BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 18 2006, pages 2262–2268
doi:10.1093/bioinformatics/btl362

Gene expression

Bayesian variable selection for the analysis of microarray data with censored outcomes

Naijun Sha[1], Mahlet G. Tadesse[2] and Marina Vannucci[3,*]

[1]Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968, USA,
[2]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA 19104, USA and
[3]Department of Statistics, Texas A&M University, College Station, TX 77843, USA

Received on May 3, 2006; revised on June 9, 2006; accepted on June 27, 2006
Advance Access publication July 15, 2006
Associate Editor: Martin Bishop

ABSTRACT

Motivation: A common task in microarray data analysis consists of identifying genes associated with a phenotype. When the outcomes of interest are censored time-to-event data, standard approaches assess the effect of genes by fitting univariate survival models. In this paper, we propose a Bayesian variable selection approach, which allows the identification of relevant markers by jointly assessing sets of genes. We consider accelerated failure time (AFT) models with log-normal and log-*t* distributional assumptions. A data augmentation approach is used to impute the failure times of censored observations and mixture priors are used for the regression coefficients to identify promising subsets of variables. The proposed method provides a unified procedure for the selection of relevant genes and the prediction of survivor functions.

Results: We demonstrate the performance of the method on simulated examples and on several microarray datasets. For the simulation study, we consider scenarios with large number of noisy variables and different degrees of correlation between the relevant and non-relevant (noisy) variables. We are able to identify the correct covariates and obtain good prediction of the survivor functions. For the microarray applications, some of our selected genes are known to be related to the diseases under study and a few are in agreement with findings from other researchers.

Availability: The Matlab code for implementing the Bayesian variable selection method may be obtained from the corresponding author.

Contact: mvannucci@stat.tamu.edu

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The practical utility of variable selection is well recognized and this topic has been the focus of much research. A number of variable selection techniques have been developed for linear regression models, some of which have been extended to deal with censored survival data. These include the stepwise selection procedure (Peduzzi *et al*., 1980) and methods based on penalized likelihood, such as the lasso method (Tibshirani, 1997), and the non-concave penalized likelihood approach (Fan and Li, 2002). In the Bayesian framework, Volinsky *et al*. (1997) have proposed the use of Bayesian model averaging, where a set of likely models chosen with the leaps-and-bound algorithm are fitted one at a time. Faraggi and Simon (1998) have extended Lindley's (1968) decision theoretical approach for linear regression to the Cox model.

In the past few years, several microarray studies with time-to-event outcomes have been collected. The analysis of these data are complicated by their high-dimensionality, i.e. the fact that the number p of covariates is substantially larger than the sample size n. The existing methods described above cannot effectively handle the analysis of such data. The frequentist approaches are not well-defined, while the Bayesian procedures lack suitable search methods for the exploration of the space of possible models for $P > 20$. A widely used procedure for identifying genes related to survival outcomes consists of fitting univariate Cox models on each gene and selecting those that pass a threshold for significance (Rosenwald *et al*., 2002). Another approach first clusters the genes then fits a Cox model using the average expression level of each cluster as covariate (Hastie *et al*., 2001). This method however can be sensitive to the choice of the clustering algorithm. Different approaches using partial least squares (Park *et al*., 2002; Nguyen and Rocke, 2002) or principal components analysis (Li and Gui, 2004) have also been proposed. These methods select linear combinations of genes rather than the original variables. Recently, Gui and Li (2005) have extended Efron *et al*.'s (2004) least angle regression (LARS) procedure for variable selection to Cox models. They called their approach the LARS-Cox algorithm and presented applications to gene expression data analysis.

In this article, we propose a Bayesian variable selection approach for censored survival data in the context of accelerated failure time (AFT) models. The proposed method closely builds upon our previous work in variable selection (Brown *et al*., 1998; Sha *et al*., 2004; Tadesse *et al*., 2005). In this adaptation to survival data, we adopt a data augmentation (Tanner and Wong, 1987) approach to impute the censored survival times and build into the model a variable selection mechanism that uses mixture priors for the regression coefficients (George and McCulloch, 1993; Brown *et al*., 1998). We work under log-normal and log-*t* failure time distributional assumptions. We specify conjugate priors for the model parameters and derive a marginalized likelihood where the regression coefficients are integrated out. This approach substantially accelerates the model fitting task and can be valuable when variable selection is a major task of the inference. This is our main reason for choosing the AFT model over the more popular Cox model. With the Cox model, the regression coefficients cannot

[*]To whom correspondence should be addressed.

be analytically integrated out, thus requiring that the p regression parameters be sampled at each MCMC iteration. Complex MCMC procedures that sample all model parameters are computationally intensive and tend to have poor mixing. One possible approach for variable selection in the Cox model is put forward by Lee and Mallick (2004). There, the linear predictor of the model is treated as a random latent variable and a residual effect is added to it. This rather ad-hoc approach conveniently reduces the model to a standard linear form, from which the regression parameters can be generated based on their full conditional distributions or they can be marginalized. In addition, the inferential strategy adopted by Lee and Mallick (2004) performs the variable selection and survival function estimation tasks in a two-stage approach: a model is first estimated based on the variable subsets visited by a first MCMC sampler, then a second MCMC chain is run to estimate the survival function.

Our MCMC procedure leads to the simultaneous selection of relevant variables and the prediction of the survivor function. Unlike univariate approaches, our method allows the evaluation of the joint effect of sets of variables. It uses stochastic search techniques to explore the high-dimensional variable space and provides joint posterior probabilities for sets of variables as well as marginal posterior probabilities for the inclusion of single variables. We present simulation studies to investigate the performance of our approach in high dimensional data with different levels of correlation among variables. We also illustrate the method with applications to microarray datasets. The paper is organized as follows. In Section 2 we present the AFT model under log-normal and log-t parametric assumptions. We also describe the Bayesian variable selection approach and discuss the MCMC procedure. In Section 3 we assess the performance of the method using simulated data. Section 4 gives a detailed analysis of several DNA microarray studies and Section 5 concludes the article with a brief discussion.

2 METHODS

2.1 Accelerated failure time models

Survival analysis is concerned with the analysis of time-to-event data. A main feature of survival data, which makes it hard to analyze with conventional regression methods and requires special treatment, is the presence of censored observations. Here, we focus on AFT models as a useful alternative to the popular Cox model (Cox, 1972). Rather than assuming a multiplicative effect on the hazard functions as in the Cox regression, AFT models assume a multiplicative effect on the survival times. The general form of an AFT model is given by

$$\log(T_i) = \alpha + x_i'\beta + \varepsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where $\log(T_i)$ is the log survival time, α is the intercept term, x_i is a p-vector of covariates, β is the vector of regression parameters and the ε_i's are independent and identically distributed (iid) random variables whose common distribution may take a parametric form or may be unspecified. Kalbfleisch and Prentice (1980) give a comprehensive treatment of parametric AFT models and Wei (1992) reviews inference procedures for nonparametric models in the frequentist setting. The AFT model has not received much attention in the Bayesian framework. A parametric Bayesian analysis was presented by Bedrick *et al.* (2000) and semiparametric Bayesian approaches have been considered using a Dirichlet process prior (Christensen and Johnson, 1998) or a Dirichlet process mixture prior (Kuo and Mallick, 1997). In this article, we consider parametric AFT models under normal and t distributional assumptions for ε_i. We can then specify conjugate priors and integrate out the regression parameters from the model. This results in a much faster and more efficient MCMC sampler, as we circumvent the need to update the p-vector β at each MCMC iteration.

In what follows, let c_i be the censoring time independent of t_i. We observe $t_i^* = \min(t_i, c_i)$ and $\delta_i = I\{t_i \le c_i\}$, where $I\{\cdot\}$ is the usual indicator function. We make use of the data augmentation approach Tanner and Wong (1987) to impute the censored values. Let $\boldsymbol{W} = (w_1, \ldots, w_n)'$, where $w_i = \log(t_i)$, be the augmented data, so that

$$\begin{cases} w_i = \log(t_i^*) & \text{if } \delta_i = 1 \\ w_i > \log(t_i^*) & \text{if } \delta_i = 0 \end{cases} \tag{2}$$

2.1.1 Log-normal model Suppose the ε_i's in model (1) are iid $\mathcal{N}(0, \sigma^2)$, so the T_i's follow a log-normal distribution. The complete data are then normally distributed, $\boldsymbol{W} \mid \boldsymbol{X}, \alpha, \beta, \sigma^2 \sim \mathcal{N}(\alpha\mathbf{J} + \mathbf{X}\beta, \sigma^2\mathbf{I})$ with $\mathbf{J}_{n\times1} = (1, \ldots, 1)'$ and $\mathbf{I}_{n\times n}$ the identity matrix. Conjugate priors for this model are given by

$$\begin{aligned} \alpha \mid \sigma^2 &\sim \mathcal{N}(\alpha_0, h_0\sigma^2) \\ \beta \mid \sigma^2 &\sim \mathcal{N}(\beta_0, \sigma^2\Sigma_0) \\ \sigma^2 &\sim \mathcal{IG}(\nu_0/2, \nu_0\sigma_0^2/2), \end{aligned} \tag{3}$$

where the hyperparameters $\alpha_0, h_0, \beta_0, \Sigma_0, \nu_0, \sigma_0^2$ need to be specified. Vague priors on α and β are obtained by choosing $\alpha_0 = 0$ and h_0 large, $\beta_0 = 0$ and $\Sigma_0 = h\mathbf{I}$ with h large. For σ^2, a weakly informative prior is obtained with a small value of ν_0.

After integrating out α, β and σ^2, the marginal likelihood of the augmented data becomes

$$\begin{aligned} \mathcal{L}(\boldsymbol{W} \mid \boldsymbol{X}) \quad \propto \quad & \{\nu_0\sigma_0^2 + (\boldsymbol{W} - \alpha_0\mathbf{J} - \boldsymbol{X}\beta_0)'(\mathbf{I} + h_0\mathbf{J}\mathbf{J}' + \boldsymbol{X}\Sigma_0\boldsymbol{X}')^{-1} \\ & (\boldsymbol{W} - \alpha_0\mathbf{J} - \boldsymbol{X}\beta_0)\}^{-\frac{n+\nu_0}{2}}. \end{aligned} \tag{4}$$

This corresponds to a multivariate t-distribution

$$\boldsymbol{W} \mid \boldsymbol{X} \sim \mathcal{T}_{\nu_0}[\alpha_0\mathbf{J} + \boldsymbol{X}\beta_0, \sigma_0^2(\mathbf{I} + h_0\mathbf{J}\mathbf{J}' + \boldsymbol{X}\Sigma_0\boldsymbol{X}')] \tag{5}$$

with truncation given by equation (2). Thus, the full conditional of a censored case, w_i with $\delta_{\mathrm{i}} = 0$, follows a univariate truncated t-distribution and it can be updated using Gibbs sampling.

2.1.2 Log-t model Suppose now that the ε_i's in (1) are iid from a t-distribution with ν degrees of freedom. The T_i's then have a log-t distribution and the augmented data, W_i, follows a $t_\nu(\alpha + x_i'\beta, \sigma)$ distribution. With the introduction of auxiliary random variables λ_i, the t-distribution can be written as the scale mixture of a normal

$$w_i = \alpha + x_i'\beta + \sigma\sqrt{\lambda_i}\tilde{\varepsilon}_i, \quad \tilde{\varepsilon}_i \sim \mathcal{N}(0,1),\ \lambda_i \sim \mathcal{IG}(\nu/2, \nu/2). \tag{6}$$

We can now adopt the same prior setting as in Section 2.1.1 and the marginal likelihood for the augmented data are given by a truncated t-distribution with mean $\alpha_0\mathbf{J} + \boldsymbol{X}\beta_0$ and variance $\sigma_0^2(\boldsymbol{\Lambda} + h_0\mathbf{J}\mathbf{J}' + \boldsymbol{X}\Sigma_0\boldsymbol{X}')$ where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$.

2.2 Mixture priors for variable selection

Bayesian methods for variable selection introduce a latent p-vector γ with binary entries, which is built into the prior of the β's. There is a vast amount of literature on Bayesian variable selection methodologies. We cite Chipman *et al.* (2001) for a comprehensive review. Briefly, the vector γ is used to search the space of variable subsets and identify the most promising models. The MCMC procedure starts from a randomly chosen initial model and is quickly drawn toward models with relatively high posterior probability. In this approach, the regression coefficients are assumed to arise from a scale mixture of a point mass at 0 and a normal density (George and McCulloch, 1993)

$$\beta_j \mid \gamma_j, \sigma^2 \sim (1 - \gamma_j)\mathcal{I}(0) + \gamma_j\mathcal{N}(0, \sigma^2\tau_j), \quad j = 1, \ldots, p, \tag{7}$$

where τ_j is the j-th diagonal element of Σ_0 in (3). We assume the γ_j's are independent Bernoulli random variables, with ω elicited as the proportion of variables expected *a priori* in the model. This prior formulation can be relaxed by specifying a Beta (a, b) hyperprior on ω. As for prediction, the

Bayesian approach allows model averaging, which averages over a range of likely models to estimate future outcomes.

2.3 Model fitting via MCMC

Using the parametric AFT models with conjugate priors, we are able to integrate out the model parameters, α, β and σ^2. The model fitting thus consists of updating the remaining parameters, namely the variable selection indicator, γ, the latent failure time for censored cases, w_i with $\delta_i = 0$, and for the log-*t* case, the auxiliary parameters $\mathbf{\Lambda}$. Our MCMC procedure iterates the following steps:

(1) Update γ from its full conditional via a Metropolis search by randomly choosing one of the following move types:

(a) add/delete: randomly pick one of the p indices in γ^{old} and change its value from 0 to 1, or 1 to 0,

(b) swap: pick independently and at random a 0 and a 1 in γ^{old} and switch their values. The proposed γ^{new} is accepted with probability

$$\min\left\{1, \frac{f(\gamma^{\text{new}} \,|\, \boldsymbol{X}, \boldsymbol{W})}{f(\gamma^{\text{old}} \,|\, \boldsymbol{X}, \boldsymbol{W}),}\right\} \quad \text{or} \quad \min\left\{1, \frac{f(\gamma^{\text{new}} \,|\, \boldsymbol{X}, \boldsymbol{W}, \mathbf{\Lambda})}{f(\gamma^{\text{old}} \,|\, \boldsymbol{X}, \boldsymbol{W}, \mathbf{\Lambda})}\right\}, \tag{8}$$

where

$$f(\gamma|\boldsymbol{X}, \boldsymbol{W}) \propto f(\boldsymbol{W}|\boldsymbol{X}, \gamma)p(\gamma)$$
$$f(\gamma|\boldsymbol{X}, \boldsymbol{W}, \mathbf{\Lambda}) \propto f(\boldsymbol{W}|\boldsymbol{X}, \mathbf{\Lambda}, \gamma)p(\gamma),$$

respectively for the log-normal and log-*t* cases.

(2) Update the censored elements of $\boldsymbol{W}$, w_i with $\delta_i = 0$, from $f(w_i \,|\, \boldsymbol{W}_{(-i)}$, $\boldsymbol{X}$, $\gamma)$ in the log-normal and $f(w_i \,|\, \boldsymbol{W}_{(-i)}, \boldsymbol{X}, \gamma, \mathbf{\Lambda})$ in the log-*t* case.

(3) For the log-*t* model only, we sample Λ using a sub-Gibbs sampler to update each λ_i from its full conditional $f(\lambda_i \,|\, \boldsymbol{X}, \boldsymbol{W}, \gamma, \mathbf{\Lambda}_{(-i)}) \propto f(\boldsymbol{W} \,|\, \boldsymbol{X}, \gamma, \mathbf{\Lambda}) f(\lambda_i)$. Since this distribution does not have a standard form, we use a Metropolis–Hastings algorithm, in which a candidate λ_i^{new} is generated from an inverse gamma, $\mathcal{IG}(\nu/2, \nu\lambda_i^{\text{old}}/2)$, which we denote by $q(\lambda_i^{\text{new}} \,|\, \lambda_i^{\text{old}})$. The acceptance probability is

$$\min\left\{1, \frac{f(\lambda_i^{\text{new}} \,|\, \boldsymbol{X}, \boldsymbol{W}, \gamma, \mathbf{\Lambda}_{(-i)})q(\lambda_i^{\text{old}} \,|\, \lambda_i^{\text{new}})}{f(\lambda_i^{\text{old}} \,|\, \boldsymbol{X}, \boldsymbol{W}, \gamma, \mathbf{\Lambda}_{(-i)})q(\lambda_i^{\text{new}} \,|\, \lambda_i^{\text{old}})}\right\}. \tag{9}$$

2.4 Posterior inference

The MCMC samples can be used to draw posterior inference. Of particular interest are the prediction of survival times for future patients, the estimation of their predictive survivor functions, and the identification of relevant variables.

2.4.1 Prediction of survival time Suppose n_f patients with covariate data $\boldsymbol{X}_f (n_f \times p)$ are available and we wish to predict their survival times. This can be accomplished via model averaging (Madigan and Raftery, 1994; Brown *et al*., 1998). For a given γ, we can evaluate the joint distribution of $\boldsymbol{W}$ and $\boldsymbol{W_f}$ then use properties of the multivariate-*t* distribution to derive the conditional distribution of $\boldsymbol{W}_f$ given $\boldsymbol{W}$. Let $\widehat{\boldsymbol{W}}$ be the augmented data with the censored failure times imputed by the mean of their sampled values, $\widehat{\boldsymbol{W}} = \frac{1}{M}\sum_{k=1}^{M} \boldsymbol{W}^{(k)}$. The log-survival times are estimated through the posterior mean of the predictive distribution weighted by the posterior probabilities of the visited models:

$$\begin{aligned} \widehat{\boldsymbol{W}}_f &= \sum_{\gamma} \boldsymbol{X}^*_{f(\gamma)} \hat{\boldsymbol{\beta}}^*_{\gamma} \cdot p(\gamma \,|\, \boldsymbol{X}, \widehat{\boldsymbol{W}}), && \text{for log-normal} \\ \widehat{\boldsymbol{W}}_f &= \sum_{\gamma} \boldsymbol{X}^*_{f(\gamma)} \hat{\boldsymbol{\beta}}^{\gamma}_{*} \cdot p(\gamma \,|\, \mathbf{X}, \widehat{\boldsymbol{W}}, \hat{\Lambda}), && \text{for log-}t, \end{aligned} \tag{10}$$

with $\hat{\Lambda} = \dfrac{1}{M}\sum_{k=1}^{M} \Lambda^{(k)}$. Here

$$\hat{\boldsymbol{\beta}}^{\gamma}_{*} = \left(\boldsymbol{X}^{*'}_{\gamma} \boldsymbol{V} \boldsymbol{X}^*_{\gamma} + \Sigma_{0}{}^{*-1}_{(\gamma)}\right)^{-1} \left(\boldsymbol{X}^{*'}_{\gamma} \boldsymbol{V} \widehat{\boldsymbol{W}} + \Sigma_{0}{}^{*-1}_{(\gamma)} \boldsymbol{\beta}_{0}{}^{*}_{\gamma}\right)$$

with

$$\boldsymbol{X}^*_{\gamma} = (\boldsymbol{J}, \boldsymbol{X}_{\gamma}), \beta_{0}{}^{*}_{\gamma} = \begin{pmatrix} \alpha_0 \\ \beta_0 \gamma \end{pmatrix}, \Sigma_{0}{}^{*}_{(\gamma)} = \begin{pmatrix} h_0 & 0 \\ 0 & \Sigma_{0(\gamma)} \end{pmatrix},$$

and $\boldsymbol{V} = \boldsymbol{I}$ or $\boldsymbol{V} = \text{diag}(1/\hat{\boldsymbol{\lambda}}_1, 1/\hat{\boldsymbol{\lambda}}_2, \ldots, 1/\hat{\boldsymbol{\lambda}}_n)$, respectively for the log-normal and log-*t* cases.

2.4.2 Predictive survivor function Let x be the covariate vector for a new subject. We implement the sampling-based approach of Gelfand (1996) to compute an estimate of the survivor function. For the log-normal case we have

$$\begin{aligned} P(T > t \,|\, x, \boldsymbol{X}, \widehat{\boldsymbol{W}}) &= P(W > w \,|\, x, \boldsymbol{X}, \widehat{\boldsymbol{W}}) \\ &= \int P(W > w \,|\, x, \boldsymbol{X}, \widehat{\boldsymbol{W}}, \gamma) p(\gamma \,|\, x, \boldsymbol{X}, \widehat{\boldsymbol{W}}) d\gamma \\ &\approx \sum_{k=1}^{M} P(W > w \,|\, x, \boldsymbol{X}, \widehat{\boldsymbol{W}}, \gamma^{(k)}) \\ &\times p(\gamma^{(k)} \,|\, \boldsymbol{X}, \widehat{\boldsymbol{W}}), \end{aligned} \tag{11}$$

where $\gamma^{(k)}$ is the model visited at the *k*-th iteration, $P(W > w \,|\, x, \boldsymbol{X}, \widehat{\boldsymbol{W}}, \gamma^{(k)})$ is the tail area of a univariate *t*-distribution, and $p(\gamma \,|\, \boldsymbol{X}, \widehat{\boldsymbol{W}})$ is used as importance sampling density for $p(\gamma \,|\, x, \boldsymbol{X}, \widehat{\boldsymbol{W}})$. For the log-*t* case this becomes

$$\begin{aligned} P(T > t \,|\, x, \boldsymbol{X}, \widehat{\boldsymbol{W}}, \hat{\mathbf{\Lambda}}) &= P(W > w \,|\, x, \boldsymbol{X}, \widehat{\boldsymbol{W}}, \hat{\mathbf{\Lambda}}) \\ &\approx \sum_{k=1}^{M} P(W > w \,|\, x, \boldsymbol{X}, \widehat{\boldsymbol{W}}, \hat{\mathbf{\Lambda}}, \gamma^{(k)}) \\ &\times p(\gamma^{(k)} \,|\, \boldsymbol{X}, \widehat{\boldsymbol{W}}, \hat{\mathbf{\Lambda}}). \end{aligned} \tag{12}$$

2.4.3 Selection of variables Inference about variable selection can be done either through the joint posterior distribution of γ or through the marginal posterior distributions of its elements. The former selects variables based on the γ vector with largest posterior probability among all vectors visited by the MCMC sampler, i.e.

$$\hat{\gamma} = \underset{1 \le k \le M}{\text{argmax}}\; p(\gamma^{(k)} \,|\, \boldsymbol{X}, \widehat{\boldsymbol{W}}) \quad \text{or} \quad \hat{\gamma} = \underset{1 \le k \le M}{\text{argmax}}\; p(\gamma^{(k)} \,|\, \boldsymbol{X}, \widehat{\boldsymbol{W}}, \hat{\mathbf{\Lambda}}) \tag{13}$$

for log-normal and log-*t*, respectively.

Alternatively, the marginal posterior probability that variable *j* is included in the model can be estimated by the empirical frequency in the Markov chain Monte Carlo output and the variables associated with the risk of failure can then be identified as those with marginal posterior probability greater than some arbitrary threshold, $\hat{\gamma}_j = I\{p(\gamma_j = 1 \,|\, \boldsymbol{X}) > \kappa\}$.

3 SIMULATION STUDY

In this section we investigate the performance of the proposed approach using simulated data with different levels of correlation among variables. We follow the strategy described in Gui and Li (2005) to generate data for n event times and p covariates, of which p_γ are chosen to be related to the survival time. The remaining $p - p_\gamma$ variables are not related to the survival time but may be correlated with the p_γ relevant predictors. As described in Gui and Li (2005), this is accomplished by first drawing a $n \times n$ matrix A from a uniform $U(-1.5,\ 1.5)$ distribution and choosing p_γ of the columns to be related to the survival time. The normalized orthogonal basis of A, $\{\vartheta_1, \ldots, \vartheta_{p_\gamma}, \varrho_1, \ldots, \varrho_{n-p_\gamma}\}$ is constructed using Gram–Schmidt orthonormalization. By Cauchy's inequality, for any $p_\gamma \times (n - p_\gamma)$ matrix T, $\text{corr}(\vartheta y, (\varrho + \vartheta T)x) \le \rho/\sqrt{1 + \rho^2}$, for $\forall y \in R^{p_\gamma}, \forall x \in R^{n-p_\gamma}$, where ρ^2 is the largest eigenvalue of $T'T$. The $p - p_\gamma$ variables not related to the risk of failure are then generated from the linear space $C = \{\varrho + \vartheta T\}$ with the appropriate choice of the maximum eigenvalue of $T'T$. We considered maximum possible correlations of 0, 0.5 and 0.8.

We first considered $n = 100$, $p = 1{,}000$ and $p_\gamma = 10$. For the regression coefficients corresponding to the relevant covariates, we generated small to medium effect sizes by drawing values from uniform $U(-3, -0.1)$ and $U(0.1, 3)$ distributions. The remaining elements were set to zero. The survival times were generated using the AFT model

$$\log(t_i) = \alpha + x_i\beta + \varepsilon_i, \quad i = 1, \ldots, 100$$

with ε_i arising from a normal or a t density. In addition, 20% of the observations were randomly censored at time $c_i \sim$ Uniform $(0,t_i)$. For each sample, we observe $[\boldsymbol{x_i}, \min(t_i, c_i), \delta_i]$, where δ_i is the censoring indicator. We simulated another 100 uncensored observations as our validation set using a different but similarly generated covariate matrix and the same regression coefficients.

We report here the results for the log-normal example. These were obtained by choosing the hyperparameters to lead to weakly informative priors. We used $\alpha_0 = 0$ and $\beta_0 = 0$. We set $h_0 = 10^6$ and $\Sigma_0 = h\boldsymbol{I}$. The results did not appear to be sensitive to the choice of h_0, but we noticed some sensitivity to the specification of h. This hyperparameter regulates the amount of shrinkage in the model induced by the mixture prior. In general, one would want to avoid values that are too small, which lead to too much regularization and poor mixing of the MCMC chains, resulting in few distinct models being visited. Large values, on the other hand, could induce non-linear shrinkage as a result of Lindley's paradox (Lindley, 1957). In Sha *et al.* (2004) we employ mixture priors in the context of probit models for classification and provide some guidelines on how to choose the shrinkage parameter, such that the ratio of prior to posterior precision is relatively small. Similarly here, a range of possible values for the hyperparameter h can be defined by the ratio of prior to posterior precision. In practice, we have found that values of h that provide good mixing of the MCMC sampler, with 20–30% distinct visited models, are appropriate. The results we present in this simulation study are obtained by setting $h = 1$ in all cases. This choice was not critical as long as h was chosen in the range [0.1, 10]. For the prior on γ, we considered a Bernoulli prior with expected number of included variables equal to 15. We set ν_0=3 for the inverse gamma prior on σ^2, which we center around 1. For all the examples, a starting model with 40 randomly selected variables was used and the MCMC chain was run for 200 000 iterations with the first 100 000 used as burn-in.

We simulated the ε_i's from a $\mathcal{N}(0, 1)$ density and fitted the AFT model described in Section 2.1.1 via the proposed variable selection MCMC inferential procedure. Figure 1a shows the trace plot for the number of variables selected at each MCMC iteration for the dataset with no correlation among variables. The chain mixed well, concentrating mostly on models with 12 to 23 covariates. Figure 1b displays the marginal posterior probabilities of inclusion of single variables, $p(\gamma_j = 1 \mid \boldsymbol{X})$. There are 8 variables with marginal posterior probabilities greater than 0.5, all in the set of 10 covariates simulated to effectively predict the survival times. The γ vector with largest posterior probability among all visited models contained 10 variables with the same 8 good variables identified by the marginal probabilities. We also evaluated the predictive performance of the method. The estimated survival times for the 100 observations in the validation set computed using formula (10) resulted in a mean squared error $\mathrm{MSE}(\boldsymbol{W}_f, \widehat{\boldsymbol{W}}_f) = 3.779$. We also estimated the survivor functions for the validation set using equation (11). The true and estimated survivor functions for two of the samples are given in Figure 2a. Figure 2b displays the true and estimated mean survivor functions, obtained by averaging over all samples, together with the Kaplan–Meier estimate. The results from our procedure provide good fit to the data.

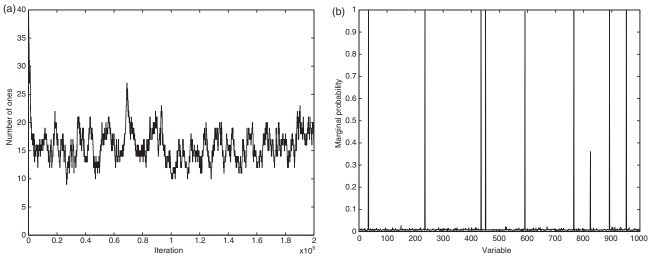

Fig. 1. Analysis of log-normal simulated data with zero correlation. (**a**) Number of included variables. (**b**) Marginal posterior probabilities of inclusion.

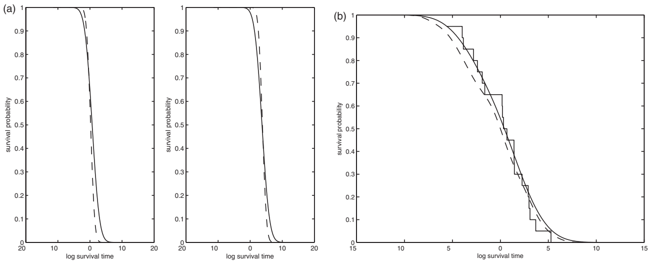

Fig. 2. True and estimated survivor functions for log-normal simulated data with zero correlation. (**a**) True (dashed line) and estimated (solid line) survivor functions for two subjects. (**b**) True (dashed), estimated mean survivor function with our approach (solid) and Kaplan–Meier estimate (steps).

Table 1. Simulated log-normal survival data with $p_\gamma = 10$: marginal posterior probabilities of inclusion of relevant variables

β_j	$n = 100$			$n = 50$		
	$\rho = 0$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.8$
−2.3991	1.0000	1.0000	1.0000	0.2216	0.3087	0.3990
2.2711	1.0000	1.0000	1.0000	0.0638	0.4427	0.3454
1.1848	1.0000	0.6672	0.8253	0.0088	0.1030	0.0302
1.7311	1.0000	0.6745	0.9962	0.5889	0.0667	0.0556
0.2118	0.0101	0.0090	0.0205	0.0051	0.0098	0.0185
−1.2105	0.9933	0.5381	0.8869	0.0188	0.0174	0.0184
1.4173	1.0000	0.8393	1.0000	0.0295	0.0117	0.6194
−0.8331	0.3611	0.8108	0.7210	0.0075	0.0383	0.0189
−1.1797	1.0000	0.7962	1.0000	0.1574	0.0120	0.0532
−2.4062	1.0000	1.0000	1.0000	0.3468	0.5154	0.3078

We repeated the analysis for the simulated examples with different correlation levels among the variables. The MCMC samplers showed similar behavior (data not shown). With a correlation of 0.5 and 0.8 among variables, we obtained results identical to those reported above. Table 1, column 1 lists the simulated regression coefficients and columns 2–4 report the marginal posterior probabilities of inclusion for the 10 variables associated with survival times under the different correlation levels.

We further investigated the performances of our method by looking at different sample sizes and different number of discriminating variables. First we lowered the sample size to $n = 50$, which is a more challenging scenario. We repeated the simulation outlined above for the three correlation levels. The results are reported in Table 1, columns 5–7. As expected, we note that the performance of the method is not as good with a smaller sample size, although a good proportion of the relevant variables are still correctly identified. This result confirms the intuition that it is harder to identify the predictive variables as the sample size decreases. Next, we doubled the number of predictive variables and considered the case $p_\gamma = 20$. We repeated the simulation study above, with the different correlation levels and with $n = 50$ and $n = 100$. The findings were essentially concordant with the observations above. A good proportion of the true variables were correctly identified for all correlation levels with better results under larger sample sizes (see Supplementary Material).

We note that a general feature of the above results is that the predictive variables that do not get selected are those with smaller regression coefficients, which intuitively are more difficult to detect. Small posterior probabilities reflect the fact that small regression coefficients are in general harder to be discriminated from the zero value. To investigate this further we re-simulated the same scenarios considered above with the non-zero regression coefficients drawn from uniform $U(-1,-0.1)$ and $U(0.1,1)$ distributions. In the Supplementary Material, we report results for the case $n = 100$ and $p_\gamma = 10$. As expected we obtained posterior probabilities, which are in general smaller than those in Table 1. There were 6 variables with fairly large probabilities, which also appeared in the best visited model. Similar findings apply to all the correlation cases.

Finally, we considered a simulation in which all the variables were just noise and were not associated with the survival time. We looked at the case of zero correlation and $n = 50$. Although, the Markov chain still visited models with 10 to 30 variables, in accordance with the prior expectation, the estimates of the marginal posterior probabilities all resulted in very small values, indicating that no variable could be selected.

4 APPLICATION TO DNA MICROARRAY DATA

We now illustrate the practical utility of our methodology using three different microarray datasets: a breast cancer data investigating genes associated with the risk of developing distant metastases within a short time interval (van't Veer *et al*., 2002), and two separate studies examining gene expression profiles predictive of survival in diffuse large B-cell lymphoma (Alizadeh *et al*., 2002; Rosenwald *et al*., 2002). The identification of these molecular signatures could lead to improved diagnosis, as patients could then be stratified into different risk groups and receive the appropriate treatment regimen. In addition, there is interest in estimating the survivor function for future patients.

In order to decide on suitable distributional assumptions, we computed non-parametric kernel density estimates of the log failure times. For all three datasets a log-normal model appeared to be a reasonable approximation to the underlying survival distribution (see Supplementary Material). For each data, in order to avoid possible dependence of the results on the initial model, we ran four MCMC chains with starting models of 1, 10, 50 and 100 randomly selected γ_j's set to one. Each MCMC sampler was run for 200 000 iterations with the first 100 000 used as burn-in.

Table 2. Breast cancer data: genes associated with time to distant metastasis

GenBank ID	Symbol	$p(\gamma_j = 1 \mid \boldsymbol{X})$
NM_001141	ALOX15B	0.27
AI352507		0.20
AI141554		0.19
AW206610		0.18
NM_003239	TGFB3	0.14
NM_003862	FGF18	0.12
NM_000793	DIO2	0.12
NM_004887	CXCL14	0.11
AI912975		0.11
NM_007203	AKAP2	0.08
NM_004490	GRB14	0.06

4.1 Breast cancer data

First, we illustrate our method using the van't Veer *et al*. (2002) study. The data consist of primary tumors from 78 lymph-node negative breast cancer patients, 34 of whom developed distant metastases within five years and 44 who continued to be disease-free. The original authors formulated the analysis in a classification framework. Here, instead, we want to take advantage of all the information available in the data and consider each patient's failure time as the outcome of interest. Patients who did not experience distant metastases within the five years constitute censored cases. Two patients had several missing gene expression levels and were removed from the analysis. The remaining data were split into a training and a test set with 38 patients in each group.

The gene expression levels were monitored using two-channel arrays with ~25 000 probes. Transcript abundance of genes were estimated using the intensity ratio with respect to a reference pool obtained by combining cRNA samples from all tissues. We used the same criteria as those described in the original paper to pre-process the data. Probes with more than a 2-fold difference and a *P*-value less than 0.01 in more than five patients were kept. This resulted in 3839 genes considered for analysis.

For the prior specification, we set $\alpha_0 = 0$, $h_0 = 100$, $\beta_0 = 0$ and chose h =1. We specified a weakly informative prior for σ^2 by setting $\nu_0 = 3$ with $\sigma_0 = 1$, a value commensurate with the residual sum of squares obtained by considering the uncensored samples only, based on all genes. We chose independent Bernoulli priors for the components of γ with the expected number of ones set to 10. This favors the selection of small sets of genes.

All four MCMC chains mostly visited models with 5 to 15 variables. We assessed the concordance of the visited models across the MCMC runs by examining the differences in marginal posterior probabilities for the inclusion of variables. There was a good agreement between pairs of MCMC chains with most differences being close to 0 (see Supplementary Material). We pooled the output of the four chains, normalized the relative posterior probabilities, and computed $p(\gamma_j = 1 \mid \boldsymbol{X})$ based on the pooled set of models. Genes with high posterior probabilities represent promising targets for further biological studies. Table 2 lists the eleven genes with largest marginal probabilities. Most of these are known to be associated with the development and progression of breast cancer. For example, we identified the transforming growth factor beta 3 (TGFB3), which acts hormonally to control the proliferation and differentiation of multiple cell types. TGFB3 is in fact known to be correlated with overall survival in human breast carcinoma and with lymph node metastasis (Ghellal *et al*., 2000). In addition, we identified fibroblast growth factor 18 (FGF18), which plays an important role in tumor growth and invasion. Another important gene known to be involved in metastasis, chemokine C-X-C motif ligand 14 (CXCL14), was also selected by our approach. CXCL14 has been demonstrated to enhance breast cancer cell growth, migration and invasion (Allinen *et al*., 2004). The growth factor receptor-bound protein 14 (GRB14), which has previously been found to express in some breast cancer cell lines and to correlate with estrogen receptor positivity (Daly *et al*., 1996) was also detected.

As we mentioned above, van't Veer *et al*. (2002) tackled the analysis as a classification problem. They identified 70 markers that discriminate between patients who experienced distant metastases within five years and those who didn't. Since our method takes into account the actual failure times, we expect our results to be more informative and to better capture genes associated with the risk of developing distant metastases. Interestingly, we found that two of the genes we identified, TGFB3 and FGF18, were among van't Veer *et al*.'s prognostic markers.

We finally assessed the predictive performance of our selected models by computing the estimated survival times and survivor curves for the 16 uncensored patients in the validation set and obtained $\mathrm{MSE}(\boldsymbol{W}_f, \widehat{\boldsymbol{W}}_f) = 1.9317$.

4.2 Diffuse large B-cell lymphoma studies

We now consider two distinct diffuse large B-cell lymphoma (DLBCL) microarray studies evaluating gene expression profiles associated with patients' survival (Alizadeh *et al*., 2000; Rosenwald *et al*., 2002). Both disease-free and overall survival are major concerns in the treatment of this disease. Although most patients respond initially to chemotherapy, fewer than half achieve lasting remission.

4.2.1 Alizadeh et al. (2000) data Gene expression levels of 42 DLBCL patients were monitored on a custom-designed 'Lymphochip' microarray. This cDNA array contains genes that are preferentially expressed in lymphoid cells and genes that are implicated in cancer or immunology. After a pre-processing step to remove unreliable expression readings, 4026 probes were kept for analysis.

We re-analyzed this data using our method with similar prior settings as the breast cancer example and ran four MCMC chains. The samplers visited models with 5 to 25 probes and there was good concordance across the chains. After pooling the output from the four chains, we identified four genes with marginal posterior probabilities greater than 0.1 (Table 3). All the selected genes play important roles in apoptotic processes and/or the development and progression of various cancers. For example, B-cell leukemia/lymphoma 2 (BCL2) is known to be an important predictor of survival in DLBCL. It has also been identified by Alizadeh *et al*. (2000) as having differential expression between the two major subgroups of DLBCL, germinal center B-cell like and activated B-cell like DLBCL, which were found to have statistically significant difference in survival. Several other studies have also shown association between BCL2 expression and disease-free survival (Kramer *et al*., 1998).

Table 3. Alizadeh *et al*. (2000) DLBCL data: genes associated with survival time

Clone ID	Symbol	$p(\gamma_j = 1 \mid \boldsymbol{X})$
18247	CASP3	0.27
20339	STP1	0.23
13603	BCL2	0.16
19384	JNK3	0.12

Table 4. Rosenwald *et al*. (2002) DLBCL data: genes associated with survival time

GenBank ID	Symbol	$p(\gamma_j = 1 \mid \boldsymbol{X})$
D42043	RAFTLIN	0.75
D88532	PIK3R3	0.72
BC012161	SEPT1	0.57
LC_33732		0.33
D13666	POSTN	0.30
AK000978		0.25
U51004	HINT1	0.17
U11791	CCNH	0.16
X00457	HLA-DPα	0.15
M29536	EIF2S2	0.13
AF017786	PPAP2B	0.11

We also compared our results to those of Lee and Mallick (2004) who proposed a Bayesian variable selection approach in survival and considered this data for analysis. However, instead of analyzing all 4026 probes, they first performed univariate *t*-tests and fitted their procedure on only 1000 probes. They selected a set of 12 genes as being associated with survival, one of which was JNK3.

4.2.2 Rosenwald et al. (2002) data In this study, 240 patients were monitored using a Lymphochip cDNA microarray with 7399 probes. The authors fitted univariate Cox proportional hazards models on each probe after dividing the data into a training and a test set. Here, we focus on the 160 patients in the training set.

Using the same prior settings as above we ran four MCMC chains and pooled the output. Table 4 reports the 11 genes with marginal posterior probabilities greater than 0.1. Among these were the HLA-DPα gene from the major histocompatibility class II family, which is associated to various cancers and was also identified by Rosenwald *et al*. (2002). Gui and Li (2005), who also analyzed this data using their LARS-Cox algorithm, identified this gene as one of the top four to be associated with survival. Another important gene we identified and that overlapped with Gui and Li (2005) analysis is osteoblast specific factor 2 (fasciclin I-like, POSTN), which belongs to the lymph node signature group defined by Rosenwald *et al*. (2002).

5 DISCUSSION

In this article, we have proposed Bayesian variable selection methods for the analysis of high-dimensional data with censored outcomes. We have considered parametric AFT models to relate failure times to covariates. We have adopted a data augmentation approach to impute the censored survival times and have built into the model a variable selection mechanism that uses mixture priors for the regression coefficients. In addition to variable selection, our method provides prediction of the survivor function, which can be estimated using the same MCMC output resulting from the model fit.

Our approach exploits the conjugacy of the priors to integrate out some parameters and define an efficient MCMC procedure. In this article, we have considered log-normal and log-*t* failure times. Other distributional assumptions require more complex MCMC sampling techniques or approximation methods because the regression coefficients cannot be analytically integrated out. For example, for log-logistic failure time, one could use the connection between logistic and *t* distributions. As discussed by Albert and Chib (1993), a logistic random variable can be approximated by 1.577 times a $t(8)$ random variable. This implies that, following the discussion of Section 2.1.2, we can write the log-logistic model in terms of a normal regression as

$$W_i \approx \alpha + x_i\beta + \sigma d\sqrt{\lambda_i}\tilde{\varepsilon}_i, \quad \tilde{\varepsilon}_i \sim \mathcal{N}(0,1), \lambda_i \sim \mathcal{IG}(\nu/2, \nu/2) \tag{14}$$

with $d = 1.577$ and $\nu = 8$.

We have staged a simulation study with high-dimensional data and possibly correlated variables to assess the performance of the procedure. The survival times were simulated as linear combinations of a small set of variables with various degrees of correlation between the relevant and non-relevant (noisy) variables. We obtained good results both in terms of the variable selection and survival function estimation. The performance of our method was only slightly affected by correlation among the covariates. We also analyzed several microarray datasets and identified important markers with known association to cancer development and progression. The proposed method has a wide range of applications and can be extended to the context of quantitative trait loci (QTL) scans or associations between traits and polymorphic markers.

ACKNOWLEDGEMENTS

The authors would like to thank Hongzhe Li for sharing the code to generate the simulated examples. N.S., M.G.T. and M.V. are supported by NIH/NHGRI grant R01HG003319. N.S.is also partially supported by BBRC/RCMI NIH grant 2G12RR08124 and M.V. by NSF CAREER award DMS-0093208.

Conflict of Interest: none declared.

REFERENCES

Albert,J. and Chib,S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.*, **88**, 669–679.

Alizadeh,A. *et al*. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Allinen,M. *et al*. (2004) Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell*, **6**, 17–32.

Bedrick,E. *et al*. (2000) Bayesian accelerated failure time analysis with application to veterinary epidemiology. *Stat. Med.*, **19**, 221–237.

Brown,P. *et al*. (1998) Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Series B*, **60**, 627–641.

Chipman,H., George,E. and McCulloch,R. (2001) *The Practical Implementation of Bayesian Model Selection.* IMS Lecture Notes—Monograph Series Volume 38.

Christensen,R. and Johnson,W. (1988) Modelling accelerated failure time with a Dirichlet process. *Biometrika*, **75**, 693–704.

Cox,D. (1972) Regression models and life tables. *J. R. Stat. Soc. Series B*, **34**, 187–220.

Daly,R. *et al*. (1996) Cloning and characterization of GRB14, a novel member of the GRB7 gene family. *J. Biol. Chem.*, **271**, 12502–12510.

Efron,B. *et al*. (2004) Least angle regression. *Ann. Stat.*, **32**, 407–451.

Fan,J. and Li,R. (2002) Variable selection for Cox's proportional hazards model and frailty model. *Ann. Stat.*, **30**, 74–99.

Faraggi,D. and Simon,R. (1998) Bayesian variable selection method for censored survival data. *Biometrics*, **54**, 1475–1485.

Gelfand,A. (1996) Model determination using sampling-based methods. *Markov chain Monte Carlo in Practice,* Chapman and Hall, London, pp. 145–161.

George,E. and McCulloch,R. (1993) Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.*, **88**, 881–889.

Ghellal,A. *et al*. (2000) Prognostic significance of tgfβ1 and tgfβ3 in human breast carcinoma. *Anticancer Res.*, **20**, 4413–4418.

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.

Hastie,T. *et al*. (2001) Supervised harvesting of expression trees. *Genome Biol.*, **2**, research 0003.1–0003.12.

Kalbfleisch,J. and Prentice,R. (1980) *The Statistical Analysis of Failure Time Data.* Wiley, NY.

Kramer,M. *et al*. (1998) Clinical relevance of BCL2, BCL6, and MYC rearrangements in diffuse large B-cell lymphoma. *Blood*, **92**, 3152–3162.

Kuo,L. and Mallick,B. (1997) Bayesian semiparametric inference for the accelerated failure time model. *Canadian J. Stat.*, **25**, 457–472.

Lee,K. and Mallick,B. (2004) Bayesian methods for variable selection in survival models with application to DNA microarray data. *Sankhya*, **66**, 756–778.

Lindley,D. (1957) A statistical paradox. *Biometrika*, **44**, 187–192.

Lindley,D. (1968) The choice of variables in multiple regression (with discussion). *J. R. Stat. Soc. Series B***34**, 31–66.

Li,H. and Gui,J. (2004) Partial Cox regression analysis for high-dimensional microarray gene expression data. *ISMB2004/Bioinformatics*, **20**, 208–215.

Madigan,D. and Raftery,A. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.*, **89**, 1535–1546.

Nguyen,D. and Rocke,D. (2002) Partial least squares proportional hazard regression for application to dna microarray survival data. *Bioinformatics*, **18**, 1625–1632.

Park,P. *et al*. (2002) Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**, 120–127.

Peduzzi,P. *et al*. (1980) A stepwise variable selection procedure for nonlinear regression models. *Biometrics*, **36**, 511–516.

Rosenwald,A. *et al*. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. *New Eng. J. Med.*, **346**, 1937–1946.

Sha,N. *et al*. (2004) Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, **60**, 812–819.

Tadesse,M. *et al*. (2005) Bayesian variable selection in clustering high-dimensional data. *J. Amer. Stat. Assoc.*, **100**, 602–617.

Tanner,T. and Wong,W. (1987) The calculation of posterior distributions by data augmentation. *J. Amer. Stat. Assoc.*, **82**, 528–549.

Tibshirani,R. (1997) The lasso method for variable selection in the Cox model. *Stat. Med.*, **16**, 385–395.

van't Veer,L. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Volinsky,C. *et al*. (1997) Bayesian model averaging in proportional hazard models: Assessing the risk of stroke. *App. Stat.*, **46**, 433–448.

Wei,L. (1992) The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat. Med.*, **11**, 1871–1879.