Title

Penalized Cox Regression Analysis in the High-Dimensional and Low-sample Size Settings, with Applications to Mi-croarray Gene Expression Data

Authors

Gui, Jiang
Li, Hongzhe

Publication Date

2004-03-01

Penalized Cox Regression Analysis in the High-Dimensional and Low-sample Size Settings, with Applications to Microarray Gene Expression Data

Jiang Gui and Hongzhe Li

Department of Statistics and Rowe Program in Human Genetics, University of California Davis School of Medicine, Davis, CA 95616, USA

Email: hli@ucdavis.edu; jgui@wald.ucdavis.edu

Running title: Penalized Cox Regression Analysis for Microarray Data

Address for correspondence:

Hongzhe Li, Ph.D.

Rowe Program in Human Genetics

University of California Davis School of Medicine

Davis, CA 95616, USA

Tel: (530) 754-9234; Fax: (530) 754-6015

E-mail:hli@ucdavis.edu

ABSTRACT

An important application of microarray technology is to relate gene expression profiles to various clinical phenotypes of patients. Success has been demonstrated in molecular classification of cancer in which the gene expression data serve as predictors and different types of cancer serve as a categorical outcome variable. However, there has been less research in linking gene expression profiles to the censored survival data such as patients' overall survival time or time to cancer relapse. Due to large variability in time to certain clinical event among patients, studying possibly censored survival phenotypes can be more informative than treating the phenotypes as categorical variables. We propose to use the L_1 penalized estimation for the Cox model to select genes that are relevant to patients' survival and to build a predictive model for future prediction. The computational difficulty associated with the estimation in the high-dimensional and low-sample size settings can be efficiently solved by using the latest developed least angle regression method. Results from our simulation studies and application to real data set on predicting survival after chemotherapy for patients with diffuse large B-cell lymphoma demonstrate that the proposed procedure, which we call the LARS-Lasso procedure, can be used for identifying important genes that are related to time to death due to cancer and for building a parsimonious model for predicting the survival of future patients. The LARS-Lasso regression gives much better predictive performance than the L_2 penalized regression or dimension-reduction based methods such as the partial Cox regression method.

Keywords: penalized estimation, least angle regression, microarray gene expression, censored survival data, Lasso.

INTRODUCTION

DNA microarray technology permits simultaneous measurements of expression levels for thousands of genes, which offers the possibility of a powerful, genome-wide approach to the genetic basis of different types of tumors. The genome-wide expression profiles can be used for molecular classification of cancers, for studying varying levels of drug responses in the area of pharmacogenomics and for predicting different patients' clinical outcomes. The problem of cancer class prediction using the gene expression data, which can be formulated as predicting binary or multi-category outcomes, has been studied extensively and has been demonstrated great promise in recent years (Alon *et al.*, 1999; Golub *et al.*, 199; Alizadeh *et al.*, 2000; Garber *et al.*, 2001; Sorlie *et al.*, 2001). However, there has been less development in relating gene expression profiles to other phenotypes, such as quantitative continuous phenotypes or censored survival phenotypes such as time to cancer recurrence or time to death. Due to large variability in time to certain clinical event such as cancer recurrence among cancer patients, studying possibly censored survival phenotypes can be more informative than treating the phenotypes as binary or categorical variables.

The Cox regression model (Cox, 1972) is the most popular method in regression analysis for censored survival data. However, due to the very high dimensional space of the predictors, i.e., the genes with expression levels measured by microarray experiments, the standard maximum Cox partial likelihood method cannot be applied directly to obtain the parameter estimates. Besides the high-dimensionality, the genes expression levels of some genes are often highly correlated, which creates the problem of high co-linearity. To deal with the problem of collinearity, the most popular approach is to use the penalized partial likelihood, including both the L_2 penalized estimation, which is often called the ridge regression, and the L_1 penalized estimation, which was proposed by Tibshirani (1995) and is called the least absolute shrinkage and selection operator (Lasso) estimation. Such Lasso procedure minimizes the negative log partial likelihood subject to the sum of the absolute value of the coefficients being less than a constant, s. Comparing to the L_2 penalized procedure with constraints on the sum of the square of the coefficients, the Lasso procedure provides method for variable selection. These penalized procedures have been investigated mainly in the setting where the sample size is greater than the number of predictors. Li and Luan (2003) was the first to investigate the L_2 penalized estimation of the Cox model in the high-dimensional low-sample size settings and applied their method to relate the gene expression profile to survival data. To avoid the inversion of large matrix, they used the kernel tricks to reduce the computation

to involving only inversion of matrix of the size of the sample size. They demonstrated that the such procedure can be applied to build a predictive model for predicting the patients's future survival times.

One limitation of the L_2 penalized estimation of the Cox model as presented in Li and Luan (2003) is that it uses all the genes in the prediction and does not provide a way of selecting relevant genes for prediction. However, from biological point of view, one should expect that only a small subset of the genes is relevant to predicting the phenotypes. Including all the genes in the predictive model introduces noises and is expected to lead to poor predictive performance. Due to the high-dimensionality, the standard variable selection methods such as stepwise and backward selection cannot be applied. Tibshirani (1997) further extended the Lasso procedure for variable selection for the Cox proportional hazard models and proposed to use the quadratic programming procedure for maximizing the L_1 penalized partial likelihood in order to obtain the parameter estimates. However, such quadratic programming procedure cannot be applied directly to the settings when the sample size is much smaller than the number of potential predictors, such as in the setting of microarray data analysis.

Recently, Efron *et al.* (2004) proposed the least angle regression (LARS) procedure for variable selection in the linear regression setting. The LARS selects predictor by its current correlation or angle with the response, where the current correlation is defined correlation between the predictor and the current residuals. If the active set is defined as the set of indices corresponding to covariates with the greatest absolute current correlations, as the constraint constant s increases, the predictors are chosen one by one without deletion into the active set. The special feature of LARS is that before a new predictor is chosen to the active set as s increases, the corresponding increment of the coefficients only depends on all predictors in the active set. Efron *et al.* (2004) further pointed out the link between LARS and Lasso, showing that LARS can be modified to provide solution for Lasso. Instead of solving Lasso discretely by quadratic programming, modified LARS can give the whole solution path of all predictors. With this powerful algorithm, Lasso can be extended to perform subset selection in the high-dimension and low-sample settings. We propose in this paper to use LARS algorithm to obtain the solutions for the Cox model with L_1 penalty in the setting of very high dimensional covariates such as the gene expression data obtained by microarrays. We call such estimation procedure the LARS-Lasso procedure.

The rest of the paper is organized as follows. We first present the model and briefly review the Lasso estimation of the regression coefficients and present a modified LARS procedure for the Lasso estimation. We then evaluate the LARS-Lasso procedure by simulation studies and

applications to real data set of diffuse large B-cell lymphoma (DLBCL) survival times and gene expression data (Rosenwald *et al.*, 2002). Comparisons of results with methods proposed previously by using simulations and analysis of real data set of patients with DLBCL are also presented. Finally, we give a brief discussion of the methods and conclusions.

STATISTICAL MODELS AND METHODS

Cox proportional hazards model and Lasso estimation

Suppose that we have a sample size of n from which to estimate the relationship between the survival time and the gene expression levels $X_1, \cdots, X_p$ of p genes. Due to censoring, for $i = 1, \cdots, n$, the ith datum in the sample is denoted by $(t_i, \delta_i, x_{i1}, x_{i2}, \cdots, x_{ip})$, where δ_i is the censoring indicator and t_i is the survival time if $\delta_i = 1$ or censoring time if $\delta_i = 0$, and $x_i = \{x_{i1}, x_{i2}, \cdots, x_{ip}\}^{'}$ is the vector of the gene expression level of p genes for the ith sample. Our aim is to build the following Cox regression model for the hazard of cancer recurrence or death at time t

$$\begin{aligned} \lambda(t) &= \lambda_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \\ &= \lambda_0(t)\exp(\beta^{'} X), \end{aligned} \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\beta = \{\beta_1, \cdots, \beta_p\}$ is the vector of the regression coefficients, and $X = \{X_1, \cdots, X_p\}$ is the vector of gene expression levels with the corresponding sample values of $x_i = \{x_{i1}, \cdots, x_{ip}\}$ for the ith sample. We define $f(X) = \beta^{'} X$ to be the linear risk score function.

Based on the available sample data, the Cox's partial likelihood (Cox, 1972) can be written as

$$L(\beta) = \prod_{r \in D} \frac{\exp(\beta^{'} x_r)}{\sum_{j \in R_r} \exp(\beta^{'} x_j)},$$

where D is the set of indices of the events (e.g., deaths) and R_r denotes the set of indices of the individuals at risk at time $t_r - 0$. Let $l(\beta) = \log L(\beta)$, then the Lasso estimate of β (Tibshirani, 1995, 1997) can be expressed as

$$\hat{\beta}(s) = \text{argmax } l(\beta), \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq s,$$

where s is a tuning parameter determining how many covariates with coefficients being zero.

Tibshirani (1997) proposed the following iterative procedure to reformulate this optimization problem with constraint as a Lasso problem for linear regression models. Specifically, let

$\eta = \beta' X$, $\mu = \partial l/\partial \eta$, $A = -\partial^2 l/\partial \eta \eta^T$ and $z = \eta + A^- \mu$. Here since the sum of all elements in each row (or column) of the matrix A is 0, A is clearly a singular matrix. We can however use the generalized inverse. Alternatively, Tibshirani proposed to replace the information matrix A with a diagonal matrix D, which has the same diagonal elements as A. However, in most of applications, n is usually small and calculation of the generalized inverse is computationally feasible. In addition, due the high-dimensionality of the predictors, it is important to make the algorithm as accurate as possible. With this reparameterization, a one-term Taylor series expansion for $l(\beta)$ has the form of

$$(z - \eta)^T A (z - \eta).$$

Although there are multiple choices of A^-, it is easy to show that if $rank(A) = n - 1$, for any A^- that satisfies $AA^-A = A$ and $z = \eta + A^- \mu$, $(z - \eta)^T A (z - \eta)$ is invariant to the choice of the generalized inverse of A. To show this, let $C_i = \{k : i \in R_k\}$ denote the risk sets containing individuals i, and $C_{ii'} = \{k : i, i' \in R_k\}$ denote the risk sets containing individuals i and i'. Define

$$B = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix},$$

then it is easy to verify that $BAB' = \begin{pmatrix} A_{(n-1)\times(n-1)} & 0 \\ 0 & 0 \end{pmatrix}$ and $(B\mu)' = (\mu_1, \cdots, \mu_{n-1}, 0)$. If $rank(A) = n - 1$, then $A^- = B'CB$, where

$$C = \begin{pmatrix} A^{-1}_{(n-1)\times(n-1)} & * \\ * & * \end{pmatrix},$$

and * is used to represent any values. Therefore,

$$(z - \eta)^T A (z - \eta) = \mu' A^- \mu = (B\mu)' C B \mu = \mu'_{-n} A^{-1}_{(n-1)\times(n-1)} \mu_{-n},$$

where $\mu'_{-n} = (\mu_1, \cdots, \mu_{n-1})$.

The iterative procedure of Tibshirani (1997) involves the following four steps,

1. Fix s and initialize $\hat{\beta} = 0$.

2. Compute η, μ, A and z based on the current value of $\hat{\beta}$.

3. Minimize $(z - \beta' X)^T A (z - \beta' X)$ subject to $\sum |\beta_j| \leq s$.

4. Repeat step 2 and 3 until $\hat{\beta}$ does not change.

Tibshrani (1997) proposed to use the quadratic programming for solving Step 3. However, in the high-dimension and low-sample size setting, i.e., in the case when $p >> n$, the quadratic programming algorithm cannot be directly applied. We propose in the next section a simple modification of the LARS algorithm of Efron *et al.* (2004) for Step 3.

LARS-Lasso procedure: a modification of LARS for solving Lasso

The LARS algorithm (Efron *et al.*, 2004) is a new model selection algorithm developed for linear regression model. The algorithm is a less greedy version of traditional forward selection methods. One of the main advantages of LARS is its computational efficiency. Efron *et al.* (2004) also provided a simple modification of the LARS in order to obtain all Lasso solutions. We propose to apply a modified LARS algorithm for solving Step 3 of the iterative procedure presented in last section. First, we apply the Choleski decomposition to obtain $T = A^{1/2}$ such that $T'T = A$, then Step 3 of the iterative procedure presented in the previous section can be rewritten as

Step 3: minimize $(y - \beta' \hat{X})^T (y - \beta' \hat{X})$ subject to $\sum |\beta_j| \leq s$,

where $y = Tz$ and $\hat{X} = TX$. The original LARS procedure requires pre-processing data by centering the response to have mean 0 and standardizing the covariates to have mean 0 and unit length. By standardizing the covariates, the LARS algorithm can be performed based only on correlation calculations. However, the algorithm still works when the predictors are not scaled. In this case, we can modify the original correlation-based LARS procedure to select those variables having the largest absolute inner product between the predictor $\hat{X}$ and the current residuals of y (we call the current inner product for the rest of the paper) instead of the largest absolute current correlation, where the current residual is defined as $y - \beta' \hat{X}$ evaluated at the current estimate of β. We call the combined procedure the LARS-Lasso procedure, which is computationally thrifty.

To determine the value of the tuning parameter s or the number of genes to be used in the final model, one can choose s which minimizes the cross-validated partial likelihood (CVPL) (Verwij and Van Houwelingen, 1993; Huang and Harrington, 2002), which is defined as

$$CVPL(s) = -\frac{1}{n} \sum_{i=1}^{n} \left[l(\hat{f}^{(-i)}(s)) - l^{(-i)}(\hat{f}^{(-i)}(s)) \right],$$

where $\hat{f}^{(-i)}(s)$ is the estimate of the score function based on the LARS-Lasso procedure with tuning parameter s from the data without the ith subject. The terms $l(f)$ and $l^{(-i)}(f)$ are the log partial likelihoods with all the subjects and without the ith subject, respectively. The optimal value of s is chosen to maximize the sum of the contributions of each subject to the log partial likelihood. This CVPL is a special case of a more general cross-validated likelihood approach for model selections (Smyth, 2001; Van Der Laan *et al.*, 2003) and has been demonstrated to perform well in prediction in the context of the penalized Cox regression (Huang and Harrington, 2002).

Evaluation of the predictive performance: the time dependent ROC curves and area under the curves

In order to assess how well the model predicts the outcome, we propose to employ the idea of time dependent receiver-operator characteristics (ROC) curve for censored data and area under the curve (AUC) as our criteria. These methods were recently developed by Heagerty *et al.* (2000) in the context of the medical diagnosis. For a given score function $f(X)$, we can define time dependent sensitivity and specificity functions as

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

and define the corresponding $\text{ROC}(t|f(X))$ curve for any time t as the plot of $\text{sensitivity}(c, t|f(X))$ vs $1 - \text{specificity}(c, t|f(X))$ with cutoff point c varying, and the AUC as the area under the $\text{ROC}(t|f(X))$ curve, denoted by $\text{AUC}(t|f(X))$. Here $\delta(t)$ is the event indicator at time t. A nearest neighbor estimator for the bivariate distribution function is used for estimating these conditional probabilities accounting for possible censoring (Akritas, 1994). Note that larger AUC at time t based on a score function $f(X)$ indicates better predictability of time to event at time t as measured by sensitivity and specificity evaluated at time t. In our application presented in the next section, we study several different methods of constructing the score function $f(X)$ in the Cox model (1) and compare their predictive performance based on the AUCs.

EVALUATION OF THE METHODS BY SIMULATION STUDIES

We performed simulation studies to evaluate how well the LARS-Lasso procedure performs in the high-dimensional and low-sample size settings. We focus on whether the important covariates that are related to survival endpoints can be selected by the LARS-Lasso procedure and how well the model can be used for predicting the survival time for future patients.

In our simulation studies, we assume that 20 out of a total of 500 genes are related to time to cancer recurrence through a Cox regression model with 10 coefficients generated from an uniform U(-1,-0.1) distribution and 10 coefficients generated from an uniform U(0.1,1) distribution (see first column of Table 1 for the coefficients generated). A Weibull distribution with the shape parameter of 5 and the scale parameter of 2 is used for the baseline hazard function, and a uniform U(2,10) is used for simulating the censoring times. Based on this setting, we would expect about 40% censoring.

In order to generate gene expression data for 500 predictors (genes), we first generate an 100×500 dataset X from an uniform U(-1.5, 1.5) distribution. We assume that the first 20 genes with expression levels $X_1, X_2, \cdots, X_{20}$ are related to patient's risk cancer recurrence through a Cox model. In order to generate gene expression data for the rest of 480 genes which are not related to the survival, we first use Gram-Schmidt orthonormalization to construct its normalized orthogonal basis $\{\alpha_1, \cdots, \alpha_{20}, \beta_1, \cdots, \beta_{80}\}$, where $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_{20}\}$ is an orthogonal basis of the linear space A expanded by $X_1, X_2, \cdots, X_{20}$ and $\beta = \{\beta_1, \beta_2, \cdots, \beta_{80}\}$ is a set of orthogonal basis of B, which is the orthogonal complement space of A. By Cauchy's inequality, it is easy to show that if $\{\alpha_1, \cdots, \alpha_{20}, \beta_1, \cdots, \beta_{80}\}$ is a set of normalized orthogonal basis, then for any 20×80 matrix T, we have $corr(\alpha y, (\beta + \alpha T)x) \leq \lambda/\sqrt{1+\lambda^2}$, for $\forall x \in R^{80}, y \in R^{20}$, where λ^2 is the largest eigenvalue of $T'T$. Based on this result, we can generate the expression levels of genes which are unrelated to survival from the linear space $C = \{\beta + \alpha T\}$ with appropriate choice of the maximum eigenvalue of $T'T$ in order to control the maximum correlation between vectors in space A and C. We considered the maximum possible correlation of 0, 0.71, 0.82 and 0.87 in our simulations.

Effects of between-gene correlation on identifying relevant genes

For each chosen maximum possible correlation between the relevant genes and non-relevant genes, we generated 100 data sets of sample size of 100 individuals. For each replication,

we applied the LARS-Lasso procedure to build a model which includes 20 genes by selecting an appropriate s value in the LARS-Lasso estimation. Table 1 summarizes the frequencies that the 20 relevant genes are among the first 20 genes that are selected by the LARS-Lasso procedure. We observe the following interesting results. First, as expected, the predictors with larger coefficients are more likely to be selected by the LARS-Lasso procedure. Second, it is interesting to observe that when the maximum possible correlation between the relevant and non-relevant genes increases, i.e., when the linear space spanned by the non-relevant genes gets close to the linear space expanded by those relevant genes, the chance of the relevant genes with smaller coefficients being selected gets smaller. This is because that at each step, the LARS-Lasso procedure only selects the gene with the largest absolute current inner product in the model. Of course, the chance of these relevant genes being selected also depends on the sample size. For example, for the maximum possible correlation of 0.85, more relevant genes are selected if the sample size is increased to 200 (see the last column of Table 1).

Predictive performance and comparison with other methods

We then examined the predictive performance of the proposed method. We simulated a sample size of 100 patients as the training data set to build the predictive model and evaluated the predictive performance based on another new data set of 100 patients (test data set). For each simulation, we generated 500 gene expression levels for each patients with the maximum possible between-gene correlation of 0.82. For each replication, we built a predictive model based on the training set. We applied the CVPL to choose the tuning parameter s used in the model. As an example, Figure 1 (a) shows the CVPL plot for one simulated data set, indicating that the tuning parameter of $s = 9$ gives the best predictive performance using the CVPL criteria, which corresponds to 38 genes. We then predicted the risk scores for the 100 patients in the test set. We repeated this procedure 100 times. We used the time-dependent AUC as a criteria to assess the predictive performance.

Figure 1 (b) shows the average of the estimated AUCs (+/-SE) over 100 replications using the predictive scores for the test sets, indicating a very good predictive performance. The AUC is over 75% at the beginning of the following-ups and remains high at later time. As a comparison, Figure 1 (c) and (d) show the AUCs plots for the predicted scores based on the L_2 penalized procedure proposed by Li and Luan (2003) and the principal-components based partial Cox regression (PC-PCR) procedure proposed by Li and Gui (2004). Note that both the L_2 penalized procedure and the PC-PCR procedure use all the genes in building

Table 1: Simulation results based on 100 replications. The first column shows the true coefficients of the 20 genes which are related to the risk of cancer recurrence. Columns 2-5 shows the frequency of each of these 20 relevant genes being selected by the LARS-Lasso procedure under four different correlation structures. The sample sizes are 100 patients for all the simulations. For the maximum possible correlation of 0.87, sample size of 200 patients was also considered and the results are presented in the last column.

	Maximum Correlation				
	0	0.71	0.82	0.87	
Coefficient	100	100	100	100	200
$\beta_1 = 0.19$	50	15	3	0	3
$\beta_2 = 0.95$	100	100	92	75	91
$\beta_3 = 0.96$	100	100	95	80	94
$\beta_4 = 0.91$	100	99	87	71	92
$\beta_5 = 0.19$	53	15	2	0	7
$\beta_6 = 0.25$	60	23	2	0	5
$\beta_7 = 0.69$	100	95	67	45	56
$\beta_8 = 0.33$	88	42	6	4	13
$\beta_9 = 0.34$	88	50	16	6	2
$\beta_{10} = 0.33$	91	53	13	1	4
$\beta_{11} = -0.92$	100	100	92	61	84
$\beta_{12} = -0.16$	40	7	5	1	0
$\beta_{13} = -0.83$	100	98	86	59	84
$\beta_{14} = -0.62$	100	91	58	26	44
$\beta_{15} = -0.65$	100	96	60	32	46
$\beta_{16} = -0.47$	98	76	38	11	22
$\beta_{17} = -0.72$	100	95	70	39	62
$\beta_{18} = -0.24$	66	19	6	5	8
$\beta_{19} = -0.41$	100	68	24	5	14
$\beta_{20} = -0.23$	64	23	3	4	4

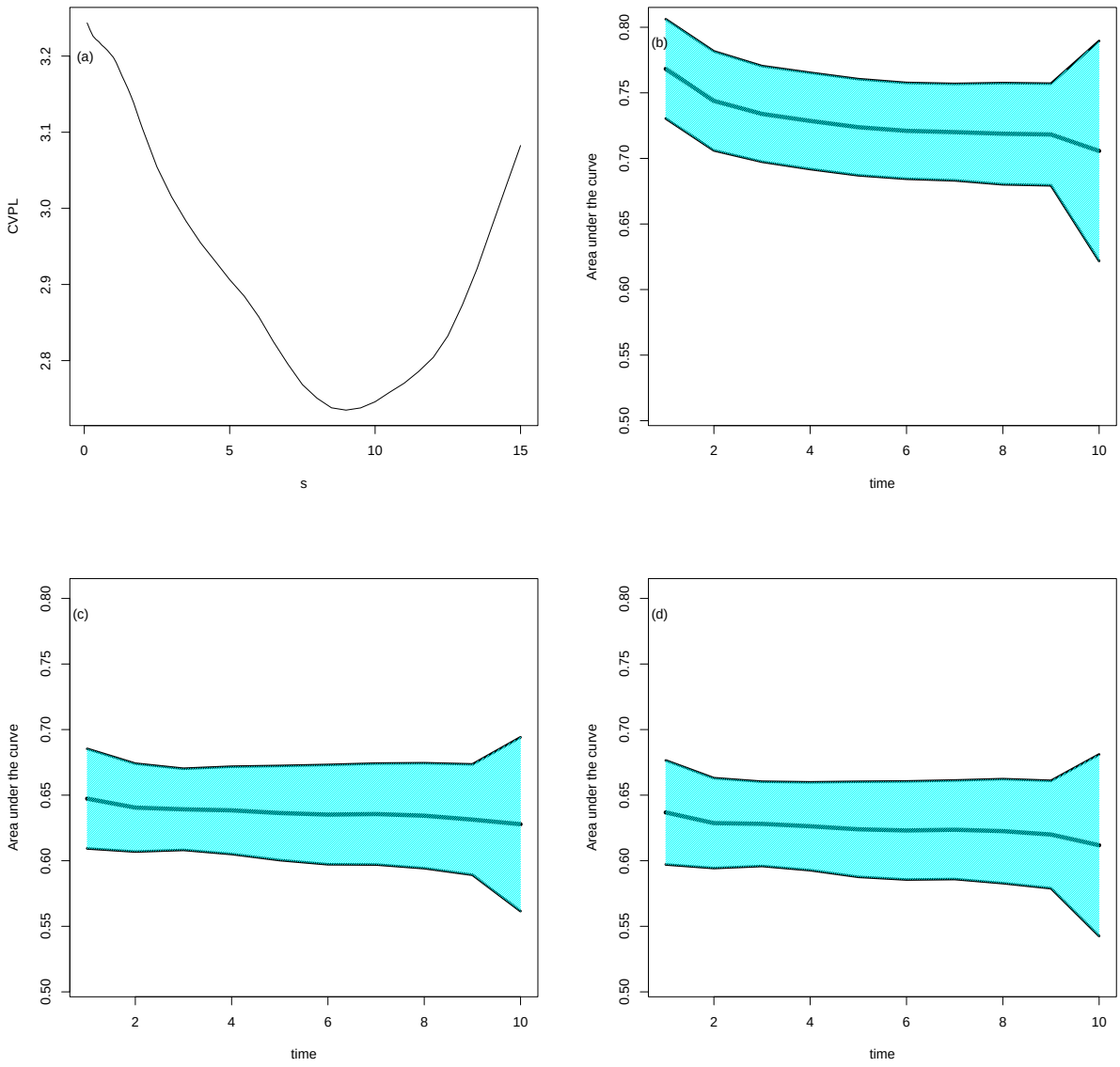

Figure 1: Results of simulations. (a) CVPL plot for one simulated data set; (b) AUCs for the test samples based on LARS-Lasso procedure; (c) AUCS for the test samples based on L_2 penalized estimation; (d) AUCS for the test samples based on the PC-PCR procedure. For each plot of (b) to (d), the three lines are the average AUCs over 100 replications together with +/- 1SE.

the predictive models. Clearly, neither of these procedures performed as well as the LARS-Lasso procedure in predicting the survival times for future patients as measure by the AUCs. We also performed L_2 procedure and the PC-PCR procedure using genes selected based on univariate Cox regression analysis and did not observe any improvement in their predictive performances.

As another way of comparing these three different methods, for each replication, we divided the patients in the test set into high and low-risk groups based on having positive or negative predictive risk scores and tested the statistical significance in the risk of cancer recurrence between the two groups. We observed that for a p-value of less than 10^{-5}, all 100 replications showed significant difference in risk between the high and low risk groups defined by the LARS-Lasso predicted scores, as compared to only 38 and 22 replications showing significant difference in risk between the high and low risk groups defined by the risk scores predicted by the L_2 penalized procedure and the PC-PCR procedure.

In summary, results from our simulation studies indicate that the LARS-Lasso procedure can indeed select genes that are related to censored phenotypes, especially those genes with relatively strong effects, although genes with smaller effects on survival are difficult to identify, especially when the correlations between the gene expression levels are high. When the correlations between the gene expression levels of the relevant genes and non-relevant genes are high, the CVPL procedure tends to select more genes in building the predictive models. However, we observed much better predictive performance of the LARS-Lasso procedure than the procedures proposed previously (Li and Luan, 2003; Li and Gui, 2004).

APPLICATION TO PREDICTION OF SURVIVAL TIME OF PATIENTS WITH DLBCL

To further demonstrate the utility of the LARS-Lasso procedure in relating microarray gene expression data to censored survival phenotypes, we re-analyzed a recently published data set of DLBCL by Rosenwald *et al.* (2002). This data set includes a total of 240 patients with DLBCL, including 138 patient deaths during the followups with median death time of 2.8 years. Rosenwald *et al.* divided the 240 patients into a training set of 160 patients and a validation set or test set of 80 patients and built a multivariate Cox model. The variables in the Cox model included the average gene expression levels of smaller sets of genes in four different gene expression signatures together with the gene expression level of BMP6. It should

be noted that in order to select the gene expression signatures, they performed a hierarchical clustering analysis for genes across all the samples (including both test and training samples). In order to compare our results with those in Rosenwald *et al.* (2002), we used the same training and test data sets in our analysis.

The gene expression measurements of 7,399 genes are available for analysis. However, there are a large number of missing gene expression values in the data set. Among the 7,399 genes, only 434 genes have no missing values. We first applied a nearest neighbor technique (Troyanskaya *et al.*, 2001) to estimate those missing values. Specifically, for each gene, we first identified 8 genes which are the nearest neighbors according to Euclidean distance. We then filled the missing with the average of the nearest neighbors. Our method is slightly different from that of Troyanskaya *et al.* (2001) in that the nearest neighbors are not restricted to those 434 genes with no missing data. We also tried the method of Troyanskaya *et al.* (2001) for filling the missing value, and the results of survival time prediction with two methods were very close.

Selection of genes related to risk of death

Although the LARS-Lasso procedure can in principle be used to fit the Cox model with $n - 1 = 159$ genes based on training sample of 160 patients, the algorithm becomes unstable when the number of variables is close to the sample size. As it was pointed out by Osborne *et al.* (1998), as s increases, when the number of nonzero coefficients are getting close to the number of observations, Lasso may not have an unique solution. In the following analysis, we only consider the Cox model with fewer than 100 genes. Figure 2 shows the path of the coefficients of the first 100 genes selected by the LARS-Lasso procedure as $s = \sum_{i=1}^{7399} |\beta_i|$ increases. Note that we only obtained coefficients at the turning points, where there is a change (addition or deletion) in the set of genes selected in the LARS iterations. In this figure, we add lines between those points to get an approximate of the full coefficient path. Note that genes are chosen in order of their relevances in predicting survival. These genes would provide a good list of candidate genes for further investigation. Table 2 shows the GenBank ID and a brief description of the first 10 genes selected. It is interesting to note that seven of these genes belong to the three gene expression signature groups defined in Rosenwald *et al.* (2002). These three signature groups include Germinal-center B-cell signature, MHC class II signature and Lymph-node signature. No genes in the proliferation signature group defined by Rosenwald *et al.*(2002) were selected by LARS-Lasso. Based on our search of GenBank

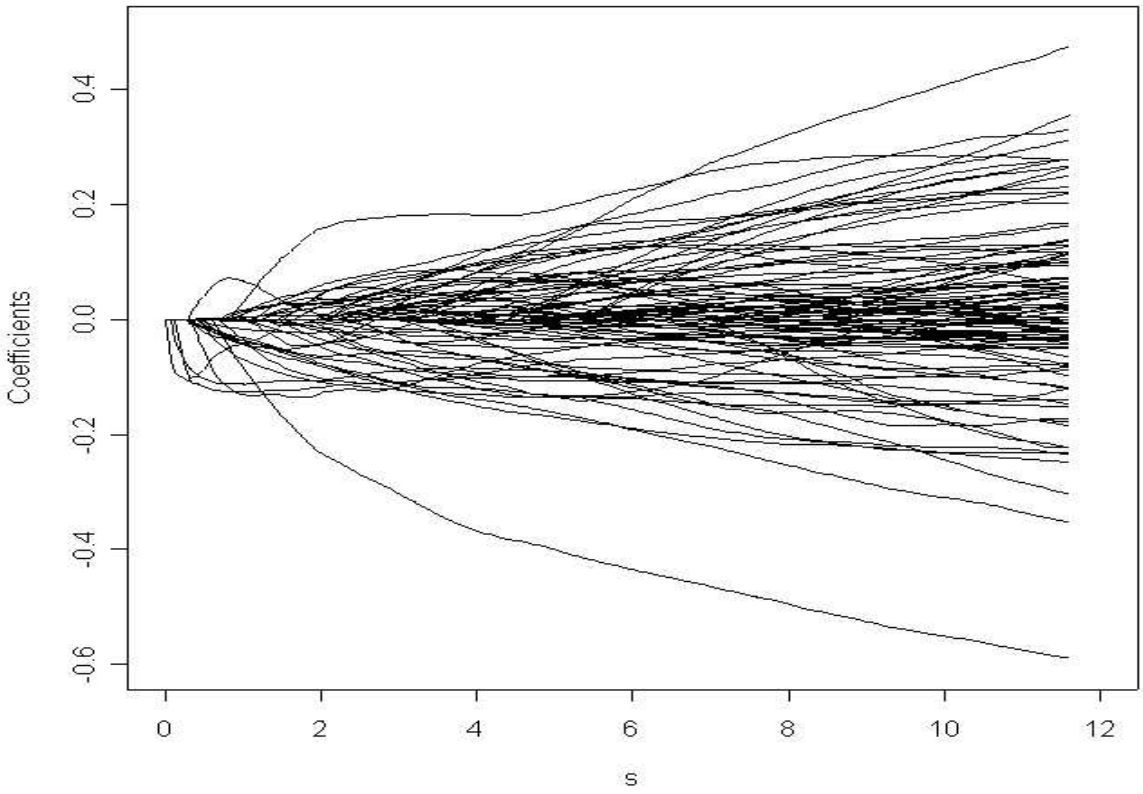

Figure 2: Approximated coefficients' path for the first 100 genes estimated by the LARS-Lasso procedure based on the 160 lymphoma patients in the training data set. Each line corresponds to the estimated coefficient for a given gene as the tuning parameter s increases. For a given s, the y-axis gives the current estimates of the coefficients of the genes selected.

(http://www.ncbi.nlm.nih.gov/Genbank/index.html), we found that the other three genes are also related to lymphoma or death. The gene AA76074 is COX15 homolog and mutations in this gene produce a defect in the mitochondrial heme biosynthetic pathway, causing early-onset fatal hypertrophic cardiomyopathy. The gene A29003 is protein coding TCL1A gene which has been demonstrated to be a powerful oncogene and when it is over-expressed in both B and T cells, it predominantly yields mature B cell lymphomas. Finally, the gene L19872 is Aryl hydrocarbon receptor (AHR), which is a ligand-activated transcription factor involved in the regulation of biological responses to planar aromatic hydrocarbons. AHR has been shown to regulate xenobiotic-metabolizing enzymes such as cytochrome P450, which belongs to the lymph-node signature group.

Table 2: GenBank ID and descriptions of the top 10 genes selected by the LARS-Lasso procedure based on the 160 patients in the training data set. As indicated are the gene expression signature groups that these genes belong to; Germ=Germinal-center B-cell signature, MHC=MHC class II signature, Lymph=Lymph-node signature. Genes AA760674, AA729003 and L19872 do not belong to these signature groups. No description was provided for gene LC_29222 by Rosenwald *et al.*(2002).

GenBank ID	Signature	Description
AA760674		cytochrome oxidase assembly protein (yeast)
X00452	MHC	major histocompatibility complex, class II, DQ alpha 1
AA729055	MHC	major histocompatibility complex, class II, DR alpha
AA714513	MHC	major histocompatibility complex, class II, DR beta 5
AA729003		T-cell leukemia/lymphoma 1A
AA805575	Germ	thyroxine-binding globulin precursor
AA598653	Lymph	osteoblast specific factor 2 (fasciclin I-like)
LC_29222	Lymph	
X59812	Lymph	cytochrome P450, subfamily XXVIIA polypeptide 1
L19872		hydrocarbon receptor

Evaluation of the predictive performance

We also examined how well the model built by the LARS-Lasso procedure predicts the survival of a future patient. Using the training set of 160 patients, we built a predictive Cox model with the LARS-Lasso procedure using the CVPL to select the optimal tuning value s. The minimum CVPL was obtained when $s = 0.28$, which corresponds to selecting four genes in the model. We also observed that the CVPL value increases by only 0.001 when the tuning parameter s increase from 0.28 to 0.33, which corresponds to nine genes in the model. Matter of fact, for s ranging from 0.28 to 0.33, the predictive performances of the resulting models are very comparable. We chose the most parsimonious model with four genes. These four genes are AA805575, LC_29222, X00452 and X59812 (see Table 2 for a description of these four genes), belonging to three of the four signature groups defined in Rosenwald *et al.* (2002).

We obtained the estimates of the coefficients of these four genes using the LARS-Lasso procedure, denoted by vector $\hat{\beta}$. The estimated coefficients for all four genes were negative, indicating that high expression levels of these genes reduce the risk of death among the patients

with DLBCL. This agrees with what Rosenwald *et al.* (2002) has found (see Table 2 of their paper). Based on the estimated model with four genes, we estimated the risk scores ($f(X) = \hat{\beta}'X$) for the 80 patients in the test data set based on their gene expression levels of the four genes in the predictive model. Figure 3 (a) shows the time-dependent AUCs for 1 to 20 years after diagnosis based on the estimated scores for the patients in the test set. The AUCs are between 0.66 and 0.68 in the first 10 years of followups, indicating a reasonable predictive performance.

To further examine whether clinically relevant groups can be identified by the model, we used zero as a cutoff point of the risk scores and divided the test patients into two groups based on whether they have positive or negative risk scores. Figure 3 (b) shows the Kaplan-Meier curves for the two groups of patients, showing very significant difference (p-value=0.0004) in overall survival between the high risk group (36 patients) and low risk group (44 patients).

A Comparison with other methods

As a comparison, we also analyzed the lymphoma data set using two other methods, the partial Cox regression methods in Li and Gui (2004) and the L_2 penalized method using linear kernels proposed by Li and Luan (2003). Figure 3 (c) and (d) show the survival curves of the two groups of the patients in the test data set defined by the scores estimated by the L_2 penalized method and the PRC method. We observe that the two risk groups defined by the LARS-Lasso estimated model showed more significant difference in risk of death than the groups defined by the other two models (p-value of 0.0004 versus 0.003). Figure 3 (a) shows the AUCs based on the risk scores estimated by the three different methods, again indicating better predictive performance of the LARS-Lasso procedure.

DISCUSSION AND CONCLUSIONS

It is clinically relevant and very important to predict patient's time to cancer relapse or time to death due to cancer after treatment using gene expression profiles of the cancerous cells prior to the treatment. Powerful statistical methods for such prediction allow microarray gene expression data to be used most efficiently. In this paper, we have proposed and studied the LARS-Lasso procedure for censored survival data in order to identify important predictive genes for survival using microarray gene expression data. To solve the computational difficulty, we modified the latest developed LARS procedure (Efron *et al.*, 2004) to obtain the solutions for the Lasso estimation of the Cox model. Since the risk of cancer recurrence or death

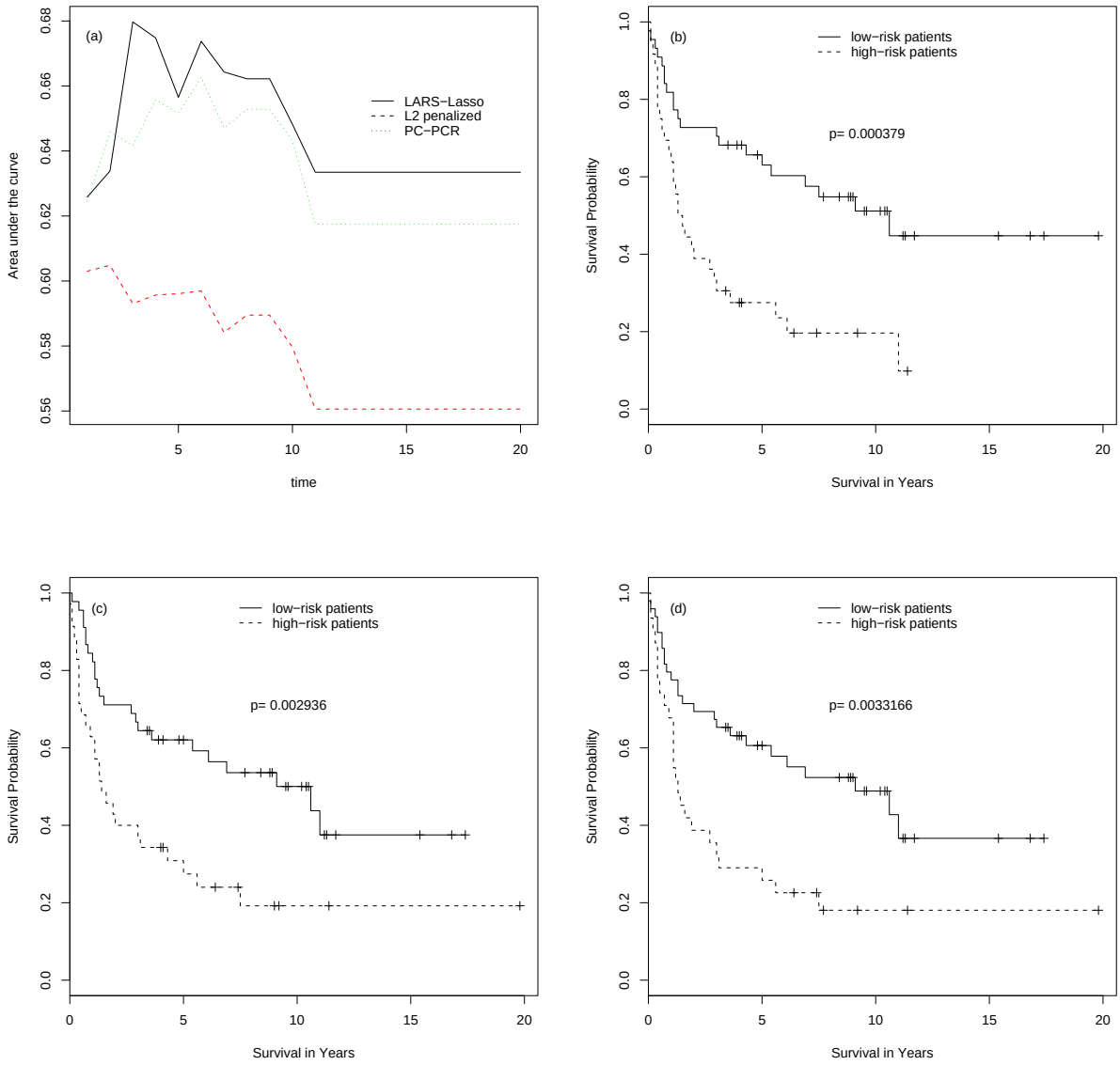

Figure 3: Results of analysis of the lymphoma data set. (a): AUCs for the test samples based on risk scores estimated by three different procedures. (b)-(d): the Kaplan-Meier curves for the high and low risk groups defined by the estimated scores for the 80 patients in the test data set. The scores are estimated based on the models estimated by the LARS-Lasso procedure (plot (b)), L_2 penalized procedure (plot (c)) and the PC-PCR procedure (plot (d)). The number of patients in the high risk group is 36, 35 and 31, respectively.

due to cancer may result from the interplay between many genes, methods which can utilize data of many genes, as in the case of our proposed procedure, are expected to show better performance in predicting risk. Our simulation studies demonstrated that the procedure can indeed be used to identify genes which are related to censored survival outcomes and to built a parsimonious model for predicting future patients survival. We have also demonstrated the applicability of our methods by analyzing time to death of the diffuse large B-cell lymphoma patients and obtained satisfactory results, as evaluated by both applying the model to the test data set and time dependent ROC curves.

While we did not compare the new procedure with all the other procedures available, we did compared the LARS-Lasso procedure with several other previously proposed methods in predictive performance and found that the new procedure performed better than the L_2 penalized or PC-PCR procedure (Li and Luan, 2003; Li and Gui, 2004) in predicting the future patients' survival. We would however expect that the LARS-Lasso procedure performs better than other dimension-reduction based procedures such as the partial least squares (Nguyen and Rocke, 2001; Park *et al.*, 2002) or the principal components Cox regression because the LARS-Lasso procedure automatically selects and utilizes only the relevant genes in building the predictive model. A comprehensive comparison of different methods warrants further research. It worth mentioning that the L_1 penalized regression was also demonstrated to perform better than other procedures in the settings of microarray gene expression data and linear models (Segal *et al.*, 2003)

The proposed LARS-Lasso procedure has no computational or methodological limitation in term of the number of genes that can potentially be used in the prediction of patient's time to clinical event. The method can in principle select $n - 1$ genes, where n is the sample size. However, when the number of predictors is close to the sample size, there is a risk of over-fitting. One drawback of the LARS-Lasso procedure is that if there is a group of variables or genes among which the pairwise correlations are very high, the LARS-Lasso tends to select only one variable from the group and does not care which one is selected. For genes sharing the same biological pathways, the correlations among them can be high (Zou and Hastie, 2003). If the LARS-Lasso procedure is used mainly for selecting important and relevant genes, one may want to include all these highly correlated genes, if one of them is selected. If the goal is to build a model with good predictive accuracy, this problem may not be severe since simple models are preferred for the scientific insight into the relationship between survival and gene expressions. However, we may expect more robust prediction if the average gene expression levels of highly correlated genes are used in the model. One way to extend the LARS-Lasso

procedure is that at each LARS variable selection step, we selected not only one single gene with the largest absolute current inner product, but a group of such genes with similar current inner product. An alternative is to use the elastic net penalty as recently proposed by Zou and Hastie (2003) for the penalized estimation.

The LARS-Lasso procedure assumes the Cox proportional hazards model, which is the most popular model for censored survival data. However, the proportional hazards assumption may not hold for gene expression data or for all diseases. It is possible to develop robust procedures under misspecified proportional hazards models along the lines of Lin and Wei (1989). In addition, model checking techniques analogues to those of Lin *et al.* (1993) can be derived. As an alternative, we can consider similar L_1 penalized estimation for the accelerated failure time models (Wei, 1992) or more general semi-parametric transformation models (Cheng *et al.*, 1995). We are currently pursuing these alternative models.

In summary, an important application of microarray technology is to relate gene expression profiles to various clinical phenotypes of patients such as time to cancer recurrent or overall survival time. The statistical model built to relate gene expression profile to the censored survival time should have the property of high predictive accuracy and parsimony. The proposed LARS-Lasso procedure in this paper can be very useful in building a parsimonious predictive model that can be used for classifying the future patients into clinically relevant high and low risk groups based on the gene expression profile and survival times of previous patients. The procedure can also be applied to select important genes which are related to patients' survival outcome.

ACKNOWLEDGMENTS

This research was supported by NIH grant R01-ES09911. We thank Dr. Yihui Luan for help on the L_2 penalized procedure.

REFERENCES

Akritas, M.G. 1994. Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics* 22,1299-1327.

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.E.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson,

J.R., 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 expression 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 tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences* 96(12),6745- 6750.

Cheng, S.C., Wei, L.J. and Ying, Z. 1995. Analysis of transformation models with censored data. *Biometrika* 82, 835-45.

Cox, D.R. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B* 34,187-220.

Efron B., Johnston I., Hastie T. and Tibshirani R. 2004. Least angle regression. *Annals of Statistics*, in press

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. *Proceeding of National Academy of Science USA* 98, 13784-13789.

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.

Heagerty, P.J., Lumley, T. and Pepe, M. 2000. Time dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 56,337-344.

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.

Li, H. and Gui, J. 2004. Partial Cox Regression Analysis for High-Dimensional Microarray Gene Expression Data. *Bioinformatics*, in press.

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.

Lin, D.Y. and Wei, L.J. 1989. The robust inference for the Cox proportional hazards model. *Journal of American Statistical Associaiton* 84, 1074-1078.

Lin, D.Y., Wei, L.J. and Ying, Z. 1993. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* 80,557-572.

Nguyen, D., Rocke, D.M. 2001. Partial Least Squares Proportional Hazard Regression for Application to DNA Microarray Data. *Technical report*, UC Davis.

Osborne, M.R., Presnell, B. and Turlach, B.A. 1998. On the Lasso and its dual. *Research Report*, Department of Statistics, University of Adelaide.

Park, P.J., Tian L, Kohane, I.S. 2002. Linking Expression Data with Patient Survival Times Using Partial Least Squares. *Bioinformatics* 18, S120-127.

Rosenwald, A., Wright, G., Chan, W., Connors, J.M., Campo, E., Fisher, R., Gascoyne, R.D., Muller-Hermelink, K., Smeland, E.B. and Staut, L.M. 2002. The Use of Molecular Profling to Predict Survival After Themotheropy for Diffuse Large-B-Cell Lymphoma. *The New England Journal of Medicine* 346,1937-1947.

Segal, M.R., Dahlquist, K.D., Conklin, B.R. 2003. Regression approaches for microarray data analysis. *Journal of Computational Biology* 10, 961-980.

Smyth, P. 2001. Model Selection of Probabilistic Clustering Using Cross-validated Likelihood. *Statistics and Computing* 10, 63-72.

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. and Brown, P.O., Botstein D, Eystein Lnning P., Brresen-Dale, A. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* 98,10869-10874.

Tibshirani, R. 1995. Regression shrinkage and selection via the Lasso. *Journal of Royal Statistical Society B* 58,267-288.

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

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17,520-525.

Van Der Laan, M.J., Dudoit, S., Keles, S. 2003. Asymptotic Optimality of likelihood-based Cross Validation, *Technical Report*, Division of Biostatistics, University of California, Berkeley.

Verwij, P.J.M. and Van Houwelingen, J.C. 1993. Cross validation in survival analysis. *Statistics in Medicine* 12,2305-2314.

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.

Zou, H. and Hastie, T. 2003. Regression shrinkage and selection via the elastic net, with applications to microarrays. *Technical report*, Department of Statistics, Stanford University.