Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian proportional hazards model for current status data with monotone splines

Bo Cai [a], Xiaoyan Lin [b], Lianming Wang [b,*]

[a] *Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, 29208, USA*

[b] *Department of Statistics, University of South Carolina, Columbia, SC, 29208, USA*

ARTICLE INFO

Article history:
Received 16 April 2010
Received in revised form 24 January 2011
Accepted 17 March 2011
Available online 5 April 2011

Keywords:
Current status data
Gibbs sampler
Monotone spline
Poisson distribution
The proportional hazards model

ABSTRACT

The proportional hazards model is widely used to deal with time to event data in many fields. However, its popularity is limited to right-censored data, for which the partial likelihood is available and the partial likelihood method allows one to estimate the regression coefficients directly without estimating the baseline hazard function. In this paper, we focus on current status data and propose an efficient and easy-to-implement Bayesian approach under the proportional hazards model. Specifically, we model the baseline cumulative hazard function with monotone splines leading to only a finite number of parameters to estimate while maintaining great modeling flexibility. An efficient Gibbs sampler is proposed for posterior computation relying on a data augmentation through Poisson latent variables. The proposed method is evaluated and compared to a constrained maximum likelihood method and three other existing approaches in a simulation study. Uterine fibroid data from an epidemiological study are analyzed as an illustration.

1. Introduction

Cox's proportional hazards (PH) model is probably the most widely used regression model in survival analysis. It assumes that covariates have a multiplicative effect on the hazard function of the failure time of interest (Cox, 1972) as follows,

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

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 denoting the covariate effects. Besides the nice interpretation of the regression parameters, the PH model has great modeling flexibility as the baseline hazard function $\lambda_0(t)$ is an unspecified nonnegative function.

In general, nonparametric estimation of the baseline hazard (or survival) function is difficult since it has an infinite dimension. For right-censored data, under the PH model, there exists partial likelihood approach (Cox, 1975), which allows one to estimate the regression parameters directly without the need for estimating the baseline hazard (survival) function. For interval-censored data, however, due to its more complex data structure, the partial likelihood under the PH model does not exist, and the counting process and martingale techniques fail to work (Sun, 2006). In interval-censored data, the failure time of interest is not observed directly but is known to fall within some interval formed by some monitoring times. For example, the onset time of HIV cannot be observed directly, and the HIV status is determined only by some laboratory test at clinical visit times. If the result is HIV positive at the first examination time, we say that the onset time is left censored; if the results are always negative, we say that the onset time is right-censored at the last examination

[*] Corresponding author. Tel.: +1 803 777 2834; fax: +1 803 777 4084.
E-mail address: wang99@mailbox.sc.edu (L. Wang).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.03.013

time; otherwise, the onset time is interval-censored and the observed interval is formed by the two closest examination times with negative and positive HIV status in this case. In general, interval-censored data is a mixture of left, interval, and right-censored observations.

In this paper, we focus our attention on analyzing case 1 interval-censored data (Groeneboom and Wellner, 1992; Sun, 2006) with the PH model. This type of data is also referred to as current status data, in which the failure times of all subjects in the study are either left-censored or right-censored (Groeneboom and Wellner, 1992; Shiboski, 1998; Sun, 2006). Current status data often occur in the cross-sectional studies and tumorigenicity studies. For example, the onset times of nonlethal tumors are never observed directly but are known to be smaller or larger than the death or sacrifice times of the animals depending on whether or not some tumors have developed at that time.

There are some existing approaches available for regression analysis of current status data under the semiparametric PH model. Among others, Huang (1996) investigated the efficient estimation and established the asymptotical results; Pan (1999) proposed a generalized gradient projection method (GGP) by reformulating the iterative convex minorant (ICM) algorithm of Groeneboom and Wellner (1992) to allow covariates; Shiboski (1998) proposed a class of generalized additive models with the PH model as a special case based on isotonic regression.

Compared to the flourishing methods focusing on the inference, the computational techniques are underdeveloped for current status data (Shiboski, 1998; Mongoué-Tchokoté and Kim, 2008). Existing software packages include `intcox` based on the Pan (1999) approach (Henschel et al., 2009a), `C1.coxph` using the generalized Gauss–Seidel algorithm (Mongoué-Tchokoté and Kim, 2008), and two Bayesian packages, `BITE` (Härkänen, 2003) and `survBayes` (Henschel et al., 2009b). Even with these statistical packages available, practitioners tend to fit parametric models or use the simpler binary regression treating time as a predictor (Shiboski, 1998). It is also a common practice to transform current status data to right-censored data by simply treating the left-censored observations as exact and then apply the well-established packages such as `phreg` in SAS or `coxph` in R for right-censored data (Mongoué-Tchokoté and Kim, 2008; Gómez et al., 2009).

In this paper, we propose a novel Bayesian approach which provides an alternative way to analyze current status data under the PH model. In general, Bayesian methods have great modeling flexibility through hierarchical structures and provide feasible computation via Markov chain Monte Carlo (MCMC) for the situations with complicated likelihoods. Also, Bayesian methods do not rely on asymptotic theory and can provide accurate results even when the sample size is small. The proposed approach takes advantage of a simple way of modeling the cumulative hazard function with monotone splines and utilizes a data augmentation with Poisson latent variables to facilitate the posterior computation. Unlike most Bayesian methods for censored data in the literature, the proposed approach does not require any imputation of failure times. The proposed Gibbs sampler is easy to implement and is more efficient than several existing approaches as shown in our simulation study.

The remainder of this paper is organized as follows. Section 2 provides the details about the proposed method, including the use of monotone splines to model $\Lambda_0(t)$, the prior specification, the data augmentation, and posterior computation. Section 3 shows the simulation results of the proposed approach and its comparison with the maximum likelihood estimation method (MLE) and three existing approaches using `intcox`, `survBayes`, and `C1.coxph`, respectively. Section 4 illustrates the proposed method with an application of fibroid data from an epidemiological study. Section 5 provides some concluding remarks.

2. The proposed approach

2.1. Modeling $\Lambda_0(t)$ with monotone splines

Suppose there are n independent subjects in the study. Let T_i denote the failure time of interest, C_i the censoring (observation) time, and $\boldsymbol{x}_i$ a $p \times 1$ covariate vector for subject i. Throughout this paper, we assume that $T_i \sim F(\cdot|\boldsymbol{x}_i)$ and given covariate $\boldsymbol{x}_i$, T_i and C_i are independent. Denote $\delta_i = I(T_i \leq C_i)$ the censoring indicator for subject i with 1 indicating left-censoring and 0 indicating right-censoring. Given the observed data $\{(c_i, \delta_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$, the likelihood function can be written as

$$L = \prod_{i=1}^{n} [1 - \exp\{-\Lambda_0(c_i)\exp(\boldsymbol{x}_i\boldsymbol{\beta})\}]^{\delta_i} \exp\{-(1-\delta_i)\Lambda_0(c_i)\exp(\boldsymbol{x}_i\boldsymbol{\beta})\}, \tag{1}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is the baseline cumulative hazard function. This likelihood is a simplified version of the full likelihood under the assumption that the failure time and the censoring time are independent given covariates (Sun, 2006).

In the above likelihood, the baseline cumulative hazard function $\Lambda_0(t)$ is a totally unspecified nondecreasing function. Using splines to model unknown functions is very common in statistics and it provides a bridge between parametric and nonparametric models. To be more specific, the resulting model is essentially a parametric model as it has a finite number of parameters after the spline basis functions are chosen; however, it also has a nonparametric nature as it does not make assumptions on the shape of the fitted curve. Here we propose to model $\Lambda_0(t)$ with a linear combination of monotone splines of Ramsay (1988) as follows,

$$\Lambda_0(t) = \sum_{l=1}^{k} \gamma_l I_l(t), \tag{2}$$

where I_ls are the integrated spline basis functions, each of which is nondecreasing from 0 to 1, and $\{\gamma_l\}_{l=1}^k$ are taken to be nonnegative values for $l \geq 1$ to ensure that $\Lambda_0(t)$ is nondecreasing. These basis functions are totally determined if the knots and degree of the monotone splines are specified. The placement of the knots determines the shape, and the degree determines the smoothness of the I splines. The number k of spline basis functions equals the number m of interior knots plus the degree d of the splines. We refer one to Ramsay (1988) for more details about I spline basis functions.

Similar expressions of (2) are used to model the unknown nondecreasing functions in Lin and Wang (2010) under the probit model and in Wang and Dunson (2010) under the proportional odds model. In general, having a large number of knots or basis functions may cause over-fitting and is usually unnecessary in using splines. Following Lin and Wang (2010) and Wang and Dunson (2010), we adopt a moderate number (10–30) of equally-spaced knots to allow for efficient computation while providing great modeling flexibility.

Our modeling $\Lambda_0(t)$ with monotone splines in expression (2) is much simpler than that used in the two Bayesian packages `BITE` and `survBayes`. Specifically, `BITE` models the hazard function with a piecewise constant function and `survBayes` models the log baseline hazard function with cubic B-spline, both requiring numerical approximation for evaluating survival functions in the likelihood. In contrast, with the specification of (2), we can calculate the likelihood directly.

2.2. Posterior computation

To facilitate the computation, we consider the following data augmentation with Poisson latent variables,

$$\delta_i = 1_{(z_i>0)}, \quad z_i \sim \text{Poisson}(\Lambda_0(c_i)\mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta}}), \tag{3}$$

$$z_i = \sum_{l=1}^{k} z_{il}, \quad z_{il} \sim \text{Poisson}(\gamma_l I_l(c_i)\mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta}}),\ l = 1, \ldots, k,$$

where in the second part z_i is decomposed as a sum of independent Poisson random variables z_{il}s for each i. Under the specification of (1) and (2) with the above data augmentation, the augmented likelihood function can be written as

$$L = \prod_i \left\{ \delta_i^{1_{(z_i>0)}} (1-\delta_i)^{1_{(z_i=0)}} \prod_{l=1}^{k} \text{Possion}(z_{il}|\gamma_l I_l(c_i)\mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta}}) \right\}.$$

We assign conditionally independent exponential priors $\mathcal{E}xp(\lambda)$ for γ_l's given λ and assign a Gamma prior $\mathcal{G}a(a_\lambda, b_\lambda)$ for λ with a mean of a_λ/b_λ and variance a_λ/b_λ^2. This specification has the following merits. First, it leads to conjugate forms for each of the full conditional distributions of γ_ls and λ. Second, adopting the hyper prior for λ allows borrowing information among the spline coefficients. Third, such prior specifications penalize large values of the coefficients γ_ls and function to shrink the coefficients of those unnecessary spline bases towards zero. For convenience, we take independent priors for β_j's with $\pi(\beta_j) = N(\mu_j, \sigma_j^2), j = 1, \ldots, p$, but note that our algorithm below allows effective sampling for any log-concave prior for β_j by using the adaptive rejection sampling (ARS) (Gilks and Wild, 1992) or adaptive rejection Metropolis sampling (ARMS) (Gilks et al., 1995). Alternatively, one may use the Metropolis random walk sampling algorithm with Gaussian proposals for β_j. After specifying initial values for the hyper parameters, the algorithm of the posterior computation iterates through the following sequential steps.

1. Sample z_i's and z_{il}'s. Set $z_i = 0$ and $z_{il} = 0$ for all i and l first. If $\delta_i = 1$, then sample $z_i \sim \text{Possion}\{\Lambda(c_i)\mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta}}\}$ such that $z_i > 0$ and then sample $(z_{i1}, \ldots, z_{ik}) \sim \text{Multinomial}(z_i, (p_{i1}, \ldots, p_{ik}))$ where $p_{il} = \gamma_l I_l(c_i)/\sum_{j=1}^k \gamma_j I_j(c_i)$.
2. Sample γ_l from $\mathcal{G}a(1 + \sum_{i=1}^n z_{il} 1_{(I_l(c_i)>0)}, \lambda + \sum_{i=1}^n I_l(c_i)\mathrm{e}^{\boldsymbol{x}_i'\boldsymbol{\beta}})$.
3. Sample each β_j with ARS or ARMS. The full conditional distribution of β_j is
$$\beta_j|\cdot \ \propto \pi(\beta_j) \exp\left(\sum_{i=1}^n z_i x_{ij}\beta_j - \sum_{i=1}^n \Lambda(c_i)\mathrm{e}^{\boldsymbol{x}_i\boldsymbol{\beta}} \right),$$
which is log-concave when $\pi(\beta_j)$ is log-concave.
4. Sample λ from $\mathcal{G}a(a_\lambda + k, b_\lambda + \sum_{l=1}^k \lambda_l)$.

The above algorithm is fast since all the parameters can be updated either from a standard distribution or by using the ARS. Our R code for this algorithm is easy to implement and the only required input is the data matrix containing the vectors of censoring times, censoring indicators, and covariates. Also, users have options to change the default settings, such as the degree and the number of the interior knots of monotone splines. The R code is available upon request.

3. A simulation study

An intensive simulation study was conducted to evaluate the proposed Bayesian method as well as several existing approaches for comparison. We generated 500 data sets with sample size $n = 200$ in each data set from the following model,

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

Table 1
Simulation results on the regression parameters for the proposed Bayesian method, the constrained maximum likelihood method (MLE), the Pan (1999) approach (`intcox`), Mongoué-Tchokoté and Kim (2008) approach (`C1.coxph`), and the Henschel et al. (2009b) approach (`survBayes`).

			Proposed				MLE				`C1.coxph`				`survBayes`				`intcox`	
β_1	β_2		Bias	SSD	ESE	CP95	Bias	SSD	ESE	CP95	Bias	SSD	ESE	CP95	Bias	SSD	ESE	CP95	Bias	SSD
1	1	$\hat{\beta}_1$	0.011	0.286	0.271	0.936	0.092	0.308	0.315	0.951	0.230	0.374	0.209	0.648	−0.054	0.255	0.270	0.948	0.128	0.310
		$\hat{\beta}_2$	0.017	0.280	0.279	0.958	0.093	0.325	0.317	0.963	0.223	0.379	0.180	0.632	−0.087	0.255	0.273	0.964	0.094	0.322
	0	$\hat{\beta}_1$	0.024	0.284	0.274	0.936	0.109	0.307	0.317	0.954	0.244	0.390	0.208	0.708	−0.060	0.248	0.255	0.944	0.145	0.326
		$\hat{\beta}_2$	0.010	0.248	0.255	0.952	0.012	0.281	0.277	0.956	0.010	0.325	0.176	0.728	0.004	0.220	0.242	0.964	0.008	0.285
	−1	$\hat{\beta}_1$	0.009	0.290	0.270	0.946	0.107	0.315	0.329	0.942	0.240	0.401	0.210	0.694	−0.085	0.274	0.263	0.934	0.137	0.328
		$\hat{\beta}_2$	−0.039	0.282	0.279	0.946	−0.120	0.331	0.324	0.959	−0.264	0.392	0.179	0.580	0.081	0.286	0.267	0.938	−0.129	0.326
0	1	$\hat{\beta}_1$	−0.018	0.244	0.239	0.946	−0.007	0.268	0.277	0.941	−0.009	0.305	0.191	0.784	0.006	0.229	0.233	0.944	0.066	0.267
		$\hat{\beta}_2$	0.038	0.252	0.268	0.962	0.117	0.303	0.287	0.957	0.124	0.328	0.192	0.674	−0.038	0.226	0.249	0.962	0.112	0.289
	0	$\hat{\beta}_1$	−0.029	0.251	0.233	0.922	−0.011	0.264	0.281	0.933	−0.018	0.308	0.189	0.776	−0.010	0.234	0.227	0.928	0.058	0.272
		$\hat{\beta}_2$	−0.006	0.232	0.238	0.948	0.002	0.259	0.255	0.957	−0.010	0.279	0.188	0.806	−0.006	0.216	0.230	0.962	−0.008	0.254
	−1	$\hat{\beta}_1$	−0.036	0.240	0.238	0.950	−0.009	0.267	0.257	0.952	−0.024	0.290	0.191	0.802	−0.013	0.223	0.233	0.962	0.057	0.255
		$\hat{\beta}_2$	−0.047	0.285	0.269	0.930	−0.125	0.302	0.334	0.925	−0.222	0.362	0.192	0.662	0.024	0.253	0.251	0.940	−0.121	0.324
−1	1	$\hat{\beta}_1$	−0.020	0.271	0.257	0.954	−0.074	0.304	0.311	0.938	−0.161	0.343	0.233	0.796	0.062	0.251	0.243	0.940	0.029	0.309
		$\hat{\beta}_2$	0.031	0.273	0.274	0.954	0.102	0.308	0.303	0.957	0.188	0.337	0.217	0.732	−0.033	0.254	0.259	0.946	0.090	0.303
	0	$\hat{\beta}_1$	−0.016	0.252	0.251	0.962	−0.073	0.291	0.281	0.950	−0.149	0.320	0.230	0.840	0.065	0.235	0.237	0.948	0.025	0.287
		$\hat{\beta}_2$	0.008	0.255	0.248	0.946	0.013	0.268	0.279	0.941	0.011	0.303	0.212	0.822	0.005	0.239	0.240	0.948	0.011	0.277
	−1	$\hat{\beta}_1$	−0.028	0.258	0.257	0.944	−0.083	0.296	0.289	0.945	−0.170	0.322	0.233	0.802	0.052	0.238	0.246	0.958	0.013	0.287
		$\hat{\beta}_2$	−0.050	0.273	0.274	0.948	−0.129	0.303	0.314	0.936	−0.215	0.337	0.214	0.712	0.010	0.244	0.259	0.964	−0.116	0.299

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 $\Lambda_0(t) = \log(1+t) + t^{3/2}$, $\beta_1 = 1, 0$ or -1, and $\beta_2 = 1, 0$, or -1. The censoring times c_is were generated independently from a truncated exponential distribution $\mathcal{E}xp(1)$ with support $(0, 10)$. The censoring indicator δ_i was then generated from a Bernoulli distribution with success probability $F(c_i|\boldsymbol{x}_i)$.

In specifying the monotone splines, we chose 3 for the degree to allow adequate smoothness and took 15 equally spaced knots within the minimum and maximum of the observation times for each generated data set. To implement the Bayesian computation, we adopt independent normal priors $N(0, 10^2)$ for β_js, and specify $a_\lambda = b_\lambda = 1$ to give a diffuse Gamma prior for the hyper-parameter λ. For each dataset, we implemented the Gibbs sampler in Section 2.2 and collected 3000 iterations after discarding the first 1000 iterations as a burn-in. Fast convergence was observed in all the setups of the simulation.

A natural benchmark method to compare with the proposed method is the maximum likelihood method under the same monotone spline specification (2). The maximum likelihood method is feasible here since there are only a finite number $(k+p)$ of parameters to estimate. Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \boldsymbol{\beta}')'$ denote the unknown parameters that need to be estimated. The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ was obtained by using the built-in Matlab function "fmincon" to minimize the negative logarithm of the likelihood function (1) over the constrained parameter space. The variance–covariance matrix of $\hat{\boldsymbol{\theta}}$ was calculated by the inverse of the observed information matrix $\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})^{-1}$ (Zeng et al., 2006; Lin and Wang, 2010). Specifically, the (s, l)th element of $\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})$ was approximated numerically by

$$\mathbf{I}(s, l) \approx -n^{-1}h_n^{-2}\left\{\log l(\hat{\boldsymbol{\theta}} + h_n\vec{e}_s - h_n\vec{e}_l) - \log l(\hat{\boldsymbol{\theta}} + h_n\vec{e}_s) - \log l(\hat{\boldsymbol{\theta}} - h_n\vec{e}_l) + \log l(\hat{\boldsymbol{\theta}})\right\},$$

where $\vec{e}_s$ is a $(1+k+p)$-dimensional vector with the sth element equal to 1 and all other elements equal to 0, $\log l$ is the log likelihood function, and h_n is a tuning constant with an order of $n^{-1/2}$. In our simulation, taking the tuning parameter $h_n = 0.1n^{-1/2}$ or $0.2n^{-1/2}$ did not produce any difference in the variance estimates. In our simulation, not all the data sets produced convergent results using "fmincon" with the default maximum number of iterations. Such non-convergence is mainly due to over-fitting when there are more than adequate number of parameters needed to fit the data. The number of non-converged data sets varies from 11 to 86 out of 500 across the 9 setups in our simulation. The summarized results are based on the converged data sets only.

Three other existing methods were also implemented for comparison including the Pan (1999) approach, Mongoué-Tchokoté and Kim (2008) approach, and the Henschel et al. (2009b) approach. These approaches were carried out by using the existing R packages `intcox`, `C1.coxph`, and `survBayes`, respectively.

Table 1 shows the frequentist operating characteristics of the regression parameters resulting from the proposed Bayesian method and the other four methods aforementioned. The Bias is calculated as the difference between the average of the 500 point estimates and the true value, ESE the average of the estimated standard errors, SSD the sample standard deviation of the 500 point estimates, and CP95 the 95% coverage probability. For Pan's method, only Bias and SSD are reported due to the unavailability of variance estimates from `intcox`.

From the results in Table 1, the proposed Bayesian method works very well with small bias for all the point estimates, the ESEs close to the SSDs, and the CP95s close to the nominal value 0.95. The Bayesian method by using `survBayes` produces comparative results as the proposed approach. However, the implementation of `survBayes` is over four times slower than

Table 2
Mean squared errors of the estimates of β_1, β_2, and F_0 from the proposed Bayesian method, the constrained maximum likelihood method (MLE), the Pan (1999) approach (`intcox`), Mongoué-Tchokoté and Kim (2008) approach (`C1.coxph`), and the Henschel et al. (2009b) approach (`survBayes`).

True		Proposed			MLE			`C1.coxph`			`survBayes`			`intcox`		
β_1	β_2	MSE (β_1)	MSE (β_2)	maxMSE (F_0)	MSE (β_1)	MSE (β_2)	maxMSE (F_0)	MSE (β_1)	MSE (β_2)	maxMSE (F_0)	MSE (β_1)	MSE (β_2)	maxMSE (F_0)	MSE (β_1)	MSE (β_2)	maxMSE (F_0)
1	1	0.082	0.079	0.006	0.108	0.109	0.009	0.192	0.193	0.013	0.068	0.072	0.007	0.112	0.113	0.008
	0	0.081	0.062	0.005	0.112	0.077	0.008	0.212	0.106	0.013	0.065	0.048	0.005	0.127	0.081	0.009
	−1	0.084	0.081	0.005	0.120	0.119	0.008	0.218	0.223	0.014	0.082	0.088	0.005	0.126	0.123	0.010
0	1	0.060	0.065	0.005	0.076	0.096	0.008	0.093	0.153	0.011	0.052	0.052	0.005	0.075	0.096	0.008
	0	0.064	0.054	0.005	0.079	0.065	0.007	0.095	0.078	0.009	0.055	0.046	0.005	0.077	0.064	0.008
	−1	0.059	0.084	0.005	0.066	0.127	0.008	0.085	0.180	0.011	0.050	0.064	0.005	0.068	0.120	0.008
−1	1	0.074	0.075	0.006	0.102	0.102	0.009	0.143	0.150	0.013	0.067	0.066	0.006	0.096	0.100	0.009
	0	0.063	0.065	0.005	0.084	0.078	0.007	0.125	0.092	0.011	0.059	0.057	0.005	0.083	0.077	0.007
	−1	0.068	0.077	0.005	0.090	0.115	0.008	0.132	0.159	0.013	0.059	0.059	0.005	0.082	0.103	0.007

the proposed method from our simulation, largely due to the simplicity of our modeling with monotone spline and the Gibbs sampler. Among the three frequentist approaches, the maximum likelihood method gives reasonable variance estimates, the `C1.coxph` developed by Mongoué-Tchokoté and Kim (2008) produces poor variance estimates, and `intcox` fails to produce variance estimates. All these three frequentist methods yields large biases in the point estimates when the true regression parameters are non-zero with `C1.coxph` showing the largest biases. It is clear that the two Bayesian methods outperform the three frequentist approaches.

To compare the proposed Bayesian approach with the other four methods more formally, we also calculated the mean squared errors (MSEs) of estimates of β_1, β_2, and F_0 for each method. The results are presented in Table 2. The maxMSE(F_0) was taken to be the maximum of all the local MSE(F_0)s, which were evaluated on a set of pre-specified evenly-spaced grid points. In Table 2, it is clear that all the MSEs from the two Bayesian methods are smaller than those from the other three frequentist methods. Another interesting result is that all methods seem to produce good estimates of F_0 in terms of the small values of $\text{maxMS}E(F_0)$ although the two Bayesian methods give more accurate estimates than the other frequentist approaches.

We also tried using different numbers of knots (10, 20 and 25) and degrees (1, 2, 3 and 4) for the monotone splines and obtained similar results. Fig. 1 depicts the average of 500 point estimates of β_1 and β_2 with different degrees of smoothness (degree $=$ 1, 2, 3 and 4) and different numbers of knots ($m = 10, 15, 20$ and 25) under the simulation with $(\beta_1, \beta_2) = (1, 1)$. It evinces that although the estimates vary slightly, the proposed method is basically robust to the choice of numbers of knots and degrees of smoothness (2, 3 and 4) while it has relatively more variation for degree of smoothness of 1. The robust behavior of using monotone splines were also reported in Lin and Wang (2010) and Wang and Dunson (2010).

4. Fibroid data application

Right from the Start (RFTS) is a prospective, cohort study of early pregnancy. Eligible women were 18 years or older, enrolled by 13 weeks of gestation based on last menstrual period (LMP), did not use assisted reproductive technology, intended to carry the pregnancy to term, spoke English or Spanish, and did not plan to move for the next 18 months. For the details of the study, we refer to Laughlin et al. (2009). One of the research interests of this study was to estimate the cumulative incidence of fibroids for black and white women and to identify risk factors of fibroids. Endovaginal ultrasound was scheduled for all participants as early as possible in the pregnancy, with a goal of the 7th gestational week. The fibroid status of whether there had been some fibroid developed was determined through the ultrasound examination for each participant. Hence, the onset time of fibroid for each woman was never observed but was known to be larger rather than smaller at the ultrasound examination time. In other words, the onset time of fibroid for each woman was either left-censored or right-censored in the study, leading to a current status data structure.

RFTS contains three separately funded subsets, RFTS 1, 2 and 3. Because RFTS 1 involves substantial measurement error and RFTS 3 has a very small sample size, here we focus on the data from RFTS 2 only, which contains 1377 white women and 227 black women. Possible risk factors of the investigators' concern include: parity status (whether a participant has given birth before), age of menarche (when a participant had her first period), BMI (body mass index) status, and race (black or white). We use $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$ to denote the covariate vector for woman i, with $x_{i1} = 1$ indicating that woman i has given birth before and 0 otherwise, x_{i2} the standardized age of menarche, $x_{i3} = 1$ indicating that the women i has a BMI at least 30, which is the cutoff point for obesity, and $x_{i4} = 1$ indicating that woman i is black.

We implemented the Gibbs sampler described in Section 2.2 and summarized the results based on 5000 iterations of MCMC samples after discarding the first 1000 iterations as a burn-in. We also tried various of numbers of equally-spaced knots and degrees in specifying the monotone splines. Table 3 presents the estimation results in terms of the regression parameters in the cases that the degree is taken to be 1, 2, 3, and 4 and the number of interior knots is taken to be 10, 15, 20. It is clear that the proposed method produces very close estimates in all these cases indicating that the method is robust to the choice of the number of knots and degree in the spline specification. From Table 3, race is clearly a significant risk factor, and the hazard rate of developing fibroids for black women is about $e^{1.35} \approx 3.86$ times as high as the hazard rate for

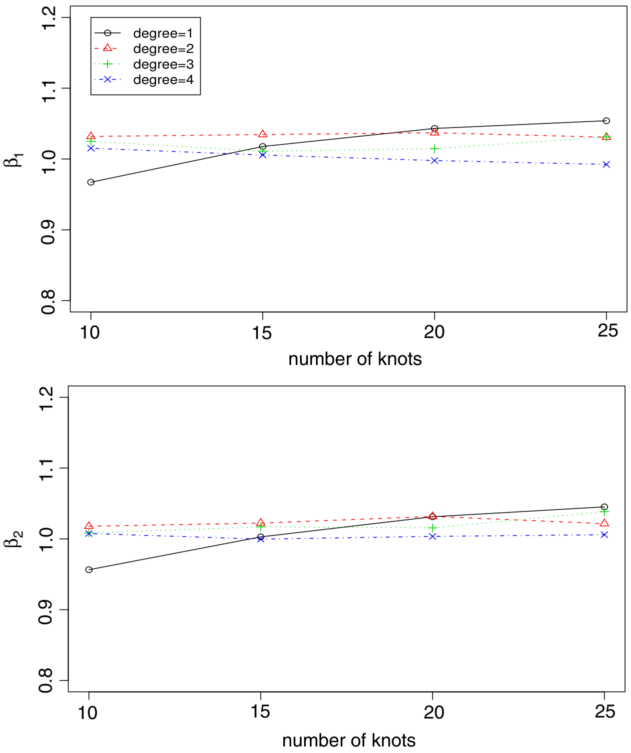

Fig. 1. The average of 500 point estimates of β_1 and β_2 with different degrees of smoothness (degree = 1, 2, 3 and 4) and different numbers of knots ($m = 10, 15, 20$ and 25) in the simulation case of $(\beta_1, \beta_2) = (1, 1)$.

Table 3
Results of the fibroid data analysis: point estimates and 95% interval estimates of the regression parameter for different numbers (m) of equally spaced knots and degrees (d) in the monotone spline specification.

		$m = 10$	$m = 15$	$m = 20$
$d = 1$	β_1	−0.343 (−0.622, −0.060)	−0.382 (−0.645, −0.107)	−0.394 (−0.665, −0.125)
	β_2	−0.149 (−0.274, −0.024)	−0.151 (−0.277, −0.026)	−0.150 (−0.273, −0.021)
	β_3	0.059 (−0.285, 0.387)	0.037 (−0.306, 0.368)	0.041 (−0.294, 0.368)
	β_4	1.374 (1.048, 1.691)	1.375 (1.049, 1.684)	1.345 (1.025, 1.662)
$d = 2$	β_1	−0.356 (−0.632, −0.074)	−0.379 (−0.650, −0.103)	−0.409 (−0.675, −0.135)
	β_2	−0.150 (−0.279, −0.027)	−0.150 (−0.280, −0.025)	−0.149 (−0.278, −0.025)
	β_3	0.045 (−0.307, 0.382)	0.042 (−0.312, 0.381)	0.028 (−0.318, 0.356)
	β_4	1.371 (1.055, 1.685)	1.354 (1.039, 1.670)	1.353 (1.036, 1.674)
$d = 3$	β_1	−0.347 (−0.620, −0.063)	−0.366 (−0.642, −0.088)	−0.403 (−0.668, −0.131)
	β_2	−0.147 (−0.274, −0.025)	−0.145 (−0.272, −0.021)	−0.147 (−0.275, −0.026)
	β_3	0.056 (−0.276, 0.387)	0.041 (−0.290, 0.374)	0.023 (−0.326, 0.355)
	β_4	1.366 (1.024, 1.674)	1.353 (1.029, 1.675)	1.346 (1.026, 1.677)
$d = 4$	β_1	−0.359 (−0.641, −0.077)	−0.383 (−0.646, −0.112)	−0.397 (−0.669, −0.120)
	β_2	−0.144 (−0.271, −0.018)	−0.147 (−0.272, −0.021)	−0.146 (−0.272, −0.020)
	β_3	0.047 (−0.310, 0.379)	0.039 (−0.300, 0.364)	0.033 (−0.317, 0.363)
	β_4	1.353 (1.037, 1.674)	1.352 (1.033, 1.660)	1.328 (1.013, 1.643)

the white race. Also, parity and the age of menarche have a statistically significant negative effect on the risk of the fibroid indicating that having given birth or having a larger age of menarche reduces the risk of developing fibroids. In contrast, the BMI status does not show a significant effect on the risk of developing fibroids. These findings agree with the expectation of epidemiologists and are consistent with the conclusion in Wang and Dunson (2010) under the proportional odds model.

Fig. 2 plots the estimated cumulative incidence functions for black women and white women, respectively, controlling all other covariates equal to 0. The corresponding 95% pointwise credible intervals are also given. From Fig. 2, it is clear that the black race has a much higher cumulative incidence curve than the white race. These curves are obtained in the case that the number of knots equals 15 and the degree is 3. Other choices of number of knots and degree give very similar results.

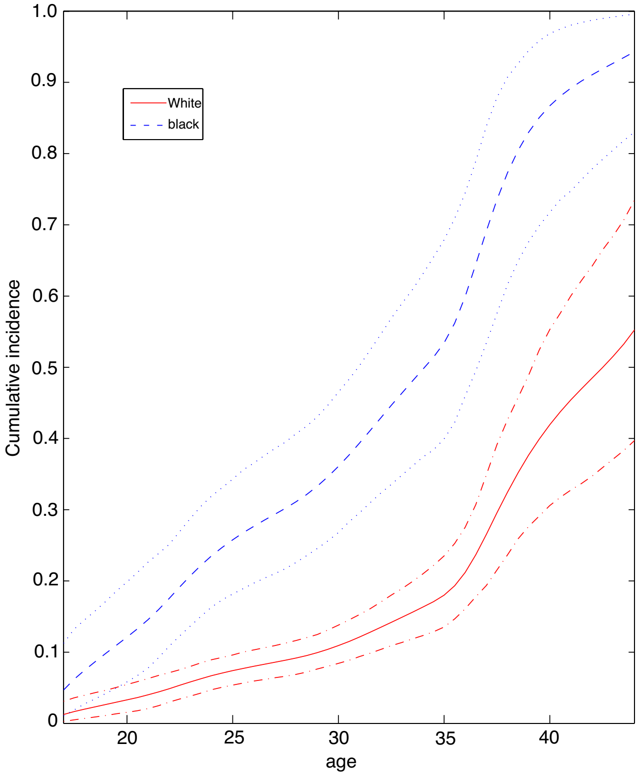

Fig. 2. The posterior means (dashed and solid) and the corresponding 95% credible intervals (dotted and dash–dotted) of the cumulative incidence of fibroids for black and white races in the baseline group (controlling all covariates equal to 0).

5. Discussion

In this paper, we propose a novel Bayesian approach for analyzing current status data under the semiparametric PH model. The use of monotone splines for modeling the baseline cumulative hazard function leads to two nice consequences: (1) there are only a finite number of parameters in the model; (2) the resulting estimate of survival (or hazard) function is smooth. Based on a data augmentation through Poisson latent variables, the posterior computation is straightforward and does not require any imputation of failure times. Our simulation results show that the proposed method works well and outperforms the likelihood method and three other existing approaches in terms of accuracy and ease of implementation.

The extension of the proposed approach to modeling the clustered or multivariate current status data is straightforward by introducing frailty terms to induce dependency among the failure times. Assuming the frailty terms to have a Gamma distribution, the proposed sampling algorithm only needs a little modification and updating the frailty terms is easy by sampling from some Gamma posterior distributions.

The proposed monotone spline specification may not foster easy computation for analyzing the right-censored data under the PH model. For the right-censored data, one can simply apply the partial likelihood method directly. To generalize the proposed Bayesian approach to general interval-censored data, one needs to develop more advanced data augmentation methods to facilitate the posterior computation due to the more complicated likelihood function. We are currently working in this direction with the hope of developing a fast and easy-to-implement statistical package that allows one to analyze any interval-censored data.

References

Cox, D., 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society Series B 34, 187–220.
Cox, D., 1975. Partial likelihood. Biometrika 62, 269–276.
Gilks, W., Best, N., Tan, K., 1995. Adaptive rejection Metropolis sampling within Gibbs sampling. Applied Statistics 44, 455–472.
Gilks, W., Wild, P., 1992. Adaptive rejection sampling for Gibbs sampling. Applied Statistics 41, 337–348.
Gómez, G., Calle, M., Oller, R., Langohr, K., 2009. Tutorial on methods for interval-censored data and their implementation in R. Statistical Modelling 9, 259–297.
Groeneboom, P., Wellner, J.A., 1992. Information Bounds and Non-Parametric Maximum Likelihood Estimation. Birkhauser, Boston.
Härkänen, T., 2003. BITE: a Bayesian intensity estimator. Computational Statistics 18, 564–583.
Henschel, V., Heiß, C., Mansmann, U., 2009a. The intcox package. Comprehensive R Archive Network.
Henschel, V., Heiß, C., Mansmann, U., 2009b. survBayes: an introduction into the package. Comprehensive R Archive Network.
Huang, J., 1996. Efficient estimation for the proportional hazards model with interval censoring. Annals of Statistics 24, 540–568.

Laughlin, S.K., Baird, D.D., Savitz, D.A., Herring, A.H., Hartmann, K.E., 2009. Prevalence of uterine leiomyomas in the first trimester of pregnancy. Obstetrics & Gynecology 113, 630–635.

Lin, X., Wang, L., 2010. Semiparametric probit model for case 2 interval-censored failure time data. Statistics in Medicine 29, 972–981.

Mongoué-Tchokoté, S., Kim, J.S., 2008. New statistical software for the proportional hazards model with current status data. Computational Statistics and Data Analysis 52, 4272–4286.

Pan, W., 1999. Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. Journal of Computational and Graphical Statistics 8, 109–120.

Ramsay, J.O., 1988. Monotone regression splines in action. Statistical Science 3, 425–441.

Shiboski, S.C., 1998. Generalized additive models for current status data. Lifetime Data Analysis 4, 29–50.

Sun, J., 2006. The Statistical Analysis of Interval-Censored Data. Springer.

Wang, L., Dunson, D.B., 2010. Semiparametric Bayes proportional odds models for current status data with under-reporting. Biometrics, doi:10.1111/j.1541-0420.2010.01532.x.

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