Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models

Brent A. Johnson, D. Y. Lin and Donglin Zeng*

We propose a general strategy for variable selection in semiparametric regression models by penalizing appropriate estimating functions. Important applications include semiparametric linear regression with censored responses and semiparametric regression with missing predictors. Unlike the existing penalized maximum likelihood estimators, the proposed penalized estimating functions may not pertain to the derivatives of any objective functions and may be discrete in the regression coefficients. We establish a general asymptotic theory for penalized estimating functions. We present suitable numerical algorithms to implement the proposed estimators. In addition, we develop a resampling technique to estimate the variances of the estimated regression coefficients when the asymptotic variances cannot be evaluated directly. Simulation studies demonstrate that the proposed methods perform well in variable selection and variance estimation. We illustrate our methods using data from the Paul Coverdell Stroke Registry.

Keywords: Accelerated failure time model; Buckley-James estimator; censoring; LASSO; Least-squares; linear regression; missing data; SCAD.

---

* Brent A. Johnson is Assistant Professor, Department of Biostatistics, Emory University, Atlanta, GA 30322 (email: bajohn3@emory.edu). D. Y. Lin is Dennis Gillings Distinguished Professor and Donglin Zeng is Associate Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7420 (emails: lin@bios.unc.edu; dzeng@bios.unc.edu). This research was supported by the NIH grants P30 ES10126, T32 ES007018, R03 AI068484 (for Johnson), R37 GM047845 (for Lin) and R01 CA082659 (for Lin and Zeng). We thank Paul Weiss for preparing the stroke data set.

1 Introduction

A major challenge in regression analysis is to decide which predictors, among many potential ones, are to be included in the model. It is customary to use stepwise selection and subset selection. These procedures, however, are unstable and ignore the stochastic errors introduced by the selection process. Several methods, including bridge regression (Frank and Friedman, 1993), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), elastic net (EN) (Zou and Hastie, 2005), and adaptive lasso (ALASSO) (Zou, 2006) have been proposed to select variables and estimate their regression coefficients simultaneously. These methods can be cast in the framework of penalized least-squares and likelihood.

Consider the linear regression model

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

where Y_i is the response variable and $\mathbf{x}_i$ a d-vector of predictors for the ith subject, $\boldsymbol{\beta}$ is a d-vector of regression coefficients, and $(\varepsilon_1, \ldots, \varepsilon_n)$ are independent and identically distributed errors. For simplicity, assume that the ε_i have zero means. Define $l(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, where $\mathbf{y} = (Y_1, \ldots, Y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$. Then the penalized least-squares estimator of $\boldsymbol{\beta}$ is the minimizer of the objective function

$$l(\boldsymbol{\beta}) + n \sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{2}$$

where $p_\lambda(\cdot)$ is a penalty function. Appropriate choices of p_λ (detailed in Section 2) yield the aforementioned variable selection procedures. For likelihood-based models, the penalized maximum likelihood estimator is obtained by setting $l(\boldsymbol{\beta})$ to the minus log-likelihood.

For many semiparametric problems, the estimation of regression coefficients (without the task of variable selection) does not pertain to the minimization of any objective function. Important examples include weighted estimating equations for missing data (Robins et al., 1994; Tsiatis, 2006) and the Buckley-James (1979) estimator for semiparametric

linear regression with censored responses. Another example arises from Lin and Ying's (2001) semiparametric regression analysis of longitudinal data. For the last example, Fan and Li (2004) proposed a variable selection method by incorporating the SCAD penalty into Lin and Ying's estimator. They noted that their estimator may be cast in the form (2), so that their earlier results (Fan and Li, 2001) for penalized least squares could be applied. In this paper, we go beyond specific problems and provide a very general theory for a broad class of penalized estimating functions. In this regard, only Fu's (2003) work on generalized estimating equations (GEE; Liang and Zeger, 1986) with bridge penalty (Frank and Friedman, 1993; Knight and Fu, 2000) is similar. That work only deals with smooth estimating functions whereas our theory applies to very general, possibly discrete estimating functions. In addition, we present general computational strategies.

The remainder of the article is organized as follows. We present our penalized estimating functions in Section 2, paying special attention to the aforementioned missing data and censored data problems. We state the asymptotic results in Section 3 and address implementation issues in Section 4. We report the results of our simulation studies in Section 5 and apply the methods to real data in Section 6.

2 Penalized estimating functions

2.1 General setting

Suppose that $\mathbf{U}(\boldsymbol{\beta}) \equiv (U_1(\boldsymbol{\beta}), \ldots, U_d(\boldsymbol{\beta}))^T$ is an estimating function for $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_d)^T$ based on a random sample of size n. For maximum likelihood estimation, $\mathbf{U}(\boldsymbol{\beta})$ is simply the score function. We are mainly interested in the situations where $\mathbf{U}(\boldsymbol{\beta})$ is not a score function or the derivative of any objective function. A penalized estimating function is defined as

$$\mathbf{U}^P(\boldsymbol{\beta}) = \mathbf{U}(\boldsymbol{\beta}) - n\mathbf{q}_\lambda(|\boldsymbol{\beta}|)\text{sgn}(\boldsymbol{\beta}),$$

where $\mathbf{q}_\lambda(|\boldsymbol{\beta}|) = (q_{\lambda,1}(|\beta_1|), \ldots, q_{\lambda,d}(|\beta_d|))^T$, $q_{\lambda,j}(\cdot),\ j = 1, \ldots, d$ are coefficient-dependent continuous functions, and the second term is the component-wise product of $\mathbf{q}_\lambda$ and $\text{sgn}(\boldsymbol{\beta})$. In most cases, $q_{\lambda,j} = p'_{\lambda,j}$ for some penalty function $p_{\lambda,j}$ and the functions $q_{\lambda,j},\ j = 1, \ldots, d$ are the same for all d components of $\mathbf{q}_\lambda(|\boldsymbol{\beta}|)$, i.e. $q_{\lambda,j} = q_{\lambda,k}, j \neq k$. When the functions $q_{\lambda,j},\ j = 1, \ldots, d$ do not vary with j, we drop the subscript for simplicity and ease of notation.

When $q_\lambda = p'_\lambda$, we consider five penalty functions: (a) the LASSO penalty (Tibshirani, 1996, 1997) $p_\lambda(|\theta|) = \lambda|\theta|$, (b) the hard thresholding penalty (Antoniadis, 1997) $p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$, (c) the SCAD penalty (Fan and Li, 2001, 2002, 2004) defined by

$$p'_\lambda(|\theta|) = \lambda \left\{ I(|\theta| < \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| \geq \lambda) \right\}$$

for $a > 2$, (d) EN penalty (Zou and Hastie, 2005) $p_\lambda(|\theta|) = \lambda_1|\theta| + \lambda_2\theta^2$, and (e) ALASSO penalty (Zou, 2006) $p_{\lambda,j}(|\theta|) = \lambda|\theta|\omega_j$, for a known, data-driven weight ω_j. In our applications, we use the weight $\omega_j = 1/|\tilde{\beta}^o_j|,\ j = 1, \ldots, d$, where $\tilde{\boldsymbol{\beta}}^o = (\tilde{\beta}^o_1, \ldots, \tilde{\beta}^o_d)^T$ refers to the d-vector of regression coefficient estimates obtained from solving the original estimating equation: $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$.

The hard thresholding penalty is important because it corresponds to best subset selection and stepwise deletion in certain cases. The LASSO (Tibshirani, 1996, 1997) is one of the most popular shrinkage estimators. However, the LASSO has deficiencies; in particular, it is inconsistent for certain designs (Meinshausen and Bühlmann, 2006; Zou, 2006). Fan and Li (2001, 2002) attempted to avoid such deficiencies by constructing a new penalty function (SCAD) which results in an estimator that achieves an *oracle* property: that is, the estimator has the same limiting distribution as an estimator which knows the true model *a priori*. Recently, Zou (2006) introduced ALASSO which, like SCAD, achieves the oracle property and may have numerical advantages for some problems. Finally, Zou and Hastie (2005) introduced the mixture penalty EN to effectively select 'grouped' variables and has been popular in the statistical analysis of large data sets.

2.2 Application to censored data

Censoring is a common phenomenon in scientific studies (cf. Kalbfleisch and Prentice, 2002, p. 12). The presence of censoring causes major complications in the implementation of the penalized least-squares approach because the values of the Y_i are unknown for the censored observations. The problem is much simpler for the proportional hazards regression because the partial likelihood (1975) plays essentially the same role as the standard likelihood (Tibshirani, 1997; Fan and Li, 2002; Cai et al, 2005). However, the proportional hazards model may not be appropriate in some applications, especially when the response variable does not pertain to failure time.

Let Y_i and C_i denote, respectively, the response variable and censoring variable for the ith subject, $i = 1, \ldots, n$. The data consist of $(\widetilde{Y}_i, \Delta_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, where $\widetilde{Y}_i = \min(Y_i, C_i)$, $\Delta_i = I(Y_i \leq C_i)$, and $\mathbf{x}_i$ is a d-vector of predictors. We relate Y_i to $\mathbf{x}_i$ through the semiparametric linear regression model given in (1), where ε_i are independent and identically distributed with an unspecified distribution function $F(\cdot)$. We assume that Y_i is independent of C_i conditional on $\mathbf{x}_i$. When the response variable pertains to failure time, both Y_i and C_i are commonly measured on the log scale and model (1) is called the accelerated failure time model (Kalbfleisch and Prentice, 2002, p. 44).

Clearly,

$$E\left\{\Delta_i Y_i + (1-\Delta_i)E(Y_i|\Delta_i = 0)\big|\mathbf{x}_i\right\} = \alpha + \boldsymbol{\beta}^T\mathbf{x}_i,$$

and

$$E(Y_i|\Delta_i = 0) = \boldsymbol{\beta}^T\mathbf{x}_i + \frac{\int_{e_i(\boldsymbol{\beta})}^{\infty}\{1 - F(s)\}\,ds}{1 - F\{e_i(\boldsymbol{\beta})\}},$$

where $\alpha = E(\varepsilon_i)$ and $e_i(\boldsymbol{\beta}) = \widetilde{Y}_i - \boldsymbol{\beta}^T\mathbf{x}_i$. Thus, Buckley and James (1979) proposed the estimating function for $\boldsymbol{\beta}$

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\mathbf{x}_i\{\xi_i(\boldsymbol{\beta}) - \boldsymbol{\beta}^T\mathbf{x}_i\}, \tag{3}$$

where

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

and $\widehat{F}(t;\boldsymbol{\beta})$ is the Kaplan-Meier estimator of $F(t)$ based on $\{e_i(\boldsymbol{\beta}), \Delta_i\}$, $i = 1, \ldots, n$. If $\Delta_i = 1$ for all i, then the penalized estimating function $\mathbf{U}^P(\boldsymbol{\beta})$ corresponding to (3) becomes the penalized least-squares estimating function arising from (2). Thus, the penalized Buckley-James estimator is a direct generalization of the penalized least-squares estimator to censored data.

2.3 Application to missing data

It is often difficult to have complete data on all study subjects. Let R_i be the missingness indicator for the ith subject, the event $\{R_i = \infty\}$ indicating that the ith subject has complete data. The observed data for the ith subject is $G_r(\mathbf{Z}_i)$, where $G_r(\cdot)$ is the missingness operator acting on the full data $\mathbf{Z}_i$ of the ith subject when $R_i = r$. In simple linear regression, for example, we may only consider $R_i \in \{1, 2, \infty\}$ corresponding to $G_1(\mathbf{Z}_i) = \{Y_i\}$, $G_2(\mathbf{Z}_i) = \{x_i\}$, and $G_\infty(\mathbf{Z}_i) = \{Y_i, x_i\} = \mathbf{Z}_i$, respectively. The observed data are represented as $\{R_i, G_{R_i}(\mathbf{Z}_i),\ i = 1, \ldots, n\}$. We focus on monotone missingness and make two assumptions: (i) $P(R_i = \infty|\mathbf{Z}_i = \mathbf{z}) > \kappa > 0$, and (ii) $P(R_i = r|\mathbf{Z}_i = \mathbf{z}) = P(R_i = r|G_r(\mathbf{z}) = g_r)$.

Consider the semiparametric linear regression model given in (1). The weighted complete-case estimating function takes the form

$$\mathbf{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{I(R_i = \infty)s_i(\boldsymbol{\beta})}{\pi(\infty, \mathbf{Z}_i)},$$

where $s_i(\boldsymbol{\beta}) = \mathbf{x}_i(Y_i - \alpha - \boldsymbol{\beta}^T\mathbf{x}_i)$, and $\pi(r, G_r(\mathbf{z})) = P(R_i = r|G_r(\mathbf{z}) = g_r)$. To improve efficiency, we adopt the strategy of Robins et al. (1994) and propose the estimating function

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{S}(\boldsymbol{\beta}) - \sum_{i=1}^{n}\sum_{r}\left[\frac{I(R_i = r) - \widetilde{\lambda}_r\{G_r(\mathbf{Z}_i), \boldsymbol{\eta}\}I(R_i \geq r)}{\widetilde{\pi}\{r, G_r(\mathbf{Z}_i), \boldsymbol{\eta}\}}\right]\widetilde{E}\{s_i(\boldsymbol{\beta})|G_r(\mathbf{Z}_i)\},$$

where $\widetilde{\lambda}_r\{G_r(\mathbf{Z}_i), \boldsymbol{\eta}\} = \{1 + \exp[-\mu_r\{G_r(\mathbf{Z}_i), \boldsymbol{\eta}\}]\}^{-1}$, $\mu_r\{G_r(\mathbf{Z}_i), \boldsymbol{\eta}\}$ is a linear predictor based on $G_r(\mathbf{Z}_i)$ and $\boldsymbol{\eta}$, $\widetilde{\pi}\{r, G_r(\mathbf{Z}_i), \boldsymbol{\eta}\} = \prod_{m=1}^{r}\widetilde{\lambda}_m\{G_m(\mathbf{Z}_i), \boldsymbol{\eta}\}$, and $\widetilde{E}\{s_i(\boldsymbol{\beta})|G_r(\mathbf{Z}_i)\}$ is the conditional expectation of $s_i(\boldsymbol{\beta})$ given $G_r(\mathbf{Z}_i)$ under a posited parametric submodel for the full-data generating process.

3 Asymptotic results

Fan and Li (2001) showed that the penalized least-squares estimator minimizing (2), or more generally the penalized maximum likelihood estimator, with the SCAD or hard thresholding penalty behaves asymptotically as if the true model is known *a priori* — the so-called *oracle* property. We show that this property holds for a very broad class of penalized estimating functions, of which the Buckley-James and weighted estimating functions with the SCAD and hard thresholding penalty functions are special cases.

Let $\boldsymbol{\beta}_0 \equiv (\beta_{01}, \ldots, \beta_{0d})^T$ denote the true value of $\boldsymbol{\beta}$. Without loss of generality, suppose that $\beta_{0j} \neq 0$ for $j \leq s$ and $\beta_{0j} = 0$ for $j > s$. We impose the following conditions.

C.1. There exists a nonsingular matrix $\mathbf{A}$ such that for any given constant M,

$$\sup_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0| \leq Mn^{-1/2}} |n^{-1/2}\mathbf{U}(\boldsymbol{\beta}) - n^{-1/2}\mathbf{U}(\boldsymbol{\beta}_0) - n^{1/2}\mathbf{A}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)| = o_p(1).$$

Furthermore, $n^{-1/2}\mathbf{U}(\boldsymbol{\beta}_0) \rightarrow_d N(\mathbf{0}, \mathbf{V})$, for $\mathbf{V}$ a $d \times d$ matrix.

C.2. The penalty function $q_{\lambda_n}(\cdot)$ possesses the following properties:

(i) For non-zero fixed θ, $\lim n^{1/2} q_{\lambda_n}(|\theta|) = 0$ and $\lim q'_{\lambda_n}(|\theta|) = 0$;

(ii) For any $M > 0$, $\lim \sqrt{n} \inf_{|\theta| \leq Mn^{-1/2}} q_{\lambda_n}(|\theta|) \rightarrow \infty$.

Remark 1. Condition C.1 is not unusual and is satisfied by many commonly used estimating functions. This condition is implied by standard conditions for Z-estimators (van der Vaart and Wellner 1996, Thm 3.3.).

Remark 2. Condition C.2 pertains to the choices of the penalty function and regularization parameter. This condition is key to obtaining the oracle property. In particular, condition C.2(i) prevents the j-th element of the penalized estimating function from being dominated by the penalty term, $q_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j)$, for $\beta_{j0} \neq 0$, because $\sqrt{n} q_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j)$ vanishes. However, if $\beta_{j0} = 0$, condition C.2(ii) implies that $\sqrt{n} q_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j)$ diverges to $+\infty$ or $-\infty$ depending on the sign of β_j in the small neighborhood of β_{j0}. Hence, the

j-element of the penalized estimating function is dominated by the penalty term so any consistent solution, say $\widehat{\boldsymbol{\beta}}$, to the estimating equation $\mathbf{U}^P(\boldsymbol{\beta}) = \mathbf{0}$ must satisfy $\widehat{\beta}_j = 0$.

Remark 3. Condition C.2 is satisfied by several commonly used penalties with proper choices of the regularization parameter λ_n.

(a) Under the hard penalty, i.e., $q_{\lambda_n}(|\theta|) = 2(\lambda_n - |\theta|)I(|\theta| < \lambda_n)$, it is straightforward to verify that condition C.2 holds if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$.

(b) Under the SCAD penalty, i.e.,

$$q_{\lambda_n}(|\theta|) = \lambda_n \left\{ I(|\theta| < \lambda_n) + \frac{(a\lambda_n - |\theta|)_+}{(a-1)\lambda_n} I(|\theta| \geq \lambda_n) \right\}$$

with $a > 2$, it is easy to see that if we choose $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, condition C.2 holds because $\sqrt{n}q_{\lambda_n}(|\theta|) = q'_{\lambda_n}(|\theta|) = 0$ for $\theta \neq 0$ and $\sqrt{n}\inf_{|\theta| \leq Mn^{-1/2}} q_{\lambda_n}(|\theta|) = \sqrt{n}\lambda_n$.

(c) For the ALASSO penalty, we assume $\sqrt{n}\lambda_n \to 0$, $n\lambda_n \to \infty$ and $q_{\lambda_n}(|\theta|) = \lambda_n \hat{w}$, for some data-dependent weight $\hat{w}$. First, $n^{1/2}q_{\lambda_n}(|\theta|) = n^{1/2}\lambda_n\hat{w} \to 0$ and $q'_{\lambda_n}(|\theta|) = 0$ for $|\hat{w}| < \infty$ and $\theta \neq 0$. Second, to obtain sparsity, we require that the weights are sufficiently large for θ sufficiently small, say $|\theta| < Mn^{-1/2}$. For simplicity, suppose the data-dependent weights are defined $\hat{w} = |\tilde{\theta}|^{-\gamma}$, for some $\gamma > 0$ and $\tilde{\theta}$ pertaining to the solutions to the unpenalized estimating equations. Then, trivally $\sqrt{n}(\tilde{\theta} - \theta_0) = O_p(1)$, which implies $\sqrt{n}\inf_{|\theta| \leq Mn^{-1/2}} \lambda_n\hat{w} = Mn\lambda_n \to \infty$, as desired. In this paper, we chose $\gamma = 1$ but Zou (2006) notes that other weights may be useful; see Zou (2006, Remarks 1-2) for additional comments on the weights.

(d) When $q_{\lambda_n}(|\theta|) = \lambda_n/|\theta|$, condition C.2 is satisfied if $\sqrt{n}\lambda_n \to 0$ and $n\lambda_n \to \infty$. To see this, note that $\sqrt{n}q_{\lambda_n}(|\theta|) = \sqrt{n}\lambda_n/|\theta| \to 0$, $q'_{\lambda_n}(|\theta|) = -\lambda_n/|\theta|^2 \to 0$, for $\theta \neq 0$, and $\sqrt{n}\inf_{|\theta| \leq Mn^{-1/2}} \lambda_n/|\theta| = Mn\lambda_n \to \infty$. An anonymous referee pointed out that $q_{\lambda_n}(|\theta|) = \lambda_n/|\theta|$ pertains to $p_{\lambda_n}(|\theta|) = \lambda_n \log(|\theta|)$ on the original scale.

(e) Condition C.2 does not hold for the LASSO and EN penalty functions.

To accommodate discrete estimating functions such as (3), we provide a formal definition of the solution to the penalized estimating equation. An estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_d)^T$ is called a zero-crossing to the penalized estimating equation if, for $j = 1, \ldots, d$,

$$\overline{\lim}_{\epsilon \to 0+} n^{-1} U_j^P(\widehat{\boldsymbol{\beta}} + \epsilon \mathbf{e}_j) U_j^P(\widehat{\boldsymbol{\beta}} - \epsilon \mathbf{e}_j) \leq 0,$$

where $\mathbf{e}_j$ is the jth canonical unit vector. Also, an estimator $\widehat{\boldsymbol{\beta}}$ is called an approximate zero-crossing if

$$\overline{\lim}_{n \to \infty} \overline{\lim}_{\epsilon \to 0+} n^{-1} U_j^P(\widehat{\boldsymbol{\beta}} + \epsilon \mathbf{e}_j) U_j^P(\widehat{\boldsymbol{\beta}} - \epsilon \mathbf{e}_j) \leq 0.$$

If $\mathbf{U}^P$ is continuous, then the zero-crossing is an exact solution to the penalized estimating equation.

The following theorem states the main theoretical results regarding the proposed penalized estimators, including the existence of a root-n consistent estimator, the sparsity of the estimator and the asymptotic normality of the estimator.

Theorem 1 *Define the number of non-zero coefficients $s = \#\{j | \beta_{j0} \neq 0\}$. Under conditions C.1 and C.2, the following results hold:*
(a) There exists a root-n consistent approximate zero-crossing of $\mathbf{U}^P(\boldsymbol{\beta})$, i.e, $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + O_p(n^{-1/2})$ such that $\widehat{\boldsymbol{\beta}}$ is an approximate zero-crossing of $\mathbf{U}^P(\boldsymbol{\beta})$.
(b) For any root-n consistent approximate zero-crossing of $\mathbf{U}^P(\boldsymbol{\beta})$, denoted by $\widehat{\boldsymbol{\beta}} \equiv (\widehat{\beta}_1, \ldots, \widehat{\beta}_d)^T$, $\lim_n P(\widehat{\beta}_j = 0 \text{ for } j > s) = 1$. Moreover, if we denote $\widehat{\boldsymbol{\beta}}_1 = (\widehat{\beta}_1, \ldots, \widehat{\beta}_s)^T$ and $\boldsymbol{\beta}_{01} = (\beta_{01}, \ldots, \beta_{0s})^T$, then

$$n^{1/2}(\mathbf{A}_{11} + \boldsymbol{\Sigma}_{11}) \left\{ \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01} + (\mathbf{A}_{11} + \boldsymbol{\Sigma}_{11})^{-1} \mathbf{b}_n \right\} \to_d N(0, \mathbf{V}_{11}),$$

where $\mathbf{A}_{11}, \boldsymbol{\Sigma}_{11}$ and $\mathbf{V}_{11}$ are the first $s \times s$ sub-matrices of $\mathbf{A}$, $diag\left\{-q'_{\lambda_n}(|\boldsymbol{\beta}_0|) sgn(\boldsymbol{\beta}_0)\right\}$ and $\mathbf{V}$, respectively, and $\mathbf{b}_n = -(q_{\lambda_n}(|\beta_{01}|) sgn(\beta_{01}), \ldots, q_{\lambda_n}(|\beta_{0s}|) sgn(\beta_{0s}))^T$.
(c) Let $\mathbf{U}_1^P(\boldsymbol{\beta})$ and $\mathbf{U}_1(\boldsymbol{\beta})$ denote the first s-components of $\mathbf{U}^P(\boldsymbol{\beta})$ and $\mathbf{U}(\boldsymbol{\beta})$, respectively, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_1$ denotes the first s-components of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_2$ denote the second

$(d-s)$-components of $\boldsymbol{\beta}$. That is, without loss of generality, $\boldsymbol{\beta}_2 = \mathbf{0}$. If $\mathbf{U}_1((\boldsymbol{\beta}_1^T, \mathbf{0}^T)^T)$ is continuous in $\boldsymbol{\beta}_1$, then there exists $\widehat{\boldsymbol{\beta}}_1$ such that

$$\mathbf{U}_1^P((\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T) = \mathbf{0}.$$

That is, the solution is exact.

The proof of Theorem 1 is relegated to Appendix A. The asymptotic results for penalized weighted estimators follow easily from this theorem. Applying this theorem to the penalized Buckley-James estimators, we obtain the following result.

Corollary 1 *Assume that condition C.2 holds in addition to the following three conditions:*

D.1. There exists a constant c_0 such that $P(\widetilde{Y} - \boldsymbol{\beta}^T\mathbf{x} < c_0) < 1$ for all $\boldsymbol{\beta}$ in some neighborhood of $\boldsymbol{\beta}_0$.

D.2. The random variable $\mathbf{x}$ has compact support.

D.3. F has finite Fisher information for location.

Then the conclusions of Theorem 1 follow.

Remark 4. Corollary 1 implies that the penalized Buckley-James estimators with the penalty functions satisfying condition C.2 possess the oracle property. Conditions D.1-D.3 are the regularity conditions given in Ritov (1990, p. 306) to ensure that condition C.1 holds. The expressions for $\mathbf{A}$ and $\mathbf{V}$ can be found in Ritov (1990) and Lai and Ying (1991a). The matrix $\mathbf{V}$ is directly estimable from the data whereas $\mathbf{A}$ is not because the latter involves the unknown density of the error term ε.

Remark 5. A result similar to Corollary 1 exists for the adaptive estimators presented in Subsection 2.3. Namely, the penalized, weighted estimators with SCAD, hard thresholding, and ALASSO penalties also possess an oracle property. Technical conditions needed to obtain a strongly consistent estimator sequence and hence establish condition C.1 are

given in Robin, Rotnizky, and Zhao (1995, Appendix A.1). Such technical conditions are assumed throughout the Tsiatis (2006) text, for example. The matrices $\mathbf{A}$ and $\mathbf{V}$ may be calculated directly; examples are given in Tsiatis (2006, chs.10-11).

Theorem 1 implies that the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_1$ is

$$\boldsymbol{\Omega}_{11} = n^{-1}(\mathbf{A}_{11} + \boldsymbol{\Sigma}_{11})^{-1}\mathbf{V}_{11}(\mathbf{A}_{11} + \boldsymbol{\Sigma}_{11})^{-1},$$

and a consistent estimator is given by

$$\widehat{\boldsymbol{\Omega}}_{11} = n^{-1}(\widehat{\mathbf{A}}_{11} + \widehat{\boldsymbol{\Sigma}}_{11})^{-1}\widehat{\mathbf{V}}_{11}(\widehat{\mathbf{A}}_{11} + \widehat{\boldsymbol{\Sigma}}_{11})^{-1}.$$

Other authors (e.g. Fu, 2003) use the following alternative estimator for $\mathrm{cov}(\widehat{\boldsymbol{\beta}}_1)$,

$$\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}_1) = \widetilde{\boldsymbol{\Omega}}_{11}, \quad \widetilde{\boldsymbol{\Omega}} = n^{-1}\left[\left(\widehat{\mathbf{A}} + \widehat{\boldsymbol{\Sigma}}\right)^{-1}\widehat{\mathbf{V}}\left(\widehat{\mathbf{A}} + \widehat{\boldsymbol{\Sigma}}\right)^{-1}\right].$$

Using the sandwich matrix $\widetilde{\boldsymbol{\Omega}}$ actually produces a standard error estimate for the entire vector $\widehat{\boldsymbol{\beta}}$, that is, both non-zero and zero coefficient estimates. On the other hand, $\widehat{\boldsymbol{\Omega}}_{11}$ implicitly sets $\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}_2) = 0$, its asymptotic value. In this paper, we use $\widehat{\boldsymbol{\Omega}}_{11}$ which agrees with the following earlier papers on variable selection: Fan and Li (2001, 2002, 2004), Cai et al. (2005) and Zou (2006). Note the matrix $\widehat{\boldsymbol{\Omega}}_{11}$ can be readily calculated when $\mathbf{A}$ and $\mathbf{V}$ can be directly evaluated. For discrete estimating functions such as the Buckley-James estimating function, $\mathbf{A}$ cannot be estimated reliably from the data. To solve this problem, we propose a resampling procedure.

Let $\mathbf{U}_1^P(\boldsymbol{\beta})$ denote the components of $\mathbf{U}^P(\boldsymbol{\beta})$ corresponding to the regression coefficients whose penalized estimating function estimates are nonzero and define $\widehat{\boldsymbol{\beta}}_1^*$ as the solution to the following estimating equation

$$\mathbf{U}_1^P(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{W}_{1i} G_i, \tag{4}$$

where $(G_1, \ldots, G_n)$ are independent standard normal variables, and $(\mathbf{W}_{11}, \ldots, \mathbf{W}_{1n})$ are given in Appendix B. We show in Appendix B that the conditional distribution of $n^{1/2}(\widehat{\boldsymbol{\beta}}_1^* -$

$\widehat{\boldsymbol{\beta}}_1$) given the observed data is the same in the limit as the unconditional distribution of $n^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01})$. Thus, we may estimate the covariance matrix of $\widehat{\boldsymbol{\beta}}_1$ and construct confidence intervals for individual regression coefficients by using the empirical distribution of $\widehat{\boldsymbol{\beta}}_1^*$.

4 Implementation

In this paper, we use a majorize-minorize (MM) algorithm to estimate the penalized regression coefficients (Hunter and Li, 2005). The MM algorithm may be viewed as a Fisher-scoring (or Newton-Raphson) type algorithm for solving a perturbed, penalized estimating equation and is closely related to the local quadratic algorithm (Tibshirani, 1996; Fan and Li, 2001). By using condition C.1 and the local quadratic approximations for penalty functions (Fan and Li, 2001, Section 3.3), the MM algorithm is:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} + \left\{ \mathbf{A}(\widehat{\boldsymbol{\beta}}^{(k)}) + \boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}^{(k)}) \right\}^{-1} \mathbf{U}^P(\widehat{\boldsymbol{\beta}}^{(k)}), \quad k \geq 0,$$

where $\widehat{\boldsymbol{\beta}}^{(0)}$ is the solution to $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$, and

$$\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}) = \text{diag}\left\{ q_\lambda(|\beta_1|)/(\epsilon + |\beta_1|), \ldots, q_\lambda(|\beta_d|)/(\epsilon + |\beta_d|) \right\},$$

for ϵ a small number ($\epsilon = 10^{-6}$ in our examples). This algorithm requires that the estimating function $\mathbf{U}(\boldsymbol{\beta})$ is continuous so that the asymptotic slope matrix $\mathbf{A}$ can be evaluated directly, as in the missing data example. For general estimating functions, we propose the iterative algorithm:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{U}(\boldsymbol{\beta}) - n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}^{(k)})\boldsymbol{\beta}\|, \quad k \geq 0,$$

where $\widehat{\boldsymbol{\beta}}^{(0)}$ is a minimizer of $\|\mathbf{U}(\boldsymbol{\beta})\|$. For the penalized Buckley-James estimator, there is a simple iterative algorithm:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \left\{ \mathbf{X}^T\mathbf{X} + n\boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}^{(k)}) \right\}^{-1} \mathbf{X}^T\boldsymbol{\xi}(\widehat{\boldsymbol{\beta}}^{(k)}), \quad k \geq 0,$$

where $\widehat{\boldsymbol{\beta}}^{(0)}$ is the original Buckley-James estimator, and $\boldsymbol{\xi}(\boldsymbol{\beta}) = [\xi_1(\boldsymbol{\beta}), \ldots, \xi_n(\boldsymbol{\beta})]^T$. In each algorithm, we iterate until convergence; the final solution is an approximate solution

to the penalized estimating equation $\mathbf{U}^P(\boldsymbol{\beta}) = \mathbf{0}$. To improve numerical stability, we standardize each predictor to have mean 0 and variance 1.

We need to choose λ for LASSO, ALASSO, and hard thresholding penalty functions, (a, λ) for the SCAD penalty, and (λ_1, λ_2) for the EN penalty. Fan and Li (2001, 2002) showed that the choice of $a \equiv 3.7$ performs well in a variety of situations; we use their suggestion throughout our numerical analyses. Zou and Hastie (2005) show that the EN estimator is equivalent to an ℓ_1-penalty on augmented data. In the rest of this section, we include the subscript λ on $\widehat{\boldsymbol{\beta}}$, i.e. $\widehat{\boldsymbol{\beta}}_\lambda$, to stress the dependence of the estimator on the regularization parameter λ. In the case of EN penalty, it is understood that cross-validation is two-dimensional.

For uncensored data, Tibshirani (1996) and Fan and Li (2001) suggested the following generalized cross-validation statistic (Wahba, 1985):

$$\mathrm{GCV}^\dagger(\lambda) = \frac{\mathrm{RSS}(\lambda)/n}{\{1 - d(\lambda)/n\}^2},$$

where $\mathrm{RSS}(\lambda)$ is the residual sum of squares $\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda\|^2$, and $d(\lambda)$ is the effective number of parameters, i.e., $d(\lambda) = \mathrm{tr}[\{\widehat{\mathbf{A}} + \boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}_\lambda)\}^{-1}\widehat{\mathbf{A}}^T]$. Note that the intercept is omitted in $\mathrm{RSS}(\lambda)$ since $\mathbf{y}$ may be centered at $n^{-1}\sum_{i=1}^n Y_i$. When the Y_i are potentially censored, $d(\lambda)$ may still be regarded as the effective number of parameters; however, $\mathrm{RSS}(\lambda)$ is unknown. We propose to estimate $n^{-1}\mathrm{RSS}(\lambda)$ by

$$\widehat{\nu}(\lambda) = \frac{\sum_{i=1}^n \Delta_i (Y_i - \widehat{\alpha} - \widehat{\boldsymbol{\beta}}_\lambda^T \mathbf{x}_i)^2 / \widehat{K}(Y_i)}{\sum_{i=1}^n \Delta_i / \widehat{K}(Y_i)},$$

where $\widehat{K}(t)$ is the Kaplan-Meier estimator for $K(t) = P(C > t)$, and $\widehat{\alpha} = n^{-1}\sum_{i=1}^n \xi_i(\widehat{\boldsymbol{\beta}}^{(0)})$. For missing data, we propose to estimate $n^{-1}\mathrm{RSS}(\lambda)$ by

$$\widehat{\nu}(\lambda) = \frac{\sum_{i=1}^n I(R_i = \infty)(Y_i - \widehat{\boldsymbol{\beta}}_\lambda^T \mathbf{x}_i)^2 / \widetilde{\pi}(\infty, \mathbf{Z}_i, \widehat{\boldsymbol{\eta}})}{\sum_{i=1}^n I(R_i = \infty) / \widetilde{\pi}(\infty, \mathbf{Z}_i, \widehat{\boldsymbol{\eta}})}.$$

Both proposals are based on large-sample arguments. Namely, $\hat{\nu}(\lambda)$ is a consistent estimator for $\lim n^{-1}\mathrm{RSS}(\lambda)$ for fixed λ under conditional independence between censoring and

failure time distribution, for censored outcome data, and under the MAR assumption for missing data (cf. Tsiatis, 2006, ch. 6). Thus, our generalized cross-validation statistic is

$$\mathrm{GCV}(\lambda) = \frac{\widehat{\nu}(\lambda)}{\{1 - d(\lambda)/n\}^2},$$

and we select $\widehat{\lambda} = \arg\min_\lambda \mathrm{GCV}(\lambda)$.

5 Simulation studies

5.1 Censored data

We simulated 1000 data sets of size n from the model

$$Y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \sigma\varepsilon_i,\ i = 1, \ldots, n,$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, ε_i and $\mathbf{x}_i$ are independent standard normal with the correlation between the jth and kth components of $\mathbf{x}$ equal to $0.5^{|j-k|}$. This model was considered by Tibshirani (1996) and Fan and Li (2001). We set the censoring distribution to be uniform$(0, \tau)$, where τ was chosen to yield approximately 30% censoring. We compared the model error $\mathrm{ME} \equiv (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\mathbf{x}\mathbf{x}^T)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ of the proposed penalized estimator to that of the original Buckley-James estimator using the median relative model error (MRME). We also compared the average numbers of regression coefficients that are correctly or incorrectly shrunk to 0. The results are presented in Table 1, where *oracle* pertains to the situation in which we know *a priori* which coefficients are non-zero.

The performance of the proposed estimator with the SCAD, hard thresholding, and ALASSO penalties approach that of the oracle estimator as n increases. When the signal-to-noise ratio is small (e.g. large n or small σ), oracle methods (SCAD, hard thresholding, ALASSO) outperform LASSO and EN in terms of model error and model complexity. On the other hand, LASSO and EN tend to perform better than the oracle methods as σ/n increases.

Table 1: Simulation results on model selection with censored data where table entries are median relative model error (MRME) and the average number of correct and incorrect zeros, C and I respectively.

		Avg. No. of 0s	
Method	MRME(%)	C	I
$n = 50, \sigma = 3$			
SCAD	69.48	4.73	0.35
Hard	73.41	4.30	0.17
LASSO	66.16	3.99	0.11
ALASSO	57.77	4.40	0.17
EN	76.48	3.54	0.08
Oracle	32.76	5	0
$n = 50, \sigma = 1$			
SCAD	40.11	4.78	0.01
Hard	69.79	4.18	0.01
LASSO	64.48	3.97	0.01
ALASSO	48.21	4.90	0.01
EN	95.55	3.49	0
Oracle	31.30	5	0

Table 2: Simulation results on standard error estimation for the non-zero coefficients $(\beta_1, \beta_2, \beta_5)$ in least squares regression with censored data. SD refers to the mean absolute deviation of the estimated regression coefficients divided by 0.6745 while SD_m refers to the median of the standard error estimates. Table entries are for a sample size $n = 100$ and (error) standard deviation $\sigma = 1$.

	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_5$	
	SD	SD_m	SD	SD_m	SD	SD_m
SCAD	0.145	0.129	0.135	0.128	0.128	0.114
Hard	0.151	0.130	0.145	0.129	0.138	0.119
LASSO	0.160	0.134	0.145	0.143	0.161	0.130
ALASSO	0.149	0.132	0.130	0.133	0.133	0.113
EN	0.172	0.113	0.151	0.111	0.155	0.103
Oracle	0.144	0.129	0.136	0.126	0.143	0.111

Table 2 reports the results on the accuracy of the proposed resampling technique in estimating the variances of the non-zero estimated regression coefficients. The standard deviation (SD) pertains to the median absolute deviation of the estimated regression coefficients divided by 0.6745. The median of the standard error estimates, denoted by SD_m, gauges the performance of the resampling procedure. Evidently, the resampling procedure yields reasonable standard error estimates, particularly for large n.

5.2 Missing data

We simulated 1000 datasets of size n from the model

$$Y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \sigma\varepsilon_i,\ i = 1, \ldots, n,$$

where ε_i and $\mathbf{x}_i$ are independent standard normal with the correlation between the jth and kth components of $\mathbf{x}$ equal to $0.5^{|j-k|}$. We considered two scenarios:

$$\text{Model 1:} \quad \boldsymbol{\beta} = (0.25, 0.5, 0, 0, 0.75, 1.5, 0.75, 0, 0, 1)^T,$$
$$\text{Model 2:} \quad \boldsymbol{\beta} = (0, 1.25, 0, 0, 0, 2, 0, 0, 0, 1.5)^T.$$

For a random design $\mathbf{X}$, define the theoretical R^2

$$R^2 = \frac{\boldsymbol{\beta}_0^T E(\mathbf{x}\mathbf{x}^T)\boldsymbol{\beta}_0}{\boldsymbol{\beta}_0^T E(\mathbf{x}\mathbf{x}^T)\boldsymbol{\beta}_0 + \sigma^2}.$$

For $\sigma = 1$ and 2, both Models 1-2 have theoretical R^2=0.89 and 0.67, respectively. Although Models 1-2 have the same theoretical R^2, they have differing numbers of non-zero coefficients; the number of non-zero coefficients over the total number of coefficients (i.e. $d = 10$) in a given model is sometimes referred to as the model fraction. The model fraction in Model 1 is 0.6 while Model 2 has a model fraction of 0.3. We simulated data such that subjects fall into one of three categories: $R = 1$ means that the subject is missing (x_1, x_2); $R = 2$ means that the subject is missing x_1; and $R = \infty$ means that the subject has complete data. The observed data $\{R, G_R(\mathbf{Z})\}$ were generated in the following sequence of steps:

1. Simulate a Bernoulli random variable B_1 with probability $\widetilde{\lambda}_1\{G_1(\mathbf{Z}_i), \boldsymbol{\eta}\}$.

2. If $B_1 = 1$, set $R = 1$; else

3. Simulate a Bernoulli random variable B_2 with probability $\widetilde{\lambda}_2\{G_2(\mathbf{Z}_i), \boldsymbol{\eta}\}$.

4. If $B_2 = 1$, set $R = 2$; else set $R = \infty$.

The missingness process was formulated by logistic models

$$\begin{aligned} \text{logit } \widetilde{\lambda}_1\{G_1(\mathbf{Z}_i)\} &= \eta_{10} + \eta_{11} Y_i + \sum_{j=3}^{10} \eta_{1j} x_{ij}, \\ \text{logit } \widetilde{\lambda}_2\{G_2(\mathbf{Z}_i)\} &= \eta_{20} + \eta_{21} Y_i + \sum_{j=2}^{10} \eta_{1j} x_{ij}, \end{aligned}$$

where

$$\boldsymbol{\eta}_1 = (-6, 0.75, 0, 0, 1.25, 1.5, 1.25, 0, 0, 1.25)^T,$$

$$\boldsymbol{\eta}_2 = (-1.5, 0.5, 1.5, 0, 0, 0.5, 0.5, 0.5, 0, 0, 0.5)^T.$$

These models yielded approximately 40% missing with subjects falling in the $R = 1, 2$ categories in roughly equal proportions.

We present the numerical results with $n = 250$ in Table 3. Oracle methods (SCAD, hard thresholding, ALASSO) perform better than LASSO and EN in terms of relative model error and complexity when there are a few strong predictors of response, as in Model 1. However, oracle methods perform worse than LASSO and EN when there are many weakly significant predictors, as in Model 2.

6 The Paul Coverdell Stroke Registry

The Paul Coverdell National Acute Stroke Registry collects demographic, quantitative and qualitative factors related to acute stroke care in four prototype states: Georgia, Massachusetts, Michigan, and Ohio (cf. The Paul Coverdell Prototype Registries Writing

Table 3: Simulation results on model selection with missing data where MRME is the median relative model error, C and I report the average number of correct and incorrect zeros, respectively. For $\sigma = 1$ and $\sigma = 2$, Models 1-2 have theoretical $R^2 = 0.89$ and 0.67, respectively; however, the number of non-zero coefficients is six in Model 1 while only three in Model 2.

	Model 1			Model 2		
		Avg. No. of 0s			Avg. No. of 0s	
Method	MRME(%)	C	I	MRME(%)	C	I
$\sigma = 1$						
SCAD	81.79	3.35	0.21	42.60	5.56	0
Hard	82.38	3.37	0.25	48.73	5.79	0.01
LASSO	87.88	2.42	0.09	66.49	4.11	0
ALASSO	82.24	3.55	0.23	37.74	6.25	0
EN	85.59	2.38	0.08	70.56	3.92	0
$\sigma = 2$						
SCAD	93.64	3.33	0.69	48.73	5.92	0.02
Hard	90.10	3.70	1.12	46.24	6.37	0.05
LASSO	82.29	2.54	0.40	59.96	4.56	0.02
ALASSO	82.01	3.37	0.70	48.87	6.08	0.02
EN	88.62	2.55	0.44	66.17	4.63	0.03

Group, 2005). The goals of the registry include a better understanding of factors associated with stroke and a general improvement in the quality of acute stroke care in the United States. For purposes of illustration, we consider a subset of 800 patients with hemorragic or ischemic stroke from the Georgia prototype registry. Our data set includes nine predictors and a hospital length-of-stay (LOS) endpoint, defined as the number of days from hospital admission to hospital discharge. Conclusions from analyses like ours would be important to investigators in health policy and management, for example. The complete registry data for all four prototypes consists of several thousand hospital admissions but has not been released publicly. A more comprehensive analysis is ongoing.

Our data include the following nine predictors: Glasgow Coma Scale (GCS; 3-15, with 15 representing excellent health), serum albumin, creatinine, glucose, age, sex (1 if male), race (1 if white), whether the patient was admitted to the intensive care unit (ICU; 1 if yes), and stroke subtype (1 if hemorrhagic, 0 if ischemic). Of 800 patients, 419 (52.4%) have complete data (i.e. $R = \infty$). A total of 94 (11.8%) patients are missing both GCS and serumn albumin (i.e. $R = 1$) and 287 (35.9%) patients are missing GCS only (i.e. $R = 2$).

Estimates for the nuisance parameter η in the stroke data are presented in Table 4. We find that in subjects missing both GCS and albumin (i.e. $R = 1$) tend to have higher creatinine and glucose levels but less likely to be admitted to the ICU upon admission to the hospital. Ischemic stroke and ICU admission were strongly associated with missing GCS score (i.e. $R = 2$) only. Because the missingness mechanism is related to other important prognostic variables, this is mild evidence that the MCAR (missing completely at random) assumption is not well-supported and variable selection techniques based on such an assumption will lead to to incorrect conclusions. Our analyses using methods described in Section 2 assuming MAR (missing at random) are displayed in Table 5.

We use $\hat{\lambda} = (0.28, 0.63, 0.11, 0.16)$ for the SCAD, Hard, LASSO, and ALASSO estimates, respectively, and use $(\hat{\lambda}_1, \hat{\lambda}_2) = (0.34, 0.9)$ for the EN estimates. Table 5 presents

Table 4: Estimates of η in the stroke data, where η pertains to the parameters in the coarsening models $\widetilde{\lambda}_1\{G_1(\mathbf{Z})\}$ and $\widetilde{\lambda}_2\{G_2(\mathbf{Z})\}$.

	η_1		η_2	
(int)	-2.342	(0.152)	0.478	(0.082)
Albumin	–	–	-0.112	(0.089)
Creatinine	-0.492	(0.291)	-0.101	(0.091)
Sex	0.172	(0.113)	0.043	(0.079)
Glucose	-0.286	(0.164)	-0.067	(0.084)
ICU	-0.470	(0.155)	-0.304	(0.091)
Age	0.045	(0.124)	0.006	(0.087)
Type	-0.101	(0.144)	-0.213	(0.094)
Race	0.084	(0.122)	-0.034	(0.085)
LOS	-0.007	(0.140)	-0.045	(0.092)

Table 5: Estimated regression coefficients and their standard errors in the stroke data.

	Full	SCAD	Hard	LASSO	ALASSO	EN
GCS	-0.762 (0.327)	-0.603 (0.434)	-0.864 (0.587)	-0.681 (0.480)	-0.584 (0.400)	-0.628 (0.424)
Albumin	-1.142 (0.306)	-0.958 (0.450)	-1.043 (0.486)	-0.984 (0.425)	-0.876 (0.402)	-0.882 (0.387)
Creatinine	-0.726 (0.331)	-0.372 (0.177)	-0.734 (0.347)	-0.529 (0.255)	-0.365 (0.179)	-0.402 (0.199)
Sex	-0.007 (0.288)	0 (-)	0 (-)	0 (-)	0 (-)	0 (-)
Glucose	-0.312 (0.310)	0 (-)	0 (-)	-0.140 (0.165)	0 (-)	-0.030 (0.039)
ICU	1.861 (0.323)	2.043 (0.442)	1.970 (0.469)	1.807 (0.419)	1.947 (0.415)	1.771 (0.392)
Age	-0.696 (0.324)	-0.293 (0.203)	-0.678 (0.465)	-0.586 (0.369)	-0.405 (0.260)	-0.516 (0.312)
Type	0.553 (0.333)	0.200 (0.155)	0 (-)	0.448 (0.335)	0.213 (0.158)	0.381 (0.273)
Race	-1.316 (0.315)	-1.403 (0.374)	-1.320 (0.366)	-1.216 (0.331)	-1.242 (0.332)	-1.151 (0.310)

the regression coefficient estimates for the stroke data. Higher levels of albumin and creatine are strongly related to shorter hospital stays while patients admitted to the ICU are associated with longer hospital stays. Older patients tend to have shorter stays in the hospital than younger patients; this is most easily explained by the fact that those older stroke patients die in the hospital quickly because their bodies are too weak to recover. Patients with hemorrhagic strokes have longer recovery periods and thus stay at the hospital for a longer duration. White stroke patients tend to have shorter hospital stays than non-whites. Finally, sex and glucose are weak predictors of hospital stays. The LASSO and EN estimates tend to retain more predictors in the final model and, hence, have more complex models than compared to the other penalized estimators. Among the SCAD, Hard, and ALASSO estimates, SCAD and ALASSO yield similar coefficient estimates while the Hard thresholding estimates yield the sparsest model. Our methods yield models that appear to have reasonable scientific interpretation and do not make a strong MCAR assumption, an assumption that is not supported by the data.

7 Remarks

We have developed a general methodology for selecting variables and simultaneously estimating their regression coefficients in semiparametric models. This development overcame two major challenges that are not present with any of the existing variable selection methods. First, $\mathbf{U}^P(\boldsymbol{\beta})$ may not correspond to the derivative of an objective function nor quasi-likelihood, so that the mathematical arguments employed by previous authors to establish the asymptotic properties of penalized maximum likelihood or penalized GEE estimators do not apply. Second, $\mathbf{U}^P(\boldsymbol{\beta})$ may be discrete in $\boldsymbol{\beta}$, which entails considerable theoretical and computational challenges. In particular, the variances of the estimated regression coefficients cannot be evaluated directly and we have developed a novel resampling procedure, which can also be used for variance estimation without the task of variable selection. Our simulation results indicate that the resampling method works well for modest sample sizes.

Rank estimators (Prentice, 1978; Tsiatis, 1990; Wei et al., 1990; Lai and Ying, 1991b; Ying, 1993) provide potential alternatives to the Buckley-James estimator, but are computationally more demanding to implement. In general, the rank estimating functions do not correspond to the derivatives of any objective functions. This is also true of estimating functions for many other semiparametric problems. In all those situations, we can use Theorem 1 to establish the asymptotic properties of the corresponding variable selection procedures and use the proposed resampling technique to estimate the variances of the selected variables.

The proportional hazards and accelerated failure time models cannot hold simultaneously unless the error distribution is extreme-value. Thus, it is useful to have variable selection methods for both models at one's disposal since one model may fit the data better than the other. A major advantage of model (1) is that the regression coefficients have direct physical interpretation. Hazard ratio can be an awkward concept, especially when the response variable does not pertain to failure time.

Appendix A. Proof of Theorem 1

To prove part (a), we consider $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T$, where $\widehat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_{01} + n^{-1}\mathbf{A}_{11}^{-1}\mathbf{U}_1(\boldsymbol{\beta}_0)$. Because $n^{1/2}q_{\lambda_n}(|\beta_{0j}|) \to 0$, $j = 1, \ldots, s$, under condition C.2 (i) and $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + O_p(n^{-1/2})$, we have

$$n^{-1/2}U_j^P(\widehat{\boldsymbol{\beta}} \pm \epsilon \mathbf{e}_j) = o_p(1) - n^{1/2}q_{\lambda_n}(|\widehat{\beta}_j \pm \epsilon|) = o_p(1).$$

Under Condition C.2 (ii), for $j = s+1, \ldots, d$, $n^{-1/2}U_j^P(\widehat{\boldsymbol{\beta}} + \epsilon \mathbf{e}_j)$ and $n^{-1/2}U_j^P(\widehat{\boldsymbol{\beta}} - \epsilon \mathbf{e}_j)$ are dominated by $-n^{1/2}q_{\lambda_n}(\epsilon)$ and $n^{1/2}q_{\lambda_n}(\epsilon)$, respectively, so they have opposite signs when ϵ goes to zero. Therefore, $\widehat{\boldsymbol{\beta}}$ is an approximate zero-crossing by definition.

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

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

Using the jth component of the penalized estimating function and the definition of the approximate zero-crossing, we obtain that on the set of $\{\widehat{\beta}_j \neq 0, |\widehat{\beta}_j| < Mn^{-1/2}\}$,

$$o_p(1) = \left\{n^{-1/2}U_j(\boldsymbol{\beta}_0) + n^{1/2}\mathbf{A}_j(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(1) - n^{1/2}q_{\lambda_n}(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j)\right\}^2,$$

where $\mathbf{A}_j$ is the jth row of $\mathbf{A}$. 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,

$$P(\widehat{\beta}_j \neq 0, |\widehat{\beta}_j| < Mn^{-1/2}, n^{1/2}q_{\lambda_n}(|\widehat{\beta}_j|) > M') < \epsilon/2.$$

Since $\lim_n \sqrt{n}\inf_{|\theta| \leq Mn^{-1/2}} q_{\lambda_n}(|\theta|) \to \infty$ by condition C.2 (ii), $\widehat{\beta}_j \neq 0$ and $|\widehat{\beta}_j| < Mn^{-1/2}$ imply that $n^{1/2}q_{\lambda_n}(|\widehat{\beta}_j|) > M'$ for large n. Thus, $P(\widehat{\beta}_j \neq 0, |\widehat{\beta}_j| < Mn^{-1/2}) = P(\widehat{\beta}_j \neq 0, |\widehat{\beta}_j| < Mn^{-1/2}, n^{1/2}q_{\lambda_n}(|\widehat{\beta}_j|) > M')$. Therefore, $P(C_j) < \epsilon/2 + P(\widehat{\beta}_j \neq 0, |\widehat{\beta}_j| < Mn^{-1/2}, n^{1/2}q_{\lambda_n}(|\widehat{\beta}_j|) > M') < \epsilon$.

To prove the second part of (b), since

$$o_p(1) = n^{-1/2}\mathbf{U}_1(\boldsymbol{\beta}_0) + n^{1/2}\mathbf{A}_{11}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) - n^{1/2}q_{\lambda_n}(|\widehat{\boldsymbol{\beta}}_1|)\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_1),$$

after the Taylor series expansion of the last term, we conclude that

$$n^{1/2}\left\{(\mathbf{A}_{11}+\mathbf{\Sigma}_{11})(\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01}+(\mathbf{A}_{11}+\mathbf{\Sigma}_{11})^{-1}\mathbf{b}_n\right\}=-n^{-1/2}\begin{pmatrix}U_1(\boldsymbol{\beta}_0)\\ \vdots \\ U_s(\boldsymbol{\beta}_0)\end{pmatrix}+o_p(1)\rightarrow_d N(0,\mathbf{V}_{11}).$$

To prove part (c), we consider $\boldsymbol{\beta}_1 \in R^s$ on the boundary of a ball around $\boldsymbol{\beta}_{01}$, i.e., $\boldsymbol{\beta}_1=\boldsymbol{\beta}_{01}+n^{-1/2}\mathbf{u}$ with $|\mathbf{u}|=r$ for a fixed constant r. From the penalized estimating function $\mathbf{U}_1^P$, we have

$$\begin{aligned}
& n^{-1/2}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})^T\mathbf{A}_{11}^T\mathbf{U}_1^P(\boldsymbol{\beta}) \\
= \;& (\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})^T\mathbf{A}_{11}^T\left\{n^{-1/2}\mathbf{U}_1(\boldsymbol{\beta})-n^{1/2}q_{\lambda_n}(|\boldsymbol{\beta}_1|)\mathrm{sgn}(\boldsymbol{\beta}_1)\right\} \\
= \;& O_p(|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}|)+n^{1/2}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})^T\mathbf{A}_{11}^T\mathbf{A}_{11}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}) \\
& -n^{1/2}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})\mathbf{A}_{11}^T\mathrm{diag}\{q'_{\lambda_n}(|\beta_j^*|)\mathrm{sgn}(\beta_{0j})\}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}),
\end{aligned}$$

where β_j^* is between β_j and β_{0j} for $j=1,\ldots,s$. Since $\mathbf{A}_{11}$ 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 $\mathbf{A}_{11}^T\mathbf{A}_{11}$. The first term is of order $rO_p(n^{-1/2})$. Since $\max_j q'_{\lambda_n}(|\beta_j^*|)\rightarrow 0$, 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

$$P\left[\min_{|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}|=n^{-1/2}r}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})^T\mathbf{A}_{11}^T\mathbf{U}_1^P((\boldsymbol{\beta}_1^T,\mathbf{0}^T)^T)>0\right]>1-\epsilon.$$

Applying the Brouwer fixed-point theorem to the continuous function $\mathbf{U}_1^P((\boldsymbol{\beta}_1^T,\mathbf{0}^T)^T)$, we see that $\min_{|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}|=n^{-1/2}r}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})^T\mathbf{A}_{11}^T\mathbf{U}^P((\boldsymbol{\beta}_1^T,\mathbf{0}^T)^T)>0$ implies that $\mathbf{A}_{11}^T\mathbf{U}_1^P((\boldsymbol{\beta}_1^T,\mathbf{0}^T)^T)$ has a solution within this ball, or equivalently, $\mathbf{U}_1^P((\boldsymbol{\beta}_1^T,\mathbf{0}^T)^T)$ has a solution within this ball. That is, we can choose an exact solution $\widehat{\boldsymbol{\beta}}=(\widehat{\boldsymbol{\beta}}_1^T,\mathbf{0}^T)^T$ to $\mathbf{U}_1^P(\boldsymbol{\beta})=\mathbf{0}$ with $\widehat{\boldsymbol{\beta}}=\boldsymbol{\beta}_0+O_p(n^{-1/2})$. Hence, $\widehat{\boldsymbol{\beta}}$ is a zero-crossing of $\mathbf{U}^P(\boldsymbol{\beta})$.

Appendix B: Conditional distribution of $(\widehat{\boldsymbol{\beta}}_1^* - \widehat{\boldsymbol{\beta}}_1)$

Here, we justify the resampling procedure for the penalized Buckley-James estimator. Similar justifications can be made for other estimators. Under conditions D.1-D.3, we have the following asymptotic linear expansion for the penalized Buckley-James estimating function:

$$n^{-1/2}\mathbf{U}_1^P(\boldsymbol{\beta}) = n^{-1/2}\mathbf{U}_1^P(\boldsymbol{\beta}_0)+(\mathbf{A}_{11}+\boldsymbol{\Sigma}_{11})n^{1/2}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01})+o(\max\{1, n^{1/2}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01}\|\}). \tag{A.3}$$

In addition,

$$n^{-1/2}\mathbf{U}_1(\boldsymbol{\beta}_0) = n^{-1/2}\sum_{i=1}^{n}\mathbf{w}_{1i} + o(1),$$

where $\mathbf{w}_{1i}$ consists of the components of $\mathbf{w}_i$ corresponding to $\boldsymbol{\beta}_1$, and $\mathbf{w}_i$, $i = 1, \ldots, n$, as given in Lin and Wei (1992), are n independent zero-mean random vectors. Replacing the unknown quantities in $\mathbf{w}_i$ with their sample estimators yields $\mathbf{W}_i$. Recall that $\widehat{\boldsymbol{\beta}}_1^*$ satisfies $\mathbf{U}_1^P(\widehat{\boldsymbol{\beta}}_1^*) = \sum_{i=1}^n \mathbf{W}_{1i}G_i$, where $\mathbf{W}_{1i}$ consists of the components of $\mathbf{W}_i$ corresponding to $\widehat{\boldsymbol{\beta}}_1$. Applying (A.3) to $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_1^*$ yields

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{W}_{1i}G_i = (\mathbf{A}_{11} + \boldsymbol{\Sigma}_{11})n^{1/2}(\widehat{\boldsymbol{\beta}}_1^* - \widehat{\boldsymbol{\beta}}_1) + o(1).$$

The conclusion then follows.

References

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

Cai, J., Fan, J., Li, R., and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika*, **92**, 303–316.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Ser. B*, **34**, 187–202.

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.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, **99**, 710–723.

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

Fu, W. J. (2003). Penalized Estimating Equations. *Biometrics*, **35**, 109–148.

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

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed. Hoboken, Wiley.

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.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lin, J. S. and Wei, L. J. (1992). Linear regression analysis for multivariate failure time observations. *Journal of the American Statistical Association*, **87**, 1091–1097.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussion). *Journal of the American Statistical Association*, **96**, 103–126.

Meinshausen, N. and Bühlmann, P. (2006) Variable selection and high-dimensional graphs with the lasso. *The Annals of Statistics*, **34**, 1436–1462.

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.

Robins, J. M., A. Rotnizky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

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.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, **18**, 354–372.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, **13**, 1378–402.

Wei, L. J., Ying, Z., and Lin, D. Y. (1990). Regression analysis of censored survival data based on rank tests. *Biometrika*, **77**, 845–851.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics*, **21**, 76–99.

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

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