Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 17, No. 2, pp. 227–250, 2021

A NEW PENALIZED LEAST ABSOLUTE DEVIATION MODEL FOR HIGH DIMENSIONAL SPARSE LINEAR REGRESSION AND AN EFFICIENT SEQUENTIAL LINEAR PROGRAMMING ALGORITHM*

Zhenzhi Qin and Liping Zhang†

Abstract: The high dimensional sparse linear regression problem has many important applications in electronic engineering, statistics, and compressed sensing. In this paper, we introduce a new penalized least absolute deviation (LAD) model for the high dimensional sparse linear regression problem. First, we adopt the LAD model because it is suitable for data in the presence of outliers and provides a powerful technique for outlier robustness. Second, we adopt a reweighted difference of ℓ_1 and ℓ_2 norms as a nonconvex regularized term which enables one to reconstruct the sparse signal of interest from substantially fewer measurements. We propose a rule for setting the penalty level and show that the new model can provide a surprisingly good estimation error without any assumptions on the moments of the noise even for Cauchy noise. We present an efficient sequential linear programming algorithm for solving the new model and prove that the generated sequence converges to a stationary point satisfying the first-order optimality condition. Numerical results are also presented to indicate the efficiency of the proposed algorithm.

Key words: *high dimensional sparse linear regression, linear programming, penalized least absolute deviation, global convergence*

Mathematics Subject Classification: *65K10, 49M29, 90C30, 90C26*

1 Introduction

Since the number of observations is much less than the number of unknown coefficients in the high dimensional sparse linear regression problem, analysis of high dimensional data poses many challenges and has attracted much recent interests in many fields such as applied mathematics, machine learning, statistics, economic, finance, and compressed sensing.

In this paper, we consider the following high dimensional linear regression problem

$$\mathbf{y} = A\mathbf{x} + \mathbf{z},$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$ is the n-dimensional vector of outcomes, A is the $n \times m$ measurement matrix, $\mathbf{x}$ is the m-dimensional vector of regression coefficients, and $\mathbf{z} = (z_1, \dots, z_n)^T$ is the n-dimensional vector of measurement noises. We assume $A = (A_1, \dots, A_m)$ where $A_i \in \mathbb{R}^n$ denotes the ith regressor or variable. Throughout the paper, we assume that each vector

*This work was supported by the National Nature Science Foundation of China (Grant No. 11771244).

†Cooresponding author.

A_i is normalized such that $\|A_i\|_2 = \sqrt{n}$ for $i = 1, \dots, m$. We are interested in the high dimensional case where $m \gg n$ and the problem is sparse in the sense that only a small proportion of the coefficients are nonzero. In such a scenario, a key task is identifying and estimating the nonzero coefficients and our goal is to reconstruct the unknown vector $\mathbf{x} \in \mathbb{R}^m$.

In general, the high dimensional linear regression problem can be formulated as

$$\text{minimize } \|\mathbf{x}\|_0 \qquad \text{subject to } \mathbf{y} = A\mathbf{x} + \mathbf{z}, \tag{1.1}$$

where $\|\mathbf{x}\|_0$ (called ℓ_0-norm) denotes the number of nonzero elements of the vector $\mathbf{x}$. For the high dimensional sparse linear regression problem, a key assumption is the sparsity of the true coefficient $\mathbf{x}^*$, which guarantees the model identifiability and enhances the model fitting accuracy and interpretability [13, 25]. Here we assume

$$\|\mathbf{x}^*\|_0 = \kappa < n. \tag{1.2}$$

The set of nonzero coefficients or significant variables of $\mathbf{x}^*$ is unknown. In what follows, the true parameter values of $\mathbf{x}$ and m and κ are implicitly indexed by the sample size n, but we omit the index in our notation whenever this does not cause confusion.

Since the ordinary least squares method is not consistent in the setting of $m \gg n$, many new methods have been advanced to solve (1.1) in recent years. The methods with ℓ_1-norm penalties have been proposed, see e.g., [2, 7, 21, 26]. Among them, a popular approach is the ℓ_1 penalized least squares model (also called Lasso [25])

$$\min \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{1.3}$$

where $\lambda > 0$ is a given penalty level. There are many state-of-the-art algorithms available for (1.3), such as ADMM [3, 29], semismooth Newton ALM [19], and others [1, 4, 16, 17, 28, 33, 32]. Although the Lasso method is analytically simple and has nice properties under the assumption of Gaussian noise and known variance [13], it is not suitable for data in the presence of outliers or heterogeneity. Moreover, the Gaussian assumption may not hold in practice and the estimation of the standard deviation is not easy.

To deal with the cases where the error distribution is unknown or may have a heavy tail, Wang [26] proposed the ℓ_1 penalized least absolute deviation (ℓ_1-LAD) model for (1.1) as follows

$$\hat{\mathbf{x}}_{\ell_1\text{-LAD}} \in \arg\min \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda\|\mathbf{x}\|_1, \tag{1.4}$$

and showed that the optimal solution in (1.4) achieves the estimation error bound

$$\|\hat{\mathbf{x}}_{\ell_1\text{-LAD}} - \mathbf{x}^*\|_2 = O\left(\sqrt{\frac{\kappa \log m}{n}}\right) \tag{1.5}$$

with a nearly universal penalty parameter and established a sure screening property for such an estimator. The ℓ_1-LAD method selects the penalty level λ not depending on noise distribution but only assuming that noises have median 0 and $P(z_i = 0) = 0$ for all $i = 1, \dots, n$. Note that the Cauchy noise satisfies this condition [26]. The LAD-type methods are important when heavy-tailed errors are present. These methods have desired robust properties in linear regression models; see, e.g., [15, 27].

In Lasso or LAD type of methods, the ℓ_1 penalization works as the convex relaxation of the ℓ_0-minimization (1.1). Such a convex relaxation works well [6] and has become widespread. This raises the question of whether a different alternative to ℓ_0 minimization might also find the correct solution, but with a lower measurement requirement than ℓ_1 minimization [8, 31]. Fortunately, it is found that the nonconvex penalized regressions enable one to reconstruct the sparse signal of interest from substantially fewer measurements [9, 30, 34]. On the computational side, researchers have observed that under certain conditions on the sensing matrix A, several nonconvex penalized regressions do produce solutions with better quality [8, 11, 12, 14, 18, 20]. A popular nonconvex penalization is ℓ_p $(0 < p < 1)$ quasi-norm, which was proposed as alternative to ℓ_0 and has been extensively studied; see, e.g., [8, 9, 11, 14, 18, 34]. We are interested in the difference of ℓ_1 and ℓ_2 norms (ℓ_{1-2}) as a nonconvex alternative to ℓ_0, which was first addressed in [12] in the context of nonnegative least squares problems and group sparsity with applications to spectroscopic imaging, and further studied in [20, 30]. The associated minimization is as follows

$$\min \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda(\|\mathbf{x}\|_1 - \|\mathbf{x}\|_2). \tag{1.6}$$

Although minimizing such an nonconvex penalization is generally more difficult, it has advantages over the convex ℓ_1 norm in the sense of enhancing the sparsity [30]. A contour plot of the ℓ_{1-2} metric can be seen in Figure 1. Numerical results reported in [30] indicates that ℓ_{1-2} outperforms $\ell_{1/2}$ and ℓ_1 for a highly coherent matrix A.

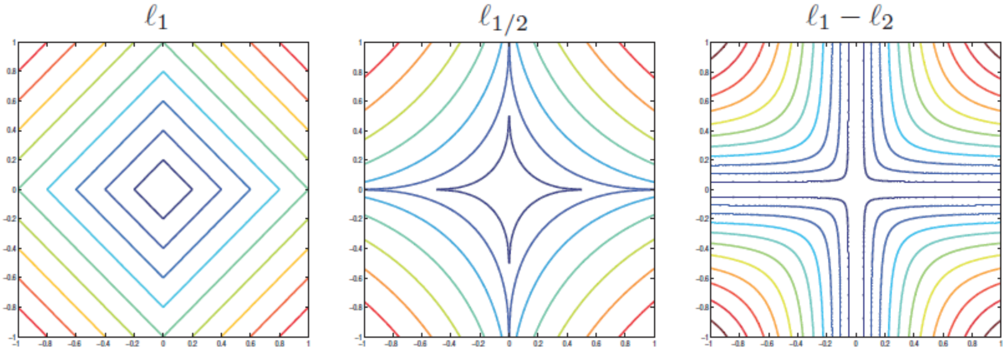


Figure 1: Contours of three sparsity metrics. The level curves of $\ell_1 - \ell_2$ approach the x and y axes as the values get small, hence promoting sparsity.

Motivated by the above observations, in this paper we combine the LAD used in [26] and the reweighted difference of ℓ_1 and ℓ_2 norms from [8, 30] to propose a new penalized regression for (1.1)

$$\hat{\mathbf{x}} \in \arg\min \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda\left(\|W\mathbf{x}\|_1 - \gamma\|W\mathbf{x}\|_2\right), \tag{1.7}$$

where $\gamma \in [0, 1)$ is a free parameter, the weighted matrix $W = \mathrm{diag}(w_1, \ldots, w_m)$ is a positive diagonal matrix given in [8] which can help to improve the signal reconstruction. This new penalized LAD method allows it to combine the sparsity of the nonconvex regularization and the outlier-robustness of the LAD regression. Some analyses show that the new model (1.7) has many advantages over the models given in [26] and [30].

The main contributions of this paper are as follows.

(a) We present a new penalized regression method (1.7) for the high dimensional sparse linear regression problem (1.1). The method combines the advantages of models (1.3) and (1.4). It is not only suitable for data in the presence of outliers but also able to enhance sparsity to improve the signal reconstruction.

(b) We propose rules for setting the penalty level λ and the parameter γ following the general principle of choosing the penalized parameter introduced in [2, 26]; We can obtain the special choice of $\lambda = 2c\sqrt{n\log m}$ and take $\gamma = \frac{t-\beta}{c}$ where $\beta = \max\{w_1^{-1}, \dots, w_m^{-1}\}$ and $c > t > \beta$ is a fixed constant. Such choices are universal and we only assume that the noises have median 0 and $P(z_i = 0) = 0$ for all $i = 1, \dots, n$.

(c) We show that the estimation error bound

$$\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2 = O\left(\sqrt{\frac{\kappa \log m}{n}}\right) \tag{1.8}$$

holds with high probability. Notice that we do not have any assumptions on the moments of the noise, we only need a scale parameter to control the tail probability of the noise. Moreover, the estimation error bound (1.8) still holds even for Cauchy distributed noise where the first order moment does not exist.

(d) We present a sequential linear programming algorithm for solving (1.7) based on the difference of convex functions (DC) algorithm and prove that the generated sequence converges to a stationary point satisfying the first-order optimality condition of (1.7). Especially, we formulate the subproblem at each iteration as a linear programming to handle, which makes the method computationally efficient and promising.

The rest of this paper is organized as follows. In Section 2, we list some existing results and discuss the choice of the penalty level. In Section 3, we present the main result about the estimation error bound. We present an iterative method for solving (1.7) based on the DC algorithm and establish the convergence analysis in Section 4. In Section 5, we report some numerical results, which show the effects of γ and λ. The reported numerical results also indicate that the model (1.7) has good numerical performance for both Gaussian noise and Cauchy noise. Some conclusions are given in Section 6.

Notation. Let us fix some notation. Let $[n] = \{1, \dots, n\}$ and $T \subset [n]$ be an index set and T^C the complementary set, and let $|T|$ be the cardinality of T. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x}^T\mathbf{y} = \langle \mathbf{x}, \mathbf{y} \rangle$ is their inner product. $supp(\mathbf{x}) = \{i \in [n] : \ x_i \neq 0\}$ denotes the support of $\mathbf{x}$, and it is obvious that $\|\mathbf{x}\|_0 = |supp(\mathbf{x})|$. $|\mathbf{x}| \in \mathbb{R}^n$ denotes the vector with the ith component $|x_i|$. $\mathbf{x}_T$ denote the n-dimensional vector such that $(\mathbf{x}_T)_i = x_i$ if $i \in T$ and otherwise $(\mathbf{x}_T)_i = 0$. In addition, we define two n-dimensional vectors by $\mathbf{x}^+ = \max\{\mathbf{0}, \mathbf{x}\}$ and $\mathbf{x}^- = \max\{\mathbf{0}, -\mathbf{x}\}$. For $x \in \mathbb{R}$, $sgn(x)$ is defined as

$$sgn(x) = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{if } x < 0, \\ [-1, 1] & \text{if } x = 0. \end{cases}$$

Finally, we denote $\mathbf{1}_n = (1, \dots, 1)^T \in \mathbb{R}^n$ as the all one vector.

$\boxed{2}$ Preliminaries and Choice of Penalty Level

In this section, we first recall some results given in [26, 30] which will be used in the sequel. Based on these results we then give a rule for choosing the penalty level λ in the penalized regression (1.7).

For any $\mathbf{x} \in \mathbb{R}^n$, it is well known that $\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{n}\|\mathbf{x}\|_2$. The following further result between the norms $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_1$ was given in [30, Lemma 2.1].

Lemma 2.1. *Let* $\mathbf{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$, $\Lambda = supp(\mathbf{x})$ *and* $\|\mathbf{x}\|_0 = s$. *Then,*

$$(s - \sqrt{s}) \min_{i \in \Lambda} |x_i| \le \|\mathbf{x}\|_1 - \|\mathbf{x}\|_2 \le (\sqrt{s} - 1)\|\mathbf{x}\|_2.$$

We now consider the choice rule of λ in (1.7). Here we assume that the measurement errors z_i's are i.i.d. random variables and satisfy $\mathcal{P}(z_i = 0) = 0$ and the median of z_i is 0 for $i \in [n]$. Since the LAD is used in the penalized model (1.7), we can follow the way given in [26]. Following the general principle of choosing the penalty introduced in [2], we know that the subdifferential of $\|\mathbf{y} - A\mathbf{x}\|_1$ evaluated at the point of true coefficient $\mathbf{x}^*$ measures the estimation error in the linear regression. For the penalized regression (1.7), we choose a penalty level λ such that it is greater than the maximum absolute value of subdifferential of $\|\mathbf{y} - A\mathbf{x}\|_1$ at $\mathbf{x}^*$ with high probability. That is, we need to find a penalty level λ for a given constant $c > 0$ and a given small probability α such that

$$\mathcal{P}(\lambda \ge c\|S\|_\infty) \ge 1 - \alpha, \tag{2.1}$$

where S is the subdifferential of $\|\mathbf{y} - A\mathbf{x}\|_1$ at the point of true coefficient $\mathbf{x}^*$, and it is specified by

$$S = A^T\mathbf{u} \text{ and } \mathbf{u} \in \mathbb{R}^n \text{ with the } i\text{th component } u_i = sgn(z_i). \tag{2.2}$$

Since z_i's are independent and have median 0, we know that $\mathcal{P}(u_i = 1) = \mathcal{P}(u_i = -1) = 0.5$ and u_i's are independent. Then the distribution of $\mathbf{u}$ is known and hence the distribution of $\|S\|_\infty$ is easy to know for any given A and does not depend on any unknown parameters. Therefore, the inequality (2.1) holds when we take $1 - \alpha$ quantile of $\|S\|_\infty$ as λ/c.

After the above discussion, we immediately obtain the following choice rule for the penalty level λ via the same proof as given in [26, Lemma 1].

Proposition 2.2. *The choice of penalty* $\lambda = c\sqrt{2V(\alpha)n\log m}$ *satisfies the inequality (2.1), where* $V(\alpha) > 0$ *is a constant such that* $2m^{1-V(\alpha)} \le \alpha$.

Taking $\alpha = 2/m$ and $V(\alpha) = 2$, by Proposition 2.2, we can easily obtain that the special choice of penalty $\lambda = 2c\sqrt{n\log m}$ satisfies the inequality

$$\mathcal{P}(\lambda \ge c\|S\|_\infty) \ge 1 - \frac{2}{m}.$$

In this way, when the data size n is larger, since the variance of measurement errors z_i is larger than zero, the value of $\|\mathbf{y} - A\mathbf{x}^*\|_1$ might be large. And then in the minimization program (1.7), it will induce that the value of $\|\mathbf{y} - A\hat{\mathbf{x}}\|_1$ much less than the value of $\|\mathbf{y} - A\mathbf{x}^*\|_1$ if the penalty level λ were small. This well result the overfitting of the minimization program (1.7), i.e. $\|\mathbf{y} - A\hat{\mathbf{x}}\|_1 \ll \|\mathbf{y} - A\mathbf{x}^*\|_1$. So we need λ increases with the increase of data size n. Clearly, the above choices of penalty level are simple and they do not require any assumptions on A, m and n, and the distribution of measurement errors z_i. As long as z_i's are independent random variables with median 0 and $\mathcal{P}(z_i = 0) = 0$, the choices satisfy our requirement (2.1) and do not rely on the Gaussian assumption. This is a big advantage over the traditional lasso method. For simplicity, we use $\lambda = 2c\sqrt{n\log m}$ as the default choice for the penalty level in (1.7).

Note that the difference between (1.4) and (1.7) is the penalization. So, in order to get the estimation error bound (1.8), following the way to show (1.5) given in [26], we need the following lemmas which were given in [26, Lemmas 4, 5 and 7].

Lemma 2.3. *For the random variable* $\mathbf{z} = (z_1, \ldots, z_n)^T \in \mathbb{R}^n$ *and any given real number* $\eta \in \mathbb{R}$*, if there is a positive constant* $a > 0$ *such that*

$$\mathcal{P}(z_i \geq \eta) \leq \frac{1}{2 + a\eta} \quad \forall \eta \geq 0; \quad \mathcal{P}(z_i \leq \eta) \leq \frac{1}{2 + a|\eta|} \quad \forall \eta < 0 \tag{2.3}$$

holds for $i \in [n]$*, then we have the expectation*

$$E(|z_i + \eta| - |z_i|) \geq \frac{a}{16}|\eta| \min\left\{|\eta|, \frac{6}{a}\right\} \quad \forall i \in [n].$$

Lemma 2.4. *For any vector* $\mathbf{z} = (z_1, \ldots, z_n)^T \in \mathbb{R}^n$ *and any real number* $u > 0$*, we have*

$$\sum_{i=1}^{n} |z_i| \min\{|z_i|, u\} \geq \begin{cases} \frac{u\|\mathbf{z}\|_1}{2} & \textit{if } \|\mathbf{z}\|_1 \geq \frac{nu}{2}, \\ \|\mathbf{z}\|_2^2 & \textit{if } \|\mathbf{z}\|_1 < \frac{nu}{2}. \end{cases}$$

Remark 2.5. The condition (2.3) introduces the scale assumptions on the measurement errors $z_i, i \in [n]$. Here a serves as a scale parameter of the distribution of z_i. This is a very weak condition and even Cauchy distribution satisfies it [26, Lemmas 4 and 5].

3 Estimation Error Bound

The purpose of this section is to establish the upper bound (1.8) for the estimation error of our model (1.7). For simplicity, it follows from Proposition 2.2 that in this section we use $\lambda = 2c\sqrt{n \log m}$ as the default choice of penalty level, and we assume that this penalty level satisfies $\lambda \geq c\|S\|_\infty$ for some fixed constant $c > 0$.

For the $m \times m$ positive diagonal matrix $W = \operatorname{diag}(w_1, \ldots, w_m)$, we define $\beta = \max\{w_i^{-1} : i \in [m]\}$. Let $\Lambda = \mathit{supp}(\mathbf{x}^*)$ and $\mathbf{h} = W(\mathbf{x}^* - \hat{\mathbf{x}})$, where $\mathbf{x}^*$ is the true coefficient in (1.2) and $\hat{\mathbf{x}}$ is defined as (1.7). Clearly, $\Lambda = \mathit{supp}(W\mathbf{x}^*)$, $\mathbf{h} = \mathbf{h}_\Lambda + \mathbf{h}_{\Lambda^C}$ with $\mathbf{h}_\Lambda = W(\mathbf{x}_\Lambda^* - \hat{\mathbf{x}}_\Lambda)$ and $\mathbf{h}_{\Lambda^C} = -W\hat{\mathbf{x}}_{\Lambda^C}$. Then, by Lemma 2.1, we have

$$\|W\hat{\mathbf{x}}\|_1 - \gamma\|W\hat{\mathbf{x}}\|_2 \geq \|W\mathbf{x}^*\|_1 - \|\mathbf{h}_\Lambda\|_1 + \|\mathbf{h}_{\Lambda^C}\|_1 - \gamma\|W\mathbf{x}^*\|_2 - \gamma\|\mathbf{h}\|_2.$$

Since $\hat{\mathbf{x}}$ is an optimal solution of (1.7), we get

$$\|\mathbf{y} - A\hat{\mathbf{x}}\|_1 + \lambda\left(\|W\hat{\mathbf{x}}\|_1 - \gamma\|W\hat{\mathbf{x}}\|_2\right) \leq \|\mathbf{z}\|_1 + \lambda\left(\|W\mathbf{x}^*\|_1 - \gamma\|W\mathbf{x}^*\|_2\right).$$

Hence, It follows from $\|\mathbf{h}\|_2 \leq \|\mathbf{h}_\Lambda\|_2 + \|\mathbf{h}_{\Lambda^C}\|_2$ that

$$\|\mathbf{z} + AW^{-1}\mathbf{h}\|_1 - \|\mathbf{z}\|_1 \leq \lambda\big(\|\mathbf{h}_\Lambda\|_1 + \gamma\|\mathbf{h}_\Lambda\|_2 - \|\mathbf{h}_{\Lambda^C}\|_1 + \gamma\|\mathbf{h}_{\Lambda^C}\|_2\big). \tag{3.1}$$

In order to get the error bound (1.8), we need the following lemma to analyze the upper bound of the right hand side of (3.1).

Lemma 3.1. *Let* $\mathbf{h}$*,* β *and* Λ *be defined as above. Then,*

$$\left(1 + \frac{\beta}{c}\right)\|\mathbf{h}_\Lambda\|_1 + \gamma\|\mathbf{h}_\Lambda\|_2 \geq \left(1 - \frac{\beta}{c}\right)\|\mathbf{h}_{\Lambda^C}\|_1 - \gamma\|\mathbf{h}_{\Lambda^C}\|_2.$$

Proof. By the convexity of $\|\mathbf{z}\|_1$ and (2.2), we have

$$\begin{aligned}\|\mathbf{z} + AW^{-1}\mathbf{h}\|_1 - \|\mathbf{z}\|_1 &\geq \langle \mathbf{u}, AW^{-1}\mathbf{h}\rangle = \langle A^T\mathbf{u}, W^{-1}\mathbf{h}\rangle \geq -\beta\|\mathbf{h}\|_1\|A^T\mathbf{u}\|_\infty \\ &= -\beta\|\mathbf{h}\|_1\|S\|_\infty \geq -\frac{\lambda\beta}{c}(\|\mathbf{h}_\Lambda\|_1 + \|\mathbf{h}_{\Lambda^C}\|_1),\end{aligned}$$

which, together with (3.1), implies that

$$\|\mathbf{h}_\Lambda\|_1 + \gamma\|\mathbf{h}_\Lambda\|_2 - \|\mathbf{h}_{\Lambda^C}\|_1 + \gamma\|\mathbf{h}_{\Lambda^C}\|_2 \geq -\frac{\beta}{c}(\|\mathbf{h}_\Lambda\|_1 + \|\mathbf{h}_{\Lambda^C}\|_1).$$

Hence, the desired result holds. □

Let c, t be two given constants such that $c > t > \beta$. Taking

$$\gamma = \frac{t-\beta}{c} \tag{3.2}$$

in Lemma 3.1, we have

$$\|\mathbf{h}_\Lambda\|_1 \geq \bar{c}\|\mathbf{h}_{\Lambda^C}\|_1, \quad \bar{c} = \frac{c-t}{c+t}, \tag{3.3}$$

which is the similar result given in [26] for the ℓ_1-LAD model (1.4). Define the following set

$$\Delta_{\bar{c}} = \{\mathbf{d} \in \mathbb{R}^m : \ \|\mathbf{d}_T\|_1 \geq \bar{c}\|\mathbf{d}_{T^C}\|_1, \text{where } T \subset [m] \text{ and } |T| \leq \kappa\}.$$

Clearly, $\mathbf{h} \in \Delta_{\bar{c}}$ from (3.3). This conclusion is pretty crucial for our discussion, and it is also important for the ℓ_1-LAD method [26] and the classical lasso method [2].

To study the performance of (1.7), we discuss some properties of matrix $\bar{A} = AW^{-1}$. For simplicity, we follow the description as in [26]. We first define two important constants λ_κ^u and λ_κ^l as

$$\lambda_\kappa^u = \sup_{\mathbf{d}\in\mathbb{R}^m, 0<\|\mathbf{d}\|_0\leq\kappa} \frac{\|\bar{A}\mathbf{d}\|_2^2}{n\|\mathbf{d}\|_2^2}, \quad \lambda_\kappa^l = \inf_{\mathbf{d}\in\mathbb{R}^m, 0<\|\mathbf{d}\|_0\leq\kappa} \frac{\|\bar{A}\mathbf{d}\|_2^2}{n\|\mathbf{d}\|_2^2}.$$

Here the vector $\mathbf{d} \in \mathbb{R}^m$ with $0 < \|\mathbf{d}\|_0 \leq \kappa$ is called a κ-sparse vector. The definition of the above constants is related to the Restricted Isometry Constants [5]. The value of $\max\{n\lambda_\kappa^u - 1, 1 - n\lambda_\kappa^l\}$ is called κ-restricted isometry property (RIP), which is widely used in compressed sensing. The (κ_1, κ_2)-restricted orthogonality constant $\theta_{\kappa_1,\kappa_2}$ is defined to be the smallest number such that

$$|\langle\bar{A}\mathbf{d}_1, \bar{A}\mathbf{d}_2\rangle| \leq n\theta_{\kappa_1,\kappa_2}\|\mathbf{d}_1\|_2\|\mathbf{d}_2\|_2$$

holds for any κ_1 and κ_2 sparse vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ with disjoint supports. In compressed sensing, the more $n\lambda_\kappa^u$ and $n\lambda_\kappa^l$ tend to 1, the more precise that the reconstruction of sparse signals will be. With overwhelming probability, it can be seen that random Gaussian, random Bernoulli, and random partial Fourier matrices satisfy the RIP with small difference between $n\lambda_\kappa^u$, $n\lambda_\kappa^l$ and 1 [10, 22]. This good properties will make our model (1.7) work well.

Following [26], we define the following constant for the matrix $\bar{A}$ [2, 26]:

$$K_\kappa^l(\bar{c}) = \min_{\mathbf{h}\in\Delta_{\bar{c}}} \frac{\|\bar{A}\mathbf{h}\|_1}{n\|\mathbf{h}_T\|_2}.$$

To show the performance of the estimator (1.7), we should ensure that $K_\kappa^l(\bar{c})$ is bounded away from 0 or approaches 0 slowly enough as in the ℓ_1-LAD case (1.4) [26]. From [26], if we set $0 \leq \frac{1}{\bar{c}} \leq \sqrt{2}$ and $\kappa\log(m) = o(n)$, when n large enough, we know that $K_\kappa^l(\bar{c})$ is bounded away from 0 with high probability where $\bar{A}$ is under gaussian random design case, i.e. the entries of A are generated by independent and identically distributed $N(0,1)$ random variables. For simplicity, we will write $K_\kappa^l(\bar{c})$ as K_κ^l whenever no confusion.

To obtain the error bound (1.8), we now establish a lemma similar to [26, Lemma 3]. From (3.1), we have

$$\|\mathbf{z}+\bar{A}\mathbf{h}\|_1-\|\mathbf{z}\|_1\leq\lambda\Big((1+\gamma)\|\mathbf{h}_\Lambda\|_1+\gamma\|\mathbf{h}_{\Lambda^C}\|_2\Big).$$

Then we try to bound the estimation error via investigating the random variable

$$\frac{1}{\sqrt{n}}(\|\mathbf{z}+\bar{A}\mathbf{h}\|_1-\|\mathbf{z}\|_1).$$

To do so, we define a random variable $B(\mathbf{d})$ for any vector $\mathbf{d}\in\mathbb{R}^m$ as

$$B(\mathbf{d})=\frac{1}{\sqrt{n}}|(\|\mathbf{z}+\bar{A}\mathbf{d}\|_1-\|\mathbf{z}\|_1)-E(\|\mathbf{z}+\bar{A}\mathbf{d}\|_1-\|\mathbf{z}\|_1)|.$$

We obtain the following important result about $B(\mathbf{d})$ via the same proof as [26, Lemma 3].

Lemma 3.2. *Suppose that z_i's are independent random variables. For $m>3\sqrt{\kappa}$, we have*

$$\mathcal{P}\left(\sup\{B(\mathbf{d}):\ \|\mathbf{d}\|_0=\kappa,\|\mathbf{d}\|_2=1\}\geq(1+2\sigma\sqrt{\lambda_\kappa^u})\sqrt{2\kappa\log m}\right)\leq 2m^{-4\kappa(\sigma^2-1)},$$

where $\sigma>1$ is a constant.

Remark 3.3. Lemma 3.2 indicates that for any κ sparse vector $\mathbf{d}\in\mathbb{R}^m$,

$$\frac{1}{\sqrt{n}}(\|\mathbf{z}+\bar{A}\mathbf{d}\|_1-\|\mathbf{z}\|_1)\geq\frac{1}{\sqrt{n}}E(\|\mathbf{z}+\bar{A}\mathbf{d}\|_1-\|\mathbf{z}\|_1)-(1+2\sigma\sqrt{\lambda_\kappa^u})\sqrt{2\kappa\log m}\|\mathbf{d}\|_2 \quad (3.4)$$

holds with probability at least $1-2m^{-4\kappa(\sigma^2-1)}$. This means that the value of the random variable $\frac{1}{\sqrt{n}}(\|\mathbf{z}+\bar{A}\mathbf{d}\|_1-\|\mathbf{z}\|_1)$ is very close to its expectation with high probability. Clearly, the expectation is not random and much easier to analyze than the random variable itself. Hence, this observation will help us to prove the error bound (1.8) by using Lemma 2.3.

Now we present our main theorem on the error bound (1.8) based on the following conditions:

$$\lambda_{2\kappa}^l>\theta_{\kappa,2\kappa}\alpha,\ \text{where}\ \alpha=\frac{c}{c-\beta}\frac{\gamma}{\sqrt{\kappa}-1}+\frac{c+\beta}{c-\beta}\frac{\sqrt{\kappa}}{\sqrt{\kappa}-1}, \quad (3.5)$$

and

$$\frac{3\sqrt{n}}{16\sqrt{2}}K_{2\kappa}^l>\frac{\lambda\sqrt{\kappa}}{\sqrt{n}}+\left(\frac{\lambda\gamma}{\sqrt{n}}+\sigma_1\sqrt{2\kappa\log m}\right)\frac{2c\sqrt{\kappa}+t-c}{(c-\beta)(\sqrt{\kappa}-1)}, \quad (3.6)$$

where $\sigma_1=1+2\sigma\sqrt{\lambda_\kappa^u}$ and $\sigma>1$, $c>t>\beta$ are given as (3.2).

We give some remarks for the above two conditions.

Remark 3.4. The condition (3.5) shows that $\theta_{\kappa,2\kappa}$ can be bounded by $\lambda_{2\kappa}^l$. Note that α decreases with κ, which means that α can be bounded by a constant based on c, β and γ. Hence, the condition (3.5) can be replaced by a number of similar RIP conditions [5].

Remark 3.5. Condition (3.6) implies that the columns of A cannot be too sparse. This is because if the columns of A are sparse then the ℓ_1 norm of columns of A will be small, hence the value $K_{2\kappa}$ will be small.

If we take $\lambda=2c\sqrt{n\log m}$ in (3.6) and assume $\kappa\log(m)=o(n)$ (which is a natural result of the sparsity of the data), then we have $K_{2\kappa}^l(\bar{c})>\Gamma\sqrt{\frac{\kappa\log m}{n}}$, where Γ is bounded by a constant. By the similar arguments in [26], under gaussian random design case, we show that $K_{2\kappa}$ is bounded away from zero for n large enough when $0\leq\frac{1}{\bar{c}}<\sqrt{2}$. This implies that the condition (3.6) must hold for n large enough if $\bar{c}>\frac{1}{\sqrt{2}}$.

Theorem 3.6. *Consider the high dimensional linear regression model (1.1). Let m, κ satisfy $m > 3\sqrt{\kappa}$ and $\bar{A} = AW^{-1}$. Suppose that $z_1, z_2, \ldots, z_n$ are i.i.d. random variables satisfying (2.3), the conditions (3.5) and (3.6) hold. Then with probability at least $1 - 2m^{-4\kappa(\sigma^2-1)+1}$, the estimator $\hat{\mathbf{x}}$ in (1.7) satisfies*

$$\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2 \leq C\sqrt{\frac{\kappa \log m}{n}},$$

where the constant C is given as

$$C = \frac{16\sqrt{2}\lambda_{2\kappa}^u(1+\alpha)\beta}{a(\lambda_{2\kappa}^l - \theta_{\kappa,2\kappa}\alpha)^2}\left(2c + \left(\frac{2c\gamma}{\sqrt{\kappa}} + \sqrt{2}\sigma_1\right)\frac{2c\sqrt{\kappa}+t-c}{(c-\beta)(\sqrt{\kappa}-1)}\right)$$

with $\sigma_1 = 1 + 2\sigma\sqrt{\lambda_\kappa^u}$ and $\sigma > 1$.

Proof. Recall that $\mathbf{h} = W(\mathbf{x}^* - \hat{\mathbf{x}})$ and $\Lambda = supp(\mathbf{x}^*)$, and then arrange the indices in Λ^C in the sequence of declining magnitude of h_{Λ^C} and divide Λ^C into subsets of size k. Then $\Lambda^C = \Lambda_1 \bigcup \Lambda_2 \bigcup \ldots \bigcup \Lambda_l$, where each Λ_i comprises k indices probably except Λ_l. Denote $\Lambda_0 = \Lambda$, $\Lambda_* = \Lambda \bigcup \Lambda_1$.

From $\Lambda \subset \Lambda_*$ and (3.3), it is shown that

$$\|\mathbf{h}_{\Lambda_*}\|_1 \geq \bar{c}\|\mathbf{h}_{\Lambda_*^C}\|_1,$$

which implies that

$$\mathbf{h} \in \Delta_{\bar{c}} := \{\mathbf{d} \in \mathbb{R}^m : \ \|\mathbf{d}_T\|_1 \geq \bar{c}\|\mathbf{d}_{T^C}\|_1, \text{where } T \subset [m] \text{ and } |T| \leq 2\kappa\}.$$

Combining the definition of $K_{2\kappa}^l(\bar{c})$, we know that

$$\|\bar{A}\mathbf{h}\|_1 \geq nK_{2\kappa}^l(\bar{c})\|\mathbf{h}_{\Lambda_*}\|_2. \tag{3.7}$$

It is easy to see that

$$\begin{aligned}\frac{1}{\sqrt{n}}(\|\bar{A}\mathbf{h} + \mathbf{z}\|_1 - \|\mathbf{z}\|_1) \geq& \frac{1}{\sqrt{n}}(\|\bar{A}\mathbf{h}_{\Lambda_0} + \mathbf{z}\|_1 - \|\mathbf{z}\|_1) \\ &+ \sum_{i\geq 1}\frac{1}{\sqrt{n}}(\|\bar{A}(\sum_{j=0}^{i}\mathbf{h}_{\Lambda_j}) + \mathbf{z}\|_1 - \|\bar{A}(\sum_{j=0}^{i-1}\mathbf{h}_{\Lambda_j}) + \mathbf{z}\|_1).\end{aligned}$$

For convenience, define

$$F(\mathbf{d}) = \frac{1}{\sqrt{n}}E(\|\bar{A}\mathbf{d} + \mathbf{z}\|_1 - \|\mathbf{z}\|_1)$$

for any fixed vector $\mathbf{d} \in \mathbb{R}^m$. It follows from (3.4) that with probability at least $1 - 2m^{-4\kappa(\sigma^2-1)}$,

$$\frac{1}{\sqrt{n}}(\|\bar{A}\mathbf{h}_{\Lambda_0} + \mathbf{z}\|_1 - \|\mathbf{z}\|_1) \geq F(\mathbf{h}_{\Lambda_0}) - (1 + 2\sigma\sqrt{\lambda_\kappa^u})\sqrt{2\kappa\log m}\|\mathbf{h}_{\Lambda_0}\|_2.$$

In the same way, for any $i \geq 1$ with probability at least $1 - 2m^{-4\kappa(\sigma^2-1)}$,

$$\begin{aligned}&\frac{1}{\sqrt{n}}(\|\bar{A}(\sum_{j=0}^{i}\mathbf{h}_{\Lambda_j}) + \mathbf{z}\|_1 - \|\bar{A}(\sum_{j=0}^{i-1}\mathbf{h}_{\Lambda_j}) + \mathbf{z}\|_1) \\ &\geq F(\mathbf{h}_{\Lambda_i}) - (1 + 2\sigma\sqrt{\lambda_\kappa^u})\sqrt{2\kappa\log m}\|\mathbf{h}_{\Lambda_i}\|_2.\end{aligned}$$

Let $\sigma_1 = 1 + 2\sigma\sqrt{\lambda_\kappa^u}$, where $\sigma > 1$ is a constant. Combining the above two inequalities, we obtain that with probability at least $1 - 2m^{-4\kappa(\sigma^2-1)+1}$,

$$\frac{1}{\sqrt{n}}(\|\bar{A}\mathbf{h} + \mathbf{z}\|_1 - \|\mathbf{z}\|_1) \geq F(\mathbf{h}) - \sigma_1\sqrt{2\kappa\log m}\sum_{i\geq 0}\|\mathbf{h}_{\Lambda_i}\|_2. \tag{3.8}$$

Then,

$$\begin{aligned}\sum_{i=2}^{l}\|\mathbf{h}_{\Lambda_i}\|_2 \leq \sum_{i=2}^{l}\frac{\|\mathbf{h}_{\Lambda_i}\|_1 - \|\mathbf{h}_{\Lambda_i}\|_2}{\sqrt{\kappa}-1} &\leq \frac{\sum_{i=1}^{l-1}\|\mathbf{h}_{\Lambda_i}\|_1 - \|\mathbf{h}_{\Lambda_i}\|_2}{\sqrt{\kappa}-1} \\ &\leq \frac{\|\mathbf{h}_{\Lambda_C}\|_1 - \|\mathbf{h}_{\Lambda_C}\|_2}{\sqrt{\kappa}-1},\end{aligned} \tag{3.9}$$

where the first inequality from Lemma 2.1, the second inequality holds since the magnitude of h_{Λ_i} is larger than those of $h_{\Lambda_{i+1}}$'s $(i \geq 1)$, the third inequality holds due to the triangle inequality.

From Lemma 3.1 and (3.2), we have

$$(1 - \frac{\beta}{c})(\|\mathbf{h}_{\Lambda_C}\|_1 - \|\mathbf{h}_{\Lambda_C}\|_2) \leq (1 + \frac{\beta}{c})\|\mathbf{h}_\Lambda\|_1 + \gamma\|h_\Lambda\|_2,$$

which, together with (3.9), yields

$$\sum_{i=2}^{l}\|\mathbf{h}_{\Lambda_i}\|_2 \leq \frac{c}{c-\beta}\frac{1}{\sqrt{k}-1}\left(\left(1 + \frac{\beta}{c}\right)\|\mathbf{h}_\Lambda\|_1 + \gamma\|\mathbf{h}_\Lambda\|_2\right). \tag{3.10}$$

Combining (3.1), (3.8) and (3.10), we have

$$\begin{aligned}F(\mathbf{h}) &\leq \frac{\lambda}{\sqrt{n}}(\|\mathbf{h}_\Lambda\|_1 + \gamma\|\mathbf{h}_\Lambda\|_2 + \gamma\|\mathbf{h}_{\Lambda^C}\|_2 - \|\mathbf{h}_{\Lambda^C}\|_1) + \sigma_1\sqrt{2\kappa\log m}\sum_{i\geq 0}\|\mathbf{h}_{\Lambda_i}\|_2 \\ &\leq \frac{\lambda}{\sqrt{n}}\|\mathbf{h}_\Lambda\|_1 + \left(\frac{\lambda\gamma}{\sqrt{n}} + \sigma_1\sqrt{2\kappa\log m}\right)\sum_{i\geq 0}\|\mathbf{h}_{\Lambda_i}\|_2 \\ &\leq \frac{\lambda\sqrt{\kappa}}{\sqrt{n}}\|\mathbf{h}_\Lambda\|_2 + \left(\frac{\lambda\gamma}{\sqrt{n}} + \sigma_1\sqrt{2\kappa\log m}\right)\Big(\|\mathbf{h}_\Lambda\|_2 + \|\mathbf{h}_{\Lambda_1}\|_2 \\ &\qquad + \frac{c}{c-\beta}\frac{1}{\sqrt{\kappa}-1}\Big((1 + \frac{\beta}{c})\sqrt{\kappa}\|\mathbf{h}_\Lambda\|_2 + \gamma\|\mathbf{h}_\Lambda\|_2\Big)\Big).\end{aligned}$$

Define

$$C_1 = \frac{\lambda\sqrt{\kappa}}{\sqrt{n}} + \left(\frac{\lambda\gamma}{\sqrt{n}} + \sigma_1\sqrt{2\kappa\log m}\right)\frac{2c\sqrt{\kappa} + t - c}{(c-\beta)(\sqrt{\kappa}-1)}, \quad C_2 = \frac{\lambda\gamma}{\sqrt{n}} + \sigma_1\sqrt{2\kappa\log m}.$$

Clearly, $C_2 \leq C_1$ and the above inequality yields

$$F(\mathbf{h}) \leq C_1\|\mathbf{h}_\Lambda\|_2 + C_2\|\mathbf{h}_{\Lambda_1}\|_2 \leq C_1(\|\mathbf{h}_\Lambda\|_2 + \|\mathbf{h}_{\Lambda_1}\|_2). \tag{3.11}$$

Next, we consider two cases for any real number $a > 0$. First, if $\|\bar{A}h||_1 \geq \frac{3n}{a}$, by Lemmas 2.3 and 2.4, we have

$$F(\mathbf{h}) \geq \frac{3}{16\sqrt{n}}\|\bar{A}\mathbf{h}\|_1.$$

And then, it follows from (3.7) that

$$F(\mathbf{h}) \geq \frac{3\sqrt{n}}{16} K^l_{2\kappa} \|\mathbf{h}_{\Lambda_*}\|_1 \geq \frac{3\sqrt{n}}{16\sqrt{2}} K^l_{2\kappa} (\|\mathbf{h}_\Lambda\|_2 + \|\mathbf{h}_{\Lambda_1}\|_2).$$

Since the condition (3.6) holds, we must have $\|\mathbf{h}_{\Lambda_*}\|_2 = 0$ and hence it follows from (3.10) that $\mathbf{h} = \mathbf{0}$.

On the other hand, if $\|\bar{A}\mathbf{h}\|_1 < \frac{3n}{a}$, by Lemmas 2.3 and 2.4, we have

$$F(\mathbf{h}) \geq \frac{a}{16\sqrt{n}} \|\bar{A}\mathbf{h}\|_2^2. \tag{3.12}$$

Similar to the proof of [5, Theorems 3.1 and 3.2], we can obtain

$$\begin{aligned} |\langle \bar{A}\mathbf{h}_{\Lambda_*}, \bar{A}\mathbf{h} \rangle| &\geq n\lambda^l_{2\kappa} \|\mathbf{h}_{\Lambda_*}\|_2^2 - n\theta_{\kappa,2\kappa} \|\mathbf{h}_{\Lambda_*}\|_2 \sum_{i \geq 2} \|\mathbf{h}_{\Lambda_i}\|_2 \\ &\geq n(\lambda^l_{2\kappa} - \theta_{\kappa,2\kappa}\alpha) \|\mathbf{h}_{\Lambda_*}\|_2^2, \end{aligned}$$

where the last inequality is from (3.10), (3.5) and $\|\mathbf{h}_\Lambda\|_1 \leq \sqrt{\kappa}\|\mathbf{h}_\Lambda\|_2$. In addition,

$$|\langle \bar{A}\mathbf{h}_{\Lambda_*}, \bar{A}\mathbf{h} \rangle| \leq \|\bar{A}\mathbf{h}_{\Lambda_*}\|_2 \|\bar{A}\mathbf{h}\|_2 \leq \|\bar{A}\mathbf{h}\|_2 \sqrt{n\lambda^u_{2\kappa}} \|\mathbf{h}_{\Lambda_*}\|_2.$$

Therefore,

$$\|\bar{A}\mathbf{h}\|_2^2 \geq n\frac{(\lambda^l_{2\kappa} - \theta_{\kappa,2\kappa}\alpha)^2}{\lambda^u_{2\kappa}} \|\mathbf{h}_{\Lambda_*}\|_2 \geq n\frac{(\lambda^l_{2\kappa} - \theta_{\kappa,2\kappa}\alpha)^2}{\sqrt{2}\lambda^u_{2\kappa}} (\|\mathbf{h}_\Lambda\|_2 + \|\mathbf{h}_{\Lambda_1}\|_2),$$

which, together with (3.11) and (3.12) and $\lambda = 2c\sqrt{n \log m}$, implies that

$$\begin{aligned} (\|\mathbf{h}_\Lambda\|_2 + \|\mathbf{h}_{\Lambda_1}\|_2) &\leq \frac{\sqrt{2}\lambda^u_{2\kappa}}{(\lambda^l_{2\kappa} - \theta_{\kappa,2\kappa}\alpha)^2} \frac{16}{a\sqrt{n}} C_1 \\ &\leq \frac{\sqrt{2}\lambda^u_{2\kappa}}{(\lambda^l_{2\kappa} - \theta_{\kappa,2\kappa}\alpha)^2} \frac{16}{a} \Big(2c + (\frac{2c\gamma}{\sqrt{\kappa}} + \sqrt{2}\sigma_1)\cdot \\ &\qquad \frac{2c\sqrt{\kappa} + t - c}{(c-\beta)(\sqrt{\kappa}-1)}\Big) \sqrt{\frac{\kappa \log m}{n}} \end{aligned} \tag{3.13}$$

holds with probability at least $1 - 2m^{-4\kappa(\sigma^2-1)+1}$. By (3.10), we obtain

$$\sum_{i=2}^{l} \|\mathbf{h}_{\Lambda_i}\|_2 \leq \alpha \|\mathbf{h}_{\Lambda_*}\|_2, \tag{3.14}$$

and

$$\frac{1}{\beta}\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2 \leq \|W(\mathbf{x}^* - \hat{\mathbf{x}})\|_2 = \|\mathbf{h}\|_2 \leq (1+\alpha)\|\mathbf{h}_{\Lambda_*}\|_2. \tag{3.15}$$

Combining (3.13), (3.14) and (3.15), we can show that the inequality

$$\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2 \leq \frac{16\sqrt{2}\lambda^u_{2\kappa}(1+\alpha)\beta}{a(\lambda^l_{2\kappa} - \theta_{\kappa,2\kappa}\alpha)^2} \Big(2c + (\frac{2c\gamma}{\sqrt{\kappa}} + \sqrt{2}\sigma_1) \frac{2c\sqrt{\kappa} + t - c}{(c-\beta)(\sqrt{\kappa}-1)}\Big) \sqrt{\frac{\kappa \log m}{n}}$$

holds with probability at least $1 - 2m^{-4\kappa(\sigma^2-1)+1}$. By the definition of the constant C, we get the desired result. □

Remark 3.7. Suppose that the conditions in Theorem 3.6 also hold. Then, using (3.13) and the decrease of the magnitude of $\mathbf{h}_{\Lambda_i}$, we easily prove that

$$\left\{ i \in [m] \middle| \, |x_i^*| \geq C\sqrt{\frac{\kappa \log m}{n}} \right\} \subset supp(\hat{\mathbf{x}}).$$

This means if the nonzero coefficients are large enough in primal signal $\mathbf{x}^*$, then the method (1.7) can select them into the model.

4 Algorithm and Convergence Analysis

In this section, we will develop a sequential linear programming (SLP) algorithm to solve (1.7) based on the DC algorithm [23, 24, 30]. We reformulate the subproblem in the proposed algorithm as a linear program which is easily solvable. The convergence analysis of the proposed algorithm will be established. At the end of this section, we will provide a rule to choose the reweighted matrix W.

4.1 A sequential linear programming algorithm

The DC algorithm is a descent method to solve an optimization problem of the difference of convex functions by Tao and An [23, 24]. Consider the following optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^m} f(\mathbf{x}) := g(\mathbf{x}) - h(\mathbf{x}), \tag{4.1}$$

where g and h are lower semicontinuous proper convex functions on $\mathbb{R}^m$. The kth iteration scheme of the DC algorithm for solving (4.1) is given as follows:

$$\begin{cases} \mathbf{v}^k \in \partial h(\mathbf{x}^k), \\ \mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^m} g(\mathbf{x}) - h(\mathbf{x}^k) - \langle \mathbf{v}^k, \mathbf{x} - \mathbf{x}^k \rangle, \end{cases}$$

where $\mathbf{v}^k \in \partial h(\mathbf{x}^k)$ means that $\mathbf{v}^k$ is a subgradient of $h(\mathbf{x})$ at $\mathbf{x}^k$.

Clearly, the problem (1.7) can be written as the form of (4.1) by letting

$$g(\mathbf{x}) = \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda\|W\mathbf{x}\|_1, \quad h(\mathbf{x}) = \lambda\gamma\|W\mathbf{x}\|_2.$$

Note that the subdifferential of $\|W\mathbf{x}\|_2$ is

$$\partial\|W\mathbf{x}\|_2 = \begin{cases} \frac{W^2\mathbf{x}}{\|W\mathbf{x}\|_2}, & \text{if } \mathbf{x} \neq \mathbf{0}, \\ \{W\mathbf{u} | \; \|\mathbf{u}\|_2 \leq 1, \mathbf{u} \in \mathbb{R}^m\}, & \text{if } \mathbf{x} = \mathbf{0}. \end{cases}$$

Thus, using the DC algorithm to solve (1.7), the kth iteration scheme is given as

$$\mathbf{x}^{k+1} = \begin{cases} \arg\min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda\|W\mathbf{x}\|_1, & \text{if } \mathbf{x}^k = \mathbf{0} \\ \arg\min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda\Big(\|W\mathbf{x}\|_1 - \gamma\left\langle \mathbf{x}, \frac{W^2\mathbf{x}^k}{\|W\mathbf{x}^k\|_2} \right\rangle\Big), & \text{if } \mathbf{x}^k \neq \mathbf{0}. \end{cases}$$

It is equivalent to solve a convex optimization which is in the form of

$$\min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda(\|W\mathbf{x}\|_1 - \gamma\langle \mathbf{x}, \mathbf{v} \rangle), \tag{4.2}$$

where $\mathbf{v} \in \mathbb{R}^m$ is a constant vector. This problem can be reformulated as a linear program. Using notation of $|\mathbf{x}|$, $\mathbf{x}^+$ and $\mathbf{x}^-$, we have $\mathbf{x} = \mathbf{x}^+ - \mathbf{x}^-$ and $|\mathbf{x}| = \mathbf{x}^+ + \mathbf{x}^-$. Denote $\mathbf{s} = |\mathbf{y} - A\mathbf{x}| \in \mathbb{R}^n$. It is easy to see that

$$\|\mathbf{y} - A\mathbf{x}\|_1 = \mathbf{1}_n^T \mathbf{s}, \quad \|W\mathbf{x}\|_1 = \mathbf{1}_m^T W(\mathbf{x}^+ + \mathbf{x}^-).$$

Hence, the problem (4.2) can be reformulated as

$$\begin{array}{ll} \min & \mathbf{1}_n^T \mathbf{s} + \lambda\Big(\mathbf{1}_m^T W(\mathbf{x}^+ + \mathbf{x}^-) - \gamma \mathbf{v}^T(\mathbf{x}^+ - \mathbf{x}^-)\Big) \\ \text{s.t.} & \mathbf{s} + A(\mathbf{x}^+ - \mathbf{x}^-) \geq \mathbf{y}, \\ & A(\mathbf{x}^+ - \mathbf{x}^-) - \mathbf{s} \leq \mathbf{y}, \\ & \mathbf{s} \geq \mathbf{0}, \mathbf{x}^+ \geq \mathbf{0}, \mathbf{x}^- \geq \mathbf{0}. \end{array}$$

Let $\mathbf{c}, \mathbf{d} \in \mathbb{R}^{n+2m}$ be defined as

$$\mathbf{c} = \begin{pmatrix} \mathbf{1}_n \\ \lambda W \mathbf{1}_m - \lambda\gamma\mathbf{v} \\ \lambda W \mathbf{1}_m + \lambda\gamma\mathbf{v} \end{pmatrix}, \quad \mathbf{d} = \begin{pmatrix} \mathbf{s} \\ \mathbf{x}^+ \\ \mathbf{x}^- \end{pmatrix}.$$

Setting

$$B = \begin{pmatrix} -I & -A & A \\ -I & A & -A \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} -\mathbf{y} \\ \mathbf{y} \end{pmatrix},$$

where I is an $n \times n$ identity matrix, we can equivalently formulate the problem as the following linear program (LP):

$$\begin{array}{ll} \min & \mathbf{c}^T \mathbf{d} \\ \text{s.t.} & B\mathbf{d} \leq \mathbf{w}, \quad \mathbf{d} \geq \mathbf{0}. \end{array} \tag{4.3}$$

This is a standard linear optimization. Clearly, it is feasible and $\mathbf{c}^T\mathbf{d} \geq 0$ for any feasible solution $\mathbf{d}$. Hence, it has at least one optimal solution which can produce the current iterative point. It is well-known that there are many solution methods, such as the interior-point method and the simplex method, to solve the LP (4.3). Moreover, the complexity of the interior-point method for linear program is $O((n + 2m)^{3.5})$, which implies that such method is very efficient.

Consequently, the details of our algorithm for solving (1.7) is described as follows.

Algorithm 4.1. *A sequential linear programming algorithm*

Step 0. Define $\epsilon > 0$, set $k := 0, \mathbf{x}^0 = \mathbf{0}$ and choose $\mathbf{x}^{-1}$ such that $\|\mathbf{x}^0 - \mathbf{x}^{-1}\|_2 \geq \epsilon$.

Step 1. If $\|\mathbf{x}^k - \mathbf{x}^{k-1}\|_2 < \epsilon$, STOP. Output the optimal solution $\mathbf{x}^k$ of (1.7).

Step 2. If $\mathbf{x}^k = \mathbf{0}$, compute

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x} \in \mathbb{R}^m} \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda\|W\mathbf{x}\|_1. \tag{4.4}$$

Otherwise, compute

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x} \in \mathbb{R}^m} \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda\Big(\|W\mathbf{x}\|_1 - \gamma\left\langle \mathbf{x}, \frac{W^2\mathbf{x}^k}{\|W\mathbf{x}^k\|_2}\right\rangle\Big). \tag{4.5}$$

Step 3. Set $k := k + 1$ and go to Step 1.

Note that Algorithm 4.1 converts the primal nonconvex program (1.7) into a linear program (4.3), the cost is that it doesn't obtain a global optimal solution in general. The experiments in Section 5 will show that Algorithm 4.1 often produces a solution that is close to global minimizer with the initial point $\mathbf{x}^0 = \mathbf{0}$. When $\mathbf{x}^0 = \mathbf{0}$ and $W = I$, (4.4) reduces to solving the ℓ_1-LAD model (1.4) given in [26], so Algorithm 4.1 will start at the solution given by the ℓ_1-LAD method and then optimize the problem (1.7), which possibly explains that our method (1.7) will outperforms the ℓ_1-LAD method.

4.2 Convergence analysis

Assuming the subproblem (4.2) at each iteration of Algorithm 4.1 is solved accurately and the optimal solution can be achieved at some extreme point of the feasible region of (4.3), we can show that the generated sequence $\{\mathbf{x}^k\}$ converges to a stationary point of (1.7).

Definition 4.2. A point $\mathbf{x} \in \mathbb{R}^n$ is said to be a stationary point of (1.7) if it satisfies

$$\mathbf{0} \in \partial\|\mathbf{y} - A\mathbf{x}\|_1 + \lambda\left(\partial\|W\mathbf{x}\|_1 - \gamma\partial\|W\mathbf{x}\|_2\right). \tag{4.6}$$

By the first-order necessary optimality condition, $\hat{\mathbf{x}}$ satisfies (4.6). Hence, $\hat{\mathbf{x}}$ is a stationary point of (1.7). Clearly, an optimal solution of (1.7) must be a stationary point, but the converse is not true due to the nonconvexity of (1.7).

For simplicity, we define

$$F(\mathbf{x}) = \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda(\|W\mathbf{x}\|_1 - \gamma\|W\mathbf{x}\|_2). \tag{4.7}$$

The following lemma shows that the level set of $F(\mathbf{x})$ is bounded and the sequence $\{F(\mathbf{x}^k)\}$ is nonincreasing. Consequently, the sequence $\{\mathbf{x}^k\}$ generated by Algorithm 4.1 is bounded.

Lemma 4.3. *Let $F(\mathbf{x})$ be defined as (4.7) with $\lambda > 0$ and $0 \leq \gamma < 1$. Let $\{\mathbf{x}^k\}$ be the sequence generated by Algorithm 4.1. Then, the sequence $\{F(\mathbf{x}^k)\}$ is nonincreasing and the level set*

$$\mathcal{L}(\tilde{\mathbf{x}}) = \{\mathbf{x} \in \mathbb{R}^m |\ F(\mathbf{x}) \leq F(\tilde{\mathbf{x}})\}$$

is bounded for any given $\tilde{\mathbf{x}} \in \mathbb{R}^m$.

Proof. It is well known that $\|W\mathbf{x}\|_1 \geq \|W\mathbf{x}\|_2$ for any $\mathbf{x} \in \mathbb{R}^m$. Then for any $\mathbf{x} \in \mathcal{L}(\tilde{\mathbf{x}})$, we have

$$\lambda(1-\gamma)\|W\mathbf{x}\|_2 \leq F(\mathbf{x}) \leq F(\tilde{\mathbf{x}}).$$

Since $W = \mathrm{diag}(w_1, \ldots, w_m)$ is a positive diagonal matrix, we get $\tau = \min\{w_i :\ i \in [m]\} > 0$. This, together with $\lambda > 0$ and $0 \leq \gamma < 1$, yields

$$\|\mathbf{x}\|_2 \leq \frac{F(\tilde{\mathbf{x}})}{\tau\lambda(1-\gamma)},$$

which implies that the level set $\mathcal{L}(\tilde{\mathbf{x}})$ is bounded for any given $\tilde{\mathbf{x}} \in \mathbb{R}^m$.

We now prove that the sequence $\{F(\mathbf{x}^k)\}$ is nonincreasing. For any $k = 0, 1, 2, \ldots$, taking $\mathbf{v}^k \in \partial\|W\mathbf{x}^k\|_2$, it follows from (4.4) and (4.5) that

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda\Big(\|W\mathbf{x}\|_1 - \gamma\langle\mathbf{x}, \mathbf{v}^k\rangle\Big).$$

By the first-order optimality condition, there exist $\mathbf{u}^{k+1} \in \partial\|\mathbf{y} - A\mathbf{x}^{k+1}\|_1$ and $\mathbf{w}^{k+1} \in \partial\|W\mathbf{x}^{k+1}\|_1$ such that

$$\mathbf{u}^{k+1} + \lambda\mathbf{w}^{k+1} - \lambda\gamma\mathbf{v}^k = 0. \tag{4.8}$$

By the convexity, we have

$$\|\mathbf{y} - A\mathbf{x}^k\|_1 - \|\mathbf{y} - A\mathbf{x}^{k+1}\|_1 \geq \langle \mathbf{x}^k - \mathbf{x}^{k+1}, \mathbf{u}^{k+1}\rangle,$$

$$\|W\mathbf{x}^k\|_1 - \|W\mathbf{x}^{k+1}\|_1 \geq \langle \mathbf{x}^k - \mathbf{x}^{k+1}, \mathbf{w}^{k+1}\rangle,$$

and

$$\|W\mathbf{x}^k\|_2 - \|W\mathbf{x}^{k+1}\|_2 \leq \langle \mathbf{x}^k - \mathbf{x}^{k+1}, \mathbf{v}^k\rangle.$$

Consequently, we obtain from (4.8) that

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq \langle \mathbf{x}^k - \mathbf{x}^{k+1}, \mathbf{u}^{k+1} + \lambda\mathbf{w}^{k+1} - \lambda\gamma\mathbf{v}^k\rangle = 0.$$

This shows that the sequence $\{F(\mathbf{x}^k)\}$ is nonincreasing. □

It follows from Lemma 4.3 that the sequence $\{\mathbf{x}^k\}$ generated by Algorithm 4.1 is bounded due to $\{\mathbf{x}^k\} \subset \mathcal{L}(\mathbf{x}^0)$. The following theorem shows that it converges to a stationary point of (1.7).

Theorem 4.4. *Let $\{\mathbf{x}^k\}$ be the sequence generated by Algorithm 4.1. Then,*

- (i) *there exists a point $\hat{\mathbf{x}}$ and an integer N large enough such that $\mathbf{x}^k = \hat{\mathbf{x}}$ when $k > N$.*
- (ii) *$\hat{\mathbf{x}}$ satisfies the first-order optimality condition (4.6), i.e., it is a stationary point of (1.7)*

Proof. (i) Recalling the LP form (4.3) of the subproblem at the kth iteration, we get

$$\langle \mathbf{c}^k, \mathbf{d}^{k+1}\rangle = \|\mathbf{y} - A\mathbf{x}^{k+1}\|_1 + \lambda(\|W\mathbf{x}^{k+1}\|_1 - \gamma\langle \mathbf{x}^{k+1}, \mathbf{v}^k\rangle), \tag{4.9}$$

where $\mathbf{v}^k \in \partial\|W\mathbf{x}^k\|_2$. It is easy to see that $\|W\mathbf{x}^k\|_2 = \langle \mathbf{x}^k, \mathbf{v}^k\rangle$. Hence,

$$F(\mathbf{x}^k) = \|\mathbf{y} - A\mathbf{x}^k\|_1 + \lambda(\|W\mathbf{x}^k\|_1 - \gamma\langle \mathbf{x}^k, \mathbf{v}^k\rangle) = \langle \mathbf{c}^k, \mathbf{d}^k\rangle. \tag{4.10}$$

Combining (4.9) and (4.10), we obtain

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) = \langle \mathbf{c}^k, \mathbf{d}^k - \mathbf{d}^{k+1}\rangle + \lambda\gamma(\|W\mathbf{x}^{k+1}\|_2 - \langle \mathbf{x}^{k+1}, \mathbf{v}^k\rangle). \tag{4.11}$$

If $\mathbf{x}^1 = \mathbf{0}$, Algorithm 4.1 will stop at $k = 1$ and produce the solution $\hat{\mathbf{x}} = \mathbf{0}$. Otherwise, it follows from (4.11) that

$$F(\mathbf{0}) - F(\mathbf{x}^1) = \langle \mathbf{c}^0, \mathbf{d}^0 - \mathbf{d}^1\rangle + \lambda\gamma\|\mathbf{x}^1\|_2 \geq \lambda\gamma\|\mathbf{x}^1\|_2 > 0,$$

where the inequality holds since $\mathbf{d}^1$ is the optimal solution of (4.3) associated with the first iteration and $\mathbf{d}^0$ is its feasible solution. Hence, by Lemma 4.3, we have $F(\mathbf{0}) > F(\mathbf{x}^k)$ for all $k = 1, 2\ldots$, which derives $\mathbf{x}^k \neq \mathbf{0}$ for all $k = 1, 2\ldots$.

By Lemma 4.3, the sequence $\{F(\mathbf{x}^k)\}$ is nonincreasing. Since $F(\mathbf{x}^k) \geq 0$ for all $k = 1, 2\ldots$, the sequence $\{F(\mathbf{x}^k)\}$ is convergent. By Cauchy-Schwartz inequality,

$$\lambda\gamma(\|W\mathbf{x}^{k+1}\|_2 - \langle \mathbf{x}^{k+1}, \mathbf{v}^k\rangle) = \lambda\gamma\left(\|W\mathbf{x}^{k+1}\|_2 - \left\langle \mathbf{x}^{k+1}, \frac{W^2\mathbf{x}^k}{\|W\mathbf{x}^k\|_2}\right\rangle\right) \geq 0. \tag{4.12}$$

Note that $\mathbf{d}^{k+1}$ minimizes the LP (4.3) with $\mathbf{c} = \mathbf{c}^k$ and then we have

$$\langle \mathbf{c}^k, \mathbf{d}^k - \mathbf{d}^{k+1}\rangle \geq 0,$$

which, together with the convergence of $\{F(\mathbf{x}^k)\}$, (4.11) and (4.12), implies that

$$\lim_{k\to\infty}\langle \mathbf{c}^k, \mathbf{d}^k-\mathbf{d}^{k+1}\rangle = 0, \quad \lim_{k\to\infty}\|W\mathbf{x}^{k+1}\|_2-\langle \mathbf{x}^{k+1}, \mathbf{v}^k\rangle = 0.$$

Since $\{\mathbf{x}^k\}$ is bounded from Lemma (4.3), then as $k\to\infty$, we have

$$(x_i^{k+1}x_j^k - x_j^{k+1}x_i^k)^2 \to 0, \quad \forall i,j\in[m]. \tag{4.13}$$

Let E^0 be the extreme point set of the feasible region

$$\{\mathbf{d}\in\mathbb{R}^{n+2m}|\ B\mathbf{d}\le \mathbf{w}, \mathbf{d}\ge \mathbf{0}\}.$$

By the representation theorem of convex set, E^0 is finite. Let

$$E=\{W(\mathbf{x}^+-\mathbf{x}^-)|\ (\mathbf{s};\mathbf{x}^+;\mathbf{x}^-)\in E^0\}.$$

So E is finite and $\mathbf{0}\notin E$. Divide E into several parts $E_1, E_2, \ldots, E_\alpha$ such that each part is a subset of one-dimensional linear space. Define

$$\delta=\min\{\|W\mathbf{x}_1\|_2^2\|W\mathbf{x}_2\|_2^2-\langle W\mathbf{x}_1, W\mathbf{x}_2\rangle^2|\ W\mathbf{x}_1, W\mathbf{x}_2 \text{ are in different parts}\}.$$

Then, $\delta>0$ from the fact that E is finite and nonempty. Hence, it follows from (4.13) that for sufficiently large k,

$$\|W\mathbf{x}^{k+1}\|_2^2\|W\mathbf{x}^k\|_2^2-\langle W\mathbf{x}^{k+1}W\mathbf{x}^k\rangle^2<\delta.$$

Note that the sequence $\{W\mathbf{x}^k\}\subset E$. By the definition of δ, there exists k_0 such that the sequence $\{W\mathbf{x}^k\}_{k\ge k_0}$ is in a same subset of one-dimensional linear space. Without loss of generality, we assume that $\{W\mathbf{x}^k\}_{k\ge k_0}\subset E_1$. Since E_1 is a subset of one-dimensional linear space, there exists only one normalized base vector $W^{-1}\mathbf{v}^*$ such that $W\mathbf{x}^k=\theta_k W^{-1}\mathbf{v}^*$ for all $k\ge k_0$, where θ_k is a representation coefficient. Consequently, we obtain for all $k\ge k_0$,

$$\mathbf{v}^k=\frac{W^2\mathbf{x}^k}{\|W\mathbf{x}^k\|_2}=\mathbf{v}^*,$$

which implies that $\mathbf{c}^k=\mathbf{c}^*$ is a constant vector for all $k\ge k_0$. This shows that when k sufficiently large, the cost coefficient vector $\mathbf{c}^*$, the coefficient matrix B and the right-hand side $\mathbf{w}$ in (4.3) at the kth iteration are constant. Hence, the optimal solution sequence $\{\mathbf{x}^k\}$ will be a constant vector when k large enough. Therefore, there exists an integer N large enough such that $\mathbf{x}^k=\hat{\mathbf{x}}$ when $k>N$.

(ii) By (i), we assume that $\{\mathbf{x}^k\}$ converges to $\hat{\mathbf{x}}$ and hence $\hat{\mathbf{x}}$ is the optimal solution of the convex optimization as follows:

$$\min_{\mathbf{x}\in\mathbb{R}^m}\quad \|\mathbf{y}-A\mathbf{x}\|_1+\lambda\Big(\|W\mathbf{x}\|_1-\gamma\left\langle \mathbf{x}, \frac{W^2\hat{\mathbf{x}}}{\|W\hat{\mathbf{x}}\|_2}\right\rangle\Big),$$

which yields

$$\mathbf{0}\in\partial\|\mathbf{y}-A\hat{\mathbf{x}}\|_1+\lambda\left(\partial\|W\hat{\mathbf{x}}\|_1-\gamma\frac{W^2\hat{\mathbf{x}}}{\|W\hat{\mathbf{x}}\|_2}\right).$$

This shows that $\hat{\mathbf{x}}$ is a stationary point of (1.7). □

Remark 4.5. From the numerical experiments, we see that the iterative number of Algorithm 4.1 is 2 or 3 in general. So, the computational complexity of Algorithm 4.1 is about $O((n+2m)^{3.5})$ when the interior-point method is used to solve the subproblem (4.3).

In addition, if $\hat{\mathbf{x}}$ is an optimal solution of (1.7), the first-order necessary optimality condition shows

$$\mathbf{0} \in \partial\|\mathbf{y} - A\hat{\mathbf{x}}\|_1 + \lambda\left(\partial\|W\hat{\mathbf{x}}\|_1 - \gamma\frac{W^2\hat{\mathbf{x}}}{\|W\hat{\mathbf{x}}\|_2}\right),$$

which is a necessary and sufficient condition for

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda\left(\|W\mathbf{x}\|_1 - \gamma\left\langle \mathbf{x}, \frac{W^2\hat{\mathbf{x}}}{\|W\hat{\mathbf{x}}\|_2}\right\rangle\right).$$

Consequently, the corresponding $\hat{\mathbf{d}}$ with $\hat{\mathbf{x}}$ is an optimal solution of (4.3). By Theorem 4.4, Algorithm 4.1 can produce an optimal solution of (1.7) with high probability.

4.3 Construction of the reweighted matrix W and DCA-RPLAD

It follows from *Remark 3.7* that the significant elements in primal signal $\mathbf{x}^*$ could be selected. To strengthen this option, we know from [8] that the reweighted method wants to give the weights to approximately become inversely proportional to the true signal magnitude, and then make the model (1.7) chooses the optimal solution $\hat{\mathbf{x}}$ such that the reweighted penalization $\|W\hat{\mathbf{x}}\|_1 - \gamma\|W\hat{\mathbf{x}}\|_2$ could be regard as an approximation of $\|\hat{\mathbf{x}}\|_{\ell_0}$.

In this subsection, we detail the way, which is referenced from [8], to construct the reweighted matrix W and describe an iteration algorithm based on Algorithm 4.1. The associated parameters are taken as the choice rule given in [8]. Numerical experiments show that $l_{\max} = 2$ is a great selection for the maximum number of iterations.

Algorithm 4.6. *DCA-RPLAD*

Step 1. Let $l_{\max}$ be a specified maximum number of iterations. Set $l := 1$ and $W^{(l)} = I$.

Step 2. Solve the following problem using Algorithm 4.1

$$\mathbf{x}^{(l)} \in \arg\min_{\mathbf{x}\in\mathbb{R}^m} \|\mathbf{y} - A\mathbf{x}\|_1 + \lambda(\|W^{(l)}\mathbf{x}\|_1 - \gamma\|W^{(l)}\mathbf{x}\|_2).$$

Let $i_0 = n/[4\log(m/n)]$ and $|\mathbf{x}^{(l)}|_{(i_0)}$ be the i_0th largest magnitude in $|\mathbf{x}^{(l)}|$. Set

$$\varepsilon = \max\{|\mathbf{x}^{(l)}|_{(i_0)}, 10^{-3}\}.$$

Step 3. Update the weighted matrix $W^{(l+1)} = \mathrm{diag}(w_1^{(l+1)}, \ldots, w_m^{(l+1)})$:

$$w_i^{(l+1)} = \frac{1}{|x_i^{(l)}| + \varepsilon}, \quad i \in [m].$$

Step 4. Terminate on convergence or when $l = l_{\max}$. Otherwise, set $l := l+1$ and go to Step 2.

5 Numerical Experiments

In this section, we present numerical experiments to demonstrate the efficiency of the DCA-RPLAD method (see Algorithms 4.1 and 4.6). We will first investigate the effect of different selections of the penalty level λ and the parameter γ, and then compare it with the popular methods such as Lasso [2, 25] and ℓ_1-LAD [26] for the high-dimensional sparse linear regression (1.1) in both Gaussian noise case and Cauchy noise case, where the Lasso method solves (1.3) and the ℓ_1-LAD method solves (1.4). Throughout this section, we use $\epsilon = 10^{-7}$, $n = 100$ and $m = 400$. We set each row of the design matrix A is generated by $N(0, \Sigma)$ distribution with Toeplitz correlation matrix $\Sigma_{ij} = 0.5^{|i-j|}$, and then normalize the jth column $A_{.j}$ of A such that $\|A_{.j}\|_2 = \sqrt{n}$ for all $j \in [m]$. In addition, all codes were written by using Matlab R2017b and the numerical experiments were done on a laptop with 16 GB RAM and a 3.60-GHz Intel Core i7 CPU.

$\boxed{5.1}$ Selection of parameters

From theoretical analysis, the choice rule of λ is discussed in Proposition 2.2 and γ is given in (3.2), but these values are hard to calculate. Now we will investigate the effect of their different choices on the DCA-RPLAD method from computational viewpoint. We consider their following special values to investigate the numerical performance of the DCA-RPLAD method:

$$\lambda_1 = \sqrt{0.01 n \log p}, \quad \lambda_2 = \sqrt{0.1 n \log p}, \quad \lambda_3 = \sqrt{0.15 n \ logp}, \quad \lambda_4 = \sqrt{0.5 n \log p},$$

and

$$\lambda_5 = \sqrt{2 n \log p}, \quad \lambda_6 = \sqrt{4 n \log p}, \quad \gamma_1 = 0.25, \quad \gamma_2 = 0.5, \quad \gamma_3 = 0.75, \quad \gamma_4 = 1$$

In this subsection, we set $\kappa = 5$ and the primal signal $\mathbf{x}^*$ with $x^*_{2i-1}, 1 \leq i \leq 5$ being random numbers between 1 and 10 and other components zero. For each setting, we run the experiment 100 times in both Gaussian noise case and Cauchy noise case and average the estimation errors $\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2$. The obtained numerical results are summarized in Table 1.

Table 1: The average of estimation error $\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2$ over 100 simulations with different λ, γ and noise distributions

	Noise distributions	γ_1	γ_2	γ_3	γ_4
λ_1	$N(0,1)$	0.8497	0.8578	0.8652	0.8774
	Cauchy	28.4543	29.8448	30.5477	30.9958
λ_2	$N(0,1)$	0.3122	0.3160	0.3177	0.3208
	Cauchy	1.2773	1.3248	1.3622	1.4112
λ_3	$N(0,1)$	0.2963	0.2956	0.2971	0.2967
	Cauchy	0.6985	0.7118	0.7386	0.7489
λ_4	$N(0,1)$	0.3215	0.3138	0.3055	0.2988
	Cauchy	0.4951	0.4719	0.4620	0.4534
λ_5	$N(0,1)$	0.5339	0.4764	0.4347	0.3912
	Cauchy	1.43199	1.1496	0.9679	0.8958
λ_6	$N(0,1)$	6.4115	4.9274	4.3036	3.9292
	Cauchy	8.8558	7.9454	7.0776	6.9795

From Table 1 we can see that λ_1 is too small and λ_6 is too large in Cauchy noise case. And in Gaussian noise case, the change in estimating error with different λ choice is the same as Cauchy noise case. Obviously, the selections of parameter γ will not change the estimating error drastically. It can be seen that among the six penalty levels, λ_3 and λ_4 have relatively better results in the estimation error $\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2$. According to the numerical results, we usually choose $\lambda = \lambda_4$ and $\gamma = \gamma_2, \gamma_4$.

$\boxed{5.2}$ Numerical performance

In this subsection, we present numerical experiments to demonstrate the efficiency of the DCA-RPLAD method for the high-dimensional sparse linear regression (1.1) in both Gaussian noise case and Cauchy noise case. We compare it with the Lasso method [2, 25] and the ℓ_1-LAD method [26]. For the parameters of the DCA-RPLAD method, we choose the following setting as the default values:

$$\lambda = \sqrt{0.5n \log m}, \quad \gamma = 0.5 \text{ or } 1, \quad l_{\max} = 2.$$

In the DCA-RPLAD method, if $\gamma = 0.5$ is adopted, we denote it “RPLAD-half”; if $\gamma = 1$ is adopted, we denote it “RPLAD-1”.

For the ℓ_1-LAD method, we use two penalty levels $\lambda = \sqrt{0.5n \log m}$ and $\lambda_0 = \sqrt{2n \log m}$ which is selected in [26]. If $\lambda = \sqrt{0.5n \log m}$ is adopted, we denote the ℓ_1-LAD method “LADL1-new”; if $\lambda = \sqrt{2n \log m}$ is adopted, we denote “LADL1-old”. For the classical Lasso method, the standard deviation σ is generally known in the Gaussian distribution case, we choose the $\sigma \times \lambda_0$ as the penalty level λ, where the method is denoted as “Lasso”. In the noiseless case or Cauchy noise case, we use $0.01 \times \lambda_0$ or the cross validation to select the penalty level, where lasso method is denoted as “CV-Lasso”.

For each simulation, we set the sparsity levels $\kappa = 5, 7, 9, \ldots, 17$ and set $\mathbf{x}^*$ as $x^*_{2i-1}, 1 \leq i \leq \kappa$ being random numbers between 1 and 10 and other components zero. For each setting, we run the simulation 200 times in both Gaussian noise $N(0, 1)$ and Cauchy noise $C(1, 0)$. We summarizes the average estimation error and variable selection results for the different distributions. The numerical results about variable selections are reported in Table 2, where “AVERR1” means the average number of significant variables which are unselected in the 200 simulations, and “AVERR2” means the average number of insignificant variables which are selected in 200 simulations. The change of the average value of estimation errors $\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2$ is depicted in Figure 2.

The numerical results reported in Table 2 show that for both Gaussian noise and Cauchy noise, the numerical performance of the DCA-RPLAD method (Algorithm 4.6) in variable selections is better than the ℓ_1-LAD method and the Lasso method. It is worth mentioning that for Gaussian noise, the performance of the DCA-RPLAD method is best in terms of “AVERR2”, which shows that this method can usually identify insignificant variables. On the other hand, for Canchy noise, the DCA-RPLAD method outperforms the other method, its performance is best on both picking out significant variables and identifying insignificant variables. In addition, the Lasso method is not suitable for dealing with Canchy noise. But for Gaussian noise, it can completely pick out significant variables. For both Cauchy noise and Gaussian noise, the ℓ_1-LAD method with the new penalty level outperforms the old one in picking out significant variables, but the converse in identifying insignificant variables.

It is to see from Figure 2 that the DCA-RPLAD method outperforms than the Lasso method [2, 25] and the ℓ_1-LAD method [26]. Especially, the reweighted new penalized LAD model (1.7) is very suitable for the more sparsity data with Cauchy noise.

Table 2: The numerical performance in variable selections of the DCA-RPLAD method, the ℓ_1-LAD method and the Lasso method for both Gaussian noise and Cauchy noise

Method	Noise	Error	$\kappa = 5$	$\kappa = 7$	$\kappa = 9$	$\kappa = 11$	$\kappa = 13$	$\kappa = 15$	$\kappa = 17$
RPLAD-half	$N(0,1)$	AVERR1	0	0	0	0	0	0.020	0.005
		AVERR2	0	0	0	0.005	0.005	0.030	0.515
	Cauchy	AVERR1	0.025	0.050	0.055	0.180	0.270	0.280	0.435
		AVERR2	0.050	0.070	0.225	0.420	0.930	2.180	8.625
RPLAD-1	$N(0,1)$	AVERR1	0	0	0	0.005	0.005	0	0.010
		AVERR2	0	0	0	0	0	0.005	0.090
	Cauchy	AVERR1	0.035	0.060	0.035	0.090	0.225	0.355	0.515
		AVERR2	0.025	0.045	0.070	0.205	0.575	1.525	5.390
LADL1-old	$N(0,1)$	AVERR1	0.010	0.685	3.185	6.195	9.045	11.61	14.02
		AVERR2	0.735	1.060	1.060	1.005	0.975	0.970	0.805
	Cauchy	AVERR1	0.360	1.760	4.355	6.965	9.450	11.90	14.27
		AVERR2	0.830	0.975	0.930	0.930	0.750	0.835	0.870
LADL1-new	$N(0,1)$	AVERR1	0	0	0	0	0	0.020	0.105
		AVERR2	20.76	20.50	20.85	20.54	20.17	19.68	19.70
	Cauchy	AVERR1	0	0.015	0.035	0.120	0.335	0.470	1.185
		AVERR2	20.62	20.78	20.51	20.23	19.78	20.18	19.85
Lasso	$N(0,1)$	AVERR1	0	0	0	0	0	0	0
		AVERR2	1.755	2.920	3.955	5.785	7.375	9.400	11.76
	Cauchy	AVERR1	0.990	1.410	1.960	2.410	2.760	3.685	3.810
		AVERR2	104.3	102.7	101.4	99.53	98.19	96.85	95.55
CV-Lasso	$N(0,1)$	AVERR1	0	0	0	0	0	0	0
		AVERR2	15.54	22.57	24.32	26.25	27.17	26.57	26.05
	Cauchy	AVERR1	1.545	2.140	2.970	3.370	3.795	5.145	5.230
		AVERR2	12.44	15.10	18.73	23.40	26.17	27.26	31.70

$\boxed{6}$ Conclusions

We have studied the high-dimensional sparse linear regression (1.1). For the purpose of proposing a new method suitable for data in the presence of outliers and with more sparsity structure, we adopt the advantages of the LAD method and the nonconvex regularization, and then present a new penalized LAD method based on the difference of ℓ_1 and ℓ_2 norms. To enhance the sparsity, we use the reweighted strategy given in [8] and propose the reweighted penalized LAD method (1.7). From theoretical analysis, we give the choice rule of the penalty level λ and the parameter γ. Especially, we establish an estimation error bound (1.8) for the proposed model (1.7). From computational analysis, we propose an iterative method (Algorithm 4.6) based on a DC-type algorithm (Algorithm 4.1) to solve (1.7). The largest advantage of Algorithm 4.1 is to solve only one linear program at each iteration. We prove that the sequence generated by Algorithm 4.1 converges to a stationary point of (1.7). Numerical results illustrate that our method outperforms some popular methods such as the Lasso method [2, 25] and the ℓ_1-LAD method [26] for solving (1.1).

Acknowledgements

The authors would like to thank the editor and the anonymous referees for their helpful constructive comments and suggestions which lead to a significantly improved version of the paper.

References

[1] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.* 2 (2009) 183–202.

[2] P.J. Bickel, Y. Ritov and A.B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, *Ann. Statist.* 37 (2009) 1705–1732.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, *Found. Trends Mach. Learn.* 3 (2011) 1–122.

[4] R.H. Byrd, G.M. Chin, J. Nocedal and F. Oztoprak, A family of second-order methods for convex l_1-regularized optimization, *Math. Program.* 159 (2016) 435–467.

[5] T. Cai, L. Wang and G. Xu, New bounds for restricted isometry constants, *IEEE Trans. Inform. Theory* 56 (2010) 4388–4394.

[6] E. Candès, J. Romberg and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Comm. Pure Appl. Math.* 59 (2006) 1207–1223.

[7] E. Candès and T. Tao, The Dantzig selector: statistical estimation when p is much larger than n, *Ann. Statist.* 35 (2007) 2313–2351.

[8] E. Candès, M. Wakin and S. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, *J. Fourier Anal. Appl.* 14 (2008) 877–905.

[9] X. Chen, F. Xu and Y. Ye, Lower bound theory of nonzero entries in solutions of $\ell_2 - \ell_p$ minimization, *SIAM J. Sci. Comput.* 32 (2010) 2832–2852.

[10] A. Cohen, W. Dahmen and R. Devore, Compressed sensing and best k-term approximation, *J. Amer. Math. Soc.* 22 (2009) 221–231.

[11] I. Daubechies, R. DeVore, M. Fornasier and C. Güntük, Iteratively reweighted least squares minimization for sparse recovery, *Commun. Pure Appl. Math.* 63 (2010) 1–38.

[12] E. Esser, Y. Lou and J. Xin, A method for finding structured sparse solutions to nonnegative least squares problems with applications, *SIAM J. Imaging Sci.* 6 (2013) 2010–2046.

[13] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.* 96 (2001) 1348–1360.

[14] S. Foucartand M. Lai, Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$, *Appl. Comput. Harmon. Anal.* 26 (2009) 395–407.

[15] X. Gao and J. Huang, Asymptotic analysis of high-dimensional LAD regression with Lasso, *Statist. Sinica* 20 (2010) 1485–1506.

[16] T. Goldstein and S. Osher, The split Bregman method for ℓ_1-regularized problems, *SIAM J. Imaging Sci.* 2 (2009) 323–343.

[17] E. Hale, W. Yin and Y. Zhang, Fixed-point continuation for ℓ_1-minimization: Methodology and convergence, *SIAM J. Optim.* 19 (2008) 1107–1130.

[18] M.-J. Lai, Y. Xu and W. Yin, Improved iteratively reweighted least squares for unconstrained smoothed ℓ_q minimization, *SIAM J. Numer. Anal.* 51 (2013) 927–957.

[19] X. Li, D. Sun and K.-C. Toh, A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems, *SIAM J. Optim.* 28 (2018) 433–458.

[20] Y. Lou, P. Yin, Q. He and J. Xin, Computing sparse representation in a highly coherent dictionary based on difference of L_1 and L_2, *J. Sci. Comput.* 64 (2015) 178–196.

[21] N. Meinshausen and B. Yu, Lasso-type recovery of sparse representations for high-dimensional data, *Ann. Statist.* 37 (2009) 246–270.

[22] H. Rauhut, Compresssive sensing and structred random matrices, *Radon Ser. Comput. Appl. Math.* 9 (2010) 1–92.

[23] P.D. Tao and L.T.H. An, Convex analysis approach to dc programming: Theory, algorithms and applications, *Acta Math. Vietnam.* 22 (1997) 289–355.

[24] P.D. Tao and L.T.H. An, A D.C. optimization algorithm for solving the trust-region subproblem, *SIAM J. Optim.* 8 (1988) 476–505.

[25] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. Roy. Statist. Soc. Ser. B.* 58 (1996) 267–288.

[26] L. Wang, L_1 penalized LAD estimator for high dimensional linear regression, *J. Multi. Anal.* 120 (2013) 135–151.

[27] H. Wang, G. Li and G. Jiang, Robust regression shrinkage and consistent variable selection via the LAD-Lasso, *J. Bus. Econom. Statist.* 25 (2007) 347–355.

[28] Z. Wen, W. Yin, D. Goldfarb and Y. Zhang, A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation, *SIAM J. Sci. Comput.* 32 (2010) 1832–1857.

[29] J. Yang and Y. Zhang, Alternating direction algorithms for L_1-problems in compressive sensing, *SIAM J. Sci. Comput.* 33 (2011) 250–278.

[30] P. Yin, Y. Lou, Q. He and J. Xin, Minimization of ℓ_{1-2} for compressed sensing, *SIAM J. Sci. Comput.* 37 (2015) A536–A563.

[31] P. Yin, E. Esser and J. Xin, Ratio and difference of L_1 and L_2 norms and sparse representation with coherent dictionaries, *Commun. Inform. Systems* 14 (2014) 87–109.

[32] W. Yin, S. Osher, D. Goldfarb and J. Darbon, Bregman iterative algorithms for l_1 minimization with applications to compressed sensing, *SIAM J. Imaging Sci.* 1 (2008) 143–168.

[33] S. Yun and K.-C. Toh, A coordinate gradient descent method for ℓ_1-regularized convex minimization, *Comput. Optim. Appl.* 48 (2011) 273–307.

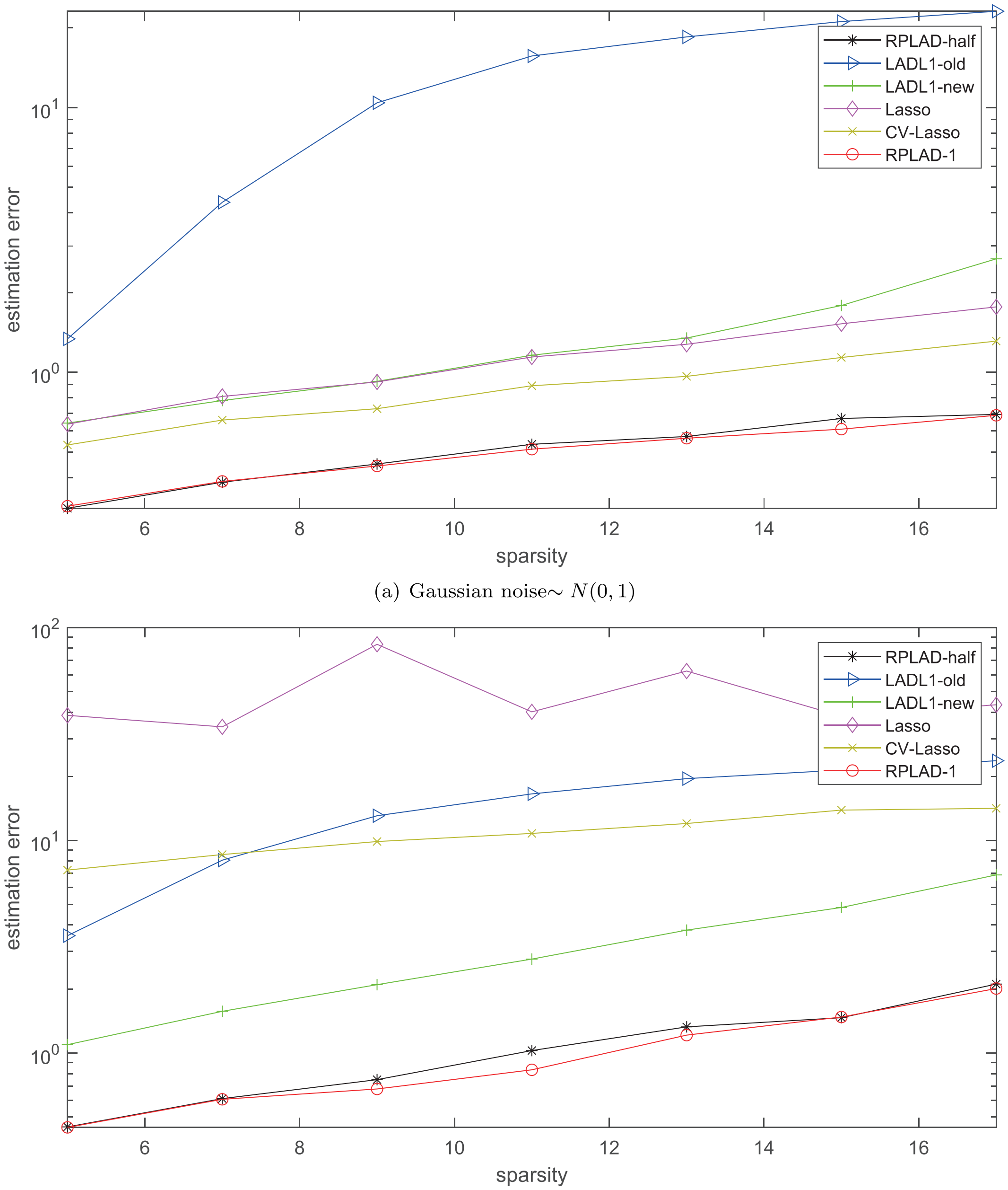
[34] J. Zeng, S. Lin, Y. Wang and Z. Xu, $L_{1/2}$ regularization: Convergence of iterative half thresholding algorithm, *IEEE Trans. Image Process.* 62 (2014) 2317–2329.

[35] Y. Zhao and D. Li, Reweighted ℓ_1-minimization for sparse solutions to underdetermined linear systems, *SIAM J. Optim.* 22 (2012) 1065–1088.

Manuscript received 4 February 2021
revised 8 May 2021
accepted for publication 9 June 2021

Zhenzhi Qin
Department of Mathematical Sciences
Tsinghua University, Beijing,100084, China
E-mail address: qzz19@mails.tsinghua.edu.cn

Liping Zhang
Department of Mathematical Sciences
Tsinghua University, Beijing, 100084, China
E-mail address: lipingzhang@tsinghua.edu.cn



(a) Gaussian noise$\sim N(0,1)$

(b) Cauchy noise$\sim C(0,1)$

Figure 2: Estimation errors provided by the six methods for (1.1) with Gaussian noise or Cauchy noise.