Variable selection in semiparametric linear regression with censored data

Address for correspondence:

Brent A. Johnson

Department of Biostatistics

Rollins School of Public Health

Emory University

1518 Clifton Rd., NE

Atlanta, GA 30322

U. S. A.

Email: bajohn3@emory.edu

Variable selection in semiparametric linear regression with censored data

Brent A. Johnson[1]

Abstract

We describe two procedures for selecting variables in the semiparametric linear regression model for censored data. One procedure penalizes a vector of estimating equations and simultaneously estimates regression coefficients and selects submodels. A second procedure controls systematically the proportion of unimportant variables through forward selection and the addition of pseudo random variables. We explore both rank-based statistics and Buckley-James statistics in the proposed setting and evaluate the performance of all methods through extensive simulation studies and one real data set.

KEY WORDS: False selection rate; Hard thresholding; Non-smooth estimating function; Rank regression; Soft thresholding; Survival analysis.

1 Introduction

Variable selection is an important problem in linear regression, with applications in many disciplines such as econometrics, biostatistics, bioinformatics, and data mining. Variable selection is a challenging topic in its own right but becomes more complicated when the outcomes may be censored. Outcome censoring occurs, for example, when the response of interest is a failure time yet the failure times for some subjects are unobserved because the follow-up period is complete. Then, the failure times for these subjects are subject to right-censoring and what is known is that their failure times would occur at some time beyond the duration of follow-up. There have been many recent advances in the area of variable selection for censored and uncensored data, which we attempt to summarize below although the summary is in no way comprehensive. Because this paper is concerned with selecting variables in a linear regression model, our introduction and summary will be presented within this context.

In the linear regression model, we assume the i-th response $Y_i,\ i = 1, \ldots, n$ is related to a d-dimensional vector of standardized, prognostic variables $\mathbf{X}_i$

$$Y_i = \boldsymbol{\beta}^T \mathbf{X}_i + \epsilon_i,\ i = 1, \ldots, n, \tag{1}$$

[1]Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, U. S. A. (Email: bajohn3@emory.edu)

where ϵ_i are independent and identically distributed according to an unspecified distribution function $F(\cdot)$. The goal of variable selection for uncensored data is usually characterized through *prediction accuracy*, that is, minimizing the prediction error, where prediction error (PE) for a new observation $(Y_h, \mathbf{X}_h)$ is defined as

$$\text{PE}(\widehat{\mu}) = E\{Y_h - \widehat{\mu}(\mathbf{X}_h)\}^2,$$

where $\hat{\mu} = \hat{\mu}(\mathbf{X}_h)$ is the predicted equation with estimated regression coefficients $\widehat{\boldsymbol{\beta}}$ replacing the unknown regression parameters $\boldsymbol{\beta}$ and the expectation is taken with respect to the joint distribution of $(Y_h, \mathbf{X}_h)$. Assuming the errors ϵ_i have mean zero and variance σ^2, the prediction error $\text{PE}(\hat{\mu})$ may be decomposed as

$$\text{PE}(\widehat{\mu}) = \sigma^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T E(\mathbf{X}_1\mathbf{X}_1^T)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \sigma^2 + \text{ME}(\widehat{\mu}),$$

where ME denotes the model error. In minimizing $\text{PE}(\widehat{\mu})$, we have no control over σ^2; hence, variable selection focuses on minimizing $\text{ME}(\widehat{\mu})$. The primary goal of variable selection procedures is to minimize $\text{ME}(\widehat{\mu})$ using the so-called "important" subset of the total d variables. Generally speaking, methods of variable selection may be split into two classes, those methods which shrink regression coefficients and those that do not. We briefly summarize these two broad strategies below.

Variable selection procedures which do not shrink coefficients include forward selection, backward and stepwise deletion, and all subsets regression methods. These methods generate a sequence of models using hypothesis testing and use goodness-of-fit statistics (GOF) for selecting the best submodel. Generally, one either (a) fixes α-to-enter and selects the model where the following model has a p-value greater than α, or (b) chooses the submodel, among the sequence of submodels, which minimizes the GOF statistic, e.g. Mallow's C_p (Mallows, 1973, 1975), the Akaike information criterion (AIC; Akaike, 1973, 1977), or the Bayesian information criterion (BIC; Schwarz, 1978). The deviance information criterion (DIC; Spiegelhalter et al., 2002) is a Bayesian alternative to AIC and BIC for posterior model selection and analogous to AIC/BIC in the sense that DIC is a composite measure of model adequacy plus a penalty for model complexity.

A criticism of the above procedures is that the stochastic nature of the model selection process is difficult to summarize and often ignored when reporting the accuracy of the regression coefficient estimates in the final model. An alternative to such procedures are methods which simultaneously shrink regression coefficients and set some coefficients to zero, thereby, removing them from the final model. For linear models, one

such shrinkage estimator is penalized least squares estimators, i.e. the minimizer of $Q_{LS}(\boldsymbol{\beta})$, where

$$Q_{LS}(\boldsymbol{\beta}) = n^{-1}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2}$$

where $p_\lambda(|\beta_j|)$ is a penalty on the absolute value of the j-th regression coefficient through a smoothing parameter λ, and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$, and $\|\cdot\|$ denotes the Euclidean norm (Tibshirani, 1996; Fan and Li, 2001). In likelihood-based models, (2) is modified by replacing the least squares objective function with minus the log-likelihood, i.e.

$$Q_{LIK}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} l_i(Y_i, \mathbf{X}_i, \boldsymbol{\beta}) + \sum_{j=1}^{d} p_\lambda(|\beta_j|),$$

where $l_i(Y_i, \mathbf{X}_i, \boldsymbol{\beta})$ is minus the log conditional density of Y_i given $\mathbf{X}$. The objective function $Q_{LIK}(\boldsymbol{\beta})$ lends itself naturally to censored data problems by writing $l_i(\cdot)$ as minus the log partial likelihood for the i-th subject (e.g. Tibshirani, 1997; Fan and Li, 2002). In this manuscript, we consider the following three penalty functions: (i) the LASSO penalty (Tibshirani, 1996, 1997), $p_\lambda(|\beta|) = \lambda|\beta|$, (ii) the hard thresholding penalty, $p_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$, and (iii) the SCAD penalty (Fan and Li, 2001, 2002) given by the continuous function

$$q_\lambda(|\beta|) = \lambda\left\{I(|\beta| < \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| \geq \lambda)\right\}, \quad \text{for } a > 2,$$

where we define

$$(\partial/\partial\beta)p_\lambda(|\beta|) = q_\lambda(|\beta|)\text{sgn}(\beta).$$

Here, our goal is to consider new methods for variable selection with censored data based on the *semiparametric linear model* for censored data rather than the methods based on $Q_{LIK}(\boldsymbol{\beta})$, the partial likelihood, and generally the proportional hazards assumption. The application of variable selection methods in model (1) is new and has not yet been studied in detail in the literature. Specifically, we aim to select variables in the linear model (1) when the observed data are $\{(Z_i, \Delta_i, \mathbf{X}_i),\ i = 1, \ldots, n\}$, where $Z_i = \min(Y_i, C_i)$, $\Delta_i = I(Y_i \leq C_i)$ for a censoring random variable C_i, $i = 1, \ldots, n$. We will assume throughout that Y_i is conditionally independent of C_i given the prognostic variables $\mathbf{X}_i$. We also assume that all variables have been standardized such that $n^{-1}\sum_{i=1}^n x_{ij} = 0$ and $n^{-1}\sum_{i=1}^n x_{ij}^2 = 1$ for all $j,\ 1 \leq j \leq d$.

In this paper, we consider two different strategies for selecting variables in the semiparametric linear model for censored data. First, we extend the class of shrinkage estimators to the current setting, where shrinkage estimator is defined through the three penalty functions, $p_\lambda(|\beta|)$, given above. Second, we

consider a recently proposed method which controls the proportion of unimportant variables in the model (Wu, Boos, and Stefanski, 2007). The latter method works by adding additional noise variables to the original d predictors and monitoring the number of variables falsely selected in forward selection. Wu et al. (2007) consider their method in the context of linear and logistic regression; below, we extend their method to the semiparametric linear model for censored outcomes. The methods are described in Section 2 and large sample properties discussed in Section 3. We evaluate the small sample performance of our methods through Monte Carlo studies in Section 5 and illustrate the utility of the methods in Section 6.

2 Methods

Two proposals for statistical inference in the semiparametric linear model (1) with censored outcomes include one based on generalized ranks (Prentice, 1978) and another based on extending the least squares estimator (Buckley and James, 1979). With no variable selection, the asymptotic properties for the rank-based and Buckley-James statistics have been described elsewhere (Ritov, 1990; Tsiatis, 1990; Wei, Ying, and Lin, 1990; Lai and Ying, 1991a; Lai and Ying, 1991b; Ying, 1993; Huang, 2002; Strawderman, 2005) as well as numerical strategies for drawing inference from a sample of data (Huang, 2002; Strawderman, 2005).

2.1 Penalized weighted log-rank statistics

Define the weighted log-rank estimating function as

$$\mathbf{U}_W(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Delta_i W\{e_i(\boldsymbol{\beta}), \boldsymbol{\beta}\}[\mathbf{X}_i - \tilde{\mathbf{X}}\{e_i(\boldsymbol{\beta}), \boldsymbol{\beta}\}], \tag{3}$$

where $e_i(\boldsymbol{\beta}) = Z_i - \boldsymbol{\beta}^T\mathbf{X}_i$, $W(\cdot)$ satisfies condition A9 in Strawderman (2005),

$$\begin{aligned}
&\tilde{\mathbf{X}}(t, \boldsymbol{\beta}) = \mathbf{S}^{(1)}(t, \boldsymbol{\beta})/S^{(0)}(t, \boldsymbol{\beta}) \\
&S^{(0)}(t, \boldsymbol{\beta}) = n^{-1}\sum_{j=1}^{n} I\{e_j(\boldsymbol{\beta}) \geq t\}, \quad \mathbf{S}^{(1)}(t, \boldsymbol{\beta}) = n^{-1}\sum_{j=1}^{n} \mathbf{X}_j I\{e_j(\boldsymbol{\beta}) \geq t\}.
\end{aligned}$$

Define the penalized weighted log-rank estimating function as

$$\mathbf{U}_W^P(\boldsymbol{\beta}) = \mathbf{U}_W(\boldsymbol{\beta}) + n\mathbf{b}_\lambda(\boldsymbol{\beta}), \tag{4}$$

where $\mathbf{b}_\lambda(\boldsymbol{\beta}) = (q_\lambda(|\beta_1|)\text{sgn}(\beta_1), \ldots, q_\lambda(|\beta_d|)\text{sgn}(\beta_d))^T$, $q_\lambda(|\beta|)$ is a continuous function; and define the estimator $\hat{\boldsymbol{\beta}}_W$ as a consistent solution to $\mathbf{U}_W^P(\boldsymbol{\beta}) = 0$. Two weight functions of substantial interest are $W(t, \boldsymbol{\beta}) = 1$ and $W(t, \boldsymbol{\beta}) = S^{(0)}(t, \boldsymbol{\beta})$, that correspond to the log-rank and Gehan weights, respectively.

It is well-known that for general weight functions, $\mathbf{U}_W(\boldsymbol{\beta})$ is neither continuous nor componentwise monotone in $\boldsymbol{\beta}$. Hence, even without the variable selection arising from the penalty term in (4), it is difficult to solve $\mathbf{U}_W = 0$ directly. In addition, assuming one can derive the asymptotic mean and covariance matrix of a consistent estimator sequence under appropriate regularity conditions, standard error estimates for $\hat{\boldsymbol{\beta}}_W$ may be difficult to calculate because the asymptotic covariance for $\hat{\boldsymbol{\beta}}_W$ will depend on the unknown density of ϵ_i in a complicated way.

2.2 Penalized Buckley-James statistics

Assuming the prognostic variables have mean zero, the Buckley-James estimating function (Buckley and James, 1979) is

$$\widetilde{\mathbf{U}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{\xi_i(\boldsymbol{\beta}) - \boldsymbol{\beta}^T \mathbf{X}_i\} \mathbf{X}_i.$$

where

$$\xi_i(\boldsymbol{\beta}) = \Delta_i Y_i + (1 - \Delta_i) \left[\boldsymbol{\beta}^T \mathbf{X}_i + \frac{\int_{e_i(\boldsymbol{\beta})}^{\infty} \{1 - \hat{F}(s, \boldsymbol{\beta})\}\, ds}{1 - \hat{F}\{e_i(\boldsymbol{\beta}), \boldsymbol{\beta}\}} \right], \tag{5}$$

where $\hat{F}(t, \boldsymbol{\beta})$ is the left-continuous version of the Kaplan-Meier estimator of $F(t)$ based on $\{e_i(\boldsymbol{\beta}), \Delta_i\}$ for $i = 1, \ldots, n$. Note that $\widetilde{\mathbf{U}}(\boldsymbol{\beta})$ reduces to the normal equations when $\Delta_i = 1$ for all $i = 1, \ldots, n$. Define the penalized Buckley-James estimating function as

$$\widetilde{\mathbf{U}}^P(\boldsymbol{\beta}) = \widetilde{\mathbf{U}}(\boldsymbol{\beta}) + n\mathbf{b}_\lambda(\boldsymbol{\beta}), \tag{6}$$

and $\tilde{\boldsymbol{\beta}}$ as a solution to $\widetilde{\mathbf{U}}^P(\boldsymbol{\beta}) = 0$. It is again well-known that the Buckley-James estimating function $\widetilde{\mathbf{U}}(\boldsymbol{\beta})$ is discontinuous and may contain multiple roots. A general technique for solving $\widetilde{\mathbf{U}}(\boldsymbol{\beta}) = 0$ is to iterate between imputing the censored outcomes and solving the "complete-data" normal equations. Extensions of this and other ideas for solving $\widetilde{\mathbf{U}}^P(\boldsymbol{\beta}) = 0$ will be explored in Section 4.

2.3 Controlling the false selection rate

Wu, Boos, and Stefanski (2007) proposed recently a new method of selecting variables by controlling the *false selection rate* (FSR), defined to be the expected proportion of falsely selected unimportant variables among the total variables selected in a variable selection procedure. For a data set $\mathcal{A}_{obs} = \{(Z_i, \Delta_i, \mathbf{X}_i),\ i = 1, \ldots, n\}$ and forward selection procedure, the fraction of falsely selected unimportant predictors is defined: $\gamma(\mathcal{A}_{obs}) = D_Z(\mathcal{A}_{obs})/\{1 + D_N(\mathcal{A}_{obs}) + D_Z(\mathcal{A}_{obs})\}$, or alternatively as the solution to the equation,

$$0 = D_Z(\mathcal{A}_{obs}) - \{1 + D_N(\mathcal{A}_{obs}) + D_Z(\mathcal{A}_{obs})\}\gamma(\mathcal{A}_{obs}), \tag{7}$$

where $D_N(\mathcal{A}_{obs})$ and $D_Z(\mathcal{A}_{obs})$ are the number of important and unimportant variables, respectively, from running forward selection on the observed data $\mathcal{A}_{obs}$. The goal of the FSR method is to maintain $\gamma(\mathcal{A}_{obs})$ near some predetermined target level, say γ_0, on average.

The target level γ_0 can be derived through equation (7) by replacing $\gamma(\mathcal{A}_{obs})$ with γ_0 and taking expectations. Then, the definition of the target FSR level γ_0 is,

$$\gamma_0 = \frac{E\{D_Z(\mathcal{A}_{obs})\}}{E\{1 + D_N(\mathcal{A}_{obs}) + D_Z(\mathcal{A}_{obs})\}}.$$

In practice, however, a user-defined α-to-enter level controls the number of important and unimportant variables in the forward selection. To reflect the dependence of the forward selection procedure on the α-to-enter level, we define $D_N(\mathcal{A}_{obs}; \alpha)$ and $D_Z(\mathcal{A}_{obs}; \alpha)$ to be the number of important and unimportant variables, respectively, after running forward selection on the observed data $\mathcal{A}_{obs}$ for a given, user-defined α-to-enter. Now, define the FSR function

$$\gamma(\alpha) = \frac{E\{D_Z(\mathcal{A}_{obs}; \alpha)\}}{E\{1 + D_N(\mathcal{A}_{obs}; \alpha) + D_Z(\mathcal{A}_{obs}; \alpha)\}}. \tag{8}$$

Assuming the continuity of $\gamma(\alpha)$, the goal of FSR is determine the size α^* such that $\gamma(\alpha^*) = \gamma_0$. Because of the discrete nature of $\gamma(\alpha)$, a precise definition of α^* is given by $\alpha^* = \sup\{\alpha | \gamma(\alpha) \leq \gamma_0\}$. Note, that both $D_N(\mathcal{A}_{obs}; \alpha)$ and $D_Z(\mathcal{A}_{obs}; \alpha)$ are unobserved quantities in any model selection procedure; only the sum $\{D_N(\mathcal{A}_{obs}; \alpha) + D_Z(\mathcal{A}_{obs}; \alpha)\}$ is known.

We approximate the FSR function $\gamma(\alpha)$ using a ratio of averages over M augmented data sets, $\mathcal{A}_1, \ldots, \mathcal{A}_M$. Define the m-th augmented data set $\mathcal{A}_m = \{(Z_i, \Delta_i, \mathbf{X}_i, \mathbf{X}^*_{i,m}),\ i = 1, \ldots, n\}$, where $\mathbf{X}_{i,m}$ is a p-dimensional vector of simulated, pseudopredictors (defined in Appendix 3) for the i-th subject. Using the augmented data $\mathcal{A}_m$, we fit the linear regression model

$$Y_i = \boldsymbol{\beta}^T \mathbf{X}_i + \boldsymbol{\beta}^{*T} \mathbf{X}^*_{i,m} + \epsilon_i,\ i = 1, \ldots, n, \tag{9}$$

where $\boldsymbol{\beta}^*$ are the regression coefficients corresponding to the pseudopredictors and other notation follows from (1). Define $D_N(\mathcal{A}_m;\alpha)$, $D_Z(\mathcal{A}_m;\alpha)$ and $D_P(\mathcal{A}_m;\alpha)$ as the number of important, unimportant, and pseudopredictors, respectively, after running forward selection at the user-defined α-to-enter level α using the augmented data set $\mathcal{A}_m$. Next, we define the ratio of averages,

$$\begin{aligned} \widehat{\gamma}_P(\alpha) &= \bar{D}_{P,M}(\alpha)/\{1+\bar{D}_M(\alpha)\}, \quad \text{where} \\ \bar{D}_{P,M}(\alpha) &= M^{-1}\sum_{m=1}^{M} D_P(\mathcal{A}_m;\alpha) \\ \bar{D}_M(\alpha) &= M^{-1}\sum_{m=1}^{M} D_N(\mathcal{A}_m;\alpha)+D_Z(\mathcal{A}_m;\alpha)+D_P(\mathcal{A}_m;\alpha) \end{aligned} \tag{10}$$

Because the pseudopredictors are known *a priori* to be unrelated to the outcome Y, $\widehat{\gamma}_P(\alpha)$ may be used to approximate the stochastic behaviour in the expected proportion of falsely selected variables as a function of α. The final step in the FSR method is to link the two functions: $\widehat{\gamma}_P(\alpha)$ and $\gamma(\alpha)$. This connection is described in the paragraph below.

A relationship between $\widehat{\gamma}_P(\alpha)$ and $\gamma(\alpha)$ is achieved by setting $\widehat{\gamma}_P(\alpha)=\zeta_0$ and $\gamma(\alpha)=\gamma_0$. Under assumptions (A1)-(A2) from Wu et al. (2007, p. 237), one can show that $\zeta_0=(p\gamma_0)/(p\gamma_0+d_Z)$, where p is the number pseudopredictors and $d_Z=\#\{\beta_j=0,\ j=1,\ldots,d\}$. Hence, $\widehat{\gamma}_P(\alpha)=(p\gamma_0)/(p\gamma_0+d_Z)$. Using the definition of $\widehat{\gamma}_P(\alpha)$ in (10) and solving for γ_0 leads one to $\gamma_0=[\{d_Z\bar{D}_{P,M}(\alpha)\}/p]/[1+\bar{D}_M(\alpha)]$. We note that under assumptions (A1)-(A2) in Wu et al. (2007), the expression in the numerator (i.e. $[\{d_Z\bar{D}_{P,M}(\alpha)\}/p]$) is estimating $E\{D_Z(\mathcal{A}_{obs};\alpha)\}$ and the denominator is estimating $E\{1+D_N(\mathcal{A}_{obs};\alpha)+D_Z(\mathcal{A}_{obs};\alpha)\}$, as desired. Using the approximation $\hat{d}_Z(\alpha)\approx\{p+d-D_N(\mathcal{A}_{obs};\alpha)+D_Z(\mathcal{A}_{obs};\alpha)\}$ allows one to construct an iterative algorithm for FSR model selection, which we describe in Subsection 4.4.

3 Asymptotic Results

Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$ and suppose that $\beta_{0j}\neq 0$ for $j\leq d_N$ and $\beta_{0j}=0$ for $j>d_N$. We impose the following conditions on the penalty function $q_{\lambda_n}(|\beta|)$.

Q1. (i) $\lim_{n\to\infty}\lambda_n=0$ and $\lim_{n\to\infty}\sqrt{n}\lambda_n=\infty$.

(ii) For non-zero fixed β, $\lim\sqrt{n}q_{\lambda_n}(|\beta|)=0$ and $\lim q'_{\lambda_n}(|\beta|)=0$;

(iii) For any $M>0$, $\lim_{n\to\infty}\sup_{|\beta|\leq Mn^{-1/2}}q'_{\lambda_n}(|\beta|)=0$ and $\lim\lambda_n^{-1}\inf_{|\beta|\leq Mn^{-1/2}}q_{\lambda_n}(|\beta|)>0$;

Remark 1. Condition Q1 pertains to the choices of the penalty function and the regularization parameter. Intuitively, condition Q1(ii) implies the penalty imposed on regression coefficients decreases as the sample size increases; so the penalized coefficient estimates will be equal to the usual (unpenalized) estimates, say $\widehat{\boldsymbol{\beta}}^{(0)}$, for sample size sufficiently large. The L_1 penalty does not satisfy condition Q1 because it penalizes large and small coefficient estimates $\widehat{\boldsymbol{\beta}}^{(0)}$ regardless of the sample size [namely, one assumes the existence of λ_0 such that $n^{-1/2}\lambda_n \to \lambda_0$ as $n \to \infty$ (e.g. Knight and Fu, 2000, Theorem 2)]. However, for sample size sufficiently large, the penalty on the penalized estimating function will be zero, and thus the SCAD and Hard threshold estimates will equal the usual (unpenalized) estimates $\widehat{\boldsymbol{\beta}}^{(0)}$.

The following theorem states the main theoretical result regarding the solution to the penalized weighted log-rank estimating equation, including the existence of a $\sqrt{n}$-consistent solution, the sparsity of the solution and the asymptotic normality of the estimator.

Theorem 1 *Under conditions Q1 above and conditions A1-A8 in Appendix 1, the following results hold:*
(i) There exists an $\sqrt{n}$-consistent solution to $\mathbf{U}_W^P(\boldsymbol{\beta})$, i.e, $\widehat{\boldsymbol{\beta}}_W = \boldsymbol{\beta}_0 + O_p(n^{-1/2})$ such that $n^{-1/2}\mathbf{U}_W^P(\widehat{\boldsymbol{\beta}}_W) = o_p(1)$.
(ii) $\lim_n pr(\widehat{\beta}_{Wj} = 0$ *for* $j > d_N) = 1$.
(iii) Let $\widehat{\boldsymbol{\beta}}_{W1} = (\widehat{\beta}_{W1}, \ldots, \widehat{\beta}_{Wd_N})^T$ *and* $\boldsymbol{\beta}_{01} = (\beta_{01}, \ldots, \beta_{0d_N})^T$. *Then*

$$\sqrt{n}(\mathbf{A}_{W11} + \boldsymbol{\Sigma}_{11}) \left\{ \widehat{\boldsymbol{\beta}}_{W1} - \boldsymbol{\beta}_{01} + (\mathbf{A}_{W11} + \boldsymbol{\Sigma}_{11})^{-1} \mathbf{b}_{\lambda_n}(\boldsymbol{\beta}_0) \right\} \to_d N(0, \mathbf{B}_{W11}),$$

where $\mathbf{A}_{W11}$ and $\mathbf{B}_{W11}$ are the first $d_N \times d_N$ sub-matrices of

$$\begin{aligned} \mathbf{A}_W &= \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\tau} W(t, \boldsymbol{\beta}_0) \{\mathbf{X}_i - \tilde{\mathbf{X}}(t, \boldsymbol{\beta}_0)\}^{\otimes 2} \{\lambda'(t)/\lambda(t)\} \, dN_i(t, \boldsymbol{\beta}_0), \\ \mathbf{B}_W &= \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\tau} \{W(t, \boldsymbol{\beta}_0)\}^2 \{\mathbf{X}_i - \tilde{\mathbf{X}}(t, \boldsymbol{\beta}_0)\}^{\otimes 2} \, dN_i(t, \boldsymbol{\beta}_0), \end{aligned}$$

$N_i(t, \boldsymbol{\beta}) = I\{e_i(\boldsymbol{\beta}) \le t, \Delta_i = 1\}$, $\lambda(t)$ is the hazard function of the errors $e_i(\boldsymbol{\beta}_0)$, $\lambda'(t) = (d/dt)\lambda(t)$, $\boldsymbol{\Sigma}_{11}$ is the first $d_N \times d_N$ sub-matrix of diag $\left\{ q'_{\lambda_n}(|\boldsymbol{\beta}_0|) sgn(\boldsymbol{\beta}_0) \right\}$, $\mathbf{A}^{\otimes 2} = \mathbf{A} \otimes \mathbf{A}$, the Kronecker product of the matrix $\mathbf{A}$ with itself, and τ is defined in Condition A6 of Appendix 1.

Remark 2. Conditions A1-A8 are conditions similar to ones given in Strawderman (2005) for deriving an asymptotic linear expression for $n^{-1/2}\mathbf{U}_W(\boldsymbol{\beta})$ and subsequently deriving a $\sqrt{n}$-consistent, asymptotically normal estimator $\widehat{\boldsymbol{\beta}}_W$. The proof for Theorem 1 is given in Appendix 2.

Based on the work of Ritov (1990), it is natural to believe that a similar conclusion stated in Theorem 1 also holds for penalized Buckley-James statistics. In particular, conditions A1-A6 in Appendix 1 have similar implications as the regularity conditions given in Ritov (1990) to derive an asymptotic linear expression for $n^{-1/2}\widetilde{\mathbf{U}}(\boldsymbol{\beta})$, where the asymptotic slope matrix $\boldsymbol{\Omega}$ is given by

$$\boldsymbol{\Omega} = \int Var(\mathbf{X}|C - \boldsymbol{\beta}_0^T\mathbf{X} \geq u)\,\{u - E[\epsilon|\epsilon > u]\}\left\{-\frac{f'(u)}{f(u)} + E[\frac{f'(\epsilon)}{f(\epsilon)}|\epsilon > u]\right\}\mathrm{pr}(C - \boldsymbol{\beta}_0^T\mathbf{X} \geq u)\,dF(u),$$

assuming $\boldsymbol{\Omega}$ nonsingular. We conjecture that an *approximate* $\sqrt{n}$-consistent solution to $\widetilde{\mathbf{U}}^P(\boldsymbol{\beta}) = 0$ exists under appropriate regularity conditions, including condition Q1. Assuming the conjecture is correct, the conclusions of Theorem 1 continue to hold where

$$\sqrt{n}(\boldsymbol{\Omega}_{11} + \boldsymbol{\Sigma}_{11})\left\{\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\boldsymbol{\Omega}_{11} + \boldsymbol{\Sigma}_{11})^{-1}\mathbf{b}_{\lambda_n}(\boldsymbol{\beta}_0)\right\} \rightarrow_d N(0, \boldsymbol{\Lambda}_{11}).$$

where

$$\boldsymbol{\Lambda} = \int Var(\mathbf{X}|C - \boldsymbol{\beta}_0^T\mathbf{X} \geq u)\,\{u - E[\epsilon|\epsilon > u]\}^2\,\mathrm{pr}(C - \boldsymbol{\beta}_0^T\mathbf{X} \geq u)\,dF(u),$$

and $f'(t) = (d/dt)f(t)$, $f(t) = (d/dt)F(t)$.

It is evident that the asymptotic covariance of $\hat{\boldsymbol{\beta}}_{W1}$ is

$$n^{-1}(\mathbf{A}_{W11} + \boldsymbol{\Sigma}_{11})^{-1}\mathbf{B}_{W11}(\mathbf{A}_{W11} + \boldsymbol{\Sigma}_{11})^{-1}.$$

Unlike the asymptotic variance formulae given in Fan and Li (2001, 2002), the expressions for the proposed estimators in the accelerated failure time model cannot be evaluated directly because they depend on unknown density functions. Nevertheless, several recent numerical techniques are now available which allow one to obtain standard error estimates without numerical derivatives or smoothing, which may be unstable. Two recent methods include inverse numerical differentiation (Huang, 2002), and resampling for $\mathbf{A}_W$ (Strawderman, 2005, Algorithm 2). In the sequel, we use Strawderman's method. Also, the above methods work as well for penalized Buckley-James statistics with $\boldsymbol{\Omega}_{11}$ and $\boldsymbol{\Lambda}_{11}$ replacing $\mathbf{A}_{W11}$ and $\mathbf{B}_{W11}$, respectively.

4 Estimation and Inference

Here, we provide details on estimation algorithms to obtain coefficient estimates from a sample of data for a fixed smoothing parameter λ. We discuss the penalized statistics in an order (beginning with penalized

Buckley-James statistics) which facilitates the flow of the manuscript. Finally, we discuss novel methods for cross-validating the smoothing parameter λ in the semiparametric linear regression model for censored outcomes.

4.1 Estimation for penalized Buckley-James statistics

First, we note that in neighborhoods of the truth β_{0j}, $|\beta_{0j}| > 0$, the derivative of the penalty function is well-approximated by

$$\frac{\partial}{\partial \beta_j}\, p_\lambda(|\beta_j|) = q_\lambda(|\beta_j|)\text{sgn}(\beta_j) \approx \{q_\lambda(|\beta_{0j}|)/|\beta_{0j}|\}\, \beta_j.$$

This is the *local quadratic approximation* discussed in Tibshirani (1996) and Fan and Li (2001). This approximation allows one to replace a potentially unstable optimization routine involving a singularity at zero with a stable optimization routine. In particular, under regularity conditions, the asymptotic linear expansion of the penalized Buckley-James estimating equation is

$$n^{-1/2}\tilde{\mathbf{U}}^P(\boldsymbol{\beta}) \approx n^{-1/2}\tilde{\mathbf{U}}(\boldsymbol{\beta}_0) + n^{1/2}\boldsymbol{\Omega}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + n^{1/2}\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta},$$

ignoring an $o_p(1)$ term and $\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}) = \text{diag}\,\{q_\lambda(|\beta_1|)/|\beta_1|, \ldots, q_\lambda(|\beta_d|)/|\beta_d|\}$. The expression follows directly from the quadratic approximation of the penalty functions on the regression coefficients. Hence, after rearranging terms, we obtain the following iterative algorithm, which we term the *Direct Algorithm*.

Direct Algorithm.

1. $\boldsymbol{\beta}^{(0)}$ solves $\tilde{\mathbf{U}}(\boldsymbol{\beta}) = 0$
2. $\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \left\{n\widehat{\boldsymbol{\Omega}}^{(k)} + n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(k)})\right\}^{-1} \tilde{\mathbf{U}}^P(\boldsymbol{\beta}^{(k)})$

where $\widehat{\boldsymbol{\Omega}}^{(k)}$ is a consistent estimate of $\boldsymbol{\Omega}$ based on the current iterate $\boldsymbol{\beta}^{(k)}$ using Strawderman's Algorithm 2 (2005) and $\boldsymbol{\beta}^{(0)}$ is a consistent solution to $\tilde{\mathbf{U}}(\boldsymbol{\beta})$.

4.2 Estimation for penalized weighted log-rank statistics

In the case where $W(t,\boldsymbol{\beta}) = S^{(0)}(t,\boldsymbol{\beta})$, the estimating function $\mathbf{U}_W^P(\boldsymbol{\beta})$ in (4) simplifies to

$$\begin{aligned}\mathbf{U}_G^P(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \Delta_i S^{(0)}\{e_i(\boldsymbol{\beta}),\boldsymbol{\beta}\}[\mathbf{X}_i - \tilde{\mathbf{X}}\{e_i(\boldsymbol{\beta}),\boldsymbol{\beta}\}] + n\mathbf{b}_\lambda(\boldsymbol{\beta}) \\ &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i(\mathbf{X}_i - \mathbf{X}_j) I\{e_i(\boldsymbol{\beta}) \le e_j(\boldsymbol{\beta})\} + n\mathbf{b}_\lambda(\boldsymbol{\beta}) \qquad (11)\end{aligned}$$

where the (11) follows from the definition of $\tilde{\mathbf{X}}\{\boldsymbol{\beta}; e_i(\boldsymbol{\beta})\}$ and straightforward algebraic manipulations. It is easy to see that $\mathbf{U}_G^P(\boldsymbol{\beta}) = 0$ in (11) is the gradient of the following function:

$$\begin{aligned}L_G^P(\boldsymbol{\beta}) &= L_G(\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \\ L_G(\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i\{e_i(\boldsymbol{\beta}) - e_j(\boldsymbol{\beta})\}^-, \qquad (12)\end{aligned}$$

where $c^- = \max(-c, 0)$. It is well-known that minimizing $L_G(\boldsymbol{\beta})$ is asymptotically equivalent to minimizing $\|\mathbf{U}_G(\boldsymbol{\beta})\|$ (Fygensen and Ritov, 1994) and so $L_G^P(\boldsymbol{\beta})$ may also be expressed as a function of $\|\mathbf{U}_G(\boldsymbol{\beta})\|$. Because $L_G(\boldsymbol{\beta})$ is convex in $\boldsymbol{\beta}$, a sufficient condition to ensure that $L_G^P(\boldsymbol{\beta})$ is convex is to require that the penalty function $p_\lambda(|\beta_j|)$ is convex. One such penalty function is $p_\lambda(|\beta_j|) = \lambda|\beta_j|$, i.e. the LASSO penalty, in which case the optimization may be rewritten in the following way: minimize $\sum_{i=1}^n\sum_{j=1}^n \Delta_i u_{ij}$ subject to the constraints $u_{ij} > 0$, $u_{ij} \ge -\{e_i(\beta) - e_j(\beta)\}$, and $\sum_{k=1}^d |\beta_k| \le k$, where k is a tuning parameter. This optimization may be accomplished through quadratic programming.

Now, of course, it would be desirable to work directly with $L_G^P(\boldsymbol{\beta})$ for arbitrary penalty functions, not just convex penalties. However, the singularity in $L_G(\boldsymbol{\beta})$ makes this task challenging. An alternative algorithm again uses the quadratic approximations defined above.

Algorithm 2. Gehan-type weight functions.

1. $\boldsymbol{\beta}^{(0)} =$ arg min $L_G(\boldsymbol{\beta})$.
2. $\boldsymbol{\beta}^{(k+1)} =$ arg min $\|\mathbf{Q}(\boldsymbol{\beta};\boldsymbol{\beta}^{(k)})\|$, where $\mathbf{Q}(\boldsymbol{\beta};\boldsymbol{\beta}^{(k)}) = \mathbf{U}_G(\boldsymbol{\beta}) + n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(k)})\boldsymbol{\beta}$.
3. Repeat step 2 until convergence.

Here, we explain the intuition behind Algorithm 2 . We begin with the premise that the Direct Algorithm could be applied to the penalized Gehan estimating function $\mathbf{U}_G^P(\boldsymbol{\beta})$ but a direct minimization of $L_G^P(\boldsymbol{\beta})$ would be more desirable because it avoids the estimation of $\mathbf{A}_G(\boldsymbol{\beta})$. Again, we note that $\boldsymbol{\beta}^{(0)}$ in Step 1 is equivalent asymptotically to $\{\arg\min \|\mathbf{U}_G(\boldsymbol{\beta})\|\}$, where $\mathbf{U}_G(\boldsymbol{\beta})$ is known to be componentwise monotone. Once $\boldsymbol{\beta}^{(0)}$ is obtained, $\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(0)})$ is fixed and the product $\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}$ is again componentwise monotone. We conjecture that, under appropriate regularity conditions, $\mathbf{Q}(\boldsymbol{\beta};\boldsymbol{\beta}^{(k)})$ is componentwise monotone for every intermediate value $\boldsymbol{\beta}^{(k)}$ and, upon convergence, the solution $\widehat{\boldsymbol{\beta}}_G$ is equivalent asymptotically to the minimizer of $L_G^P(\boldsymbol{\beta})$. We also note that the calculations in Algorithm 2 are similar to the minimizations required in Huang's (2002) 'inverse numerical differentiation' algorithm where $n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{(k)})\boldsymbol{\beta}$ correspond to his perturbation vectors; however, his perturbation vectors do not depend on $\boldsymbol{\beta}$ as they do here in Algorithm 2. For general weight functions, we assume that $\hat{\boldsymbol{\beta}}_G$ is a consistent root of $\mathbf{U}_G^P(\boldsymbol{\beta})$ and use the asymptotic linearity of $\mathbf{U}_W^P(\boldsymbol{\beta})$ similar to what was done in the Direct Algorithm.

Algorithm 3. General weight function

1. Set $\boldsymbol{\beta}_W^{(0)} = \hat{\boldsymbol{\beta}}_G$.
2. Let $\widehat{\mathbf{A}}_W^{(k)}$ be a consistent estimate of $\mathbf{A}_W$ based on the current iterate $\boldsymbol{\beta}^{(k)}$.
3. $\boldsymbol{\beta}_W^{(k+1)} = \boldsymbol{\beta}_W^{(k)} - \{n\widehat{\mathbf{A}}_W^{(k)} + n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_W^{(k)})\}^{-1}\mathbf{U}_W^P(\boldsymbol{\beta}_W^{(k)})$.
4. Repeat steps 2-3 until convergence.

Here we note that step 2 requires resampling (Strawderman, 2005, Algorithm 2) and a single pass through Algorithm 3 is similar to Strawderman's one-step estimator.

4.3 Selection of λ

To implement our proposed algorithm, we require a choice of λ for the LASSO and Hard thresholding penalty functions, and of (a, λ) for the SCAD penalty. Fan and Li (2001, 2002) suggest using $a = 2+\sqrt{3} \approx 3.7$ and showed that this selection performs well in small samples. We use their suggested estimate for a and simplify the cross-validation for all estimators to one involving the scalar λ. In this subsection, we include the subscript λ on $\widehat{\boldsymbol{\beta}}$, i.e. $\widehat{\boldsymbol{\beta}}_\lambda$, respectively, to stress the dependence of the estimator on the regularization parameter λ. Finally, $\hat{\boldsymbol{\beta}}_\lambda$ will refer to a weighted log-rank estimator unless specified otherwise.

For uncensored data in the multiple linear regression model, Tibshirani (1996) suggested the following generalized cross-validation statistic:

$$\mathrm{GCV}_{\mathrm{LS}}(\lambda) = \frac{\mathrm{RSS}(\lambda)/n}{\{1 - e(\lambda)/n\}^2}$$

where $\mathrm{RSS}(\lambda)$ is the residual sum of squares $\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda\|$ and $e(\lambda)$ is the effective number of parameters, $e(\lambda) = \mathrm{tr}[\mathbf{X}(\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda)^{-1}\mathbf{X}^T]$. Then, the optimal λ is given by arg min $\mathrm{GCV}_{\mathrm{LS}}(\lambda)$ (Tibshirani, 1996; Fan and Li, 2001). Note that for uncensored data, the intercept may be safely ignored in $\mathrm{RSS}(\lambda)$ as $\mathbf{X}$ has mean zero and $\mathbf{Y}$ may be standardized by $\sum_i^n Y_i/n$. For general penalized estimating equations, in particular ones with missing data, the effective number of parameters $e(\lambda)$ may be approximated using a technique by Tibshirani (1997). Upon convergence, we note that $\widehat{\boldsymbol{\beta}}_\lambda$ may be approximated by the ridge regression estimate $\widehat{\boldsymbol{\beta}}_r = \{n\widehat{\mathbf{A}}_W(\widehat{\boldsymbol{\beta}}_\lambda) + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)\}^{-1}\mathbf{V}^T\mathbf{Y}^*$, where $\mathbf{V}$ is the Cholesky root of $n\widehat{\mathbf{A}}_W(\widehat{\boldsymbol{\beta}}_\lambda)$ and the pseudo response $\mathbf{Y}^* = (\mathbf{V}^T)^{-1}\{n\widehat{\mathbf{A}}_W(\widehat{\boldsymbol{\beta}}_\lambda)\widehat{\boldsymbol{\beta}}_\lambda - \mathbf{U}_W(\widehat{\boldsymbol{\beta}}_\lambda)\}$. So, in the semiparametric linear model with censored outcomes, the effective number of parameters in the ridge estimate (Hoerl and Kennard, 1970) is: $e(\lambda) = \mathrm{tr}[\{\widehat{\mathbf{A}}_W(\widehat{\boldsymbol{\beta}}_\lambda) + \boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)\}^{-1}\widehat{\mathbf{A}}_W(\widehat{\boldsymbol{\beta}}_\lambda)]$. In the penalized Buckley-James estimator, $\widehat{\mathbf{A}}_W(\widehat{\boldsymbol{\beta}}_\lambda)$ is replaced by $\widehat{\boldsymbol{\Omega}}(\widehat{\boldsymbol{\beta}}_\lambda)$. While we can propose a substitute for $e(\lambda)$ in $\mathrm{GCV}_{\mathrm{LS}}(\lambda)$ when some outcomes may be censored, replacing $\mathrm{RSS}(\lambda)$ effectively remains challenging.

For the estimators proposed in this article, our cross-validation statistic is given by

$$\mathrm{GCV}(\lambda) = \frac{L_G(\widehat{\boldsymbol{\beta}}_\lambda)/n}{\{1 - e(\lambda)/n\}^2},$$

where the residual sum of squares loss function in $\mathrm{GCV}_{\mathrm{LS}}(\lambda)$ is replaced by $L_G(\boldsymbol{\beta})$. It is well-known that for uncensored data, $L_G(\boldsymbol{\beta})$ yields a proper distance measure and so, geometrically speaking, it is analogous to least squares (McKean, 2004). Because the convexity of $L_G(\boldsymbol{\beta})$ is preserved in the presence of censoring, as defined in (12), the geometric interpretation of $L_G(\boldsymbol{\beta})$ extends naturally to censored data. Finally, because of the well-known relation between Buckley-James statistics and weighted log-rank statistics (Ritov, 1990), one may regard $\mathrm{GCV}(\lambda)$ as a general goodness-of-fit statistic for choosing smoothing parameters in the semiparameteric linear model for censored outcomes. Therefore, in the sequel, we define $\hat{\lambda} = \text{arg min } \mathrm{GCV}(\lambda)$. Our experience suggests that the $\mathrm{GCV}(\lambda)$ criterion performs well in simulation studies. Additional comments regarding cross-validation are relegated to Section 7.

4.4 FSR algorithm

The success of the FSR algorithm relies heavily on the estimated $\widehat{\gamma}_P(\alpha)$. This function is calculated by running forward selection on every augmented data set $\mathcal{A}_1, \ldots, \mathcal{A}_M$ and for every α-to-enter level in a fine grid of $[0, 1]$. Calculating $\widehat{\gamma}_P(\alpha)$ is computationally expensive but it is only done once. Once the function $\widehat{\gamma}_P(\alpha)$ is calculated, the FSR algorithm ultimately seeks to find the α-to-enter level $\hat{\alpha}_0$ which satisfies

$$\hat{\alpha}_0 = \sup\{\alpha | \widehat{\gamma}_P(\alpha) \leq \zeta_0, \alpha \in [0, 1]\}, \tag{13}$$

for the cutoff ζ_0 which leads to the target false selection rate γ_0. This is accomplished through the following two-step iterative process: (a) $\zeta^{(k+1)} = (\gamma_0 p)/\{\gamma_0 p + \widehat{D}_Z(\mathcal{A}_{obs}; \hat{\alpha}^{(k)})\}$, where $\widehat{D}_Z(\mathcal{A}_{obs}; \hat{\alpha}^{(k)})$ is the number of original d predictors left out of the final model; (b) update $\hat{\alpha}^{(k+1)}$ via (13) and $\zeta^{(k)}$. The process is initialized with $\zeta^{(0)} = \gamma_0 p/(\gamma_0 p + d)$ and finished with a final forward selection run with α-to-enter level $\hat{\alpha}_0$.

Remark 3. We use generalized Wald tests (cf. Rotnizky and Jewell, 1990; Boos, 1992) to implement forward selection for Buckley-James and weighted log-rank statistics. Finally, when we calculate $\widehat{\gamma}_P(\alpha)$ over a fine grid of $[0, 1]$, we use the partition $\alpha_k = k/500,\ k = 1, \ldots, 500$ in simulation studies and $\alpha_k = k/1000,\ k = 1, \ldots, 1000$ in data analyses.

5 Simulation Studies

In this section, we evaluate the proposed methods through simulation studies. The simulation design below is adapted from ones considered earlier in the literature (Tibshirani and Knight, 1999; Wu et al., 2007). Here, we simulate $M = 100$ datasets from the model

$$Y_i = \boldsymbol{\beta}_0^T \mathbf{X}_i + \epsilon_i,\ i = 1, \ldots, n,$$

where $\mathbf{X}_i$ is multivariate normal with the correlation between any X_k and X_j equal to $\rho^{|k-j|}$, $\rho = 0.5$ and the sample size $n = 100$. Here, the errors ϵ_i independently follow a standard normal distribution and the censoring distribution is independent of Y_i and taken to be uniform$(0, \tau)$ with τ chosen to yield 40% censoring. The true regression coefficients are chosen to yield a theoretical $R^2 = 0.75$, where for a random vector $\mathbf{X}_1$, define

$$R^2 = \frac{\beta_0^T E(\mathbf{X}_1 \mathbf{X}_1^T) \beta_0}{\beta_0^T E(\mathbf{X}_1 \mathbf{X}_1^T) \beta_0 + \sigma^2}.$$

We consider four different simulated models H1-H4 with different numbers of non-zero, true regression coefficients. The true regression coefficients are clustered in two groups, centered at X_4 and X_{13}, and generated in the following manner.

1. For $h = 1, 2, 3, 4$, set the initial coefficient values

$$\beta_{4+k,h} = \beta_{13+k,h} = (h - k)^2, \quad \text{for } |k| < h,$$

2. multiply initial coefficient values by common constant to yield $R^2 = 0.75$.

Model H1 has 16 noise variables and two very strong predictors of outcome while Model H4 has four noise variables and 14 weak to moderate predictors of outcome.

We evaluate the variable selection procedures through four statistics: relative median model errors, number of correct and incorrect zeros, and the estimated false selection rate. Relative median model error (RMME) is the ratio of the median model errors from the variable selection procedure over the model error from the unpenalized statistic using all d variables. Our definition of relative median model error is similar to Tibshirani (1996, 1997) but differs from Fan and Li (2001, 2002) in that the latter use median of the ratios of model errors whereas we use the ratio of median model errors. In addition, the average number of variables which are correctly shrunk to zero is termed a "correct" zero and non-zero variables which are incorrectly shrunk to zero termed an "incorrect" zero. We call the model which knows *a priori* which variables are non-zero but with unknown regression coefficients β, the "true" model. Finally, we also monitor the proportion of unimportant variables included in the final model, that is, the false selection rate.

In Figure 1, we summarize the results from 100 Monte Carlo data sets using weighted log-rank statistics and $\rho = 0.5$ in the random design matrix $\mathbf{X}$. We use a solid black line to represent the ideal curve resulting from the true model. We find that among the methods considered, FSR and SCAD perform better than LASSO when the *model fraction* (i.e. proportion of significant variables among total variables) is low while LASSO performs best in models with many weak to moderate variables. The tradeoff among methods is illustrated in Figures 1(b)-(d): FSR and SCAD tend to be too exclusive thereby eliminating some important variables whereas LASSO is less discriminate and pays dividends when the model fraction is high. We repeated the above simulation exercise with uncorrelated predictors but noted only modest differences.

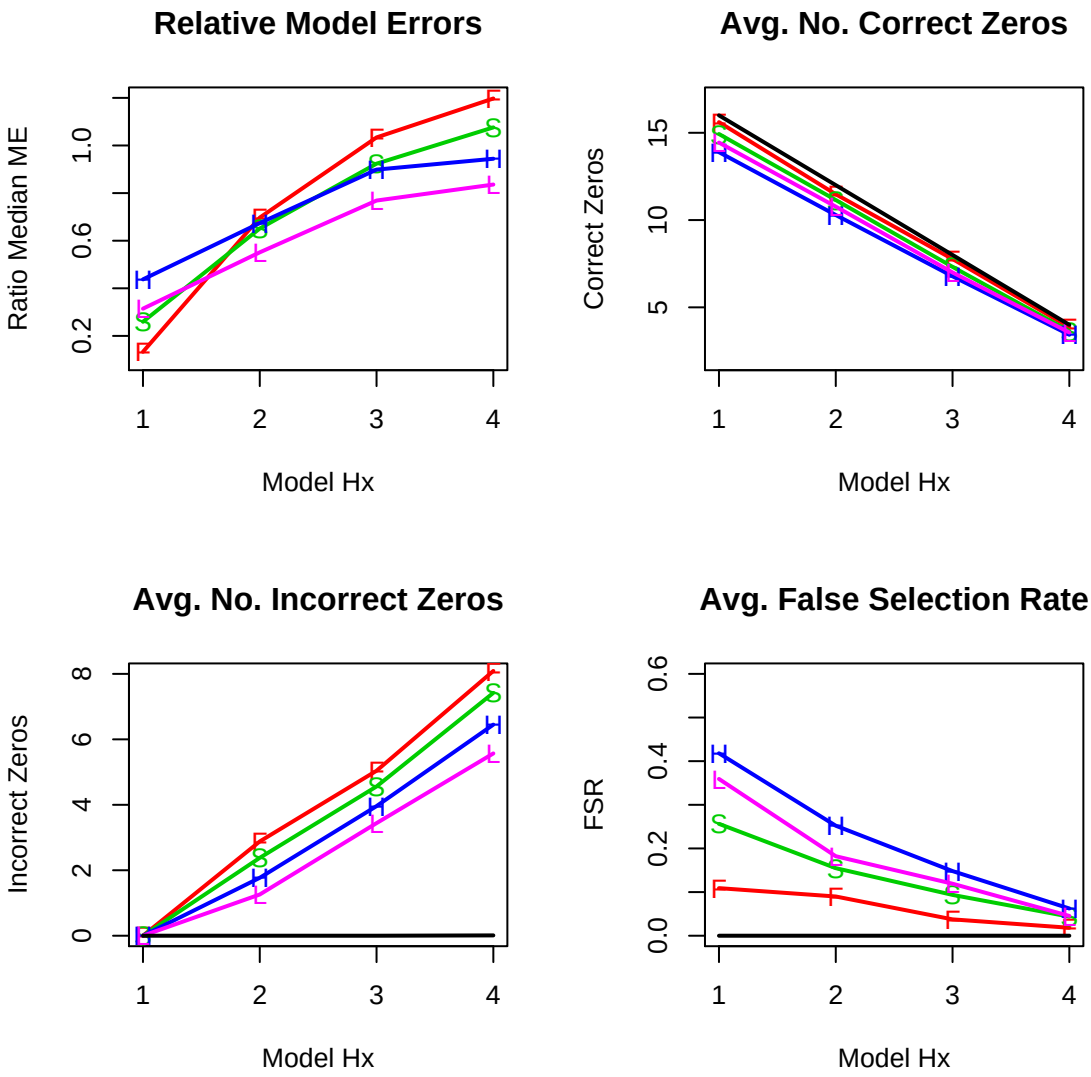

Figure 1: Simulation summary of 100 Monte Carlo data sets using weighted log-rank statistics with Gehan weight and $\text{corr}(x_j, x_k) = (1/2)^{|j-k|}$ for two predictors x_j and x_k. Different variable selection procedures are indicated by a different symbol: FSR (F), SCAD (S), Hard (H), and LASSO (L).

	Variable	Buckley-James	Gehan	log-rank
1.	Age	−0.22(0.07)	−0.22(0.07)	−0.21(0.07)
2.	Albumin	0.16(0.08)	0.19(0.10)	0.19(0.08)
3.	Alk. Phos.	−0.01(0.05)	−0.01(0.08)	−0.04(0.07)
4.	Ascites	−0.12(0.07)	−0.13(0.06)	−0.07(0.08)
5.	Bilirubin	−0.49(0.07)	−0.50(0.13)	−0.52(0.08)
6.	Edema	−0.15(0.08)	−0.15(0.07)	−0.18(0.09)
7.	Hepatomegaly	−0.06(0.06)	−0.06(0.08)	−0.10(0.09)
8.	Prothrombin	−0.21(0.07)	−0.23(0.09)	−0.16(0.09)
9.	Sex	0.07(0.05)	0.08(0.05)	0.05(0.07)
10.	Spiders	−0.11(0.07)	−0.13(0.09)	−0.02(0.10)

Table 1: Full model results for Mayo primary biliary cirrhosis data using the accelerated failure time model and Buckley-James and weighted log-rank statistics

6 Example

Here, we consider the Mayo primary biliary cirrhosis data (Fleming and Harrington, 1991, Appendix D.1). The data contains information about the survival time and prognostic variables for 418 patients who met standard eligibility criteria for a study of the drug D-penicillamine. We consider the following ten variables in our analysis: age, albumin, alkaline phosphatase, ascites, bilirubin, edema, hepatomegaly, prothrombin time, sex and vascular spiders. Albumin, alkaline phosphatase, bilirubin, and prothrombin time have all been transformed on the natural logarithmic scale (see Fleming and Harrington, 1991, Ch. 4). Of 418 eligible patients, 312 patients were included in a randomized study and used to build a Cox proportional hazards model for the natural history of PBC (Dickson et al., 1989) which includes the five variables: age, albumin, bilirubin, edema, and prothrombin time. We use 312 randomized patients in our analyses below, that is, the same data used to select the five important variables in the natural history model under the proportional hazards assumption (e.g. Fleming and Harrington, 1991, Ch. 4, p. 156). Our statistical analysis of the Mayo PBC begins in Table 1 where we provide parameter estimates from fitting the full accelerated failure time model with ten variables and give numeric labels for independent variables used in subsequent figures.

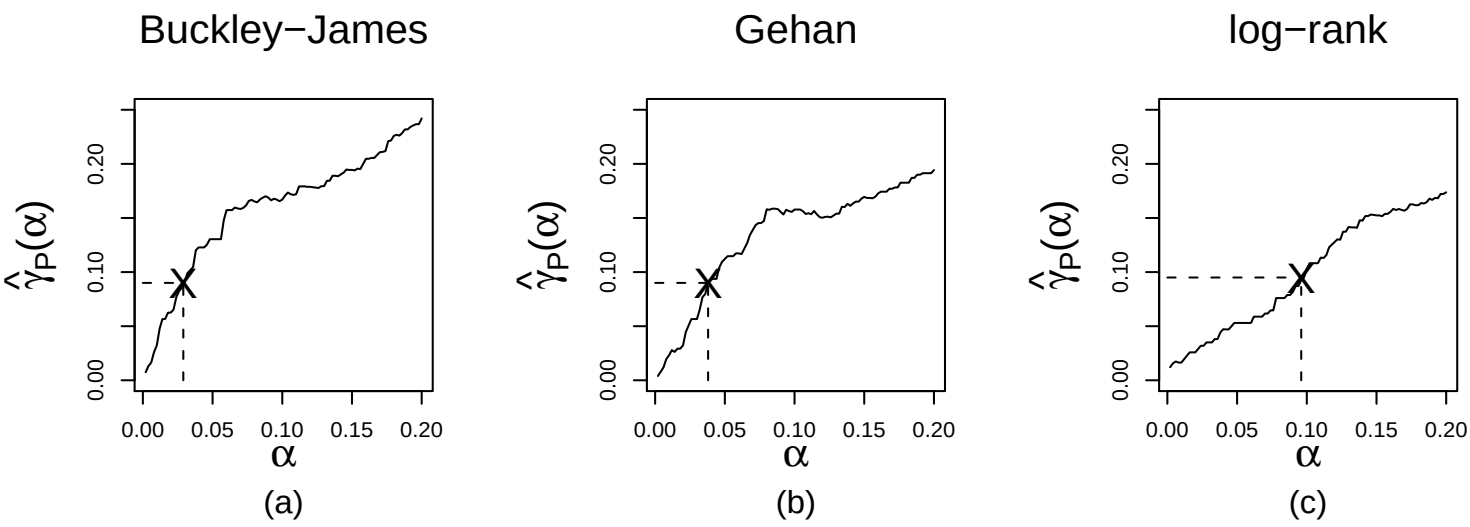

Figure 2: The proportion of pseudo variables as a function of α-to-enter for the Mayo PBC data. The final cutoffs ζ_0 leads to $\hat{\alpha}_0 = 0.032,\ 0.038$ and 0.096 for Buckley-James, Gehan, and log-rank statistics, respectively.

In Table 1, we see that age, albumin, bilirubin, edema, and prothrombin time are strong predictors of mortality for untreated PBC amidst other important variables, which is consistent with the findings of Dickson et al. (1989). In addition, ascites, sex, and vascular spiders appear to be weakly important variables for Buckley-James and Gehan statistics, but not as important for log-rank statistics.

A summary of the FSR procedure is given in Table 2 and displays of the proportion of pseudopredictors $\widehat{\gamma}_P(\alpha)$ given in Figure 2. Table 2 gives the sequence of variables entering the model, the generalized Wald test statistics, and approximate p-values. Our implementation of FSR with $\gamma_0 = 0.05$ leads to the same five-variable model (including age, albumin, bilirubin, edema, and prothrombin time) for all three statistics. The order of variables entering the final model agree generally: bilirubin and edema enter first or second, age enters third, albumin and prothrombin time enter fourth or fifth. Three variables appear to be weakly associated with mortality (ascites, sex, spiders) in Buckley-James and Gehan statistics but do not enter in the final model at the $\gamma = 0.05$ level. The estimated α-to-enter for Buckley-James, Gehan, and log-rank statistics were $\hat{\alpha}_0 = 0.032,\ 0.038$ and 0.096, respectively. We found it surprising that $\hat{\alpha}_0$ for the log-rank statistic was more than twice as large than that of Buckley-James and Gehan when $\gamma_0 = 0.05$. This difference may be related, in part, to the larger gap in significance level between the fifth and sixth

Table 2: Forward selection for Mayo primary biliary cirrhosis data via generalized Wald tests with an underline indicating the FSR stopping rule for 5% falsely selected variables, i.e. $\gamma_0 = 0.05$. Variable index number (Index) is given in the sequence of variables entered via forward selection with indices found in Table 1. The test statistic T and p-value are also given for each step in the sequence.

Buckley-James			Gehan			log-rank		
Index	T	$P(T > \chi^2_1)$	Index	T	$P(T > \chi^2_1)$	Index	T	$P(T > \chi^2_1)$
5.	83.78	$< 10^{-4}$	6.	92.89	$< 10^{-4}$	6.	66.92	$< 10^{-4}$
6.	38.25	$< 10^{-4}$	5.	75.41	$< 10^{-4}$	5.	85.77	$< 10^{-4}$
1.	21.96	$< 10^{-4}$	1.	25.38	$< 10^{-4}$	1.	17.80	$< 10^{-4}$
8.	11.24	$< 10^{-3}$	8.	15.56	$< 10^{-4}$	2.	9.40	< 0.01
2.	7.03	0.01	2.	6.29	0.01	8.	7.78	< 0.01
4.	2.17	0.14	10.	2.61	0.11	9.	1.97	0.16
10.	2.45	0.12	4.	2.60	0.11	4.	1.03	0.31
9.	2.21	0.14	9.	2.70	0.10	7.	0.99	0.32
7.	0.95	0.33	7.	0.90	0.34	10.	0.17	0.68
3.	0.05	0.83	3.	0.01	0.98	3.	0.13	0.72

variable entered sequentially using the log-rank statistic compared with either Buckley-James or Gehan. We found that FSR with $\gamma_0 = 0.1$ will result in a final model which includes ascites, sex, and spiders for Buckley-James and Gehan statistics. Even at $\gamma_0 = 0.1$, FSR using log-rank statistics did not include the next variable in the sequence, i.e. sex. This phenomenon may be partially explained by Figure 2, where for α in a neighborhood about 0.10, $\widehat{\gamma}_P(\alpha)$ appears to flatten out in panels (a)-(b).

In Figure 3, we compare the final model results among FSR, SCAD, Hard thresholding, and LASSO. We display approximate 95% percent confidence intervals of coefficient estimates for each independent variable entering the final model. The variable index number is defined in Table 1 and given across the ordinate. We find that for the ten main effects considered, FSR and SCAD agree generally on five variables in final model. The hard thresholding penalty and LASSO include at least one more variable depending on the statistic. The log-rank statistic with hard or LASSO penalty includes sex in addition to the five variables in the FSR/SCAD model. Both Buckley-James and Gehan statistics include ascites, sex, and spiders using the hard thresholding penalty but drop sex when using the LASSO penalty.

7 Remarks

This paper describes methods for selecting variables in the semiparametric linear regression model for censored outcomes and is a summary of research presented in several talks by the author, starting with one given at Brown University in January 2005. The methods are novel and differ fundamentally from other methods (e.g. Tibshirani, 1996; Fan and Li, 2001) for censored data in that we do not require a proportional hazards assumption. Under certain regularity conditions on the penalty functions, we derive the large sample properties for penalized weighted log-rank statistics and give an approximate result for penalized Buckley-James statistics. We note that the technical proofs and numerical algorithms are quite general and may be applicable to wider class of non-smooth estimating functions under suitable regularity conditions. In addition to penalized estimating function methods, we also describe a variable selection method based on controlling the false selection rate (FSR). The large sample properties of FSR for Buckley-James and weighted log-rank statistics are difficult to derive, in part, because they rely the asymptotic properties of forward selection. In simulation studies, we found that FSR and SCAD have similar operating characteristics and perform better than LASSO when the model fraction is low while LASSO performs better than FSR and SCAD when the model fraction is high. In the (generalized)

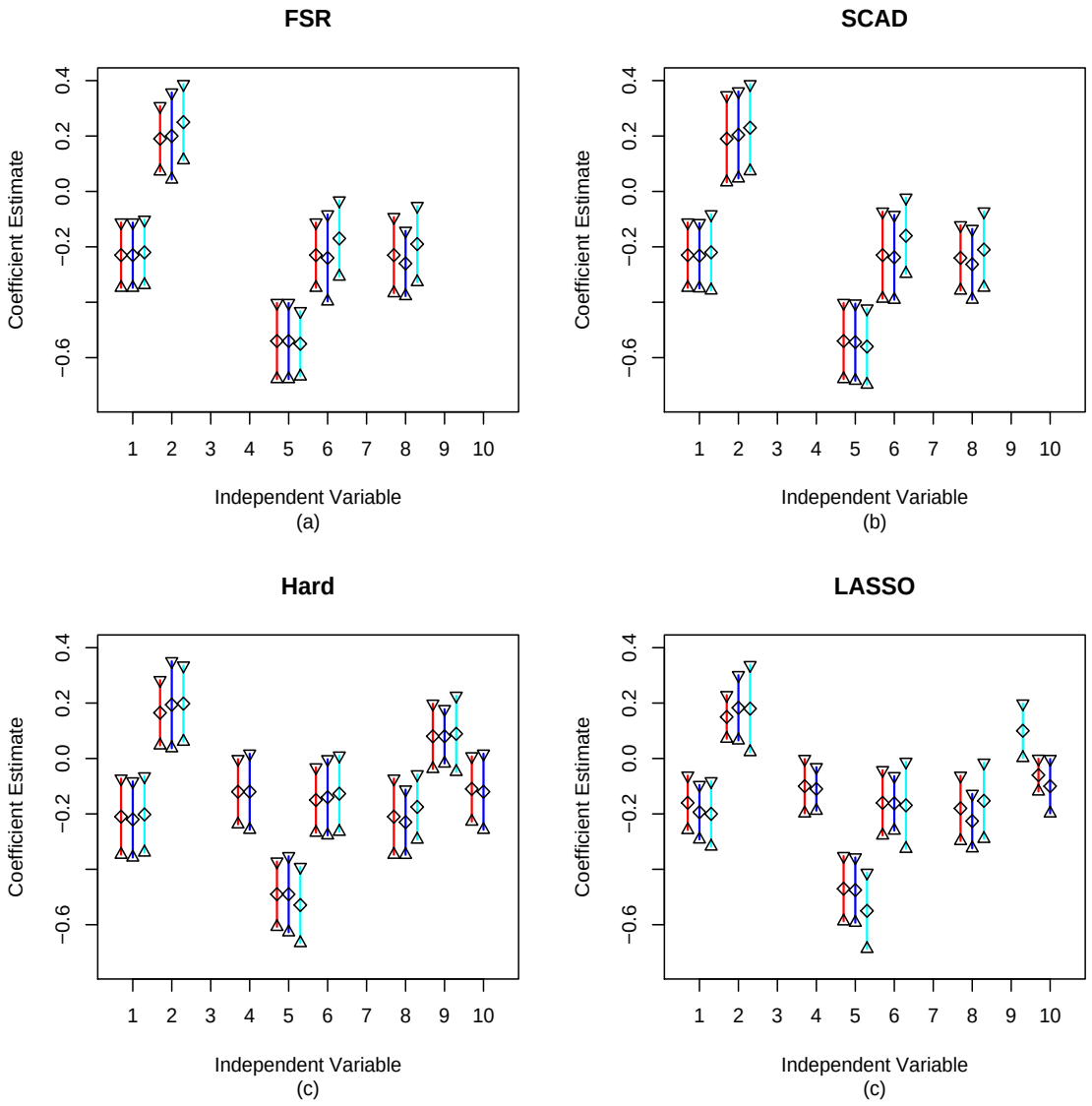

Figure 3: Analysis results for Mayo primary biliary cirrhosis data. Vertical lines represent approximate 95% confidence intervals for each of ten potential independent variables included in the final model. Three lines for each independent variable represent estimates from Buckley-James (left), Gehan (center), and log-rank (right) statistics

linear regression model setting, Wu, Boos, and Stefanski (2007) give heuristic arguments that FSR is a reliable, approximate method to control the false selection rate where no similar methods exist. Simulation studies suggest that the FSR method has the desired properties and the algorithm, albeit computationally intensive, is relatively straightforward to implement.

The penalized estimating function methods presented here (as with all penalized likelihood and penalized least squares methods) depend on the selection of a regularization parameter λ that controls the shrinking of parameter estimates. For a given data set, we proposed selecting λ by minimizing the function GCV(λ). The numerator of GCV(λ) is the loss function $L_G(\boldsymbol{\beta})$, which reduces to the well-known Wilcoxon objective function for uncensored data. An anonymous referee noted that other loss functions may also be substituted for $L_G(\boldsymbol{\beta})$; in particular, one can attempt to replace RSS(λ)/n in $\mathrm{GCV}_{\mathrm{LS}}(\lambda)$ with, say, its conditional expectation given the observed data. In earlier drafts, we considered inversely-weighted complete case estimators for RSS(λ)/n while an anonymous referee suggested the Kaplan-Meier estimator based on $\{(Y_i - \hat{\alpha} - \widehat{\boldsymbol{\beta}}_\lambda^T \mathbf{X}_i, \Delta_i),\ i = 1, \ldots, n\}$ (for $\hat{\alpha}$ an estimate of $\alpha = E(\epsilon_1)$) or the bivariate Kaplan-Meier estimator of Stute (1993). Generally, these methods must wrestle with thorny issues including (i) the estimation of $E(\epsilon_1)$, or (ii) stable tail behaviour in the distribution of the errors $e_i(\boldsymbol{\beta})$ in (3), the censoring distribution, or both. Each of (i) and (ii) can become problematic and lead to theoretical and numerical challenges. At the same time, under certain conditions and depending on the particular goals of model selection for a given data set, there may be several reasonable cross-validation measures which deserve attention. A careful comparison between such measures is of interest (to the author, anyway) but beyond the scope of the current paper. In any case, the asymptotic properties in Section 3 and numerical algorithms in Section 4 are valid irrespective of the cross-validation technique.

On a related topic, deriving explicit expressions for optimal smoothing parameters has received little attention in the model selection literature. We take this as mild evidence that it is a difficult problem, even in least squares regression with uncensored data. In penalized least squares regression with L_1 penalty, however, some related work has been completed; notable papers include Huang (2003), Rosset and Zhu (2004). These papers argue that, with probability tending to one, there exists a range of regularization parameters λ such that the average squared prediction error of the lasso estimator is less than that of ordinary least squares. The proofs of these results generally rest on the property that $\widehat{\boldsymbol{\beta}}_\lambda$ is piece-wise linear (Osborne et al., 2000; Efron et al. 2004). In this paper, we considered penalized estimators which do not correspond necessarily to be the minimizer of any loss function. Even with a proper penalized loss

function with L_1 penalty, e.g. generalized linear models, the coefficients paths $\widehat{\boldsymbol{\beta}}_\lambda$ are no longer piece-wise linear. In short, the earlier arguments of Rosset and Zhu (2004) do not extend to the current model nor to many other models where cross-validation is in common usage.

Acknowledgements

We acknoweldge the comments of the associate editor and two anonymous referees which improved the manuscript significantly. The research of the author was supported in part by grants from the National Institutes for the Environmental Health Sciences (P30ES10126, T32ES007018), the National Institutes of Allergies and Infectious Diseases (R03AI068484), and Emory's Center for AIDS Research. We also acknowledge the assistance of Donglin Zeng with technical details related to the proof of Theorem 1 and thank Eugene Huang and Dennis Boos for many helpful discussions. Finally, we thank Wu, Boos, and Stefanski for allowing us to view a version of their manuscript before publication.

Appendix 1

Regularity Conditions

We impose the following regularity conditions to prove statements in Section 3. These are given as Conditions A1, A2, A4-A7,A9-A10 in Strawderman (2005, p.662) to yield a consistent, asymptotically normal estimator sequence $\widehat{\boldsymbol{\beta}}_W$. We also note that for inference in the semiparametric accelerated failure time model, similar assumptions appear as conditions (A)-(F) in Tsiatis (1990, p.357-358) and conditions (A.1)-(A.3) in Ritov (1990, p. 306).

A1. We require the error distribution F in (1) satisfy $F(t) = \int_0^t f(u)\,du$, where $f(\cdot)$ has continuous first and bounded second derivatives.

A2. We require that random vectors $\{N_i(t, \boldsymbol{\beta}_0), I(e_i(\boldsymbol{\beta}_0) \geq t), \mathbf{X}_i, t \geq 0\}$ are independent and identically distributed.

A3. Censoring is noninformative.

A4. We require that $\mathbf{X}_1$ is bounded with probability one.

A5. Let $s^{(j)}(t,\boldsymbol{\beta}) = \lim_{n\to\infty} S^{(j)}(t,\boldsymbol{\beta}),\ j = 0,1,2$ defined in Section 2.1. We require that $s^{(j)}(t,\boldsymbol{\beta}),\ j = 0,1,2$ are continuous for $(t,\boldsymbol{\beta}) \in \Re^+ \times \mathcal{N}(\boldsymbol{\beta}_0)$, where $\mathcal{N}(\boldsymbol{\beta}_0)$ is some neighborhood of $\boldsymbol{\beta}_0$.

A6. The upper limit of integration τ satisfies $\inf\{t \in [0,\tau] : s^{(0)}(t,\boldsymbol{\beta}_0)\} > 0$.

A7. For $(t,\boldsymbol{\beta}_0) \in [0,\tau] \times \mathcal{N}(\boldsymbol{\beta}_0)$, $W(t,\boldsymbol{\beta})$ is of bounded variation and there exists a continuous, bounded deterministic function $w(t,\boldsymbol{\beta})$ such that $|W(t,\boldsymbol{\beta}) - w(t,\boldsymbol{\beta})|$ converges to zero in probability uniformly.

A8. The asymptotic slope matrix $\mathbf{A}_W$ is nonsingular.

Condition A1 constrains the density of the errors ϵ to be sufficiently smooth. Condition A2 ensures that statistics defined as sample averages follow an ordinary weak law of large numbers. Condition A3 allows consistent estimation of $\boldsymbol{\beta}_0$ without restrictive assumptions (and knowledge) on the underlying censoring mechanism. Conditions A1-A4 allow one to construct martingale processes with desirable features (e.g. orthogonal, square-integrable, mean zero) that lead to an elegant way for studying the limiting behaviour of the score function $n^{-1/2}\mathbf{U}_W(\boldsymbol{\beta})$. Condition A5 is used for placing bounds on the score function in neighbourhoods of $\boldsymbol{\beta}_0$. Condition A6 controls unstable tail behaviour and ensures that the denominator of $\tilde{\mathbf{X}}\{t,\boldsymbol{\beta}\}$ is positive for all $t, t \le \tau$. Condition A7 restricts the class of weight functions which lead to a well-behaved score function. Conditions A1-A7 imply that $\mathbf{A}_W(\boldsymbol{\beta})$ exists and is continuous at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Conditions A5 plus A8 ensure that $\boldsymbol{\beta}_0$ is unique. Additional details on the roles of these assumptions may be found in Strawderman (2005).

Appendix 2

Proof of Theorem 1

We first prove part (i). Note that $\sqrt{n}\mathbf{b}_{\lambda_n}(\boldsymbol{\beta}_0) \to 0$ by condition Q1(ii). Consider $\boldsymbol{\beta}$ such that $|\boldsymbol{\beta} - \boldsymbol{\beta}_0| \le Mn^{-1/2}$. By conditions A1-A8 in Appendix 1 and condition Q1(ii),

$$n^{-1/2}\mathbf{U}_W^P(\boldsymbol{\beta}) = n^{-1/2}\mathbf{U}_W(\boldsymbol{\beta}_0) + \sqrt{n}\mathbf{A}_W(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + r_n - \sqrt{n}\{\mathbf{b}_{\lambda_n}(\boldsymbol{\beta}) - \mathbf{b}_{\lambda_n}(\boldsymbol{\beta}_0)\} \tag{A.1}$$

for some random variable $r_n = o_p(1)$. If $\beta_{0j} \neq 0$, then $\mathrm{sgn}(\beta_j) = \mathrm{sgn}(\beta_{0j})$, so that

$$q_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) - q_{\lambda_n}(|\beta_{0j}|)\mathrm{sgn}(\beta_{0j}) = \{q_{\lambda_n}(|\beta_j|) - q_{\lambda_n}(|\beta_{0j}|)\}\mathrm{sgn}(\beta_j);$$

if $\beta_{0j} = 0$, the above equation holds naturally. Thus, it follows from the mean-value theorem that

$$n^{-1/2}\mathbf{U}_W^P(\boldsymbol{\beta}) = n^{-1/2}\mathbf{U}_W(\boldsymbol{\beta}_0) + \sqrt{n}\mathbf{A}_W(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + r_n$$

$$-\sqrt{n}\text{diag}\left\{q'_{\lambda_n}(|\beta_j^*|)\text{sgn}(\beta_j)\right\}(\boldsymbol{\beta} - \boldsymbol{\beta}_0). \tag{A.2}$$

where β_j^* lies between β_j and β_{j0}. Let $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 - n^{-1/2}\mathbf{A}_W^{-1}n^{-1/2}\mathbf{U}_W(\boldsymbol{\beta}_0)$. By conditions A1-A8 from Appendix 1, $\widehat{\boldsymbol{\beta}}_n$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}_0$. Thus, for any $\epsilon > 0$ and $\delta > 0$, there exists an M such that $P(|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0| < Mn^{-1/2}$ and $|r_n| < \delta) > 1 - \epsilon$. On this set, we have

$$|n^{-1/2}\mathbf{U}_W^P(\widehat{\boldsymbol{\beta}}_n)| \leq \sqrt{n}\sum_{j=1}^{d}|q'_{\lambda_n}(|\beta_j^*|)||\beta_j - \beta_{0j}| + |r_n|.$$

On the other hand, if $\beta_{0j} \neq 0$, $q'_{\lambda_n}(|\beta_j^*|) = q'_{\lambda_n}(|\beta_{0j}|) + o(1) \to 0$ by condition Q1(ii); if $\beta_{0j} = 0$, $q'_{\lambda_n}(|\beta_j^*|) \leq \sup_{|\beta|\leq Mn^{-1/2}}|q'_{\lambda_n}(|\beta|)| \to 0$ by condition Q1(iii). As a result, when n is large, $|n^{-1/2}\mathbf{U}_W^P(\widehat{\boldsymbol{\beta}}_n)| \leq 2\delta$. That is, $P(|n^{-1/2}\mathbf{U}_W^P(\widehat{\boldsymbol{\beta}}_n)| > 2\delta) < \epsilon$. Therefore, $n^{-1/2}\mathbf{U}_W^P(\widehat{\boldsymbol{\beta}}_n) = o_p(1)$.

To prove the second half of part (i), we consider $\boldsymbol{\beta}$ on the boundary of a ball around $\boldsymbol{\beta}_0$, i.e., $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}$ with $|\mathbf{u}| = r$ for a fixed constant r. By equation (A.2), we have

$$\begin{aligned}
& n^{-1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{A}_W^T\mathbf{U}_W^P(\boldsymbol{\beta}) \\
= \;& (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{A}_W^T\left\{n^{-1/2}\mathbf{U}_W(\boldsymbol{\beta}) - n^{1/2}\mathbf{b}_{\lambda_n}(\boldsymbol{\beta})\right\} \\
= \;& O_p(|\boldsymbol{\beta} - \boldsymbol{\beta}_0|) + n^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{A}_W^T\mathbf{A}_W(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
& -n^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\mathbf{A}_W^T\text{diag}\{q'_{\lambda_n}(|\beta_j^*|)\text{sgn}(\beta_j)\}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).
\end{aligned}$$

Since $\mathbf{A}_W$ is nonsingular, the second term on the right-hand side is larger than $a_0r^2n^{-1/2}$, where a_0 is the smallest eigenvalue of $\mathbf{A}_W^T\mathbf{A}_W$. The first term is of order $rO_p(n^{-1/2})$. As before, $\max_j q'_{\lambda_n}(|\beta_j^*|) \to 0$, so the third term is dominated by the second term. Therefore, for any ϵ, if we choose r large enough so that for large n, the probability that the absolute value of the first term is larger than the second term is less than ϵ, then we have

$$P\left[\min_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|=n^{-1/2}r}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{A}_W^T\mathbf{U}_W^P(\boldsymbol{\beta}) > 0\right] > 1 - \epsilon.$$

Applying the Brouwer fixed-point theorem to the continuous function $\mathbf{U}_W^P(\boldsymbol{\beta})$, we see that

$$\min_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|=n^{-1/2}r}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\mathbf{A}_W^T\mathbf{U}_W^P(\boldsymbol{\beta}) > 0$$

implies that $\mathbf{A}_W^T\mathbf{U}_W^P(\boldsymbol{\beta})$ has a solution within this ball, or equivalently, $\mathbf{U}_W^P(\boldsymbol{\beta})$ has a solution within this ball. That is, we can choose an exact solution $\widehat{\boldsymbol{\beta}}_n$ to $\mathbf{U}_W^P(\boldsymbol{\beta}) = \mathbf{0}$ with $\widehat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + O_p(n^{-1/2})$.

To prove part (ii), we consider the sets in the probability space: $C_{nj} = \left\{\widehat{\beta}_{nj} \neq 0\right\}$, $j = d_N + 1, \ldots, d$. It suffices to show that for any $\epsilon > 0$, when n is large enough, $P(C_{nj}) < \epsilon$. Since $\widehat{\beta}_{nj} = O_p(n^{-1/2})$, there exists some M such that when n is large enough,

$$P(C_{nj}) < \epsilon/2 + P\left\{\widehat{\beta}_{nj} \neq 0, |\widehat{\beta}_{nj}| < Mn^{-1/2}\right\}.$$

Using the jth component of (A.1), we obtain

$$o_p(1) = n^{-1/2}\mathbf{U}_{Wj}(\widehat{\boldsymbol{\beta}}_0) + \sqrt{n}\mathbf{A}_{Wj}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + o_p(1) - \sqrt{n}q_{\lambda_n}(|\widehat{\beta}_{nj}|)\mathrm{sgn}(\widehat{\beta}_{nj}).$$

The first three terms on the right-hand side are of order $O_p(1)$. As a result, there exists some M' such that for large n,

$$P(\sqrt{n}q_{\lambda_n}(|\widehat{\beta}_{nj}|) > M') < \epsilon/2.$$

Since

$$\sqrt{n}\lambda_n \lim_n \lambda_n^{-1} \inf_{|\beta| \leq Mn^{-1/2}} q_{\lambda_n}(|\beta|) \to \infty$$

by condition Q1, $\widehat{\beta}_{nj} \neq 0$ and $|\widehat{\beta}_{nj}| < Mn^{-1/2}$ imply that $\sqrt{n}q_{\lambda_n}(|\widehat{\beta}_{nj}|) > M'$ for large n. Therefore,

$$P(C_{nj}) < \epsilon/2 + P(\sqrt{n}q_{\lambda_n}(|\widehat{\beta}_{nj}|) > M') < \epsilon.$$

To prove part (iii), we use the following fact:

$$o_p(1) = n^{-1/2}\mathbf{U}_W^P(\widehat{\boldsymbol{\beta}}_n) = o_p(1) + n^{-1/2}\mathbf{U}_W(\boldsymbol{\beta}_0) + \sqrt{n}\mathbf{A}_W(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) - \sqrt{n}\mathbf{b}_{\lambda_n}(\widehat{\boldsymbol{\beta}}_n).$$

We consider the first d_N components of the above expression. After the Taylor series expansion of the last term, we conclude that

$$\sqrt{n}\left\{\mathbf{A}_{W11} + \boldsymbol{\Sigma}_{11})(\widehat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{01} + (\mathbf{A}_{W11} + \boldsymbol{\Sigma}_{11})^{-1}\mathbf{b}_{\lambda_n}(\boldsymbol{\beta}_0)\right\} = \sqrt{n}\begin{pmatrix} U_{W1}^P(\boldsymbol{\beta}_0) \\ \vdots \\ U_{Wd_N}^P(\boldsymbol{\beta}_0) \end{pmatrix} + o_p(1) \to_d N(0, \mathbf{B}_{W11}).$$

Appendix 3

Simulated pseudo predictors.

The FSR method uses p simulated *pseudo variables* for each of n subjects over M augmented data sets $\mathcal{A}_1, \ldots, \mathcal{A}_M$. Here, we give details on how vectors of pseudo variables $\mathbf{X}_i^*$ are defined. Rather than simply simulating standard normal variates, numerical studies in Wu et al. (2007) suggest there is some advantage in defining pseudo variables through random permutations of the original design matrix $\mathbf{X}$. For simplicity, assume that $p = d$ and let $\mathcal{P}(\mathbf{X})$ be a $n \times d$ randomly permuted design matrix. If $\mathbf{x}_j = (X_{1j}, \ldots, X_{nj})^T$, then the j-th column of $\mathcal{P}(\mathbf{X})$ is perm$(\mathbf{x}_j)$, where perm$(\cdot)$ is the random permutation. Let $\mathbf{X}^* = (\mathbf{X}_1^{*T}, \ldots, \mathbf{X}_n^{*T})^T$ be an $n \times d$ block of pseudo variables defined through

$$\begin{aligned} \mathbf{X}^* &= (I_n - P_X)\mathcal{P}(\mathbf{X}) \\ P_X &= (\mathbf{1}_n, \mathbf{X}) \left\{ (\mathbf{1}_n, \mathbf{X})^T (\mathbf{1}_n, \mathbf{X}) \right\}^{-1} (\mathbf{1}_n, \mathbf{X})^T \end{aligned}$$

where I_n is an n-dimensional identity matrix and $\mathbf{1}_n$ is an n-dimensional column vector of ones. Defined in this way, the pseudo variables have desirable characteristics such as having the same sample moments as the original columns in the design matrix and orthogonality between the new pseudo variables and the original d covariables. As Wu et al. allow $p > d$, then defining the columns of $\mathcal{P}(\mathbf{X})$ through modular arithmetic allows for a precise definition of $\mathbf{X}^*$. Finally, Wu et al. note that having more than d pseudo variables does not significantly improve FSRs performance in small samples and, in the current presentation, the additional notational complexity is unnecessary.

References

Buckley, J. and James, I. (1979) "Linear Regression With Censored Data," *Biometrika*, **66**, 429-436.

Boos, D. D. (1992) "On Generalized Score Tests," *The American Statistician*, **46**, 327-333.

Cox, D. R. (1972) "Regression Models and Life-Tables" (with Discussion), *Journal of the Royal Statistical Society, Ser. B*, **34**, 187-202.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., and Langworth, A. (1989). "Prognosis in Primary Biliary Cirrhosis: Model for Decision Making," *Hepatology*, **10**, 1–7.

Efron, B., Hastie, T., Johnstone, I. M., and Tibshirani, R. (2004) "Least Angle Regression (with Discussion)," *The 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**, 1348-1360.

Fan, J. and Li, R. (2002) "Variable Selection for Cox's Proportional Hazards Model and Frailty Model," *The Annals of Statistics*, **30**, 74-99.

Faraggi, D. and Simon, R. (1998) "Bayesian Variable Selection Method for Censored Survival Data," *Biometrics*, **54**, 1475-1485.

Fleming, T. A. and Harrington, D. P. (1991). *Counting Processes and Survival Analyses*. New York: Wiley.

Fygenson, M. and Ritov, Y. (1994) "Monotone Estimating Equations for Censored Data," *Annals of Statistics*, **22**, 732-46.

Hoerl, A. E. and Kennard, R. W. (1970) "Ridge Regression: Biased Estimation for Non-Orthogonal Problems," *Technometrics*, **12**, 55-67.

Huang, F. (2003) "Prediction Error Property of the LASSO Estimator and its Generalization," *Aust. N. Z. J. Stat.* **45**, 217-228.

Huang, Y. (2002) "Calibrated Regression of Censored Lifetime Medical Cost," *Journal of the American Statistical Association*, **97**, 318-327.

Knight, K. and Fu, W. (2000) "Asymptotics for LASSO-Type Estimators," *The Annals of Statistics*, **28**, 1356-1378.

Lai, T. L. and Ying, Z. (1991a) "Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis With Censored Data," *The Annals of Statistics*, **19**, 1370-1402.

Lai, T. L. and Ying, Z. (1991b) "Rank Regression Methods for Left Truncated and Right Censored Data," *The Annals of Statistics*, **19**, 531-556.

McKean, J. W. (2004) "Robust Analysis of Linear Models," *Statistical Science*, **19**, 562-570.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000) "On the LASSO and its dual," *J. Comput. Graph. Statist.*, **9**, 319-337.

Prentice, R. L. (1978) "Linear Rank Tests With Right-Censored Data," *Biometrika*, **65**, 167-179.

Ritov, Y. (1990) "Estimation in a Linear Regression Model With Censored Data," *The Annals of Statistics*, **18**, 303-328.

Rosset, S. and Zhu, J. (2004) "Corrected Proof of the Result of 'A Prediction Error Property of the LASSO Estimator and Its Generalization' by Huang (2003)," *Aust. N. Z. J. Stat.*, **46**, 505-510.

Rotnizky, A., and Jewell, N. P. (1990) "Hypothesis Testing of Regression Parameters in Semiparametric Generalized Linear Models for Cluster Correlated Data," *Biometrika*, **77**, 485-497.

Strawderman, R. L. (2005) "The Accelerated Gap Times Model," *Biometrika*, **92**, 647-666.

Stute, W. (1993) "Consistent Estimation Under Random Censorship When Covariables Are Present," *J. Multiv. Analysis*, **45**, 89-103.

Tibshirani, R. J. (1996) "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society, Ser. B*, **58**, 267-288.

Tibshirani, R. J. (1997) "The LASSO Method for Variable Selection in the Cox Model," *Statistics in Medicine*, **16**, 385-395.

Tibshirani, R. J. and Knight, K. (1999) "The Covariance Inflation Criterion for Adaptive Model Selection," *Journal of the Royal Statistical Society, Ser. B*, **61**, 529-546.

Tsiatis, A. A. (1990) "Estimating Regression Parameters Using Linear Rank Tests for Censored Data," *The Annals of Statistics*, **18**, 354-372.

Wei, L. J., Ying, Z., and Lin, D. Y. (1990) "Regression Analysis of Censored Survival Data Based on Rank Tests," *Biometrika*, **77**, 845-851.

Wu, Y., Boos, D. B., and Stefanski, L. A. (2007) "Controlling Variable Selection By the Addition of Pseudo Variables," *Journal of the American Statistical Association*, **102**, 235-243.

Ying, Z. (1993) "A Large Sample Study of Rank Estimation for Censored Regression Data," *The Annals of Statistics*, **21**, 76-99.