*University of Michigan School of Public Health*
The University of Michigan Department of Biostatistics Working Paper Series

*Year* 2006 *Paper* 62

# Doubly Penalized Buckley-James Method for Survival Data with High-Dimensional Covariates

Sijian Wang* Bin Nan†
Ji Zhu‡ David G. Beer**

*University of Michigan, sijwang@umich.edu
†University of Michigan, bnan@umich.edu
‡University of Michigan, jizhu@umich.edu
**University of Michigan, dgbeer@umich.edu

# Doubly Penalized Buckley-James Method for Survival Data with High-Dimensional Covariates

Sijian Wang, Bin Nan, Ji Zhu, and David G. Beer

**Abstract**

Recent interest in cancer research focuses on predicting patients' survival by investigating gene expression profiles based on microarray analysis. We propose a doubly penalized Buckley-James method for the semiparametric accelerated failure time model to relate high-dimensional genomic data to censored survival outcomes, which uses a mixture of $L_1$-norm and $L_2$-norm penalties. Similar to the elastic-net method for linear regression model with uncensored data, the proposed method performs automatic gene selection and parameter estimation, where highly correlated genes are able to be selected (or removed) together. The two-dimensional tuning parameter is determined by cross-validation and uniform design. The proposed method is evaluated by simulations and applied to the Michigan squamous cell lung carcinoma study.

# Doubly Penalized Buckley-James Method for Survival Data with High-Dimensional Covariates

Sijian Wang[1], Bin Nan[2], Ji Zhu[3], and David G. Beer[4]

[1,2] Department of Biostatistics
[3] Department of Statistics
[4] Departments of Surgery and Radiation Oncology
University of Michigan, Ann Arbor, MI 48109
[2] *email*: bnan@umich.edu

November 6, 2006

SUMMARY. Recent interest in cancer research focuses on predicting patients' survival by investigating gene expression profiles based on microarray analysis. We propose a doubly penalized Buckley-James method for the semiparametric accelerated failure time model to relate high-dimensional genomic data to censored survival outcomes, which uses a mixture of $L_1$-norm and $L_2$-norm penalties. Similar to the elastic-net method for a linear regression model with uncensored data, the proposed method performs automatic gene selection and parameter estimation, where highly correlated genes are able to be selected (or removed) together. The two-dimensional tuning parameter is determined by cross-validation and uniform design. The proposed method is evaluated by simulations and applied to the Michigan squamous cell lung carcinoma study.

KEY WORDS: Accelerated failure time model; Buckley-James method; Censored survival data; Elastic-net; High-dimensional covariate; Lung cancer; Microarray analysis; Variable selection.

# 1 Introduction

Microarray technologies, including cDNA and oligonucleotide arrays, simultaneously obtain thousands of gene expressions for each sample. Although a large number of genes are believed to be mostly inactive, there are many genes whose activities are associated with various physiological or environmental effects. An interesting and important task in analyzing human genomic data is to relate gene activities to phenotypic or clinical information.

The work of this article is motivated by the analysis of lung cancer using oligonucleotide arrays that initially involved the examination of lung adenocarcinomas (Beer et al., 2002), which has been more recently expanded to squamous cell carcinomas of the lung (Raponi et al., 2006). These tumors are strongly associated with tobacco use and along with adenocarcinomas account for the majority of non-small cell type lung cancer. Since histopathology is insufficient for prediction of disease progression and clinical outcomes in patients with both types of non-small cell type lung cancer, a goal of this study is to predict patients' survival utilizing gene expression data among 129 patients who presented with squamous cell carcinomas of the lung (Raponi et al., 2006). The RNA from each patient's tumor is examined using Affymetrix 133A microarrays containing over 22,000 probe sets. The patients are randomly divided into two groups: a training set with 65 patients and a test set with 64 patients. We want to select relevant genes from the training set and then use these genes to predict survival for patients in the test set.

In the past few years, there has been extensive research on applications of microarray data to cancer studies. Many investigators have developed methods to predict cancer classes using gene expression data, and demonstrated that analyzing microarray data can be very helpful and promising in cancer research, see e.g. Alon et al. (1999), Golub et al. (1999), Alizadeh et al. (2000), Garber et al. (2001), and Sorlie et al. (2001). There has also been active methodological research in relating gene expression profiles to censored survival

phenotypes. In addition to the challenge of high dimensionality of the gene expression data that all statistical methods need to deal with, another major challenge is the incomplete survival outcome due to limited follow-up time in such studies. While much work is based on the Cox model (e.g. Tibshirani, 1997; Li and Luan, 2003; Li and Gui, 2004; Li and Li, 2004; Gui and Li, 2005), other survival models have also be applied to the gene expression data. Among those, Ma, Kosorok and Fine (2006) studied the additive hazards model and Huang, Ma and Xie (2006) studied the accelerated failure time model.

In a classical regression setting, parameters of interest are often estimated by maximizing (or minimizing) an objective function, e.g. the partial likelihood function for the Cox model. Let $p$ be the number of genes and $n$ be the sample size of the training set. When $p > n$, regularization techniques are needed, in other words, the objective function needs to be penalized.

For censored survival data, Li and Luan (2003) investigate the $L_2$-norm penalized partial likelihood estimation based on the Cox model, where the penalty term is $\sum_{j=1}^{p} \beta_j^2$, $\beta$ is a $p$-dimensional parameter of interest (relative hazards in the Cox model). This method include all variables and does not provide a way of selecting a small set of relevant genes. Tibshirani (1997), Gui and Li (2005), and Park and Hastie (2006) propose the LASSO method that uses the $L_1$-norm penalty $\sum_{j=1}^{p} |\beta_j|$ to the partial likelihood function. Tibshirani (1997) uses the quadratic programming method for the optimization, Gui and Li (2005) use a modification of the LARS algorithm by Efron et al. (2004), and Park and Hastie (2006) propose the predictor-corrector algorithm for convex optimization that generalizes the LARS algorithm. However, the $L_1$-norm penalty suffers from two drawbacks (Zou and Hastie, 2005):

1. When there are several genes that share one biological pathway, it is possible that their expression levels are highly correlated. The $L_1$-norm penalty, however, can usually only select one gene. The ideal method should be able to automatically select the whole group of

relevant and yet highly correlated genes while eliminating trivial ones.

2. As shown in Rosset, Zhu, Hastie (2004), the $L_1$-norm penalty can select at most $n$, the sample size, input variables. But for microarray data, the sample size $n$ is usually on the order of 10s or 100s, while the number of attributes $p$ is typically on the order of 10,000s. So claiming that no more than $n$ genes are involved in a complicated biological process seems to be unrealistic for many biomedical studies. The ideal method should be able to select an arbitrary number of genes relevant to the clinical outcome.

On the other hand, for censored survival data, a linear regression model is a viable alternative to the Cox model, because it models failure time directly and thus has a simpler and more intuitive interpretation. Let $T_i$ be the random failure time and $X_i$ be the covariate vector for subject $i$, then

$$g(T_i) = \alpha + X_i'\beta + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $g$ is a pre-specified monotone function, $\epsilon_i$ is the error term with an unknown distribution that is assumed to have zero mean and bounded variance and be independent for all $i$. When $g(\cdot) = \log(\cdot)$, the above model is called the accelerate failure time (AFT), see e.g. Kalbfleisch and Prentice (2002).

When $T_i$ are subject to right censoring, Huang and Harrington (2005) applied the partial least squares (PLS) method based on the Buckley-James estimating equation to estimate the covariates' effects. But similar to the principle component approach, their method in fact involves all the genes for prediction and can not directly specify relevant genes that are associated with survival time. Huang et al. (2006) proposed a regularized method for the above linear model based on an inverse probability of censoring weighted loss function.

In this article, we propose a doubly penalized Buckley-James method for variable selection, parameter estimation and prediction for survival time using high-dimensional gene expression data. It extends the elastic-net regression for linear models developed by Zou

and Hastie (2005) to right censored survival data. It has several attractive features that make it a proper tool for analyzing microarray data with survival outcomes. First, it carries out variable selection and estimation simultaneously. It selects genes that are most relevant to the survival outcome, and excludes all other genes from the analysis. Secondly, it can select an arbitrary number of genes with non-zero coefficients, which is more flexible than using only the $L_1$-norm penalty. Thirdly, it automatically selects highly correlated genes altogether that are likely to be in the same biological pathway. This feature not only helps us possibly understand biological processes more clearly, but also very much improves the prediction performance. Furthermore, in contrast to the usual belief that the intercept $\alpha$ is not estimable, we conjecture that $\alpha$ can be consistently estimated by relaxing the commonly used assumption of bounded covariate support, which is supported by our simulation studies. Theoretical verification is still under exploration. We also introduce the uniform design method of Fang and Wang (1994) for selecting multi-dimensional tuning parameters, which is much more efficient than the simple grid search that is commonly used in the literature.

The rest of the article is organized as follows. In Section 2, we introduce the doubly penalized Buckley-James method for model (1), and discuss the intercept estimation. In Section 3, we introduce the method of selecting tuning parameters by using the uniform design. We present simulation studies in Section 4, and report the analysis of the Michigan squamous cell lung carcinoma study in Section 5. We provide a discussion in Section 6.

# 2 Doubly Penalized Buckley-James Method

## 2.1 Buckley-James Method

The linear model plays a fundamental role in statistical analysis. Even when $p < n$, the situation we consider in this subsection, however, the least squares approach can not be directly applied when the response variable is subject to censoring. In the past three decades,

many researchers (Miller, 1976; Buckley and James, 1979; Koul et al., 1981, among many others) extended the least-square principle in order to accommodate censoring. Later, the rank based estimating method drew great attention, see e.g. Tsiatis (1990) and Wei et al. (1990). Ritov (1990) establishes the equivalence of the Buckley-James method and the weighted ranked based method, and Tsiatis (1990), Ritov (1990), Lai and Ying (1991) and Ying (1993) provide the asymptotic properties of either the rank based estimator or the Buckley-James estimator. A nice summary can be found in Chapter 7 of Kalbfleisch and Prentice (2002). Wei (1992) discussed some advantages of the Buckley-James method over the Cox regression model, including simpler interpretation and better fits for some data sets.

For notational simplicity, here let $T_i$ denote the transformed failure time, e.g. the logarithm of the failure time. Then model (1) becomes

$$T_i = \alpha + X_i'\beta + \epsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

When $T_i$ is subject to right censoring, we can only observe $(Y_i, \delta_i, X_i)$, where $Y_i = \min(T_i, C_i)$, $C_i$ is the transformed censoring time by the same transformation for $T_i$, and $\delta_i = 1_{\{T_i \leq C_i\}}$ is the censoring indicator. Suppose we observe a random sample of $n$ observations $(Y_i, \delta_i, X_i)$, $i = 1, \ldots, n$.

If there is no censoring, the least squares method can be applied to estimate the parameters in model (2). For censored data, the key idea of the Buckley-James method is to recover those censored $T_i$ by their conditional expectations given corresponding censoring times and covariates. This is the same idea as the single imputation of Little and Rubin (2002). Define the "imputed" failure time $Y_i^*$ as

$$Y_i^* = Y_i\delta_i + E(T_i|T_i > Y_i, X_i)(1 - \delta_i). \tag{3}$$

Clearly $Y_i^* = T_i$ if $\delta_i = 1$ and $Y_i^* = E(T_i|T_i > Y_i, X_i)$ when $\delta_i = 0$. Absorbing the unknown

intercept $\alpha$ into $\epsilon_i$ in model (2) and set the new error term to be

$$\xi_i = \alpha + \epsilon_i = T_i - X_i'\beta,$$

with the true $\beta$, the quantity $E(T_i|T_i > Y_i, X_i)$ for a censored subject $i$ can be calculated by

$$\begin{aligned} E(T_i|T_i > Y_i, X_i) &= X_i'\beta + E(\xi_i|\xi_i > Y_i - X_i'\beta) \\ &= X_i'\beta + \int_{Y_i - X_i'\beta}^{\infty} \frac{tdF(t)}{1 - F(Y_i - X_i'\beta)}, \end{aligned} \tag{4}$$

where $F$ is the distribution function of $\xi = T - X'\beta$ in which the intercept is absorbed. That $X_i$ disappears from the conditional expectation of $\xi$ is due to a common assumption of independence between the error term and covariates in linear regression. Buckley and James (1979) substitute the above $F$ by its Kaplan-Meier estimator $\hat{F}$ in order to estimate $\beta$. Then the least squares method can be applied to the following regression model

$$Y_i^* = \alpha + X_i'\beta + \epsilon_i^*, \tag{5}$$

where $\epsilon_i^*$ are independent with zero mean.

Denote $Y^* = (Y_1^*, Y_2^*, \ldots, Y_n^*)'$, $X_i^* = X_i - \bar{X}$, where $\bar{X} = \sum_{i=1}^n X_i/n$, and $X^* = (X_1^*, X_2^*, \ldots, X_n^*)'$. Then the least squares estimator of $\beta$ in model (5) is

$$\hat{\beta} = (X^{*'}X^*)^{-1}X^{*'}Y^*. \tag{6}$$

The final solution requires an iterative procedure since $Y_i^*$ defined in (3) contain $\beta$. When the iterated algorithm converges, the intercept $\alpha$ can be estimated by $\hat{\alpha} = \bar{Y}^* - \sum_{i=1}^p \bar{X}_i'\hat{\beta}$, where $\bar{Y}^* = \sum_{i=1}^n Y_i^*/n$. Clearly whether $\alpha$ can be consistently estimated directly affects the prediction of survival time for additional independent samples.

## 2.2 Estimation of Intercept

Buckley and James (1979) claim that the intercept can not be estimated consistently due to the existence of censoring. In some of their simulations, however, Schneider and Weissfeld (1986) and Heller and Simonoff (1990) find that the intercept can be estimated quite

well using the Buckley-James method. Based on the work of Susarla and Ryzin (1980), we conjecture that the intercept can be consistently estimated when the supports of some covariates are not restricted to finite intervals, which seems suitable to the gene expression data. Detailed discussion is given below.

Let $\eta_i = C_i - X_i'\beta$. For the true $\beta$, it is clear from model (2) that the intercept $\alpha$ is the mean of survival time $\xi_i = T_i - X_i'\beta$ that is subject to right censoring. Hence $\alpha$ needs to be estimated from the "observed" data $(\xi_i \wedge \eta_i, \delta_i)$, $i = 1, \ldots, n$, here $\xi_i \wedge \eta_i = \min(\xi_i, \eta_i)$. We put quotes on the word *observed* because $\beta$ is actually unknown and thus $(e_i, \delta_i)$ are not really observed. Susarla and Ryzin (1980) provide a set of sufficient conditions for consistence and asymptotic normality of a mean survival time estimator, and this estimator is equivalent to the Buckley-James estimator shown by Susarla, Tsai and Ryzin (1984). The fundamental assumption of Susarla and Ryzin (1980), in our notation, is

$$\begin{aligned} \tau &= \sup\{t | t \text{ is in the support of } \xi\} \\ &\leq \sup\{t | t \text{ is in the support of } \eta\}. \end{aligned} \tag{7}$$

The above assumption may be violated if $\xi$ and $\eta$ are replaced by $T$ and $C$, respectively, because the maximum follow-up time for a biomedical study is often much shorter than the lifespan of study subjects. If the supports of some covariates with nonzero coefficients in model (2) are not bounded, then the supports of both $\xi$ and $\eta$ are dominated by the support of those covariates and hence (7) is satisfied. Special care is needed when $\beta$ is replaced by its estimator $\hat{\beta}$. The theoretical issues of estimating $\beta$ and $\alpha$ under the relaxed assumption on covariates will be discussed elsewhere. The results of the following simulation studies provide numerical evidence to support our conjecture.

Consider the following model

$$T = 2 + 1 \times X + \epsilon, \tag{8}$$

where $\epsilon \sim N(0.0.5^2)$. Four different settings of the support of $x$ are investigated. In the first model, $X \sim N(0, 1.96/3)$; in the second model, $X \sim U(-1, 1)$; in the third model $X \sim U(-0.5, 0.5)$; and in the fourth model, $X \sim U(-0.25, 0.25)$. The censoring distribution is $C \sim U(0, 4) \wedge V$, here $V$ is the truncation time. For the first two models, we tried four different $V$: 1, 1.5, 2, and 3. For last two models, we tried three different $V$: 1.5, 2, and 3. We drop V=1 because it yields a very high censoring rate that causes numerical trouble. For each setting, we simulated 1000 runs with a sample size of 500. The simulation results are summarized in Table 1.

The first setting corresponds to unbounded covariate support, and it is clearly seen that the bias of the intercept estimator is minimal even for a very short follow-up time. The bias of the intercept estimator exists in other three settings that have finite covariate supports, but is diminishing with wider covariate support and extended follow-up time. It suggests that the intercept estimator can be numerically satisfactory if covariates have wide support, even though the unbounded covariate support is required by theory. The bias for the slope parameter $\beta$ is minimal across all simulation settings.

## 2.3 Buckley-James Method with Double Penalization

In microarray data analysis, the number of covariates $p$ is usually much greater than the sample size $n$, and the classical Buckley-James method fails since the estimation for $\beta$ in model (5) is not unique. So some regularization is needed to obtain a stable estimator of $\beta$ with smaller prediction error. We propose a modified Buckley-James approach by using penalized least squares with both the $L_1$-norm and the $L_2$-norm penalty terms. To be specific, we consider the following minimization problem

$$\min_{\beta} \frac{1}{2} \sum_{i=1}^{n} (Y_i^* - X_i^{*\prime}\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2, \tag{9}$$

where $\lambda_1$ and $\lambda_2$ are prespecified constants and are called the tuning parameters in the machine learning field.

This type of regularization method with double penalties was originally developed by Zou and Hastie (2005) for linear models with uncensored data. They called it the elastic-net regression. By using the mixture of the $L_1$-norm and the $L_2$-norm penalties, it combines good features of the two. Similar to the regression with the $L_1$-norm penalty, the elastic-net method simultaneously performs automatic variable selection and continuous shrinkage. The added advantages by including a $L_2$-norm penalty are that groups of correlated variables now can be selected together and the number of selected variables is no longer limited by $n$. The proposed doubly penalized Buckley-James method extends these good features to the linear regression with censored data. Following are the major steps of the algorithm for a given pair of $(\lambda_1, \lambda_2)$.

*Algorithm.* Doubly Penalized Buckley-James method

1. Initialize $\beta = \beta^{(0)}$.

2. At the $m$-th iteration,

   (a) compute

$$Y_i^* = \delta_i Y_i + (1-\delta_i)\left\{ X_i'\beta^{(m-1)} + \int_{Y_i - X_i'\beta^{(m-1)}}^{\infty} \frac{t d\hat{F}(t)}{1-\hat{F}(Y_i - X_i'\beta^{(m-1)})} \right\};$$

   (b) compute $\beta^{(m)}$ by

$$\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}(Y_i^* - X_i^{*'}\beta)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2; \tag{10}$$

   (c) stop the iteration if $|\beta^{(m)} - \beta^{(k)}| < \epsilon$ for some $k \in \{0, 1, \ldots, m-1\}$, here $\epsilon$ is a prespecified precision.

3. When convergence is claimed, rescale $\hat{\beta}$ obtained from the last iteration to be $(1+\lambda_2)\hat{\beta}$, and compute $\hat{\alpha} = \bar{Y}^* - \sum_{i=1}^{p} \bar{X}_i' \hat{\beta}$.

The optimization in Step 2(b) is a standard elastic-net problem and can be carried out by the method of Zou and Hastie (2005). The stopping rule given in Step 2(c) considers possible oscillation among iterations, a common problem of the Buckley-James method due to the discontinuity feature of the estimating function for $\beta$ caused by the Kaplan-Meier estimator. When oscillation occurs, the current solution can be chosen as the final solution, see e.g. Huang and Harrington (2006). The final rescale step is very important. We can see in step 2(b) that $\beta$ is doubly shrunken by both the $L_1$-norm and the $L_2$-norm penalties. This double shrinkage actually introduces unnecessary extra bias comparing to using either the $L_1$-norm or the $L_2$-norm penalty only. Following Zou and Hastie (2005), we rescale $\hat{\beta}$ by multiplying the amplifying factor $1 + \lambda_2$.

Similar to the elastic-net method for the linear regression, the doubly penalized Buckly-James model can select correlated genes, and the number of selected genes can exceed the sample size.

# 3 Tuning Parameter Selection

In practice, it is important to select appropriate tuning parameters $\lambda_1$ and $\lambda_2$ in order to obtain a good prediction precision. It is anticipated that the computing cost for the optimization in Subsection 2.3 is high. Thus an efficient way of choosing $(\lambda_1, \lambda_2)$ is of particular interest, given that the Buckley-James method may involve a lot of iterations. A commonly used method is to specify a fine two-dimensional grid that covers a desirable wide range of $(\lambda_1, \lambda_2)$ uniformly, then use either a separate validation data set or cross-validation to search all the points on the grid for the optimal pair of $(\lambda_1, \lambda_2)$. But this equi-lattice search method is very inefficient. Another approach is to search the optimal $\lambda_1$ for a fixed

$\lambda_2$, and then search the optimal $\lambda_2$ by fixing $\lambda_1$ at the previously found point. This method is computationally more efficient, but it is very easy to miss the optimal pair of $(\lambda_1, \lambda_2)$ in the two-dimensional search region due to the nonuniform feature of the searched points.

We propose to use the uniform design approach of Fang and Wang (1994) to generate candidate points of $(\lambda_1, \lambda_2)$. The method actually works for a tuning parameter with arbitrary dimension. Let $D$ be the search region. Using the concept of discrepancy that measures uniformity on $D \subset R^d$ with arbitrary dimension $d$, $d = 2$ for our case, which is basically the Kolmogorov statistic for a uniform distribution on $D$, Fang and Wang (1994) point out that the discrepancy of the good lattice point set from a uniform design converges to zero with a rate of $O(m^{-1}(\log m)^d)$, here $m$ (a prime number) denotes the number of generated points on $D$. They also point out that the sequence of equi-lattice points on $D$ has a rate of $O(m^{-1/d})$ and the sequence of uniformly distributed random numbers on $D$ has a rate of $O(m^{-1/2}(\log\log \mathrm{m})^{1/2})$. Thus the uniform design has an optimal rate when $d \geq 2$. Fang and Wang (1994) provide useful tables of the uniform design in Appendix A, which makes it trivial to implement their method.

For uncensored data, cross-validation (CV) and generalized cross-validation (GCV) are commonly used to choose tuning parameters from a set of candidate points. For survival data, O'Sullivan (1988) and Nan et al. (2005) propose CV and GCV for choosing the smoothing parameter for the smoothing spline estimator. To choose the smoothing parameter, they suggest recording the derived dependent variables and covariates at the last iteration step and treating them as a linear regression problem. Then GCV for linear models can be applied to evaluate the current smoothing parameter. Following the same idea, we derive GCV for doubly penalized Buckly-James model.

Given a pair of $\lambda_1$ and $\lambda_2$, we fit the doubly penalized Buckley-James model. Let $\hat{\alpha}$, $\hat{\beta}$, and $Y_i^*$ be the values obtained in the last iteration. Assuming $\hat{\beta}_{s_1}, \ldots, \hat{\beta}_{s_m}$ are non-zero, and

other $\hat{\beta}_j$'s are all zero. Let $X_0$ be the matrix consisted of columns $s_1, \ldots, s_m$ in $X$, which are corresponding columns for non-zero $\hat{\beta}_j$'s. Denote $q = \text{Trace}(X_0(X_0'X_0 + \lambda_2 I)^{-1}X_0')$ that is discussed in Zou, Hastie and Tibshirani (2005), then GCV is given by

$$\text{GCV} = \sum_{i=1}^{n}(Y_i^* - \hat{\alpha} - X_i'\hat{\beta})^2/(n-q)^2.$$

We calculate GCV for each pair of $(\lambda_1, \lambda_2)$ determined by the uniform design, and then select the one that yields the smallest GCV.

# 4 Simulation Studies

In this section, we assess the group selection feature and the prediction performance of the proposed doubly penalized Buckley-James method by simulation studies.

## 4.1 Group Selection of Correlated Covariates

The purpose of this simulation is to show that the doubly penalized Buckley-James method tends to select highly correlated and predictable covariates in groups. We consider two examples with the same simulation settings studied by Zou and Hastie (2005) for censored survival data. For both examples, the logarithm of true survival time is simulated by

$$T = X\beta + \sigma\epsilon, \ \epsilon \sim N(0,1). \tag{11}$$

In the first example, we choose $\sigma = 15$ and generate 50 observations and 40 covariates for each simulated data set. The coefficients are set to be $\beta_j = 3$ when $1 \le j \le 15$ and $\beta_j = 0$ when $j \ge 16$. The covariates are generated as follows:

$$\begin{aligned}
\epsilon_j^x &\sim N(0, 0.01), \quad j = 1, \ldots, 15, \\
X_j &= Z_1 + \epsilon_j^x, \quad Z_1 \sim N(0,1), \quad j = 1, \ldots, 5 \\
X_j &= Z_2 + \epsilon_j^x, \quad Z_2 \sim N(0,1), \quad j = 6, \ldots, 10
\end{aligned}$$

$$
\begin{aligned}
X_j &= Z_3 + \epsilon_j^x, \quad Z_3 \sim N(0,1), \quad j = 11, \ldots, 15 \\
X_j &\sim N(0,1), \quad j = 16, \ldots, 40.
\end{aligned}
$$

The logarithm of censoring time $C$ is generated from a uniform distribution $U(-\tau, \tau)$, where $\tau$ is chosen to yield 50% censoring rate. The observed log transformed survival time is $Y = T \wedge C$.

The second example has the same setting as the first one except $p > n$. We now choose 100 observations with 120 covariates. Again only the first 15 elements of $\beta$ are equal to 3, and the rest are equal to 0.

In each example, we have three equally important groups that are related to survival time, and there are five covariates within each group. An ideal variable selection method would keep only the first 15 covariates and set the coefficients of all others to be 0.

For both examples, 200 runs are simulated. For each covariate, we evaluate the frequency of being selected among 200 simulation runs and the sample average of its coefficient. Due to space limitation, we only summarize the intercept and the first 15 slope parameters in Table 2. In the first example, 95 percent of the rest 25 variables are selected less than 41 times, and 95 percent of their estimates are within the range (-0.1495, 0.1159). In the second example, 95 percent of the rest 105 variables are selected less than 41 times, and 95 percent of their estimates are within the range (-0.0568, 0.0580). We see from this simple simulation study that the doubly penalized Buckley-James method is able to select highly correlated and predictable covariates in groups.

## 4.2 Comparisons to Other Regularization Methods

In this Subsection, we conduct a simulation study to compare the doubly penalized Buckley-James method to either the $L_1$-norm or the $L_2$-norm penalized Buckley-James method. The log transformed survival times are also generated from model (11). Log transformed

censoring times are generated from a uniform distribution that yields 50% censoring rate. Since the true survival time for each subject is available in simulated data, we use the relative prediction error (RPE) obtained from an independent test data set to evaluate the prediction performance, where $RPE \approx (1/n)\sum_{i=1}^{n}(T_i - X_i'\hat{\beta})^2/\sigma^2$ and $\hat{\beta}$ is obtained from the training data set.

For each of the following simulations, we generate an independent validation data set to choose tuning parameter(s). So in fact we generate three independent data sets for each simulation: a training set for model fitting, a validation set for tuning parameter selection, and a test set for RPE calculation. Their corresponding sample sizes are denoted as $(n_1, n_2, n_3)$.

Four examples are considered. The first two examples have the same settings as that in Tibshirani (1996) with an exception that we consider censored data here. The last two examples consider situations with several groups of correlated covariates.

Example 1 considers a few large effects with sample sizes (50,50,400) for the three data sets . We choose $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and $\sigma = 3$. The pairwise correlation between two predictors $X_{j_1}$ and $X_{j_2}$ is $\rho(j_1, j_2) = 0.5^{|j_1 - j_2|}$.

Example 2 considers many small effects with sample sizes (50,50,400). The only difference to method 1 is that $\beta_j = 0.85$ for all $j$.

Example 3 considers a group of highly correlated covariates in the case of $p > n$ with sample sizes (100, 100, 400). This example has the same setting as the second example in the previous subsection.

Example 4 considers two groups of moderate correlated covariates in the case of $p > n$ with sample sizes (100, 100, 400). As in Example 3, we also have 120 predictors. We choose $\sigma = 5$ and set the first six slope parameters to be $(3, 3, 2, 3, 3, 2)$ and all other 114 slope parameters to be zero. The first three covariates consist of a group and the next three

consist of another group. Within each group, the pairwise correlation between any two predictors $X_{j_1}$ and $X_{j_2}$ is 0.5. Unlike Example 3, the coefficients within each group are set to be unequal.

Due to the computing limitation, we conduct 200 simulation runs for Examples 1 and 2 and 50 simulation runs for Examples 3 and 4. The RPE values and corresponding standard deviations are listed in Table 3. We can see that in all examples, the doubly penalized Buckley-James method has not only the smallest RPE but also the smallest standard deviation.

# 5 Squamous Cell Lung Carcinoma Data Analysis

The goal of the Michigan squamous cell lung carcinoma study is to predict the survival of early-stage lung cancer patients using microarray gene expression data. The study has enrolled 129 subjects with squamous cell lung carcinoma. RNA samples are analyzed by using Affymetrix U133A microarray chips. Subjects are divided into a training set that has 65 subjects and a test set that has 64 subjects. We use the proposed doubly penalized Buckley-James method to select relevant genes from a linear model based on the training set, and then use the model to predict survival for subjects in the test set.

Gene expression values are log transformed. Those genes with very low expression levels or very small variabilities are excluded. This step is done by the Bioconductor package "genefilter". Then the rest of the genes are assessed by running univariate AFT models using the Buckley-James method, and 1000 genes with the smallest p-values are selected.

Starting with these 1000 genes, the AFT model fitted by the proposed doubly penalized Buckley-James method has selected 59 probe sets using the training set, see Table 4. Tuning parameters are determined by the method of uniform design and the generalized cross-validation described in Section 3. We start with 233 points in the region $[10, 200] \times [0.001, 100]$

for the tuning parameters $\lambda_1$ and $\lambda_2$, where $\lambda_2$ is uniformly spread using the uniform design on the log scale. The optimal pair of tuning parameters determined by the training set is $(\lambda_1, \lambda_2)_{opt} = (18.56, 9.54)$. Among those 59 probe sets, there are 4 duplicated genes and 5 anonymous probe sets.

The model with these selected 59 probe sets is then used to predict the survival times for subjects in the test set. A subject is assigned to the high risk group if the predicted survival time is less than 3 years, or to the low risk group if the predicted survival time is no less than 3 years. Kaplan-Meier curves for these two groups are plotted in the left panel of Figure 1. We can see that the two curves are separated well. The log rank test yields a p-value of 0.02.

We have also analyzed the data using the Cox model. Instead of fitting univariate AFT models by the Buckley-James method, we fit univariate Cox models to select 1000 genes to start with. We then fit a Cox model by using the doubly penalized partial likelihood with both the $L_1$-norm and the $L_2$-norm penalties, which minimizes the following objective function for $\beta$:

$$-\log \prod_{i=1}^{n} \left\{ \frac{\exp(X_i'\beta)}{\sum_{k=1}^{n} 1_{\{Y_k \geq Y_i\}} \exp(X_k'\beta)} \right\}^{\delta_i} + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \,.$$

An iterative approach is used for solving the above optimization problem. At each iteration, the partial likelihood is linearized and then the elastic-net method is applied.

The doubly penalized Cox model has selected 204 genes using the training set. The cumulative baseline hazard function is estimated by the Breslow estimator. Then survival probabilities for subjects in the training set are calculated and the risk score $X'\hat{\beta}$ that yields a 50% survival probability at 3 years is chosen to be the threshold for high/low risk groups. Kaplan-Meier curves for the two groups classified by such a threshold in the test set are plotted in the right panel of Figure 1. The p-value of the log rank test is 0.03.

From Figure 1 we see that the doubly penalized Buckley-James method uses less number of genes to yield a similar separation of the high/low risk groups in the test set. These two methods achieve agreements on 49 out of 64 subjects in terms of risk group assignments. Among the 59 genes selected by the doubly penalized Buckley-James method, 44 are also selected by the doubly penalized partial likelihood approach.

Several of the genes identified using the proposed method are consistent with prior analysis of survival-related genes in squamous cell carcinoma lung cancer. The increased expression of the tyrosine kinase FGFR2 was observed to be associated with better survival (Raponi et al., 2006), which is also demonstrated in this study. The biological basis for this relationship is not established however the role of fibroblast growth factor signaling is associated with normal lung development and the interaction between the epithelial and mesenchymal-derived cellular components of the lung (De Langhe et al., 2006). Loss or decreased expression of FGFR2 may allow lung squamous carcinoma cells to escape from this interaction and affect differentiated function or cell proliferation. In analysis of the other main type of non-small cell lung cancer namely lung adenocarcinomas, the increased expression of both KRT7 (Gharib et al., 2002) and the angiogenic molecule VEGF (Beer et al., 2002; Gharib et al., 2004) at the mRNA and protein levels were investigated and shown to be related to poor patient outcome. Both genes in the present study are also associated with increased expression and a reduced survival consistent with these earlier studies. Interestingly, increased expression of several of the other genes including DNA methyltransferase (DNMT3B), dynamin 2 (DNM2) and DNA polymerase delta (POLD3) are suggestive of more DNA replication and thus more highly proliferative tumors and observed in the present study to demonstrate increased expression in patient's tumors with reduced survival. Additional studies will be required to establish the direct relationships between the expression of these genes and tumor behavior in squamous cell carcinomas of the lung.

# 6 Discussion

A set of regularity conditions needs to be developed for the consistent estimation of the intercept parameter in the linear model for censored survival data. A relaxation of the requirement of bounded support for covariates will affect the existing asymptotic theory for the slope estimators developed by Tsiatis (1990), Ritov (1990), Lai and Ying (1991), and Ying (1993), and a uniform extension of Susarla and Van Ryzin (1980) is important for obtaining an intercept estimator with nice asymptotic features. All these theoretical issues are under investigation and will be presented elsewhere.

A possible alternative approach of estimating the slope parameters is to use the rank based estimating equations. When $p < n$, using Gehen weights yields a monotone rank based estimating function that is an important feature for developing sound numeric algorithms. Penalized method is needed for the situation that $p > n$, however. Then an interesting question would be: how to construct an objective function using the rank based approach, which allows utilizing the $L_1$-norm and the $L_2$-norm penalties and yet still can be optimized by a feasible numerical algorithm.

# References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expresion profiling. *Nature* **403,** 503-511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissures probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America* **96,** 6745-6750.

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B., and Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* **8,** 816-824.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika* **66,** 429-436.

Cox, D. R. (1972). Regression models and lifetables. *Journal of the Royal Statistical Society, Series B* **34,** 187-220.

De Langhe, S. P., Carraro, G., Warburton, D., Hajhosseini, M. K., and Bellusci, S. (2006). Levels of mesenchymal FGFR2 signaling modulate smooth muscle progenitor cell commitment in the lung. *De Biol* **299,** 52-62.

Efron, B., Johnston, I., Hastie, T., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32,** 407-499.

Fang, K.-T. and Wang, Y. (1994). *Number-Theoretic Methods in Statistics*. Chapman and Hall: London.

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., Van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D., and Petersen, I. (2001). Diversity of gene expression

in adenocarcinoma of the lung. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 13784-13789.

Gharib, T. G., Chen, G., Wang, H., Huang, C. C., Prescott, M. S., Sheddon, K. A., Misek, D. E., Thomas, D. G., Giordano, T. J., Taylor, J. M. G., Yee, J., Orringer, M. B., Hanash, S., and Beer, D. G. (2002). Proteomic analysis of cytokeratin isoforms associated with survival in lung adenocarcinoma. *Neoplasia* **4,** 440-448.

Gharib, T. G., Chen, G., Huang, C. C., Misek, D. E., Iannettoni, M. D., Orringer, M. B., Hanash, S., and Beer, D. G. (2004). Genomic and proteomic analyses of VEGF and IGFBP3 in lung adenocarcinomas. *Clinical Lung Cancer* **5,** 307-312.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531-537.

Gui, J. and Li, H. (2005). Penalized Cox Regression Analysis in the HighDimensional and Lowsample Size Settings, with Applications to Microarray Gene Expression Data. *Bioinformatics* **21,** 3001-3008.

Heller, G. and Simonoff, J. S. (1990). A comparison of Estimators for Regression with a Censored Response Variable. *Biometrika* **77,** 515-520.

Huang, J. and Harrington, D. (2002). Penalized partial likelihood regression for right censored data with bootstrap selection of the penalty parameter. *Biometrics* **58,** 781-791.

Huang, J. and Harrington, D. (2005). Iterative Partial Least Squares with Right-Censored Data Analysis: A Comparison to Other Dimension Reduction Techniques. *Biometrics*

**61,** 17-24.

Huang, J., Ma, S. and Xie, H. (2006). Regularized Estimation in the Accelerated Failure Time Model with High Dimensional Covariates. *Biometrics* **62,** 813-820.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. Hoboken, New Jersey: John Wiley & Sons.

Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *Annals of Statistics* **9,** 1276-1288.

Lai, T. L. and Ying, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *Annals of Statistics* **10,** 1370-1402.

Li, H. and Gui, J. (2004). Partial Cox regression analysis for highdimensional microarray gene expression data. *Bioinformatics* **20,** i208-i215.

Li, L. and Li, H. (2004). Dimension Reduction Methods for Microarrays with Application to Censored Survival Data. *Bioinformatics* **20,** 3406-3412.

Li, H. and Luan, Y. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium of Biocomputing* **8,** 65-76.

Little R. J. A. and Rubin D. B. (2002). *Statistical Analysis with Missing Data.* New Jersey: John Wiley & Sons.

Ma, S., Kosorok, M. R. and Fine, J. P. (2006). Additive risk models for survival data with high-dimensional covariates. *Biometrics* **62**, 202-210.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrika* **63,** 449-464.

Nan, B., Lin, X., Lisabeth, L. D. and Harlow, S. D. (2005). A varying-coefficient Cox model for the effect of age at a marker event on age at menopause. *Biometrics* **61**, 576-583.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing* **9,** 531-542.

Park, M. Y. and Hastie, T. (2006). An $L_1$ regularizationpath algorithm for generalized linear models. Technical report, Stanford University, 2006.

Raponi, M., Zhang, Y., Yu, J., Chen, G., Lee, G., Taylor, J.M.G., MacDonald, J., Thomas, D., Moskaluk, C., Wang, Y. and Beer, D. G. (2006). Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Research*, 66, 7466-7472.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Annals of Statistics* **18,** 303-328.

Rosset, S., Zhu, J., and Hastie, T. (2004). Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research* **5,** 941-973.

Schneider, H. and Weissfeld, L. (1986). Estimation in Linear Models with Censored Data. *Biometrika* **73,** 741-745.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., Van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Eystein Lonning, P., and Borresen-Dale A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 13784-13789.

Susarla, V. and Van Ryzin, J. (1990). Large Sample Theory for an Estimator of the Mean Survival Time from Censored Samples. *Annals of Statistics* **8,** 1002-1016.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal*

*Statistical Society, Series B* **58,** 267-288.

Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine* **16,** 385-395.

Tsiatis, A. A. (1990). Estimating regresion parameters using linear rank tests for censored data. *Annals of Statistics* **18,** 354-372.

Wei, L. J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* **11,** 1871-1879.

Wei, L. J., Ying, Z., and Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77,** 845-851.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics* **21,** 76-99.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67,** 301-320.

Zou, H., Hastie, T., and Tibshirani, R. (2005). On the degrees of freedom of the Lasso. Technical Report, Department of Statistics, Stanford University.

Table 1: Intercept and slope estimation for four univariate Buckley-James models with different covariate support.

| Truncation time | Censoring rate | Empirical mean (std. dev.) for $\hat{\alpha}$ | Empirical mean (std. dev.) for $\hat{\beta}$ |
|---|---|---|---|
| | | $X \sim N(0, 1.96/3)$ | |
| 1.0 | 0.90 | 1.972 (0.071) | 0.999 (0.084) |
| 1.5 | 0.79 | 1.987 (0.071) | 0.997 (0.084) |
| 2.0 | 0.67 | 1.995 (0.045) | 0.998 (0.063) |
| 3.0 | 0.52 | 1.999 (0.032) | 1.000 (0.045) |
| | | $X \sim U(-1, 1)$ | |
| 1.0 | 0.92 | 1.811 (0.026) | 1.018 (0.037) |
| 1.5 | 0.80 | 1.955 (0.067) | 1.001 (0.103) |
| 2.0 | 0.67 | 1.994 (0.041) | 1.003 (0.067) |
| 3.0 | 0.52 | 1.999 (0.031) | 1.001 (0.052) |
| | | $X \sim U(-0.5, 0.5)$ | |
| 1.5 | 0.86 | 1.800 (0.063) | 1.011 (0.173) |
| 2 | 0.69 | 1.957 (0.032) | 1.003 (0.118) |
| 3 | 0.51 | 1.999 (0.032) | 0.999 (0.100) |
| | | $X \sim U(-0.25, 0.25)$ | |
| 1.5 | 0.88 | 1.650 (0.059) | 1.006 (0.310) |
| 2 | 0.70 | 1.899 (0.031) | 1.010 (0.227) |
| 3 | 0.51 | 1.999 (0.029) | 1.005 (0.196) |

Table 2: Frequency of being selected by the doubly penalized Buckley-James method for the first 15 nonzero slopes and the summery statistics of their estimates from 200 simulation runs.

| | $p = 40$ | | $p = 120$ | |
|---|---|---|---|---|
| | Frequency | Empirical mean (Standard Deviation) | Frequency | Empirical mean (Standard Deviation) |
| $\alpha$ | — | 0.001 (2.873) | — | -0.008 (1.126) |
| $\beta_1$ | 195 | 2.746 (1.119) | 200 | 2.899 (0.589) |
| $\beta_2$ | 194 | 2.887 (1.477) | 200 | 2.868 (0.555) |
| $\beta_3$ | 196 | 2.689 (1.021) | 200 | 2.850 (0.568) |
| $\beta_4$ | 195 | 2.744 (1.036) | 200 | 2.869 (0.553) |
| $\beta_5$ | 196 | 2.803 (1.179) | 200 | 2.876 (0.567) |
| $\beta_6$ | 193 | 2.602 (0.966) | 200 | 2.874 (0.603) |
| $\beta_7$ | 195 | 2.650 (1.114) | 200 | 2.879 (0.575) |
| $\beta_8$ | 192 | 2.583 (1.196) | 200 | 2.873 (0.588) |
| $\beta_9$ | 196 | 2.683 (1.115) | 200 | 2.877 (0.584) |
| $\beta_{10}$ | 196 | 2.630 (1.185) | 200 | 2.862 (0.615) |
| $\beta_{11}$ | 197 | 2.706 (0.696) | 200 | 2.899 (0.611) |
| $\beta_{12}$ | 191 | 2.620 (1.011) | 200 | 2.875 (0.612) |
| $\beta_{13}$ | 195 | 2.788 (0.966) | 200 | 2.866 (0.602) |
| $\beta_{14}$ | 196 | 2.811 (0.791) | 200 | 2.897 (0.619) |
| $\beta_{15}$ | 197 | 2.688 (0.802) | 200 | 2.858 (0.604) |

Table 3: Comparison of different regularization methods in terms of average relative prediction error (empirical standard deviation) calculated from simulated test sets, where the Buckley-James method does not apply in Examples 3 and 4 due to $p > n$. BJ: Buckley-James; DP-BJ: Doubly penalized Buckley-James; $L_1$-BJ: $L_1$-norm penalized Buckley-James; $L_2$-BJ: $L_2$-norm penalized Buckley-James.

| Example | BJ | DP-BJ | $L_1$-BJ | $L_2$-BJ |
|---|---|---|---|---|
| 1 | 0.462 (0.279) | 0.280 (0.186) | 0.283 (0.196) | 0.342 (0.199) |
| 2 | 0.477 (0.338) | 0.230 (0.146) | 0.353 (0.193) | 0.256 (0.160) |
| 3 | — | 0.452 (0.189) | 0.501 (0.210) | 1.580 (0.281) |
| 4 | — | 0.200 (0.103) | 0.400 (0.187) | 0.722 (0.179) |

Table 4: Probe set ID, Gene symbol, and estimated coefficient for each of the 59 probe sets selected by the doubly penalized Buckley-James model based on 65 subjects in the training set.

| Probe set | Gene symbol | Coef. | Probe set | Gene symbol | Coef. |
|---|---|---|---|---|---|
| 218433_at | PANK3 | 0.624 | 219957_at | — | -0.130 |
| 209220_at | GPC3 | 0.543 | 210512_s_at | VEGF | -0.137 |
| 219128_at | FLJ20558 | 0.389 | 214791_at | LOC93349 | -0.139 |
| 211578_s_at | RPS6KB1 | 0.303 | 208862_s_at | CTNND1 | -0.142 |
| 203638_s_at | FGFR2 | 0.274 | 212080_at | MLL | -0.143 |
| 214190_x_at | GGA2 | 0.201 | 202005_at | ST14 | -0.181 |
| 211084_x_at | PRKD3 | 0.159 | 218245_at | TSK | -0.182 |
| 203895_at | PLCB4 | 0.119 | 204027_s_at | METTL1 | -0.187 |
| 203639_s_at | FGFR2 | 0.101 | 203040_s_at | HMBS | -0.193 |
| 208228_s_at | FGFR2 | 0.084 | 201003_x_at | — | -0.211 |
| 207551_s_at | MSL3L1 | 0.068 | 213240_s_at | KRT4 | -0.212 |
| 222099_s_at | C19orf13 | 0.012 | 212680_x_at | PPP1R14B | -0.226 |
| 201545_s_at | PABPN1 | -0.001 | 212076_at | MLL | -0.228 |
| 203082_at | BMS1L | -0.010 | 212836_at | POLD3 | -0.234 |
| 201613_s_at | AP1G2 | -0.013 | 201059_at | — | -0.255 |
| 219217_at | FLJ23441 | -0.013 | 204385_at | KYNU | -0.290 |
| 218810_at | FLJ23231 | -0.029 | 202978_s_at | ZF | -0.292 |
| 209457_at | DUSP5 | -0.043 | 209709_s_at | HMMR | -0.329 |
| 204218_at | DKFZP564M082 | -0.056 | 209446_s_at | — | -0.347 |
| 209016_s_at | KRT7 | -0.057 | 211240_x_at | CTNND1 | -0.380 |
| 217253_at | — | -0.080 | 202253_s_at | DNM2 | -0.406 |
| 203545_at | ALG8 | -0.080 | 217014_s_at | AZGP1 | -0.456 |
| 221989_at | RPL10 | -0.086 | 203431_s_at | RICS | -0.472 |
| 200747_s_at | NUMA1 | -0.093 | 51192_at | SSH3 | -0.486 |
| 203212_s_at | MTMR2 | -0.094 | 36552_at | DKFZP586P0123 | -0.510 |
| 219919_s_at | SSH3 | -0.102 | 219241_x_at | SSH3 | -0.683 |
| 220668_s_at | DNMT3B | -0.109 | 213700_s_at | PKM2 | -0.762 |
| 202887_s_at | DDIT4 | -0.118 | 218136_s_at | MSCP | -0.770 |
| 212669_at | CAMK2G | -0.119 | 202471_s_at | IDH3G | -0.914 |
| 212568_s_at | DLAT | -0.126 | | | |

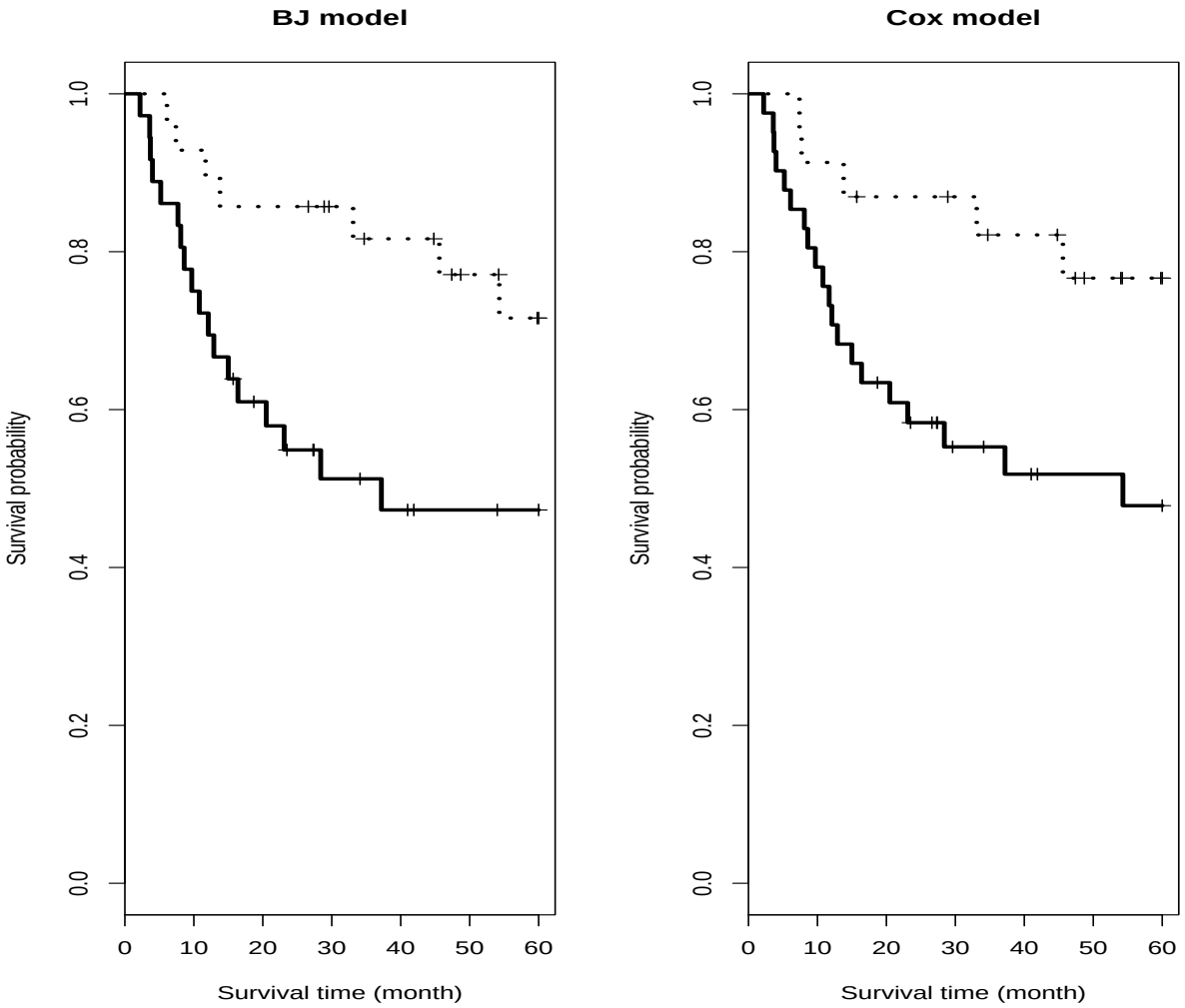


Figure 1: Lung cancer survival curves (Kaplan-Meier) of the test set high/low risk groups classified by the doubly penalized Buckley-James method and the doubly penalized partial likelihood method fitted from the training set: —— high risk group; **- - -** low risk group. Log rank p-value = 0.02 for the doubly penalized Buckley-James method; Log rank p-value = 0.03 for the doubly penalized partial likelihood method.