The International Journal of Biostatistics

Volume 7, Issue 1

2011

Article 21

Bayesian Variable Selection in Semiparametric Proportional Hazards Model for High Dimensional Survival Data

Kyu Ha Lee, *University of Missouri - Columbia*

Sounak Chakraborty, *University of Missouri - Columbia*

Jianguo Sun, *University of Missouri - Columbia*

Recommended Citation:

Lee, Kyu Ha; Chakraborty, Sounak; and Sun, Jianguo (2011) "Bayesian Variable Selection in Semiparametric Proportional Hazards Model for High Dimensional Survival Data," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 21.

DOI: 10.2202/1557-4679.1301

Bayesian Variable Selection in Semiparametric Proportional Hazards Model for High Dimensional Survival Data

Kyu Ha Lee, Sounak Chakraborty, and Jianguo Sun

Abstract

Variable selection for high dimensional data has recently received a great deal of attention. However, due to the complex structure of the likelihood, only limited developments have been made for time-to-event data where censoring is present. In this paper, we propose a Bayesian variable selection scheme for a Bayesian semiparametric survival model for right censored survival data sets. A special shrinkage prior on the coefficients corresponding to the predictor variables is used to handle cases when the explanatory variables are of very high-dimension. The shrinkage prior is obtained through a scale mixture representation of Normal and Gamma distributions. Our proposed variable selection prior corresponds to the well known lasso penalty. The likelihood function is based on the Cox proportional hazards model framework, where the cumulative baseline hazard function is modeled a priori by a gamma process. We assign a prior on the tuning parameter of the shrinkage prior and adaptively control the sparsity of our model. The primary use of the proposed model is to identify the important covariates relating to the survival curves. To implement our methodology, we have developed a fast Markov chain Monte Carlo algorithm with an adaptive jumping rule. We have successfully applied our method on simulated data sets under two different settings and real microarray data sets which contain right censored survival time. The performance of our Bayesian variable selection model compared with other competing methods is also provided to demonstrate the superiority of our method. A short description of the biological relevance of the selected genes in the real data sets is provided, further strengthening our claims.

KEYWORDS: Bayesian variable selection, Cox proportional hazards model, gamma process, Gibbs sampler, lasso, microarray

Author Notes: We would like to thank the University of Missouri Bioinformatics Consortium for letting us use their Dell EM64T cluster system (Lewis). We would also like to thank Mr. Adam Lane, graduate student, Department of Statistics, University of Missouri - Columbia for grammatical reviews and language editing.

1 Introduction

In many microarray experiments, a major objective is to find a subset of genes, which are strongly correlated with the survival time (or the time to events of interest). The identified genes in the subset are then used to build a predictive model, which allows us to predict the survival times for new patients. The survival prediction based on discovery of the relationship between the survival time and gene expression profile could help medical doctors make good prognoses and accurate diagnoses, and determine appropriate treatment procedures.

There are several aspects of the microarray data analysis with survival time which require special attention. First, survival times are typically incomplete or censored. This is commonly due to the time limit or cost considerations which may require the investigator to terminate the study before all patients realize the event of interest. Second, owing to the nature of microarray gene expression data such as the diffuse large B-cell lymphoma (DLBCL) of Rosenwald et al.(2002), it is naturally presumed that the number of genomic covariates (p) exceeds the number of subjects (n). In recent studies of variable selection given these conditions most of the proposed methods used Cox's proportional hazards model with a prior dimension reduction (Bair et al., 2006) or with a penalty function (Tibshirani, 1996).

In the proportional hazards model (Cox, 1972), the hazard rate for the survival time T of a subject with covariate vector $\mathbf{x} = (x_1, \dots, x_p)'$ can be expressed as

$$h(t|\mathbf{x}) = h_0(t) \exp(\mathbf{x}'\boldsymbol{\beta}), \quad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is a vector of p regression parameters, and $h_0(t)$ is unspecified arbitrary baseline hazard function. When $n > p$, estimates of $\boldsymbol{\beta}$ can be obtained by maximizing the partial likelihood (Cox, 1975),

$$L(\boldsymbol{\beta}) = \prod_{k \in \mathcal{D}} \frac{\exp(\mathbf{x}'_k \boldsymbol{\beta})}{\sum_{l \in \mathcal{R}_k} \exp(\mathbf{x}'_l \boldsymbol{\beta})}, \quad (2)$$

where $\mathcal{R}_k$ is the set of indices of subjects at risk at time right before t_k , and $\mathcal{D}$ is the set of indices of the failure times. However, the solution that maximizes (2) is not unique for the cases when $p \gg n$. One possible way of solving this problem is by including a penalty or a regularization term into the model. Verweij and van Houwelingen (1994) proposed a L_2 penalized Cox regression model, where estimates of the regression parameters are obtained by maximizing the penalized log partial likelihood $l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^p \beta_j^2$. (where $l(\boldsymbol{\beta})$ is the log partial likelihood from (2)). The $\sum_{j=1}^p \beta_j^2$ is the L_2 -penalty function and λ is the tuning parameter, which determines the amount of shrinkage. In a comparative study by Bøvelstad et al.

(2007), the L_2 -penalty (ridge regression) method was shown to have the best overall prediction performance among methods they considered for the Cox regression model. However, the L_2 -penalty or ridge regression do not give a sparse solution.

To achieve sparsity, Tibshirani (1996) introduced the lasso penalty based on the L_1 -norm. The lasso shrinks the regression coefficients toward zero by imposing a penalty on their absolute values and hence achieves sparsity. Based on this, Tibshirani (1997) suggested a penalized log partial likelihood $l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^p |\beta_j|$. (Notations are the same as before). However, the lasso can select at most n variables. For gene expression microarray data sets, where the number of covariates (p) is very large and the number of subjects (n) is limited this may be a problem. Realistically, if the number of truly relevant genes is more than n the lasso will fail to identify them. In similar studies Gui and Li (2005) proposed to use the L_1 penalized estimation for the Cox model to select genomic covariates that are relevant to survival time of patients. They developed LARS-Cox procedure to solve the computational difficulty associated with the estimation in the case where $n \ll p$ but, like the lasso, it can select at most n variables. Adapting ideas of dimension reduction and penalized regression, Bair et al. (2006) proposed the supervised principal components regression (SPC). More recently Tibshirani (2009) proposed the *Cox univariate shrinkage* (CUS) estimator, which under the assumption of feature independence, obtains a simple estimate through a set of one-dimensional maximizations with the lasso penalty. The CUS estimator is the maximizer of the penalized partial log-likelihood,

$$J(\boldsymbol{\beta}) = \sum_{j=1}^p g_j(\beta_j) - \lambda \sum |\beta_j|, \quad (3)$$

where $g_j(\boldsymbol{\beta}) \equiv \sum_{k=1}^K (x_{kj}\beta_j - \log \sum_{m \in \mathcal{R}_k} \exp(x_{mj}\beta_j))$, K is the number of distinct failure times, and $\lambda \geq 0$ is the tuning parameter. A variable selection scheme for Cox models with semiparametric relative risk is also introduced by Du et al. (2010).

Variable selection under the Bayesian framework traditionally has been done using the SSVS (Stochastic Search Variable Selection) procedure proposed by George and McCulloch (1993). In this procedure, after identifying the promising subsets of variables based on their posterior probability, Gibbs sampling is used for further identification. The SSVS type of techniques are also adopted in discrete response models (Lee et al. 2003; Tadesse et al., 2005). In recent years Chakraborty et al. (2007) and Chakraborty (2009) extended Bayesian gene selection for non-linear binary and multiclass problems. Annett et al. (2009) developed the iterative Bayesian Model Average (BMA) algorithm for the application of survival analysis to microarray data. However, iterative BMA cannot select more than the size of the BMA window (maximum 30). This could be problematic if there exist 31 or more relevant variables.

Under the Bayesian framework the lasso estimate can be viewed as a Bayesian posterior mode estimate under independent Laplace priors for the regression parameters. The Laplace distribution can be expressed as a scale-mixture of normals with a gamma mixing density (Andrews and Mallows, 1974),

$$\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}}e^{-\frac{z^2}{2s}} \cdot \frac{a^2}{2}e^{-\frac{a^2 s}{2}} ds, \quad (4)$$

where $a > 0$. Based on this scale mixture representation (4) Park and Casella (2008) proposed a full hierarchical Bayes method for variable selection and estimation of a linear regression model.

In this paper, we extend the idea of the Bayesian lasso to the Cox proportional hazards model for survival data in the situation where there are many more genomic variables than subjects. In our proposed model, we adopt the grouped data likelihood (Burridge, 1981), where the cumulative baseline hazard function is modeled nonparametrically by a discrete gamma process (Kalbfleish, 1978). A big merit of our approach is that the regression coefficients and the baseline hazard can be jointly modeled and estimated via MCMC methods. Furthermore, grouped data likelihood will help us to suitably handle grouped survival data, or survival data with many failure time ties. We construct the Bayesian hierarchical model with a lasso-type shrinkage prior based on the scale mixture representation of the Laplace distribution. In our approach the tuning parameter is controlled adaptively by assigning a prior on it. The proposed model is shown to be better than the existing methods in terms of variable selection and prediction accuracy.

The rest of the paper is organized as follows. In Section 2, we introduce and describe our *Penalized semiparametric Bayesian Cox model* (PSBC). In Section 3 we give a step by step MCMC algorithm to fit our model. In Section 4, we present simulation studies and compare the performance of our proposed model against several competing methods, like CUS, SPC, and iterative BMA. In Section 5, we illustrate the performance of our method on real-life data sets. In Section 6, we provide some discussion and the possible future direction of our research. The proof regarding the unimodality of our posterior is given in the Appendix.

2 Penalized Semiparametric Bayesian Cox Model (PSBC)

Suppose a data set consists of n subjects. For the i^{th} subject we record the survival time t_i , covariates $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$, and the right censoring indicators c_i . We denote the survival times from all n subjects by $\mathbf{T} = (t_1, \dots, t_n)'$ and the $n \times p$ covariate

matrix by X . Under the Cox model (1), the joint survival probability of n subjects given the matrix of covariates X is given by

$$P(\mathbf{T} > \mathbf{t} | \boldsymbol{\beta}, X, H_0) = \exp\left\{-\sum_{i=1}^n \exp(\mathbf{x}_i' \boldsymbol{\beta}) H_0(t_i)\right\}, \quad (5)$$

where $H_0(t)$ is the cumulative baseline hazard function. There are several types of nonparametric prior processes we can consider for the baseline hazard or cumulative hazard function in the Cox model. One of the most popular choices in semi-parametric Bayesian survival analysis is a gamma process. In model (5) we can assign the gamma process prior to the cumulative baseline hazard function $H_0(t)$ (Kalbfleish, 1978) as follows.

$$H_0 \sim \mathcal{GP}(c_0 H^*, c_0), \quad (6)$$

where $H^*(t)$ is an increasing function with $H^*(0) = 0$, and c_0 is a positive constant. The specified H^* can be viewed as an initial guess of H_0 and c_0 is a specification of weight or confidence attached to that guess. In general, H^* is taken to be a known parametric function, such as exponential or Weibull distribution. For example, if H^* corresponds to the Weibull distribution, then $H^*(t) = \eta_0 t^{\kappa_0}$, where (η_0, κ_0) is a specified vector of hyperparameters. The values of the hyperparameters should be carefully selected to avoid a convergence problem or floating problem in MCMC sampling.

In practical situations, the occurrence of grouped data is quite common. It has been shown that using grouped data leads to more satisfactory results than by using continuous data by Burrige (1981) under the empirical Bayes setting. Therefore, we will construct our model based on the group data likelihood (Ibrahim et al., 2005, chapter 3.2.2).

In order to construct the grouped data likelihood, let us consider a finite partition of the time axis, $0 = s_0 < s_1 < s_2 < \dots < s_J$, with $s_J > t_i$ for all $i = 1, \dots, n$. Then, we have J disjoint intervals, $I_j = (s_{j-1}, s_j]$ for $j = 1, 2, \dots, J$. The survival time t_i of the i^{th} subject falls in one of those J disjoint intervals. We assume that the observed data $D = (X, \mathcal{R}_j, \mathcal{D}_j : j = 1, 2, \dots, J)$ is available as grouped within I_j 's. The $\mathcal{R}_j$ and $\mathcal{D}_j$ are the risk sets and the failure sets corresponding to the j^{th} interval I_j . Here, X is the $n \times p$ matrix of covariates with i^{th} row $\mathbf{x}_i'$ denoting the usual vector of predictors for the i^{th} subject. We also assume that the x_{ij} s are standardized so that $\sum_i x_{ij}/n = 0$ and $\sum_i x_{ij}^2/n = 1$. Let h_j denote the increment in the cumulative baseline hazard in the interval I_j , as follows

$$h_j = H_0(s_j) - H_0(s_{j-1}), \quad j = 1, 2, \dots, J. \quad (7)$$

The gamma process prior in (6) implies that the h_j 's follow independent gamma distributions, that is

$$h_j \sim \mathcal{G}(\alpha_{0j} - \alpha_{0,j-1}, c_0), \quad (8)$$

where $\alpha_{0j} = c_0 H^*(s_j)$. Therefore, the conditional probability of the i^{th} subject failing in the interval I_j is given by

$$\begin{aligned} P(T_i \in I_j | \mathbf{h}) &= \exp\{-\exp(\mathbf{x}'_i \boldsymbol{\beta}) H(t_{j-1})\} - \exp\{\exp(\mathbf{x}'_i \boldsymbol{\beta}) H(t_i)\} \\ &= \exp\{-\exp(\mathbf{x}'_i \boldsymbol{\beta}) \sum_{m=1}^{j-1} h_m\} [1 - \exp\{-h_j \exp(\mathbf{x}'_i \boldsymbol{\beta})\}], \end{aligned} \quad (9)$$

where $\mathbf{h} = (h_1, h_2, \dots, h_J)'$. This leads to our grouped data likelihood function (Ibrahim et al., 2005, chapter 3.2.2)

$$L(\boldsymbol{\beta}, \mathbf{h} | D) \propto \prod_{j=1}^J G_j, \quad (10)$$

where $G_j = \exp\{-h_j \sum_{k \in R_j - D_j} \exp(\mathbf{x}'_k \boldsymbol{\beta})\} \prod_{l \in D_j} [1 - \exp\{-h_j \exp(\mathbf{x}'_l \boldsymbol{\beta})\}]$.

The Cox's partial likelihood (2) can be viewed as a limiting case of the marginal posterior of $\boldsymbol{\beta}$ in the Cox model with a gamma process prior on the cumulative baseline hazard (See Chapter 3.2.3, Ibrahim et al., 2005). It can be shown, if c_0 is near 0 (meaning placing very little faith in the prior estimate H^* of $H(t)$), the marginal likelihood obtained from (10) is precisely the Cox's partial likelihood. On the other hand, if we let $c_0 \rightarrow \infty$, our likelihood (10) takes the form of the likelihood function based on $(\boldsymbol{\beta}, h_0)$ (Ibrahim et al., 2005, eq. 3.2.12), where $h_0(t) = \frac{d}{dt} H^*(t)$.

The regression coefficient vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$, plays a major role in selecting the covariates in the model (10). So we introduce a prior that will result in shrinkage in our model. The variable selection will be done through BIC thresholding. More details about BIC thresholding will be given in Section 3 and Section 4. Note that, the $|\beta_j|$ in the lasso penalty is proportional to the (minus) log-density of Laplace distribution. Laplace prior (11) for the regression coefficients is given by,

$$\pi(\boldsymbol{\beta}) = \prod_{j=1}^p \frac{\lambda}{2} \exp(-\lambda |\beta_j|). \quad (11)$$

The Laplace distribution puts more probability mass near zero and in its tails than standard normal distribution. Therefore, we can expect that the prior (11) will bring about the posterior estimates (mode), which are either 0 or large as the lasso. This

ensures us a sparse solution. To guarantee unimodality (Appendix A), instead of (11) we use the conditional Laplace prior of the form (Park and Casella, 2008),

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^p \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left(-\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}}\right). \quad (12)$$

A noninformative marginal prior $\pi(\sigma^2) = 1/\sigma^2$ is assigned on σ^2 . The hierarchical representation of the Bayesian lasso with prior (12) can thus be obtained by utilizing the representation of the Laplace distribution as a scale mixture of normals with an exponential mixing density as described in (4). Combining (4), (8), (10), and (12), the hierarchical representation of our full model can be described as,

$$\begin{aligned} L(D|\boldsymbol{\beta}, \mathbf{h}) &\propto \prod_{j=1}^J \exp\{-h_j \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\mathbf{x}'_k \boldsymbol{\beta})\} \prod_{l \in \mathcal{D}_j} [1 - \exp\{-h_j \exp(\mathbf{x}'_l \boldsymbol{\beta})\}], \\ h_j &\sim \mathcal{G}(\alpha_{0j} - \alpha_{0,j-1}, c_0), \\ \boldsymbol{\beta}|\sigma^2, \tau_1^2, \dots, \tau_p^2 &\sim \mathbf{N}_p(\mathbf{0}_p, \sigma^2 \mathbf{D}_{\boldsymbol{\tau}}), \quad \mathbf{D}_{\boldsymbol{\tau}} = \text{diag}(\tau_1^2, \dots, \tau_p^2), \\ \tau_1^2, \dots, \tau_p^2 | \lambda^2 &\sim \prod_{j=1}^p \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2 / 2} d\tau_j^2, \\ \pi(\sigma^2) &= \frac{1}{\sigma^2} d\sigma^2, \quad \sigma^2, \tau_1^2, \dots, \tau_p^2 > 0. \end{aligned} \quad (13)$$

The λ is a tuning parameter, which controls the sparsity of our model. Large value of λ will induce higher sparsity. In general, the tuning parameter of the ordinary lasso is chosen by cross-validation or generalized cross-validation (Tibshirani, 1996). Likewise, the tuning parameter for the Bayesian lasso can be selected using empirical Bayes approach which involves maximizing the marginal maximum likelihood estimates of the hyperparameters (Casella, 2001; Park and Casella, 2008). However, it requires an EM-algorithm type step within each MCMC iteration and hence it is computationally slow for large p cases. We use here the hierarchical Bayes method to choose λ by using a diffuse hyperprior on it (Park and Casella, 2008). From (13), we can see that the tuning parameter λ enters in our model as λ^2 . Therefore, we assign gamma prior directly on λ^2 of the form

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} e^{-\delta \lambda^2}. \quad (14)$$

The r and δ are the shape and the rate parameters of the gamma distribution respectively. To avoid mixing problems, r and δ should be determined such that the prior density (14) moves toward 0 relatively fast as $\lambda^2 \rightarrow \infty$. At the same time, the prior

(14) should be able to put a large enough probability mass near the maximum likelihood estimate and should be comparatively flat. One big advantage of using the full Bayes approach for λ is that it can be easily incorporated into our hierarchical model (13) structure and it does not affect the full conditional distribution of any other parameters. Moreover, it is computationally less expensive than the empirical Bayes approach. This full Bayesian approach results in adaptive selection of the tuning parameter λ . Therefore, the final joint posterior distribution about PSBC model can be written as,

$$\begin{aligned}
 \pi(\boldsymbol{\beta}, \mathbf{h}, \sigma^2, \boldsymbol{\tau}, \lambda^2 | D) &\propto L(D | \boldsymbol{\beta}, \mathbf{h}) \times \pi(\mathbf{h}) \times \pi(\boldsymbol{\beta} | \sigma^2, \boldsymbol{\tau}) \times \pi(\boldsymbol{\tau} | \lambda^2) \times \pi(\sigma^2) \times \pi(\lambda^2) \\
 &\propto \prod_{j=1}^J \exp\{-h_j \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\mathbf{x}'_k \boldsymbol{\beta})\} \prod_{l \in \mathcal{D}_j} [1 - \exp\{-h_j \exp(\mathbf{x}'_l \boldsymbol{\beta})\}] \\
 &\quad \times \prod_{j=1}^J \left[h_j^{(\alpha_{0j} - \alpha_{0,j-1}) - 1} \exp(-c_0 h_j) \right] \exp \left\{ -\frac{1}{2\sigma^2} \boldsymbol{\beta}' \mathbf{D}_{\boldsymbol{\tau}}^{-1} \boldsymbol{\beta} \right\} \\
 &\quad \times \frac{1}{\sigma^2} (\lambda^2)^{r-1} \exp(-\delta \lambda^2) \prod_{j=1}^p \lambda^2 \exp(-\frac{1}{2} \lambda^2 \tau_j^2), \tag{15}
 \end{aligned}$$

where $\mathbf{h} = (h_1, \dots, h_J)'$ and $\boldsymbol{\tau} = (\tau_1^2, \dots, \tau_p^2)'$.

3 Computational Scheme

In this section, we describe an MCMC algorithm to fit our PSBC model described in the previous section. The final output from the MCMC simulation is used for variable selection, parameter estimation, and prediction in our PSBC model. From the joint posterior distribution in (15) we derived full conditional posterior distributions for the model parameters $(\boldsymbol{\beta}, \mathbf{h}, \sigma^2, \boldsymbol{\tau}, \lambda^2)$. For the parameter $\boldsymbol{\beta}$, the conditional posterior distribution does not have a standard form, so we use Metropolis-Hastings algorithm with an adaptive jumping rule. The conditional posterior distributions of other parameters have standard forms or can be well approximated by certain distribution, so Gibbs sampler can be used to get posterior samples. Our detailed step by step MCMC algorithm, the conditional distributions, and justifications are given as follows.

- **Step 1.** Start with initial values, $\boldsymbol{\beta}^{(0)}, \mathbf{h}^{(0)}, \boldsymbol{\tau}^{(0)}, \sigma^{2(0)}$, and $\lambda^{2(0)}$. The index g denote the g^{th} iteration, and set $g = 1$.

- **Step 2.** Update $\boldsymbol{\beta}$ by using Metropolis-Hastings algorithm with an adaptive jumping rule. The full conditional posterior distribution for β_j , $j = 1, \dots, p$, is

$$\begin{aligned} & \pi(\beta_j | \boldsymbol{\beta}^{(-j)}, \mathbf{h}, \sigma^2, \boldsymbol{\tau}, \lambda^2, D) \\ & \propto \prod_{j=1}^J \exp\{-h_j \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\mathbf{x}'_k \boldsymbol{\beta})\} \prod_{l \in \mathcal{D}_j} [1 - \exp\{-h_j \exp(\mathbf{x}'_l \boldsymbol{\beta})\}] \\ & \quad \times \exp\left\{-\frac{1}{2\sigma^2} \boldsymbol{\beta}' \mathbf{D}_{\boldsymbol{\tau}}^{-1} \boldsymbol{\beta}\right\}, \end{aligned} \quad (16)$$

where $\boldsymbol{\beta}^{(-j)}$ denotes the $\boldsymbol{\beta}$ vector without the j^{th} element.

The conditional posterior $\pi(\boldsymbol{\beta} | \mathbf{h}, \sigma^2, \boldsymbol{\tau}, \lambda^2, D)$ is log-concave in $\boldsymbol{\beta}$, so the derivative-free adaptive rejection sampling (ARS) (Gilks, 1992) could be an efficient way to sample from (16). We also tested the Adaptive Rejection Metropolis Sampling (ARMS) (Gilks et al., 1995), available through R package HI. When we update one component of $\boldsymbol{\beta}$ at a time the ARMS reduces to ARS for log-concave densities. However, since available R code for ARMS enables multivariate updating for $\boldsymbol{\beta}$, it is much faster than single-parameter updating. Three sampling methods, ARS, ARMS, and the usual Metropolis-Hastings algorithm are implemented for performance comparison. In our simulation study, we learned that if p (number of covariates) is larger than 200 ARS does no better than the Metropolis-Hastings algorithm. It also turns out that, ARMS does not work properly for multivariate updating even with a moderate number of parameters (Gilks et al., 1995). This is primarily because, constructing efficient proposal densities for a high dimensional parameter vector is not an easy problem in ARMS. To overcome this, we designed a special random walk Metropolis-Hastings method with adaptive jumping rules to update the $\boldsymbol{\beta}$ parameter.

The most important factor in our MH-algorithm is to find an efficient Metropolis jumping rule. We propose a simple adaptive jumping rule to force the pace of the algorithm. Let μ_{β_j} denote the mean of the normal jumping kernel density. The μ_{β_j} is adjusted only after the proposal values are not accepted during many consecutive iterations. That means, μ_{β_j} is tuned only when the kernel density with μ_{β_j} proposes the optimal sample, β_j^* , which in turn results in almost the highest posterior probability in (16) among all the proposals from it. It is fairly intuitive that, β_j^* is closer to the posterior mode of (16) than μ_{β_j} because (16) is a log-concave function. Then we take a normal jumping kernel with mean equal to β_j^* for the next iteration. This simple idea of jumping rules is derivative free and thus speeds up the algorithm for the high dimensional parameter space. A graphical illustration of the jumping rules is shown in Figure 1. The variance of the normal kernel is tuned to obtain reasonable acceptance rates.

Figure 1: Jumping rules for μ_{β_j} . (a) After β^* being sampled, jumping densities with mean $\mu_{\beta_j}^{(g-21)} = \mu_{\beta_j}^{(g-20)} = \dots = \mu_{\beta_j}^{(g-1)}$ have not produced accepted proposals during 20 consecutive iterations. (b) Then we take β^* as the new mean for jumping density for next iteration g .

Our Metropolis-Hastings algorithm for updating β is as follows,

- (i) Set $j = 1$.
- (ii) Sample a proposal $\beta_j^{(prop)}$ from a proposal distribution, $J_g(\beta_j^{(prop)} | \beta_j^{(g-1)}) = N(\beta_j^{(prop)} | \mu_{\beta_j}^{(g-1)}, \mathbf{v}_{\beta_j}^{(g-1)})$.
- (iii) Calculate the ratio of ratios:

$$r_j = \frac{\pi(\beta_j^{(prop)} | \beta^{(-j)(g-1)}, \mathbf{h}^{(g-1)}, \sigma^{2(g-1)}, \boldsymbol{\tau}^{(g-1)}, \lambda^{2(g-1)}, D) / J_g(\beta_j^{(prop)} | \beta_j^{(g-1)})}{\pi(\beta_j^{(g-1)} | \beta^{(-j)(g-1)}, \mathbf{h}^{(g-1)}, \sigma^{2(g-1)}, \boldsymbol{\tau}^{(g-1)}, \lambda^{2(g-1)}, D) / J_g(\beta_j^{(g-1)} | \beta_j^{(prop)})} \quad (17)$$

- (iv) Accept the proposal $\beta_j^{(prop)}$ with acceptance probability $\min(r_j, 1)$.
- (v) Tune the mean and variance of the proposal distribution in step (ii) with the above aforementioned adaptive jumping rules and obtain $\mu_{\beta_j}^{(g)}$ and $\mathbf{v}_{\beta_j}^{(g)}$.
- (vi) Stop if $j = p$. Otherwise set $j = j + 1$, and go to step (ii).

In Appendix A we provide a brief mathematical proof of the unimodality of the posterior distribution of β .

- **Step 3.** The original full conditional distribution for h_j is,

$$\begin{aligned} & \pi(h_j | \mathbf{h}^{(-j)}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}, \lambda^2, D) \\ & \propto h_j^{(\alpha_{0j} - \alpha_{0,j-1}) - 1} \prod_{l \in \mathcal{D}_j} \left\{ 1 - \exp(-h_j e^{\mathbf{x}'_l \boldsymbol{\beta}}) \right\} \exp \left\{ -h_j \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} e^{\mathbf{x}'_k \boldsymbol{\beta}} + c_0 \right\}. \end{aligned} \quad (18)$$

However, (18) can be well approximated by the gamma distribution in (19) (See Chapter 3.2.2, Ibrahim et al., 2005). So instead of (18) we sample from

$$h_j | \mathbf{h}^{(-j)}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}, \lambda^2, D \sim \mathcal{G} \left(\alpha_{0j} - \alpha_{0,j-1} + d_j, c_0 + \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\mathbf{x}'_k \boldsymbol{\beta}) \right), \quad (19)$$

for $j = 1, \dots, J$, where d_j is the number of subjects in the set $\mathcal{D}_j$. This direct sampling from (19) provides a more efficient Gibbs sampling scheme in our algorithm.

- **Step 4.** The full conditional of $1/\tau_j^2$ is inverse-Gaussian, so we sample from

$$\frac{1}{\tau_j^2} | \boldsymbol{\tau}^{(j-1)}, \boldsymbol{\beta}, \mathbf{h}, \sigma^2, \lambda^2, D \sim \text{Inv-Gaussian} \left(\frac{\lambda \sigma}{|\boldsymbol{\beta}_j|}, \lambda^2 \right), \quad (20)$$

for $j = 1, \dots, p$.

- **Step 5.** Since the full conditional distribution of σ^2 is inverse-gamma, we draw a sample from

$$\sigma^2 | \boldsymbol{\beta}, \mathbf{h}, \boldsymbol{\tau}, \lambda^2, D \sim \mathcal{IG} \left(\frac{p}{2}, \sum_{j=1}^p \frac{\beta_j^2}{2\tau_j^2} \right). \quad (21)$$

- **Step 6.** The full conditional for λ^2 is gamma so we sample from

$$\lambda^2 | \boldsymbol{\beta}, \mathbf{h}, \sigma^2, \boldsymbol{\tau}, D \sim \mathcal{G} \left(p + r, \frac{\sum_{j=1}^p \tau_j^2}{2} + \delta \right). \quad (22)$$

- **Step 7.** Let M denote the number of posterior samples desired. If $g < M$, then $g = g + 1$, and go to **Step 2**. Otherwise, go to the next step.
- **Step 8.** Calculate the posterior mean and median of $\boldsymbol{\beta}$ to obtain estimators.

It is worth noting that the posterior estimates $\hat{\boldsymbol{\beta}}$ in our PSBC will be biased toward zero when the posterior estimate of the tuning parameter λ is large.

Therefore, if the estimate of β is of primary interest, it is recommended to obtain the scaled estimate, given by $\hat{\gamma}\hat{\beta}$. This is done by fitting the single predictor $\gamma x_i \beta$ in a Cox model on the data (see Tibshirani, 2009). However, Bayesian lasso formulation (with posterior mean and median) cannot shrink the β estimates to exact zeros. Therefore, for variable selection some thresholding is necessary. We arrange the absolute posterior estimates of β_j in decreasing order, and compute the Bayesian Information Criterion (BIC) (23) of the sequentially entering each β_j . From the plotted BIC curve we can easily find the best set of covariates that should be included in the model. More clear illustration is given in the simulation studies (Section 4) and analysis of real data sets (Section 5). Using the posterior samples, one can also calculate the credible intervals for each β_j and select the covariates that have significantly larger (or smaller) value than zero. To be noted here, if we use the posterior mode the solution exactly matches with the lasso and shrinks toward zero. However, due to computation complexity we used the posterior mean and median.

4 Simulation Study

We performed simulation studies to compare the performance of our PSBC with CUS, SPC, and iterative BMA in terms of prediction accuracy and variable selection. R packages, `uniCox`, `superpc`, and `iterativeBMA` are used to implement CUS, SPC, and iterative BMA respectively. For accurate estimation with CUS, we set the `n.lam` (the number of λ values to consider) option in `uniCox` equal to p (the number of variables in the data set). The `n.threshold` (the number of thresholds to consider) option for SPC in `superpc` is also set to p . For iterative BMA, we set `maxNvar` equal to p for $p \leq 25$ and 25 otherwise. If `maxNvar` is larger than 25, iterative BMA gives an error due to the limitation mentioned previously. Other user-specified input parameters, `nbest` and `#top` ranked variables to consider, in iterative BMA are set to 50 and p respectively.

Selection of the prior parameters $(r, \delta, \kappa_0, \eta_0, c_0)$, plays a crucial role in our PSBC model. For the prior on λ^2 in (14), we set r and δ such that the prior has the mean equal to 10 times the $\hat{\lambda}_{CUS}^2$ (the cross validation estimate of λ from CUS). This formulation results in reasonable initial guess on λ^2 and gives a relatively flat noninformative prior. We set η_0 and κ_0 so that the increasing function $H^*(t)$ looks similar to the Nelson-Aalen estimator of $H_0(t)$. This has to do with the fact that $H^*(t)$ is an initial guess at $H_0(t)$. For simulated data sets, clear separation of significant variables from insignificant ones is observed as the confidence parameter c_0 gets smaller. However, the convergence speed is quite slow if c_0 is kept too small. After examining the trace plots from all simulations with several choices of c_0 , we

decided to use $c_0 = 2$ to get reasonably fast convergence and good performance of isolating the true non-zero predictors. Additionally, the intervals I_j for the grouped data likelihood we considered should be carefully selected. We fixed $J - 1$ to be the number of uniquely ordered failure times from the simulated data set so that one subject failure falls in each interval. A different choice of intervals may be necessary, especially when there are many failure observations in the data set.

To avoid any problem with multi-modality the posterior estimates of the parameters are calculated using two independent MCMC chains. Each chain ran for 10,000 iterations with the first quarter as burn-in. We monitor the convergence of our MCMC chain through trace plots of the posterior samples and also by the Gelman and Rubin's diagnostic (calculated using R package coda).

Since we have assigned continuous priors for β our posterior mean estimate does not give exact zeros. In this paper, we use the BIC thresholding techniques for our PSBC model. The absolute posterior estimates of β_j are arranged in descending order, and then we compute the BIC values in a stepwise manner by sequentially adding important covariates in (23). The formula for BIC with j largest β s is written as

$$BIC_j = -2\{l_j(\hat{\beta}_{(1:j)}) - l_0(\mathbf{0})\} + j\log(n), \quad j = 1, \dots, p, \quad (23)$$

where $l_j(\hat{\beta}_{(1:j)})$ is the maximized log likelihood under a model $\mathcal{M}_j$ that includes covariates corresponding to largest j $|\beta_j|$ -s given by $(\hat{\beta}_{(1:j)})$, and $l_0(\mathbf{0})$ is the log likelihood under the null model. Volinsky and Raftery (2000) showed that BIC provides a closer approximation to twice the log Bayes factor when the number of events d , is used instead of n in (23). We calculated BIC based on both n and d , but decided to use n in (23) as it gave a little better result in variable selection. However, the difference in the BICs (computed using n and p) in our simulation is not significant.

We design two different simulation scenarios, (i) covariates are independent of each other and (ii) covariates are correlated among each other. The detailed description of our simulation settings is as follows:

- **Simulation 1.** We simulated Gaussian samples of size $n = 100$ with $p = 20, 200, 500$, and 1000 predictors. The pairwise correlation between the predictors x_i and x_j is denoted by $C(x_i, x_j) = \rho^{|i-j|}$. In this setting the correlation $\rho = 0$ is used to generate the data set. The survival time t is generated from an exponential distribution with rate parameter $1/\exp(\sum_{j=1}^p x_j \beta_j)$. The censoring time is simulated such that 40% censoring is expected. We assume that only a randomly chosen 10 out of a total of p variables are correlated with the survival time. The 10 coefficients (β_j) corresponding to those significant x_j -s are set equal to 4. The remaining coefficients (β_j) are set equal

to 0. This simulation setting is quite similar to that of Tibshirani (2009). The underlying true model is very sparse as p gets larger ($p = 20, 200, 500$, and 1000 predictors) and the variable selection becomes increasingly difficult.

- **Simulation 2.** The second scenario is same as the first one except $\rho = 0.5$ is used to generate the data set.

We randomly generate data under each combination of simulation setting and number of covariates 100 times. The variable selection capability of our model is judged by computing four operating characteristics based on the parameter estimates. They are true positive rate (TPR or sensitivity), true negative rate (TNR or specificity), positive predictive value (PPV), and negative predictive value (NPV). The formulas for these operating characteristics are as follows.

$$\begin{aligned} TPR &= \frac{TP}{10} \times 100 (\%), & TNR &= \frac{TN}{p-10} \times 100 (\%), \\ PPV &= \frac{TP}{TP+FP} \times 100 (\%), & NPV &= \frac{TN}{TN+FN} \times 100 (\%). \end{aligned} \quad (24)$$

The TP denotes the number of significant variables correctly entered into a model. The TN denotes the number of insignificant variables correctly excluded from the model. On the other hand, FP and FN represent the number of irrelevant variables mistakenly entered into the model and the number of relevant variables mistakenly excluded from the model respectively. Ideally, TPR, TNR, PPV, and NPV should be 100%, indicating only the truly relevant variables are included in the model.

In Table 1 and Table 2 we report the overall averages and standard deviations of the four measures for PSBC, CUS, SPC, and iterative BMA. We have also included the average number of variables selected by all methods. From Table 1 and 2, we can see that our PSBC model outperforms the competing methods in terms of variable selection accuracy. In almost all the cases our PSBC gives higher TPR, TNR, PPV, and NPV. As in our simulation studies we assume only 10 significant variables out of p , missing only one significant variable results in 10% lower TPR. On the other hand, 10% lower in TNR means including $(p-10) \times 0.1$ many insignificant variables into the model. Therefore, for the better assessment of the overall performance of variable selection, we also look at the PPV, NPV, and the number of selected variables. Our PSBC model consistently gave a higher PPV and NPV (Table 1 and Table 2) than all other competing methods. In the case of $p = 1000$ even though CUS and SPC tend to yield slightly higher TPR, the average number of variables chosen by them are 84 (123 for Simulation 2) and 92 (96 for Simulation 2), respectively. This lack of sparsity of CUS and SPC is also noticed for all other values of p . The iterative BMA selects 25 variables for all the cases when $p > 25$ because its fitting algorithm can only choose 25 or fewer variables.

Table 1: Simulation results based on 100 replications when $C(x_i, x_j) = 0$: Overall average of operating characteristics (standard deviations in parentheses). The best operating characteristic value for each case is highlighted in bold.

No. of variables in data set (p)	Method	TPR %	TNR %	PPV %	NPV %	No. of variables selected
20	PSBC	100 (0)	97.3 (6.0)	97.7 (5.1)	100 (0)	10.3(0.6)
	CUS	99.7(1.7)	16.8(22.4)	55.5(8.4)	99.5(3.2)	18.3(2.3)
	SPC	76.7(15.6)	94.8(8.9)	95(7.7)	81.9(10.3)	8.2(2.1)
	BMA	98.7(7.6)	8.1(23.7)	53.1(9.4)	99.0(5.8)	19.1(2.9)
200	PSBC	98.9 (4.7)	99.3 (1.2)	90.8 (13.2)	99.9 (0.2)	11.2(2.3)
	CUS	83.6(19.9)	84.2(16.7)	37.0(25.0)	99.1(1.0)	38.4(32.7)
	SPC	61.6(24.2)	95.8(6.7)	68.9(29.2)	98.0(1.2)	14.2(14.5)
	BMA	41.6(16.6)	89.5(1.8)	18.1(11.0)	96.7(0.9)	24.1(3.0)
500	PSBC	86.5 (12.6)	99.3 (0.5)	75.3 (14.5)	99.7 (0.3)	11.9(2.8)
	CUS	73.3(22.8)	92.2(12.8)	36.0(27.3)	99.4(0.5)	45.6(63.8)
	SPC	69.8(21.8)	91.7(16.2)	45.5(28.9)	99.4(0.4)	47.4(80.8)
	BMA	45.5(11.4)	96.0(0.3)	19.0(5.7)	98.9(0.2)	24.3(2.3)
1000	PSBC	63.1(18.5)	99.7 (0.3)	68.8 (16.2)	99.6 (0.2)	9.7(3.6)
	CUS	61.4(30.3)	92.2(16.1)	42.1(36.6)	99.6 (0.3)	83.8(161.5)
	SPC	67.2 (24.0)	91.4(17.6)	38.5(28.4)	99.6 (0.2)	91.6(175.9)
	BMA	39.2(10.8)	98.0(0.3)	16.7(6.5)	99.4(0.1)	24.2(2.6)

On the other hand, our PSBC selects only 10 variables (also 10 for Simulation 2) on average and consistently provides much sparser model than CUS, SPC, and iterative BMA. This property is also noticed when $p = 20, 200$, and 500 . The standard deviations for the number of variables selected by CUS and SPC are alarmingly large for $p = 500$ and 1000 . This phenomenon is primarily because, CUS and SPC often end up selecting all the variables. Our PSBC does not have this problem.

The BIC curves (averaged over 100 replications) for all the four fitting methods are generated, and are given in Figure 2 and Figure 3. To generate BIC curves for the competing methods, we computed BIC values (23) based on their parameter estimates over the grid of tuning parameter values. The lowest point on the average BIC curve corresponds to the best choice of covariates. From figures 2 and 3 we can see that, based on the BIC curve, PSBC selects the correct number of covariates. On the other hand, only a very small subset of variables is chosen by CUS and SPC, giving rise to false negatives. Iterative BMA tends to select more variables than necessary in simulation 1 and very few in simulation 2. This indicates if BIC is used as the sole criterion for thresholding (variable selection), CUS, SPC, and iterative BMA can produce a sparse model like PSBC. However, the overall accu-

Table 2: Simulation results based on 100 replications when $C(x_i, x_j) = 0.5^{|i-j|}$: Overall average of operating characteristics (standard deviations in parentheses). The best operating characteristic value for each case is highlighted in bold.

No. of variables in data set (p)	Method	TPR %	TNR %	PPV %	NPV %	No. of variables selected
20	PSBC	100 (0)	93.4 (11.3)	94.7 (8.3)	100 (0)	10.7(1.1)
	CUS	99.9(1.0)	6.8(13.4)	52.0(4.4)	99.9(1.4)	19.3(1.4)
	SPC	89.6(13.4)	48.5(26.8)	66.0(12.6)	86.9(15.8)	14.1(3.5)
	BMA	83.8(30.4)	29.3(41.4)	59.9(18.7)	90.0(18.5)	15.5(6.8)
200	PSBC	96.2 (6.6)	98.9 (1.6)	86.7 (16.4)	99.8 (0.4)	11.6(3.1)
	CUS	88.9(15.4)	73.4(22.1)	23.7(16.2)	99.3(0.8)	59.4(42.9)
	SPC	76.8(20.5)	88.0(11.0)	37.7(22.7)	98.7(1.1)	30.5(22.3)
	BMA	40.0(16.0)	89.4(1.7)	17.1(8.1)	96.6(0.9)	24.2(3.1)
500	PSBC	86.5 (13.7)	99.3 (0.5)	74.2 (15.7)	99.7 (0.3)	12.1(2.9)
	CUS	78.6(21.7)	85.6(19.6)	26.0(23.5)	99.5(0.4)	78.4(97.2)
	SPC	74.0(23.0)	87.4(17.7)	30.5(25.1)	99.5(0.5)	69.2(88.3)
	BMA	42.2(13.9)	95.9(0.6)	17.9(8.6)	98.8(0.3)	24.3(2.5)
1000	PSBC	63.7(16.3)	99.7 (0.3)	71.9 (18.2)	99.7 (0.2)	9.5(3.8)
	CUS	71.0 (28.3)	88.3(20.2)	28.2(29.9)	99.7 (0.3)	122.5(201.7)
	SPC	68.9(24.6)	91.0(13.4)	29.1(25.9)	99.7 (0.2)	95.9(134)
	BMA	38.4(12.3)	97.9(0.3)	16.1(6.5)	99.4(0.1)	24.3(2.1)

racy is compromised, as sparsity for CUS, SPC, and iterative BMA is achieved at the cost of many false negatives.

Finally, our PSBC and the other three competing methods are also evaluated in terms of the prediction accuracy. Test sets of size $n = 100$ are generated (for all simulation scenarios) and the difference in test set deviance between a fitted model and the null model is computed using

$$DEV_j = -2\{l_j^{(test)}(\hat{\beta}_{(1:j)train}) - l_0^{(test)}(\mathbf{0})\}, \quad j = 1, \dots, p, \quad (25)$$

where $l_j^{(test)}(\hat{\beta}_{(1:j)train})$ and $l_0^{(test)}(\mathbf{0})$ are the log likelihoods for test data set evaluated by estimates from training set under the model $\mathcal{M}_j$, and the null model respectively. The small value of difference in test set deviance (25) indicates a good prediction performance. The average drop in test set deviance based on 100 random replications are presented in Figure 4 and Figure 5. We can see that the drop in test deviance plots of our PSBC model are always below the plots for CUS, SPC, and iterative BMA except for the case when $p = 20$. For $p = 20$, iterative BMA seems to be working a little better than PSBC, but performances of two methods are comparable. This validates the better prediction performance of our PSBC model.

Figure 2: Simulation results based on 100 replications when $C(x_i, x_j) = 0$: Average BIC curves for four different methods over four different number of variables in data set. The number of variables that results in minimum BIC is marked with a vertical line.

Figure 3: Simulation results based on 100 replications when $C(x_i, x_j) = 0.5^{|i-j|}$: Average BIC curves for four different methods over four different number of variables in data set. The number of variables that results in minimum BIC is marked with a vertical line.

Figure 4: Simulation results based on 100 replications when $C(x_i, x_j) = 0$: Average drop in test set deviance curves for four different methods over four different number of variables in data set.

Figure 5: Simulation results based on 100 replications when $C(x_i, x_j) = 0.5^{|i-j|}$: Average drop in test set deviance curves for four different methods over four different number of variables in data set.

5 Applications to Real-life Data Sets

We applied our PSBC model along with CUS, SPC, and iterative BMA to three different real life data sets. A detailed description of the data sets we used is given in Table 3. In all three data sets we have more genetic covariates than the subjects ($p \gg n$). In Data 1 (Norway/Stanford breast cancer data) and Data 3 (DLBCL data) the missing gene expressions are imputed by using the 10-nearest neighbors in the space of genes. Each data set is randomly divided into a training and a test set 10 different times. The number of patients used in the training and the test sets for each data is also provided in Table 3. We calculated the difference in test set deviance (25) between the fitted model and the null model over 10 random splits as we sequentially include a gene (or a variable) into the model based on its significance as done in section 4. The significance is measured by the estimated $|\hat{\beta}_j|$ for the corresponding gene. The average difference in test set deviance plot is given in Figure 6. From Figure 6 we can see that our PSBC works best among the four prediction methods for all three data sets. Despite variability in sample size and covariate dimension from data set to data set, our PSBC consistently gives the best prediction performance.

Table 3: Data description: The number of subjects in training set / the number of subjects in test set are presented in parentheses.

Data ID	Description	No. of subjects, n (training / test)	No. of genes or variables, p
Data 1. NSBC	Norway/Stanford breast cancer data (Sørli et al., 2003)	115 (76 / 39)	552
Data 2. DutchBC	Dutch breast cancer data (van Houwelingen et al., 2006)	295 (200 / 95)	4919
Data 3. DLBCL	Diffuse large-B-cell lymphoma data (Rosenwald et al., 2002)	240 (120 / 120)	7399

For the microarray data such as Data 2 (Dutch breast cancer data) and Data 3 (DLBCL data), the difference in test set deviance plots for iterative BMA cannot be clearly seen as it cannot complete the curve (iterative BMA cannot select more than 25 genes). Therefore, we considered another model evaluation criterion to overcome this difficulty in comparison. The relationship between a covariate X and a binary disease-status indicator D can be summarized in ROC curves by plotting estimates of the sensitivity of the test $X > c$ versus 1-specificity for all possible values c . Heagerty et al. (2000) extended this concept to time-dependent binary variables such as vital status, when censoring is present. Their method is known as the time-dependent receiver operator characteristic (ROC) curves. The sensitivity and specificity for cutoff c at time t for a given score function $f(X)$ is defined as

Figure 6: Results for real life data sets: Average difference in test set deviance for CUS, SPC, iterative BMA, and PSBC over 10 random splits.

$$\begin{aligned}\text{sensitivity}(c, t|f(X)) &= P[f(X) > c|\delta(t) = 1], \\ \text{specificity}(c, t|f(X)) &= P[f(X) \leq c|\delta(t) = 0],\end{aligned}\quad (26)$$

where $\delta(t)$ is the event indicator at time t . Based on estimates from the training set we estimate the risk scores (prognostic index) $f(X_{test}) = \mathbf{x}_{test}\hat{\boldsymbol{\beta}}_{train}$ for subjects in the test set by using four fitting methods. The higher value of prognostic index is indicative of a disease. A ROC curve can be drawn at each time t and overall predictive performance can be assessed by the area under the curve (AUC). The larger AUC at time t indicates better predictability at that particular time point. The time-dependent AUCs for all three data sets are calculated and displayed in Figure 7. For all three data sets, the AUC plot for our PSBC model is consistently above the plots for other competing methods, which indicates a higher prediction performance of our PSBC.

Additionally, the BIC curves for the data sets are computed as in Section 4. In Table 4 we report the number of genes selected by the four methods. For CUS and SPC, the optimal tuning parameters are obtained by 5-fold and 10-fold cross-validation, respectively. For iterative BMA, user-specified parameters are determined by 10-run/10-fold cross-validation.

Table 4: Results for real life data sets: The number of genes selected by PSBC, CUS, SPC, and iterative BMA.

Data ID	PSBC	CUS	SPC	iterative BMA
Data 1. NSBC	8	311	552	12
Data 2. DutchBC	11	288	130	20
Data 3. DLBCL	10	2735	286	25

To validate the biological significance of the selected genes in the DLBCL data set by our PSBC model, we report the GenBank ID and description of the corresponding top ten genes selected (by PSBC with BIC thresholding) in Table 5. It is important to note that four genes among these ten genes belong to Lymph-node signature group, proliferation signature group, and MHC class II group defined by Rosenwald et al. (2002). None of these top ten genes selected by PSBC is included in the Germinal-center B-cell group. However, when we investigated the list of top 14 genes by PSBC, the gene *ESTs* (AA505184) from Germinal-center B-cell group appeared in the model. Furthermore, out of the top ten genes selected by PSBC in Table 5, genes with GenBank ID D42043 and BC012161 are also selected by CUS, SPC, and iterative BMA. The other six genes, GenBank IDs X15804, X77743, NM_014456, AF127481, U15552, and M20430 selected by our PSBC are also marked as important by at least one of the competing methods as indicated in Table 5.

Figure 7: Results for real life data sets: Average AUC for CUS, SPC, iterative BMA, and PSBC over 10 random splits.

Table 5: DLBCL data set: The top 10 genes selected by PSBC with BIC thresholding based on the 240 patients.

GenBank ID	Signature	Description
X15804*♣	Lymph node	Actinin, alpha 1
D42043*♣♦♠		KIAA0084 protein
X90858*	Proliferation	Uridine phosphorylase
X77743*♦		Cyclin-dependent kinase 7 (MO15 homolog, <i>Xenopus laevis</i> , cdk-activating kinase)
BC012161*♣♦♠	Proliferation	Septin 1
NM_014456*♣♦		programmed cell death 4 (neoplastic transformation inhibitor)
AF127481*♣	MHC class II	lymphoid blast crisis oncogene
U15552*♣♦		acidic 82 kDa protein mRNA
M20430*♣♦		major histocompatibility complex, class II, DR beta 5
U46767*		small inducible cytokine subfamily A (Cys-Cys), member 13

*: The gene selected by PSBC, ♣: The gene selected by CUS, ♦: The gene selected by SPC, ♠: The gene selected by iterative BMA

6 Discussion

One major goal of microarray technology is to build a predictive model, by relating the censored survival time to gene expression profiles. The predictive model is required to have high predictive accuracy and also be parsimonious. In this paper, we proposed a penalized semiparametric Bayesian Cox (PSBC) model for such purpose. The performance of variable selection and prediction of our model is compared with three other methods (CUS, SPC, and iterative BMA) on both simulated and real data sets. We used variable selection criteria, like TPR, TNR, PPV, and NPV to show the superiority of our PSBC in the simulated data sets. Furthermore graphical criteria, like drop in test set deviance and time dependent AUC curves, were used to evaluate the prediction power of our model. In almost all the simulations and real data cases, our PSBC outperforms the other three competing methods.

The tuning parameter in our PSBC model is adaptively chosen by putting a prior distribution on λ^2 , whereas, in CUS and SPC λ is determined by a cross validation method. Cross-validation procedures often result in an unstable choice of tuning parameter and it is heavily dependent on the choice of the range of λ . Although we used the other method (CUS estimates) to choose the prior parameter values of λ in our numerical studies, it is not necessary. In fact, one could consider a non-informative prior on λ . This can be done since the posterior distribution of

λ is largely affected by the number of variables (p). In other words, if p is very large the prior parameters are dominated and they do not have a significant effect on the posterior distribution of λ . In our PSBC model we get a clear picture of λ by obtaining the full posterior distribution (22). The distribution of the estimates of the tuning parameter λ from the simulated data sets are shown in Figure 8 and Figure 9. From Figure 8 and 9, we can see that the estimate of λ gets larger as we have more variables in a data set. This ensures that PSBC will have parsimony for high-dimensional data. Note that, in the Figure 8 and 9, we have nonzero estimates for the tuning parameter for the case when $p = 20$. This indicates that, PSBC has good variable selection capability for the low dimensional data as well. Unlike PSBC, the method CUS often results in $\hat{\lambda}_{CUS} = 0$ for $p = 20$, so it selects all the variables giving rise to false positives as we can see in Table 1 and Table 2.

Although PSBC gives very impressive results in the simulated and real life data sets, there are still several matters that could be taken into account. For example, the genes are often grouped together by their role play, and therefore, it is desired to select or drop the whole group of genes together in a variable selection method. Group lasso (Yuan and Lin, 2006), fused lasso (Tibshirani et al., 2005), and elastic-net (Zou and Hastie, 2005) are proposed for grouped variable selection. Bayesian priors corresponding to these shrinkage and grouping methods are introduced by Kyung et al. (2010) under linear model setup. By incorporating the Bayesian priors into our PSBC, we can generalize it in such a way that a highly pairwise-correlated group of variables can be suitably picked or dropped together. In many clinical trials and medical settings, it is quite common to have interval censored survival data as well as right censored data. Similar to our PSBC, we can develop a model to handle interval censored data by replacing the likelihood in (10) by a beta process model (Sinha, 1997). Moreover, under the multiplicative assumption similar Bayesian variable selection technique can also be developed for the accelerated failure time (AFT) model. These are some future directions of our research.

Figure 8: Simulation results based on 100 replications when $C(x_i, x_j) = 0$: Histogram of $\hat{\lambda}_{PSBC}$.

Figure 9: Simulation results based on 100 replications when $C(x_i, x_j) = 0.5^{|i-j|}$: Histogram of $\hat{\lambda}_{PBC}$.

Appendix A Unimodality Under Prior (12)

In this section, we prove unimodality of the joint posterior distribution $\pi(\boldsymbol{\beta}, \sigma^2|D)$ of $\boldsymbol{\beta}$ and $\sigma^2 > 0$ under the prior (12). The unimodality is established for common choices of $\pi(\sigma^2)$ and any choice of $\lambda \geq 0$, in the sense that for every $c > 0$, the upper level set $\{(\boldsymbol{\beta}, \sigma^2) | \pi(\boldsymbol{\beta}, \sigma^2) > 0\}$ is connected.

The marginal joint posterior distribution $\pi(\boldsymbol{\beta}, \sigma^2|D)$ is obtained by integrating out $\mathbf{h}$ from $\pi(\boldsymbol{\beta}, \mathbf{h}, \sigma^2|D)$ as follows.

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma^2|D) &= \int_{\mathbb{R}_+^J} \pi(\boldsymbol{\beta}, \mathbf{h}, \sigma^2|D) d\mathbf{h} \\ &\propto \pi(\boldsymbol{\beta}|\sigma^2) \pi(\sigma^2) \int_{\mathbb{R}_+^J} L(D|\boldsymbol{\beta}, \mathbf{h}) \pi(\mathbf{h}) d\mathbf{h} \\ &\propto \pi(\boldsymbol{\beta}|\sigma^2) \pi(\sigma^2) \int_{\mathbb{R}_+^J} \prod_{j=1}^J h_j^{(\alpha_{0j}-\alpha_{0,j-1})-1} \exp \left\{ -h_j \left(c_0 + \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\mathbf{x}'_k \boldsymbol{\beta}) \right) \right\} \\ &\quad \times \prod_{l \in \mathcal{D}_j} [1 - \exp\{-h_j \exp(\mathbf{x}'_l \boldsymbol{\beta})\}] d\mathbf{h}. \end{aligned} \quad (27)$$

Using the first order Taylor expansion we have,

$$\prod_{l \in \mathcal{D}_j} [1 - \exp\{-h_j \exp(\mathbf{x}'_l \boldsymbol{\beta})\}] \approx h_j^{d_j} \prod_{l \in \mathcal{D}_j} \exp(\mathbf{x}'_l \boldsymbol{\beta}). \quad (28)$$

Therefore,

$$\begin{aligned} &\int_{\mathbb{R}_+^J} L(D|\boldsymbol{\beta}, \mathbf{h}) \pi(\mathbf{h}) d\mathbf{h} \\ &\approx \int_{\mathbb{R}_+^J} \prod_{j=1}^J h_j^{(\alpha_{0j}-\alpha_{0,j-1}+d_j)-1} \exp \left\{ -h_j \left(c_0 + \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\mathbf{x}'_k \boldsymbol{\beta}) \right) \right\} \prod_{l \in \mathcal{D}_j} \exp(\mathbf{x}'_l \boldsymbol{\beta}) d\mathbf{h} \\ &= \prod_{j=1}^J \frac{(\alpha_{0j} - \alpha_{0,j-1} + d_j - 1)!}{\{c_0 + \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\mathbf{x}'_k \boldsymbol{\beta})\}^{\alpha_{0j} - \alpha_{0,j-1} + d_j}} \prod_{l \in \mathcal{D}_j} \exp(\mathbf{x}'_l \boldsymbol{\beta}). \end{aligned} \quad (29)$$

Combining (27) and (29), the marginal joint posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ can be written as,

$$\pi(\boldsymbol{\beta}, \sigma^2|D) \propto \pi(\boldsymbol{\beta}|\sigma^2) \pi(\sigma^2) \prod_{j=1}^J \frac{(\alpha_{0j} - \alpha_{0,j-1} + d_j - 1)!}{\{c_0 + \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\mathbf{x}'_k \boldsymbol{\beta})\}^{\alpha_{0j} - \alpha_{0,j-1} + d_j}} \prod_{l \in \mathcal{D}_j} \exp(\mathbf{x}'_l \boldsymbol{\beta}). \quad (30)$$

Only considering the terms that involve either $\boldsymbol{\beta}$ or σ^2 , the log posterior is given by,

$$\begin{aligned} \ln(\pi(\sigma^2)) &= \frac{p}{2} \ln \sigma^2 - \frac{\lambda}{\sqrt{\sigma^2}} \|\boldsymbol{\beta}\|_1 \\ &+ \sum_{j=1}^J \left\{ -(\alpha_{0j} - \alpha_{0,j-1} + d_j) \ln \left(c_0 + \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} e^{\mathbf{x}'_k \boldsymbol{\beta}} \right) + \sum_{l \in \mathcal{D}_j} \mathbf{x}'_l \boldsymbol{\beta} \right\}. \end{aligned} \quad (31)$$

We can see the second and third terms are trivially concave in $(\boldsymbol{\beta}, \sigma^2)$. We denote the last term in (31) by $g(\boldsymbol{\beta})$. To show $g(\boldsymbol{\beta})$ is concave in each component of $\boldsymbol{\beta}$, it suffices to show that $\frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_r^2} \leq 0, \forall r = 1, \dots, p$.

$$\begin{aligned} \frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_r^2} &= \frac{(\alpha_{0j} - \alpha_{0,j-1} + d_j)}{(c_0 + e^{\mathbf{x}'_k \boldsymbol{\beta}})^2} \\ &\times \left\{ \left(\sum_{k \in \mathcal{R}_j - \mathcal{D}_j} x_{kr} e^{\mathbf{x}'_k \boldsymbol{\beta}} \right)^2 - \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} x_{kr}^2 e^{\mathbf{x}'_k \boldsymbol{\beta}} \left(c_0 + \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} e^{\mathbf{x}'_k \boldsymbol{\beta}} \right) \right\}. \end{aligned} \quad (32)$$

By the Cauchy-Schwarz inequality, we get

$$\left(\sum_{k \in \mathcal{R}_j - \mathcal{D}_j} x_{kr} e^{\mathbf{x}'_k \boldsymbol{\beta}} \right)^2 \leq \left(\sum_{k \in \mathcal{R}_j - \mathcal{D}_j} x_{kr}^2 e^{\mathbf{x}'_k \boldsymbol{\beta}} \right) \left(\sum_{k \in \mathcal{R}_j - \mathcal{D}_j} e^{\mathbf{x}'_k \boldsymbol{\beta}} \right) \quad (33)$$

Since $(\alpha_{0j} - \alpha_{0,j-1} + d_j) > 0$, using (33) we get $\frac{\partial^2 g(\boldsymbol{\beta})}{\partial \beta_r^2} \leq 0, \forall r = 1, \dots, p$. Therefore, $g(\boldsymbol{\beta})$ is concave in each component of $\boldsymbol{\beta}$. That implies (31) is concave in $(\boldsymbol{\beta}, \sigma^2)$ provided that $\ln \pi(\sigma^2)$ is concave. Hence the posterior is unimodal with any noninformative scale-invariant marginal prior $\pi(\sigma^2)$.

REFERENCES

- Annest, A., Bumgarner, R.E., Raftery, A.E., and Yeung, K.Y. (2009), "Iterative Bayesian Model Averaging: a method for the application of survival analysis to high-dimensional microarray data," *BMC Bioinformatics*, **10**:72.
- Bair, E., Paul, D., and Tibshirani, R. (2006), "Prediction by supervised principal components," *Journal of the American Statistical Association*, **101**, 119.137.
- Bøvelstad, H.M et al. (2007), "Predicting Survival from Microarray Data - a comparative study," *Bioinformatics*, **23**, 2080-2087.
- Burridge, J. (1981), "Empirical Bayes Analysis for Survival Time Data," *Journal of the Royal Statistical Society, Series B*, **43**, 65-75.
- Casella, G. (2001), "Empirical Bayes Gibbs Sampling," *Biostatistics*, **2**, 485-500.
- Chakraborty, S. et al. (2007), "Gene Expression-Based Glioma Classification Using Hierarchical Bayesian Vector Machines," *Sankhya*, **69**, 514-547.
- Chakraborty, S. (2009), "Simultaneous Cancer Classification and Gene Selection with Bayesian Nearest Neighbor Method: An Integrated Approach," *Computational Statistics and Data Analysis*, **53**, 1462-1474.
- Cox, D.R. (1972), "Regression Models and Life Tables (with discussion)," *Journal of the Royal Statistical Society, Series B*, **34**, 187-220.
- Cox, D.R. (1975), "Partial Likelihood," *Biometrika*, **62**, 269-276.
- Du, P., Ma, S., and Liang, H. (2010), "Penalized Variable Selection Procedure for Cox Models with Semiparametric Relative Risk," *Annals of Statistics*, **38**, Number 4, 2092-2117.
- Engler, D. and Li, Y. (2009), "Survival Analysis with High-Dimensional Covariates: An Application in Microarray Studies," *Statistical Applications in Genetics and Molecular Biology*, **8**, Article 14.
- George, E. and McCulloch, R. (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, **88**, 881-889.
- Gilks, W. R. (1992), "Derivative-free Adaptive Rejection Sampling for Gibbs Sampling," *Bayesian Statistics 4*, Oxford University Press.
- Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995), "Adaptive Rejection Metropolis Sampling," *Applied Statistics*, **44**, 455-472.

- Gui, J. and Li, H. (2005), "Penalized Cox Regression Analysis in the High-dimensional and Low-sample Size Settings, with Applications to Microarray Gene Expression Data," *Bioinformatics*, **21**, 3001-3008.
- Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000), "Time dependent ROC curves for censored survival data and a diagnostic marker," *Bioinformatics*, **56**, 337-344.
- Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001), *Bayesian Survival Analysis*, Springer.
- Kalbfleisch, J. D. (1978), "Nonparametric Bayesian Analysis of Survival Time Data," *Journal of the Royal Statistical Society, Series B*, **40**, 214-221.
- Kalish, L. A. (1992), "Phase III Multiple Myeloma: Evaluation of Combination Chemotherapy in Previously Untreated Patients," *Technical Report #726E*, Department of Biostatistics, Dana-Farber Cancer Institute.
- Kyung, M., Gill, J., Ghosh, M., and Casella, J. (2010), "Penalized Regression, Standard Errors, and Bayesian Lassos," *International Society for Bayesian Analysis*, **5**, 1-44.
- Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003), "Gene Selection: A Bayesian Variable Selection Approach," *Bioinformatics*, **19**, 90-97.
- Park, T. and Casella, G. (2008), "Bayesian Lasso," *Journal of the American Statistical Association*, **103**, 681-686.
- Rosenwald, A. et al. (2002), "The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large B-Cell Lymphoma," *The New England Journal of Medicine*, **346**, 1937-1947.
- Schumacher, M. et al. (1994), "Randomized 2 x 2 Trial Evaluating Hormonal Treatment and the Duration of Chemotherapy in Node-Positive Breast Cancer Patients. German Breast Cancer Study Group," *Journal of Clinical Oncology*, **12**, 2086-2093.
- Sha et al. (2004), "Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage," *Biometrics*, **60**, 812-819.
- Shinha (1997), "Time-discrete beta process model for interval-censored survival data," *Canadian Journal of Statistics*, **25**, 445-446.

- Sørbye, T. et al. (2003), "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proceedings of the National Academy of Sciences of the USA*, **100**, 8418-8423.
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, **58**, 267-288.
- Tibshirani, R. (1997), "The Lasso Method for Variable Selection in the Cox Model," *Statistics in Medicine*, **16**, 385-395.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society, Series B*, **67**, 91-108.
- Tibshirani, R. (2009), "Univariate Shrinkage in the Cox Model for High Dimensional Data," *Statistical Applications in Genetics and Molecular Biology*, **8**, Article 21.
- van de Vijver, M. J. et al. (2002), "A gene-expression signature as a predictor of survival in breast cancer," *The New England Journal of Medicine*, **347**, 1999-2009.
- van Houwelingen, H. C. et al. (2006), "Cross-validated Cox Regression on Microarray Gene Expression Data," *Statistics in Medicine*, **25**, 3201-3216.
- Verweij, P. and van Houwelingen, H. (1993), "Cross-validation in Survival Analysis," *Statistics in Medicine*, **12**, 2305-2314.
- Volinsky, C.T., Raftery, A.E. (2000), "Bayesian Information Criterion for Censored Survival Models," *Biometrics*, **56**, 256-262.
- Yuan, M. and Lin, Y. (2005), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society, Series B*, **68**, 49-67.
- Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, **67**, 301-320.