Nonparametric Statistics
Vol. 00, No. 00, Month-Month 200x, 1–11

Rank-based variable selection

Brent A. Johnson* and Limin Peng

Department of Biostatistics, Emory University, Atlanta, Georgia 30322, USA

(18 September 2007)

This note considers variable selection in the robust linear model via R-estimates. The proposed rank-based approach is a generalization of the penalized least squares estimators where we replace the least squares loss function with Jaeckel's (1972) dispersion function. Our rank-based method is robust to outliers in the errors and has roots in traditional nonparametric statistics for simple location-shift problems. We establish the theoretical properties of our estimators which ensure desirable asymptotic behaviour of setting coefficient estimates to zero for unimportant variables and consistently estimating coefficients for important variables. Numerical studies indicate that the rank-based methods perform well for both light- and heavy-tailed error distributions.

KEY WORDS: Lasso; Oracle property; Penalized least squares; Robust linear model.

AMS MATHEMATICS SUBJECT CLASSIFICATION CODES: 62J07 (Ridge regression, shrinkage estimator); 62G30 (Order statistics; empirical distribution functions); 62G35 (Robustness).

1 Introduction

Recently, a substantial amount of attention has been paid to variable selection in the linear model through so-called penalized least squares (PLS) estimators [1–4]. Consider the linear regression model

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i, \ (i = 1, \ldots, n), \tag{1}$$

where y_i is the response variable and $\mathbf{x}_i$ is a d-vector of fixed predictors for the ith subject, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)'$ is a d-vector of regression coefficients, and $(\varepsilon_1, \ldots, \varepsilon_n)$ are independent and identically distributed errors with absolutely continuous density f. Here, we assume that the predictors have been standardized to have mean zero and unit variance. A PLS estimator is defined as the minimizer of

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

where $\text{RSS} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, $p_{\lambda,j}$ is a penalty function on the j-th coefficient and λ is a regularization parameter. Penalty functions $p_{\lambda,j}$ that shrink some coefficient estimates to zero are also considered variable selection procedures and we only discuss those penalty functions in the sequel. In the presence of outliers, it is desirable to replace the residual sum of squares (RSS) in (2) with a robust statistic. Naturally, one could replace RSS with a statistic based on robust M-estimators [5]. Alternatively, one could replace RSS with a rank-based statistic. The latter rank-based method has not been described in the literature and is the focus of this paper. Two interesting extensions of the methods discussed here are high-breakdown and censored outcome variable selection through the high-breakdown

*Corresponding author. Email: bajohn3@emory.edu

Nonparametric Statistics
ISSN 1048-5252 print/ ISSN 1029-0311 online
http://www.tandf.co.uk/journals
DOI: 10.1080/10485250xxxxxxxxxxxxx

rank regression estimator [6] and the rank-based accelerated failure time model [7–9], respectively. Hence, we believe the methods described in the sequel may be of general interest and utility.

In this note, we suggest a robust variable selection approach by replacing the RSS in (2) with Jaeckel's [10] dispersion function. Hence, our rank-based objective function is defined as

$$Q_n(\boldsymbol{\beta}) = D_n(\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_{\lambda,j}(|\beta_j|), \tag{3}$$

where $D_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \phi[R\{e_i(\boldsymbol{\beta})\}/(n+1)-(1/2)]e_i(\boldsymbol{\beta})$, $e_i(\boldsymbol{\beta}) = y_i - \boldsymbol{\beta}'\mathbf{x}_i$, $\phi(\cdot)$ is a weight function assumed to be non-decreasing, and $R\{e_i(\boldsymbol{\beta})\}$ is the rank of $e_i(\boldsymbol{\beta})$ among $\{e_1(\boldsymbol{\beta}), \ldots, e_n(\boldsymbol{\beta})\}$. It is well-known that when the weight function is the identity function, $D_n(\boldsymbol{\beta})$ is equivalent to the usual Wilcoxon statistic. The proposed estimator for $\boldsymbol{\beta}$ is defined as $\widehat{\boldsymbol{\beta}}_n(\lambda) =$ arg min $Q_n(\boldsymbol{\beta})$.

In this paper, we consider several penalty functions $p_{\lambda,j}$ of recent interest. The hard and soft thresholding rules from wavelet thresholding [18] lead to the hard penalty function, $p_{\lambda,j}(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$ (for all j), and the lasso [1] penalty function $p_{\lambda,j}(|\beta_j|) = \lambda|\beta|$ (for all j), respectively. The scad penalty [2] was presented as a compromise between the hard and soft penalty and is defined

$$p_{\lambda,j}(|\beta|) = \lambda|\beta|\left\{I(|\beta| \le \lambda) + \frac{(a - |\beta|/2\lambda)}{(a-1)} I(\lambda < |\beta| \le a\lambda) + \frac{a^2\lambda}{(a-1)2|\beta|} I(|\beta| > a\lambda)\right\}, \quad j = 1, \ldots, d,$$

where $a > 2$. Elastic net (en) penalty, $p_{\lambda,j}(|\beta|) = \lambda_1|\beta| + \lambda_2\beta^2$ (for all j), was introduced by Zou and Hastie [3], and like the scad penalty, is a mixing of two penalties: the ℓ_1 and ridge penalty [11]. Adaptive lasso (alasso) is a consistent version of the ℓ_1 penalty [4] and is defined $p_{\lambda,j}(|\beta|) = \lambda|\beta|w_j$, for a data-dependent weight w_j on the j-th coefficient. In our analyses, we use the weight $w_j = 1/|\tilde{\beta}_j|$, where $\tilde{\boldsymbol{\beta}}_n = (\tilde{\beta}_1, \ldots, \tilde{\beta}_d)'$ is the d-vector of usual rank coefficient estimates [12,13] defined as the minimizer of the unpenalized dispersion function $D_n(\boldsymbol{\beta})$. Note that some penalty functions are defined in terms of more than one regularization parameter (e.g. (a, λ) for scad and (λ_1, λ_2) for elastic net). Without loss of generality, we use λ to refer to regularization parameters, regardless of their dimension.

The remainder of our article is organized as follows. The main results are described in Section 2 and an efficient algorithm for calculating the penalized rank statistics in Section 3. We illustrate the utility of our method through simulation studies in Section 4 and one real data set in Section 5.

2 Main results

In this section, we study the large sample properties of the proposed rank estimators. We show that the scad, hard, and adaptive lasso rank-based estimators have the the so-called *oracle* property [18], that is, under appropriate regularity conditions, the minimizer of $Q_n(\boldsymbol{\beta})$ behaves asymptotically as if the true model were known *a priori*. To describe the large sample properties, we use the short-hand $\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\beta}}_n(\lambda)$ and introduce the following new notation.

Without loss of generality, suppose that $\beta_{0j} \neq 0$ for $j \le s$ and $\beta_{0j} = 0$ for $j > s$. Then, a definition for the partitioned vector of true regression parameters follows naturally, i.e. $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}', \boldsymbol{\beta}_{02}')'$ where $\boldsymbol{\beta}_{02} = 0$. The estimator is similarly partitioned as $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2')'$ where $\widehat{\boldsymbol{\beta}}_1 = (\widehat{\beta}_1, \ldots, \widehat{\beta}_s)'$ and $\widehat{\boldsymbol{\beta}}_2 = (\widehat{\beta}_{s+1}, \ldots, \widehat{\beta}_d)'$. The derivatives of the penalty functions play important roles in the operating characteristics of the penalized estimators. We use the dummy argument θ for one of β_j, $j = 1, \ldots, d$, and adopt the following notation $(d/d\theta)p_{\lambda_n,j}(|\theta|) = q_{\lambda_n,j}(|\theta|)\text{sgn}(\theta)$, and $(d/d\theta)q_{\lambda_n,j}(|\theta|) = \dot{q}_{\lambda_n,j}(|\theta|)\text{sgn}(\theta)$. For simplicity in exposition, we drop the subscript j when the penalty function does not depend on j.

We begin with some preliminary results on rank statistics adopted in this paper [10,14,15]. Throughout the paper we assume that the regularity conditions in Jurečková [15] hold (See also, Heiler and Willers [16]).

Define the d-dimensional score function $\nabla D_n(\boldsymbol{\beta}) = (-1)\mathbf{U}_n(\boldsymbol{\beta})$ where

$$\mathbf{U}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})\phi\left[\frac{R\{e_i(\boldsymbol{\beta})\}}{n+1} - \frac{1}{2}\right],$$

where $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$ and is exactly zero when the predictors are standardized. Define the matrix $\mathbf{A} = \lim_n \tau^{-1}(\mathbf{X}'\mathbf{X})/n$ where

$$\tau^{-1} = \int_{-\infty}^{\infty}\phi\{F(u)\}\left\{\frac{d}{du}f(u)\right\}du.$$

First, we have the usual asymptotic linearity for rank statistics [14, 15, 20]; that is, for any $M > 0$,

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le M\cdot n^{-1/2}}\|n^{-1/2}\mathbf{U}_n(\boldsymbol{\beta}) - n^{-1/2}\mathbf{U}_n(\boldsymbol{\beta}_0) + \mathbf{A}\{n^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\}\| \to_P 0.$$

Second, $D_n(\boldsymbol{\beta})$ has a quadratic approximation in any root-n neighborhood of $\boldsymbol{\beta}_0$ [10]. Specifically, for any $M > 0$,

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le M\cdot n^{-1/2}}|D_n(\boldsymbol{\beta}) - D_n(\boldsymbol{\beta}_0) + n^{-1/2}\mathbf{U}_n(\boldsymbol{\beta}_0)'\{n^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\} - \frac{1}{2}\{n^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\}'\mathbf{A}\{n^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\}| \to_p 0.$$

In addition, $n^{-1/2}\mathbf{U}_n(\boldsymbol{\beta}_0) \to_d N(0,\mathbf{V})$ with $\mathbf{V} = \kappa^2\tau\mathbf{A}$, where $\kappa = \left[\int_0^1\{\phi(u) - \int_0^1\phi(v)dv\}^2du\right]^{1/2}$. It follows easily that $n^{1/2}(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to_d N(0,\kappa^2\tau\mathbf{A}^{-1})$. In the special case of Wilcoxon score (i.e. $\phi(u) = u - 1/2$), it is easy to check that $\kappa = 12^{-1/2}$, $\tau = \left[\int_{-\infty}^{\infty}\{f(u)\}^2du\right]^{-1}$, and $\mathbf{V} = (12)^{-1}\lim_n(\mathbf{X}'\mathbf{X})/n$.

The following theorem states the main theoretical result regarding the scad and hard penalized rank estimators — including the existence of an $n^{1/2}$-consistent estimator, the sparsity of the estimator (that is, shrinking some coefficient estimates exactly to zero) and the asymptotic normality of the estimator.

THEOREM 2.1 *Let $(y_1,\mathbf{x}_1),\ldots,(y_n,\mathbf{x}_n)$ be independent and identically distributed. Assume that $q_{\lambda_n,j} = q_{\lambda_n},\ j = 1,\ldots,d$, and*
(i) for non-zero fixed θ, $\lim n^{1/2}q_{\lambda_n}(|\theta|) = 0$ and $\lim \dot{q}_{\lambda_n}(|\theta|) = 0$;
(ii) for any $M > 0$, $\lim\lambda_n^{-1}\inf_{|\theta|\le Mn^{-1/2}}q_{\lambda_n}(|\theta|) > 0$;
(iii) $\lambda_n \to 0$ and $n^{1/2}\lambda_n \to \infty$.
Then the following conclusions hold:

(a) there exists a local minimizer $\widehat{\boldsymbol{\beta}}_n$ of $Q_n(\boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.
(b) $\lim_{n\to\infty}\Pr(\widehat{\boldsymbol{\beta}}_2 = 0) = 1$ and

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

where $\mathbf{A}_{11},\boldsymbol{\Sigma}_{11}$ and $\mathbf{V}_{11}$ are the first $s\times s$ sub-matrices of $\mathbf{A},\boldsymbol{\Sigma} = diag\{\dot{q}_{\lambda_n}(|\boldsymbol{\beta}_0|)sgn(\boldsymbol{\beta}_0)\}$ and $\mathbf{V}$, respectively, and $\mathbf{b}_n = (q_{\lambda_n}(|\beta_{01}|)sgn(\beta_{01}),\ldots,q_{\lambda_n}(|\beta_{0s}|)sgn(\beta_{0s}))^T$.

Remark 1. In Theorem 2.1, we impose assumptions on the penalty function and the regularization parameter in order to simultaneously achieve the root-n consistency of the regularized rank estimation and the

consistency of variable selection. To see that see that scad penalty satisfies conditions (i)-(ii), note that

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

where $c_+ = cI(c \geq 0)$. If we let $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, then $\lim_n \sqrt{n} q_{\lambda_n}(|\theta|) = \lim_n \dot{q}_{\lambda_n}(|\theta|) = 0$ for $\theta \neq 0$ and $\sqrt{n} \inf_{|\theta| \leq M n^{-1/2}} q_{\lambda_n}(|\theta|) = \sqrt{n}\lambda_n$. For the hard penalty, we have $q_{\lambda_n}(|\theta|) = 2(\lambda_n - |\theta|) I(|\theta| < \lambda_n)$ which again satisfies conditions (i)-(ii) if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$.

In Theorem 2.2, we show that the oracle property of the rank-based variable selection approach persists with the adaptive lasso penalty, albeit under slightly different conditions.

THEOREM 2.2 *Let $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ be independent and identically distributed and consider adaptive lasso penalty, i.e. $p_{\lambda,j}(|\beta_j|) = \lambda |\beta_j| / |\tilde{\beta}_j|$.*
(a) If $\sqrt{n}\lambda_n = O_p(1)$, then $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.
(b) If $\sqrt{n}\lambda_n \to \lambda_0$ with $0 \leq \lambda_0 < \infty$ and $n\lambda_n \to \infty$, then the adaptive alasso estimator satisfies $\lim_n P(\widehat{\boldsymbol{\beta}}_2 = 0) = 1$ and

$$n^{1/2} \mathbf{A}_{11} \left\{ \widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01} + n^{-1/2} \mathbf{A}_{11}^{-1} \lambda_0 \mathbf{b}_1 \right\} \to_d N(0, \mathbf{V}_{11}),$$

where $\mathbf{A}_{11}$ and $\mathbf{V}_{11}$ are the first $s \times s$ sub-matrices of $\mathbf{A}$ and $\mathbf{V}$, respectively, and $\mathbf{b}_1 = (sgn(\beta_{01})/|\beta_{01}|, \ldots, sgn(\beta_{0s})/|\beta_{0s}|)'$.

The proofs for Theorems 2.1-2.2 are given in the Appendix. Finally, a consistent estimator for the asymptotic variance-covariance matrix of $\widehat{\boldsymbol{\beta}}_1$ is given by

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

where $\mathbf{X}_1$ refers to the first s columns of the design matrix, $\widehat{\mathbf{A}}_{11} = \hat{\tau}^{-1} \mathbf{X}_1' \mathbf{X}_1 / n$, $\hat{\tau}$ a consistent estimate of τ [17], $\widehat{\mathbf{V}}_{11} = \mathbf{X}_1' \mathbf{X}_1 / n$, and $\widehat{\Sigma}_{11}$ is given below in (5).

3 Implementation

In this paper, we estimate the penalized regression coefficients using a majorize-minorize (MM) algorithm [22]. If we let $\boldsymbol{\beta}^{[k]}$ denote the k-th iterate for fixed λ and $\boldsymbol{\beta}^{[0]} = \tilde{\boldsymbol{\beta}}_n$, then the iterative MM algorithm is written:

$$\boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} - \rho_k \left[\nabla^2 D_n(\boldsymbol{\beta}^{[k]}) - n\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}^{[k]}) \right]^{-1} \left[\nabla D_n(\boldsymbol{\beta}^{[k]}) - n\mathbf{q}_\lambda(\boldsymbol{\beta}^{[k]}) \right], \quad k > 0, \tag{4}$$

where $\nabla D_n(\boldsymbol{\beta})$ and $\nabla^2 D_n(\boldsymbol{\beta})$ denote the gradient vector and Hessian matrix, respectively, ρ_k is some positive scalar, and

$$\begin{aligned} \mathbf{q}_\lambda(\boldsymbol{\beta}) &= \{q_{\lambda,1}(|\beta_1|)\mathrm{sgn}(\beta_1), \ldots, q_{\lambda,d}(|\beta_d|)\mathrm{sgn}(\beta_d)\}', \\ \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}) &= \mathrm{diag}\left\{ q_{\lambda,1}(|\beta_1|)/(\epsilon + |\beta_1|), \ldots, q_{\lambda,d}(|\beta_d|)/(\epsilon + |\beta_d|) \right\}, \end{aligned} \tag{5}$$

with ϵ chosen to be a small number. The matrix $\nabla^2 D_n(\boldsymbol{\beta})$ depends on the unknown density f but may be estimated using existing methods [17]. Our MM algorithm continues until successive iterates are less a user-defined threshold; in our case, we continue until $\max_{1 \leq j \leq d} |\beta_j^{[k+1]} - \beta_j^{[k]}| < 10^{-8}$. Interested readers may find exemplary **R** programs that implement the the above MM algorithm (4) on the first author's website: `http://userwww.service.emory.edu/~bajohn3`.

We tune our estimators by minimizing a generalized cross-validation (GCV) statistic. In the penalized least squares setup, the GCV statistic is given by:

$$\mathrm{GCV_{LS}}(\lambda) = \frac{\mathrm{RSS}(\lambda)/n}{\{1 - e(\lambda)/n\}^2}, \tag{6}$$

where $\mathrm{RSS}(\lambda) = \|\mathbf{y} - \hat{\mathbf{y}}(\lambda)\|^2$, $\hat{\mathbf{y}}(\lambda) = \hat{\alpha} + \mathbf{X}\widehat{\boldsymbol{\beta}}_n(\lambda)$, $\hat{\alpha} = \bar{\mathbf{y}}$ [an estimate of the intercept $\alpha = E(\varepsilon_1)$], $e(\lambda)$ is the effective number of parameters, i.e., $e(\lambda) = \mathrm{tr}[\mathbf{X}\{\mathbf{X}'\mathbf{X} + \boldsymbol{\Sigma}_\lambda(\widehat{\boldsymbol{\beta}}_n(\lambda))\}^{-1}\mathbf{X}']$. Naturally, one can extend (6) by replacing $\mathrm{RSS}(\lambda)$ with the sum of absolute deviations, $\|\mathbf{y} - \hat{\mathbf{y}}(\lambda)\|_1$ and using a robust estimator for α. To avoid estimating α, we propose a different cross-validation statistic. In the rank-regression framework, we optimize the convex dispersion function, $D_n(\boldsymbol{\beta})$. Because $D_n(\boldsymbol{\beta})$ is itself a norm, it has a similar geometric interpretation to the residual sum of squares. Thus, we define our new criterion

$$\mathrm{GCV}(\lambda) = \frac{D_n[\widehat{\boldsymbol{\beta}}_n(\lambda)]/n}{\{1 - e(\lambda)/n\}^2}.$$

The final regularization parameter is chosen to minimize $\mathrm{GCV}(\lambda)$. Our experience suggests the proposed cross-validation statistic and one based on absolute deviations perform equally well for symmetric error distributions. We conjecture that $\mathrm{GCV}(\lambda)$ may offer some advantages for asymmetric error distributions although we have not fully investigated this claim.

4 Simulation Studies

We simulated 100 datasets of size n from the model

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \sigma\varepsilon_i,\ i = 1, \ldots, n, \tag{7}$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'$, $\mathbf{x}_i$ are standard normal with the correlation between the jth and kth components of $\mathbf{x}$ equal to $0.5^{|j-k|}$. The errors ε_i are iid following one of standard normal, Laplace, Cauchy, or t_5 distribution. This model has been used elsewhere in the variable selection literature when considered with normal errors ε_i [1, 2]. In our simulation studies, we only consider the Wilcoxon weight function. Define the prediction $\hat{\mathbf{y}}(\lambda) = \mathbf{X}\widehat{\boldsymbol{\beta}}_n(\lambda)$ in the linear model (7). It is easy to see that the prediction error $\mathrm{PE} = \|\mathbf{y} - \hat{\mathbf{y}}(\lambda)\|^2$ can be decomposed: $\mathrm{PE} = \mathrm{ME} + n\sigma^2$, where the model error is $\mathrm{ME} \equiv (\widehat{\boldsymbol{\beta}}_n(\hat{\lambda}) - \boldsymbol{\beta})'E(\mathbf{x}\mathbf{x}')(\widehat{\boldsymbol{\beta}}_n(\hat{\lambda}) - \boldsymbol{\beta})$. Hence, the statistic ME is a summary measure of the variable selection procedure. As in Fan and Li [2], we compare the proposed rank-based variable selection procedures using the median of relative model error (MRME), where the ratio of model error (RME) is defined as the model error of the rank-based penalized estimators over the model error of the full-model rank regression estimator for each Monte Carlo data set. We also compared the average numbers of regression coefficients that are correctly (C) or incorrectly (I) shrunk to 0. The results for normal errors 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 rank-based estimators with the scad and alasso penalty approach that of the oracle estimator as n increases. When the signal-to-noise ratio (e.g. σ/n in our example) is small, scad and alasso methods perform better than lasso in terms of model error and model complexity. However, lasso and elastic net estimators perform better than oracle methods as σ/n increases.

Table 2 reports the results on the accuracy of the proposed sandwich formulae in estimating the variances of the non-zero estimated regression coefficients: here, we only summarize standard errors estimates for $\hat{\beta}_1(\hat{\lambda})$. 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 (SEE) gauges the performance of the sandwich estimator. The median of the absolute deviation of the standard error estimates divided by 0.6745 is reported in parentheses. At $n = 75$, the standard error estimates already match reasonably well the true standard errors of the regression coefficient estimates. When the sample size is increased to

Table 1. Simulation results on model selection with normal errors.

Method	MRME (%)	Avg. No. of 0 coefficients C	I
$n = 50, \sigma = 3$			
scad	63.60	4.65	0.29
hard	78.63	4.76	0.47
lasso	60.62	3.54	0.04
alasso	63.49	4.50	0.13
en	61.90	3.46	0.03
oracle	27.37	5	0
$n = 50, \sigma = 1$			
scad	35.88	4.05	0
hard	42.72	4.70	0
lasso	56.12	3.05	0
alasso	35.70	4.90	0
en	67.08	2.86	0
oracle	27.54	5	0
$n = 65, \sigma = 1$			
scad	40.09	4.11	0
hard	44.27	4.67	0
lasso	64.95	3.03	0
alasso	38.39	4.88	0
en	74.73	2.88	0
oracle	30.53	5	0

Table 2. Simulation results on standard error estimation with normal errors.

	normal		t_5		Laplace		Cauchy	
	SD	SEE(SD)	SD	SEE(SD)	SD	SEE(SD)	SD	SEE(SD)
$n = 75, \sigma = 1$								
scad	0.144	0.127 (0.022)	0.160	0.141 (0.021)	0.134	0.157 (0.032)	0.258	0.268 (0.060)
hard	0.146	0.127 (0.024)	0.158	0.141 (0.023)	0.132	0.156 (0.031)	0.238	0.268 (0.060)
lasso	0.158	0.120 (0.022)	0.153	0.132 (0.020)	0.137	0.145 (0.026)	0.272	0.259 (0.067)
alasso	0.149	0.123 (0.021)	0.157	0.136 (0.021)	0.152	0.150 (0.029)	0.258	0.262 (0.062)
en	0.158	0.119 (0.022)	0.164	0.130 (0.019)	0.130	0.143 (0.027)	0.294	0.258 (0.066)
oracle	0.146	0.138 (0.019)	0.151	0.152 (0.025)	0.137	0.165 (0.023)	0.231	0.268 (0.056)

$n = 100$ and error distribution normal, the difference $(SD - SEE)$ is -0.009, -0.005, 0.007, -0.005, 0.003, and 0.010 for the scad, hard, lasso, alasso, elastic net (en) and oracle estimates, respectively. Results for other error distributions improve similarly.

To illustrate the effects of long-tailed error distributions, we give the median relative model errors comparing least squares to rank estimators for each of five penalty functions in Table 3. Table entries are summarized for $\sigma = 1$ over 100 Monte Carlo data sets. We find that penalized least squares estimators have smaller model error than penalized rank estimators when the true error distribution is normal but the opposite is true for error distributions with heavy tails. Moreover, the relative difference between penalized least squares and rank estimators grows with sample size — for example, 32.2% increase for the scad penalty when the sample size is doubled from $n = 50$ to $n = 100$; the analogous increase for en penalty is 118.1%. This result confirms our intuition that least squares estimators are inferior to rank estimators for heavy-tailed error distributions.

Next, we compare Fan and Li's [2] robust M-estimator to our proposed rank-based estimator. To compare robust variable selection methods, we used the linear model in (7) with $n = 60$ and a variation on Fan and Li's simulation exercise in their Example 4 [2]. Now, assume the errors ε_i are iid from the bivariate mixture distribution:

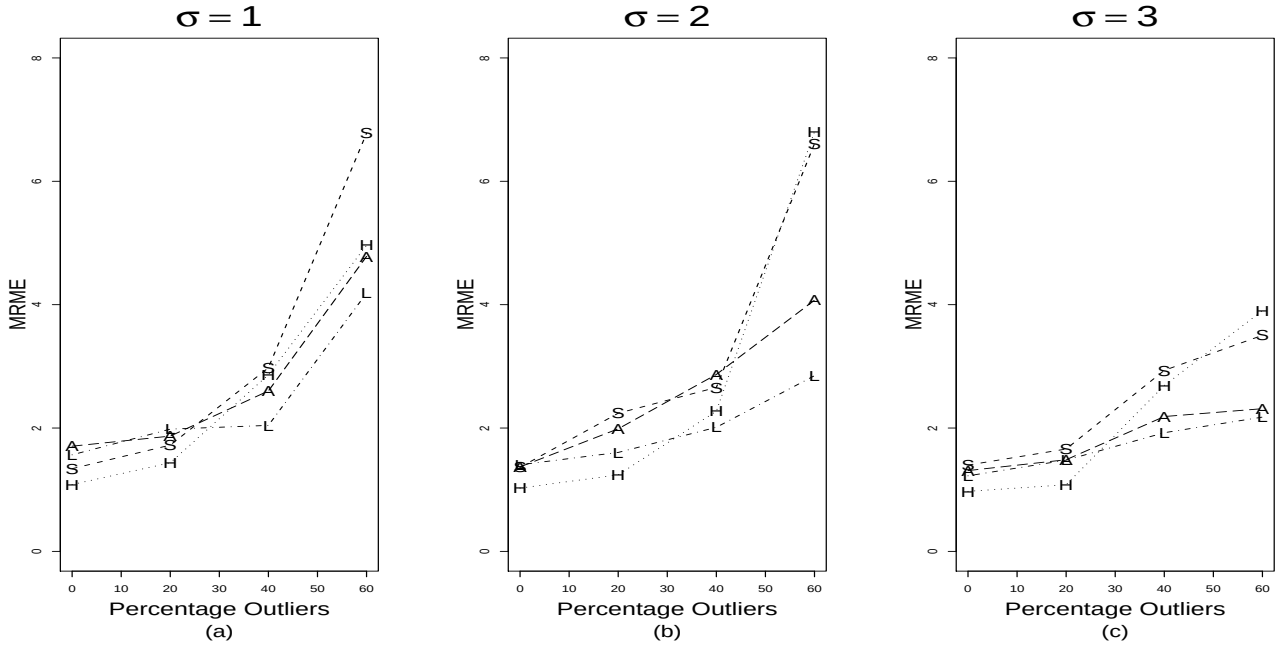
$$F_\varepsilon(t) = p\Phi(t) + (1-p)\left[\frac{1}{\pi}\arctan(t) + \frac{1}{2}\right],$$

where $\Phi(t)$ denotes the standard normal distribution and the expression in square brackets is the standard Cauchy distribution. Small values of the proportion p lead to a contaminated normal distribution and is often useful for studying the effect of outliers. Fan and Li studied the case of 10% outliers, i.e. $p = 0.10$, whereas our simulation exercise lets the mixing proportion p vary from $0 - 0.60$. The statistic of interest

Table 3. Median relative model errors: penalized least squares versus penalized rank regression.

	scad	hard	lasso	alasso	en
n=50					
normal	0.96	0.96	1.02	1.05	0.98
t_5	1.22	1.19	1.14	1.30	1.13
Laplace	1.33	1.19	1.29	1.41	1.22
Cauchy	16.14	17.48	9.99	15.88	7.23
n=75					
normal	0.95	1.03	0.93	1.04	0.96
t_5	1.38	1.37	1.28	1.34	1.27
Laplace	1.29	1.26	1.38	1.47	1.34
Cauchy	19.77	23.76	9.85	14.81	9.62
n=100					
normal	0.90	0.91	0.91	0.97	0.95
t_5	1.20	1.22	1.17	1.22	1.18
Laplace	1.38	1.49	1.24	1.45	1.31
Cauchy	21.34	26.20	15.33	18.51	15.77

Figure 1. Penalized R-estimates versus M-estimates. Median relative model errors (MRME) are displayed as a function of p, where p is the proportion of errors from a standard normal distribution in a bivariate mixture error distribution F_ε. The remaining $n \times (1-p)$ errors follow a standard Cauchy distribution. MRME is displayed for each of lasso (L), adaptive lasso (A), hard (H), and scad (S) penalty and values close to one suggest the rank-based estimator is better.



in our simulation study is again the median relative model error, defined as the median of the ratios $\mathrm{ME}\{\widehat{\boldsymbol{\beta}}(\text{M-estimator})\}/\mathrm{ME}\{\widehat{\boldsymbol{\beta}}(\text{R-estimator})\}$.

Figure 1 displays our simulation results. We display the simulation summary statistic MRME for each of scad (S), hard thresholding (H), lasso (L), and adaptive lasso (A) penalties and across increasing error variance. When the percentage of outlying observations is small, we find that the penalized R- and M-estimates perform similarly, on average. However, when the proportion of outlying observations gets larger, the proposed rank-based variable selection methods tend to yield a model with smaller model error. Moreover, this result seems to hold across penalties with the ℓ_1 and scad penalties yielding the smallest and largest change, respectively. Interestingly, we found that as the error variance decreased, the relative gain of penalized R-estimators over Fan and Li's penalized, robust M-estimator increased.

Table 4. Full model comparisons in the diabetes data: ordinary least squares (OLS), median and rank regression.

		OLS		Median		Rank	
1.	Age	-0.48	(2.84)	0.46	(3.54)	-0.88	(2.94)
2.	Sex	-11.41	(2.91)	-15.60	(3.77)	-12.77	(3.02)
3.	BMI	24.73	(3.16)	22.00	(4.37)	25.07	(3.28)
4.	BP	15.43	(3.11)	19.49	(4.18)	15.99	(3.22)
5.	S1	-37.68	(19.80)	-40.90	(29.19)	-37.96	(20.53)
6.	S2	22.68	(16.11)	20.24	(23.16)	21.88	(16.71)
7.	S3	-4.81	(10.10)	-6.79	(14.40)	-4.51	(10.47)
8.	S4	8.42	(7.67)	12.26	(12.50)	8.52	(7.96)
9.	S5	35.73	(8.17)	36.23	(10.45)	36.91	(8.47)
10.	S6	3.22	(3.13)	2.41	(4.37)	2.40	(3.25)

Table 5. Penalized rank estimates for diabetes data.

	scad	hard	lasso	alasso	en
Age	0 (-)	0 (-)	0 (-)	0 (-)	0 (-)
Sex	-12.77 (2.95)	-12.77 (2.95)	-10.29 (2.50)	-11.87 (2.80)	-10.45 (2.49)
BMI	25.05 (3.32)	25.05 (3.29)	25.11 (3.04)	25.67 (3.26)	26.24 (3.10)
BP	15.99 (3.14)	15.99 (3.15)	14.24 (2.76)	15.69 (3.05)	14.85 (2.78)
S1	-37.81 (20.06)	-37.81 (9.14)	-5.66 (2.45)	-27.04 (6.56)	-5.36 (2.28)
S2	21.78 (15.97)	21.78 (10.02)	0 (-)	13.74 (6.60)	0 (-)
S3	-4.42 (10.05)	0 (-)	10.10 (2.52)	0 (-)	10.60 (2.48)
S4	8.48 (8.02)	8.48 (5.98)	0 (-)	6.08 (3.88)	0 (-)
S5	36.86 (7.97)	36.86 (5.55)	25.92 (3.27)	34.01 (4.20)	26.91 (3.25)
S6	0 (-)	0 (-)	1.25 (1.30)	0 (-)	1.09 (1.01)

5 Analysis of Diabetes Data

Here, we apply our methods to the diabetes data used in Efron et al. [23]. The data set consists of ten predictors — age, sex, body mass index (BMI), blood pressure (BP), and six blood serum measurements, S1-S6 — for each of $n = 442$ patients. The endpoint of interest is a quantitative measure of disease progression after one year of follow-up. The statistical goal is to build a model that includes important prognostic variables of disease progression. In this section, the predictors are scaled to have mean zero and unit variance. Our scale differs from that of Efron et al. [23] in that they scale the predictors to have unit ℓ_2-norm.

We use the Wilcoxon weight function in our analysis of the diabetes data. The estimated coefficients in the full linear model with ten predictors are summarized in Table 4. We compare the ordinary least squares fit to the estimated coefficients of median and rank regression. We found the point estimates to agree, however, the standard error estimates for the estimated rank coefficients were uniformly smaller than their estimated median coefficient counterparts.

We summarize the estimated regression coefficients using our penalized rank regression estimator in Table 5. We find that scad and hard thresholding estimators yield very similar models, the only difference being the variable S3; however, we note the standard error estimates do differ between these two methods. The adaptive lasso model is similar to the hard thresholding model, although the coefficient estimates are generally shrunk closer to zero for the adaptive lasso penalty. The models from the so-called oracle estimators (scad, hard, alasso) may be contrasted with the models resulting from lasso and elastic net penalties. The models from lasso and elastic net include S3 and S6 and exclude S2 and S4 whereas, in general, the oracle models switch these pairs of variables.

6 Remarks

We have developed a robust estimator for selecting variables in the linear model. Our estimator extends the usual rank regression estimator [12, 13] by minimizing an objective function defined as the sum of an appropriate penalty term and the usual dispersion function $D_n(\boldsymbol{\beta})$. The resulting penalized rank estimators simultaneously select variables and estimate their regression coefficients, a feature which allows one to study the operating characteristics under certain regularity conditions. Like robust M-estimators, our

penalized rank-based estimator is robust to outliers and heavy-tailed or asymmetric error distributions. Hence, rank-based variable selection methods proposed here offer investigators another robust method for variable selection in the linear model and can be implemented using standard software.

Appendix A: Proof of Theorem 2.1

We define the partitioned score vector $\mathbf{U}_n(\boldsymbol{\beta}) = (\mathbf{U}_{n,1}(\boldsymbol{\beta})', \mathbf{U}_{n,2}(\boldsymbol{\beta})')'$ similar to that of $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2')'$. To prove part (a), by the continuity of $D_n(\boldsymbol{\beta})$, it suffices to show that for any given $\epsilon > 0$, there exists a large constant M such that

$$\Pr\left\{\inf_{\|\mathbf{u}\|=M} W_n(\mathbf{u}) > 0\right\} \geq 1 - \epsilon, \tag{A1}$$

where $W_n(\mathbf{u}) = Q_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - Q_n(\boldsymbol{\beta}_0)$. This implies the existence of a local minimizer in the M-ball around $\boldsymbol{\beta}_0$ and thus completes the proof of part (a).

By condition (i), using the quadratic approximation of $D_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u})$ and the fact that $p_{\lambda_n}(0) = 0$ and $n^{-1/2}\mathbf{U}_n(\boldsymbol{\beta}_0) = O_p(1)$, we get

$$\begin{aligned}
W_n(\mathbf{u}) &\approx -\{n^{-1/2}\mathbf{U}_n(\boldsymbol{\beta}_0)\}'\mathbf{u} + \frac{1}{2}(\mathbf{u}'\mathbf{A}\mathbf{u}) + n\sum_{j=1}^{d}\{p_{\lambda_n}(|\beta_j + n^{-1/2}u_j| - p_{\lambda_n}(|\beta_j|)\} \\
&\geq -\{n^{-1/2}\mathbf{U}_n(\boldsymbol{\beta}_0)\}'\mathbf{u} + \frac{1}{2}(\mathbf{u}'\mathbf{A}\mathbf{u}) + n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_j + n^{-1/2}u_j| - p_{\lambda_n}(|\beta_j|)\} \\
&= -\{n^{-1/2}\mathbf{U}_n(\boldsymbol{\beta}_0)\}'\mathbf{u} + \frac{1}{2}(\mathbf{u}'\mathbf{A}\mathbf{u}) + n^{1/2}\sum_{j=1}^{s} q_{\lambda_n}(|\beta_{0j}|)u_j + \sum_{j=1}^{s}\dot{q}_{\lambda_n}(|\beta_{0j}|)u_j^2\{1 + o(1)\} \\
&\approx -O_p(1)\sum_{j=1}^{d}|u_j| + \frac{1}{2}(\mathbf{u}'\mathbf{A}\mathbf{u}) \geq -O_p(1)\sum_{j=1}^{d}|u_j| + \frac{1}{2}a_0\|\mathbf{u}\|^2
\end{aligned}$$

where $\approx$ represents the asymptotic equivalence uniformly in $\mathbf{u} \in \{\mathbf{u} : \ \|\mathbf{u}\| \leq M\}$ and a_0 is the smallest eigenvalue of $\mathbf{A}$. Since $\mathbf{A}$ is positive definite, $a_0 > 0$. Therefore (A1) holds by choosing a sufficiently large M.

We now prove part (b). It is sufficient to show that with probability tending to 1, for any $M > 0$ and for each $\boldsymbol{\beta}^*$ satisfying $\|\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_{01}\| = O_p(n^{-1/2})$ and $\beta_j^* \in (-Mn^{-1/2}, Mn^{-1/2})$ with $j = s+1, \ldots, d$, $\partial Q_n(\boldsymbol{\beta})/\partial\beta_j|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ and β_j^* have the same sign. By the asymptotic linearity of $\mathbf{U}_n(\boldsymbol{\beta})$, we have

$$\begin{aligned}
n^{-1/2}\frac{\partial Q_n(\boldsymbol{\beta})}{\partial\beta_j}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} &= -n^{-1/2}U_{n,j}(\boldsymbol{\beta}_0) + \mathbf{A}_{(j)}\{n^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)\} + n^{1/2}q_{\lambda_n}(|\beta_j^*|)\mathrm{sgn}(\beta_j^*) + o(1) \\
&= O_p(1) + \left(n^{1/2}\lambda_n\right)\cdot\{q_{\lambda_n}(|\beta_j^*|)\mathrm{sgn}(\beta_j^*)/\lambda_n\}, \quad j = s+1, \ldots, d,
\end{aligned}$$

where $\mathbf{A}_{(j)}$ denotes the jth row of $\mathbf{A}$. It follows from conditions (ii) and (iii) that

$$\left(n^{1/2}\lambda_n\right)\cdot\{q_{\lambda_n}(|\beta_j^*|)/\lambda_n\} \to \infty.$$

This then implies that the sign of $\frac{\partial Q_n(\boldsymbol{\beta})}{\partial\beta_j}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ is completely determined by that of β_j^* as n is large enough.

By the definition of $\widehat{\boldsymbol{\beta}}_n$, we see from previous arguments that $n^{-1/2}\frac{\partial Q_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_1}|_{\boldsymbol{\beta}=(\widehat{\boldsymbol{\beta}}_1', \mathbf{0}')'} = o_p(1)$. It follows

from the asymptotic linearity of $\mathbf{U}_n(\boldsymbol{\beta})$ that

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

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

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

Appendix B: Proof of Theorem 2.2

The proof of Theorem 2 bears similarity with that of Theorem 1. The major distinctions are the use of the convexity of $D_n(\boldsymbol{\beta})$ and the way of controlling the asymptotic order of the adaptive lasso penalty.

To show part (a), we first obtain that $(|\tilde{\boldsymbol{\beta}}_j|)^{-1} = (|\boldsymbol{\beta}_{0j}|)^{-1} + n^{-1/2}O_p(1)$ because $\|\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$. Because $n^{1/2}\lambda_n = O_p(1)$, straightforward manipulations similar to that in the proof of Theorem 1 give

$$\begin{aligned}
W_n(\mathbf{u}) &\approx -\{n^{-1/2}\mathbf{U}_n(\boldsymbol{\beta}_0)\}'\mathbf{u} + \frac{1}{2}(\mathbf{u}'\mathbf{A}\mathbf{u}) + n\lambda_n\sum_{j=1}^{d}\left(\frac{|\beta_{0j}+n^{-1/2}u_j|}{|\tilde{\beta}_j|} - \frac{|\beta_{0j}|}{|\tilde{\beta}_j|}\right) \\
&\geq -O_p(1)\sum_{j=1}^{d}|u_j| + \frac{1}{2}(\mathbf{u}'\mathbf{A}\mathbf{u}) - n^{1/2}\lambda_n\sum_{j=1}^{s}|u_j|/|\tilde{\beta}_j| \\
&\approx -O_p(1)\sum_{j=1}^{d}|u_j| + \frac{1}{2}(\mathbf{u}'\mathbf{A}\mathbf{u}) - O_p(1)\sum_{j=1}^{s}|u_j|/|\beta_{0j}| \\
&\geq -O_p(1)M + \frac{1}{2}a_0\|\mathbf{u}\|^2 - O_p(1)\sum_{j=1}^{s}|u_j|/|\beta_{0j}|.
\end{aligned}$$

Adopting the similar arguments for Theorem 1, we have $W_n(\mathbf{u}) > 0$ for all $\|\mathbf{u}\| = M$ when M is chosen to be sufficiently large. By the convexity of $D_n(\boldsymbol{\beta})$, it follows that $W_n(\mathbf{u}) > 0$ for all $\|\mathbf{u}\| \geq M$. This indicates that all global minimizers of $Q_n(\boldsymbol{\beta})$ must lie in the M-ball around $\boldsymbol{\beta}_0$. Therefore, the adaptive lasso rank estimator $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.

To show part (b), we note from the asymptotic linearity of $\mathbf{U}_n(\boldsymbol{\beta})$ that

$$n^{-1/2}\frac{\partial Q_n(\boldsymbol{\beta})}{\partial\beta_j}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = -n^{-1/2}U_{n,j}(\boldsymbol{\beta}_0) + \mathbf{A}_{(j)}\{n^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)\} + n\lambda_n\frac{\mathrm{sgn}(\beta_j^*)}{|n^{1/2}\tilde{\beta}_j|} + o(1),$$

where $\boldsymbol{\beta}^*$ is the same as that defined in the proof of Theorem 1. Because $n^{1/2}\mathbf{U}_n(\boldsymbol{\beta}_0) = O_p(1)$ and $n^{1/2}|\tilde{\beta}_j| = O_p(1)$ for $j = s+1, \ldots, d$, we have

$$n^{-1/2}\frac{\partial Q_n(\boldsymbol{\beta})}{\partial\beta_j}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = O_p(1) + n\lambda_n\frac{\mathrm{sgn}(\beta_j^*)}{O_p(1)}, \quad j = s+1, \ldots, d.$$

Because $n\lambda_n \to \infty$, the sign of $\partial Q_n(\boldsymbol{\beta})/\partial\beta_j|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ is the same as that of β_j^* $(j = s+1, \ldots, d)$ as n is large

enough. This implies that $\Pr(\widehat{\boldsymbol{\beta}}_2 = 0) \to 1$. Coupled with the definition of $\widehat{\boldsymbol{\beta}}_n$, this also implies that

$$\begin{aligned} o_p(1) &= n^{-1/2} \frac{\partial Q_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_1}|_{\boldsymbol{\beta}=(\widehat{\boldsymbol{\beta}}_1', \mathbf{0}')'} \\ &= -n^{-1/2}\mathbf{U}_{n,1}(\begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ 0 \end{pmatrix}) + n^{1/2}\lambda_n \left(\frac{\mathrm{sgn}(\hat{\beta}_1)}{|\tilde{\beta}_1|}, \ldots, \frac{\mathrm{sgn}(\hat{\beta}_s)}{|\tilde{\beta}_s|} \right)' \\ &= -n^{-1/2}\mathbf{U}_{n,1}(\boldsymbol{\beta}_0) + \mathbf{A}_{11} n^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \beta_{01}) + n^{1/2}\lambda_n \left(\frac{\mathrm{sgn}(\beta_{01})}{|\tilde{\beta}_1^o|}, \ldots, \frac{\mathrm{sgn}(\beta_{0s})}{|\tilde{\beta}_s|} \right)' + o_p(1). \end{aligned}$$

Because $n^{1/2}\lambda_n \to \lambda_0$ and $\tilde{\beta}_j \to_p \beta_{0j} \neq 0$ for $j = 1, \ldots, s$, $n^{1/2}\mathbf{A}_{11}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01} + n^{-1/2}\mathbf{A}_{11}^{-1}\lambda_0 \mathbf{b}_1) = n^{-1/2}\mathbf{U}_{n,1}(\boldsymbol{\beta}_0) + o_p(1)$. Given that $n^{-1/2}\mathbf{U}_{n,1}(\boldsymbol{\beta}_0) \to_d N(0, V_{11})$, applying the Slutsky's Theorem completes the proof of part (b).

References

[1] Tibshirani, R. J., 1996, Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**, 267–288.
[2] Fan, J. & Li, R., 2001, Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.*, **96**, 1348–1360.
[3] Zou, H. & Hastie, T., 2005, Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, **67**, 301–320.
[4] Zou, H., 2006, The adaptive lasso and its oracle properties. *J. Am. Statist. Assoc.*, **101**, 1418–1429.
[5] Huber, P.J., 1981, *Robust Statistics* Wiley, New York.
[6] Chang, W. H., McKean, J. W., Naranjo, J. D., and Sheather, S. J., 1999, High-breakdown rank regression. *J. Am. Statist. Assoc.*, **94**, 205–219.
[7] Prentice, R. L., 1978, Linear rank tests with right-censored data, *Biometrika*, **65**, 167-179.
[8] Tsiatis, A. A., 1990, Estimating regression parameters using linear rank tests for censored data, *Ann. Statist.*, **18**, 354–372.
[9] Johnson, B. A., 2007, Variable selection in semiparametric linear regression with censored data, *J. R. Statist. Soc. B*, (In press).
[10] Jaeckel, L.A., 1972, Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.*, **43**, 1449–1458.
[11] Frank, I. E. & Friedman, J. H., 1993, A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148.
[12] Hettsmansperger, T. P. & McKean, J. W., 1977, A robust alternative based on ranks to least squares in analyzing linear models. *Technometrics*, **19**, 275–84.
[13] McKean, J. W. & Hettsmansperger, T. P., 1978, A robust analysis of the general linear model based on one step *R*-estimates. *Biometrika*, **65**, 571–579.
[14] Jurečková, J., 1969, Asymptotic linearity of a rank statistic in regression parameter. *Ann. Math. Statist.*, **40**, 1889–1900.
[15] Jurečková, J., 1971, Nonparametric estimate of regression coefficients. *Ann. Math. Statist.*, **42**, 109–148.
[16] Heiler, S., and Willers, R., 1988, Asymptotic normality of *R*-estimation in the linear model. *Statistics*, **19**, 173–184.
[17] Koul, H. L., Sievers, G. L. & McKean, J. W., 1987, An estimator of the scale parameter for the rank analysis of linear models under general score functions. *Scand. J. Statist.*, **14**, 131–141.
[18] Donoho, D. L. & Johnstone, I. M., 1994, Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**, 425–455.
[19] Antoniadis, A., 1997, Wavelets in statistics: A review (with discussion). *J. It. Statist. Assoc.*, **6**, 97–144.
[20] Hájek, J., Šidák, Z. & Sen, P. K., *Theory of Rank Tests*. San Diego: Academic Press, 1999.
[21] Knight, K. & Fu, W., 2000, Asymptotics for lasso-type estimators. *Ann. Statist.*, **28**, 1356–1378.
[22] Hunter, D. R. & Li, R., 2005, Variable selection using MM algorithms. *Ann. Statist.*, **33**, 1617–1642.
[23] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, T., 2004, Least angle regression *Ann. Statist.*, **32**, 407–499.