Ann Inst Stat Math (2010) 62:487–514
DOI 10.1007/s10463-008-0184-2

Simultaneous estimation and variable selection in median regression using Lasso-type penalty

Jinfeng Xu · Zhiliang Ying

Received: 22 March 2007 / Revised: 2 April 2008 / Published online: 10 July 2008

Abstract We consider the median regression with a LASSO-type penalty term for variable selection. With the fixed number of variables in regression model, a two-stage method is proposed for simultaneous estimation and variable selection where the degree of penalty is adaptively chosen. A Bayesian information criterion type approach is proposed and used to obtain a data-driven procedure which is proved to automatically select asymptotically optimal tuning parameters. It is shown that the resultant estimator achieves the so-called oracle property. The combination of the median regression and LASSO penalty is computationally easy to implement via the standard linear programming. A random perturbation scheme can be made use of to get simple estimator of the standard error. Simulation studies are conducted to assess the finite-sample performance of the proposed method. We illustrate the methodology with a real example.

Keywords Variable selection · Median regression · Least absolute deviations · Lasso · Perturbation · Bayesian information criterion

1 Introduction

In the general linear model with independent and identically distributed errors, the Least Absolute Deviation (LAD) or L_1 method has been a viable alternative to the

J. Xu (✉)
Department of Statistics and Applied Probability, National University of Singapore,
Singapore 117546, Singapore
e-mail: staxj@nus.edu.sg

Z. Ying
Department of Statistics, Columbia University, New York, NY 10027, USA
e-mail: zying@stat.columbia.edu

least squares method especially for its superior robustness properties. Consider the linear regression model

$$Y_i = \beta_0^T x_i + e_i, \quad 1 \le i \le n, \tag{1}$$

where x_i are known p-vectors, β_0 the unknown p-vector of regression coefficients, and e_i the $i.i.d$ random errors with a common distribution F.

The L_1 estimator $\hat{\beta}_{L_1}$ is defined as a minimizer of the L_1 loss function

$$L_n(\beta) = \sum_{i=1}^{n} |Y_i - \beta^T x_i|. \tag{2}$$

Although there is no explicit analytic form for $\hat{\beta}_{L_1}$, the minimization may be carried out easily via linear programming (see for example, Koenker and D'Orey 1987). The more recent paper by Portnoy and Koenker (1997) gives speedy ways to compute the L_1 minimization, even for very large problems.

An important aspect in (regression) model building is model (variable) selection. For the least squares-based regression, there are a number of well established methods, including the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), Mallows's C_p and etc. These approaches are characterized by two basic elements: a goodness of fit measure and a complexity index. The selection criteria are typically based upon various ways of balancing the two elements, so the resultant prediction errors are minimized. Remarkably, for the L_1 based regression, there have been very few works on model selection. Hurvich and Tsai (1990) proposed a (double-exponential) likelihood function-based criterion and studied its small sample properties via simulations. Robust modification of Mallows's C_p were proposed by Ronchetti and Staudte (1994) under the framework of M estimation. Developing general variable selection methods with sound theoretical foundation and feasible implementation for L_1-based regression remains a great challenge.

An intriguing and novel recent advancement in variable selection is known as Basis Pursuit, proposed by Chen and Donoho (1994) or the Least Absolute Shrinkage and Selection Operator (LASSO), proposed by Tibshirani (1996). In it, the estimation and model selection are simultaneously treated as a single minimization problem. Knight and Fu (2000) established some asymptotic results for LASSO-type estimators. Fan and Li (2001) introduced Smoothly Clipped Absolute Deviation (SCAD) approach and proved its optimal properties. Efron et al. (2004) introduced Least Angel Regression (LARS) algorithm and its close connection to LASSO is extensively discussed.

With the fixed p and the invertible $X^T X$, the least squares estimate $\beta^{LS} = (X^T X)^{-1} X^T Y$ uniquely minimizes the squared loss

$$\sum_{i=1}^{n} (Y_i - \beta^T x_i)^2.$$

LASSO estimate is defined as the minimizer of

$$\sum_{i=1}^{n}(Y_i - \beta^T x_i)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \le s * \sum_{j=1}^{p}|\beta_j^{LS}|,$$

where $0 \le s \le 1$ controls the amount of shrinkage that is applied to the estimates.

LASSO is similar in form to the ridge regression where the term in the constraint is β_j^2 rather than $|\beta_j|$. A remarkable feature of LASSO , as a result of the L_1 constraint, is that for some β_j's, their fitted values are exactly 0. In fact, as the shrinkage parameter s goes from 1 to 0, the estimates go from no 0 to all 0. LASSO can also be regarded as a penalized least squares estimator with L_1 penalty: a minimizer of the objective function:

$$\sum_{i=1}^{n}(Y_i - \beta^T x_i)^2 + \lambda_n \sum_{j=1}^{p}|\beta_j|,$$

where λ_n is the tuning parameter.

In the present paper, we propose a parallel approach borrowing the ideas from LASSO by using the L_1 penalty, but with the least squares loss replaced by the L_1 loss. In doing so, we gain advantages in two fronts. First, it allows us to penetrate the difficult problem of variable selection for the L_1 regression. Appealingly, the shrinkage property of the LASSO estimator continues to hold in L_1 regression, see Fig. 1. Second, the single criterion function with both components being of L_1-type reduces (numerically) the minimization to a strictly linear programming problem, making any resulting methodology extremely easy to implement.

To be specific, our proposed estimator is a minimizer of the following criterion function

$$\sum_{i=1}^{n}|Y_i - \beta^T x_i| + \lambda_n \sum_{j=1}^{p}|\beta_j|.$$

It can be equivalently defined as a minimizer of the objective function

$$\sum_{i=1}^{n}|Y_i - \beta^T x_i| \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \le s * \sum_{j=1}^{p}|\beta_j^{L_1}|,$$

where β^{L_1} is the usual L_1 estimator.

As pointed out by Fan and Li (2001), LASSO does not possess the so-called oracle property in the sense that it cannot simultaneously have the best rate of convergence while correctly, with probability tending to one as sample size increases, set all unnecessary coefficients to 0. With this in mind, they proposed a variant of the penalty, called SCAD, smoothly clipped absolute deviation. Using such penalty, they were able to achieve the oracle property for the resulting estimator. Unfortunately, if we modify our approach by replacing the L_1 penalty function with the SCAD function, then the resultant minimization will be much more complicated. In particular, it is no longer numerically solvable by the linear programming.

To maintain numerical simplicity and uniqueness of solution of the linear programming, and to achieve the desirable oracle property, it is necessary for us to modify and extend the LASSO-type objective function. The tuning parameter λ_n there plays a crucial role of striking a balance between estimation of β and variable selection. Large values of λ_n tend to remove variables and increase bias in the estimation aspect while small values tend to retain variables. Thus it would be ideal that a large λ_n be used if a regression parameter is 0 (to be removed) and a small value be used if it is not 0. To this end, it becomes clear that we need a separate λ_n for each parameter component β_j. In other words, we need to consider the estimator as a minimizer of the objective function:

$$\sum_{i=1}^{n} |Y_i - \beta' X_i| + \sum_{j=1}^{p} \lambda_{nj} |\beta_j|.$$

In particular, we are interested in the case that $\lambda_{nj} = \eta_n \xi_{nj}$, where ξ_{nj} are fixed weighting parameter. Likewise, the estimator can be regarded as a minimizer of

$$\sum_{i=1}^{n} |Y_i - \beta^T x_i| \quad \text{subject to} \quad \sum_{j=1}^{p} \xi_{nj} |\beta_j| \le s * \sum_{j=1}^{p} \xi_{nj} |\beta_j^{L_1}|.$$

In the Bayesian view, the β_js have independent prior distributions- double exponential

$$f(\beta_j) = \frac{\lambda_{nj}}{2} \exp(-\lambda_{nj} |\beta_j|).$$

With proper choice of tuning parameters, we will show the resultant penalized estimator exhibit optimal properties. After we obtained our results (Xu 2005), we noticed a recent work of Adaptive LASSO by Zou (2006) which has the same spirit in proposing different scaling parameters in LASSO for fixed p and squared loss. However, our work is motivated by a unified L_1 based approach for simultaneous estimation and variable selection and focus on theoretical investigation of the resultant data-driven procedure for the absolute deviation loss.

The rest of the paper is organized as follows. In Sect. 2, we introduce some notations for L_1 regression, list some conditions under which our main results hold and establish a useful proposition. In Sect. 3, asymptotics for the estimator are considered. The conditions under which consistency or $\sqrt{n}$-consistency hold are given and limiting distribution results are proved. In Sect. 4, for properly chosen tuning parameters, we establish the oracle property of the estimator and use the perturbation method to estimate the standard error of the estimator. A two-stage data-driven procedure is also provided and proved to automatically select asymptotically optimal tuning parameters. In Sect. 5, simulation study as well as real data application are conducted to examine the performance of the proposed approach.

2 Differentially penalized L_1 estimator

We define the differentially penalized L_1 estimator $\hat{\beta}$ as a minimizer of the objective function

$$Z_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \beta^T x_i| + \frac{1}{n}\sum_{j=1}^{p} \lambda_{nj}|\beta_j|, \tag{3}$$

where λ_{nj}, $1 \le j \le p$ are regularization parameters.

We need to make the following assumptions on the error distribution and the covariates. These assumptions are essentially the same as those made in Pollard (1991) and Rao and Zhao (1992).

(C.1) $\{e_i\}$ are $i.i.d$ with median 0 and a density function $f(\cdot)$ which is continuous and strictly positive in a neighborhood of 0.

(C.2) $\{x_i\}$ is a deterministic sequence and there exists a positive definite matrix V for which $\frac{1}{n}V_n^2 = \frac{1}{n}\sum_{i\le n} x_i x_i^T \to V^2$.

Now we introduce some notations. Let the true coefficient vector $\beta_0 = \binom{\beta_0^1}{\beta_0^2}$, where β_0^1 is s-vector and β_0^2 is $(p-s)$-vector. Without loss of generality, assume $1 \le s < p$, $\beta_0^2 = 0$. Considering only the first s covariates, by (C.2), we have $\frac{1}{n}V_{n1}^2 = \frac{1}{n}\sum_{i\le n} x_i^1 x_i^{1^T} \to V_1^2$, where x_i^1 is the subvector of x_i which contains the first s components.

Denote $G_n(\beta) = \frac{1}{n}\sum\limits_{i=1}^{n} E(|Y_i - x_i^T\beta| - |Y_i - x_i^T\beta_0|)$, $R_i(\beta) = |Y_i - x_i^T\beta| - |Y_i - x_i^T\beta_0| + (\beta-\beta_0)^T x_i \operatorname{sgn}(e_i)$, then

$$\begin{aligned}\frac{1}{n}(L_n(\beta) - L_n(\beta_0)) &= -\frac{1}{n}\sum_{i=1}^{n}(\beta-\beta_0)^T x_i \operatorname{sgn}(e_i) + G_n(\beta)\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}[R_i(\beta) - ER_i(\beta)].\end{aligned}$$

In order to study the asymptotical properties of the penalized estimator, we need to establish the Local Asymptotic Quadratic (LAQ) property of the loss function (2).

Proposition 1 *Under* (C.1)–(C.2)*, for every sequence $d_n > 0$ with $d_n \to 0$ in probability, we have*

$$\begin{aligned}n^{-1}L_n(\beta) - n^{-1}L_n(\beta_0) &= -n^{-1}\sum_{i=1}^{n}(\beta-\beta_0)^T x_i \operatorname{sgn}(e_i)\\ &\quad + f(0)(\beta-\beta_0)^T V^2(\beta-\beta_0) + o_p(\|\beta-\beta_0\|^2 + n^{-1})\end{aligned} \tag{4}$$

holds uniformly in $\|\beta-\beta_0\| \le d_n$.

Proof It is easy to see that

$$|R_i(\beta)| \le 2|x_i^T(\beta-\beta_0)|I(|Y_i - x_i^T\beta| \le |x_i^T(\beta-\beta_0|)),$$

so

$$\sup_{\|\beta-\beta_0\|\le d_n} \frac{R_i(\beta)}{\|\beta-\beta_0\|} \le 2\|x_i\|I(|Y_i - x_i^T\beta_0| \le 2d_n\|x_i\|).$$

Since for any compact set B, the class of functions $\{\frac{R_i(\beta)}{\|\beta-\beta_0\|} : \beta \in B\}$, is Euclidean with an integrable envelope in the sense of Pakes and Pollard (1989), we can apply the maximal inequality of Pollard (1990, p. 38) to get, for some $C > 0$,

$$E\left[\sup_{\|\beta-\beta_0\|\le d_n} \frac{\sum_{i=1}^n (R_i(\beta) - ER_i(\beta))}{\sqrt{n}\|\beta-\beta_0\|}\right]^2$$
$$\le \frac{C}{n}\sum_{i=1}^n \|x_i\|^2 I(|\epsilon_i| \le 2d_n\|x_i\|) = o(1)$$

as $n \to \infty$ and $d_n \to 0$. Thus uniformly in $\|\beta-\beta_0\| \le d_n$,

$$\frac{1}{n}\sum_{i=1}^n (R_i(\beta) - ER_i(\beta)) = o(\|\beta-\beta_0\|/\sqrt{n}). \tag{5}$$

and

$$G_n(\beta) = \frac{1}{n}\sum_{i=1}^n \int_0^{-x_i^T(\beta-\beta_0)} \mathrm{Esgn}(\epsilon_i + u)\mathrm{d}u,$$

where $\mathrm{Esgn}(\epsilon_i + u) = 2\int_{-u}^0 f(x)\mathrm{d}x$. Since $G_n(\beta)$ is a convex function, it has derivative 0 at β_0, and its second derivative at β_0 is $\frac{1}{n}\sum_{i=1}^n x_i x_i^T 2f(0)$, by Taylor expansion,

$$G_n(\beta) = f(0)(\beta-\beta_0)^T V^2(\beta-\beta_0) + o(\|\beta-\beta_0\|^2). \tag{6}$$

Hence, (4) follows directly from (5) and (6). □

3 Large sample properties

Knight and Fu (2000) studied the limiting distributions of LASSO-type estimator in least squares setting. In this section, we establish similar large sample properties for the proposed estimator $\hat{\beta}_n$. The key tools we use are the LAQ property of the loss function and a novel inequality. The following result shows that $\hat{\beta}_n$ is consistent provided $\lambda_{nj} = o_p(n)$.

Theorem 1 *Under* (C.1)–(C.2) *and* $\lim_{n\to\infty} \lambda_{nj}/n \xrightarrow{p} \lambda_{0j} \geq 0, 1 \leq j \leq p$, $\hat{\beta}_n \to$ argmin (Z), *where*

$$Z(\beta) = \lim_{n\to\infty} G_n(\beta) + \sum_{j=1}^{p} \lambda_{0j}|\beta_j|.$$

In particular, since β_0 is the minimizer of $G_n(\beta)$, $\hat{\beta}_n$ is consistent, provided $\lambda_{nj} = o_p(n)$.

Proof By the uniform law of large numbers (Pollard 1990), $n^{-1}L_n(\beta) - n^{-1}L_n(\beta_0) - G_n(\beta) = o(1)$ uniformly for β in any compact set K, hence $Z_n(\beta) - Z(\beta) = o_p(1)$. Since $Z_n(\beta) \geq \frac{1}{n}L_n(\beta)$ and argmin $(L_n) = O_p(1)$, we know that argmin $(Z_n) = O_p(1)$. It follows that $\hat{\beta}_n \to$ argmin (Z). □

In order to establish the root-n consistency of $\hat{\beta}_n$, we need to study the following object function:

$$C(u) = u^T Du/2 - a^T u + \sum_{j=1}^{s} \lambda_j u_j + \sum_{j=s+1}^{p} \lambda_j |u_j|$$

where $u \in R^p$, D is a positive definite matrix, $\lambda_1, \ldots, \lambda_s$ are constants, $\lambda_{s+1}, \ldots, \lambda_p$ are nonnegative constants, and suppose that $\hat{u}$ is a minimizer of $c(u)$, then we have the following proposition

Proposition 2 *For any u, we have* $C(u) - C(\hat{u}) \geq (u - \hat{u})^T D(u - \hat{u})/2$

Proof First, let us look at the case when $s = 0$, and without loss of generality, assumes $\hat{u} = \binom{\hat{u}_1}{0}$,,where $\hat{u}_1 \in R^r$,and $\hat{u}_{1i} \neq 0, 1 \leq i \leq r$
denote

$$D = \begin{pmatrix} D_{11} & D_{21} \\ D_{12} & D_{22} \end{pmatrix}$$

where $D_{11}, D_{12}, D_{21}, D_{22}$ are $r \times r, r \times (p-r), (p-r) \times r, (p-r) \times (p-r)$ matrices, and $D_{12}^T = D_{21}$.
denote $a = \binom{a_1}{a_2}$, a_1, a_2 are r, $(p-r)$ dimensional vectors respectively.

Since the $\binom{\hat{u}_1}{0}$ are minimizer of $C(u)$, we have

$$(D_{11}\hat{u}_1)_i - a_{1i} + \lambda_i \text{sgn}(\hat{u}_{1i}) = 0, \quad 1 \leq i \leq r \tag{7}$$

$$|(D_{21}\hat{u}_1)_i - a_{2i}| \leq \lambda_i, \quad r+1 \leq i \leq p \tag{8}$$

the first equality holds because $\hat{u}_1$ is not zero componentwise and is the minimizer of the objective function

$$u_1^T D_{11} u_1/2 - a_1^T u_1 + \sum_{j=1}^{r} \lambda_j |u_j|$$

so the derivative of the objective function at the minimizer will be 0. The second inequality holds because generally $0 \in R^p$ is the minimizer of

$$u^T M u - b^T u + \sum_{i=1}^{p} \lambda_i |u_i|$$

is equivalent to say that

$$|b_i| \le \lambda_i$$

To show the proposition holds, we need to prove the inequality

$$\begin{aligned} &u_1^T D_{11} u_1/2 + u_2^T D_{22} u_2/2 + u_2^T D_{21} u_1 - a_1^T u_1 - a_2^T u_2 + \sum_{i=1}^{p} \lambda_j |u_j| \\ &\ge \hat{u}_1^T D_{11} \hat{u}_1/2 - a_1^T \hat{u}_1 + \sum_{i=1}^{r} \lambda_i |\hat{u}_{1i}| + (u_1 - \hat{u}_1)^T D_{11} (u_1 - \hat{u}_1)/2 \\ &\quad + (u_1 - \hat{u}_1)^T D_{12} u_2 + u_2^T D_{22} u_2/2 \end{aligned}$$

equivalent to

$$\begin{aligned} &u_2^T D_{21} u_1 - a_1^T u_1 - a_2^T u_2 + \sum_{i=1}^{p} \lambda_j |u_j| \\ &\ge \hat{u}_1^T D_{11} \hat{u}_1 - a_1^T \hat{u}_1 + \sum_{i=1}^{r} \lambda_i |\hat{u}_{1i}| - \hat{u}_1^T D_{11} u_1 + (u_1 - \hat{u}_1)^T D_{12} u_2 \end{aligned}$$

by (7)

$$\hat{u}_1^T D_{11} \hat{u}_1 - a_1^T \hat{u}_1 + \sum_{i=1}^{r} \lambda_i |\hat{u}_{1i}| = 0$$

and

$$-a_1^T u_1 + \sum_{i=1}^{r} \lambda_j |u_j| \ge -\hat{u}_1^T D_{11} u_1$$

so we only need to prove that

$$-a_2^T u_2 + \sum_{i=r+1}^{p} \lambda_i |u_{2(i-r)}| \geq -\hat{u}_1^T D_{12} u_2$$

which is apparent by (8).

Now we consider the general case when the $s > 0$,denote

$$D = \begin{pmatrix} C & B \\ B^T & A \end{pmatrix}$$

where C, B, B^T, A are $s \times s, s \times (p-s), (p-s) \times s, (p-s) \times (p-s)$matrices. denote $a = \binom{a_1}{a_2}$, a_1, a_2 are $s, (p-s)$ dimensional vectors respectively. denote $b = (\lambda_1, \ldots, \lambda_s)^T$ we take the transformation $v_1 = u_1 + C^{-1}Bu_2$, the objective function becomes

$$v_1^T C v_1/2 + u_2^T (A - B^T C^{-1} B) u_2/2 - (a_2^T - a_1^T C^{-1} B) u_2$$
$$-a_1^T v_1 + \sum_{i=s+1}^{p} \lambda_i |u_{2(i-s)}| + b^T v_1 - b^T C^{-1} B u_2$$

and when minimizing this function u_2 and v_1 can be separated. for the function of u_2 we apply the result we just got above, and rewrite the function of v_1 in its quadratic form and it is straightforward to have

$$\begin{aligned} & c(u) - c(\hat{u}) \\ & \quad \geq (u_2 - \hat{u}_2)^T (A - BC^{-1}B^T)(u_2 - \hat{u}_2)/2 \\ & \qquad + (v_1 - \hat{v}_1)^T C (v_1 - \hat{v}_1)/2 = (u - \hat{u})^T D (u - \hat{u})/2 \end{aligned}$$

Hence the proposition holds in the general case. □

Suppose that $\hat{u}_n$ is a minimizer of the objective function

$$B_n(u) = -n^{-1/2} x_i^T \operatorname{sgn}(\epsilon_i) u + f(0) u^T V^2 u + \sum_{i=1}^{s} \frac{\lambda_{ni}}{\sqrt{n}} \operatorname{sgn}(\beta_{0i}) u_i + \sum_{i=s+1}^{p} \frac{\lambda_{ni}}{\sqrt{n}} |u_i|. \tag{9}$$

The following result shows that $\hat{\beta}_n$ is $\sqrt{n}$-consistent provided $\lambda_{nj} = O_p(\sqrt{n})$.

Theorem 2 *Under* (C.1)–(C.2) *and* $\lambda_{nj}/\sqrt{n} \xrightarrow{p} \lambda_{0j} \geq 0, 1 \leq j \leq p$, *then* $\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} \text{argmin}\,(V)$, *where*

$$V(u) = W^T u + f(0) u^T V^2 u + \sum_{i=1}^{s} \lambda_{0i} sgn(\beta_{0i}) u_i + \sum_{i=s+1}^{p} \lambda_{0i} |u_i|,$$

and W has a $N(0, V^2)$ distribution. In particular if $\lambda_{nj} = o_p(\sqrt{n})$, the penalized estimator behaves like the full L_1 estimator $\hat{\beta}_{L_1}$.

Proof It follows from Theorem 1 that, $\hat{\beta}_n$ is consistent. By (4)

$$\begin{aligned}
&Z_n(\hat{\beta}_n) - Z_n(\beta_0)\\
&\quad= -n^{-1}x_i^T \operatorname{sgn}(e_i)(\hat{\beta}_n - \beta_0) + f(0)(\hat{\beta}_n - \beta_0)^T V^2(\hat{\beta}_n - \beta_0)/2 + o_p(\|\hat{\beta}_n - \beta_0\|^2 + n^{-1})\\
&\qquad + \sum_{i=1}^{s} \frac{\lambda_{ni}}{n}(|\hat{\beta}_n^i| - |\beta_{0i}|) + \sum_{i=s+1}^{p} \frac{\lambda_{ni}}{n}|\hat{\beta}_n^i|\\
&\quad= -n^{-1}x_i^T \operatorname{sgn}(e_i)(\hat{\beta}_n - \beta_0) + f(0)(\hat{\beta}_n - \beta_0)^T V^2(\hat{\beta}_n - \beta_0)/2 + o_p(\|\hat{\beta}_n - \beta_0\|^2 + n^{-1})\\
&\qquad + \sum_{i=1}^{s} \frac{\lambda_{ni}}{n}\operatorname{sgn}(\beta_{0i})(\hat{\beta}_n^i - \beta_{0i}) + \sum_{i=s+1}^{p} \frac{\lambda_{ni}}{n}|\hat{\beta}_n^i - \beta_{0i}|
\end{aligned}$$

let $n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) = \tilde{u}_n$, then

$$Z_n(\hat{\beta}_n) - Z_n(\beta_0) = n^{-1}B_n(\tilde{u}_n) + o_p(n^{-1}\|\tilde{u}_n\|^2 + n^{-1})$$

It is easy to see that $\hat{u}_n \xrightarrow{p}$ argmin $(V) = O_p(1)$. So $n^{-\frac{1}{2}}\hat{u}_n$ converges to 0 in probability, again by (4), we have $Z_n(\beta_0 + n^{-\frac{1}{2}}\hat{u}_n) - Z_n(\beta_0) = n^{-1}B_n(\hat{u}_n) + o_p(n^{-1}\|\hat{u}_n\|^2 + n^{-1})$.
Furthermore, since $\hat{\beta}_n$ is a minimizer of $Z_n(\beta)$, we have

$$0 \geq n^{-1}(B_n(\tilde{u}_n) - B_n(\hat{u}_n)) + o_p(n^{-1}\|\tilde{u}_n\|^2 + n^{-1}\|\hat{u}_n\|^2 + n^{-1})$$

by the Proposition 2, we have,

$$\frac{1}{n}f(0)(\tilde{u}_n - \hat{u}_n)^T V^2(\tilde{u}_n - \hat{u}_n) + o_p(n^{-1}\|\tilde{u}_n\|^2 + n^{-1}\|\hat{u}_n\|^2 + n^{-1}) \leq 0$$

It says that $n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0)$ and $\hat{u}_n$ has the same asymptotic distribution, and completes the proof. □

4 Adaptive two-stage procedure

4.1 Oracle property

In this section, we show that for properly chosen tuning parameters, the resultant penalized estimator exhibits the so-called oracle property. Suppose the $\{\lambda_{nj}\}$ satisfy the following conditions:

$$\text{(C.3)}\quad \frac{\lambda_{nj}}{\sqrt{n}} \xrightarrow{p} 0 \quad \text{for} \quad 1 \leq j \leq s \quad \text{and} \quad \frac{\lambda_{nj}}{\sqrt{n}} \xrightarrow{p} \infty \quad \text{for} \quad s+1 \leq j \leq p.$$

The first part of (C.3) tries to preserve the $\sqrt{n}$-consistency of the estimator, and the second part of it does the work of shrinking the zero coefficients directly to zero. Notice that the rates of regularization parameters are different between zero coefficients and nonzero ones. Practically, we do not know beforehand about such information, and actually this is exactly the task that the variable selection procedure is trying to accomplish. However, since we can estimate the coefficients with some precision, we can choose data-driven tuning parameters with asymptotically correct rates, and then the penalized estimator can exhibit the same asymptotic properties as the one with ideal tuning parameters. An approach based on this idea is given and illustrated in Sect. 4.3.

Perturbation methods are used to estimate the covariance matrix, define a new loss function

$$Z_n^*(\beta) := \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \beta^T x_i\right|\omega_i + \frac{1}{n}\sum_{j=1}^{p}\lambda_{nj}\left|\beta_j\right|$$

where $\omega_i (i = 1, \dots, n)$ are independent positive random variables with $E(\omega_i) = Var(\omega_i) = 1$ and are independent of the data $(Y_i, X_i)(i = 1, \dots, n)$, let the $\hat{\beta}_n^*$ be a minimizer $Z_n^*(\beta)$. We will show in Sect. 4.2 that conditional on the data $(Y_i, x_i)(i = 1, \dots, n)$, $n^{\frac{1}{2}}(\hat{\beta}_n^* - \hat{\beta}_n)$ has the same asymptotic distribution as $n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0)$, hence the realizations of $\hat{\beta}_n^*$ by repeatedly generalizing the random sample $(\omega_1, \dots, \omega_n)$ can be used to estimate the covariance matrix.

We need to establish the following $\sqrt{n}$-consistency for later use.

Proposition 3 *($\sqrt{n}$-consistency) Under* (C.1)–(C.3)*, we have* $n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) = O_P(1)$.

Proof We only need to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$P\left\{\inf_{\|u\|=C} Z_n\left(\beta_0 + \frac{u}{\sqrt{n}}\right) > Z_n(\beta_0)\right\} \geq 1 - \epsilon \tag{10}$$

together with the convexity of Z_n, this implies that $n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) = O_P(1)$.
And

$$\begin{aligned} n\left\{Z_n\left(\beta_0 + \frac{u}{\sqrt{n}}\right) - Z_n(\beta_0)\right\} &\geq L_n\left(\beta_0 + \frac{u}{\sqrt{n}}\right) - L_n(\beta_0) \\ &+ \sum_{j=1}^{s}\lambda_{nj}\left(\left|\beta_{j0} + \frac{u_j}{\sqrt{n}}\right| - |\beta_{j0}|\right) - \sum_{i=1}^{n}\frac{u^T x_i}{\sqrt{n}}\operatorname{sgn}(e_i) + \frac{f(0)}{2}u^T V^2 u \\ &+ o_p(1 + \|u\|^2) \end{aligned} \tag{11}$$

for sufficiently large C, the second term of (11) dominates the rest terms, hence (10) holds. It completes the proof. □

The following Proposition is needed to establish the oracle property of the estimator.

Proposition 4 *Under conditions* (C.1)–(C.3)*, with probability tending to one, for any given β^1 satisfying that $\|\beta^1 - \beta_0^1\| = O_p(n^{-1/2})$ and any constant C,*

$$Z_n\left\{\begin{pmatrix}\beta^1\\0\end{pmatrix}\right\} = \min_{\|\beta^2\|\le Cn^{-1/2}} Z_n\left\{\begin{pmatrix}\beta^1\\\beta^2\end{pmatrix}\right\}$$

Proof Denote the gradient of $L_n(\beta)$ by $U_n(\beta) = -\sum_{i=1}^n x_i \mathrm{sgn}(Y_i - \beta^T x_i)$. It is sufficient to show that with probability tending to one as $n \to \infty$, for any β^1 satisfying that $\beta^1 - \beta_0^1 = O_P(n^{-1/2})$, and $\|\beta^2\| \le Cn^{-1/2}$, $U_{nj}(\beta) + \lambda_{nj}\mathrm{sgn}(\beta_j)$ and β_j have the same signs for $\beta_j \in (-Cn^{-1/2}, Cn^{-1/2})$ for $j = s+1, \dots, p$.

Similarly as in Proposition 1, we have

$$U_n(\beta) = U_n(\beta_0) + 2f(0)nV^2(\beta - \beta_0) + o_p(n^{1/2} + n\|\beta - \beta_0\|) \tag{12}$$

it follows that

$$\begin{aligned} U_{nj}(\beta) + \lambda_{nj}\mathrm{sgn}(\beta_j) &= n^{1/2}\Big\{n^{-1/2}U_j(\beta_0) + 2f(0)n^{1/2}[V^2(\beta - \beta_0)]_j \\ &\qquad + o(1 + n^{1/2}\|\beta - \beta_0\|) + \frac{\lambda_{nj}}{\sqrt{n}}\mathrm{sgn}(\beta_j)\Big\} \\ &= n^{1/2}(O_P(1) + \frac{\lambda_{nj}}{\sqrt{n}}\mathrm{sgn}(\beta_j)), \end{aligned}$$

since $\frac{\lambda_{nj}}{\sqrt{n}} \to \infty$, for $j = s+1, \dots, p$, the sign of the $U_j(\beta) + \lambda_{nj}\mathrm{sgn}(\beta_j)$ is completely determined by the sign of the β_j, this completes the proof. □

Now, we can establish the following main theorem. The first component is exactly zero and the second component is estimated as well as if the correct model were known. This is the so-called oracle property.

Theorem 3 *(Oracle property) Under* (C.1)–(C.3)*,with probability tending to one, the penalized estimator $\hat\beta_n = \begin{pmatrix}\hat\beta_n^1\\\hat\beta_n^2\end{pmatrix}$ has the following properties:*

(a) $\hat\beta_n^2 = 0$

(b) $n^{\frac{1}{2}}(\hat\beta_n^1 - \beta_0^1) \to N(0, \frac{1}{4f(0)^2}V_1^{-2})$.

Proof It follows from Proposition 4 that part (a) holds. To prove part (b), notice that $\hat\beta_n^1$ is the minimizer of the object function

$$Z_{n1}(\beta^1) := \frac{1}{n}\sum_{i=1}^n \left|Y_i - {\beta^1}^T x_i^1\right| + \frac{1}{n}\sum_{j=1}^s \lambda_{nj}\left|\beta_j^1\right|, \tag{13}$$

it is the penalized estimator considering only the first s covariates. Since $\hat\beta_n^1$ is the minimizer and $\lambda_{nj} = o(n^{1/2}),\ \ 1 \le j \le s$, by Theorem 2, we know that $\sqrt{n}(\hat\beta_n^1 -$

$\beta_0^1) \to$ argmin (V_1), where $V_1(u_1) = W_1^T u_1 + f(0) u_1^T V_1^2 u_1$, and W_1 has a $N(0, V_1^2)$ distribution. argmin $(V_1) = (2f(0))^{-1} V_1^{-2} W_1 \to N(0, \frac{1}{4f(0)^2} V_1^{-2})$. It completes the proof. □

Remark 1 By Theorem 2, we can see that although with positive probability, the LASSO estimate shrinks some coefficients to zero, it does necessarily shrink the true zero coefficients and thus may erroneously retain insignificant variables in the model and in the meantime increase biases in the estimation of true nonzero coefficients. However, by Theorem 3 (b), with differentially scaled tuning parameters, the MLASSO performs the correct variable selection and achieves the asymptotic efficiency.

4.2 Distributional approximation

Now we establish the asymptotic properties of the perturbed penalized estimator, $\hat{\beta}_n^*$ is a minimizer of the loss function

$$Z_n^*(\beta) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \beta^T x_i| \omega_i + \frac{1}{n} \sum_{j=1}^{p} \lambda_{nj} |\beta_j|, \tag{14}$$

We are able to show conditional on data, the randomly perturbed estimator can be used to approximate the distribution of the estimator. To be more specific, we have the following theorem:

Theorem 4 *Under conditions* (C.1)–(C.3)*, with probability tending to one, conditional on the data* $(Y_i, x_i)(i = 1, \ldots, n)$, $\hat{\beta}_n^* = \begin{pmatrix} \hat{\beta}_{n1}^* \\ \hat{\beta}_{n2}^* \end{pmatrix}$ *has the following properties:*

(a) $\hat{\beta}_{n2}^* = 0$

(b) $n^{\frac{1}{2}} (\hat{\beta}_{n1}^* - \hat{\beta}_n^1) \to N\left(0, \frac{1}{4f(0)^2} V_1^{-2}\right)$.

Proof Using the same arguments as in proving Proposition 1, denote $L_n^*(\beta) = \sum_{i=1}^n |Y_i - x_i^T \beta| \omega_i$, for every sequence $d_n > 0$ with $d_n \to 0$ in probability, we can prove

$$\begin{aligned} & n^{-1} L_n^*(\beta) - n^{-1} L_n^*(\beta_0) = \\ & \quad -n^{-1} \sum_{i=1}^{n} (\beta - \beta_0)^T x_i \mathrm{sgn}(e_i) \omega_i + f(0)(\beta - \beta_0)^T V^2 (\beta - \beta_0) \\ & \quad + o_p(\|\beta - \beta_0\|^2 + n^{-1}) \end{aligned} \tag{15}$$

holds uniformly in $\|\beta - \beta_0\| \leq d_n$, Then as in Proposition 3, we can prove that conditionally on the original data, $n^{\frac{1}{2}} (\hat{\beta}_n^* - \hat{\beta}_n) = O_P(1)$, and as in Proposition 4, for any given β^1 satisfying that $\|\beta^1 - \beta_0^1\| = O_p(n^{-1/2})$ and any constant C, conditionally on the original data, we have

$$Z_n^* \left\{ \begin{pmatrix} \beta^1 \\ 0 \end{pmatrix} \right\} = \min_{\|\beta^2\| \leq C n^{-1/2}} Z_n^* \left\{ \begin{pmatrix} \beta^1 \\ \beta^2 \end{pmatrix} \right\}$$

By Theorem 1, with probability tending to one , $\hat{\beta}_n^2 = 0$, it follows that conditionally on the original data, $\hat{\beta}^*_{n2} = 0$, then considering only the first s covariates, $\hat{\beta}^*_{n1}$ is a minimizer of function

$$Z^*_{n1}(\beta^1) = \frac{1}{n}\sum_{i=1}^{n}|Y_i - {\beta^1}^T x_i^1|\omega_i + \frac{1}{n}\sum_{j=1}^{s}\lambda_{nj}|\beta_j^1|,$$

By Proposition 1 and Theorem 2, we have

$$V_{n1}(\hat{\beta}_n^1 - \beta_0^1) = \frac{1}{2f(0)}\sum_{i=1}^{n}{x_i^1}^T V_{n1}^{-1}\mathrm{sgn}(e_i) + o_p(1) \tag{16}$$

and similarly we have

$$V_{n1}(\hat{\beta}^*_{n1} - \beta_0^1) = \frac{1}{2f(0)}\sum_{i=1}^{n}{x_i^1}^T V_{n1}^{-1}\mathrm{sgn}(e_i)\omega_i + o_p(1).$$

Thus,

$$V_{n1}(\hat{\beta}^*_{n1} - \hat{\beta}_n^1) = \frac{1}{2f(0)}\sum_{i=1}^{n}{x_i^1}^T V_{n1}^{-1}\mathrm{sgn}(e_i)(\omega_i - 1) + o_p(1).$$

It suffices to show conditionally on (Y_i, x_i), $i = 1, \ldots, n$,

$$\frac{1}{2f(0)}\sum_{i=1}^{n}{x_i^1}^T V_{n1}^{-1}\mathrm{sgn}(e_i)(\omega_i - 1) \to N\left(0, \frac{I_s}{4f(0)^2}\right) \tag{17}$$

Since $\sum_{i=1}^{n} V_{n1}^{-1}x_i^1{x_i^1}^T V_{n1}^{-1}\mathrm{sgn}(e_i)^2 = I_s$ and $\max\limits_{1\le i\le n}|{x_i^1}^T V_{n1}\mathrm{sgn}(e_i)| \to 0$, by the central limit theorem, (17) follows, it completes the proof. □

4.3 Selection of tuning parameters

Although the theoretical result in Sects. 4.1 and 4.2 is interesting, it is impractical if we can not find λ_{nj} which satisfy C.3. Noticing that the rates of tuning parameters for nonzero and zero coefficients are different, we must base our selection of λ_{nj} on some preliminary estimation procedure. Denote the components of the L_1 estimator $\hat{\beta}_{L_1}$ by a_j, $j = 1, \ldots, p$. The estimates of their standard errors are denoted by b_j, $j = 1, \ldots, p$. Let

$$\lambda_{nj} = \eta * \left(\frac{\sqrt{n} * |b_j|}{|a_j|}\right)^{\gamma}, \gamma > 1, \eta > 0. \tag{18}$$

For the λ_{nj} defined in (18), we have the following proposition:

Proposition 5 *For a fixed* (η, γ)*, the* λ_{nj} *defined in* (18) *satisfy C.3.*

Proof (i) If the jth component of the β_0 is zero, then a_j/b_j converges to a $N(0,1)$ r.v.. Denoting the sequence by Z_n and the latter by Z, we have $\lambda_{nj}/\sqrt{n} = \eta * n^{\gamma/2-1/2}/|Z_n|^\gamma \to \infty$
(ii) If the jth component of the β_0 is non-zero, denote it by θ, it is known that $n^{\frac{1}{2}}(a_j - \theta) \to N(0, \sigma^2)$, where σ is a fixed positive constant, denote the sequence by Z_n and the latter by Z. So $\sqrt{n} * b_j \to \sigma$ and $a_j = Z_n/\sqrt{n} + \theta$, then $\lambda_{nj}/\sqrt{n} = \eta * \left(\frac{\sqrt{n}*|b_j|}{|Z_n/\sqrt{n}+\theta|}\right)^\gamma/\sqrt{n} \to \eta * \left(\frac{\sigma}{|Z/\sqrt{n}+\theta|}\right)^\gamma/\sqrt{n} \to 0.$ □

Now if the λ_{nj} is chosen as above to estimate the β_0, we also have to select parameters η and γ. Resampling-based model selection methods such as cross-validation or generalized degrees of freedom (Shen and Ye 2002) can be employed which are computationally intensive. Tibshirani (1996) constructed the generalized cross-validation style statistic to select the tuning parameter. A key statistic therein is the number of effective parameters or the degrees of freedom. An interesting property of the LASSO-type procedure is that the number of nonzero coefficients of the estimator is an unbiased estimate of its degrees of freedom; see, for example, Efron et al. (2004) or Zou et al. (2004). In this connection, generalized cross-validation (GCV) can be modified to choose η and γ where the residual sums of squares is replaced with the sum of absolute residuals and the degrees of freedom d is taken as the number of nonzero coefficients of the differentially penalized L_1 estimator. This allows us to reduce computational burden greatly. In traditional subset selection, GCV criterion is not consistent in the sense of choosing true subset model with probability tending to 1 as sample size goes to infinity while the bayesian information criterion (BIC) is. To come up with a consistent data-driven variable selection procedure, a BIC type criterion seems necessary. For tradition subset selection in the L_1 regression, the BIC is defined as

$$\text{BIC} = L_n\{\hat{\beta}\}/\hat{\sigma} + \frac{\log n}{2}\alpha,$$

where $\hat{\sigma} = \sum_{i=1}^{n} |Y_i - \hat{\beta}_{L_1} x_i|/n$ is the maximum likelihood estimate of scale parameter σ in the full model with all the candidate variables and α is the size of subset model. The size of subset model is its degrees of freedom while the degrees of freedom of the Lasso-type procedure can be unbiasedly estimated by the number of nonzero coefficients of the estimator. Thus the number of nonzero coefficients is denoted by d and the following BIC-type criterion is proposed for the selection of tuning parameters η and γ.

$$\text{BIC} = L_n\{\hat{\beta}\}/\hat{\sigma} + \frac{\log n}{2}d.$$

The selected (η_n, γ_n) minimizes the BIC function in the region $(\eta, \gamma) \in (0, \infty) \times (1, \infty)$.

The proposed selection criterion has at least two advantages. First, since the optimal tuning parameters are found by a grid search, computationally it is feasible. Second, the asymptotic optimality such as the oracle property is usually established for the penalized estimator with fixed tuning parameters (Fan and Li 2002). Theoretically, it is more worthy to investigate theoretical properties of the penalized estimator with data-driven tuning parameters. Interestingly, in the following theorem, the differentially penalized L_1 estimator with the BIC-based data-driven tuning parameters indeed exhibits the oracle property.

Theorem 5 *If we define the tuning parameters* $\{\lambda_{nj}\}$ *as in* (18)*, and select* (η, γ) *by the BIC function defined above, the resultant penalized estimator exhibits the oracle property.*

Proof It suffices to prove that the selected tuning parameters satisfy the C.3 or the penalized estimator is asymptotically equivalent to the estimator with tuning parameters satisfying the C.3. Without loss of generality, we assume that $0 \le \lim_{n\to\infty} \frac{\lambda_{nj}}{\sqrt{n}} \le \infty$ and $0 \le \lim_{n\to\infty} \frac{\lambda_{nj}}{n} \le \infty$. In the case where the limits do not exist, similar arguments can be applied for subsequences of λ_{nj}. It is easy to see that tuning parameters $\{\lambda_{nj}\}$ can fall into the following five cases.

(1) $\lim_{n\to\infty} \frac{\lambda_{nj}}{\sqrt{n}} = 0, 1 \le j \le p,$

(2) $\lim_{n\to\infty} \frac{\lambda_{nj}}{\sqrt{n}} = 0, 1 \le j \le s, 0 < \lim_{n\to\infty} \frac{\lambda_{nj}}{\sqrt{n}} < \infty, s+1 \le j \le p,$

(3) $\lim_{n\to\infty} \frac{\lambda_{nj}}{\sqrt{n}} = 0, 1 \le j \le s, \lim_{n\to\infty} \frac{\lambda_{nj}}{\sqrt{n}} = \infty, s+1 \le j \le p,$

(4) $0 < \lim_{n\to\infty} \frac{\lambda_{nj}}{\sqrt{n}} < \infty, 1 \le j \le s, \lim_{n\to\infty} \frac{\lambda_{nj}}{n} = \infty, s+1 \le j \le p,$

(5) $0 < \lim_{n\to\infty} \frac{\lambda_{nj}}{\sqrt{n}} = \infty, 1 \le j \le s, \lim_{n\to\infty} \frac{\lambda_{nj}}{n} = \infty, s+1 \le j \le p,$

(i) For tuning parameters in case 1, the estimator have no zero components, denote it by $\hat{\beta}(p)$, while the penalized estimator with tuning parameters in case 3 only have s true nonzero coefficients, denote it by $\hat{\beta}(s)$, denote the usual L_1 estimator with all the p covariates and only the first s covariates by $\hat{\beta}_{L_1}(p)$ and $\hat{\beta}_{L_1}(s)$, respectively. Hence

$$\begin{aligned} \mathrm{BIC}(\hat{\beta}(p)) &= \sum_{i=1}^{n} |Y_i - x_i^T \hat{\beta}(p)|/\hat{\sigma} + \frac{\log n}{2} p \\ &\ge \sum_{i=1}^{n} |Y_i - x_i^T \hat{\beta}_{L_1}^T(p)|/\hat{\sigma} + \frac{\log n}{2} p \end{aligned}$$

and

$$\begin{aligned} \mathrm{BIC}(\hat{\beta}(s)) &= \sum_{i=1}^{n} |Y_i - {x_i^1}^T \hat{\beta}(s)|/\hat{\sigma} + \frac{\log n}{2} s \\ &\le \sum_{i=1}^{n} |Y_i - {x_i^1}^T \hat{\beta}_{L_1}(s)|/\hat{\sigma} + \frac{\log n}{2} s + \sum_{j=1}^{s} \lambda_{nj}^0 (|\hat{\beta}_{L_1}(s)^j| - |\hat{\beta}(s)^j|, \end{aligned}$$

since $\hat{\beta}_{L_1}(s)$ and $\hat{\beta}(s)$ are both $\sqrt{n}-$ consistent and $\{\lambda_{nj}^0\}$ satisfy C.3, $\text{BIC}(\hat{\beta}(s)) \leq \sum_{i=1}^n |Y_i - x_i^{1^T} \hat{\beta}_{L_1}(s)|/\hat{\sigma} + \frac{\log n}{2} s + o_p(1)$. And by Proposition 1, we know that

$$\sum_{i=1}^n (|Y_i - x_i^{1^T} \hat{\beta}_{L_1}(p)| - |e_i|)/\hat{\sigma}$$

and

$$\sum_{i=1}^n (|Y_i - x_i^{1^T} \hat{\beta}_{L_1}(s)| - |e_i|)/\hat{\sigma}$$

are both $O_p(1)$, so with probability tending to 1, $\text{BIC}(\hat{\beta}(p)) > \text{BIC}(\hat{\beta}(s))$, hence the selected tuning parameters are not in case 1.

(ii) For tuning parameters in case 2, by consistency, true nonzero components of the estimator remain nonzero, and some of true zero components might be zero. If all true zero components are zero, then the estimator is asymptotically equivalent to the penalized estimator with tuning parameters in case 3, otherwise, apply the same argument as before, with probability tending to 1, those tuning parameters are not chosen.

$L_n(\beta) = O_p(n)$, with probability tending to 1, true zero components of the estimator are zero, some of true nonzero components might be zero. If all true nonzero components are nonzero, then its degrees of freedom is the same as in the case 3. Suppose that the estimator is $\tilde{\beta}(s)$ and the estimator corresponding to tuning parameters in case 3 is $\hat{\beta}(s)$. Since its tuning parameters are larger than the ones in the case 3, if they are not asymptotically equivalent, $L_n(\tilde{\beta}(s))$ is strictly larger than $L_n(\hat{\beta}(s))$ and tuning parameters are not chosen. In the other situation, the subset of nonzero components of the resultant estimator is a true subset of $\{1, 2, \ldots, s\}$. Without loss of generality, assume the subset of nonzero components of the resultant estimator is $\{1, 2, \ldots, s-1\}$, denote it by $\hat{\beta}(s-1)$, to show these tuning parameters are not chosen, it suffices to prove that, with probability tending to 1, $\text{BIC}(\hat{\beta}(s-1)) > \text{BIC}(\hat{\beta}(s))$, denote the L_1 estimator with only the first $s-1$ covariates by $\hat{\beta}_{L_1}(s-1)$, we only need to prove that $\sum_{i=1}^n (|Y_i - x_i^{1^T} \hat{\beta}_{L_1}(s-1)| - |e_i|)/\hat{\sigma} > \log n + O_p(1)$.

Actually, we will show that for sufficiently large n,

$$\sum_{i=1}^n (|Y_i - x_i^{1^T} \hat{\beta}_{L_1}(s-1)| - |e_i|) > \delta n, \tag{19}$$

where δ is a positive number.

Denote $\beta_0^{(s)} = (\beta_{01}, \ldots, \beta_{0s})^T$, let $\mathcal{B} = \left\{\beta^{(s)} \in R^s : |\beta^{(s)} - \beta_0^{(s)}| = \frac{|\beta_{0s}|}{2}\right\}$. By the uniform law of large number, we have

$$\sup_{\{\beta^{(s)}\in\mathcal{B}\}} |\frac{1}{n}\sum_{i=1}^{n}\{(|Y_i - {x_i^1}^T\beta^{(s)}| - |e_i|) - E(|Y_i - {x_i^1}^T\beta^{(s)}| - |e_i|)\}| \xrightarrow{p} 0 \quad (20)$$

For large n, $|\hat{\beta}_{L_1}(s) - \beta_0^{(s)}| < \frac{|\beta_{0s}|}{2}$, thus there exists a point $\hat{\beta}_{(s)} \in \mathcal{B}$ satisfying,

$$\sum_{i=1}^{n}|Y_i - {x_i^1}^T\hat{\beta}_{L_1}(s-1)| \geq \sum_{i=1}^{n}|Y_i - {x_i^1}^T\hat{\beta}^{(s)}|,$$

Also there exists $\bar{\beta}^{(s)}$ such that

$$\inf_{\beta^{(s)}\in\mathcal{B}}\sum_{i=1}^{n}E|Y_i - {x_i^1}^T\beta^{(s)}| = \sum_{i=1}^{n}E|Y_i - {x_i^1}^T\bar{\beta}^{(s)}|.$$

So

$$\begin{aligned}&\sum_{i=1}^{n}(|Y_i - {x_i^1}^T\hat{\beta}^{(s)}| - |e_i|)\\ &= \sum_{i=1}^{n}\{(|Y_i - {x_i^1}^T\hat{\beta}^{(s)}| - |e_i|) - E(|Y_i - {x_i^1}^T\hat{\beta}^{(s)}| - |e_i|)\} + E(|Y_i - {x_i^1}^T\hat{\beta}^{(s)}| - |e_i|)\}\\ &\geq o_p(n) + \sum_{i=1}^{n}E(|Y_i - {x_i^1}^T\bar{\beta}^{(s)}| - |e_i|) \geq \delta n,\end{aligned}$$

thus (19) holds, hence concludes the proof. □

Remark 2 Other model selection criteria such as GCV and AIC can also be modified accordingly to choose parameters η and γ. However, as GCV and AIC criteria are not consistent in the sense of choosing the true model with probability tending to 1 as n goes to infinity, the GCV or AIC based differentially scaled L_1 penalized estimator do not enjoy the desired oracle property.

Simulation results in Sect. 5 shows that the finite-sample performance of the estimates does not vary much with different values of γ. As a referee pointed out, it is worthy to explore how the tuning parameter γ changes the behavior of the estimator. To gain some insights into this, here we study how the estimate changes with γ in the orthonormal case. Denote $\frac{\sqrt{n}*|b_j|}{|a_j|}$ by z_{nj} and $\lambda_{nj} = \lambda z_{nj}^{\gamma}$. In the orthonormal case, the covariate vector $x_1, \ldots, x_p$ are mutually orthogonal, and $\hat{\beta}_j = \mathrm{sgn}(x_i^T Y)(|x_i^T Y| - \frac{\lambda}{2}z_{nj}^{\gamma})I(|x_i^T Y| \geq \frac{\lambda}{2}z_{nj}^{\gamma})$. For true zero coefficients, $z_{nj} = O_p(\sqrt{n})$; for true nonzero coefficients, $z_{nj} = O_p(1)$. The magnitudes of z_{nj} of true zero coefficients are much larger than those of true nonzero coefficients and this contrast is power factored by γ into λ_{nj} to differentially shrink coefficients. Large γ itself leads to large tuning parameters for nonzero coefficients although it magnifies the difference of tuning parameters for true zero coefficients and nonzero coefficients. Thus the ideal γ should be slightly larger than 1 to not only assure the correct asymptotic rate and but also minimize its impact on inflating the tuning parameters. From the explicit formula for the estimates, we see that small change of γ will not in general influence much on the behavior of the estimates.

5 Numerical studies

We have conducted extensive numerical studies to compare our proposed method with LASSO, traditional subset selection methods and the oracle least absolute deviations estimates. We denote our method by MLASSO since it is the natural generalization of LASSO by considering multiple tuning parameters. All simulations are conducted using IMSL's routine for L_1 regression RLAV in Fortran. We select the tuning parameter of LASSO and MLASSO by BIC or GCV function. For MLASSO, the results between selecting both η and γ and selecting only η with a fixed $\gamma = 1.5$ are very similar. So we let $\gamma = 1.5$ in most of the examples. For subset selection methods, the best subset is chosen as the one which minimizes BIC or GCV function. Following Tibshirani (1996) and Fan and Li (2001), we report simulation results in terms of model error instead of prediction error (PE). In the setting of linear models, suppose that

$$Y = \beta^T X + \epsilon,$$

from the (Y, X) we get $\hat{\beta}$ as an estimate of β and use the $\hat{\beta}^T X_{\text{future}}$ to predict the future response Y_{future}, where $(Y_{\text{future}}, X_{\text{future}})$ is an independent copy of (Y, X). The mean-squared error (ME) is defined by

$$\text{ME} = E(\hat{\beta}^T X_{\text{future}} - \beta^T X_{\text{future}})^2 = (\hat{\beta} - \beta)^T R(\hat{\beta} - \beta),$$

The prediction error (PE) is defined as

$$\text{PE} = E(Y - \hat{\beta}^T X_{\text{future}})^2 = ME + \sigma^2,$$

where R is the population covariance matrix of X, and σ^2 is the variance of the error.

5.1 Normal error case

Considering the simulation scenario of Fan and Li (2001), we simulate 100 datasets. Each of them consists of n observations from the model

$$Y_i = \beta^T X_i + \sigma \varepsilon_i \tag{21}$$

$\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, the components of x and ε are standard normal. The correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$. We compare the model error of each variable selection procedure to that of the full L_1 estimator. The relative value is called the relative model error (RME). We do the simulations for different sample sizes and σ values. In Table 1, we summarize the results in terms of the median of relative model errors (MRME), the average correct 0 coefficients and the average incorrect 0 coeffients over 100 simulated datasets. Our resampling procedure uses $1,000$ random samples from standard exponential distribution. The results are similar for the other distributions. From Table 1, we see that in the situation of small sample size and big noise, Best subset performs the best in reducing the model error while LASSO tends to identify the least incorrect zero components. When the noise level is decreased, even in the situation of small sample size, all procedures do not identify any nonzero component to be zero and in terms of both the number of correctly identified zero components and the reduction of the model error, MLASSO and Best subset perform much better than LASSO. When the sample size is increased, MLASSO tends to perform better than Best subset and closer to the true oracle estimator. It is also interesting to notice that the same criterion (BIC or GCV) based Best subset and MLASSO perform very similarly. In all the simulations, the methods of fixing $\gamma = 1.5$ and selecting γ give almost the identical performance. Hence in the later examples, we will let $\gamma = 1.5$.

We also use the simulations to test the accuracy of the estimated standard error of the estimator via the perturbation method. The standard error of 100 estimates (SD) is regarded as the true standard error of the estimator. The mean and the standard error of 100 estimated standard errors of the estimator via the perturbation method (SD_m, SD_s) are used to assess the performance of the perturbation method. In Table 2, we summarize the results for the situation of $n = 60, \sigma = 1.0$. From Table 2, it can be seen that SD and SD_m are very close and hence the perturbation method performs very well.

To demonstrate the consistent property of BIC-type MLASSO, we increase the sample size. In Table 3, we summarize the results about the proportion of the procedure selecting the true model. It can be seen from Table 3 BIC-based MLASSO and BIC-based Best subset tend to select the true model with the proportion increases to 1 as the sample size increases. The other procedures do not exhibit this good property. It can also be seen that BIC-based MLASSO performs even better than BIC-based Best subset when the sample size is large.

5.2 Laplace error case

In this example and the next example, we change the error distribution in model (21) to explore the robustness of the proposed estimator. We simulate 100 datasets consisting of 60 observations from model (21) with the error distribution now drawn from the standard double exponential (Laplace) distribution. The σ is set to be 1.0. Table 4 and Table 5 summarize the results of the simulations. From Table 4, it can be seen that the MLASSO performs favorably compared to the other methods. Form Table 5, we see that the perturbation method indeed gives a very accurate estimate of the standard error for the estimator.

Table 1 Variable selection in normal error case

Method	MRME (%)	Avg. no. of 0 coefficients	
		Correct	Incorrect
$n = 40, \sigma = 3.0$			
LASSO (BIC)	72.49	3.34	0.15
LASSO (GCV)	75.76	3.55	0.17
LASSO (AIC)	75.69	2.60	0.15
MLASSO^1 (BIC)	81.12	3.85	0.32
MLASSO^1 (GCV)	78.81	4.21	0.42
MLASSO^1 (AIC)	77.33	3.20	0.13
MLASSO^2 (BIC)	80.87	3.90	0.32
MLASSO^2 (GCV)	79.54	4.21	0.41
MLASSO^2 (AIC)	76.22	3.24	0.13
Subset (BIC)	69.40	4.11	0.34
Subset (GCV)	65.09	4.44	0.40
Subset (AIC)	68.52	3.44	0.29
Oracle	37.95	5	0
$n = 40, \sigma = 1.0$			
LASSO (BIC)	72.49	3.25	0
LASSO (GCV)	72.49	3.46	0
LASSO (AIC)	73.85	2.60	0
MLASSO^1 (BIC)	59.63	4.19	0
MLASSO^1 (GCV)	51.83	4.51	0
MLASSO^1 (AIC)	73.58	3.46	0
MLASSO^2 (BIC)	59.63	4.19	0
MLASSO^2 (GCV)	50.92	4.52	0
MLASSO^2 (AIC)	72.40	3.48	0
Subset (BIC)	56.45	4.16	0
Subset (GCV)	52.21	4.51	0
Subset (AIC)	68.86	3.40	0
Oracle	37.95	5	0
$n = 60, \sigma = 1.0$			
LASSO (BIC)	69.19	3.45	0
LASSO (GCV)	68.53	3.53	0
LASSO (AIC)	73.82	2.38	0
MLASSO^1 (BIC)	53.30	4.35	0
MLASSO^1 (GCV)	52.13	4.44	0
MLASSO^1 (AIC)	72.79	3.28	0
MLASSO^2 (BIC)	54.35	4.37	0
MLASSO^2 (GCV)	54.06	4.42	0
MLASSO^2 (AIC)	71.60	3.32	0

Table 1 Continued

		Avg. no. of 0 coefficients	
Method	MRME (%)	Correct	Incorrect
Subset (BIC)	65.07	4.26	0
Subset (GCV)	55.33	4.39	0
Subset (AIC)	81.48	3.48	0
Oracle	33.63	5	0

The value of γ in MLASSO[1] is selected, whereas the value of γ in MLASSO[2] is 1.5

Table 2 Estimation in normal error case ($n = 60$, $\sigma = 1.0$)

	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_5$		
Method	SD	SD_m	(SD_s)	SD	SD_m	(SD_s)	SD	SD_m	(SD_s)
LASSO (BIC)	0.208	0.224	0.055	0.216	0.230	0.050	0.193	0.231	0.058
LASSO (GCV)	0.207	0.225	0.055	0.214	0.231	0.050	0.191	0.231	0.058
LASSO (AIC)	0.207	0.224	0.056	0.215	0.230	0.049	0.191	0.232	0.058
MLASSO[1] (BIC)	0.200	0.205	0.053	0.225	0.206	0.051	0.191	0.191	0.054
MLASSO[1] (GCV)	0.195	0.206	0.054	0.209	0.205	0.051	0.193	0.190	0.055
MLASSO[1] (AIC)	0.196	0.205	0.053	0.217	0.205	0.050	0.192	0.190	0.054
MLASSO[2] (BIC)	0.198	0.205	0.053	0.224	0.207	0.050	0.191	0.190	0.053
MLASSO[2] (GCV)	0.198	0.206	0.055	0.211	0.205	0.050	0.190	0.189	0.053
MLASSO[2] (AIC)	0.199	0.206	0.054	0.212	0.206	0.049	0.190	0.190	0.053
Subset (BIC)	0.199	0.201	0.050	0.226	0.201	0.054	0.181	0.182	0.044
Subset (GCV)	0.194	0.202	0.051	0.215	0.201	0.051	0.181	0.182	0.041
Subset (AIC)	0.195	0.203	0.052	0.217	0.203	0.055	0.183	0.182	0.042
Oracle	0.192	0.209	0.052	0.196	0.205	0.051	0.155	0.180	0.039

The value of γ in MLASSO[1] is selected, whereas the value of γ in MLASSO[2] is 1.5

Table 3 Performance on consistency

sample size (n)	60	100	200	500	1,000	2,000
Subset (BIC)	0.45 (4.26)	0.66 (4.58)	0.64 (4.61)	0.81 (4.79)	0.81 (4.80)	0.80 (4.77)
Subset (GCV)	0.53 (4.39)	0.60 (4.52)	0.56 (4.44)	0.62 (4.54)	0.56 (4.42)	0.52 (4.34)
Subset (AIC)	0.16 (3.48)	0.19 (3.55)	0.18 (3.50)	0.23 (3.60)	0.21 (3.54)	0.22 (3.56)
LASSO (BIC)	0.24 (3.45)	0.27 (3.83)	0.28 (3.75)	0.38 (4.11)	0.44 (3.97)	0.19 (3.53)
LASSO (GCV)	0.29 (3.53)	0.27 (3.79)	0.20 (3.53)	0.25 (3.65)	0.29 (3.47)	0.17 (3.18)
LASSO (AIC)	0.05 (2.38)	0.06 (2.39)	0.08 (2.42)	0.07 (2.40)	0.10 (2.50)	0.10 (2.52)
MLASSO (BIC)	0.65 (4.35)	0.68 (4.59)	0.83 (4.78)	0.84 (4.80)	0.85 (4.83)	0.90 (4.90)
MLASSO (GCV)	0.67 (4.44)	0.67 (4.57)	0.67 (4.52)	0.68 (4.51)	0.71 (4.53)	0.61 (4.40)
MLASSO (AIC)	0.25 (3.28)	0.26 (3.29)	0.28 (3.35)	0.29 (3.36)	0.32 (3.38)	0.32 (3.34)

The number in the parenthesis is the average number of correctly identified nonzero coefficients

Table 4 Variable selection in Laplace error case

		Avg. no. of 0 coefficients	
Method	MRME (%)	Correct	Incorrect
LASSO (BIC)	61.17	3.67	0
LASSO (GCV)	60.94	3.81	0
LASSO (AIC)	58.55	2.71	0
MLASSO (BIC)	42.89	4.64	0
MLASSO (GCV)	41.20	4.74	0
MLASSO (AIC)	61.32	3.88	0
Subset (BIC)	43.73	4.69	0
Subset (GCV)	42.59	4.71	0
Subset (AIC)	59.32	3.89	0
Oracle	25.32	5	0

Table 5 Estimation in Laplace error case

	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_5$		
Method	SD	SD_m	(SD_s)	SD	SD_m	(SD_s)	SD	SD_m	(SD_s)
LASSO (BIC)	0.213	0.252	0.059	0.225	0.270	0.072	0.214	0.257	0.069
LASSO (GCV)	0.212	0.253	0.060	0.222	0.270	0.072	0.213	0.258	0.068
LASSO (AIC)	0.212	0.253	0.060	0.224	0.271	0.071	0.214	0.259	0.068
MLASSO (BIC)	0.212	0.222	0.060	0.203	0.229	0.068	0.191	0.204	0.058
MLASSO (GCV)	0.213	0.222	0.0584	0.203	0.229	0.067	0.183	0.204	0.0580
MLASSO (AIC)	0.212	0.222	0.059	0.202	0.228	0.065	0.183	0.205	0.057
Subset (BIC)	0.210	0.217	0.064	0.218	0.217	0.056	0.162	0.187	0.046
Subset (GCV)	0.210	0.218	0.0638	0.218	0.217	0.056	0.164	0.188	0.046
Subset (AIC)	0.211	0.217	0.064	0.219	0.218	0.057	0.164	0.189	0.047
Oracle	0.201	0.224	0.067	0.210	0.221	0.057	0.153	0.190	0.047

5.3 Mixed error case

In this example, we do the same simulations as in the previous example except that we now draw the error distribution from the standard normal distribution with 30% outliers from standard Cauchy distribution. Table 6 and Table 7 summarize the simulation results. It can be seen that the MLASSO performs the best in this situation and the perturbation method still performs very well.

5.4 Prostate cancer example

In this example, we apply the proposed approach to the prostate cancer data. The dataset comes from a study by Stamey et al. (1989). It consists of 97 patients who

Table 6 Variable selection in mixed error case

		Avg. no. of 0 coefficients	
Method	MRME (%)	Correct	Incorrect
LASSO (BIC)	71.35	3.90	0.00
LASSO (GCV)	71.29	4.02	0.00
LASSO (AIC)	56.43	3.24	0.00
MLASSO (BIC)	37.84	4.76	0.00
MLASSO (GCV)	36.38	4.80	0.00
MLASSO (AIC)	45.45	4.28	0.00
Subset (BIC)	41.28	4.82	0.03
Subset (GCV)	41.00	4.86	0.03
Subset (AIC)	29.02	4.09	0.00
Oracle	37.27	5	0

Table 7 Estimation in mixed error case

	$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_5$		
Method	SD	SD_m	(SD_s)	SD	SD_m	(SD_s)	SD	SD_m	(SD_s)
LASSO (BIC)	0.241	0.272	0.091	0.239	0.271	0.072	0.222	0.269	0.085
LASSO (GCV)	0.251	0.274	0.090	0.237	0.273	0.072	0.213	0.268	0.085
LASSO (AIC)	0.247	0.274	0.090	0.236	0.272	0.073	0.214	0.269	0.086
MLASSO (BIC)	0.231	0.236	0.067	0.244	0.233	0.063	0.171	0.204	0.051
MLASSO (GCV)	0.230	0.236	0.067	0.245	0.234	0.064	0.170	0.204	0.050
MLASSO (AIC)	0.234	0.238	0.066	0.247	0.233	0.064	0.171	0.204	0.050
Subset (BIC)	0.267	0.245	0.066	0.322	0.234	0.075	0.280	0.202	0.053
Subset (GCV)	0.267	0.245	0.066	0.322	0.234	0.074	0.279	0.202	0.053
Subset (AIC)	0.269	0.245	0.067	0.322	0.233	0.074	0.280	0.203	0.054
Oracle	0.215	0.245	0.064	0.239	0.242	0.069	0.188	0.203	0.050

were about to receive a radical prostatectomy. A number of clinical measures for each patient were recorded. The purpose of the study was to examine the correlation between the level of prostate specific antigen and eight factors. The factors are log (cancer volume) (lcavol), log (prostate weight) (lweight), age, log (benign prostaic hyperplasia amount) (lbph), seminal vesicle invasion (svi), log (capsular penetration) (lcp), Gleason score (gleason) and percentage Gleason scores 4 or 5 (pgg45). First we standardize the predictors and center the response variable, then we fit a linear model that relates the log (prostate specific antigen) (lpsa) to the predictors. We use the full LAD, the LASSO and the MLASSO method to estimate the coefficients in the model. The results are summarized in Table 8. With BIC or GCV, LASSO and Best subset result in the identical model and both of them exclude variable lcp and pgg45. With AIC, Best subset excludes only variable pgg45 while LASSO selects $\eta = 0$ and results in the full LAD. With GCV, MLASSO selects $\eta = 0.16$ and

Table 8 Prostate cancer example

Predictor	1 lcavol	2 lweight	3 age	4 lbph	5 svi	6 lcp	7 gleason	8 pgg45
LAD	0.63(0.13)	0.25 (0.12)	−0.16 (0.08)	0.20 (0.11)	0.35 (0.12)	−0.17 (0.14)	0.14 (0.11)	0.09 (0.13)
Subset (BIC)	0.58 (0.11)	0.23 (0.11)	−0.19 (0.08)	0.25 (0.11)	0.32 (0.13)	0.00 (−)	0.12 (0.09)	0.00 (−)
Subset (GCV)	0.58 (0.11)	0.23 (0.11)	−0.19 (0.08)	0.25 (0.11)	0.32 (0.13)	0.00 (−)	0.12 (0.09)	0.00 (−)
Subset (AIC)	0.61 (0.12)	0.22 (0.11)	−0.17 (0.09)	0.21 (0.10)	0.38 (0.12)	−0.15 (0.08)	0.19 (0.11)	0.00 (−)
LASSO (BIC)	0.59 (0.11)	0.24 (0.11)	−0.11 (0.08)	0.17 (0.11)	0.23 (0.13)	0.00 (0.07)	0.04 (0.07)	0.00 (0.09)
LASSO (GCV)	0.59 (0.11)	0.24 (0.11)	−0.11 (0.08)	0.17 (0.11)	0.23 (0.13)	0.00 (0.07)	0.04 (0.07)	0.00 (0.09)
LASSO (AIC)	0.63 (0.13)	0.25 (0.12)	−0.16 (0.08)	0.20 (0.11)	0.35 (0.12)	−0.17 (0.14)	0.14 (0.11)	0.09 (0.13)
MLASSO (BIC)	0.64 (0.12)	0.18 (0.10)	0.00 (0.05)	0.11 (0.07)	0.22 (0.11)	0.00 (0.04)	0.00 (0.03)	0.00 (0.00)
MLASSO (GCV)	0.60 (0.12)	0.22 (0.11)	−0.18 (0.08)	0.25 (0.11)	0.35 (0.12)	−0.06 (0.09)	0.14 (0.08)	0.00 (0.05)
MLASSO (AIC)	0.63 (0.13)	0.25 (0.12)	−0.16 0.08)	0.20 (0.11)	0.35 (0.12)	−0.17 (0.14)	0.14 (0.11)	0.09 (0.13)

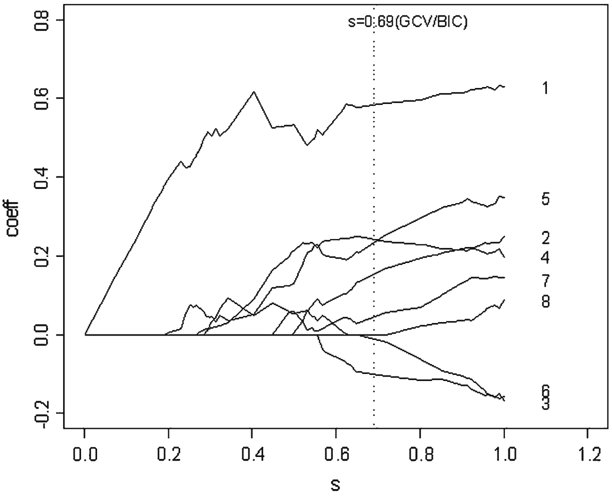

Fig. 1 Graphical display of LASSO shrinkage of eight coefficients as a function of shrinkage parameter s in the prostate cancer example. The broken line $s = 0.69$ is selected by both BIC and GCV criterion

excludes only variable pgg45 while with BIC it selects $\eta = 0.84$ and results in a very parsimonious model that retains only the four variables (lcavol, lweight, lbph and svi). With AIC, η is again selected to be 0 and MLASSO produces the full LAD. In Fig. 1, we show the LASSO estimates as a function of shrinkage parameter s, both BIC-based and GCV-based approach select the shrinkage parameter $s = 0.69$. In Fig. 2, we show the MLASSO estimates as a function of shrinkage parameter s, the BIC-based and GCV-based approach select the shrinkage parameter $s = 0.28, 0.71$ respectively.

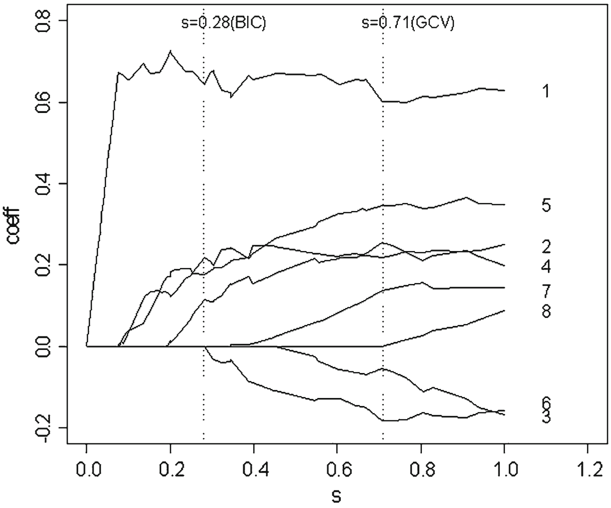

Fig. 2 Graphical display of MLASSO shrinkage of eight coefficients as a function of shrinkage parameter s in the prostate cancer example. The broken line $s = 0.28$ is selected by BIC and $s = 0.71$ is selected by GCV

6 Discussion

Variable selection is a fundamental problem in statistical modeling. A variety of methods have been well developed in the least squares-based regression while their counterparts in the median regression are much less understood. With the recent advancement of the linear programming techniques for the L_1 minimization, numerical simplicity is now also a nice property for the methods in the median regression. In this article, we consider the problem of simultaneous estimation and variable selection in the median regression model via penalizing the L_1 loss function via the L_1 (Lasso) penalty. Combining L_1 loss function with LASSO-type penalty, the penalized estimator can be solved easily by standard linear programming packages. Differentially scaled L_1 penalty are used to achieve desirable properties in terms of both identifying zero coefficients and estimating nonzero coefficients. Large sample properties of the proposed estimator are established by using local asymptotic quadratic property of the L_1 loss function and a novel inequality. Standard error of the estimator is obtained by using the random perturbation method. It is shown that for properly chosen tuning parameters, the differentially penalized L_1 estimator exhibits the oracle property. More interestingly, a modified BIC function is employed to obtain data-driven tuning parameters and the resultant two-stage procedure is proved to enjoy optimal properties. Extensive numerical studies show that the unified L_1 method fares comparably well in terms of simultaneous estimation and variable selection and retains the appealing robustness of L_1 estimator. The numerical simplicity of the proposed methodology gains extra benefits in real data analysis.

In spirit, the differentially scaled L_1 penalty is similar to the Adaptive Lasso proposed by Zou (2006) though the latter is developed for the squared loss while our investigation is conducted under the L_1 loss setting. A practical issue for both the differentially scaled L_1 penalty and the Adaptive Lasso is the construction of tuning parameters which behaves differently for the true nonzero coefficients and the true zero coefficients. A slight difference between our construction and the Adaptive Lasso is that we standardize the preliminarily estimated coefficients via their standard errors while the Adaptive Lasso uses the unstandardized ones. As the magnitudes of the standard errors for the least squares estimates or the median estimates may differ substantially in practice especially when the predictor variables are highly correlated, the differential or adaptive weights should be standardized to decrease their impact on the tuning parameters.

Acknowledgments This research was supported by grants from the U.S. National Institutes of Health, the U.S. National Science Foundation and the National University of Singapore (R-155-000-075-112). We are very grateful to the editor, the associate editor and the referees for their helpful comments which have greatly improved the paper.

References

Chen, S., Donoho, D. (1994). Basis pursuit. In *28th Asilomar Conference Signals*. Asilomar: Systems Computers.

Efron, B., Johnstone, I., Hastie, T., Tibshirani, R. (2004). Least angle regression (with discussions). *Annals of Statistics 32*, 407–499.

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

Fan, J., Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics 30*, 74–99.

Hurvich, C. M., Tsai, C. L. (1990). Model selection for Least absolute Deviations Regressions in Small Samples. *Statistics and Probability Letters 9*, 259–265.

Knight, K., Fu, W. J. (2000). Asymptotics for Lasso-type estimators. *Annals of Statistics 28*, 1356–1378.

Koenker, R., D'Orey, V. (1987). Computing regression quantiles. *Applied Statistics 36*, 383–393.

Shen, X., Ye, J. (2002). Adaptive model selection. *Journal of the American Statistical Association 97*, 210–221.

Pakes, A., Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*, 1027–1057.

Pollard, D. (1990). *Empirical Processes: Theory and Applications, Reginal Conference Series Probability and Statistics: Vol. 2*. Hayward: Institute of Mathematical Statistics.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory 7*, 186–199.

Portnoy, S., Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statistical Science 12*, 279–296.

Rao, C. R., Zhao, L. C. (1992). Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap. *Sankhyā A 54*, 323–331.

Ronchetti, E., Staudte, R. G. (1994). A Robust Version of Mallows's C_p. *Journal of the American Statistical Association 89*, 550–559.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate, ii: Radical prostatectomy treated patients. *Journal of Urology 16*, 1076–1083.

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

Xu, J. (2005). Parameter estimation, model selection and inferences in L_1-based linear regression. PhD dissertation. Columbia University.

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

Zou, H., Hastie, T., Tibshirani, R. (2007). On the “degrees of freedom” of the LASSO. *Annals of Statistics 35*, 2173–2192.