2009-02-11

Survival Analysis with High-Dimensional Covariates: An Application in Microarray Studies

David Engler
david_engler@byu.edu

Yi Li

Original Publication Citation

Engler, David and Li, Yi (29) "Survival Analysis with High-Dimensional Covariates: An Application in Microarray Studies," Statistical Applications in Genetics and Molecular Biology: Vol. 8 : Iss. 1, Article 14.

Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 14

Survival Analysis with High-Dimensional Covariates: An Application in Microarray Studies

David Engler* Yi Li†

*Brigham Young University, engler@byu.edu

†Harvard University and Dana Farber Cancer Institute, yili@jimmy.harvard.edu

Survival Analysis with High-Dimensional Covariates: An Application in Microarray Studies

David Engler and Yi Li

Abstract

Use of microarray technology often leads to high-dimensional and low-sample size (HDLSS) data settings. A variety of 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, adaptations of the elastic net approach are 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.

KEYWORDS: survival analysis, microarray, elastic net, variable selection

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.

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 et al., 1973) and BIC (Schwarz, 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). Implementation of a number of these methods, however is not feasible in HDLSS environments.

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.

Several authors have proposed variable selection methods for HDLSS time-to-event data under the Cox proportional hazards model (Cox, 1972). For example, Cox-based methods utilizing kernel transformations (Li and Luan, 2003), threshold gradient descent minimization (Gui and Li, 2005a), and lasso penalization (see Gui and Li, 2005b; Segal, 2005; Park and Hastie, 2007) have been proposed.

Likewise, a few authors have proposed variable selection methods based on accelerated failure time models (see Wei, 1992). Methods based on the lasso penalization and the threshold gradient descent (Huang et al., 2006) have been proposed as well as an approach based on Bayesian variable selection (Sha et al., 2006).

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, 2005a; Gui and Li, 2005b). 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. It should be noted that the adaptive lasso, a recent improvement to the lasso, has been proposed by Zhang and Lu (2007) for censored data. While the approach overcomes a number of the drawbacks of lasso, use of the adaptive lasso may not be appropriate in high-dimensional data settings without reliance upon ridge regression (see Zhang and Lu, 2007; Lu and Zhang, 2007).

Modification of the elastic net penalization approach 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 (2005), the elastic net is computationally efficient. To date, the only attempt to employ the elastic net penalization approach to HDLSS censored data under the AFT model (Wang et al., 2008) employs an imputation approach based on the Buckley and James algorithm (Buckley and James, 1979). However, the Buckley-James approach entails an iterative least squares procedure that is known to suffer from convergence problems (see Wu and Zubovic, 1995) 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 propor-

DOI: 10.2202/1544-6115.1423

tional hazards model and utilizes modifications of the algorithms proposed by Tibshirani (1997) and Gui and Li (2005b). Second, an accelerated failure time 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.

2 Methods

2.1 Elastic Net

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

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

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 (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

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

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,

$y_i = \min(t_i, c_i)$ 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 $\lambda(t_i) = \lambda_0(t_i)\exp(\boldsymbol{\beta}'\mathbf{x_i})$, 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

$$\ell(\beta) = \frac{1}{n}\sum_{r\in D} \ln\left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{(r)})}{\sum_{j\in R_r}\exp(\boldsymbol{\beta}'\mathbf{x}_j)}\right), \tag{3}$$

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 (see 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 (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 $(\mathbf{z}_0 - \boldsymbol{\eta})'\mathbf{A}(\mathbf{z}_0 - \boldsymbol{\eta})$.

Modifying the approach of Gui and Li (2005b), 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 $\mathbf{Q} = \mathbf{A^{1/2}}$. 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). That is, the optimal $\hat{\boldsymbol{\beta}}$ is formulated as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left[\boldsymbol{\beta}'\left(\frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + \lambda_2\mathbf{I}}{1+\lambda_2}\right)\boldsymbol{\beta} - 2\tilde{\mathbf{z}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \lambda_1\sum_{j=1}^{p}|\beta_j|\right]. \tag{4}$$

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

1. Set tuning parameters and initialize $\hat{\boldsymbol{\beta}} = \mathbf{0}$.

DOI: 10.2202/1544-6115.1423

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 $\tilde{\mathbf{X}} = \mathbf{QX}$ and $\tilde{\mathbf{z}}_0 = \mathbf{Qz}_0$.
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 $\mathbf{x}_i$:

$$h(t_i) = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{5}$$

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 (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

$$h(y_i^*) = (\delta_i)h(y_i) + (1 - \delta_i)\{\hat{S}(y_i)\}^{-1} \sum_{t_{(r)} > t_i} h(t_{(r)})\Delta\hat{S}(t_{(r)}), \tag{6}$$

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. 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 (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. Derivation of Proposition 1 and 2 is provided in the Appendix.

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 $\hat{\beta}_i = \hat{\beta}_j$.

DOI: 10.2202/1544-6115.1423

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 $\text{sign}(\hat{\beta}_i) = \text{sign}(\hat{\beta}_j)$. Then for fixed λ_1 and λ_2

$$\frac{|\hat{\beta}_i - \hat{\beta}_j|}{|\tilde{\mathbf{z}}|} \leq \frac{\sqrt{2\{1 - (\mathbf{x}_i' \mathbf{A} \mathbf{x}_j)\}}}{\lambda_2}, \tag{7}$$

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 3.2.

2.5 Selection of Tuning Parameters

The elastic net requires the selection of two tuning parameters, λ_1 and λ_2. 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. Tuning parameter selection can be implemented through use of cross-validation methods over a rough grid of candidate values for λ_2 and k. In the current setting, selection of λ_2 and k under both EN-Cox and EN-AFT is conducted through use of a cross validation score (CVS) (Huang, 2006; see also Verwij and Van Houwelingen (1993), Huang

and Harrington (2002)):

$$CVS(\mathbf{X}, \lambda_2, k) = \ell(\mathbf{X}, \hat{\boldsymbol{\beta}}_{\lambda_2,k,-(i)}) - \ell(\mathbf{X}_{-(i)}, \hat{\boldsymbol{\beta}}_{\lambda_2,k,-(i)}), \tag{8}$$

where $\hat{\boldsymbol{\beta}}_{\lambda_2,k,-(i)}$ consists of the coefficient estimates (for a given variable selection approach) obtained while excluding the i^{th} subject for fixed values λ_2 and k and where $\mathbf{X}_{-(i)}$ denotes the complete data set, absent the i^{th} subject. Under the Cox-based models, the function $\ell(.)$ represents the negative log partial likelihood (3). Under the AFT-based models, $\ell(.)$ represents the AFT objective function (*i.e.*, $\sum_i[(y_i^* - \hat{\boldsymbol{\beta}}'\mathbf{x}_i)'(y_i^* - \hat{\boldsymbol{\beta}}'\mathbf{x}_i)]$). Values of λ_2 and k that correspond to the minimization of (8) are identified and selected.

Potentially viable alternatives to the above approach include, but are not limited to, BIC (see Wang et al., 2007) as well as the approaches outlined by Zhang and Lu (2007) and by Wang et al. (2008).

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

$$\text{sensitivity}(c, t|f(X)) = P[f(X) > c|\delta(t) = 1] \tag{9}$$

$$\text{specificity}(c, t|f(X)) = P[f(X) \leq c|\delta(t) = 0], \tag{10}$$

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.

2.7 Software

Analyses were performed using the R software package (http://www.r-project.org). The R implementation of the Cox-based and AFT-based elastic net models presented in this paper is available at http://statweb.byu.edu/engler/ENET.

DOI: 10.2202/1544-6115.1423

3 Results

3.1 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. 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).

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) 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 (Wang et al., 2008). 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.

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

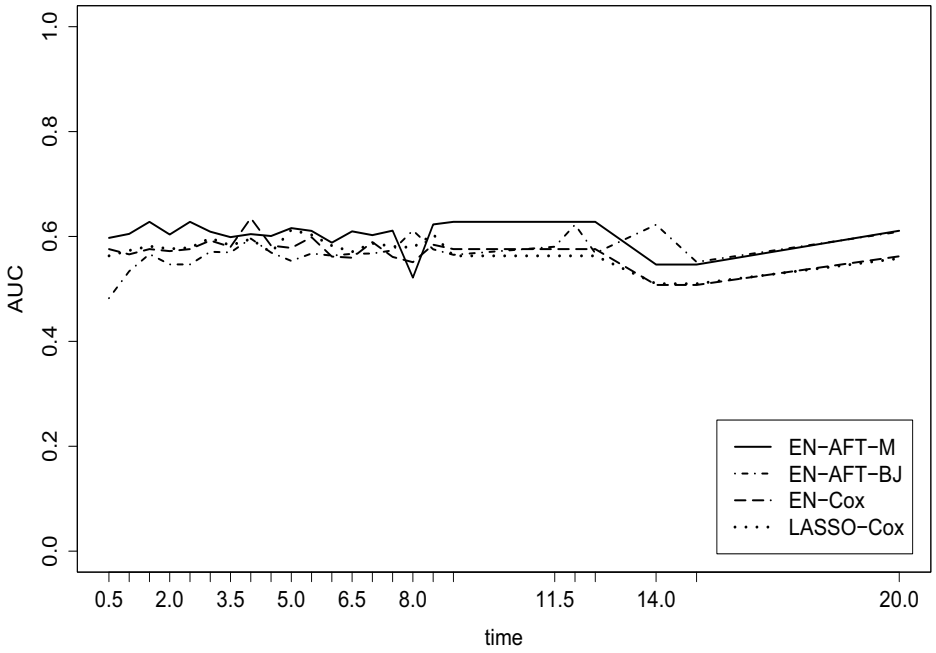

Figure 1: Comparison of predictive performance (area under the ROC curve, over time) for the Rosenwald DLBCL data set.

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.

In summary, both EN-Cox and EN-AFT-M (EN-AFT based on mean imputation) perform as well or better than the lasso-based method and EN-AFT-BJ (EN-AFT based on the Buckley-James imputation) in terms of predictive power. It is additionally important to note that the elastic-net based methods

DOI: 10.2202/1544-6115.1423

are much more computationally efficient than their Cox-based and AFT-based counterparts (see Section 3.3); completion of the Lasso-Cox method exceeded several days while EN-AFT-M (including parameter selection through cross-validation) completed in well under an hour.

3.2 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 (2005b) 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. 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 results for the three scenarios are presented in Tables 1, 2, and 3. Listed in each table are the non-zero coefficient values along with the relative frequency of selection across all simulations for these coefficients. The average frequency of selection (across all simulations and across all zero-valued coefficients) of the remaining coefficients is also listed.

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 "bins". The average AUC in each time-bin was then calculated. Figure 3.2 contains the plotted average AUCs over time for each of the three

DOI: 10.2202/1544-6115.1423

Table 1: Variable selection results (frequency of selection) for LASSO-Cox (L-Cox), EN-Cox, EN-AFT (BJ: Buckley-James imputation, M: mean imputation) methods for independent variables, 40% censoring. The column "Actual" dentotes the true parameter value. The remaining columns consist of the frequency of selection (across all simulations) by the respective methods. Average false positive (FP): relative frequency (across all simulations) of selection of $\beta_j = 0$, averaged across all $j \in \{7, \ldots, 200\}$.

	Actual	L-Cox	EN-Cox	EN-AFT-BJ	EN-AFT-M
β_1	0.96	0.82	0.84	1.00	1.00
β_2	-0.65	0.79	0.73	0.81	0.98
β_3	-0.54	0.78	0.70	0.73	0.98
β_4	-0.57	0.80	0.68	0.77	0.97
β_5	0.95	0.82	0.84	0.99	1.00
β_6	0.24	0.32	0.10	0.17	0.49
FP		0.025	0.170	0.024	0.025

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 of Wang et al. (2008).

Results for the first scenario (*i.e.*, independent covariates, 40% censoring) are presented in Table 1 and in Figure 3.2 (under "Scenario 1"). For this simple scenario, the Cox-based methods seem roughly equivalent in terms of performance results; both EN-Cox and LASSO-Cox have a median AUC (across all times) of 0.80. With regard to the AFT-based methods, both EN-AFT-M and EN-AFT-BJ appear to outperform the Cox-based models in this setting, more frequently identifying variables of interest. The AFT approach based on mean imputation (EN-AFT-M) performs particularly well. For coefficients with moderate or high absolute effects ($\boldsymbol{\beta}_{1-5}$), the mean frequency of selection of EN-AFT-M is 0.986. With regard to the selection of the remaining non-zero coefficient (β_6), EN-AFT-M outperforms the Cox-based methods as well as the method based on Buckley-James imputation (EN-AFT-BJ).

Results for the second scenario (*i.e.*, grouped covariates with high correlation within groups, 40% censoring) are presented in Table 2 and in Figure 3.2 (under "Scenario 2"). With regard to variable selection (Table 2), the AFT-based selection methods exhibit the highest accuracy, followed by EN-Cox and

Table 2: Variable selection results (frequency of selection) for LASSO-Cox (L-Cox), EN-Cox, EN-AFT (BJ: Buckley-James imputation, M: mean imputation) methods for correlated variables, 40% censoring. 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 ($\rho \in [0.85, 0.95]$). The column "Actual" dentotes the true parameter value. The remaining columns consist of the frequency of selection (across all simulations) by the respective methods. Average false positive (FP): relative frequency (across all simulations) of selection of $\beta_j = 0$, averaged across all $j \in \{7, \ldots, 200\}$.

	Actual	L-Cox	EN-Cox	EN-AFT-BJ	EN-AFT-M
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
FP		0.002	0.002	0.016	0.017

Table 3: Variable selection results (frequency of selection) for LASSO-Cox (L-Cox), EN-Cox, EN-AFT (BJ: Buckley-James imputation, M: mean imputation) methods for independent variables, 60% censoring. The column "Actual" dentotes the true parameter value. The remaining columns consist of the frequency of selection (across all simulations) by the respective methods. Average false positive: relative frequency (across all simulations) of selection of $\beta_j = 0$, averaged across all $j \in \{7, \ldots, 200\}$.

	Actual	L-Cox	EN-Cox	EN-AFT-BJ	EN-AFT-M
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
FP[3]		0.022	0.035	0.028	0.030

DOI: 10.2202/1544-6115.1423

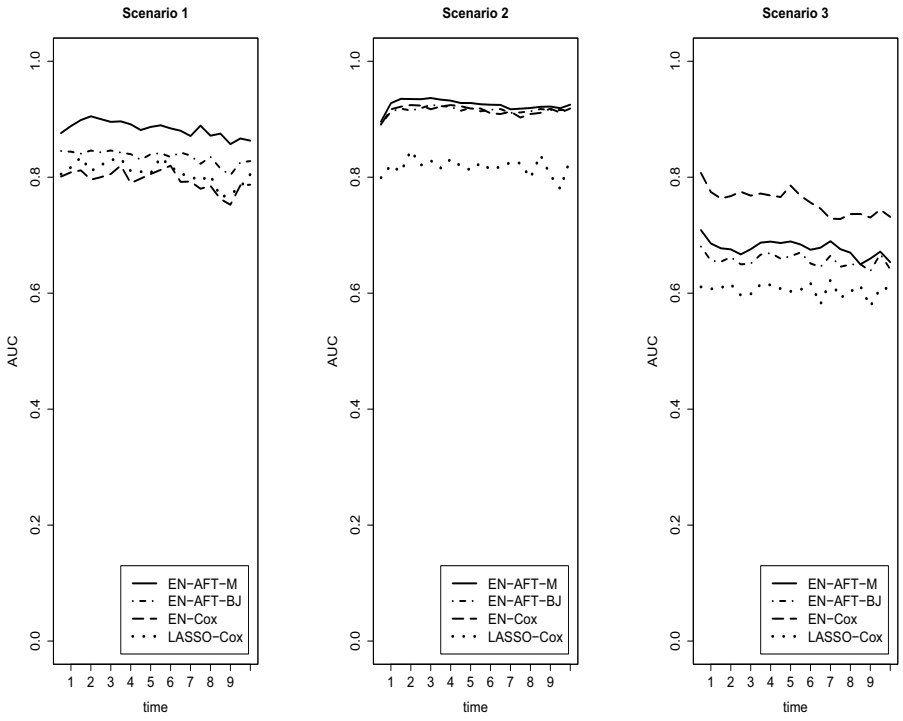

Figure 2: Comparison of predictive performance (area under the ROC curve, over time) for Scenario 1: independent covariates, 40% censoring, Scenario 2: correlated subsets of covariates (*i.e.*, grouping effect), 40% censoring. Scenario 3: independent covariates, 60% censoring

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, 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 setting, the performance of EN-AFT-M with regard to frequency of variable selection is superior to the Cox-based methods and to EN-AFT-BJ. Regarding predictive performance (Figure 3.2), 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 3 and Figure 3.2 (under "Scenario 3"). For this scenario in which a high level of censoring is present, the three elastic net

Table 4: Comparison of computation times for LASSO-Cox (L-Cox), EN-Cox, EN-AFT (BJ: Buckley-James imputation, M: mean imputation) methods (in seconds). p: number of variables, N: number of subjects.

p	N	L-Cox	EN-Cox	EN-AFT-BJ	EN-AFT-M
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

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 imputation in the AFT models is based on fewer observed events and is therefore less accurate.

In summary, then, EN-Cox performs as well or better than LASSO-Cox in each of the three scenarios. The improvement of EN-Cox is particularly notable when correlated covariates are present. Moreover, the computational efficiency of EN-Cox exceeds that of LASSO-Cox. With regard to the AFT-based methods, EN-AFT-M performs as well or better than EN-AFT-BJ in all three scenarios, particularly wtih regard to frequency of variable selection. Additionally, the improvement in computational efficiency is substantial.

3.3 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 4 for various data set dimensionalities.

Note that the run times listed in Table 4 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 $p = 1000$,

DOI: 10.2202/1544-6115.1423

the total run time for EN-AFT-M is 47.6 seconds and is 106280.7 seconds for EN-AFT-BJ.

4 Discussion

Adaptation of the elastic net penalization criterion for use in high-dimensional and low-sample size censored data settings leads to computationally 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 (Wang et al., 2008) in terms of both frequency of variable selection and computational efficiency.

Several features of the EN-Cox and EN-AFT implementations may warrant further investigation. Some have proposed methods for improving the computational efficiency of the LASSO-Cox (Segal, 2005). While EN-Cox was shown to perform efficiently in comparison to LASSO-Cox, improvements might be made. For example, utilization of the penalized likelihood approach of Park and Hastie (2007) may be of particular interest under the Cox model.

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 (1) can be extended to

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_2 \sum_{j=1}^{p} W_{2j}\beta_j^2 + \lambda_1 \sum_{j=1}^{p} W_{1j}|\beta_j|, \qquad (11)$$

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. To date, such an approach has not been investigated, however, and may not be optimal. An alternative approach might entail modification of the adaptive elastic net (see Ghosh, 2007; Zou and Zhang, 2009) for censored data settings. In HDLSS settings, the Zou and Zhang approach may be of particular interest. Likewise, assessment of the performance of the adaptive lasso of Zhang and Lu (2007) in high-dimensional data settings is warranted.

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) can be rewritten as

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}}\Bigg[\boldsymbol{\beta}'\left(\frac{\mathbf{X}'\mathbf{X}+\lambda_2\mathbf{I}}{1+\lambda_2}\right)\boldsymbol{\beta} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \\ &\qquad \lambda_1\{\sum_{j=1}^{p}|\beta_{j0}| + \frac{1}{2|\beta_{j0}|}(\beta_j^2-\beta_{j0}^2)\}\Bigg], \qquad (12)\end{aligned}$$

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 (12), a closed form solution for $\boldsymbol{\beta}$ can be written as

$$\hat{\boldsymbol{\beta}} = (1+\lambda_2)\{\mathbf{X}_m'\mathbf{X}_m + \lambda_2\mathbf{I} + \lambda_1\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}_m'\mathbf{y}, \qquad (13)$$

where $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0) = \text{diag}(\frac{1}{\beta_1},\ldots,\frac{1}{\beta_m})$. Equation (13) can then be utilized to obtain the sandwich estimator for the covariance matrix for $\boldsymbol{\beta}_m$.

Finally, a current drawback of the elastic net is that, like the lasso, it may not always yield consistent results (see Ghosh, 2007). The adaptive elastic net, proposed by Zou and Zhang (2009), resolves this issue for HDLSS data. Adaptation of this new approach for HDLSS censored data settings will be of future interest.

5 Appendix

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

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

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, $\hat{\beta}_i = \hat{\beta}_j$.

DOI: 10.2202/1544-6115.1423

Proposition 2: By definition,

$$\frac{\partial L(\lambda_1, \lambda_2, \boldsymbol{\beta})}{\partial \beta_k}|_{\beta=\hat{\beta}} = 0 \qquad \text{for } \hat{\beta}_k \neq 0. \tag{14}$$

Also, note that

$$L(\lambda_1, \lambda_2, \hat{\boldsymbol{\beta}}) \leq L(\lambda_1, \lambda_2, \boldsymbol{\beta} = \mathbf{0}). \tag{15}$$

By (14) (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 (15),

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

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 Akailseoniaikiudo, 267–281.

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

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

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. *Annals of Statistics*, **32**, 407–499.

Fan,J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1359.

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

Ghosh, S. (2007). Adaptive elastic net: an improvement of elastic net to achieve oracle properties. IUPUI Tech report no. pr07-01.

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., 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. *Annals of Statistics*, **33**, 1617–1642.

Kaplan,E.L., and Meier,P. (1958). Nonparametric estimation from incompete observations. *Journal of the American Statistical Association*, **53**, 457–481.

DOI: 10.2202/1544-6115.1423

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.

Lu,W., and Zhang,H.H. (2007). Variable selection for proportional odds model. *Statistics in Medicine*, **26**, 3771–3781.

Park,M.Y., and Hastie,T. (2007). An L_1 regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society, Series B*, **69**, 659–677.

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

Schwarz,G. (1978). Estimating the dimension of a model. *Annals of 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.

Tibshirani,R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

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

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

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.

Wang,H., Li,R., Tsai,C.L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553–568.

Wang,S., Nan,B., Zhu,J., Beer,D.G. (2008). Doubly penalized Buckley-James method for survival data with high-dimensional covariates. *Biometrics*, **64**, 132–140.

Wu,C.S.P., Zubovic,Y. (1995). A large-scale monte carlo study of the buckley-james estimator with censored data. *Journal of Statistical Computation and Simulation*, **51**, 97–119.

Zhang,H.H., and Lu,W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika*, **94**, 691–703.

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

Zou,H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou,H., and Zhang,H.H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, to appear.

DOI: 10.2202/1544-6115.1423