Published in final edited form as:
J Am Stat Assoc. 2013 April 1; 108(502): 632–643. doi:10.1080/01621459.2013.766613.

Robust Variable Selection with Exponential Squared Loss

Xueqin Wang [Professor],
Department of Statistical Science, School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, 510275, China; and Zhongshan School of Medicine, Sun Yat-Sen University, Guangzhou, 510080, China; and Xinhua College, Sun Yat-Sen University, Guangzhou, 510520, China

Yunlu Jiang [Research Assistant],
Department of Statistical Science, School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, 510275, China

Mian Huang [Assistant Professor], and
School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433, China

Heping Zhang [Susan Dwight Bliss Professor]
Department of Biostatistics, Yale University School of Public Health, New Haven, Connecticut 06520-8034, U.S.A. and Changjiang Scholar, Department of Statistical Science, School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou, 510275, China

Xueqin Wang: wangxq88@mail.sysu.edu.cn; Yunlu Jiang: jiangyunlu2011@gmail.com; Mian Huang: huang.mian@mail.shufe.edu.cn; Heping Zhang: heping.zhang@yale.edu

Abstract

Robust variable selection procedures through penalized regression have been gaining increased attention in the literature. They can be used to perform variable selection and are expected to yield robust estimates. However, to the best of our knowledge, the robustness of those penalized regression procedures has not been well characterized. In this paper, we propose a class of penalized robust regression estimators based on exponential squared loss. The motivation for this new procedure is that it enables us to characterize its robustness that has not been done for the existing procedures, while its performance is near optimal and superior to some recently developed methods. Specifically, under defined regularity conditions, our estimators are $\sqrt{n}-\text{consistent}$ and possess the oracle property. Importantly, we show that our estimators can achieve the highest asymptotic breakdown point of 1/2 and that their influence functions are bounded with respect to the outliers in either the response or the covariate domain. We performed simulation studies to compare our proposed method with some recent methods, using the oracle method as the benchmark. We consider common sources of influential points. Our simulation studies reveal that our proposed method performs similarly to the oracle method in terms of the model error and the positive selection rate even in the presence of influential points. In contrast, other existing procedures have a much lower non-causal selection rate. Furthermore, we re-analyze the Boston Housing Price Dataset and the Plasma Beta-Carotene Level Dataset that are commonly used examples for regression diagnostics of influential points. Our analysis unravels the discrepancies of using our robust method versus the other penalized regression method, underscoring the importance of developing and applying robust penalized regression methods.

Keywords

Robust regression; Variable selection; Breakdown point; Influence function

1. INTRODUCTION

Selecting important explanatory variables is one of the most important problems in statistical learning. To this end, there have been many important progresses in the use of penalized regression methods for variable selection. Those penalized regression methods have a unified framework for theoretical properties and enjoy great flexibility in allowing different choices of penalty such as the bridge regression (Frank and Friedman, 1993), LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and adaptive LASSO (Zou, 2006). It is important to note that many of those methods are closely related to the least squares method. It is well known that the least squares method is sensitive to outliers in the finite samples, and consequently, the outliers can present serious problems for the least squares based methods in variable selection. Therefore, in the presence of outliers, it is desirable to replace the least squares criterion with a robust one.

In a seminal work, Fan and Li (2001) introduced a general framework of penalized robust regression estimators, i.e., to minimize

$$\Gamma_n(\beta)=\sum_{i=1}^{n}\phi(Y_i-\mathbf{x}_i^T\beta)+n\sum_{j=1}^{p}p_{\lambda_{\mathrm{nj}}}(|\beta_j|) \quad (1.1)$$

with respect to β, where $\phi(\cdot)$ is the Huber's function. Since then, penalized robust regression has attracted increased attention, and different loss functions $\phi(\cdot)$ and different penalty functions have been proposed and examined. For example, Wang et al. (2007) proposed the LAD-LASSO where $\phi(t) = |t|$, $p_{\lambda_{nj}}(|\beta_j|) = \lambda_{nj}|\beta_j|$; Wu and Liu (2009) introduced the penalized quantile regression when $\phi(t) = t\{\tau - I(t < 0)\}$ with $0 \le \tau \le 1$, and $p_{\lambda_{nj}}(|\beta_j|)$ is either the SCAD penalty or the adaptive LASSO penalty; Kai et al. (2011) studied the variable selection in semiparametric varying-coefficient partially linear model via a penalized composite quantile loss (Zou and Yuan, 2008); Johnson and Peng (2008) evaluated a rank-based variable selection; Wang and Li (2009) examined a weighted Wilcoxon-type SCAD method for robust variable selection; Leng (2010) presented the variable selection via regularized rank regression; and Bradic et al. (2011) investigated the penalized composite quasi-likelihood for ultrahigh dimensional variable selection.

Despite these progresses, to the best of our knowledge, the robustness (e.g., the breakdown point and influence function) of these variable selection procedures has not been well characterized or understood. For example, we do not know the answers to the basic questions: What is the breakdown point for a penalized robust regression estimator? Is its influence function bounded? In the regression setting, the choice of loss function determines the robustness of the resulting estimators. Thus, for model (1.1), the loss function $\rho(\cdot)$ is critical to the robustness, and for this reason, a loss function with a superior robustness performance is of great interest.

The motivation of this work is to study the robustness of variable selection procedures, which necessitates us to introduce a new robust variable selection procedure. As discussed above, the robustness has not been well characterized for the existing procedures. Besides the robustness, as presented below, our new procedure performs at a level that is near optimal and superior to two competing methods in practical settings. The numerical performance of our method is not entirely surprising, as the exponential loss function has been used in Adaboost for classification problem with similar success (Friedman et al., 2000).

We begin with the following loss function

$$\phi_\gamma(t)=1-\exp(-t^2/\gamma),$$

which is an exponential squared loss with a tuning parameter γ. The tuning parameter γ controls the degree of robustness for the estimators. When γ is large, $\phi_\gamma(t) \approx t^2/\gamma$, and therefore the proposed estimators are similar to the least squares estimators in the extreme case. For a small γ, observations with large absolute values of $t_i=Y_i-\mathbf{x}_i^T\beta$ will result in large losses of $\phi_\gamma(t_i)$, and therefore have a small impact on the estimation of β. Hence, a smaller γ would limit the influence of an outlier on the estimators, although it could also reduce the sensitivity of the estimators.

In this paper, we discuss how to select γ so that the corresponding penalized regression estimators are robust and possess desirable finite and large sample properties. We show that our estimators satisfy selection consistency and asymptotic normality. We characterize the finite sample breakdown point and show that our estimators possess the highest asymptotic addition breakdown point for a wide range of penalties, including L_q penalty with $q > 0$, adaptive LASSO, and elastic net. In addition, we derive the influence functions and show that the influence functions of our estimators are bounded with respect to the outliers in either the response or the covariate domain.

The rest of this paper is organized as follows. In Section 2, we introduce the penalized robust regression estimators with the exponential squared loss, and investigate the sampling properties. In Section 3, we study the robustness properties by deriving the influence functions and finite sample breakdown point. In Section 4, numerical simulations are conducted to compare the performance of the proposed method with composite quantile loss and L_1 loss using the oracle method as the benchmark. We conclude with some remarks in Section 5. The proofs are given in the Appendix.

2. PENALIZED ROBUST REGRESSION ESTIMATOR WITH EXPONENTIAL SQUARED LOSS

Assume that $\{\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ is a random sample from population $(\mathbf{x}, Y)$. Here, Y is a univariate response, $\mathbf{x}$ is a d-dimensional predictor, and $(\mathbf{x}, Y)$ has joint distribution F. Suppose that $(\mathbf{x}_i, Y_i)$ satisfying a linear regression model

$$Y_i=\mathbf{x}_i^T\beta+\varepsilon_i, i=1,\ldots,n, \quad (2.1)$$

where β is a d-dimensional vector of unknown parameters, and the error terms ε_i are i.i.d. with unknown distribution G, $E(\varepsilon_i) = 0$, and ε_i is independent of $\mathbf{x}_i$. An intercept term is included if the first elements of all $\mathbf{x}_i$'s are 1. Let $D_i = (\mathbf{x}_i, Y_i)$, and $\mathbf{D}_n = (D_1, \cdots, D_n)$ be the observed data. Variable selection is necessary because some of the coefficients in β are zero; that is, some of the variables in $\mathbf{x}_i$ do not contribute to Y_i. Without loss of generality, let $\beta=\left(\beta_1^T,\beta_2^T\right)^T$, where $\beta_1 \in \mathbb{R}^s$ and $\beta_2 \in \mathbb{R}^{d-s}$. The true regression coefficients are $\beta_0=\left(\beta_{01}^T,\beta_{02}^T\right)^T$ with each element of β_{01} being nonzero, and $\beta_{02} = \mathbf{0}$. Let $\mathbf{x}_i=\left(\mathbf{x}_{i1}^T,\mathbf{x}_{i2}^T\right)^T$, where $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ are the covariates corresponding to β_1 and β_2.

The objective function of penalized robust regression consists of a loss function and a penalty term. In this paper, we propose maximizing

$$\ell_n(\beta)=\sum_{i=1}^{n}\exp\{-(Y_i-\mathbf{x}_i^T\beta)^2/\gamma_n\}-n\sum_{j=1}^{d}p_{\lambda_{\mathrm{nj}}}(|\beta_j|)\quad(2.2)$$

with respect to $\boldsymbol{\beta}$, which is a special case of (1.1) for any $\gamma_n \in (0, +\infty)$. Because γ_n is a tuning parameter and needs to be chosen adaptively according to the data, our objective function (2.2) is distinct from (1.1). This additional feature is critical for our estimators to have a high breakdown point and high efficiency.

Let $\hat{\beta}=\left(\hat{\beta}_{n1}^T,\hat{\beta}_{n2}^T\right)^T$ be the resulting estimator of (2.2),

$$a_n=\max\{p'_{\lambda_{\mathrm{nj}}}(|\beta_{0j}|):\beta_{0j}\neq 0\},$$

$$b_n=\max\{p''_{\lambda_{\mathrm{nj}}}(|\beta_{0j}|):\beta_{0j}\neq 0\},$$

and

$$I(\beta,\gamma)=\frac{2}{\gamma}\int \mathrm{xx}^T e^{-r^2/\gamma}\left(\frac{2r^2}{\gamma}-1\right)\mathrm{dF}(\mathbf{x},y),$$

where $r = Y - \mathbf{x}^T\boldsymbol{\beta}$.

We assume the following regularity condition:

(C1) $\Sigma = E(\mathbf{xx}^T)$ is positive definite, and $E\|\mathbf{x}\|^3 < \infty$.

Condition (C1) ensures that the main term dominates the remainder in the Taylor expansion. It warrants further examination as to whether this condition can be weakened. With these preparations, we present the following sampling properties for our proposed estimators.

Theorem 1 *Assume that condition (C1) holds, $b_n = o_p(1)$, and $I(\boldsymbol{\beta}_0, \gamma_0)$ is negative definite.*

- **i.** *If $\gamma_n - \gamma_0 = o_p(1)$ for some $\gamma_0 > 0$, there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(n^{-1/2} + a_n)$.*
- **ii.** *(Oracle property) If $\sqrt{n}a_n=O_p(1)$, $1/\min_{s+1\le j\le d}(\sqrt{n}\lambda_{\mathrm{nj}})=o_p(1)$, $\sqrt{n}(\gamma_n-\gamma_0)=o_p(1)$, and with probability 1,*

$$\liminf_{n\to\infty}\liminf_{t\to 0+}\left\{\min_{s+1\le j\le d}p'_{\lambda_{\mathrm{nj}}}(|t|)/\lambda_{\mathrm{nj}}\right\}>0,\quad(2.3)$$

then we have: (a) sparsity, i.e., $\hat{\boldsymbol{\beta}}_{n2} = \mathbf{0}$ with probability 1; (b) asymptotic normality,

$$\sqrt{n}(I_1(\beta_{01},\gamma_0)+\Sigma_1)\left\{\hat{\beta}_{n1}-\beta_{01}+(I_1(\beta_{01},\gamma_0)+\Sigma_1)^{-1}\Delta\right\}\xrightarrow{D}N(\mathbf{0},\Sigma_2),\quad(2.4)$$

where $\Sigma_1=\mathrm{diag}\{p''_{\lambda_{n1}}(|\beta_{01}|),\cdots,p''_{\lambda_{\mathrm{ns}}}(|\beta_{0s}|)\}$, $\Sigma_2=\mathrm{cov}(\exp(-r^2/\gamma_0)\frac{2r}{\gamma_0}\mathbf{x}_{i1})$,

$$\Delta=\left(p'_{\lambda_{n1}}(|\beta_{01}|)\mathrm{sign}(\beta_{01}),\cdots,p'_{\lambda_{\mathrm{ns}}}(|\beta_{0s}|)\mathrm{sign}(\beta_{0s})\right)^T,$$

$$I_1(\beta_{01},\gamma_0)=\frac{2}{\gamma_0}E\left[\exp\left(-r^2/\gamma_0\right)\left(\frac{2r^2}{\gamma_0}-1\right)\right](E\mathbf{x}_{i1}\mathbf{x}_{i1}^T).$$

Remark 1 *Note that not all penalties satisfy the conditions in Theorem 1. For example, LASSO is inconsistent, and the oracle property does not hold.* Zou (2006) *proposed the adaptive LASSO, and showed that it enjoys the oracle property. The adaptive LASSO penalty has the form of* $p_{\lambda_{nj}}(|\beta_j|) = \lambda_{nj}|\beta_j|$ *with* $\lambda_{nj} = \tau_{nj}/|\tilde{\beta}_j|^k$ *for some* $k > 0$*, where* $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_d)^T$ *is a* $\sqrt{n}-\mathrm{consistent}$ *estimator of* $\boldsymbol{\beta}_0$*, and* τ_{nj}*'s are the regularization parameters. With this penalty in the penalized robust regression* (2.2)*, we can prove that our estimators are* $\sqrt{n}-\mathrm{consistent}$ *and have the oracle property under the following condition (C2) in addition to the regularity condition (C1). The detailed proof is omitted here.*

(C2): $\max_{1\le j\le s}(\sqrt{n}\lambda_{\mathrm{nj}})=o_p(1)\ \text{and}\ 1/\min_{s+1\le j\le d}(\sqrt{n}\lambda_{\mathrm{nj}})=o_p(1).$

In fact, $\max_{1\le j\le s}(\sqrt{n}\lambda_{\mathrm{nj}})=o_p(1)\ \text{implies}\ \sqrt{n}a_n=O_p(1)$ *by the definition of* a_n*. Note that some data-driven methods for selecting* λ_{nj}*, e.g., cross-validation, may not satisfy condition (C2). Here, we utilize a BIC criterion for which (C2) holds.*

3. ROBUSTNESS PROPERTIES AND IMPLEMENTATION

3.1 Finite sample breakdown point

Finite sample breakdown point is used to measure the maximum fraction of outliers in a sample that an estimator can tolerate before returning arbitrary values. It is a global measurement of robustness in terms of resistance to outliers. Several definitions of the finite sample breakdown point have been proposed in literature [see Hampel (1971); Donoho (1982); Donoho and Huber (1983)].

Here, let us describe the addition measure defined by Donoho and Huber (1983). Recall the notations $D_i = (\mathbf{x}_i, Y_i)$, and $\mathbf{D}_n = (D_1, \cdots, D_n)$. We assume that $\mathbf{D}_n$ contains m bad points and $n - m$ good points. Denote the bad points by $\mathbf{D}_m = \{D_1, \cdots, D_m\}$ and the good points by $\mathbf{D}_{n-m} = \{D_{m+1}, \cdots, D_n\}$. The fraction of bad points in $\mathbf{D}_n$ is m/n. Let $\hat{\boldsymbol{\beta}}(\mathbf{D}_n)$ denote a regression estimator based on sample $\mathbf{D}_n$. The finite sample addition breakdown point of an estimator is defined as

$$\mathrm{BP}=(\hat{\beta}_n;\mathbf{D}_{n-m})=\min\left\{\frac{m}{n}:\sup_{\mathbf{D}_m}\|\hat{\beta}(\mathbf{D}_n)-\hat{\beta}(\mathbf{D}_{n-m})\|=\infty\right\},$$

where ‖ · ‖ is the Euclidean norm. In the regression setting, many estimators such as S-estimator (Rousseeuw and Yohai, 1984), MM-estimator, τ-estimator (Yohai and Zamar, 1988), and REWLS-estimator (Gervini and Yohai, 2002), can achieve the highest asymptotic breakdown point of 1/2.

Next, we derive the finite sample breakdown point for our proposed penalized robust estimators with the exponential squared loss. We first take an initial estimator $\tilde{\boldsymbol{\beta}}_n$. For a contaminated sample $\mathbf{D}_n$, let

$$\zeta(\gamma_n)=\frac{2m}{n}+\frac{2}{n}\sum_{i=m+1}^{n}\phi_{\gamma_n}\left\{r_i(\tilde{\beta}_n)\right\}, \quad (3.1)$$

where $r_i(\beta)=Y_i-\mathbf{x}_i^T\beta$. It is easy to see that $\zeta(\gamma_n)$ is a real number ranged in (0, 2]. Let

$$a_{\mathrm{nm}}=(n-m)^{-1}\max_{\beta\in\mathbb{R}^d}\#\{i: m+1\le i\le n \text{ and } \mathbf{x}_i^T\beta=0\}.$$

Note that if a set of d regressor variables is linearly independent, then $a_{nm}=(d-1)/(n-m)$. Denote by $BP(\hat{\boldsymbol{\beta}}_n;\mathbf{D}_{n-m},\gamma_n)$ the breakdown point for the penalized robust regression estimators $\hat{\boldsymbol{\beta}}_n$ with the tuning parameter γ_n.

Theorem 2 *For any penalty function of the form* $p_{\lambda_{nj}}(|\beta_j|)=\lambda_{nj}g(|\beta_j|)$*, where* $g(\cdot)$ *is a strictly increasing and unbounded function defined on* $[0,\infty]$*, and the weight* λ_{nj} *is positive for all* $j=1,\cdots,d$*. If* $m/n\le\varepsilon<(1-2a_{nm})/(2-2a_{nm})$, $a_{nm}<0.5$*, and* $\zeta(\gamma_n)<(1-\varepsilon)(2-2a_{nm})$ *hold, then, for any initial estimator* $\tilde{\boldsymbol{\beta}}_n$ *of* $\boldsymbol{\beta}_0$*, we have*

$$\mathrm{BP}=(\hat{\beta}_n;\mathbf{D}_{n-m},\gamma_n)\ge\min\left\{\mathrm{BP}(\tilde{\beta}_n;\mathbf{D}_{n-m}),\frac{1-2a_{\mathrm{nm}}}{2-2a_{\mathrm{nm}}},1-\frac{\zeta(\gamma_n)}{2-2a_{\mathrm{nm}}}\right\}. \quad (3.2)$$

Theorem 2 provides the lower bound of breakdown point of the proposed penalized robust estimators. The breakdown point depends on the breakdown point of an initial estimate and the tuning parameter γ_n. If $\tilde{\boldsymbol{\beta}}_n$ is a robust estimator with asymptotic breakdown point 1/2, and γ_n is chosen such that $\zeta(\gamma_n)\in(0,1]$, then $BP(\hat{\boldsymbol{\beta}}_n;\mathbf{D}_{n-m},\gamma_n)$ is asymptotically 1/2. This observation guides us in selecting γ_n to reach the highest efficiency. We should note that the breakdown point is a limiting measure of the bias robustness (He and Simpson, 1993).

There are many commonly used penalties that utilize the penalty form of Theorem 2, such as LASSO, adaptive LASSO, the L_q penalty with $q>0$, logarithm penalty, elastic-net penalty, and adaptive elastic-net penalty. Hence the breakdown point result holds for these penalized robust regression estimators. However, it is an open question to extend the result of Theorem 2 to bounded penalties such as SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). As a practical solution, for those bounded penalties, we may employ the one-step or k-step local linear approximation (LLA) method proposed by Zou and Li (2008). Additional discussions will be given in Section 5.

3.2 Influence function

The influence function introduced by Hampel (1968) measures the stability of estimators given an infinitesimal contamination. Denote by $\delta_{\mathbf{z}}$ the point mass probability distribution at a fixed point $\mathbf{z}=(\mathbf{x}_0,y_0)^T\in\mathbb{R}^{d+1}$. Given the distribution F of $(\mathbf{x},Y)$ in $\mathbb{R}^{d+1}$ and proportion $\varepsilon\in(0,1)$, the mixture distribution of F and $\delta_{\mathbf{z}}$ is $F_\varepsilon=(1-\varepsilon)F+\varepsilon\delta_{\mathbf{z}}$. Suppose that λ_{nj}'s have the limit point λ_{0j}'s, let

$$\beta_0^*=\arg\min_{\beta}\left[\left\{\int\left(1-e^{-(y-\mathbf{x}^T\beta)^2/\gamma_0}\right)\mathrm{dF}\right\}+\sum_{j=1}^{d}p_{\lambda_{0j}}(|\beta_j|)\right],$$

$$\beta_\varepsilon^*=\arg\min_\beta \Big[\Big\{\int \Big(1-e^{-(y-\mathbf{x}^T\beta)^2/\gamma_0}\Big)\ \mathrm{dF}_\varepsilon\Big\}+\sum_{j=1}^{d} p_{\lambda_{0j}}(|\beta_j|).$$

Note that β_0^* is a shrinkage of the true coefficient $\boldsymbol{\beta}_0$ to 0, and $\hat{\beta}_n \xrightarrow{P} \hat{\beta}_0^*$ under the conditions in Theorem 1. For the exponential-type penalized robust estimators, the influence function at a point $\mathbf{z} \in \mathbb{R}^{d+1}$ is defined as $\mathrm{IF}(\mathbf{z};\beta_0^*)=\lim_{\varepsilon\to 0+}(\beta_\varepsilon^*-\beta_0^*)/\varepsilon$, provided that the limit exists.

Theorem 3 *For the penalized robust regression estimators with exponential squared loss, the j-th element of the influence function* $\mathrm{IF}_j(\mathbf{z};\beta_0^*)$ has the following form:

$$\mathrm{IF}_j(\mathbf{z};\beta_0^*)=\begin{cases} 0 & \text{if } \beta_{0j}^*=0, \\ -\Gamma_j\left\{2\exp(-r_0^2/\gamma_0)\, r_0\mathbf{x}_0/\gamma_0+\nu_2\right\}, & \text{otherwise,}\end{cases} \tag{3.3}$$

where Γ_j *denotes the j-th row of* $\{2A(\gamma_0)/\gamma_0 - B\}^{-1}$, $r_0=y_0-\mathbf{x}_0^T\beta_0^*$,

$$\nu_2=\left\{p'_{\lambda_{01}}(|\beta_{01}^*|)\mathrm{sign}(\beta_{01}^*),\cdots,p'_{\lambda_{0d}}(|\beta_{0d}^*|)\mathrm{sign}(\beta_{0d}^*)\right\}^T,$$

$$B=\mathrm{diag}\left\{p''_{\lambda_{01}}(|\beta_{01}^*|),\cdots,p''_{\lambda_{0d}}(|\beta_{0d}^*|)\right\},$$

and

$$A(\gamma)=\int \mathrm{xx}^T\ \exp\{-(y-\mathbf{x}^T\beta_0^*)^2/\gamma\}\left\{\frac{2(y-\mathbf{x}^T\beta_0^*)^2}{\gamma}-1\right\}\ \mathrm{dF}(\mathbf{x},y). \tag{3.4}$$

For the case of adaptive LASSO penalty, with the regularization parameter selected by the BIC described in Section 3.3, by the condition (C2), we have $\lambda_{0j} = 0$ for $j = 1, \cdots, s$, and $\lambda_{0j} = +\infty$ for $j = s + 1, \cdots, d$. Therefore, the corresponding influence functions of the zero coefficients are zero. For the nonzero coefficients, the influence functions have the form

$$\mathrm{IF}_j(\mathbf{z};\beta_0^*)=-\Gamma_j\{2\exp(-r_0^2/\gamma_0)\, r_0\mathbf{x}_{01}/\gamma_0\}. \tag{3.5}$$

3.3 Algorithm and the choice of tuning parameters

The penalty function facilitates variable selection in regression. Theorem 2 implies that there may be many penalties that can facilitate robust variable selection; however, we believe it is a lesser issue to compare the performance of different penalties for a given type of loss function. A more important issue is to compare the performance of different loss functions for a given form of penalty.

In the simulations, we focus on the adaptive LASSO penalty $p_{\lambda_{nj}}(|\beta_j|) = \tau_{nj}|\beta_j|/|\tilde{\beta}_{nj}|^k$ with $k = 1$, $\lambda_{nj} = \tau_{nj}/|\tilde{\beta}_{nj}|$, and $\tilde{\beta}_{nj}$ is an initial robust regression estimator. The maximization of (2.2) with an adaptive LASSO penalty involves nonlinear weighted L_1 regularization. To facilitate the computation, we use a quadratic approximation to replace the loss function. Let

$$\ell^*(\beta)=\sum_{i=1}^{n}\exp\{-(Y_i-\mathbf{x}_i^T\beta)^2/\gamma_n\}. \quad (3.6)$$

Suppose that $\tilde{\boldsymbol{\beta}}$ is an initial estimator, then the loss function is approximated as $\ell^*(\beta) \approx \ell^*(\tilde{\beta})+\frac{1}{2}(\beta-\tilde{\beta})^T\nabla^2\ell^*(\tilde{\beta})(\beta-\tilde{\beta})$. Next, we maximize

$$\frac{1}{2}(\beta-\tilde{\beta})^T\nabla^2\ell^*(\tilde{\beta})\,(\beta-\tilde{\beta})-n\sum_{j=1}^{d}\lambda_{\mathrm{nj}}\,|\beta_j|/|\tilde{\beta}_j$$

with respect to $\boldsymbol{\beta}$, which leads to an approximated solution of (2.2).

There exist efficient computing algorithms for this L_1 regularization problem. Popular algorithms include the least angle regression (Efron et al., 2004) and coordinate descent procedures such as the pathwise coordinate optimization introduced by Friedman et al. (2007), coordinate descent algorithms proposed by Wu and Lange (2008), and the block coordinate gradient descent (BCGD) algorithm proposed by Tseng and Yun (2009). In our paper, we use BCGD algorithm for computation.

To implement our methodology, we need to select both types of tuning parameters λ_{nj} and γ_n. Since λ_{nj} and γ_n depend on each other, it could be treated as a bivariate optimization problem. In this paper, we consider a simple selection method for λ_{nj}, and a data-driven procedure for γ_n.

The choice of the regularization parameter $\boldsymbol{\lambda_{nj}}$—In general, many methods can be used to select λ_{nj}, such as cross-validation, AIC, BIC. To reduce intensive computation, and guarantee consistent variable selection, we consider the regularization parameter by minimizing a BIC-type objective function (see Wang et al. (2007)):

$$\sum_{i=1}^{n}[1-\exp\{-(Y_i-\mathbf{x}_i^T\beta)^2/\gamma_n\}]+n\sum_{j=1}^{d}\tau_{\mathrm{nj}}\,|\beta_j|/|\tilde{\beta}_{\mathrm{nj}}|-\sum_{j=1}^{d}\log(0.5n\tau_{\mathrm{nj}})\,\log(n).$$

This leads to $\lambda_{nj} = \hat{\tau}_{nj}/|\tilde{\beta}_{nj}|$, where

$$\hat{\tau}_{\mathrm{nj}}=\frac{\log(n)}{n}.$$

Note that this simple choice satisfies the conditions for oracle property of the adaptive LASSO, as described in the following corollary.

Corollary 1 *If the regularization parameter is chosen as* $\hat{\tau}_{nj} = \log(n)/n$*, then the solution of* (2.2) *with adaptive LASSO penalty possess the oracle property.*

The choice of tuning parameter $\boldsymbol{\gamma_n}$—The tuning parameter γ_n controls the degree of robustness and efficiency of the proposed robust regression estimators. To select γ_n, we propose a data-driven procedure which yields both high robustness and high efficiency simultaneously. We first determine a set of the tuning parameters such that the proposed penalized robust estimators have asymptotic breakdown point at 1/2, and then select the tuning parameter with the maximum efficiency. We describe the whole procedure in the following steps:

1. **Find the pseudo outlier set of the sample**

   Let $\mathbf{D}_n = \{(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)\}$. Calculate ** and $S_n = 1.4826 \times median_i\,|r_i(\hat{\beta}_n) - median_j(r_j(\hat{\beta}n))|$. Then, take the pseudo outlier set $\mathbf{D}_m = \{(\mathbf{x}_i, Y_i) : |r_i(\hat{\beta}_n)| \geq 2.5S_n\}$, set $m = \#\{1 \leq i \leq n : |r_i(\hat{\beta}_n)| \leq 2.5S_n\}$, and $\mathbf{D}_{n-m} = \mathbf{D}_n/\mathbf{D}_m$.

2. **Update the tuning parameter γ_n**

   Let γ_n be the minimizer of $\det(\hat{V}(\gamma))$ in the set $G = \{\gamma : \zeta(\gamma) \in (0, 1]\}$, where $\zeta(\gamma)$ is defined in (3.1), $\det(\cdot)$ denotes the determinant operator, $\hat{V}(\gamma) = \{\hat{I}_1(\hat{\beta}_n)\}^{-1}\tilde{\Sigma}_2\{\hat{I}_1(\hat{\beta}_n)\}^{-1}$, and

$$\hat{I}_1(\hat{\beta}_n)=\frac{2}{\gamma}\left\{\frac{1}{n}\sum_{i=1}^{n}\exp\left(-r_i^2(\hat{\beta}_n)/\gamma\right)\left(\frac{2r_i^2(\hat{\beta}_n)}{\gamma}-1\right)\right\}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T\right),$$

$$\tilde{\Sigma}_2=\text{cov}\left\{\exp\left(-r_1^2(\hat{\beta}_n)/\gamma\right)\frac{2r_1(\hat{\beta}_n)}{\gamma}\mathbf{x}_1,\cdots,\exp\left(-r_n^2(\hat{\beta}_n)/\gamma\right)\frac{2r_n(\hat{\beta}_n)}{\gamma}\mathbf{x}_n\right\}.$$

3. **Update $\hat{\beta}_n$:**

   With fixed $\lambda_{nj} = \log(n)/(n|\beta_{nj}|)$, and the selected γ_n in step 2, and update $\hat{\beta}_n$ by maximizing (2.2).

In practice, we set the MM-estimator $\tilde{\beta}_n$ as the initial estimate, i.e., $\hat{\beta}_n = \tilde{\beta}_n$, then repeat Steps 1 – 3 until $\hat{\beta}_n$ and γ_n converges. Note that λ_{nj} is fixed within the iterations. One may update $\hat{\beta}_n$ in step 3 with λ_{nj} selected by cross-validation, however, this approach requires huge computation. See discussions in section 5.

Remark 2 *In the initial cycle of our algorithm, we set* $\hat{\beta}_n$ *as the MM estimator and detect the outliers in step 1. This allows us to calculate* $\zeta(\gamma_n)$ *by* (3.1). *Based on Theorem 2, the proposed penalized robust estimators achieve asymptotic breakdown point 1/2 if two conditions hold: (1) the initial robust estimator possesses asymptotic breakdown point 1/2; and (2) the tuning parameter* γ_n *is chosen such that* $\zeta(\gamma_n) \in (0, 1]$. Therefore, $\hat{\beta}_n$ *achieves the asymptotic breakdown point 1/2 at all iterations.*

To attain high efficiency, we choose the tuning parameter γ_n *by minimizing the determinant of asymptotic covariance matrix as in step 2. Since the calculation of of* $\det(\hat{V}(\gamma))$ *depends on estimate* $\hat{\beta}_n$*, we update* $\hat{\beta}_n$ *in step 3, and repeat the algorithm until convergence. Based on our limited experience in simulation, the computing algorithm converges very quickly. In practice, we repeat Steps 1–3 once.*

4. SIMULATION AND APPLICATION

4.1 Simulation study

In this section, we conduct simulation studies to evaluate the finite-sample performance of our estimator. We first illustrate how to select the tuning parameter γ. We choose $n = 200$, $d = 8$, and $\beta = (1, 1.5, 2, 1, 0, 0, 0, 0)^T$. We generate $\mathbf{x}_i = (x_{i1}, \cdots, x_{id})^T$ from a multi-normal distribution $N(\mathbf{0}, \Omega_2)$, where the (i, j)-th element of Ω_2 is $\rho^{|i-j|}$, $\rho = 0.5$. The error term follows a Cauchy distribution.

Following the procedure described in section 3.3, we obtain the values of $\zeta(\gamma)$ and det($\hat{V}(\gamma)$), and plot them against the tuning parameter γ as depicted in Figure 1. Set G can be determined from Figure 1(a). The tuning parameter γ is selected by minimizing the det($\hat{V}(\gamma)$) over G as illustrated in Figure 1(b).

Next, we evaluate the performance of various loss functions with different sample sizes. For each setting, we simulate 1000 data sets from model (2.1) with sample sizes of $n = 100, 150, 200, 400, 600, 800$. We choose $d = 8$ and $\boldsymbol{\beta} = (1, 1.5, 2, 1, 0, 0, 0, 0)^T$. We use the following three mechanisms to generate influential points:

1. Influential points in the predictors. Covariate $\mathbf{x}_i$ follows a mixture of d-dimensional normal distributions $0.8N(\mathbf{0}, \Omega_1) + 0.2N(\boldsymbol{\mu}, \Omega_2)$, $\Omega_1 = I_{d\times d}$, $\boldsymbol{\mu} = \mathbf{3}\mathbf{1}_d$, $\mathbf{1}_d$ is d-dimensional vector of ones, and the error term follows a standard normal distribution;

2. Influential points in the response. Covariate $\mathbf{x}_i$ follows a multi-normal distribution $N(\mathbf{0}, \Omega_2)$, and the error term follows a mixture normal distribution $0.8N(0, 1) + 0.2N(10, 6^2)$;

3. Influential points in both the predictors and response. Covariate $\mathbf{x}_i$ follows a mixture of d-dimensional normal distributions $0.8N(\mathbf{0}, \Omega_1) + 0.2N(\boldsymbol{\mu}, \Omega_2)$, and the error term follows a Cauchy distribution.

For each mechanism mentioned above, we compare the performance of four methods: CQR-LASSO (CQR is the shortening of the composite quantile regression introduced by Zou and Yuan (2008)); LAD-LASSO proposed by Wang et al. (2007); the oracle method based on MM-estimator; and our method (ESL-LASSO). For CQR-LASSO, we set the quantiles $\tau_k = k/10$ for $k = 1, 2, \cdots, 9$. The performance is represented by the positive selection rate (PSR), the non-causal selection rate (NSR), and the median and median absolute deviation (MAD) of the model error advocated by Fan and Li (2001), where PSR(Chen and Chen, 2008) is the proportion of causal features selected by one method in all causal features, NSR(Fan and Li, 2001) is the average restricted only to the true zero coefficients, and the model error is defined by

$$\mathrm{ME}=\left(\hat{\beta}_n-\beta_0\right)^T E\left[\mathrm{xx}^T\right]\left(\hat{\beta}_n-\beta_0\right).$$

The tuning parameter γ is selected for each simulated sample, and $\bar{\gamma}_n$ and $\bar{\zeta}(\gamma_n)$ are the averages of the selected values in 1000 simulations.

From Tables 1, 2, and 3, we find that ESL-LASSO yields larger model error than LAD-LASSO and CQR-LASSO when the sample size is small, because it involves some consistent estimators in its selection procedure. With the increase of the sample size, the medians and MADs of the model error decrease in all three settings. Especially in the view of the median of the model error, although ESL-LASSO's is always larger than CQR-LASSO's but shall be smaller than LAD-LASSO's if the sample size is at least 200 in the first setting, while ESL-LASSO's shall be smaller than both LAD-LASSO's and CQR-LASSO's if the sample size is large enough in the other two settings.

The PSR is around 1 for all three methods in all settings. Therefore, the performance of the three methods are comparable in terms of the model error and PSR. What distinguishes ESL-LASSO from LAD-LASSO and CQR-LASSO is NSR. Indeed, the NSR of the ESL-LASSO estimator is as close 1 as that of the oracle estimator, while the NSR of the LAD-LASSO and CQR-LASSO ranges from 0.431 to 0.738, and from 0.675 to 0.988, respectively. These simulation experiments suggest that the ESL-LASSO estimator leads to a consistent variable selection in common situations where outliers result from either the response or the covariate domain.

So far, we only considered $d = 8$. Next, we run a simulation study for $d = 100$. The true coefficients are set to be $\boldsymbol{\beta} = (1, 1.5, 2, 1, \mathbf{0}_p)^T$, where $\mathbf{0}_p$ is p-dimensional row vector of zeros, and $p = 96$. Covariate $\mathbf{x}_i$ follows a multi-normal distribution $N(\mathbf{0}, \Omega_2)$, and the error term follows a mixture normal distribution $0.7N(0, 1) + 0.3N(5, 3^2)$. Since our proposed method requires that p is fixed and less than n, we take $n = 1000$. We replicate the simulation 500 times to evaluate the finite-sample performance of ESL-LASSO. The simulation results are summarized in Table 4. This table reveals that ESL-LASSO still performs better in variable selection and has much lower model error than LAD-LASSO and CQR-LASSO.

4.2 Applications

Example 1—We apply the proposed methodology to analyze the Boston Housing Price Dataset which is available on http://lib.stat.cmu.edu/datasets/boston. This dataset was presented by Harrison and Rubinfeld (1978), and is commonly used as an example for regression diagnostics (Belsley et al. (1980)).

The data contains the following 14 variables: crim (per capita crime rate by town), zn (proportion of residential land zoned for lots over 25,000 sq.ft), indus (proportion of non-retail business acres per town), chas (Charles River dummy variable: equal to 1 if tract bounds river; 0 otherwise), nox (nitrogen oxides concentration: parts per 10 million), rm (average number of rooms per dwelling), age (proportion of owner-occupied units built prior to 1940), dis (weighted mean of distances to five Boston employment centres), rad (index of accessibility to radial highways), tax (full-value property-tax rate per ten thousand dollar), ptratio (pupil-teacher ratio by town), black $(1000(Bk - 0.63)^2$, where Bk is the proportion of blacks by town), lstat (lower status of the population (percent)), and medv (median value of owner-occupied homes in thousand dollars). There are 506 observations in the dataset. The response variable is medv, and the rest are the predictors. In this section, the predictors are scaled to have mean zero and unit variance.

Table 5 compares the estimates of the regression coefficients from the MM method and ordinary least-squares (OLS) besides ESL-LASSO, CQR-LASSO and LAD-LASSO. The selected variables and their coefficients are clearly different among the five methods.

With the proposed ESL-LASSO procedure, Steps 1 and 2 of the selection procedure first chooses the tuning γ_n at 2.1, which corresponds to $m = 8$, the number of ”bad points” in the observations. Table 5 reveals that ESL-LASSO selects four of the six variables selected by both CQR-LASSO and LAD-LASSO. The four variables are rm, tax, ptratio, and lstat. The two variables selected by LAD-LASSO but not by CQR-LASSO and ESL-LASSO are dis and black.

Example 2—As another illustration, we analyze the Plasma Beta-carotene Level Dataset from a cross-sectional study. This dataset consists of 315 sample, in which there are 273 female patients, and can be downloaded at http://lib.stat.cmu.edu/datasets/Plasma Retinol.

In this example, we only analyze the 273 female patients. Of interest are to study the relationships between the plasma beta-carotene level (betaplasma) and the following 10 covariates: age, smoking status (smokstat), quetelet, vitamin use (vituse), number of calories consumed per day (calories), grams of fat consumed per day (fat), grams of fiber consumed per day (fiber), number of alcoholic drinks consumed per week (alcohol), cholesterol consumed (cholesterol), and dietary beta-carotene consumed (betadiet).

We plot a histogram of betaplasma and cholesterol in Figure 2. Figure 2 indicates that there are unusual points in either the response or the covariate domain. In the following, the predictors are scaled to have zero mean and unit variance. We use the first 200 sample as a training data set to fit the model, and then use the remaining 73 sample to evaluate the predictive ability of the selected model. The prediction performance is measured by the median absolute prediction error (MAPE).

Next, we apply ESL-LASSO, CQR-LASSO and LAD-LASSO to analyze the plasma beta-carotene level dataset. For ESL-LASSO, we first apply the proposed procedure to choose the tuning parameter γ_n, and obtain $\gamma_n = 0.190$. The results are summarized in Table 6. This table reveals that ESL-LASSO selects “fiber” that is also selected by CQR-LASSO and LAD-LASSO. In addition, CQR-LASSO selects “quetelet” and LAD-LASSO selects “batadiet”, and these two different choices are not selected by ESL-LASSO. Although ESL-LASSO selects one fewer covariate than LAD-LASSO, ESL-LASSO has a slightly smaller MAPE. ESL-LASSO also selects one fewer covariate than CQR-LASSO, the MAPE of ESL-LASSO is 10% higher.

For further comparisons, we apply a combination of the bootstrap and cross validation Figure 2 method to obtain the standard errors of the estimates for the number of non-zeros and the model errors for two real datasets. For each bootstrap sample, we randomly split the 506 observations into training and testing sets in the Boston Housing Price Dataset of sizes 300 and 206, respectively. For the plasma beta-carotene level dataset, we randomly split the 273 observations into training and testing datasets of sizes 200 and 73, respectively. For ESL-LASSO, CQR-LASSO and LAD-LASSO, we calculate the median of absolute prediction error based on $|y - x^T\hat{\beta}|$ and MAD of prediction error based on $y - x^T\hat{\beta}$ for the test set, respectively. The average errors and the average numbers of estimated nonzero coefficients over 200 repetitions are summarized in Table 7. The standard deviations are given in their corresponding parentheses. ESL-LASSO tends to select the fewest number of non-zeros.

5. DISCUSSION

In this paper, we proposed a robust variable selection procedure via a penalized regression with the exponential squared loss. We investigated the sampling properties and studied the robustness properties of the proposed estimators. Through the theoretical and simulation results, we demonstrate the merits of our proposed method. We also illustrate that our proposed method can result in notable difference in real data analysis. Specifically, we show that our estimators possessed the highest finite sample breakdown point, and the influence functions are bounded with respect to outliers in either the response or the covariate domain.

Although Theorem 2 requires that the penalty has the form $p_{\lambda_{nj}}(|\beta_j|) = \lambda_{nj}g(|\beta_j|)$, where $g(\cdot)$ is a strictly increasing and unbounded function in $[0, \infty]$. We conjecture that the breakdown point result also holds for bounded penalties. Intuitively, we may use the one-step or k-step local linear approximation (LLA) proposed by Zou and Li (2008), where $p_{\lambda_{\mathrm{nj}}}(|\beta_j|) \approx p_{\lambda_{\mathrm{nj}}}(|\beta_j^{(0)}|)+p'_{\lambda_{\mathrm{nj}}}(|\beta_j^{(0)}|)(|\beta_j| - |\beta_j^{(0)}|), \text{ for } \beta_j \approx \beta_j^{(0)}$. LLA provides a reasonable approximation in the sense that it posses the ascent property of the MM algorithm (Hunter and Lange, 2004), and yields the best convex majorization. Furthermore, the one-step LLA enjoys the oracle property. For the case where all $p'_{\lambda_{\mathrm{nj}}}(|\beta_j^{(0)}|)>0$, Theorem 2 is applicable to LLA. For penalties where the derivative equals zero for some j (e.g., SCAD), the LLA can be implemented by using Algorithm 2 as suggested in Section 4 of Zou and Li (2008).

How to select both γ_n and regularization parameters λ_{nj} in a data-driven way is a difficult problem, since selection of γ_n depends on the choice of λ_{nj} and an estimate of $\boldsymbol{\beta}$. Although the data-adaptive methods such as cross-validation can be applied, it could cause huge computation and may not satisfy condition (C2) of Remark 1. To implement our proposed ESL-LASSO, we consider a relative simple approach. We first choose regularization parameters λ_{nj} via a simple BIC criterion, and then proposed a data-driven approach to selecting the tuning parameter γ_n. We demonstrate the advantages of our methodology via simulation study and application. According to our simulation studies, the performance of our ESL-LASSO implementation is comparable to the oracle procedure irrespective of the presence and the mechanisms of outliers. We measure the performance by the model error, positive selection rate, and the non-causal selection rate. In the presence of outliers (regardless of the mechanisms), LAD-LASSO and CQR-LASSO perform poorly in terms of the non-causal selection rate.

After re-analyzing the Boston Housing Price Dataset and the Plasma Beta-Carotene Level Dataset, we find that the results and interpretation can be quite different with different methods. Although we do not know the real model from which the real data were generated, our simulation results suggest that a plausible cause of the discrepancies among ESL-LASSO, CQR-LASSO and LAD-LASSO is that there are outliers in the datasets. We refer to Harrison and Rubinfeld (1978) on the identification of influential points for these datasets. In short, our proposed ESL-LASSO is expected to produce a more reliable model without the need to identifying the specific influential points.

Acknowledgments

Huang's research is partially supported by a funding through Projet 211 Phase 3 of SHUFE, and Shanghai Leading Academic Discipline Project, B803. Wang's research is partially supported by NSFC(11271383), RFDP(20110171110037), SRF for ROCS, SEM, Fundamental Research Funds for the Central Universities. Zhang's research is partially supported by grant R01DA016750-08 from the U.S. National Institute on Drug Abuse.

APPENDIX

Proof of Theorem 1 (i). Let $\xi_n = n^{-1/2} + a_n$. We first prove that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|\mathbf{u}\|=C}\ell_n(\beta_0+\xi_n\mathbf{u})<\ell_n(\beta_0)\right\} \geq 1-\varepsilon, \quad \text{(A.1)}$$

where $\mathbf{u}$ is d-dimensional vector such that $\|\mathbf{u}\| = C$. Then, we prove that there exists a local maximizer $\hat{\boldsymbol{\beta}}_n$ such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(\xi_n)$. Let

$$D_n(\beta,\gamma)=\sum_{i=1}^{n}\exp\left\{-(Y_i-\mathbf{x}_i^T\beta)^2/\gamma\right\}\frac{2(Y_i-\mathbf{x}_i^T\beta)}{\gamma}\mathbf{x}_i.$$

Since $p_{\lambda_{nj}}(0) = 0$ for $j = 1, \cdots, d$ and $\gamma_n - \gamma_0 = o_p(1)$, by Taylor's expansion, we have

$$\begin{aligned}&\ell_n(\beta_0+\xi_n\mathbf{u})\\ &-\ell_n(\beta_0)\leq\xi_nD_n(\beta_0,\gamma_n)^T\mathbf{u}\\ &-\tfrac{1}{2}\mathbf{u}^T\left[-I\left(\beta_0,\gamma_n\right)\right]\mathbf{u}n\xi_n^2\{1\\ &+o_p(1)\}\quad-\sum_{j=1}^{s}\left[n\xi_np'_{\lambda_{nj}}(|\beta_{0j}|)\text{sign}(\beta_{0j})u_j\quad+n\xi_n^2p''_{\lambda_{nj}}(|\beta_{0j}|)u_j^2\{1\quad+o(1)\}\right]=\xi_n\left\{D_n(\beta_0,\gamma_0)+o_p(\sqrt{n})\right\}^T\mathbf{u}-\tfrac{1}{2}\mathbf{u}^T\left[-I(\beta_0,\gamma_0)+o(1\right.\end{aligned}$$

Note that $n^{-1/2}D_n(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = O_p(1)$. Therefore, the order of the first term on the right side is equal to $O_p(n^{1/2}\xi_n)=O_p(n\xi_n^2)$ in the last equation of (A.2). By choosing a sufficiently large C, the second term dominates the first term uniformly in $\|\mathbf{u}\| = C$. Meanwhile, the third term in (A.2) is bounded by

$$\sqrt{s}n\xi_n a_n\|\mathbf{u}\|+n\xi_n^2 b_n\|\mathbf{u}\|^2.$$

Since $b_n = o_p(1)$, the third term is also dominated by the second term of (A.2). Therefore, (A.1) holds by choosing a sufficiently large C. The proof of Theorem 1 (i) is completed.

Lemma 1 *Assume that the penalty function satisfies* (2.3). *If* $\sqrt{n}(\gamma_n-\gamma_0)=o_p(1)$ *for some* $\boldsymbol{\gamma}_0 > 0$, $b_n = o_p(1)$, $\sqrt{n}a_n=o_p(1), \text{ and } 1/\min_{s+1\le j\le d}(\sqrt{n}\lambda_{\text{nj}})=o_p(1)$*, then with probability 1, for any given* $\boldsymbol{\beta}$ *satisfying* $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| = O_p(n^{-1/2})$ *and any constant C,*

$$\ell_n(\beta_1,0)=\max_{\|\beta_2\|\le \text{C}n^{-1/2}}\ell_n(\beta_1,\beta_2).$$

Proof of Lemma 1. We will show that with probability 1, for any $\boldsymbol{\beta}_1$ satisfying $\boldsymbol{\beta}_1-\boldsymbol{\beta}_{01} = O_p(n^{-1/2})$, and for some small $\varepsilon_n = Cn^{-1/2}$ and $j = s + 1, \cdots, d$, we have $\partial\ell_n(\boldsymbol{\beta})/\partial\beta_j < 0$, for $0 < \beta_j < \varepsilon_n$, and $\partial\ell_n(\boldsymbol{\beta})/\partial\beta_j > 0$, for $-\varepsilon_n < \beta_j < 0$. Let

$$Q_n(\beta,\gamma)=\sum_{i=1}^{n}\exp\{-(Y_i-\mathbf{x}_i^T\beta)^2/\gamma\}. \qquad (5.3)$$

By Taylor's expansion, we have

$$\frac{\partial\ell_n(\beta)}{\partial\beta_j} = \frac{\partial Q_n(\beta,\gamma_n)}{\partial\beta_j} - \text{np}'_{\lambda_{\text{nj}}}(|\beta_j|)\text{sign}(\beta_j) = \frac{\partial Q_n(\beta_0,\gamma_n)}{\partial\beta_j} + \sum_{l=1}^{p}\frac{\partial^2 Q_n(\beta_0,\gamma_n)}{\partial\beta_j\partial\beta_l}(\beta_l - \beta_{l0}) + \sum_{l=1}^{p}\sum_{k=1}^{p}\frac{\partial^3 Q_n(\beta^{*},\gamma_n)}{\partial\beta_j\partial\beta_l\partial\beta_k}(\beta_l-\beta_{l0})(\beta_k - \beta_{k0}) - \text{np}'_{\lambda_{\text{nj}}}(|\beta_j|)\text{sign}($$

where $\boldsymbol{\beta}^{*}$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Note that

$$n^{-1}\frac{\partial Q_n(\beta_0,\gamma_0)}{\partial\beta_j}=O_p(n^{-1/2}),$$

and

$$n^{-1}\frac{\partial^2 Q_n(\beta_0,\gamma_0)}{\partial\beta_j\partial\beta_l}=E\left\{\frac{\partial^2 Q_n(\beta_0)}{\partial\beta_j\partial\beta_l}\right\}+o_p(1).$$

Since $b_n = o_p(1)$ and $\sqrt{n}a_n=o_p(1)$, we obtain $\boldsymbol{\beta}-\boldsymbol{\beta}_0 = O_p(n^{-1/2})$. By $\sqrt{n}(\gamma_n-\gamma_0)=o_p(1)$, we have

$$\frac{\partial\ell_n(\beta)}{\partial\beta_j}=n\lambda_{\text{nj}}\left\{-\lambda_{\text{nj}}^{-1}p'_{\lambda_{\text{nj}}}(|\beta_j|)\text{sign}(\beta_j)+O_p(n^{-1/2}/\lambda_{\text{nj}})\right\}.$$

Since $1/\min_{s+1\le j\le d}(\sqrt{n}\lambda_{\mathrm{nj}})=o_p(1)$, and $\lim_{n\to\infty}\inf\lim_{t\to 0+}\inf\left\{\min_{s+1\le j\le d}p'_{\lambda_{\mathrm{nj}}}(|t|)/\lambda_{\mathrm{nj}}\right\}>0$ with probability 1, the sign of the derivative is completely determined by that of β_j. This completes the proof of Lemma 1.

Proof of Theorem 1 (ii). Part (a) holds by Lemma 1. We have showed that there exists a root-n consistent local maximizer of $\ell_n\{(\beta_1, 0)\}$, satisfying that

$$\frac{\partial\ell_n\{(\hat{\beta}_{n1},0)\}}{\partial\beta_j}=0, \text{ for } j=1,\cdots,s. \quad (5.4)$$

Since $\hat{\beta}_{n1}$ is a consistent estimator, we have

$$\frac{\partial Q_n\{(\hat{\beta}_{n1},0),\gamma_n\}}{\partial\beta_j} - np'_{\lambda_{\mathrm{nj}}}(|\hat{\beta}_j|)\mathrm{sign}(\hat{\beta}_j) = \frac{\partial Q_n(\beta_0,\gamma_n)}{\partial\beta_j} + \sum_{l=1}^{s}\left\{\frac{\partial^2 Q_n(\beta_0,\gamma_n)}{\partial\beta_j\partial\beta_l} + o_p(1)\right\}(\hat{\beta}_l-\beta_{01}) - n\left[p'_{\lambda_{\mathrm{nj}}}(|\beta_{0j}|)\mathrm{sign}(\beta_{0j}) + \left\{p''_{\lambda_{\mathrm{nj}}}(|\beta_{0j}|)+o_p(1)\right\}(\hat{\beta}_j-\beta_{0j})\right],$$

where $Q_n(\beta, \gamma)$ is defined in (5.3). Since $\sqrt{n}(\gamma_n-\gamma_0)=o_p(1)$, the proof of part (b) is completed by Slutsky Lemma and the central limit theorem.

Lemma 2 *Let* $\mathbf{D}_n = \{D_1, \cdots, D_n\}$ *be any sample of size n,* $\mathbf{D}_m = \{D_1, \cdots, D_m\}$ *be a contaminating sample of size m, and* $a_{\mathrm{nm}}=\max_{\beta\in\mathbb{R}^d}\#\{i: m+1\le i\le n \text{ and } \mathbf{x}_i^T\beta=0\}/(n-m)$. Assume

$$a_{\mathrm{nm}}<0.5, \varepsilon<(1-2a_{\mathrm{nm}})/(2-2a_{\mathrm{nm}}) \text{ and } \zeta(\gamma_n)<(1-\varepsilon)(2-2a_{\mathrm{nm}}).$$

For the weighted vector $\lambda = (\lambda_{n1}, \cdots, \lambda_{nd})$*, if*

$$0<\min_{\{j:1\le j\le d\}}\lambda_{\mathrm{nj}}<+\infty,$$

there exists a C such that $m/n \le \varepsilon$ *implies*

$$\inf_{\|\beta\|\ge C}\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_{\gamma_n}\{r_i(\beta)\}+\sum_{j=1}^{d}\lambda_{\mathrm{nj}}g(|\beta_j|)\right\}>\frac{1}{n}\sum_{i=1}^{n}\phi_{\gamma_n}\{r_i(\check{\beta}_n)\}+\sum_{j=1}^{d}\lambda_{\mathrm{nj}}g(|\check{\beta}_{\mathrm{nj}}|), \quad (5.5)$$

where $r_i(\beta)=Y_i-\mathbf{x}_i^T\beta$, $\check{\beta}_n$ *is an initial estimator of* β *and* $\phi_\gamma(t) = 1 - \exp(-t^2/\gamma)$.

Proof of Lemma 2. By definition of a_{nm}, we have

$$\#\{i: m+1\le i\le n \text{ and } |\mathbf{x}_i^T\beta|>0\}/(n-m)\ge 1-a_{\mathrm{nm}},$$

for all β. Since $\varepsilon < (1 - 2a_{nm})/(2 - 2a_{nm})$ and $\zeta(\gamma_n) < (1 - \varepsilon)(2 - 2a_{nm})$, there exist $a_{n1} > a_{nm}$ and $a_{n2} > a_{nm}$ such that $\varepsilon < (1 - 2a_{n1})/(2 - 2a_{n1})$ and $\zeta(\gamma_n) < (1 - \varepsilon)(2-2a_{n2})$.

Take $a_n^*=\min\{a_{n1},a_{n2}\}>a_{\mathrm{nm}}$, we have

$$\varepsilon<(1-2a_n^*)/(2-2a_n^*)\ \text{and}\ \zeta(\gamma_n)<(1-\varepsilon)\,(2-2a_n^*).$$

Since $a_n^*>a_{\mathrm{nm}}$, by using a compacity argument (Yohai (1987)), we can find δ > 0 such that

$$\inf_{\|\beta\|=1}\#\{i:m+1\leq i\leq n\ \text{and}\ |\mathbf{x}_i^T\beta|>\delta\}/(n-m)\geq 1-a_n^*.$$

Since $1-\varepsilon>1/(2-2a_n^*)$, we can find η such that $(1-\varepsilon)(1-a_n^*)>1-\eta>1/2$. Take Δ which satisfies

$$1<1+\Delta<\min\left\{\frac{(1-\varepsilon)\,(2-2a_n^*)}{\zeta(\gamma_n)},\frac{(1-\varepsilon)\,(1-a_n^*)}{1-\eta}\right\},$$

and take $a_0=\dfrac{(1-\eta)\zeta(\gamma_n)(1+\Delta)}{(1-\varepsilon)(2-2a_n^*)}$. We then have $a_0<\min\{1-\eta,\zeta(\gamma_n)/2\}$. Then $m/n\leq\varepsilon$ implies

$$a_0(n-m)/n\geq(1-\varepsilon)a_0>(1-\eta)\zeta(\gamma_n)/(2-2a_n^*).$$

Because $\phi_{\gamma_n}(t)$ is a continuous, bounded, and even function, there exists a $k_2\geq 0$ such that $\phi_{\gamma_n}(k_2)=a_0/(1-\zeta)$. Let $C_1\geq(k_2+\max_{m+1\leq i\leq n}|Y_i|)/\delta$. Then $m/n\leq\varepsilon$ implies

$$\inf_{\|\beta\|\geq C_1}\sum_{i=1}^{n}\phi_{\gamma_n}\{r_i(\beta)\}\geq\inf_{\|\beta\|=1}\sum_{i\in A}\phi_{\gamma_n}(|Y_i|-C_1|\mathbf{x}_i^T\beta|)\geq(n-m)(1-a_n^*)\phi_{\gamma_n}(k_2)=(n-m)(1-a_n^*)a_0/(1-\eta)>n\zeta(\gamma_n)/2\geq\sum_{i=1}^{n}\phi_{\gamma_n}\{r_i(\breve{\beta}_n)\},\quad(5.6)$$

where $A=\{i:m+1\leq i\leq n\ \text{and}\ |\mathbf{x}_i^T\beta|>\delta\}$.

Let $C_2=\sqrt{p}q^{-1}\{\sum_{j=1}^{d}\lambda_{\mathrm{nj}}g(|\breve{\beta}_j|)/(\min\{\lambda_{\mathrm{ni}}\})\}$. Take $C=\max\{C_1,C_2\}$. We have

$$\inf_{\|\beta\|\geq C}\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_{\gamma_n}\{r_i(\beta)\}\qquad+\sum_{j=1}^{d}\lambda_{\mathrm{nj}}g(|\beta_j|)\right\}\geq\inf_{\|\beta\|\geq C}\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_{\gamma_n}\{r_i(\beta)\}\right\}+\inf_{\|\beta\|\geq C}\left\{\sum_{j=1}^{d}\lambda_{\mathrm{nj}}g(|\beta_j|)\right\}\geq\inf_{\|\beta\|\geq C_1}\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_{\gamma_n}\{r_i(\beta)\}\right\}\qquad+\inf_{\|\beta\|\geq C_2}\left\{\sum_{j=}^{d}\right.$$

By (5.6), we only need to deal with the second term. Note that $\|\beta\|\geq C_2$ implies that there exists an element β_j of **β** such that $|\beta_j|\geq C_2/\sqrt{p}$ for some *j*. Hence,

$$\inf_{\|\beta\|\geq C_2}\left\{\sum_{j=1}^{d}\lambda_{\mathrm{nj}}g(|\beta_j|)\right\}\geq\lambda_{\mathrm{nj}}g(|\beta_j|)\geq\sum_{j=1}^{d}\lambda_{\mathrm{nj}}g(|\breve{\beta}_{\mathrm{nj}}|).$$

Proof of Theorem 2. For a contaminated sample $\mathbf{D}_n$, and $m/n\leq\varepsilon$, according to Lemma 2, if $\|\hat{\boldsymbol{\beta}}_n\|\geq C$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\phi_{\gamma_n}\{r_i(\hat{\beta}_n)\}+\sum_{j=1}^{d}\lambda_{\mathrm{nj}}g(|\hat{\beta}_{\mathrm{nj}}|)>\frac{1}{n}\sum_{i=1}^{n}\phi_{\gamma_n}\{r_i(\check{\beta}_n)\}+\sum_{j=1}^{d}\lambda_{\mathrm{nj}}g(|\check{\beta}_{\mathrm{nj}}|).$$

This is a contradiction to the fact that $\hat{\beta}_n$ minimizes $\{\frac{1}{n}\sum_{i=1}^{n}\phi_{\gamma_n}\{r_i(\beta)\}+\sum_{j=1}^{d}\lambda_{\mathrm{nj}}g(|\beta_j|)\}$ for $\beta \in \mathbb{R}^d$. Therefore, we have

$$\mathrm{BP}(\hat{\beta}_n;\mathbf{D}_{n-m},\gamma_n)\geq\min\left\{\mathrm{BP}(\check{\beta}_n;\mathbf{D}_{n-m}),(1-2a_{\mathrm{nm}})/(2-2a_{\mathrm{nm}}),1-\frac{\zeta(\gamma_n)}{2-2a_{\mathrm{nm}}}\right\}.$$

Proof of Theorem 3. Note that

$$(1-\varepsilon)\int\left[\exp\left\{-(y-\mathbf{x}^T\beta_\varepsilon^*)^2/\gamma_0\right\}\left(\frac{2}{\gamma_0}(y-\mathbf{x}^T\beta_\varepsilon^*)\right)\mathbf{x}\right]\mathrm{dF}(\mathbf{x},y)+\varepsilon\exp\left\{-(y_0-\mathbf{x}_0^T\beta_\varepsilon^*)^2/\gamma_0\right\}\left(\frac{2}{\gamma_0}(y_0-\mathbf{x}_0^T\beta_\varepsilon^*)\right)\mathbf{x}_0-\nu_1(\varepsilon)=0,\quad(5.7)$$

where $\nu_1(\varepsilon)=(p'_{\lambda_{01}}(|\beta_{\varepsilon1}|)\mathrm{sign}(\beta_{\varepsilon1}),\cdots,p'_{\lambda_{0d}}(|\beta_{\varepsilon d}|)\mathrm{sign}(\beta_{\varepsilon d}))^T$.

Let $r_0=y_0-\mathbf{x}_0^T\beta_0^*$. Differentiating with respect to ε in both sides of (5.7) and letting ε → 0, we obtain

$$\begin{aligned}&\int\left[\exp\{-(y-\mathbf{x}^T\beta_\varepsilon^*)^2/\gamma_0\}\left(\frac{-4(y-\mathbf{x}^T\beta_\varepsilon^*)^2}{\gamma_0^2}\right)\frac{\partial}{\partial\varepsilon}(y-\mathbf{x}^T\beta_\varepsilon^*)\,\mathbf{x}+\exp\{-(y-\mathbf{x}^T\beta_\varepsilon^*)^2/\gamma_0\}\frac{\partial}{\partial\varepsilon}\left(\frac{2(y-\mathbf{x}^T\beta_\varepsilon^*)}{\gamma_0}\right)\mathbf{x}\right]\mathrm{dF}(\mathbf{x},y)\Big|_{\varepsilon=0}\\&-\frac{\partial\nu_1(\varepsilon)}{\partial\varepsilon}=-\frac{\exp\{-r_0^2/\gamma_0\}2r_0x_0}{\gamma_0}-\nu_2,\end{aligned}\quad(5.8)$$

where $\nu_2=\left(p'_{\lambda_{01}}(|\beta_{01}^*|)\mathrm{sign}(\beta_{01}^*),\cdots,p'_{\lambda_{0d}}(|\beta_{0d}^*|)\mathrm{sign}(\beta_{0d}^*)\right)^T$. By using (5.7) and (5.8), it can be shown that

$$\left(\frac{2A(\gamma_0)}{\gamma_0}-B_1\right)[\mathrm{IF}\{(\mathbf{x}_0,y_0),\beta_0^*\}]=-\frac{\exp\{-r_0^2/\gamma_0\}2r_0x_0}{\gamma_0}-\nu_2,\quad(5.9)$$

where

$$A(\gamma)=\int\mathrm{xx}^T\ \exp\{-(y-\mathbf{x}^T\beta_0^*)^2/\gamma\}\left(\frac{2(y-\mathbf{x}^T\beta_0^*)^2}{\gamma}-1\right)\mathrm{dF}(\mathbf{x},y),$$

$$B_1=\mathrm{diag}\left\{p''_{\lambda_{01}}(|\beta_{01}^*|)+p'_{\lambda_{01}}(|\beta_{01}^*|)\delta(\beta_{01}^*),\cdots,p''_{\lambda_{0d}}(|\beta_{0d}^*|)+p'_{\lambda_{0d}}(|\beta_{0d}^*|)\delta(\beta_{0d}^*)\right\},$$

with

$$\delta(x)=\begin{cases}+\infty & \text{if } x=0\\ 0, & \text{otherwise.}\end{cases}$$

This completes the proof of Theorem 3.

Proof of Corollary 1. Since $\tilde{\boldsymbol{\beta}}_n$ is a root-n consistent estimator and $\hat{\boldsymbol{\tau}}_{nj} = \log(n)/n$, we obtain

$$\max_{1 \le j \le s}\left(\sqrt{n}\hat{\tau}_{\mathrm{nj}}/|\tilde{\beta}_{\mathrm{nj}}|\right) = o_p(1), \text{ and } 1/\min_{s+1 \le j \le d}\left(\sqrt{n}\hat{\tau}_{\mathrm{nj}}/|\tilde{\beta}_{\mathrm{nj}}|\right) = o_p(1), \quad (5.10)$$

which satisfies the conditions of oracle property in Theorem 1. This completes the proof of Corollary 1.

REFERENCES

Belsley, D.; Kuh, E.; Welsch, R. Regression Diagnostics: Identifying Influential Data And Sources Of Collinearity. Wiley; 1980.

Bradic J, Fan J, Wang W. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011; 73(3): 325–349.

Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika. 2008; 95(3):759–771.

Donoho, D. Technical report. Boston: Harvard University; 1982. Breakdown properties of multivariate location estimators. Technical report

Donoho D, Huber P. The notion of breakdown point. A Festschrift for Erich L. Lehmann. 1983:157–184.

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of statistics. 2004; 32(2):407–499.

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

Frank I, Friedman J. A Statistical View of Some Chemometrics Regression Tools. Technometrics. 1993; 35(2):109–135.

Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. The Annals of Applied Statistics. 2007; 1(2):302–332.

Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. The Annals of Statistics. 2000; 28(2):337–407.

Gervini D, Yohai V. A class of robust and fully efficient regression estimators. The Annals of Statistics. 2002; 30(2):583–616.

Hampel, F. PhD thesis. Berkeley: University of California; 1968. Contributions to the theory of robust estimation.

Hampel F. A general qualitative definition of robustness. The Annals of Mathematical Statistics. 1971; 42(6):1887–1896.

Harrison D, Rubinfeld D. Hedonic prices and the demand for clean air. J. Environ. Economics and Management. 1978; 5:81–102.

He X, Simpson D. Lower bounds for contamination bias: Globally minimax versus locally linear estimation. The Annals of Statistics. 1993; 21(1):314–337.

Hunter D, Lange K. A tutorial on MM algorithms. The American Statistician. 2004; 58(1):30–37.

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

Kai B, Li R, Zou H. New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models. The Annals of Statistics. 2011; 39(1):305–332.

Leng C. Variable selection and coefficient estimation via regularized rank regression. Statistica Sinica. 2010; 20:167–181.

Rousseeuw P, Yohai V. Robust regression by means of S-estimators. Robust and Nonlinear Time Series. 1984; 26:256–272.

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

Tseng P, Yun S. A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming. 2009; 117(1):387–423.

Wang H, Li G, Jiang G. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. Journal of Business and Economic Statistics. 2007; 25(3):347–355.

Wang L, Li R. Weighted Wilcoxon-Type Smoothly Clipped Absolute Deviation Method. Biometrics. 2009; 65(2):564–571. [PubMed: 18647294]

Wu T, Lange K. Coordinate descent algorithms for lasso penalized regression. The Annals of Applied Statistics. 2008; 2(1):224–244.

Wu Y, Liu Y. Variable selection in quantile regression. Statistica Sinica. 2009; 19(2):801–817.

Yohai V. High breakdown-point and high efficiency robust estimates for regression. The Annals of statistics. 1987; 15(2):642–656.

Yohai V, Zamar R. High breakdown-point estimates of regression by means of the minimization of an efficient scale. Journal of the American Statistical Association. 1988; 83(402):406–413.

Zhang C. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics. 2010; 38(2):894–942.

Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101(476):1418–1429.

Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. The Annals of Statistics. 2008; 36(4):1509–1533.

Zou H, Yuan M. Composite quantile regression and the oracle model selection theory. The Annals of Statistics. 2008; 36(3):1108–1126.

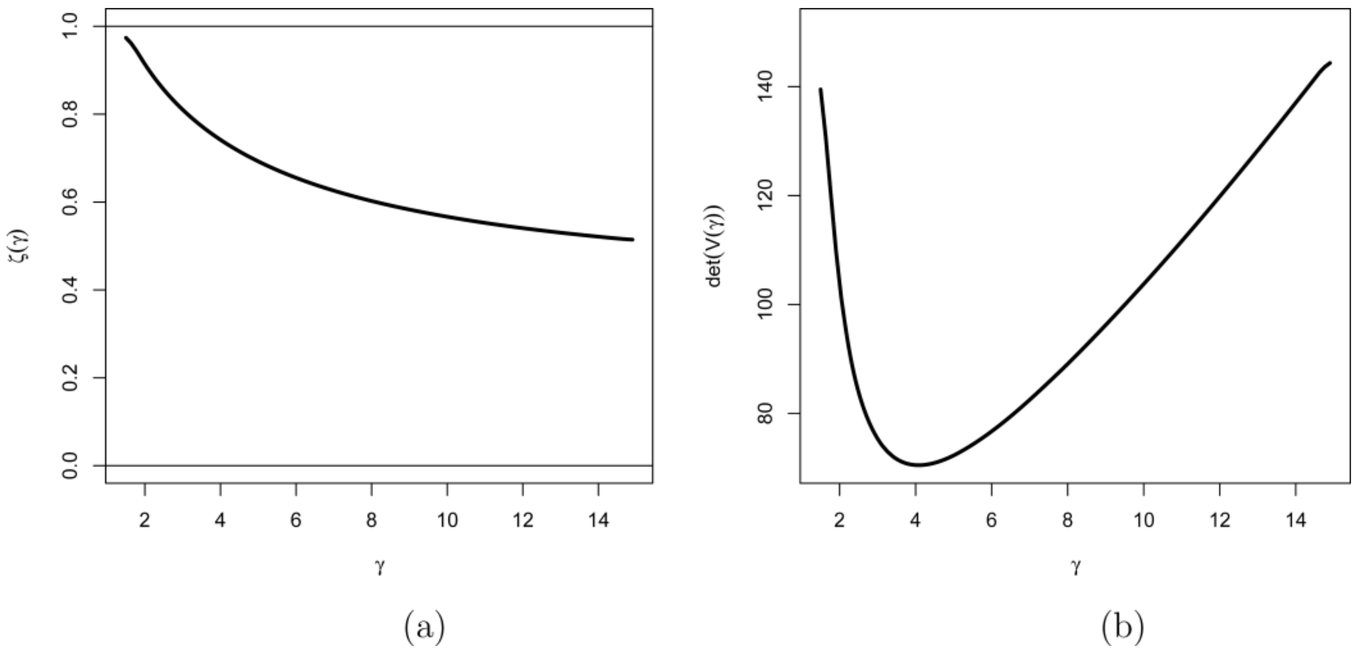

Figure 1.

(a) $\zeta(\gamma)$ against γ; (b) The determinant of matrix $\hat{V}(\gamma)$ against γ

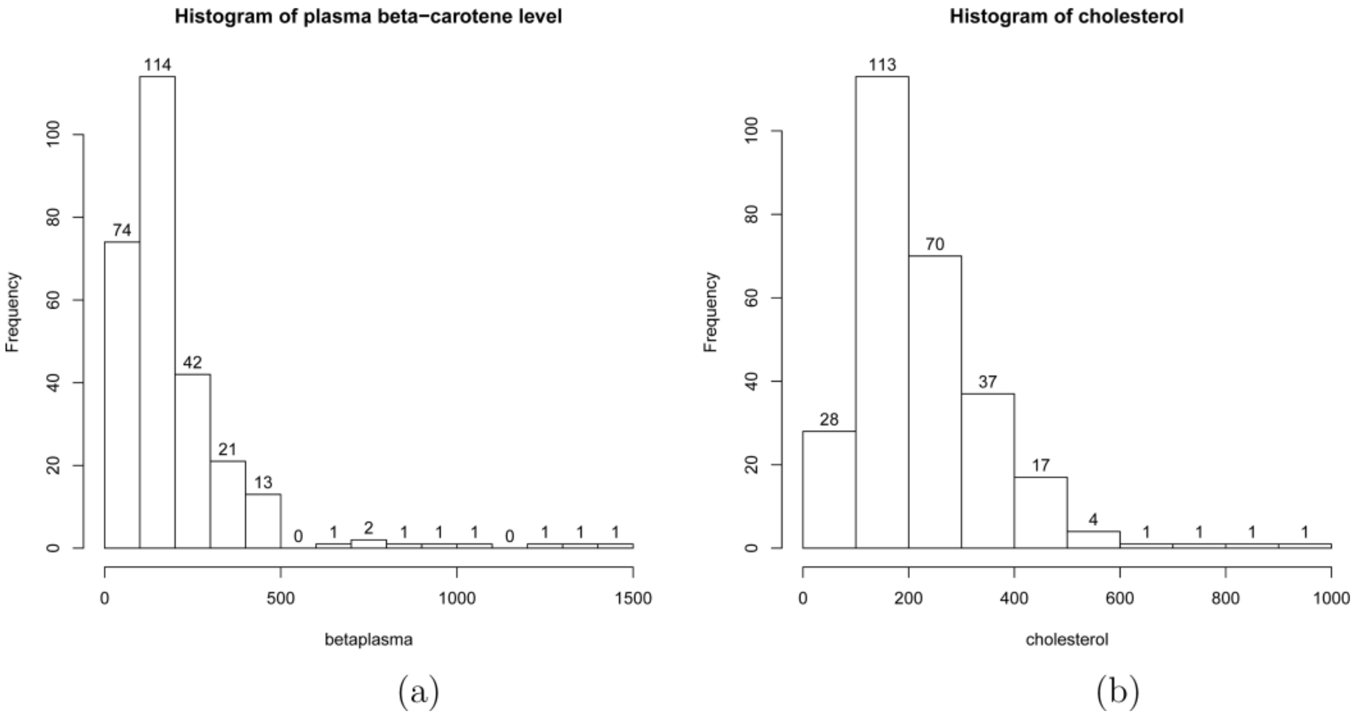

Figure 2.

Histogram of betaplasma (a) and cholesterol (b).

Table 1

Simulation results in the first setting

						Model error	
n	Method	$\bar{\gamma}_n$	$\zeta(\gamma_n)$	PSR	NSR	Median	MAD
100	ESL-LASSO	3.965	0.260	0.982	0.999	0.076	0.040
	CQR-LASSO	– –	– –	1.000	0.877	0.041	0.021
	LAD-LASSO	– –	– –	1.000	0.581	0.057	0.030
	Oracle	– –	– –	1.000	1.000	0.034	0.018
150	ESL-LASSO	4.303	0.282	0.999	1.000	0.038	0.020
	CQR-LASSO	– –	– –	1.000	0.906	0.026	0.013
	LAD-LASSO	– –	– –	1.000	0.554	0.035	0.016
	Oracle	– –	– –	1.000	1.000	0.022	0.011
200	ESL-LASSO	4.450	0.309	1.000	1.000	0.027	0.013
	CQR-LASSO	– –	– –	1.000	0.935	0.019	0.010
	LAD-LASSO	– –	– –	1.000	0.539	0.028	0.012
	Oracle	– –	– –	1.000	1.000	0.017	0.009
400	ESL-LASSO	4.500	0.331	1.000	1.000	0.012	0.006
	CQR-LASSO	– –	– –	1.000	0.966	0.010	0.005
	LAD-LASSO	– –	– –	1.000	0.498	0.0142	0.007
	Oracle	– –	– –	1.000	1.000	0.009	0.005
600	ESL-LASSO	4.500	0.337	1.000	1.000	0.008	0.004
	CQR-LASSO	– –	– –	1.000	0.980	0.006	0.003
	LAD-LASSO	– –	– –	1.000	0.480	0.009	0.005
	Oracle	– –	– –	1.000	1.000	0.006	0.003
800	ESL-LASSO	4.500	0.338	1.000	1.000	0.005	0.003
	CQR-LASSO	– –	– –	1.000	0.988	0.005	0.002
	LAD-LASSO	– –	– –	1.000	0.498	0.007	0.003
	Oracle	– –	– –	1.000	1.000	0.004	0.002

Table 2

Simulation results in the second setting

						Model error	
n	**Method**	$\bar{\gamma}_n$	$\zeta(\gamma_n)$	**PSR**	**NSR**	**Median**	**MAD**
100	ESL-LASSO	4.315	0.454	0.939	1.000	0.352	0.231
	CQR-LASSO	– –	– –	1.000	0.781	0.066	0.033
	LAD-LASSO	– –	– –	1.000	0.738	0.113	0.061
	Oracle	– –	– –	1.000	1.000	0.051	0.026
150	ESL-LASSO	4.375	0.629	0.995	1.000	0.148	0.075
	CQR-LASSO	– –	– –	1.000	0.739	0.066	0.033
	LAD-LASSO	– –	– –	1.000	0.737	0.070	0.034
	Oracle	– –	– –	1.000	1.000	0.035	0.017
200	ESL-LASSO	4.449	0.633	1.000	1.000	0.080	0.039
	CQR-LASSO	– –	– –	1.000	0.789	0.046	0.023
	LAD-LASSO	– –	– –	1.000	0.712	0.050	0.026
	Oracle	– –	– –	1.000	1.000	0.025	0.013
400	ESL-LASSO	4.496	0.638	1.000	1.000	0.027	0.012
	CQR-LASSO	– –	– –	1.000	0.864	0.021	0.010
	LAD-LASSO	– –	– –	1.000	0.686	0.023	0.011
	Oracle	– –	– –	1.000	1.000	0.012	0.006
600	ESL-LASSO	4.499	0.642	1.000	1.000	0.015	0.007
	CQR-LASSO	– –	– –	1.000	0.897	0.015	0.007
	LAD-LASSO	– –	– –	1.000	0.650	0.016	0.008
	Oracle	– –	– –	1.000	1.000	0.008	0.004
800	ESL-LASSO	4.499	0.642	1.000	1.000	0.009	0.005
	CQR-LASSO	– –	– –	1.000	0.910	0.010	0.006
	LAD-LASSO	– –	– –	1.000	0.633	0.011	0.005
	Oracle	– –	– –	1.000	1.000	0.006	0.003

Table 3

Simulation results in the third setting

						Model error	
n	**Method**	$\bar{\gamma}_n$	$\zeta(\gamma_n)$	**PSR**	**NSR**	**Median**	**MAD**
100	ESL-LASSO	4.565	0.662	1.000	0.989	0.174	0.101
	CQR-LASSO	– –	– –	1.000	0.675	0.173	0.097
	LAD-LASSO	– –	– –	1.000	0.488	0.113	0.061
	Oracle	– –	– –	1.000	1.000	0.098	0.050
150	ESL-LASSO	3.646	0.731	1.000	1.000	0.081	0.046
	CQR-LASSO	– –	– –	1.000	0.730	0.103	0.055
	LAD-LASSO	– –	– –	1.000	0.492	0.067	0.036
	Oracle	– –	– –	1.000	1.000	0.066	0.036
200	ESL-LASSO	3.654	0.722	1.000	1.000	0.058	0.030
	CQR-LASSO	– –	– –	1.000	0.778	0.068	0.031
	LAD-LASSO	– –	– –	1.000	0.487	0.051	0.025
	Oracle	– –	– –	1.000	1.000	0.049	0.023
400	ESL-LASSO	3.589	0.850	1.000	1.000	0.022	0.012
	CQR-LASSO	– –	– –	1.000	0.856	0.031	0.016
	LAD-LASSO	– –	– –	1.000	0.459	0.023	0.011
	Oracle	– –	– –	1.000	1.000	0.023	0.011
600	ESL-LASSO	3.590	0.717	1.000	1.000	0.014	0.007
	CQR-LASSO	– –	– –	1.000	0.897	0.020	0.010
	LAD-LASSO	– –	– –	1.000	0.431	0.014	0.007
	Oracle	– –	– –	1.000	1.000	0.016	0.008
800	ESL-LASSO	3.525	0.833	1.000	1.000	0.010	0.005
	CQR-LASSO	– –	– –	1.000	0.912	0.015	0.008
	LAD-LASSO	– –	– –	1.000	0.435	0.011	0.005
	Oracle	– –	– –	1.000	1.000	0.012	0.006

Table 4

Simulation results when $d = 100$

						Model error	
n	**Method**	$\bar{\gamma}_n$	$\zeta(\gamma_n)$	**PSR**	**NSR**	**Median**	**MAD**
1000	ESL-LASSO	4.441	0.736	1.000	1.000	0.010	0.005
	CQR-LASSO	– –	– –	1.000	0.866	0.030	0.013
	LAD-LASSO	– –	– –	1.000	0.661	0.022	0.008
	Oracle	– –	– –	1.000	1.000	0.010	0.005

Table 5

Estimated regression coefficients from the Boston Housing Price Data

	Method				
Variable	ESL-LASSO	CQR-LASSO	LAD-LASSO	MM	OLS
crim	0	0	0	−0.097	−0.101
zn	0	0	0	0.072	0.118
indus	0	0	0	−0.005	0.015
chas	0	0	0	0.038	0.074
nox	0	0	0	−0.097	−0.224
rm	0.590	0.422	0.503	0.491	0.291
age	0	0	0	−0.117	0.002
dis	0	−0.057	−0.013	−0.235	−0.338
rad	0	0	0	0.156	0.290
tax	−0.105	−0.133	−0.058	−0.208	−0.226
ptratio	−0.076	−0.153	−0.155	−0.179	−0.224
black	0	0.040	0.085	0.124	0.092
lstat	−0.131	−0.334	−0.243	−0.174	−0.408

Table 6

Estimated regression coefficients from the plasma beta-carotene level data

	Method		
Variable	**ESL-LASSO**	**CQR-LASSO**	**LAD-LASSO**
age	0	0	0
smokstat	0	0	0
quetelet	0	−0.057	0
vituse	0	0	0
calories	0	0	0
fat	0	0	0
fiber	0.114	0.077	0.058
alcohol	0	0	0
cholesterol	0	0	0
betadiet	0	0	0.075
MAPE	0.559	0.503	0.568

Table 7

Bootstrap results

			Model error	
Dataset	**Method**	**No. of non-zeros**	**Median**	**MAD**
Boston Housing Price	ESL-LASSO	3.710(0.830)	0.381(0.021)	0.180(0.069)
	CQR-LASSO	7.025(1.015)	0.286(0.016)	0.258(0.020)
	LAD-LASSO	5.020(0.839)	0.277(0.017)	0.113(0.075)
Plasma Beta-Carotene Level	ESL-LASSO	0.305(0.462)	0.459(0.030)	0.180(0.054)
	CQR-LASSO	2.915(1.026)	0.453(0.032)	0.299(0.050)
	LAD-LASSO	2.570(1.020)	0.429(0.036)	0.176(0.161)