Lifetime Data Analysis (2022) 28:282–318
https://doi.org/10.1007/s10985-022-09549-5

# Bayesian penalized Buckley-James method for high dimensional bivariate censored regression models

Wenjing Yin[1] · Sihai Dave Zhao[1] · Feng Liang[1]

Received: 25 June 2020 / Accepted: 22 January 2022 / Published online: 3 March 2022

## Abstract

For high dimensional gene expression data, one important goal is to identify a small number of genes that are associated with progression of the disease or survival of the patients. In this paper, we consider the problem of variable selection for multivariate survival data. We propose an estimation procedure for high dimensional accelerated failure time (AFT) models with bivariate censored data. The method extends the Buckley-James method by minimizing a penalized $L_2$ loss function with a penalty function induced from a bivariate spike-and-slab prior specification. In the proposed algorithm, censored observations are imputed using the Kaplan-Meier estimator, which avoids a parametric assumption on the error terms. Our empirical studies demonstrate that the proposed method provides better performance compared to the alternative procedures designed for univariate survival data regardless of whether the true events are correlated or not, and conceptualizes a formal way of handling bivariate survival data for AFT models. Findings from the analysis of a myeloma clinical trial using the proposed method are also presented.

**Keywords** Buckley-James estimator · Bayesian penalization · Multivariate survival data · Variable selection

## 1 Introduction

Emerging developments in sequencing technology have made it easier to collect massive amount of gene expression data that can be used for understanding cancer

✉ Feng Liang
liangf@illinois.edu

Wenjing Yin
wyin5@illinois.edu

Sihai Dave Zhao
sdzhao@illinois.edu

[1] Department of Statistics, University of Illinois, Urbana-Champaign, Champaign, IL, USA

genomics. One popular model that associates survival times with covariates is the proportional hazards model, or often known as the Cox model (Cox 1972). It is well-known that the Cox model may not fit the survival data well if the proportional hazards assumption is not satisfied. As a useful alternative to the Cox model, the accelerated failure time (AFT) model has recently received great attention (Miller 1976; Buckley and James 1979; Koul et al. 1981; Schneider and Weissfeld 1986; Wei et al. 1990; Tsiatis 1990; Wei 1992; Stute and Wang 1993; Jin et al. 2003, 2006a; Huang et al. 2007, 2009; Kalbfleisch and Prentice 2011; Wang and Song 2011; Khan and Shaw 2016; Huang et al. 2020).

In biomedical and epidemiologic studies, it is common that multiple events could have happened during the study such that more than one event time is collected from the same group of patients. For such multivariate survival data, since those events are about the same patient, it is nature to utilize all available data in the analysis even though not all events are of major interest. One challenge in associating multiple event times with covariates through AFT models is that the true multivariate survival distribution is very difficult to estimate. Meanwhile, it is also very difficult to model the correlation structure between the events. To address these concerns, both parametric and nonparametric methods have been considered for modeling either the marginal survival distribution or the joint survival distribution. Parametric methods such as bivariate Weibull distribution and bivariate normal distribution have been suggested for modeling joint survival distribution for bivariate survival data (Hanagal 2006; Yi and He 2006; Lu 2007). Nonparametric approaches such as extensions of Buckley-James estimator (Hornsteiner and Hamerle 1996; Pan and Kooperberg 1999; Chiou et al. 2014; Pan and Louis 2000), rank-based estimating equations (Jin et al. 2006b; Li and Yin 2009; Wang and Fu 2011), and other nonparametric modeling methods (Lu 2007; Yin and Cai 2005; Visser 1996; He and Lawless 2005; Huang 2002; Chang 2004) have been proposed.

The problem of variable selection often arises in analyzing gene expression data, which is usually high dimensional with the number of features $p$ being greater than the sample size $n$. None of the methods reviewed above for multivariate AFT models can handle high-dimensional data or address the problem of variable selection. For univariate AFT models, some variable selection methods have been developed based on Buckley-James method or rank-based approaches (Johnson et al. 2009; Wang et al. 2008; Li et al. 2014; Cai et al. 2009; Xu et al. 2010) while others have been proposed based on the Stute's weighted least squares (Huang et al. 2006; Huang and Ma 2010; Wang and Song 2011; Hu and Chai 2013; Khan and Shaw 2016; Khan et al. 2019; Khan and Shaw 2019; Huang et al. 2020). From Bayesian aspects, Sha et al. (2006) and Lee et al. (2017) developed variable selection methods for log-normal AFT models, while Lee and Mallick (2004) and Duan et al. (2018) studied Weibull AFT models. Nonparametric Bayesian AFT models have also been explored (Ahmed et al. 2012; Konrath et al. 2015).

In this paper, motivated by the lack of such variable selection approaches, we consider high-dimensional bivariate survival data with the same design matrix and develop a Bayesian variable selection framework assuming that the two events are independent.

The relationship between the two events can be studied through the prior specification of regression coefficients. The remaining of this paper is organized as follows. In Sect. 2, we introduce the imputation of the censored observations using Kaplan-Meier estimator under univariate AFT models which is equivalent to the Buckley-James estimator using an iterative procedure. In Sect. 3, we elaborate the variable selection approach inspired by the Bayesian framework for bivariate survival data and discuss the details of the computation process along with some remarks. Sect. 4 assesses empirical results of the proposed method using three simulation studies, as well as a detailed analysis of a myeloma clinical trial in Sect. 4.3. Finally Sect. 5 concludes the paper with a brief discussion.

## 2 Background

### 2.1 Bivariate AFT model

Let $\{(\boldsymbol{T}_i, \boldsymbol{C}_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$ be independent and identically distributed random vectors where $\boldsymbol{T}_i = (T_{i1}, T_{i2})$ denotes the log-transformed bivariate survival time, $\boldsymbol{C}_i = (C_{i1}, C_{i2})$ denotes the log-transformed bivariate censoring time, and $\boldsymbol{X}_i$'s are the $p$-dimensional covariate vectors. For ease of exposition, we assume that the design matrices for two columns of survival times are the same. Due to the right censoring of survival time $\boldsymbol{T}_i$, the observed data consists of the triplets $\{(\boldsymbol{Y}_i, \boldsymbol{\delta}_i, \boldsymbol{X}_i), i = 1, \ldots, n\}$, where $\boldsymbol{Y}_i = (\min\{T_{i1}, C_{i1}\}, \min\{T_{i2}, C_{i2}\})$ is the bivariate censored survival time and $\boldsymbol{\delta}_i = (\boldsymbol{I}\{T_{i1} \le C_{i1}\}, \boldsymbol{I}\{T_{i2} \le C_{i2}\})$ is the bivariate censoring indicator such that a zero value indicates censoring for the corresponding survival time for $i$-th observation. Throughout we assume that censoring time $C_{ik}$ is independent of survival time $T_{ik}$ conditioning on covariates $\boldsymbol{X}_i$ for $k = 1, 2$ and $i = 1, \ldots, n$.

In this paper, we focus on the AFT model for the analysis of bivariate time-to-event data, which has the general form:

$$\boldsymbol{T}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \ i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right)$ is a $p$-by-2 matrix of regression coefficients, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2})$ denotes independent and identically distributed bivariate random error term. We further assume that $\varepsilon_{i1}$ and $\varepsilon_{i2}$ follow unspecified distribution functions $F_1$ and $F_2$, which have zero means and finite variances $\sigma_1^2$ and $\sigma_2^2$, respectively. We denote the first and the second survival time as $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$.

Popular parametric choices for $F_1$ and $F_2$ include Gaussian distribution, log-normal distribution, or Weibull distribution. For example, Hanagal (2006) considered a bivariate Weibull regression model and derived a maximum likelihood estimator for regression coefficients while Yi and He (2006) discussed estimation method for bivariate normal AFT models. For nonparametric approaches, Lu (2007) proposed a $\chi^2$-type test statistic for testing independence structure between the bivariate survival data with unshared covariates by assuming a joint unspecified distribution and utilizing a Gehan weighted log-rank estimating function.

### 2.2 Buckley-James estimator for univariate AFT models

The main challenge with the AFT model is that the actual survival time $\boldsymbol{T}_i$ may not be observable for some cases due to censoring. Next, we review prior work, especially the Buckley-James estimator for regression coefficient $\boldsymbol{\beta}$. With slight abuse of notation, in this particular subsection, we consider a marginal univariate AFT model and let $\boldsymbol{T}, \boldsymbol{Y}, \boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ be vectors of length $n$ and $\boldsymbol{\beta}$ be a vector of length $p$.

When there is no censoring and the survival times $T_i$'s are fully observed, the most natural estimator for $\boldsymbol{\beta}$ is the least squares estimator $\tilde{\boldsymbol{\beta}}$ given by

$$\tilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{T}$$

which is obtained from solving the score equation

$$\sum_{i=1}^{n} \boldsymbol{X}_i' \left(T_i - \boldsymbol{X}_i\boldsymbol{\beta}\right) = 0.$$

When $\boldsymbol{T}$ is censored, the least-square principle has been extended to accommodate censoring by many researchers such as Miller (1976); Buckley and James (1979); Koul et al. (1981) and many others. One popular estimator is the Buckley-James estimator (Buckley and James 1979) which imputes the censored observations by their conditional expectations given corresponding censoring times and covariates as follows:

$$\begin{aligned}
\hat{T}_i &= \boldsymbol{E}\left[T_i \mid \delta_i = 0, \boldsymbol{X}_i\right] \\
&= \boldsymbol{E}\left[T_i \mid \left(T_i > Y_i\right), \boldsymbol{X}_i\right] \\
&= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{E}\left[\varepsilon_i \mid \left(\varepsilon_i > Y_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)\right] \\
&= \boldsymbol{X}_i\boldsymbol{\beta} + \int_{Y_i - \boldsymbol{X}_i\boldsymbol{\beta}}^{\infty} \frac{t dF\left(t\right)}{1 - F\left(Y_i - \boldsymbol{X}_i\boldsymbol{\beta}\right)},
\end{aligned} \tag{2}$$

where $F\left(\cdot\right)$, the distribution function of error terms $\varepsilon_i = T_i - \mathbf{X}_i\boldsymbol{\beta}$, is estimated by its Kaplan-Meier estimator $\hat{F}_{\boldsymbol{\varepsilon},\boldsymbol{\beta}}$ defined on $\varepsilon_i$'s. Then the least squares estimator with censored survival times can be computed as

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\hat{\boldsymbol{T}}, \tag{3}$$

where $\hat{T}_i = Y_i$, if $\delta_i = 1$, and $\hat{T}_i$ is defined in Equation (2) if $\delta_i = 0$.

Since $\hat{T}_i$ contains $\boldsymbol{\beta}$, estimating $\boldsymbol{\beta}$ requires an iterative procedure. In the spirit of the EM algorithm, Schneider and Weissfeld (1986) suggested the following computation procedure: at the $t$-th iteration, given the current estimate $\hat{\boldsymbol{\beta}}^{(t)}$, operate the E-step by computing

$$\hat{T}_i^{(t)} = \delta_i Y_i + \left(1 - \delta_i\right) \boldsymbol{E}_{\hat{F}^{(t)}} \left[T_i \mid \left(T_i > Y_i\right), \boldsymbol{\beta}^{(t)}\right], \tag{4}$$

where $\hat{F}^{(t)}$ denotes the Kaplan-Meier estimator $\hat{F}_{\boldsymbol{\varepsilon},\hat{\boldsymbol{\beta}}^{(t)}}(\cdot)$ based on errors at the $t$-th iteration $\varepsilon_{i,\hat{\boldsymbol{\beta}}^{(t)}} = Y_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(t)}$:

$$\hat{F}_{\boldsymbol{\varepsilon},\hat{\boldsymbol{\beta}}^{(t)}}(s) = 1 - \prod_{i:\varepsilon_{i,\hat{\boldsymbol{\beta}}^{(t)}}<s} \left[1 - \frac{\delta_i}{\sum_{j=1}^{n} \boldsymbol{I}\left\{\varepsilon_{j,\hat{\boldsymbol{\beta}}^{(t)}} \geq \varepsilon_{i,\hat{\boldsymbol{\beta}}^{(t)}}\right\}}\right];$$

then operates the M-step by updating $\hat{\boldsymbol{\beta}}^{(t+1)}$ via Equation (3) using $\hat{T}_i^{(t)}$'s.

Let $\varepsilon_{(i)}$ be the order statistic of $\varepsilon_{i,\hat{\boldsymbol{\beta}}^{(t)}}$'s and $\delta_{(i)}$ be the corresponding ordered vector for the censoring indicator $\delta_i$'s. Then $\hat{F}_{\boldsymbol{\varepsilon},\hat{\boldsymbol{\beta}}^{(t)}}(s)$ can be written as a step-wise function in the following form

$$\boldsymbol{P}\left(\varepsilon_{i,\hat{\boldsymbol{\beta}}^{(t)}} \leq s\right) = \hat{F}^{(t)}_{\boldsymbol{\varepsilon},\hat{\boldsymbol{\beta}}}(s) = \sum_{i=1}^{n} w_i \boldsymbol{I}\left\{\varepsilon_{(i)} \leq s\right\},$$

where $w_i$'s are weights or jumps for $\boldsymbol{T}$ which can be computed as

$$w_1 = \frac{\delta_{(1)}}{n}, w_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}}.$$

## 3 Bayesian variable selection for bivariate AFT models

In gene expression data, it is common that the dimension $p$ is large and potentially larger than sample size $n$. Many approaches have been proposed under univariate AFT models for both low and high dimensional problems under the conditional independence assumption. Extending variable selection to bivariate survival data is difficult due to the unknown dependence structure between $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$. Parametric methods for jointly modeling two survival times have been proposed (Hanagal 2006; Yi and He 2006). However, when the parametric distribution is misspecified for the survival data, the results are usually unsatisfactory. Assuming an unspecified joint distribution $F(\boldsymbol{T}_{\cdot 1}, \boldsymbol{T}_{\cdot 2})$ is one option for nonparametric methods but finding the nonparametric estimator of the joint distribution function is very difficult.

To address the above issues, we develop a new methodology inspired by the Bayesian framework and present a computational procedure that borrows ideas from the EM algorithm. The proposed method extends the Buckley-James approach by using penalized least squares with a penalty function $pen(\boldsymbol{\beta})$ induced from the so-called spike-and-slab prior (Ročková and George 2014). To be specific, we consider the minimization problem of the following penalized $L_2$ loss function

$$\min_{\boldsymbol{\beta}} \sum_{k=1}^{2} \sum_{i=1}^{n} \left(\hat{T}_{ik} - \boldsymbol{X}_i \boldsymbol{\beta}_k\right)^2 + pen\left(\boldsymbol{\beta}\right),$$

where $\hat{T}_{ik}$'s are imputed censored observations for bivariate survival data computed based on Equation (4) for $k = 1, 2$ following Buckley-James approach described in Sect. 2.2. Here we assume that $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$ are marginally independent which allows us to find the nonparametric estimators of $F_1$ and $F_2$ marginally. To utilize shared information across $\boldsymbol{\beta}$, we design a new penalty function from the Bayesian perspective, i.e., the penalty function $pen(\boldsymbol{\beta})$ is induced by the negative logarithm of a prior specification $\pi(\boldsymbol{\beta})$. The connection between penalty function and negative logarithm of a prior specification has been mentioned in Park and Casella (2008), and more deeply discussed by Van Erp et al. (2019). Then the loss function $l(\boldsymbol{\beta})$ can be defined as

$$l(\boldsymbol{\beta}) = \sum_{k=1}^{2} \sum_{i=1}^{n} \left( \hat{T}_{ik} - \boldsymbol{X}_i \boldsymbol{\beta}_k \right)^2 - \log \pi(\boldsymbol{\beta}). \tag{5}$$

Minimizing Equation (5) is equivalent to maximizing $\exp\{-l(\boldsymbol{\beta})\}$, namely,

$$\begin{aligned} L\left(\hat{\boldsymbol{T}} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right) &= \exp\{-l(\boldsymbol{\beta})\} \\ &\propto \prod_{k=1}^{2} \prod_{i=1}^{n} \exp\left\{-(\hat{T}_{ik} - \boldsymbol{X}_i \boldsymbol{\beta}_k)^2\right\} \times \pi(\boldsymbol{\beta}). \end{aligned} \tag{6}$$

In other words, the minimization problem is naturally transformed into a maximization problem involving a product of a working Gaussian likelihood and a prior distribution of $\boldsymbol{\beta}$.

Next we describe the details of our bivariate prior specification $\pi(\boldsymbol{\beta})$, and the algorithm for the maximization problem, and end Sect. 3 with some remarks.

## 3.1 Bivariate spike and slab priors

### 3.1.1 Univariate spike and slab

When there is only one response variable, a useful prior that applies to the Bayesian variable selection framework for high-dimensional problems is the spike-and-slab prior. The key idea of spike-and-slab prior is to zero out non-relevant coefficients by making the posterior mean values of those coefficients small. The hyper-variances control the magnitudes of posterior mean values for relevant and non-relevant coefficients which correspond to the spike and the slab.

We define a binary latent variable $\gamma_j$ to indicate whether the $j^{th}$ covariate is active. A binary index vector $\boldsymbol{\gamma} = \left(\gamma_1, \ldots, \gamma_p\right)$ is provided for the model under consideration. A hierarchical prior structure for $\boldsymbol{\beta}$ starts with a prior distribution $\pi(\boldsymbol{\gamma})$ and then a prior distribution of $\boldsymbol{\beta}$ conditional on $\gamma$ denoted by $\pi\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}\right)$.

For univariate cases, popular choices for the hierarchical prior structure include the Bernoulli families for $\gamma_j$ and the "spike and slab" prior (Mitchell and Beauchamp

1988) for $\beta_j$ such that

$$\pi\left(\beta_j \mid \gamma_j\right) = \gamma_j g_1\left(\beta_j\right) + \left(1 - \gamma_j\right) g_0\left(\beta_j\right),$$

where $g_1(\cdot)$ is usually the density function of a symmetric distribution such as a normal distribution with mean 0, and $g_0(\cdot)$ denotes the distribution function of a point mass at 0. Another popular choice is to consider a combination of two normal distributions such that

$$\pi\left(\beta_j \mid \gamma_j\right) = \left(1 - \gamma_j\right) \mathsf{N}\left(0, v_{0_{\gamma_j}}\right) + \gamma_j \mathsf{N}\left(0, v_{1_{\gamma_j}}\right),$$

where $0 < v_0 < v_1$ (George and McCulloch 1997; Narisetty et al. 2014; Ročková and George 2014). By construction, setting $v_0$ small and $v_1$ large results in $\mathsf{N}(0, v_0)$ as concentrated and $\mathsf{N}(0, v_1)$ as diffused. The likelihood and priors induce a joint distribution over the data, parameters, and the model space. The selection of any model can be made based on the posterior distribution over $\boldsymbol{\gamma}$ given the observed data.

The selection of $j$-th variable is determined using the Barbieri-Berger median model (Barbieri and Berger 2004). That is, let $\gamma_j = 1$, if

$$\pi\left(\gamma_j = 1 \mid \boldsymbol{Y}, \beta_j, \sigma^2\right) > 0.5.$$

### 3.1.2 Bivariate spike and slab

Extending the same idea to bivariate cases, we propose a bivariate spike-and-slab prior as the following:

$$\begin{aligned} \boldsymbol{\beta}_j \mid \boldsymbol{\gamma}_j \sim\ & \gamma_{j1}\gamma_{j2} \mathsf{N}(0, v_1)\, \mathsf{N}(0, v_1) \\ & + \gamma_{j1}\left(1 - \gamma_{j2}\right) \mathsf{N}(0, v_1)\, \mathsf{N}(0, v_0) \\ & + \left(1 - \gamma_{j1}\right) \gamma_{j2} \mathsf{N}(0, v_0)\, \mathsf{N}(0, v_1) \\ & + \left(1 - \gamma_{j1}\right)\left(1 - \gamma_{j2}\right) \mathsf{N}(0, v_0)\, \mathsf{N}(0, v_0), \\ \boldsymbol{\gamma}_j \mid (\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})) \sim\ & \text{Bivariate Bernoulli}\left(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}\right), \end{aligned}$$

where $\boldsymbol{\beta}_j = \left(\beta_{j1}, \beta_{j2}\right)$ and $\boldsymbol{\gamma}_j = \left(\gamma_{j1}, \gamma_{j2}\right)$ for $j = 1, \ldots, p$.

We consider a matrix of binary indices $\boldsymbol{\gamma}$ of size $p$ by 2 such that $\boldsymbol{\gamma}_j = \left(\gamma_{j1}, \gamma_{j2}\right)$ corresponds to the binary index vector of pair $\boldsymbol{\beta}_j = \left(\beta_{j1}, \beta_{j2}\right)$. A bivariate Bernoulli distribution is assigned to $\boldsymbol{\gamma}_j$ such that all four different outcomes are modeled through hyper-parameters $\pi_{11}, \pi_{10}, \pi_{01}$, and $\pi_{00}$. By design, the prior distribution $\pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma})$ follows a bivariate Gaussian distribution where $\mathsf{N}(0, v_0)$ and $\mathsf{N}(0, v_1)$ are two independent Gaussian distributions.

Similar to univariate cases, setting $v_0$ small and $v_1$ large results in a concentrated or a diffused bivariate Gaussian distribution in both directions of a two-dimensional space. The selection of any model can be made based on the posterior distribution over $\boldsymbol{\gamma}$ given the observed data. With the bivariate Bernoulli distribution, four posterior probabilities corresponding to four outcomes of pair $\boldsymbol{\beta}_j = \left(\beta_{j1}, \beta_{j2}\right)$ can be computed. Let

$$\phi_{max} = \max_{l,m} \pi \left(\gamma_{j1} = l, \gamma_{j2} = m \mid \boldsymbol{Y}, \boldsymbol{\beta}_j, \sigma_1^2, \sigma_2^2\right).$$

The selection of $j$-th variable is determined by:

$$\begin{cases} \gamma_{j1} = 1, \gamma_{j2} = 1 & \text{if } \phi_{max} = \pi \left(\gamma_{j1} = 1, \gamma_{j2} = 1 \mid \boldsymbol{Y}, \boldsymbol{\beta}_j, \sigma_1^2, \sigma_2^2\right) \\ \gamma_{j1} = 1, \gamma_{j2} = 0 & \text{if } \phi_{max} = \pi \left(\gamma_{j1} = 1, \gamma_{j2} = 0 \mid \boldsymbol{Y}, \boldsymbol{\beta}_j, \sigma_1^2, \sigma_2^2\right) \\ \gamma_{j1} = 0, \gamma_{j2} = 1 & \text{if } \phi_{max} = \pi \left(\gamma_{j1} = 0, \gamma_{j2} = 1 \mid \boldsymbol{Y}, \boldsymbol{\beta}_j, \sigma_1^2, \sigma_2^2\right) \\ \gamma_{j1} = 0, \gamma_{j2} = 0 & \text{otherwise.} \end{cases}$$

To summarize, we consider the following priors for fully observed data, For $j = 1, ..., p$, consider the hierarchical model:

$$\begin{aligned} \hat{T}_{ik} &= \boldsymbol{X}_i \boldsymbol{\beta}_k + \varepsilon_{ik}, \\ \varepsilon_{ik} &\overset{\text{i.i.d}}{\sim} F_k \text{ with } \boldsymbol{E}\left(\varepsilon_{ik}\right) = 0, \boldsymbol{VAR}\left(\varepsilon_{ik}\right) = \sigma_k^2, \\ \boldsymbol{\beta}_j \mid \boldsymbol{\gamma}_j &\overset{\text{i.i.d}}{\sim} \gamma_{j1}\gamma_{j2}\mathsf{N}\left(0, v_1\right)\mathsf{N}\left(0, v_1\right) \\ &\quad + \gamma_{j1}\left(1 - \gamma_{j2}\right)\mathsf{N}\left(0, v_1\right)\mathsf{N}\left(0, v_0\right) \\ &\quad + \left(1 - \gamma_{j1}\right)\gamma_{j2}\mathsf{N}\left(0, v_0\right)\mathsf{N}\left(0, v_1\right) \\ &\quad + \left(1 - \gamma_{j1}\right)\left(1 - \gamma_{j2}\right)\mathsf{N}\left(0, v_0\right)\mathsf{N}\left(0, v_0\right), \\ \boldsymbol{\gamma} \mid \left(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}\right) &\sim \text{Bivariate Bernoulli}\left(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}\right), \\ \left(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}\right) &\sim \text{Dirichlet}(\alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{00}), \\ \sigma_k^2 &\sim \mathsf{IG}\left(\lambda_{0,k}/2, \lambda_{0,k}\sigma_{0,k}^2/2\right), \end{aligned} \tag{7}$$

where $\mathsf{IG}\left(\lambda_{0,k}/2, \lambda_0\sigma_{0,k}^2/2\right)$ stands for an inverse gamma distribution with parameters $\lambda_{0,k}$ and $\sigma_{0,k}^2$. In practice, we recommend to consider $v_1 = 1, \lambda_{0,1} = \lambda_{0,2} = 1$, and $\sigma_{0,1}^2 = \sigma_{0,2}^2 = 1$. For the hyper-parameters $(\alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{00})$, we also recommend choosing $\alpha_{11} = \alpha_{10} = \alpha_{01} = \alpha_{00}$.

### 3.2 Computation

Utilizing the Gaussian working likelihood induced by penalized $L_2$ loss proposed in Equation (6), we want to derive the computational process for the maximization problem. Following computations for Buckley-James estimator from Schneider and Weissfeld (1986) and borrowing ideas from the EM algorithm suggested by Ročková and George (2014) for Bayesian variable selection, we propose a "full" working likelihood involving censored observations. By performing an EM-like algorithm, we are able to handle imputations of censored observations and to compute expectations of latent variables in the Bayesian framework while maximizing the working likelihood and the posterior distribution within the same iteration. That is, in each iteration, we first perform E-step computation to impute the censored observations based on current estimate of $\boldsymbol{\beta}$ and update expectations of latent variables, then at M-step we maximize the working likelihood defined in Equation (6) with corresponding $\hat{T}_{ik}$'s and the updated values of latent variables.

Let the censored survival time $Y_{ik} = \min\{T_{ik}, C_{ik}\}$ and censoring indicator $\delta_{ik} = \boldsymbol{I}\{T_{ik} \leq C_{ik}\}$ for $k = 1, 2$ be observed for the bivariate censored survival data. We first adopt the augmentation approach proposed by Tanner and Wong (1987) and consider the augmented data $\boldsymbol{W}_k = (W_{1k}, W_{2k}, \ldots, W_{nk})'$ for each survival time such that

$$\begin{cases} W_{ik} = Y_{ik} & \text{if } \delta_{ik} = 1 \\ W_{ik} > Y_{ik} & \text{if } \delta_{ik} = 0. \end{cases}$$

We want to impute $W_{ik}$'s with $\delta_{ik} = 0$ marginally using the predictive distribution $\pi(\boldsymbol{T}_{\cdot k} \mid \boldsymbol{Y}_{\cdot k})$ without borrowing information from the other event and use the imputations to approximate the posterior distribution $\pi(\cdot \mid \boldsymbol{Y})$. For $k = 1, 2$, define $\Delta_{0,k} = \{i : \delta_{ik} = 0\}$ and $\Delta_{1,k} = \{i : \delta_{ik} = 1\}$ such that $\Delta_{0,k}$ is the index set for censored data points and $\Delta_{1,k}$ is the index set the uncensored data points for survival time $\mathbf{T}_{\cdot k}$.

Let $\boldsymbol{\pi} = (\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})'$ and $\boldsymbol{\alpha} = (\alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{00})'$. The "full" working likelihood involving censored observations corresponding to the hierarchical model defined in Equation (7) can be expressed as

$$\begin{aligned} L\left(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\pi}, \sigma_1^2, \sigma_2^2 \mid \boldsymbol{X}\right) = & \prod_{k=1}^{2}\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{1}{2\sigma_k^2}\left(W_{ik} - \mathbf{X}_i\boldsymbol{\beta}_k\right)^2\right\} \\ & \times \prod_{j=1}^{p} \pi\left(\boldsymbol{\beta}_j | \boldsymbol{\gamma}_j\right) \prod_{j=1}^{p} q_1\left(\boldsymbol{\gamma}_j \mid \boldsymbol{\pi}\right) q_2\left(\boldsymbol{\pi} \mid \boldsymbol{\alpha}\right) \\ & \times q_3\left(\sigma_1^2 \mid \lambda_{0,1}, \sigma_{0,1}^2\right) q_4\left(\sigma_2^2 \mid \lambda_{0,2}, \sigma_{0,2}^2\right), \end{aligned}$$

where $\pi(\cdot), q_1(\cdot), q_2(\cdot), q_3(\cdot)$and $q_4(\cdot)$ are density functions of the priors assigned on $\boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, \boldsymbol{\pi}, \sigma_1^2$, and $\sigma_2^2$ respectively.

By the design of the augmented data $\boldsymbol{W}_k$ for $k = 1, 2$, we need to impute $\{W_{ik} : i \in \Delta_{0.k}\}$ and $\left\{W_{ik}^2 : i \in \Delta_{0,k}\right\}$ and also compute expectations for latent variables $\left\{\boldsymbol{\gamma}_j : j = 1, \ldots, p\right\}$.

### E-step

Suppose after $t$-th iteration, $\boldsymbol{\beta}^{(t)} = \left(\boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)}\right)$, $\boldsymbol{\pi}^{(t)}$, $\sigma_1^{2(t)}$ and $\sigma_2^{2(t)}$ are obtained as the current estimates of $\boldsymbol{\beta}, \boldsymbol{\pi}, \sigma_1^2$, and $\sigma_2^2$ respectively. We want to compute expected values of $\left\{W_{ik} : i \in \Delta_{0,k}\right\}$, $\left\{W_{ik}^2 : i \in \Delta_{0,k}\right\}$ for $k = 1, 2$, and $\left\{\boldsymbol{\gamma}_j : j = 1, \ldots, p\right\}$ with respect to the current conditional distribution of $\boldsymbol{T}$ given observed outcome data

$\boldsymbol{Y}$, covariates $\boldsymbol{X}$, and the current estimates of the parameters $\left(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_1^{2^{(t)}}, \sigma_2^{2^{(t)}}\right)$ :

$$\begin{aligned}
&\boldsymbol{E}\left[W_{ik} \mid Y_{ik}, \boldsymbol{\beta}_k^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_k^{2^{(t)}}\right],\\
&\boldsymbol{E}\left[W_{ik}^2 \mid Y_{ik}, \boldsymbol{\beta}_k^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_k^{2^{(t)}}\right], \text{ and}\\
&\boldsymbol{E}\left[\boldsymbol{\gamma}_j \mid \boldsymbol{Y}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_1^{2^{(t)}}, \sigma_2^{2^{(t)}}\right].
\end{aligned}$$

Finding the conditional distribution of $\boldsymbol{T}_{\cdot k}$ given $\boldsymbol{X}$ and $\left(\boldsymbol{\beta}_k^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_k^{2^{(t)}}\right)$ is equivalent to finding the conditional distribution of $\boldsymbol{\varepsilon}_{\cdot k}$ given $\boldsymbol{X}$ and current estimates at $t$-th iteration. In Sect. 2.2, we have shown that, given $\boldsymbol{\beta}_k^{(t)}$ as an estimate of $\boldsymbol{\beta}_k$, one is able to use the Kaplan-Meier estimator to compute the conditional expectations $\boldsymbol{E}\left[W_{ik} \mid Y_{ik}, \boldsymbol{\beta}_k^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_k^{2^{(t)}}\right]$ marginally as

$$W_{ik}{}^{(t)} = \boldsymbol{E}\left[W_{ik} \mid Y_{ik}, \boldsymbol{\beta}_k^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_k^{2^{(t)}}\right] = \sum_{h=1}^{n} w_{i,h}^k \varepsilon_{(h)k} + \boldsymbol{X}_i^T \boldsymbol{\beta}_k^{(t)}, \tag{8}$$

where $\varepsilon_{(h)k}$ is the order statistic of residual $\varepsilon_{\cdot k}^{(t)} = \left(\varepsilon_{1k}^{(t)}, \ldots, \varepsilon_{nk}^{(t)}\right)$ at $t$-th iteration. The sub-weights $w_{i,h}^k$'s for $k$-th survival time are defined as

$$w_{i,h}^k = \begin{cases} \dfrac{w_h^k}{1-\hat{F}_{\varepsilon_{\cdot k}, \boldsymbol{\beta}_k^{(t)}}\left(Y_{ik} - X_i^T \boldsymbol{\beta}_k^{(t)}\right)} & \text{if } \varepsilon_{(h)k} > \varepsilon_{ik}^{(t)} \\ 0 & \text{otherwise,} \end{cases}$$

such that $\hat{F}_{\varepsilon_{\cdot k}, \boldsymbol{\beta}_k^{(t)}}$ is the Kaplan-Meier estimator at $t$-th iteration computed marginally based on $\boldsymbol{\varepsilon}_{\cdot k}^{(t)}$, and $w_h^k$'s are Kaplan-Meier weights following computation described in Sect. 2.2 for $k = 1, 2$.

Similarly, we can find the conditional expectation of $W_{ik}^2$ given $Y_{ik}$, $\boldsymbol{X}_i$, and $\boldsymbol{\beta}_k^{(t)}$ by the Kaplan-Meier estimator:

$$\begin{aligned}
W_{ik}^{2\,(t)} &= \boldsymbol{E}\left[W_{ik}^2 \mid Y_{ik}, \boldsymbol{\beta}_k^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_k^{2^{(t)}}\right] \\
&= \sum_{h=1}^{n} w_{i,h}^k \varepsilon_{(h)k}^2 + \left(\boldsymbol{X}_i^T \boldsymbol{\beta}_k^{(t)}\right)^2 + 2\boldsymbol{X}_i^T \boldsymbol{\beta}_k^{(t)} \sum_{h=1}^{n} w_{i,h}^k \varepsilon_{(h)k}.
\end{aligned} \tag{9}$$

After imputing censored observations marginally for the bivariate survival data, our next step is to compute the posterior distribution of $\boldsymbol{\gamma}_j = (\gamma_{j1}, \gamma_{j2})$ for $j = 1, \ldots, p$. Since the relationship between $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$ are modeled through the prior specification of $\boldsymbol{\beta}$, the posterior distribution of $\boldsymbol{\gamma}_j$ is able to capture such information and pass it on to the next iteration. The posterior distribution of $\boldsymbol{\gamma}_j$ also follows a bivariate Bernoulli distribution such that

$$\begin{aligned}\pi\left(\boldsymbol{\gamma}_j \mid \cdot\right) \propto & \left[\pi_{11}\frac{1}{2\sqrt{v_1 \times v_1}}\exp\left\{-\frac{1}{2v_1}\beta_{j1}^2 - \frac{1}{2v_1}\beta_{j2}^2\right\}\right]^{\gamma_{j1}\gamma_{j2}} \\ & + \left[\pi_{10}\frac{1}{2\sqrt{v_1 \times v_0}}\exp\left\{-\frac{1}{2v_1}\beta_{j1}^2 - \frac{1}{2v_0}\beta_{j2}^2\right\}\right]^{\gamma_{j1}(1-\gamma_{j2})} \\ & + \left[\pi_{01}\frac{1}{2\sqrt{v_0 \times v_1}}\exp\left\{-\frac{1}{2v_0}\beta_{j1}^2 - \frac{1}{2v_1}\beta_{j2}^2\right\}\right]^{(1-\gamma_{j1})\gamma_{j2}} \\ & + \left[\pi_{00}\frac{1}{2\sqrt{v_0 \times v_0}}\exp\left\{-\frac{1}{2v_0}\beta_{j1}^2 - \frac{1}{2v_0}\beta_{j2}^2\right\}\right]^{(1-\gamma_{j1})(1-\gamma_{j2})}.\end{aligned}$$

Let $\boldsymbol{\phi}_j = \left(\phi_{j,11}, \phi_{j,10}, \phi_{j,01}, \phi_{j,00}\right)'$ be the probabilities of four possible outcomes for $j$-th variable. Based on the posterior distribution of $\boldsymbol{\gamma}_j$ , the conditional expectation of four combinations of $\gamma_{j1}$ and $\gamma_{j2}$ given $\boldsymbol{Y}$, $\boldsymbol{X}$, and the current estimates of the parameters $\left(\boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_1^{2^{(t)}}, \sigma_2^{2^{(t)}}\right)$ can be computed as

$$\phi_{j,lm}^{(t)} = \boldsymbol{E}\left[\gamma_{j1} = l, \gamma_{j2} = m \,\middle|\, \boldsymbol{Y}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_1^{2^{(t)}}, \sigma_2^{2^{(t)}}\right]$$

with

$$\begin{aligned}\phi_{j,11}^{(t)} &\propto \pi_{11}\frac{1}{2\sqrt{v_1 \times v_1}}\exp\left\{-\frac{1}{2v_1}\beta_{j1}^2 - \frac{1}{2v_1}\beta_{j2}^2\right\}, \\ \phi_{j,10}^{(t)} &\propto \pi_{10}\frac{1}{2\sqrt{v_1 \times v_0}}\exp\left\{-\frac{1}{2v_1}\beta_{j1}^2 - \frac{1}{2v_0}\beta_{j2}^2\right\}, \\ \phi_{j,01}^{(t)} &\propto \pi_{01}\frac{1}{2\sqrt{v_0 \times v_1}}\exp\left\{-\frac{1}{2v_0}\beta_{j1}^2 - \frac{1}{2v_1}\beta_{j2}^2\right\}, \\ \phi_{j,00}^{(t)} &\propto \pi_{00}\frac{1}{2\sqrt{v_0 \times v_0}}\exp\left\{-\frac{1}{2v_0}\beta_{j1}^2 - \frac{1}{2v_0}\beta_{j2}^2\right\},\end{aligned} \tag{10}$$

where $\phi_{j,11}^{(t)}, \phi_{j,10}^{(t)}, \phi_{j,01}^{(t)}$, and $\phi_{j,00}^{(t)}$ are all normalized for each $j = 1, \ldots, p$.

We define two diagonal matrices $\boldsymbol{D}_{\boldsymbol{\gamma},1}$ and $\boldsymbol{D}_{\boldsymbol{\gamma},2}$ such that both $\boldsymbol{D}_{\boldsymbol{\gamma},1}$ and $\boldsymbol{D}_{\boldsymbol{\gamma},2}$ have size $p \times p$:

$$\begin{aligned}\boldsymbol{D}_{\boldsymbol{\gamma},1} &= \operatorname{diag}\left\{\frac{\gamma_{j1}\gamma_{j2}}{v_1} + \frac{\gamma_{j1}(1-\gamma_{j2})}{v_1} + \frac{(1-\gamma_{j1})\gamma_{j2}}{v_0} + \frac{(1-\gamma_{j1})(1-\gamma_{j2})}{v_0}\right\}_{j=1}^{p}, \\ \boldsymbol{D}_{\boldsymbol{\gamma},2} &= \operatorname{diag}\left\{\frac{\gamma_{j1}\gamma_{j2}}{v_1} + \frac{\gamma_{j1}(1-\gamma_{j2})}{v_0} + \frac{(1-\gamma_{j1})\gamma_{j2}}{v_1} + \frac{(1-\gamma_{j1})(1-\gamma_{j2})}{v_0}\right\}_{j=1}^{p}.\end{aligned}$$

Then the conditional expectations of $\boldsymbol{D}_{\boldsymbol{\gamma},k}^{(t)} = \boldsymbol{E}\left[\boldsymbol{D}_{\boldsymbol{\gamma},k} \mid \boldsymbol{Y}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_k^{2(t)}\right]$ for $k = 1, 2$ given $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\boldsymbol{\beta}^{(t)}$ can be computed as

$$\begin{aligned} \boldsymbol{D}_{\boldsymbol{\gamma},1}^{(t)} &= \operatorname{diag}\left\{\frac{\phi_{j,11}^{(t)}}{v_1} + \frac{\phi_{j,10}^{(t)}}{v_1} + \frac{\phi_{j,01}^{(t)}}{v_0} + \frac{\phi_{j,00}^{(t)}}{v_0}\right\} = \operatorname{diag}\left\{\frac{\phi_{j,1\cdot}^{(t)}}{v_1} + \frac{\phi_{j,0\cdot}^{(t)}}{v_0}\right\}, \\ \boldsymbol{D}_{\boldsymbol{\gamma},2}^{(t)} &= \operatorname{diag}\left\{\frac{\phi_{j,11}^{(t)}}{v_1} + \frac{\phi_{j,10}^{(t)}}{v_0} + \frac{\phi_{j,01}^{(t)}}{v_1} + \frac{\phi_{j,00}^{(t)}}{v_0}\right\} = \operatorname{diag}\left\{\frac{\phi_{j,\cdot 1}^{(t)}}{v_1} + \frac{\phi_{j,\cdot 0}^{(t)}}{v_0}\right\}, \end{aligned} \tag{11}$$

where $\phi_{j,1\cdot}^{(t)}, \phi_{0\cdot}^{(t)}, \phi_{\cdot 1}^{(t)}$, and $\phi_{\cdot 0}^{(t)}$ are marginal posterior probabilities for $\gamma_{j1}$ and $\gamma_{j2}$ at $t$-th iteration.

For $k = 1, 2$, let $\widehat{\boldsymbol{W}}_k = \left(\hat{W}_{1k}, \ldots, \hat{W}_{nk}\right)$ and $\widehat{\boldsymbol{W}}_k^2 = \left(\hat{W}^2{}_{1k}, \ldots, \hat{W}^2{}_{nk}\right)$ be the augmented data for $k$-th survival time such that censored outcomes are imputed marginally using the Kaplan-Meier estimator. After $t$-th iteration, we can update the augmented data $\widehat{\boldsymbol{W}}^{(t)} = \left(\widehat{\boldsymbol{W}}_1^{(t)}, \widehat{\boldsymbol{W}}_2^{(t)}\right)$ and $\widehat{\boldsymbol{W}^2}^{(t)} = \left(\widehat{\boldsymbol{W}^2}_1^{(t)}, \widehat{\boldsymbol{W}^2}_2^{(t)}\right)$ with imputations as

$$\widehat{W_{ik}}^{(t)} = \begin{cases} Y_{ik} & \text{if } \delta_{ik} = 1 \\ W_{ik}^{(t)} & \text{if } \delta_{ik} = 0 \end{cases}, \quad \widehat{W_{ik}^2}^{(t)} = \begin{cases} Y_{ik}^2 & \text{if } \delta_{ik} = 1 \\ W_{ik}^{2\,(t)} & \text{if } \delta_{ik} = 0 \end{cases}.$$

### M-step

Given $\widehat{\boldsymbol{W}}^{(t)}$, $\widehat{\boldsymbol{W}^2}^{(t)}$, $\boldsymbol{D}_{\boldsymbol{\gamma},1}^{(t)}$, $\boldsymbol{D}_{\boldsymbol{\gamma},2}^{(t)}$ and $\left\{\boldsymbol{\gamma}_j^{(t)} : j = 1, \ldots, p\right\}$, we are able to maximize the objective function $\boldsymbol{Q}\left(\boldsymbol{\beta}, \boldsymbol{\pi}, \sigma_1^2, \sigma_2^2 \mid \boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)}\right)$ with respect to the parameters $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right)$, $\boldsymbol{\pi} = (\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})'$, $\sigma_1^2$, and $\sigma_2^2$. That is, we want to find the MAP estimators of $\boldsymbol{\beta}$, $\boldsymbol{\pi}$, $\sigma_1^2$, and $\sigma_2^2$ with the objective function

$$\begin{aligned} \boldsymbol{Q}\left(\boldsymbol{\beta}, \boldsymbol{\pi}, \sigma_1^2, \sigma_2^2 \mid \boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)}\right) = -\sum_{k=1}^{2} \left\| \widehat{\boldsymbol{W}}_k^{(t)} - \boldsymbol{X}^T \boldsymbol{\beta}_k \right\|^2 \\ - \boldsymbol{\beta}_k^T D_{\boldsymbol{\gamma},k}^{(t)} \boldsymbol{\beta}_k + \text{Const.} \end{aligned}$$

We first maximize $\boldsymbol{Q}\left(\boldsymbol{\beta}, \boldsymbol{\pi}, \sigma_1^2, \sigma_2^2 \mid \boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)}\right)$ with respect to $\boldsymbol{\beta}_k$ and $\sigma_k^2$ for $k = 1, 2$:

$$\boldsymbol{\beta}_k^{(t+1)} = \left(\boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{D}_{\boldsymbol{\gamma},k}\right)^{-1} \boldsymbol{X}^T \widehat{\boldsymbol{W}_k}^{(t)}, \tag{12}$$

$$\sigma_k^{2\,(t+1)} = \frac{\sum_{i=1}^{n} \left[\widehat{W_{ik}^2}^{(t)} + \left(\boldsymbol{X}_i^T \boldsymbol{\beta}_k^{(t+1)}\right)^2 - 2\widehat{W_{ik}}^{(t)} \boldsymbol{X}_i^T \boldsymbol{\beta}_k^{(t+1)}\right] + \lambda_{0,k}\sigma_{0,k}^2}{n + \lambda_{0,k} + 2}. \tag{13}$$

Next we maximize $\boldsymbol{Q}\left(\boldsymbol{\beta}, \boldsymbol{\pi}, \sigma_1^2, \sigma_2^2 \mid \boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_1^{2^{(t)}}, \sigma_2^{2^{(t)}}\right)$ with respect to hyper-parameter $\pi_{11}, \pi_{10}, \pi_{01}$, and $\pi_{00}$ subject to constraint $\pi_{11}+\pi_{10}+\pi_{01}+\pi_{00} = 1$, and obtain

$$\begin{aligned}
{\pi_{11}}^{(t)} &= \frac{\sum_{j=1}^{p} \phi_{j,11}^{(t)} + \alpha_{11} - 1}{p + \alpha_{11} + \alpha_{10} + \alpha_{01} + \alpha_{00} - 4}, \\
{\pi_{10}}^{(t)} &= \frac{\sum_{j=1}^{p} \phi_{j,10}^{(t)} + \alpha_{10} - 1}{p + \alpha_{11} + \alpha_{10} + \alpha_{01} + \alpha_{00} - 4}, \\
{\pi_{01}}^{(t)} &= \frac{\sum_{j=1}^{p} \phi_{j,01}^{(t)} + \alpha_{01} - 1}{p + \alpha_{11} + \alpha_{10} + \alpha_{01} + \alpha_{00} - 4}, \\
{\pi_{00}}^{(t)} &= \frac{\sum_{j=1}^{p} \phi_{j,00}^{(t)} + \alpha_{00} - 1}{p + \alpha_{11} + \alpha_{10} + \alpha_{01} + \alpha_{00} - 4}.
\end{aligned} \tag{14}$$

**Algorithm 1:** An Algorithm Using Kaplan-Meier Estimator for Bivariate Survival Data

```
Initialize $\boldsymbol{\phi}_j^{(0)}$, $\boldsymbol{\beta}_k^{(0)}$, $\boldsymbol{\pi}^{(0)}$, and $\sigma_k^{2^{(0)}}$ for $j = 1, \ldots, p$ and $k = 1, 2$;
for $t = 1 : maxIter$ do
    Update the approximate posterior parameters $\widehat{W_{ik}}^{(t)}$, $\widehat{W_{ik}^2}^{(t)}$, $\boldsymbol{\phi}_j^{(t)}$, and $\boldsymbol{D}_{\boldsymbol{\gamma},k}^{(t)}$ from Equation (8), (9), (10), and (11);
    Update the MAP estimators $\boldsymbol{\beta}_k^{(t+1)}$, $\sigma_k^{2^{(t+1)}}$, and $\left(\pi_{11}^{(t)}, \pi_{10}^{(t)}, \pi_{01}^{(t)}, \pi_{00}^{(t)}\right)'$ from Equation (12), (13) and (14);
    if stopping criterion is satisfied then
        break;
    end
end
```

### Deterministic annealing

Since the method proposed above is analogous to the EM algorithm, a potential drawback of the algorithm is that it can be prone to entrapment in local maximum modes for multi-modal posterior landscapes (Ročková and George 2014). To migrate this issue, we follow Ročková and George (2014) and further propose a deterministic annealing variant of the iterative algorithm. That is, consider a tempered version of the objective function derived from Equation (6) which embeds Equation (6) as a special case:

$$l_t\left(\boldsymbol{W}, \boldsymbol{\beta}, \boldsymbol{\pi}, \sigma_1^2, \sigma_2^2 \mid \boldsymbol{X}\right) = \frac{1}{t} \log \sum_{k=1}^{2} \sum_{i=1}^{n} \left\{- \left(W_{ik} - \boldsymbol{X}_i \boldsymbol{\beta}_k\right)^2 \times \pi\left(\boldsymbol{\beta}\right)\right\}^t, \tag{15}$$

where $0 < t \leq 1$ and $1/t$ determines the degree of separation between the multiple modes of $l_t$. The M-step for maximizing Equation (15) stays the same as Algorithm 1. The E-step, on the other hand, can be obtained following Algorithm 1 except for Equation (10),

$$\phi_{j,lm}^{(t)} = \left\{ \boldsymbol{E} \left[ \gamma_{j1} = l, \gamma_{j2} = m \mid \boldsymbol{Y}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)} \right] \right\}^t. \tag{16}$$

In practice, to optimize Equation (15), we consider a decreasing sequence $1/t_1 > 1/t_2 > \cdots > 1/t_{maxIter}$ suggested by Ročková and George (2014), where the solution at $1/t_i$ serves as the starting point for the computation at $1/t_{i+1}$. When $t$ approaches 0, the unique solution can be found for Equation (15),

$$\boldsymbol{\beta}_k^{init} = \left[ \boldsymbol{X}^T \boldsymbol{X} + \frac{v_0 + v_1 + 1}{2 v_0 v_1} \boldsymbol{I}_p \right]^{-1} \boldsymbol{X}^T \widehat{\boldsymbol{W}_k},$$

where $\widehat{\boldsymbol{W}}$'s are the imputed censored observations calculated from one-step Kaplan-Meier weights. Following Ročková and George (2014), $\boldsymbol{\beta}_k^{init}$ can be served as a very promising general initialization value for the proposed algorithm.

### 3.3 Remarks

There are a few remarks on the algorithm described above for bivariate survival data.

- In our framework, we assume that the design matrix $\boldsymbol{X}$ is shared by both columns of survival times. The data structure fits many biomedical studies where multiple types of event times are collected. The framework cannot be applied to recurrent events but allows $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$ to have any kind of correlation structures. In our real data analysis, we first consider event-free time which characterizes the length of time that the patient remains free of disease after primary treatment of the study for the disease ends. These events may include the return of the cancer or the onset of certain symptoms. We also consider overall survival which stands for the length of time that the patient has survived from the start of the study for a disease. In such cases, there is only one set of covariates collected for two different types of event times. However, our framework can be easily extended to fit cases that when two studies consist of the same group of patients but collect overlapping covariates. Therefore the dimension $p$ for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are also not required to be the same.
- If we assume that both $F_1$ and $F_2$ follow specified distributions such as independent Gaussian distributions with mean zero and finite variance $\sigma_1^2$ and $\sigma_2^2$ respectively, the estimation steps of Equation (8) and Equation (9) can be replaced by computing the first and the second moment of truncated Gaussian distributions.

  It is easy to show that, for $k = 1, 2$, the conditional distribution $W_{ik}$ given $Y_{ik}$, $\boldsymbol{X}_i$, and current estimates $\left(\boldsymbol{\beta}_k^{(t)}, \boldsymbol{\pi}^{(t)}, \sigma_k^{2(t)}\right)$ follows a Gaussian distribution with mean $\boldsymbol{X}_i^T \boldsymbol{\beta}_k^{(t)}$ and variance $\sigma_k^{2(t)}$ truncating at $Y_{ik}$.

Let $\alpha_k = \frac{Y_{ik} - X_i^T \boldsymbol{\beta}_k^{(t)}}{\sigma_k^{(t)}}$ and $\Phi(\cdot), \phi(\cdot)$ be the distribution function and the density function for a standard Gaussian respectively. Specifically, Equation (8) and Equation (9) can be replaced by

$$\begin{aligned} W_{ik}^{(t)} &= \mathbf{X}_i^T \boldsymbol{\beta}_k^{(t)} + \frac{\phi(\alpha_k)}{1 - \Phi(\alpha_k)} \sigma_k^{(t)}, \\ W_{ik}^{2\,(t)} &= \sigma_k^{2\,(t)} \left[ 1 + \frac{\alpha_k \phi(\alpha_k)}{1 - \Phi(\alpha_k)} - \left( \frac{\phi(\alpha_k)}{1 - \Phi(\alpha_k)} \right)^2 \right] + \left[ W_{ik}^{(t)} \right]^2. \end{aligned} \tag{17}$$

Other parametric distributions such as Weibull distribution or log Student's $t$-distribution can also be considered and the corresponding parameters can be estimated in a similar manner.

## 4 Empirical results

### 4.1 Simulation setups

The performance of the proposed method is examined using three simulation studies. For each study, we generate the design matrix $\boldsymbol{X}$ from multivariate normal distribution with mean zero and covariance matrix with elements $\boldsymbol{\Sigma}_{ij} = 0.5^{|i-j|}$. Then we generate two survival times $\boldsymbol{U}$ and $\boldsymbol{V}$ independently following either normal distribution or exponential distribution. To better mimic the pattern of different types of bivariate survival times in real life, we consider a weight variable $c$ such that $c$ controls how much the two survival times $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$ are correlated. That is, we generate $\boldsymbol{T} = (\boldsymbol{T}_{\cdot 1}, \boldsymbol{T}_{\cdot 2})$ such that

$$\boldsymbol{T}_{\cdot 1} = \boldsymbol{U}, \ \boldsymbol{T}_{\cdot 2} = (1 - c)\boldsymbol{U} + c\boldsymbol{V},$$

where $0 < c \leq 1$.

When $c$ takes value 1, we end up with two independent survival times $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$. On the other hand, when $0 < c < 1$, two survival times $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$ fit into the case of event-free survival and overall survival which are the types of data we encounter in the real data analysis. Censoring times $\boldsymbol{C}_{\cdot 1} = \boldsymbol{C}_{\cdot 2}$ are generated from $\mathsf{Unif}(0, \eta)$. When $0 < c < 1$, the value of $\eta$ is chosen such that 40% and 60% censoring rates are achieved for two events respectively. When $c = 1$, the value of $\eta$ will be chosen such that 40% of censoring is achieved by both events.

Let sample size $n$ be 100 and dimension $p$ be 100, 500, or 800. For each $n, p$ combination, 10 nonzero coefficients are randomly selected for each column of the true $\boldsymbol{\beta}$ matrix. All of the simulation studies are repeated by 200 times. We generate nonzero coefficients independently from $\mathsf{N}(3, 0.5)$ and fix the true coefficient values for all simulation runs.

In total four methods are compared to demonstrate the performance of the proposed method: Univariate Coxnet which is a $L_1$ penalized method of the Cox model developed by Tibshirani (1997) for univariate survival data with $\lambda_{min}$ or $\lambda_{1se}$; Univariate

Adaptive Elastic Net (Univariate AEnet) for AFT models which is a penalized method developed for univariate AFT models under the marginal independence assumption (Khan and Shaw 2016); Univariate Bayesian penalized Buckley-James estimator (Univariate BP-BJ) which is the same framework as the proposed method developed under univariate spike-and-slab prior; and the proposed Bivariate Bayesian penalized Buckley-James estimator (Bivariate BP-BJ) for bivariate survival data. Since there are no readily implemented variable selection methods for multivariate survival data, we take these approaches for comparison purpose.

For methods developed for univariate survival data, we will apply the method separately on each survival time and evaluate the combined results based on false positive (FP), false negative (FN), sensitivity, specificity, and Matthews correlation coefficient (MCC) which are defined as the following

$$\text{sensitivity} = \frac{TP}{TP+FN}, \quad \text{specificity} = \frac{TN}{TN+FP},$$
$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

Higher scores in sensitivity, specificity, and MCC indicate better performance in model selection. For all of the simulation studies, values of $\upsilon_0$ are obtained by tuning parameter selection procedure proposed in Sect. 4.2 while fixing $\upsilon_1 = 1$.

### [Study 1] No-sharing

We randomly pick the positions of nonzero coefficients in the true $\boldsymbol{\beta}$ matrix and ensure that there is no overlap between the relevant variables in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ by design. We only consider the case $c = 1$ such that $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$ are independent.

The results of Study 1 are summarized in Tables 1 and 2. We see that in terms of false positives and false negatives, treating the survival times separately results in frequent selections of the noise variables but is able to recognize most of the relevant variables for both low and high-dimensional designs. The Univariate AEnet method selects fewer noise variables and misses more relevant variables. Since our design of the simulation study invalidates the marginal independence assumption of the AEnet method, we observe different behaviors of false positives and false negatives for the method compared with other univariate competing methods. The proposed bivariate method, on the other hand, is able to achieve the best MCC score for all of the designs of Study 1 with relatively low sensitivity scores and high specificity scores. The algorithm is more strict in selecting relevant variables and is able to help recognizing zero elements in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ by updating the posterior distribution.

### [Study 2] All-sharing

WLOG, we assume that the first 10 variables are relevant variables for both $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$. That is, the nonzero positions for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are exactly the same. Here we only consider $c = 1$ as a special case.

The performance of four methods are reported in Tables 3 and 4. Study 2 represents an extreme case that the two survival time are independent but they are both affected by the exact same set of covariates. By applying the method separately, large number of false negatives suggests that marginal models fail to capture the connection between the two events. However, in the proposed method, if we select a variable that is relevant with one of the events, with high probability that the variable is also a signal variable for the other event. The proposed method indicates great performance comparing to any of the univariate models in both false positives and false negatives with lower sensitivity but higher specificity scores while achieving the best MCC scores for all of the designs.

### [Study 3] Some-sharing

For most of the real world scenarios, only partial variables are shared by the two events. For both $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$, we assume that the first 5 variables are relevant and then randomly selected another 5 variables from the rest such that only the first 5 variables are shared by $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$. For the some-sharing cases, we consider $c = 0.3, 0.5, 0.7$, or 1. By design, the number of true relevant variables in total will be 25 for $c = 0.3, 0.5$ and 0.7. If $c = 1$, there are still 20 relevant variables.

The results of Study 3 can be found in Tables 5 and 6. When $c \neq 1$, the proposed method fails to recognize about half of the relevant variables for all of the high-dimensional designs. The reason is that, once a variable is recognized as irrelevant in one of the columns, it is highly likely to be also recognized as a noise variable unless the signal is strong enough to be identified in its own column. The proposed method tends to have lower false positives with the increase in dimensionality but is still able to achieve the best MCC scores among all competing methods. We also see that the nonparametric imputation is able to give robust estimates for censored observations as there is almost no much difference between designs with exponential distribution and Gaussian distribution.

Overall, the proposed method not only provides a procedure with how to deal with bivariate survival data, it also demonstrates better performance comparing to treating the event one at a time. By applying univariate methods to bivariate survival data, we also intend to conceptualize a way of dealing with bivariate survival data more formally. Note that all of the datasets generated for empirical studies are available from the corresponding author on request.

## 4.2 Tuning parameter selection

Recall the Bayesian hierarchical framework defined in Eq. (7), the tuning parameters are identified as $(\upsilon_0, \upsilon_1)$, $\lambda_{0,k}$, $\sigma^2_{0,k}$ for $k = 1, 2$, and $(\alpha_{11}, \alpha_{10}, \alpha_{01}, \alpha_{11})$. In practice, we recommend to choose $\lambda_{0,k} = 1$, $\sigma^2_{0,k} = 1$, $\alpha_{11} = \alpha_{10} = \alpha_{01} = \alpha_{11} = 2$ and to fix $\upsilon_1 = 1$.

We see that the $\upsilon_0$ in spike-and-slab priors controls the regularization power of the variable selection, which is similar to the penalty parameter $\lambda$ in LASSO. Sabourin et al. (2015) proposed a data-driven method for choosing penalty parameter $\lambda$ in LASSO

**Table 1** False positives and false negatives reported for five different methods with no sharing positions of relevant variables in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$

| $p$ | dist. | Univariate coxnet with $\lambda_{1se}$ | | Univariate coxnet with $\lambda_{min}$ | | Univariate AEnet | | Univariate BP-BJ | | Bivariate BP-BJ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FP | FN | FP | FN | FP | FN | FP | FN | FP | FN |
| $c = 1$ | | | | | | | | | | | |
| 100 | exp. | 20.0 | 0.0 | 23.6 | 0.0 | 6.4 | 13.8 | 11.6 | 0.1 | 0.77 | 0.0 |
| | norm. | 15.5 | 0.0 | 17.7 | 0.0 | 6.0 | 13.0 | 9.8 | 0.3 | 0.42 | 0.0 |
| 500 | exp. | 34.5 | 1.0 | 56.2 | 0.3 | 8.2 | 15.8 | 26.3 | 2.6 | 1.9 | 2.8 |
| | norm. | 35.5 | 0.9 | 54.9 | 0.2 | 8.1 | 15.9 | 24.8 | 2.9 | 1.7 | 2.9 |
| 800 | exp. | 32.3 | 2.6 | 62.5 | 0.5 | 9.0 | 16.6 | 21.4 | 4.4 | 1.3 | 7.2 |
| | norm. | 34.8 | 2.1 | 60.1 | 0.6 | 9.0 | 16.6 | 20.3 | 4.4 | 1.3 | 7.2 |

**Table 2** Sensitivity, specificity, and MCC scores reported for five different methods with no sharing positions of relevant variables in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$

| $p$ | dist. | Univariate coxnet with $\lambda_{1se}$ | | | Univariate coxnet with $\lambda_{min}$ | | | Univariate AEnet | | | Univariate BP-BJ | | | Bivariate BP-BJ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC |
| $c = 1$ | | | | | | | | | | | | | | | | |
| 100 | exp. | 1.0 | 0.89 | 0.67 | 1.0 | 0.87 | 0.64 | 0.31 | 0.96 | 0.34 | 0.99 | 0.94 | 0.78 | 1.0 | 0.99 | **0.98** |
| | norm. | 1.0 | 0.91 | 0.72 | 1.0 | 0.90 | 0.70 | 0.35 | 0.97 | 0.38 | 0.98 | 0.95 | 0.80 | 1.0 | 0.99 | **0.99** |
| 500 | exp. | 0.95 | 0.96 | 0.57 | 0.99 | 0.94 | 0.49 | 0.21 | 0.99 | 0.25 | 0.87 | 0.97 | 0.58 | 0.86 | 0.99 | **0.88** |
| | norm. | 0.96 | 0.96 | 0.57 | 0.99 | 0.94 | 0.50 | 0.21 | 0.99 | 0.25 | 0.86 | 0.97 | 0.58 | 0.85 | 0.99 | **0.88** |
| 800 | exp. | 0.87 | 0.98 | 0.55 | 0.98 | 0.96 | 0.47 | 0.17 | 0.99 | 0.20 | 0.78 | 0.99 | 0.57 | 0.64 | 1.0 | **0.76** |
| | norm. | 0.90 | 0.98 | 0.55 | 0.97 | 0.96 | 0.48 | 0.17 | 0.99 | 0.21 | 0.78 | 0.99 | 0.58 | 0.64 | 1.0 | **0.76** |

The best MCC score is bolded for each simulation design among all competing methods

**Table 3** False positives and false negatives reported for five different methods with all sharing positions of relevant variables in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$

| $p$ | dist. | Univariate coxnet with $\lambda_{1se}$ | | Univariate coxnet with $\lambda_{min}$ | | Univariate AEnet | | Univariate BP-BJ | | Bivariate BP-BJ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FP | FN | FP | FN | FP | FN | FP | FN | FP | FN |
| $c = 1$ | | | | | | | | | | | |
| 100 | exp. | 17.6 | 0.1 | 20.8 | 0.1 | 6.4 | 13.4 | 9.8 | 1.0 | 0.7 | 0.0 |
| | norm. | 16.8 | 0.0 | 18.9 | 0.0 | 5.8 | 13.2 | 10.0 | 0.2 | 0.4 | 0.0 |
| 500 | exp. | 33.7 | 1.0 | 56.2 | 0.2 | 8.3 | 16.5 | 26.0 | 2.5 | 2.3 | 0.9 |
| | norm. | 36.7 | 0.7 | 56.3 | 0.2 | 8.4 | 16.0 | 21.6 | 2.5 | 2.0 | 0.9 |
| 800 | exp. | 32.3 | 2.4 | 61.7 | 0.6 | 8.8 | 16.9 | 15.9 | 4.1 | 1.6 | 5.7 |
| | norm. | 36.0 | 1.4 | 60.4 | 0.2 | 9.1 | 16.3 | 20.6 | 3.1 | 1.8 | 4.6 |

**Table 4** Sensitivity, specificity, and MCC scores reported for five different methods with all sharing positions of relevant variables in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$

| $p$ | dist. | Univariate coxnet with $\lambda_{1se}$ | | | Univariate coxnet with $\lambda_{min}$ | | | Univariate AEnet | | | Univariate BP-BJ | | | Bivariate BP-BJ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC |
| $c = 1$ | | | | | | | | | | | | | | | | |
| 100 | exp. | 1.0 | 0.90 | 0.70 | 1.0 | 0.88 | 0.66 | 0.33 | 0.96 | 0.35 | 0.95 | 0.95 | 0.78 | 1.0 | 1.0 | **0.98** |
| | norm. | 1 | 0.91 | 0.71 | 1.0 | 0.90 | 0.68 | 0.34 | 0.97 | 0.37 | 0.99 | 0.94 | 0.80 | 1.0 | 1.0 | **0.99** |
| 500 | exp. | 0.95 | 0.97 | 0.58 | 0.99 | 0.94 | 0.50 | 0.18 | 0.99 | 0.21 | 0.88 | 0.97 | 0.59 | 0.95 | 1.0 | **0.92** |
| | norm. | 0.96 | 0.96 | 0.57 | 0.99 | 0.94 | 0.50 | 0.20 | 0.99 | 0.24 | 0.87 | 0.98 | 0.62 | 0.96 | 1.0 | **0.93** |
| 800 | exp. | 0.88 | 0.98 | 0.56 | 0.97 | 0.96 | 0.47 | 0.15 | 0.99 | 0.19 | 0.81 | 0.99 | 0.63 | 0.72 | 1.0 | **0.80** |
| | norm. | 0.93 | 0.98 | 0.56 | 0.99 | 0.96 | 0.49 | 0.18 | 0.99 | 0.22 | 0.85 | 0.99 | 0.61 | 0.77 | 1.0 | **0.83** |

The best MCC score is bolded for each simulation design among all competing methods

**Table 5** False positives and false negatives reported for five different methods with some sharing positions of relevant variables in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Four different $c$ values are considered

| $p$ | dist. | Univariate coxnet with $\lambda_{1se}$ | | Univariate coxnet with $\lambda_{min}$ | | Univariate AEnet | | Univariate BP-BJ | | Bivariate BP-BJ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FP | FN | FP | FN | FP | FN | FP | FN | FP | FN |
| $c = 0.3$ | | | | | | | | | | | |
| 100 | exp. | 18.1 | 5.1 | 27.4 | 2.6 | 5.8 | 20.1 | 11.5 | 6.1 | 4.1 | 4.0 |
| | norm. | 14.3 | 4.8 | 23.1 | 2.3 | 5.2 | 19.3 | 14.9 | 5.2 | 4.5 | 3.2 |
| 500 | exp. | 21.7 | 12.3 | 47.9 | 7.6 | 7.8 | 22.4 | 18.8 | 10.4 | 3.7 | 10.7 |
| | norm. | 21.9 | 12.0 | 47.8 | 7.4 | 7.6 | 22.4 | 14.3 | 10.5 | 2.7 | 11.1 |
| 800 | exp. | 17.6 | 14.0 | 46.5 | 9.1 | 8.1 | 22.6 | 16.6 | 12.3 | 1.9 | 15.6 |
| | norm. | 19.5 | 13.8 | 48.2 | 9.2 | 8.2 | 22.5 | 11.8 | 12.7 | 1.2 | 16.4 |
| $c = 0.5$ | | | | | | | | | | | |
| 100 | exp. | 16.4 | 6.6 | 26.5 | 3.0 | 5.1 | 20.6 | 13.4 | 5.4 | 5.4 | 3.5 |
| | norm. | 16.1 | 5.9 | 25.4 | 2.5 | 5.3 | 20.4 | 17.0 | 5.3 | 6.0 | 3.5 |
| 500 | exp. | 21.0 | 10.7 | 45.1 | 7.1 | 7.1 | 22.0 | 20.1 | 9.9 | 2.9 | 10.7 |
| | norm. | 21.8 | 11.6 | 47.3 | 7.6 | 7.1 | 22.4 | 12.9 | 10.9 | 2.3 | 12.0 |
| 800 | exp. | 20.4 | 13.0 | 49.2 | 8.8 | 8.2 | 22.8 | 14.6 | 13.2 | 1.4 | 15.7 |
| | norm. | 20.2 | 13.8 | 48.2 | 9.5 | 7.7 | 22.9 | 12.6 | 13.3 | 1.2 | 16.9 |

**Table 5** (continued)

| $p$ | dist. | Univariate coxnet with $\lambda_{1se}$ | | Univariate coxnet with $\lambda_{min}$ | | Univariate AEnet | | Univariate BP-BJ | | Bivariate BP-BJ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FP | FN | FP | FN | FP | FN | FP | FN | FP | FN |
| $c = 0.7$ | | | | | | | | | | | |
| 100 | exp. | 14.7 | 5.1 | 23.5 | 2.7 | 5.0 | 19.7 | 16.5 | 5.6 | 5.0 | 3.7 |
| | norm. | 16.9 | 4.9 | 25.1 | 2.4 | 5.8 | 20.1 | 15.3 | 5.2 | 4.4 | 3.2 |
| 500 | exp. | 20.9 | 11.8 | 47.0 | 7.5 | 7.0 | 22.1 | 20.8 | 9.9 | 3.4 | 10.9 |
| | norm. | 21.4 | 12.1 | 46.8 | 7.4 | 7.9 | 22.5 | 15.9 | 10.5 | 2.8 | 11.0 |
| 800 | exp. | 20.5 | 12.6 | 49.6 | 8.6 | 8.3 | 22.5 | 12.3 | 12.6 | 1.2 | 15.7 |
| | norm. | 19.5 | 12.6 | 47.6 | 8.1 | 7.6 | 22.2 | 11.7 | 12.1 | 0.8 | 16.5 |
| $c = 1$ | | | | | | | | | | | |
| 100 | exp. | 18.5 | 0.0 | 21.3 | 0.0 | 6.1 | 13.7 | 10.8 | 0.5 | 0.8 | 0.0 |
| | norm. | 15.4 | 0.0 | 17.1 | 0.0 | 6.1 | 13.7 | 10.2 | 0.3 | 0.8 | 0.0 |
| 500 | exp. | 36.5 | 0.8 | 60.2 | 0.1 | 8.8 | 16.4 | 21.1 | 2.2 | 1.8 | 1.4 |
| | norm. | 37.0 | 0.8 | 57.6 | 0.2 | 8.8 | 16.4 | 21.8 | 2.4 | 1.7 | 1.8 |
| 800 | exp. | 33.0 | 2.5 | 62.0 | 0.6 | 9.0 | 16.9 | 19.2 | 4.0 | 1.5 | 6.5 |
| | norm. | 34.9 | 1.8 | 61.8 | 0.3 | 9.0 | 16.7 | 17.5 | 3.7 | 1.5 | 6.7 |

**Table 6** Sensitivity, specificity, and MCC scores reported for five different methods with some sharing positions of relevant variables in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Four different $c$ values are considered

| $p$ | dist. | Univariate coxnet with $\lambda_{1se}$ | | | Univariate coxnet with $\lambda_{min}$ | | | Univariate AEnet | | | Univariate BP-BJ | | | Bivariate BP-BJ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC |
| $c = 0.3$ | | | | | | | | | | | | | | | | |
| 100 | exp. | 0.80 | 0.90 | 0.59 | 0.90 | 0.84 | 0.57 | 0.20 | 0.97 | 0.24 | 0.76 | 0.93 | 0.64 | 0.84 | 0.98 | **0.82** |
| | norm. | 0.81 | 0.92 | 0.64 | 0.91 | 0.87 | 0.61 | 0.23 | 0.97 | 0.29 | 0.79 | 0.91 | 0.62 | 0.87 | 0.97 | **0.83** |
| 500 | exp. | 0.51 | 0.98 | 0.42 | 0.70 | 0.95 | 0.41 | 0.11 | 0.99 | 0.15 | 0.59 | 0.98 | 0.49 | 0.57 | 1.0 | **0.67** |
| | norm. | 0.52 | 0.98 | 0.42 | 0.70 | 0.95 | 0.41 | 0.10 | 0.99 | 0.15 | 0.58 | 0.99 | 0.53 | 0.55 | 1.0 | **0.67** |
| 800 | exp. | 0.44 | 0.99 | 0.40 | 0.64 | 0.97 | 0.39 | 0.09 | 0.99 | 0.14 | 0.51 | 0.99 | 0.46 | 0.38 | 1.0 | **0.55** |
| | norm. | 0.45 | 0.99 | 0.39 | 0.63 | 0.97 | 0.38 | 0.10 | 0.99 | 0.14 | 0.49 | 0.99 | 0.50 | 0.34 | 1.0 | **0.54** |
| $c = 0.5$ | | | | | | | | | | | | | | | | |
| 100 | exp. | 0.74 | 0.91 | 0.56 | 0.88 | 0.85 | 0.56 | 0.18 | 0.97 | 0.23 | 0.78 | 0.92 | 0.64 | 0.86 | 0.97 | **0.81** |
| | norm. | 0.77 | 0.91 | 0.59 | 0.90 | 0.85 | 0.59 | 0.18 | 0.97 | 0.23 | 0.79 | 0.90 | 0.60 | 0.86 | 0.97 | **0.79** |
| 500 | exp. | 0.57 | 0.98 | 0.47 | 0.72 | 0.95 | 0.43 | 0.12 | 0.99 | 0.17 | 0.60 | 0.98 | 0.57 | 0.53 | 0.99 | **0.68** |
| | norm. | 0.54 | 0.98 | 0.44 | 0.70 | 0.95 | 0.41 | 0.10 | 0.99 | 0.15 | 0.56 | 0.99 | 0.52 | 0.42 | 1.0 | **0.66** |
| 800 | exp. | 0.48 | 0.99 | 0.41 | 0.65 | 0.97 | 0.39 | 0.09 | 0.99 | 0.13 | 0.47 | 0.99 | 0.45 | 0.37 | 1.0 | **0.56** |
| | norm. | 0.45 | 0.99 | 0.39 | 0.62 | 0.97 | 0.38 | 0.09 | 1.00 | 0.13 | 0.47 | 0.99 | 0.47 | 0.32 | 1.0 | **0.52** |

The best MCC score is bolded for each simulation design among all competing methods

**Table 6** (continued)

| $p$ | dist. | Univariate coxnet with $\lambda_{1se}$ | | | Univariate coxnet with $\lambda_{min}$ | | | Univariate AEnet | | | Univariate BP-BJ | | | Bivariate BP-BJ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC | sens. | spec. | MCC |
| $c = 0.7$ | | | | | | | | | | | | | | | | |
| 100 | exp. | 0.79 | 0.92 | 0.63 | 0.89 | 0.87 | 0.60 | 0.21 | 0.97 | 0.27 | 0.77 | 0.91 | 0.59 | 0.85 | 0.97 | **0.81** |
| | norm. | 0.80 | 0.90 | 0.61 | 0.90 | 0.86 | 0.59 | 0.20 | 0.97 | 0.24 | 0.79 | 0.91 | 0.62 | 0.87 | 0.97 | **0.83** |
| 500 | exp. | 0.53 | 0.98 | 0.44 | 0.70 | 0.95 | 0.41 | 0.12 | 0.99 | 0.17 | 0.61 | 0.98 | 0.49 | 0.56 | 1.0 | **0.67** |
| | norm. | 0.52 | 0.98 | 0.43 | 0.70 | 0.95 | 0.42 | 0.10 | 0.99 | 0.14 | 0.58 | 0.98 | 0.51 | 0.56 | 1.0 | **0.68** |
| 800 | exp. | 0.50 | 0.99 | 0.43 | 0.66 | 0.97 | 0.39 | 0.10 | 0.99 | 0.14 | 0.50 | 0.99 | 0.50 | 0.37 | 1.0 | **0.56** |
| | norm. | 0.50 | 0.99 | 0.43 | 0.68 | 0.97 | 0.41 | 0.11 | 1.00 | 0.17 | 0.52 | 0.99 | 0.52 | 0.34 | 1.0 | **0.55** |
| $c = 1$ | | | | | | | | | | | | | | | | |
| 100 | exp. | 1.0 | 0.90 | 0.69 | 1.0 | 0.88 | 0.66 | 0.31 | 0.97 | 0.34 | 0.97 | 0.94 | 0.78 | 1.0 | 1.0 | **0.98** |
| | norm. | 1.0 | 0.91 | 0.72 | 1.0 | 0.91 | 0.70 | 0.32 | 0.97 | 0.35 | 0.98 | 0.94 | 0.79 | 1.0 | 1.0 | **0.99** |
| 500 | exp. | 0.96 | 0.96 | 0.57 | 0.99 | 0.94 | 0.48 | 0.18 | 0.99 | 0.21 | 0.89 | 0.98 | 0.63 | 0.93 | 1.0 | **0.92** |
| | norm. | 0.96 | 0.96 | 0.56 | 0.99 | 0.94 | 0.49 | 0.18 | 0.99 | 0.21 | 0.88 | 0.98 | 0.62 | 0.91 | 1.0 | **0.91** |
| 800 | exp. | 0.88 | 0.98 | 0.55 | 0.97 | 0.96 | 0.47 | 0.15 | 0.99 | 0.19 | 0.80 | 0.99 | 0.60 | 0.67 | 1.0 | **0.77** |
| | norm. | 0.91 | 0.98 | 0.55 | 0.98 | 0.96 | 0.48 | 0.16 | 0.99 | 0.20 | 0.82 | 0.99 | 0.62 | 0.66 | 1.0 | **0.77** |

problems using permutation, which motivates us to design a similar algorithm for choosing $v_0$. Sabourin et al. (2015) proposed that, by randomly permuting the response variable $\boldsymbol{Y}$, an underlying null model can be achieved as the true model, and one would be able to find the point of $\lambda$ such that the first variable enters the null model with a pre-specified sequence of $\lambda$ choices, and record the previous step of $\lambda$ as the tuning parameter selection.

In this paper, we propose a similar data-driven algorithm with permutation for selecting the appropriate value of $v_0$ within the context of survival analysis. By randomly permuting pairs of $(\boldsymbol{Y}_i, \boldsymbol{\delta}_i)$, we tune $v_0$ with a pre-specified sequence, re-scaled with $\sqrt{\log(p)/n}$ for cooperating with the relationship between $p$ and $n$. For each $v_0$ candidate, we repeat the permutation process for at least 50 times to obtain the averaged posterior probabilities and make sure that the posterior probabilities are stable and not affected by randomness. Since for each $p$, the tuple of posterior probabilities has four components, we check in total how many variables have $\phi_{00}$ as the maximum value within the tuple and find the value of $v_0$ such that the first variable enters the null model.

The algorithm is designed to start with any pre-specified sequence $V_0$. In practice, we recommend to start with a relatively long sequence but with larger gaps between each step for $V_0$. The permutation selection method can also be applied to any Bayesian variable selection methods.

**Algorithm 2:** Tuning Parameter Selection

```
initialize $\lambda_{0,k} = \sigma^2_{0,k} = 1$ and fix $v_1 = 1$;
initialize $\alpha_{11} = \alpha_{10} = \alpha_{01} = \alpha_{00} = 2$;
initialize $V_0 = V_0 \times \sqrt{\frac{\log p}{n}}$;
for $j = 1 : length(V_0)$ do
    for Repeat in 1:50 do
        Permute $(\boldsymbol{Y}_i, \boldsymbol{\delta}_i)$ to get $\left(\tilde{\boldsymbol{Y}}_i, \tilde{\boldsymbol{\delta}}_i\right)$;
        run the proposed algorithm on $\left(\boldsymbol{X}, \tilde{\boldsymbol{Y}}_i, \tilde{\boldsymbol{\delta}}_i\right)$ with $v_0 = V_{0j}$;
        get posterior probabilities;
    end
    Compute averaged posterior probabilities;
    if the first variable enters the null model then
        break;
        record $j^* = j$;
    end
end
Return $V_{0j^*-1}$;
Special case: If $j^* = 1$, take $V_{01}$.
```

### 4.3 Real data analysis

Multiple myeloma is a cancer of plasma cells which can be developed from monoclonal gammopathy. The abnormal plasma cells produce abnormal antibodies, which can cause kidney malfunctions or even form a mass in the bone marrow or soft tissue. The myeloma patients have been widely studied in clinical trails and have been observed to survive from a few months to more than 10 years after diagnosis. Gene expression profiling of myeloma plasma cells can be obtained to identify a gene signature associated with short survival in myeloma patients (Shaughnessy Jr et al. 2007). Understanding the cancer genomics and identifying risk groups with a high predictive power could also contribute to selecting patients for personalized therapy.

To examine the performance of the proposed method, we studied event-free survival and overall survival from newly diagnosed multiple myeloma patients enrolled in clinical trials UARK 98-026 and UARK 2003-33 (Zhan et al. 2006; Shaughnessy Jr et al. 2007). Two treatment regimes, total therapy II (TT2) and total therapy III (TT3), are compared in the clinical trails. In total there were 340 patients in TT2 with 191 events in event-free survival and 126 death in overall survival. The average follow-up time is 47.1 months for event-free survival and 55.8 months for overall survival. In TT3, among 214 patients there were 55 events for event-free survival and 43 death for overall survival. The average follow-up time in TT3 is 35.6 months for event-free survival and 37 months for overall survival. If the patient is censored for event-free survival, the patient is also censored for overall survival. Gene expression values of 54675 probesets were measured for each patient using Affymetrix U133Plus2.0 microarrays. The data was retrieved from the MicroArray Quality Control Consortium II GEO entry (GSE24080) (Shi et al. 2010).

We apply our proposed method along with competing methods suggested in the empirical studies to the bivariate survival data of the TT2 patients to select significant genes. Then by fitting a Buckley-James regression model to the TT2 patients with selected genes, we develop the risk scores for the TT3 patients and further estimate the C-statistics of those models on the TT3 patients (Uno et al. 2011). We first pre-process the data with the screening procedure proposed by Zhu et al. (2011). The procedure is able to provide sure screening for any single-index model including the AFT model. Since the screening procedure is designed for unviariate single-index model, we apply the procedure to marginal AFT models for $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$ following criteria suggested by Zhu et al. (2011) and Li et al. (2014). Using the combined soft- and hard- thresholding rule, we choose up to $n/\log(n)$ covariates for each event with a procedure involving randomly generated auxiliary variables. After combining retained variables for event-free survival and overall survival, in total 98 covariates are kept after pre-processing. Note that the Univariate AEnet selects an empty model therefore we skip reporting the result here.

Instead of performing heavy tuning on the proposed method, we adopt the knockoff framework proposed in Candès et al. (2018) with the recommended $v_0$ choice to select relevant variables. The knockoff framework creates a fixed fake feature matrix $\tilde{\boldsymbol{X}}$ of size $n$ by $p$ such that one is able to perform variable selection purely based on model results of $\left(\boldsymbol{Y}, \boldsymbol{X}, \tilde{\boldsymbol{X}}\right)$ without any advanced tuning procedures. The selection can be

**Table 7** Validation C-statistics on TT3 for univariate Coxnet with $\lambda_{1se}$ and $\lambda_{min}$, univariate BP-BJ, and the proposed bivariate BP-BJ

| Method | Model size of $\boldsymbol{\beta}_1$ | Model size of $\boldsymbol{\beta}_2$ | C-statistic of event-free survival | C-statistic of overall survival |
|---|---|---|---|---|
| Univariate Coxnet | 10 | 10 | 62.7% | 62.3% |
| Univariate BP-BJ | 10 | 10 | 59.7% | 54.7% |
| Bivariate BP-BJ | 3 | 3 | 62.2% | 63.8% |

**Table 8** Genes selected by the proposed method

| Probesets | Gene name |
|---|---|
| 225834_at | FAM72A /// FAM72B /// FAM72C /// FAM72D |
| 236641_at | KIF14 |
| 218672_at | SCNM1 /// TNFAIP8L2-SCNM1 |

made by observing the difference in model results, such as estimated coefficients for frequentist methods or posterior probabilities for Bayesian frameworks, between the real variables and the fake ones. Intuitively if the difference for a real variable is huge, then the variable is very likely to be relevant, otherwise the variable is very likely to be a noise variable. For the proposed method, we generate a fixed knockoff design matrix $\tilde{\boldsymbol{X}}$ following Barber et al. (2015) and apply the proposed method with the combined covariate matrix $\left(\boldsymbol{X}, \tilde{\boldsymbol{X}}\right)$. To measure the differences between the real variables and the fake ones, we first compute differences in all four posterior probabilities, then for each variable, we select the outcome that corresponds to the maximum difference. We consider a difference between the real and the fake variable as huge if the magnitude of the difference is above the 90-th percentile of all maximum differences. We apply the proposed method under the knockoff framework and evaluate the C-statistics of both event-free survival and overall survival for TT3 patients. The univariate BP-BJ results can be obtained in a similar manner. For Coxnet, we also avoid tuning for $\lambda$ and instead of comparing differences of coefficients magnitudes, we compare $\lambda_j$ and $\tilde{\lambda}_j$ which are values at which the $j$-th variable and its knockoff enter the model. Then we consider the following signed maximum statistic suggested by Barber et al. (2015)

$$W_j = \max\left(\lambda_j, \tilde{\lambda}_j\right) \operatorname{sign}\left(\lambda_j, \tilde{\lambda_j}\right).$$

If $w_j$ is above the 90-th percentile of all of the $w_j$'s, then the $j$-th variable is selected.

The real data analysis results are reported in Table 7. We see that both Coxnet and the proposed method have similar performances with validation C-statistics around 62% for both event-free survival and overall survival. Coxnet performs slightly better for event-free survival while the proposed method performs better for overall survival. Unfortunately the univariate BP-BJ algorithm gives the worst results but the its performance can be improved with fine tuning in other hyper-parameters. On the other hand, the proposed method is able to achieve the same performance with Coxnet with only 6 selected variables, demonstrating that the selections of the proposed method have significant effect in survival times of myeloma patients.

Table 8 reports the genes identified by the proposed method. In a recent study FAM72D is identified to be associated with cell proliferation in multiple myeloma (Chatonnet et al. 2020; Noll et al. 2015) while KIF14 has been recognized to be one of the signature genes related with survival for multiple myeloma patients in various analysis (Shaughnessy 2005; Shaughnessy Jr et al. 2007; Hawley et al. 2013).

## 5 Discussion

In this paper, we have studied the high-dimensional estimation procedure for bivariate AFT models by utilizing the Buckley-James estimator under the Bayesian framework. By applying a bivariate spike-and-slab prior, we proposed a variable selection method which minimizes the penalized $L_2$ loss function with penalty induced from the prior specification. Being inspired by the EM algorithm, we suggested an iterative process for computing the proposed method using an EM-like algorithm with a working Gaussian likelihood. The method has demonstrated the ability to be scalably applied to high-dimensional bivariate censored regression models and have shown outstanding performance compared with treating the event one at a time using empirical studies. We applied the proposed method to study the data from the multiple myeloma clinical trial, and showed that our method could achieve comparable validation C-statistics with less selected genes.

Many methods have been developed for multivariate survival data analysis under the Cox model. However, multivariate AFT models have received fewer attention due to the difficulty of estimating the joint survival distribution or handling unknown correlation structures. In the proposed method, we considered another approach and assumed that the two events are independent and the connections are carried and learnt in the prior knowledge of unknown coefficients. We have shown that, even the true survival times are correlated, the algorithm is able to capture the connection in the posterior distribution while keeping the imputation steps for censored observations simple to deal with. The design of the method also allows multiple structures of multivariate survival data to be handled.

We do realize that the design of the bivariate spike-and-slab prior can be hard to generalize to multivariate data with more than two event types. However, we also recognize the advantage of the proposed method, both in computation and in performance, which motivates us to look into more accessible designs of prior specifications to deal with high-dimensional multivariate survival data.

## A Multicollinearity design

Let sample size $n$ be 100 and dimension $p$ be 100. Let the first 10 variables be independently generated from standard normal distribution. Then for $j = 11, \cdots, 20$, consider

$$\boldsymbol{X}_j = \boldsymbol{X}_{j-10} + \tau,$$

where $\tau$ is a random error from a standard normal distribution. The rest of the variables are further generated from multivariate normal distribution with mean zero and covariance matrix with elements $\boldsymbol{\Sigma}_{ij} = 0.5^{|i-j|}$. Following Sect. 4.1, generate $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$ and corresponding censoring times and censoring indicators. Furthermore, we assume the relevant variables as the following

- no sharing: $\left\{j : \boldsymbol{\beta}_{j1} \neq 0\right\} = \{1, \cdots, 10\}, \left\{j : \boldsymbol{\beta}_{j2} \neq 0\right\} = \{21, \cdots, 30\}$
- all sharing: $\left\{j : \boldsymbol{\beta}_{jk} \neq 0\right\} = \{1, \cdots, 10\}$
- some sharing: $\left\{j : \boldsymbol{\beta}_{j1} \neq 0\right\} = \{1, \cdots, 10\}, \left\{j : \boldsymbol{\beta}_{j2} \neq 0\right\} = \{1, \cdots, 5, \cdots, 21, \cdots, 25\}$

All of the true relevant variables are generated independently from $\mathsf{N}(3, 0.5)$. We repeat all simulation setups for 200 times and fix the true coefficient values for all simulation runs.

The results of the multicollinearity design can be found in Tables 9 and 10. For this simulation design, the univariate AEnet failed to give any results due to the issue with singular matrix computation, therefore we only report results from the other four competing methods. We see that all of the methods tend to recognize the ten irrelevant variables as signals. For no-sharing and all-sharing cases, the proposed method is able to give the smallest number of false positives while being able to recognize almost all of the relevant variables, giving almost zero false negatives. For some-sharing cases, we observe more obvious trade-off between false positives and false negatives for using $\lambda_{min}$ and $\lambda_{1se}$ while the proposed method selects the variables more strictly, returning with lower false positive scores and higher false negative scores. However, in terms of MCC score as an overall measure, the proposed method is able to achieve the highest MCC scores for all setups, demonstrating that the proposed method is able to outperform existing methods and to handle complicated data examples.

## B Dense design

Let $n = 100$ and $p = 100$. Following Sect. 4.1, we generate design matrix $\boldsymbol{X}$ from multivariate normal distribution with mean zero and covariance matrix with elements $\boldsymbol{\Sigma}_{ij} = 0.5^{|i-j|}$. Then we generate $\boldsymbol{T}_{\cdot 1}$ and $\boldsymbol{T}_{\cdot 2}$ and corresponding censoring times and censoring indicators in a similar manner. In this simulation design, we assume that for each column of the true coefficient matrix, there are 20 relevant variables. That is, for some-sharing setups, we will have in total 45 relevant variables. All of the true relevant variables are generated independently from $\mathsf{N}(3, 0.5)$. We repeat all simulation setups for 200 times and fix the true coefficient values for all simulation runs.

The results of the dense design can be found in Tables 11 and 12. We see that the proposed method gives consistent performance to have the best MCC scores among all competing methods. For no-sharing and all-sharing setups, the proposed method is able to give the best combination of false positives and false negatives, achieving highest sensitivity and specificity scores. For some-sharing setups, when $c \neq 1$, the

**Table 9** False positives and false negatives reported for multicollinearity design

| sharing type | $c$ | dist. | Univariate Coxnet with $\lambda_{1se}$ | | Univariate Coxnet with $\lambda_{min}$ | | Univariate BP-BJ | | Bivariate BP-BJ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | FP | FN | FP | FN | FP | FN | FP | FN |
| no | 1 | exp. | 24.3 | 0.0 | 32.4 | 0.0 | 21.2 | 0.6 | 10.8 | 0.0 |
| | | norm. | 22.6 | 0.0 | 29.3 | 0.0 | 17.2 | 1.8 | 10.0 | 0.4 |
| all | 1 | exp. | 44.6 | 0.0 | 57.1 | 0.0 | 24.6 | 1.5 | 20.4 | 0.0 |
| | | norm. | 41.7 | 0.0 | 51.4 | 0.0 | 22.7 | 2.7 | 20.2 | 0.0 |
| some | 0.3 | exp. | 32.3 | 5.6 | 49.1 | 2.8 | 27.8 | 8.8 | 24.0 | 4.1 |
| | | norm. | 31.5 | 5.7 | 46.5 | 2.8 | 26.4 | 8.7 | 23.7 | 4.1 |
| | 0.5 | exp. | 31.3 | 6.2 | 47.8 | 2.4 | 30.5 | 7.8 | 26.0 | 3.0 |
| | | norm. | 31.2 | 5.8 | 46.4 | 2.8 | 27.2 | 8.0 | 24.2 | 3.3 |
| | 0.7 | exp. | 32.9 | 5.5 | 50.4 | 2.3 | 25.2 | 8.0 | 24.0 | 3.4 |
| | | norm. | 31.1 | 6.0 | 46.5 | 2.4 | 27.8 | 7.6 | 24.3 | 3.0 |
| | 1 | exp. | 33.3 | 0.0 | 43.8 | 0.0 | 23.0 | 0.7 | 15.3 | 0.0 |
| | | norm. | 31.3 | 0.0 | 39.3 | 0.0 | 19.8 | 2.1 | 14.8 | 0.4 |

**Table 10** Sensitivity, specificity, and MCC scores reported for multicollinearity design

| sharing type | $c$ | dist. | Univariate Coxnet with $\lambda_{1se}$ | | | Univariate Coxnet with $\lambda_{min}$ | | | Univariate BP-BJ | | | Bivariate BP-BJ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | sen. | spec. | MCC | sens. | spec. | MCC | sen. | spec. | MCC | sen. | spec. | MCC |
| no | 1 | exp. | 1.0 | 0.87 | 0.63 | 1.0 | 0.82 | 0.56 | 0.97 | 0.88 | 0.64 | 1.00 | 0.94 | **0.78** |
| | | norm. | 1.0 | 0.87 | 0.64 | 1.0 | 0.84 | 0.59 | 0.91 | 0.90 | 0.65 | 0.98 | 0.94 | **0.78** |
| all | 1 | exp. | 1.0 | 0.75 | 0.48 | 1.0 | 0.68 | 0.42 | 0.92 | 0.86 | 0.58 | 1.00 | 0.89 | **0.66** |
| | | norm. | 1.0 | 0.77 | 0.50 | 1.0 | 0.71 | 0.45 | 0.86 | 0.87 | 0.56 | 1.00 | 0.89 | **0.67** |
| some | 0.3 | exp. | 0.78 | 0.82 | 0.45 | 0.89 | 0.72 | 0.42 | 0.65 | 0.84 | 0.40 | 0.84 | 0.86 | **0.55** |
| | | norm. | 0.77 | 0.82 | 0.45 | 0.89 | 0.73 | 0.44 | 0.65 | 0.85 | 0.41 | 0.84 | 0.86 | **0.56** |
| | 0.5 | exp. | 0.75 | 0.82 | 0.44 | 0.90 | 0.73 | 0.44 | 0.69 | 0.83 | 0.40 | 0.88 | 0.85 | **0.57** |
| | | norm. | 0.77 | 0.82 | 0.45 | 0.89 | 0.73 | 0.44 | 0.68 | 0.84 | 0.42 | 0.87 | 0.86 | **0.57** |
| | 0.7 | exp | 0.78 | 0.81 | 0.45 | 0.91 | 0.71 | 0.43 | 0.68 | 0.86 | 0.44 | 0.87 | 0.86 | **0.57** |
| | | norm. | 0.76 | 0.82 | 0.45 | 0.91 | 0.73 | 0.45 | 0.70 | 0.84 | 0.43 | 0.88 | 0.86 | **0.58** |
| | 1 | exp | 1.0 | 0.82 | 0.56 | 1.0 | 0.76 | 0.49 | 0.97 | 0.87 | 0.62 | 1.00 | 0.91 | **0.72** |
| | | norm. | 1.0 | 0.83 | 0.57 | 1.0 | 0.78 | 0.51 | 0.90 | 0.89 | 0.61 | 0.98 | 0.92 | **0.71** |

The best MCC score is bolded for each simulation design among all competing methods

**Table 11** False positives and false negatives reported for dense design

| sharing type | $c$ | dist. | Univariate Coxnet with $\lambda_{1se}$ | | Univariate Coxnet with $\lambda_{min}$ | | Univariate AEnet | | Univariate BP-BJ | | Bivariate BP-BJ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | FP | FN | FP | FN | FP | FN | FP | FN | FP | FN |
| no | 1 | exp. | 26.0 | 1.9 | 34.6 | 0.5 | 5.3 | 33.1 | 39.1 | 2.9 | 8.8 | 1.3 |
| | | norm. | 26.5 | 2.5 | 34.9 | 0.7 | 5.3 | 33.5 | 39.7 | 3.5 | 8.6 | 1.4 |
| all | 1 | exp. | 25.8 | 3.1 | 36.4 | 0.7 | 5.4 | 33.2 | 39.2 | 3.5 | 9.2 | 0.9 |
| | | norm. | 25.9 | 1.5 | 33.5 | 0.3 | 5.5 | 33.1 | 40.9 | 3.1 | 8.5 | 0.6 |
| some | 0.3 | exp. | 17.3 | 27.5 | 28.5 | 18.2 | 4.3 | 49.8 | 36.2 | 17.1 | 16.4 | 19.0 |
| | | norm. | 16.6 | 28.6 | 27.7 | 19.1 | 4.7 | 49.6 | 32.8 | 17.9 | 14.8 | 19.7 |
| | 0.5 | exp. | 16.5 | 29.1 | 27.9 | 18.0 | 4.7 | 49.6 | 34.9 | 18.1 | 16.1 | 19.7 |
| | | norm. | 16.7 | 28.6 | 26.9 | 18.6 | 4.3 | 49.7 | 32.1 | 16.8 | 14.6 | 19.3 |
| | 0.7 | exp. | 15.4 | 27.8 | 26.0 | 18.5 | 3.9 | 49.2 | 31.5 | 17.8 | 14.8 | 20.0 |
| | | norm. | 17.4 | 28.3 | 29.2 | 17.9 | 5.1 | 50.0 | 33.1 | 16.2 | 14.5 | 18.2 |
| | 1 | exp. | 25.4 | 1.4 | 33.6 | 0.2 | 5.5 | 32.8 | 43.9 | 2.4 | 10.7 | 1.1 |
| | | norm. | 25.8 | 1.7 | 33.5 | 0.3 | 5.5 | 32.6 | 38.9 | 2.8 | 7.9 | 1.0 |

**Table 12** Sensitivity, specificity, and MCC scores reported for dense design

| sharing type | $c$ | dist. | Univariate Coxnet with $\lambda_{1se}$ | | | Univariate Coxnet with $\lambda_{min}$ | | | Univariate AEnet | | | Univariate BP-BJ | | | Bivariate BP-BJ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | sen. | spec. | MCC | sens. | spec. | MCC | sen. | spec. | MCC | sen. | spec. | MCC | sen. | spec. | MCC |
| no | 1 | exp. | 0.95 | 0.84 | 0.68 | 0.99 | 0.78 | 0.64 | 0.17 | 0.97 | 0.23 | 0.93 | 0.76 | 0.56 | 0.97 | 0.94 | **0.86** |
| | | norm. | 0.94 | 0.83 | 0.66 | 0.98 | 0.78 | 0.64 | 0.16 | 0.97 | 0.22 | 0.91 | 0.75 | 0.55 | 0.97 | 0.95 | **0.86** |
| all | 1 | exp. | 0.92 | 0.84 | 0.66 | 0.98 | 0.77 | 0.63 | 0.17 | 0.97 | 0.23 | 0.91 | 0.76 | 0.55 | 0.98 | 0.94 | **0.86** |
| | | norm. | 0.96 | 0.84 | 0.69 | 0.99 | 0.79 | 0.65 | 0.17 | 0.97 | 0.22 | 0.92 | 0.74 | 0.55 | 0.98 | 0.95 | **0.88** |
| some | 0.3 | exp. | 0.50 | 0.88 | 0.41 | 0.67 | 0.80 | 0.45 | 0.09 | 0.97 | 0.13 | 0.69 | 0.75 | 0.41 | 0.65 | 0.89 | **0.55** |
| | | norm. | 0.48 | 0.89 | 0.40 | 0.65 | 0.81 | 0.44 | 0.10 | 0.97 | 0.13 | 0.67 | 0.77 | 0.42 | 0.64 | 0.90 | **0.56** |
| | 0.5 | exp. | 0.47 | 0.89 | 0.39 | 0.66 | 0.81 | 0.44 | 0.10 | 0.97 | 0.13 | 0.67 | 0.76 | 0.40 | 0.64 | 0.89 | **0.54** |
| | | norm. | 0.48 | 0.89 | 0.40 | 0.66 | 0.81 | 0.46 | 0.10 | 0.97 | 0.14 | 0.69 | 0.78 | 0.44 | 0.65 | 0.90 | **0.57** |
| | 0.7 | exp | 0.49 | 0.89 | 0.42 | 0.66 | 0.82 | 0.47 | 0.10 | 0.97 | 0.16 | 0.68 | 0.78 | 0.43 | 0.64 | 0.90 | **0.55** |
| | | norm. | 0.49 | 0.88 | 0.39 | 0.67 | 0.80 | 0.45 | 0.09 | 0.97 | 0.11 | 0.70 | 0.77 | 0.45 | 0.67 | 0.90 | **0.58** |
| | 1 | exp | 0.97 | 0.84 | 0.69 | 0.99 | 0.79 | 0.65 | 0.18 | 0.97 | 0.23 | 0.94 | 0.73 | 0.54 | 0.97 | 0.93 | **0.84** |
| | | norm. | 0.96 | 0.84 | 0.68 | 0.99 | 0.79 | 0.65 | 0.18 | 0.97 | 0.24 | 0.93 | 0.76 | 0.57 | 0.98 | 0.95 | **0.88** |

The best MCC score is bolded for each simulation design among all competing methods

proposed method is more strict in selecting signals which results in missing almost half of the relevant variables. However the proposed method is still able to correctly identify more relevant variables and noise variables compared with other competing methods, achieving the highest MCC scores.

## References

Ahmed SE, Hossain S, Doksum KA (2012) Lasso and shrinkage estimation in weibull censored regression models. J Stat Plan Inference 142(6):1273–1284

Barber RF, Candès EJ et al (2015) Controlling the false discovery rate via knockoffs. Annal Stat 43(5):2055–2085

Barbieri MM, Berger JO (2004) Optimal predictive model selection. Annal Stat 32(3):870–897

Buckley J, James I (1979) Linear regression with censored data. Biometrika 66(3):429–436

Cai T, Huang J, Tian L (2009) Regularized estimation for the accelerated failure time model. Biometrics 65(2):394–404

Candès E, Fan Y, Janson L, Lv J (2018) Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. J R Stat Soc: Ser B (Stat Methodol) 80(3):551–577

Chang SH (2004) Estimating marginal effects in accelerated failure time models for serial sojourn times among repeated events. Lifetime Data Anal 10(2):175–190

Chatonnet F, Pignarre A, Sérandour AA, Caron G, Avner S, Robert N, Kassambara A, Laurent A, Bizot M, Agirre X et al (2020) The hydroxymethylome of multiple myeloma identifies fam72d as a 1q21 marker linked to proliferation. Haematologica 105(3):774–783

Chiou SH, Kang S, Kim J, Yan J (2014) Marginal semiparametric multivariate accelerated failure time model with generalized estimating equations. Lifetime Data Anal 20(4):599–618

Cox DR (1972) Regression models and life-tables. J R Stat Soc: Ser B (Methodol) 34(2):187–202

Duan W, Zhang R, Zhao Y, Shen S, Wei Y, Chen F, Christiani DC (2018) Bayesian variable selection for parametric survival model with applications to cancer omics data. Human Genom 12(1):49

George EI, McCulloch RE (1997) Approaches for bayesian variable selection. Stat Sinica 7(2):339–373

Hanagal DD (2006) Bivariate weibull regression model based on censored samples. Stat Papers 47(1):137–147

Hawley TS, Riz I, Yang W, Wakabayashi Y, DePalma L, Chang YT, Peng W, Zhu J, Hawley RG (2013) Identification of an abcb1 (p-glycoprotein)-positive carfilzomib-resistant myeloma subpopulation by the pluripotent stem cell fluorescent dye cdy1. Am J Hematol 88(4):265–272

He W, Lawless JF (2005) Bivariate location-scale models for regression analysis, with applications to lifetime data. J R Stat Soc: Ser B (Stat Methodol) 67(1):63–78

Hornsteiner U, Hamerle A (1996) A combined gee/buckley-james method for estimating an accelerated failure time model of multivariate failure times. Discussion Paper 47, Ludwig-Maximillians Universitat, Munchen. Also available from http://stat.unimuenchen.de/sfb386/publikation.html

Hu J, Chai H (2013) Adjusted regularized estimation in the accelerated failure time model with high dimensional covariates. J Multiv Anal 122:96–114

Huang J, Ma S (2010) Variable selection in the accelerated failure time model via the bridge method. Lifetime Data Anal 16(2):176–195

Huang J, Ma S, Xie H (2006) Regularized estimation in the accelerated failure time model with high-dimensional covariates. Biometrics 62(3):813–820

Huang J, Ma S, Xie H (2007) Least absolute deviations estimation for the accelerated failure time model. Stat Sinica 17(4):1533–1548

Huang J, Ma S, Xie H, Zhang CH (2009) A group bridge approach for variable selection. Biometrika 96(2):339–355

Huang L, Kopciuk K, Lu X (2020) Adaptive group bridge selection in the semiparametric accelerated failure time model. J Multiv Anal 175:104562

Huang Y (2002) Censored regression with the multistate accelerated sojourn times model. J R Stat Soc: Ser B (Stat Methodol) 64(1):17–29

Jin Z, Lin D, Wei L, Ying Z (2003) Rank-based inference for the accelerated failure time model. Biometrika 90(2):341–353

Jin Z, Lin D, Ying Z (2006) On least-squares regression with censored data. Biometrika 93(1):147–161

Jin Z, Lin D, Ying Z (2006) Rank regression analysis of multivariate failure time data based on marginal linear models. Scandinavian J Stat 33(1):1–23

Johnson BA et al (2009) On lasso for censored data. Electron J Stat 3:485–506

Kalbfleisch JD, Prentice RL (2011) The statistical analysis of failure time data. Wiley, New Jersey

Khan MHR, Shaw JEH (2016) Variable selection for survival data with a class of adaptive elastic net techniques. Stat Comput 26(3):725–741

Khan MHR, Shaw JEH (2019) Variable selection for accelerated lifetime models with synthesized estimation techniques. Stat Methods Med Res 28(3):937–952

Khan MHR, Bhadra A, Howlader T (2019) Stability selection for lasso, ridge and elastic net implemented with aft models. Stat Appl Genet Mol Biol 18(5):742

Konrath S, Fahrmeir L, Kneib T (2015) Bayesian accelerated failure time models based on penalized mixtures of gaussians: regularization and variable selection. AStA Adv Stat Anal 99(3):259–280

Koul H, Vv Susarla, Van Ryzin J et al (1981) Regression analysis with randomly right-censored data. Annal Stat 9(6):1276–1288

Lee KE, Mallick BK (2004) Bayesian methods for variable selection in survival models with application to dna microarray data. Sankhyā: Ind J Stat 66(4):756–778

Lee KH, Chakraborty S, Sun J (2017) Variable selection for high-dimensional genomic data with censored outcomes using group lasso prior. Comput Stat Data Anal 112:1–13

Li H, Yin G (2009) Generalized method of moments estimation for linear regression with clustered failure time data. Biometrika 96(2):293–306

Li Y, Dicker L, Zhao SD (2014) The dantzig selector for censored linear regression models. Stat Sinica 24(1):251

Lu W (2007) Tests of independence for censored bivariate failure time data. Lifetime Data Anal 13(1):75–90

Miller RG (1976) Least squares regression with censored data. Biometrika 63(3):449–464

Mitchell TJ, Beauchamp JJ (1988) Bayesian variable selection in linear regression. J Am Stat Assoc 83(404):1023–1032

Narisetty NN, He X et al (2014) Bayesian variable selection with shrinking and diffusing priors. Annal Stat 42(2):789–817

Noll JE, Vandyke K, Hewett DR, Mrozik KM, Bala RJ, Williams SA, Kok CH, Zannettino AC (2015) Pttg1 expression is associated with hyperproliferative disease and poor prognosis in multiple myeloma. J Hematol Oncol 8(1):106

Pan W, Kooperberg C (1999) Linear regression for bivariate censored data via multiple imputation. Stat Med 18(22):3111–3121

Pan W, Louis TA (2000) A linear mixed-effects model for multivariate censored data. Biometrics 56(1):160–166

Park T, Casella G (2008) The bayesian lasso. J Am Stat Assoc 103(482):681–686

Ročková V, George EI (2014) Emvs: the em approach to bayesian variable selection. J Am Stat Assoc 109(506):828–846

Sabourin JA, Valdar W, Nobel AB (2015) A permutation approach for selecting the penalty parameter in penalized model selection. Biometrics 71(4):1185–1194

Schneider H, Weissfeld L (1986) Estimation in linear models with censored data. Biometrika 73(3):741–745

Sha N, Tadesse MG, Vannucci M (2006) Bayesian variable selection for the analysis of microarray data with censored outcomes. Bioinformatics 22(18):2262–2268

Shaughnessy J (2005) Amplification and overexpression of cks1b at chromosome band 1q21 is associated with reduced levels of p27 kip1 and an aggressive clinical course in multiple myeloma. Hematology 10:117–126

Shaughnessy JD Jr, Zhan F, Burington BE, Huang Y, Colla S, Hanamura I, Stewart JP, Kordsmeier B, Randolph C, Williams DR et al (2007) A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. Blood 109(6):2276–2284

Shi L, Campbell G, Jones W, Campagne F, Wen Z, Walker S, Su Z, Chu T, Goodsaid F, Pusztai L et al (2010) The maqc-ii project: a comprehensive study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnol 28:827–838

Stute W, Wang JL (1993) The strong law under random censorship. Annal Stat 36:1591–1607

Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82(398):528–540

Tibshirani R (1997) The lasso method for variable selection in the cox model. Stat Med 16(4):385–395

Tsiatis AA (1990) Estimating regression parameters using linear rank tests for censored data. Annal Stat 90:354–372

Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei L (2011) On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med 30(10):1105–1117

Van Erp S, Oberski DL, Mulder J (2019) Shrinkage priors for bayesian penalized regression. J Math Psychol 89:31–50

Visser M (1996) Nonparametric estimation of the bivariate survival function with an application to vertically transmitted aids. Biometrika 83(3):507–518

Wang S, Nan B, Zhu J, Beer DG (2008) Doubly penalized buckley-james method for survival data with high-dimensional covariates. Biometrics 64(1):132–140

Wang X, Song L (2011) Adaptive lasso variable selection for the accelerated failure models. Commun Stat-Theory Methods 40(24):4372–4386

Wang YG, Fu L (2011) Rank regression for accelerated failure time model with clustered and censored data. Comput Stat Data Anal 55(7):2334–2343

Wei LJ (1992) The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. Stat Med 11(14–15):1871–1879

Wei LJ, Ying Z, Lin D (1990) Linear regression analysis of censored survival data based on rank tests. Biometrika 77(4):845–851

Xu J, Leng C, Ying Z (2010) Rank-based variable selection with censored data. Stat Comput 20(2):165–176

Yi GY, He W (2006) Methods for bivariate survival data with mismeasured covariates under an accelerated failure time model. Commun Stat-Theory Methods 35(8):1539–1554

Yin G, Cai J (2005) Quantile regression models with multivariate failure time data. Biometrics 61(1):151–161

Zhan F, Huang Y, Colla S, Stewart JP, Hanamura I, Gupta S, Epstein J, Yaccoby S, Sawyer J, Burington B et al (2006) The molecular classification of multiple myeloma. Blood 108(6):2020–2028

Zhu LP, Li L, Li R, Zhu LX (2011) Model-free feature screening for ultrahigh-dimensional data. J Am Stat Assoc 106(496):1464–1475