NIH Public Access
Author Manuscript
J Am Stat Assoc. Author manuscript; available in PMC 2010 April 6.

Published in final edited form as:
J Am Stat Assoc. 2008 June 1; 103(482): 672–680. doi:10.1198/016214508000000184.

Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models

Brent A. Johnson,
Assistant Professor, Department of Biostatistics, Emory University, Atlanta, GA 30322 (E-mail: bajohn3@emory.edu)

D. Y. Lin, and
Dennis Gillings Distinguished Professor (E-mail: lin@bios.unc.edu)

Donglin Zeng
Associate Professor (E-mail: dzeng@bios.unc.edu), Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599

Abstract

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 and 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; Least absolute shrinkage and selection operator; Least squares; Linear regression; Missing data; Smoothly clipped absolute deviation

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. But these procedures 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

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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

where Y_i is the response variable, $\mathbf{x}_i$ is 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's have means 0. 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

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

where $p_\lambda(\cdot)$ is a penalty function. Appropriate choices of p_λ (detailed in Sec. 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, Rotnitzky, and Zhao 1994; Tsiatis 2006) and the Buckley–James estimator for semiparametric linear regression with censored responses (Buckley and James 1979). Another example arises from Lin and Ying's (2001) semiparametric regression analysis of longitudinal data. For this 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 of (2), so that their earlier results (Fan and Li 2001) for penalized least squares could be applied. In this article 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 (GEEs) (Liang and Zeger 1986) with bridge penalty (Frank and Friedman 1993; Knight and Fu 2000) is similar. That work deals only 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 interested mainly in the situations in which $\mathbf{U}(\boldsymbol{\beta})$ is not a score function or the derivative of any objective function. A penalized estimating function is defined as

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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

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 componentwise 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 $\mathbf{q}_\lambda(|\boldsymbol{\beta}|)$, that is, $q_{\lambda,j} = q_{\lambda,k}, j \neq k$. 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, $p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta|<\lambda)$; (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) the EN penalty (Zou and Hastie 2005), $p_\lambda(|\theta|) = \lambda_1|\theta| + \lambda_2\theta^2$; and (e) the 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 $\omega_j=1/|\tilde{\beta}_j^{o}|$, $j = 1, \ldots, d$, where $\tilde{\beta}^{o}=(\tilde{\beta}_1^{o},\ldots,\tilde{\beta}_d^{o})^T$ 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, but it has some 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) that results in an estimator that achieves an *oracle* property: that is, the estimator has the same limiting distribution as an estimator that 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; this penalty is popular in the statistical analysis of large data sets.

2.2 Application to Censored Data

Censoring is a common phenomenon in scientific studies (see Kalbfleisch and Prentice 2002, p. 12). The presence of censoring causes major complications in 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 (Cox 1972) plays essentially the same role as the standard likelihood (Tibshirani 1997; Fan and Li 2002; Cai, Fan, and Zhou 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 the response variable and censoring variable for the ith subject, $i = 1, \ldots, n$. The data consist of $(\tilde{Y}_i, \Delta_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, where $\tilde{Y}_i = \min(Y_i, C_i)$, $\Delta_i = I(Y_i \leq C_i)$ 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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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,

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

and

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

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

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

where

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

and $\hat{F}(t;\boldsymbol{\beta})$ 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 often is difficult to have complete data on all study subjects. Let R_i be the missingness indicator for the ith subject, with the event $\{R_i = \infty\}$ indicating that the ith subject has complete data. The observed data for the ith subject are $G_r(\mathbf{Z}_i)$, 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 consider only $R_i \in \{1, 2, \infty\}$ corresponding to $G_1(\mathbf{Z}_i) = \{Y_i\}$, $G_2(\mathbf{Z}_i) = \{x_i\}$, and $G_\infty(\mathbf{Z}_i) = \{Y_i, x_i\} = \mathbf{Z}_i$. 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: (a) $P(R_i = \infty|\mathbf{Z}_i = \mathbf{z}) > \kappa > 0$ and (b) $P(R_i = r|\mathbf{Z}_i = \mathbf{z}) = P(R_i = r\,|G_r(\mathbf{z}) = g_r)$.

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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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

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

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

where $\tilde{\lambda}_r\{G_r(\mathbf{Z}_i), \boldsymbol{\eta}\} = \{1 + \exp[-\mu_r\{G_r(\mathbf{Z}_i), \boldsymbol{\eta}\}]\}^{-1}$, $\mu_r\{G_r(\mathbf{Z}_i), \boldsymbol{\eta}\}$ is a linear predictor based on $G_r(\mathbf{Z}_i)$ and $\boldsymbol{\eta}$, $\tilde{\pi}\{r, G_r(\mathbf{Z}_i), \eta\}=\prod_{m=1}^{r}\tilde{\lambda}_m\{G_m(\mathbf{Z}_i), \eta\}$, and $\tilde{E}\{s_i(\boldsymbol{\beta})|G_r(\mathbf{Z}_i)\}$ 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,

$$\sup_{|\beta-\beta_0|\leq Mn^{-1/2}}|n^{-1/2}\mathbf{U}(\beta)-n^{-1/2}\mathbf{U}(\beta_0)-n^{1/2}\mathbf{A}(\beta-\beta_0)|=o_p(1).$$

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

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

a. For nonzero fixed θ, $\lim n^{1/2}q_{\lambda_n}(|\theta|) = 0$ and $\lim q'_{\lambda_n}(|\theta|)=0$.

b. 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).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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.2a prevents the j th element of the penalized estimating function from being dominated by the penalty term, $q_{\lambda_n}(|\beta_j|)\,\mathrm{sgn}(\beta_j)$, for $\beta_{j0} \neq 0$, because $\sqrt{n}q_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j)$ vanishes. But if $\beta_{j0} = 0$, then condition C.2b implies that $\sqrt{n}q_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j)$ diverges to $+\infty$ or $-\infty$, depending on the sign of β_j in the small neighborhood of β_{j0}. Thus the j -element of the penalized estimating function is dominated by the penalty term, so that any consistent solution, say $\hat{\boldsymbol{\beta}}$, to the estimating equation $\mathbf{U}^P(\boldsymbol{\beta}) = \mathbf{0}$ must satisfy $\hat{\beta}_j = 0$.

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, that is,

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

  with $a > 2$, it is easy to see that if we choose $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, then 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 that $\sqrt{n}\lambda_n \to 0$, $n\lambda_n \to \infty$ 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 be sufficiently large for θ sufficiently small, say $|\theta| < Mn^{-1/2}$. For simplicity, suppose that the data-dependent weights are defined as $\hat{w} = |\tilde{\theta}|^{-\gamma}$ for some $\gamma > 0$ and $\tilde{\theta}$ pertaining to the solutions to the unpenalized estimating equations. Then, trivially, $\sqrt{n}(\tilde{\theta} - \theta_0)=O_p(1)$, which implies that $\sqrt{n}\inf_{|\theta|\leq Mn^{-1/2}}\lambda_n\hat{w}=Mn\lambda_n \to \infty$, as desired. In this article we chose $\gamma = 1$ but Zou (2006, remarks 1 and 2) noted that other weights may be useful.
- **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 $\sqrt{n} \times \inf_{|\theta|\leq Mn^{-1/2}}\lambda_n/|\theta|=Mn\lambda_n \to \infty$. 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 $\hat{\boldsymbol{\beta}}= (\hat{\beta}_1, \ldots, \hat{\beta}_d)^T$ is called a zero-crossing to the penalized estimating equation if, for $j =1, \ldots, d$,

$$\varlimsup_{\varepsilon\to 0+} n^{-1}U_j^P(\widehat{\beta}+\varepsilon\mathbf{e}_j)U_j^P(\widehat{\beta}-\varepsilon\mathbf{e}_j)\leq 0,$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

where $\mathbf{e}_j$ is the j th canonical unit vector. In addition, an estimator $\hat{\boldsymbol{\beta}}$ is called an approximate zero-crossing if

$$\overline{\lim_{n\to\infty}}\ \overline{\lim_{\varepsilon\to 0+}}\ n^{-1}U_j^P(\widehat{\beta}+\varepsilon\mathbf{e}_j)U_j^P(\widehat{\beta}-\varepsilon\mathbf{e}_j)\le 0.$$

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 nonzero coefficients $s = \#\{j \mid \beta_{j0} \neq 0\}$. Under conditions C.1 and C.2, the following results hold:

a. There exists a root-n–consistent approximate zero-crossing of $\mathbf{U}^P(\boldsymbol{\beta})$, that is, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + O_p(n^{-1/2})$, such that $\hat{\boldsymbol{\beta}}$ 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 $\hat{\boldsymbol{\beta}} \equiv (\hat{\beta}_1, \ldots, \hat{\beta}_d)^T$, $\lim_n P(\hat{\beta}_j = 0 \text{ for } j > s) = 1$. Moreover, if we write $\hat{\boldsymbol{\beta}}_1 = (\hat{\beta}_1, \ldots, \hat{\beta}_s)^T$ and $\boldsymbol{\beta}_{01} = (\beta_{01}, \ldots, \beta_{0s})^T$, then

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

where $\mathbf{A}_{11}$, $\boldsymbol{\Sigma}_{11}$, and $\mathbf{V}_{11}$ are the first $s \times s$ submatrices of $\mathbf{A}$, $\mathrm{diag}\{-q'_{\lambda_n}(|\beta_{01}|)\mathrm{sgn}(\beta_{01})\}$, and $\mathbf{V}$, and $\mathbf{b}_n = -(q_{\lambda_n}(|\beta_{01}|)\times \mathrm{sgn}(\beta_{01}),\ldots, q_{\lambda_n}(|\beta_{0s}|)\,\mathrm{sgn}(\beta_{0s}))^T$.

c. Let $\mathbf{U}_1^P(\beta)$ and $\mathbf{U}_1(\boldsymbol{\beta})$ denote the first s-components of $\mathbf{U}^P(\boldsymbol{\beta})$ and $\mathbf{U}(\boldsymbol{\beta})$, and let $\beta=(\beta_1^T,\beta_2^T)^T$, 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 $\mathbf{U}_1((\beta_1^T, \mathbf{0}^T)^T)$ is continuous in $\boldsymbol{\beta}_1$, then there exists $\hat{\boldsymbol{\beta}}_1$ such that

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

that is, the solution is exact.

The proof of Theorem 1 is relegated to Appendix A. The asymptotic results for penalized weighted estimators readily follow 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 $P(\tilde{Y} - \boldsymbol{\beta}^T\mathbf{x} < c_0) < 1$ for all $\boldsymbol{\beta}$ in some neighborhood of $\boldsymbol{\beta}_0$.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

D.2. The random variable **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 have the oracle property. Conditions D.1–D.3 are the regularity conditions given by Ritov (1990, p. 306) to ensure that condition C.1 holds. The expressions for **A** and **V** were given by Ritov (1990) and Lai and Ying (1991a). The matrix **V** is directly estimable from the data, whereas **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 Section 2.3—namely, the penalized weighted estimators with SCAD, hard thresholding, and ALASSO penalties also have an oracle property. Technical conditions needed to obtain a strongly consistent estimator sequence and hence establish condition C.1 are given by Robins et al. (1994). Such technical conditions are assumed throughout the text of Tsiatis (2006), for example. The matrices **A** and **V** may be calculated directly; examples were given by Tsiatis (2006, chaps. 10 and 11).

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

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

and that a consistent estimator is given by

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

Other authors (e.g., Fu 2003) used the following alternative estimator for $\operatorname{cov}(\hat{\boldsymbol{\beta}}_1)$:

$$\widehat{\operatorname{cov}}(\hat{\beta}_1)=\tilde{\boldsymbol{\Omega}}_{11},\quad \tilde{\boldsymbol{\Omega}}=n^{-1}\left|(\widehat{\mathbf{A}}+\widehat{\textstyle\sum})^{-1}\widehat{\mathbf{V}}(\widehat{\mathbf{A}}+\widehat{\textstyle\sum})^{-1}\right|.$$

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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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

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

where $(G_1, \ldots, G_n)$ are independent standard normal variables and $(\mathbf{W}_{11}, \ldots, \mathbf{W}_{1n})$ are as given in Appendix B. In Appendix B we show that the conditional distribution of $n^{1/2}(\widehat{\beta}_1^* - \widehat{\beta}_1)$ given the observed data is the same in the limit as the unconditional distribution of $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{01})$ Thus we may estimate the covariance matrix of $\hat{\boldsymbol{\beta}}_1$ and construct confidence intervals for individual regression coefficients using the empirical distribution of $\widehat{\beta}_1^*$.

4. IMPLEMENTATION

In this article 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). Using condition C.1 and the local quadratic approximations for penalty functions (Fan and Li 2001, sec. 3.3), the MM algorithm is

$$\widehat{\beta}^{(k+1)}=\widehat{\beta}^{(k)}+\{\mathbf{A}(\widehat{\beta}^{(k)})+\sum\nolimits_{\lambda}(\widehat{\beta}^{(k)})\}^{-1}\mathbf{U}^P(\widehat{\beta}^{(k)}), \quad k \geq 0,$$

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

$$\sum\nolimits_{\lambda}(\beta)=\mathrm{diag}\{q_\lambda(|\beta_1|)/(\varepsilon+|\beta_1|), \ldots, q_\lambda(|\beta_d|)/(\varepsilon+|\beta_d|)\}$$

for ϵ a small number ($\epsilon = 10^{-6}$ in our examples). This algorithm requires that the estimating function $\mathbf{U}(\boldsymbol{\beta})$ be 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

$$\widehat{\beta}^{(k+1)}=\underset{\beta}{\mathrm{argmin}}\|\mathbf{U}(\beta) - n\sum\nolimits_{\lambda}(\widehat{\beta}^{(k)})\beta\|, \quad k \geq 0,$$

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

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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

where $\hat{\boldsymbol{\beta}}^{(0)}$ 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) showed 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 $\hat{\boldsymbol{\beta}}$ (i.e., $\hat{\boldsymbol{\beta}}_\lambda$) 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 (GCV) statistic (Wahba 1985):

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

where $RSS(\lambda)$ is the residual sum of squares $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\|^2$, and $d(λ)$ is the effective number of parameters, that is, $d(λ) = \text{tr}[\{\hat{\mathbf{A}} + \boldsymbol{\Sigma}_\lambda(\hat{\boldsymbol{\beta}}_\lambda)\}^{-1}\hat{\mathbf{A}}^T]$. Note that the intercept is omitted in $RSS(\lambda)$, because $\mathbf{y}$ may be centered at $n^{-1}\sum_{i=1}^n Y_i$. When the Y_i 's are potentially censored, $d(\lambda)$ still may be considered the effective number of parameters; however, $RSS(\lambda)$ is unknown. We propose estimating $n^{-1}\, RSS(\lambda)$ by

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

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

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

Both proposals are based on large-sample arguments—namely, $\hat{v}(\lambda)$is a consistent estimator for $\lim n^{-1}\, RSS(\lambda)$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, chap. 6). Thus our GCV statistic is

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

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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

5. SIMULATION STUDIES

5.1 Censored Data

We simulated 1,000 data sets of size n from the model

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

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and ε_i and $\mathbf{x}_i$ are independent standard normal with the correlation between the j th and kth components of $\mathbf{x}$ equal to $.5^{|j-k|}$. This model was considered by Tibshirani (1996) and Fan and Li (2001). We set the censoring distribution to be uniform(0,τ), where τ was chosen to yield approximately 30% censoring. We compared the model error, ME $\equiv (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T E(\mathbf{x}\mathbf{x}^T)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$, of the proposed penalized estimator with 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 approached that of the oracle estimator as n increases. When the signal-to-noise ratio was small (e.g., large n or small σ), oracle methods (SCAD, hard thresholding, ALASSO) outperformed LASSO and EN in terms of model error and model complexity. On the other hand, LASSO and EN tended to perform better than the oracle methods as σ/n increased.

Table 2 reports the results on the accuracy of the proposed re-sampling technique in estimating the variances of the nonzero estimated regression coefficients. The standard deviation (SD) pertains to the median absolute deviation of the estimated regression coefficients divided by .6745. The median of the standard error estimates, denoted by SD_m, gauges the performance of the resampling procedure. Evidently, the resampling procedure yielded reasonable standard error estimates, particularly for large n.

5.2 Missing Data

We simulated 1,000 datasets of size n from the model

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

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 $.5^{|j-k|}$. We considered two scenarios:

Model 1:

$$\beta=(.25, .5, 0, 0, .75, 1.5, .75, 0, 0, 1)^T$$

and

Model 2:

$$\beta=(0, 1.25, 0, 0, 0, 2, 0, 0, 0, 1.5)^T.$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

For a random design **X**, define the theoretical R^2

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

For σ= 1 and 2, both models 1 and 2 have theoretical R^2 =.89 and.67. Although models 1 and 2 have the same theoretical R^2, they have differing numbers of nonzero coefficients; the number of nonzero 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.6, whereas model 2 has a model fraction of.3. We simulated data such that subjects fall into one of three categories: R = 1 means that the subject was missing (x_1, x_2), R = 2 means that the subject was missing x_1, and $R = \infty$ means that the subject had 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 $\tilde{\lambda}_1\{\mathrm{G}_1(\mathbf{Z}_i), \boldsymbol{\eta}\}$.
2. If B_1 = 1, then set R = 1; otherwise, continue.
3. Simulate a Bernoulli random variable B_2 with probability $\tilde{\lambda}_2\{\mathrm{G}_2(\mathbf{Z}_i), \boldsymbol{\eta}\}$.
4. If B_2 = 1, then set R = 2; otherwise, set $R = \infty$.

We formulated the missingness process by logistic models

$$\text{logit}\,\tilde{\lambda}_1\{G_1(\mathbf{Z}_i)\}=\eta_{10}+\eta_{11}Y_i+\sum_{j=3}^{10}\eta_{1j}x_{ij}$$

and

$$\text{logit}\,\tilde{\lambda}_2\{G_2(\mathbf{Z}_i)\}=\eta_{20}+\eta_{21}Y_i+\sum_{j=2}^{10}\eta_{1j}x_{ij},$$

where

$$\eta_1=(-6,.75,0,0,1.25,1.5,1.25,0,0,1.25)^T$$

and

$$\eta_2=(-1.5,.5,1.5,0,0,.5,.5,.5,0,0,.5)^T.$$

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

Table 3 presents the numerical results with n = 250. Oracle methods (SCAD, hard thresholding, ALASSO) performed better than LASSO and EN in terms of relative model

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

error and complexity when there were a few strong predictors of response, as in model 1; however, oracle methods performed 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 (Paul Coverdell Prototype Registries Writing Group 2005). The goals of the registry include gaining a better understanding of factors associated with stroke and generally improving the quality of acute stroke care in the United States. For the purpose 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 consist of several thousand hospital admissions and 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 or not the patient was admitted to the intensive care unit (ICU; 1 if yes), and stroke subtype (1 if hemorrhagic; 0 if ischemic). Of the 800 patients, 419 (52.4%) had complete data (i.e., $R = \infty$), 94 (11.8%) were missing both GCS and serumn albumin (i.e., $R = 1$), and 287 (35.9%) were missing only GCS (i.e., $R = 2$).

Table 4 presents estimates for the nuisance parameter η in the stroke data. We see that the subjects missing both GCS and albumin (i.e., $R = 1$) tended to have higher creatinine and glucose levels but were less likely to be admitted to the ICU on 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 missing completely at random (MCAR) assumption is not well supported, and variable selection techniques based on such an assumption will lead to incorrect conclusions. Our analyses using methods described in Section 2 assuming data missing at random (MAR) are displayed in Table 5.

We use $\hat{\lambda}$= (.28,.63,.11,.16) for the SCAD, Hard, LASSO, and ALASSO estimates, and use $(\hat{\lambda}_1, \hat{\lambda}_2)$ = (.34,.9) for the EN estimates. Table 5 presents the regression coefficient estimates for the stroke data. Higher levels of albumin and creatine are strongly related to shorter LOS, whereas admission to the ICU is associated with longer LOS. Older patients tend to have LOS than younger patients; this is most easily explained by the fact that many older stroke patients quickly die in the hospital because their bodies are too weak to recover. Patients with hemorrhagic strokes have longer recovery periods and thus longer LOS. White stroke patients tend to have shorter LOS than non-whites. Finally, sex and glucose are weak predictors of LOS. The LASSO and EN estimates tend to retain more predictors in the final model and, thus have more complex models compared with the other penalized estimators. Among the SCAD, Hard, and ALASSO estimates, SCAD and ALASSO yielded similar coefficient estimates, whereas the Hard thresholding estimates yielded the sparsest model. Our methods yielded 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.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

7. REMARKS

We have developed a general methodology for selecting variables and simultaneously estimating their regression coefficients in semiparametric models. This development overcomes 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 or to quasi-likelihood, so that the mathematical arguments used 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 also can be used for variance estimation without the need for variable selection. Our simulation results indicate that the resampling method works well for modest sample sizes.

Rank estimators (Prentice 1978; Tsiatis 1990; Wei, Ying, and Lin 1990; Lai and Ying 1991b; Ying 1993) provide potential alternatives to the Buckley–James estimator but are computationally more demanding to implement (cf. Johnson 2008). In general, 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 of 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, because one model may fit the data better than another. A major advantage of model (1) is that the regression coefficients have a direct physical interpretation. Hazard ratio can be an awkward concept, especially when the response variable does not pertain to failure time.

Acknowledgments

This research was supported by National Institutes of Health grants P30 ES10126, T32 ES007018, and R03 AI068484 (B.J.); R37 GM047845 (D.L.); and R01 CA082659 (D.L. and D.Z.). The authors thank Paul Weiss for preparing the stroke data set.

References

Buckley J, James I. Linear Regression With Censored Data. Biometrika 1979;66:429–436.

Cai J, Fan J, Li R, Zhou H. Variable Selection for Multivariate Failure Time Data. Biometrika 2005;92:303–316. [PubMed: 19458784]

Cox DR. Regression Models and Life-Tables (with discussion). Journal of the Royal Statistical Society, Ser B 1972;34:187–202.

Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. Journal of the American Statistical Association 2001;96:1348–1360.

Fan J, Li R. Variable Selection for Cox's Proportional Hazards Model and Frailty Model. The Annals of Statistics 2002;30:74–99.

Fan J, Li R. New Estimation and Model Selection Procedures for Semi-parametric Modeling in Longitudinal Data Analysis. Journal of the American Statistical Association 2004;99:710–723.

Frank IE, Friedman JH. A Statistical View of Some Chemometrics Regression Tools. Technometrics 1993;35:109–148.

Fu WJ. Penalized Estimating Equations. Biometrics 2003;35:109–148.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Hunter DR, Li R. Variable Selection Using MM Algorithms. The Annals of Statistics 2005;33:1617–1642.

Johnson BA. Variable Selection in Semiparametric Linear Regression With Censored Data. Journal of the Royal Statistical Society, Ser B 2008;70:351–370.

Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. 2. Hoboken, NJ: Wiley; 2002.

Knight K, Fu W. Asymptotics for Lasso-Type Estimators. The Annals of Statistics 2000;28:1356–1378.

Lai TL, Ying Z. Large Sample Theory of a Modified Buckley–James Estimator for Regression Analysis With Censored Data. The Annals of Statistics 1991a;19:1370–1402.

Lai TL, Ying Z. Rank Regression Methods for Left-Truncated and Right Censored Data. The Annals of Statistics 1991b;19:531–556.

Liang KY, Zeger SL. Longitudinal Data Analysis Using Generalized Linear Models. Biometrika 1986;73:13–22.

Lin JS, Wei LJ. Linear Regression Analysis for Multivariate Failure Time Observations. Journal of the American Statistical Association 1992;87:1091–1097.

Lin DY, Ying Z. Semiparametric and Nonparametric Regression Analysis of Longitudinal Data. Journal of the American Statistical Association 2001;96:103–126. (with discussion).

Meinshausen N, Bühlmann P. Variable Selection and High-Dimensional Graphs With the Lasso. The Annals of Statistics 2006;34:1436–1462.

Paul Coverdell Prototype Registries Writing Group. Acute Stroke Care in the US: Results From 4 Pilot Prototypes of the Paul Coverdell National Acute Stroke Registry. Stroke 2005;36:1232–1240. [PubMed: 15890989]

Prentice RL. Linear Rank Tests With Right-Censored Data. Biometrika 1978;65:167–179.

Ritov Y. Estimation in a Linear Regression Model With Censored Data. The Annals of Statistics 1990;18:303–328.

Robins JM, Rotnitzky A, Zhao LP. Estimation of Regression Coefficients When Some Regressors Are not Always Observed. Journal of the American Statistical Association 1994;89:846–866.

Tibshirani RJ. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Ser B 1996;58:267–288.

Tibshirani RJ. The Lasso Method for Variable Selection in the Cox Model. Statistics in Medicine 1997;16:385–395. [PubMed: 9044528]

Tsiatis AA. Estimating Regression Parameters Using Linear Rank Tests for Censored Data. The Annals of Statistics 1990;18:354–372.

Tsiatis, AA. Semiparametric Theory and Missing Data. New York: Springer; 2006.

van der Vaart, AW.; Wellner, JA. Weak Convergence and Empirical Processes. New York: Springer; 1996.

Wahba G. A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem. The Annals of Statistics 1985;13:1378–1402.

Wei LJ, Ying Z, Lin DY. Regression Analysis of Censored Survival Data Based on Rank Tests. Biometrika 1990;77:845–851.

Ying Z. A Large Sample Study of Rank Estimation for Censored Regression Data. The Annals of Statistics 1993;21:76–99.

Zou H. The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association 2006;101:1418–1429.

Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society, Ser B 2005;67:301–320.

APPENDIX A: PROOF OF THEOREM 1

To prove part a, we consider $\widehat{\beta}=(\widehat{\beta}_1^T, \mathbf{0}^T)^T$, where $\widehat{\beta}_1=\beta_{01}+n^{-1}\mathbf{A}_{11}^{-1}\mathbf{U}_1(\beta_0)$. Because $n^{1/2}q_{\lambda_n}(|\beta_{0j}|) \to 0$, $j = 1, \ldots, s$, under condition C.2.a and $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0+O_p(n^{-1/2})$, we have

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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

Under condition C.2b, for $j = s + 1, \ldots, d$, $n^{-1/2}U_j^P(\widehat{\beta}+\varepsilon\mathbf{e}_j)$ and $n^{-1/2}U_j^P(\widehat{\beta}-\varepsilon\mathbf{e}_j)$ are dominated by $-n^{1/2}q_{\lambda_n}(\varepsilon)$and $n^{1/2}q_{\lambda_n}(\varepsilon)$, so they have opposite signs when ε goes to 0. Therefore, $\hat{\boldsymbol{\beta}}$ is an approximate zero-crossing by definition.

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

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

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

$$\begin{aligned}o_p(1)=\{&n^{-1/2}U_j(\beta_0)+n^{1/2}\mathbf{A}_j(\widehat{\beta}-\beta_0)\\&+o_p(1)-n^{1/2}q_{\lambda_n}(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j)\}^2,\end{aligned}$$

where $\mathbf{A}_j$ is the j th row of $\mathbf{A}$. The first three terms on the right side are of order $O_p(1)$. As a result, there exists some M' such that for large n,

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

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

To prove the second part of part b, because

$$\begin{aligned}o_p(1)=&n^{-1/2}\mathbf{U}_1(\beta_0)+n^{1/2}\mathbf{A}_{11}(\widehat{\beta}_1-\beta_{01})\\&-n^{1/2}q_{\lambda_n}(|\widehat{\beta}_1|)\mathrm{sgn}(\widehat{\beta}_1),\end{aligned}$$

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

$$\begin{aligned}&n^{1/2}\{(\mathbf{A}_{11}+\textstyle\sum_{11})(\widehat{\beta}_1-\beta_{01}+(\mathbf{A}_{11}+\textstyle\sum_{11})^{-1}\mathbf{b}_n\}\\&=-n^{-1/2}\begin{pmatrix}U_1(\beta_0)\\\vdots\\U_s(\beta_0)\end{pmatrix}+o_p(1)\to_d\mathrm{N}(0,\mathbf{V}_{11}).\end{aligned}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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

$$\begin{aligned} & n^{-1/2}(\beta_1 - \beta_{01})^T \mathbf{A}_{11}^T \mathbf{U}_1^P(\beta) \\ =&(\beta_1 - \beta_{01})^T \mathbf{A}_{11}^T \left\{ n^{-1/2}\mathbf{U}_1(\beta) - n^{1/2} q_{\lambda_n}(|\beta_1|)\mathrm{sgn}(\beta_1) \right\} \\ =&O_p(|\beta_1 - \beta_{01}|) + n^{1/2}(\beta_1 - \beta_{01})^T \mathbf{A}_{11}^T \mathbf{A}_{11}(\beta_1 - \beta_{01}) \\ -&n^{1/2}(\beta_1 - \beta_{01})\mathbf{A}_{11}^T \mathrm{diag}\left\{ q'_{\lambda_n}(|\beta_j^*|)\mathrm{sgn}(\beta_{0j}) \right\}(\beta_1 - \beta_{01}), \end{aligned}$$

where β_j^* is between β_j and β_{0j} for $j = 1, \ldots, s$. Because $\mathbf{A}_{11}$ is non-singular, the second term on the right side is larger than $a_0 r^2 n^{-1/2}$, where a_0 is the smallest eigenvalue of $\mathbf{A}_{11}^T \mathbf{A}_{11}$. The first term is of order $rO_p(n^{-1/2})$. Because $\max_j q'_{\lambda_n}(|\beta_j^*|) \to 0$, the third term is dominated by the second term. Therefore, for any ε, if we choose r sufficiently large 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, we then have

$$P\left[\min_{|\beta_1 - \beta_{01}| = n^{-1/2} r} (\beta_1 - \beta_{01})^T \mathbf{A}_{11}^T \mathbf{U}_1^P((\beta_1^T, \mathbf{0}^T)^T) > 0 \right] > 1 - \varepsilon.$$

Applying the Brouwer fixed-point theorem to the continuous function $\mathbf{U}_1^P((\beta_1^T, \mathbf{0}^T)^T)$, we see that $\min_{|\beta_1 - \beta_{01}| = n^{-1/2} r} (\beta_1 - \beta_{01})^T \mathbf{A}_{11}^T \times \mathbf{U}^P((\beta_1^T, \mathbf{0}^T)^T) > 0$ implies that $\mathbf{A}_{11}^T \mathbf{U}_1^P((\beta_1^T, \mathbf{0}^T)^T)$ has a solution within this ball or, equivalently, $\mathbf{U}_1^P((\beta_1^T, \mathbf{0}^T)^T)$ has a solution within this ball. That is, we can choose an exact solution $\widehat{\beta} = (\widehat{\beta}_1^T, \mathbf{0}^T)^T$ to $\mathbf{U}_1^P(\beta) = \mathbf{0}$ with $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_0 + O_p(n^{-1/2})$. Thus $\hat{\boldsymbol{\beta}}$ is a zero-crossing of $\mathbf{U}^P(\boldsymbol{\beta})$.

APPENDIX B: CONDITIONAL DISTRIBUTION OF (β^1∗−β^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:

$$\begin{aligned} n^{-1/2}\mathbf{U}_1^P(\beta) = & n^{-1/2}\mathbf{U}_1^P(\beta_0) + (\mathbf{A}_{11} + \Sigma_{11}) n^{1/2}(\beta_1 - \beta_{01}) \\ & + o(\max\{1, n^{1/2}\|\beta_1 - \beta_{01}\|\}). \end{aligned} \tag{B.1}$$

In addition,

$$n^{-1/2}\mathbf{U}_1(\beta_0) = n^{-1/2}\sum_{i=1}^{n} \mathbf{w}_{1i} + o(1),$$

where $\mathbf{w}_{1i}$ comprises the components of $\mathbf{w}_i$ corresponding to $\boldsymbol{\beta}_1$, and $\mathbf{w}_i$, $i = 1, \ldots, n$, as given by Lin and Wei (1992), are n independent mean-0 random vectors. Replacing the unknown

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

quantities in $\mathbf{w}_i$ with their sample estimators yields $\mathbf{W}_i$. Recall that $\widehat{\beta}_1^*$ satisfies $\mathbf{U}_1^P(\widehat{\beta}_1^*)=\sum_{i=1}^n \mathbf{W}_{1i}G_i$, where $\mathbf{W}_{1i}$ comprises the components of $\mathbf{W}_i$ corresponding to $\hat{\boldsymbol{\beta}}_1$. Applying (B.1) to $\hat{\boldsymbol{\beta}}_1$ and $\widehat{\beta}_1^*$ yields

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

The conclusion then follows.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Simulation results on model selection with censored data: MRME and the average number of correct (c) and incorrect (I) 0's

		Average number 0's	
Method	**MRME (%)**	**C**	**I**
$n = 50, \sigma = 3$			
SCAD	69.48	4.73	.35
Hard	73.41	4.30	.17
LASSO	66.16	3.99	.11
ALASSO	57.77	4.40	.17
EN	76.48	3.54	.08
Oracle	32.76	5	0
n = 50, $\sigma = 1$			
SCAD	40.11	4.78	.01
Hard	69.79	4.18	.01
LASSO	64.48	3.97	.01
ALASSO	48.21	4.90	.01
EN	95.55	3.49	0
Oracle	31.30	5	0

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 2

Simulation results on standard error estimation for the nonzero coefficients ($\beta_1, \beta_2, \beta_5$) in least squares regression with censored data

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
	SD	**SD$_m$**	**SD**	**SD$_m$**	**SD**	**SD$_m$**
SCAD	.145	.129	.135	.128	.128	.114
Hard	.151	.130	.145	.129	.138	.119
LASSO	.160	.134	.145	.143	.161	.130
ALASSO	.149	.132	.130	.133	.133	.113
EN	.172	.113	.151	.111	.155	.103
Oracle	.144	.129	.136	.126	.143	.111

NOTE: SD refers to the mean absolute deviation of the estimated regression coefficients divided by .6745; SD$_m$, to the median of the standard error estimates. The table entries are for a sample size $n = 100$ and (error) standard deviation $\sigma = 1$.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 3

Simulation results on model selection with missing data: MRME and the average number of correct (C) and incorrect (I) 0's

	Model 1			Model 2		
		Average number of 0's			Average number of 0's	
Method	MRME (%)	C	I	MRME (%)	C	I
σ= 1						
SCAD	81.79	3.35	.21	42.60	5.56	0
Hard	82.38	3.37	.25	48.73	5.79	.01
LASSO	87.88	2.42	.09	66.49	4.11	0
ALASSO	82.24	3.55	.23	37.74	6.25	0
EN	85.59	2.38	.08	70.56	3.92	0
σ= 2						
SCAD	93.64	3.33	.69	48.73	5.92	.02
Hard	90.10	3.70	1.12	46.24	6.37	.05
LASSO	82.29	2.54	.40	59.96	4.56	.02
ALASSO	82.01	3.37	.70	48.87	6.08	.02
EN	88.62	2.55	.44	66.17	4.63	.03

NOTE: For $\sigma = 1$ and $\sigma = 2$, models 1 and 2 have theoretical R^2 =.89 and.67; however, the number of nonzero coefficients is six in model 1 but only three in model 2.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 4

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

	η_1	η_2
(int)	$-2.342_{(.152)}$	$.478_{(.082)}$
Albumin		$-.112_{(.089)}$
Creatinine	$-.492_{(.291)}$	$-.101_{(.091)}$
Sex	$-172_{(.113)}$	$.043_{(.079)}$
Glucose	$-.286_{(.164)}$	$-.067_{(.084)}$
ICU	$-.470_{(.155)}$	$-.304_{(.091)}$
Age	$.045_{(.124)}$	$.006_{(.087)}$
Type	$-.101_{(.144)}$	$-.213_{(.094)}$
Race	$.084_{(.122)}$	$-.034_{(.085)}$
LOS	$-.007_{(.140)}$	$-.045_{(.092)}$

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 5

Estimated regression coefficients and their standard errors in the stroke data

	Full	SCAD	Hard	LASSO	ALASSO	EN
GCS	$-.762_{(.327)}$	$-.603_{(.434)}$	$-.864_{(.587)}$	$-.681_{(.480)}$	$-.584_{(.400)}$	$-.628_{(.424)}$
Albumin	$-1.142_{(.306)}$	$-.958_{(.450)}$	$-1.043_{(.486)}$	$-.984_{(.425)}$	$-.876_{(.402)}$	$-.882_{(.387)}$
Creatinine	$-.726_{(.331)}$	$-.372_{(.177)}$	$-.734_{(.347)}$	$-.529_{(.255)}$	$-.365_{(.179)}$	$-.402_{(.199)}$
Sex	$-.007_{(.288)}$	$0_{(-)}$	$0_{(-)}$	$0_{(-)}$	$0_{(-)}$	$0_{(-)}$
Glucose	$-.312_{(.310)}$	$0_{(-)}$	$0_{(-)}$	$-.140_{(.165)}$	$0_{(-)}$	$-.030_{(.039)}$
ICU	$1.861_{(.323)}$	$2.043_{(.442)}$	$1.970_{(.469)}$	$1.807_{(.419)}$	$1.947_{(.415)}$	$1.771_{(.392)}$
Age	$-.696_{(.324)}$	$-.293_{(.203)}$	$-.678_{(.465)}$	$-.586_{(.369)}$	$-.405_{(.260)}$	$-.516_{(.312)}$
Type	$.553_{(.333)}$	$.200_{(.155)}$	$0_{(-)}$	$.448_{(.335)}$	$.213_{(.158)}$	$.381_{(.273)}$
Race	$-1.316_{(.315)}$	$-1.403_{(.374)}$	$-1.320_{(.366)}$	$-1.216_{(.331)}$	$-1.242_{(.332)}$	$-1.151_{(.310)}$