Published in final edited form as:
Lifetime Data Anal. 2008 December ; 14(4): 496–520. doi:10.1007/s10985-008-9101-5.

Bayesian variable selection for the Cox regression model with missing covariates

Joseph G. Ibrahim,
Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA, e-mail:ibrahim@bios.unc.edu

Ming-Hui Chen, and
Department of Statistics, University of Connecticut, Storrs, CT 06269, USA, e-mail: mhchen@stat.uconn.edu

Sungduk Kim
Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, NIH, Rockville, MD 20852, USA, e-mail: kims2@mail.nih.gov

Abstract

In this paper, we develop Bayesian methodology and computational algorithms for variable subset selection in Cox proportional hazards models with missing covariate data. A new joint semi-conjugate prior for the piecewise exponential model is proposed in the presence of missing covariates and its properties are examined. The covariates are assumed to be missing at random (MAR). Under this new prior, a version of the Deviance Information Criterion (DIC) is proposed for Bayesian variable subset selection in the presence of missing covariates. Monte Carlo methods are developed for computing the DICs for all possible subset models in the model space. A Bone Marrow Transplant (BMT) dataset is used to illustrate the proposed methodology.

Keywords

Conjugate prior; Deviance information criterion; Missing at random; Proportional hazards models

1 Introduction

Bayesian variable selection in survival analysis is still one of the most challenging problems encountered in practice due to issues regarding (i) prior elicitation, (ii) evaluation of a model selection criterion due to the complication of censoring, and (iii) numerical computation of the criterion for all possible models in the model space. In the context of survival analysis, these issues have been discussed in Ibrahim et al. (1999a, 2001a) and the many references therein. There have been numerous papers in the statistical literature on Bayesian variable selection and model comparison, including articles by George and McCulloch (1993, 1997); Laud and Ibrahim (1995); George et al. (1996); Raftery (1996); Smith and Kohn (1996); Raftery et al. (1997); Brown et al. (1998, 2002); Clyde (1999); Chen et al. (1999, 2003, 2008); Dellaportas and Forster (1999); Chipman et al. (1998, 2001, 2003); George (2000); George and Foster (2000); Ibrahim et al. (2000); Ntzoufras et al. (2003) and Clyde and George (2004). However, the literature on Bayesian variable selection in the presence of

Correspondence to: Joseph G. Ibrahim.

missing data and in particular, for survival data in the presence of missing covariates, is still quite sparse. Part of the reason for this is that in the presence of missing covariate data, models can become quite complex and closed forms are not even available in the simplest of models. Thus, computing quantities such as Bayes factors, posterior model probabilities, the Aikiake Information Criterion (AIC) (Akaike 1973), the Bayesian Information Criterion (BIC) (Schwarz 1978), and Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002), for example, become serious computational challenges. For example, to compute BIC in the presence of missing covariate data, one would need to maximize the observed data likelihood. There are two challenging issues with this: (i) the observed data likelihood does not have a closed form for most models, even the linear model when the covariates are not normally distributed, and suitable approximation is often not available, and (ii) maximizing the observed data likelihood can be a huge challenge even if it is available in closed form. There are also several technical issues for computing AIC and BIC in the presence of missing covariates. One could argue that these measures are not well defined in the context of missing covariate data since the penalty term is not clearly defined. In particular, if we use the observed data likelihood obtained by averaging over the possible missing values of the covariates according to the missing covariate distribution, it is not clear how to appropriately define the dimensional penalty for AIC and BIC. We elaborate more on this issue in Sect. 5.

This issue becomes even more complex when computing Bayes factors, as one has to integrate over a very large space and the integrals easily become of very high dimension even in the simplest missing data problems. Specifically, it is well known that methods based on Bayes factors or posterior model probabilities, proper prior distributions are needed. It is a major task to specify prior distributions for all models in the model space, especially if the model space is large. For survival models with missing covariates, it becomes even more challenging to specify prior distributions, as in this case, one needs to specify priors not only for the regression coefficients in the survival model, but also the parameters involved in the models for missing covariates. The prior elicitation issue has been discussed in detail by several authors including Laud and Ibrahim (1995); Chen et al. (1999) and Ibrahim and Chen (2000). In addition, it is well known that Bayes factors and posterior model probabilities are generally sensitive to the choices of prior hyperparameters, and thus one cannot simply select vague proper priors to get around the elicitation issue. Even when informative prior distributions are available, computing Bayes factors and posterior model probabilities is difficult and expensive as one needs to compute prior and posterior normalizing constants for each model in the model space. It may be practically infeasible to compute these quantities in the context of variable subset selection for survival models with missing data. Alternatively, criterion based methods can be attractive in the sense that they do not require proper prior distributions in general, and thus have an advantage over posterior model probabilities in this sense. Several recent papers advocating the use of Bayesian criteria for model assessment include Geisser and Eddy (1979); Gelfand et al. (1992); Gelfand and Dey (1994); Ibrahim and Laud (1994); Laud and Ibrahim (1995); Gelman et al. (1996); Dey et al. (1997); Gelfand and Ghosh (1998); Ibrahim et al. (2001b); Spiegelhalter et al. (2002); Chen et al. (2004); Huang et al. (2005); Hanson (2006); Celeux et al. (2006) and Kim et al. (2007).

To overcome some of the methodologic and computational issues mentioned above, we develop two methodologies in this paper: (i) a class of semi-conjugate priors in the presence of MAR covariate data, and (ii) a variation of DIC for survival models with missing covariates. The proposed class of priors overcome the elicitation issues mentioned above as well as the computational challenges. The proposed priors make elicitation easier than other conventional informative priors by basing the elicitation on observable quantities rather than the parameters themselves, along with a scalar quantifying the confidence in that prediction.

This is an especially attractive approach in variable selection contexts since in this context the regression coefficients for every model in the model space have a different contextual meaning and interpretation, and thus specifying hyperparameters for all of the models in the model space is a monumental task. This elicitation challenge can be overcome by focusing on constructing a prior based on a prediction for the response variable, as pointed out by Laud and Ibrahim (1995) and Chen and Ibrahim (2003). They are also computationally attractive in that they lead to full conditionals that are log-concave and hence easily sampled via Adaptive Rejection Sampling (ARS) (Gilks and Wild 1992) within Gibbs. Thus, sampling the posterior with these priors is computationally very efficient.

The proposed version of DIC is an extension of a version of DIC discussed in Huang et al. (2005) for generalized linear models with missing covariates. For survival data with censored observations and missing covariates, DIC has a computational advantage over other criterion-based methods, such as AIC or BIC. With the computational methods developed in Sect. 4, the DIC measures can be easily computed for all models in the model space for a moderate number of covariates. In contrast, computation of AIC or BIC becomes quite difficult and challenging for variable subset selection for survival data with censored observations and missing covariates.

The rest of this paper is organized as follows. Section 2 presents a detailed development of the semi-conjugate prior under the piecewise exponential model in the presence of MAR covariates. Section 3 sets up all necessary formulas for the survival models, priors, and posteriors in the context of variable subset selection and presents a novel version of DIC for survival data with missing covariates. Section 4 presents the computational algorithms for computing the DIC measures for all models in the model space. A detailed analysis of the BMT data is given in Sect. 5. We conclude the article with brief remarks in Sect. 6.

2 The model, prior and posterior

2.1 The model

Let y_i denote the minimum of the censoring time C_i and the survival time T_i, and let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})'$ be the $k \times 1$ vector of covariates associated with y_i for the ith subject. Denote by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ the $k \times 1$ vector of regression coefficients. Also, $\nu_i = 1\{T_i = y_i\}$ is the indicator for the event for $i = 1, 2, \ldots, n$, where n is the total number of observations. As usual, we assume throughout that $\boldsymbol{x}_i$ does not include an intercept, since the intercept is not estimable in the Cox proportional hazards model, and that given $\boldsymbol{x}_i$, T_i and C_i are independent. In the presence of missing covariates, the missing data mechanism is defined as the distribution of the $k \times 1$ random vector $\boldsymbol{r}_i = (r_{i1}, r_{i2}, \ldots, r_{ik})'$, where $r_{ij} = 0$ when x_{ij} is missing and $r_{ij} = 1$ when x_{ij} is observed for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$. We assume that any missingness in covariates x_{ij} is missing at random (MAR) (Rubin 1976; Little and Rubin 2002). As discussed in Ibrahim et al. (2005), for MAR covariates x_{ij} we do not need to model the missing data mechanism.

We consider the Cox proportional piecewise exponential hazards model for $[y_i|\boldsymbol{x}_i]$, which has the survival function given by

$$S(y_i|\boldsymbol{x}_i, \beta, \boldsymbol{\lambda}) = \exp\{-\exp(\boldsymbol{x}_i'\beta) H_0(y_i|\boldsymbol{\lambda})\}, \tag{2.1}$$

where $H_0(t|\boldsymbol{\lambda})$ is the baseline cumulative hazard function. The piecewise exponential model is assumed for the baseline hazard function $h_0(t)$. Specifically, we first partition the time axis into J intervals: $(s_0, s_1]$, $(s_1, s_2]$, …, $(s_{J-1}, s_J]$, where $s_0 = 0 < s_1 < s_2 < \cdots < s_J$. In practice, it is sufficient to choose s_J to be greater than the largest follow-up time. We then assume a constant hazard λ_j over the jth interval $I_j = (s_{j-1}, s_j]$. That is, $h_0(y)=\lambda_j$ if $y \in I_j$ for $j = 1, 2, \ldots, J$. Then the corresponding baseline cumulative hazard function, $H_0(y|\boldsymbol{\lambda})$, is given by

$$H_0(y|\boldsymbol{\lambda})=\lambda_j(y - s_{j-1})+\sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1}) \tag{2.2}$$

for $s_{j-1} < y \le s_j$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)$. We note that when $J = 1$, $H_0(y|\boldsymbol{\lambda})$ reduces to the parametric exponential model.

We write $\boldsymbol{x}_i'=(\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i}')'$, where $\boldsymbol{x}_{1i}$ is a $k_1 \times 1$ vector of covariates that are observed for all n observations, $\boldsymbol{x}_{2i}$ is a $k_2 \times 1$ vector of covariates that have at least one missing value in the n observations, and $k_1 + k_2 = k$ with $k_1 \ge 0$ and $k_2 \ge 1$. Furthermore, we let $\boldsymbol{x}_{2i,mis}$ denote the vector of covariates that are missing for the ith case and let $\boldsymbol{x}_{2i,obs}$ be the vector of covariates that are observed for the ith case. Let $D = \{(y_i, \nu_i, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i,mis}, \boldsymbol{x}_{2i,obs}), i = 1, 2, \ldots, n\}$ denote the complete data. Then, the complete data likelihood function is given by

$$\begin{aligned} L(\beta,\boldsymbol{\lambda}|D)= & \prod_{i=1}^{n}\prod_{j=1}^{J}\{\lambda_j\exp(\boldsymbol{x}_i'\beta)\}^{\delta_{ij}\nu_i} \\ & \times\exp\left[-\delta_{ij}\exp(\boldsymbol{x}_i'\beta)\left\{\lambda_j(y_i - s_{j-1})+\sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right\}\right], \end{aligned} \tag{2.3}$$

where $\delta_{ij} = 1$ if the ith subject failed or was right censored in the jth interval $(s_{j-1}, s_j]$, and 0 otherwise.

2.2 Prior and posterior

We first specify a prior distribution for $(\boldsymbol{\beta}, \boldsymbol{\lambda})$. To this end, we extend the conjugate prior proposed by Chen and Ibrahim (2003) for the generalized linear model (GLM) to the piecewise exponential model in (2.1). Let X denote the $n \times k$ matrix with its ith row equal to $\boldsymbol{x}_i'$. Given X, we propose a semi-conjugate prior as follows:

$$\begin{aligned} \pi(\beta,\boldsymbol{\lambda}|\boldsymbol{y}_0,X,a_0) \propto & \left(\prod_{i=1}^{n}\prod_{j=1}^{J}\{\lambda_j\exp(\boldsymbol{x}_i'\beta)\}^{a_0\delta_{0ij}}\right. \\ & \left.\times\exp\left[-a_0\delta_{0ij}\exp(\boldsymbol{x}_i'\beta)\left\{\lambda_j(y_{0i} - s_{j-1})+\sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right\}\right]\right)\pi_0(\boldsymbol{\lambda}), \end{aligned} \tag{2.4}$$

where $a_0 > 0$ is a scalar prior parameter, $\boldsymbol{y}_0 = (y_{01}, \ldots, y_{0n})'$ is an $n \times 1$ vector of prior parameters, $\delta_{0ij} = 1$ if $s_{j-1} < y_{0i} \le s_j$ and 0 otherwise, and $\pi_0(\boldsymbol{\lambda})$ is an initial prior for $\boldsymbol{\lambda}$.

The prior (2.4) is called semi-conjugate since, by ignoring $\pi_0(\boldsymbol{\lambda})$, the prior has an identical form as the complete data likelihood given in (2.3). As discussed in Chen and Ibrahim (2003), y_{0i} can be viewed as a prior prediction for the marginal mean of y_i. Since y_{0i} is the prior prediction of y_i, we assume that y_{0i} is an "observed" failure time. Thus, in eliciting $\boldsymbol{y}_0$, we must focus on a prediction (or guess) for $E(y_i)$, which narrows the possibilities for choosing y_{0i}. To obtain a noninformative prior for $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, we specify all the y_{0i} equal. As shown in Chen and Ibrahim (2003), this specification under the GLM yields a prior in which the prior modes of the slopes in the regression model are the same. For the piecewise exponential model, we consider $y_{01} = \cdots = y_{0n} = y_0$, where $0 < y_0 \leq s_1$. Under this specification of $\boldsymbol{y}_0$, (2.4) reduces to

$$\pi(\beta, \boldsymbol{\lambda}|y_0, X, a_0) \propto \prod_{i=1}^{n} \{\lambda_1 \exp(\boldsymbol{x}_i'\beta)\}^{a_0} \exp\{-a_0 y_0 \lambda_1 \exp(\boldsymbol{x}_i'\beta)\}\pi_0(\boldsymbol{\lambda}). \tag{2.5}$$

We further specify $\pi_0(\boldsymbol{\lambda})$ as follows

$$\pi_0(\boldsymbol{\lambda}) \propto \frac{1}{\lambda_1} \prod_{j=2}^{J} \lambda_j^{b_1 - 1} \exp(-b_2 \lambda_j), \tag{2.6}$$

where $b_1 > 0$ and $b_2 > 0$. Note that in (2.5), we assume an improper uniform initial prior for $\boldsymbol{\beta}$ and an improper Jeffreys-type initial prior for λ_1. Thus, $\pi_0(\boldsymbol{\lambda})$ introduced in (2.4) and further specified in (2.6) is an improper prior. However, under some mild conditions, the prior (2.5) is proper and $(\log \lambda_1, \boldsymbol{\beta})$ has a prior mode of $(-\log y_0, 0, \ldots, 0)'$. We formally characterize these properties in the following theorem.

Theorem 2.1—*Let X_{obs} denote a submatrix of X with rows consisting of those completely observed $\boldsymbol{x}_i$'s and $\boldsymbol{x}_{mis} = (\boldsymbol{x}_{2i,mis}', i = 1, 2, \ldots, n)'$. Also, let $X_{obs}^* = (\mathbf{1}, X_{obs})$. Assume that X_{obs}^* is of full rank $(k + 1)$ and $\pi_0(\boldsymbol{\lambda})$ is given by (2.6). Then, for any given $\boldsymbol{x}_{mis}$, (i) $(\log \lambda_1, \boldsymbol{\beta})$ has a unique prior mode of $(-\log y_0, 0, \ldots, 0)'$ and (ii) $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|y_0, X, a_0)$ is proper.*

The proof of Theorem 2.1 is given in Appendix A. We note that the conditions of Theorem 2.1 require at least $k + 1$ complete observations with linearly independent covariate vectors including an intercept. From Theorem 2.1, we see that when $y_{01} = \cdots = y_{0n} = y_0$, the prior mode of $\boldsymbol{\beta}$ is $\mathbf{0}$ and with this prior prediction for the y_i, both $\boldsymbol{\beta}$ and λ_1 are identifiable in the sense that the joint prior is proper. Note that if we assume a general gamma prior instead of a Jeffreys-type prior for λ_1 in (2.6), we can show that the prior mode of $\boldsymbol{\beta}$ is still $\mathbf{0}$, but the prior mode of $\log \lambda_1$ is no longer $-\log y_0$. Thus, a different specification of the initial prior for λ_1 only changes the prior mode of the “intercept” in the survival model. Although we assume $y_0 \leq s_1$ in Theorem 2.1, we can show that the prior mode of $\boldsymbol{\beta}$ is still $\mathbf{0}$ even when $y_0 > s_1$. This is intuitively appealing since, in this case, the prior prediction y_{0i} does not depend on the ith subject’s specific covariate information. We further note that the parameter a_0 in (2.4) or (2.5) can be generally viewed as a precision parameter that quantifies the strength or confidence of our prior belief in $\boldsymbol{y}_0$. From Theorem 2.1, we see that the prior mode of $\boldsymbol{\beta}$ does not depend on a_0. Thus, a_0 controls only the prior precision of $\boldsymbol{\beta}$. This is an attractive feature that allows us to do sensitivity analyses by varying a_0 in the prior.

Next, we specify the distribution for the missing covariates. Since we are primarily interested in inferences about $\boldsymbol{\beta}$, we only need to model $\boldsymbol{x}_{2i}$ since $\boldsymbol{x}_{1i}$ is observed for all n observations. Therefore, we model $\boldsymbol{x}_{2i}$ conditioning on the completely observed covariates $\boldsymbol{x}_{1i}$ throughout. Using a sequence of one-dimensional conditional distributions proposed by Lipsitz and Ibrahim (1996) and Ibrahim et al. (1999b), we specify the distribution of the k_2-dimensional covariate vector $\boldsymbol{x}_{2i} = (x_{2i1}, x_{2i2}, \ldots, x_{2ik_2})'$ as

$$\begin{aligned} f(\boldsymbol{x}_{2i}|\boldsymbol{x}_{1i}, \alpha) = \; & f(x_{2i1}|\boldsymbol{x}_{i1}, \alpha_1) f(x_{2i2}|x_{2i1}, \boldsymbol{x}_{i1}, \alpha_2) \ldots \\ & f(x_{2ik_2}|x_{2i,k_2-1}, \ldots, x_{2i1}, \boldsymbol{x}_{i1}, \alpha_{k_2}), \end{aligned} \tag{2.7}$$

where $\boldsymbol{\alpha}_l$ is a vector of parameters for the lth conditional distribution, the $\boldsymbol{\alpha}_l$'s are distinct, and moreover, $\alpha=(\alpha_1', \alpha_2', \ldots, \alpha_{k_2}')'$. To complete the prior specification, we take independent priors for $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$ so that

$$\pi(\alpha)=\prod_{l=1}^{k_2}\pi(\alpha_l). \tag{2.8}$$

Let $\boldsymbol{x}_{obs}=((\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i,obs}'), i=1, 2, \ldots, n)'$. Using (2.5)–(2.8), the joint prior for $\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{x}_{mis}$, and $\boldsymbol{\alpha}$ is given by

$$\pi(\beta, \boldsymbol{\lambda}, \boldsymbol{x}_{mis}, \alpha|y_0, a_0, \boldsymbol{x}_{obs}) \propto \left[\prod_{i=1}^{n}\{\lambda_1 \exp(\boldsymbol{x}_i'\beta)\}^{a_0} \exp\{-a_0 y_0 \lambda_1 \exp(\boldsymbol{x}_i'\beta)\}\right] \times \left[\prod_{i=1}^{n} f(\boldsymbol{x}_{2i}|\boldsymbol{x}_{1i}, \alpha)\right] \pi_0(\boldsymbol{\lambda})\pi(\alpha). \tag{2.9}$$

Let $D_{obs} = (\boldsymbol{y}, \boldsymbol{\nu}, \mathbf{x}_{obs})$ denote the completely observed data, where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ and $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots, \nu_n)'$. Then, the joint posterior distribution is given by

$$\pi(\beta, \boldsymbol{\lambda}, \boldsymbol{x}_{mis}, \alpha|y_0, a_0, D_{obs}) \propto L(\beta, \boldsymbol{\lambda}|D)\pi(\beta, \boldsymbol{\lambda}, \boldsymbol{x}_{mis}, \alpha|y_0, a_0, \boldsymbol{x}_{obs}), \tag{2.10}$$

where $L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D)$ and $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{x}_{mis}, \boldsymbol{\alpha}|y_0, a_0, \boldsymbol{x}_{obs})$ are given by (2.3) and (2.9), respectively. Although the posterior distribution in (2.10) is analytically intractable, a Gibbs sampling algorithm can be easily developed to sample from this posterior distribution. The implementational details of the Gibbs sampling algorithm are discussed in Appendix B.

3 Bayesian variable subset selection

Let $\mathscr{M}$ denote the model space. We enumerate the models in $\mathscr{M}$ by $m = 1, 2, \ldots, \mathscr{K}$, where $\mathscr{K}$ is the dimension of $\mathscr{M}$ and model $\mathscr{K}$ denotes the full model. Also, let $\boldsymbol{\beta}^{(\mathscr{K})} = (\beta_1, \beta_2, \ldots, \beta_k)'$ denote the regression coefficients for the full model including an intercept, and let $\boldsymbol{x}_i^{(m)}$ and $\boldsymbol{\beta}^{(m)}$ denote $k_m \times 1$ vectors of covariates and regression coefficients for model m, and a specific choice of k_m covariates. We write $\boldsymbol{x}_i=(\boldsymbol{x}_i^{(m)\prime}, \boldsymbol{x}_i^{(-m)\prime})'$, and $\boldsymbol{\beta}^{(\mathscr{K})} = (\boldsymbol{\beta}^{(m)\prime}, \boldsymbol{\beta}^{(-m)\prime})'$, where $\boldsymbol{x}_i^{(-m)}$ is $\boldsymbol{x}_i$ with $\boldsymbol{x}_i^{(m)}$ deleted, and $\boldsymbol{\beta}^{(-m)}$ is $\boldsymbol{\beta}^{(\mathscr{K})}$ with $\boldsymbol{\beta}^{(m)}$ deleted. We also write $\boldsymbol{x}_{1i}=(\boldsymbol{x}_{1i}^{(m)\prime}, \boldsymbol{x}_{1i}^{(-m)\prime})'$ and $\boldsymbol{x}_{2i}=(\boldsymbol{x}_{2i}^{(m)\prime}, \boldsymbol{x}_{2i}^{(-m)\prime})'$, where $\boldsymbol{x}_{1i}^{(m)}$ is a $k_{1m}(\leq k_1)$ dimensional vector, $\boldsymbol{x}_{2i}^{(m)}$ is a k_{2m} $(\leq k_2)$ dimensional vector, and $\boldsymbol{x}_{1i}^{(-m)}$ and $\boldsymbol{x}_{2i}^{(-m)}$ are $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ with $\boldsymbol{x}_{1i}^{(m)}$ and $\boldsymbol{x}_{2i}^{(m)}$ deleted, respectively. Furthermore, we write $\boldsymbol{x}_{2i,mis}=(\boldsymbol{x}_{2i,mis}^{(m)\prime}, \boldsymbol{x}_{2i,mis}^{(-m)\prime})'$ and $\boldsymbol{x}_{2i,obs}=(\boldsymbol{x}_{2i,obs}^{(m)\prime}, \boldsymbol{x}_{2i,obs}^{(-m)\prime})'$, where $\boldsymbol{x}_{2i,mis}^{(-m)}$ and $\boldsymbol{x}_{2i,obs}^{(-m)}$ are $\boldsymbol{x}_{2i,mis}$ and $\boldsymbol{x}_{2i,obs}$ with $\boldsymbol{x}_{2i,mis}^{(m)}$ and $\boldsymbol{x}_{2i,obs}^{(m)}$ deleted, respectively.

Under model m, let $D_m=\{(y_i, \nu_i, \boldsymbol{x}_{1i}^{(m)}, \boldsymbol{x}_{2i,mis}^{(m)}, \boldsymbol{x}_{2i,obs}^{(m)}), i=1, 2, \ldots, n\}$ denote the complete data and then the complete data likelihood function is given by

$$L(\beta^{(m)}, \boldsymbol{\lambda}|D_m) = \prod_{i=1}^{n}\prod_{j=1}^{J}\left\{\lambda_j \exp((\boldsymbol{x}_i^{(m)})'\beta^{(m)})\right\}^{\delta_{ij}\nu_i} \times \exp\left[-\delta_{ij}\exp\left\{(\boldsymbol{x}_i^{(m)})'\beta^{(m)}\right\}\left\{\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right\}\right], \tag{3.1}$$

where δ_{ij} is defined in (2.3). Using exactly the same order of the sequence of one-dimensional conditional distributions for the covariates $\boldsymbol{x}_{2i}$ in (2.7) by deleting $\boldsymbol{x}_{2i}^{(-m)}$, we specify the distribution of the k_{2m}-dimensional covariate vector $\boldsymbol{x}_{2i}^{(m)} = (x_{2i1}^{(m)}, x_{2i2}^{(m)}, \ldots, x_{2ik_{2m}}^{(m)})'$ as

$$f(\boldsymbol{x}_{2i}^{(m)}|\boldsymbol{x}_{1i}^{(m)}, \alpha^{(m)}) = f(x_{2i1}^{(m)}|\boldsymbol{x}_{i1}^{(m)}, \alpha_1^{(m)}) f(x_{2i2}^{(m)}|x_{2i1}^{(m)}, \boldsymbol{x}_{i1}^{(m)}, \alpha_2^{(m)}) \times \cdots \times f(x_{2ik_{2m}}^{(m)}|x_{2i,k_{2m}-1}^{(m)}, \ldots, x_{2i1}^{(m)}, \boldsymbol{x}_{i1}^{(m)}, \alpha_{k_{2m}}^{(m)}), \tag{3.2}$$

where $\alpha^{(m)} = (\alpha_1^{(m)\prime}, \alpha_2^{(m)\prime}, \ldots, \alpha_{k_{2m}}^{(m)\prime})'$. It is important to note that in (3.2), $\boldsymbol{\alpha}^{(m)}$ is a subvector of $\boldsymbol{\alpha}$ in (2.7). We further write $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^{(m)\prime}, \boldsymbol{\alpha}^{(-m)\prime})'$ where $\boldsymbol{\alpha}^{(-m)}$ is $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}^{(m)}$ deleted. Similar to (2.8), the prior for $\boldsymbol{\alpha}^{(m)}$ is specified as $\pi(\alpha^{(m)}) = \prod_{l=1}^{k_{2m}} \pi(\alpha_l^{(m)})$.

Let $\boldsymbol{x}_{obs}^{(m)} = ((\boldsymbol{x}_{1i}^{(m)\prime}, \boldsymbol{x}_{2i,obs}^{(m)\prime}), i = 1, 2, \ldots, n)'$ and $\boldsymbol{x}_{mis}^{(m)} = (\boldsymbol{x}_{2i,mis}^{(m)\prime}, i = 1, 2, \ldots, n)'$. By applying the semi-conjugate prior (2.5) to model m, we have the joint prior for $\boldsymbol{\beta}^{(m)}$, $\boldsymbol{\lambda}$, $\boldsymbol{x}_{mis}^{(m)}$, and $\boldsymbol{\alpha}^{(m)}$ given by

$$\pi\left(\beta^{(m)}, \boldsymbol{\lambda}, \boldsymbol{x}_{mis}^{(m)}, \alpha^{(m)}|y_0, a_0, \boldsymbol{x}_{obs}^{(m)}\right) \propto \left(\prod_{i=1}^{n}[\lambda_1 \exp\{\boldsymbol{x}_i^{(m)\prime}\beta^{(m)}\}]^{a_0} \exp[-a_0 y_0 \lambda_1 \exp\{\boldsymbol{x}_i^{(m)\prime}\beta^{(m)}\}]\right) \times \left[\prod_{i=1}^{n} f\left(\boldsymbol{x}_{2i}^{(m)}|\boldsymbol{x}_{1i}^{(m)}, \alpha^{(m)}\right)\right] \pi_0(\boldsymbol{\lambda})\pi(\alpha^{(m)}), \tag{3.3}$$

where $\pi_0(\boldsymbol{\lambda})$ and $f(\boldsymbol{x}_{2i,mis}^{(m)}, \boldsymbol{x}_{2i,obs}^{(m)}|\boldsymbol{x}_{1i}^{(m)}, \alpha^{(m)})$ are defined by (2.6) and (3.2), respectively. Note that all models in the model space share the same prior for $\boldsymbol{\lambda}$, that is, the prior for $\boldsymbol{\lambda}$ is the same for all models in the model space. Let $D_{m,obs} = (\boldsymbol{y}, \nu, \boldsymbol{x}_{obs}^{(m)})$ denote the completely observed data. Under model m, the joint posterior distribution is given by

$$\pi(\beta^{(m)}, \boldsymbol{\lambda}, \boldsymbol{x}_{mis}^{(m)}, \alpha^{(m)}|y_0, a_0, D_{m,obs}) \propto L(\beta^{(m)}, \boldsymbol{\lambda}|D_m) \times \pi(\beta^{(m)}, \boldsymbol{\lambda}, \boldsymbol{x}_{mis}^{(m)}, \alpha^{(m)}|y_0, a_0, \boldsymbol{x}_{obs}^{(m)}), \tag{3.4}$$

where $L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\lambda}|D_m)$ and $\pi(\beta^{(m)}, \boldsymbol{\lambda}, \boldsymbol{x}_{mis}^{(m)}, \alpha^{(m)}|y_0, a_0, \boldsymbol{x}_{obs}^{(m)})$ are given by (3.1) and (3.3), respectively.

We carry out Bayesian variable selection via DIC, originally proposed by Spiegelhalter et al. (2002). The use of DIC for missing data models has been discussed in detail in Celeux et al. (2006). Let $\theta^{(m)}=(\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis}^{(m)})$. DIC is defined as follows:

$$\mathrm{DIC}_m=\mathrm{Dev}_m(\overline{\theta}^{(m)})+2p_m, \tag{3.5}$$

where $\mathrm{Dev}_m(\boldsymbol{\theta}^{(m)})$ is a deviance function and $\bar{\boldsymbol{\theta}}^{(m)}$ is the posterior mean of $\boldsymbol{\theta}^{(m)}$. In (3.5), p_m is the effective number of model parameters, which is calculated as

$$p_m=\overline{\mathrm{Dev}}_m(\theta^{(m)})-\mathrm{Dev}_m(\overline{\theta}^{(m)}), \tag{3.6}$$

where

$$\overline{\mathrm{Dev}}_m(\theta^{(m)})=E[\,\mathrm{Dev}_m(\theta^{(m)})|D_{m,obs}] \tag{3.7}$$

and the expectation is taken with respect to the posterior distribution given in (3.4). Since we are primarily interested in inferences about the survival model, we define the deviance function, $\mathrm{Dev}_m(\boldsymbol{\theta}^{(m)})$ in (3.5) as follows:

$$\mathrm{Dev}_m(\theta^{(m)})=-2\log L(\beta^{(m)},\boldsymbol{\lambda}|D_m),$$

where $L(\boldsymbol{\beta}^{(m)},\boldsymbol{\lambda}|D_m)$ is given by (3.1). Following Huang et al. (2005), we compute $\mathrm{Dev}_m(\boldsymbol{\theta}^{-(\mathrm{m})})$ as

$$\begin{aligned}\mathrm{Dev}_m(\overline{\theta}^{(m)})=-2\sum_{i=1}^{n}\sum_{j=1}^{J}\ \Big(&\delta_{ij}\nu_i[\,\log E[\,\lambda_j|D_{m,obs}]+E[\,(\boldsymbol{x}_i^{(m)})'\beta^{(m)}|D_{m,obs}]\\&-\delta_{ij}\exp\{E[\,(\boldsymbol{x}_i^{(m)})'\beta^{(m)}|D_{m,obs}]\}\Big\{E[\,\lambda_j|D_{m,obs}]\\&\times(y_i-s_{j-1})+\sum_{g=1}^{j-1}E[\,\lambda_g|D_{m,obs}](s_g-s_{g-1})\Big\}\Big),\end{aligned} \tag{3.8}$$

where all expectations are taken with respect to the posterior distribution in (3.4). In (3.8), instead of computing $(E[\,\boldsymbol{x}_i^{(m)}|D_{m,obs}])'E[\beta|D_{m,obs}]$, we compute $E[\,(\boldsymbol{x}_i^{(m)})'\beta^{(m)}|D_{m,obs}]$ in the presence of missing covariates, which yields a more appropriate dimensional penalty term p_m.

The DIC defined above is a Bayesian measure of predictive model performance, which is decomposed into a measure of fit and a measure of model complexity (p_m). The smaller the value of DIC, the better the model will predict new observations generated in the same way as the data. As discussed and shown in Chen et al. (2008), the performance of DIC is similar to AIC. Moreover, the DIC defined in (3.5) has a nice computational property for Bayesian variable selection, which will be discussed in detail in the next section.

4 Computation of DIC measures

To carry out Bayesian variable selection, we need to compute DIC_m in (3.5) for $m = 1, 2, \ldots, \mathscr{K}$. Due to the complexity of the survival model in (3.1), analytical evaluation of DIC_m does not appear possible. Thus, a Monte Carlo (MC) method is needed to compute all DIC$_m$'s in the model space. To this end, we propose two approaches for computing the DIC$_m$'s. The first approach, called "the direct sampling method", is based on direct Monte Carlo samples from each model in the model space. The second approach is "the single MC sample method," which was proposed by Chen et al. (2008). The latter method requires only one Markov chain Monte Carlo (MCMC) sample from the posterior distribution under the full model and computes the Bayesian criterion simultaneously for all possible subset models in the model space. From (3.7) and (3.8), we observe that for DIC$_m$ in (3.5), we need to compute the following quantities: (i) $E[\mathrm{Dev}_m(\boldsymbol{\theta}^{(m)})|D_{m,obs}]$; (ii) $E[(\boldsymbol{x}_i^{(m)})'\beta^{(m)}|D_{m,obs}]$; and (iii) $E[\lambda_j|D_{m,obs}]$ for $j = 1, 2, \dots, J$. For (ii), we note that when $\boldsymbol{x}_i^{(m)}$ is completely observed, then $E[(\boldsymbol{x}_i^{(m)})'\beta^{(m)}|D_{m,obs}]=(\boldsymbol{x}_i^{(m)})'E[\beta^{(m)}|D_{m,obs}]$. Thus, for (ii), we may further consider (iia) $E[\boldsymbol{\beta}^{(m)}|D_{m,obs}]$ and (iib) $E[(\boldsymbol{x}_i^{(m)})'\beta^{(m)}|D_{m,obs}]$ with at least one missing covariate in $\boldsymbol{x}_i^{(m)}$. It is interesting to observe that there is a common feature among (i), (iia), (iib), and (iii). That is, all of these quantities can be written as

$$g_m=E[g(\theta^{(m)})|D_{m,obs}], \tag{4.1}$$

for various functions g, where $\theta^{(m)}=(\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis}^{(m)})$ and the expectation is taken with respect to the joint posterior distribution in (3.4) under model m.

First, we discuss the direct sampling method. Using the Gibbs sampling algorithm given in Appendix B, we generate a Monte Carlo sample $\{\theta_q^{(m)}, q=1,2,\dots,Q\}$ from the joint posterior distribution in (3.4) under model m. Then, a Monte Carlo estimate of g_m is given by

$$\widehat{g}_m=\frac{1}{Q}\sum_{q=1}^{Q}g(\theta_q^{(m)}) \tag{4.2}$$

for all g's listed in (i)–(iii). Then, plugging various $\hat{g}_m$'s in (3.5) gives a Monte Carlo estimate of DIC$_m$.

Next, we discuss the single MC sample method. Using the notation given in Sect. 3, we write $\gamma=(\beta,\boldsymbol{\lambda},\boldsymbol{x}_{mis},\alpha)$, $\gamma^{(m)}=(\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis}^{(m)},\alpha^{(m)})$, and $\gamma^{(-m)}=(\beta^{(-m)},\boldsymbol{x}_{mis}^{(-m)},\alpha^{(-m)})$, where $\boldsymbol{x}_{mis}^{(-m)}$ is $\boldsymbol{x}_{mis}$ with $\boldsymbol{x}_{mis}^{(m)}$ deleted and $\boldsymbol{\gamma}^{(-m)}$ is $\boldsymbol{\gamma}$ with $\boldsymbol{\gamma}^{(m)}$ deleted. Thus, the marginal likelihood under model m is given by

$$C_m=\int\pi^*(\gamma^{(m)}|y_0,a_0,D_{m,obs})d\gamma^{(m)}, \tag{4.3}$$

where

$$\begin{aligned}
&\pi^*(\gamma^{(m)}|y_0,a_0,D_{m,obs})\\
&=L(\beta^{(m)},\boldsymbol{\lambda}|D_m)\left(\prod_{i=1}^{n}[\lambda_1\exp\{\boldsymbol{x}_i^{(m)\prime}\beta^{(m)}\}]^{a_0}\exp[-a_0y_0\lambda_1\exp\{\boldsymbol{x}_i^{(m)\prime}\beta^{(m)}\}]\right)\\
&\quad\times\left[\prod_{i=1}^{n}f(\boldsymbol{x}_{2i}^{(m)}|\boldsymbol{x}_{1i}^{(m)},\alpha^{(m)})\right]\pi_0^*(\boldsymbol{\lambda})\pi(\alpha^{(m)}),
\end{aligned} \tag{4.4}$$

$\pi_0^*(\boldsymbol{\lambda})=\frac{1}{\lambda_1}\prod_{j=2}^{J}\lambda_j^{b_1-1}\exp(-b_2\lambda_j)$, and $L(\boldsymbol{\beta}^{(m)}, \boldsymbol{\lambda}|D_m)$, $\pi_0(\boldsymbol{\lambda})$ and $f(\boldsymbol{x}_{2i}^{(m)}|\boldsymbol{x}_{1i}^{(m)}, \alpha^{(m)})$ are defined by (3.1), (2.6) and (3.2), respectively. Then, for a given function g, we have

$$g_m=E[\,g(\theta^{(m)})|D_{m,obs}]=\int g(\theta^{(m)})\frac{\pi^*(\gamma^{(m)}|y_0,a_0,D_{m,obs})}{C_m}d\beta^{(m)}, \tag{4.5}$$

where C_m is defined in (4.3).

For any given function g such that $E[|g(\boldsymbol{\theta}^{(m)})||D_{m,obs}] < \infty$, we have the following identity

$$g_m=\frac{C_{\mathscr{K}}}{C_m}E\left[g(\theta^{(m)})w(\gamma^{(-m)}|\gamma^{(m)})\frac{\pi^*(\gamma^{(m)}|y_0,a_0,D_{m,obs})}{\pi^*(\gamma|y_0,a_0,D_{obs})}\middle|D_{obs}\right], \tag{4.6}$$

where $C_{\mathscr{K}}$ is the marginal likelihood given in (4.3) under the fill model, $\pi^*(\boldsymbol{\gamma}|y_0, a_0, D_{obs})$ is given in (4.4) corresponding to the full model, which is essentially the kernel of the joint posterior distribution in (2.10), and the expectation is taken with respect to the joint posterior distribution in (2.10) under the full model. In (4.6), $w(\boldsymbol{\gamma}^{(-m)}|\,\boldsymbol{\gamma}^{(m)})$ is a completely known conditional density, whose support is contained in, or equal to, the support of the conditional density of $\boldsymbol{\gamma}^{(-m)}$ given $\boldsymbol{\gamma}^{(m)}$ with respect to the joint posterior distribution in (2.10) under the full model.

Observe that as a special case of (4.1), we have $g_m = 1$ when $g \equiv 1$. Using this result, we further obtain that

$$\frac{C_m}{C_{\mathscr{K}}}=E\left[w(\gamma^{(-m)}|\gamma^{(m)})\frac{\pi^*(\gamma^{(m)}|y_0,a_0,D_{m,obs})}{\pi^*(\gamma|y_0,a_0,D_{obs})}\middle|D_{obs}\right]. \tag{4.7}$$

Using (4.6) and (4.7), we have

$$g_m=\frac{E\left[g(\theta^{(m)})w(\gamma^{(-m)}|\gamma^{(m)})\frac{\pi^*(\gamma^{(m)}|y_0,a_0,D_{m,obs})}{\pi^*(\gamma|y_0,a_0,D_{obs})}\middle|D_{obs}\right]}{E\left[w(\gamma^{(-m)}|\gamma^{(m)})\frac{\pi^*(\gamma^{(m)}|y_0,a_0,D_{m,obs})}{\pi^*(\gamma|y_0,a_0,D_{obs})}\middle|D_{obs}\right]}. \tag{4.8}$$

We note that as the dimension of $\boldsymbol{\lambda}$ does not change across all models, $\pi^*(\boldsymbol{\lambda})$ cancels out in the ratio $\frac{\pi^*(\gamma^{(m)}|y_0,a_0,D_{m,obs})}{\pi^*(\gamma|y_0,a_0,D_{obs})}$.

Let $\{\boldsymbol{\gamma}_q = (\boldsymbol{\beta}_q, \boldsymbol{\lambda}_q, \boldsymbol{x}_{mis,q}, \boldsymbol{\alpha}_q), q = 1, 2, \ldots, Q\}$ denote an MCMC sample of size Q from the joint posterior distribution (2.10) under the full model. Write $\gamma_q=(\gamma_q^{(m)},\gamma_q^{(-m)})$, where $\gamma_q^{(m)}=(\beta_q^{(m)},\boldsymbol{\lambda}_q,\boldsymbol{x}_{mis,q}^{(m)},\alpha_q^{(m)})$, and $\gamma_q^{(m)}=(\beta_q^{(-m)},\boldsymbol{x}_{mis,q}^{(-m)},\alpha_q^{(-m)})$. Also let $\theta_q^{(m)}=(\beta_q^{(m)},\boldsymbol{\lambda}_q,\boldsymbol{x}_{mis,q}^{(m)})$. Then, an MC estimate of g_m is given by

$$\widehat{g}_m=\frac{\sum_{q=1}^{Q}g(\theta_q^{(m)})w(\gamma_q^{(-m)}|\gamma_q^{(m)})\frac{\pi^*(\gamma_q^{(m)}|y_0,a_0,D_{m,obs})}{\pi^*(\gamma_q|y_0,a_0,D_{obs})}}{\sum_{q=1}^{Q}w(\gamma_q^{(-m)}|\gamma_q^{(m)})\frac{\pi^*(\gamma_q^{(m)}|y_0,a_0,D_{m,obs})}{\pi^*(\gamma_q|y_0,a_0,D_{obs})}}. \tag{4.9}$$

Under certain regularity conditions, such as ergodicity, we have

$$\lim_{Q\to\infty}\widehat{g}_m=g_m,$$

implying that $\hat{g}_m$ is consistent.

As shown in Chen et al. (2008) the optimal choice of $w(\boldsymbol{\gamma}^{(-m)} \mid \boldsymbol{\gamma}^{(m)})$ is the conditional posterior distribution of $\boldsymbol{\gamma}^{(-m)}$ given $\boldsymbol{\gamma}^{(m)}$ under the full model in the sense that $\hat{g}_m$ achieves the minimum asymptotic variance. However, the optimal choice of $w(\boldsymbol{\gamma}^{(-m)} \mid \boldsymbol{\gamma}^{(m)})$ is not computationally feasible. Thus, we propose the following weight function

$$w(\gamma^{(-m)}|\gamma^{(m)})=w(\beta^{(-m)}|\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis})w(\alpha^{(-m)}|\alpha^{(m)},\boldsymbol{x}_{mis})w(\boldsymbol{x}_{mis}^{(-m)}|\boldsymbol{x}_{mis}^{(m)},\alpha^{(m)}). \tag{4.10}$$

Note that in (4.10), when model m includes all missing covariates $\boldsymbol{x}_{2i}$, we do not need to compute $w(\boldsymbol{x}_{mis}^{(-m)}|\boldsymbol{x}_{mis}^{(m)},\alpha^{(m)})$ as in this case, $\boldsymbol{x}_{mis}^{(-m)}$ is a null vector in the sense that it has zero dimension. In (4.10), a good $w(\boldsymbol{\beta}^{(-m)}|\boldsymbol{\beta}^{(m)}, \boldsymbol{\lambda}, \boldsymbol{x}_{mis})$, which is close to the optimal choice, can be constructed based on the asymptotic approximation to the joint posterior posterior. Let $\hat{\boldsymbol{\beta}}^{(-m)}$ $(\boldsymbol{\lambda}, \boldsymbol{x}_{mis})$ denote the conditional posterior mode of $\boldsymbol{\beta}^{(-m)}$ given $\boldsymbol{\beta}^{(m)}$, $\boldsymbol{\lambda}$ and $\boldsymbol{x}_{mis}$ under the full model. Specifically, we first compute

$$\widehat{\beta}^{(-m)}(\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis})=\arg\max_{\beta^{(-m)}}\log\pi^*(\beta|\boldsymbol{\lambda},\boldsymbol{x}_{mis},y_0,a_0,D_{obs}), \tag{4.11}$$

where

$$\begin{aligned}&\log\pi^*(\beta|\boldsymbol{\lambda},\boldsymbol{x}_{mis},y_0,a_0,D_{obs})\\&=\sum_{i=1}^{n}\sum_{j=1}^{J}\left[\delta_{ij}\nu_i\boldsymbol{x}_i'\beta-\delta_{ij}\exp(\boldsymbol{x}_i'\beta)\left\{\lambda_j(y_i-s_{j-1})+\sum_{g=1}^{j-1}\lambda_g(s_g-s_{g-1})\right\}\right]\\&+\sum_{i=1}^{n}[a_0\boldsymbol{x}_i'\beta-a_0y_0\lambda_1\exp(\boldsymbol{x}_i'\beta)]\end{aligned} \tag{4.12}$$

and then compute

$$\begin{aligned}&\widehat{\Sigma}^{(-m)}(\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis})\\&=\left[-\frac{\partial^2\log\pi^*(\beta|\boldsymbol{\lambda},\boldsymbol{x}_{mis},y_0,a_0,D_{obs})}{\partial\beta^{(-m)}\partial\beta^{(-m)\prime}}\bigg|_{\beta^{(-m)}=\widehat{\beta}^{(-m)}(\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis})}\right]^{-1}.\end{aligned}$$

Thus, a good $w(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, \boldsymbol{\lambda}, \boldsymbol{x}_{mis})$ can be constructed as follows:

$$\begin{aligned}&w(\beta^{(-m)}|\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis})\\&=(2\pi)^{-\frac{k-k_m}{2}}|\widehat{\Sigma}^{(-m)}(\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis})|^{-\frac{1}{2}}\exp\left\{-\tfrac{1}{2}(\beta^{(-m)}-\widehat{\beta}^{(-m)}(\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis}))'\right.\\&\qquad\left.\times[\widehat{\Sigma}^{(-m)}(\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis})]^{-1}(\beta^{(-m)}-\widehat{\beta}^{(-m)}(\beta^{(m)},\boldsymbol{\lambda},\boldsymbol{x}_{mis}))\right\},\end{aligned} \tag{4.13}$$

which approximates the joint conditional posterior $\pi(\boldsymbol{\beta}^{(-m)} \mid \boldsymbol{\beta}^{(m)}, \boldsymbol{\lambda}, \boldsymbol{x}_{mis}, y_0, a_0, D_{obs})$ under the full model. Similarly, we can construct a good $w(\boldsymbol{\alpha}^{(-m)}|\boldsymbol{\alpha}^{(m)}, \boldsymbol{x}_{mis})$ in (4.10). For $w(\boldsymbol{x}_{mis}^{(-m)}|\boldsymbol{x}_{mis}^{(m)},\alpha^{(m)})$, we use a Monte Carlo estimate given by

$$w(\boldsymbol{x}_{mis}^{(-m)}|\boldsymbol{x}_{mis}^{(m)},\alpha^{(m)})=\frac{1}{Q}\sum_{q=1}^{Q}w(\boldsymbol{x}_{mis}^{(-m)}|\boldsymbol{x}_{mis}^{(m)},\alpha^{(m)},\alpha_q^{(-m)}), \tag{4.14}$$

where

$$w(\boldsymbol{x}_{mis}^{(-m)}|\boldsymbol{x}_{mis}^{(m)},\alpha^{(m)}\alpha_q^{(-m)}) \propto \prod_{i=1}^{n} f(\boldsymbol{x}_{2i}|\boldsymbol{x}_{1i},\alpha^{(m)},\alpha_q^{(-m)}),$$

and $f(\boldsymbol{x}_{2i}|\boldsymbol{x}_{1i},\alpha^{(m)},\alpha_q^{(-m)})$ is given by (2.7) under the full model.

5 Analysis of the BMT data

The BMT data set consists of $n = 2397$ cases who received HLA-identical sibling transplant from 1995 to 2004 for AML or ALL in CR1 (pre-transplant status = 1st complete remission) with graft source of BM or PB/PB+BM. Infants were excluded (age < 2 year old). The outcome variable, y_i in years, was the time from transplant to death or end of follow up, and ν_i denotes the censoring indicator which equals 1 if the ith subject died, and is 0 otherwise. The median follow-up was 5.1 years with interquartile range of 3.0 to 7.8 years. There were 904 deaths in the data set. We consider ten covariates: disease (disease type: AML, ALL), age, yeartx (transplant year), karnofprg (Karnofsky score at pre-transplant), gsource (graftype: BM, PB/PB+BM), sexmatch (Donor-Patient sex match: MM, MF, FM, FF), regimprg (conditioning regimen: CY+TBI±oth, TBI + other, Busulf + CY ± oth, Other/Unknown), prevgvh1 (GVHD prophylaxis: mtx ± other, csa ± other, mtx + csa ± other, tdep ± other, Other/Unknown), cytoabnew (cytogenetics: Poor, InterMed, Normal, Good), and wbcdx (WBC at diagnosis ($10^9/l$)). The covariates age, yeartx, karnofprg, and wbcdx are continuous, and the covariates disease and gsource are binary. We dichotomize sexmatch as sexmatch1, sexmatch2, and sexmatch3, where (sexmatch1, sexmatch2, sexmatch3) takes values (0, 0, 0), (1, 0, 0), (0, 1, 0), and (0, 0, 1), which correspond to MM, MF, FM, and FF, respectively. In exactly the same fashion, we dichotomize regimprg, prevgvh1, and cytoabnew as (regimprg1, regimprg2, regimprg3), (prevgvh11, prevgvh12, prevgvh13, prevgvh14) and (cytoabnew1, cytoabnew2, cytoabnew3). For instance, the values (0,0,0), (1,0,0), (0,1,0), and (0,0,1) for (cytoabnew1, cytoabnew2, cytoabnew3) correspond to Poor, InterMed, Normal, and Good for cytoabnew, respectively.

Let x_1 = disease, x_2 = age, x_3 = yeartx, x_4 = karnofprg, x_5 = gsource, $\boldsymbol{x}_6$ = (sexmatch1, sexmatch2, sexmatch3)′, $\boldsymbol{x}_7$ = (regimprg1, regimprg2, regimprg3)′ $\boldsymbol{x}_8$ = (prevgvh11, prevgvh12, prevgvh13, prevgvh14)′, $\boldsymbol{x}_9$=(sexmatch1, sexmatch2, sexmatch3)′ and x_{10} = log(wbcdx). For these 10 covariates, $x_1, x_2, \ldots, \boldsymbol{x}_8$ were completely observed for all cases and $\boldsymbol{x}_9$ and x_{10} had missing information. There were 488 (20.36%) individuals with cytogenetics ($\boldsymbol{x}_9$) missing and 230 (9.6%) individuals with WBC missing, and 96 individuals with both cytogenetics and WBC missing. Overall, there were 623 (25.99%) individuals with at least one covariate missing. We assume that the missing covariates are MAR. In all computations, we standardized all completely observed covariates.

For the BMT data, we fit the piecewise exponential model given by (2.1) and (2.2) for the outcome variable y_i, where s_j is chosen to be the (j/J)th quantile of the failure times y_i, for j =

1, 2, …, $J-1$. Since x_1, x_2, …, $\boldsymbol{x}_8$ are always observed, they do not need to be modeled, and thus we condition on those covariates throughout. We then use a proportional odds logistic regression model for $\boldsymbol{x}_9$ and a normal regression model for x_{10}. Specifically, under the full model with all ten covariates, $f(\boldsymbol{x}_9|x_1, x_2, \ldots, \boldsymbol{x}_8, \boldsymbol{\alpha}_9)$ is specified as follows:

$$\begin{array}{lcl}
P(\boldsymbol{x}_9 & = & (0,0,0)'|x_1,x_2,\ldots,\boldsymbol{x}_8,\alpha_9) \\
 & = & F(\alpha_{9,10}+\alpha_{91}x_1+\cdots+\alpha_{95}x_5+\alpha'_{96}\boldsymbol{x}_6+\alpha'_{97}\boldsymbol{x}_7+\alpha'_{98}\boldsymbol{x}_8), \\
P(\boldsymbol{x}_9 & = & (1,0,0)'|x_1,x_2,\ldots,\boldsymbol{x}_8,\alpha_9) \\
 & = & F(\alpha_{9,20}+\alpha_{91}x_1+\cdots+\alpha_{95}x_5+\alpha'_{96}\boldsymbol{x}_6+\alpha'_{97}\boldsymbol{x}_7+\alpha'_{98}\boldsymbol{x}_8) \\
 & & -F(\alpha_{9,10}+\alpha_{91}x_1+\cdots+\alpha_{95}x_5+\alpha'_{96}\boldsymbol{x}_6+\alpha'_{97}\boldsymbol{x}_7+\alpha'_{98}\boldsymbol{x}_8), \\
P(\boldsymbol{x}_9 & = & (0,1,0)'|x_1,x_2,\ldots,\boldsymbol{x}_8,\alpha_9) \\
 & = & F(\alpha_{9,30}+\alpha_{91}x_1+\cdots+\alpha_{95}x_5+\alpha'_{96}\boldsymbol{x}_6+\alpha'_{97}\boldsymbol{x}_7+\alpha'_{98}\boldsymbol{x}_8) \\
 & & -F(\alpha_{9,20}+\alpha_{91}x_1+\cdots+\alpha_{95}x_5+\alpha'_{96}\boldsymbol{x}_6+\alpha'_{97}\boldsymbol{x}_7+\alpha'_{98}\boldsymbol{x}_8),
\end{array}$$

and $P(\boldsymbol{x}_9=(0,0,1)'|x_1,x_2,\ldots,\boldsymbol{x}_8,\alpha_9)=1-F(\alpha_{9,30}+\alpha_{91}x_1+\cdots+\alpha_{95}x_5+\alpha'_{96}\boldsymbol{x}_6+\alpha'_{97}\boldsymbol{x}_7+\alpha'_{98}\boldsymbol{x}_8)$, where $F(u) = \exp(u)/\{1+\exp(u)\}$, $\alpha_{9,10} \le \alpha_{9,20} \le \alpha_{9,30}$, $\alpha'_{96}=(\alpha_{96,1},\alpha_{96,2},\alpha_{96,3})$, $\alpha'_{97}=(\alpha_{97,1},\alpha_{97,2},\alpha_{97,3})$, $\alpha'_{98}=(\alpha_{98,1},\alpha_{98,2},\alpha_{98,3},\alpha_{98,4})$ and $\alpha_9=(\alpha_{9,10},\alpha_{9,20},\alpha_{9,30},\alpha_{91},\ldots,\alpha_5,\alpha'_{96},\alpha'_{97},\alpha'_{98})'$. We note that $\alpha_{9,10}$, $\alpha_{9,20}$, and $\alpha_{9,30}$ are three intercepts in the proportional odds logistic regression model. Furthermore, $f(x_{10}|x_1, x_2, \ldots, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{\alpha}_{10})$ is taken to be the density of a $N(\alpha_{10,0}+\alpha_{10,1}x_1+\cdots+\alpha_{10,5}x_5+\alpha'_{10,6}\boldsymbol{x}_6+\alpha'_{10,7}\boldsymbol{x}_7+\alpha'_{10,8}\boldsymbol{x}_8+\alpha'_{10,9}\boldsymbol{x}_9,\alpha_{10,10})$ distribution, where $\alpha_{10,10} > 0$ denotes the variance. The prior for $(\boldsymbol{\beta}, \lambda)$ is given by (2.5) and (2.6). In (2.5), we consider several values for a_0 such as $a_0 = 0.1$, 0.01, 0.001, and 0.0001 and in (2.6), we take $b_1 = b_2 = 0.001$. For the parameters in the models for the missing covariates, an inverse gamma prior with scale and shape parameters equal to 0.001 is specified for $\alpha_{10,10}$, and independent normal priors, $N(0, 1000)$, are specified for all other parameters. We wish to compare the following $2^{10} = 1024$ models: no covariates (null model), (x_1), …, (x_{10}), (x_1, x_2), …, $(x_1, x_2, x_3, x_4, x_5, \boldsymbol{x}_6, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, x_{10})$ (full model). In all computations, the Gibbs sampling algorithm given in Appendix B was used to sample from the posterior distributions and 10,000 Gibbs samples after a burn-in of 1,000 iterations were used to compute all DIC measures and other posterior estimates. The convergence of the Gibbs sampling algorithm was checked using several diagnostic procedures as recommended by Cowles and Carlin (1996).

We first carry out the complete case (CC) analysis of the BMT data. There were $n^* = 1{,}774$ subjects with all ten covariates completely observed. In the CC analysis, we first perform subset variable selection using the AIC and BIC criteria, since with no missing data, these two criteria can be easily computed. Let $L_{cc}(\boldsymbol{\beta}^{(m)}, \boldsymbol{\lambda}|D_{cc,m})$ denote the likelihood function given in (3.1) with the completely observed data $D_{cc,m}$ under model m in a model space that consists of 2^{10} possible subset models. Then, AIC and BIC are given by

$$\mathrm{AIC}_m=-2\log L_{cc}(\widehat{\beta}^{(m)},\widehat{\boldsymbol{\lambda}}|D_{cc,m})+2p_m, \tag{5.1}$$

where $\hat{\boldsymbol{\beta}}^{(m)}$ and $\hat{\boldsymbol{\lambda}}$ are the maximum likelihood estimates of $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\lambda}$ and $p_m = k_m + J$, and

$$\mathrm{BIC}_m=-2L_{cc}(\widehat{\beta}^{(m)},\widehat{\boldsymbol{\lambda}}|D_{cc,m})+[\log(n^*)]p_m. \tag{5.2}$$

Table 1 shows the best three AIC or BIC models for $J = 10$, 15, and 20. From Table 1, we see that the best three AIC models are (x_1, x_2, x_4, $\boldsymbol{x}_9$), (x_1, x_2, x_4, x_5, $\boldsymbol{x}_9$), and (x_1, x_2, x_3, x_4, x_5, $\boldsymbol{x}_9$), and the best three BIC models are (x_1, x_2, $\boldsymbol{x}_9$), (x_1, x_2, x_4, $\boldsymbol{x}_9$), and (x_1, x_2, x_5, $\boldsymbol{x}_9$). In Table 1, under the model (x_1, x_2, x_3, x_4, x_5, $\boldsymbol{x}_9$), the values of $p_m = k_m + J = 8 + J$ are 18, 23, and 28 for $J = 10$, 15, and 20, respectively while the values of p_m become 15, 20, and 25 for $J = 10$, 15, and 20, respectively, under the model (x_1, x_2, $\boldsymbol{x}_9$). Thus, (x_1, x_2, $\boldsymbol{x}_9$) is the smallest model while (x_1, x_2, x_3, x_4, x_5, $\boldsymbol{x}_9$) is the largest model among the five models listed in Table 1. We note that the order of the best three models under either AIC or BIC remains the same for $J = 10$, $J = 15$, and $J = 20$. Thus, subset variable selection under both AIC and BIC is robust to the choice of J.We also see from Table 1 that the lowest values of AIC are attained at $J = 15$ for all five models while the lowest values of BIC are attained at $J = 10$. This result is expected since BIC favors smaller and more parsimonious models than AIC, due to a larger dimensional penalty imposed by BIC.

For the CC case, we used the direct sampling method to compute the DIC values for all 1024 models under various choices of J and a_0. The results based on the best three DIC models under various choices of J and a_0 are shown in Table 2. From Table 2, we see that when a_0 is small, for example, $a_0 = 0.001$ or $a_0 = 0.0001$, the DIC values are very close to the corresponding values of AIC and the values of p_m in DIC are also very close to those in AIC. We also observe that there is a convex pattern in the DICs as functions of J and a_0. Specifically, the DIC values with $J = 15$ are smaller than those with either $J = 10$ or $J = 20$ for all the best three models, and, in addition, the DIC values with $a_0 = 0.001$ are smaller than those with $a_0 = 0.1$, $a_0 = 0.01$, and $a_0 = 0.0001$ under these same models, though the DIC values with $a_0 = 0.001$ are close to those with $a_0 = 0.0001$. These results are quite desirable as they empirically show that DIC may be used to guide the choices of J and a_0 in achieving the best predictive model performance. In this CC case, among the values of J and a_0 being considered, based on the DIC measure, the best choices of J and a_0 are $J = 15$ and $a_0 = 0.001$. In the CC case, we also implemented the single MC sample method discussed in Sect. 4 for computing the DIC measures. Using a Gibbs sample of 10,000 iterations after a burn-in of 1,000 iterations from the posterior distribution under the full model, the Monte Carlo estimates of DIC_m and p_m are 3595.25 and 20.99 for model (x_1, x_2, x_4, $\boldsymbol{x}_9$), 3595.40 and 21.94 for model (x_1, x_2, x_4, x_5, $\boldsymbol{x}_9$), and 3595.74 and 22.99 for model (x_1, x_2, x_3, x_4, x_5, $\boldsymbol{x}_9$). These estimates are very similar to those given in Table 2 using the direct sampling method.

For the best two DIC models with $J = 15$ and $a_0 = 0.001$, we also computed the posterior means (Estimates), the posterior standard deviations (SD's), and 95% highest posterior density (HPD) intervals of the model parameters. The results are shown in Table 3. Under model (x_1, x_2, x_4, $\boldsymbol{x}_9$), all 95% HPD intervals do not contain 0, indicating the importance of all these covariates. The results given in Table 3 indicate that an ALL patient has a higher risk of death compared to an AML patient, an older patient has a higher risk of death, a higher Karnofsky score at pre-transplant leads to a lower risk of death, and a patient with poor cytogenetics is likely to have a high risk of death. Under model (x_1, x_2, x_4, x_5, $\boldsymbol{x}_9$), all covariates except for gsource have 95% HPD intervals that do not contain 0.

Next, we carry out an all case (AC) analysis of the BMT data, that is, an analysis incorporating all of the cases. In the AC case, due to the additional complication of modeling the missing covariates, AIC and BIC are computationally infeasible, as discussed earlier and in fact, one could even argue that these measures are not well defined here since the penalty term is not clearly defined. In particular, if we use the marginal likelihood $L(\hat{\boldsymbol{\beta}}^{(m)},\hat{\boldsymbol{\lambda}}|D_m)$ and then average over all of the possible missing values of the covariates according to the missing covariate distribution, it is not clear how to appropriately define the dimensional penalty p_m for AIC and BIC. Thus, for the AC case, we used DIC as the criterion for performing variable subset selection. To this end, we used the the direct sampling method to compute the DIC values for all 1,024 models under various choices J and a_0. The DIC values for the best three models are presented in Table 4. Note that models $(x_1, x_2, x_4, x_5, \boldsymbol{x}_8, \boldsymbol{x}_9)$ and $(x_1, x_2, x_4, x_5, \boldsymbol{x}_9)$ are consistently the best and second best models for all the values of J and a_0 considered in Table 4, while model $(x_1, x_2, x_3, x_4, x_5, \boldsymbol{x}_9)$ is the third best for most combinations of J and a_0 except for $(J, a_0) = (15, 0.1)$ and $(J, a_0) = (10, 0.001)$. For $(J, a_0) = (15, 0.1)$ the third best model is $(x_1, x_2, x_3, x_4, x_5, \boldsymbol{x}_9, x_{10})$ with $\text{DIC}_m = 4973.88$ and $p_m = 24.75$, and for $(J, a_0) = (10, 0.001)$ the third best model is $(x_1, x_2, x_4, x_5, \boldsymbol{x}_9, x_{10})$ with $\text{DIC}_m = 4743.10$ and $p_m = 22.46$. From Table 4, we also see that the second best DIC model $(x_1, x_2, x_4, x_5, \boldsymbol{x}_9)$ in the CC analysis remains the second best DIC model in the AC analysis. In the AC analysis, when $a_0 = 0.001$, the values of DIC_m and p_m for the best CC analysis model $(x_1, x_2, x_4, \boldsymbol{x}_9)$ now become 4744.66 and 20.71 for $J = 10$, 4731.20 and 25.69 for $J = 15$, and 4750.68 and 30.80 for $J = 20$, which are much larger than the corresponding DIC values under the best AC model $(x_1, x_2, x_4, x_5, \boldsymbol{x}_8, \boldsymbol{x}_9)$ and the second best AC model $(x_1, x_2, x_4, x_5, \boldsymbol{x}_9)$. When $a_0 = 0.001$, the best CC model $(x_1, x_2, x_4, \boldsymbol{x}_9)$ is the ninth best AC model for $J = 10$ and the tenth best model for both $J = 15$ and $J = 20$. Interestingly, similar to the CC analysis, the "optimal" choices of J and a_0 are $J = 15$ and $a_0 = 0.001$. Compared to the CC analysis, another noticeable change in the AC analysis is that the values of the dimensional penalty p_m are larger than the corresponding values in the CC analysis, which is expected since the dimension of those missing covariates leads to the additional dimensional penalty in p_m.

For the best two DIC models with $J = 15$ and $a_0 = 0.001$, we also computed the posterior estimates of the model parameters, and the results are shown in Table 5. Under both the best two DIC models, all covariates except for gsource in the survival model for the time from transplant to death have 95% HPD intervals that do not contain 0. As $\boldsymbol{x}_9$ is the only missing covariate in both models, using (3.2), we only need to model $\boldsymbol{x}_9$ via the proportional odds logistic regression model conditional on the other covariates, namely, disease, age, karnofprg, gsource, and prevgvh1 for model $(x_1, x_2, x_4, x_5, \boldsymbol{x}_8, \boldsymbol{x}_9)$ and disease, age, karnofprg, and gsource for model $(x_1, x_2, x_4, x_5, \boldsymbol{x}_9)$. The corresponding posterior estimates for these two missing covariate models are also shown in Table 5. Under these two models for the missing covariate cytoabnew $(\boldsymbol{x}_9)$, we see that all covariates except for karnofprg have 95% HPD intervals that do not contain 0. Under the second best model, Table 3 compares the posterior estimates from the AC analysis to those of the CC analysis. In Table 3, we see that the AC analysis leads to smaller posterior standard deviations and shorter HPD intervals for all parameters in the survival model. In particular, gsource is nearly "significant" in the response model and "significant" in the covariate model in the AC analysis, where significance means that the 95% HPD interval does not contain 0.

6 Discussion

We have proposed a joint semi-conjugate prior for the regression coefficients $\boldsymbol{\beta}$ and piecewise hazard parameters $\boldsymbol{\lambda}$ and examined their theoretical properties in the piecewise exponential model for right censored survival data. The proposed prior is quite attractive in the context of variable subset selection for survival data with missing covariates. It is proper and the functional form of the prior is immediately determined for all models once the functional form of the prior is written for the full model. In addition, the prior is completely specified by only one hyper-parameter, namely, a_0. This indeed makes the elicitation of priors for all models in the model space much easier. Otherwise, prior elicitation would be an enormous task. In addition, we have empirically shown that the DIC measure can be used to guide the choice of a_0 to achieve the best posterior predictive performance. In Sect. 5, for the BMT data, we see that the best model for the AC is different than the one based on a CC analysis. This empirical result demonstrates that one cannot do variable selection just based on the completely observed cases. In fact, it is important to use all cases in performing variable selection.

Our computational methods in this paper are intended for situations where the number of models in the model space can be enumerated, so with this in mind, our proposed procedure works best when the number of covariates is 9–15. We have considered two Monte Carlo methods for computing the DIC measures. The direct sampling method is easy to implement. However, care needs to be taken in monitoring convergence of the Gibbs sampling algorithm for each model in the model space. On the other hand, the single MC sample method requires only one Gibbs sample from the posterior distribution under the full model. Thus, one needs to monitor convergence of the Gibbs sampling algorithm only once. However, in this case, one needs to construct a “good” weight function $w(\gamma_q^{(-m)}|\gamma_q^{(m)})$ to obtain an efficient single MC sample method. The choice of $w(\gamma_q^{(-m)}|\gamma_q^{(m)})$ proposed in Sect. 4 works well. However, it requires finding the conditional posterior modes, which may be computationally expensive. Finding a less efficient but less computationally expensive weight function is an important future project, which is currently under investigation. We note that both Monte Carlo methods can be easily implemented using multiple computers. Thus, a parallel computing system can greatly speed up the computation of the DIC measures for variable selection. With a Linux cluster, the proposed computational procedure can work well when the number of covariates is up to 20.

Another important criterion used in model assessment is Bayesian Model Averaging (BMA). Since we have focused this paper on variable selection and selecting a set of top models, we have not addressed the issue of BMA at all, as this is an entirely different topic with different inferential goals and different computational strategies. The performance of the proposed semi-conjugate priors in the presence of MAR covariates and the effects of covariates such as sexmatch within the BMA context will be explored in future work.

Acknowledgments

The authors wish to thank Dr. Mei-Jie Zhang for providing the BMT data. The authors also wish to thank the Editor-in-Chief, the Editor, and a referee for their helpful comments and suggestions, which have improved the paper. This research was partially supported by NIH grants #GM 70335 and #CA 74015.

Appendix A: proofs of Theorem 2.1

Observe that the marginal prior of (λ_1, $\boldsymbol{\beta}$) is of the form

$$\pi(\lambda_1,\beta|y_0,X,a_0) \propto \prod_{i=1}^{n}\{\lambda_1\exp(\boldsymbol{x}_i'\beta)\}^{a_0}\exp\{-a_0y_0\lambda_1\exp(\boldsymbol{x}_i'\beta)\}\frac{1}{\lambda_1}.$$

Let $\beta_0 = \log \lambda_1$. We have

$$\pi(\beta_0,\beta|y_0,X,a_0) \propto \prod_{i=1}^{n}\left\{\exp(\beta_0+\boldsymbol{x}_i'\beta)\right\}^{a_0}\exp\left\{-a_0y_0\exp(\beta_0+\boldsymbol{x}_i'\beta)\right\}. \tag{A.1}$$

Since X_{obs}^{*} is of full rank, then $X^{*} = (\mathbf{1}, X)$ is of full rank. It is easy to show that $\pi(\beta_0, \boldsymbol{\beta}|y_0, X, a_0)$ is log-concave in $(\beta_0, \boldsymbol{\beta}')'$. This implies that $\pi(\beta_0, \boldsymbol{\beta}|y_0, X, a_0)$ has a unique mode. Set

$$\frac{\partial}{\partial(\beta_0,\beta')'}\log \pi(\beta_0,\beta|y_0,X,a_0)=a_0\sum_{i=1}^{n}[\,1-y_0\exp(\beta_0+\boldsymbol{x}_i'\beta)](1,\boldsymbol{x}_i')'=0. \tag{A.2}$$

Thus, $(-\log y_0, 0, \ldots, 0)'$ is the unique solution of (A.2). This implies that $(-\log y_0, 0, \ldots, 0)'$ is the unique prior mode of $(\log \lambda_1, \boldsymbol{\beta})$.

For (ii), it suffices to prove that $\pi(\beta_0, \boldsymbol{\beta}|y_0, X, a_0)$ in (A.1) is proper since $\pi_0(\lambda_2, \ldots, \lambda_J)$ is a proper prior. We write

$$\pi*(\beta_0,\beta|y_0,X,a_0)=\prod_{i=1}^{n}\{\exp(\beta_0+\boldsymbol{x}_i'\beta)\}^{a_0}\exp\{-a_0y_0\exp(\beta_0+\boldsymbol{x}_i'\beta)\}.$$

It is easy to observe that

$$\{\exp(\beta_0+\boldsymbol{x}_i'\beta)\}^{a_0}\exp\{-a_0y_0\exp(\beta_0+\boldsymbol{x}_i'\beta)\}\le K_0$$

for $i = 1, 2, \ldots, n$, where $K_0 > 0$ is a constant. Since X_{obs}^* is of full rank, there exist $i_1 < i_2 < \cdots < i_{k+1}$ such that $\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, \ldots, \boldsymbol{x}_{i_{k+1}}$ are completely observed and the $(k+1) \times (k+1)$ matrix $X^{**}=((1,\boldsymbol{x}_{i_g}'), g=1,2,\ldots,k+1)$ is of full rank. Let $\boldsymbol{u} = (u_1, u_2, \ldots, u_{k+1})'$. Taking a one-to-one transformation $\boldsymbol{u} = X^{**}(\beta_0, \boldsymbol{\beta}')'$ leads to

$$\begin{aligned}&\int\pi^*(\beta_0,\beta|y_0,X,a_0)d\beta_0d\beta\\&\le K_0^{n-k-1}\int\prod_{g=1}^{k=1}\{\exp(\beta_0+\boldsymbol{x}_{i_g}'\beta)\}^{a_0}\exp\{-a_0y_0\exp(\beta_0+\boldsymbol{x}_{i_g}'\beta)\}d\beta_0d\beta\\&=K_1\prod_{g=1}^{k=1}\int\exp(a_0u_g)\exp\{-a_0y_0\exp(u_g)\}du_g<\infty,\end{aligned} \tag{A.3}$$

which completes the proof of Theorem 2.1.

Appendix B: posterior sampling

In this appendix, we discuss how to sample from the posterior distribution under the full model given in (2.10). To this end, we propose a Gibbs sampling algorithm, which requires sampling from the following full conditional distributions in turn:

i. $[\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{x}_{mis}, a_0, D_{obs}]$;

ii. $[\boldsymbol{\lambda}|\boldsymbol{\beta}, \boldsymbol{x}_{mis}, a_0, D_{obs}]$;

iii. $[\boldsymbol{x}_{mis}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, a_0, D_{obs}]$;

iv. $[\boldsymbol{\alpha}|\boldsymbol{x}_{mis}, a_0, D_{obs}]$.

We briefly discuss how we sample from each of the above full conditional distributions. For (i), the full conditional density of $\boldsymbol{\beta}$ given $\boldsymbol{\lambda}$, $\boldsymbol{x}_{mis}$, a_0, and D_{obs} is of the form

$$\begin{aligned}\pi(\beta|\boldsymbol{\lambda}, \boldsymbol{x}_{mis}, a_0, D_{obs}) \propto {} & \prod_{i=1}^{n}\exp\{(\nu_i+a_0)\boldsymbol{x}_i'\beta - a_0 y_0 \lambda_1 \exp(\boldsymbol{x}_i'\beta)\} \\ & \times \exp\left[-\sum_{j=1}^{J}\delta_{ij}\exp\,(\boldsymbol{x}_i'\beta)\left\{\lambda_j(y_i - s_{j-1})+\sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right\}\right].\end{aligned}$$

It is easy to show that $\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{x}_{mis}, a_0, D_{obs})$ is log-concave in **β**. Thus, we can sample the β_j's via the adaptive rejection algorithm of Gilks and Wild (1992). For (ii), given **β** and $\boldsymbol{x}_{mis}$, λ_1, $\lambda_2, \ldots, \lambda_J$ are conditionally independent. Let $h_{ij}=\delta_{ij}(y_i - s_{j-1})+\sum_{g=j+1}^{J}\delta_{ig}(s_j - s_{j-1})$. Then, we have

$$\lambda_1 \sim \text{Gamma}\left(na_0+\sum_{i=1}^{n}\delta_{i1}\nu_{i,}\sum_{i=1}^{n}\left\{(a_0y_0+h_{i1})\exp(\boldsymbol{x}_i'\beta)\right\}\right), \tag{B.1}$$

and

$$\lambda_j \sim \text{Gamma}\left(b_1+\sum_{i=1}^{n}\delta_{ij}\nu_i, b_2+\sum_{i=1}^{n}h_{ij}\exp(\boldsymbol{x}_i'\beta)\right), \tag{B.2}$$

for $j = 2, \ldots, J$. Hence, sampling the λ_j from (B.1) and (B.2) is straightforward.

For (iii), given **β**, **λ**, and **α**, the $\boldsymbol{x}_{2i,mis}$'s are conditionally independent, and the conditional distribution for $\boldsymbol{x}_{2i,mis}$ is

$$\begin{aligned}\pi\,(\boldsymbol{x}_{2i,mis}|\beta, \boldsymbol{\lambda}, \alpha, a_0, D_{obs}) \propto {} & f(\boldsymbol{x}_{2i}|\boldsymbol{x}_{1i}, \alpha)\exp\{(\nu_i+a_0)\boldsymbol{x}_i'\beta\}\exp\{-a_0y_0\lambda_1 \\ & \times\exp(\boldsymbol{x}_i'\beta)\}\exp\left[-\sum_{j=1}^{J}\delta_{ij}\exp(\boldsymbol{x}_i'\beta)\left\{\lambda_j(y_i - s_{j-1})+\sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right\}\right].\end{aligned}$$

Thus, the conditional distribution of $\boldsymbol{x}_{2i,mis}$ depends on the form of $f(\boldsymbol{x}_{2i}|\boldsymbol{x}_{1i}, \boldsymbol{\alpha})$. In Sect. 5, for the BMT data, $f(\boldsymbol{x}_{2i}|\boldsymbol{x}_{1i}, \boldsymbol{\alpha})$ is a product of a proportional odds logistic density and a normal density, and hence, sampling $\boldsymbol{x}_{2i,mis}$ is relatively straightforward. In fact, the conditional distribution for (cytoabnew1, cytoabnew2, cytoabnew3) is multinomial while the conditional distribution for log(wbcdx) is log-concave, which can be sampled via the adaptive rejection algorithm of Gilks and Wild (1992). For (iv), the full conditional distribution is $\pi\,(\alpha|\boldsymbol{x}_{mis}, a_0, D_{obs}) \propto \prod_{i=1}^{n} f(\boldsymbol{x}_{2i}|\boldsymbol{x}_{1i}, \alpha)\pi(\alpha)$. For various covariate distributions specified through a series of one dimensional conditional distributions, sampling **α** is straightforward. For example, in Section 5, the full conditional distribution for each component of $\boldsymbol{\alpha}_9$ is log-concave, and hence we can sample these α_{9j}'s via the adaptive rejection algorithm of Gilks and Wild (1992), and the full conditional distributions for the components of $\boldsymbol{\alpha}_{10}$ are either normal or inverse gamma, which are easy to sample from.

References

Akaike, H. Information theory and an extension of themaximum likelihood principle. In: Petrov, BN.; Csaki, F., editors. International symposium on information theory; Budapest: Akademia Kiado; 1973. p. 267-281.

Brown PJ, Vanucci M, Fearn T. Multivariate Bayesian variable selection and prediction. J R Stat Soc B 1998;60:627–641.

Brown PJ, Vanucci M, Fearn T. Bayes model averaging with selection of regresors. J R Stat Soc B 2002;64:519–536.

Celeux G, Forbes F, Robert CP, Titterington DM. Deviance information criteria for missing data models (with discussion). Bayesian Anal 2006;1:651–674.

Chen MH, Ibrahim JG. Conjugate priors for generalized linear models. Stat Sinica 2003;13:461–476.

Chen MH, Ibrahim JG, Yiannoutsos C. Prior elicitation, variable selection, and Bayesian computation for logistic regression models. J R Stat Soc B 1999;61:223–242.

Chen MH, Ibrahim JG, Shao QM, Weiss RE. Prior elicitation for model selection and estimation in generalized linear mixed models. J Stat Plan Inference 2003;111:57–76.

Chen MH, Dey DK, Ibrahim JG. Bayesian criterion based model assessment for categorical data. Biometrika 2004;91:45–63.

Chen MH, Huang L, Ibrahim JG, Kim S. Bayesian variable selection and computation for generalized linear models with conjugate priors. Bayesian Anal 2008;3:585–614. [PubMed: 19436774]

Chipman HA, George IE, McCulloch RE. Bayesian CART model search (with discussion). J Am Stat Assoc 1998;93:935–960.

Chipman, HA.; George, IE.; McCulloch, RE. The practical implementation of Bayesian model selection (with discussion). In: Lahiri, P., editor. Model selection. Beachwood: Institute of Mathematical Statistics; 2001. p. 63-134.

Chipman, HA.; George, IE.; McCulloch, RE. Bayesian treed generalized linear models (with discussion). In: Bernardo, JM.; Bayarri, M.; Berger, JO.; Dawid, AP.; Heckerman, D.; Smith, AFM., editors. Bayesian statistics. Vol. vol 7. Oxford: Oxford University Press; 2003. p. 85-103.

Clyde, M. Bayesian model averaging and model search strategies (with discussion). In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. Bayesian statistics. Vol. vol 6. Oxford: Oxford University Press; 1999. p. 157-185.

Clyde M, George IE. Model uncertainty. Stat Sci 2004;19:81–94.

Cowles MK, Carlin BP. Markov chain Monte Carlo convergence diagnostics: a comparative review. J Am Stat Assoc 1996;91:883–904.

Dellaportas P, Forster JJ. Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. Biometrika 1999;86:615–633.

Dey DK, Chen MH, Chang H. Bayesian approach for the nonlinear random effects models. Biometrics 1997;53:1239–1252.

Geisser S, Eddy W. A predictive approach to model selection. J Am Stat Assoc 1979;74:153–160.

Gelfand AE, Dey DK. Bayesian model choice: asymptotics and exact calculations. J R Stat Soc B 1994;56:501–514.

Gelfand, AE.; Dey, DK.; Chang, H. Model determinating using predictive distributions with implementation via sampling-based methods (with discussion). In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. Bayesian statistics. Vol. vol 4. Oxford: Oxford University Press; 1992. p. 147-167.

Gelfand AE, Ghosh SK. Model choice: a minimum posterior predictive loss approach. Biometrika 1998;85:1–13.

Gelman A, Meng XL, Stern HS. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). Stat Sinica 1996;6:733–807.

George EI. The variable selection problem. J Am Stat Assoc 2000;95:1304–1308.

George EI, Foster DP. Calibration and empirical Bayes variable selection. Biometrika 2000;87:731–747.

George EI, McCulloch RE. Variable selection via Gibbs sampling. J Am Stat Assoc 1993;88:881–889.

George EI, McCulloch RE. Approaches for Bayesian variable selection. Stat Sinica 1997;7:339–374.

George, EI.; McCulloch, RE.; Tsay, R. Two approaches to Bayesian model selection with applications. In: Berry, D.; Chaloner, K.; Geweke, J., editors. Bayesian analysis in statistics and econometrics: essays in honor of Arnold Zellner. New York: Wiley; 1996. p. 339-348.

Gilks WR, Wild P. Adaptive rejection sampling for Gibbs sampling. J R Stat Soc C (Appl Stat) 1992;41:337–348.

Hanson TE. Inference for mixtures of finite polya tree models. J Am Stat Assoc 2006;101:1548–1565.

Huang L, Chen MH, Ibrahim JG. Bayesian analysis for generalized linear models with nonignorably missing covariates. Biometrics 2005;61:767–780. [PubMed: 16135028]

Ibrahim JG, Chen MH. Power prior distributions for regression models. Stat Sci 2000;15:46–60.

Ibrahim JG, Laud PW. A Predictive approach to the analysis of designed experiments. J Am Stat Assoc 1994;89:309–319.

Ibrahim JG, Chen MH, McEachern SN. Bayesian variable selection for proportional hazards models. Can J Stat 1999a;27:701–717.

Ibrahim JG, Lipsitz SR, Chen MH. Missing covariates in generalized linear models when the missing data mechanism is nonignorable. J R Stat Soc B 1999b;61:173–190.

Ibrahim JG, Chen MH, Ryan LM. Bayesian variable selection for time series count data. Stat Sinica 2000;10:971–987.

Ibrahim, JG.; Chen, MH.; Sinha, D. Bayesian survival analysis. New York: Springer-Verlag; 2001a.

Ibrahim JG, Chen MH, Sinha D. Criterion based methods for Bayesian model assessment. Stat Sinica 2001b;11:419–443.

Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing data methods in regression models. J Am Stat Assoc 2005;100:332–346.

Kim S, Chen MH, Dey DK, Gamerman D. Bayesian dynamic models for survival data with a cure fraction. Lifetime Data Anal 2007;13:17–35. [PubMed: 17136621]

Laud PW, Ibrahim JG. Predictive model selection. J R Stat Soc B 1995;57:247–262.

Lipsitz SR, Ibrahim JG. A conditional model for incomplete covariates in parametric regression models. Biometrika 1996;83:916–922.

Little, RJA.; Rubin, DB. Statistical analysis with missing data. 2nd edn. New York: Wiley; 2002.

Ntzoufras I, Dellaportas P, Forster JJ. Bayesian variable and link determination for generalised linear models. J Stat Plan Inference 2003;111:165–180.

Raftery AE. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika 1996;83:251–266.

Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. J Am Stat Assoc 1997;92:179–191.

Rubin DB. Inference and missing data. Biometrika 1976;63:581–592.

Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6:461–464.

Smith M, Kohn R. Nonparametric regression using Bayesian variable selection. J Econom 1996;75:317–343.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). J R Stat Soc B 2002;64:583–639.

Table 1

Values of AIC and BIC under best three AIC or BIC models for the completely observed BMT data

Model	***J* = 10**			***J* = 15**			***J* = 20**		
	AIC$_m$	**BIC$_m$**	p_m	**AIC$_m$**	**BIC$_m$**	p_m	**AIC$_m$**	**BIC$_m$**	p_m
$(x_1, x_2, x_4, \boldsymbol{x}_9)$	3607.09	3694.78	16	3595.19	3710.29	21	3614.49	3756.99	26
$(x_1, x_2, x_4, x_5, \boldsymbol{x}_9)$	3607.09	3700.27	17	3595.46	3716.05	22	3614.63	3762.62	27
$(x_1, x_2, x_3, x_4, x_5, \boldsymbol{x}_9)$	3607.90	3706.56	18	3595.70	3721.76	23	3615.16	3768.63	28
(x_1, x_2, x_9)	3610.60	3692.82	15	3598.82	3708.44	20	3618.05	3755.07	25
$(x_1, x_2, x_5, \boldsymbol{x}_9)$	3610.68	3698.38	16	3599.15	3714.25	21	3618.26	3760.76	26

Table 2

Values of DIC under best three DIC models for the completely observed BMT data

Model	**a_0 = 0.001**					
	***J* = 10**		***J* = 15**		***J* = 20**	
	DIC$_m$	**p_m**	**DIC$_m$**	**p_m**	**DIC$_m$**	**p_m**
$(x_1, x_2, x_4, \mathbf{x}_9)$	3607.18	16.02	3595.26	21.00	3614.65	26.05
$(x_1, x_2, x_4, x_5, \mathbf{x}_9)$	3606.96	16.91	3595.47	21.97	3615.12	26.94
$(x_1, x_2, x_3, x_4, x_5, \mathbf{x}_9)$	3607.76	18.43	3595.68	22.96	3615.31	28.07
	J = 15					
	a_0 = 0.1		a_0 = 0.01		a_0 = 0.0001	
	DIC$_m$	p_m	DIC$_m$	p_m	DIC$_m$	p_m
$(x_1, x_2, x_4, \mathbf{x}_9)$	3769.45	19.20	3599.56	20.73	3595.35	21.07
$(x_1, x_2, x_4, x_5, \mathbf{x}_9)$	3769.71	20.21	3599.76	21.70	3595.62	22.07
$(x_1, x_2, x_3, x_4, x_5, \mathbf{x}_9)$	3769.66	21.18	3599.82	22.64	3595.87	22.89

Table 3

Posterior estimates of **β** under best two DIC models with $J = 15$ and $a_0 = 0.001$ for the completely observed BMT data

Model	Variable	Estimate	SD	95% HPD interval
$(x_1, x_2, x_4, \boldsymbol{x}_9)$	Disease	0.164	0.039	(0.086, 0.239)
	Age	0.269	0.040	(0.193, 0.350)
	Karnofprg	−0.086	0.036	(−0.152, −0.013)
	Cytoabnew1	−0.400	0.113	(−0.621, −0.185)
	Cytoabnew2	−0.586	0.107	(−0.802, −0.382)
	Cytoabnew3	−0.638	0.209	(−1.065, −0.241)
$(x_1, x_2, x_4, x_5, \boldsymbol{x}_9)$	Disease	0.167	0.039	(0.091, 0.246)
	Age	0.250	0.042	(0.167, 0.331)
	Karnofprg	−0.086	0.036	(−0.154, −0.014)
	Gsource	0.054	0.041	(−0.027, 0.133)
	Cytoabnew1	−0.399	0.111	(−0.626, −0.190)
	Cytoabnew2	−0.580	0.108	(−0.786, −0.372)
	Cytoabnew3	−0.621	0.207	(−1.037, −0.217)

Table 4

Values of DIC under best three DIC models for the BMT data with all cases

Model	**a_0 = 0.001**					
	***J* = 10**		***J* = 15**		***J* = 20**	
	DIC$_m$	**p_m**	**DIC$_m$**	**p_m**	**DIC$_m$**	**p_m**
(x_1, x_2, x_4, x_5, $\boldsymbol{x}_8$, $\boldsymbol{x}_9$)	4742.43	25.07	4729.11	30.09	4748.67	35.09
(x_1, x_2, x_4, x_5, $\boldsymbol{x}_9$)	4742.52	21.39	4729.30	26.48	4748.95	31.53
(x_1, x_2, x_3, x_4, x_5, $\boldsymbol{x}_9$)	4743.19	22.40	4729.73	27.44	4749.27	32.43
	J = 15					
	a_0 = 0.1		a_0 = 0.01		a_0 = 0.0001	
	DIC$_m$	p_m	DIC$_m$	p_m	DIC$_m$	p_m
(x_1, x_2, x_4, x_5, $\boldsymbol{x}_8$, $\boldsymbol{x}_9$)	4973.47	27.01	4735.55	29.61	4729.31	30.17
(x_1, x_2, x_4, x_5, $\boldsymbol{x}_9$)	4973.85	23.28	4735.63	25.76	4729.33	26.49
(x_1, x_2, x_3, x_4, x_5, $\boldsymbol{x}_9$)	4974.25	24.29	4735.87	26.97	4729.84	27.57

Table 5

Posterior estimates of **β** and **α** under best DIC model with J = 15 and a_0 = 0.001 for the completely observed BMT data

Model	Parameters	Variable	Estimate	SD	95% HPD interval
$(x_1, x_2, x_4, x_5, x_8, x_9)$	**β**	Disease	0.130	0.034	(0.062, 0.196)
		Age	0.233	0.037	(0.164, 0.308)
		Karnofprg	−0.090	0.032	(−0.152, −0.028)
		Gsource	0.062	0.037	(−0.009, 0.136)
		Prevgvh11	0.127	0.063	(0.008, 0.253)
		Prevgvh12	0.062	0.071	(−0.071, 0.207)
		Prevgvh13	0.014	0.041	(−0.067, 0.093)
		Prevgvh14	0.010	0.044	(−0.076, 0.096)
		Cytoabnew1	−0.424	0.109	(−0.633, −0.206)
		Cytoabnew2	−0.614	0.103	(−0.812, −0.405)
		Cytoabnew3	−0.726	0.205	(−1.150, −0.342)
	α	Intercept 1	−2.158	0.071	(−2.302, −2.024)
		Intercept 2	−0.428	0.046	(−0.518, −0.338)
		Intercept 3	2.950	0.097	(2.754, 3.135)
		Disease	0.363	0.043	(0.275, 0.444)
		Age	0.105	0.047	(0.011, 0.194)
		Karnofprg	−0.073	0.043	(−0.155, 0.011)
		Gsource	0.144	0.046	(0.056, 0.238)
		Prevgvh11	−0.021	0.073	(−0.159, 0.127)
		Prevgvh12	−0.213	0.080	(−0.367, −0.056)
		Prevgvh13	0.036	0.049	(−0.058, 0.134)
		Prevgvh14	−0.136	0.054	(−0.245, −0.032)
$(x_1, x_2, x_4, x_5, x_9)$	**β**	Disease	0.127	0.034	(0.060, 0.195)
		Age	0.235	0.037	(0.163, 0.308)
		Karnofprg	−0.088	0.031	(−0.151, −0.027)
		Gsource	0.068	0.036	(−0.004, 0.135)
		Cytoabnew1	−0.423	0.108	(−0.623, −0.200)
		Cytoabnew2	−0.618	0.102	(−0.809, −0.409)

Model	Parameters	Variable	Estimate	SD	95% HPD interval
		Cytoabnew3	−0.743	0.204	(−1.140, −0.351)
	$\boldsymbol{\alpha}$	Intercept 1	−2.136	0.070	(−2.277, −2.003)
		Intercept 2	−0.420	0.046	(−0.508, −0.329)
		Intercept 3	2.928	0.096	(2.743, 3.116)
		Disease	0.353	0.043	(0.269, 0.437)
		Age	0.113	0.046	(0.018, 0.198)
		Karnofprg	−0.069	0.042	(−0.150, 0.014)
		Gsource	0.149	0.045	(0.063, 0.240)