Journal of the Korean Statistical Society (2023) 52:361–381
https://doi.org/10.1007/s42952-023-00204-9

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

The expectation–maximization approach for Bayesian additive Cox regression with current status data

Di Cui[1] · **Clarence Tee**[2]

Received: 7 August 2022 / Accepted: 11 January 2023 / Published online: 22 January 2023

Abstract
In this paper, we propose a Bayesian additive Cox model for analyzing current status data based on the expectation–maximization variable selection method. This model concurrently estimates unknown parameters and identifies risk factors, which efficiently improves model interpretability and predictive ability. To identify risk factors, we assign appropriate priors on the indicator variables which denote whether the risk factors are included. By assuming partially linear effects of the covariates, the proposed model offers flexibility to account for the relationship between risk factors and survival time. The baseline cumulative hazard function and nonlinear effects are approximated via penalized B-splines to reduce the dimension of parameters. An easy to implement expectation–maximization algorithm is developed using a two-stage data augmentation procedure involving latent Poisson variables. Finally, the performance of the proposed method is investigated by simulations and a real data analysis, which shows promising results of the proposed Bayesian variable selection method.

Keywords Additive Cox model · Bayesian variable selection · Current status data · EM algorithm · Splines

1 Introduction

Current status data, also known as type II interval-censored data, is frequently encountered in cross-sectional studies across a range of disciplines from epidemiological to the social sciences (Chan et al., 2021; Koley & Dewanji, 2022). The true

✉ Di Cui
dicui6-c@my.cityu.edu.hk

Clarence Tee
teewtc@ihpc.a-star.edu.sg

1 Department of Advanced Design and Systems Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR

2 Institute of High-Performance Computing, A*STAR, Singapore, Singapore

time of failure can be larger or smaller than the observation time, since the survival status of a subject is assessed only once. For example, in the Worcester Heart Attack Study (WHAS) (White & Chew, 2008), patients' death time is either left or right censored since the survival states of patients were followed up only once after hospital discharge.

Traditionally, it is of main interest to study the effect of covariates on survival time. Cox's proportional hazards (PH) model (Cox, 1972) is one of the most popular frameworks for regression analysis of time-to-event data. Therefore, a number of methods have been developed to analyze current status data under the PH assumption and its numerous variants. Zhang and Sun (2010) reviewed the literature prior to 2010. As the baseline hazard function is unspecified, several methods were introduced to reduce the dimensionality. Zeng et al. (2016) and Gao and Chan (2019) treated the baseline hazard function as a step function to study nonparametric maximum likelihood estimation for a broad class of semi-parametric transformation models. Cai et al. (2011) discussed the regression analysis for current status data in a Bayesian framework. McMahan et al. (2013) developed an expectation–maximization (EM) algorithm under the same model specification with Cai et al. (2011). Gamage et al. (2020) extended McMahan et al. (2013)'s work to observed data consisting of instantaneous failures.

In real world applications, the rapidly paced development of technology allows more information to be collected and made available for analysis, resulting in a large number of candidate covariates. Keeping all covariates may result in overfitting, which poses a problem in estimation accuracy and model interpretability. Many statistical methods were proposed for variable selection, such as penalization procedures LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001) and Adaptive LASSO (Zou, 2006). In addition, Bayesian methods have also gained popularity (George & McCulloch, 1993, 1997; Liang et al., 2008; Narisetty & He, 2014). Held et al. (2016) combined test-based Bayes factors into the PH model for right-censored data. Nikooienejad et al. (2020) assigned a nonlocal prior on the regression coefficients for the PH model in the presence of right-censored data. A horseshoe prior was considered by Mu et al. (2021) to control the sparsity of covariates under PH assumptions. Despite the extensive literature on variable selection for right-censored data, little attention has been paid to the interval censored data in the Bayesian framework. The latest work concentrated on variable selection for such outcomes are mainly developed via frequentist-based regularization methods (Du et al., 2021, 2022; Li et al., 2020; Scolas et al., 2016; Wu & Cook, 2015; Zhao et al., 2019), which may be challenging when the model or underlying data structure becomes complicated.

Our reasons for revisiting this topic are twofold. Firstly, the assumption of linear dependencies between explanatory covariates and survival time may not be flexible enough to describe its complicated relationship, and imposing it incorrectly could cause biased estimators. By combining linear effect terms and nonlinear effect terms, which are described by unknown smooth functions, additive Cox models avoid the aforementioned drawbacks and allow each nonlinear effect to have its own distinguished effect curve and risk region. Lu and McMahan (2018) conducted statistical inference on current status data using a partially linear proportional hazards model with without variable selection. Secondly, expectation–maximization variable selection (EMVS) (Ročková & George, 2014) has been shown to be a deterministic Bayesian variable selection method from its efficiency at identifying related covariates and it has also been successfully applied to various model structures (Dai & Jin, 2022; Koslovsky et al., 2018a, 2018b; Zhao & Lian, 2016). Since this method selects all covariates simultaneously, it avoids multiple model comparison, which is the main challenge faced by traditional pairwise comparison methods using Bayesian model comparison statistics. Moreover, EMVS can simply conduct posterior inference through an EM algorithm, which outperforms the traditional Markov Chain Monte Carlo (MCMC) algorithm in terms of computational time (Ročková & George, 2014).

In this paper, we leverage on EMVS's validity in selecting covariates by extending it to additive Cox models to identify complex relationships between risk factors and survival time with current status data. To reduce the model dimensionality, penalized B-splines are utilized to approximate the baseline hazard function and the nonlinear effects. We develop an efficient EM algorithm that obtains parameter estimates through a two-stage data augmentation procedure involving latent Poisson variables and performs variable selection through latent index variables simultaneously. Furthermore, the additional constrained optimization procedure is avoided since the constraint of monotonicity of the baseline hazard function is satisfied directly during the closed form updates.

The rest of the paper is organized as follows. In Sect. 2, we describe the additive Cox models and propose an EMVS method for selecting covariates. Section 3 reports the results of simulation studies to illustrate the performance of the proposed method. In Sect. 4, we apply the proposed method to data from WHAS. Section 5 concludes with a summary and discussion.

2 The model framework

2.1 Model setup

For the event time T, the cumulative hazard function of an additive Cox model takes the form:

$$\Lambda(t|\boldsymbol{x},\boldsymbol{z}) = \Lambda_0(t)\exp\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{q} \phi_j\left(z_j\right)\right), \tag{1}$$

where $\Lambda_0(\cdot)$ is unknown baseline cumulative hazard function, $\boldsymbol{x}$ and $\boldsymbol{z}$ are $p \times 1$ and $q \times 1$ covariates respectively, $\boldsymbol{\beta}$ is the vector of regression coefficients and $\phi_j(\cdot), j = 1, \dots, q$ are unknown smooth functions. Given $\boldsymbol{x}$ and $\boldsymbol{z}$, the conditional cumulative distribution function (CDF) of event time T is written as

$$F(t|\boldsymbol{x},\boldsymbol{z}) = 1 - \exp\left\{-\Lambda_0(t)\exp\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{q} \phi_j\left(z_j\right)\right)\right\}. \tag{2}$$

To hold the conditions of a proper CDF, it is required that $\Lambda_0(\cdot)$ is a non-negative and monotone function with $\Lambda_0(0) = 0$. Furthermore for identifiability purposes, it is often assumed that $\phi_j(0) = 0$ for each j. Note that any constant effect is absorbed into $\Lambda_0(\cdot)$.

In the setting of current status data, the event time T is not observed directly. Let $t_i, i = 1, \ldots, n$ denote the event time for n subjects, and $c_i, i = 1, \ldots, n$ the observation time. The censoring indicator is defined as $\delta_i = I(t_i \leq c_i), i = 1, \ldots, n$. Then the likelihood function of all observed data $\mathcal{D} = \left\{ (c_i, \delta_i, x_i^{\mathrm{T}}, z_i^{\mathrm{T}})^{\mathrm{T}} \right\}, i = 1, \ldots, n$ is given by

$$L(\beta, \Lambda_0, \phi_1, \ldots, \phi_q | \mathcal{D}) = \prod_{i=1}^n F(c_i | \boldsymbol{x}_i, \boldsymbol{z}_i)^{\delta_i} \left[1 - F(c_i | \boldsymbol{x}_i, \boldsymbol{z}_i)\right]^{1-\delta_i}, \tag{3}$$

which is under the assumption that the failure time and censoring time are independent given covariates.

In the above likelihood, baseline cumulative hazard function $\Lambda_0(\cdot)$ and smooth function $\left\{\phi_j(\cdot)\right\}_{j=1}^q$ are unspecified with parameters of infinite dimensions. Splines are a feasible technique to reduce the dimension while maintaining model flexibility as it makes no assumptions about the shape of the fitted curve. In this article, $\Lambda_0(\cdot)$ and $\left\{\phi_j(\cdot)\right\}_{j=1}^q$ are all approximated via B-splines because it has good properties both in theoretical and computational aspects (Höllig & Hörner, 2013). These functions are constructed as follows,

$$\Lambda_0(\cdot) = \sum_{k=1}^K \alpha_k B_k(\cdot), \tag{4}$$

and

$$\phi_j(\cdot) = \sum_{l=1}^L h_{jl} B_{jl}(\cdot), \tag{5}$$

where $\left\{B_k(\cdot)\right\}_{k=1}^K$ and $\left\{B_{jl}(\cdot)\right\}_{l=1}^L$ are B-spline basis functions and $\left\{\alpha_k\right\}_{k=1}^K$ as well as $\left\{h_{jl}\right\}_{l=1}^L$, their corresponding spline coefficients, respectively. In order to guarantee the non-negativity and monotonicity of $\Lambda_0(\cdot)$, a constraint is put on the spline coefficients, that is $0 \leq \alpha_1 \leq \cdots \leq \alpha_K$. Schumaker (2007) indicated that this constraint is sufficient to guarantee the non-negativity and monotonicity. $\Lambda_0(\cdot)$ and $\left\{\phi_j(\cdot)\right\}_{j=1}^q$ are all required to be 0 when the independent variable being 0, which is satisfied automatically if the intercept term of the basis functions is omitted (which is the only function take the nonzero value at 0). To construct the basis functions, the number and locations of interior knots need to be specified in order to determine the shape, as well as the degree, which controls the smoothness of the model. The total number of basis functions K and L is the sum of the number of their corresponding interior knots and the degree.

It is well known that the order as well as the number and location of interior knots have an impact on model fit. In general, cubic spline (of order 3) is smooth enough to fit the curve, whereas too many (few) knots lead to over (under) fitting. To overcome the difficulty, Eliers and Marx (1996) recommend a relatively large number of interior knots, and then apply a discrete roughness penalty in the B-spline coefficients to counterbalance the flexibility of the generous basis functions. The penalized B-splines (also known as P-splines) are introduced in detail in Sect. 2.3. Once the number of interior knots has been determined, the knots are equidistantly placed at the quantiles of support of splines.

Denoting $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$, $\boldsymbol{B}(c_i) = (B_1(c_i), \ldots, B_K(c_i))^{\mathrm{T}}$, $\boldsymbol{h}_j = (h_{j1}, \ldots, h_{jL})^{\mathrm{T}}$ and $\boldsymbol{B}_j(z_{ij}) = (B_{j1}(z_{ij}), \ldots, B_{jL}(z_{ij}))^{\mathrm{T}}$. The cumulative hazard function (1) can be written in vector notation as:

$$\Lambda\left(t|\boldsymbol{x}_i, \boldsymbol{z}_i\right) = \boldsymbol{\alpha}^{\mathrm{T}} B\left(c_i\right) \exp\left(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}} \boldsymbol{B}\left(\boldsymbol{z}_i\right)\right), \tag{6}$$

where $\boldsymbol{h} = (h_1^{\mathrm{T}}, \ldots, h_q^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{B}(z_i) = (B_1^{\mathrm{T}}(z_{i1}), \ldots, B_q^{\mathrm{T}}(z_{iq}))^{\mathrm{T}}$.

2.2 Data augmentation for the EM algorithm

Direct maximization of Eq. (3) is intractable because of its complex form. Hence, in the spirit of Wang et al. (2016), an EM algorithm is proposed to estimate the unknown parameters. In order to derive the algorithm, a two-stage data augmentation involving latent Poisson random variables is utilized based on the relationship between the Cox model and a non-homogeneous Poisson process.

The first stage is to associate the censoring indicator δ_i with a non-homogeneous Poisson process w_i with mean $\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{B}\left(c_i\right) \exp\left(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}} \boldsymbol{B}\left(\boldsymbol{z}_i\right)\right)$ so that $\delta_i = I\left(w_i > 0\right)$. At the second stage, w_i is decomposed into a summation of independent Poisson processes w_{ik}. Therefore, the data augmentation procedure is summarized as follows,

$$\delta_i = I\left(w_i > 0\right), w_i \sim \text{Poisson}\left(\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{B}\left(c_i\right) \exp\left\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}} \boldsymbol{B}\left(\boldsymbol{z}_i\right)\right\}\right),$$

$$w_i = \sum_{k=1}^{K} w_{ik}, w_{ik} \sim \text{Poisson}\left(\alpha_k B_k\left(c_i\right) \exp\left\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}} \boldsymbol{B}\left(\boldsymbol{z}_i\right)\right\}\right), k = 1, \ldots, K. \tag{7}$$

Denoting $\boldsymbol{w} = (w_{11}, \ldots, w_{1K}, \ldots, w_{n1}, \ldots, w_{nk})^{\mathrm{T}}$ and $P_{w_{ik}}(\cdot)$ as the Poisson mass function associated with the random variable w_{ik}, the augmented data likelihood can be expressed as:

$$L_{aug}(w|D) = \prod_{i=1}^{n} \prod_{k=1}^{K} P_{w_{ik}}\left(w_{ik}\right)\left(\delta_i I\left(w_i > 0\right) + \left(1 - \delta_i\right) I\left(w_i = 0\right)\right). \tag{8}$$

which is the Poisson distribution function subject to the constraints $w_i > 0$ if $\delta_i = 1$ and $w_i = \sum_{k=1}^{K} w_{ik}$ otherwise. Full details of the data augmentation procedure are provided in Appendix 1. The first stage transfers the complex likelihood function into Poisson distribution. The second stage splits $\boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{B}\left(c_i\right)$ into independent $\left\{\alpha_k B_k\left(c_i\right)\right\}_{k=1}^{K}$, which simplifies the calculation in the following EM algorithm.

2.3 Prior specification

To facilitate Bayesian variable selection, the well-known spike-and-slab prior is assigned to the regression coefficients $\boldsymbol{\beta}$. An indicator variable $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^{\mathrm{T}}$ is introduced to identify i.e., $\beta_m = 0$ if $\gamma_m = 0$ and $\beta_m \neq 0$ otherwise for $m = 1, \ldots, p$. Thus, the prior we assigned to $\boldsymbol{\beta}$ is

$$P\left(\boldsymbol{\beta}|\boldsymbol{\gamma},\sigma^2,v_1\right)=N\left(0,\Sigma_p\right)\text{with}\Sigma_p=\sigma^2\text{diag}\left(d_1,\ldots,d_p\right),$$

where $N(\cdot,\cdot)$ is normal density function and $d_m=\left(1-\gamma_m\right)\cdot v_0+\gamma_m\cdot v_i$ for $0<v_0<v_1$. Though v_0 is often set to 0 in practice, George and McCulloch (1997) recommended setting a small and positive v_0 to exclude unimportant nonzero effects. v_0 of the spike distribution serves to pull coefficients estimates toward zero. The increase of v_0 enlarges the variance of the spike component, which has the effect of shrinking the small effect without drastically affecting the significant effects. To ensure large coefficients possibly unaffected by the shrunken spike prior, a heavy-tailed slab prior suggested by Ročková and George (2014) is induced for v_1,

$$P\left(v_1\right)=\frac{v_1^{b_1}\left(1+v_1\right)^{-a_1-b_1-2}}{B\left(a_1+1,b_1+1\right)}I\left(v_1>0\right),$$

where $B(\cdot,\cdot)$ is a Beta function. Referring to Cui and George (2008), Liang et al. (2008) and Maruyama and George (2011), the hyper-parameters a_1 and b_1 are set to be 0 and $-3/4$ for better performance. The setting of a_1 and b_1 makes flatter proper prior, resulting in stable estimation.

Without extra structural information about the predictors, the i.i.d Bernoulli prior is chosen for $\boldsymbol{\gamma}$,

$$P(\boldsymbol{\gamma}|\omega)=\omega^{|\gamma|}(1-\omega)^{p-|\gamma|},$$

where $|\boldsymbol{\gamma}|=\sum_{m=1}^{q}\gamma_m$ and ω is a hyperparameter following the uniform distribution $U(0,1)$.

The roughness penalty for the B-spline coefficients $\boldsymbol{\alpha}$ is based on squared differences of the coefficients of adjacent splines to ensure sufficient smoothness of the fitted curve. A commonly used second-order difference penalty is $\sum_k(\alpha_k-2\alpha_{k-1}+\alpha_{k-2})^2=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha}$, where

$$\boldsymbol{Q}=\begin{pmatrix}1 & -2 & 1 & 0 & \cdots & 0\\ 0 & 1 & -2 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots\\ 0 & 0 & \cdots & 1 & -2 & 1\end{pmatrix}_{(K-2)\times K}.$$

The roughness penalty, given a hyperparameter λ, can be translated into a prior from differences between the coefficients (Lang & Brezger, 2004),

$$P\left(\alpha_k-\alpha_{k-1}|\lambda\right)=N(0,\lambda),k=1,\ldots,K.$$

The elements of $\boldsymbol{\alpha}$ are restricted as $0=\alpha_0\le\alpha_1\le\cdots\le\alpha_K$ to obtain the non-negativity and monotonicity of $\Lambda_0(\cdot)$. The constraint is imposed by incorporating the indicator functions (Brezger & Steiner, 2008) into the prior. Therefore, the prior for $\boldsymbol{\alpha}$ is formulated as follows:

$$P(\boldsymbol{\alpha}) \propto \lambda^{\frac{K}{2}} \exp\left(-\frac{\lambda}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{\alpha}\right) \prod_{k=1}^{K} I(\alpha_k \geq \alpha_{k-1}),$$

where $\boldsymbol{P} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} + \varepsilon I$ is a full rank matrix with a small, positive constant ε.

Similarly, we assign a second-order difference penalty to $\boldsymbol{h}_j$. The prior is given by

$$P(\boldsymbol{h}_j) \propto \tau_j^{\frac{L}{2}} \exp\left(-\frac{\tau_j}{2}\boldsymbol{h}_j^{\mathrm{T}}\boldsymbol{P}\boldsymbol{h}_j\right),$$

where denotes τ_j as the penalty parameter.

The priors for σ^2, λ and $\{\tau_j\}_{j=1}^{q}$ are set as IG(a_2, b_2), Ga(a_3, b_3) and Ga(a_4, b_4), resulting in uninformative priors, where Ga denote the Gamma distribution with probability density function (pdf)

$$f(x;a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$$

and IG denotes the inverse Gamma distribution with pdf

$$f(x;a,b) = \frac{b^a}{\Gamma(a)} \left(\frac{1}{x}\right)^{a+1} \exp\left(-\frac{b}{x}\right),$$

respectively. The hyperparameters $a_2, b_2, a_3, b_3, a_4, b_4$ are set to be 0.5 in all numerical experiments.

Finally, the joint posterior distribution of all the parameters is given by

$$\begin{aligned} &L_c\left(\boldsymbol{\alpha}, \beta, h, \gamma, \sigma^2, v_1, \omega, \lambda, \tau | \mathcal{D}, w\right) \propto P(\mathcal{D}, w|\boldsymbol{\alpha}, \beta, h) P\left(\beta|\gamma, \sigma^2, v_1\right) P\left(\sigma^2\right) P\left(v_1\right) \\ &\times P(\gamma|\omega) P(\omega) P(\boldsymbol{\alpha}|\lambda) P(\lambda) P(h|\tau) P(\tau), \end{aligned} \tag{9}$$

where the term $P(\mathcal{D}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{h})$ is $L_{aug}(\boldsymbol{w}|\mathcal{D})$.

2.4 EM algorithm

An EM algorithm is derived to find the posterior maximized parameters iteratively as an alternative to the conventional MCMC approach, which possesses computational efficiency over stochastic search alternatives.

The EM algorithm begins with the Expectation (E-step) of the logarithm of L_c with respect to the latent variables ($\boldsymbol{w}$ and $\boldsymbol{\gamma}$) conditioned on the observed data $\mathcal{D}$ and current parameter estimate, whereafter, the maximum (M-step) likelihood estimators of the expected log-posterior likelihood resulting from E-step are calculated. Each parameter is estimated under the condition that the remaining parameters are fixed in M-step. The two steps are repeated iteratively until convergence is achieved.

Denoting $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{h}, \sigma^2, v_1, \omega, \lambda, \tau\}$. In the $(u+1)$th iteration, the expected log likelihood in E-step is given as:

$$E\left[\log L_c | \mathcal{D}, \boldsymbol{\theta}^{(u)}\right] = Q_1\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{h}, \sigma^2, v_1, \lambda, \tau | \mathcal{D}, \boldsymbol{\theta}^{(u)}\right) + Q_2\left(\omega | \mathcal{D}, \boldsymbol{\theta}^{(u)}\right) + L\left(\boldsymbol{\theta}^{(u)}\right),$$

where $L\left(\boldsymbol{\theta}^{(u)}\right)$ is a function of $\boldsymbol{\theta}^{(u)}$ but is irrelevant of $\boldsymbol{\theta}$ and

$$Q_1\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{h}, \sigma^2, v_1, \lambda, \tau | \mathcal{D}, \boldsymbol{\theta}^{(u)}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K}\left\{E\left(w_{ik}|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)\left[\log\alpha_k + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}}\boldsymbol{B}\left(\boldsymbol{z}_i\right)\right]\right.$$

$$\left.-\alpha_k B_k\left(c_i\right)\exp\left[\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}}\boldsymbol{B}\left(\boldsymbol{z}_i\right)\right]\right\} - \frac{1}{2}\sum_{m=1}^{p} E(\log d_m|\mathcal{D}, \boldsymbol{\theta}^{(u)})$$

$$-\left(\frac{p}{2} + a_2 + 1\right)\log\sigma^2 - \frac{1}{\sigma^2}\left[b_2 + \frac{1}{2}\sum_{m=1}^{p} E\left(\frac{1}{d_m}\middle|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)\beta_m^2\right]$$

$$+b_1\log v_1 - \left(a_1 + b_1 + 2\right)\log\left(1 + v_1\right) + \left(\frac{K}{2} + a_3 - 1\right)\log\lambda - \left(\frac{\lambda\boldsymbol{\alpha}^{\mathrm{T}}P\boldsymbol{\alpha}}{2} + b_3\right)$$

$$+\sum_{j=1}^{q}\left[\left(\frac{L}{2} + a_4 - 1\right)\log\tau_j - \left(\frac{\tau_j\boldsymbol{h}_j^{\mathrm{T}}P\boldsymbol{h}_j}{2} + b_4\right)\right],$$

$$Q_2\left(\omega|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right) = \sum_{m=1}^{p} E\left(\gamma_m|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)\log\omega + \left(p - \sum_{m=1}^{p} E\left(\gamma_m|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)\right)\log(1 - \omega).$$

The E-steps proceeds by computing the conditional expectation $E\left(w_{ik}|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)$, $E\left(\gamma_m|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)$, $E\left(\log d_m|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)$ and $E\left(1/d_m|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)$ from Q_1 and Q_2. The next step is to find $\boldsymbol{\theta}^{(u+1)}$ which maximizes Q_1 and Q_2. The details of the EMVS algorithm are discussed in Appendix 2.

In summary, the EMVS algorithm proceeds as follows:

Step 1. Initialize the parameters $\boldsymbol{\theta}^{(0)}$;

Step 2. Evaluate the conditional expectations $E\left(\gamma_m|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)$, $E\left(\log d_m|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)$, $E\left(1/d_m|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)$ and $E\left(w_{ik}|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)$;

Step 3. Obtain $\theta^{(u+1)}$ by maximizing $Q\left(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)$;

Step 4. Iterate between steps 2 and 3 until the maximum absolute difference of $\boldsymbol{\theta}$ between two successive iterations is smaller than 10^{-5}.

The decision of variable selection is based on the probability $P(\gamma_m = 1)$, that is, the conditional expectation p_m^* calculated in the E-step. The default threshold value is 0.5, i.e., x_m is selected if $p_m^* > 0.5$.

2.5 Variance estimation

The covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be estimated based on the profile likelihood, where $\widehat{\boldsymbol{\beta}}$ is defined as estimates of $\boldsymbol{\beta}$. Denoting $\boldsymbol{\psi} = \left\{\boldsymbol{\alpha}, \boldsymbol{h}, \sigma^2, v_1, \omega, \lambda, \tau\right\}$, then the profile log-likelihood is defined as:

$$l_{pn}(\boldsymbol{\beta}) = \max_{\boldsymbol{\psi}} \log\left(L \times P\left(\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2, v_1\right)P\left(\sigma^2\right)P\left(v_1\right)P(\boldsymbol{\gamma}|\omega)P(\boldsymbol{\alpha}|\lambda)P(\lambda)P(\boldsymbol{h}|\tau)\right)P(\tau).$$

Referring to Zeng et al. (2016), the covariance matrix of $\widehat{\boldsymbol{\beta}}$ is calculated by the negative inverse of the information matrix $I(\widehat{\boldsymbol{\beta}})$. As $l_{pn}(\boldsymbol{\beta})$ does not have a closed form solution, the (s,t)th element of $I(\widehat{\boldsymbol{\beta}})$ is approximated by a second order numerical difference:

$$\frac{l_{pn}\left(\widehat{\boldsymbol{\beta}}\right) - l_{pn}\left(\widehat{\boldsymbol{\beta}} + \epsilon_n \boldsymbol{e}_s\right) - l_{pn}\left(\widehat{\boldsymbol{\beta}} + \epsilon_n \boldsymbol{e}_t\right) + l_{pn}\left(\widehat{\boldsymbol{\beta}} + \varepsilon_n \boldsymbol{e}_s + \varepsilon_n \boldsymbol{e}_t\right)}{\epsilon_n^2},$$

where $\boldsymbol{e}_s$ is a p-dimensional vector. The sth element is 1 and the remaining is 0, and ϵ_n is a tuning constant of order $n^{-1/2}$. The value of $l_{pn}(\boldsymbol{\beta})$ can be computed using the EM algorithm again with $\boldsymbol{\beta}$ held fixed.

3 Simulation studies

In this section, simulation studies are conducted to illustrate the performance of the proposed method. We independently generate 400 data sets with the sample size of $n = 200$ and 500, respectively, from the following model:

$$F(t|\boldsymbol{x}, z) = 1 = \exp\left[-\Lambda_0(t)\exp\left\{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + \phi(z)\right\}\right],$$

where $\boldsymbol{\beta} = (1, 0, 2, 0, -1, 0, 0, 0)$ is a 8×1 vector. For the discrete covariates vector $\boldsymbol{x}_d = (x_1, x_2)^{\mathrm{T}}$, we independently sample x_1 and x_2 from Bernoulli(p) with $p = 0.5$. The remaining continuous covariates vector $\boldsymbol{x}_c = (x_3, \dots, x_8)^{\mathrm{T}}$ follows a multi-normal distribution with mean 0 and covariance matrix $\left(0.5^{|k-l|}\right)_{1 \le k,l \le 6}$. $\boldsymbol{z} = (z_1, z_2)^{\mathrm{T}}$ is independently sampled from $U(-1, 1)$. The censoring time C is generated from an exponential distribution with mean 1, and the censoring indicator is $\delta = I(t \le c)$. We consider two scenarios: (i) $\Lambda_0(t) = \log(1 + t)$, $\phi(z) = \sin\left(\pi z_1\right)$; (ii) $\Lambda_0(t) = t$, $\phi(z) = \sin\left(\pi z_1\right) + z_2^2 - 1$. For each scenario, the proposed EMVS method for additive Cox model is taken into account.

For specification of the B-splines used in estimating $\Lambda_0(\cdot)$ and $\phi_j(\cdot)$, the cubic basis functions are utilized to ensure adequate smoothness. We set 12 equally spaced interior knots for each function within the minimum and maximum of c and z_j (Bremhorst & Lambert, 2016; Çetinyürek & Lambert, 2010). Then K and L are equal to 15. v_0 of the spike distribution is set to be 0.01, and we empirically find that the estimation results are robust for the variation of v_0 within $[0.001, 0.1]$. This program is implemented in R on core i7 CPU, 2.60 GHz, 6 cores with Windows 10. Running the EMVS algorithm once for parameter and variance estimation spends around 1 min and 2 min with 200 and 500 sample sizes for scenario (i), respectively. And it takes around 1.2 min and 2.5 min with 200 and 500 sample sizes for scenario (ii), respectively.

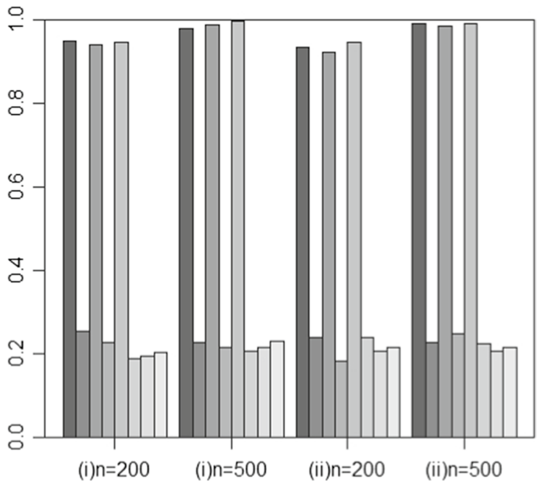

Fig. 1 The posterior mean of $\gamma_m, m = 1, \dots, 8$ for two scenarios

Table 1 FPR and FNR of EMVS and lasso for scenario (i)

n	(K, L)	FPR	FNR
EMVS			
200	(15, 15)	0.012	0.024
500	(15, 15)	0.001	0.001
LASSO			
200	(5, 5)	0.089	0.093
	(10, 10)	0.091	0.075
	(15, 15)	0.085	0.071
500	(5, 5)	0.069	0.005
	(10, 10)	0.049	0.005
	(15, 15)	0.027	0.001

We use LASSO as a benchmark to compare the variable selection accuracy with the proposed method under the similar model specifications. The LASSO estimates are calculated based on Eq. (8) using EM algorithm. The details of the benchmark are discussed further in Appendix 3.

The decision on the variable selection is made based on the posterior probability $P(\gamma_m = 1)$. Figure 1 displays the posterior mean of γ_m for two scenarios. The false positive rate (FPR) and false negative rate (FNR) are important indexes to evaluate the variable selection accuracy. They are defined as FPR = FP/(FP + TN) and FNR = FN/(FN + TP), where FP is the number of false positives, FN is the number of false negatives, TP is the number of true positives and TN is the number of true negatives. We independently generated 400 data sets for each scenario. Tables 1 and 2 reported the average FPR and FNR for the two scenarios. We tried different K and L for LASSO. It is shown that EMVS outperforms LASSO in different settings. EMVS exhibits a considerable accuracy of variable selection even in a small sample

Table 2 FPR and FNR of EMVS and lasso for scenario (ii)

n	(K, L)	FPR	FNR
EMVS			
200	(15, 15)	0.033	0.047
500	(15, 15)	0.005	0.000
LASSO			
200	(5, 5)	0.098	0.087
	(10, 10)	0.087	0.090
	(15, 15)	0.086	0.083
500	(5, 5)	0.053	0.005
	(10, 10)	0.042	0.023
	(15, 15)	0.020	0.040

Table 3 Summary on the estimation of the non-zero coefficients for scenario (i)

n	Effect	Bias	MSE	MCE	SEE
	β_1	− 0.0872	0.0867	0.3549	0.3503
200	β_3	− 0.0811	0.0661	0.2962	0.2849
	β_5	0.0623	0.0572	0.2123	0.2258
	β_1	− 0.0351	0.0383	0.1973	0.1934
500	β_3	0.0258	0.0427	0.1853	0.1882
	β_5	− 0.0249	0.0199	0.1292	0.1228

Table 4 Summary on the estimation of the non-zero coefficients for scenario (ii)

n	Effect	Bias	MSE	MCE	SEE
	β_1	− 0.0922	0.0824	0.3549	0.3663
200	β_3	− 0.0768	0.0733	0.3271	0.3058
	β_5	0.0675	0.0627	0.2334	0.2349
	β_1	− 0.0457	0.0436	0.2065	0.2123
500	β_3	0.0362	0.0289	0.1832	0.1854
	β_5	− 0.0187	0.0216	0.1216	0.1187

size. The results of LASSO are more conservative as it inclined to retain more variables.

The estimation results using EMVS of linear effects based on the 400 data sets are summarized in Tables 3 and 4, including bias, mean square error (MSE) between the estimated parameters and the true values, the Monte Carlo standard error (MCE) and the average of the numerical standard error (SEE). In these tables, SEE is computed using the square root of the negative inverse of $I(\widehat{\boldsymbol{\beta}})$ defined in Sect. 2.5, and MCE is the standard deviation of the 400 estimates for each parameter. The estimation accuracy improves with increasing sample sizes and performance did not appear to depend on the number of interior knots. For the variance estimation, we set $\epsilon_n = 10n^{-1/2}$ for all cases. The variance method is quite accurate even in small sample sizes. Also, the increase of the log-likelihood values for scenario (i) is shown in Appendix 4 as an example to illustrate the efficiency of the proposed method.

To assess the estimation of the cumulative baseline hazard function $\Lambda_0(\cdot)$ and the unknown smooth function $\phi_j(\cdot)$, Fig. 2 presents the point-wise medians of the estimated curves obtained by EMVS based on the 400 data sets for scenario (i). We show part of the different settings and it can be seen that the estimated curves coincide with the true curves.

The prediction performance of the proposed method is evaluated by the receiver operating characteristic (ROC) curve, which is a commonly used graphical tool in binary classification problems. The ROC curve plots the TRP against FRP given the prediction time point t. Prediction accuracy is assessed by the area under the curve (AUC). We followed the method proposed by Díaz-Coto et al. (2020) to calculate the AUC for current status data. Figure 3 displays the ROC curves for scenario (i) with different prediction time point t.

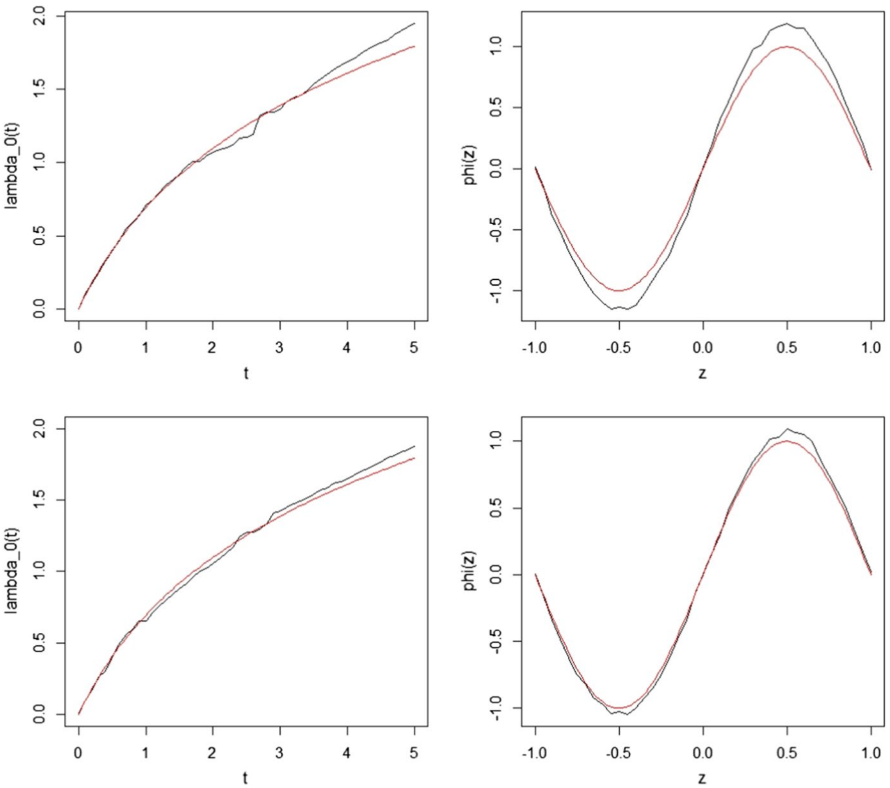

Fig. 2 Summary of 400 estimates of $\Lambda_0(\cdot)$ (left) and $\phi(\cdot)$ (right) obtained by EMVS for scenario (i), when $n = 200$ (top) and $n = 500$ (bottom). The results include point-wise median (black curve) and the true function curve (red curve)

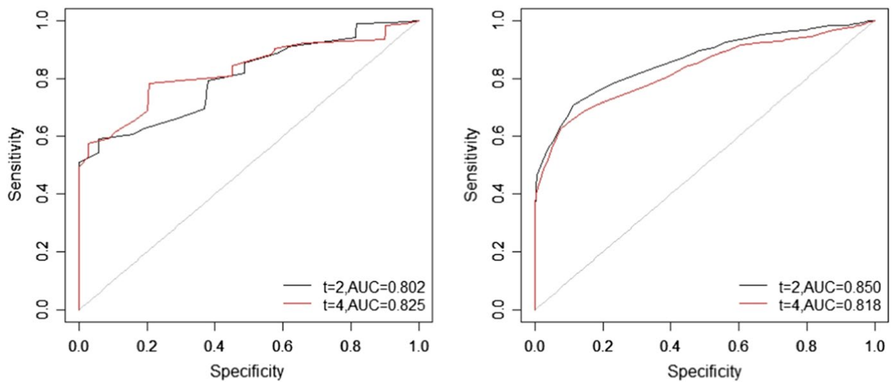

Fig. 3 ROC curves for scenario (i) with $n = 200$ (left) and $n = 500$ (right)

4 Real data analysis

In this section, we applied the proposed Bayesian variable selection procedure for the additive Cox model to the Worcester Heart Attack Study (WHAS) data set used in Hosmer and Lemeshow (2002). The goal of WHAS is to describe the time trend associated with risk factors in long-term survival among residents following acute myocardial infarction (MI). We aim to identify the risk factors that affect the survival time of patients after hospital discharge. A total of 461 patients were followed up from the hospital discharge years 1997, 1999 and 2001. As only the status of patients at last follow-up were reported, for those patients who died before the last follow-up, the true survival time is unknown but is known to be earlier than the date of the last follow-up, which is left-censored. Among these 461 patients, there are 176 deaths, about censoring level of 38%. The response variable is the number of days between the date of last follow-up and the date of hospital discharge. This dataset contains 14 risk factors, age at hospital admission(age), gender (0 = male, 1 = female), initial heart rate (hr), initial systolic blood pressure (sysbp), initial diastolic blood pressure (diasbp), body mass index (bmi), history of cardiovascular disease (cvd, 0 = no, 1 = yes), atrial fibrillation (afb, 0 = no, 1 = yes), cardiogenic shock (sho, 0 = no, 1 = yes), congestive heart complications (chf, 0 = no, 1 = yes), complete heart block (av3, 0 = no, 1 = yes), MI order (miord, 0 = first, 1 = recurrent), MI type (mitype, 0 = non Q-wave, 1 = Q-wave) and cohort year (year, 1 = 1997, 2 = 1999, 3 = 2001).

To model survival time, Hosmer and Lemeshow (2002) suggested fitting the Cox Proportional Hazards model with six explanatory variables: age, bmi, hr, diasbp, gender and chf, where bmi indicates a statistically significant departure from linearity. For better modeling performance, all the explanatory variables are standardized to $[0, 1]$. Based on this result, the Additive Cox model for the WHAS data set is given by

$$\begin{aligned}\Lambda(t|\text{covariates}) =& \Lambda_0(t)\exp\big\{\text{age}\beta_1 + \text{gender}\beta_2 + \text{hr}\beta_3 + \text{sysbp}\beta_4 \\ & + \text{diasbp}\beta_5 + \text{cvd}\beta_6 + \text{afb}\beta_7 + \text{sho}\beta_8 + \text{chf}\beta_9 \\ & + \text{av3}\beta_{10} + \text{miord}\beta_{11} + \text{mitype}\beta_{12} + \text{year}\beta_{13} + \phi(\text{bmi})\big\}.\end{aligned}$$

Table 5 The estimation results of WHAS data based on EMVS methods

Factor	Estimate	SEE	Posterior mean of γ	Index[a]
Age	4.0514	1.2815	1.000	1
Gender	− 0.2160	0.1362	0.095	0
hr	0.2441	0.2268	0.109	0
sysbp	0.0233	0.2073	0.055	0
diasbp	− 0.9092	0.1888	0.999	1
cvd	− 0.0462	0.1518	0.056	0
afb	0.0639	0.1487	0.058	0
sho	0.0247	0.2004	0.055	0
chf	1.0247	0.1813	1.000	1
av3	0.0308	0.1967	0.055	0
miord	0.0666	1.1332	0.058	0
mitype	− 0.2199	0.1457	0.096	0
Year	− 0.2088	0.1519	0.091	0

[a]Index = 1 means that the factor is selected, index = 0 otherwise

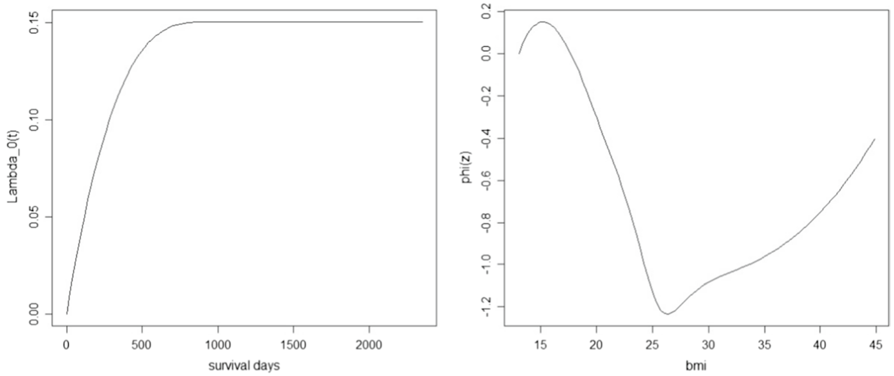

Fig. 4 The estimated baseline cumulative hazard function (left) and the unknown smooth function of bmi (right) of WHAS data set

The censoring indicator δ is set to be 1 if the onset time of patients is left censored, and 0 otherwise. In the model implementation, the baseline hazard function $\Lambda_0(\cdot)$ and the smooth regression function both are approximated by cubic B-splines with 12 interior knots, that is, $K = L = 15$.

Data analysis results are reported in Table 5. Three factors: age, diasbp and chf are identified as risk factors, which happen to be a subset of the significant covariates from Hosmer and Lemeshow (2002). Age is the dominant factor that affect the survival time of a patient. Older age and the existence of congestive heart complications will decrease the survival probability of people, and higher diastolic blood pressure means a lower risk of death event. Figure 4 displays the fitted curves of $\Lambda_0(\cdot)$ and $\phi(\cdot)$. It can be seen that the hazard rate monotonically increases with time to a threshold and plateaus. The factor bmi exhibits a clear nonlinear effect of the survival time. Too high or too low bmi both greatly increase the survival hazard.

5 Conclusion

In this paper, we develope an EMVS algorithm for variable selection of additive Cox models in the context of survival analysis. Based on the spike-and-slab prior and the two-stage data augmentation procedure, the proposed method is efficient and easy to implement. Both the simulation studies and the WHAS data analysis demonstrate the good performance of the proposed method. This method can be extended to other types of censored data, such as the type II interval-censored data, where the true failure time is known to fall into a particular interval. The two-stage data augmentation procedure can also be applied to a type II interval-censored data directly. Furthermore, this Bayesian framework can also be applied to other model structures for current status data to meet different needs, for example, the proportional odds model, the additive hazards model and cure model.

Appendix 1: Derivation of the augmented likelihood function

The likelihood function of all observed data is

$$L = \prod_{i=1}^{n} \left[1 - \exp\left\{-\boldsymbol{\alpha}^{\mathrm{T}} B(c_i)\exp\left(x_i^{\mathrm{T}}\beta + h^{\mathrm{T}}B(z_i)\right)\right\}\right]^{\delta_i} \times \exp\left\{-\boldsymbol{\alpha}^{\mathrm{T}} B(c_i)\exp\left(x_i^{\mathrm{T}}\beta + h^{\mathrm{T}}B(z_i)\right)\right\}^{1-\delta_i}.$$

In the first stage of data augmentation, we assume that $w_i \sim \text{Poisson}\left(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{B}(c_i)\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}}\boldsymbol{B}(z_i)\right)\right)$. Obviously,

$$P_{w_i}\left(w_i > 0\right) = 1 - \exp\left\{-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{B}\left(c_i\right)\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}}\boldsymbol{B}\left(z_i\right)\right)\right\},$$

and

$$P_{w_i}\left(w_i = 0\right) = \exp\left\{-\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{B}\left(c_i\right)\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}}\boldsymbol{B}\left(z_i\right)\right)\right\}.$$

In the second stage of data augmentation, w_i is represented as the summation of K independent random variables w_{ik}, where $w_{ik} \sim \text{Poisson}\left(\alpha_k B_k(c_i)\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}}\boldsymbol{B}(z_i)\right)\right)$, $k = 1, \ldots, K$. Therefore,

$$P_{w_i}\left(w_i\right) = \prod_{k=1}^{K} P_{w_{ik}}\left(w_{ik}\right) I\left(w_i = \sum\nolimits_{k=1}^{K} w_{ik}\right).$$

Combination of both, the augmented likelihood function is given by

$$
\begin{aligned}
L_{\text{aug}} &= \prod_{i=1}^{n} \left[1 - \exp\left\{-\boldsymbol{\alpha}^{\mathrm{T}} B(c_i) \exp\left(x_i^{\mathrm{T}} \beta + h^{\mathrm{T}} B(z_i)\right)\right\}\right]^{\delta_i} \times \exp\left\{-\boldsymbol{\alpha}^{\mathrm{T}} B(c_i) \exp\left(x_i^{\mathrm{T}} \beta + h^{\mathrm{T}} B(z_i)\right)\right\}^{1-\delta_i} \\
&= \prod_{i=1}^{n} P_{w_i}\left(w_i > 0\right)^{\delta_i} P_{w_i}\left(w_i = 0\right)^{1-\delta_i} = \prod_{i=1}^{n} P_{w_i}\left(w_i\right) \delta_i^{I(w_i>0)} \left(1 - \delta_i\right)^{I(w_i=0)} \\
&= \prod_{i=1}^{n} \prod_{k=1}^{K} P_{w_{ik}}\left(w_{ik}\right) \delta_i^{I(w_i>0)} \left(1 - \delta_i\right)^{I(w_i=0)} I\left(w_i = \sum\nolimits_{k=1}^{K} w_{ik}\right).
\end{aligned}
$$

Appendix 2: Details of the EMVS algorithm

In $(u+1)$th iteration, noting that w_{ik} follows a binomial distribution given that w_i, and w_i is truncated and follows a Poisson distribution, the expected values can be expressed as:

$$
E\left(w_{ik} | \mathcal{D}, \boldsymbol{\theta}^{(u)}\right) = \frac{\alpha_k^{(u)} B_k(c_i)}{\sum_{k=1}^{K} \alpha_k^{(u)} B_k(c_i)} E\left(w_i | \mathcal{D}, \boldsymbol{\theta}^{(u)}\right),
$$

$$
E\left(w_i | \mathcal{D}, \boldsymbol{\theta}^{(u)}\right) = \frac{\sum_{k=1}^{K} \alpha_k^{(u)} B_k(c_i) \exp\left\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(u)} + \left(\boldsymbol{h}^{(u)}\right)^{\mathrm{T}} \boldsymbol{B}(z_i)\right\} \delta_i}{1 - \exp\left\{-\exp\left\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(u)} + \left(\boldsymbol{h}^{(u)}\right)^{\mathrm{T}} \boldsymbol{B}(z_i)\right\}\right\}}.
$$

Since γ_m follows a binomial distribution, the conditional expectations as from Ročková and George (2014) are:

$$
E\left(\gamma_m | \mathcal{D}, \boldsymbol{\theta}^{(u)}\right) = \mathrm{P}\left(\gamma_m = 1 | \cot\right) = p_m^* = \frac{r_m}{r_m + s_m},
$$

$$
E\left(\log d_m | \mathcal{D}, \boldsymbol{\theta}^{(u)}\right) = \left(1 - p_m^*\right) \log v_0 + p_m^* \log v_1^{(u)},
$$

$$
E\left(\frac{1}{d_m} \middle| \mathcal{D}, \boldsymbol{\theta}^{(u)}\right) = \frac{1 - p_m^*}{v_0} + \frac{p_m^*}{v_1^{(u)}},
$$

where $r_m = P(\beta_m | \left(\sigma^2\right)^{(u)}, v_1^{(u)}, \gamma_m = 1) \mathrm{P}(\gamma_m = 1 | \omega^{(u)})$ and $s_m = P(\beta_m | \left(\sigma^2\right)^{(u)}, v_1^{(u)}, \gamma_m = 0) \mathrm{P}\left(\gamma_m = 0 | \omega^{(u)}\right)$.

The next step is to find $\boldsymbol{\theta}^{(u+1)}$ which maximizes Q_1 and Q_2. Consider the partial derivative of Q_1 and Q_2 with respect to each parameter, for $\boldsymbol{\beta}$ and $\boldsymbol{h}_j$,

$$
\frac{\partial Q_1}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{k=1}^{K} \boldsymbol{x}_i \left[E\left(w_{ik} | \mathcal{D}, \boldsymbol{\theta}^{(u)}\right) - \alpha_k^{(u)} B_k\left(c_i\right) \exp\left\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}} \boldsymbol{B}\left(z_i\right)\right\}\right] - \Sigma \boldsymbol{\beta},
$$

$$\frac{\partial Q_1}{\partial \boldsymbol{h}_j} = \sum_{i=1}^{n}\sum_{k=1}^{K} \boldsymbol{B}_j(z_{ij})\left[E\left(w_{ik}|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right) - \alpha_k^{(u)} B_k\left(c_i\right)\exp\left\{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}}\boldsymbol{B}\left(z_i\right)\right\}\right] - \tau_j P\boldsymbol{h}_j,$$

where $\Sigma = \frac{1}{\sigma^2}\mathrm{diag}\left(E\left(1/d_1\middle|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right), \dots, E\left(1/d_p\middle|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)\right)$. Setting the partial derivative to 0, the new maximized $\boldsymbol{\beta}^{(u+1)}$ can be obtained from the equation. Then, replacing $\boldsymbol{\beta}$ in the second equation with $\boldsymbol{\beta}^{(u+1)}$, the solution is $\boldsymbol{h}_j^{(u+1)}$. Roots of the above equations can be solved for by using the standard root finding routine. The support of $\alpha_k^{(u+1)}$ is $\left[\alpha_{k-1}^{(u+1)}, \alpha_{k+1}^{(u)}\right]$. Denote

$$\alpha_k^{(u+1)*} = \frac{-D_k \pm \sqrt{D_k^2 - 4\lambda P_{kk}\sum_{i=1}^{n} E\left(w_{ik}|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)}}{2\lambda P_{kk}},$$

where P_{kk} is the (k,k)th element of matrix $\boldsymbol{P}$, $D_k = \boldsymbol{P}_{k,-k}^{\mathrm{T}}\boldsymbol{\alpha}_{-k}^{(u+1)} + \sum_{i=1}^{n} B_k\left(c_i\right)\exp\left[\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(u+1)} + \left(\boldsymbol{h}^{(u+1)}\right)^{\mathrm{T}}\boldsymbol{B}\left(\boldsymbol{z}_i\right)\right]$ with $\boldsymbol{P}_{k,-k} = \left(P_{k1}, \dots, P_{k,(k-1)},\right.$
$\left.P_{k,(k+1)}, \dots, P_{kK}\right)^{\mathrm{T}}$ and $\boldsymbol{\alpha}_{-k}^{(u+1)} = \left(\alpha_1^{(u+1)}, \dots, \alpha_{k-1}^{(u+1)}, \alpha_k^{(u)}, \dots, \alpha_K^{(u)}\right)^{\mathrm{T}}$. $\alpha_k^{(u+1)} = \alpha_k^{(u+1)*}$ if $\alpha_k^{(u+1)*}$ belongs to the support. Otherwise, we maximize the following formula

$$\sum_{i=1}^{n}\left\{E\left(w_{ik}|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)\log\alpha_k - \alpha_k B_k\left(c_i\right)\exp\left[\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}}\boldsymbol{B}\left(z_i\right)\right]\right\} - \frac{\lambda\boldsymbol{\alpha}^{\mathrm{T}}P\boldsymbol{\alpha}}{2}$$

with the support $\left[\alpha_{k-1}^{(u+1)}, \alpha_{k+1}^{(u)}\right]$, which can be solved by R function *maxLik* with linear constraints.

As the new maximized expressions of the remaining parameters have closed-form expressions, the parameters are directly updated by,

$$\left(\sigma^2\right)^{(u+1)} = \frac{b_2 + \frac{1}{2}\sum_{m=1}^{p} E\left(\frac{1}{d_m}\middle|\mathcal{D}, \boldsymbol{\theta}^{(u)}\right)\left(\beta_m^{(u+1)}\right)^2}{\frac{p}{2} + a_2 + 1},$$

$$v_1^{(u+1)} = \frac{A - B \pm \sqrt{(B-A)^2 + 4A\left(\frac{5}{4} + B\right)}}{2\left(\frac{5}{4} + B\right)},$$

$$\lambda^{(u+1)} = \frac{K + a_3 - 1}{\sum_{k=1}^{K}\alpha_k^{(u+1)} + b_3},$$

$$\tau_j^{(u+1)} = \frac{\frac{L}{2} + a_4 + 1}{\frac{\left(\boldsymbol{h}_j^{(u+1)}\right)^{\mathrm{T}} \boldsymbol{P}\boldsymbol{h}_j^{(u+1)}}{2} + b_4},$$

where $\mathrm{A} = \sum_{m=1}^{p} p_m^* \left(\beta_m^{(u+1)}\right)^2 / \left(2\left(\sigma^2\right)^{(u+1)}\right)$ and $\mathrm{B} = \sum_{m=1}^{p} p_m^*/2$.

In simulation studies and real data analysis, we recommended to set the initial values of $\boldsymbol{\beta}$ as $(1, 0, \ldots, 0)^{\mathrm{T}}$. The algorithm is robust to the initial values of other parameters.

Appendix 3: LASSO for variable selection of the current status data based on additive Cox model

Different from the EMVS, we use monotone I-splines (Ramsay, 1988) to approximate $\Lambda_0(\cdot)$, that is,

$$\Lambda_0(\cdot) = \sum_{k=1}^{K} \alpha_k I_k(\cdot),$$

where $\left\{I_k(\cdot)\right\}_{k=1}^{K}$ are integrated spline basis functions, each of which is nondecreasing from 0 to 1, and the spline coefficients $\left\{\alpha_k\right\}_{k=1}^{K}$ are taken to be non-negative to guarantee the monotonicity. We also use cubic splines to provide enough smoothness. The other model specification keeps same.

Based on Eq. (8), variables w_i and w_{ik} are treated as missing data and $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{h}\}$ are the parameters to be estimated. In $(u+1)$th iteration, the expected values of w_i and w_{ik} in E-step take the same value of the proposed method. In M-step, first we update α as

$$\boldsymbol{\alpha}_k^{(u+1)} = \frac{\sum_{i=1}^{n} E\left(w_{ik} | \mathcal{D}, \boldsymbol{\beta}^{(u)}, \boldsymbol{h}^{(u)}\right)}{\sum_{i=1}^{n} I_k(c_i) \exp\left\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}^{(u+1)} + \left(\boldsymbol{h}^{(u+1)}\right)^{\mathrm{T}} \boldsymbol{B}(z_i)\right\}}.$$

Denote $\boldsymbol{\Psi} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{h}^{\mathrm{T}})^{\mathrm{T}}$, then wen update β and h by maximizing

$$Q\left(\boldsymbol{\beta}, \boldsymbol{h} | \mathcal{D}, \boldsymbol{\beta}^{(u)}, \boldsymbol{h}^{(u)}, \boldsymbol{\alpha}^{(u+1)}\right) - n\vartheta\left(\sum_{m=1}^{p} \left|\beta_m\right| + \sum_{j=1}^{q} \sum_{l=1}^{L} \left|h_{jl}\right|\right),$$

where

$$Q\left(\boldsymbol{\beta}, \boldsymbol{h} | \mathcal{D}, \boldsymbol{\beta}^{(u)}, \boldsymbol{h}^{(u)}, \boldsymbol{\alpha}^{(u+1)}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} E\left(w_{ik} | \mathcal{D}, \boldsymbol{\beta}^{(u)}, \boldsymbol{h}^{(u)}\right) \left[\log \alpha_k^{(u+1)} + \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}} \boldsymbol{B}(z_i)\right]$$

$$-\alpha_k^{(u+1)} I_k\left(c_i\right) \exp\left\{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{h}^{\mathrm{T}} \boldsymbol{B}(z_i)\right\}$$

and ϑ is a tuning parameter. Define $\nabla Q\left(\boldsymbol{\Psi}^{(u)}\right) = -\partial Q / \partial \boldsymbol{\Psi}|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(u)}}$ and $\nabla^2 Q\left(\boldsymbol{\Psi}^{(u)}\right) = -\partial^2 Q / \partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^{\mathrm{T}}|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(u)}}$. Through a second-order Taylor expansion around $\boldsymbol{\Psi}^{(u+1)}$, $-Q\left(\boldsymbol{\beta}, \boldsymbol{h} | \mathcal{D}, \boldsymbol{\beta}^{(u)}, \boldsymbol{h}^{(u)}, \boldsymbol{\alpha}^{(u+1)}\right)$ can be written as $1/2(\boldsymbol{Y} - \boldsymbol{R}\boldsymbol{\Psi})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{R}\boldsymbol{\Psi})$, where $\boldsymbol{R}$ is from Cholesky decomposition

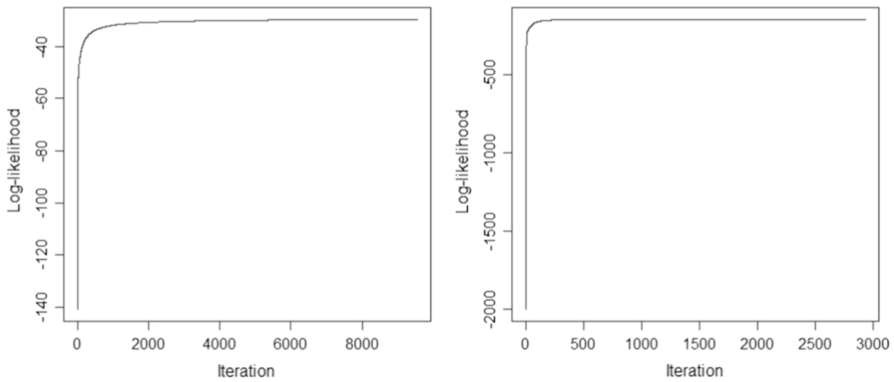

Fig. 5 The increase of log-likelihood values over iterations for scenario (i) when $n = 200$ (left) and $n = 500$ (right)

of $\nabla^2 Q(\boldsymbol{\Psi}^{(u)})$ satisfying $\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \nabla^2 Q(\boldsymbol{\Psi}^{(u)})$ and pseudo response $\boldsymbol{Y} = (\boldsymbol{R}^{\mathrm{T}})^{-1}\{\nabla^2 Q(\boldsymbol{\Psi}^{(u)})\boldsymbol{\beta} - \nabla Q(\boldsymbol{\Psi}^{(u)})\}$ (Zhang & Lu, 2007). Thus, we need to minimize

$$-\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{R\Psi})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{R\Psi}) + n\vartheta \sum_{b=1}^{p+q\times L} \Psi_b.$$

To obtain the LASSO regressor $\boldsymbol{\Psi}^{(u+1)}$, we use *glmnet* package in R. The EM algorithm stops if the maximum absolute difference of the parameters between two successive iterations is smaller than 10^{-5}.

Appendix 4: The increase of log-likelihood values

See Fig. 5.

Author contribution All authors contributed to the study conception and design. Material preparation, data analysis was performed by Di Cui and Clarence Tee. The first draft of the manuscript was written by Di Cui and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability The datasets generated during and analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Bremhorst, V., & Lambert, P. (2016). Flexible estimation in cure survival models using Bayesian P-splines. *Computational Statistics & Data Analysis, 93*, 270–284.

Brezger, A., & Steiner, W. J. (2008). Monotonic regression based on Bayesian p-splines: An application to estimating price response functions from store-level scanner data. *Journal of Business & Economic Statistics, 26*(1), 90–104.

Cai, B., Lin, X., & Wang, L. (2011). Bayesian proportional hazards model for current status data with monotone splines. *Computational Statistics & Data Analysis, 55*(9), 2644–2651.

Çetinyürek, Y. A., & Lambert, P. (2011). Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines. *Statistics in Medicine, 30*(1), 75–90.

Chan, S., Wang, X., Jazić, I., Peskoe, S., Zheng, Y., & Cai, T. (2021). Developing and evaluating risk prediction models with panel current status data. *Biometrics, 77*(2), 599–609.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (methodological), 34*(2), 187–202.

Cui, W., & George, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference, 138*(4), 888–900.

Dai, W., & Jin, B. (2022). Bayesian inference of clustering and multiple Gaussian graphical models selection. *Journal of the Korean Statistical Society, 51*(2), 422–440.

Díaz-Coto, S., Martínez-Camblor, P., & Corral-Blanco, N. O. (2020). Cumulative/dynamic ROC curve estimation under interval censorship. *Journal of Statistical Computation and Simulation, 90*(9), 1570–1590.

Du, M., Zhao, H., & Sun, J. (2021). A unified approach to variable selection for Cox's proportional hazards model with interval-censored failure time data. *Statistical Methods in Medical Research, 30*(8), 1833–1849.

Du, M., Zhao, X., & Sun, J. (2022). Variable selection for case-cohort studies with informatively interval-censored outcomes. *Computational Statistics & Data Analysis, 172*, 107484.

Eliers, P., & Marx, B. (1996). Flexible Smoothing using B-splines and Penalized Likelihood (with comments and rejoinder). *Statistical Science, 11*(2), 89–121.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.

Gamage, P. W. W., Chaudari, M., McMahan, C. S., Kim, E. H., & Kosorok, M. R. (2020). An extended proportional hazards model for interval-censored data subject to instantaneous failures. *Lifetime Data Analysis, 26*(1), 158–182.

Gao, F., & Chan, K. C. G. (2019). Semiparametric regression analysis of length-biased interval-censored data. *Biometrics, 75*(1), 121–132.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association, 88*(423), 881–889.

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*(2), 339–373.

Held, L., Gravestock, I., & Sabanés Bové, D. (2016). Objective Bayesian model selection for Cox regression. *Statistics in Medicine, 35*(29), 5376–5390.

Höllig, K., & Hörner, J. (2013). *Approximation and modeling with B-splines*. Society for Industrial and Applied Mathematics.

Hosmer, D. W., & Lemeshow, S. (2002). *Applied survival analysis: Regression modelling of time to event data*. Wiley.

Koley, T., & Dewanji, A. (2022). Current status data with two competing risks and missing failure types: A parametric approach. *Journal of Applied Statistics, 49*(7), 1769–1783.

Koslovsky, M. D., Swartz, M. D., Chan, W., Leon-Novelo, L., Wilkinson, A. V., Kendzor, D. E., & Businelle, M. S. (2018). Bayesian variable selection for multistate Markov models with interval-censored data in an ecological momentary assessment study of smoking cessation. *Biometrics, 74*(2), 636–644.

Koslovsky, M. D., Swartz, M. D., Leon-Novelo, L., Chan, W., & Wilkinson, A. V. (2018). Using the EM algorithm for Bayesian variable selection in logistic regression models with related covariates. *Journal of Statistical Computation and Simulation, 88*(3), 575–596.

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

Li, C., Pak, D., & Todem, D. (2020). Adaptive lasso for the Cox regression with interval censored and possibly left truncated data. *Statistical Methods in Medical Research, 29*(4), 1243–1255.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*(481), 410–423.

Lu, M., & McMahan, C. S. (2018). A partially linear proportional hazards model for current status data. *Biometrics, 74*(4), 1240–1249.

Maruyama, Y., & George, E. I. (2011). Fully Bayes factors with a generalized g-prior. *The Annals of Statistics, 39*(5), 2740–2765.

McMahan, C. S., Wang, L., & Tebbs, J. M. (2013). Regression analysis for current status data using the EM algorithm. *Statistics in Medicine, 32*(25), 4452–4466.

Mu, J., Liu, Q., Kuo, L., & Hu, G. (2021). Bayesian variable selection for the Cox regression model with spatially varying coefficients with applications to Louisiana respiratory cancer data. *Biometrical Journal, 63*(8), 1607–1622.

Narisetty, N. N., & He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics, 42*(2), 789–817.

Nikooienejad, A., Wang, W., & Johnson, V. E. (2020). Bayesian variable selection for survival data using inverse moment priors. *The Annals of Applied Statistics, 14*(2), 809.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, *3*(4), 425–441.

Ročková, V., & George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association, 109*(506), 828–846.

Schumaker, L. (2007). *Spline functions: Basic theory*. Cambridge University Press.

Scolas, S., El Ghouch, A., Legrand, C., & Oulhaj, A. (2016). Variable selection in a flexible parametric mixture cure model with interval-censored data. *Statistics in Medicine, 35*(7), 1210–1225.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (methodological), 58*(1), 267–288.

Wang, L., McMahan, C. S., Hudgens, M. G., & Qureshi, Z. P. (2016). A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data. *Biometrics, 72*(1), 222–231.

White, H. D., & Chew, D. P. (2008). Acute myocardial infarction. *Lancet, 372*(9638), 570–584.

Wu, Y., & Cook, R. J. (2015). Penalized regression for interval-censored times of disease progression: Selection of HLA markers in psoriatic arthritis. *Biometrics, 71*(3), 782–791.

Zeng, D., Mao, L., & Lin, D. Y. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika, 103*(2), 253–271.

Zhang, H. H., & Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika, 94*(3), 691–703.

Zhang, Z., & Sun, J. (2010). Interval censoring. *Statistical Methods in Medical Research, 19*(1), 53–70.

Zhao, H., Wu, Q., Li, G., & Sun, J. (2019). Simultaneous estimation and variable selection for interval-censored data with broken adaptive ridge regression. *Journal of the American Statistical Association*, *115*(529), 204–216.

Zhao, K., & Lian, H. (2016). The expectation–maximization approach for Bayesian quantile regression. *Computational Statistics & Data Analysis, 96*, 1–11.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.