Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2020-0250

Vol. **30**, No. 1, pp. 190-209
July 2021

Truncated L_1 Regularized Linear Regression: Theory and Algorithm

Mingwei Dai[1], Shuyang Dai[2,3], Junjun Huang[2], Lican Kang[2] and Xiliang Lu[2,3,*]

[1] *Center of Statistical Research and School of Statistics, Southwestern University of Finance and Economics, Chengdu, P.R. China.*
[2] *School of Mathematics and Statistics, Wuhan University, Wuhan, P.R. China.*
[3] *Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan, P.R. China.*

Received 9 December 2020; Accepted (in revised version) 26 February 2021

Abstract. Truncated L_1 regularization proposed by Fan in [5], is an approximation to the L_0 regularization in high-dimensional sparse models. In this work, we prove the non-asymptotic error bound for the global optimal solution to the truncated L_1 regularized linear regression problem and study the support recovery property. Moreover, a primal dual active set algorithm (PDAS) for variable estimation and selection is proposed. Coupled with continuation by a warm-start strategy leads to a primal dual active set with continuation algorithm (PDASC). Data-driven parameter selection rules such as cross validation, BIC or voting method can be applied to select a proper regularization parameter. The application of the proposed method is demonstrated by applying it to simulation data and a breast cancer gene expression data set (bcTCGA).

AMS subject classifications: 62J99

Key words: High-dimensional linear regression, sparsity, truncated L_1 regularization, primal dual active set algorithm.

1 Introduction

In this paper, we consider the high-dimensional sparse linear regression model

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}^*+\boldsymbol{\epsilon}, \tag{1.1}$$

where $\mathbf{y}\in\mathbb{R}^n$ is the response vector, $\mathbf{X}=[\mathbf{x}_1,\cdots,\mathbf{x}_p]\in\mathbb{R}^{n\times p}$ is the covariance matrix, $\boldsymbol{\beta}^*\in\mathbb{R}^p$ is the underlying regression coefficients vector, $\boldsymbol{\epsilon}=(\epsilon_1,\cdots,\epsilon_n)^T\in\mathbb{R}^n$ is the random noise.

[*]Corresponding author. *Email addresses:* `daimw@swufe.edu.cn` (M. Dai), `shuyang_dai@whu.edu.cn` (S. Dai), `hjj_wd@whu.edu.cn` (J. Huang), `kanglican@whu.edu.cn` (L. Kang), `xllv.math@whu.edu.cn` (X. Lu)

Without loss generality, we assume that $\mathbf{X}$ is normalized such that each column of $\mathbf{X}$ is $\sqrt{n}$-length. We focus on the case $n \ll p$ and $\|\boldsymbol{\beta}^*\|_0 < n$ for the high dimensional and sparsity assumptions for (1.1), where $\|\boldsymbol{\beta}^*\|_0$ denotes the cardinality of nonzero element of $\boldsymbol{\beta}^*$.

There are various convex and non-convex regularization methods for variable estimation and selection of model (1.1). The popular convex regularization methods include the least absolute shrinkage and selection operator method (Lasso) [19], the adaptive Lasso [28] and Elastic net [29]. Thanks to the convexity of these regularizers, people have designed a lot of efficient numerical algorithms to solve above models, see e.g. [4, 21]. The convex model also has its drawback: it produces biased estimates for large coefficients [14] and lacks oracle property [6]. Some useful nonconvex regularization methods are proposed to circumvent this drawback, such as the bridge penalty method [9,10], the truncated L_1 regularization [5], the smoothly clipped absolute deviation (SCAD) penalty [7], the Dantzig selector [3], the minimax concave penalty (MCP) [23], the capped-L_1 penalty [26], etc.

The above mentioned nonconvex regularizers can be viewed as an approximation of original L_0 penalty ($\|\cdot\|_0$). Among these regularization methods, the truncated L_1 regularization has an attractive property: its thresholding operator is exactly same as the thresholding for L_0 regularizer. In this work, we will consider the truncated L_1 regularization for variable estimation and selection, i.e., we want to solve

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\sum_{i=1}^{p}\rho_\lambda(\beta_i), \tag{1.2}$$

where $\lambda > 0$ is the regularization parameter and $\rho_\lambda(\cdot)$ is defined by

$$\rho_\lambda(t)=\begin{cases} \lambda|t|, & \text{if } |t|<\lambda, \\ \frac{\lambda^2}{2}, & \text{if } |t|\geq\lambda. \end{cases} \tag{1.3}$$

We will prove that if the covariance matrix $\mathbf{X}$ satisfies a certain incoherence condition, then one can obtain the nonasymptotic error bound for the global optimal solution to (1.2). And the support recovery property is also studied. Due to the non-convex and non-smooth structure of the truncated L_1 regularization, (1.2) is a non-convex and non-smooth optimization problem. Then it is very difficult to design a stable and efficient numerical algorithm.

Inspired by [8, 12, 15, 17], we will propose a primal dual active set algorithm (PDAS) to compute the optimal solution to (1.2). PDAS can be viewed as a generalized Newton method, which involves two steps for each iteration. The active set is first determined using the summation of the primal and dual variables. Then the primal variable is updated by solving an optimization problem on the active set with small size, and the dual variable is updated based on a closed-form expression. Combining PDAS with a continuation strategy on the regularization parameter λ can make the whole algorithm more efficient. The regularization parameter λ can be determined by a data-driven method such as cross validation, Bayesian information criterion or the voting method [12].

The rest of this paper is organized as follows. At the end of this section we give some notations which will be used throughout this paper. In Section 2, we present the theoretical analysis for the global optimal solution to (1.2). Under some conditions, we establish the non-asymptotic error bound for the global optimal solution, and show that its support set coincides with the target support set with high probability. In Section 3 we introduce the PDAS algorithm and its globalization with continuation strategy. In Section 4 we conduct extensive numerical experiments to evaluate the performance of PDAS and illustrate its application by analyzing a gene expression data set. We conclude and summarize in Section 5. Proofs for all the lemmas and theorems are provided in the appendix.

Let $\|\boldsymbol{\beta}\|_q = (\sum_{i=1}^p |\beta_i|^q)^{\frac{1}{q}}$ for $q(q\in[1,\infty])$ be L_q norm of a vector $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)^T\in\mathbb{R}^p$. Denote by $\rho(\boldsymbol{\beta},\lambda)=(\rho_\lambda(\beta_1),\cdots,\rho_\lambda(\beta_p))^T$, and let $\|\boldsymbol{\beta}\|_{\min}$ be the minimum absolute value of $\boldsymbol{\beta}$. Let $S=\{1,\cdots,p\}$, and for any $A\subseteq S$ with size $|A|$, we use $\boldsymbol{\beta}_A$ (or $\mathbf{X}_A\in\mathbb{R}^{n\times|A|}$) to represent the subvector (or submatrix) whose entries (or columns) are listed in A, and $\mathbf{X}_{AB}$ denotes the submatrix of $\mathbf{X}$ whose rows and columns are listed in A and B, respectively. The true active set and inactive set are given by $A^*=\mathrm{supp}(\boldsymbol{\beta}^*)$ and $I^*=(A^*)^c$.

2 Theoretical properties of global optimal solutions

The truncated L_1 penalty $\rho_\lambda(\cdot)$ in (1.3) possesses some similar properties with other general non-convex penalties including SCAD [7], MCP [23] and the capped-L_1 penalty [26]. For example, with fixed λ, $\rho_\lambda(\cdot)$ is one symmetric function about the ordinate axis, and vanishes at zero and satisfies subadditivity, that is, $\rho_\lambda(u+v)\le\rho_\lambda(u)+\rho_\lambda(v)$ for all $u,v\ge 0$. See [25] for details analysing these properties about the non-convex regularization. Furthermore, its estimators also admit the sparsity, unbiasedness and continuity, advocated and characterized by [1,7]. Therefore, due to these excellent properties, we can derive the oracle nonasymptotic error bound for the global solution and study its support recovery property by following [25].

Define $\boldsymbol{\beta}^\diamond$ as the global minimization of problem (1.2). To estimate the error between $\boldsymbol{\beta}^\diamond$ and the true solution $\boldsymbol{\beta}^*$, we need the restricted invertibility factor and η-NC condition [25] which are defined as follows.

Definition 2.1. For $q\ge 1$, $\xi>0$ and $A\subset S$, we define the restricted invertibility factor as

$$\mathrm{RIF}_q(\xi,A)=\inf\left\{\frac{|A|^{1/q}\left\|\mathbf{X}^T\mathbf{X}\mathbf{u}\right\|_\infty}{n\|\mathbf{u}\|_q}:\ \|\rho(\mathbf{u}_{A^c},\lambda)\|_1<\xi\|\rho(\mathbf{u}_A,\lambda)\|_1\right\}. \tag{2.1}$$

Let $\eta\in(0,1]$. We say that the truncated L_1 regularization method (1.2) satisfies the η

null consistency condition (η-NC) if the following equality holds:

$$\min_{\mathbf{b}\in\mathbb{R}^p}\left(\|\boldsymbol{\epsilon}/\eta-\mathbf{X}\mathbf{b}\|_2^2/2n+\|\rho(\mathbf{b},\lambda)\|_1\right)=\frac{\|\boldsymbol{\epsilon}/\eta\|_2^2}{2n}. \tag{2.2}$$

Theorem 2.1. *Assume that the η-NC condition* (2.2) *holds with $\eta\in(0,1)$. Let $\xi=(\eta+1)/(1-\eta)$ in* (2.1) *and $\|\mathbf{X}^T\boldsymbol{\epsilon}/n\|\le\lambda$. Then for all $q\ge 1$,*

$$\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^\diamond\|_q\le\frac{2\lambda|A^*|^{1/q}}{\mathrm{RIF}_q(\xi,A^*)}.$$

The proof can be find in the appendix. Next, we study the probabilistic and nonasymptotic error bounds of the minimizer $\boldsymbol{\beta}^\diamond$ under the following two assumptions.

(C1) The error terms $\epsilon_1,\cdots,\epsilon_n$ are independent and identically distributed with mean 0 and subgaussian tails, that is, there exists one constant $\sigma>0$ such that $E[\exp(t\epsilon_i)]\le\exp(\sigma^2t^2/2)$ for $t\in\mathbb{R}$, $i=1,\cdots,n$.

(C2) $\|\boldsymbol{\beta}^*_{A^*}\|_{\min}\ge\frac{2\gamma_n}{\mathrm{RIF}_\infty(\xi,A^*)}$, where $\gamma_n=\sigma\sqrt{\frac{2\log(p/\alpha)}{n}}$ with $\alpha\in(0,\frac{1}{2})$, η and ξ are defined in Theorem 2.1.

Remark 2.1. Condition (C1) on the subgaussion tails of the error terms is standard in high-dimensional regression models. Condition (C2) assumes that the signal is not too small, which is needed for the target signal to be detectable.

Theorem 2.2. *Assume that the η-NC condition* (2.2) *holds with $\eta\in(0,1)$, and* (C1)-(C2) *hold. Set $\xi=(\eta+1)/(1-\eta)$ in* (2.1). *Then for all $q\ge1$ and any $\alpha\in(0,\frac{1}{2})$ defined in* (C2), *with probability at least $1-2\alpha$,*

$$\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^\diamond\|_q\le\frac{2\gamma_n|A^*|^{1/q}}{\mathrm{RIF}_q(\xi,A^*)}.$$

The following theorem establishes the support recovery property of the global solution $\boldsymbol{\beta}^\diamond$.

Theorem 2.3. *Assume that the η-NC condition* (2.2) *holds with $\eta\in(0,1)$, and* (C1)-(C2) *hold. Set $\xi=(\eta+1)/(1-\eta)$ in* (2.1). *Then for any $\alpha\in(0,\frac{1}{2})$ defined in* (C2), *with probability at least $1-2\alpha$, $A^*\subseteq supp(\boldsymbol{\beta}^\diamond)$.*

The proof to Theorems 2.2 and 2.3 are in the appendix.

3 Primal dual active set (PDAS) algorithm with continuation

As above theoretical analysis, the global solution $\boldsymbol{\beta}^\diamond$ of (1.2) is one oracle estimator of the target regression coefficients $\boldsymbol{\beta}^*$. But the minimization problem (1.2) is one non-convex and non-smooth optimization problem, it is not easy to design numerical algorithm to obtain this oracle estimator. Inspired by [12, 13, 17], we propose a primal dual active set algorithm (PDAS) for fixed regularization parameter λ. Then coupled with a warm-start strategy as its globalization, we have PDAS with continuation (PDASC) method.

3.1 PDAS algorithm

We first give a necessary condition to the global minimization $\boldsymbol{\beta}^\diamond$.

Lemma 3.1. *If $\boldsymbol{\beta}^\diamond$ is the global minimizer of* (1.2)*, then it satisfies*

$$\begin{cases} \mathbf{d}^\diamond = \mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\diamond)/n, \\ \boldsymbol{\beta}^\diamond = \Gamma_\lambda(\boldsymbol{\beta}^\diamond+\mathbf{d}^\diamond), \end{cases} \tag{3.1}$$

where the i-th element of $\Gamma_\lambda(\cdot)$ is defined by

$$(\Gamma_\lambda(\boldsymbol{\beta}))_i = \begin{cases} 0, & |\beta_i| \le \lambda, \\ \beta_i, & |\beta_i| > \lambda. \end{cases} \tag{3.2}$$

Conversely, if $\boldsymbol{\beta}^\diamond$ and $\mathbf{d}^\diamond$ satisfy (3.1)-(3.2)*, then $\boldsymbol{\beta}^\diamond$ is a local minimizer of* (1.2).

The proof can be find in the appendix. Lemma (3.1) is similar to Lemma 1 in [11] and Lemma 3.4 in [12], with different penalty functions substituting to the truncated L_1 penalty.

Denote by $A^\diamond = \text{supp}(\boldsymbol{\beta}^\diamond)$ and $I^\diamond = (A^\diamond)^c$, then from the definition of $\boldsymbol{\beta}^\diamond$ and $\mathbf{d}^\diamond$ defined in (3.1) and $\Gamma_\lambda(\cdot)$ in (3.2), we can conclude that

$$A^\diamond = \{i \in S : |\beta_i^\diamond + d_i^\diamond| > \lambda\}, \quad I^\diamond = \{i \in S : |\beta_i^\diamond + d_i^\diamond| \le \lambda\},$$

and

$$\begin{cases} \boldsymbol{\beta}^\diamond_{I^\diamond} = \mathbf{0}, \\ \mathbf{d}^\diamond_{A^\diamond} = \mathbf{0}, \\ \boldsymbol{\beta}^\diamond_{A^\diamond} = (\mathbf{X}^T_{A^\diamond}\mathbf{X}_{A^\diamond})^{-1}\mathbf{X}^T_{A^\diamond}\mathbf{y}, \\ \mathbf{d}^\diamond_{I^\diamond} = \mathbf{X}^T_{I^\diamond}(\mathbf{y}-\mathbf{X}_{A^\diamond}\boldsymbol{\beta}^\diamond_{A^\diamond})/n. \end{cases}$$

For fixed λ, let $\{\boldsymbol{\beta}^k, \mathbf{d}^k\}$ be the value at k-th iteration, and denote the active set and inactive set as $\{A^k, I^k\}$ based on $\{\boldsymbol{\beta}^k, \mathbf{d}^k\}$, where $\{A^k, I^k\}$ is expressed as

$$A^k = \{i \in S : |\beta_i^k + d_i^k| > \lambda\}, \quad I^k = \{i \in S : |\beta_i^k + d_i^k| \le \lambda\}. \tag{3.3}$$

Therefore we can get a new approximation pair $\{\boldsymbol{\beta}_{I^k}^{k+1},\mathbf{d}_{A^k}^{k+1},\boldsymbol{\beta}_{A^k}^{k+1},\mathbf{d}_{I^k}^{k+1}\}$ showed as follow:

$$\begin{cases} \boldsymbol{\beta}_{I^k}^{k+1}=\mathbf{0}, \\ \mathbf{d}_{A^k}^{k+1}=\mathbf{0}, \\ \boldsymbol{\beta}_{A^k}^{k+1}=(\mathbf{X}_{A^k}^T\mathbf{X}_{A^k})^{-1}\mathbf{X}_{A^k}^T\mathbf{y}, \\ \mathbf{d}_{I^k}^{k+1}=\mathbf{X}_{I^k}^T(\mathbf{y}-\mathbf{X}_{A^k}\boldsymbol{\beta}_{A^k}^{k+1})/n. \end{cases} \tag{3.4}$$

The proposed PDAS algorithm is described in the following Algorithm 1.

Algorithm 1 PDAS Algorithm

```
 1: Input: $\boldsymbol{\beta}^0$, $\mathbf{d}^0$, $\lambda$, $K$
 2: for $k=0,1,\cdots,K$, do
 3:     $A^k=\{j\in S:|\beta_j^k+d_j^k|>\lambda\}$, $I^k=(A^k)^c$.
 4:     $\boldsymbol{\beta}_{I^k}^{k+1}=\mathbf{0}$.
 5:     $\mathbf{d}_{A^k}^{k+1}=\mathbf{0}$.
 6:     $\boldsymbol{\beta}_{A^k}^{k+1}=(\mathbf{X}_{A^k}^T\mathbf{X}_{A^k})^{-1}\mathbf{X}_{A^k}^T\mathbf{y}$.
 7:     $\mathbf{d}_{I^k}^{k+1}=\mathbf{X}_{I^k}^T(\mathbf{y}-\mathbf{X}_{A^k}\boldsymbol{\beta}_{A^k}^{k+1})/n$.
 8:     if $A^k=A^{k+1}$ or $k\geq K$, then
 9:         Stop and denote the last iteration $\boldsymbol{\beta}_{\widehat{A}}$, $\boldsymbol{\beta}_{\widehat{I}}$, $\mathbf{d}_{\widehat{A}}$, $\mathbf{d}_{\widehat{I}}$.
10:     else
11:         $k=k+1$
12:     end if
13: end for
14: Output: $\widehat{\boldsymbol{\beta}}(\lambda)=(\boldsymbol{\beta}_{\widehat{A}}^{\mathrm{T}},\boldsymbol{\beta}_{\widehat{I}}^{\mathrm{T}})^{\mathrm{T}}$ and $\widehat{\mathbf{d}}(\lambda)=(\mathbf{d}_{\widehat{A}}^{\mathrm{T}},\mathbf{d}_{\widehat{I}}^{\mathrm{T}})^{\mathrm{T}}$ as the estimation at $\lambda$.
```

PDAS algorithm (Algorithm 1) terminates computation when the sequential estimated support coincides with each other or the maximum iteration number exceeds the given iteration number K large enough. In PDAS algorithm, step 3 selects the active predictors by combining the primal part with dual part. Then, it obtains the solution limited to the selected active set A^k as described in step 6, where the solver is the least square estimator limited to the selected active set A^k.

PDAS algorithm only obtains the solution $\widehat{\boldsymbol{\beta}}(\lambda)$ for the fixed regularization parameter λ, and we generally concentrate more on the solution path with different λ belonging to a finite interval. Thence we propose one sequential version of PDAS with a warm-start strategy to get the desirable solution path and use one appropriate variable selection criterion to choose the optimal solution in the next subsection.

3.2 PDASC algorithm

Combining PDAS algorithm with a continuation strategy to provide good initial guesses, we have PDASC algorithm to output a solution path. From the Lemma 3.1, **0** will be one minimizer of the optimization problem (1.2) if $\lambda \geq \|\mathbf{X}^T\mathbf{y}/n\|_\infty$. Therefore we can let $\lambda_m = \lambda_0 \alpha^m$, for $\alpha \in (0,1)$, be a decreasing sequence of regularization parameters, where we set $\lambda_0 = \|\mathbf{X}^T\mathbf{y}/n\|_\infty$ such that

$$\widehat{\boldsymbol{\beta}}(\lambda_0) = \mathbf{0} \quad \text{and} \quad \widehat{\mathbf{d}}(\lambda_0) = \mathbf{X}^T\mathbf{y}/n.$$

Then we can run Algorithm 1 on the sequence $\{\lambda_m\}_m$, and get the solution path $\{\widehat{\boldsymbol{\beta}}(\lambda_m), \widehat{\mathbf{d}}(\lambda_m)\}_m$. In PDASC algorithm, we set the initial values be $\{\widehat{\boldsymbol{\beta}}(\lambda_m), \widehat{\mathbf{d}}(\lambda_m)\}$ in Algorithm 1 with $\lambda = \lambda_{m+1}$. In addition, we can terminate the PDASC algorithm and obtain a solution path until $\|\widehat{\boldsymbol{\beta}}_{\lambda_m}\|_0 > \lfloor \frac{n}{\log p} \rfloor$ for some m. Last, the optimal λ can be determined by a data-driven method such as cross validation, Bayesian information criterion or the voting method [12] without any extra computational overhead. The pseudocode of PDASC algorithm is described in the following Algorithm 2.

Algorithm 2 PDASC Algorithm

```
Input: $\widehat{\boldsymbol{\beta}}(\lambda_0) = \mathbf{0}$, $\widehat{\mathbf{d}}(\lambda_0) = \mathbf{X}^T\mathbf{y}/n$, $\lambda_0 = \|\mathbf{X}^T\mathbf{y}/n\|_\infty$, $M$, $\alpha$.
for $m = 1, \cdots, M$ do
    $\lambda = \lambda_m = \lambda_0 \alpha^m$, $\boldsymbol{\beta}^0 = \widehat{\boldsymbol{\beta}}(\lambda_{m-1})$, $\mathbf{d}^0 = \widehat{\mathbf{d}}(\lambda_{m-1})$.
    Run Algorithm 1 to get $\widehat{\boldsymbol{\beta}}(\lambda_m)$ and $\widehat{\mathbf{d}}(\lambda_m)$.
    if $\|\widehat{\boldsymbol{\beta}}(\lambda_m)\|_0 > \lfloor \frac{n}{\log p} \rfloor$, stop.
end for
Output: $\left\{\widehat{\boldsymbol{\beta}}(\lambda_0), \widehat{\boldsymbol{\beta}}(\lambda_1), \cdots, \widehat{\boldsymbol{\beta}}(\lambda_M)\right\}$.
```

4 Numerical examples

In this section, we use simulation data set and real data set to illustrate the effectiveness of the proposed PDASC algorithm to truncated L_1 penalty. As comparison the solvers to Lasso, MCP and SCAD are R package ncvreg [2].

In the computations, the $n \times p$ covariates matrix $\mathbf{X}$ is generated according to the following three settings:

(I) The rows of $\mathbf{X}$ are independently distributed from $N(0,\Sigma)$, where $\Sigma_{i,j} = \rho^{|i-j|}$ for $1 \leq i,j \leq p$, and ρ is the correlation parameter.

(II) We first generate a $n \times p$ random Gaussian matrix $\widetilde{\mathbf{X}}$ whose entries are i.i.d. $\sim N(0,1)$. Then the covariates matrix $\mathbf{X}$ is generated with $\mathbf{x}_1 = \widetilde{\mathbf{x}}_1$, $\mathbf{x}_p = \widetilde{\mathbf{x}}_p$, and $\mathbf{x}_j = \widetilde{\mathbf{x}}_j + \rho(\widetilde{\mathbf{x}}_{j+1} + \widetilde{\mathbf{x}}_{j-1})$, $j = 2, \cdots, p-1$. Here ρ is a measure of the correlation among covariates.

(III) The rows of $\mathbf{X}$ are independently distributed from $N(0,\Sigma)$, where the diagonal and off-diagonal elements of Σ are 1 and $\rho=\frac{1}{1+C\cdot T}$ for $C>0$ and $T=\|\boldsymbol{\beta}^*\|_0$, respectively. See [27] for details.

The support A^* is chosen uniformly from S with $|A^*|=T<n$. The nonzero elements of $\boldsymbol{\beta}^*$ are generated via $\beta_i^*=\theta_i R^{\kappa_i}$, where θ_i are i.i.d. Bernoulli random variables, κ_i are i.i.d. uniform random variables in $[0,1]$ and $R>1$. Then the response vector is generated based on $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}^*+\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}\sim N(0,\sigma^2 I_p)$.

4.1 Accuracy and efficiency

In this section, we compare PDASC with Lasso, MCP and SCAD in terms of the average ℓ_∞ absolute error (AE), the average ℓ_2 relative error (RE), the average exact support recovery probability (RP), the mean size of the estimated supports (MSES), and the average CPU time (Time) (in seconds). Let J denote the number of independent replications. Then above criteria can be defined as

$$\mathrm{AE}=\frac{1}{J}\sum_{j=1}^{J}\|\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}^*\|_\infty,\quad \mathrm{RE}=\frac{1}{J}\sum_{j=1}^{J}\|\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}^*\|/\|\boldsymbol{\beta}^*\|,$$
$$\mathrm{RP}=\frac{1}{J}\sum_{j=1}^{J}\mathbf{1}\{\widehat{A}^{(j)}=A^*\},\quad \mathrm{MSES}=\frac{1}{J}\sum_{j=1}^{J}|\widehat{A}^{(j)}|,$$
$$\mathrm{Time}=\frac{1}{J}\sum_{j=1}^{J}t^{(j)},$$

where $\widehat{\boldsymbol{\beta}}^{(j)}$ is the estimator at j-th simulation, $\widehat{A}^{(j)}$ is the estimated support, and $t^{(j)}$ is the j-th running time. We consider following three scenarios:

- $\mathbf{X}$ is generated according to (I), and $\sigma=0.5$, 1, $\rho=0.2{:}0.2{:}0.8$, $R=10$, $n=400$, $p=4000$, $T=20$.

- $\mathbf{X}$ is generated according to (II), and $\sigma=0.5$, 1, $\rho=0.2{:}0.2{:}0.8$, $n=1000$, $p=10000$, $T=40$.

- $\mathbf{X}$ is generated according to (III), and $\sigma=0.5$, 1, $C=2{:}2{:}8$, $n=400$, $p=4000$, $T=20$.

The results reported in Tables 1-3 are based on 100 independent replications. As shown in Tables 1-3, PDASC is more accurate in terms of estimation error measured by AE and RE, exact support recovery probability (RP), and mean length of the estimated supports (MSES) than Lasso, MCP and SCAD in all the settings considered here. As for computational efficiency, PDASC is about 5-10 times faster than Lasso, MCP and SCAD.

Table 1: Numerical results with $n=400$, $p=4000$, $T=20$, $R=10$, $\sigma=0.5$ and 1, $\rho=0.2:0.2:0.8$ and $\mathbf{X}$ follows (I).

ρ	σ	Method	AE	RE (10^{-2})	Time(s)	RP	MSES
0.2	0.5	Lasso	0.675	10.95	3.86	0.96	20.04
		MCP	0.173	1.28	4.18	1	20
		SCAD	0.384	2.98	3.89	1	20
		PDASC	0.055	0.57	0.71	1	20
	1	Lasso	0.698	10.96	3.43	0.7	20.35
		MCP	0.215	1.73	4.18	1	20
		SCAD	0.413	3.25	4.14	1	20
		PDASC	0.111	1.14	0.70	1	20
0.4	0.5	Lasso	0.714	11.14	3.84	0.85	20.2
		MCP	0.170	1.26	4.17	1	20
		SCAD	0.395	3.06	3.83	1	20
		PDASC	0.057	0.58	0.72	1	20
	1	Lasso	0.723	11.20	3.40	0.63	20.53
		MCP	0.204	1.69	4.08	1	20
		SCAD	0.413	3.30	4.18	1	20
		PDASC	0.115	1.16	0.73	1	20
0.6	0.5	Lasso	0.724	11.20	3.91	0.48	20.8
		MCP	0.189	1.41	4.14	0.99	19.99
		SCAD	0.412	3.15	3.83	0.99	19.99
		PDASC	0.055	0.56	0.86	1	20
	1	Lasso	0.734	11.17	3.41	0.23	21.46
		MCP	0.227	1.87	4.01	0.99	19.99
		SCAD	0.432	3.40	4.18	0.99	19.99
		PDASC	0.111	1.13	0.69	1	20
0.8	0.5	Lasso	0.876	12.15	3.85	0.01	23.95
		MCP	0.208	1.57	4.24	0.98	19.98
		SCAD	0.464	3.55	3.95	0.98	19.97
		PDASC	0.056	0.57	0.92	1	20
	1	Lasso	0.886	12.23	3.28	0.01	25.44
		MCP	0.247	2.00	3.85	0.98	19.97
		SCAD	0.482	3.82	4.24	0.96	19.98
		PDASC	0.111	1.15	0.61	1	20

Table 2: Numerical results with $n=1000$, $p=10000$, $T=40$, $R=10$, $\sigma=0.5$ and 1, $\rho=0.2{:}0.2{:}0.8$ and $\mathbf{X}$ follows (II).

ρ	σ	Method	AE	RE (10^{-2})	Time	RP	MSES
0.2	0.5	Lasso	0.766	11.62	20.55	0.94	40.06
		MCP	0.249	1.51	20.60	1	40
		SCAD	0.512	3.74	22.59	1	40
		PDASC	0.037	3.31	2.49	1	40
	1	Lasso	0.773	11.63	22.00	0.91	40.09
		MCP	0.264	1.67	22.01	1	40
		SCAD	0.523	3.82	18.66	1	40
		PDASC	0.074	0.66	2.26	1	40
0.4	0.5	Lasso	0.820	11.83	20.59	0.41	40.9
		MCP	0.287	1.73	20.65	0.98	40.01
		SCAD	0.534	3.85	22.60	0.99	40.01
		PDASC	0.033	0.30	2.63	1	40
	1	Lasso	0.825	11.85	22.61	0.38	41.08
		MCP	0.300	1.86	22.64	0.98	40.01
		SCAD	0.544	3.92	18.59	0.99	40.01
		PDASC	0.067	0.60	2.29	1	40
0.6	0.5	Lasso	0.865	12.02	20.69	0.18	41.86
		MCP	0.6123	3.27	20.65	0.82	40.05
		SCAD	0.615	4.23	21.96	0.91	40.04
		PDASC	0.041	0.32	2.69	0.98	40.04
	1	Lasso	0.867	12.04	22.78	0.17	41.96
		MCP	0.613	3.32	22.91	0.82	40.04
		SCAD	0.622	4.28	18.30	0.92	40.03
		PDASC	0.103	0.76	2.43	0.96	40.09
0.8	0.5	Lasso	0.853	11.98	20.69	0.15	41.88
		MCP	0.312	1.81	20.70	0.96	39.99
		SCAD	0.567	4.00	22.00	0.96	40.02
		PDASC	0.025	0.23	2.79	1	40
	1	Lasso	0.855	11.99	23.51	0.17	41.95
		MCP	0.319	1.89	23.50	0.96	39.99
		SCAD	0.573	4.05	18.41	0.96	40.02
		PDASC	0.051	0.46	2.49	0.99	40.03

Table 3: Numerical results with $n=400$, $p=4000$, $T=20$, $R=10$, $\sigma=0.5$ and 1, $C=2:2:8$ and $\mathbf{X}$ follows (III).

C	σ	Method	AE	RE (10^{-2})	Time	RP	MSES
2	0.5	Lasso	0.703	11.03	3.95	0.81	20.25
		MCP	0.178	1.32	3.86	1	20
		SCAD	0.399	3.09	4.16	1	20
		PDASC	0.056	0.57	0.81	1	20
	1	Lasso	0.707	11.03	3.95	0.54	20.74
		MCP	0.212	1.78	4.06	1	20
		SCAD	0.415	3.36	4.41	1	20
		PDASC	0.112	1.15	0.69	1	20
4	0.5	Lasso	0.691	10.97	3.95	0.89	20.12
		MCP	0.173	1.28	3.95	1	20
		SCAD	0.399	3.07	4.21	1	20
		PDASC	0.057	0.58	0.82	1	20
	1	Lasso	0.727	11.18	3.86	0.63	20.52
		MCP	0.206	1.73	4.11	1	20
		SCAD	0.418	3.33	4.27	1	20
		PDASC	0.113	1.16	0.71	1	20
6	0.5	Lasso	0.694	11.08	4.02	0.89	20.11
		MCP	0.169	1.29	3.93	1	20
		SCAD	0.397	3.10	4.31	1	20
		PDASC	0.055	0.57	0.85	1	20
	1	Lasso	0.701	11.08	3.93	0.68	20.47
		MCP	0.200	1.72	4.09	1	20
		SCAD	0.416	3.37	4.38	1	20
		PDASC	0.110	1.13	0.64	1	20
8	0.5	Lasso	0.687	10.89	3.96	0.91	20.1
		MCP	0.161	1.20	4.20	1	20
		SCAD	0.365	2.82	4.18	1	20
		PDASC	0.057	0.58	0.81	1	20
	1	Lasso	0.711	11.15	4.01	0.68	20.4
		MCP	0.194	1.65	4.10	1	20
		SCAD	0.385	3.10	4.34	1	20
		PDASC	0.114	1.17	0.57	1	20

4.2 Influence of the model parameters

In this subsection, we consider the influence of model parameters, including sample size n, ambient dimension p, correlation ρ, and noise level σ, on the performance of PDASC and another three alternative methods in terms of computational speed and support recovery. To this end, we test all the four methods with $\mathbf{X}$ generated according to setting (I). The sample size n, the covariate dimension p, the sparsity level T, the correlation ρ, and the noise level σ are set as following:

- $n=100:50:600$, $p=600$, $T=10$, $R=5$, $\sigma=1$, $\rho=0.5$.
- $n=200$, $p=300:300:3000$, $T=10$, $R=5$, $\sigma=1$, $\rho=0.5$.
- $n=200$, $p=600$, $T=10$, $R=5$, $\sigma=1$, $\rho=0.1:0.1:0.8$.
- $n=200$, $p=600$, $T=10$, $R=5$, $\sigma=0.1:0.1:2.5$, $\rho=0.5$.

The evaluation measures how RP and time change with respect to n,p,ρ and σ. The results are shown in Figs. 1-2. For example, the four sub-figures in Fig. 1 show the performance of RP of all the four methods represented with four solid lines with different colors as n,p,ρ and σ vary, respectively. We can see that PDASC (the black solid line) is on the top of each sub-figures in Fig. 1, and is at the bottom of each sub-figures in Fig. 2, which indicates that PDASC achieves higher support recovery probability, and faster speed than those of Lasso, MCP and SCAD.

4.3 Real data example

We further illustrate the application PDASC by analyzing the Breast cancer gene expression data set (bcTCGA), which have been studied by [16,18,22] and can be downloaded from `http://myweb.uiowa.edu/pbreheny/data/bcTCGA.html`.

This data set comes from breast cancer tissue samples deposited to The Cancer Genome Atlas (TCGA) project. In this data set, expression measurements of 17814 genes, including BRCA1, from 536 patients are available Among all genes in bcTCGA, BRCA1 is the first gene identified that increases the risk of early onset breast cancer. BRCA1 is also likely to interact with many other genes, including tumor suppressors and regulators of the cell division cycle. Hence we let BRCA1 be the response vector $\mathbf{y}$. There are 491 genes with missing data, which are excluded from the analysis. Hence, the dimension of the covariate matrix $\mathbf{X}$ is 536×17322. We use PDASC to fit this data set with the linear model. We also apply Lasso, MCP and SCAD to this data set by using the R package ncvreg [2]. The detailed results are showed in Table 4.

In Table 4, Lasso selects more genes than another three methods, and MCP selects the fewest genes. The coefficients of the common selected genes using these four methods have same sign. Some estimated coefficients of Lasso and SCAD are nearly close to zero such as *KIAA0101, LSM12, MFGE8, UHRF1, CENPQ, SPRY2* and *CDC6*. PDASC yields

Table 4: The estimation of bcTCGA.

Gene name	number	Lasso	MCP	SCAD	PDASC
ABHD13	82	-	-0.022	-	-
C17orf53	1743	0.082	-	0.091	-
CCDC56	2739	0.056	-	0.039	-
CDC25C	2964	0.028	-	0.027	-
CDC6	2987	0.011	-	0.005	0.069
CEACAM6	3076	-	-	-	0.023
CENPK	3105	0.018	-	0.011	-
CRBN	3676	-	-0.057	-	-
DTL	4543	0.091	0.355	0.089	-
FABP1	5081	-	-	-	-0.126
FAM77C	5261	-	-	-	0.017
FGFRL1	5481	-	-0.022	-	-
HBG1	6616				0.069
HIST2H2BE	6811	-	-0.012	-	-
KHDRBS1	7709	-	0.112	-	-
KIAA0101	7719	0.007	-	-	-
KLHL13	8002	-	-0.013	-	-
LSM12	8782	0.006	-	-	-
MFGE8	9230	-0.005	-	-	-
MIA	9359				-0.006
NBR2	9941	0.273	0.504	0.235	0.555
NPY1R	10311				0.008
PSME3	12146	0.085	-	0.074	-
RDM1	12615				0.058
SETMAR	13518	-	-0.063	-	-
SLC25A22	13833	-	0.017	-	-
SLC6A4	14021	-	-	-	0.013
SPAG5	14296	0.024	0.048	0.013	0.180
SPRY2	14397	-0.012	-	-0.005	-
TIMELESS	15122	0.033	-	0.036	-
TMPRSS4	15432	-	-	-	0.031
TOP2A	15535	0.035	-	0.035	0.128
TUBA1B	15882	0.021	-	-	-
TYR	15953	-	-	-	0.132
UHRF1	16087	0.003	-	-	-
VPS25	16315	0.106	0.307	0.108	-

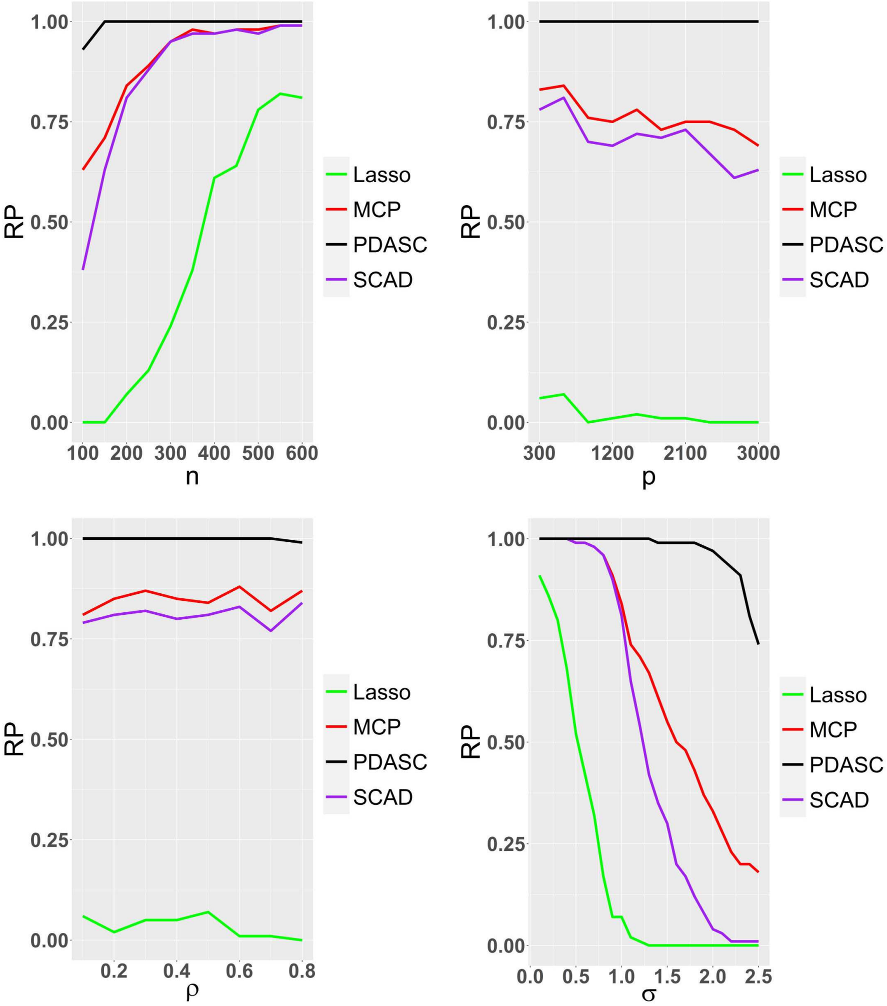

Figure 1: RP versus n,p,ρ and σ.

similar values of the estimated coefficients to these of MCP for genes *NBR2* and *SPAG5*, and yields the similar value of the estimated coefficients with Lasso and SCAD for gene *CDC6, NBR2, SPAG5, TOP2A*.

5 Conclusion

In this paper, we consider the truncated L_1 regularization [5] to spare linear regression model. We establish the nonasymptotic error bounds and study its support recovery property. Moreover, a primal dual active set algorithm with continuation strategy (PDASC) is proposed for variable estimation and selection. Both simulation data and

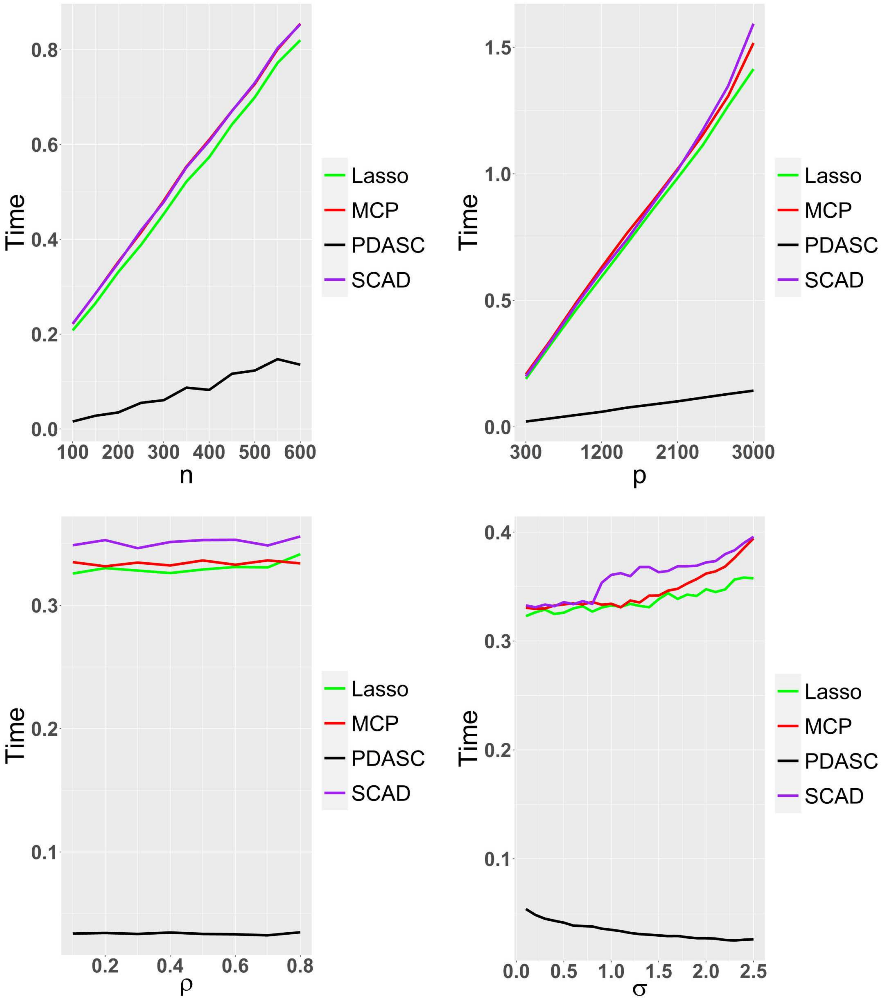

Figure 2: Time versus n,p,ρ and σ.

real data demonstrates the superior performance of the PDASC in terms of accuracy, support recovery and computational efficiency in comparison with the lasso, MCP and SCAD methods. How to extend these results to the nonlinear models can be our further research.

Acknowledgments

The work of Shuyang Dai is partially supported by the National Key Research and Development Program of China (No. 12071363), and the work of Xiliang Lu is partially supported by the National Key Research and Development Program of China (No.

2020YFA0714200), the National Science Foundation of China (No. 11871385) and the Natural Science Foundation of Hubei Province (No. 2019CFA007). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Appendix

In this appendix, we will prove Theorems 2.1-2.3 and Lemma 3.1.

To prove Theorems 2.1-2.3, we first introduce three lemmas (Lemmas A.1-A.3). These three Lemmas are available in the literatures, we scratch the proof for the completeness.

Lemma A.1 (Lemma 1 in [25]). *If $\boldsymbol{\beta}^\diamond$ is the global solution of the optimization problem* (1.2), *then*

$$\left\|\mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\diamond)/n\right\|_\infty \le \lambda.$$

Proof. As $\boldsymbol{\beta}^\diamond$ is the minimizer of (1.2). Then it yields that for all real $t\in\mathbb{R}$,

$$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\diamond\|_2^2/(2n)+\rho_\lambda(\beta_j^\diamond)\le\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\diamond-\mathbf{x}_j t\right\|_2^2/(2n)+\rho_\lambda(\beta_j^\diamond+t).$$

Moreover, $\rho_\lambda(t)$ satisfies the subadditive property in t, then

$$t\mathbf{x}_j^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\diamond)/n\le t^2\left\|\mathbf{x}_j\right\|_2^2/(2n)+\rho_\lambda(\beta_j^\diamond+t)-\rho_\lambda(\beta_j^\diamond)\le t^2/2+\rho_\lambda(t).$$

Thus we have

$$\left\|\mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\diamond)/n\right\|_\infty\le\inf_{t>0}\,[t/2+\rho_\lambda(t)/t]=\lambda.$$

This completes the proof. □

Lemma A.2 (Lemma 2 in [25]). *Assume the η-NC condition* (2.2) *with $\eta\in(0,1)$. Suppose $\boldsymbol{\beta}^\diamond$ is the global solution of* (1.2). *Let $\boldsymbol{\Delta}=\boldsymbol{\beta}^\diamond-\boldsymbol{\beta}^*$ and $\xi=(1+\eta)/(1-\eta)$. Then,*

$$\|\mathbf{X}\boldsymbol{\Delta}\|_2^2/(2n)+\|\rho(\boldsymbol{\Delta}_{I^*},\lambda)\|_1\le\xi\|\rho(\boldsymbol{\Delta}_{A^*},\lambda)\|_1.$$

Proof. From the definition of $\boldsymbol{\beta}^\diamond$, we have

$$\begin{aligned}0&\le\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^*\|_2^2/(2n)+\|\rho(\boldsymbol{\beta}^*,\lambda)\|_1-\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\diamond\|_2^2/(2n)-\|\rho(\boldsymbol{\beta}^\diamond,\lambda)\|_1\\&=-\|\mathbf{X}\boldsymbol{\Delta}\|_2^2/(2n)+\boldsymbol{\epsilon}^T\mathbf{X}\boldsymbol{\Delta}/n+\|\rho(\boldsymbol{\beta}^*,\lambda)\|_1-\|\rho(\boldsymbol{\beta}^*+\boldsymbol{\Delta},\lambda)\|_1.\end{aligned}$$

By $\eta-NC$ condition (2.2), we have $\|\boldsymbol{\epsilon}/\eta\|_2^2/(2n)\le\|\boldsymbol{\epsilon}/\eta-t\mathbf{X}\boldsymbol{\Delta}\|_2^2/(2n)+\|\rho(t\boldsymbol{\Delta},\lambda)\|_1$ for all $t>0$, which can be written as

$$\boldsymbol{\epsilon}^T\mathbf{X}\boldsymbol{\Delta}/n\le\eta t\|\mathbf{X}\boldsymbol{\Delta}\|_2^2/(2n)+(\eta/t)\|\rho(t\boldsymbol{\Delta},\lambda)\|_1.$$

The above two displayed inequalities yield

$$(1-\eta t)\|\mathbf{X}\boldsymbol{\Delta}\|_2^2/(2n)\le(\eta/t)\|\rho(t\boldsymbol{\Delta},\lambda)\|_1+\|\rho(\boldsymbol{\beta}^*,\lambda)\|_1-\|\rho(\boldsymbol{\beta}^*+\boldsymbol{\Delta},\lambda)\|_1.\qquad\text{(A.1)}$$

Set $t=1$ in (A.1). Then it follows that $\boldsymbol{\beta}^*_{I^*}=\mathbf{0}$, and the sub-additivity of $\rho_\lambda(t)$ that

$$\begin{aligned}&(1-\eta)\|\mathbf{X}\boldsymbol{\Delta}\|_2^2/(2n)\\ \le&\eta\|\rho(\boldsymbol{\Delta},\lambda)\|_1+\|\rho(\boldsymbol{\beta}^*_{A^*},\lambda)\|_1-\|\rho(\boldsymbol{\beta}^*_{A^*}+\boldsymbol{\Delta}_{A^*},\lambda)\|_1-\|\rho(\boldsymbol{\Delta}_{I^*},\lambda)\|_1\\ \le&(\eta+1)\|\rho(\boldsymbol{\Delta}_{A^*},\lambda)\|_1+(\eta-1)\|\rho(\boldsymbol{\Delta}_{I^*},\lambda)\|_1.\end{aligned}$$

Thus we can get

$$\|\mathbf{X}\boldsymbol{\Delta}\|_2^2/(2n)+\|\rho(\boldsymbol{\Delta}_{I^*},\lambda)\|_1\le\xi\|\rho(\boldsymbol{\Delta}_{A^*},\lambda)\|_1.$$

This completes the proof. □

Lemma A.3. *Suppose* (C1) *holds. Then for any* $\alpha\in(0,\frac{1}{2})$, *we have*

$$\mathbb{P}\left(\|\mathbf{X}^T\boldsymbol{\epsilon}/n\|_\infty\le\gamma_n\right)\ge1-2\alpha, \tag{A.2}$$

where $\gamma_n=\sigma\sqrt{\frac{2\log(p/\alpha)}{n}}$.

Proof. This lemma follows from standard probabilities calculations, see, [20, 24]. □

A.1 Proof of Theorem 2.1

Proof. Let $\boldsymbol{\Delta}=\boldsymbol{\beta}^\diamond-\boldsymbol{\beta}^*$. By Lemma A.2, we have

$$\|\mathbf{X}\boldsymbol{\Delta}\|_2^2/(2n)+\|\rho(\boldsymbol{\Delta}_{I^*},\lambda)\|_1\le\xi\|\rho(\boldsymbol{\Delta}_{A^*},\lambda)\|_1.$$

Thus, by (2.1), we can get

$$\|\boldsymbol{\Delta}\|_q\le\left\|\mathbf{X}^\top\mathbf{X}\boldsymbol{\Delta}\right\|_\infty|A^*|^{1/q}/\left\{n\mathrm{RIF}_q(\xi,A^*)\right\}. \tag{A.3}$$

It follows from Lemma A.1 that $\left\|\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\diamond)/n\right\|_\infty\le\lambda$. Besides, we can chose λ such that $\left\|\mathbf{X}^T\boldsymbol{\epsilon}/n\right\|_\infty\le\lambda$. Thus, we have

$$\left\|\mathbf{X}^\top\mathbf{X}\boldsymbol{\Delta}/n\right\|_\infty=\left\|\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^\diamond-\boldsymbol{\epsilon})/n\right\|_\infty\le2\lambda. \tag{A.4}$$

Combing (A.3) with (A.4), it yields that

$$\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^\diamond\|_q\le\frac{2\lambda|A^*|^{1/q}}{\mathrm{RIF}_q(\xi,A^*)}. \tag{A.5}$$

This completes the proof. □

A.2 Proof of Theorems 2.2

Proof. By Lemma A.3, for any $\alpha\in(0,\frac{1}{2})$, we have

$$\mathbb{P}\left(\|\mathbf{X}^T\boldsymbol{\epsilon}/n\|_\infty\geq\gamma_n\right)\leq 2\alpha,$$

where $\gamma_n=\sigma\sqrt{\frac{2\log(p/\alpha)}{n}}$. By (A.5), with probability at least $1-2\alpha$,

$$\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^\diamond\|_q\leq\frac{2\gamma_n|A^*|^{1/q}}{\mathrm{RIF}_q(\xi,A^*)}. \tag{A.6}$$

This completes the proof. □

A.3 Proof of Theorem 2.3

Proof. Set $q=\infty$ in (A.6). Then the condition (C2) shows that

$$\|\boldsymbol{\beta}^*-\boldsymbol{\beta}^\diamond\|_\infty<\|\boldsymbol{\beta}^*_{A^*}\|_{\min}.$$

It implies that $A^*\subseteq\mathrm{supp}(\boldsymbol{\beta}^\diamond)$. □

A.4 Proof of Lemma 3.1

Proof. Let $L_\lambda(\boldsymbol{\beta})=\frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\sum_{i=1}^p\rho_\lambda(\beta_i)$. Assume that the vector $\boldsymbol{\beta}^\diamond=\left(\beta_1^\diamond,\cdots,\beta_p^\diamond\right)\in\mathbb{R}^p$ is the global minimizer of $L_\lambda(\cdot)$. Then, we have

$$\begin{aligned}
&\beta_i^\diamond\in\underset{t\in\mathbb{R}}{\mathrm{argmin}}\ L_\lambda(\beta_1^\diamond,\cdots,\beta_{i-1}^\diamond,t,\beta_{i+1}^\diamond,\cdots,\beta_p^\diamond)\\
\Leftrightarrow\ &\beta_i^\diamond\in\underset{t\in\mathbb{R}}{\mathrm{argmin}}\ \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta}^\diamond-\mathbf{Y}+(t-\beta_i^\diamond)\mathbf{x}_i\|^2+\rho_\lambda(t)\\
\Leftrightarrow\ &\beta_i^\diamond\in\underset{t\in\mathbb{R}}{\mathrm{argmin}}\ \frac{1}{2}(t-\beta_i^\diamond)^2+(t-\beta_i^\diamond)\mathbf{x}_i^T(\mathbf{X}\boldsymbol{\beta}^\diamond-\mathbf{Y})+\rho_\lambda(t)\\
\Leftrightarrow\ &\beta_i^\diamond\in\underset{t\in\mathbb{R}}{\mathrm{argmin}}\ \frac{1}{2}\left(t-\beta_i^\diamond-\mathbf{x}_i^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^\diamond)\right)^2+\rho_\lambda(t).
\end{aligned}$$

Then, by Lemmas 3.3-3.4 of [12], we can conclude that

$$\begin{cases}\mathbf{d}^\diamond=\mathbf{X}^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}^\diamond)/n,\\ \boldsymbol{\beta}^\diamond=\Gamma_\lambda(\boldsymbol{\beta}^\diamond+\mathbf{d}^\diamond),\end{cases}$$

where the i-th element of $\Gamma_\lambda(\cdot)$ is defined as

$$(\Gamma_\lambda(\boldsymbol{\beta}))_i=\begin{cases}0, & |\beta_i|\leq\lambda,\\ \beta_i, & |\beta_i|>\lambda.\end{cases}$$

Conversely, assume that $\boldsymbol{\beta}^\diamond$ and $\mathbf{d}^\diamond$ satisfy (3.1) and (3.2). Denote

$$A^\diamond=\{i\in S: |\beta_i^\diamond+d_i^\diamond|>\lambda\},\quad I^\diamond=(A^\diamond)^c.$$

By (3.1) and (3.2), we can conclude that $|\beta_i^\diamond|>\lambda$ and $d_i^\diamond=0$ for $i\in A^\diamond$, and $|d_j^\diamond|\geq\lambda$ for $j\in I^\diamond$. Then we will show that $L_\lambda(\boldsymbol{\beta}^\diamond+\mathbf{h})\geq L_\lambda(\boldsymbol{\beta}^\diamond)$ if $\mathbf{h}$ is small enough with $\|\mathbf{h}\|_\infty<\lambda$. By some simple computation, it yields that

$$\begin{aligned}L_\lambda(\boldsymbol{\beta}^\diamond+\mathbf{h})-L_\lambda(\boldsymbol{\beta}^\diamond)&=\frac{1}{2n}\|\mathbf{X}\boldsymbol{\beta}^\diamond-\mathbf{Y}+\mathbf{X}\mathbf{h}\|_2^2-\frac{1}{2n}\|\mathbf{X}\boldsymbol{\beta}^\diamond-\mathbf{Y}\|_2^2+\sum_{i=1}^p(\rho_\lambda(\beta_i^\diamond+h_i)-\rho_\lambda(\beta_i^\diamond))\\&\geq\frac{\|\mathbf{X}h\|_2^2}{2n}-\langle\mathbf{h},\mathbf{d}^\diamond\rangle+\sum_{i\in I^\diamond}(\rho_\lambda(\beta_i^\diamond+h_i)-\rho_\lambda(\beta_i^\diamond))\\&\geq\frac{\|\mathbf{X}h\|_2^2}{2n}+\sum_{i\in I^\diamond}\lambda|h_i|-|\langle\mathbf{h}_{I^\diamond},\mathbf{d}_{I^\diamond}^\diamond\rangle|\\&\geq 0,\end{aligned}$$

where the first inequality holds due to $i\in A^\diamond$, $\rho_\lambda(\beta_i^\diamond+h_i)=\rho_\lambda(\beta_i^\diamond)=\frac{\lambda^2}{2}$ for small enough $\mathbf{h}$, the last inequality holds by $|d_i^\diamond|\leq\lambda$ for $i\in I^\diamond$. Therefore $\boldsymbol{\beta}^\diamond$ is a local minimizer of (1.2). □

References

[1] Antoniadis A and Fan J, Regularization of wavelet approximations. Journal of the American Statistical Association, 96(455):939-967, 2001.

[2] Breheny P and Huang J, Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. The Annals of Applied Statistics, 5(1):232, 2011.

[3] Candes E and Tao T, The dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics, 35(6):2313–2351, 2007.

[4] Daubechies I, Defrise M and De Mol C, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on Pure and Applied Mathematics, 57(11):1413–1457, 2010.

[5] Fan J, Comments on ”Wavelets in statistics: A review” by A. Antoniadis. Journal of the Italian Statistical Society, 6(2):131, 1997.

[6] Fan J and Li R, Variable selection via nonconcave penalized likelihood and its oracle properties. Publications of the American Statal Association, 96(456):1348–1360, 2001a.

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

[8] Fan Q, Jiao Y and Lu X, A primal dual active set algorithm with continuation for compressed sensing. IEEE Transactions on Signal Processing, 62(23):6276–6285, 2014.

[9] Frank LE and Friedman JH, A statistical view of some chemometrics regression tools. Technometrics, 35(2):109–135, 1993.

[10] Fu WJ, Penalized regressions: the bridge versus the lasso. Journal of Computational and Graphical Statistics, 7(3):397–416, 1998.

[11] Huang J, Jiao Y, Liu Y and Lu X, A constructive approach to l 0 penalized regression. The Journal of Machine Learning Research, 19(1):403–439, 2018.

[12] Huang J, Jiao Y, Jin B, Liu J, Lu X and Yang C, A unified primal dual active set algorithm for nonconvex sparse recovery. Statistical Science (in press), 2020a.

[13] Huang J, Jiao Y, Kang L, Liu J, Liu Y, Lu X and Yang Y, On newton screening. ArXiv preprint arXiv:200110616,2020b.

[14] Huang ZJ, The sparsity and bias of the lasso selection in high-dimensional linear regression. Annals of Statistics, 36(4):1567–1594, 2008.

[15] Jiao Y, Jin B and Lu X, A primal dual active set with continuation algorithm for the l(0)-regularized optimization problem. Applied and Computational Harmonic Analysis, 39(3):400–426, 2014.

[16] Lv S, Lin H, Lian H and Huang Ja, Oracle inequalities for sparse additive quantile regression in reproducing kernel Hilbert space. The Annals of Statistics, 46(2):781–813, 2018.

[17] Shi Y, Huang J, Jiao Y and Yang Q, A semismooth newton algorithm for high-dimensional nonconvex sparse learning. IEEE Transactions on Neural Networks and Learning Systems, 2019.

[18] Tan A and Huang J, Bayesian inference for high-dimensional linear regression under mnet priors. Canadian Journal of Statistics, 44(2):180–197, 2016.

[19] Tibshirani R, Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.

[20] Wainwright MJ, Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). IEEE Transactions on Information Theory, 55(5):2183–2202, 2009.

[21] Wright SJ, Nowak RD and Figueiredo MAT, Sparse reconstruction by separable approximation. IEEE Transactions on Signal Processing, 57(7):2479–2493, 2009.

[22] Yi C and Huang J, Semismooth newton coordinate descent algorithm for elastic-net penalized huber loss regression and quantile regression. Journal of Computational and Graphical Statistics, 26(3):547–557, 2017.

[23] Zhang CH, Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 38(2):894–942, 2010a.

[24] Zhang CH and Huang J, The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics, 36(4):1567–1594, 2008.

[25] Zhang CH and Zhang T, A general theory of concave regularization for high-dimensional sparse estimation problems. Statistical Science, 27(4):576–593, 2012.

[26] Zhang T, Analysis of multi-stage convex relaxation for sparse regularization. Journal of Machine Learning Research, 11(3), 2010b.

[27] Zhao P and Yu B, On model selection consistency of lasso. Journal of Machine Learning Research, 7:2541–2563, 2006.

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

[29] Zou H and Zhang HH, On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics, 37(4):1733, 2009.