Bayesian variable selection and estimation in semiparametric joint models of multivariate longitudinal and survival data

An-Min Tang[1], **Xingqiu Zhao**[2,3], and **Nian-Sheng Tang***[,1]

[1] Department of Statistics, Yunnan University, Kunming 650091, China
[2] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong
[3] Shenzhen Research Institute, Hong Kong Polytechnic University, Shenzhen 518057, China

Received 2 May 2015; revised 1 February 2016; accepted 16 February 2016

This paper presents a novel semiparametric joint model for multivariate longitudinal and survival data (SJMLS) by relaxing the normality assumption of the longitudinal outcomes, leaving the baseline hazard functions unspecified and allowing the history of the longitudinal response having an effect on the risk of dropout. Using Bayesian penalized splines to approximate the unspecified baseline hazard function and combining the Gibbs sampler and the Metropolis–Hastings algorithm, we propose a Bayesian Lasso (BLasso) method to simultaneously estimate unknown parameters and select important covariates in SJMLS. Simulation studies are conducted to investigate the finite sample performance of the proposed techniques. An example from the International Breast Cancer Study Group (IBCSG) is used to illustrate the proposed methodologies.

Keywords: Bayesian Lasso; Bayesian penalized splines; Joint models; Mixture of normals; Survival analysis.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

Joint models of longitudinal and survival data (JMLSs) represent a flexible class of models for describing the interrelationships among longitudinal variables and survival variables, and they are widely applied to cancer and HIV/AIDS clinical studies; see, for example, Chi and Ibrahim (2006), De Gruttola and Tu (1994), Hu et al. (2009), Rizopoulos et al. (2009), Song and Wang (2008), Tsiatis and Davidian (2004), Tsiatis et al. (1995), Wang and Taylor (2001), Zhu et al. (2012), and references cited therein. Unlike the two-stage model for longitudinal and survival data proposed by Tsiatis et al. (1995), a JMLS consists of a longitudinal submodel and a survival submodel (Ibrahim et al., 2002, 2010), which share common random effects for capturing the individual characteristics. The longitudinal submodel is used to account for the association among the longitudinal responses and the related covariates, while the survival submodel is employed to investigate the relationship among the event time, the longitudinal processes, and the time-independent covariates.

Basic JMLSs have been widely studied under the normality assumption of longitudinal responses and the shared parameter model, in which the longitudinal outcome and the time to event share a latent Gaussian random effect, due to mathematical tractability and computational convenience. But, when the normality assumption is violated, the existing approaches to analyze basic JMLSs may lead to unreasonable or even misleading conclusions (Rizopoulos and Ghosh, 2011; Li et al., 2012; Baghfalaki

*Corresponding author: e-mail: nstang@ynu.edu.cn, Phone: +86-871-65032416, Fax: +86-871-65033700

et al., 2013). To this end, some alternative methods for analyzing JMLSs have been proposed in recent years. For example, Huang et al. (2010, 2014) proposed a relatively robust estimation approach to a univariate JMLS with longitudinal responses following the skew distribution; Baghfalaki et al. (2013, 2014) presented a robust inference on JMLSs under the assumption that the longitudinal responses are normally distributed and the time to event shares a common Gaussian random effect. However, the above-mentioned approaches are not flexible enough to capture the feature of longitudinal responses having bimodal or multimodal distributions in a JMLS. Another important limitation of the above-mentioned methods is that they did not consider longitudinal information, which, if appropriately used, could offer a better insight into the dynamics of the disease's progression (Rizopoulos et al., 2014). Also, an approach to accommodate the above-mentioned issue has not been studied in the JMLS literature, although we often encounter bimodal or multimodal data in longitudinal studies. Hence, to relax the normality assumption of the longitudinal outcomes and allow for the effect of the history of the longitudinal response, this article proposes a novel semiparametric JMLS (SJMLS) for multivariate longitudinal and survival data by assuming that the longitudinal responses are distributed as a finite mixture of normal distributions, the baseline hazard functions are unknown, and the history of longitudinal response up to the current time, which is defined by the current expectation of longitudinal response, may have an effect on the risk of dropout.

Generally, the piecewise constant hazard model could be employed to specify the prior distribution of the unknown baseline hazard (Zhu et al., 2012; Huang et al., 2014; Tang et al., 2014). But it might lead to a nonsmooth survival function, especially when the time axis is divided into a small number of intervals. This feature might not be desirable in some applications, in which a smooth but still flexible enough baseline hazard function should be postulated. To address the issue, this article uses the well-known Bayesian penalized splines (Lang and Brezger, 2004) to approximate the unknown log baseline hazard functions in the considered SJMLS.

In addition, covariate selection is another issue to be addressed in a SJMLS. Traditionally, the important covariates in a regression model can be identified by the forward selection method, backward elimination method, stepwise selection method (Hocking, 1976), or model comparison via Bayes factor (Kass and Raftery, 1995; Lee and Tang, 2006) or some information criterion such as the Akaike information criterion (Akaike, 1974), but these approaches are computationally expensive and unstable for the complicated models with a large number of covariates. As an alternative, some penalized likelihood methods have been proposed for simultaneous variable selection and parameter estimation in multiple linear regression. Notable methods include the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), the adaptive Lasso (Zou, 2006), and boosting algorithm (Buhlmann and Hothorn, 2007), which has received considerable attention in various regression frameworks (Buhlmann and Yu, 2003; Buhlmann, 2006; Hofner et al., 2011). In particular, in a Bayesian framework, Park and Casella (2008) proposed a BLasso by imposing the double exponential prior on the regression coefficients and the gamma distribution on the shrinkage parameter. The BLasso approach has been extended to various models including semiparametric structural equation models (Guo et al., 2012) and linear regression models (Hans, 2009; Lykou and Ntzoufras, 2013), due to its stability and computationally efficiency. However, to our knowledge, there was little literature yet that addressed covariate selection in SJMLSs via the BLasso approach. Hence, the second main purpose of this article is to extend BLasso approach to the considered SJMLSs.

This research was motivated by a clinical trial from the International Breast Cancer Study Group (IBCSG), which is devoted to an innovative clinical cancer study for improving the outcome of women with breast cancer. In this trial, each premenopausal woman with a node-positive breast cancer was randomly assigned to either the adjuvant chemotherapy or the reintroduction of three single courses of delayed chemotherapy. In addition to the adjuvant treatment effects, patients' quality of life (QOL) was assumed to have prognostic information and to be predictive of breast cancer progression. Cancer progression was monitored over time via two failure time random variables: disease-free survival (DFS), which is defined as the time duration of staying free of disease after a particular treatment for

a patient suffering from a cancer, and overall survival (OS), which is defined as the time duration of staying alive for a patient suffering from a cancer. In this study, the median of DFS is 7.611 years with a censoring proportion of 46.39%, while the median of OS is 9.255 years with a censoring proportion of 63.10%. Therapeutic method has a direct effect on DFS and OS, and the toxicity of therapeutic method may adversely affect a patient's QOL, which is specifically related to DFS and OS. Four indicators of health-related QOL including physical well-being (lousy-good), mood (miserable-happy), appetite (none-good), and perceived coping ('how much effort does it cost you to cope with your illness?" (a great deal-none)) were measured at the baseline and at months 3 and 18 after randomization for each of 832 patients. There were a total of 2154 QOL observations in the data set. Chi and Ibrahim (2006), Zhu et al. (2012), and Tang et al. (2014) analyzed the data set via various parametric/semiparametric JMLSs. However, they did not consider the selection of the potentially important covariates including therapy designs and individual characteristics. To this end, a BLasso approach is developed to simultaneously estimate unknown parameters and identify the significant effect of the potentially covariates on QOL, DFS, and OS in a framework of SJMLS.

The rest of this article is organized as follows. In Section 2, we describe a SJMLS with longitudinal outcomes following a finite mixture of normal distributions. Section 3 proposes a Bayesian Lasso (BLasso) approach to identify the important covariates in a SJMLS. Simulation studies are conducted to investigate the performance of the proposed methods in Section 4. An example is analyzed in Section 5. Some concluding remarks are given in Section 6. Technical details are presented in all appendices.

2 A SJMLS

2.1 Model and notation

Consider a data set from n individuals. For the i-th individual ($i = 1, \ldots, n$), let y_{ijk} be the k-th longitudinal outcome observed at time t_{ij} for $j = 1, \ldots, n_i$ and $k = 1, \ldots, K$; let T^*_{im} be the true survival time of the m-th time-to-event outcome, C_{im} the censoring time, and $T_{im} = \min(T^*_{im}, C_{im})$ the corresponding observed event time. Also, denote $\delta_{im} = 1(T^*_{im} \leq C_{im})$ as the event indicator for $i = 1, \ldots, n$, $m = 1, \ldots, M$, where $1(A)$ is the indicator function of an event A.

Denote $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijK})^{\mathrm{T}}$, $\boldsymbol{T}_i = (T_{i1}, \ldots, T_{iM})^{\mathrm{T}}$, and $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{iM})^{\mathrm{T}}$. Let $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})^{\mathrm{T}}$ be time-independent random effects underlying both the longitudinal and survival processes for the i-th individual. Given $\boldsymbol{b}_i$, it is assumed that $\boldsymbol{y}_{ij}$'s are conditionally independent of each other. Under the above assumptions, we consider the following linear model for longitudinal response vector $\boldsymbol{y}_{ij}$:

$$\boldsymbol{y}_{ij} = \boldsymbol{\eta}(\boldsymbol{R}_i(t_{ij}), W_i(t_{ij}), \boldsymbol{b}_i) + \boldsymbol{\varepsilon}_{ij}, \tag{1}$$

where $\boldsymbol{\eta}(\boldsymbol{R}_i(t_{ij}), W_i(t_{ij}), \boldsymbol{b}_i) = \beta^{\mathrm{T}}\boldsymbol{R}_i(t_{ij}) + W_i(t_{ij})\boldsymbol{b}_i$ is the trajectory function vector of longitudinal response vector $\boldsymbol{y}_{ij}$ for the i-th individual at time t_{ij}, $\boldsymbol{R}_i(t_{ij})$ is an $(r+1) \times 1$ time-dependent design vector at time point t_{ij} whose first element is set to be 1 for allowing a more convenient formulation of the model, β is an $(r+1) \times K$ unknown parameter matrix with the k-th column being $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kr})^{\mathrm{T}}$ for $k = 1, \ldots, K$, $W_i(t_{ij})$ is a $K \times q$ design matrix corresponding to the random effects $\boldsymbol{b}_i$, and $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \ldots, \varepsilon_{ijK})^{\mathrm{T}}$ is a $K \times 1$ vector of measurement errors whose distribution is assumed to follow a finite mixture of normal distributions rather than a classical normal distribution, which is specified in Section 2.2. Similar to a common assumption for the random effects $\boldsymbol{b}_i$ in a mixed-effects model, it is assumed that $\boldsymbol{b}_i$ is independent and identically distributed as a multivariate normal distribution with zero mean and covariance matrix $\Omega = (\Omega_{jk})_{q\times q}$, that is, $\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} N_q(\boldsymbol{0}, \Omega)$. Also, we assume that $\boldsymbol{\varepsilon}_{ij}$'s are independent of $\boldsymbol{b}_i$.

To incorporate the history information of longitudinal response up to current time and time-independent covariates $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{ip})^{\mathrm{T}}$, we consider an M-dimensional survival model for the i-th individual under the assumption that all components of the time-to-event outcomes are independent.

Let $\lambda_m(t|\boldsymbol{b}_i)$ be the conditional hazard function of the m-th time-to-event outcome given $\boldsymbol{b}_i$ for the i-th individual, which is defined as

$$\lambda_m(t|\boldsymbol{b}_i) = \lambda_{m0}(t)\exp\left\{\boldsymbol{\psi}_m^{\mathrm{T}}\boldsymbol{\eta}(\boldsymbol{R}_i(t), W_i(t), \boldsymbol{b}_i) + \boldsymbol{\gamma}_m^{\mathrm{T}}\boldsymbol{\xi}_i\right\} \quad \text{for} \quad t > 0, \tag{2}$$

where $\boldsymbol{\psi}_m = (\psi_{m1}, \ldots, \psi_{mK})^{\mathrm{T}}$ quantifies the association between the true value of the longitudinal trajectories at time t and the hazard of an event at the same time point, $\boldsymbol{\gamma}_m = (\gamma_{m1}, \ldots, \gamma_{mp})^{\mathrm{T}}$ is a vector of regression coefficients corresponding to covariate vector $\boldsymbol{\xi}_i$, and $\lambda_{m0}(t)$ is an unknown baseline hazard function. Because $\lambda_{m0}(t)$ is nonnegative, it can be written as $\lambda_{m0}(t) = \exp\{\lambda^*_{m0}(t)\}$, which implies that equation (2) can be rewritten as

$$\lambda_m(t|\boldsymbol{b}_i) = \exp\left\{\lambda^*_{m0}(t) + \boldsymbol{\psi}_m^{\mathrm{T}}\boldsymbol{\eta}(\boldsymbol{R}_i(t), W_i(t), \boldsymbol{b}_i) + \boldsymbol{\gamma}_m^{\mathrm{T}}\boldsymbol{\xi}_i\right\}, \tag{3}$$

where $\lambda^*_{m0}(t)$ is referred to as the log baseline hazard function. Then, for the i-th individual, the conditional probability density function of $(\boldsymbol{T}_i, \boldsymbol{\delta}_i)$ given $\boldsymbol{b}_i$ is given by

$$\Pr(\boldsymbol{T}_i, \boldsymbol{\delta}_i|\boldsymbol{b}_i) = \prod_{m=1}^{M} S_m(T_{im}|\boldsymbol{b}_i)\{\lambda_m(T_{im}|\boldsymbol{b}_i)\}^{\delta_{im}}, \tag{4}$$

where $S_m(t|\boldsymbol{b}_i) = \exp\{-\int_0^t \lambda_m(u|\boldsymbol{b}_i)du\}$ is the m-th conditional survival function.

2.2 Specifying the distribution of measurement error

In classical longitudinal data models, it is usually assumed that measurement error vector $\boldsymbol{\varepsilon}_{ij}$ follows a multivariate normal distribution, which may be questionable in practice. Moreover, the violation of the basic assumption would lead to biased estimates of parameters or even misleading conclusions. To this end, it is desirable to develop an approach to relax the basic normality assumption. Similar to Escobar and West (1995) and Müller et al. (1996), here we assume that $\boldsymbol{\varepsilon}_{ij}$ follows the following finite mixture of normal distributions: $\boldsymbol{\varepsilon}_{ij} \sim \sum_{g=1}^{G} \pi_g N_K(\boldsymbol{\mu}_g, \Sigma_g)$, where π_g is a random probability weight between 0 and 1 such that $0 \le \pi_g \le 1$ and $\sum_{g=1}^{G} \pi_g = 1$, G is an integer that specifies the number of normal distributions possibly used in approximating $\boldsymbol{\varepsilon}_{ij}$'s distribution. As Ishwaran and Zarepour (2000) pointed out that increasing G may not significantly improve the accuracy of parameter estimations and a large value G may lead to an increase in computing time. Hence, a moderate value of G such as 20 or 50, which might be enough to capture a good approximation in application, is recommended for Bayesian inference. More details on the selection of G can refer to Ishwaran and Zarepour (2000) and Ohlssen et al. (2007). Generally, it is rather difficult and inefficient to present a Bayesian procedure to make inference on the above specified model because of a finite mixture model of normal distributions involved. An efficient approach to address the issue in a Markov chain Monte Carlo (MCMC) framework is to introduce a latent variable L_{ij} for recording each $\boldsymbol{\varepsilon}_{ij}$'s cluster membership and then take its distribution to be

$$\boldsymbol{\varepsilon}_{ij} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, L_{ij} \sim N_K(\boldsymbol{\mu}_{L_{ij}}, \Sigma_{L_{ij}}), \tag{5}$$

where $\Sigma_{L_{ij}}$ is the L_{ij}-th element of the set of covariance matrices $\boldsymbol{\Sigma} = \{\Sigma_g : g = 1, \ldots, G\}$ with $\Sigma_g = \mathrm{diag}(\sigma_g^{11}, \ldots, \sigma_g^{KK})$, $\boldsymbol{\mu}_{L_{ij}}$ is the L_{ij}-th element of the set of mean vectors $\boldsymbol{\mu} = \{\boldsymbol{\mu}_g : g = 1, \ldots, G\}$ with $\boldsymbol{\mu}_g \sim N_K(\boldsymbol{\mu}_\mu, \Sigma_g)$. In fact, the latent variable L_{ij} is a set of "pointers" for identifying the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ associated with individual i and the measured time point t_{ij} so that the distribution of $\boldsymbol{\varepsilon}_{ij}$ is known when L_{ij} is known, which is similar to the Dirichlet process approximation (DP) to unknown distribution (Chow et al., 2011). Motivated by Chow et al. (2011), the latent variable L_{ij} can be specified by the following Dirichlet process:

$$L_{ij}|\boldsymbol{\pi} \overset{\text{i.i.d.}}{\sim} \text{Multinomial}(\pi_1, \ldots, \pi_G), \tag{6}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)^{\mathrm{T}}$ is defined by the following stick-breaking procedure:

$$\pi_1 = \kappa_1 \text{ and } \pi_g = \kappa_g \prod_{\ell=1}^{g-1}(1-\kappa_\ell) \text{ for } g = 2, \ldots, G, \tag{7}$$

where $\kappa_g \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \tau)$ for $g = 1, \ldots, G-1$, and $\kappa_G = 1$ so that $\sum_{g=1}^{G} \pi_g = 1$. Under the above assumptions, equation (1) can reformulated by

$$\boldsymbol{y}_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, L_{ij} \sim N_K\left(\boldsymbol{\eta}(\boldsymbol{R}_i(t_{ij}), W_i(t_{ij}), \boldsymbol{b}_i) + \boldsymbol{\mu}_{L_{ij}}, \Sigma_{L_{ij}}\right). \tag{8}$$

2.3 Modeling log baseline hazard functions

Following Lang and Brezger (2004), a penalized splines approximation to log baseline hazard function $\lambda^*_{m0}(t)$ in equation (3) is given by

$$\lambda^*_{m0}(t) = \varphi_{m0} + \varphi_{m1}t + \cdots + \varphi_{ms}t^s + \sum_{j=1}^{h_m} \varphi_{m,s+j}(t - \mathcal{K}_{mj})^s_+ = \boldsymbol{\varphi}_m^{\mathrm{T}} \mathbb{B}_m(t), \tag{9}$$

where s is the degree of the polynomial components, h_m is the number of knots (h_m knots define $h_m + 1$ regression intervals because the ending points are not used as knots), $\boldsymbol{\varphi}_m = (\varphi_{m0}, \ldots, \varphi_{m,s+h_m})^{\mathrm{T}}$ is a vector of parameters, and $\mathbb{B}_m(t) = (1, t, \ldots, t^s, (t - \mathcal{K}_{m1})^s_+, \ldots, (t - \mathcal{K}_{mh_m})^s_+)^{\mathrm{T}}$ with $a^s_+ = \{\max(a, 0)\}^s$, $\mathcal{K}_{mj}$ is the location of the j-th knot that can be taken to be the $((j+1)/(h_m+2))$-th quantile of the unique data set $\{T_{im} : i = 1, \ldots, n\}$ for $j = 1, \ldots, h_m$ and $m = 1, \ldots, M$. Generally, one can use the Akaike information criterion or Bayesian information criterion to select the optimal degree of regression splines and number of knots, that is, the optimal sizes of s and h_m. Here, following the argument of Eilers and Marx (1996), a moderate number of knots (usually between 20 and 40) and a small value of s (e.g., s=2 or 3) are recommended for Bayesian analysis.

Clearly, it is rather difficult and complicated to compute equation (4) via the above presented formulae because of an intractable integral involved. To overcome the difficulty, we first construct a finite partition for the m-th time-to-event outcome time axis for $m = 1, \ldots, M$. To this end, we let $0 = C_{m0} < C_{m1} < C_{m2} < \cdots < C_{m\mathcal{L}_m}$, which leads to $\mathcal{L}_m$ intervals $(C_{m0}, C_{m1}], (C_{m1}, C_{m2}], \ldots, (C_{m,\mathcal{L}_m-1}, C_{m\mathcal{L}_m}]$, where $C_{m\mathcal{L}_m}$ can be taken to be some value that is greater than $\max(T_{1m}, \ldots, T_{nm})$ and $\mathcal{L}_m$ is a prespecified integer (e.g., 100 or 150). Generally, one can select subintervals $(C_{m,\ell-1}, C_{m\ell}]$ with equal lengths, or approximately equal lengths subject to the restriction that at least one failure occurs in each interval, or equal numbers of failures or censored observations (Ibrahim et al., 2002) for $m = 1, \ldots, M$. Then, the conditional survival probability $S_m(T_{im}|\boldsymbol{b}_i)$ can be written as

$$S_m(T_{im}|\boldsymbol{b}_i) = \exp\left\{-\sum_{\ell=1}^{\mathcal{L}_m} \mathcal{D}_{im\ell}\right\}, \tag{10}$$

where $\mathcal{D}_{im\ell} = \int_{C_{m,\ell-1}}^{C_{m\ell}} \lambda_m(u|\boldsymbol{b}_i)1(u \le T_{im})du$, and $1(u \le T_{im})$ is a generic indicator function taking the value 1 if $u \le T_{im}$ and 0 otherwise. According to the theory of rectangular integral approximation, when $\mathcal{L}_m$ is sufficiently large, $\mathcal{D}_{iml}$ can be approximated by

$$\begin{aligned}\mathcal{D}_{im\ell} \approx\ & (C_{m\ell} - C_{m,\ell-1})\lambda_m(u_{m\ell}|\boldsymbol{b}_i)1(C_{m\ell} < T_{im}) + \\ & + (T_{im} - C_{m,\ell-1})\lambda_m(u^*_{im\ell}|\boldsymbol{b}_i)1(C_{m,\ell-1} < T_{im} \le C_{m\ell}),\end{aligned} \tag{11}$$

where $u_{m\ell} = (C_{m\ell} + C_{m,\ell-1})/2$ and $u^*_{im\ell} = (T_{im} + C_{m,\ell-1})/2$. Clearly, $\mathcal{D}_{im\ell} = 0$ if $C_{m,\ell-1} > T_{im}$. Based on equations (9)–(11), it is feasible to facilitate the computation of equations (3)–(4).

2.4 Prior specification

To develop Bayesian inference on the considered models, we need specifying the prior distributions for covariance matrix Ω of random effects, $\boldsymbol{\mu}_\mu$ and σ_g^{kk} ($k = 1, \ldots, K$, $g = 1, \ldots, G$) related to equation (5) and τ related to equation (7). Following the arguments of Chow et al. (2011) and Zhu et al. (2012), we consider the following priors for Ω, $\boldsymbol{\mu}_\mu$, σ_g^{kk}, and τ:

$$\boldsymbol{\Omega} \sim \mathrm{IW}_q(\boldsymbol{R}^0, \varrho), \quad \boldsymbol{\mu}_\mu \sim N_K\left(\boldsymbol{\zeta}_\mu^0, H_\mu^0\right), \quad \left(\sigma_g^{kk}\right)^{-1} \sim \Gamma(c_1, c_2), \quad \tau \sim \Gamma(a_\tau, b_\tau),$$

where $\boldsymbol{R}^0$, ϱ, $\boldsymbol{\zeta}_\mu^0$, H_μ^0, c_1, c_2, a_τ, and b_τ are the pregiven hyperparameters, $\mathrm{IW}_q(\cdot, \cdot)$ represents the inverted Wishart distribution, and $\Gamma(a, b)$ denotes the gamma distribution with parameters a and b. The hyperparameters a_τ and b_τ should be carefully selected because they directly affect estimate of τ controlling the behavior of $\boldsymbol{\varepsilon}_{ij}$. The details for the selection of a_τ and b_τ can refer to Chow et al. (2011).

In a Bayesian framework, we require specifying the prior of φ_{mj} related to equation (9). Following Lang and Brezger (2004), we consider the following second-order difference for specifying φ_{mj}'s prior:

$$\varphi_{mj} = 2\varphi_{m,j-1} - \varphi_{m,j-2} + u_{mj} \text{ with } u_{mj} \sim N\left(0, \varsigma_m^2\right) \text{ for } j = 2, \ldots, s + h_m,$$

and the diffuse prior for φ_{m0} and $\varphi_{m1} \propto$ constant, where ς_m^2 is introduced to control the amount of smoothness. The prior for ς_m^{-2} is assumed to follow a Gamma distribution, that is, $\varsigma_m^{-2} \sim \mathrm{Gamma}(a_\varsigma^m, b_\varsigma^m)$ with the pregiven hyperparameters a_ς^m and b_ς^m. A common selection for the hyperparameters is $a_\varsigma^m = 1$ and a small value for b_ς^m, for example, $b_\varsigma^m = 0.005$, leading to an almost diffuse prior for ς_m^2.

For the above-defined models together with the above given priors, our major interest is to estimate parameters $\boldsymbol{\beta}$, Ω, $\boldsymbol{\psi}_m$, and $\boldsymbol{\gamma}_m$ and to identify the important covariates. To this end, we consider a BLasso approach as follows.

3 Bayesian Lasso

Tibshirani (1996) showed that the Lasso estimates for linear regression parameters via the ℓ_1-penalized least-squares criterion can be interpreted as Bayesian posterior mode estimates when the regression parameters have independent Laplace (i.e., double-exponential) priors. Motivated by the idea, Bae and Mallick (2004) and Yuan and Lin (2006) subsequently proposed the Laplace-like prior for linear regression parameter, Park and Casella (2008) proposed a Bayesian framework for Lasso, and Guo et al. (2012) extended BLasso approach to a semiparametric structural equation model. However, to our knowledge, there is little work developed on covariate selection for the considered SJMLS in a Bayesian framework.

Following Park and Casella (2008) and Guo et al. (2012), a BLasso procedure can be proposed to identify the important covariates in equations (1) and (2) by imposing the following conditional Laplace priors on $\boldsymbol{\beta}_k$, $\boldsymbol{\gamma}_m$, and $\boldsymbol{\psi}_m$:

$$p(\boldsymbol{\beta}_k|\vartheta_k) = \prod_{j=0}^{r} \frac{\vartheta_k}{2} \exp(-\vartheta_k|\beta_{kj}|), \; p(\boldsymbol{\gamma}_m|\upsilon_m) = \prod_{j=1}^{p} \frac{\upsilon_m}{2} \exp(-\upsilon_m|\gamma_{mj}|),$$

$$p(\boldsymbol{\psi}_m|\nu_m) = \prod_{j=1}^{K} \frac{\nu_m}{2} \exp(-\nu_m|\psi_{mj}|),$$

for $k = 1, \ldots, K$ and $m = 1, \ldots, M$, respectively, where ϑ_k, υ_m, and ν_m are the regularization parameters that control the tail decay. Because the masses of the above presented Laplace priors are quite highly concentrated around zero with a distinct peak at zero, posterior means or modes of β_{kj}'s, γ_{mj}'s, and ψ_{mj}'s are shrunk toward zero, which is the key principle in using BLasso method to select the important covariates. Following Tibshirani (1996), the Laplace distribution with the form $a\exp(-a|z|)/2$ can be represented as a scale mixture of normal distributions with independent exponentially distributed variance, that is,

$$\frac{a}{2}\exp\left(-a|z|\right) = \int_0^\infty \frac{1}{\sqrt{2\pi u}}\exp\left(-\frac{z^2}{2u}\right)\frac{a^2}{2}\exp\left(-\frac{a^2u}{2}\right)du \quad \text{for } a > 0,$$

which shows that the prior on $\boldsymbol{\beta}_k$ or $\boldsymbol{\gamma}_m$ or $\boldsymbol{\psi}_m$ can be written as a tractable hierarchical formulation by introducing a latent variable. Therefore, the above specified priors for $\boldsymbol{\beta}_k$, $\boldsymbol{\gamma}_m$, and $\boldsymbol{\psi}_m$ can be reformulated as the following hierarchical models:

$$\begin{aligned}
&\boldsymbol{\beta}_k|H_{\beta_k} \sim N_{r+1}(\mathbf{0}, H_{\beta_k}) \text{ with } H_{\beta_k} = \operatorname{diag}\left(h^2_{\beta_{k0}}, \ldots, h^2_{\beta_{kr}}\right),\\
&\boldsymbol{\gamma}_m|H_{\gamma_m} \sim N_p(\mathbf{0}, H_{\gamma_m}) \text{ with } H_{\gamma_m} = \operatorname{diag}\left(h^2_{\gamma_{m1}}, \ldots, h^2_{\gamma_{mp}}\right),\\
&\boldsymbol{\psi}_m|H_{\psi_m} \sim N_K(\mathbf{0}, H_{\psi_m}) \text{ with } H_{\psi_m} = \operatorname{diag}\left(h^2_{\psi_{m1}}, \ldots, h^2_{\psi_{mK}}\right),\\
&p\left(h^2_{\beta_{k0}}, \ldots, h^2_{\beta_{kr}}\right) = \prod_{j=0}^{r}\frac{\vartheta_k^2}{2}\exp\left(-\frac{\vartheta_k^2}{2}h^2_{\beta_{kj}}\right),\\
&p\left(h^2_{\gamma_{m1}}, \ldots, h^2_{\gamma_{mp}}\right) = \prod_{j=1}^{p}\frac{\upsilon_m^2}{2}\exp\left(-\frac{\upsilon_m^2}{2}h^2_{\gamma_{mj}}\right),\\
&p\left(h^2_{\psi_{m1}}, \ldots, h^2_{\psi_{mK}}\right) = \prod_{j=1}^{K}\frac{\nu_m^2}{2}\exp\left(-\frac{\nu_m^2}{2}h^2_{\nu_{mj}}\right).
\end{aligned} \tag{12}$$

The above hierarchical representation greatly simplifies the computation because all the full conditional distributions have the closed expressions. Thus, one can directly draw observations from these conditional distributions using the Gibbs sampler (Geman and Geman, 1984).

To implement the above presented BLasso procedure, it is necessary to select ϑ_k^2, υ_m^2, and ν_m^2. Generally, one can specify ϑ_k^2, υ_m^2, and ν_m^2 by using the empirical Bayes method or the fully Bayes method with the appropriate hyperprior. Inspired by Park and Casella (2008), we consider the following conjugate priors for ϑ_k^2, υ_m^2, and ν_m^2:

$$\vartheta_k^2 \sim \Gamma\left(a_\vartheta^k, b_\vartheta^k\right), \quad \upsilon_m^2 \sim \Gamma\left(a_\upsilon^m, b_\upsilon^m\right), \quad \text{and} \quad \nu_m^2 \sim \Gamma\left(a_\nu^m, b_\nu^m\right), \tag{13}$$

where a_ϑ^k, b_ϑ^k, a_υ^m, b_υ^m a_ν^m, and b_ν^m are the prespecified hyperparameters. Thus, it follows from equations (12) and (13) that the conditional distributions of ϑ_k^2, υ_m^2, and ν_m^2 are given by

$$\vartheta_k^2|\boldsymbol{\beta}_k, H_{\beta_k} \sim \Gamma\left(a_\vartheta^k + r + 1, b_\vartheta^k + \frac{1}{2}\sum_{j=0}^{r}h^2_{\beta_{kj}}\right), \upsilon_m^2|\boldsymbol{\gamma}_m, H_{\gamma_m} \sim \Gamma\left(a_\upsilon^m + p, b_\upsilon^m + \frac{1}{2}\sum_{j=1}^{p}h^2_{\gamma_{mj}}\right),$$

$$\upsilon_m^2|\boldsymbol{\gamma}_m, H_{\psi_m} \sim \Gamma\left(a_\upsilon^m + K, b_\upsilon^m + \frac{1}{2}\sum_{j=1}^{K} h_{\psi_{mj}}^2\right),$$

respectively. The conditional distributions for $h_{\beta_{kj}}^{-2}$ $(j = 1, \ldots, r)$, $h_{\gamma_{m\ell}}^{-2}$ $(\ell = 1, \ldots, p)$, and $h_{\psi_{m\iota}}^{-2}$ $(\iota = 1, \ldots, K)$ are given by

$$h_{\beta_{kj}}^{-2}|\beta_{kj}, \vartheta_k^2 \sim \mathrm{IG}\left(\left|\vartheta_k/\beta_{kj}\right|, \vartheta_k^2\right), \quad h_{\gamma_{m\ell}}^{-2}|\gamma_{m\ell}, \upsilon_m^2 \sim \mathrm{IG}\left(\left|\upsilon_m/\gamma_{m\ell}\right|, \upsilon_m^2\right),$$
$$h_{\psi_{m\iota}}^{-2}|\psi_{m\iota}, \upsilon_m^2 \sim \mathrm{IG}\left(\left|\upsilon_m/\psi_{m\iota}\right|, \upsilon_m^2\right),$$

respectively, where $\mathrm{IG}(a, b)$ represents the inverse Gaussian distribution with the scale parameter a and the shape parameter b. For the details for sampling observations from the inverse Gaussian distribution one can refer to Appendix B.

Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_Y, \boldsymbol{\theta}_T, \boldsymbol{\theta}_\varepsilon\}$, where $\boldsymbol{\theta}_Y = \{\beta, \Omega\}$, $\boldsymbol{\theta}_T = \{(\boldsymbol{\varphi}_m, \boldsymbol{\psi}_m, \boldsymbol{\gamma}_m) : m = 1, \ldots, M\}$ and $\boldsymbol{\theta}_\varepsilon$ contains all unknown parameters related to $\boldsymbol{\varepsilon}_{ij}$'s distribution. Let $\boldsymbol{B} = \{\boldsymbol{b}_i : i = 1, \ldots, n\}$ be the set of random effects, and $\boldsymbol{D}_o = \{(\boldsymbol{y}_{ij}, \boldsymbol{T}_i, \boldsymbol{R}_i(t_{ij}), \boldsymbol{W}_i(t_{ij}), \boldsymbol{\xi}_i, \boldsymbol{\delta}_i) : i = 1, \ldots, n, j = 1, \ldots, n_i\}$ be the observed data set. Bayesian statistical inference including parameter estimation and covariate selection on $\boldsymbol{\theta}$ and $\boldsymbol{B}$ is focused on the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{B}|\boldsymbol{D}_o)$. The Gibbs sampler (Geman and Geman, 1984) together with the Metropolis–Hastings (MH) algorithm is adopted to simulate a sequence of random observations from the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{B}|\boldsymbol{D}_o)$, and then the Bayesian estimates are obtained from the mean of the generated random observations. The conditional distributions required in implementing the above proposed BLasso procedure are presented in Appendix A.

4 Simulation studies

In this section, we conducted several simulation studies to investigate the finite performance of the above proposed methods.

We considered the model defined in equations (1) and (2) with $K = 2$, $M = 2$, $r = 7$, $p = 6$, $q = 2$, $W_i(t_{ij}) = I_2$, $\boldsymbol{R}_i(t_{ij}) = (1, R_{i1}, \ldots, R_{i6}, t_{ij})^{\mathrm{T}}$, and sample size $n = 200$. The data were generated as follows: covariate vectors $(R_{i1}, \ldots, R_{i6})^{\mathrm{T}}$ and $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{i6})^{\mathrm{T}}$ were independently generated from the multivariate normal distribution $N_6(\mathbf{1}, I)$, $\boldsymbol{b}_i$ was generated from a bivariate normal distribution $N_2(\mathbf{0}, \Omega)$ with $(\Omega_{11}, \Omega_{12}, \Omega_{21}, \Omega_{22}) = (0.25, 0.10, 0.10, 0.25)$, the censoring time was taken to be $C_{im} = \mathbf{1}(u_{im} > 1.0) + u_{im}\mathbf{1}(u_{im} \le 1.0)$ in which u_{im} was generated from a uniform distribution $U(0.8, 1.2)$, $T_{im} = \min(T_{im}^*, C_{im})$, and $t_{ij} = 0.25(j-1)$ for $j = 1, \ldots, n_i$, where n_i satisfies $t_{in_i} \le \max(T_{i1}, T_{i2})$. The true values of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\gamma}_1$, and $\boldsymbol{\gamma}_2$ were taken to be $\boldsymbol{\beta}_1 = (1.0, 0.8, 0.2, -0.2, 0.0, 0.0, 0.0, 0.4)^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (0.4, 0.9, -0.2, 0.2, 0.0, 0.0, 0.0, 0.6)^{\mathrm{T}}$, $\boldsymbol{\gamma}_1 = (0.45, -0.35, 0.35, 0.00, 0.00, 0.00)^{\mathrm{T}}$, $\boldsymbol{\gamma}_2 = -\boldsymbol{\gamma}_1$, respectively, which indicated that variables R_{i4}, R_{i5}, R_{i6}, ξ_{i4}, ξ_{i5}, and ξ_{i6} were six unimportant covariates in the model considered here. Our main purpose was to use the proposed approach to identify the unimportant covariates and estimate nonzero coefficients. Bayesian results were obtained from 200 replications.

To show that the proposed methods can capture the feature of various longitudinal measurement error distributions and cover the feature of various log baseline hazard functions, we considered two scenarios for $\boldsymbol{\varepsilon}_{ij}$ and $\lambda_{m0}^*(t)$ as follows.

Scenario 1. The log baseline hazard functions were specified by

$$\lambda_{10}^*(t) = 2t^2 - 1.6t, \quad \lambda_{20}^*(t) = \log\left(1 + 0.7\sin\left(\frac{2\pi}{3}t\right)\right),$$

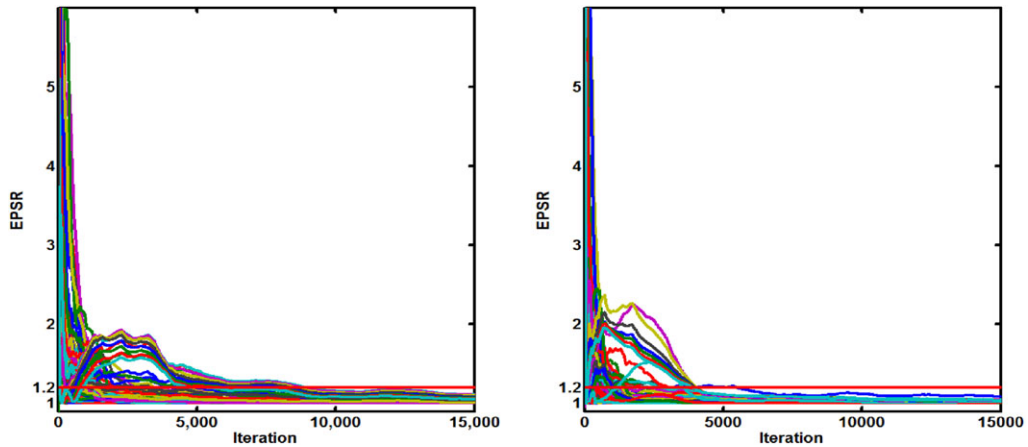

Figure 1 EPSR (i.e., estimated potential scale reduction) values of all parameters against iteration numbers for a randomly selected replication in Scenario 1 (left panel) and Scenario 2 (right panel).

which correspond to a quadratic function and a nonlinear function, respectively; and the true distribution of $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{ij1}, \varepsilon_{ij2})^{\mathrm{T}}$ was taken to be

$$\varepsilon_{ij1} \sim N(0, 0.25), \quad \varepsilon_{ij2} \sim 0.4N(0, 0.3) + 0.3N(-1.5, 0.3) + 0.3N(1.5, 0.4),$$

which corresponded to unimodal and trimodal distributions, respectively.

Scenario 2. The log baseline hazard functions were taken to be

$$\lambda^*_{10}(t) = \log(2), \quad \lambda^*_{20}(t) = \log(1 + 0.5t),$$

which correspond to a constant function adopted by Zhu et al. (2012) and a nonlinear function, respectively; and the true distribution of $\boldsymbol{\varepsilon}_{ij}$ was specified by

$$\varepsilon_{ij1} = \varepsilon^*_{ij1} - 2 \text{ with } \varepsilon^*_{ij1} \overset{\text{i.i.d.}}{\sim} \Gamma(4, 2), \quad \varepsilon_{ij2} \overset{\text{i.i.d.}}{\sim} 0.6N(-0.4, 0.04) + 0.4N(0.6, 0.04),$$

which correspond to a right skewed distribution and a bimodal distribution, respectively. The average censoring proportions of the survival times were about 29% and 35% for the Scenarios 1 and 2, respectively.

For each of the above generated data sets, the proposed semiparametric Bayesian procedure was used to simultaneously evaluate Bayesian estimates of unknown parameters, error distributions and log baseline hazard functions $\lambda^*_{m0}(t)$, and select the important covariates. The hyperparameters $\boldsymbol{R}^0$ and $\boldsymbol{\psi}^0_m$ were taken to be their corresponding true values, and we set $\varrho = 1, c_1 = 11, c_2 = 2, \boldsymbol{\zeta}^0_\mu = \mathbf{0}_2, H^0_\mu = 10I_2$, and $a_\tau = b_\tau = 0.1$ relating to the DP mixture of normal distributions, $a^m_\varsigma = 1$ and $b^m_\varsigma = 0.005$ relating to the second-order difference, $a^k_\vartheta = a^m_\upsilon = a^m_\nu = 1$ and $b^k_\vartheta = b^m_\upsilon = b^m_\nu = 0.1$ corresponding to diffuse priors in equation (13). To approximate $\mathcal{D}_{iml}$ defined in equation (11), we equably divided the time axis into 100 (i.e., $\mathcal{L}_m = 100$) subintervals. Following the argument given in Section 2.2, we set $G = 20$, the degree of splines $s = 2$, and the number of knots $h_m = 20$. To investigate the convergence of the proposed algorithm, we calculated the estimated potential scale reduction (EPSR) values of parameters (Gelman et al., 1996) based on three parallel sequences of observations that were generated from three different starting values. For the randomly selected five test runs, we observed that the EPSR values were less than 1.2 after about 10,000 iterations (e.g., see Fig. 1). Thus, 5000 observations were collected after 10,000 iterations in producing Bayesian results for each of 200 replications. For comparison, we calculated Bayesian estimates of parameters under noninformative priors of the regression parameters.

Table 1 Bayesian estimates of parameters in the simulation studies.

Par.	True	Scenario I			Scenario II					
		Laplace prior			Laplace prior			Noninformative prior		
		Bias	RMS	F0(%)	Bias	RMS	F0(%)	Bias	RMS	F0(%)
ψ_{11}	0.00	0.019	0.111	96.0	−0.003	0.114	95.0	−0.051	0.160	89.0
ψ_{12}	0.50	−0.030	0.123	3.5	−0.051	0.136	4.0	0.021	0.140	2.5
ψ_{21}	0.00	0.024	0.118	97.0	−0.003	0.102	97.5	−0.055	0.147	90.0
ψ_{22}	0.60	−0.063	0.139	0.0	−0.039	0.115	0.0	0.038	0.139	0.5
γ_{11}	0.45	0.010	0.093	0.0	−0.029	0.099	0.5	0.021	0.107	0.0
γ_{12}	−0.35	0.032	0.092	4.5	0.024	0.092	5.0	0.001	0.091	3.5
γ_{13}	0.35	0.005	0.093	1.5	−0.028	0.100	7.0	0.009	0.096	2.5
γ_{14}	−0.00	0.013	0.074	97.0	−0.002	0.075	98.0	−0.010	0.102	90.0
γ_{15}	0.00	0.012	0.072	96.5	−0.001	0.077	96.5	−0.009	0.087	91.5
γ_{16}	0.00	0.011	0.079	96.0	−0.003	0.080	96.0	0.006	0.098	89.5
γ_{21}	−0.45	0.034	0.100	0.0	0.040	0.110	0.5	−0.001	0.099	0.0
γ_{22}	0.35	−0.026	0.096	4.5	−0.016	0.085	2.0	0.012	0.101	3.5
γ_{23}	−0.35	0.035	0.095	4.5	0.039	0.096	7.5	0.001	0.103	4.5
γ_{24}	0.00	0.002	0.072	97.5	0.006	0.076	97.0	0.016	0.103	91.0
γ_{25}	0.00	−0.004	0.077	95.5	0.000	0.075	97.0	0.014	0.100	91.0
γ_{26}	0.00	−0.003	0.077	95.5	−0.002	0.076	98.0	0.013	0.085	90.0
β_{10}	1.00	−0.002	0.041	0.0	0.004	0.067	0.0	0.004	0.069	0.0
β_{11}	0.80	−0.007	0.040	0.0	−0.008	0.052	0.0	0.010	0.054	0.0
β_{12}	0.20	−0.003	0.036	0.0	−0.001	0.049	3.0	−0.004	0.051	5.0
β_{13}	−0.20	0.002	0.040	0.5	0.011	0.058	8.0	−0.002	0.053	4.5
β_{14}	0.00	−0.001	0.038	94.0	−0.002	0.046	96.5	−0.005	0.055	87.5
β_{15}	0.00	−0.001	0.034	93.0	−0.006	0.048	97.0	0.006	0.048	91.0
β_{16}	0.00	−0.003	0.036	95.0	−0.004	0.048	96.0	−0.025	0.050	89.0
β_{17}	0.40	−0.005	0.035	0.0	−0.031	0.115	3.0	−0.008	0.107	4.5
β_{20}	0.40	−0.010	0.070	0.0	−0.005	0.046	0.0	0.002	0.044	0.0
β_{21}	0.90	−0.011	0.052	0.0	−0.010	0.046	0.0	0.001	0.041	0.0
β_{22}	−0.20	0.008	0.055	5.5	0.006	0.044	0.5	−0.003	0.041	0.0
β_{23}	0.20	−0.010	0.054	6.0	−0.002	0.044	0.0	−0.009	0.043	2.0
β_{24}	0.00	0.000	0.043	97.0	−0.002	0.038	94.0	−0.001	0.041	88.5
β_{25}	0.00	0.004	0.044	98.0	−0.001	0.041	91.5	−0.001	0.042	87.5
β_{26}	0.00	−0.001	0.046	96.5	0.000	0.041	91.5	0.001	0.041	87.0
β_{27}	0.60	−0.006	0.082	0.0	−0.006	0.034	0.0	0.002	0.037	0.0
Ω_{11}	0.25	0.028	0.039	–	0.091	0.100	–	0.093	0.103	–
Ω_{12}	0.10	−0.027	0.036	–	−0.027	0.037	–	−0.023	0.035	–
Ω_{22}	0.25	0.070	0.086	–	0.033	0.044	–	0.040	0.050	–

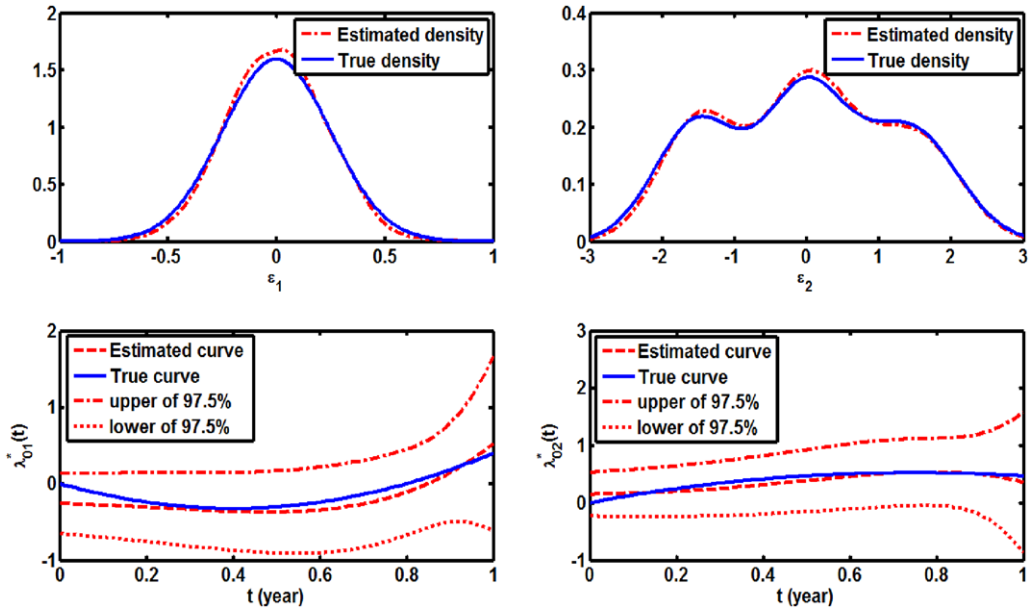

Figure 2 Estimated versus true densities of residual ε_{ij1} (upper left panel) and ε_{ij2} (upper right panel), and estimated versus true values of log baseline hazard function λ_{10}^{*} (lower left panel) and λ_{20}^{*} (lower right panel) in Scenario 1.

Results were reported in Table 1, where "bias" is the difference between the true value and the mean of the estimates based on 200 replications, "RMS" is the root mean square between the estimates based on 200 replications and its true value, and "F0" is the proportion that parameter was identified to be zero in 200 replications in terms of the criterion that a parameter was identified to be 0 if its 95% confidence interval contains zero.

Examination of Table 1 indicated that (i) Bayesian estimates of parameters were reasonably accurate regardless of $\boldsymbol{\varepsilon}_{ij}$'s distributions and prior inputs of parameters because their absolute biases were less than 0.10 and their RMS values were less than 0.15, especially for parameters corresponding to unimportant covariates, their corresponding absolute biases, and RMS values were obviously less in most cases; (ii) BLasso could identify the correct models in most cases regardless of prior inputs of parameters because the F0 values corresponding to the important covariates were less than 8%, but the F0 values corresponding to unimportant covariates were more than 90%; (iii) estimates obtained with the Laplace priors of parameters were better than those obtained with the noninformative priors of parameters in terms of the RMS values; (iv) BLasso method behaves better than a general Bayesian method with the noninformative priors of parameters in terms of the F0 values. Figures 2 and 3 plotted the estimated densities of ε_{ij1} and ε_{ij2} against their corresponding true densities, the estimated curves of $\lambda_{10}^{*}(t)$ and $\lambda_{20}^{*}(t)$ against their corresponding true curves for a randomly selected replication under the above considered two scenarios, respectively. Inspection of Figs. 2 and 3 showed that (i) the finite mixture of normal distributions was flexible enough to capture the general shapes of our considered two distribution assumptions for $\boldsymbol{\varepsilon}_{ij}$; (ii) the proposed P-splines approximation to nonparametric function was flexible enough to recover the true log baseline hazard function, and the slight difference between the estimated and true curves was observed at some time points; (iii) the estimated 95% confidence region for the baseline hazard function could cover its true curve with a reasonably narrow region. The performance of the proposed approach to recover the true baseline hazard function in a multivariate survival model can be measured by the root mean square error of function $\lambda_{m0}^{*}(t)$:

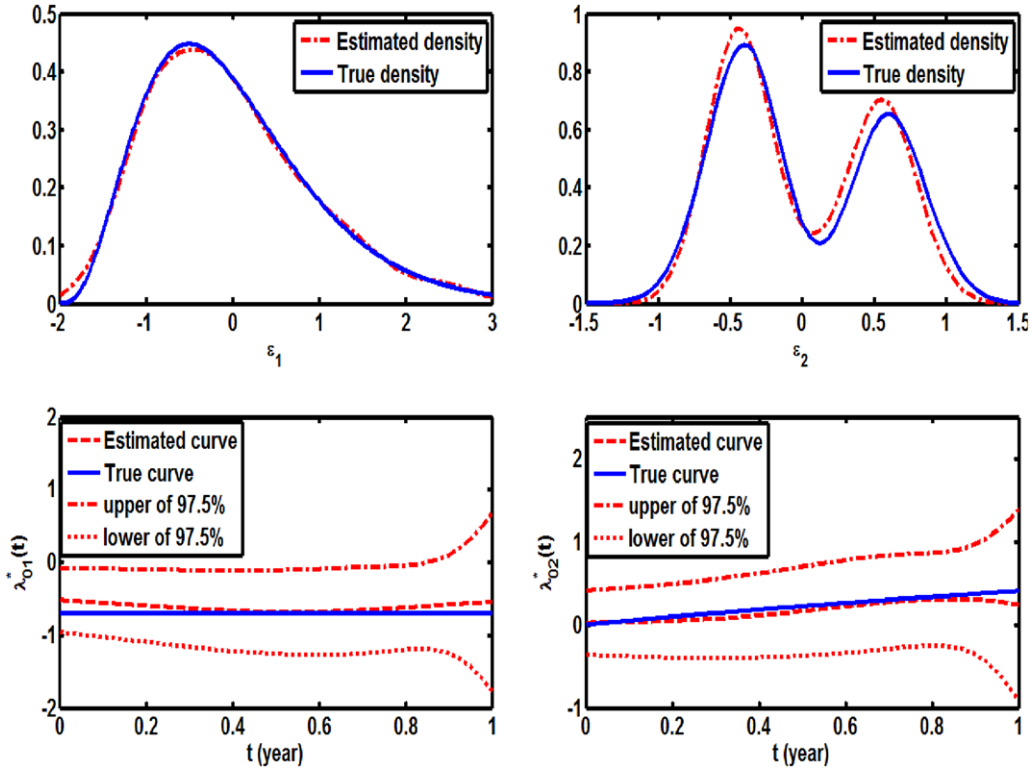

Figure 3 Estimated versus true densities of residual ε_{ij1} (upper left panel) and ε_{ij2} (upper right panel), and estimated versus true values of log baseline hazard function λ_{10}^{*} (lower left panel) and λ_{20}^{*} (lower right panel) in Scenario 2.

$\text{RMSE}(\lambda_{m0}^{*}) = \sqrt{\sum_{l=0}^{\mathcal{L}_m}(\lambda_{m0}^{*}(c_{ml}) - \hat{\lambda}_{m0}^{*}(c_{ml}))^2/(\mathcal{L}_m+1)}$ with $\hat{\lambda}_{m0}^{*}(t) = \hat{\boldsymbol{\varphi}}_m^{\mathrm{T}}\mathbb{B}_m(t)$, where $\hat{\boldsymbol{\varphi}}_m$ was the mean of 200 estimates for $\boldsymbol{\varphi}_m$. $\text{RMSE}(\lambda_{10}^{*})$ and $\text{RMSE}(\lambda_{20}^{*})$ were 0.101 and 0.068 under Scenario 1, respectively, and 0.090 and 0.059 under Scenario 2, respectively, which indicated that our proposed P-splines approximation to $\lambda_{m0}^{*}(t)$ performed well. All these findings indicated that (i) our proposed Bayesian procedure could well capture the true information of $\boldsymbol{\varepsilon}_{ij}$ and $\lambda_{m0}(t)$ regardless of their true distributions and forms, and (ii) BLasso could identify the true model with a high probability.

5 Application to the IBCSG data

To illustrate applications of the proposed approach, we considered a data set from a clinical trial conducted by IBCSG for 832 premenopausal women from Switzerland, Sweden, and New Zealand/Australia. Our major interest is to investigate the relationship between longitudinal outcome (i.e., QOL) and survival time (i.e., DFS and OS) and to identify important factors (i.e., covariates), which have a significant effect on QOL and/or DFS and OS. Chi and Ibrahim (2006) and Zhu et al. (2012) analyzed the data set via a JMLS with longitudinal measurement error following a multivariate normal distribution and the fixed covariates. Unlike Chi and Ibrahim (2006) and Zhu et al. (2012), we fitted the IBCSG data via a SJMLS defined in (1) and (2) by using the above developed BLasso procedure. For each of four longitudinal QOL indicators (appetite, y_1; perceived coping, y_2; mood, y_3; and physical well-being, y_4, more details could refer to Appendix C), we transformed its corresponding

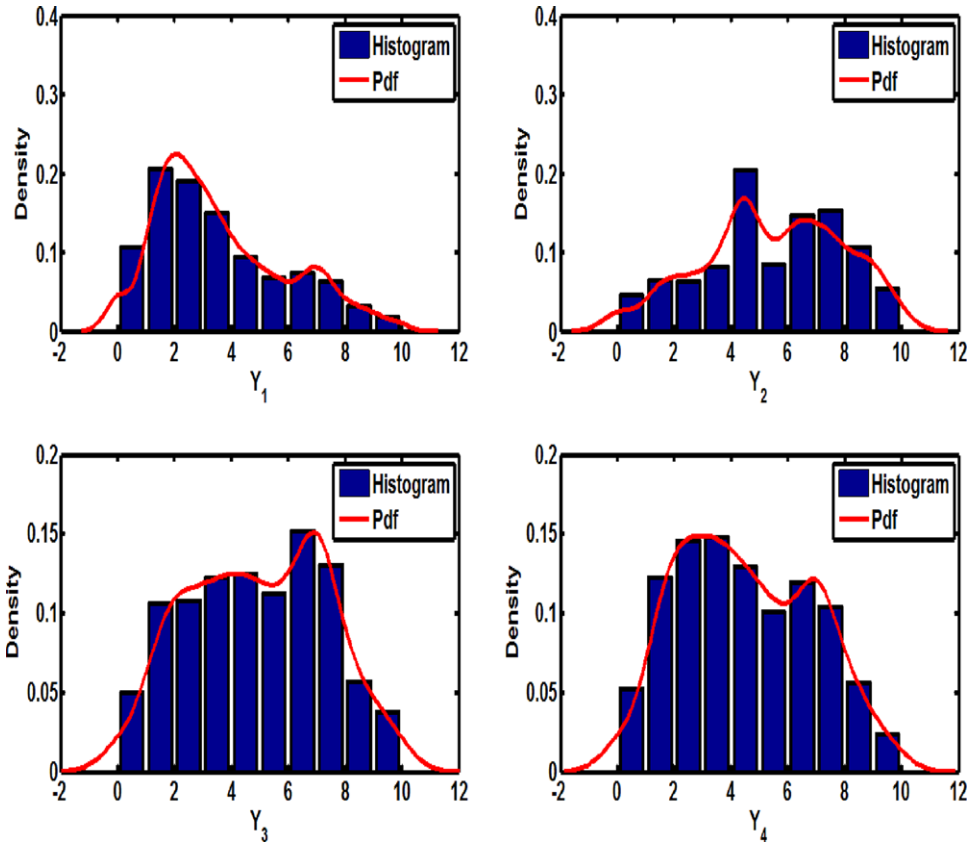

Figure 4 Histograms and estimated densities of y_1 (upper left panel), y_2 (upper right panel), y_3 (lower left panel), and y_4 (lower right panel): IBCSG data.

observed value to $\sqrt{100 - \text{QOL}}$ (Chi and Ibrahim, 2006). The transformed QOLs decreased over time and were scaled tween 0 and 10 with smaller values reflecting better QOL (Zhu et al., 2012). Their corresponding densities and histograms were shown in Fig. 4. Examination of Fig. 4 indicated that the within-individual error might not follow a normal distribution but some multimodal and asymmetric distribution, for example, a finite mixture of normal distributions.

Let $y_{ij1}, \ldots, y_{ij4}$ be the observed values of $y_1, \ldots, y_4$ for the i-th woman at time point t_{ij} for $i = 1, \ldots, 832$ and $j = 1, \ldots, n_i$ with $n_i \in \{1, 2, 3\}$. Similar to Chi and Ibrahim (2006) and Zhu et al. (2012), we fitted the IBCSG data set to the following SJMLS:

$$\begin{cases} \boldsymbol{y}_{ij} = \boldsymbol{\eta}(\boldsymbol{R}_i(t_{ij}), W_i(t_{ij}), \boldsymbol{b}_i) + \boldsymbol{\varepsilon}_{ij}, \ i = 1, \ldots, 832, \ j = 1, \ldots, n_i, \\ \lambda_m(t|\boldsymbol{b}_i) = \exp\{\lambda^*_{m0}(t) + \boldsymbol{\psi}_m^{\mathrm{T}} \boldsymbol{\eta}(\boldsymbol{R}_i(t), W_i(t), \boldsymbol{b}_i) + \boldsymbol{\gamma}_m^{\mathrm{T}} \boldsymbol{\xi}_i\}, \ m = 1, 2, \end{cases}$$

where $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ij4})^{\mathrm{T}}$, $\boldsymbol{\eta}(\boldsymbol{R}_i(t), W_i(t), \boldsymbol{b}_i) = \beta \boldsymbol{R}_i(t) + W_i(t)\boldsymbol{b}_i$ with $W_i(t) = I_4$ and $\boldsymbol{R}_i(t) = (1, R_{i1}, \ldots, R_{i8}, t)^{\mathrm{T}}$ in which covariates $R_{i1}, \ldots, R_{i8}$ were listed in Appendix C, $\beta = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_4)^{\mathrm{T}}$ with $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{k9})^{\mathrm{T}}$ for $k = 1, \ldots, 4$, $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{i8})^{\mathrm{T}}$ in which $\xi_{i\ell} = R_{i\ell}$ for $\ell = 1, \ldots, 8$, $\boldsymbol{\psi}_m = (\psi_{m1}, \ldots, \psi_{m4})^{\mathrm{T}}$, and $\boldsymbol{\gamma}_m = (\gamma_{m1}, \ldots, \gamma_{m8})^{\mathrm{T}}$. Here, we assumed that New Zealand/Australia was the reference category. Moreover, it was assumed that the random effects $\boldsymbol{b}_i$'s were independent and identically distributed as $N_4(\boldsymbol{0}, \Omega)$, and the longitudinal measurement errors $\boldsymbol{\varepsilon}_{ij}$'s were independent and identically distributed as a finite mixture of normal distributions.

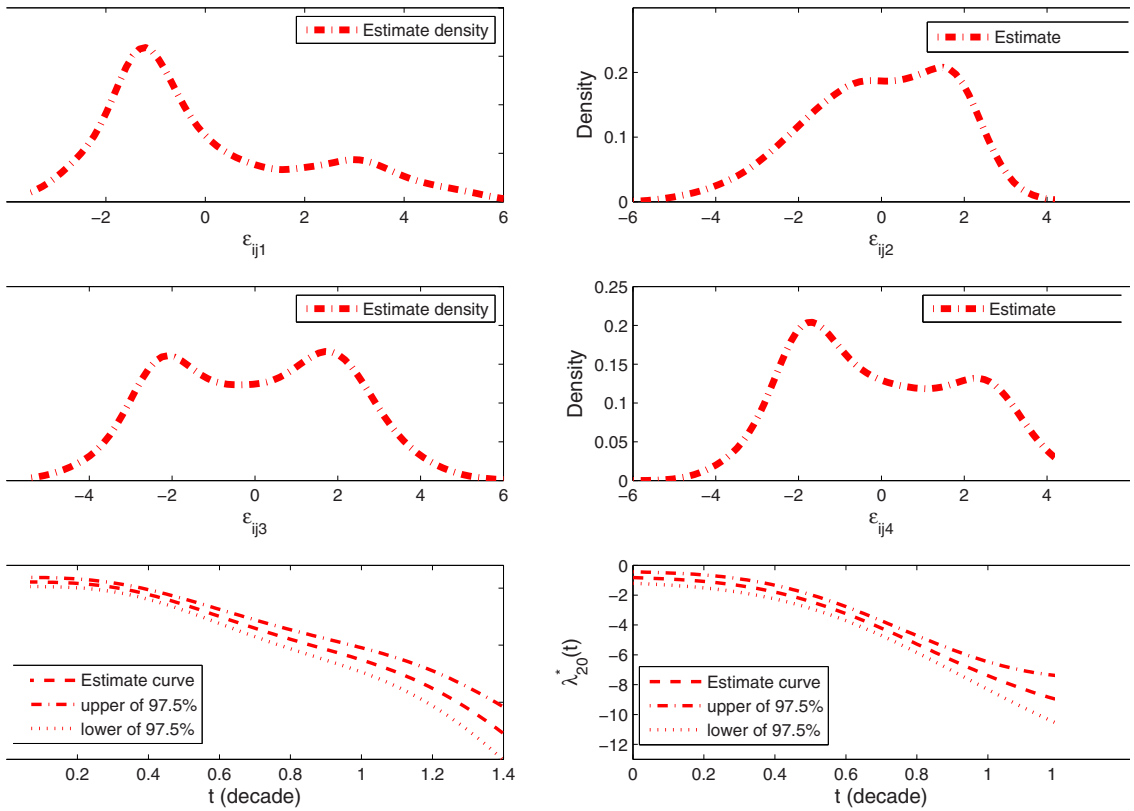

Figure 5 (a) Estimated densities of residual ε_{ijk} for $k = 1$ (upper left panel), $k = 2$ (upper right panel), $k = 1$ (middle left panel), and $k = 4$ (middle right panel): IBCSG data. (b) Estimated log baseline hazard functions of $\lambda^*_{m0}(t)$ for $m = 1$ (lower left panel) and $m = 2$ (lower right panel): IBCSG data.

To use the proposed BLasso procedure to analyze the IBCSG data set, we took $G = 20$, $s = 2$, $h_m = 40$, and $\mathcal{L}_m = 200$ with the equal-length subintervals when using P-splines to approximate the log baseline hazard $\lambda^*_{m0}(t)$ for $m = 1, 2$. The same priors and hyperparameters are specified as in simulation studies. Based on the above settings, we calculated the EPSR values for parameters in the above specified SJMLS, which is not presented to save space. The EPSR values showed that the MCMC algorithm converged about 10,000 iterations because the EPSR values of parameters were less than 1.2 about 10,000 iterations. Thus, 5000 observations were collected to evaluate Bayesian estimates and standard deviations of parameters after 10,000 iterations. Results were presented in Tables 2, 3 and Fig. 5. Matlab program for implementing the proposed BLasso can be seen in the Supporting Information on the journal's website.

Examination of Fig. 5 indicated that (i) the estimated densities of ε_{ij1} and ε_{ij4} were skew, and the estimated density of ε_{ij3} was bimodal, which implied that it might be unreasonable to specify a symmetric normal distribution for $\boldsymbol{\varepsilon}_{ij}$; (ii) the estimated log baseline hazard functions $\hat{\lambda}^*_{10}(t)$ and $\hat{\lambda}^*_{20}(t)$ monotonically decreased with respect to t, and were located within their corresponding 95% confidence regions. From Table 2, we have the following observations: (i) the estimated correlations r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, and r_{34} were 0.535, 0.930, 0.851, 0.634, 0.684, and 0.869, respectively, which showed that components of $\boldsymbol{b}_i$ were positively correlated, where r_{jk} is the correlation coefficient of b_{ij} and b_{ik}; (ii) the number of positive nodes, the number of initial cycles, the reintroduction of CMF, residence

Table 2 Bayesian estimates (BEs) and 95% confidence intervals (CIs) of parameters in the longitudinal model of the IBCSG data.

	Appetite	Perceived coping	Mood	Physical well-being
Par.	BE (95%CI)	BE (95%CI)	BE (95%CI)	BE (95%CI)
Intercept	3.444 (3.246, 3.642)	5.411 (5.148, 5.674)	4.568 (4.407, 4.729)	4.327 (4.131, 4.523)
#Positive nodes > 4	−0.058 (−0.209, 0.093)	0.134 (−0.089, 0.357)	0.025 (−0.130, 0.180)	−0.007 (−0.142, 0.128)
#Initial cycle	0.085 (−0.091, 0.261)	0.166 (−0.134, 0.466)	0.197 (0.023, 0.371)	0.173 (−0.023, 0.369)
Reintroduction	−0.045 (−0.217, 0.127)	0.061 (−0.208, 0.330)	0.023 (−0.155, 0.201)	−0.087 (−0.263, 0.089)
#INIR	−0.063 (−0.294, 0.168)	−0.063 (−0.457, 0.331)	−0.119 (−0.352, 0.114)	−0.039 (−0.288, 0.210)
Residence: Switzerland	−0.048 (−0.193, 0.097)	0.160 (−0.110, 0.430)	−0.018 (−0.173, 0.137)	−0.077 (−0.244, 0.090)
Residence: Sweden	0.103 (−0.122, 0.328)	0.157 (−0.147, 0.461)	0.454 (0.211, 0.697)	0.264 (0.011, 0.517)
#Age > 40	0.297 (0.162, 0.432)	0.114 (−0.149, 0.377)	0.391 (0.250, 0.532)	0.287 (0.113, 0.461)
ER (1 = positive)	−0.044 (−0.215, 0.127)	−0.007 (−0.225, 0.211)	−0.131 (−0.292, 0.030)	−0.146 (−0.277,−0.015)
Time (in decades)	−0.284 (−0.892, 0.324)	−5.233 (−6.364,−4.102)	−0.839 (−1.211,−0.467)	0.536 (0.087, 0.985)
Covariate matrix $\boldsymbol{\Omega}$	0.498 (0.390, 0.606)	0.492 (0.355, 0.629)	0.514 (0.406, 0.622)	0.535 (0.423, 0.647)
		1.700 (1.426, 1.974)	0.647 (0.506, 0.788)	0.794 (0.631, 0.957)
			0.613 (0.480, 0.746)	0.606 (0.473, 0.739)
				0.793 (0.650, 0.936)

Notes #INIR: interaction of the number of initial cycles and reintroduction.

Table 3 Bayesian estimates (BEs) and 95% confidence intervals (CIs) for parameters in the survival model of the IBCSG data.

	DFS	OS
	BE (95%CI)	BE (95%CI)
Appetite	−0.986 (−1.733,−0.239)	−1.249 (−2.156,−0.342)
Perceived coping	−0.669 (−0.957,−0.381)	−0.703 (−1.115,−0.291)
Mood	−3.123 (−3.774,−2.472)	−3.778 (−4.548,−3.008)
Physical well-being	4.442 (3.913, 4.971)	5.242 (4.630, 5.854)
#Positive nodes > 4	1.706 (1.183, 2.229)	1.991 (1.350, 2.632)
#Initial cycle	0.050 (−0.405, 0.505)	0.143 (−0.427, 0.713)
Reintroduction	0.063 (−0.290, 0.416)	0.029 (−0.422, 0.480)
#INIR	−0.279 (−0.798, 0.240)	−0.136 (−0.699, 0.427)
Residence: Switzerland	0.335 (0.023, 0.647)	−0.050 (−0.499, 0.399)
Residence: Sweden	0.342 (0.032, 0.652)	0.326 (−0.309, 0.961)
#Age > 40	−0.527 (−1.009,−0.045)	−0.322 (−0.818, 0.174)
ER(1 = positive)	−0.069 (−0.441, 0.303)	−0.381 (−0.851, 0.089)

Notes #INIR: interaction of the number of initial cycles and reintroduction.

from Switzerland as well as the interaction between the number of initial cycles, and the reintroduction of the procedure did not have effect on QOL because the 95% confidence intervals of these effects did not exclude zero; (iii) "#Age" was identified to be an important covariate having a significantly positive effect on QOL because their corresponding 95% confidence intervals did not include zero, which showed that younger patients (under 40 years) had a better QOL than older patients (over 40 years); (iv) "time" was detected to be an important covariate having a significantly negative effect on QOL except for appetite and physical well-being variables because the 95% confidence intervals of the effect excluded zero, which implied that patients' QOL could be improved after initial surgery; (v) patients living in Sweden have a better QOL than those living in Switzerland, Australia, and New Zealand because their estimated coefficients are positive.

For the bivariate survival model, it followed from Table 3 that (i) DFS and OS were consistently affected by the longitudinal QOL covariates (e.g., appetite, perceived coping, mood, physical well-being) as well as the number of positive nodes of the tumor > 4 because their corresponding confidence intervals excluded zero, (ii) covariates related to residence: Switzerland and residence: Sweden and #Age > 40 would only affect DFS, (iii) neither the number of the initial CMF cycles nor the reintroduction of another cycle or the estrogen receptor status would affect DFS and OS. To wit, patients having a better physical well-being, appetite, perceived coping, or mood were less likely to have either cancer relapse or death; patients having the number of positive nodes being less than 4 might have a relapse or not survive than those having the number of positive nodes being more than 4; younger patients were more likely to have a relapse or death than older patients in terms of DFS; patients from Switzerland and Sweden might have neither cancer relapse nor death in terms of DFS.

6 Discussion

This paper presented a novel semiparametric joint model for multivariate longitudinal and survival data by relaxing the commonly adopted normality assumption of the longitudinal outcomes and leaving the baseline hazard functions completely unspecified. The advantages of the proposed model include: (i) it enhances the modeling flexibility and allows practitioners to make statistical inference

on longitudinal and survival data in a wide variety of considerations; (ii) it can capture the feature of unimodal, bimodal, and multimodal distribution for the longitudinal outcomes in a SJMLS; (iii) it does not require specifying the mean and covariance matrix of normal distribution involved in a finite mixture of normal distributions but regards them as random parameters; (iv) it can be written as a hierarchical model that allows one to develop a computationally feasible algorithm via the MH within the Gibbs sampler; (v) it requires fewer knots than smoothing splines in approximating log baseline hazard functions and is easier to implement using a data augmentation algorithm; (vi) the computational burden is not heavy, for example, it takes about 300 s to run a replication in the above conducted simulation studies, and about 2 h to run the IBCSG data set.

Although BLasso method developed by Park and Casella (2008) has been extended to various models including semiparametric structural equation models (Guo et al., 2012) and linear regression models (Hans, 2009; Lykou and Ntzoufras, 2013), little work has been developed on a SJMLS. Motivated by a data set from a clinical trial conducted by IBCSG, we presented a BLasso method to simultaneously estimate parameters and implement both shrinkage and variable selection for the considered SJMLS. Our simulation results suggested that the proposed BLasso procedure worked well under our considered settings in the sense that (i) the absolute biases of Bayesian estimates of parameters were less than 0.1 and their corresponding RMS values were less than 0.15; (ii) the average frequencies of correctly identifying unimportant covariates were more than 90%. But more simulation studies found that Bayesian variable selection and estimation strongly depend on the censoring percentage. The proposed BLasso can be easily extended to a complicated SJMLS with ordinal and nonignorable missing data in the longitudinal measurements and nonparametric random effects that are commonly encountered in practice.

Future work with the proposed SJMLS or the above-mentioned complicated SJMLS includes (i) simultaneous selection of fixed and random effects via the boosting algorithm (Buhlmann and Hothorn, 2007; Hofner et al., 2013), which is an interesting topic; (ii) a robust inference procedure, which does not depend on the normality assumption of the random effects; (iii) nonlinear effects of the covariates on each of the models; (iv) more sophisticated spline models with knots automatically selected may be used to improve the performance of the proposed procedures.

Acknowledgments The authors are grateful to the Editor, an Associate Editor, and two referees for their valuable suggestions and comments that greatly improved the manuscript. Zhao's research was partly supported by the Research Grant Council of Hong Kong (No. 503513), the National Natural Science Foundation of China (No. 11371299), and The Hong Kong Polytechnic University. Tang's research was partly supported by grants from the National Science Fund for Distinguished Young Scholars of China (No. 11225103), the National Natural Science Foundation of China (No. 11561074), and the Scientific Research Innovation Team of Yunnan Province (No. 2015HC028).

Conflict of interest
The authors have declared no conflict of interest.

A Appendix

Bayesian inference on SJMLS

To simultaneously obtain Bayesian estimates of unknown parameters, baseline hazard functions and random effects and select covariates in the considered SJMLS, the Gibbs sampler is employed to draw a sequence of random observations from the joint posterior distribution $p(\boldsymbol{\theta}_Y, \boldsymbol{\theta}_T, \boldsymbol{\theta}_\varepsilon, \boldsymbol{b}|\boldsymbol{D}_o)$. The block Gibbs sampler is conducted by iteratively sampling observations from the following conditional distributions: $p(\boldsymbol{\theta}_Y|\boldsymbol{\theta}_T, \boldsymbol{\theta}_\varepsilon, \boldsymbol{b}, \boldsymbol{D}_o)$, $p(\boldsymbol{\theta}_T|\boldsymbol{\theta}_Y, \boldsymbol{\theta}_\varepsilon, \boldsymbol{b}, \boldsymbol{D}_o)$, $p(\boldsymbol{\theta}_\varepsilon|\boldsymbol{\theta}_Y, \boldsymbol{b}, \boldsymbol{D}_o)$, and

$p(\boldsymbol{b}|\boldsymbol{\theta}_Y, \boldsymbol{\theta}_T, \boldsymbol{\theta}_\varepsilon, \boldsymbol{D}_o)$. The conditional distributions required in implementing the Gibbs sampler are presented as follows.

Block Gibbs Sampler (A): Conditional distribution related to $\boldsymbol{\theta}_y$

Let $\boldsymbol{\theta}_y = \{\beta, \Omega\}$, where $\beta = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ in which $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kr})^{\mathrm{T}}$ for $k = 1, \ldots, K$. It follows from equations (8), (10), and (11) that the conditional distribution $p(\boldsymbol{\beta}_k|\boldsymbol{\theta}_T, \boldsymbol{\theta}_\varepsilon, \boldsymbol{b}, \boldsymbol{D}_o)$ is proportional to

$$\exp\left\{\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\delta_{im}\psi_{mk}\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{R}_i(T_{im}) - \sum_{\ell=1}^{\mathcal{L}_m}\mathcal{D}_{im\ell}\right) - \frac{1}{2}\boldsymbol{\beta}_k^{\mathrm{T}}H_{\beta_k}^{-1}\boldsymbol{\beta}_k + \right.$$
$$\left. - \sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{1}{2\sigma_{L_{ij}}^{kk}}\left(y_{ijk} - \boldsymbol{\beta}_k^{T}\boldsymbol{R}_i(t_{ij}) - \boldsymbol{W}_{ki}^{\mathrm{T}}(t_{ij})\boldsymbol{b}_i\right)^2\right\},$$

which is not a familiar distribution, where $\sigma_{L_{ij}}^{kk}$ is the (k,k)-th element of covariance matrix $\Sigma_{L_{ij}}$, and $\boldsymbol{W}_{ki}^{\mathrm{T}}(t_{ij})$ is the k-th row vector of design matrix $W_i(t_{ij})$. Thus, it is rather difficult to directly sample observations from $p(\boldsymbol{\beta}_k|\boldsymbol{\theta}_T, \boldsymbol{\theta}_\varepsilon, \boldsymbol{b}, \boldsymbol{D}_o)$. Therefore, the well-known MH algorithm is adopted to simulate observations from the above conditional distribution, which is implemented as follows. Given the current value $\boldsymbol{\beta}_k^{(\ell)}$, a new candidate $\boldsymbol{\beta}_k$ is generated from the proposal distribution $N_p(\boldsymbol{\beta}_k^{(\ell)}, \sigma_{\beta_k}^2\Upsilon_{\beta_k})$ with $\sigma_{\beta_k}^2$ set to control the acceptance rate, and then the generated candidate $\boldsymbol{\beta}_k$ is accepted with probability

$$\min\left\{1, \frac{p(\boldsymbol{\beta}_k|\boldsymbol{\theta}_T, \boldsymbol{\theta}_\varepsilon, \boldsymbol{b}, \boldsymbol{D}_o)}{p(\boldsymbol{\beta}_k^{(\ell)}|\boldsymbol{\theta}_T, \boldsymbol{\theta}_\varepsilon, \boldsymbol{b}, \boldsymbol{D}_o)}\right\},$$

where $\Upsilon_{\beta_k}^{-1} = H_{\beta_k}^{-1} + \Sigma_{i=1}^{n}\Sigma_{m=1}^{M}\Sigma_{\ell=1}^{\mathcal{L}_m}\psi_{mk}^2\boldsymbol{R}_i(V_{im\ell}^*)\boldsymbol{R}_i(V_{im\ell}^*)^{\mathrm{T}}\mathcal{D}_{im\ell} + \Sigma_{i=1}^{n}\Sigma_{j=1}^{n_i}\boldsymbol{R}_i(t_{ij})\boldsymbol{R}_i(t_{ij})^{\mathrm{T}}/\sigma_{L_{ij}}^{kk}$ with $V_{im\ell}^* = 0.5(C_{m\ell} + C_{m,\ell-1})1(C_{m\ell} \le T_{im}) + 0.5(T_{im} + C_{m,\ell-1})1(C_{m,\ell-1} < T_{im} \le C_{m\ell})$.

From the prior distribution of Ω and the fact that $\boldsymbol{b}_i \sim N_q(\boldsymbol{0}, \Omega)$, it is easily shown that the conditional distribution of Ω is given by $p(\Omega|\boldsymbol{b}) \sim \mathrm{IW}_q(\varrho + n, \Sigma_{i=1}^{n}\boldsymbol{b}_i\boldsymbol{b}_i^{\mathrm{T}} + \boldsymbol{R}^0)$.

Block Gibbs Sampler (B): Conditional distribution related to $\boldsymbol{\theta}_T$

Let $\boldsymbol{\theta}_T = \{(\boldsymbol{\psi}_m, \boldsymbol{\gamma}_m, \boldsymbol{\varphi}_m) : m = 1, \ldots, M\}$. Then, $\boldsymbol{\psi}_m$, $\boldsymbol{\gamma}_m$, and $\boldsymbol{\varphi}_m$ can be iteratively sampled from their corresponding conditional distributions, which are given as follows. The conditional distribution $p(\boldsymbol{\psi}_m|\boldsymbol{\beta}, \boldsymbol{\gamma}_m, \boldsymbol{\varphi}_m, \boldsymbol{b}, \boldsymbol{D}_o)$ is proportional to

$$\exp\left\{\sum_{i=1}^{n}\left(\delta_{im}\boldsymbol{\psi}_m^{\mathrm{T}}\boldsymbol{\eta}(\boldsymbol{R}_i(T_{im}), W_i(T_{im}), \boldsymbol{b}_i) - \sum_{\ell=1}^{\mathcal{L}_m}\mathcal{D}_{im\ell}\right) - \frac{1}{2}\boldsymbol{\psi}_m^{\mathrm{T}}H_{\psi_m}^{-1}\boldsymbol{\psi}_m\right\},$$

where $\mathcal{D}_{im\ell}$ is defined in equation (11).

Conditional distribution $p(\boldsymbol{\gamma}_m|\boldsymbol{\beta}, \boldsymbol{\psi}_m, \boldsymbol{\varphi}_m, \boldsymbol{b}, \boldsymbol{D}_o)$ is proportional to

$$\exp\left\{\sum_{i=1}^{n}\left(\delta_{im}\boldsymbol{\gamma}_m^{\mathrm{T}}\boldsymbol{\xi}_i - \sum_{\ell=1}^{\mathcal{L}_m}\mathcal{D}_{im\ell}\right) - \frac{1}{2}\boldsymbol{\gamma}_m^{\mathrm{T}}H_{\gamma_m}^{-1}\boldsymbol{\gamma}_m\right\}.$$

Conditional distribution $p(\boldsymbol{\varphi}_m|\boldsymbol{\beta}, \boldsymbol{\psi}_m, \boldsymbol{\gamma}_m, \boldsymbol{b}, \boldsymbol{D}_o)$ is proportional to

$$\exp\left\{\sum_{i=1}^{n}\left(\delta_{im}\boldsymbol{\varphi}_m^{\mathrm{T}}\mathbb{B}_m(T_{im}) - \sum_{\ell=1}^{\mathcal{L}_m}\mathcal{D}_{im\ell}\right) - \frac{1}{2\varsigma_m^2}\boldsymbol{\varphi}_m^{\mathrm{T}}H_{\varphi_m}^{-1}\boldsymbol{\varphi}_m\right\},$$

where H_{φ_m} is a $(h_m + s + 1) \times (h_m + s + 1)$ second difference penalized matrix with rank $h_m + s - 1$ (Lang and Brezger, 2004). The MH algorithm for sampling observations from the above conditional distribution is similar to that for sampling $\boldsymbol{\beta}_k$. Thus, the details are omitted. The conditional distribution of ς_m^2 is given by $p(\varsigma_m^{-2}|\boldsymbol{\varphi}_m) \sim \text{Gamma}(a_\varsigma^m + 0.5(h_m + s - 1), b_\varsigma^m + 0.5\boldsymbol{\varphi}_m^\top H_{\varphi_m}^{-1}\boldsymbol{\varphi}_m)$.

Block Gibbs Sampler (C): Conditional distribution related to $\boldsymbol{b}$

Conditional distribution $p(\boldsymbol{b}_i|\boldsymbol{\theta}_Y, \boldsymbol{\theta}_T, \boldsymbol{\theta}_\varepsilon, \boldsymbol{D}_o)$ is proportional to

$$\exp\left\{\sum_{i=1}^{n}\sum_{m=1}^{M}\left(\delta_{im}\boldsymbol{\psi}_m^\top W_i(T_{im})\boldsymbol{b}_i - \sum_{\ell=1}^{\mathcal{L}_m}\mathcal{D}_{im\ell}\right) - \frac{1}{2}\boldsymbol{b}_i^\top\Omega^{-1}\boldsymbol{b}_i + \right.$$

$$\left. - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\boldsymbol{y}_{ij} - \boldsymbol{\eta}(\boldsymbol{R}_i(t_{ij}), \boldsymbol{b}_i))^\top\Sigma_{L_{ij}}^{-1}(\boldsymbol{y}_{ij} - \boldsymbol{\eta}(\boldsymbol{R}_i(t_{ij}), \boldsymbol{b}_i))\right\}.$$

Similarly, the MH algorithm is used to sample $\boldsymbol{b}_i$ from the above conditional distribution for $i = 1, \ldots, n$.

Block Gibbs Sampler (D): Conditional distribution related to $\boldsymbol{\theta}_\varepsilon$

Let $\boldsymbol{\theta}_\varepsilon$ denote all unknown parameters associated with distribution of $\boldsymbol{\varepsilon}_{ij}$, $\boldsymbol{\theta}_\varepsilon$ can be iteratively sampled by using the following steps:

Step (a). Let $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_G\}$ and $L = \{L_{ij} : i = 1, \ldots, n, j = 1, \ldots, n_i\}$. It is easily shown that the conditional distribution $p(\boldsymbol{\pi}|L, \tau)$ is distributed as the following generalized Dirichlet distribution: $p(\boldsymbol{\pi}|L, \tau) \sim \text{Dir}(a_1^*, b_1^*, \ldots, a_{G-1}^*, b_{G-1}^*)$, where $a_g^* = 1 + d_g$, $b_g^* = \tau + \sum_{j=g+1}^{G} d_j$ for $g = 1, \ldots, G-1$, and d_g is the number of L_{ij}'s whose value equals to g. Simulating observations from the conditional distribution $p(\boldsymbol{\pi}|L, \tau)$ can be implemented as follows. First, κ_g^* is independently generated from a Beta distribution $\text{Beta}(a_g^*, b_g^*)$. Then, $\pi_1, \ldots, \pi_G$ are obtained by

$$\pi_1 = \kappa_1^*,\ \pi_G = 1 - \sum_{g=1}^{G-1}\pi_g,\ \text{and}\ \pi_g = \prod_{j=1}^{g-1}\left(1 - \kappa_j^*\right)\kappa_g^*\ \text{for}\ g \neq 1\ \text{or}\ G.$$

Step (b). The conditional distribution of τ given $\boldsymbol{\pi}$ is given by $p(\tau|\boldsymbol{\pi}) \sim \text{Gamma}(a_1 + G - 1, a_2 - \sum_{g=1}^{G-1}\log(1 - \kappa_g^*))$.

Step (c). The conditional distribution of L_{ij} given $(\boldsymbol{\pi}, \boldsymbol{\mu}, \Omega, \boldsymbol{b})$ is given by

$$p(L_{ij}|\boldsymbol{\pi}, \boldsymbol{\mu}, \Omega, \boldsymbol{y}_{ij}) \overset{\text{i.i.d.}}{\sim} \text{Multinomial}\left(\pi_{ij1}^*, \ldots, \pi_{ijG}^*\right),$$

where π_{ijg}^* is proportional to $\pi_g p(\boldsymbol{y}_{ij}|\boldsymbol{\mu}_g, \Omega_g)$ with $\boldsymbol{y}_{ij}|\boldsymbol{\mu}_g, \Omega_g \sim N_K(\boldsymbol{\mu}_g, \Omega_g)$, $\Sigma = \{\Sigma_g : g = 1, \ldots, G\}$, and $\boldsymbol{\mu} = \{\boldsymbol{\mu}_g : g = 1, \ldots, G\}$.

Step (d). Let $L_1^*, \ldots, L_d^*$ be the d unique values of L_{ij}'s (i.e., unique number of "clusters"). The conditional distribution of $(\sigma_g^{kk})^{-1}$ $(k = 1, \ldots, K)$ is given by

$$\left(\sigma_g^{kk}\right)^{-1} \sim \text{Gamma}\left(c_1 + 0.5, c_2 + 0.5\left(\mu_g^k - \mu_\mu^k\right)^2\right)\ \text{for}\ g \notin \{L_1^*, \ldots, L_d^*\},$$

$$(\sigma_g^{kk})^{-1} \sim \text{Gamma}\left(c_1 + \frac{d_g + 1}{2}, c_2 + \frac{1}{2}\left\{(\mu_g^k - \mu_\mu^k)^2 + \mathcal{A}\right\}\right)\ \text{for}\ g \in \{L_1^*, \ldots, L_d^*\},$$

where $\mathcal{A} = \sum_{\{(i,j):L_{ij}=g\}}(\hat{\boldsymbol{\varepsilon}}_{ij} - \boldsymbol{\mu}_g)^2$, μ_g^k and μ_μ^k are the k-th element of vector $\boldsymbol{\mu}_g$ and $\boldsymbol{\mu}_\mu$, respectively.

Step (e). The conditional distribution of $\boldsymbol{\mu}_g$ is given by $\boldsymbol{\mu}_g|\boldsymbol{\mu}_\mu, \Sigma_g \sim N_K(\boldsymbol{\mu}_\mu, \Sigma_g)$ for $g \notin \{L_1^*, \ldots, L_d^*\}$, and

$$p(\boldsymbol{\mu}_g|\boldsymbol{\mu}_\mu, \boldsymbol{\Sigma}, L, \boldsymbol{\varepsilon}) \sim N_K\left(\frac{\sum_{\{(i,j):L_{ij}=g\}} \boldsymbol{\varepsilon}_{ij} + \boldsymbol{\mu}_\mu}{d_g + 1}, \frac{\Sigma_g}{d_g + 1}\right) \text{ for } g \in \{L_1^*, \ldots, L_d^*\},$$

where $\boldsymbol{\varepsilon} = \{\boldsymbol{\varepsilon}_{ij} : i = 1, \ldots, n, j = 1, \ldots, n_i\}$.

Step (f). The conditional distribution for $\boldsymbol{\mu}_\mu$ is given by $\boldsymbol{\mu}_\mu|\boldsymbol{\mu}_g, \Sigma_g \sim N_K(\mathbb{E}, \mathbb{F})$, where $\mathbb{F} = (\sum_{g=1}^G \Sigma_g^{-1} + H_\mu^{0-1})^{-1}$ and $\mathbb{E} = \mathbb{F}(H_\mu^{0-1}\boldsymbol{\zeta}_\mu^0 + \sum_{g=1}^G \Sigma_g^{-1}\boldsymbol{\mu}_g)$.

B Appendix

Sampling from the inverse Gaussian distribution

An inverse Gaussian distribution IG(a, b) (also known as the Wald distribution) with the mean $a > 0$ and the shape parameter $b > 0$ has the following probability density function $f(x; a, b) = \{b/(2\pi x^3)\}^{1/2} \exp\{-b(x-a)^2/(2a^2x)\}$ for $x > 0$. An algorithm (Michael et al., 1976) for simulating observation X from IG(a, b) is given as follows. First, we generate a random variable η^* from the standard normal distribution (e.g., $\eta^* \sim N(0, 1)$), and denote $v = a + a^2\eta^{*2}/(2b) - a/\sqrt{4ab\eta^{*2} + a^2\eta^{*4}}/(2b)$. Second, we sample a random number u from a uniform distribution $U(0, 1)$. Let $X = v$ if $u \leq a/(a + v)$, and $X = a^2/v$ otherwise.

C Appendix

Variables in IBCSG data

1. Four untransformed longitudinal QOL indicators
   y_1: physical well-being on a scale of zero (lousy) to hundred (good);
   y_2: mood on a scale of zero (miserable) to hundred (happy);
   y_3: appetite on a scale of zero (none) to hundred (good);
   y_4: perceived coping (how much effort does it cost you to cope with your illness?) on a scale of zero (a great deal) to hundred (none).
2. Observed event times in survival submodel
   T_{i1}: the monitored disease-free survival time, abbreviated as "DFS";
   T_{i2}: the monitored overall survival time, abbreviated as "OS".
3. Covariates in the considered SJMLS
   R_{i1}: the number of positive nodes of the tumor, abbreviated as "#Positive nodes";
   R_{i2}: three versus six initial cycles of oral cyclophosphamide, methotrexate, and fluorouracil (CMF), abbreviated as "#Initial cycle";
   R_{i3}: the reintroduction of three single courses of delayed chemotherapy, abbreviated as "Reintroduction";
   R_{i4}: the interaction of the number of initial cycles and reintroduction, abbreviated as "#INIR";
   R_{i5}: whether the residence is Switzerland, abbreviated as "Residence: Switzerland";
   R_{i6}: whether the residence is Sweden, abbreviated as "Residence: Sweden";
   R_{i7}: the age of premenopausal woman, abbreviated as "#Age";
   R_{i8}: the estrogen receptor (ER) status (negative/positive), abbreviated as "ER."

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Bae, K. and Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* **20**, 3423–3430.

Baghfalaki, T., Ganjali, M. and Berridge, D. (2013). Robust joint modeling of longitudinal measurements and time to event data using normal/independent distributions: a Bayesian approach. *Biometrical Journal* **55**, 844–865.

Baghfalaki, T., Ganjali, M. and Hashemi, R. (2014). Bayesian joint modeling of longitudinal measurements and time-to-event data using robust distributions. *Journal of Biopharmaceutical Statistics* **24**, 834–855.

Buhlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics* **34**, 559–583.

Buhlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statistical Science* **22**, 477–505.

Buhlmann, P. and Yu, B. (2003). Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association* **98**, 324–339.

Chi, Y. Y. and Ibrahim, J. G. (2006). Joint models for multivariate longitudinal and multivariate survival data. *Biometrics* **62**, 432–445.

Chow, S. M., Tang, N. S., Yuan, Y., Song, X. Y. and Zhu, H. T. (2011). Bayesian estimation of semiparametric nonlinear dynamic factor analysis models using the Dirichlet process prior. *British Journal of Mathematical and Statistical Psychology* **64**, 69–106.

De Gruttola, V. and Tu, X. M. (1994). Modeling progression of CD4–lymphocyte count and its relationship to survival time. *Biometrics* **50**, 1003–1014.

Eilers, P. and Marx, B. (1996). Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). *Statistical Science* **11**, 89–121.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

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

Gelman, A., Meng, X. L. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Guo, R., Zhu, H., Chow, S. M. and Ibrahim, J. G. (2012). Bayesian lasso for semiparametric structural equation models. *Biometrics* **68**, 567–577.

Hans, C. (2009). Bayesian Lasso regression. *Biometrika* **96**, 835–845.

He, Z., Tu, W., Wang, S., Fu, H. and Yu, Z. (2015). Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics* **71**, 178–187.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32**, 1–51.

Hofner, B., Hothorn, T. and Kneib, T (2013). Variable selection and model choice in structural survival models. *Computational Statistics* **28**, 1079–1101.

Hofner, B., Hothorn, T., Kneib, T. and Schmid, M. (2011). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics* **34**, 559–583.

Hu, W. H., Li, G. and Li, N. (2009). A Bayesian approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in Medicine* **29**, 1601–1619.

Huang, Y. X., Dagne, G. and Wu, L. (2010). Bayesian inference on joint models of HIV dynamics for time-to-event and longitudinal data with skewness and covariate measurement errors. *Statistics in Medicine* **30**, 2930–2946.

Huang, Y. X., Hu, X. J. and Dagne, G. A. (2014). Jointly modeling time-to-event and longitudinal data: a Bayesian approach. *Statistical Methods & Applications* **23**, 95–121.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2002). *Bayesian Survival Analysis*. Springer-Verlag, New York.

Ibrahim, J. G., Chu, H. and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology* **28**, 2796–2801.

Ishwaran, H. and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87**, 371–390.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.
Lee, S. Y. and Tang, N. S. (2006). Bayesian analysis of nonlinear structural equation models with nonignorable missing data. *Psychometrika* **71**, 541–564.
Li, N., Elashoff, R. M., Li, G. and Tseng, C. H. (2012). Joint analysis of bivariate longitudinal ordinal outcomes and competing risks survival times with nonparametric distributions for random effects. *Statistics in Medicine* **31**, 1707–1721.
Lykou, A. and Ntzoufras, I. (2013). On Bayesian lasso variable selection and the specification of the shrinkage parameter. *Statistics and Computing* **23**, 361–390.
Michael, J. R., Schucany, W. R. and Haas, R. W. (1976). Generating random variates using transformations with multiple roots. *American Statistician* **30**, 88–90.
Müller, P., Erkanli, A. and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.
Ohlssen, D. I., Sharples, L. D. and Spiegelhalter, D. J. (2007). Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. *Statistics in Medicine* **26**, 2088–2112.
Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.
Rizopoulos, D. and Ghosh, P. (2011). A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine* **30**, 1366–1380.
Rizopoulos, D., Hatfield, L. A., Carlin, B. P. and Takkenberg, J. J. M. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *Journal of the American Statistical Association* **109**, 1385–1397.
Rizopoulos, D., Verbeke, G. and Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society Series B* **71**, 637–654.
Song, X. and Wang, C. Y. (2008). Semiparametric approaches for joint modeling of longitudinal and survival data with time-varying. *Biometrics* **64**, 557–566.
Tang, N. S., Tang, A. M. and Pan, D. D. (2014). Semiparametric Bayesian joint models of multivariate longitudinal and survival data. *Computational Statistics & Data Analysis* **77**, 113–129.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.
Tsiatis, A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica* **14**, 809–834.
Tsiatis, A. A., Degruttola, V. and Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error: applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association* **90**, 27–37.
Wang, Y. and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association* **96**, 895–905.
Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B* **68**, 49–67.
Zhu, H. T., Ibrahim, J. G., Chi, Y. Y. and Tang, N. S. (2012). Bayesian influence measures for joint models for longitudinal and survival data. *Biometrics* **68**, 954–964.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

