Biometrics 64, 132–140
March 2008

DOI: 10.1111/j.1541-0420.2007.00877.x

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

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

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

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 the elastic-net penalty that is a mixture of L_1- 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 generalized crossvalidation. 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 expression measurements 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 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 nonsmall-cell-type lung cancer. Because histopathology is insufficient for prediction of disease progression and clinical outcomes in patients with both types of nonsmall-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 U133A 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. 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 (AFT) model.

For censored survival data, Li and Luan (2003) investigated the L_2-norm penalized partial likelihood estimation based on the Cox model (Cox, 1972),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 includes 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) proposed the least absolute shrinkage and selection operator (LASSO)method that uses the L_1-norm penalty $\sum_{j=1}^{p} |\beta_j|$ to the partial likelihood function. Tibshirani (1997)

used the quadratic programming method for the optimization, Gui and Li (2005) used a modification of the least angle regression (LARS) algorithm by Efron et al. (2004), and Park and Hastie (2006) proposed 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):

(i) 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.

(ii) As shown in Rosset, Zhu, and 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 in the order of 10s or 100s, while the number of attributes p is typically in 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 prespecified monotone function, ϵ_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 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 (Buckley and James, 1979) to estimate the covariates' effects. But similar to the principle component approach, their method in fact involves all the genes for prediction and cannot 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 a 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. Secondly, it can select an arbitrary number of genes with nonzero coefficients, which is more flexible than using only the L_1-norm penalty. Thirdly, it automatically selects highly correlated genes together 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 α is not estimable, we conjecture that α 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.

2. Doubly Penalized Buckley–James Method

2.1 *Buckley–James Method*

The linear model plays a fundamental role in statistical analysis. In the past three decades, many researchers (Miller, 1976; Buckley and James, 1979; Koul, Susarla, and Van Ryzin 1981, among many others) extended the least-square principle in order to accommodate censoring of the response variable. Later, the rank-based estimating method drew great attention; see, e.g., Tsiatis (1990) and Wei, Ying, and Lin (1990). Ritov (1990) established the equivalence between the Buckley–James method and the weighted rank based method; Tsiatis (1990), Ritov (1990), Lai and Ying (1991), and Ying (1993) provided 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, δ_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 $\delta_i = 1_{\{T_i \le C_i\}}$ is the censoring indicator.

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

$$Y_i^* = \begin{cases} Y_i & \delta_i = 1, \\ E(T_i \,|\, T_i > Y_i, X_i) & \delta_i = 0. \end{cases} \tag{3}$$

Absorbing the unknown intercept α into ϵ_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 β, the quantity $E(T_i \,|\, T_i > Y_i, X_i)$ for a censored subject i can be calculated by

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

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 ξ is due to a common assumption of independence between the error term and covariates in linear

regression. Buckley and James (1979) substituted the above F by its Kaplan–Meier estimator $\hat{F}$ in order to estimate β. Then the least-squares method can be applied to the following regression model

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

where ϵ_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 β in model (5) is

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

The final solution requires an iterative procedure since values of Y_i^* defined in (3) contain β. When the iterated algorithm converges, the intercept α 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 α can be consistently estimated directly affects the prediction of survival time for additional independent samples.

2.2 *Estimation of Intercept*

Buckley and James (1979) claimed that the intercept cannot be estimated consistently due to the existence of censoring. In some of their simulations, however, Schneider and Weissfeld (1986) and Heller and Simonoff (1990) found that the intercept can be estimated quite well using the Buckley–James method. Based on the work of Susarla and Van Ryzin (1980) and of Susarla, Tsai, and Van Ryzin (1984), we conjecture that the intercept can be consistently estimated when the supports of some covariates are not restricted to finite intervals. Under such an assumption, the supports of $\eta = C - X'\beta$ and $\xi = T - X'\beta$ are equivalent, which is a sufficient condition of Susarla and Van Ryzin (1980) to obtain a consistent mean survival time (equivalent to intercept in our case for a fixed β) from censored samples. The assumption seems suitable to the gene expression data. The theoretical issues of estimating β and α 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

$$T = 2 + X + \epsilon, \tag{7}$$

where $\epsilon \sim N(0, 0.5^2)$. Four different settings of the support of X are investigated. In the first setting, $X \sim N(0, 1.96/3)$; in the second setting, $X \sim U(-1, 1)$; in the third setting $X \sim U(-0.5, 0.5)$; and in the fourth setting, $X \sim U(-0.25, 0.25)$. The censoring distribution is $C \sim U(0, 4) \wedge V$, here V is a truncation time. For the first two settings, we tried four different V: 1, 1.5, 2, and 3. For the last two settings, we tried three different V: 1.5, 2, and 3 because $V = 1$ yields a very high censoring rate that causes numerical instability. For each setting, we simulated 1000 runs with two different sample sizes: 50 and 500. For the case of sample size 50, we also drop $V = 1$ for all four settings due to the same aforementioned reason. 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 the 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. The bias for the slope parameter β is minimal across all simulation settings.

Table 1

Intercept and slope estimation for four univariate accelerated failure time models obtained by the Buckley–James method with different covariate support. The true intercept $\alpha = 2$ and the true slope $\beta = 1$. The empirical mean (standard deviation) for each of the two parameters is provided in the table.

Truncation time	Censoring rate	Sample size = 50		Sample size = 500	
		α	β	α	β
			$X \sim N(0, 1.96/3)$		
1.0	0.90	—	—	1.972 (0.071)	0.999 (0.084)
1.5	0.79	1.981 (0.271)	1.023 (0.318)	1.987 (0.071)	0.997 (0.084)
2.0	0.67	1.982 (0.132)	0.999 (0.205)	1.995 (0.045)	0.998 (0.063)
3.0	0.52	1.992 (0.100)	1.003 (0.145)	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	1.986 (0.310)	1.082 (0.435)	1.955 (0.067)	1.001 (0.103)
2.0	0.67	1.993 (0.150)	1.029 (0.247)	1.994 (0.041)	1.003 (0.067)
3.0	0.52	1.997 (0.099)	1.007 (0.170)	1.999 (0.031)	1.001 (0.052)
			$X \sim U(-0.5, 0.5)$		
1.5	0.86	1.803 (0.194)	1.059 (0.530)	1.800 (0.063)	1.011 (0.173)
2.0	0.69	1.941 (0.095)	1.002 (0.330)	1.957 (0.032)	1.003 (0.118)
3.0	0.51	1.995 (0.080)	0.989 (0.270)	1.999 (0.032)	0.999 (0.100)
			$X \sim U(-0.25, 0.25)$		
1.5	0.88	1.676 (0.180)	1.121 (1.027)	1.650 (0.059)	1.006 (0.310)
2.0	0.70	1.893 (0.080)	1.022 (0.618)	1.899 (0.031)	1.010 (0.227)
3.0	0.51	1.996 (0.075)	1.017 (0.563)	1.999 (0.029)	1.005 (0.196)

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. Some regularization is needed to obtain a stable estimator of β 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

$$\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}(Y_i^* - X_i^{*\prime}\beta)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2, \quad (8)$$

where λ_1 and λ_2 are called the tuning parameters in the machine learning field and will be determined by crossvalidation. In fact, every centered covariate is also scaled by its sample standard deviation in order to make the numerical implementation more stable.

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 the 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 (λ_1, λ_2).

Algorithm. Doubly Penalized Buckley–James method

1. Let $\beta^{(0)}$ be the initial value of β.
2. At the m-th iteration,

   (a) compute

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

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

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

   (c) stop the iteration if $|\beta^{(m)} - \beta^{(k)}| < d$ for some $k \in \{0, 1, \ldots, m-1\}$, here d 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 $\hat{\alpha} = \bar{Y}^* - \sum_{i=1}^{p}\bar{X}_i'\hat{\beta}$.

Note that in the mth iteration, all the remaining mass is placed at the last $Y_i - X_i'\beta^{(m-1)}$ when calculating the conditional expectation of residuals. 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 phenomenon for a discrete estimating function for which there is no clearly defined root except a region where the estimating function changes sign. Oscillation occurs when the numerical procedure reiterates among a few points in that region, so numerical convergence can be claimed and the current solution can be chosen as the final solution; see, e.g., Huang and Harrington (2005). Yu and Nan (2006) provided a detailed discussion on numerical convergence of the rank-based estimating method that has the similar problem as the Buckley–James method. Our experience is that the oscillation is rare when the sample size is large and the number of coefficients is small. It is not rare, however, for the simulation settings in this article. In the following Section 4.1, 40% of 200 simulation runs for both examples achieve numerical convergence without oscillation, 1.5% of the second example reach the prespecified maximum iteration number and the results of the final iterations are claimed as solutions, and all other simulation runs have oscillations. Among all the simulation runs with oscillations, the median and 95 percentile of the cycle size (excluding the starting and ending iterations of an oscillation cycle) are 3 and 14, respectively, for the first example and 6 and 59, respectively, for the second example. Within each oscillation cycle, we calculate the maximum of the maximum absolute differences of all values of coefficients obtained at all iterations in the cycle to that obtained at the ending iteration (the so-claimed solution). The median and 95 percentile of such quantity are 0.021 and 0.087, respectively, for the first example and 0.023 and 0.109, respectively, for the second example. As what we would expect, for almost all the oscillation cases in these simulations, the numerical values of coefficients within an oscillation cycle are indeed very close to each other and to the claimed solution comparing to the biases reported in Table 2.

The final rescale step in the algorithm is very important. We can see in Step 2(b) that β 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

Given a pair of λ_1 and λ_2, we fit model (2) by the proposed doubly penalized Buckley–James method. 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 nonzero, 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 nonzero $\hat{\beta}_j$'s. Denote $q = \mathrm{Trace}(X_0(X_0'X_0 + \lambda_2 I)^{-1}X_0')$ that is discussed in Zou, Hastie, and Tibshirani (2005). Then following the ideas of O'Sullivan (1988) and Nan et al. (2005) in choosing smoothing parameters for censored survival data, we define the following generalized crossvalidation (GCV):

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

Table 2

Selection frequency and estimation bias for the doubly penalized Buckley–James method. The summary statistics are based on 200 *simulation runs. Bias here refers to the absolute bias and the relative bias is calculated by using the absolute bias. Here* * *stands for the blocks of informative covariates* (*with nonzero coefficients*)*, and* ** *stands for the blocks of noninformative covariates* (*with zero coefficients*)*.*

Parameters	Frequency (%) (min, median, max)	Relative bias (%) (min, median, max)	Bias ($\lvert\cdot\rvert$) (min, median, max)
		$n = 50$, $p = 40$, $p_{-\text{nonzero}} = 15$	
α	—	—	0.103
Block1*	(93.0, 95.5, 98.0)	(8.33%, 9.71%, 20.24%)	—
Block2**	(16.0, 21.5, 24.5)	—	(0.003, 0.015, 0.064)
Block3*	(93.0, 95.5, 97.0)	(3.31%, 15.09%, 22.41%)	—
Block4**	(21.0, 22.5, 27.0)	—	(0.143, 0.162, 0.191)
Block5*	(91.5, 93.0, 96.0)	(11.26%, 18.61%, 23.71%)	—
Block6**	(23.0, 26.0, 28.0)	—	(0.117, 0.169, 0.201)
Block7**	(25.0, 28.0, 31.5)	—	(0.001, 0.028, 0.069)
		$n = 50$, $p = 120$, $p_{-\text{nonzero}} = 60$	
α	—	—	0.052
Block1*	(86.0, 91.0, 95.5)	(0.26%, 7.87%, 36.44%)	—
Block2*	(86.5, 90.0, 95.5)	(0.64%, 8.83%, 30.99%)	—
Block3**	(16.5, 25.0, 28.0)	—	(0.001, 0.045, 0.170)
Block4**	(14.5, 23.8, 30.0)	—	(0.003, 0.030, 0.082)

We calculate GCV for each pair of candidate (λ_1, λ_2) determined by the uniform design of Fang and Wang (1994), and then select the pair that yields the smallest GCV.

4. Simulation Studies

4.1 *Group Selection of Correlated Covariates*

We consider two examples with different settings, $p < n$ and $p > n$, for assessing the group selection feature of the proposed method. For the example with $p > n$, the number of nonzero coefficients is in fact greater then n. For both examples, the logarithm of true survival time is simulated by

$$T = X'\beta + \sigma\epsilon, \text{ where } \epsilon \sim N(0, 1). \tag{10}$$

In the first example, we have $n = 50$, $p = 40$, and $\sigma = 8$. There are seven blocks of covariates and each of the first six blocks contains five correlated covariates. Coefficients β_j's for $j \in \{1, \ldots, 5\} \cup \{11, \ldots, 15\} \cup \{21, \ldots, 25\}$ are nonzero and drawn randomly from $N(3, 0.5)$. Once these coefficients are drawn, their values are fixed for all the simulation runs. The other 25 β_j,s are set to be zero. The covariate matrix X is generated from a multivariate normal distribution with zero mean and covariance matrix as

$$\Sigma = \begin{pmatrix} \Sigma_0 & & & & & & \\ & \Sigma_0 & & & & & \\ & & \Sigma_0 & 0.2J & & & \\ & & 0.2J & \Sigma_0 & & & \\ & & & & \Sigma_0 & 0.2J & \\ & & & & 0.2J & \Sigma_0 & \\ & & & & & & \Sigma_1 \end{pmatrix},$$

where Σ_0 is a 5×5 matrix with diagonal elements to be 1 and off-diagonal elements to be 0.7, Σ_1 is a 10×10 identity matrix, and J is a 5×5 matrix with all elements to be 1. The logarithm of censoring time C is generated from a uniform distribution $U(-\tau, \tau)$, where τ is chosen to yield 50% censoring rate. The observed log-transformed survival time is $Y = T \wedge C$.

In the second example, 50 samples are also simulated from model (10), but with $p = 120$ and $\sigma = 15$. The first 60 coefficients are nonzero and drawn from $N(3, 0.5)$, and their values are then fixed for all simulation runs. The remaining 60 coefficients are set to be zero. The covariate matrix X is generated from a multivariate normal distribution with zero mean and covariance matrix as

$$\Sigma = \begin{pmatrix} \Sigma_0 & & & \\ & \Sigma_0 & 0.2J & \\ & 0.2J & \Sigma_0 & \\ & & & \Sigma_0 \end{pmatrix},$$

where Σ_0 is a 30×30 matrix with diagonal elements to be 1 and off-diagonal elements to be 0.7, and J is a 30×30 matrix with all elements to be 1. The censoring time is generated in the same way as in the first example.

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, and summarize the results in Table 2. We see from this simulation study that the doubly penalized Buckley–James method tends to select highly correlated informative covariates (with nonzero coefficients) in groups with very high selection frequencies, and meanwhile exclude noninformative covariates (with zero coefficients) with reasonable frequencies for the given sample size, even when the number of nonzero coefficients is greater than the sample size.

Table 3

Comparison of different regularization methods in terms of average RPE (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	n	p	$p_{-\text{nonzero}}$	BJ	DP-BJ	L_1-BJ	L_2-BJ
1	50	8	3	0.46 (0.28)	0.28 (0.19)	0.28 (0.20)	0.34 (0.20)
2	50	8	8	0.48 (0.34)	0.23 (0.15)	0.35 (0.19)	0.26 (0.16)
3	100	120	6	—	0.45 (0.19)	0.50 (0.21)	1.58 (0.28)
4	50	120	60	—	0.63 (0.28)	3.08 (0.96)	0.86 (0.68)

4.2 *Comparisons to Other Regularization Methods*

In this section, we compare the doubly penalized Buckley–James method to either the L_1-norm or the L_2-norm penalized Buckley–James method. Log-transformed survival times are also generated from model (10). 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 $\text{RPE} \approx (1/n)\sum_{i=1}^{n}(T_i - \hat{\alpha} - X_i'\hat{\beta})^2/\sigma^2$ 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). In practice, however, data are expensive and one would want to use all data for both model fitting and tuning parameter selection.

Four examples are considered here in this section. 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 of $p > n$ with several groups of correlated covariates, and in the last example, the number of nonzero coefficients is greater than n.

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 two groups of moderate correlated covariates in the case of $p > n$ with sample sizes (100, 100, 400) and $p = 120$. 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.

Example 4 has the same simulation setting as the second example in Section 4.1 with sample sizes (50, 50, 400).

We conduct 200 simulations for Examples 1 and 2 and 50 simulations 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. 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" on the training set. Then the rest of the genes are assessed by running univariate AFT models using the Buckley–James method, again on the training set, 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 from the training data set has selected 59 probe sets using the training set, see Table 4. Among those 59 probe sets, there are four duplicated genes and five anonymous probe sets. Tuning parameters (λ_1, λ_2) are determined by uniform design and generalized crossvalidation described in Section 3. We start with 233 points in the region $[10, 200] \times [0.001, 100]$ for (λ_1, λ_2), where λ_2 is uniformly spread on the log scale. The optimal pair of (λ_1, λ_2) determined by the training set is (18.56, 9.54).

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 otherwise. 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 β:

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			

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

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 probe sets 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 a smaller number of genes to yield a similar separation of the high-/low-risk groups to the Cox model in the test set. These two methods achieve agreements on 49 out of 64 subjects in terms of risk group assignments. Among the 15 disconcordant assignments, 5 subjects are classified to be high risk by the Buckley–James method and low risk by the Cox model. Among the 59 probe sets selected by the proposed method, 44 are also selected by the Cox model.

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 based on a subset of the data in Raponi et al. (2006). 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 nonsmall-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

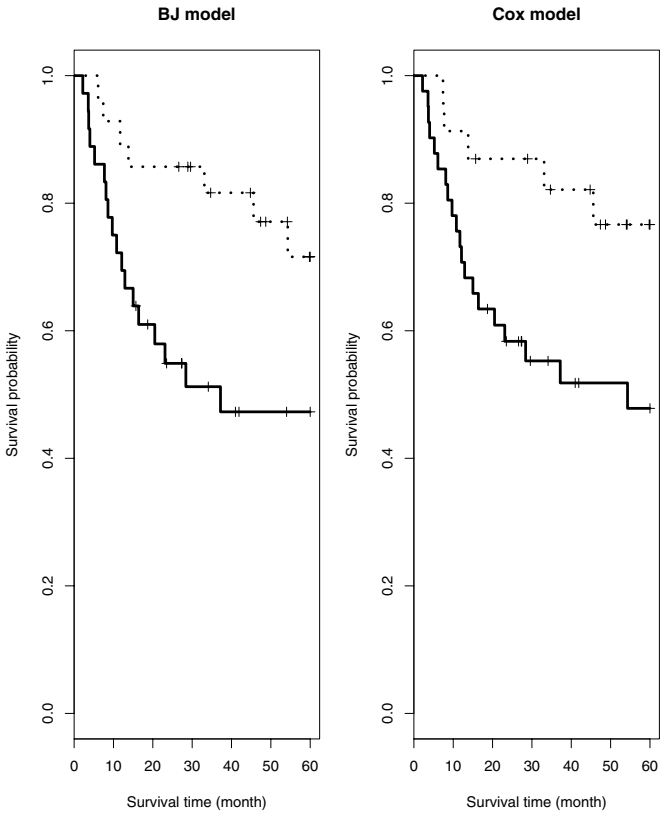

Figure 1. Lung cancer survival curves (Kaplan–Meier) of the test set high-/low-risk groups classified by the doubly penalized Buckley–James method and the doubly penalized partial likelihood method fitted from the training set: —— high-risk group; **- - -** low-risk group. Log rank p-value = 0.02 for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

of several of the other genes including DNA methyltransferase (DNMT3B), dynamin 2 (DNM2), and DNA polymerase delta (POLD3) are suggestive of more DNA replications and thus more highly proliferative tumors and are 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. A 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- and the L_2-norm penalties and yet still can be optimized by a feasible numerical algorithm.

Gene pre-filtering is a common practice in analyzing microarray data. As a reviewer pointed out, this step would affect the final gene list. Ideally the pre-filtering would be implemented in each iteration of a standard crossvalidation procedure. In this article, however, we used the GCV approach to reduce the computing cost, and there is no data elimination involved, thus the iterative pre-filtering becomes unnecessary. Note that our pre-filtering stage is completely based on the training data, which should yield an honest prediction to the test data. Understanding the effect of pre-filtering is an interesting problem and clearly deserves further investigation.

Acknowledgements

The authors are grateful for the very helpful suggestions of Professor Naisyin Wang, the associate editor, and two referees.

References

Beer, D. G., Kardia, S. L., Huang, C. C., et al. (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*. London: Chapman and Hall.

Gharib, T. G., Chen, G., Wang, H., et al. (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.

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

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.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken, 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 crossvalidation. *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, Department of Statistics, Stanford University.

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.

Susarla, V. and Van Ryzin, J. (1980). Large sample theory for an estimator of the mean survival time from censored samples. *Annals of Statistics* **8,** 1002–1016.

Susarla, V., Tsai, W. Y., and Van Ryzin, J. (1984). A Buckley–James-type estimator for the mean with censored data. *Biometrika* **71,** 624–625.

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 regression 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.

Yu, M. and Nan, B. (2006). A hybrid Newton-type method for censored survival data using double weights in linear models. *Lifetime Data Analysis* **12,** 345–364.

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.

Received November 2006. *Revised April* 2007.
Accepted May 2007.