The Annals of Applied Statistics
2020, Vol. 14, No. 2, 809–828
https://doi.org/10.1214/20-AOAS1325

BAYESIAN VARIABLE SELECTION FOR SURVIVAL DATA USING INVERSE MOMENT PRIORS

BY AMIR NIKOOIENEJAD[1], WENYI WANG[2] AND VALEN E. JOHNSON[3]

[1]*Department of Global Statistical Sciences, Eli Lilly and Company, nikooienejad_amir@lilly.com*
[2]*Department of Bioinformatics and Computational Biology, MD Anderson Cancer Center, wwang7@mdanderson.org*
[3]*Department of Statistics, Texas A&M University, vjohnson@stat.tamu.edu*

Efficient variable selection in high-dimensional cancer genomic studies is critical for discovering genes associated with specific cancer types and for predicting response to treatment. Censored survival data is prevalent in such studies. In this article we introduce a Bayesian variable selection procedure that uses a mixture prior composed of a point mass at zero and an inverse moment prior in conjunction with the partial likelihood defined by the Cox proportional hazard model. The procedure is implemented in the R package BVSNLP, which supports parallel computing and uses a stochastic search method to explore the model space. Bayesian model averaging is used for prediction. The proposed algorithm provides better performance than other variable selection procedures in simulation studies and appears to provide more consistent variable selection when applied to actual genomic datasets.

1. Introduction. Recent developments in sequencing technology have made it easier to collect massive genomic datasets that can be used to study cancer and other diseases. Given such data, there is great interest in linking genomic data to patient outcomes, and in many cases such outcomes are censored survival times.

Survival times for patients generally represent either the time to death or disease progression, the time to study termination or the time until the subject is lost to follow up. In the latter cases the subject's survival time is *censored*. The relation between survival times and covariates is modeled through the hazard function, which is the limiting rate of death in the interval $(t, t + \Delta t)$ as Δt becomes small, given patient covariates. More precisely, the hazard function h for patient i may be defined as

$$h(t \mid \mathbf{x}_i) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T \leq t + \Delta t \mid T \geq t, \mathbf{x}_i), \tag{1.1}$$

where $\mathbf{x}_i$ is a p vector of covariates thought to influence survival. We denote by $\mathbf{X}$ the $n \times p$ design matrix obtained by stacking n patient covariate vectors. Proportional hazard models take the form

$$h(t \mid \mathbf{x}_i) = h_0(t)\Phi(\mathbf{x}_i) \tag{1.2}$$

with an identifiability constraint of $\Phi(\mathbf{0}) = 1$. In this formula, $h_0(t)$ denotes the baseline hazard function. The Cox proportional hazards model (Cox (1972)) is defined by taking $\Phi(\mathbf{x}_i) = \exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}$, leading to

$$h(t \mid \mathbf{x}_i) = h_0(t) e^{\mathbf{x}_i^T \boldsymbol{\beta}}. \tag{1.3}$$

Here, $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients.

Received June 2018; revised January 2020.
Key words and phrases. Bayesian variable selection, nonlocal prior, high-dimensional data, survival data analysis, Cox proportional hazard model, cancer genomics.

An important feature of the proportional hazards model is that it yields a partial likelihood function that is independent of the baseline hazard function, h_0. For complete survival analyses, however, the baseline hazard function is necessary for predicting survival times and can be estimated nonparametrically. Further details regarding the Cox proportional hazard model may be found in Cox and Oakes (1984), Kalbfleisch and Prentice (1980) or Cox (1972).

Gene expression datasets usually contain measurements on thousands of genes collected for only hundreds of subjects. Biologically, it seems plausible that only a relatively small number of these genes contribute significantly to survival. This implies that most of the elements in the vector $\boldsymbol{\beta}$ are small or close to zero. The challenge is to find covariates with nonzero coefficients or, equivalently, those genes that contribute the most in determining the survival outcome.

Many common penalized likelihood methods originally introduced for linear regression have been extended to survival data. These methods include LASSO (Tibshirani et al. (1997)), in which an L_1 penalty is imposed on regression coefficients. Zhang and Lu (2007) utilized adaptive LASSO methodology for time to event data, while Antoniadis, Fryzlewicz and Letué (2010) adopted the Dantzig selector for survival outcomes. The extension of nonconvex penalized likelihood approaches, in particular SCAD, to the Cox proportional hazard model is discussed in Fan and Li (2002). The Iterative Sure Independence Screening (ISIS) approach introduced by Fan and Lv (2008) is also extended for ultrahigh dimensional survival data in Fan, Feng and Wu (2010), where it is used on Cox proportional hazard models and the SCAD penalty is employed for variable selection.

Some Bayesian approaches have also been introduced. Faraggi and Simon (1998) proposed a method based on approximating the posterior distribution of the parameters in the proportional hazard model by defining a Gaussian prior on regression coefficients. A loss function was then imposed to select a parsimonious model. A semiparametric Bayesian approach was utilized by Ibrahim, Chen and MacEachern (1999), who employed a discrete gamma process for the baseline hazard function and a multivariate Gaussian prior for the coefficient vector. Sha, Tadesse and Vannucci (2006) considered Accelerated Failure Time (AFT) models along with data augmentation to impute failure times. A mixture prior proposed by George and McCulloch (1997) was used to impose sparsity. In more recent work, Held, Gravestock and Sabanés Bové (2016) proposed the use of a g-prior model for the coefficient vector and employed test-based Bayes factors (Johnson (2005)) to the Cox proportional hazard models. However, this method is intended for use only when the number of covariates is less than the number of observations, that is, when $p < n$.

To our knowledge, all previous Bayesian procedures for variable selection in survival data have used local priors on model coefficients. In Bayesian hypothesis tests, local priors put a positive probability on the null value of the parameter, in this case zero, whereas nonlocal priors put zero probability on the null value. Johnson and Rossell (2010) can be consulted for more discussion on properties of local and nonlocal priors in the context of Bayesian testing. In this article we propose a Bayesian method based on a mixture prior comprised of a point mass at zero and a nonlocal prior on the regression coefficients. To handle the computational burden of implementing the resulting procedure, we employ a stochastic search method, S5 (Shin, Bhattacharya and Johnson (2018)), which we implement in an R package BVSNLP. We also discuss a general procedure for setting the tuning parameter of the nonlocal prior.

This article is structured as follows. In Section 2 we introduce notation and discuss the modeling of the problem in a Bayesian framework. Section 3 discusses the proposed method, with details of parameter selection, model search and assessment of the accuracy of the proposed variable selection procedure. Sections 4 and 5 provide simulation and real data analyses with various predictive performance measures to demonstrate how the proposed method compares to several other competing methods. Section 6 concludes with discussion.

2. Problem modeling.

2.1. *Preliminaries.* Let T_i denote the survival time and C_i denote the censoring time for individual i. Each element in the observed vector of survival times, $\mathbf{y}$, is defined as $y_i = \min\{T_i, C_i\}$. The status for each individual is defined as $\delta_i = I(T_i \le C_i)$. The status vector is represented by $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)^T$. We assume that the censoring mechanism is "at random," meaning that C_i and T_i are conditionally independent given $\mathbf{x}_i$, where $\mathbf{x}_i \in \mathbb{R}^p$ are the covariates for individual i and comprise the ith row of $\mathbf{X}$. The observed data is of the form $\{(y_i, \delta_i, \mathbf{x}_i); i = 1, 2, \ldots, n\}$.

Model $\mathbf{k}$ is defined as $\mathbf{k} = \{k_1, \ldots, k_j\}$, where $(1 \le k_1 < \cdots < k_j \le p)$, and it is assumed that $\beta_{k_1} \ne 0, \ldots, \beta_{k_j} \ne 0$ and all other elements of $\boldsymbol{\beta}$ are 0. The design matrix corresponding to model $\mathbf{k}$ is denoted by $\mathbf{X}_\mathbf{k}$ and the regression vector by $\boldsymbol{\beta}_\mathbf{k} = (\beta_{k_1}, \beta_{k_2}, \ldots, \beta_{k_j})^T$.

Let $\mathcal{R}(t) = \{i : y_i \le t\}$ represent the *risk set* at time t, the set of all individuals who are still present in the study at time t and are neither dead nor censored. We assume throughout this article that the failure times are distinct. In other words, only one individual fails at a specific failure time. With this assumption and letting $\xi_{\mathbf{k}_i} = \exp\{\mathbf{X}_{\mathbf{k}_i}^T \boldsymbol{\beta}_\mathbf{k}\}$, the partial likelihood (Cox (1972)) for $\boldsymbol{\beta}_\mathbf{k}$ in model $\mathbf{k}$ can be written as

$$L_p(\boldsymbol{\beta}_\mathbf{k}) = \prod_{i=1}^{n} \left[\frac{\xi_{\mathbf{k}_i}}{\sum_{j \in R(y_i)} \xi_{\mathbf{k}_j}} \right]^{\delta_i}. \tag{2.1}$$

Our method uses this partial likelihood as the sampling distribution in our Bayesian model selection procedure. We acknowledge that there is some information loss in (2.1) with respect to $\boldsymbol{\beta}_\mathbf{k}$. For instance, Basu (Ghosh (1988)) argues that partial likelihoods cannot usually be interpreted as sampling distributions. On the other hand, Berger, Liseo and Wolpert (1999) encourage the use of partial likelihoods when the nuisance parameters are marginalized out.

Sorting the observed unique survival times in ascending order and, consequently, reordering the status vector $\boldsymbol{\delta}$ as well as the design matrix $\mathbf{X}$ with respect to the ordered $\mathbf{y}$, the sampling distribution of $\mathbf{y}$ for model $\mathbf{k}$ can be written as

$$\pi(\mathbf{y} \mid \boldsymbol{\beta}_\mathbf{k}) = \prod_{i=1}^{n} \left[\frac{e^{\mathbf{X}_{\mathbf{k}_i}^T \boldsymbol{\beta}_\mathbf{k}}}{\sum_{j=i}^{n} e^{\mathbf{X}_{\mathbf{k}_j}^T \boldsymbol{\beta}_\mathbf{k}}} \right]^{\delta_i}. \tag{2.2}$$

A Bayesian hierarchical model can be defined in which $\pi(\mathbf{y} \mid \boldsymbol{\beta}_\mathbf{k})$ in (2.2) represents the sampling distribution, $\pi_\mathbf{k}(\boldsymbol{\beta}_\mathbf{k})$ is the prior of model coefficients $\boldsymbol{\beta}_\mathbf{k}$ and $P(\mathbf{k})$ is the prior for model $\mathbf{k}$. Using Bayes rule, the posterior probability for model $\mathbf{j}$ is written as

$$P(\mathbf{j} \mid \mathbf{y}) = \frac{P(\mathbf{j}) m_\mathbf{j}(\mathbf{y})}{\sum_{k \in \mathcal{J}} P(\mathbf{k}) m_\mathbf{k}(\mathbf{y})}, \tag{2.3}$$

where $\mathcal{J}$ is the set of all possible models and the marginal probability of the data under model $\mathbf{k}$ is defined by

$$m_\mathbf{k}(\mathbf{y}) = \int \pi(\mathbf{y} \mid \boldsymbol{\beta}_\mathbf{k}) \pi_\mathbf{k}(\boldsymbol{\beta}_\mathbf{k}) \, d\boldsymbol{\beta}_\mathbf{k}. \tag{2.4}$$

The prior density for $\boldsymbol{\beta}_\mathbf{k}$ and the prior on the model space impact the overall performance of the selection procedure and the amount of sparsity imposed on candidate models. Note that the sampling distribution in (2.2) is continuous in $\boldsymbol{\beta}_\mathbf{k}$, and in Section 2.3 we define an inverse moment prior (Johnson and Rossell (2010)) on each of the coefficients in model $\mathbf{k}$.

2.2. *Prior on model space.* Let $\boldsymbol{\gamma}_{\mathbf{k}} = \{\gamma_1, \ldots, \gamma_p\}$ denote a binary vector indicating which covariates are included in model $\mathbf{k}$. Suppose the size of model $\mathbf{k}$ is k. That is, there are k nonzero indices in $\boldsymbol{\gamma}_{\mathbf{k}}$. The nonzero indices of $\boldsymbol{\gamma}_{\mathbf{k}}$ represent the indices of the nonzero elements in the coefficient vector, $\boldsymbol{\beta}$, which a priori are modeled as independent Bernoulli random variables with success probability $P(\gamma_i = 1) = \theta$ for every $1 \le i \le p$. As discussed in Scott and Berger (2010), no fixed value for θ adjusts for multiplicity. As a result, it is necessary to define a prior on θ, say $\pi(\theta)$. The resulting marginal probability for model $\mathbf{k}$ in a fully Bayesian approach may then be written as

$$p(\mathbf{k}) \propto \int \theta^k (1-\theta)^{p-k} \pi(\theta), d\theta. \tag{2.5}$$

A common choice for $\pi(\theta)$ is the beta distribution, $\theta \sim \text{Beta}(a, b)$, where in the special case of $a = b = 1$, $\pi(\theta)$ is a uniform distribution. The marginal probability for model $\mathbf{k}$ derived from (2.5) is then equal to

$$p(\mathbf{k}) = \frac{B(a+k, b+p-k)}{B(a,b)}, \tag{2.6}$$

where $B(\cdot)$ is the Beta function. A priori, the model size, k, thus follows a Beta-binomial distribution. By choosing $b = p - a$, the mean and variance of the selected model size k is

$$\mathbb{E}(k) = a, \qquad \text{Var}(k) = \frac{2p^2 a(p-a)}{p^3} \approx 2a. \tag{2.7}$$

The approximation in the variance formula follows from a large p and a fairly small a under the sparsity assumption on the true model size. To incorporate the belief that the optimal predictive models are sparse, we recommend setting $a = 1$ and $b = p - a$. The resulting prior assigns comparatively small prior probabilities to models that contain many covariates.

2.3. *Product inverse MOMent (piMOM) prior.* We impose nonlocal prior densities on the nonzero coefficients, $\boldsymbol{\beta}_{\mathbf{k}}$. Specifically, we assume the prior densities on the nonzero coefficients in model $\mathbf{k}$ take the form of a product of independent iMOM priors, or piMOM densities (Johnson and Rossell (2012)), expressible as

$$\pi(\boldsymbol{\beta}_{\mathbf{k}} \mid \tau, r) = \frac{\tau^{rk/2}}{\Gamma(r/2)^k} \prod_{i=1}^{k} |\beta_i|^{-(r+1)} \exp\left(-\frac{\tau}{\beta_i^2}\right), \quad r, \tau > 0. \tag{2.8}$$

The hyperparameter τ represents a scale parameter that determines the dispersion of the prior around $\mathbf{0}$, while r determines the tail behavior of the density. These priors have two symmetric modes with Cauchy-like tails when $r = 1$, and assign negligible probability to a region around zero. In comparison to local priors, this characteristic of nonlocal priors potentially leads to smaller false positive rates in selection procedures by discouraging the selection of variables with small coefficients. In addition, piMOM priors possess Cauchy-like tails which introduce comparatively small penalties on large coefficients. Unlike many penalized likelihood methods, large values of regression coefficients are thus not heavily penalized by these priors. As a result, they do not necessarily impose significant penalties on nonsparse models, provided that the estimated coefficients in those models are not small. For these reasons piMOM priors work well as a default choice of priors on nonnegligible coefficients in variable selection problems. An example of an iMOM prior is depicted in Figure 1 for $r = 1$ and $\tau = 0.5$.

Another nonlocal prior that might be considered as a potential candidate for the prior densities on the nonzero coefficients in model $\mathbf{k}$ is the product of independent MOM priors, or the pMOM densities (Johnson and Rossell (2012)). A detailed discussion on the

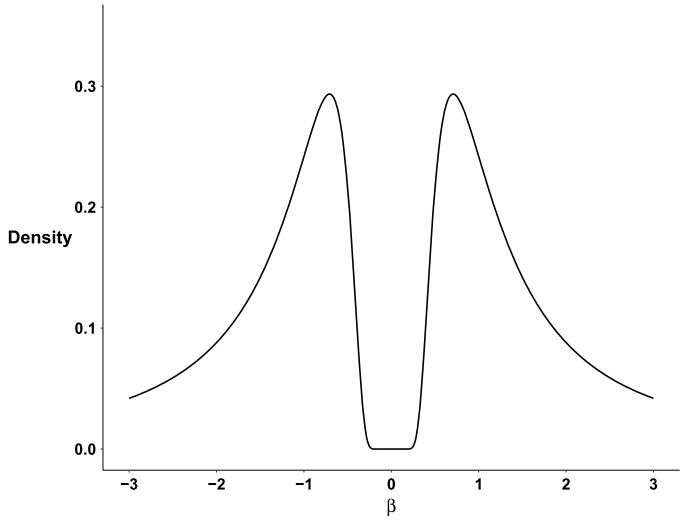


FIG. 1. *iMOM prior with* $r = 1$ *and* $\tau = 0.5$.

pMOM priors and their properties is provided in Section 3 of the Supplementary Material (Nikooienejad, Wang and Johnson (2020)). A simulation analysis to compare the performance of piMOM and pMOM in the selection process is also provided. For the reasons discussed there, piMOM-based procedures are more effective for variable selection in $p \gg n$ settings and, therefore, is our choice for the analyses of the simulation and real data in this article.

3. Methods.

3.1. *Selection of hyperparameters.* We use the procedure described in Nikooienejad, Wang and Johnson (2016) to select hyperparameter values for the piMOM prior. In that method the null distribution of the maximum likelihood estimator for $\boldsymbol{\beta}_{\mathbf{k}}$ (i.e., all components of $\boldsymbol{\beta}_{\mathbf{k}}$ are 0), obtained from randomly selected design matrices $\mathbf{X}_{\mathbf{k}}$, is compared to the prior density on $\boldsymbol{\beta}_{\mathbf{k}}$ for various values of (r, τ). Fixing $r = 1$ to achieve Cauchy-like tails, a value of τ is chosen so that the overlap between the two densities is less than a specified threshold, $1/\sqrt{p}$, and is denoted by τ_1. It can be shown that the maximum of the iMOM prior occurs at $\pm\sqrt{\tau}$. We also allow users to input a prior parameter α that controls where the modes in the prior occur. This can be useful in constraining the prior density when covariates are highly correlated (resulting in an over-dispersed prior when the sampling distribution of the null MLE under the null model becomes overly broad). We then set the value of τ according to

$$\tau = \min\bigl(\tau_1, \alpha^2\bigr). \tag{3.1}$$

To implement the procedure for computing τ_1 for survival models, we generate response vectors under the null model using the procedure described by Bender, Augustin and Blettner (2005). Survival times are sampled from a standard exponential model.

Let $\boldsymbol{t}^s$ and $\boldsymbol{c}^s$ be the vector of sampled survival times and censoring times, respectively. The sampled survival time and status for each observation is then computed as

$$y_i^s = \min\bigl\{t_i^s, c_i^s\bigr\} \quad \text{and} \quad \delta_i^s = I\bigl(t_i^s \le c_i^s\bigr) \tag{3.2}$$

which comprise $\mathbf{y}^s$ and $\boldsymbol{\delta}^s$ under the null model. Using the pair $(\mathbf{y}^s, \boldsymbol{\delta}^s)$, the MLE from a Cox model is computed. It should be noted that the asymptotic distribution of the MLE for

the Cox model under the null hypothesis is $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\mathbf{0}, I(\hat{\boldsymbol{\beta}}))$, where $I(\boldsymbol{\beta})$ is the information matrix of the partial likelihood function. Thus, it is appropriate to approximate the pooled estimated coefficients in that algorithm with a normal density function. When the sample size gets large, the variance of the MLE decreases and causes the overlap to become small and, consequently, small values of τ are selected.

In general, we find that $r = 1$ and $\tau = 0.25$ are good default values if one chooses not to run the hyperparameter selection algorithm. When $r = 1$, the peaks of the iMOM prior occur at $-\sqrt{\tau}$ and $\sqrt{\tau}$. By equating $\sqrt{\tau}$ to the absolute value of the expected effect size for a given application, insight can be gained on what value of τ is appropriate. Further details regarding this algorithm can be found in Nikooienejad, Wang and Johnson (2016).

3.2. *Computing posterior probability of models.* Computing the posterior probability for each model requires the marginal probability of observed survival times under each model, as shown in (2.3) and (2.4). The marginal probability is approximated using the Laplace approximation, where the regression coefficients in $\boldsymbol{\beta}_{\mathbf{k}}$ are integrated out. This leads to

$$m_{\mathbf{k}}(\mathbf{y}_n) = \pi(\mathbf{y}_n \mid \hat{\boldsymbol{\beta}}_{\mathbf{k}})\pi(\hat{\boldsymbol{\beta}}_{\mathbf{k}})(2\pi)^{k/2}|G_{\hat{\boldsymbol{\beta}}_{\mathbf{k}}}|^{-1/2}. \tag{3.3}$$

Here, $\hat{\boldsymbol{\beta}}_{\mathbf{k}}$ is the maximum a posteriori (MAP) estimate of $\boldsymbol{\beta}_{\mathbf{k}}$, $G_{\hat{\boldsymbol{\beta}}_{\mathbf{k}}}$ is the Hessian of the negative of the log posterior function,

$$g(\boldsymbol{\beta}_{\mathbf{k}}) = -\log\big(\pi(\mathbf{y} \mid \boldsymbol{\beta}_{\mathbf{k}})\big) - \log\big(\pi(\boldsymbol{\beta}_{\mathbf{k}})\big), \tag{3.4}$$

computed at $\hat{\boldsymbol{\beta}}_{\mathbf{k}}$, and k is the size of model $\mathbf{k}$. Finding the MAP of $\boldsymbol{\beta}_{\mathbf{k}}$ is equivalent to finding the minimum of $g(\boldsymbol{\beta}_{\mathbf{k}})$.

The details of computing the gradient and Hessian matrix of $g(\boldsymbol{\beta}_{\mathbf{k}})$ are discussed in Section 1 of the Supplementary Material (Nikooienejad, Wang and Johnson (2020)). The gradient and Hessian matrix, described by equations (3) to (7) there, are used to find the MAP and to compute the Laplace approximation of the marginal probability of $\mathbf{y}$. We use the limited memory version of the Broyden–Fletcher–Goldfarb–Shanno optimization algorithm (L-BFGS) (Liu and Nocedal (1989)) to find the MAP. The initial value for the algorithm is $\hat{\boldsymbol{\beta}}_{\mathbf{k}}$, the MLE for the Cox proportional hazard model.

Having all the components of formula (2.3), it is possible to define a MCMC framework to sample from the posterior distribution on the model space. A birth-death scheme, similar to that used in Nikooienejad, Wang and Johnson (2016), could be used for this purpose. However, for computational reasons we use another stochastic algorithm to search the model space; this algorithm is described in the next section.

The highest posterior probability model (HPPM) is defined as the model having the highest posterior probability among all visited models. In practice, many models may be assigned probabilities that are close to the probability achieved by the HPPM. For this reason and for predictive purposes, it is useful to obtain the Median Probability Model (MPM) (Barbieri and Berger (2004)) which is the model containing covariates that have posterior inclusion probabilities of at least 0.5. According to Barbieri and Berger (2004), the posterior inclusion probability for covariate i is defined as

$$p_i = \sum_{\mathbf{k}:\boldsymbol{\gamma}_{\mathbf{k}i}=1} P(M_{\mathbf{k}} \mid \mathbf{y}). \tag{3.5}$$

That is, the sum of posterior probabilities of all models that have covariate i as one of their variables. In this expression, $\boldsymbol{\gamma}_{\mathbf{k}i}$ is a binary value determining the inclusion of the ith covariate in model $\mathbf{k}$.

3.2.1. *Stochastic search algorithm.* To increase the efficiency of exploring the model space, we use the S5 algorithm. S5 was proposed by Shin, Bhattacharya and Johnson (2018) for variable selection in linear regression problems, and we adapt it here for survival models. It is a stochastic search method that screens covariates at each step. The algorithm is scalable and its computational complexity is only linearly dependent on p (Shin, Bhattacharya and Johnson (2018)).

Screening is the essential part of the S5 algorithm. In linear regression, screening is based on the correlation between excluded covariates and the residuals of the regression using the current model (Fan and Lv (2008)). The concept of screening covariates for survival response data is proposed in Fan, Feng and Wu (2010) and is defined based on the marginal utility for each covariate.

To illustrate the screening technique, suppose that the current model is $\mathbf{k}$. Let $\mathbf{k}^c$ denote the complement of set $\mathbf{k}$ containing columns of the design matrix that are not in the current model, $\mathbf{k}$. The conditional utility of covariate $m \in \mathbf{k}^c$ represents the amount of information covariate m contributes to the survival outcome, given model $\mathbf{k}$, and is defined as

$$u_{m|\mathbf{k}} = \max_{\substack{\beta_m \\ m\in\mathbf{k}^c}} \boldsymbol{\delta}^T \Bigg[(\beta_m \mathbf{X}_{(m)} + \mathbf{X}_{\mathbf{k}}\boldsymbol{\beta}_{\mathbf{k}}) - \log\Bigg\{\sum_{j=i}^{n} \exp(\beta_m x_{jm} + \mathbf{X}_{\mathbf{k}_j}\boldsymbol{\beta}_{\mathbf{k}})\Bigg\}\Bigg]. \tag{3.6}$$

By comparing $u_{m|\mathbf{k}}$ to the Cox log-likelihood equation (Formula (1) in the Supplementary Material (Nikooienejad, Wang and Johnson (2020))), it follows heuristically that the conditional utility is the maximum likelihood for covariate m after accounting for the information provided by model $\mathbf{k}$. Finding $u_{m|\mathbf{k}}$ is a univariate optimization procedure that can be computed rapidly.

With this background the S5 algorithm for survival data works as follows. At each step the d covariates with highest conditional utility are candidates to be added to the current model $\mathbf{k}$ and comprise the addition set, Γ^+. The deletion set, Γ^-, contains the current model, except that one variable is removed. From the current model, $\mathbf{k}$, we consider moves to each of its neighbors in Γ^+ and Γ^- with a probability proportional to the marginal probabilities of these neighboring models.

To avoid local maxima, the model probabilities used in S5 are raised to the power of $1/t_l$, where t_l is the l^{th} temperature in an annealing schedule in which "temperatures" decrease. To increase the number of visited models, a specified number of iterations are performed at each temperature. At the end of the procedure, the model with the highest posterior probability of visited models is identified as the HPPM.

In our version of the S5 algorithm, we used 10 equally spaced temperatures varying from 3 to 1 and 30 iterations within each temperature. Section 4 of the Supplementary Material (Nikooienejad, Wang and Johnson (2020)) provides some discussion on how these values are chosen for this application. To increase the number of visited models, we parallelized the S5 procedure so that it could be distributed to multiple CPUs. Each CPU executes the S5 algorithm independently with a different starting model. All visited models are pooled together at the end, and the HPPM and MPM are determined. Using posterior probabilities of the visited models, the posterior inclusion probability for each covariate can be computed using (3.5). In our simulations we used 120 CPUs to explore the model space for design matrices with $O(10^4)$ covariates.

3.3. *Predictive accuracy assessment.* In addition to looking at the selected genes and their pathways to determine their biological relevance in analyzing the real datasets, we used the time dependent AUC, obtained from time dependent ROC curves as introduced by Heagerty, Lumley and Pepe (2000), for survival times to summarize and compare the predictive performance of the various algorithms. This measure has a relatively straightforward interpretation and, unlike other summary measures such as the c-index (Harrell et al. (1982)), can be computed without requiring specific conditions or additional assumptions to hold (Blanche, Kattan and Gerds (2019)). However, predictive performance measures, including the c-index, Integrated Brier Score (IBS) (Gerds and Schumacher (2006)) and prediction error curves, are investigated and reported in Sections 4 and 5 for both simulation and real datasets.

There are different methods to estimate time dependent sensitivity and specificity. In our algorithm we adapted a method proposed by Uno et al. (2007), henceforth called Uno's method. In that method, after splitting data into training and test sets, sensitivity is estimated by

$$\widehat{\mathrm{SE}}_{\mathbf{k}}(t,c)=\frac{\sum_{i=1}^{n}\delta_i I(\mathbf{X}_{\mathbf{k}_i}^T\hat{\boldsymbol{\beta}}_{\mathbf{k}}>c,T_i\le t)/\hat{G}(T_i)}{\sum_{i=1}^{n}\delta_i I(T_i\le t)/\hat{G}(T_i)}, \tag{3.7}$$

and specificity is estimated by

$$\widehat{\mathrm{SP}}_{\mathbf{k}}(t,c)=\frac{\sum_{i=1}^{n} I(\mathbf{X}_{\mathbf{k}_i}^T\hat{\boldsymbol{\beta}}_{\mathbf{k}}\le c,T_i>t)}{\sum_{i=1}^{n} I(T_i>t)}. \tag{3.8}$$

These values are estimated for the test set. Therefore, in the equations above, n is the number of observations in the test set, δ_i is the status of observation i and T_i is the observed time for that observation in the test set. The variable c is the discrimination threshold that is varied to obtain the ROC curve. The function $\hat{G}$ is the Kaplan–Meier estimate of the survival function obtained from the training set. For each observation i in the test set with observed time T_i, $\hat{G}(T_i)$ is computed by a basic interpolation procedure. That is,

$$\hat{G}(T_i)=\hat{G}\big(T_{\mathrm{tr}}^*\big)\quad\text{where } T_{\mathrm{tr}}^*=\underset{T\in T_{\mathrm{tr}}}{\arg\min}\,|T-T_i|. \tag{3.9}$$

Here, T_{tr} is the set of all observed survival times in the training set. In (3.7) and (3.8), $\hat{\boldsymbol{\beta}}$ represents the estimated coefficient under a specific model.

3.3.1. *Bayesian model averaging* (*BMA*). BMA can be used to improve the predictive accuracy by accounting for the uncertainty in selected models. From (3.7) and (3.8) the final sensitivity and specificity using BMA may be defined as

$$\widehat{\mathrm{SE}}_{\mathrm{BMA}}(t,c)=\sum_{j=1}^{\mathcal{N}}\widehat{\mathrm{SE}}_{\mathbf{k}^j}(t,c)P(\mathcal{M}_{\mathbf{k}^j}\mid\mathbf{y}_n) \tag{3.10}$$

and

$$\widehat{\mathrm{SP}}_{\mathrm{BMA}}(t,c)=\sum_{j=1}^{\mathcal{N}}\widehat{\mathrm{SP}}_{\mathbf{k}^j}(t,c)P(\mathcal{M}_{\mathbf{k}^j}\mid\mathbf{y}_n), \tag{3.11}$$

where, $P(\mathcal{M}_{\mathbf{k}^i}\mid\mathbf{y}_n)$ is the posterior probability of model $\mathcal{M}_{\mathbf{k}^i}$ The value of $\mathcal{N}$ depends on what type of BMA is used. We use Occam's window, which means only models that have posterior probability of at least $w\times p(\mathcal{M}_{\mathrm{HPPM}}\mid\mathbf{y}_n)$ are used in model averaging. We set $w=0.01$ for our applications.

In the proposed method individual survival curves are estimated using the highest posterior probability model. Section 2 of the Supplementary Material (Nikooienejad, Wang and Johnson (2020)) provides the details of this procedure. Similar approaches were also adopted by Held, Gravestock and Sabanés Bové (2016) in estimating the survival curve for each individual in a study.

4. Simulation results. To investigate the performance of the proposed model selection procedure, we applied our method to simulated datasets. We followed the guidance of Morris, White and Crowther (2019) as a basis for our simulation protocol. In particular, the simulation design was based on the ADEMP structure (Aims, Data generating mechanism, Estimands, Methods, and Performance measures) discussed in that article. We refer to each of those elements as we explain different parts of the simulation design below.

Regarding *"Methods,"* we compared the performance of our algorithm to ISIS-SCAD (Fan, Feng and Wu (2010)) and GLMNET (Friedman, Hastie and Tibshirani (2010)), two of the most highly used algorithms for high-dimensional variable selection for survival data. We used the published R packages of those two methods to run the simulations. We also performed a comparison with a case when pMOM priors are used as the prior for nonzero coefficients instead of piMOM priors.

The *"Aim"* of the simulation study was to compare the performance of our method with the other two methods with respect to the correlation structure between covariates in the design matrix. More specifically, we reported three different simulation settings that consider different combinations of correlation structure, true model size and the magnitude of true coefficients. This was the basis of our *"Data generating mechanism."* The correlation structure used in those settings are similar to the simulations reported in Fan, Feng and Wu (2010).

For Case 1, $X_1, \ldots, X_p$ were multivariate Gaussian random variables with mean 0 and marginal variance of 1. The correlation structure was $\text{corr}(X_i, X_5) = 0$ for all $i \neq 4, 5$, $\text{corr}(X_5, X_4) = 1/\sqrt{2}$ and $\text{corr}(X_i, X_j) = 0.5$ for $i, j \in \{1, \ldots, p\} \setminus \{4, 5\}$. The size of the true model was 5 with nonzero regression coefficients $\beta_1 = -1.5389$, $\beta_2 = 0.6839$, $\beta_3 = -0.8498$, $\beta_4 = -1.2716$, $\beta_5 = -1.1045$ and $\beta_i = 0$ for $i > 5$. The number of observations and covariates were $n = 400$ and $p = 1000$. The censoring rate for this case was approximately 27.6%. The survival and censoring times are both sampled from an exponential distribution. The rate parameter for the distribution of censoring times was set to 0.1.

For Case 2, $X_1, \ldots, X_p$ were multivariate Gaussian random variables with mean 0 and marginal variance of 1. The correlation structure between variables was $\text{corr}(X_i, X_j) = 0.5$; $i \neq j$. The size of the true model was 6 with nonzero regression coefficients $\beta_1 = 1.1201$, $\beta_2 = 0.8322$, $\beta_3 = -1.9620$, $\beta_4 = -1.7639$, $\beta_5 = 1.6782$, $\beta_6 = 1.8995$, and $\beta_i = 0$ for $i > 6$. The number of observations and covariates were $n = 400$ and $p = 1000$. In this case the survival times were sampled from a Weibull distribution with rate parameter $\lambda = 0.1$ and shape parameter $k = 15$. The censoring times were sampled uniformly from $[0, 8]$, and the resulting censoring rate for this case was approximately 14.8%.

For Case 3 the design matrix and correlation structure between variables was the same as Case 2, where $\text{corr}(X_i, X_j) = 0.5$, $i \neq j$. The size of the true model was 20 with nonzero regression coefficients $(\beta_1, \ldots, \beta_{20})$ equal to $(-1.6802, -1.2483, 2.9430, -2.6458, -2.5173, -2.8493, -2.0070, -1.5931, 0.8800, -0.9387, 1.6599, -2.9288, -1.2495, -2.6298, -2.3434, 1.9075, -1.1044, -0.7873, 2.6722, -0.6340)$, and $\beta_i = 0$ for $i > 20$. The number of observations and covariates were $n = 400$ and $p = 1000$. The censoring rate for this case was approximately 34.1%. The survival and censoring times were both sampled from an exponential distribution. The rate parameter for the distribution of censoring times was set to 0.1.

Each simulation case was then repeated 50 times, $n_{\text{iter}} = 50$, and each time with different random seed numbers in order to generate different datasets.

The primary targets of our simulation study, or the *"Estimands,"* according to Morris, White and Crowther (2019), were identifying the true model as well as estimating the vector of coefficients of the true model. Accordingly, we reported four different quantities as *"Performance measures"* for those estimands. The first two quantities are the mean l_1 norm of the error in estimating the vector of coefficients and the mean squared error (MSE). The mean l_1 norm was computed as $\frac{1}{n_{\text{iter}}}\sum_{i=1}^{p}|\hat{\beta}_i - \beta_i|$, and the MSE was computed as $\frac{1}{n_{\text{iter}}}\sum_{i=1}^{p}(\hat{\beta}_i - \beta_i)^2$. The third quantity is the mean model size of the selected models and was denoted by MMS. MTP and MFP denote mean false positive and mean true positive values for each algorithm. Formal definitions of MFP, MTP are provided in Section 3 of the Supplementary Material Nikooienejad, Wang and Johnson (2020).

Table 1 compares the performance of our method, BVSNLP, the default settings of ISIS-SCAD and GLMNET algorithms. The λ parameter in GLMNET was picked by cross validation. Table 2 compares the Monte Carlo standard errors (Morris, White and Crowther (2019)) of the MSEs for all three methods.

In the S5 algorithm 30 iterations were used within each temperature. The parameter d was chosen as $2\lceil\log(p)\rceil$. As described in Section 3.2.1, d represents the number of candidate covariates that were added to the current model to make the addition set, Γ^{+}. Each S5 algo-

TABLE 1
Comparison between BVSNLP, ISIS-SCAD and GLMNET for simulation Cases 1, 2 *and* 3 *with* $n = 400$ *and* $p = 1000$

	BVSNLP	ISIS-SCAD	GLMNET
Case 1:			
MSE	0.141	0.792	1.441
Mean l_1 norm	0.488	2.200	4.072
MMS	4.96	8.84	51.46
MTP	4.92	4.62	4.00
MFP	0.04	4.22	47.46
Case 2:			
MSE	0.141	0.792	1.441
Mean l_1 norm	0.505	0.552	3.891
MMS	6	5.94	50.94
MTP	6	5.88	5.92
MFP	0	0.06	45.02
Case 3:			
MSE	0.602	5.287	4.701
Mean l_1 norm	2.680	22.962	22.824
MMS	20.08	14.76	105.62
MTP	19.94	12.80	19.96
MFP	0.14	1.96	85.66

TABLE 2
Monte Carlo Standard Errors for the MSE of the coefficient vector for all three methods

	Case 1	Case 2	Case 3
BVSNLP	0.064	0.008	0.065
ISIS-SCAD	0.098	0.015	0.142
GLMNET	0.007	0.024	0.055

rithm was run in parallel on 120 CPUs for all three simulation cases. The beta-binomial prior was imposed on the model space with $a = 1$, $b = p - a$. The hyperparameters of the piMOM prior were selected using the algorithm discussed in Section 3.1 with $\alpha = 0.8$ for all three simulation cases, imposed as the prior mode.

Finally, the average runtimes of the BVSNLP algorithm for the three simulation cases were 29, 20.23 and 27.99 seconds, respectively.

As demonstrated in Table 1, our method performs better than the other two methods according to all selected metrics, regardless of the size of the true model. The difference between BVSNLP and ISIS-SCAD is best illustrated as the size of the true model increases. GLMNET has significantly higher mean false positive rates than the other two methods.

Figures 2, 3(a) and 3(b) compare the average IBS over 50 iterations between the methods discussed above. IBS is computed using the R package `pec` (Mogensen, Ishwaran and Gerds (2012)) based on a five-fold cross-validation. A benchmark model based on Kaplan–Meier estimate, which includes no covariates, is also added to the figures as a reference for the comparison. The average c-index measures for all the methods are also reported in Table 3. The c-index measures are computed based on the method discussed in van Houwelingen and Putter (2011), using the `dynpred` package in R. Because a new dataset was created at each iteration, it was not possible to get the average prediction errors, due to the fact that the times points where prediction errors change were different for different data sets.

As shown in the IBS plots, all three methods perform better than the reference. BVSNLP and ISIS-SCAD have a very similar performance. For Case 3, where the true model has 20 covariates, BVSNLP outperforms the other two methods, whereas in Case 2, GLMNET has the best performance. The c-index is similar for all methods and seems to provide a smaller penalty for model size. This feature of the c-index is discussed further in Section 6.

5. Application to real data. We applied our method to selected genes associated with patient survival times for two common cancer types using datasets from The Cancer Genome Atlas (TCGA) projects: kidney renal clear cell carcinoma (KIRC) (Cancer Genome Atlas Research Network (2013)) and kidney renal papillary cell carcinoma (KIRP) (Cancer Genome Atlas Research Network (2016)). We compared the performance of our algorithm to ISIS-SCAD (Fan, Feng and Wu (2010)), GLMNET (Friedman, Hastie and Tibshirani (2010)) and Stability Selection (Meinshausen and Bühlmann (2010)). Stability Selection was combined

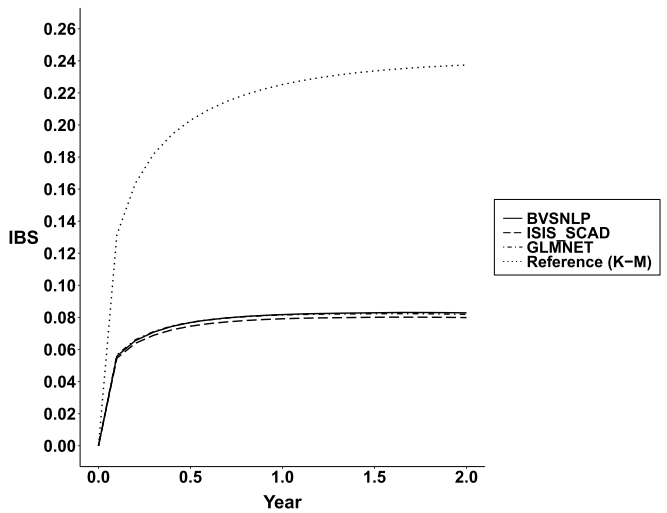
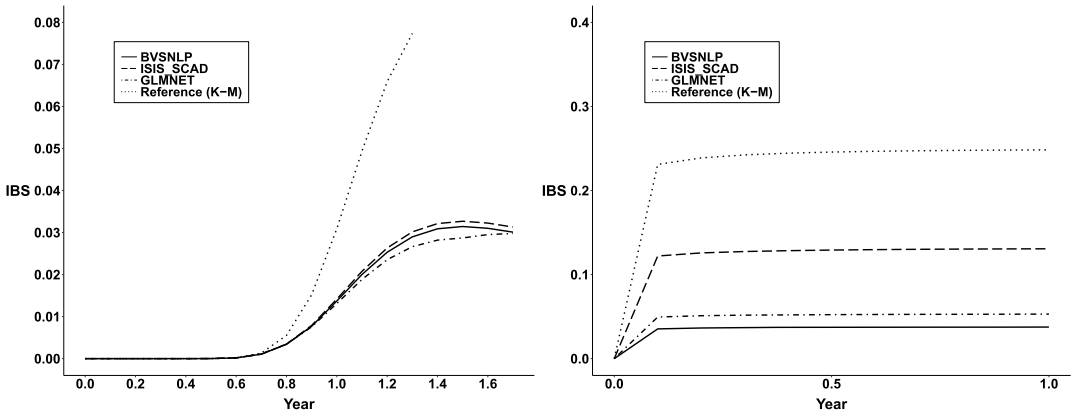


FIG. 2. *Average IBS for all methods in simulation Case* 1.



(a) Simulation case 2.

(b) Simulation case 3.

FIG. 3. *IBS plots for simulation Cases* 2 *and* 3.

with a high-dimensional selection algorithm, such as GLMNET, and selects the most stable features for a given level of Type I error. To run the Stability Selection method, we used the `c060` R package (Sill et al. (2014)) and the recommended values for function arguments.

We included patients' "Age," "Gender" and a clinical stage variable, "Stage," in the design matrix. On the advice of a clinician, the "Stage" variable was developed by combining the histological stage, pathological stage and clinical stage into one variable that is a summary of how advanced each subject's cancer was when the tissue sample was taken."Stage," like "Gender," is a categorical variable but with three levels, where "Stage i" represents the ith class of that variable; "Stage 3" represents the most advanced stage. To remove stromal contaminations from the gene expression data, the DeMixT algorithm (Wang et al. (2017)) was performed on the design matrix, and the tumor-specific expression data were used in the analyses for all algorithms.

Predictive performance was measured by a time-dependent AUC, as discussed in Section 3.3, based on a five-fold cross-validation. The observations in each fold were randomly chosen under a constraint which balanced censoring rate between folds. The AUC values were computed for the test set using the model that was obtained by performing variable selection on the training set. The selected covariates for each cancer type were also compared. For our method we report the covariates associated with the highest posterior probability model. The hyperparameter τ of the piMOM prior was selected using the algorithm in Section 3.1, with $\alpha = 0.1$ as the mode of the piMOM prior. This is our choice of α for real datasets. The results for each cancer type are discussed in separate sections below. Note that GLMNET has a random output when the hyperparameter is selected by cross-validation. As a result, based on the recommendation of the inventors of that algorithm, we ran it 100 times for each fold and took the average of results as the outcome for that fold.

TABLE 3
Average c-index measures over 50 *iterations in each simulation case*

	Case 1	Case 2	Case 3
BVSNLP	0.890	0.881	0.960
ISIS-SCAD	0.895	0.876	0.841
GLMNET	0.911	0.908	0.970

We treated categorical variables "Stage" and "Gender," as well as the continuous variable "Age," as fixed covariates in our model. However, available ISIS-SCAD and Stability Selection software packages are not able to fix preselected covariates to be included in all models. For this reason, dummy variables associated with "Stage" and "Gender" were manually added to the design matrix and were subject to the selection procedure for those methods.

5.1. *Kidney renal clear cell carcinoma* (*KIRC*). After removing covariates with missing expressions and observations with missing survival times, the KIRC dataset (Cancer Genome Atlas Research Network (2013)) contains 490 observations with 13,267 covariates. The censoring rate for this dataset is 66.94%. Table 4 shows the covariates selected by each method. As mentioned previously, GLMNET produces random outputs at each run, and therefore, for this table, only the output for one of the runs are indicated; other runs produced a similar number of selected covariates.

In addition to the categorical covariate "Stage," BVSNLP selects "AR" and "SUDS3" in the HPPM as the most significant covariates in the design matrix. The posterior inclusion probabilities for "AR" and "SUDS3" are 0.80 and 0.08, respectively. The "Age, Gender" and 'Stage' were fixed in all models and, thus, were selected with probability 1. The MAP estimates for the coefficients of "Age, Gender Male, Stage 2, Stage 3, AR" and "SUDS3" were 0.33, -0.11, 0.45, 1.61, -0.60 and 0.36, respectively. These coefficients indicate that patients with the most advanced stages of cancer had the poorest survival rates, and that a patient with a tumor sample characterized as advanced has a hazard rate that was $\exp(1.61) \approx 5$ times higher than a patient with tumor sample characterized as localized, when all other covariates were the same. These coefficients also show that the hazard rate in females is 1.12 times that in males, and age has an unfavorable impact on the hazard rate, as expected. Moreover, the negative sign for the "AR" gene indicates it has a favorable impact on survival for KIRC. "AR," the Androgen Receptor gene, functions as a steroid-hormone activated transcription factor. It has been well documented that "AR" promotes the progression of renal cell carcinoma (RCC) through hypoxia-inducible factors HIF-2α and vascular endothelial growth factor regulation (Fenner (2016)). The favorable impact of the "AR" gene was also studied

TABLE 4
Selected genes and covariates for KIRC across different variable selection algorithms

BVSNLP	Age	Gender	Stage
	SUDS3	AR	
ISIS-SCAD	Stage 3	AR	Age
	HEBP1	ATP2C1	GADD45A
	MTERF2	ADGRL3	GPSM1
	SERPINI1	SP6	ZNF815P
	INAFM2		
GLMNET	Stage	AR	Age
	HEBP1	SEC61A2	TRMT6
	PCBP4	FAHD2A	MCM8
	E2F5	SLC5A6	NARF
	RAB28	DONSON	GPSM1
	HACD1	MARS	FASN
	TRAIP	RPL17P50	SLC26A6
	GPR162	INAFM2	ACACA
Stability selection	Stage 3	AR	Age
	INAFM2		

by Hata et al. (2017) in bladder cancer. "SUDS3" is a regulatory protein that is part of the SIN3A corepressor complex component that potentially has a role in tumor suppressor pathways through regulation of apoptosis. There was previous evidence of the down-regulation of the SIN3A gene in tumorigenesis of lung cancer (Suzuki et al. (2008)).

It is noteworthy that the algorithm selected the same highest posterior probability model for different values of the hyperparameter τ in the range $[0.01, 0.9]$. This shows the robustness of the proposed variable selection algorithm to the choice of hyperparameter τ for a range of plausible values.

For this particular run of GLMNET, a much larger model was selected with 24 variables including two of the variables reported by BVSNLP. ISIS-SCAD selected 13 covariates, which included the four covariates that were selected by the Stability Selection method. "AR" and the last level of "Stage" are the common covariates among all methods.

The time dependent AUC plot for all four methods, obtained by performing a five-fold cross validation, is depicted in Figure 4. BVSNLP has slightly better predictive accuracy than GLMNET and Stability Selection. However, it achieves this accuracy with a much sparser model. We investigated the covariates that were selected by each of those algorithms in all five folds and found that BVSNLP, in addition to those fixed covariates, selects only 10 unique genes in total, where 'AR' is selected in three of the five folds.

GLMNET selected 160 different covariates across all five folds. Only five out of 24 selected covariates in Table 4 were selected in all five training datasets in cross-validation. Those include "Age, Stage" and "AR." GLMNET was run 100 times for each fold. ISIS-SCAD selected 45 different covariates, and only 'Stage 3' was selected in all training datasets in cross-validation. The Stability Selection method selected sparser models compared to ISIS-SCAD and GLMNET by selecting 13 different covariates. It picked only "Age" and "Stage 3" in all five folds.

Figures 5(a) and 5(b) compare IBS and prediction error curves, respectively, between different methods for the KIRC dataset. These two measures were computed based on a five-fold cross-validation. Computation of IBS and prediction error were done using the R package `pec` (Mogensen, Ishwaran and Gerds (2012)). A benchmark model based on the Kaplan–Meier estimate, which includes no covariates, was also added to the figures as a reference for the comparison. The c-index measures are also reported in Table 5. The c-index was computed as it was in Section 4 using the `dynpred` package in R.

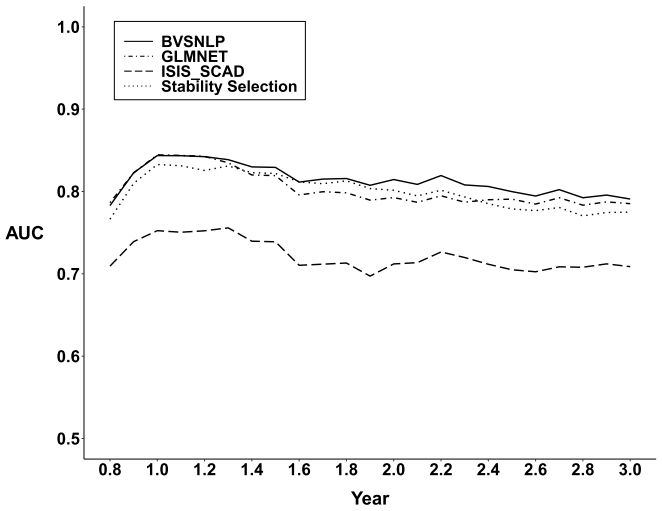
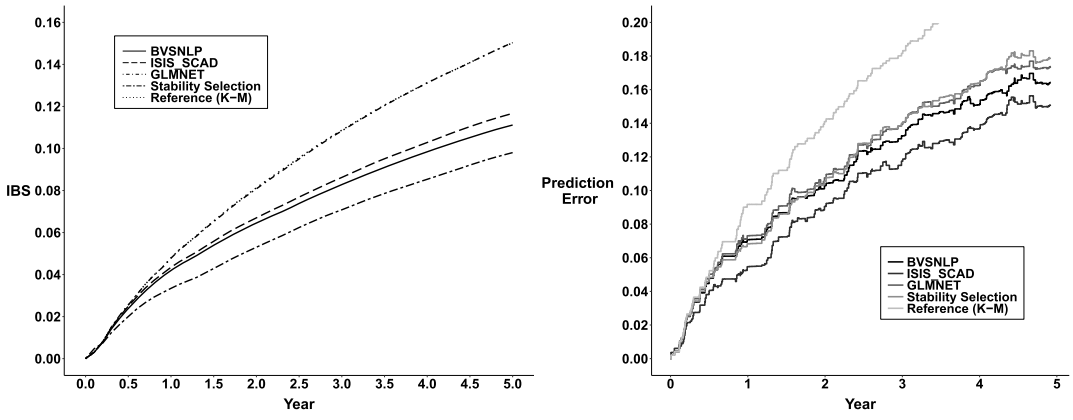


FIG. 4. *Average AUC of different variable selection methods based on a five fold cross-validation for KIRC dataset.*



(a) IBS (b) Prediction Error

FIG. 5. *IBS and prediction error of BVSNLP for the KIRC dataset.*

GLMNET has almost the same IBS curve as the reference Kaplan–Meier curve. BVSNLP outperforms ISIS-SCAD, and Stability Selection has the best IBS performance among all. For prediction error curves, BVSNLP is second to ISIS-SCAD, and GLMNET and Stability Selection have almost the same performance. A different behavior can be seen for c-index measures where GLMNET and ISIS-SCAD have higher c-indices than BVSNLP.

BVSNLP was run on 120 CPUs, Stability Selection was run on four CPUs, while GLMNET and ISIS-SCAD were run on a single CPU. The average run-time for different methods in each fold of the cross validation was 6.4, 180, five and 1.3 minutes for BVSNLP, GLMNET, ISIS-SCAD and Stability Selection.

In our previous study of binary outcomes using the same dataset (Nikooienejad, Wang and Johnson (2016)), we performed hierarchical clustering on the deconvolved tumor-specific expression matrix and identified two clusters of patient samples. We saw these two groups of patients present significantly different survival outcomes and, therefore, assigned good vs. bad survival to the groups. The dichotomization was based solely on the clustering results of deconvolved gene expression levels. Survival times and censoring did not play any role in that process. However, there was a loss of information in dichotomizing a survival dataset and analyzing it with logistic regression. Now, with BVSNLP we are able to use the original survival time to event with censoring information. To compare the biological implications between the two analyses, we looked for known expression regulation networks between the gene sets found in the binary analysis, SAV1 and NUMBL, and the new genes found in this analysis, AR and SUDS3, using Pathway Studio® (Nikitin et al. (2003), Elsevier (2018)). We found that the well-studied cancer genes TGFB1, BCL2, PPARG, NEDD4 and CTNNB1, and a regulatory microRNA, MIR21, constitute the shortest paths between SAV1 and AR. Similarly, we found cancer genes CDKN1A, WNT3A, two genes that determine cell fate (SOX17 (connected with CTNNB1) and NANOG) and PAX6 that regulates transcription, to constitute the shortest paths between NUMBL and SUDS3. These are depicted in Section 6 of the Supplementary Material (Nikooienejad, Wang and Johnson (2020)). These findings

TABLE 5
Average c-index measure of different methods for the KIRC dataset

	BVSNLP	GLMNET	ISIS-SCAD	Stability selection
c-index measure	0.804	0.816	0.846	0.797

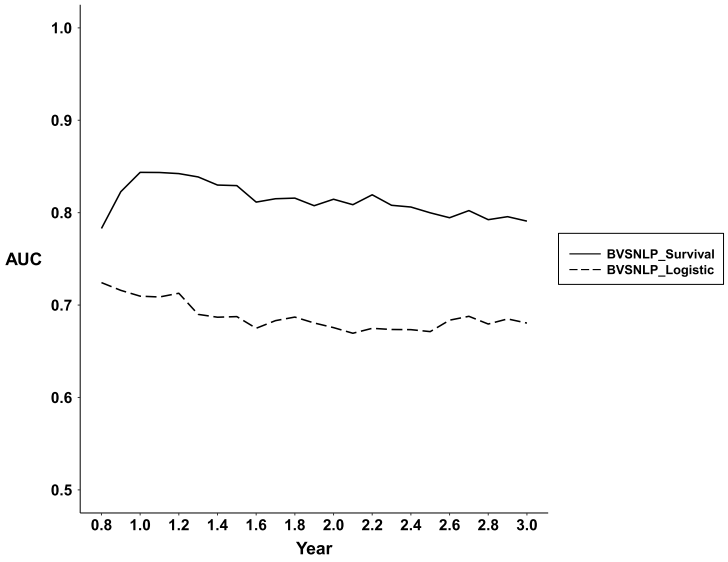


FIG. 6. *Comparison between BVSNLP model selection using survival and dichotomized versions of the KIRC dataset.*

suggest a high biological consistency between our two analyses that used BVSNLP to select features for binary and survival outcomes.

In summary, the binary model using SAV1 and NUMBL to predict overall survival of patients with kidney cancer is not as effective as the model using AR and SUDS3, as shown in Figure 6. Thus, although the findings of Nikooienejad, Wang and Johnson (2016) were biologically justified, some limitations were associated with those findings due to the information loss incurred by clustering and dichotomizing the data, and the BVSNLP survival model provides better insight on the genes associated with this cancer type.

5.2. *Kidney renal papillary cell carcinoma* (*KIRP*). The KIRP dataset (Cancer Genome Atlas Research Network (2016)) contains 244 samples with 13,335 covariates (after necessary data cleaning) and has a fairly high censoring rate of 85.7%. The covariates selected by each method are summarized in Table 6.

In addition to the fixed covariates "Age, Gender" and "Stage," BVSNLP selects the "CDK1" gene in the HPPM as the most significant covariate in the design matrix. The posterior inclusion probability for "CDK1" was 0.12. The MAP estimates for the coefficients of "Age, Gender Male, Stage 2, Stage 3" and "CDK1" were 0.12, −0.10, 0.11, 0.79 and 1.13, respectively. This shows that a unit increase in "CDK1" (Cyclin dependent kinase 1) gene expression increases the hazard rate by a factor of three for given values of the other covariates. CDK1 is a cell cycle regulator and has been reported previously as a prognostic marker

TABLE 6
Selected covariates for KIRP across different variable selection algorithms

BVSNLP	Age CDK1	Gender	Stage
ISIS-SCAD	CDK1	COL6A1	C19orf33
GLMNET	*No covariates were selected*		
Stability selection	Stage 3	MTC02P12	RPL39P3

gene for various cancer types. Many experimental studies have been performed to further understand the molecular mechanism behind the complex functions of CDK1 (Malumbres and Barbacid (2009)). This is the first time, however, that CDK1 has been reported as a prognostic marker gene in human data for papillary renal cell carcinoma. As expected, patients at the most advanced stage cancer have a hazard rate that is 2.2 times higher than patients at a localized stage of cancer, given the values of all other covariates. As in the case of KIRC patients, age and male gender have unfavorable and favorable impacts on the hazard rate, respectively.

Surprisingly, GLMNET does not select any covariates. Stability Selection picked three covariates, with only "Stage 3" in common with BVSNLP. As in the previous dataset, we tested BVSNLP for different choices of τ in the interval $[0.01, 0.9]$, and the same model was selected for all values within this range. The total runtime of BVSNLP for this dataset was around five minutes using 120 CPUs.

Figure 7 shows the predictive accuracy for the proposed method based on a five-fold cross-validation. The outcomes for GLMNET, ISIS-SCAD and Stability Selection are not displayed in the plot because those methods did not converge or failed to produce results for at least one of the five folds in the cross-validation experiment. The small AUC values in this plot for $t < 1$ warrant comment. Because there were few events soon after entry of tissue samples into the TCGA database, the AUC for early timepoints falls close to the 50% benchmark reflecting no predictive value.

Figures 8(a) and 8(b), respectively, depict IBS and prediction error curves of the BVSNLP method, based on a five-fold cross validation for the KIRP dataset, and compares it to the reference curve obtained by the Kaplan–Meier method. The c-index measure for the BVSNLP method is 0.876. The average runtime for BVSNLP in each fold of the cross-validation was 3.6 minutes on 120 CPUs.

6. Discussion. In this article a Bayesian variable selection method, BVSNLP, was proposed for selecting variables in high and ultrahigh dimensional datasets with survival time as outcomes. BVSNLP uses an inverse moment nonlocal prior density on nonzero regression coefficients. Analyses of simulated and real data suggest that BVSNLP performs comparably or better than other existing methods for variable selection for survival data. Moreover, the real data results indicated that the proposed algorithm is robust to the choice of the hyperparameter τ in the piMOM prior for values of τ in the range $[0.01, 0.9]$.

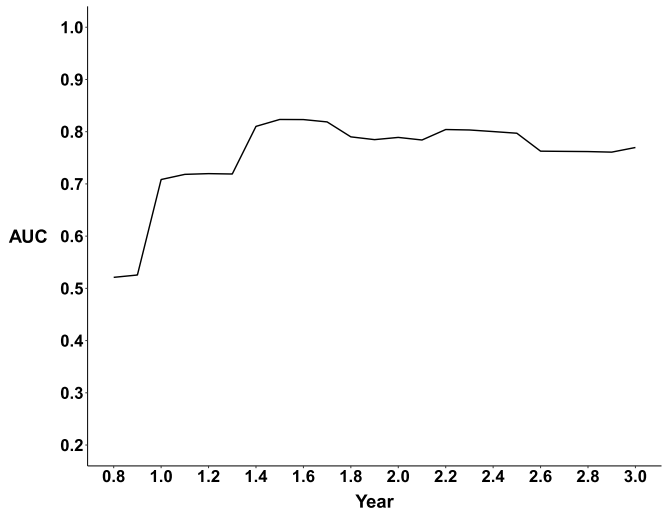
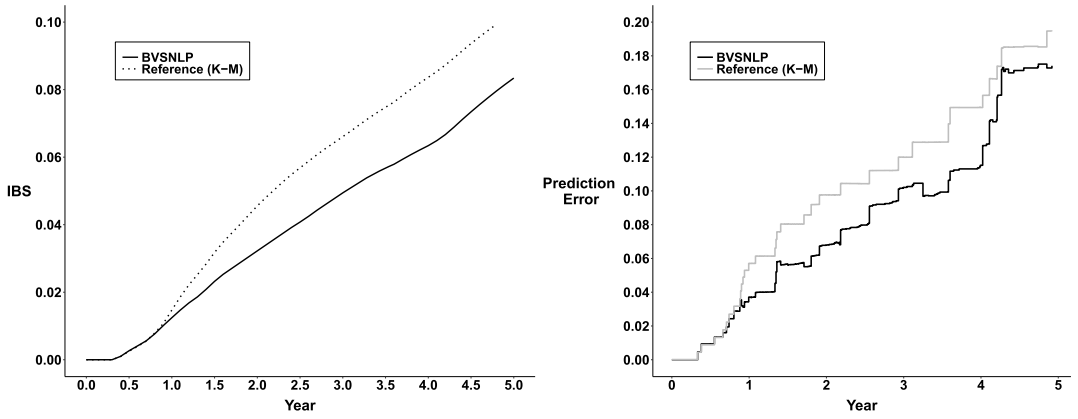


FIG. 7. *Average AUC of BVSNLP based on a five fold cross-validation for the KIRP dataset.*



(a) IBS (b) Prediction Error

FIG. 8. *IBS and prediction error of BVSNLP for the KIRP dataset.*

Various outputs are provided by the algorithm. These include the HPPM, MPM and the posterior inclusion probability for each covariate in the model. For real datasets, Bayesian model averaging is used to incorporate uncertainty in selected models when computing time dependent AUC plots using Uno's method (Uno et al. (2007)). Finally, an R package named BVSNLP has been implemented to make the algorithm freely available and adaptable to interested researchers. The package can be run in parallel fashion where hundreds of CPUs can be exploited in order to increase the number of visited models in the search for the highest posterior probability model. The BVSNLP package is available in the R repository, CRAN, at https://CRAN.R-project.org/package=BVSNLP. The user manual for the package is also available from this site.

Two real cancer genomic datasets from the TCGA website were considered in this article. Compared to other methods, BVSNLP found sparser models with biologically relevant genes. The proposed method showed a reliable predictive accuracy as measured by AUC using substantially fewer variables.

We have based our assessments on time dependent AUC and biological interpretation of the results, but other measures, like IBS, prediction error and the concordance index (also known as the c-index or Harrell's c-index), were also reported. Difficulties associated with such measures are identified in Blanche, Kattan and Gerds (2019). In particular, the authors of that article demonstrate that the concordance index can favor misspecified models over the correctly specified model because it is based on the order of event times rather than the event status at the prediction horizon. This may explain the slightly higher c-index values for GLMNET in both simulation and real datasets. The time dependent AUC does not suffer from this deficiency. Of course, different evaluation criteria can be expected to result in different rankings of models, and criteria that emphasize prediction error over low false positive rates can be expected to favor larger models. Similarly, criteria that place a higher premium on eliminating false positives will tend to select smaller models.

Acknowledgments. The authors would like to thank the editor and anonymous referees for their numerous comments that improved presentation of the material in this article. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing (HPRC). The first author was supported by NIH grant R01CA158113 prior to joining Eli Lilly and Company. The second author was supported by 1R01CA174206, 1R01CA183793, 5R01CA158113 and P30CA016672. The third author was supported by NIH grant R01CA158113.

SUPPLEMENTARY MATERIAL

Supplement to "Bayesian variable selection for survival data using inverse moment priors" (DOI: 10.1214/20-AOAS1325SUPP; .pdf). The supplementary material to this article contains details regarding the methodology, algorithm for implementing that methodology, and extended discussion of the results.

REFERENCES

ANTONIADIS, A., FRYZLEWICZ, P. and LETUÉ, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.* **37** 531–552. https://doi.org/10.1111/j.1467-9469.2009.00685.x

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. https://doi.org/10.1214/009053604000000238

BENDER, R., AUGUSTIN, T. and BLETTNER, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Stat. Med.* **24** 1713–1723. https://doi.org/10.1002/sim.2059

BERGER, J. O., LISEO, B. and WOLPERT, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.* **14** 1–28. https://doi.org/10.1214/ss/1009211803

BLANCHE, P., KATTAN, M. W. and GERDS, T. A. (2019). The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics* **20** 347–357. https://doi.org/10.1093/biostatistics/kxy006

CANCER GENOME ATLAS RESEARCH NETWORK (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499** 43–49. https://doi.org/10.1038/nature12222

CANCER GENOME ATLAS RESEARCH NETWORK (2016). Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* **374** 135–145. https://doi.org/10.1056/nejmoa1505917

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data* **21**. CRC Press, Boca Raton, FL.

ELSEVIER (2018). PathwayStudio® pathwaystudio.com. Elsevier.

FAN, J., FENG, Y. and WU, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength*: *Theory Powering Applications—a Festschrift for Lawrence D. Brown. Inst. Math. Stat.* (*IMS*) *Collect.* **6** 70–86. IMS, Beachwood, OH. https://doi.org/10.1214/10-imscoll606

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. https://doi.org/10.1214/aos/1015362185

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

FARAGGI, D. and SIMON, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics* **54** 1475–1485. https://doi.org/10.2307/2533672

FENNER, A. (2016). Kidney cancer: AR promotes RCC via lncRNA interaction. *Nat. Rev. Urol.* **13** 242. https://doi.org/10.1038/nrurol.2016.61

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22. https://doi.org/10.18637/jss.v033.i01

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.

GERDS, T. A. and SCHUMACHER, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom. J.* **48** 1029–1040. https://doi.org/10.1002/bimj.200610301

GHOSH, J. K. (1988). *Statistical Information and Likelihood*: *A Collection of Critical Essays by Dr. D. Basu. Lect. Notes in Statist* **45**.

HARRELL, F. E. JR., CALIFF, R. M., PRYOR, D. B., LEE, K. L., ROSATI, R. A. et al. (1982). Evaluating the yield of medical tests. *JAMA* **247** 2543–2546. https://doi.org/10.1001/jama.247.18.2543

HATA, S., ISE, K., AZMAHANI, A., KONOSU-FUKAYA, S., MCNAMARA, K. M., FUJISHIMA, F., SHIMADA, K., MITSUZUKA, K., ARAI, Y. et al. (2017). Expression of AR, 5αR1 and 5αR2 in bladder urothelial carcinoma and relationship to clinicopathological factors. *Life Sci.* **190** 15–20. https://doi.org/10.1016/j.lfs.2017.09.029

HEAGERTY, P. J., LUMLEY, T. and PEPE, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56** 337–344. https://doi.org/10.1111/j.0006-341x.2000.00337.x

HELD, L., GRAVESTOCK, I. and SABANÉS BOVÉ, D. (2016). Objective Bayesian model selection for Cox regression. *Stat. Med.* **35** 5376–5390. https://doi.org/10.1002/sim.7089

IBRAHIM, J. G., CHEN, M.-H. and MACEACHERN, S. N. (1999). Bayesian variable selection for proportional hazards models. *Canad. J. Statist.* **27** 701–717. https://doi.org/10.2307/3316126

JOHNSON, V. E. (2005). Bayes factors based on test statistics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 689–701. https://doi.org/10.1111/j.1467-9868.2005.00521.x

JOHNSON, V. E. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 143–170. https://doi.org/10.1111/j.1467-9868.2009.00730.x

JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. https://doi.org/10.1080/01621459.2012.682536

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data. Wiley Series in Probability and Mathematical Statistics*. Wiley, New York.

LIU, D. C. and NOCEDAL, J. (1989). On the limited memory BFGS method for large scale optimization. *Math. Program.* **45** 503–528. https://doi.org/10.1007/BF01589116

MALUMBRES, M. and BARBACID, M. (2009). Cell cycle, CDKs and cancer: A changing paradigm. *Nat. Rev. Cancer* **9** 153–166. https://doi.org/10.1038/nrc2602

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

MOGENSEN, U. B., ISHWARAN, H. and GERDS, T. A. (2012). Evaluating random forests for survival analysis using prediction error curves. *J. Stat. Softw.* **50** 1–23. https://doi.org/10.18637/jss.v050.i11

MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat. Med.* **38** 2074–2102. https://doi.org/10.1002/sim.8086

NIKITIN, A., EGOROV, S., DARASELIA, N. and MAZO, I. (2003). Pathway studio—the analysis and navigation of molecular networks. *Bioinformatics* **19** 2155–2157. https://doi.org/10.1093/bioinformatics/btg290

NIKOOIENEJAD, A., WANG, W. and JOHNSON, V. E. (2016). Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. *Bioinformatics* **32** 1338–1345. https://doi.org/10.1093/bioinformatics/btv764

NIKOOIENEJAD, A., WANG, W. and JOHNSON, V. E. (2020). Supplement to "Bayesian variable selection for survival data using inverse moment priors." https://doi.org/10.1214/20-AOAS1325SUPP

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. https://doi.org/10.1214/10-AOS792

SHA, N., TADESSE, M. G. and VANNUCCI, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* **22** 2262–2268. https://doi.org/10.1093/bioinformatics/btl362

SHIN, M., BHATTACHARYA, A. and JOHNSON, V. E. (2018). Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings. *Statist. Sinica* **28** 1053–1078. https://doi.org/10.5705/ss.202016.0167

SILL, M., HIELSCHER, T., BECKER, N. and ZUCKNICK, M. (2014). c060: Extended inference with Lasso and elastic-net regularized Cox and generalized linear models. *J. Stat. Softw.* **62** 1–22. https://doi.org/10.18637/jss.v062.i05

SUZUKI, H., OUCHIDA, M., YAMAMOTO, H., YANO, M., TOYOOKA, S., AOE, M., SHIMIZU, N., DATE, H. and SHIMIZU, K. (2008). Decreased expression of the SIN3A gene, a candidate tumor suppressor located at the prevalent allelic loss region 15q23 in non-small cell lung cancer. *Lung Cancer* **59** 24–31. https://doi.org/10.1016/j.lungcan.2007.08.002

TIBSHIRANI, R. et al. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395. https://doi.org/10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

UNO, H., CAI, T., TIAN, L. and WEI, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *J. Amer. Statist. Assoc.* **102** 527–537. https://doi.org/10.1198/016214507000000149

VAN HOUWELINGEN, H. C. and PUTTER, H. (2011). *Dynamic Prediction in Clinical Survival Analysis. Monographs on Statistics and Applied Probability* **123**. CRC Press, Boca Raton, FL. https://doi.org/10.1201/b11311

WANG, Z., MORRIS, J. S., CAO, S., AHN, J., LIU, R., TYEKUCHEVA, S., LI, B., LU, W., TANG, X. et al. (2017). Transcriptome deconvolution of heterogeneous tumor samples with immune infiltration. *BioRxiv*. https://doi.org/10.1101/146795

ZHANG, H. H. and LU, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94** 691–703. https://doi.org/10.1093/biomet/asm037