Survival Analysis with Large Dimensional Covariates: An Application in Microarray Studies

David A. Engler[*] Yi Li[†]

[*] Harvard University, engler@fas.harvard.edu

[†] Dana-Farber Cancer Institute, yili@hsph.harvard.edu

Survival Analysis with Large Dimensional Covariates: An Application in Microarray Studies

David A. Engler and Yi Li

Abstract

Use of microarray technology often leads to high-dimensional and low- sample size data settings. Over the past several years, a variety of novel approaches have been proposed for variable selection in this context. However, only a small number of these have been adapted for time-to-event data where censoring is present. Among standard variable selection methods shown both to have good predictive accuracy and to be computationally efficient is the elastic net penalization approach. In this paper, adaptation of the elastic net approach is presented for variable selection both under the Cox proportional hazards model and under an accelerated failure time (AFT) model. Assessment of the two methods is conducted through simulation studies and through analysis of microarray data obtained from a set of patients with diffuse large B-cell lymphoma where time to survival is of interest. The approaches are shown to match or exceed the predictive performance of a Cox-based and an AFT-based variable selection method. The methods are moreover shown to be much more computationally efficient than their respective Cox- and AFT- based counterparts.

Survival Analysis With Large Dimensional Covariates: An Application In Microarray Studies

David Engler
Department of Statistics
Brigham Young University
230 TMCB, Provo, UT 84664
email: engler@byu.edu

Yi Li
Department of Biostatistics
Dana-Farber Cancer Institute
375 Longwood Avenue, Boston, MA 02115
email: yili@hsph.harvard.edu

Abstract

Use of microarray technology often leads to high-dimensional and low-sample size data settings. Over the past several years, a variety of novel approaches have been proposed for variable selection in this context. However, only a small number of these have been adapted for time-to-event data where censoring is present. Among standard variable selection methods shown both to have good predictive accuracy and to be computationally efficient is the elastic net penalization approach. In this paper, adaptation of the elastic net approach is presented for variable selection both under the Cox proportional hazards model and under an accelerated failure time (AFT) model. Assessment of the two methods is conducted through simulation studies and through analysis of microarray data obtained from a set of patients with diffuse large B-cell lymphoma where time to survival is of interest. The approaches are shown to match or exceed the predictive performance of a Cox-based and an AFT-based variable selection method. The methods are moreover shown to be much more computationally efficient than their respective Cox- and AFT-based counterparts.

1 Introduction

Analysis of high-dimensional and low-sample size (HDLSS) data is increasingly an objective of interest. Such analyses are of particular interest in the analysis of DNA microarray data where the number of genes typically far exceeds sample size. In this setting, a frequent objective is the identification of a subset of genes whose expression levels are significantly correlated with a given clinical outcome or classification. Estimation of the effect of each identified gene is also usually desired. Identified genes are then often employed to build a predictive model in which prediction of outcome for new patients is conducted.

Over the years, a number of variable selection and estimation methodologies based on the maximization of a penalized likelihood have been proposed. Methods of penalization include traditional approaches such as AIC (Akaike, 1973) and BIC (Schwartz, 1978) as well as more recent developments including bridge regression (Frank and Friedman, 1993), the LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), LARS (Efron et al., 2004), the elastic net (Zou and Hastie, 2005), and MM algorithms (Hunter and Li, 2005).

A number of these penalization methods, however, are not practical for HDLSS environments. For example, metrics such as AIC and BIC are typically employed in stepwise or backward selection algorithms. When the number of potential variable combinations is very large, the such algorithms are computationally burdensome. To improve efficiency, these algorithms typically do not reintroduce a variable to the search path once it has been dropped. This practice may be particularly problematic when the data set includes a large number of variables, many of which may be correlated. Others, such as the LASSO, often rely on quadratic programming algorithms that are not tractable when the number of variables is greater than the number of subjects.

Microarray data analysis is further complicated when the outcome of interest is a time to an event. In these cases, either dropout or study termination may occur prior to event occurrence for a number of subjects. Typically, then, a number of the outcome variables are censored. When the number of subjects exceeds the number of variables, a variety of approaches are available for identification of variables significantly associated with outcome. Maximization of the Cox partial likelihood (Cox, 1972), for example, can be conducted under the assumption of proportional hazards. Alternatively, the accelerated failure time (AFT) model (see Wei, 1992) is frequently employed, often through use of estimating equations, when the proportional hazards assumption is not tenable. In HDLSS settings, however, use of these models requires modification.

Several authors have proposed variable selection methods for HDLSS time-to-event data under the Cox proportional hazards model. Li and Luan (2003), treating the negative log Cox partial likelihood as a loss function, propose use of kernel transformations (Kimeldorf and Wahba, 1971) to overcome the difficulties (*e.g.,* inversion of large matrices) introduced when the number of variables is much greater than the number of subjects. Gui and Li (2005a), also using the loss function based on the Cox partial likelihood, propose estimating the Cox regression parameters through a threshold gradient descent (TGD) minimization where the absolute values of the gradients are constrained by a specified threshold (Friedman and Popescu, 2004). A third approach based on the Cox model has been proposed by Gui and Li (2005b) and is based on the least angle regression (LARS) algorithm outlined by Efron et al. (2004). The LARS procedure performs variable selection in the linear regression setting by iteratively identifying the variable with the greatest correlation with the current residuals. Efron et al. show that variable selection procedures such as the LASSO and forward stagewise linear regression can be treated as special cases of the LARS algorithm. Gui and Li suggest modification of the transformation outlined by Tibshirani (1997) as the foundation for a Cox proportional hazards variable selection approach based on the LARS formulation of the LASSO penalty. Segal (2005) proposed methods for increasing the efficiency of LASSO under the Cox model, that achieve similar levels of performance as the Gui and Li (2005b) approach.

Likewise, a few authors have proposed variable selection methods based on AFT models. Huang et al. (2006), for example, propose two separate AFT model-based methods, a LASSO penalization approach and a threshold gradient directed (TGD) approach. Both methods entail an iterative approach which requires the calculation of the derivative of the objective function based on Stute's weighted least squares estimator (Stute, 1993; Stute, 1996). Sha et al. (2006) propose a Bayesian variable selection method based on the AFT model in which utilization of distributional assumptions and conjugate priors allows the regression parameters of interest to be integrated out. An MCMC algorithm is then employed to conduct estimation. In a comparison study, Datta et al. (2007) examined the performance of a partial least squares (PLS) approach to that of a LASSO approach under the AFT model. The LASSO was found to outperform PLS in microarray settings where a large number of non-significant, or noise, variables were present.

There are a number of drawbacks to current methods of variable selection in HDLSS settings when censored data is present. The Li and Luan (2003) method is limited, for example, in that for prediction, all genes in the data set are included; a straightforward method of gene selection for prediction is not outlined. The TGD approaches of Gui and Li (2005a) and Huang et al. (2006) seem to be limited in that, at least in initial data analyses, very small changes in the threshold parameter dramatically altered the number of variables selected. Hence, effective identification of the optimal threshold might be unwieldy. A second drawback is that in the same analyses, the TGD method appeared to have less predictive power than alternative methods (see Gui and Li, 2005ab). Use of the LASSO in the methods proposed by Gui and Li (2005b) and Huang et al. (2006) might also lead to difficulties. For one, when the number of variables p is larger than the number of subjects n, the number of variables selected by the LASSO is at most n. This restriction may be problematic for gene expression data where $p \gg n$. A second drawback of the LASSO is a result of its convexity. Zou and Hastie (2005) show that for non-strictly convex penalty functions such as the LASSO, performance is suboptimal when highly correlated variables are present. Given a set of highly correlated variables associated with outcome, procedures that employ a penalty function that is not strictly convex often will identify only one of the variables and ignore the others. This limitation might be particularly problematic in the analysis of gene expression data where identification of an entire set of correlated genes may lead to an improved understanding of the biological pathway.

Modification of the elastic net penalization approach (Zou and Hastie, 2005) may be useful for the analysis of HDLSS time-to-event data. First, the elastic net approach is not limited in the number of variables selected by the number of available subjects. That is, the number of variables selected can be greater than the number of subjects. Second, the elastic net penalty function is strictly convex and therefore will more frequently identify an entire set of correlated genes than do methods based on penalty functions that are not strictly convex. Finally, as shown by Zou and Hastie, the elastic net is computationally efficient. To date, the only attempt to employ the elastic net penalization approach to HDLSS censored data has been proposed by Wang et al. (2006) under the AFT model, employing an imputation approach based on the Buckley and James algorithm (1979). However, the Buckley-James approach entails an iterative least squares procedure that is known to suffer from convergence problems and is more computationally intensive than other methods.

In this paper, two elastic net based variable selection methods for high-dimensional low sample size time-to-event data are presented. First, a Cox elastic net (EN-Cox) approach is outlined that is based on the Cox proportional hazards model and utilizes modifications of the algorithms proposed by Tibshirani (1997) and Gui and Li (2005b). Second, an AFT elastic net (EN-AFT) approach is presented which employs a mean imputation approach for the estimation of AFT model parameters. The approaches are shown to be an improvement over existing methods in terms of prediction accuracy and computational efficiency. The elastic net penalization approach is outlined in Section 2.1, followed by presentation of the EN-Cox (Section 2.2) and EN-AFT (Section 2.3) models and computational algorithms. Discussion of properties resultant from the convexity of EN-Cox and EN-AFT is contained in Section 2.4. Selection of tuning parameters and a method for the evaluation of predictive performance are discussed in Sections 2.5 and 2.6, respectively. Performance of the EN-Cox and EN-AFT approaches is evaluated in Sections 3 and 4. Results of a gene expression data analysis are provided in Section 3 and Section 4 consists of several simulation studies. The computation efficiency of the methods is briefly assessed in Section 5. Finally, areas of potential future research are discussed in Section 6.

2 Statistical Model and Methods

2.1 Elastic Net

In the linear regression setting, the elastic net objective function is defined by Zou and Hastie (2005) as

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \tag{2.1}$$

for some fixed, non-negative λ_1 and λ_2, where $\mathbf{y} = (y_1, \ldots, y_n)$ is the centered response vector for n subjects and $\mathbf{X}$ is the design matrix based on p standardized (*i.e.*, location and scale transformed) variables. Notably, for $0 < \lambda_2 \leq 1$, the penalty function is strictly convex and hence is not restricted in its ability to identify entire sets of highly correlated variables. The elastic net estimator of $\boldsymbol{\beta}$, then, is the minimizer of (2.1).

To adjust for HDLSS data settings (and the resultant difficulties in the estimation of $\boldsymbol{\beta}$), Zou and Hastie employ two simple modifications to the elastic net model. First, an augmentation of $\mathbf{X}$ and $\mathbf{y}$ is utilized which leads to a sparse data matrix $\mathbf{X}^*$ with rank p. Hence, through use of the augmentation, selection of up to p variables is possible even when $p \gg n$. Additionally, the sparse data matrix $\mathbf{X}^*$ leads to a computationally efficient algorithm. Second, a scaled $\hat{\boldsymbol{\beta}}$ is employed to overcome a problem of double shrinkage (*i.e.,* the shrinking of coefficient estimates to increase stability). Following data augmentation and the rescaling of $\hat{\boldsymbol{\beta}}$, the resultant elastic net estimator $\hat{\boldsymbol{\beta}}$ is defined as

$$\hat{\boldsymbol{\beta}} = \text{arg min}_{\boldsymbol{\beta}} \left[\boldsymbol{\beta}' \left(\frac{\mathbf{X}'\mathbf{X} + \lambda_2 \mathbf{I}}{1 + \lambda_2} \right) \boldsymbol{\beta} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right]. \tag{2.2}$$

Of interest, then, is the elastic net estimator when the outcome is time to an event and censoring is present. Let time T_i for subject $i = 1, \ldots, n$ depend upon p gene expression levels $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. Due to censoring, $Y_i = \text{min}(T_i, C_i)$ is observed where C_i is the time to the first censoring event (*e.g.,* study conclusion, date of final follow up) for subject i. Let $\delta_i = 0$ indicate censoring and $\delta_i = 1$ otherwise.

2.2 A Cox-based Adaptation of Elastic Net

Under the Cox proportional hazards model, the hazard function for individual i is specified as $\lambda(t_i) = \lambda_0(t_i)\text{exp}(\boldsymbol{\beta}'\mathbf{x_i})$, where covariate matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ and where baseline hazard $\lambda_0(\mathbf{t})$ is common to all subjects but is unspecified or unknown. Let ordered risk set at time $t_{(r)}$ be denoted by $R_r = \{j \in 1, \ldots, n : Y_j \geq t_{(r)}\}$. Assume that censoring is noninformative and that there are no tied event times. The Cox log partial likelihood can then be defined as

$$\ell(\beta) = \frac{1}{n} \sum_{r \in D} \ln \left(\frac{\text{exp}(\boldsymbol{\beta}'\mathbf{x}_{(r)})}{\sum_{j \in R_r} \text{exp}(\boldsymbol{\beta}'\mathbf{x}_j)} \right), \tag{2.3}$$

where D denotes the set of indices for observed events. The Cox elastic net estimate of $\boldsymbol{\beta}$ in this setting can be obtained through adaptation of a quadratic programming approach outlined by Tibshirani and Hastie (Hastie and Tibshirani 1990; Tibshirani, 1997). Namely, let $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, $u = \partial\ell/\partial\boldsymbol{\eta}$, $\mathbf{A} = -E[\partial^2\ell/\partial\boldsymbol{\eta}\boldsymbol{\eta}']$, and $\mathbf{z} = (\boldsymbol{\eta} + \mathbf{A}^{-1}\mathbf{u})$. A modified Newton-Raphson iterative procedure can then be employed to optimize (2.3). Specifically, the usual Newton-Raphson update is expressed as an iterative reweighted least squares step. The weighted least squares step is then replaced by a constrained weighted least squares procedure. Let, for each step, $\mathbf{z}_0 = (\boldsymbol{\eta}_0 + \mathbf{A}^{-1}\mathbf{u})$, where $\boldsymbol{\eta}_0$ is based on the $\boldsymbol{\beta}$ estimate of the previous step. A one-term Taylor series expansion for each step can then be represented as $(\mathbf{z}_0 - \boldsymbol{\eta})'\mathbf{A}(\mathbf{z}_0 - \boldsymbol{\eta})$.

As noted by Gui and Li (2005b), however, this approximation can be rewritten as $(\tilde{\mathbf{z}}_0 - \tilde{\mathbf{X}}\boldsymbol{\beta})'(\tilde{\mathbf{z}}_0 - \tilde{\mathbf{X}}\boldsymbol{\beta})$, where $\tilde{\mathbf{z}}_0 = \mathbf{Q}\mathbf{z}_0$ and $\tilde{\mathbf{X}} = \mathbf{Q}\mathbf{X}$, where $\mathbf{Q} = \mathbf{A}^{\mathbf{1/2}}$. An estimate, $\hat{\mathbf{A}}$, of $\mathbf{A}$ can be obtained using the observed Fisher information. Under this formulation, the problem of obtaining an elastic net estimate for $\boldsymbol{\beta}$ is akin to the problem posed in (2.2). That is, the optimal $\hat{\boldsymbol{\beta}}$ is formulated as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[\boldsymbol{\beta}' \left(\frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + \lambda_2 \mathbf{I}}{1 + \lambda_2} \right) \boldsymbol{\beta} - 2\tilde{\mathbf{z}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right]. \tag{2.4}$$

Estimation of $\hat{\boldsymbol{\beta}}$ is accomplished through the following algorithm:

1. Set tuning parameters and initialize $\hat{\boldsymbol{\beta}} = \mathbf{0}$.
2. Compute $\boldsymbol{\eta}$, $\mathbf{u}$, $\hat{\mathbf{A}}$, and $\mathbf{Q}$ based on the current value of $\hat{\boldsymbol{\beta}}$.
3. Let $\mathbf{z}_0 = \mathbf{z}$ for the first iteration, otherwise compute $\mathbf{z}_0$.
4. Compute $\tilde{\mathbf{X}} = \mathbf{Q}\mathbf{X}$ and $\tilde{\mathbf{z}}_0 = \mathbf{Q}\mathbf{z}_0$.
5. Minimize $(\tilde{\mathbf{z}}_0 - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}})'(\tilde{\mathbf{z}}_0 - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}})$ subject to the elastic net constraints.
6. Update $\hat{\boldsymbol{\beta}}$.
7. Repeat steps 2–6, subject to the elastic net constraints, until $\hat{\boldsymbol{\beta}}$ does not change.

Of note, $\mathbf{Q}$ can then be obtained through the Cholesky decomposition of $\hat{\mathbf{A}}$. Selection of tuning parameters in Step 1 and their effect on the elastic net constraints in Steps 5 and 7 is discussed in Section 2.5.

2.3 An AFT Adaptation of Elastic Net

When the assumption of proportional hazards is not tenable, the accelerated failure time (AFT) model can be utilized. The AFT model is a linear regression model in which the logarithm of response T_i is related linearly to covariates $\mathbf{x}_i$:

$$h(T_i) = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.5}$$

where $h(.)$ is the log transformation or some other monotone function. In this case, the Cox assumption of multiplicative effect on hazard function is replaced with the assumption of multiplicative effect on outcome. In other words, it is assumed that the variables $\mathbf{x}_i$ act multiplicatively on time and therefore affect the rate at which individual i proceeds along the time axis.

Because censoring is present, the standard least squares approach cannot be employed to estimate the regression parameters in (2.5) even when $p < n$. One approach for AFT model implementation entails the replacement of censored Y_i with imputed values. One such approach is that of mean imputation in which each censored Y_i is replaced with the conditional expectation of T_i given $T_i > C_i$. The imputed value $h(Y_i^*)$ can then be given (see Datta, 2005) by

$$h(Y_i^*) = (\delta_i)h(Y_i) + (1 - \delta_i)\{\hat{S}(y_i)\}^{-1} \sum_{t_{(r)} > t_i} h(t_{(r)}) \Delta\hat{S}(t_{(r)}), \tag{2.6}$$

where $\hat{S}$ is the Kaplan-Meier estimator (Kaplan and Meier, 1958) of the survival function and where the $\Delta\hat{S}(t_{(r)})$ is the step of $\hat{S}$ at time $t_{(r)}$.

Datta et al. (2007) recently assessed the performance of several approaches to AFT model implementation, including reweighting the observed T_i, replacement of each censored T_i with an imputed observation, drawn from the conditional distribution of T (multiple imputation), and mean imputation. Datta et al. (2007) found that in the HDLSS setting, the mean imputation approach outperformed reweighting and multiple imputation under the LASSO penalization.

Of interest, then, is the elastic net estimate of $\boldsymbol{\beta}$ for settings when $p \gg n$. Using the imputed values (2.6), estimation of the elastic net parameters can be conducted through use of the following algorithm:

1. Set tuning parameters and initialize $\hat{\boldsymbol{\beta}} = \mathbf{0}$.
2. Minimize $\sum_i (y_i^* - \hat{\boldsymbol{\beta}}'\mathbf{x}_i)'(y_i^* - \hat{\boldsymbol{\beta}}'\mathbf{x}_i)$ subject to the elastic net constraints.
3. Update $\hat{\boldsymbol{\beta}}$.
4. Repeat steps 2–3, subject to the elastic net constraints, until $\hat{\boldsymbol{\beta}}$ does not change.

Selection of tuning parameters in Step 1 and their effect on the elastic net constraints in Steps 2 and 4 is discussed in Section 2.5.

2.4 The Grouping Effect in EN-Cox and EN-AFT

Zou and Hastie (2005) show that the elastic net is superior to the LASSO in its ability to identify entire groups of highly correlated variables in the linear regression setting. This characteristic can be referred to as a grouping effect. A variable selection method, then, that exhibits the grouping effect will assign non-zero coefficients to an entire set of highly correlated variables. This characteristic is especially important in analysis of gene expression data where identification of an entire set of correlated genes may lead to an improved understanding of the biological pathway.

Both EN-Cox and EN-AFT exhibit the grouping effect. Because EN-AFT is based on a linear regression model, this follows by the same reasoning outlined by Zou and Hastie (2005). By similar reasoning, it is also easy to show that EN-Cox exhibits the grouping effect for $0 < \lambda_2 \leq 1$. Proposition 1 describes the expected behavior of EN-Cox for an extreme case and Proposition 2 provides a general property of EN-Cox when correlated variables are present. Details of Proposition 1 and 2 are provided in Section 7.

Proposition 1: Let $\mathbf{x}_i = \mathbf{x}_j$ for some $i, j \in \{1, \ldots, p\}$. Let $\hat{\boldsymbol{\beta}}$ be the EN-Cox estimate of the Cox regression parameter $\boldsymbol{\beta}$. Then $\hat{\beta}_i = \hat{\beta}_j$.

Proposition 1 states that given identical covariate vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, the EN-Cox estimate of $\boldsymbol{\beta}$ will assign identical values to $\hat{\beta}_i$ and $\hat{\beta}_j$.

Proposition 2: Let transformed response vector $\tilde{\mathbf{z}}$ and covariate matrix $\tilde{\mathbf{X}}$ be mean-centered and standardized. Let original covariate vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ be highly correlated. Without loss of generality, assume $\rho > 0$. Let $\hat{\boldsymbol{\beta}}$ be the EN-Cox estimate of the Cox regression parameter $\boldsymbol{\beta}$ and assume sign$(\hat{\beta}_i) =$ sign$(\hat{\beta}_j)$. Then for fixed λ_1 and λ_2

$$\frac{|\hat{\beta}_i - \hat{\beta}_j|}{|\tilde{\mathbf{z}}|} \leq \frac{\sqrt{2(1 - (\mathbf{x}_i'\mathbf{A}\mathbf{x}_j))}}{\lambda_2}, \tag{2.7}$$

where $\mathbf{x}_i'\mathbf{A}\mathbf{x}_j$ is equal to the correlation between transformed covariate vectors $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$.

Proposition 2 states that the standardized difference between the EN-Cox estimates $\hat{\beta}_i$ and $\hat{\beta}_j$ corresponding to correlated variables $\mathbf{x}_i$ and $\mathbf{x}_j$ is bounded above by a function of the correlation between transformed covariate vectors $\tilde{\mathbf{x}}_i = \tilde{\mathbf{x}}_j$. Of note, Proposition 1 and 2 extend the results of Zou and Hastie (2005) to settings in which censored data is present. Further examination of the grouping effect of EN-Cox and EN-AFT is provided in Section 4.

2.5 Selection of Tuning Parameters

The elastic net requires the selection of two tuning parameters, λ_1 and λ_2. Zou and Hastie (2005) note that alternatives to λ_1 are possible. The various choices correspond to different methods of identifying the stopping point of the procedure and hence affect Steps 4 and 6 of the algorithms outlined in Sections 2.2 and 2.3. Among those alternatives proposed is the maximum number of steps k allowable in the entire solution path where one iteration of the above algorithms constitutes a single step. The choice of k is useful as its selection requires no prior knowledge (or guesswork) regarding the actual values of the regression coefficients and is employed in both EN-Cox and EN-AFT.

Evaluation of the two parameters λ_2 and k across a two-dimensional surface of parameter values is required. Potential values of λ_2 should span a wide range, *e.g.*, $\boldsymbol{\lambda}_2 = (0, 0.01, 0.1, 1, 10, 100)$. The potential values of k will depend on the size of the data set. Selection of λ_2 and k in both EN-Cox and EN-AFT is conducted simultaneously using a cross validation score (CVS) (Verwij and Van Houwelingen, 1993; Huang and Harrington, 2002; Gui and Li, 2005b; Huang et al., 2006):

$$CVS(\lambda_2, k) = \frac{1}{n}\sum_{i=1}^{N}\left[\ell(\hat{\boldsymbol{\beta}}^{(-i)}_{\lambda_2,k}) - \ell^{(-i)}(\hat{\boldsymbol{\beta}}^{(-i)}_{\lambda_2,k})\right], \tag{2.8}$$

where $\hat{\boldsymbol{\beta}}^{(-i)}_{\lambda_2,k}$ is the EN-Cox or EN-AFT estimator without the i^{th} subject, based on penalty parameters λ_2 and k. The function $\ell(.)$ is the negative log partial likelihood (2.3) in the case of EN-Cox and is the AFT objective function in the case of EN-AFT. The function $\ell^{(-i)}(.)$ is similarly defined, differing only in that subject i is excluded in its calculation.

2.6 Predictive Performance

Assessment of EN-Cox and EN-AFT can be conducted through analysis of predictive performance using time-dependent receiver operator characteristic (ROC) curves (Heagerty et al., 2000). In general, for dichotomous disease-status indicator D and continuous diagnostic test outcome X, an ROC curve is defined as the plot of the sensitivity of the test $X > c$ versus (1 - specificity) over $c \in (-\infty, \infty)$. Heagery et al. extend this formulation to time-to-event data when censoring is present. Given linear risk score function $f(X) = \boldsymbol{\beta}'\mathbf{X}$, sensitivity and specificity for cutoff c at time t are defined as

$$\text{sensitivity}(c, t|f(X)) = P[f(X) > c|\delta(t) = 1] \tag{2.9}$$

$$\text{specificity}(c, t|f(X)) = P[f(X) \leq c|\delta(t) = 0], \tag{2.10}$$

where $\delta(t)$ is the event indicator at time t. At each time t, an ROC curve is generated for $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and an associated area under the curve (AUC) is calculated. The plot of AUC over time is then helpful in assessing the predictive performance of a given variable selection method. Heagerty et al. suggest use of the nearest neighbor approach of Akritas (1994) for estimation of (2.9) and (2.10).

3 Data Analysis

Diffuse large-B-cell lymphoma (DLBCL) is a common type of non-Hodgkin's lymphoma in adults. Heterogeneity in response to treatment has suggested the existence of clinically distinct subypes. Through hierarchical clustering of DNA microarray data, Alizadeh et al. (2000) identified two distinct DLBCL subtypes which were found to correspond to differences in B-cell differentiation stages. The authors futher determined that overall survival significantly differed between the two groups.

More recently, Rosenwald et al. (2002) utilized Lymphochip DNA microarrays to collect and analyze gene expression data from 240 biopsy samples of DLBCL tumors. For each subject, 7399 gene expression measurements were obtained. During the time of follow-up, 138 patient deaths were observed (*i.e.*, 42.5% censoring). Rosenwald et al. utilized hierarchical clustering to identify three subtypes. Two of the groups corresponded roughly with the groups identified by Alizadeh et al. (2000), while a third appeared to be novel. However, Rosenwald el al. noted that despite overall differences in survival between the three groups, within-group variablility was substantial; the clustering did not fully account for differences in survival among patients. Using a training set of 160 individuals, the authors employed a Cox proportional hazards model, testing for association between survival and each of the 7399 microarray features separately. Notably, no adjustments were made for multiple testing. A total of 670 gene expressions were found to be significantly associated with outcome and were subsequently clustered into "signature" groups. A final Cox proportional hazards model was then constructed using expression averages from selected feature subsets within each signature group. Using the Cox model regression coefficient estimates, an outcome-predictor score was calculated for each of the remaining 80 individuals (the validation set). Rosenwald et al. found the outcome-predictor scores to be significantly associated with outcome.

Analysis of the Rosenwald et al. DLBCL data was conducted using both EN-Cox and EN-AFT. For comparison purposes, analysis was also conducted using the Gui and Li (2005b) LARS-based LASSO (LASSO-Cox) method. To assess the effect of differing imputation methods under the AFT model, separate analyses were conducted using the mean imputation method described in Section 2.3 and the Buckley-James imputation method. Comparison was also conducted between the A training set of 160 randomly selected subjects was utilized. Selection of tuning parameters for each method was conducted using half of the training set while model fit (*i.e.*, variable selection and coefficient estimation) was conducted using the other half. Predictive performance was assessed using a validation set composed of the 80 subjects not in the training set.

The methods varied in the number of gene expressions identified as significantly associated with survival. Both EN-Cox and EN-AFT identified a greater number of signficant features than LASSO-Cox. EN-AFT computed under mean imputation (EN-AFT-M) identified 13 genes, EN-AFT computed under Buckley-James imputation (EN-AFT-BJ) identified 18 genes, EN-Cox identified 16 genes, and LASSO-Cox identified 7 genes.

To assess predictive performance, the median AUC for each six month interval (for which there was data) was then calculated and plotted for each method. Results are presented in Figure 1. For the first ten years of follow-up, the median AUC for EN-AFT-M is 0.61 and is 0.56 for EN-AFT-BJ. Use of the Cox model results in a median AUC of 0.58 for both EN-Cox and LASSO-Cox. Instability in AUC estimates for subsequent times (post year 10) appears to be due to sparsity of event times. For this analysis, then, EN-AFT-M outperformed EN-AFT-BJ (in terms of prediction) using a smaller set of identified genes. The predictive performance of EN-AFT-M was also slightly superior to EN-Cox and LASSO-Cox in this data analysis.

The complete index of genes selected by each method is presented in Table 1. Several features of the variable selection process for this data set are notable. First, EN-COX, EN-AFT-M, and EN-AFT-BJ each select genes not identified by any of the other three variable selection methods. In part, this is due to the noise of gene expression data. Such results are also indicative of the stochastic nature of the variable selection process.

Second, the methods based on the elastic net penalization do exhibit the grouping effect discussed in Section 2.4 while LASSO-Cox does not. For example, both EN-Cox and LASSO-Cox select gene 5442, but EN-Cox also selects gene 5301 which is moderately correlated with gene 5442 ($\rho = 0.43$). EN-AFT-M and EN-AFT-BJ each identify correlated gene expressions. For example, gene 5254 and gene 5296 (uniquely identified by EN-AFT-M) are correlated ($\rho = 0.57$). Likewise, genes 1671, 2154, and 5773 (uniquely identified by EN-AFT-BJ) are correlated ($\rho \geq 0.51$). With regard to LASSO-Cox, $\rho \leq 0.30$ for any two identified gene expressions.

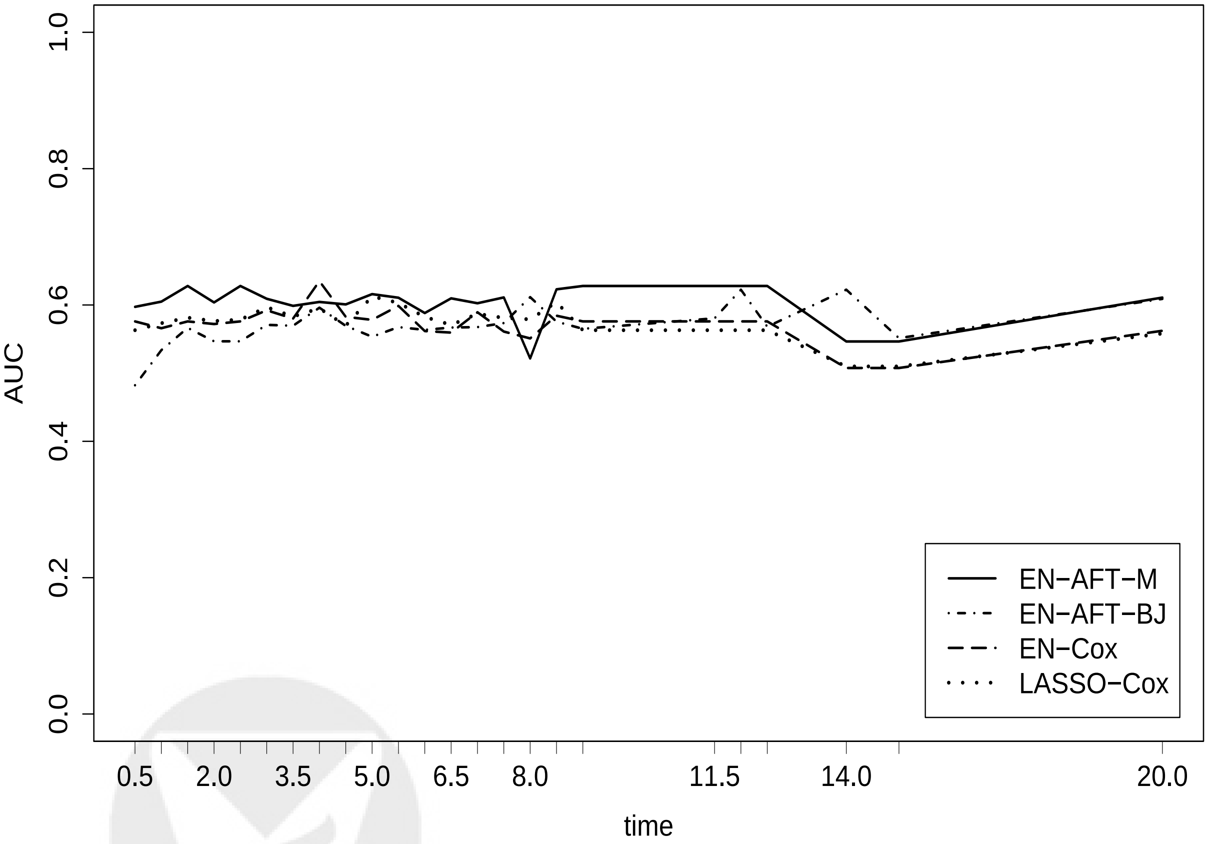

Figure 1: Comparison of predictive performance (area under the ROC curve, over time) for the Rosenwald DLBCL data set.

Of note, the LASSO-Cox AUC results are lower than those reported by Gui and Li (2005b). They had found that the LASSO-Cox variable selection method resulted in an AUC of around 0.67 for the first 10 years of follow up for the same DLBCL data set. Additionally, the tuning parameter identified by Gui and Li as optimal corresponded to a selection of four genes instead of the seven identified in the current analysis. The lower AUC scores

Table 1: *Variable selection results for LASSO-Cox, EN-Cox, EN-AFT methods for the DLBCL data set*

Method	Genes selected (index)
LASSO-Cox	80, 992, 1456, 3216, 5442, 6402, 6909
EN-Cox	952, 992, 1029, 1456, 3216, 3322, 5301, 5442, 6247, 6280, 6402, 6591, 6814, 6909, 7022, 7098
EN-AFT-BJ[1]	80, 876, 1188, 1409, 1456, 1671, 1969, 2154, 2582, 3322, 4007, 5773, 6215, 6402, 6519, 6743, 6921, 7069
EN-AFT-M[2]	290, 382, 1825, 1984, 2582, 3216, 3387, 3461, 4931, 5254, 5296, 6519, 7254

1: EN-AFT based on Buckley-James imputation
2: EN-AFT based on mean imputation

and the difference in variables selected in the current analysis may merely be due to use of different training and prediction sets. It is also possible that the differences are due to the difference in training set usage. Gui and Li appeared to have used the entire training set of 160 individuals to both select the tuning parameters and fit the model. It is possible, then, that the model may have been overfit. To avoid problems of overfitting in the current analysis, the tuning parameters were selected using half of the training set and the model fit was conducted using the other half.

4 Simulation Studies

In order to assess performance of EN-Cox and EN-AFT, several simulation studies were conducted under different data scenarios. For each scenario, covariate data was simulated following the strategy for generating gene expressions proposed by Gui and Li (2005b) which allows for correlation between certain subsets of the data. In essence, an $n \times n$ array B is initially generated from a uniform $U(-1.5, 1.5)$ distribution. A second set of data C can then be generated utilizing the normalized, orthogonal basis of the initial array. Gui and Li demonstrate that the maximum correlation between any two data vectors selected from B and C, respectively, can be specified during the data generation process. Implementation of this procedure can be conducted by prespecifying p_γ genes significantly associated with outcome. The gene expression data associated with these p_γ variables are drawn from the initial array B. The data for the remaining $p - p_\gamma$ variables are then drawn from the subsequent set of data C.

For each of the following three data scenarios, 100 simulations were conducted in which, for each simulation, data for $n = 150$ subjects and $p = 200$ gene expressions were generated. For each data set, subjects were randomly divided into two training sets of $n_t = 50$ each and one prediction set of $n_p = 50$. The first training set was utilized to select the tuning parameter(s) for the respective variable selection methods. Model fit was conducted using the second training set along with the identified tuning parameter(s). Additionally, it was assumed that the first $p_\gamma = 6$ genes were significantly associated with survival and that the remaining $p - p_\gamma$ were not.

It was first of interest to establish baseline performance for EN-Cox and EN-AFT in a relatively simple setting in which no correlation existed between any of the covariate vectors and where, on average, about 40% of the event times were censored. For this first data scenario, then, data for the first p_γ gene expression were drawn from a uniform $U(1.5, -1.5)$ distribution. That is, $\mathbf{x}_1, \ldots, \mathbf{x}_6$ were drawn from B. Following the approach of Gui and Li, data for the remaining $p - p_\gamma$ were drawn from the resultant C matrix. A Weibull distribution with scale parameter 2 and shape parameter 5 was used for the baseline hazard function and censoring times were generated using a uniform $U(2, 10)$ distribution, resulting in the desired level of censoring. Finally, half of the p_γ coefficient vector $\boldsymbol{\beta}_\gamma$ was generated from a uniform $U(-1, -0.1)$ distribution while the other half was generated from $U(0.1, 1)$. The remaining $p - p_\gamma$ coefficients were assigned a value of 0. Of note, use of the Weibull distribution ensures the appropriate use of the Cox proportional hazards model and the AFT model.

For the second data scenario, it was of interest to assess the grouping effect of EN-Cox and EN-AFT. That is, the performance of EN-Cox and EN-AFT was assessed for a scenario in which subsets of the p_γ variables were highly correlated. First, data for $\mathbf{x}_1$ and $\mathbf{x}_4$ (two of the six p_γ) were drawn from B (*i.e.*, from a uniform $U(-1.5, 1.5)$ distribution). Using the orthonormal basis of B, two sets of data, C_1 and C_2 were generated. For C_1, data were generated such that a number of the vectors in C_1 were highly correlated with vectors in B. Alternatively, vectors in B and C_2 were uncorrelated. Data for $\mathbf{x}_2$ and $\mathbf{x}_3$ were randomly drawn from the subset of C_1 highly correlated (*i.e.*, $0.85 < \rho < 0.95$) with $\mathbf{x}_1$. Data for $\mathbf{x}_5$ and $\mathbf{x}_6$ were randomly drawn from the subset of C_1 highly correlated with $\mathbf{x}_4$. The correlation between $\{\mathbf{x}_2, \mathbf{x}_3\}$ and $\{\mathbf{x}_5, \mathbf{x}_6\}$ was minimal ($|\rho| < 0.10$). Data for the remaining $p - p_\gamma$ variables were drawn from C_2. Hence, for this scenario, the p_γ genes were comprised of two groups of highly correlated variables. Also, $\boldsymbol{\beta}_\gamma$ was selected to reflect the high correlation between the p_γ gene subsets: $\boldsymbol{\beta}_j = 0.9$ for $j = 1, \ldots, 6$. The baseline hazard function and level of censoring were identical to Scenario 1.

Finally, it was of interest to assess the performance of EN-Cox and EN-AFT when an elevated level of censoring was present. For this third data scenario, gene expression data were generated as described above for Scenario 1. Likewise, the same $\boldsymbol{\beta}_\gamma$ parameter vector was used. The level of censoring, however, was increased to 60%.

For each of the three scenarios, performance of EN-Cox and EN-AFT was assessed in two ways. First, the relative frequency of selection of significant variables (*i.e.*, β_j, $j = 1, \ldots, 6$) was assessed. The average (across the remaining $p - p_\gamma$ variables) relative frequency of the selection of non-significant variables (*i.e.*, $\beta_j = 0$, $j = 7, \ldots, 200$) was also assessed. Variable selection and parameter estimation results for the three scenarios are presented in Tables 2, 3, 4, 5, 6, and 7. Second, predictive performance was assessed as described in Section 2.6. For each simulation, the AUC was calculated at each unique event time. Because unique times varied across simulations, the time scale was divided into equal sized

Table 2: *Variable selection results (frequency of selection) for LASSO-Cox, EN-Cox, EN-AFT methods for independent variables, 40% censoring*

	True value	LASSO-Cox	EN-Cox	EN-AFT-BJ[1]	EN-AFT-M[2]
β_1	0.957	0.82	0.84	1.00	1.00
β_2	-0.650	0.79	0.73	0.81	0.98
β_3	-0.539	0.78	0.70	0.73	0.98
β_4	-0.566	0.80	0.68	0.77	0.97
β_5	0.953	0.82	0.84	0.99	1.00
β_6	0.237	0.32	0.10	0.17	0.49
Average FP[3]		0.025	0.170	0.024	0.025

1: EN-AFT based on Buckley-James imputation
2: EN-AFT based on mean imputation
3: Average false positive: relative frequency (across all simulations) of selection of $\beta_j = 0$, averaged across all $j \in \{7, \dots, 200\}$

Table 3: *Variable selection results (mean and standard error of parameter estimates) for LASSO-Cox, EN-Cox, EN-AFT methods for independent variables, 40% censoring*

	True value	LASSO-Cox	EN-Cox	EN-AFT-BJ[1]	EN-AFT-M[2]
β_1	0.957	1.03 (0.027)	0.80 (0.030)	1.00 (0.034)	0.62 (0.011)
β_2	-0.650	-0.60 (0.028)	-0.47 (0.029)	-0.53 (0.035)	-0.34 (0.012)
β_3	-0.539	-0.46 (0.021)	-0.33 (0.021)	-0.37 (0.031)	-0.26 (0.010)
β_4	-0.566	-0.45 (0.024)	-0.37 (0.021)	-0.44 (0.032)	-0.28 (0.011)
β_5	0.953	1.07 (0.033)	0.84 (0.034)	1.00 (0.039)	0.62 (0.013)
β_6	0.237	0.11 (0.016)	0.14 (0.021)	0.15 (0.038)	0.07 (0.007)

1: EN-AFT based on Buckley-James imputation
2: EN-AFT based on mean imputation

"bins". The average AUC in each time-bin was then calculated. Figures 2, 3, 4 consist of the plotted average AUCs over time for each of the three scenarios. For comparison purposes, the same sets of data were also analyzed using the Gui and Li (2005b) LASSO-Cox procedure for censored data. To assess the effect of imputation method under the AFT model, separate analyses were conducted using the mean imputation method of Section 2.3 and the Buckley-James imputation method.

Results for the first scenario (*i.e.*, independent covariates, 40% censoring) are presented

Table 4: *Variable selection results (frequency of selection) for LASSO-Cox, EN-Cox, EN-AFT methods for correlated variables*[4]*, 40% censoring*

	True value	LASSO-Cox	EN-Cox	EN-AFT-BJ[1]	EN-AFT-M[2]
β_1	0.90	0.45	0.62	0.93	0.95
β_2	0.90	0.58	0.81	0.91	0.98
β_3	0.90	0.01	0.71	0.85	0.93
β_4	0.90	0.53	0.70	0.93	0.96
β_5	0.90	0.55	0.76	0.89	0.90
β_6	0.90	0.01	0.69	0.86	0.87
Average FP[3]		0.002	0.002	0.016	0.017

1: EN-AFT based on Buckley-James imputation

2: EN-AFT based on mean imputation

3: Average false positive: relative frequency (across all simulations) of selection of $\beta_j = 0$, averaged across all $j \in \{7, \ldots, 200\}$

4: Variables 1–6 are grouped into two sets: $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, $\{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$; within each set, variables are highly correlated ($\rho \in [0.85, 0.95]$)

Table 5: *Variable selection results (mean and standard error of parameter estimates) for LASSO-Cox, EN-Cox, EN-AFT methods for correlated variables*[3]*, 40% censoring*

	True value	LASSO-Cox	EN-Cox	EN-AFT-BJ[1]	EN-AFT-M[2]
β_1	0.90	1.05 (0.094)	0.93 (0.080)	0.95 (0.054)	0.50 (0.026)
β_2	0.90	1.55 (0.092)	1.01 (0.071)	1.32 (0.068)	0.74 (0.031)
β_3	0.90	0.73 (*)	0.89 (0.060)	1.32 (0.069)	0.71 (0.029)
β_4	0.90	1.20 (0.092)	0.96 (0.082)	0.99 (0.056)	0.61 (0.031)
β_5	0.90	1.41 (0.106)	1.04 (0.077)	1.36 (0.079)	0.77 (0.038)
β_6	0.90	0.19 (*)	1.01 (0.076)	1.29 (0.068)	0.77 (0.039)

1: EN-AFT based on Buckley-James imputation

2: EN-AFT based on mean imputation

3: Variables 1–6 are grouped into two sets: $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, $\{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$;

$*$: Insufficient information

Table 6: *Variable selection results (frequency of selection) for LASSO-Cox, EN-Cox, EN-AFT methods for independent variables, 60% censoring*

	True value	LASSO-Cox	EN-Cox	EN-AFT-BJ[1]	EN-AFT-M[2]
β_1	0.957	0.41	0.71	0.83	0.83
β_2	-0.650	0.30	0.41	0.42	0.39
β_3	-0.539	0.23	0.29	0.22	0.34
β_4	-0.566	0.27	0.38	0.28	0.39
β_5	0.953	0.45	0.74	0.77	0.90
β_6	0.237	0.09	0.14	0.09	0.13
Average FP[3]		0.022	0.035	0.028	0.030

1: EN-AFT based on Buckley-James imputation

2: EN-AFT based on mean imputation

3: Average false positive: relative frequency (across all simulations) of selection of $\beta_j = 0$, averaged across all $j \in \{7, \ldots, 200\}$

Table 7: *Variable selection results (mean and standard error of parameter estimates) for LASSO-Cox, EN-Cox, EN-AFT methods for independent variables, 60% censoring*

	True value	LASSO-Cox	EN-Cox	EN-AFT-BJ[1]	EN-AFT-M[2]
β_1	0.957	0.69 (0.067)	0.50 (0.034)	0.52 (0.037)	0.22 (0.015)
β_2	-0.650	-0.41 (0.046)	-0.29 (0.027)	-0.31 (0.039)	-0.13 (0.014)
β_3	-0.539	-0.21 (0.045)	-0.17 (0.028)	-0.26 (0.044)	-0.10 (0.013)
β_4	-0.566	-0.32 (0.052)	-0.25 (0.038)	-0.34 (0.055)	-0.12 (0.017)
β_5	0.953	0.63 (0.049)	0.49 (0.033)	0.56 (0.035)	0.21 (0.013)
β_6	0.237	0.29 (0.077)	0.13 (0.070)	0.28 (0.128)	0.11 (0.030)

1: EN-AFT based on Buckley-James imputation

2: EN-AFT based on mean imputation

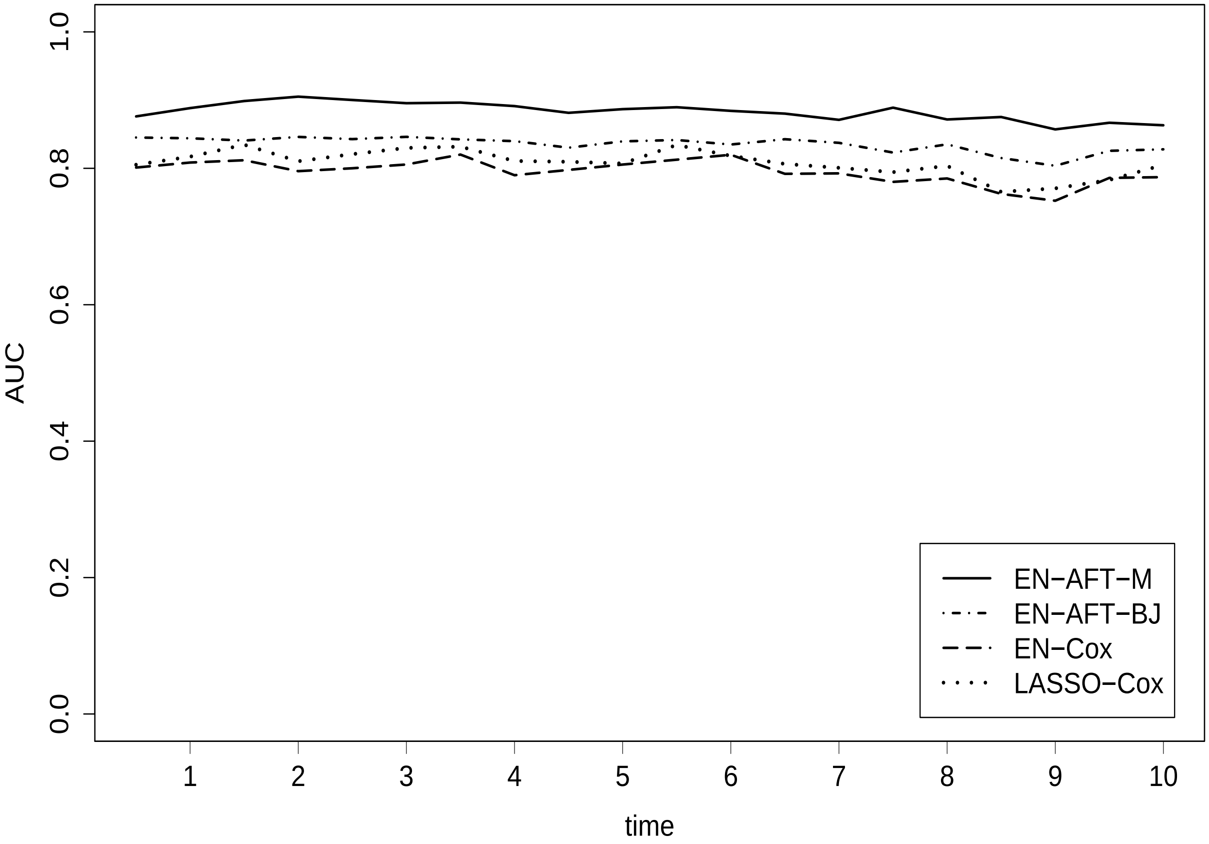

Figure 2: Comparison of predictive performance (area under the ROC curve, over time) for independent covariates, 40% censoring.

in Tables 2 and 3 and in Figure 2. For this simple scenario, the Cox-based methods seem roughly equivalent in terms of performance results, despite some apparent differences in variable selection and parameter estimation. EN-Cox has a slightly lower false positive rate (0.17 vs. 0.25) but also seems to underestimate the coefficient values in comparison to LASSO-Cox. Nevertheless, both EN-Cox and LASSO-Cox have a median AUC (across all times) of 0.80. With regard to the AFT-based methods, EN-AFT-M appears to underestimate the coefficient values in comparison to EN-AFT-BJ. However, the standard errors of EN-AFT-M are smaller than those of the other three methods. Both EN-AFT-M and EN-AFT-BJ appear to slightly outperform the Cox-based models in this setting, more fre-

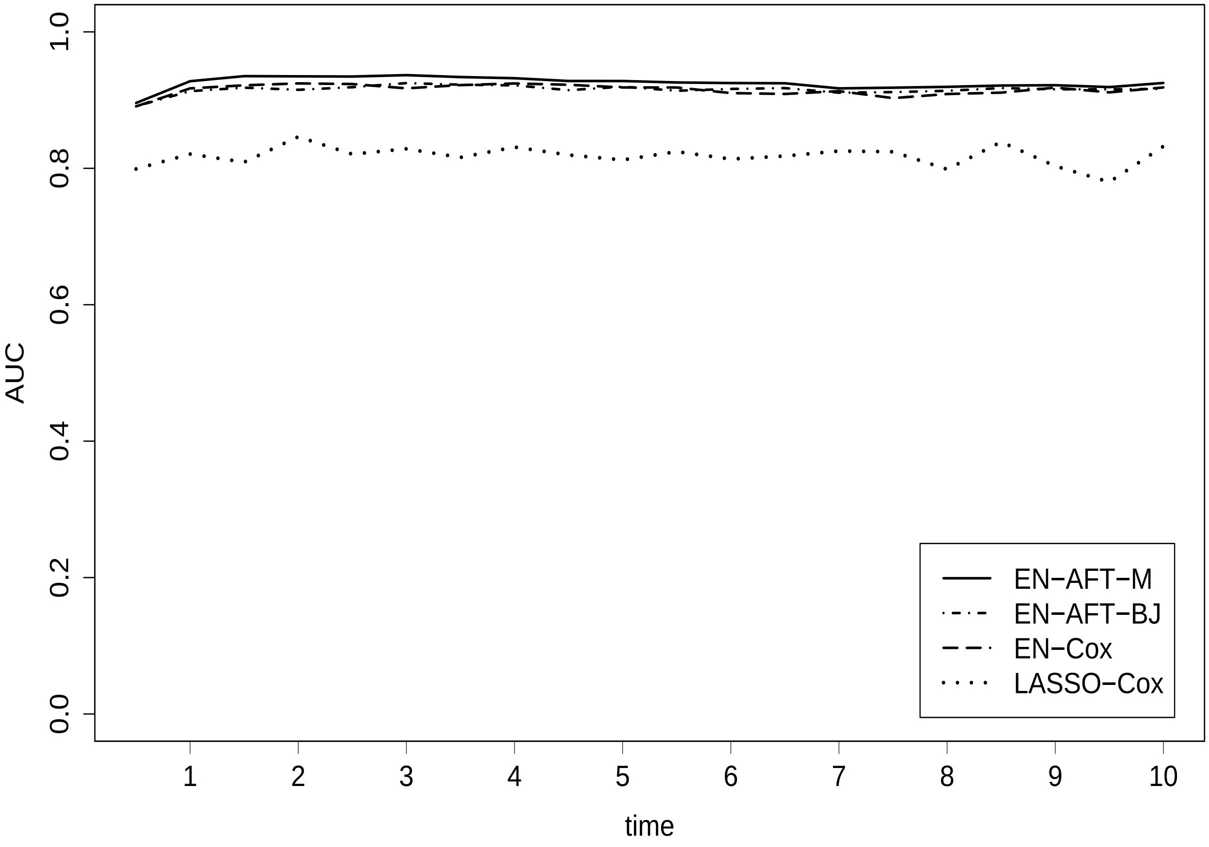

Figure 3: Comparison of predictive performance (area under the ROC curve, over time) for correlated subsets of covariates (*i.e.*, grouping effect), 40% censoring.

quently identifying variables of interest. The median AUC for EN-AFT-M is 0.89 and is 0.84 for EN-AFT-BJ.

Results for the second scenario (*i.e.*, grouped covariates with high correlation within groups, 40% censoring) are presented in Tables 4 and 5 and in Figure 3. With regard to variable selection (Table 4), the AFT-based selection methods exhibit the highest accuracy, followed by EN-Cox and then LASSO-Cox. The LASSO-Cox does not exhibit the grouping effect but instead appears to select one of several highly correlated variables and ignores the others. For example, in about half the simulations, LASSO-Cox selects β_1,

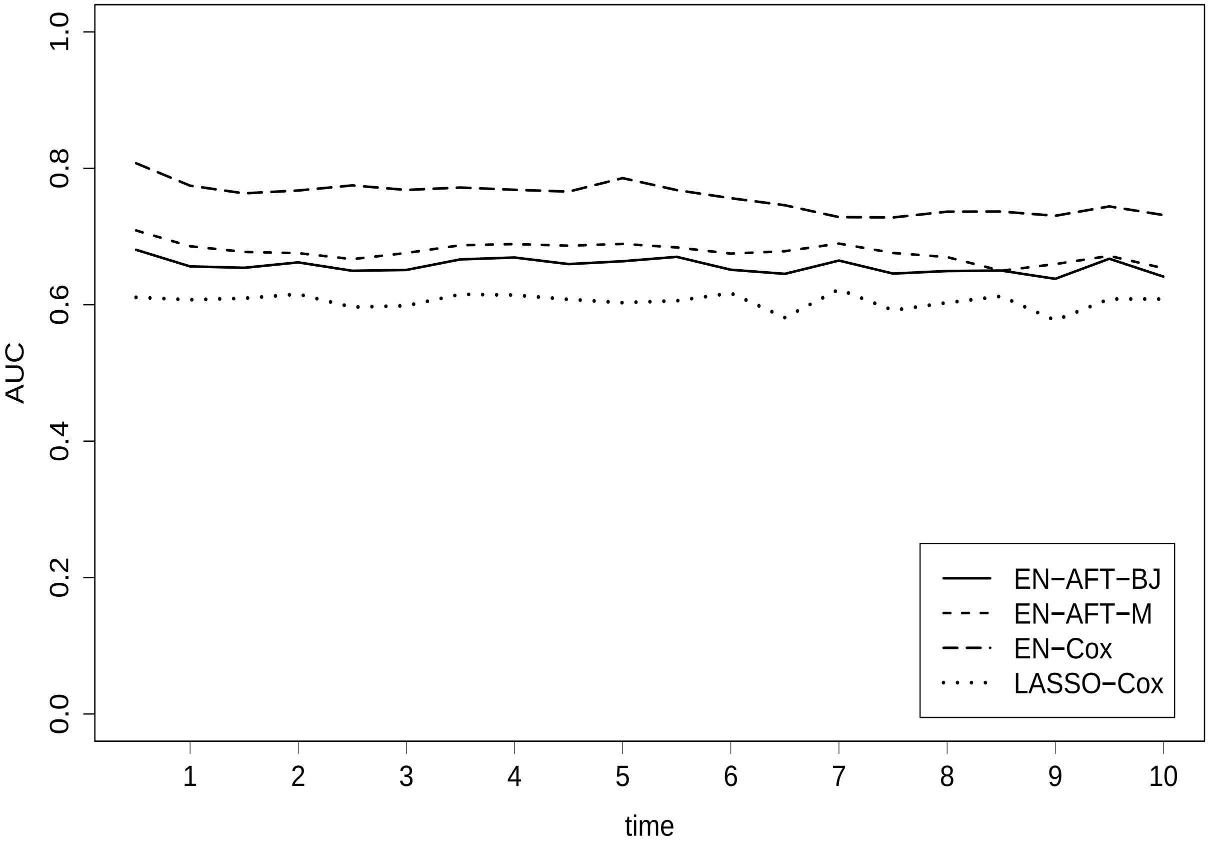

Figure 4: Comparison of predictive performance (area under the ROC curve, over time) for independent covariates, 60% censoring.

ignoring β_2 and β_3 whereas in the remaining simulations, LASSO-Cox selects β_2, ignoring β_1 and β_3. A similar pattern is observed for the second group of correlated variables, β_4, β_5, and β_6. As in the first scenario, EN-AFT-M appears to underestimate the coefficient values, but exhibits less variability about those estimates. Regarding predictive performance (Figure 3), all three EN-AFT-M, EN-AFT-BJ and EN-Cox perform well both with a median AUC (across all times) of 0.92. The over-time average AUC of LASSO-Cox in this setting is 0.82.

Results for the third scenario (*i.e.*, independent covariates, 60% censoring) are presented in Table 6 and Figure 4. For this scenario in which a high level of censoring is present, the three elastic net methods outperform LASSO-Cox in both variable selection accuracy and in predictive performance. Interestingly, while the three elastic net methods are roughly equivalent with regard to variable selection, EN-Cox (median AUC: 0.76) appears to slightly outperform the two AFT-based methods (EN-AFT-M median AUC: 0.68, EN-AFT-BJ median AUC: 0.66) in terms of predictive performance. The poorer predictive performance of the AFT-based methods may be due, in part, to the fact that the required AFT model imputation is based on fewer observed events and is therefore less accurate. Hence, while the AFT methods still correctly identify the variables of interest, the estimation of the coefficient values associated with those variables is less precise. The parameters estimates for EN-AFT-BJ (see Table 7), for example, are approximately equal to those of EN-Cox, but with higher standard errors. The EN-AFT-M estimates, on the other hand, have smaller standard errors but are more strongly biased. The median AUC of LASSO-Cox is 0.61 in this setting.

5 Computational Efficiency

Use of the elastic net penalty leads to computationally efficient algorithms. Typical run times (3.2Ghz Xeon Linux workstation) for EN-AFT-M, EN-AFT-BJ, EN-Cox, and LASSO-Cox are listed in Table 8 for various data set dimensionalities.

Note that the run times listed in Table 8 are for fixed tuning parameters and that differences in run times are even more pronounced when time of cross-validation is included. For example, a typical total run-time (cross-validation and model fitting) for $N = 150$ and $p = 200$ for EN-AFT-M is 25.0 seconds whereas the EN-AFT-BJ time is 2716.6 seconds. For $N = 150$ and $p = 1000$, the total run time for EN-AFT-M is 47.6 seconds and is 106280.7 seconds for EN-AFT-BJ.

Table 8: *Comparison of computation times for LASSO-Cox, EN-Cox, EN-AFT methods (in seconds)*

p	N	LASSO-Cox	EN-Cox	EN-AFT-BJ[1]	EN-AFT-M[2]
200	50	164.57	62.65	0.98	0.05
200	100	200.04	110.41	1.39	0.06
200	150	648.53	133.61	5.63	0.08
500	150	1107.85	217.95	6.33	0.10
1000	150	1134.76	508.79	11.02	0.29

1: EN-AFT based on Buckley-James imputation
2: EN-AFT based on mean imputation

6 Discussion

Adaptation of the elastic net penalization criterion for use in high-dimensional and low-sample size censored data settings leads to computational efficient variable selection methods with good predictive performance. Through simulation studies, EN-Cox and EN-AFT were shown to perform well in comparison to the Gui and Li (2005b) LASSO-Cox approach in simple settings with low censoring and independent covariates. The two methods were also shown to outperform LASSO-Cox in settings with a high degree of censoring and in settings where sets of highly correlated variables were present. The EN-AFT approach entailing mean imputation was also shown to outperform the approach based on Buckley-James imputation in terms of both predictive performance and computational efficiency.

It should be noted that the presented models can also be adapted to situations in which it is of interest to assign separate penalty functions to different coefficients or groups of coefficients. That is, equation (2.1) can be extended to

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_2 \sum_{j=1}^{p} W_{2j}\beta_j^2 + \lambda_1 \sum_{j=1}^{p} W_{1j}|\beta_j|, \tag{6.1}$$

where the W_{mj}, $m = 1, 2$ are covariate-specific weights. For example, if it is *a-priori* known that a group of genes are associated with outcome and identification of additional genetic regions is desired, optimization in EN-Cox and EN-AFT can be modified to allow separate penalization of the two groups. For the current analyses, the penalty term did not vary across coefficients.

Several features of the EN-Cox and EN-AFT implementations may warrant further investigation. As noted, Segal (2005) proposed methods for improving the computational efficiency of the LARS-based LASSO-Cox. The EN-Cox is LARS-based and while it was shown to perform efficiently in comparison to the LARS-based LASSO-Cox of Gui and Li (2005b) in the current analyses, improvements might be made.

It may also be of interest to obtain standard error estimates for the EN-Cox or EN-AFT regression coefficients. One possible approach is based on an adaptation of the LASSO local quadratic approximation (LQA) proposed by Fan and Li (2001) (see also Zou, 2006). First, assume the nonzero elements of $\boldsymbol{\beta}$ have been identified, perhaps through an initial EN-Cox or EN-AFT analysis. Let $\boldsymbol{\beta}_0$ be an estimate of $\boldsymbol{\beta}$ (presumably close to $\boldsymbol{\beta}$), again perhaps obtained through an initial EN-Cox or EN-AFT analysis. Equation (2.2) can be rewritten as

$$\hat{\boldsymbol{\beta}} = \text{arg min}_{\boldsymbol{\beta}} \left[\boldsymbol{\beta}' \left(\frac{\mathbf{X}'\mathbf{X} + \lambda_2 \mathbf{I}}{1+\lambda_2} \right) \boldsymbol{\beta} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \lambda_1 \left(\sum_{j=1}^{p} |\beta_{j0}| + \frac{1}{2|\beta_{j0}|}(\beta_j^2 - \beta_{j0}^2)\right) \right], \quad (6.2)$$

where $\mathbf{y}$ and $\mathbf{X}$ are replaced with $\tilde{\mathbf{z}}$ and $\tilde{\mathbf{X}}$ for EN-Cox and where $\mathbf{y}$ is replaced with $\mathbf{y}^*$ for EN-AFT. Let $\boldsymbol{\beta}_m$ consist of the m nonzero elements of $\boldsymbol{\beta}$ and let $\mathbf{X}_m$ consist of the corresponding columns of $\mathbf{X}$. By differentiating (6.2), a closed form solution for $\boldsymbol{\beta}$ can be written as

$$\hat{\boldsymbol{\beta}} = (1+\lambda_2)(\mathbf{X}_m'\mathbf{X}_m + \lambda_2\mathbf{I} + \lambda_1\boldsymbol{\Sigma}(\boldsymbol{\beta}_0))^{-1}\mathbf{X}_m'\mathbf{y}, \quad (6.3)$$

where $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0) = \text{diag}(\frac{1}{\beta_1}, \ldots, \frac{1}{\beta_m})$. Equation (6.3) can then be utilized to obtain the sandwich estimator for the covariance matrix for $\boldsymbol{\beta}_m$.

7 Appendix A: Proofs

Proposition 1: Assume that $\hat{\beta}_i \neq \hat{\beta}_j$. Define estimator $\hat{\boldsymbol{\beta}}^*$: let $\hat{\beta}^*_k = \hat{\beta}_k$ for all $k \neq i, j$, otherwise let $\hat{\beta}^*_k = p\hat{\beta}_i + (1-p)\hat{\beta}_j$ for $p = 1/2$. Since $\mathbf{x}_i = \mathbf{x}_j$, clearly $T\mathbf{x}_i = \tilde{\mathbf{x}}_i = \tilde{\mathbf{x}}_j = T\mathbf{x}_j$, $\tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}^* = \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}$, and $|\tilde{\mathbf{z}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}^*|^2 = |\tilde{\mathbf{z}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}|^2$. However, because the elastic net penalization function $f(\boldsymbol{\beta}) = \lambda_2 \sum_{k=1}^p \beta_k^2 + \lambda_1 \sum_{k=1}^p |\beta_k|$ is strictly convex, it is the case that

$$f(\hat{\boldsymbol{\beta}}^*_{i,j}) \quad = \quad f(p\hat{\beta}_i + (1-p)\hat{\beta}_j) \quad < \quad pf(\hat{\beta}_i) + (1-p)f(\hat{\beta}_j) \quad < \quad f(\hat{\boldsymbol{\beta}}_{i,j}).$$

Because $f(\hat{\boldsymbol{\beta}}^*) = f(\hat{\boldsymbol{\beta}})$ for $i \neq j$, and because $f(.)$ is additive, $f(\hat{\boldsymbol{\beta}}^*) < f(\hat{\boldsymbol{\beta}})$ and it therefore cannot be the case that $\hat{\boldsymbol{\beta}}$ is a minimizer. Hence, $\hat{\beta}_i = \hat{\beta}_j$.

Proposition 2: By definition,

$$\frac{\partial L(\lambda_1, \lambda_2, \boldsymbol{\beta})}{\partial \beta_k}|_{\beta=\hat{\beta}} = 0 \qquad \text{for } \hat{\beta}_k \neq 0. \tag{7.1}$$

Also, note that

$$L(\lambda_1, \lambda_2, \hat{\boldsymbol{\beta}}) \leq L(\lambda_1, \lambda_2, \boldsymbol{\beta} = \mathbf{0}). \tag{7.2}$$

By (7.1) (for non-zero $\hat{\beta}_i$ and $\hat{\beta}_j$),

$$-2\tilde{\mathbf{x}}_i'(\tilde{\mathbf{z}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}) + \lambda_1 \text{sign}(\hat{\beta}_i) + 2\lambda_2\hat{\beta}_i = 0,$$

and

$$-2\tilde{\mathbf{x}}_j'(\tilde{\mathbf{z}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}) + \lambda_1 \text{sign}(\hat{\beta}_j) + 2\lambda_2\hat{\beta}_j = 0.$$

Hence,

$$\hat{\beta}_i - \hat{\beta}_j = \frac{1}{\lambda_2}(\tilde{\mathbf{x}}_j' - \tilde{\mathbf{x}}_i')(\tilde{\mathbf{z}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}) \leq \frac{1}{\lambda_2}|\tilde{\mathbf{x}}_j - \tilde{\mathbf{x}}_i||\tilde{\mathbf{z}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}|,$$

where $|\mathbf{x}| = \sqrt{\mathbf{x}'\mathbf{x}}$. By (7.2),

$$|\tilde{\mathbf{z}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}|^2 \leq |\tilde{\mathbf{z}}|^2,$$

since $\tilde{\mathbf{z}}$ is centered. Hence,

$$\frac{|\hat{\beta}_i - \hat{\beta}_j|}{|\tilde{\mathbf{z}}|} \leq \frac{1}{\lambda_2}|\tilde{\mathbf{x}}_j - \tilde{\mathbf{x}}_i|\frac{|\tilde{\mathbf{z}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}|}{|\tilde{\mathbf{z}}|} \leq \frac{1}{\lambda_2}|\tilde{\mathbf{x}}_j - \tilde{\mathbf{x}}_i| \leq \frac{1}{\lambda_2}\sqrt{2(1 - \mathbf{x}_i\mathbf{A}\mathbf{x}_j)},$$

where $\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_j = \mathbf{x}_i\mathbf{A}\mathbf{x}_j$ is the correlation between standardized variables $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$.

References

Akaike, H. (1973). Information theory and the extension of the maximum likelihood principle. In Petrov, V. and Csaki, F. (Eds.). *Proceedings of the Second International Symposium on Information Theory*, Budapest Akailseoniai-kiudo, 267–281.

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Ann. Statistics* **22**, 1299–1327.

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

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

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

Dai, H., van't Veer, L., Lamb, J., He, Y. D., Mao, M., Fine, B. M., Bernards, R., van de Vijver, M., Deutsch, P., Sachs, A., Stoughton, R., Friend, S. (2005). A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. *Cancer Res.* **65**, 4059–4066.

Datta, S. (2005). Estimating the mean life time using right censored data. *Statistical Methodology* **2**, 65–69.

Datta, S., Le-Rademacher, J., Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics* **63**, 259–271.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression. *Ann. Statistics* **32**, 407–499.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *JASA* **96**, 1348–1359.

Frank, I. E., and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–148.

Friedman, J. H., and Popescu, B. E. (2004). Gradient directed regularization for linear regression and classification. Stanford University, Department of Statistics. Technical Report.

Gui, J., and Li, H. (2005a). Threshold gradient descent method for censored data regression, with applications in pharmacogenomics. *Pacific Symposium on Biocomputing* **10**, 272–283.

Gui, J., and Li, H. (2005b). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21**, 3001–3008.

Hastie, T., and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* **46**, 1005–1016.

Heagerty, P. J., Lumley, T., Pepe, M. S. (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. *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., Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62**, 813–820.

Hunter, D., and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statistics* **33**, 1617–1642.

Kaplan, E. L., and Meier, P. (1958). Nonparametric estimation from incompete observations. *JASA* **53**, 457–481.

Kimeldorf, G. S., and Wahba, G. (1971). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Statistics* **2**, 495–502.

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.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R., Gascoyne, R. D., Muller-Hermelink, K., Smeland, E. B., Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-Cell lymphoma. *N. Eng. J. Med.* **346**, 1937–1947.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statistics* **6**, 461–464.

Segal, M. R. (2005). Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisted. *Biostatistics* **7**, 268–285.

Sha, N., Tadesse, M. G., Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* **22**, 2262–2268.

Stute, W. (1993). Consistent estimation under random censorship when covariables are available. *Journal of Multi- variate Analysis* **45**, 89–103.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics* **23**, 461–471.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc. B* **58**, 267–288.

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

Verwij, P., and Van Houwelingen, H. (1993). Cross validation in survival analysis. *Statist. Med.*, **12**, 2305–2314[ISI].

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

Wang, S., Nan B., Zhu, J., Beer, D. G. (2006). Doubly penalized Buckley-James method for survival data with high-dimensional covariates *http://www.bepress.com/umichbiostat/paper62*.

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J.R. Statist. Soc. B* **67**, 301–320.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *JASA* **101**, 1418–1429.