Chapter 2
Robust Signed-Rank Variable Selection in Linear Regression

Asheber Abebe and Huybrechts F. Bindele

Abstract The growing need for dealing with big data has made it necessary to find computationally efficient methods for identifying important factors to be considered in statistical modeling. In the linear model, the Lasso is an effective way of selecting variables using penalized regression. It has spawned substantial research in the area of variable selection for models that depend on a linear combination of predictors. However, work addressing the lack of optimality of variable selection when the model errors are not Gaussian and/or when the data contain gross outliers is scarce. We propose the weighted signed-rank Lasso as a robust and efficient alternative to least absolute deviations and least squares Lasso. The approach is appealing for use with big data since one can use data augmentation to perform the estimation as a single weighted L_1 optimization problem. Selection and estimation consistency are theoretically established and evaluated via simulation studies. The results confirm the optimality of the rank-based approach for data with heavy-tailed and contaminated errors or data containing high-leverage points.

Keywords Adaptive Lasso • Wilcoxon estimation • Oracle property • Penalized least squares • LAD regression

2.1 Introduction

The growing need for dealing with 'big data' has made it necessary to find ways of determining the few important factors to consider in the statistical modeling. In the linear and generalized linear models, this translates to identifying the covariates

A. Abebe (✉)
Department of Mathematics and Statistics, Auburn University, 221 Parker Hall, Auburn, AL 36849, USA
e-mail: ash@auburn.edu

H.F. Bindele
Department of Mathematics and Statistics, University of South Alabama, 411 University Blvd. N., ILB 316, Mobile, AL 36688-0002, USA
e-mail: hbindele@southalabama.edu

R.Y. Liu, J.W. McKean (eds.), *Robust Rank-Based and Nonparametric Methods*, Springer Proceedings in Mathematics & Statistics 168,
DOI 10.1007/978-3-319-39065-9_2

that are most needed in the prediction of the outcome. In this regard, the Lasso method introduced in Tibshirani (1996) has garnered significant attention in the past two decades. The Lasso method takes advantage of the singularity of the L_1 penalty to effectively select variables via the penalized least squares procedure. This work has been refined and extended in various directions. See, for example, Fan and Li (2001), Zou and Hastie (2005), Zou (2006), Wang and Leng (2008), and references therein. Much of the focus has been in establishing the so-called "Oracle" property Fan and Li (2001) that consists of selection consistency and estimation efficiency. These are both asymptotic properties where selection consistency refers to ones ability to correctly identify the zero regression coefficients while estimation efficiency refers to ones ability to provide a $\sqrt{n}$-consistent estimator of the non-zero coefficients.

However, there are not too many results that address the lack of optimality of these variable selection procedures when the model errors are not Gaussian and/or when the data contain gross outliers. An approach based on penalized Jaeckel-type rank-regression was discussed in Johnson and Peng (2008), Johnson et al. (2008), Johnson (2009), Leng (2010) and Xu et al. (2010). The computation is complicated and, as in unpenalized rank-regression, the approach used in these papers will only result in robustness in the response space. For variable selection, however, getting a handle on leverage is crucial. One paper that discussed this issue and tried to address the influence of high leverage points is Wang and Li (2009), where they considered penalized weighted Wilcoxon estimation. Our proposed approach based on minimization of a penalized weighted signed-rank norm is much simpler to compute and provides protection against outliers and high-leverage points. It also allows one flexibility through choice of score generating functions. One limitation of our proposed approach is that it requires symmetry of the error density. In this case, the estimates are equivalent to Jaeckel-type rank-regression estimates.

Consider the linear regression model given by

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_0 + e_i, \qquad 1 \le i \le n, \tag{2.1}$$

where $\boldsymbol{\beta}_0 \in \mathscr{B} \subset \mathbb{R}^d$ is a vector of parameters, $\boldsymbol{x}_i$ is a vector of independent variables in a vector space $\mathbb{X}$, and the errors e_i are assumed to be i.i.d. with a distribution function F. Let $\mathbf{V}_n = \{(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)\}$ be the set of sample data points. Note that $\mathbf{V}_n \subset \mathbb{V} \equiv \mathbb{R} \times \mathbb{X}$. We shall assume that $\mathscr{B}$ is a compact subspace of $\mathbb{R}^d$, $\boldsymbol{\beta}_0$ is an interior point of $\mathscr{B}$.

Rank-based approaches have been shown to possess a high breakdown property resulting on robust and efficient estimators. The rank-based approach considered in this paper is based on the so-called the weighted signed-rank (WSR) norm proposed in Bindele and Abebe (2012) for estimation of coefficients of general nonlinear models. Here we consider WSR with added penalty for simultaneous estimation and variable selection in linear models. That is, we obtained an estimator $\hat{\boldsymbol{\beta}}_n$ of $\boldsymbol{\beta}_0$ satisfying

$$\hat{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta} \in \mathscr{B}}{\operatorname{Argmin}}\, Q(\boldsymbol{\beta}), \tag{2.2}$$

where $Q(\boldsymbol{\beta})$ is a penalized WSR objective function

$$Q(\boldsymbol{\beta}) = D_n(\mathbf{V}_n, w, \boldsymbol{\beta}) + n\sum_{j=1}^{d} P_{\lambda_j}(|\beta_j|). \tag{2.3}$$

and $D_n(\mathbf{V}_n, w, \boldsymbol{\beta})$ is the WSR dispersion function defined by

$$D_n(\mathbf{V}_n, w, \boldsymbol{\beta}) = \sum_{i=1}^{n} w(\boldsymbol{x}_i) a_n(i) |z(\boldsymbol{\beta})|_{(i)}. \tag{2.4}$$

Here $z_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, $|z(\boldsymbol{\beta})|_{(i)}$ is the ith ordered value among $|z_1(\boldsymbol{\beta})|, \ldots, |z_n(\boldsymbol{\beta})|$, and the numbers $a_n(i)$ are scores generated as $a_n(i) = \varphi^+(i/(n+1))$, for some bounded and non-decreasing score function $\varphi^+ : (0,1) \to \mathbb{R}^+$ that has at most a finite number of discontinuities. The function $w : \mathbb{X} \to \mathbb{R}^+$ is a continuous weight function. The penalty function $P_{\lambda_j}(\cdot)$ is defined on $\mathbb{R}^+$. When the penalty function is the Lasso penalty Tibshirani (1996) $P_{\lambda_j}(|t|) = \lambda|t|$ for all j, we will refer to the resulting estimator as the WSR-Lasso (WSR-L), and when the penalty function is the adaptive Lasso Zou (2006) $P_{\lambda_j}(|t|) = \lambda_j|t|$, we will refer to the estimator as WSR-Adaptive Lasso (WRS-AL) estimator. We should point out that for $\varphi^+ \equiv 1$, the objective function in (2.3) reduces to the WLAD-Lasso discussed in Arslan (2012). If additionally $w \equiv 1$, then we obtain the LAD-lasso discussed in Wang et al. (2007). While these LAD based estimators are easy to compute and provide robust estimators, they lack efficiency especially when the error density at zero is small (Hettmansperger and McKean 2011; Leng 2010). Note that, while not stressed in our notation, $\hat{\boldsymbol{\beta}}_n$ depends on the tuning parameter $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)'$.

Using the same idea in Wang et al. (2007), either under WSR-L or WSR-AL, one can write $Q(\boldsymbol{\beta})$ as

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n+d} v_i |z_i^*(\boldsymbol{\beta})|, \tag{2.5}$$

where $z_i^*(\boldsymbol{\beta}) = y_i^* - \boldsymbol{x}_i^{*\prime}\boldsymbol{\beta}$ with

$$(y_i^*, \boldsymbol{x}_i^*)' = \begin{cases} (y_i, \boldsymbol{x}_i)', & \text{for } 1 \le i \le n, \\ (0, n\lambda_i \mathbf{e}_i)', & \text{for } n+1 \le i \le n+d. \end{cases} \tag{2.6}$$

and

$$v_i = \begin{cases} w(\boldsymbol{x}_i)\varphi^+\left(\frac{R(z_i(\boldsymbol{\beta}))}{n+1}\right), & \text{for } i \le n, \\ 1, & \text{for } i > n. \end{cases}$$

Here $\mathbf{e}_i$ is the d-dimensional vector with ith component equal to 1 and all the others equal to 0. To this end, Eq. (2.5) can be seen as the weighted L_1 objective function. In Eq. (2.6) the WSR-L objective function is obtained by putting $\lambda_i = \lambda$ for all i. To avoid any possible confusion, we will use $Q^w_\ell(\cdot)$ and $Q^w_{a\ell}(\cdot)$ for WSR-L and WSR-AL objective functions, respectively.

Remark 2.1. Considering the unpenalized objective function $D_n(\mathbf{V}_n, w, \boldsymbol{\beta})$ defined in Eq. (2.4), asymptotic properties (consistency and $\sqrt{n}$-asymptotic normality) of the WSR estimator with $w \equiv 1$ were established under mild regularity conditions in Hössjer (1994). Considering the weighted case, analogous asymptotic results were obtained by Bindele and Abebe (2012) for general nonlinear regression model.

2.2 Asymptotics

In this section, we provide the asymptotic properties of the WSR-AL estimator defined in (2.2) under regularity conditions. Consider the following assumptions

(I_1) $P\big(\boldsymbol{x}'\boldsymbol{\beta} = \boldsymbol{x}'\boldsymbol{\beta}_0\big) < \alpha$ for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, $0 < \alpha \leq 1$, and $E_G[|\boldsymbol{x}|^r] < \infty$ for some $r > 1$, G being the distribution of $\boldsymbol{x}$.

(I_2) The density f of ε is symmetric about zero, strictly decreasing on $\mathbb{R}^+$, and absolutely continuous with finite Fisher information. Its derivative f' is bounded and $E_F(|\varepsilon|^r) < \infty$ for some $r > 1$.

These two assumptions ensure the strong consistency of $\tilde{\boldsymbol{\beta}}_n = \text{Argmin}_{\boldsymbol{\beta}}\, D_n(\mathbf{V}_n, w, \boldsymbol{\beta})$.

2.2.1 Consistency and Asymptotic Normality

We shall assume that $p_0 \leq d$ of the true regression parameters are nonzero. Thus, without loss of generality, we assume $\beta_{0j} \neq 0$ for $j \leq p_0$ and $\beta_{0j} = 0$ for $j > p_0$. Thus $\boldsymbol{\beta}_0$ can be partitioned as $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}'_{0a}, \boldsymbol{\beta}'_{0b})'$ with $\boldsymbol{\beta}_{0b} = \mathbf{0}$. Also, $\hat{\boldsymbol{\beta}}_n$ can be similarly partitioned as $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}'_{na}, \hat{\boldsymbol{\beta}}'_{nb})'$ with $\hat{\boldsymbol{\beta}}_{na} = (\hat{\beta}_{n,1}, \ldots, \hat{\beta}_{n,p_0})'$, and $\hat{\boldsymbol{\beta}}_{nb} = (\hat{\beta}_{n,p_0+1}, \ldots, \hat{\beta}_{n,d})'$.

Following Johnson and Peng (2008), we define

$$H_{\lambda_j}(|t|)sgn(t) = \frac{d}{dt}P_{\lambda_j}(|t|) \quad \textit{and} \quad \dot{H}_{\lambda_j}(|t|)sgn(t) = \frac{d}{dt}H_{\lambda_j}(|t|).$$

Also, under Eq. (2.5), taking the negative gradient with respect to $\boldsymbol{\beta}$, we obtain

$$S(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \sum_{i=1}^{n+d} v_i \boldsymbol{x}_i sgn(z_i^*(\boldsymbol{\beta})) = S_n(\boldsymbol{\beta}) + n\sum_{j=1}^{d} H_{\lambda_j}(|\beta_j|)sgn(\beta_j),$$

where $S_n(\boldsymbol{\beta}) = -\nabla_{\boldsymbol{\beta}} D_n(\mathbf{V}_n, w, \boldsymbol{\beta})$. In addition to $(I_1) - (I_2)$, we will need the following assumption:

(I_3) Define $a_n = \max_{1\le j\le p_0} H_{\lambda_j}(|t|)$ and $b_n = \min_{j>p_0} H_{\lambda_j}(|t|)$, $\forall\ t$ fixed, and assume that

(i) $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$ as $n \to \infty$
(ii) $\lim_{n\to\infty} \inf_{|t|\le c/\sqrt{n}} \{\lambda_n^{-1} H_{\lambda_j}(|t|)\} > 0$ for any $c > 0$.

Remark 2.2. Note that for the adaptive Lasso case where $P_{\lambda_j}(|t|) = \lambda_j|t|$, and in assumption (I_3), a_n and b_n are reduced to $a_n = \max_{1\le j\le p_0} \lambda_j$ and $b_n = \min_{p_0+1\le j\le d} \lambda_j$, as $H_{\lambda_j}(|t|) = \lambda_j$. It is worth pointing out the Lasso penalty does not satisfy assumption (I_3) which is not surprising as it is well-known that the Lasso estimator does not have the oracle property, and (I_3) is key to ensuring the oracle property of the resulting estimator.

Theorem 2.1. *Under assumptions* $(I_1) - (I_3)$, $\hat{\boldsymbol{\beta}}_n$ *exists and is a* $\sqrt{n}$*-consistent estimator of* $\boldsymbol{\beta}_0$.

The proof this theorem is provided in Appendix.

Next consider the following assumption commonly imposed in the framework of signed-rank estimation, see Hössjer (1994) and Abebe et al. (2012):

(I_4) $\varphi^+ \in C^2((0,1) \setminus E)$ with bounded derivatives, where E is a finite set of discontinuities.

Following Hössjer (1994), set

$$\gamma_{\varphi^+} = \int_0^1 \left(\varphi^+(t)\right)^2 dt \quad \text{and} \quad \zeta_{\varphi^+} = \int_0^1 \varphi^+(t) h_F(t) dt = -\int_{-\infty}^{\infty} \varphi^+(F^{-1}(u)) f'(u) du,$$

where $h_F(u) = -f'(F^{-1}(u))/f(F^{-1}(u))$. As it is pointed out in Hössjer (1994), (I_1) and (I_2) imply that $\zeta_{\varphi^+} > 0$. Also, letting J denote the joint distribution of $(y, \boldsymbol{x})$ and by symmetry of f, one can define a corresponding symmetric distribution as follows:

$$\begin{aligned} H_{\boldsymbol{\beta}}(t) &= \frac{1}{2}\left[P_J(z_i(\boldsymbol{\beta}) \le t) + P_J(-z_i(\boldsymbol{\beta}) \le t)\right] \\ &= \frac{1}{2}\left[E_G\{F(t) + \boldsymbol{x}^\tau\boldsymbol{\beta}\} + E_G\{F(t - \boldsymbol{x}^\tau\boldsymbol{\beta})\}\right]. \end{aligned} \tag{2.7}$$

Now setting $F_{\boldsymbol{\beta},i}(t) = \frac{1}{2}E_G\{\boldsymbol{x}_i F(t + \boldsymbol{x}^\tau\boldsymbol{\beta})\}$ and $\boldsymbol{\xi}(\boldsymbol{\beta}) = (\xi_1(\boldsymbol{\beta}), \ldots, \xi_n(\boldsymbol{\beta}))^\tau$, where

$$\xi_i(\boldsymbol{\beta}) = 2\int_{-\infty}^{\infty} \varphi^+(H_{\boldsymbol{\beta}}(t)) dF_{\boldsymbol{\beta},i}(t),$$

it is shown under $(I_1) - (I_3)$ in Hössjer (1994) that $S_n(\boldsymbol{\beta}) - \boldsymbol{\xi}(\boldsymbol{\beta}) \to 0 \quad a.s.$ as $n \to \infty$. Let $W(\boldsymbol{x}) = diag\{w_1(\boldsymbol{x}), \ldots, w_n(\boldsymbol{x})\}$ and define the expected weighted Gram matrix $\Sigma = E_G[\boldsymbol{x}'W(\boldsymbol{x})\boldsymbol{x}]$. Now partition $\boldsymbol{x}$ as $\boldsymbol{x} = (\boldsymbol{x}_a, \boldsymbol{x}_b)$, according to nonzero and zero coefficients, and let Σ_a denote the top left $p_0 \times p_0$ sub-matrix of Σ. We will assume that Σ_a is positive definite. The following main result gives the asymptotic properties (oracle property) of the penalized WSR estimator given in (2.2). Its proof is provided in Appendix.

Theorem 2.2. *Under assumptions* (I_1) *to* (I_4)*, we have* $\lim_{n\to\infty} P(\hat{\boldsymbol{\beta}}_{nb} = \mathbf{0}) = 1$*, and*

$$\sqrt{n}\big(\hat{\boldsymbol{\beta}}_{na} - \boldsymbol{\beta}_{0a}\big) \xrightarrow{\mathscr{D}} N\big(0,\ \zeta_{\varphi+}^{-2}\gamma_{\varphi+}\Sigma_a\big),$$

where Σ_a *is a* $p_0 \times p_0$ *positive definite matrix.*

Remark 2.3. From the two theorems above, (i) and (ii) in assumption (I_3) together with (I_1), I_2 and (I_4) are imposed to ensure the $\sqrt{n}$-consistency, the oracle property and the $\sqrt{n}$-asymptotic normality of the proposed estimator. Note that although Theorem 2.2 is similar to that of Johnson and Peng (2008), the definitions of a_n and b_n given here are more general and the assumptions needed for the asymptotic normality of the gradient function $S_n(\boldsymbol{\beta})$ are very different.

2.3 Some Practical Considerations

2.3.1 Estimation of the Tuning Parameter $\boldsymbol{\lambda}$

Another important issue in the estimation of $\boldsymbol{\beta}_0$ in model (2.1), is the choice of the λ_j's in Eq. (2.3). As proposed by Johnson et al. (2008) $\boldsymbol{\lambda}$ can be estimated as follows

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\text{Argmin}} \frac{D_n(\mathbf{V}_n, w, \hat{\boldsymbol{\beta}}_n(\boldsymbol{\lambda}))/n}{\{1 - e(\boldsymbol{\lambda})\}^2}, \tag{2.8}$$

where $e(\boldsymbol{\lambda}) = tr\big[\mathbf{X}\{\mathbf{X}'\mathbf{X} + \Sigma_{\boldsymbol{\lambda},\hat{\boldsymbol{\beta}}_n(\boldsymbol{\lambda})}\}^{-1}\mathbf{X}'\big]$ and $\mathbf{X}$ is the $n \times d$ matrix with column vectors $\boldsymbol{x}_i$ and $\Sigma_{\boldsymbol{\lambda},\hat{\boldsymbol{\beta}}_n(\boldsymbol{\lambda})}$ a diagonal matrix with entries

$$H_{\lambda_j}(|\hat{\beta}_{nj}(\lambda)|)\text{sgn}(\hat{\beta}_{nj}(\lambda)).$$

This cross validation procedure was considered by Johnson et al. (2008) and was shown to have advantage over the least squares cross valuation criterion that is obtained by replacing the numerator of the right hand side of Eq. (2.8) by the least squares objective function. Note that although the idea similar, the objective function $D_n(\mathbf{V}_n, w, \boldsymbol{\beta})$ considered in this paper is very different to the one considered in Johnson et al. (2008). If we restrict ourselves to WSR-AL, another alternative to

estimating $\boldsymbol{\lambda}$ is to consider the AIC and BIC approaches discussed in Wang et al. (2007) based on the considered objective function. That is, obtain $\hat{\boldsymbol{\lambda}}$ as,

$$\hat{\boldsymbol{\lambda}} = \underset{\lambda}{\text{Argmin}} \left\{ Q^w_{a\ell}(\tilde{\boldsymbol{\beta}}_n) - \sum_{j=1}^{d} \log(n\lambda_j) \right\} \text{ the for AIC approach,} \tag{2.9}$$

which leads to $\hat{\lambda}_j = 1/(n|\tilde{\beta}_{nj}|)$, and

$$\hat{\boldsymbol{\lambda}} = \underset{\lambda}{\text{Argmin}} \left\{ Q^w_{a\ell}(\tilde{\boldsymbol{\beta}}_n) - \sum_{j=1}^{d} \log(n\lambda_j) \log n \right\} \text{ the for BIC approach,} \tag{2.10}$$

which leads to $\hat{\lambda}_j = \log n/(n|\tilde{\beta}_{nj}|)$, where $\tilde{\boldsymbol{\beta}}_n = \underset{\boldsymbol{\beta}\in\mathscr{B}}{\text{Argmin}}\, D_n(\mathbf{V}_n, w, \boldsymbol{\beta})$.

2.3.2 Choice of Weights

In our analysis, we choose the weight function $w(\boldsymbol{x})$ to be

$$w(\boldsymbol{x}) = \min\left[1, \frac{\eta}{d(\boldsymbol{x})}\right],$$

where $d(\boldsymbol{x}) = (\boldsymbol{x}-_{\boldsymbol{x}})'\mathbf{C}_{\boldsymbol{x}}^{-1}(\boldsymbol{x}-_{\boldsymbol{x}})$ is a robust Mahalanobis distance, with $_{\boldsymbol{x}}$ and $\mathbf{C}_{\boldsymbol{x}}$ being robust estimates of location and covariance of $\boldsymbol{x}$, respectively and η being some positive constant usually set at $\chi^2_{0.95}$ in practice. Under this choice, it is shown in Bindele and Abebe (2012) that the resulting estimator has a bounded influence function.

2.3.3 Computational Algorithm

For computation purposes, the following steps can be followed:

1. Obtain the unpenalized (W)SR estimator $\hat{\boldsymbol{\beta}}_{\varphi+}$.
2. Use $\hat{\boldsymbol{\beta}}_{\varphi+}$.
   - Estimate $\hat{v}_i$ as $v_i(\hat{\boldsymbol{\beta}}_{\varphi+})$.
   - Use AIC/BIC in Eq. (2.9) or (2.10) with $\tilde{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}_{\varphi+}$ to estimate $\boldsymbol{\lambda}$, say $\hat{\boldsymbol{\lambda}}$.
3. Form $z^*(\boldsymbol{\beta}, \hat{\boldsymbol{\lambda}}) = y^* - \boldsymbol{x}^{*\prime}_{\hat{\boldsymbol{\lambda}}}\boldsymbol{\beta}$, where $\boldsymbol{x}^*_{\hat{\boldsymbol{\lambda}}}$ is as defined in Eq. (2.6) with $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$.
4. Find

$$\underset{\boldsymbol{\beta}}{\text{Argmin}} \sum_{i=1}^{n+d} \hat{v}_i |z^*_i(\boldsymbol{\beta}, \hat{\boldsymbol{\lambda}})|$$

using any weighted LAD software (e.g. `quanteg, rfit` in `R`).

2.4 Simulation and Real Data Studies

To demonstrate the performance of our proposed method, several simulation scenarios and a real data set are considered.

2.4.1 Low Dimensional Simulation

The setting for the low-dimensional simulation is taken from Tibshirani (1996). We take a sample of size $n = 50$ where the number of predictor variables is $d = 8$ and $\boldsymbol{\beta}_0$ is set at $\boldsymbol{\beta}_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)'$. Thus $p_0 = 3$. To study the effect of tail thickness, contamination, and leverage, we considered three different scenarios:

Scenario 1: The vector of predictor variables $\boldsymbol{x}$ is generated as $\boldsymbol{x} \sim N_8(\mathbf{0}, V)$, where $V = (v_{ij})$ and $v_{ij} = 0.5^{|i-j|}$. The error distributions are t and contaminated normal. That is, the errors are generated as $e \sim t_{df}$ for several degrees of freedom (df) and $e \sim (1-\epsilon)N(0,1) + \epsilon N(0, 3^2)$ for several levels of contamination ϵ. These distributions allow us to investigate the effect of tail thickness and the rate of contamination on the proposed method.

Scenario 2: The vector of predictors $\boldsymbol{x}$ is generated as $\boldsymbol{x} \sim (1-\epsilon)N_8(\mathbf{0}, V) + \epsilon N_8(\mathbf{1}\mu, V)$, with $\mu = 5$ and the errors are generated as $e \sim N(0, 1)$. This enables us to study the effect of contamination (such us gross outliers and leverage points) in the design space.

Scenario 3: This scenario considers a partial model misspecification similar to the one in Arslan (2012). In this case, we take $\boldsymbol{\beta}_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ and $\boldsymbol{\beta}_0^* = (3, \ldots, 3)'$. Then $\boldsymbol{x}$ and y are generated as follows: for $i = 1, \ldots, n - [n\epsilon]$, $\boldsymbol{x}_i \sim N_8(\mathbf{0}, V)$ and $y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_0 + N(0, 1)$, for $i = n - [n\epsilon] + 1, \ldots, n$, $\boldsymbol{x}_i \sim N_8(\mathbf{1}\mu, V)$, $\mu = 5$, and $y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_0^c + N(0, 1)$. Varying ϵ in $[0, 1)$ allows us to study the effect of various levels of model contamination.

In all cases, we considered the adaptive lasso penalty where the tuning parameter is computed using the BIC criterion. The estimators studied were least squares (LS-AL), least absolute deviations (LAD-AL), signed-rank (SR-AL), weighted LAD (WLAD-AL), and weighted SR (WSR-AL). The weights were computed as discussed above using minimum covariance determinant (MCD) of Rousseeuw (1984). We performed 1000 replications and calculated the average number of correct zeros (true negatives), the average number of incorrect zeros (false negatives), the percentage of correct models identified, and relative efficiencies versus LS-AL of the proposed estimators for estimating β_1 based on estimated MSEs. The results of Scenario 1 are given in Figs. 2.1 and 2.2 while the results of Scenarios 2 and 3 are given in Figs. 2.3 and 2.4, respectively.

Figure 2.1 shows that LAD-AL and SR-AL (unweighted) estimators are not very good at identifying zeroes (left panels) compared to WLAD-AL and WSR-AL.

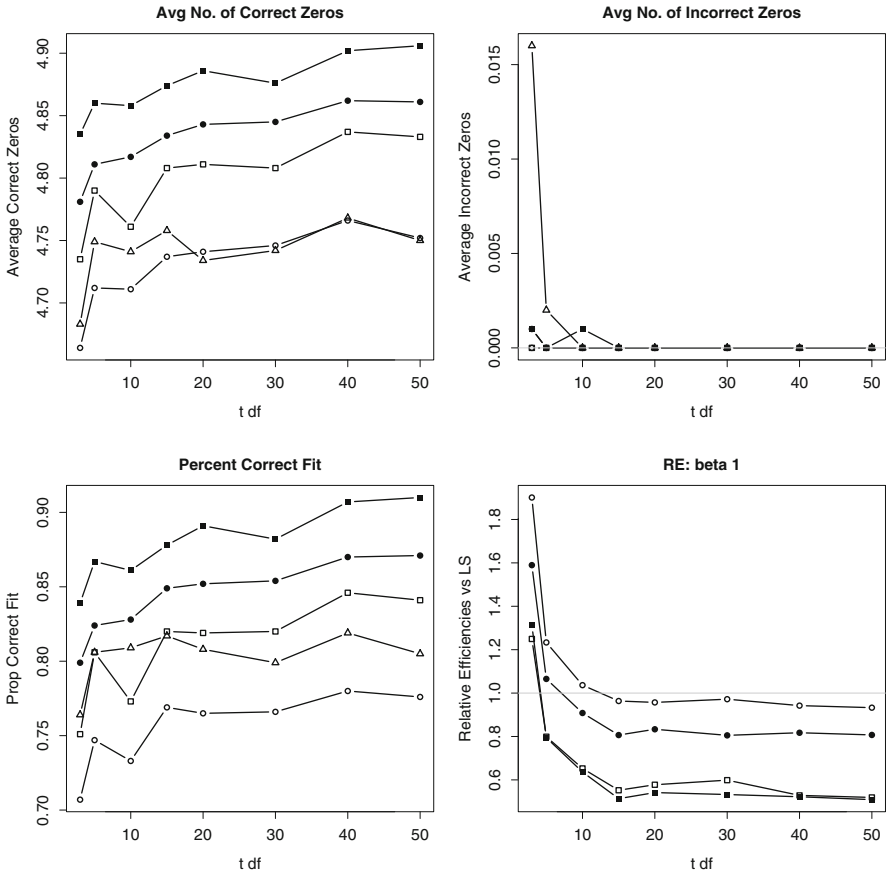

Fig. 2.1 Average number of correct and incorrect zeroes, Relative model error, Percentage of correct fit, relative efficiencies (RE) against t distribution df (Scenario 1). The symbols in the plots are LS-AL (*open triangle*), LAD-AL (*open square*), SR-AL (*open circle*), WLAD-AL (*filled square*) and WSR-AL (*filled circle*)

They are, slightly more efficient than their weighted counterpart in estimating nonzero coefficients. Their relative efficiencies versus LS-AL stabilize towards the theoretical relative efficiencies of 0.955 and 0.63 as the tails of the t distribution approach the tails of the standard normal distribution.

Figure 2.2 shows that with the exception of LS-AL, the performance in detecting true zeroes of all other estimators deteriorates as the proportion of contamination increases (left panels). On the other hand, the false negatives of LS-AL increase with increasing contamination (top right panel). Taken together, these indicate that LS-AL increasingly over-penalizes when the proportion of outliers in the data increases. and SR-AL (unweighted) estimators are not very good at identifying zeros (left panels) compared to WLAD-AL and WSR-AL. Once again the unweighted

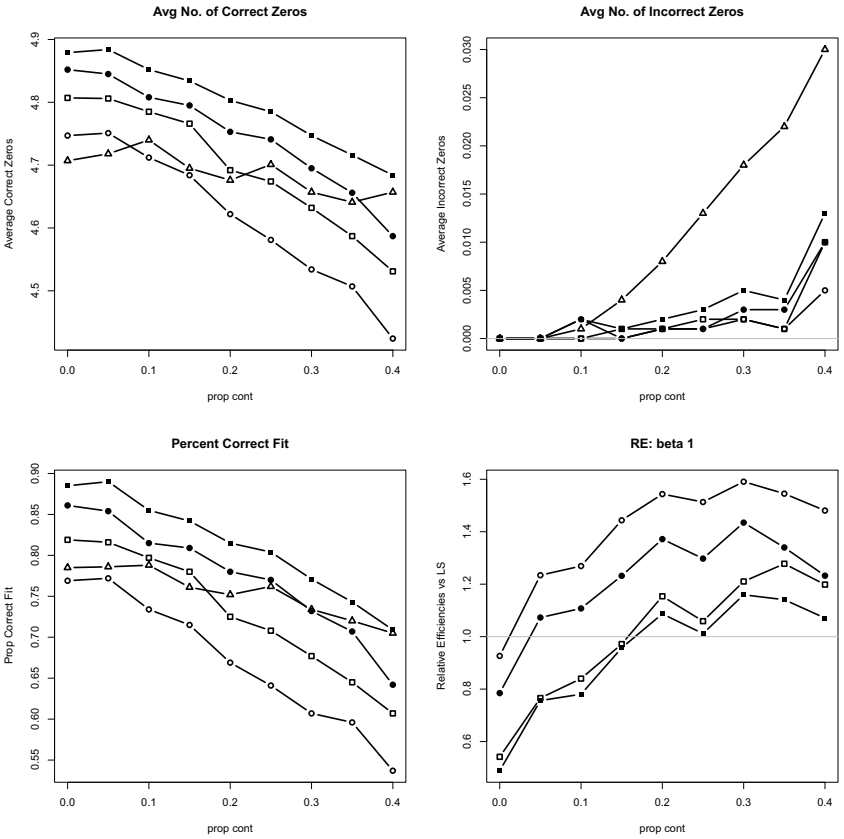

Fig. 2.2 Average number of correct and incorrect zeroes, Relative model error, Percentage of correct fit, relative efficiencies (RE) against contamination proportion (ϵ) of the contaminated normal distribution (Scenario 1). The symbols in the plots are LS-AL (*open triangle*), LAD-AL (*open square*), SR-AL (*open circle*), WLAD-AL (*filled square*) and WSR-AL (*filled circle*)

LAD and SR are slightly more efficient in estimating nonzero coefficients than their weighted counterparts while the relative efficiencies of both weighted and unweighted estimators increases with increasing proportion of error contamination.

Figure 2.3 shows that, even when the model is correctly specified, high leverage points have a detrimental effect on model selection. While the number of true positives decrease, the weighted cases appear to provide some resistance for low percentage of high-leverage points. With respect to the estimation of nonzero coefficients, the false negative rates of LS-AL increase sharply compare to all other estimators (top right panel). Once again, LS-AL is increasingly over-penalizing the model with increasing proportion of high-leverage points. It is not surprising that LS-AL is also inefficient in the estimation of nonzero coefficients, especially

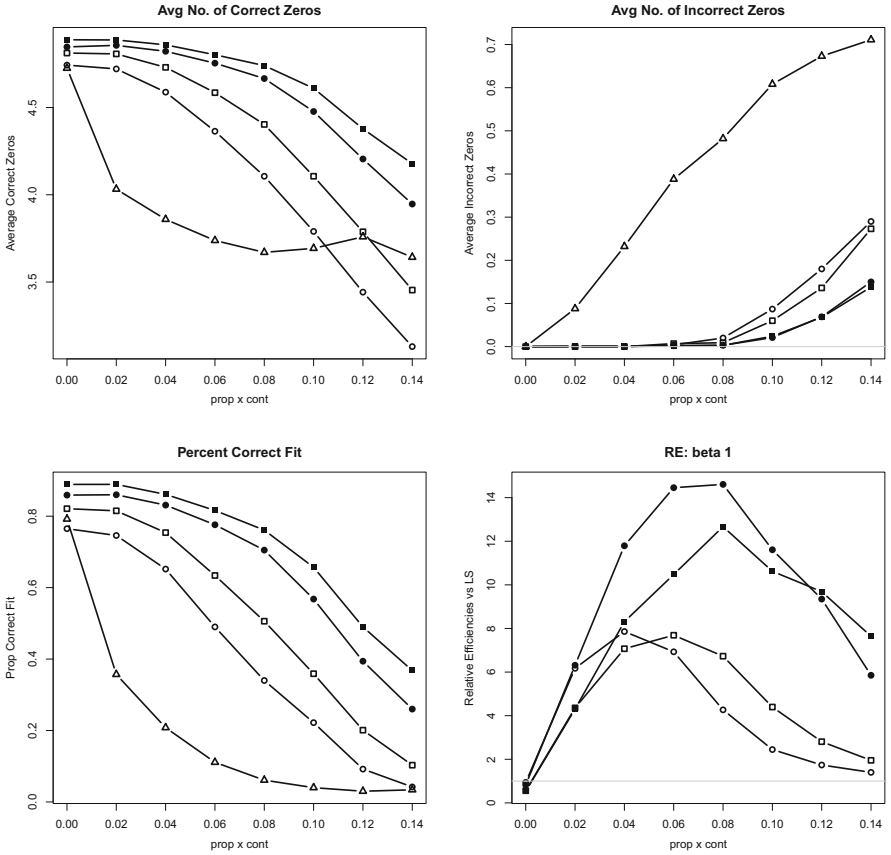

Fig. 2.3 Average number of correct and incorrect zeroes, Relative model error, Percentage of correct fit, relative efficiencies (RE) against contamination proportion (ϵ) of the distribution of the predictor $\boldsymbol{x}$ (Scenario 2). The symbols in the plots are LS-AL (*open triangle*), LAD-AL (*open square*), SR-AL (*open circle*), WLAD-AL (*filled square*) and WSR-AL (*filled circle*)

compared to WLAD-AL and WSR-AL, especially for moderate proportion (4–8 %) of high-leverage points.

Our observations remain similar to the above for model misspecification (Scenario 3). In this case, the performance of all the estimators deteriorates quite rapidly with increasing contamination. LS-AL is once again the worst offender and WLAD-AL and WSR-AL provide the highest relative efficiency. The unweighted forms are much less efficient in comparison.

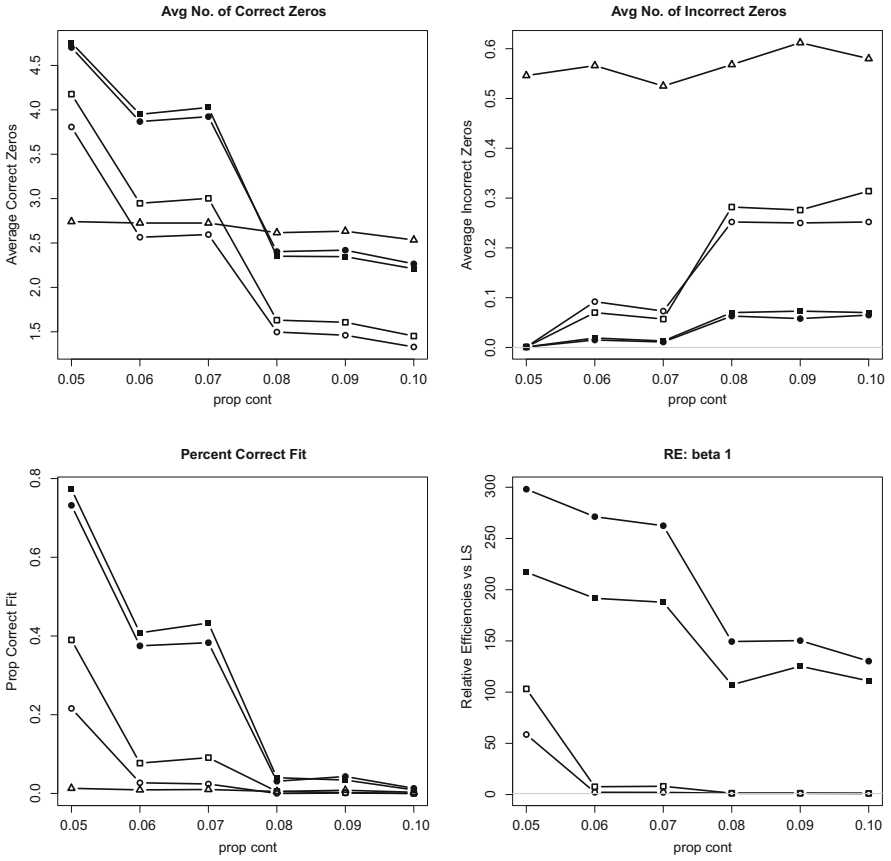

Fig. 2.4 Average number of correct and incorrect zeroes, Relative model error, Percentage of correct fit, relative efficiencies (RE) against contamination proportion (ϵ) of the model contamination (Scenario 3). The symbols in the plots are LS-AL (*open triangle*), LAD-AL (*open square*), SR-AL (*open circle*), WLAD-AL (*filled square*) and WSR-AL (*filled circle*)

2.4.2 High-Dimensional Simulation

Again as in Tibshirani (1996), consider the linear model (2.1), where $\boldsymbol{x}$ is a 100×40 matrix with entries $x_{ij} = z_{ij} + z_i$ such that z_{ij} and z_i are independent and generated from standard normal distributions. This setting makes the x_{ij}'s to be pairwise correlated with correlation coefficient of about 0.5. The random error in Eq. (2.1) is generated from two different distributions: the contaminated normal distribution with different rates of contamination and the t distribution with different degrees of freedom. The regression coefficient vector is set at $\boldsymbol{\beta} = (0, \ldots, 0, 2, \ldots, 2, 0, \ldots, 0, 2, \ldots, 2)$, where there are ten repeats in each block. From 1000 replications, average numbers of correct zeroes, average number of

incorrect zeroes and percentage of correct fit are reported. The simulation results are displayed in Fig. 2.5, where for clarity of presentation we only report results of LS-AL, SR-AL, and WSR-AL fits.

Our observations are quite similar to the low-dimensional case. LS-AL over-penalizes with increasing proportion of high leverage points, even when the model is correctly specified. SR-AL and WSR-AL provide superior performance in high leverage situations (rows three and four of Fig. 2.5). WSR-AL is clearly the best among the three for heavier tailed errors (top row). The percentage of correctly estimated models deteriorates with increasing error contamination (second row) for all the methods.

2.4.3 Boston Housing Data

The data considered here is the Boston Housing dataset which contains median values of housing in 506 census tracts and 13 predictors comprised of characteristics of the census tract. The full description of the data can be found in Leng (2010) and the dataset is available in the `R` library `MASS`. So, for sake of brevity, the description will not be included here. We first fit unpenalized regression models using the LS and SR procedures. The results are given in Table 2.1. We then fit penalized regression models using LS-AL, SR-AL, and WSR-AL. These results are displayed in Table 2.2.

The results in Table 2.1 indicate that both LS and SR find the variables INDUS and AGE insignificant while ZN is marginally significant. However, the LS and SR estimated coefficients are quite different in some cases outside of two standard errors of each other. Also, the residual plot given in Fig. 2.6 indicates the presence of heavy tails casting doubt on the LS results. In fact, observing the plot of studentized residuals of LS and SR in Fig. 2.6 plotted on the same scale, it is clear that the SR fit identifies many more outlying observations than the LS estimator. The results of penalized regressions given in Table 2.2 show that LS-AL eliminates the two insignificant variables (INDUS, AGE) from the model while SR-AL and WSR-AL eliminate a third variable (ZN) from the model. Thus, our observations are in line with those of Leng (2010).

The obvious question is if this reduction in model is associated with loss in prediction accuracy. To evaluate this, we performed cross validation where we randomly split the data into a training set containing approximately 90% of the data and a testing set containing the remaining 10 %. We fit the models using the training sets and calculated the absolute error for the test sets $|y - \hat{\alpha} - \boldsymbol{x}'\hat{\boldsymbol{\beta}}|$, where $\hat{\alpha}$ is estimated using the mean (for LS) and median (for LAD and SR) of the training set residuals $y - \boldsymbol{x}'\hat{\boldsymbol{\beta}}$. Table 2.3 gives the mean absolute error and the median model size over 100 iterations. The estimators considered all use the adaptive lasso penalty. Weights were computed using three different versions of the Mahalanobis distance: classic (Mah), minimum volume ellipsoid (MVE) of Rousseeuw (1984), and minimum covariance determinant (MCD) of Rousseeuw (1984).

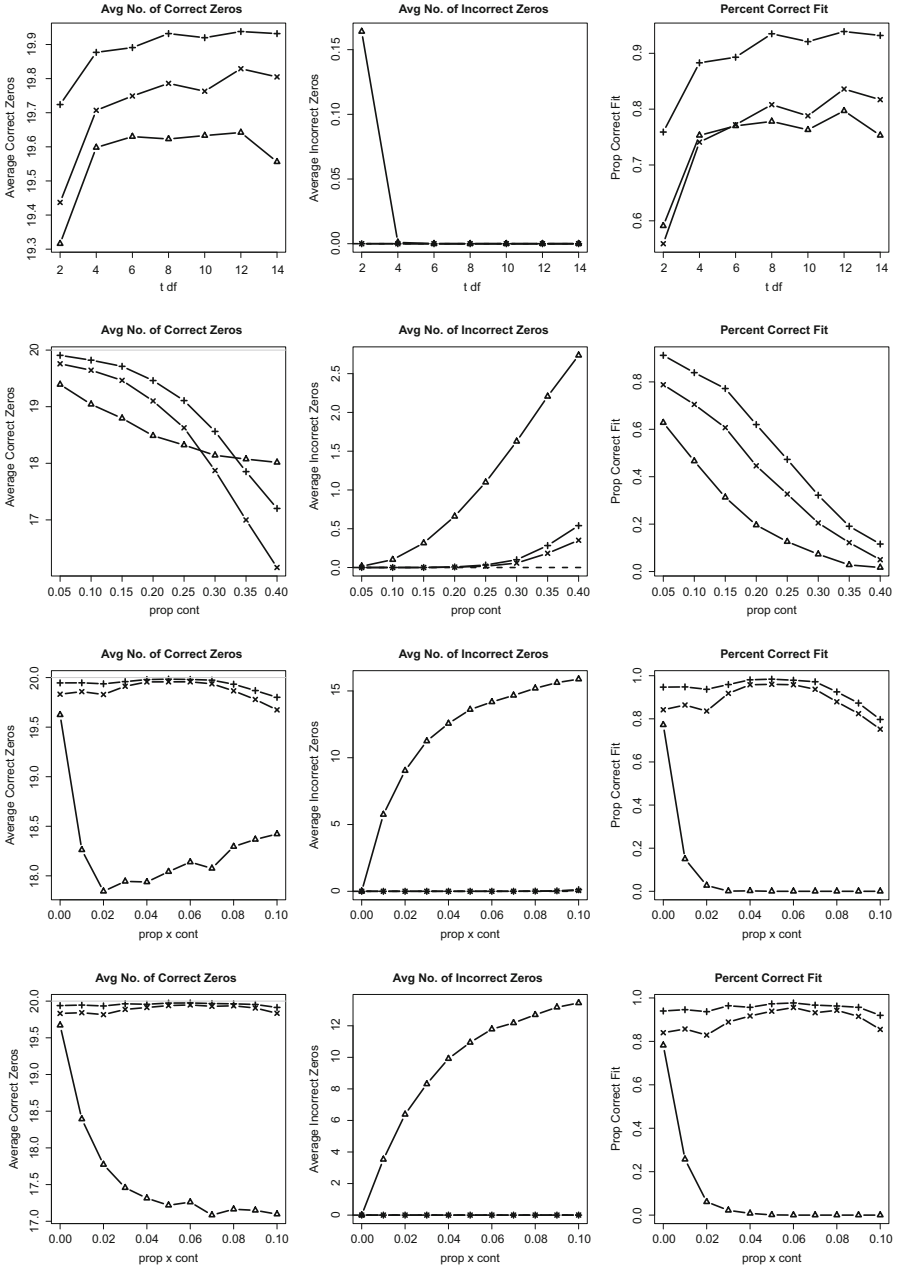

Fig. 2.5 Average number of correct and incorrect zeroes and percentage of correct fit for the high-dimensional simulation. The symbols in the plots are LS-AL (*open triangle*), SR-AL (*times*) and WSR-AL (*plus*). *First row* represents t distributed errors, *second row* represents contaminated normal, *third row* represents high-leverage points, and the *last row* represents model misspecification

Table 2.1 Estimated coefficients using LS and SR

	LS			SR		
	Coef	se	t	Coef	se	t
CRIM	−0.0103	0.0013	−7.8083	−0.0089	0.0010	−8.8709
ZN	0.0012	0.0005	2.1338	0.0008	0.0004	2.0147
INDUS	0.0025	0.0025	1.0022	0.0023	0.0019	1.2355
CHAS	0.1009	0.0345	2.9255	0.0781	0.0263	2.9691
NOX	−0.7784	0.1529	−5.0912	−0.3925	0.1166	−3.3662
RM	0.0908	0.0167	5.4300	0.1766	0.0128	13.8376
AGE	0.0002	0.0005	0.3983	−0.0006	0.0004	−1.5251
DIS	−0.0491	0.0080	−6.1486	−0.0359	0.0061	−5.9025
RAD	0.0143	0.0027	5.3725	0.0094	0.0020	4.6518
TAX	−0.0006	0.0002	−4.1574	−0.0005	0.0001	−4.6360
PTRATIO	−0.0383	0.0052	−7.3086	−0.0300	0.0040	−7.5140
B	0.0004	0.0001	3.8468	0.0006	0.0001	7.5509
LSTAT	−0.0290	0.0020	−14.3036	−0.0229	0.0015	−14.8047

Table 2.2 Estimated regression coefficients using LS, SR, LS-AL, SR-AL, and WSR-AL

	LS	LS-AL	SR	SR-AL	WSR-AL
CRIM	−0.0103	−0.0101	−0.0089	−0.0077	−0.0088
ZN	0.0012	0.0009	0.0008	0.0000	0.0000
INDUS	0.0025	0.0000	0.0023	0.0000	0.0000
CHAS	0.1009	0.0975	0.0781	0.0506	0.0546
NOX	−0.7784	−0.6990	−0.3925	−0.3238	−0.3171
RM	0.0908	0.0911	0.1766	0.1696	0.1708
AGE	0.0002	0.0000	−0.0006	0.0000	0.0000
DIS	−0.0491	−0.0489	−0.0359	−0.0251	−0.0263
RAD	0.0143	0.0126	0.0094	0.0056	0.0053
TAX	−0.0006	−0.0005	−0.0005	−0.0003	−0.0003
PTRATIO	−0.0383	−0.0376	−0.0300	−0.0317	−0.0329
B	0.0004	0.0004	0.0006	0.0005	0.0006
LSTAT	−0.0290	−0.0288	−0.0229	−0.0249	−0.0245

It is evident from Table 2.3 that while the model performances remain relatively similar, the median model sizes of the MCD and MVE weighted adaptive lasso estimation required far fewer variables. For comparable model sizes, SR-AL estimator provides lower absolute error than LS-AL, LAD-AL, WLAD-AL (Mah), and WSR-AL (Mah). Also a comparison of WLAD-AL (Arslan 2012) and WSR-AL shows that on average WSR-AL achieves a lower mean absolute error using a slightly smaller model.

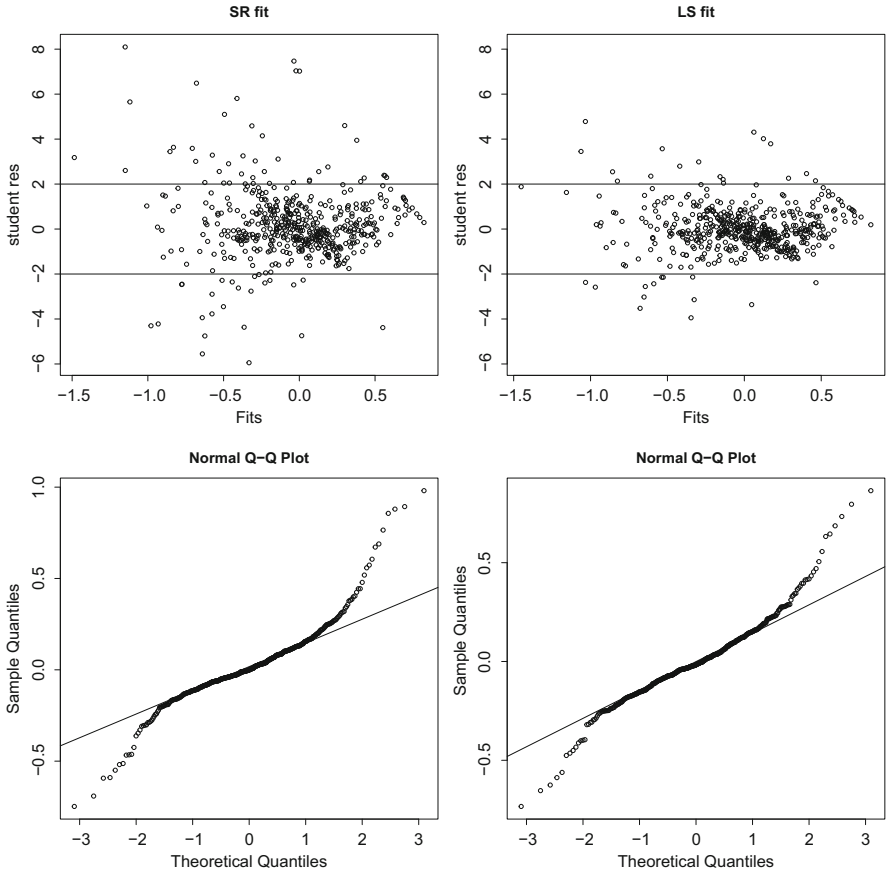

Fig. 2.6 Plots of studentized residuals versus fitted values as well as residual Q-Q plots of LS and SR fits

Table 2.3 Results of cross validation

Method	Mean absolute error (St dev)	Median model size
LS-AL	0.1408 (0.0184)	11.0
LAD-AL	0.1368 (0.0226)	11.0
SR-AL	0.1356 (0.0214)	11.0
WLAD-AL (Mah)	0.1365 (0.0213)	11.0
WSR-AL (Mah)	0.1360 (0.0210)	11.0
WLAD-AL (MVE)	0.1474 (0.0232)	10.0
WSR-AL (MVE)	0.1452 (0.0219)	10.0
WLAD-AL (MCD)	0.1523 (0.0244)	8.5
WSR-AL (MCD)	0.1490 (0.0224)	8.0

2.5 Discussion

This paper considered variable selection for linear models using penalized weighted signed-rank objective functions. It is demonstrated that the method provides selection and estimation consistency in the presence of outliers and high-leverage points. Our simulation study considered both low and high-dimensional data. In both cases, it was shown that compared to penalized least squares, penalized rank-based estimators provided more accurate true negative and false negatives identification while providing higher efficiency in estimating true positives when the error distribution is heavy tailed or contaminated. The weighted versions of the rank-based estimators provided protection against high leverage points, even when the model is incorrectly specified for the high-leverage points as long as the proportion of high-leverage points is moderate.

While the results are encouraging, an interesting extension involves regression when the data are ultra-high dimensional; that is, the dimension of the predictor also goes to infinity. This is currently under consideration by the authors. Another interesting extension involves generalized linear and single index models or even functional data analysis. Variable selection remains a valid exercise in these cases, where the last case is usually dealt with using group-selection methods.

Acknowledgements We dedicate this work to Joseph W. McKean on the occasion of his 70th birthday. We are thankful for his mentorship and guidance over the years. We also thank the anonymous referee for suggestions that improved the presentation.

Appendix

This Appendix provides some lemmas and the proofs of the main results (Theorems 2.1 and 2.2). In the proofs we have taken $W = I$ to simplify notation. The general case follows by taking $W^{1/2}\boldsymbol{x}$ in place of $\boldsymbol{x}$ in the proofs.

Proofs

The following three lemmas, whose proofs follow from slight modifications of those given in Hössjer (1994) and Hettmansperger and McKean (2011), are key to deriving the proof of the main results.

Lemma 2.1. *Under assumptions* (I_1) *and* (I_2)*, we have* $\tilde{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0$ *a.s.*

The proof of this lemma is given in Hössjer (1994) for $w \equiv 1$ and in Abebe et al. (2012) for any positive w, and a more general regression model. Also, as in Wu (1981), the proof of this lemma is obtained by showing that

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta}\in B^c} \big(D_n(\mathbf{v}_n, w, \boldsymbol{\beta}) - D_n(\mathbf{v}_n, w, \boldsymbol{\beta}_0)\big) > 0 \ \ a.s. \tag{2.11}$$

where B is an open subset of $\mathscr{B}$ and $\boldsymbol{\beta}_0 \in Int(B)$.

Lemma 2.2. *Putting* $U_n(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \dfrac{\|S_n(\boldsymbol{\gamma}) - S_n(\boldsymbol{\beta}) - \boldsymbol{\xi}(\boldsymbol{\gamma}) + \boldsymbol{\xi}(\boldsymbol{\beta})\|_1}{n^{-1/2} + \|\boldsymbol{\xi}(\boldsymbol{\gamma})\|_1}$, *we have for small enough* $\delta > 0$ *that*

$$\sup_{\|\boldsymbol{\gamma}\|\le\delta} U_n(\boldsymbol{\gamma}, \boldsymbol{\beta}_0) \xrightarrow{a.s} 0 \quad as \quad n \to \infty.$$

This lemma ensures that $n^{-1/2}S_n(\boldsymbol{\beta}_0)$ converges in distribution to a multivariate normal distribution with mean zero and covariance matrix $\gamma_{\varphi+}\Sigma$. It also results in the following asymptotic linearity established in Hettmansperger and McKean (2011).

Lemma 2.3. *Under the assumption of the errors having a finite Fisher information, we have for all* $\epsilon > 0$ *and* $C > 0$

$$P\left[\sup_{\sqrt{n}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1\le C} \|n^{-1/2}(S_n(\boldsymbol{\beta})-S_n(\boldsymbol{\beta}_0))+\zeta_{\varphi+}\sqrt{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\|_1 \ge \epsilon\right] \to 0 \quad as\ n \to \infty.$$

From this asymptotic linearity follows that for all $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_1 \le C/\sqrt{n}$, we have

$$n^{-1/2}S_n(\boldsymbol{\beta}) = n^{-1/2}S_n(\boldsymbol{\beta}_0) - \zeta_{\varphi+}\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + o(1) \tag{2.12}$$

Proof of Theorem 2.1. Set $B = \{\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u} : \ \|\mathbf{u}\|_1 < C\}$. Clearly B is an open neighborhood of $\boldsymbol{\beta}_0$ and therefore B^c is a closed subset of $\mathscr{B}$ not containing $\boldsymbol{\beta}_0$. To complete the proof, it is then sufficient to show that

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta}\in B^c} \big(Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_0)\big) > 0 \ \ a.s.$$

which from Lemma 1 of Wu (1981) will result in the $\sqrt{n}$-consistency of $\hat{\boldsymbol{\beta}}_n$. Indeed,

$$Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_0) = D_n(\mathbf{v}_n, w, \boldsymbol{\beta}) - D_n(\mathbf{v}_n, w, \boldsymbol{\beta}_0) + n\sum_{j=1}^{d}\big[P_{\lambda_j}(|\beta_j|) - P_{\lambda_j}(|\beta_{0j}|)\big]. \tag{2.13}$$

Now by the mean value theorem, assuming without loss of generality the $|\beta_{0j}| < |\beta_j|$, there exits $\alpha_j \in (|\beta_{0j}|, |\beta_j|)$ such that

$$P_{\lambda_j}(|\beta_j|) - P_{\lambda_j}(|\beta_{0j}|) = H_{\lambda_j}(|\alpha_j|)sgn(\alpha_j)(|\beta_j| - |\beta_{0j}|),$$

and therefore

$$|P_{\lambda_j}(|\beta_j|) - P_{\lambda_j}(|\beta_{0j}|)| \leq H_{\lambda_j}(|\alpha_j|)|\beta_j - \beta_{0j}|.$$

This together with Eq. (2.13) imply that

$$\begin{aligned} Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_0) &= D_n(\mathbf{v}_n, w, \boldsymbol{\beta}) - D_n(\mathbf{v}_n, w, \boldsymbol{\beta}_0) + n\sum_{j=1}^{d} H_{\lambda_j}(|\alpha_j|)sgn(\alpha_j)(|\beta_j| - |\beta_{0j}|) \\ &\geq D_n(\mathbf{v}_n, w, \boldsymbol{\beta}) - D_n(\mathbf{v}_n, w, \boldsymbol{\beta}_0) - \sqrt{n}a_n \sum_{j=1}^{p_0} |u_j|, \end{aligned} \tag{2.14}$$

as $\boldsymbol{\beta} \in B^c$ implies that $\boldsymbol{\beta}$ can be written as $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}$ with $\|\mathbf{u}\|_1 \geq C$. Being a closed subset of a compact space, B^c is compact, and hence, is closed and bounded. Then, there exists a constant M such that $C \leq \|\mathbf{u}\|_1 \leq M$. From the last term of equation (2.14), note that $\sum_{j=1}^{p_0} |u_j| \leq \|\mathbf{u}\|_1 \leq M$ from which, we have $-\sqrt{n}a_n \sum_{j=1}^{p_0} |u_j| \geq -\sqrt{n}a_n M$. Thus,

$$Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_0) \geq D_n(\mathbf{v}_n, w, \boldsymbol{\beta}) - D_n(\mathbf{v}_n, w, \boldsymbol{\beta}_0) - \sqrt{n}a_n M,$$

and so,

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta}\in B^c} \big(Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_0)\big) \geq \lim_{n\to\infty} \inf_{\boldsymbol{\beta}\in B^c} \big(D_n(\mathbf{v}_n, w, \boldsymbol{\beta}) - D_n(\mathbf{v}_n, w, \boldsymbol{\beta}_0)\big) - \lim_{n\to\infty} \Big[\sqrt{n}a_n M\Big].$$

By assumption (I_3), $\lim_{n\to\infty} \Big[\sqrt{n}a_n M\Big] = 0$, and by Lemma 2.1, we have

$$\lim_{n\to\infty} \inf_{\boldsymbol{\beta}\in B^c} \big(Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_0)\big) > 0 \ \ a.s.$$

Proof of Theorem 2.2. From the proof of Theorem 2.1 to obtain the oracle property, it is sufficient to show that for any $\boldsymbol{\beta}^*$ satisfying $\|\boldsymbol{\beta}_a^* - \boldsymbol{\beta}_{0a}\|_1 = O_p(n^{-1/2})$ and $|\beta_j^*| < Cn^{-1/2}$ for $j = p_0 + 1, \ldots, d$, $\dfrac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ and β_j^* have the same sign. Indeed,

$$\begin{aligned} n^{-1/2}\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} &= -n^{-1/2}S_n^j(\boldsymbol{\beta}_0) + \zeta_{\varphi+}\sqrt{n}(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) + \sqrt{n}H_{\lambda_j}(|\beta_j^*|)\text{sgn}(\beta_j^*) + o(1) \\ &= O_P(1) + \sqrt{n}H_{\lambda_j}(|\beta_j^*|)\text{sgn}(\beta_j^*) \ \ \text{for}\, j = p_0 + 1, \ldots, d, \end{aligned}$$

where $S_n^j(\boldsymbol{\beta}_0)$ is the j^{th} component of $S_n(\boldsymbol{\beta}_0)$. Note that by assumption (I_3), $\sqrt{n}H_{\lambda_j}(|\beta_j^*|) \geq \sqrt{n}b_n \to \infty$ as $n \to \infty$, and thus the sign of $\dfrac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ is fully determined by that of β_j^* for n large enough. This together with Theorem 2.1 implies that $\lim\limits_{n\to\infty} P(\hat{\boldsymbol{\beta}}_{nb} = \mathbf{0}) = 1$.

Moreover, by definition of $\hat{\boldsymbol{\beta}}_n$, it is obtained in a straightforward manner that $\dfrac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_a}\Big|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}_a,0)} = o_P(1)$. From this, partitioning $S_n(\boldsymbol{\beta}_0)$ as $(S_{n,a}(\boldsymbol{\beta}_0), S_{n,b}(\boldsymbol{\beta}_0))$, it follows from Eq. (2.12) that

$$o_P(1) = n^{-1/2}S_{n,a}(\boldsymbol{\beta}_0) - \zeta_{\varphi+}\sqrt{n}(\hat{\boldsymbol{\beta}}_{na} - \boldsymbol{\beta}_{0a}) + \sqrt{n}\sum_{j=1}^{p_0} H_{\lambda_j}(|\hat{\beta}_{na,j}|)\mathrm{sgn}(\hat{\beta}_{na,j}),$$

and $|\sqrt{n}\sum_{j=1}^{p_0} H_{\lambda_j}(|\hat{\beta}_{na,j}|)\mathrm{sgn}(\hat{\beta}_{na,j})| \leq p_0\sqrt{n}a_n \to 0$ as $n \to \infty$ by assumption (I_3). Hence,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{na} - \boldsymbol{\beta}_{0a}) = \zeta_{\varphi+}^{-1}n^{-1/2}S_{n,a}(\boldsymbol{\beta}_0) + o_P(1).$$

As $n^{-1/2}S_{n,a}(\boldsymbol{\beta}_0) \xrightarrow{\mathscr{D}} N\big(0,\ \gamma_{\varphi+}\Sigma_a\big)$, we have

$$\sqrt{n}\big(\hat{\boldsymbol{\beta}}_{na} - \boldsymbol{\beta}_{0a}\big) \xrightarrow{\mathscr{D}} N\big(0,\ \zeta_{\varphi+}^{-2}\gamma_{\varphi+}\Sigma_a\big).$$

References

Abebe, A., McKean, J. W., & Bindele, H. F. (2012). On the consistency of a class of nonlinear regression estimators. *Pakistan Journal of Statistics and Operation Research, 8*(3), 543–555.

Arslan, O. (2012). Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. *Computational Statistics & Data Analysis, 56*(6), 1952–1965.

Bindele, H. F., & Abebe, A. (2012). Bounded influence nonlinear signed-rank regression. *Canadian Journal of Statistics, 40*(1), 172–189.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360.

Hettmansperger, T. P., & McKean, J. W. (2011). Robust nonparametric statistical methods. In *Monographs on statistics and applied probability* (Vol. 119, 2nd ed.). Boca Raton, FL: CRC Press.

Hössjer, O. (1994). Rank-based estimates in the linear model with high breakdown point. *Journal of the American Statistical Association, 89*(425), 149–158.

Johnson, B. A. (2009). Rank-based estimation in the ℓ_1-regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data. *Biostatistics, 10*(4), 659–666.

Johnson, B. A., Lin, D., & Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association, 103*(482), 672–680.

Johnson, B. A., & Peng, L. (2008). Rank-based variable selection. *Journal of Nonparametric Statistics, 20*(3), 241–252.

Leng, C. (2010). Variable selection and coefficient estimation via regularized rank regression. *Statistica Sinica, 20*(1), 167.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871–880.

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

Wang, H., & Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis, 52*(12), 5277–5286.

Wang, H., Li, G., & Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics, 25*(3), 347–355.

Wang, L., & Li, R. (2009). Weighted Wilcoxon-type smoothly clipped absolute deviation method. *Biometrics, 65*(2), 564–571.

Wu, C. F. (1981). Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics, 9*(3), 501–513.

Xu, J., Leng, C., & Ying, Z. (2010). Rank-based variable selection with censored data. *Statistics and Computing, 20*(2), 165–176.

Zou, H. (206). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.

http://www.springer.com/978-3-319-39063-5

Robust Rank-Based and Nonparametric Methods
Michigan, USA, April 2015: Selected, Revised, and
Extended Contributions
Liu, R.; McKean, J.W. (Eds.)
2016, XIV, 277 p. 31 illus., 6 illus. in color., Hardcover
ISBN: 978-3-319-39063-5