type 2). CR refers to the average right-censoring rate of the 500 simulated data sets. Bias refers to the difference between the average of the 500 point estimates and the true value, SE the sample standard deviation of the 500 point estimates, SD the average of the 500 posterior standard deviations, and CP95 the 95 % coverage probability

β_1	β_2	CR (%)		Results on β_1				Results on β_2			
				Bias	SE	SD	CP95	Bias	SE	SD	CP95
1	0	6.9	Type 1	0.030	0.230	0.240	0.966	0.003	0.237	0.212	0.928
			Type 2	−0.018	0.230	0.239	0.956	0.004	0.229	0.209	0.942
1	1	8.0	Type 1	0.033	0.248	0.239	0.932	0.053	0.249	0.243	0.948
			Type 2	−0.030	0.241	0.235	0.930	0.012	0.237	0.238	0.958
0	0	10.8	Type 1	−0.011	0.189	0.184	0.940	0.000	0.195	0.187	0.944
			Type2	−0.039	0.178	0.179	0.942	0.000	0.188	0.184	0.950
0	1	12.2	Type 1	−0.015	0.196	0.188	0.928	0.054	0.227	0.217	0.942
			Type 2	−0.041	0.184	0.185	0.920	0.008	0.218	0.213	0.950
-1	0	20.7	Type 1	−0.060	0.190	0.191	0.938	−0.000	0.187	0.184	0.940
			Type 2	−0.029	0.181	0.186	0.950	−0.001	0.182	0.183	0.940
-1	1	22.1	Type 1	−0.032	0.192	0.196	0.958	0.032	0.205	0.209	0.962
			Type 2	−0.006	0.183	0.191	0.962	−0.008	0.197	0.206	0.960

equally-spaced method, we simply took 20 equally spaced knots between 0 and the maximum value of the finite endpoints of all the observed intervals for each simulated data set. For the quantile-based method, we took the finite end points of all the observed intervals as real observations and chose 0 and the sample quantiles (0.1–1 increased by 0.05) as knots.

Table 1 presents the frequentist operating characteristics of the proposed Bayesian method for the two knot selection strategies. For both the strategies, we chose 3 for the degree of basis functions to allow adequate smoothness. In the table, Bias denotes the average of the 500 point estimates minus the true value, SE the sample standard deviation of the 500 point estimates, SD the average of the 500 posterior standard devations, and CP95 the 95 % coverage probability. The point estimates were taken to be posterior means, and the coverage probabilities were based on the 95 % credible intervals. It is clear from Table 1 that the proposed Bayesian method works well for both the knot selection methods, because Biases are close to zero, SEs are close to the corresponding SDs, and the CP95s are close to 0.95 in all configurations.

To further compare the two knot selection methods, we summarize their DIC difference and LPML difference from the 500 simulated data sets in boxplots. Boxplots (see Fig. 1) show again that the two knot selection methods are comparable with a small spread of DIC difference between −10 and 10, and a smaller spread of LPML difference around between −5 and 5. However, it is clear that the equally-spaced knot selection outperforms the quantile-based knot selection with consistently overall smaller DICs and larger LPMLs across the 6 configurations. Note that from now on, all the following results are only based on the equally-spaced knots for the proposed method.

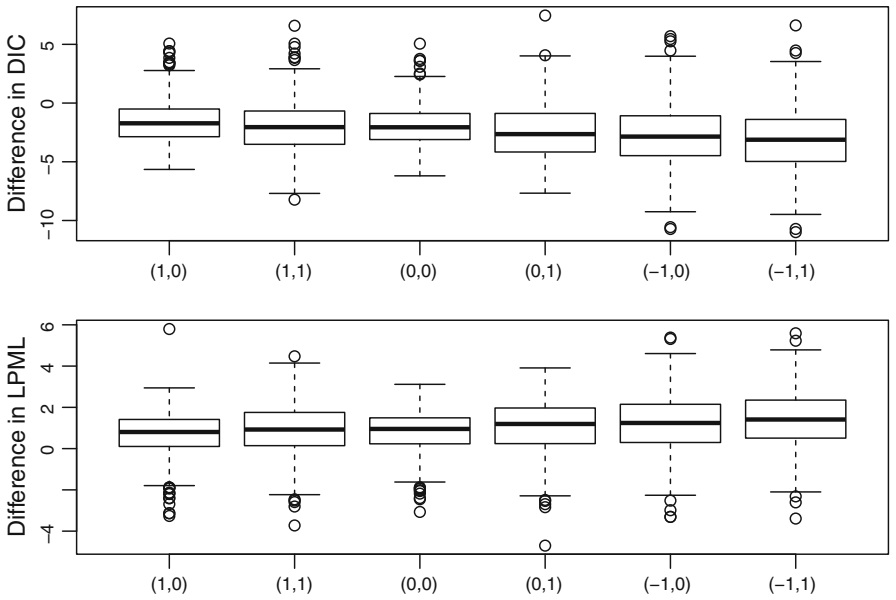

Fig. 1 *Boxplots* of DIC and LPML difference between equally-spaced knots and quantile-based knots based on 500 simulated data sets for the six (β_1, β_2) configurations. *The values in the parentheses* show the true values of (β_1, β_2)

4.2 Comparison with other existing methods

Table 2 compares the results of the proposed Bayesian method with a constrained maximum likelihood method and the three existing R packages `intcox`, `survBayes`, and `dynsurv`, on the same data sets. The constrained maximum likelihood method taking the same monotone spline specification (3) is a natural benchmark of the proposed method and is feasible here since the model only involves a finite number of unknown parameters. The point estimates were obtained by minimizing the negative logarithm of the observed likelihood function (2) over the constrained parameter space using the optimization function "fminsearch" in Matlab. The variance estimates were obtained by numerically approximating the observed information matrix formed by the second derivatives of the log-likelihood (Zeng et al. 2006; Lin and Wang 2010). However, from our observation, there are a few data sets failed to produce converged results with the constraint maximum likelihood method. Thus only the results for converged data sets were summarized and presented in Table 2. For using the three R packages, we basically adopted the default settings in the packages. For `dynsurv`, we used the time-independent coefficient model with independent vague gamma priors for the piecewise constants of the baseline hazard function and independent vague normal priors for the regression coefficients. For `survBayes`, the default prior, an autoregressive process of order one (AR(1)) prior (Henschel et al. 2009b), was used for the coefficients of the cubic B-spline approximation of the log baseline hazard function; the dispersion parameters and the regression coefficients were assigned vague gamma and normal priors, respectively.

Table 2 Estimation of the regression parameters based on 500 simulated data sets from the proposed Bayesian method, the constrained maximum likelihood method (MLE), and the R packages `survBayes`, `dynsurv`, and `intcox`. Bias refers to the difference between the average of the 500 point estimates and the true value, SE the sample standard deviation of the 500 point estimates, SD the average of the 500 posterior standard deviations (For the MLE and `intcox` methods, SD the average of the 500 estimated standard deviations), and CP95 the 95 % coverage probability

β_1	β_2		Results on β_1				Results on β_2			
			Bias	SE	SD	CP95	Bias	SE	SD	CP95
1	0	proposed	0.030	0.230	0.240	0.966	0.003	0.237	0.212	0.928
		MLE	0.052	0.243	0.250	0.970	0.000	0.244	0.222	0.947
		survBayes	0.007	0.297	0.344	0.960	−0.003	0.211	0.205	0.924
		dynsurv	0.179	0.237	0.247	0.914	0.003	0.230	0.217	0.948
		intcox	0.197	0.268	–	–	0.004	0.242	–	–
1	1	proposed	0.033	0.248	0.239	0.932	0.053	0.249	0.243	0.948
		MLE	0.061	0.257	0.256	0.955	0.061	0.262	0.295	0.957
		survBayes	0.006	0.310	0.332	0.951	−0.033	0.324	0.304	0.940
		dynsurv	0.177	0.249	0.244	0.914	0.023	0.234	0.247	0.960
		intcox	0.181	0.270	–	–	0.052	0.248	–	–
0	0	proposed	−0.011	0.189	0.184	0.940	0.000	0.195	0.187	0.944
		MLE	0.002	0.192	0.193	0.951	−0.001	0.195	0.193	0.953
		survBayes	0.029	0.207	0.188	0.923	0.003	0.201	0.187	0.940
		dynsur	0.124	0.167	0.177	0.914	−0.002	0.187	0.189	0.956
		intcox	0.130	0.200	–	–	−0.002	0.199	–	–
0	1	proposed	−0.015	0.196	0.188	0.928	0.054	0.227	0.217	0.942
		MLE	0.001	0.202	0.208	0.957	0.061	0.236	0.243	0.955
		survBayes	0.020	0.204	0.194	0.929	0.088	0.225	0.210	0.920
		dynsurv	0.125	0.174	0.181	0.916	0.015	0.215	0.217	0.954
		intcox	0.136	0.204	–	–	0.037	0.230	–	–
-1	0	proposed	−0.060	0.190	0.191	0.938	−0.000	0.187	0.184	0.940
		MLE	−0.047	0.196	0.242	0.957	0.001	0.189	0.191	0.941
		survBayes	−0.050	0.200	0.189	0.912	−0.003	0.189	0.185	0.937
		dynsurv	0.092	0.159	0.177	0.936	0.000	0.179	0.187	0.964
		intcox	0.138	0.195	–	–	−0.002	0.180	–	–
-1	1	proposed	−0.032	0.192	0.196	0.958	0.032	0.205	0.209	0.962
		MLE	−0.021	0.202	0.266	0.964	0.034	0.212	0.243	0.964
		survBayes	−0.034	0.192	0.194	0.958	0.063	0.206	0.206	0.947
		dynsurv	0.122	0.159	0.181	0.910	−0.004	0.196	0.210	0.966
		intcox	0.173	0.202	–	–	−0.016	0.200	–	–

From Table 2, `intcox` and `dynsurv` clearly perform worse with larger biases in the point estimates of β_1. The constrained likelihood method and the Bayesian method of `survBayes` produce comparable results to the proposed method in terms of Bias and CP95 but yield larger variance estimates. We did not obtain the variance estimates from `intcox` since it does not provide such estimates. We also note that

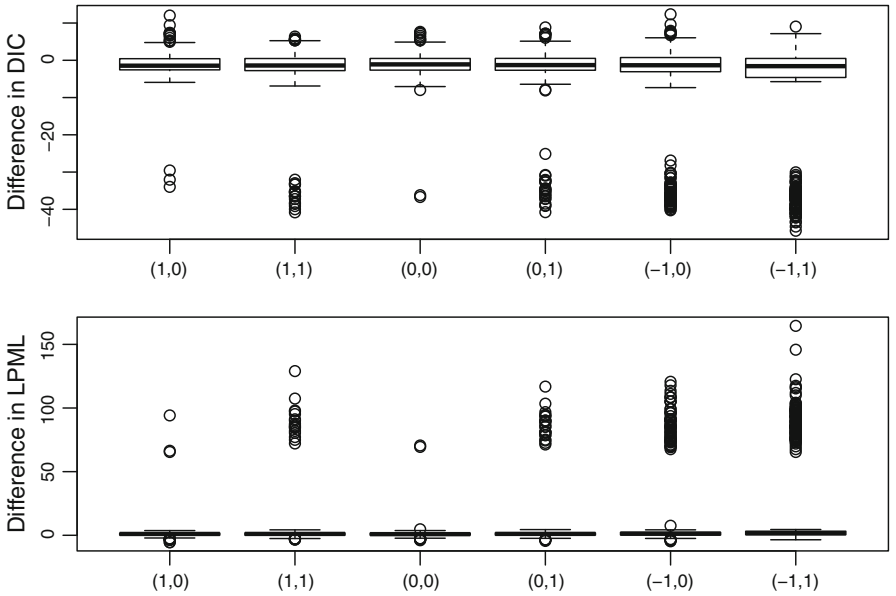

Fig. 2 *Boxplots* of DIC and LPML difference between the proposed method and the method of `dynsurv` based on 500 simulated data sets for the six (β_1, β_2) configurations. *The values in the parentheses* show the true values of (β_1, β_2)

the performance of `survBayes` depends on the choice of initial values and for some data sets, it yields non-converged results. The Bayesian method of `survBayes` is four times slower than our method in our simulations. This is not surprising as their posterior computation requires many complicated and time-consuming Metropolis-Hasting steps in addition to many numerical approximations of the integration in calculating survival functions. It takes about five minutes for our method to estimate all parameters per data set on a desktop with Intel(R)Xeon(R)CPU E5645@2.40 GHz processor.

Figure 2 presents the boxplots of the DIC and LPML difference between the proposed method and the time-independent coefficient model method in `dynsurv`. It clearly shows that the proposed method outperforms the method in `dynsurv` with smaller DICs and larger LPMLs.

4.3 Other results

One advantage of the proposed method is to provide a smooth estimate of the baseline survival function. Figure 3 presents the estimation of the baseline survival and cumulative hazard functions and their corresponding credible intervals for all the configurations. It is clear that the estimation of the baseline survival function is very good as the average of the estimates overlaps the true curve. The estimation of the cumulative hazard function is also good except at the right tail area where there is only a sparse amount of information in the observed data.

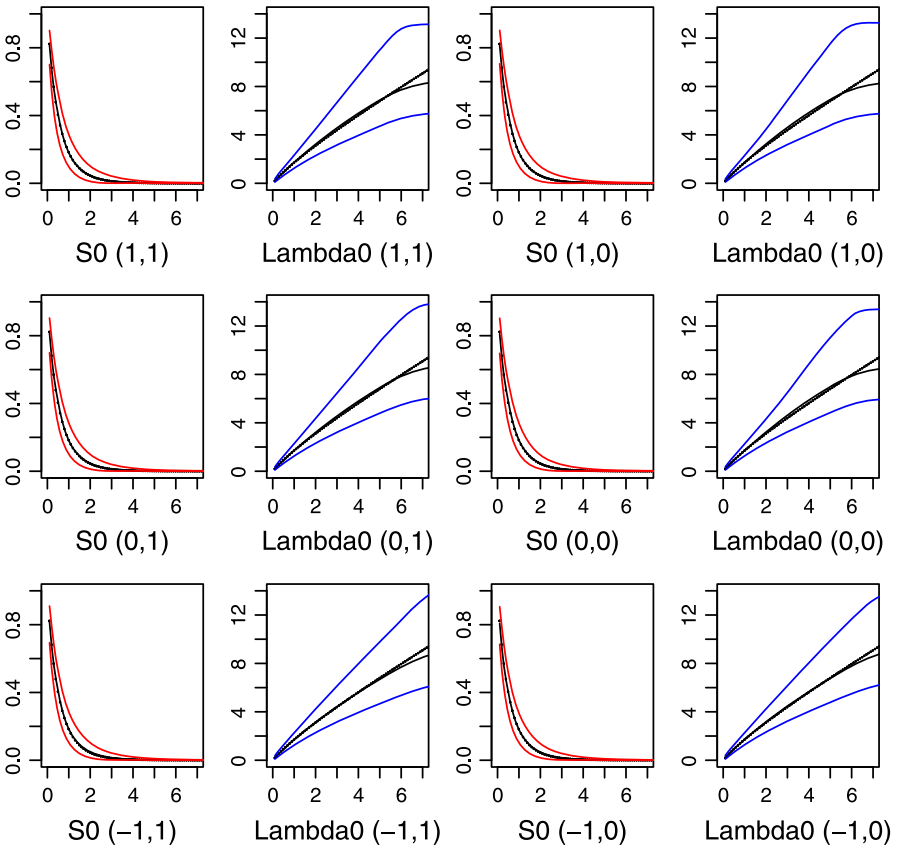

Fig. 3 Estimates of the *baseline* survival S_0 and the baseline cumulative hazard function Λ_0 and the corresponding credible lines for the six (β_1, β_2) configurations, each based on 500 simulated data sets. *Dotted lines* are true functions; *solid lines* are average estimates. *The values in the parentheses* show the true values of (β_1, β_2)

For the proposed method, besides specifying the variances in the normal prior for the regression coefficients, we only need to specify the hyperparameter in the gamma prior for λ. We ran additional simulation studies to investigate the effect of the hyperparameters in the prior $\mathcal{G}(a_\lambda, b_\lambda)$ of λ on the proposed method. We implemented our method using more vague priors by taking (a_λ, b_λ) to be $(0.1, 0.1)$, $(0.01, 0.01)$, and $(0.001, 0.001)$, respectively. Estimation results of the regression parameters are shown in Table 3. The results suggest that the proposed method is robust to the choice of (a_λ, b_λ) and that using $a_\lambda = b_\lambda = 1$ is not overly informative.

5 Two real-life data applications

We illustrate our methods with two real-life interval-censored data sets, one involving a continuous covariate and the other involving categorical covariates. Both of the two data sets have been analyzed under the PH model (Heller 2011; Zhang et al. 2005).

Table 3 Estimation of the regression parameters for using different values of a_λ and b_λ based on 500 simulated data sets. Bias refers to the difference between the average of the 500 point estimates and the true value, SE the sample standard deviation of the 500 point estimates, SD the average of the 500 posterior standard deviations, and CP95 the 95 % coverage probability

β_1	β_2	$a_\lambda = b_\lambda$	Results on β_1				Results on β_2			
			Bias	SE	SD	CP95	Bias	SE	SD	CP95
1	0	1	0.030	0.230	0.240	0.966	0.003	0.237	0.212	0.928
		0.1	0.038	0.231	0.241	0.962	0.004	0.234	0.212	0.944
		0.01	0.037	0.227	0.239	0.966	0.003	0.234	0.211	0.936
		0.001	0.036	0.227	0.240	0.968	0.003	0.233	0.211	0.938
1	1	1	0.033	0.248	0.239	0.932	0.053	0.249	0.243	0.948
		0.1	0.037	0.245	0.238	0.946	0.041	0.246	0.243	0.946
		0.01	0.039	0.246	0.239	0.948	0.039	0.245	0.243	0.954
		0.001	0.039	0.246	0.240	0.950	0.039	0.247	0.243	0.950
0	0	1	−0.011	0.189	0.184	0.940	0.000	0.195	0.187	0.944
		0.1	0.003	0.189	0.184	0.948	−0.000	0.193	0.187	0.950
		0.01	0.004	0.188	0.185	0.944	−0.000	0.193	0.187	0.952
		0.001	0.004	0.189	0.184	0.944	−0.000	0.193	0.187	0.946
0	1	1	−0.015	0.196	0.188	0.928	0.054	0.227	0.217	0.942
		0.1	−0.002	0.195	0.189	0.940	0.045	0.225	0.217	0.944
		0.01	0.001	0.196	0.189	0.938	0.042	0.225	0.217	0.940
		0.001	0.000	0.197	0.189	0.930	0.044	0.225	0.217	0.948
-1	0	1	−0.060	0.190	0.191	0.938	−0.000	0.187	0.184	0.940
		0.1	−0.042	0.191	0.193	0.950	−0.000	0.185	0.184	0.946
		0.01	−0.040	0.191	0.193	0.952	−0.000	0.185	0.185	0.946
		0.001	−0.040	0.192	0.193	0.950	−0.000	0.185	0.184	0.944
-1	1	1	−0.032	0.192	0.196	0.958	0.032	0.205	0.209	0.962
		0.1	−0.015	0.195	0.198	0.954	0.024	0.203	0.209	0.964
		0.01	−0.013	0.195	0.198	0.958	0.024	0.204	0.209	0.960
		0.001	−0.012	0.196	0.198	0.956	0.023	0.203	0.209	0.966

5.1 Colon cancer data

A total of 90 locally advanced colorectal cancer patients underwent surgery at the Memorial Sloan-Kettering Cancer Center and were followed for their time to recurrence, which was determined by radiographical scans. These patients were supposed to pay visits to their physicians every 6 or 12 months, but their actual visit times varied from patient to patient due to patients noncompliance. The failure time of interest is the time span from surgery to the recurrence of colorectal cancer, for which only interval-censored data are available. Among the 90 patients, 46 had tumor recurrence, among which 2 were left-censored. The remaining 44 patients were not observed to have tumor recurrence at the last follow-up and thus were right-censored. No patient

Table 4 Posterior means ($\overline{\text{mean}}$) and the corresponding 95 % credible intervals (CI) of the CEA effect when using different values of the degree (d) and the number (m) of interior knots for the monotone spline specification. The third and fourth rows of each cell present the corresponding DIC and LPML

		$m = 5$	$m = 10$	$m = 15$	$m = 20$
$d = 1$	$\overline{\text{mean}}$	0.289	0.295	0.295	0.283
	CI	(0.022, 0.550)	(0.036, 0.552)	(0.036, 0.551)	(0.023, 0.527)
	DIC	298.589	302.150	296.445	300.506
	LPML	−149.727	−151.783	−149.176	−151.301
$d = 2$	$\overline{\text{mean}}$	0.293	0.300	0.286	0.286
	CI	(0.032, 0.566)	(0.032, 0.542)	(0.038, 0.528)	(0.024, 0.541)
	DIC	304.125	293.703	295.620	296.932
	LPML	−152.503	−147.479	−148.529	−149.360
$d = 3$	$\overline{\text{mean}}$	0.302	0.294	0.290	0.276
	CI	(0.046, 0.565)	(0.045, 0.557)	(0.032, 0.532)	(0.023,0.533)
	DIC	299.109	296.000	295.798	296.549
	LPML	−150.136	−148.595	−148.614	−148.962
$d = 4$	$\overline{\text{mean}}$	0.295	0.289	0.283	0.270
	CI	(0.049, 0.561)	(0.043, 0.540)	(0.017, 0.527)	(0.038,0.500)
	DIC	296.191	295.935	296.024	296.964
	LPML	−148.661	−148.652	−148.632	−149.202

died before recurrence so it is reasonable to believe that the noninformative assumption holds here.

For the regression analysis, we focus on an important clinical factor called Carcinoembryonic Antigen (CEA), which is a glycoprotein involved in cell adhesion. It has been found in patients diagnosed with various types of carcinoma and is being used as a biochemical marker for such cancers. In this study, CEA was measured as a continuous variable at the baseline, i.e., surgery, and we wish to explore its association with the time to tumor recurrence.

We implemented our method using a standardized baseline CEA as a single covariate. Table 4 presents the posterior means and corresponding credible intervals of the CEA effect when using different values of the degree d and the number m of knots for the monotone spline specifications. Table 4 shows clearly that using different values for the degree and the number of knots produced similar estimates of the CEA effect with similar DICs and LPMLs. The smallest DIC and the largest LPML occur when $d = 2$ and $m = 10$. These results suggest that there is a significant effect of baseline CEA on the recurrence of colon cancer because all of the posterior means of the log hazard ratio of the baseline CEA are positive and none of the corresponding credible intervals contains zero. This indicates that patients with larger values of the baseline CEA tend to have recurrence of colon cancer earlier than patients with lower values of the baseline CEA. This conclusion agrees with that in Heller (2011) under the same model.

Table 5 Estimation results for the HIV data: posterior means and 95 % credible intervals for the regression parameters. The presented results are based on the best specification of monotone spline basis with the degree $d = 2$ and the number of knots $m = 10$

Parameter	Mean	95% CI
β_1	1.570	(1.191, 1.952)
β_2	2.770	(2.373, 3.171)
β_3	3.161	(2.741, 3.590)
$\beta_2 - \beta_1$	1.200	(0.879, 1.541)
$\beta_3 - \beta_2$	0.391	(0.035, 0.705)

5.2 HIV-1 infection data

A multi-center prospective study was conducted in the 1980's to investigate HIV-1 infection rate among people with hemophilia. The patients were at risk of HIV-1 infection from blood products made from donors' plasma. Based on the average annual dose of the blood products they received, 544 patients were assigned to four groups: high-, medium-, low-, or non-dose group. The research aims of the study were to compare the HIV-1 infection rates among those dose groups and to quantify the dose effects. In the study, the exact HIV-1 infection times were never observed and only interval-censored data were available. Among all patients, 63 were left-censored, 204 were interval-censored, and 277 were right-censored. More details about this study can be found in Goedert et al. (1989) and Kroner et al. (1994). This typical interval-censored data set has been analyzed by Sun (2006) and Zhang et al. (2005) among others.

To analyze the data, we generated three dummy variables x_{i1}, x_{i2}, and x_{i3} to indicate subject i being in the low-, medium-, and high-dose group, respectively. We tried different values of the degree d and the number m of equal-spaced knots within (0, 60) to specify monotone splines for the analysis. They all produced similar results with $d = 3$ and $m = 10$ yielding the smallest DIC and the largest LPML. Table 5 presents the estimated dose effects and the corresponding 95 % credible intervals when using $d = 3$ and $m = 10$. These results indicate that there is a significant dose effect between each dose group and the non-dose group. Also, a significant difference exists between the adjacent dose groups. Figure 4 plots the estimated survival functions for all four groups.

6 Discussion

In this paper, we propose a novel Bayesian method to analyze general interval-censored data with the PH model. Our method allows one to estimate the regression parameters and the baseline survival function jointly. The use of monotone splines leads to a finite number of parameters in the model and produces a smooth estimate of the baseline survival function. The use of shrinkage priors for the spline coefficients naturally prevents over-fitting when a large number of knots are used. The novel Poisson data augmentation strategy is specifically designed for dealing with interval-censored data using the PH model properties. Our approach has great computational advantages over the existing Bayesian methods in that it does not require imputing the unobserved failure times or contain any complicated Metropolis steps, and all the sampling steps are straightforward or automatic.

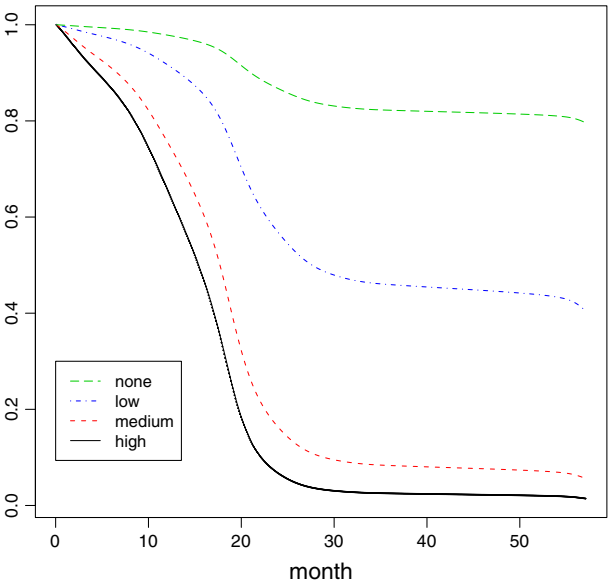

Fig. 4 Estimates of the survival functions for different dose groups in the HIV data

Extending the proposed method to handle time-varying regression coefficients through some spline modeling of the regression coefficients, such as using B-splines or piecewise constants, is possible. The similar data augmentation structure can still be applied. The resulted Gibbs sampler will be similar as the current one except that the updates of the regression coefficients $\boldsymbol{\beta}$ will be much more complicated.

The proposed method can be modified to fit the PO model (Rabinowitz et al. 2000) and the generalized odds-rate hazards models (Banerjee et al. 2007) by introducing a gamma frailty term in the frailty PH model. It can also be extended to handle clustered or multivariate interval-censored data under the frailty PH and PO models, for which existing methodological research is relatively limited. Our future effort will be devoted to such extensions, and we will incorporate these new methods into our R package `ICBayes` to provide practitioners a comprehensive and efficient statistical package for analyzing various types of interval-censored data.

References

Banerjee T, Chen MH, Dey DK, Kim S (2007) Bayesian analysis of generalized odds-rate hazards models for survival data. Lifetime Data Anal 13:241–260

Cai B, Lin X, Wang L (2011) Bayesian proportional hazards model for current status data with monotone splines. Comput Statist Data Anal 55:2644–2651

Cai T, Betensky RA (2003) Hazard regression for interval-censored data with penalized spline. Biometrics 59:570–579

Cox D (1972) Regression models and life tables (with discussion). J Royal Statist Soc Ser B 34:187–220

Cox D (1975) Partial likelihood. Biometrika 62:269–276

Finkelstein DM (1986) A proportional hazards model for interval-censored failure time data. Biometrics 42:845–854

Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Statist Assoc 85:398–409
Geman S, Geman D (1984) Stochastic relaxiation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intel 6:721–741
Gilks W, Wild P (1992) Adaptive rejection sampling for Gibbs sampling. Appl Statist 41:337–348
Goedert J, Kessler C, Adedort L, Biggar R, Andes W, White G, Drummond J, Vaidya K, Mann D, Eyster M et al (1989) A progressive-study of human immunodeficiency virus type-1 infection and the development of AIDS in subjects with hemophilia. New Engl J Med 321:1141–1148
Goggins W, Finkelstein DM, Schoenfeld DA, Zaslavsky M (1998) A Markov chain Monte Marlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. Biometrics 54:1498–1507
Gomez G, Calle ML, Oller R, Langohr K (2009) Tutorial on methods for interval-censored data and their implementation in R. Statist Model 9:259–297
Heller G (2011) Proportional hazards regression with interval censored data using an inverse probability weight. Lifetime Data Anal 17:373–385
Henschel V, Heiß C, Mansmann U (2009a) The intcox package. Comprehensive R archive network
Henschel V, Heiß C, Mansmann U (2009b) survBayes: A introduction into the package. Comprehensive R archive network.
Ibrahim JG, Chen MH, Sinha D (2001) Bayesian survival analysis. Springer, New York
Joly P, Commenges D, Letenneur L (1998) A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. Biometrics 54:185–194
Kroner B, Rosenberg P, Adedort L, Alvord W, Goedert J (1994) HIV-1 infection incidence among people with hemophilia in the United States and Western Europe, 1978–1990. J Acquired Immune Defic Syndr 7:279–286
Lin X, Wang L (2010) A semiparametric Probit model for case 2 interval-censored failure time data. Statist Med 29:972–981
Lin X, Wang L (2011) Bayesian proportional odds models for analyzing current status data: univariate, clustered, and multivariate. Commun Statist Simul Comput 40:1171–1181
Pan W (1999) Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. J Comput Graph Statist 8:109–120
Pan W (2000) A multiple imputation approach to Cox regression with interval-censored data. Biometrics 56:199–203
Park T, Casella G (2008) The Bayesian Lasso. J Am Statist Assoc 103:681–686
Plummer M, Best N, Cowles K, Vines K (2006) CODA: convergence diagnosis and output analysis for MCMC. R News 6:7–11
Ramsay JO (1988) Monotone regression splines in action. Statist Sci 3:425–441
Rabinowitz D, Betensky RA, Tsiatis AA (2000) Using conditional logistic regression to fit proportional odds models to interval censored data. Biometrics 56:511–518
Robert CP, Casella G (2004) Monte Carlo statistical methods. Springer, New York
Satten GA (1996) Rank based inference in the proportional hazards model for interval-censored data. Biometrika 83:355–370
Satten GA, Datta S, Williamson JM (1998) Inference based on imputed failure times for the proportional hazards model with interval-censored data. J Am Statist Assoc 93:318–327
Sinha D, Chen MH, Ghosh SK (1999) Bayesian analysis and model selection for interval-censored survival data. Biometrics 55:585–590
Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (Pkg: P583–639). J Royal Statist Soc Ser B 64:583–616
Sun J (2006) The statistical analysis of interval-censored data. Springer, New York
Wang L, Dunson DB (2011) Semiparametric Bayes proportional odds models for current status data with under-reporting. Biometrics 67:1111–1118
Wang L, Lin X (2011) A Bayesian approach for analyzing case 2 interval-censored failure time data under the semiparametric proportional odds model. Statist Probab Lett 81:876–883
Wang X, Chen MH, Yan J (2013) Bayesian dynamic regression models for interval censored survival data with application to children dental health. Lifetime Data Anal 19:297–316
Wang X, Yan J, Chen MH (2013) The dynsurv package. Comprehensive R archive network
Yavuz AC, Lambert P (2011) Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines. Statist Med 30:75–90

Zeng D, Cai J, Shen Y (2006) Semiparametric additive risks model for interval-censored data. Statistica Sinica 16:287–302

Zhang Y, Hua L, Huang J (2010) A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. Scandinavian J Statist 37:338–354

Zhang ZG, Sun J (2010) Interval censoring. Statist Methods Med Res 19:53–70

Zhang ZG, Sun L, Zhao X, Sun J (2005) Regression analysis of interval-censored failure time data with linear transformation models. Can J Statist 33:61–70