FUSED LASSO PENALIZED LEAST ABSOLUTE DEVIATION ESTIMATOR FOR HIGH DIMENSIONAL LINEAR REGRESSION

YANQING LIU
Community Health Center
Beijing Jiaotong University, Beijing 100044, China

JIYUAN TAO
Department of Mathematics and Statistics
Loyola University Maryland
Baltimore, MD 21210, USA

HUAN ZHANG, XIANCHAO XIU AND LINGCHEN KONG
Department of Applied Mathematics
Beijing Jiaotong University, Beijing 100044, China

Abstract. The least absolute shrinkage and selection operator (LASSO) has been playing an important role in variable selection and dimensionality reduction for high dimensional linear regression under the zero-mean or Gaussian assumptions of the noises. However, these assumptions may not hold in practice. In this case, the least absolute deviation is a popular and useful method. In this paper, we focus on the least absolute deviation via Fused LASSO, called Robust Fused LASSO, under the assumption that the unknown vector is sparsity for both the coefficients and its successive differences. Robust Fused LASSO estimator does not need any knowledge of standard deviation of the noises or any moment assumptions of the noises. We show that the Robust Fused LASSO estimator possesses near oracle performance, i.e. with large probability, the \( \ell_2 \) norm of the estimation error is of order \( O(\sqrt{k \log p/n}) \). The result is true for a wide range of noise distributions, even for the Cauchy distribution. In addition, we apply the linearized alternating direction method of multipliers to find the Robust Fused LASSO estimator, which possesses the global convergence. Numerical results are reported to demonstrate the efficiency of our proposed method.

2010 Mathematics Subject Classification. Primary: 65K05, 90C90; Secondary: 97K80.

Key words and phrases. Fused LASSO, least absolute deviation, high dimensional linear regression, linearized alternating direction method of multipliers.

The work was supported in part by National Natural Science Foundation of China (11671029), and the Fundamental Research Funds for the Central Universities (2016JBM081)

This paper was prepared at the occasion of The 10th International Conference on Optimization: Techniques and Applications (ICOTA 2016), Ulaanbaatar, Mongolia, July 23-26, 2016, with its Associate Editors of Numerical Algebra, Control and Optimization (NACO) being Prof. Dr. Zhiyou Wu, School of Mathematical Sciences, Chongqing Normal University, Chongqing, China, Prof. Dr. Changjun Yu, Department of Mathematics and Statistics, Curtin University, Perth, Australia, and Shanghai University, China, and Prof. Gerhard-Wilhelm Weber, Middle East Technical University, Ankara, Turkey.

Corresponding author: Lingchen Kong.

97
1. **Introduction.** High dimensional data are frequently collected in a large variety of research areas such as information technology, genomic, functional magnetic resonance imaging, medical imaging and diagnosis, and finance [5, 15, 18]. Analysis of high dimensional data poses many challenges for statisticians, which calls for new methods and theories in the high dimensional linear regression model where the number \( n \) of observations is much less than the number \( p \) of unknown coefficients. For a high dimensional linear regression problem, a key assumption is the sparsity of the true coefficient parameters. That is, the number \( k \) of nonzero elements in the true regression coefficients is less than \( p \). This is the so-called the high dimensional sparse linear model. The ordinary least squares method is not consistent in this setting since it is ill-posed and may not lead to sparse solution in this context. Many other methods have been proposed to solve this problem. In particular, methods based on \( \ell_1 \) penalized least squares or constrained \( \ell_1 \) minimization have been extensively studied. A remarkable model is the least absolute shrinkage and selection operator (LASSO) model proposed originally by Tibshirani [20]. The LASSO model has inspired a number of impressive articles [8, 11, 23, 31, 32], where they demonstrated the fundamental result that \( \ell_1 \) penalized least squares estimators achieve the rate \( O(\sqrt{k \log p/n}) \), which is very close to the oracle rate \( O(\sqrt{k/n}) \). However, the LASSO ignores the ordering of the features. To overcome this problem, an interesting general LASSO model, the Fused LASSO (FLASSO) model, is introduced by Tibshirani et al. [21], which focuses on the sparsity for both the coefficients and their successive differences. Some asymptotic properties of the FLASSO model have been established in Tibshirani et al. [21] and Rinaldo [19]. They also suggest to reformulate it as a second-order cone programming or quadratic programming, and then apply standard convex optimization tools to solve it, see, e.g., [9, 21, 22]. MARS (Multi Adaptive Regression Splines) is also a popular method, see, e.g., [7, 12, 17, 28]. Due to extremely expensive computation, this approach is generally not applicable for large scale data set. Li et al. [13] proposed the linearized alternating direction method of multipliers (LADMM) approach to solve the FLASSO model, and numerical experiments demonstrated its efficiency.

Note that the LASSO and FLASSO methods have nice properties under the zero-mean and a known variance, or Gaussian assumptions. However, these assumptions may not hold in practice and the estimation of the standard deviation is not easy. Thus, the least absolute deviation (LAD) method was proposed when data sets are subject to heavy-tailed errors or outliers which are commonly encountered in applications. See, e.g., [24, 25, 26]. Based on the LAD model, Wang et al. [24] combined the usual LAD criterion and the LASSO-type penalty to produce the LAD-LASSO method. Compared with the LAD regression, LAD-LASSO can do parameter estimation and variable selection simultaneously. In contrast to LASSO, LAD-LASSO is consistent to heavy-tailed errors or outliers in the response. Furthermore, Wang [25] showed that the LAD-LASSO method achieves near oracle performance under some mild conditions, which implies that the LAD-LASSO estimator enjoys the same asymptotic efficiency as the oracle estimator in high-dimensional setting. However, it cannot deal with the sparsity for both the coefficients and their successive differences. Motivated by the above analysis, in order to consider the above complex sparsity of the successive differences under the non-Gaussian assumption, an interesting question naturally occurs: can we propose a new model by combining the usual LAD criterion and the FLASSO type penalty?
This paper will focus on this issue and give an affirmative answer. That is, we develop the least absolute deviation via Fused LASSO, called Robust Fused LASSO (RFLASSO), which replaces the least square by the least absolute deviation in the FLASSO model. We consider the choice of the penalty level for the RFLASSO, which does not depend on any unknown parameters or the noise distribution. Our analysis shows that with high probability, the RFLASSO estimator has near oracle performance as the LAD-LASSO estimator. Similarly, we do not have any assumptions on the moments of the noise and we only need a scale parameter to control the tail probability of the noise. The result is true for a wide range of noise distributions, even for Cauchy distributed noise, where the first order moment does not exist. Moreover, we introduce LADMM to find the estimator of the RFLASSO model, and show that the global convergence result is maintained for the proposed method. By numerical experiments, it turns out that the simple LADMM is quite efficient to solve the RFLASSO model.

The rest of this paper is organised as follows. We introduce the RFLASSO model for high dimensional linear model and discuss its statistical properties in Section 2. In Section 3 we give the optimization method for the RFLASSO model by applying the LADMM and prove its convergence. Section 4 reports numerical results to demonstrate the efficiency of the proposed method. Finally, we make some conclusions in Section 5.

2. Statistical Properties of the RFLASSO Method. We first introduce the RFLASSO model for high dimensional linear regression, and then we discuss the choice of the penalty level and near oracle property of the RFLASSO estimator.

We begin with the following high dimensional linear regression model

\[ y = X\beta + \varepsilon, \]

where \( y = (y_1, y_2, \ldots, y_n)^T \in \mathbb{R}^n \) is a response vector, \( X = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^{n \times p} \) is a fixed design matrix, \( \beta = (\beta_1, \beta_2, \ldots, \beta_p)^T \in \mathbb{R}^p \) is a regression coefficient vector, and \( \varepsilon \in \mathbb{R}^n \) is an error vector. Throughout the paper, we assume the median of \( \varepsilon_i \) is 0. Considering the ordering of the features, the FLASSO model was introduced by Tibshirani et al. [21] under the Gaussian assumption, which focuses on the sparsity for both the coefficients and their successive differences. The FLASSO model is

\[
\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \| y - X\beta \|_2^2 + \lambda_1 \| \beta \|_1 + \lambda_2 \| A\beta \|_1,
\]

where

\[
A = \begin{bmatrix}
-1 & 1 & 0 & \cdots & 0 \\
0 & -1 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1 & 0 \\
0 & 0 & \cdots & -1 & 1
\end{bmatrix}
\]

However, for large scale data in the high dimensional setting, the Gaussian assumption may not hold in practice and the estimation of the standard deviation is not easy. LAD method was proposed when data sets subject to heavy-tailed errors or outliers are commonly encountered in applications. Motivated by the above work, in order to deal with the big data with sparsity for both the coefficients and
their successive differences, we combine the LAD and FLASSO penalty. We define the Robust Fused LASSO as
\[
\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_1 + \lambda_1 \|\beta\|_1 + \lambda_2 \|A\beta\|_1.
\] (3)

Correspondingly, we define the FLASSO penalized LAD regression estimator (the RFLASSO estimator):
\[
\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \{ \beta : \|y - X\beta\|_1 + \lambda_1 \|\beta\|_1 + \lambda_2 \|A\beta\|_1 \}.
\]

We will consider the statistical property of the RFLASSO estimator under some regularization condition, which states the upper bound for estimation error \( h = \hat{\beta} - \beta^* \) under \( \ell_2 \)-norm, where \( \beta^* \) is the true value of model (3). In what follows, for any set \( E \subset \{1, 2, \ldots, p\} \) and vector \( h \in \mathbb{R}^p \), let \( h_E = hI(E) \) denote the \( p \) dimensional vector such that we only keep the coordinates of \( h \) when their indices are in \( E \) and replace others by 0. We assume \( T = \text{supp}(\beta^*) \) with the cardinality of \( T \), \( |T| = k < n \). The set \( T \) of nonzero coefficients or significant variables of \( \beta^* \) is unknown.

2.1. Choice of Penalty. We discuss the choice of the penalties of \( \lambda_1 \) and \( \lambda_2 \) for the RFLASSO estimator \( \hat{\beta} \), which dominates the estimation error with large probability. This penalty level does not depend on any unknown parameters.

The principle of selecting the penalties \( \lambda_1 \) and \( \lambda_2 \) is motivated by [25], which is closely related to the first-order optimality condition of the RFLASSO estimator. In fact, the subdifferential of the RFLASSO evaluated at the point of true coefficient \( \beta^* \) measures the estimation error in the linear regression model. The subdifferential of \( \|Y - X\beta\|_1 \) at point \( \beta = \hat{\beta} \) can be written as \( S = X^T \text{sign}(\varepsilon) \), where \( \text{sign}(\varepsilon) \) denotes the sign of \( \varepsilon \), i.e. \( \text{sign}(\varepsilon_i) = 1 \) if \( \varepsilon_i > 0 \), \( \text{sign}(\varepsilon_i) = -1 \) if \( \varepsilon_i < 0 \) and \( \text{sign}(\varepsilon_i) = 0 \) if \( \varepsilon_i = 0 \). Since \( \varepsilon_i \) are independent and have median 0, we have \( P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 0.5 \). And the subdifferential of \( \lambda_1 \|\beta\|_1 \) and \( \lambda_2 \|A\beta\|_1 \) at point \( \beta = \hat{\beta} \) are \( \lambda_1 \text{sign}(\hat{\beta}) \) and \( \lambda_2 A^T \text{sign}(A\hat{\beta}) \), respectively. So from the first-order optimality condition, we know
\[
0 \in S + \lambda_1 \text{sign}(\hat{\beta}) + \lambda_2 A^T \text{sign}(A\hat{\beta}).
\]
Since \( \lambda_1 \|\text{sign}(\hat{\beta})\|_\infty \leq \lambda_1 \) and \( \lambda_2 A^T \|\text{sign}(A\hat{\beta})\|_\infty \leq 2\lambda_2 \), we choose the penalty level of \( \lambda_1 \) and \( \lambda_2 \) satisfying
\[
P(\lambda_1 + 2\lambda_2 \geq c\|S\|_\infty) \geq 1 - \alpha,
\]
for a given constant \( c > 1 \) and a given \( \alpha > 0 \). When the distribution of \( \text{sign}(\varepsilon) \) is known, the distribution of \( \|S\|_\infty \) is known for any given \( X \) and does not depend on any unknown parameters. In the next subsection we will give further analysis on this penalty level, which shows its connection with the error bound of the RFLASSO estimator.

2.2. Near Oracle Property. We will show that the RFLASSO estimator possesses near oracle performance, which states that with high probability the \( \ell_2 \) norm of the estimation error is of order \( O(\sqrt{k\log p/n}) \). To do so, we need present a useful lemma.

Lemma 2.1. Suppose that \( \lambda_1 = u\|S\|_\infty, 2\lambda_2 = v\|S\|_\infty, u > 1, v > 0, u - v > 1, \) and \( S = X^T \text{sign}(\varepsilon) \) is the subdifferential of \( \|Y - X\beta\|_1 \). Let
\[
\Delta_t = \{ \delta \in \mathbb{R}^p : \|\delta\|_1 \geq t\|\delta_{T^c}\|_1 \},
\]
where $T \subset \{1, 2, \ldots, p\}$ and $T$ contains at most $k$ elements, where $t = \frac{u-v-1}{u+v+1}$, $u > 1$, $v > 0$, $u-v > 1$. Then $h \in \Delta_t$.

**Proof.** Due to $\hat{\beta} \in \arg \min \{\beta : \|Y - X\beta\|_1 + \lambda_1\|\beta\|_1 + 2\lambda_2\|A\beta\|_1\}$, we have that

$$\|Y - X\hat{\beta}\|_1 + \lambda_1\|\hat{\beta}\|_1 + 2\lambda_2\|A\hat{\beta}\|_1 \leq \|Y - X\beta^*\|_1 + \lambda_1\|\beta^*\|_1 + 2\lambda_2\|A\beta^*\|_1.$$ 

This, together with the fact $Y = X\beta^* + \varepsilon$ and $h = \hat{\beta} - \beta^*$, means

$$\|Xh + \varepsilon\|_1 + \lambda_1\|\hat{\beta}\|_1 + \lambda_2\|A\hat{\beta}\|_1 \leq \|\varepsilon\|_1 + \lambda_1\|\beta^*\|_1 + \lambda_2\|A\beta^*\|_1.$$ 

Then we obtain

$$\|Xh + \varepsilon\|_1 - \|\varepsilon\|_1 \leq \lambda_1(\|\beta^*\|_1 - \|\hat{\beta}\|_1) + \lambda_2(\|A\beta^*\|_1 - \|A\hat{\beta}\|_1)$$

$$= \lambda_1(\|\beta^*\|_1 - \|\hat{\beta}\|_1) + \lambda_2 \sum_{j=2}^{p} (\|\beta^*_j - \hat{\beta}^*_j\|_1 - \|\hat{\beta}^*_j - \hat{\beta}^*_{j-1}\|_1).$$

Due to $\beta^*_2 = 0$, we have that

$$\|\beta^*\|_1 - \|\hat{\beta}\|_1 = \|\beta^*_1\|_1 - \|\hat{\beta}_1\|_1 - \|\beta^*_2\|_1 - \|\hat{\beta}_2\|_1$$

$$\leq \|\beta^*_1 - \hat{\beta}_1\|_1 + \|\beta^*_2 - \hat{\beta}_2\|_1$$

$$= \|h_T\|_1 - \|h_{T^c}\|_1,$$

and also

$$\|A\beta^*\|_1 - \|A\hat{\beta}\|_1 \leq \|A\beta^* - A\hat{\beta}\|_1 \leq 2\|\beta^* - \hat{\beta}\|_1 = 2\|h_T\|_1 + 2\|h_{T^c}\|_1.$$ 

Thus we have

$$\|Xh + \varepsilon\|_1 - \|\varepsilon\|_1 \leq \lambda_1(\|h_T\|_1 - \|h_{T^c}\|_1) + 2\lambda_2(\|h_T\|_1 + \|h_{T^c}\|_1)$$

$$= (\lambda_1 + 2\lambda_2)\|h_T\|_1 - (\lambda_1 - 2\lambda_2)\|h_{T^c}\|_1.$$ 

(4)

Since the subdifferential of $\|Y - X\beta\|_1$ is $S = X^T\text{sign}(\varepsilon)$, it is obvious that

$$\|Xh + \varepsilon\|_1 - \|\varepsilon\|_1 \geq \langle X^T\text{sign}(\varepsilon), h \rangle \geq -\|h_T\|_1\|X^T\text{sign}(\varepsilon)\|_{\infty}$$

$$= -\|h_T\|_1\|S\|_{\infty} = -\|S\|_{\infty}(\|h_T\|_1 + \|h_{T^c}\|_1).$$ 

(5)

It follows from (4) and (5) that for any $\lambda_1 > 0$, $\lambda_2 > 0,$

$$(\lambda_1 + 2\lambda_2)\|h_T\|_1 - (\lambda_1 - 2\lambda_2)\|h_{T^c}\|_1 \geq -\|S\|_{\infty}(\|h_T\|_1 + \|h_{T^c}\|_1).$$ 

(6)

Thus under the condition $\lambda_1 = u\|S\|_{\infty}$, $2\lambda_2 = v\|S\|_{\infty}$, $u > 1$, $v > 0$, $u-v > 1$, we have

$$\|h_T\|_1 \geq \frac{u - v - 1}{u + v + 1}\|h_{T^c}\|_1.$$ 

(7)

Therefore we prove the assertion. \(\square\)

In order to study the near oracle property of the RFLASSO estimator, we also need to introduce some restricted eigenvalue concepts on the design matrix $X$. First, we define some important quantities of design matrix $X$ based on $\ell_2$ norm. Set

$$\lambda_k^u = \max_{0 < \|d\|_0 \leq k} \frac{\|Xd\|_2^2}{\|d\|_2^2},$$

where $\|d\|_0 \leq k$ means $d \in \mathbb{R}^p$ is a $k$ sparse vector with at most $k$ nonzero coordinates. Similarly, set

$$\lambda_k^l = \min_{0 < \|d\|_0 \leq k} \frac{\|Xd\|_2^2}{\|d\|_2^2}, \quad \theta_{k_1} = \max_{0 < \|d_1\|_0 \leq k_1} \frac{\|\langle Xd_1, Xd_2 \rangle\|_2}{\|d_1\|_2 \|d_2\|_2},$$

$$\theta_{k_2} = \max_{0 < \|d_i\|_0 \leq k_2, i = 1, 2} \frac{\|\langle Xd_1, Xd_2 \rangle\|_2}{\|d_1\|_2 \|d_2\|_2}.$$
Then, we obtain that

\begin{equation}
\text{Proof. Without loss of generality, assume where } \eta \\text{ we in addition define the restricted eigenvalue of design matrix } X \text{ based on } \ell_1 \text{ norm as in [25], using the following formula}
\end{equation}

\begin{equation}
K_k^l(t) = \min_{h \in \Delta, n} \|Xh\|_1 / \|h\|_2.
\end{equation}

We write \( K_k^l(t) \) as \( K_k^l \) for abbreviation when it does not cause any confusion.

To bound the estimation error of the RFLASSO model, we introduce the scale assumption on the errors \( \varepsilon_i \). Suppose there exists a constant \( a > 0 \) such that

\begin{equation}
P(\varepsilon_i \geq x) \leq \frac{1}{2 + ax} \text{ for all } x \geq 0,
\end{equation}

\begin{equation}
P(\varepsilon_i \leq x) \leq \frac{1}{2 + a|x|} \text{ for all } x < 0.
\end{equation}

Here \( a > 0 \) is a scale parameter of the distribution of \( \varepsilon_i \). This is a weak condition. For instance, Cauchy distribution satisfies the above assumption, see [25] for more details. That is a advantage over the traditional method, which significantly relies on the Gaussian assumption and the variance of the errors.

We are ready to establish our main result.

**Theorem 2.2.** Consider model (1), assume \( \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \) are independent and identically distributed random variables satisfying (8) and \( \lambda_1 = u\|S\|_\infty, 2\lambda_2 = v\|S\|_\infty, u - v > 1, u > 1, v > 0 \). Suppose \( \lambda_k > \theta_k^l \left( \frac{1}{7} + \frac{1}{4} \right) \) and \( \frac{3\sqrt{n}}{16} K_k^l > (\lambda_1 + 2\lambda_2) \sqrt{k/n} + C_1 \sqrt{2k \log p} \left( \frac{5}{4} + \frac{1}{t} \right) \) hold, then the RFLASSO estimator \( \hat{\beta} \) satisfies with probability at least \( 1 - 2p^{-4k(C_2^2 - 1) + 1} \)

\begin{equation}
\|\hat{\beta} - \beta^*\|_2 \leq \sqrt{1 + \frac{1}{t}} \left( \frac{16(\lambda_1 + 2\lambda_2)\sqrt{k}}{\sqrt{n} \eta_k^l} + \sqrt{\frac{2k \log p}{n}} \frac{16C_1(5/4 + 1/t)}{\eta_k^l} \right),
\end{equation}

where \( \eta_k^l = (\lambda_k - \theta_k^l \left( \frac{1}{7} + \frac{1}{4} \right))^2, C_1 = 1 + 2C_2 \sqrt{\frac{n}{k}}, \) and \( C_2 > 1 \) is a constant.

**Proof.** Without loss of generality, assume \( |h_1| \geq |h_2| \geq \ldots \geq |h_p| \). Let \( S_0 = \{1, 2, \ldots, k\} \), due to \( h \in \Delta_t = \{ \delta \in \mathbb{R}^p : \|\delta_T\|_1 \geq t\|\delta_{T^c}\|_1 \} \), we have \( \|h_{S_0}\|_1 \geq t\|h_{S_0}\|_1 \). Partition \( \{1, 2, \ldots, p\} \) into the following sets:

\( S_0 = \{1, 2, \ldots, k\}, S_1 = \{k+1, k+2, \ldots, 2k\}, S_2 = \{2k+1, 2k+2, \ldots, 3k\}, \ldots \)

Since, for any \( x \in \mathbb{R}^n, \lambda = \lambda_1 + 2\lambda_2 \), we have

\begin{equation}
\|x\|_2 - \|x\|_1 \geq \frac{\sqrt{n}}{4} \left( \max_{1 \leq i \leq n} |x_i| - \min_{1 \leq i \leq n} |x_i| \right).
\end{equation}

Then, we obtain that

\begin{equation}
\sum_{i \geq 1} \|h_{S_i}\|_2 \leq \sum_{i \geq 1} \|h_{S_i}\|_1 + \sqrt{k} \frac{1}{4} |h_{k+1}| \leq \frac{1}{\sqrt{k}} \|h_{S_0}\|_1 + \frac{1}{4\sqrt{k}} \|h_{S_0}\|_1.
\end{equation}
Again by Lemma 3 of [25], for any $\|Xh + \varepsilon\|_1 - \|\varepsilon\|_1 = \frac{1}{\sqrt{n}} (\|Xh_S_0 + \varepsilon\|_1 - \|\varepsilon\|_1)$

$$\frac{1}{\sqrt{n}} (\|Xh + \varepsilon\|_1 - \|\varepsilon\|_1) = \frac{1}{\sqrt{n}} (\|Xh_S_0 + \varepsilon\|_1 - \|\varepsilon\|_1) \leq M(h) - C_1 \sqrt{2k \log p} \|h_S_0\|_2. \tag{9}$$

By Lemma 3 of [25], we know that with probability at least $1 - 2p^{-4k(C_2^2 - 1)}$,

$$\frac{1}{\sqrt{n}} (\|Xh + \varepsilon\|_1 - \|\varepsilon\|_1) \geq M(h) - C_1 \sqrt{2k \log p} \|h_S_0\|_2. \tag{10}$$

Again by Lemma 3 of [25], for any $i \geq 1$ with probability at least $1 - 2p^{-4k(C_2^2 - 1)}$, 

$$\left(\|X(\sum_{j=0}^{i} h_{S_j}) + \varepsilon\|_1 - \|X(\sum_{j=0}^{i-1} h_{S_j}) + \varepsilon\|_1\right) \geq M(h) - C_1 \sqrt{2k \log p} \|h_S_i\|_2,$$ 

where $C_1 = 1 + 2C_2 \sqrt{\lambda_1}$ and $C_2 > 1$ is a constant. From the above inequalities, we know that with probability at least $1 - 2p^{-4k(C_2^2 - 1) + 1}$,

$$\frac{1}{\sqrt{n}} (\|Xh + \varepsilon\|_1 - \|\varepsilon\|_1) \geq M(h) - C_1 \sqrt{2k \log p} \sum_{i \geq 0} \|h_S_i\|_2. \tag{11}$$

By this and inequalities (4) and (9), we have that with probability at least $1 - 2p^{-4k(C_2^2 - 1) + 1}$,

$$M(h) \leq \frac{\lambda_1 + 2\lambda_2 \sqrt{k}}{\sqrt{n}} \|h_{S_0}\|_2 + C_1 \sqrt{2k \log p} \left(\frac{5}{4} + \frac{1}{t}\right) \|h_{S_0}\|_2. \tag{12}$$

Next, we consider two cases. First, if $\|Xh\|_1 \geq 3n/a$, then from Lemma 4 and Lemma 5 of [25],

$$M(h) = \frac{1}{\sqrt{n}} (\|Xh + \varepsilon\|_1 - \|\varepsilon\|_1) \geq \frac{3}{16 \sqrt{n}} \|Xh\|_1 \geq \frac{3 \sqrt{n}}{16} k \|h_{S_0}\|_2. \tag{13}$$

From assumption $\frac{3 \sqrt{n}}{16} K_k^2 > (\lambda_1 + 2\lambda_2 \sqrt{k/n} + C_1 \sqrt{2k \log p} \left(\frac{5}{4} + \frac{1}{t}\right)$, we have $\|h_{S_0}\|_2 = 0$ and hence $\beta^* = \hat{\beta}$. On the other hand, if $\|Xh\|_1 \leq 3n/a$, from Lemma 4 and Lemma 5 of [25],

$$M(h) = \frac{1}{\sqrt{n}} (\|Xh + \varepsilon\|_1 - \|\varepsilon\|_1) \geq \frac{a}{16 \sqrt{n}} \|Xh\|_2. \tag{14}$$

By the same argument as in the proof of Theorem 3.1 and 3.2 in [4], we know that $|\langle Xh_{S_0}, Xh \rangle| \geq n \lambda_k^1 \|h_{S_0}\|_2^2 - n \theta^*_{k} \|h_{S_0}\|_2 \sum_{i \geq 1} \|h_{S_i}\|_2 \geq n \left(\lambda_k^1 - \theta^*_{k} \left(\frac{1}{t} + \frac{1}{4}\right)\right) \|h_{S_0}\|_2^2$, and

$$|\langle Xh_{S_0}, Xh \rangle| \leq \|Xh_{S_0}\|_2 \|Xh\|_2 \leq \|Xh\|_2 \sqrt{n \lambda_k^2} \|h_{S_0}\|_2.$$
Therefore,
\[ \|Xh\|_2^2 \geq n \left( \frac{\lambda_k^\alpha - \theta_k^\alpha (\frac{1}{t} + \frac{1}{4})}{\lambda_k^\alpha} \right)^2 \|h_{S_0}\|^2. \]

Hence by (12) and (14), we have that with probability at least \(1 - 2p^{-4k(C_2^2 - 1) + 1},\)
\[ \|h_{S_0}\|_2 \leq \frac{16(\lambda_1 + 2\lambda_2)\sqrt{k}}{n\eta_k^l} + \sqrt{\frac{2k\log p}{n} 16C_1 (5/4 + 1/t)}, \] (15)
where \(\eta_k^l = \frac{(\lambda_k^\alpha - \theta_k^\alpha (\frac{1}{t} + \frac{1}{4}))^2}{\lambda_k^\alpha}.\) Since
\[ \sum_{i \geq 1} \|h_{S_i}\|_2^2 \leq |h_{k+1}| \sum_{i \geq 1} \|h_{S_i}\|_1 \leq \frac{1}{t} \|h_{S_0}\|_2^2, \]
we have that with probability at least \(1 - 2p^{-4k(C_2^2 - 1) + 1},\)
\[ \|\hat{\beta} - \beta^*\|_2 = \sqrt{\sum_{i \geq 1} \|h_{S_i}\|_2^2} \leq \sqrt{1 + \frac{1}{t} \left( \frac{16(\lambda_1 + 2\lambda_2)\sqrt{k}}{n\eta_k^l} + \sqrt{\frac{2k\log p}{n} 16C_1 (5/4 + 1/t)} \right)}, \]
where \(\eta_k^l = \frac{(\lambda_k^\alpha - \theta_k^\alpha (\frac{1}{t} + \frac{1}{4}))^2}{\lambda_k^\alpha}, C_1 = 1 + 2C_2\sqrt{\lambda_k^\alpha} \text{ and } C_2 > 1 \) is a constant. Thus we complete the proof. \(\square\)

Notice that from the above proof and (6), if we replace \(\|S\|_\infty \text{ with } \sqrt{2A(\alpha)n\log p},\) we obtain that for any \(\lambda_1 > 0, \lambda_2 > 0,\)
\[(\lambda_1 + 2\lambda_2)\|h_T\|_1 - (\lambda_1 - 2\lambda_2)\|h_{T^c}\|_1 \geq -\sqrt{2A(\alpha)n\log p} (\|h_T\|_1 + \|h_{T^c}\|_1). \] (16)
Furthermore, if we choose \(\lambda_1 = u \sqrt{2A(\alpha)n\log p}, \lambda_2 = v \sqrt{2A(\alpha)n\log p}, u - v > 1, u > 1, v > 0,\) we have that
\[ \|h_T\|_1 \geq \frac{u - v - 1}{u + v + 1} \|h_{T^c}\|_1 \text{ and } \lambda_1 + 2\lambda_2 \geq c\sqrt{2A(\alpha)n\log p}, \]
where \(A(\alpha) > 0\) is a constant such that \(2p^{-(A(\alpha) - 1)} \leq \alpha.\) Then, it follows from Lemma 1 in [25] that
\[ P(\lambda_1 + 2\lambda_2 \geq c\|S\|_\infty) \geq 1 - \alpha. \] (17)
Similarly, setting \(A(\alpha) = 2,\) we can obtain
\[ P(\lambda_1 + 2\lambda_2 \geq c\|S\|_\infty) \geq 1 - \frac{2}{p}. \]

Hence, we can easily obtain the following corollary.

Corollary 1. Consider model (1), and assume \(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\) are independent and identically distributed random variables satisfying (8) and \(\lambda_1 = 2u\sqrt{n\log p}, 2\lambda_2 = 2v\sqrt{n\log p}, u - v > 1, u > 1, v > 0.\) Suppose \(\lambda_k^\alpha > \theta_k^\alpha \left( \frac{1}{t} + \frac{1}{4} \right) \text{ and } \frac{3\sqrt{n}}{16} K_k^l > (\lambda_1 + 2\lambda_2) \sqrt{k/n + C_1 \sqrt{2k\log p}} \left( \frac{5}{4} + \frac{1}{t} \right)\) hold, then the RFLASSO estimator \(\hat{\beta}\) satisfies with probability at least \(1 - 2p^{-4k(C_2^2 - 1) + 1}\)
\[ \|\hat{\beta} - \beta^*\|_2 \leq \sqrt{\frac{2k\log p}{n} \left( \frac{16\sqrt{2}(u + v)}{\eta_k^l} + \frac{16C_1 (5/4 + 1/t)}{\eta_k^l} \right)} \sqrt{1 + \frac{1}{t}}, \]
where \( \eta_k = \left( \frac{\lambda_k^2 - \theta_k^2 (1 + \frac{1}{2})}{\lambda_k^2} \right)^2 \), \( C_1 = 1 + 2C_2 \sqrt{X_k} \) and \( C_2 > 1 \) is a constant.

From the above Corollary we can easily obtain that with high probability,

\[
\| \tilde{\beta} - \beta^* \|_2 = O \left( \frac{2k \log p}{n} \right). \tag{18}
\]

This represents that the RFLASSO penalized least absolute deviation estimator has near oracle performance.

3. Optimal Algorithm. We will apply the linearized alternating direction method of multipliers (LADMM) to find the RFLASSO estimator. The global convergence result is also given. This is inspired by the least-squares with FLASSO penalty [13].

3.1. LADMM Algorithm for RFLASSO. We first reformulate the RFLASSO model (3) by introducing auxiliary variables \( \tau = (\tau_1, \tau_2, \cdots, \tau_n)^T \in \mathbb{R}^n \), \( \gamma = (\gamma_1, \gamma_2, \cdots, \gamma_{p-1}) \in \mathbb{R}^{p-1} \). That is, the RFLASSO model (3) can be rewritten as

\[
\min_{\beta \in \mathbb{R}^p, \gamma \in \mathbb{R}^{p-1}, \tau \in \mathbb{R}^n} \left\{ \lambda_1 \| \beta \|_1 + \lambda_2 \| \gamma \|_1 + \| \tau \|_1 + \alpha \| A \beta - \gamma \|_2^2 + \frac{\mu}{2} \| Y - X \beta - \tau \|_2^2 \right\}, \tag{19}
\]

The augmented Lagrangian function of (19) is

\[
L_{\mu, \nu}^R(\beta, \gamma, \tau, \alpha, \theta) = \lambda_1 \| \beta \|_1 + \lambda_2 \| \gamma \|_1 + \| \tau \|_1 + \alpha \| A \beta - \gamma \|_2^2 + \frac{\mu}{2} \| Y - X \beta - \tau \|_2^2,
\]

and the iterative scheme of ADMM is

\[
\begin{cases}
\beta^{k+1} = \arg \min_{\beta \in \mathbb{R}^p} L_{\mu, \nu}^R(\beta, \gamma^k, \tau^k, \alpha^k, \theta^k), \\
\gamma^{k+1} = \arg \min_{\gamma \in \mathbb{R}^{p-1}} L_{\mu, \nu}^R(\beta^{k+1}, \gamma, \tau^k, \alpha^k, \theta^k), \\
\tau^{k+1} = \arg \min_{\tau \in \mathbb{R}^n} L_{\mu, \nu}^R(\beta^{k+1}, \gamma^{k+1}, \tau, \alpha^k, \theta^k), \\
\alpha^{k+1} = \alpha^k - \nu(\lambda_k \beta^{k+1} - \gamma^{k+1}), \\
\theta^{k+1} = \theta^k - \mu(Y - X \beta^{k+1} - \tau^{k+1}).
\end{cases} \tag{20}
\]

Now, let us look at the resulting subproblems in (20). First, for the \( \beta \)-subproblem in (20), after trivial manipulation, it can be written as

\[
\beta^{k+1} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \lambda_1 \| \beta \|_1 + \frac{\mu}{2} \| y - X \beta - \tau^k - \theta^k \|_2^2 + \frac{\nu}{2} \| A \beta - \gamma^k - \alpha^k / \nu \|_2^2 \right\}.
\]

Let \( \tilde{X} = (\sqrt{\mu} X^T, \sqrt{\nu} A^T)^T \) and \( \tilde{y}_k = (\sqrt{\mu}(y - \tau^k - \theta^k / \mu)^T, \sqrt{\nu} (\gamma^k + \alpha^k / \nu)^T)^T \), then we have

\[
\beta^{k+1} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \lambda_1 \| \beta \|_1 + \frac{1}{2} \| \tilde{X} \beta - \tilde{y}_k \|_2^2 \right\}. \tag{21}
\]

This subproblem does not have a closed-form solution because of the non-identity matrix \( \tilde{X} \). As is shown in [27], we can linearize the quadratic form \( \| \tilde{X} \beta - \tilde{y}_k \|_2^2 / 2 \) in (21) and replace it by

\[
(\tilde{X}^T (\tilde{X} \beta^k - \tilde{y}_k)) (\beta - \beta^k) + \frac{\eta}{2} \| \beta - \beta^k \|_2^2,
\]
where the parameter \( \eta > 0 \) controls the proximity to \( \beta^k \) (see Theorem 3.5 for precise condition of \( \eta \) to ensure the convergence). Overall, we solve the following subproblem

\[
\beta^{k+1} = \arg \min_{\beta \in \mathbb{R}^n} \left\{ \lambda_1 \| \beta \|_1 + \bar{X}^T (\bar{X} \beta^k - \bar{y}_k)^T (\beta - \beta^k) + \frac{\eta}{2} \| \beta - \beta^k \|_2^2 \right\}
\]

and obtain an approximate solution of the \( \beta \)-subproblem in (20). Then, it follows from [13] that the closed-form solution of (22) is given by

\[
\beta^{k+1} = \text{shrink}_{1,2} (\beta^k - \bar{X}^T (\bar{X} \beta^k - \bar{y}_k)/\eta, \lambda_1/\eta) = \tilde{u}^k - P_C (\tilde{u}^k),
\]

where \( \tilde{u}^k = \beta^k - \bar{X}^T (\bar{X} \beta^k - \bar{y}_k)/\eta \) and \( C = [-\lambda_1/\eta, \lambda_1/\eta] \).

Second, for the \( \gamma \)-subproblem in (20), the closed-form solution is easy to given by

\[
\gamma^{k+1} = \arg \min_{\gamma \in \mathbb{R}^{n-1}} \left\{ \lambda_2 \| \gamma \|_1 + \frac{\nu}{2} \| \gamma - A \beta^{k+1} + \alpha^k/\nu \|_2^2 \right\}
\]

\[
\gamma^{k+1} = \text{shrink}_{1,2} (A \beta^{k+1} - \alpha^k/\nu, \lambda_2/\nu)
\]

\[
\gamma^{k+1} = w^k - P_F (w^k),
\]

where \( w^k = A \beta^{k+1} - \alpha^k/\nu \) and \( F = [-\lambda_2/\nu, \lambda_2/\nu] \).

Third, similarly, for the \( \tau \)-subproblem in (20), its closed-form solution is

\[
\tau^{k+1} = \arg \min_{\tau \in \mathbb{R}^n} \left\{ \| \tau \|_1 + \frac{\mu}{2} \| \tau - y + X \beta^{k+1} - \theta^k/\mu \|_2^2 \right\}
\]

\[
\tau^{k+1} = \text{shrink}_{1,2} (y - X \beta^{k+1} + \theta^k/\mu, 1/\mu)
\]

\[
\tau^{k+1} = m^k - P_\mathcal{P} (m^k),
\]

where \( m^k = y - X \beta^{k+1} + \theta^k/\mu \) and \( \mathcal{P} = [-1/\mu, 1/\mu] \).

Generally, the iterative scheme of LADMM algorithm for RFLASSO can be described as follows in Table 1.

| LADMM Algorithm for RFLASSO |
|-----------------------------|
| **Input**: \( X, Y, \tau, \lambda_1 > 0, \lambda_2 > 0, \mu > 0, \nu > 0 \) and \( \eta > \rho (\mu X^T X + \nu A^T A) \), where \( \rho (\cdot) \) denotes the spectral radius. |
| Select \( (\beta^0, \gamma^0, \tau^0, \alpha^0, \theta^0) = (1, 1, 1, 1, 1) \); |
| **For** \( k = 1, 2, \ldots, n \) |
| **Do** |
| Compute \( \beta^{k+1} \) by (23), |
| Compute \( \gamma^{k+1} \) by (24), |
| Compute \( \tau^{k+1} \) by (25), |
| Update \( \alpha^{k+1} = \alpha^k - \nu (A \beta^{k+1} - \gamma^{k+1}) \), |
| Update \( \theta^{k+1} = \theta^k - \mu (y - X \beta^{k+1} - \tau^{k+1}) \). |
| **End** |
| **Output**: \( (\beta^n, \gamma^n, \tau^n, \alpha^n, \theta^n) \) as an approximate solution of (3). |
3.2. Convergence Analysis. In this section, we establish the convergence for the aforementioned LADMM algorithm. Our analysis is motivated by some existing work in different settings [1, 6, 10, 16, 27, 29, 30]. While the existing work deals with FLASSO penalty least square model, we focus on the fused lasso penalty least absolute deviation context based on the variational characterization of the RFLASSO model.

The Lagrange function of the RFLASSO model (19) is

\[
L^RF = \lambda_1\|\beta\|_1 + \lambda_2\|\gamma\|_1 + \|\tau\|_1 - \alpha^T(A\beta - \gamma) - \theta^T(Y - X\beta - \tau),
\]

where \(\theta \in \mathbb{R}^n, \alpha \in \mathbb{R}^{p-1}\) are the Lagrange multipliers.

Using the first-order optimality condition of (26), we have that solving (19) is equivalent to finding \((\beta^*, \gamma^*, \tau^*, \alpha^*, \theta^*) \in S := \mathbb{R}^p \times \mathbb{R}^{p-1} \times \mathbb{R}^n \times \mathbb{R}^{p-1} \times \mathbb{R}^n, \ g_1(\beta^*) \in \partial(\|\beta^*\|_1), \ g_2(\gamma^*) \in \partial(\|\gamma^*\|_1),\) and \(g_3(\tau^*) \in \partial(\|\tau^*\|_1)\) such that

\[
\begin{align*}
\lambda_1 g_1(\beta^*) + X^T\theta^* - A^T\alpha^* &= 0, \\
\lambda_2 g_2(\gamma^*) + \alpha^* &= 0, \\
g_3(\tau^*) + \theta^* &= 0, \\
A\beta^* - \gamma^* &= 0, \\
Y - X\beta^* - \tau^* &= 0.
\end{align*}
\]

Note that \(\partial(\cdot)\) denotes the subdifferential operator of a nonsmooth convex function. Denote by \(S^*\) all the elements in \(S\) satisfying (27). Then using the notation \(w^* = (\beta^*, \gamma^*, \tau^*, \alpha^*, \theta^*) \in S^*, \ g_1(\beta) \in \partial(\|\beta\|_1), \ g_2(\gamma) \in \partial(\|\gamma\|_1),\) and \(g_3(\tau) \in \partial(\|\tau\|_1),\) (27) is equivalent to a variational problem: to find a \(w^* \in S^*; \ g_1(\beta^*) \in \partial(\|\beta^*\|_1), \ g_2(\gamma^*) \in \partial(\|\gamma^*\|_1)\) and \(g_3(\tau^*) \in \partial(\|\tau^*\|_1)\) such that the following variational inequality holds:

\[
\text{VI}(S, F) : (w - w^*)^TF(w^*) \geq 0, \ \forall w \in S,
\]

where

\[
w = \begin{pmatrix} \beta \\ \gamma \\ \tau \\ \alpha \\ \theta \end{pmatrix} \quad \text{and} \quad F(w) = \begin{pmatrix} \lambda_1 g_1(\beta) + X^T\theta - A^T\alpha \\ \lambda_2 g_2(\gamma) + \alpha \\ g_3(\tau) + \theta \\ A\beta - \gamma \\ Y - X\beta - \tau \end{pmatrix}.
\]

In order to establish convergence of LADMM, we characterize the \((k+1)\)-th iteration of the Algorithm as a VI problem in the following lemma. Here we define the block matrix \(G\) as

\[
G = \begin{pmatrix} \eta I_p - \tilde{X}^T\tilde{X} & 0 & 0 & 0 & 0 \\
0 & \nu I_{p-1} & 0 & 0 & 0 \\
0 & 0 & \mu I_n & 0 & 0 \\
0 & 0 & 0 & 1/\nu & 0 \\
0 & 0 & 0 & 0 & 1/\mu \end{pmatrix},
\]

where \(\tilde{X} = (\sqrt{\mu}X^T, \sqrt{\nu}A^T)^T\). The positive definiteness of \(G\) can be ensured by the condition \(\eta > \rho(\mu X^TX + \nu A^TA)\).
Lemma 3.1. Let \( \{w^k\} \) be the sequence generated by the LADMM algorithm. We have
\[
(w^k - w^{k+1})^T [F(w^{k+1}) + M(\gamma^k - \gamma^{k+1}) + N(\tau^k - \tau^{k+1}) - G(w^k - w^{k+1})] \geq 0, \forall w' \in S,
\]
where
\[
M = \begin{pmatrix}
-\nu A^T \\
\nu I_{p-1} \\
0_{(p-1) \times (p-1)} \\
0_{n \times (p-1)}
\end{pmatrix}
\quad \text{and} \quad
N = \begin{pmatrix}
\mu X^T \\
0_{(p-1) \times n} \\
\mu I_n \\
0_{n \times (p-1)}
\end{pmatrix}.
\]

Proof. First, we have
\[
X^T(\tilde{X}(\beta^{k+1} - \gamma_k) = \left( \sqrt{\nu}X^T, \sqrt{\nu}A^T \right) \left( \sqrt{\nu}X\beta^{k+1} - \sqrt{\nu}(Y - \tau^k - \theta^k) / \mu \right)
\]
\[
= \nu X^T X\beta^{k+1} - \mu X^T Y - \mu X^T \tau^k + X^T \theta^k - \nu A^T \gamma^k - A^T \alpha^k
\]
\[
= \nu A^T \gamma^k - A^T (A^{k+1} + A\beta^{k+1} - \mu) + \nu A^T \gamma^{k+1} - A^T \beta^{k+1} + \nu A^T \alpha^k + X^T \theta^k - \nu A^T \gamma^k - A^T \alpha^k
\]
\[
= X^T \theta^k + X^T X(\beta^{k+1} - \beta^k) - X^T \tau^k - \nu A^T \gamma^k + A^T \alpha^k + X^T \theta^k - \nu A^T \gamma^k - A^T \alpha^k.
\]
It follows from (20) that
\[
\begin{cases}
\alpha^k = \alpha^{k+1} + \nu (A\beta^{k+1} - \gamma^{k+1}), \\
\theta^k = \theta^{k+1} + \mu (Y - X\beta^{k+1} - \tau^{k+1}).
\end{cases}
\]
From the first order optimality condition of the minimization problems (22), (24) and (25), the iterative scheme (20) is equivalent to finding \( w^{k+1} = (\beta^{k+1}, \gamma^{k+1}, \tau^{k+1}, \theta^{k+1}) \in S, \ g_1(\beta^{k+1}) \in \partial(\|\beta^{k+1}\|_1), \ g_2(\gamma^{k+1}) \in \partial(\|\gamma^{k+1}\|_1) \) and \( g_3(\tau^{k+1}) \in \partial(\|\tau^{k+1}\|_1) \) such that
\[
\begin{cases}
0 = \lambda_1 g_1(\beta^{k+1}) + X^T(\tilde{X}(\beta^{k+1} - \gamma^{k+1}) + X^T \tilde{X}(\beta^{k+1} - \gamma^{k+1}) + X^T \tilde{X}(\beta^{k+1} - \beta^{k+1}), \\
0 = \lambda_2 g_2(\gamma^{k+1}) + \nu(\gamma^{k+1} - A\beta^{k+1} + \alpha^k) / \nu, \\
0 = g_3(\tau^{k+1}) + X^T \theta^k + X^T X(\beta^{k+1} - \beta^k) - X^T \tau^k - \nu A^T \gamma^k + A^T \alpha^k + X^T \theta^k - \nu A^T \gamma^k + A^T \alpha^k, \\
0 = Y - X\beta^{k+1} - \tau^{k+1} - (\theta^k - \theta^{k+1}) / \mu.
\end{cases}
\]
The definition of \( G \) yields that (33) can be rewritten as
\[
\begin{cases}
0 = \lambda_1 g_1(\beta^{k+1}) + X^T \theta^k + X^T \alpha^{k+1} + \eta(\beta^{k+1} - \beta^k) + \mu X^T(\tau^k - \tau^{k+1}) + \nu A^T(\gamma^{k+1} - \gamma^k) + \mu X^T(\tau^k - \tau^{k+1}), \\
0 = \lambda_2 g_2(\gamma^{k+1}) + \alpha^{k+1}, \\
0 = g_3(\tau^{k+1}) + \theta^{k+1}, \\
0 = A\beta^{k+1} - \gamma^{k+1} - (\theta^k - \theta^{k+1}) / \mu, \\
0 = Y - X\beta^{k+1} - \tau^{k+1} - (\theta^k - \theta^{k+1}) / \mu.
\end{cases}
\]
Then
\[
F(w^{k+1}) = \begin{pmatrix}
\lambda_1 g_1(\beta^{k+1}) + X^T\theta^{k+1} - A^T\alpha^{k+1} \\
\lambda_2 g_2(\gamma^{k+1}) + \alpha^{k+1} \\
g_3(\tau^{k+1}) + \theta^{k+1} \\
A\beta^{k+1} - \gamma^{k+1} \\
Y - X\beta^{k+1} - \tau^{k+1}
\end{pmatrix}.
\]

So
\[
M(\gamma^k - \gamma^{k+1}) = \begin{pmatrix}
-\nu A^T(\gamma^k - \gamma^{k+1}) \\
\nu(\gamma^k - \gamma^{k+1}) \\
0_{p \times 1} \\
0_{(p-1) \times 1} \\
0_{n \times 1}
\end{pmatrix},
\]
\[
N(\tau^k - \tau^{k+1}) = \begin{pmatrix}
\mu X^T(\tau^k - \tau^{k+1}) \\
0_{(p-1) \times 1} \\
\mu(\tau^k - \tau^{k+1}) \\
0_{(p-1) \times 1} \\
0_{n \times 1}
\end{pmatrix},
\]
and
\[
G(w^k - w^{k+1}) = \begin{pmatrix}
\eta I_p - \tilde{X}^T\tilde{X} & 0 & 0 & 0 & 0 \\
0 & \nu I_{p-1} & 0 & 0 & 0 \\
0 & 0 & \mu I_n & 0 & 0 \\
\eta I_p - \tilde{X}^T\tilde{X}(\beta^k - \beta^{k+1}) & \nu(\gamma^k - \gamma^{k+1}) \\
\mu(\tau^k - \tau^{k+1}) & 1(\alpha^k - \alpha^{k+1}) \\
\frac{1}{\nu}(\theta^k - \theta^{k+1}) & \frac{1}{\mu}(\theta^k - \theta^{k+1})
\end{pmatrix}.
\]

As a consequence,
\[
F(w^{k+1}) + M(\gamma^k - \gamma^{k+1}) + N(\tau^k - \tau^{k+1}) - G(w^k - w^{k+1}) = 0. \tag{35}
\]
The variational inequality yields (35) can be rewritten as
\[
(w' - w^{k+1})^T[F(w^{k+1})] + M(\gamma^k - \gamma^{k+1}) + N(\tau^k - \tau^{k+1}) - G(w^k - w^{k+1})
\geq 0, \quad \forall w' \in S. \tag{36}
\]

The following lemma can be easily derived by Lemma 3.1. For completeness, we prove it in detail.

**Lemma 3.2.** Let \( \{w^k\} \) be the sequence generated by the LADMM algorithm. Then for any \( w^* \in S^* \),
\[
(w^k - w^*)^T G(w^k - w^{k+1}) \geq (w^k - w^{k+1})^T G(w^k - w^{k+1}) - (\alpha^k - \alpha^{k+1})^T (\gamma^k - \gamma^{k+1}) - (\theta^k - \theta^{k+1})^T (\tau^k - \tau^{k+1}). \tag{37}
\]
Proof. Let $w^*$ be an arbitrary solution point in $S^*$. The inequality (30) in Lemma 3.1 yields
\[(w^* - w^{k+1})^T[F(w^{k+1})] + M(\gamma^k - \gamma^{k+1}) + N(\tau^k - \tau^{k+1}) - G(w^k - w^{k+1}) \geq 0. \tag{38}\]
Since $w^* \in S^*$, we have $Y - X\beta^* - \tau^* = 0$ and $A\beta^* - \gamma^* = 0$. Thus (38) leads to
\[
(w^{k+1} - w^*)^T G(w^k - w^{k+1}) \\
\geq (w^{k+1} - w^*)^T F(w^{k+1}) + (w^{k+1} - w^*)^T M(\gamma^k - \gamma^{k+1}) \\
+ (w^{k+1} - w^*)^T N(\tau^k - \tau^{k+1}). \tag{39}\]
Since
\[\begin{align*}
(w^{k+1} - w^*)M &= \begin{pmatrix}
\beta^{k+1} - \beta^* \\
\gamma^{k+1} - \gamma^* \\
\tau^{k+1} - \tau^* \\
\alpha^{k+1} - \alpha^* \\
\theta^{k+1} - \theta^*
\end{pmatrix}^T \begin{pmatrix}
-\nu A^T \\
\nu I_{p-1} \\
O_{p \times (p-1)} \\
O_{(p-1) \times (p-1)} \\
O_{n \times (p-1)}
\end{pmatrix} = -\nu(A\beta^{k+1} - \gamma^{k+1})^T,
\end{align*}\]
and
\[\begin{align*}
(w^{k+1} - w^*)N &= \begin{pmatrix}
\beta^{k+1} - \beta^* \\
\gamma^{k+1} - \gamma^* \\
\tau^{k+1} - \tau^* \\
\alpha^{k+1} - \alpha^* \\
\theta^{k+1} - \theta^*
\end{pmatrix}^T \begin{pmatrix}
\mu X^T \\
\nu I_{n \times n} \\
O_{(p-1) \times n} \\
O_{n \times (p-1)} \\
O_{n \times n}
\end{pmatrix} = \mu(X\beta^{k+1} + \tau^{k+1} - Y)^T,
\end{align*}\]
we have
\[
(w^{k+1} - w^*)^T G(w^k - w^{k+1}) \\
\geq (w^{k+1} - w^*)^T F(w^{k+1}) - (\alpha^k - \alpha^{k+1})^T (\gamma^k - \gamma^{k+1}) \\
- (\theta^k - \theta^{k+1})^T (\tau^k - \tau^{k+1}). \tag{40}\]
On the other hand, the convexity of $\| \cdot \|_1$ yields that the mapping $F(w)$ defined in (28) is monotone. We thus have
\[(w^{k+1} - w^*)^T F(w^{k+1}) - F(w^*) \geq 0,\]
which shows that
\[(w^{k+1} - w^*)^T F(w^{k+1}) \geq (w^{k+1} - w^*)^T F(w^*) \geq 0.\]
Replacing $(w^{k+1} - w^*)$ by $(w^{k+1} - w^*) + (w^k - w^*)$ in (40), we get for any $w^* \in S^*$,
\[
(w^{k+1} - w^*)^T G(w^k - w^{k+1}) \\
\geq (w^{k+1} - w^*)^T G(w^k - w^{k+1}) - (\alpha^k - \alpha^{k+1})^T (\gamma^k - \gamma^{k+1}) \\
- (\theta^k - \theta^{k+1})^T (\tau^k - \tau^{k+1}). \tag{41}\]

Combining with the above two lemmas, we can now show that the sequence $\{w^k\}$ generated by the LADMM algorithm is contractive with respect to the solution set $S^*$.

Lemma 3.3. Let $\{w^k\}$ be the sequence generated by the LADMM algorithm. Then for any $w^* \in S^*$, we have
\[
\|w^{k+1} - w^*\|_2^2 \leq \|w^k - w^*\|_2^2 - \|w^k - w^{k+1}\|_2^2.
\]
Proof. It follows from (37) in Lemma 3.2 that for all \( w^* \in S^* \),
\[
\|w^{k+1} - w^*\|^2_G
\]
\[
= \|w^k - w^*\|^2_G + \|w^k - w^{k+1}\|^2_G - 2(w^k - w^*)^T G (w^k - w^{k+1})
\]
\[
\leq \|w^k - w^*\|^2_G - \|w^k - w^{k+1}\|^2_G + 2(\alpha^k - \alpha^{k+1})^T (\gamma^k - \gamma^{k+1})
\]
\[
- 2(\theta^k - \theta^{k+1})^T (\tau^k - \tau^{k+1}).
\]  
(42)

As we have showed that \( \lambda_2 g(\gamma^k) + \alpha^k = 0, g_3(\tau^k) + \theta^k = 0 \) for any \( k \). Thus
\[
(\gamma^k - \gamma^{k+1})^T (\lambda_2 g_2(\gamma^{k+1}) + \alpha^{k+1}) \geq 0, \ (\gamma^{k+1} - \gamma^k)^T (\lambda_2 g_2(\gamma^k) + \alpha^k) \geq 0,
\]
and
\[
(\tau^k - \tau^{k+1})^T (g_3(\tau^{k+1}) - \theta^{k+1}) \geq 0, \ (\tau^{k+1} - \tau^k)^T (g_3(\tau^k) - \theta^k) \geq 0.
\]
Therefore,
\[
(\alpha^k - \alpha^{k+1})^T (\gamma^k - \gamma^{k+1}) \leq \lambda_2 (\gamma^k - \gamma^{k+1})^T (g_2(\gamma^{k+1}) - g_2(\gamma^k)) \leq 0,
\]  
(43)
and
\[
(\theta^k - \theta^{k+1})^T (\tau^k - \tau^{k+1}) \leq (\tau^k - \tau^{k+1})^T (g_3(\tau^{k+1}) - g_3(\tau^k)) \leq 0.
\]  
(44)

Inserting (43) and (44) into (42), we prove the assertion. \( \square \)

Lemma 3.3 implies that the sequence generated by the LADMM algorithm is contractive with respect to the solution set \( S \). The assertions summereized in the following corollary are trivial based on Lemma 3.3.

**Corollary 2.** Let \( \{w^k\} \) be the sequence generated by the LADMM algorithm. We have
1. \( \lim_{k \to \infty} \|w^k - w^{k+1}\|_G = 0; \)
2. The sequence \( \{w^k\} \) is bounded;
3. For all \( w^* \in S^* \), the sequence \( \{\|w^k - w^*\|_G\} \) is non-increasing.

Now we show the convergence of the LADMM algorithm based on Corollary 2.

**Theorem 3.4.** For any \( \mu \geq 0, \nu \geq 0, \eta \geq \rho(\mu X^T X + \nu A^T A) \), the sequence \( \{w^k = (\beta^k, \gamma^k, \tau^k, \theta^k, \alpha^k)\} \) generated by LADMM algorithm with an arbitrary initial iterate \( (\beta^0, \gamma^0, \tau^0, \theta^0, \alpha^0) \in S \) converges to \( w^\infty = (\beta^\infty, \gamma^\infty, \tau^\infty, \theta^\infty, \alpha^\infty) \), where \( (\beta^\infty, \gamma^\infty, \tau^\infty, \theta^\infty, \alpha^\infty) \) is a solution point of the RFLASSO model (3).

**Proof.** The property 1 in Corollary 2 means that
\[
\begin{align*}
\lim_{k \to \infty} \|\beta^k - \beta^{k+1}\| &= 0, \\
\lim_{k \to \infty} \|\gamma^k - \gamma^{k+1}\| &= 0, \\
\lim_{k \to \infty} \|\tau^k - \tau^{k+1}\| &= 0, \\
\lim_{k \to \infty} \|\alpha^k - \alpha^{k+1}\| &= 0, \\
\lim_{k \to \infty} \|\theta^k - \theta^{k+1}\| &= 0.
\end{align*}
\]

In addition, the property 2 in Corollary 2 implies that the sequence \( \{w^k\} \) has at least one cluster point. We denote it by \( w^\infty = (\beta^\infty, \gamma^\infty, \tau^\infty, \theta^\infty, \alpha^\infty) \), and let \( \{w^j\} \)
be a subsequence converging to \( w^\infty \). Thus we have
\[
\begin{align*}
\beta^{kj} & \to \beta^\infty, \\
\gamma^{kj} & \to \gamma^\infty, \\
\tau^{kj} & \to \tau^\infty, \\
\theta^{kj} & \to \theta^\infty, \\
\alpha^{kj} & \to \alpha^\infty,
\end{align*}
\]
and
\[
\begin{align*}
\lim_{j \to \infty} \|\beta^{kj} - \beta^{kj+1}\| &= 0, \\
\lim_{j \to \infty} \|\gamma^{kj} - \gamma^{kj+1}\| &= 0, \\
\lim_{j \to \infty} \|\tau^{kj} - \tau^{kj+1}\| &= 0, \\
\lim_{j \to \infty} \|\alpha^{kj} - \alpha^{kj+1}\| &= 0, \\
\lim_{j \to \infty} \|\theta^{kj} - \theta^{kj+1}\| &= 0.
\end{align*}
\] (45) and (46)

Now, we show that the cluster point \( w^\infty \) satisfies the optimality condition (27). Due to Lemma 3.1 and (46), we have
\[
\lim_{j \to \infty} (w' - w^{kj})^T F(w^{kj}) \geq 0, \quad \forall w' \in \mathcal{S}.
\]
Then from (45), the above inequality becomes
\[
(w' - w^\infty)^T F(w^\infty) \geq 0, \quad \forall w' \in \mathcal{S}.
\]
Thus, the cluster point \( w^\infty \) satisfies (27), i.e., \( w^\infty \in S^* \). From property 3 in Corollary 2, we have
\[
\|w^{k+1} - w^\infty\|^2_G \leq \|w^k - w^\infty\|^2_G, \quad \forall k \geq 0.
\]
Therefore, the sequence \( \{w^k\} \) has a unique cluster point \( w^\infty \). That is, \( \{w^k\} \) converges to \( w^\infty \) where \((\beta^\infty, \gamma^\infty, \tau^\infty, \theta^\infty, \alpha^\infty)\) is a solution point of the RFLASSO model (3). This completes the proof.

4. Numerical Results. In this section, we apply the proposed LADMM algorithm to solve the RFLASSO model, and compare with the SLEP package [14]. The numerical experimemt is conducted on a desktop computer with the Intel (R) Core (TM) i5 2.80 GHz CPU and 3.88 GB of RAM running windows 8.1. Our codes for implementing LADMM is written in MATLAB (R2014a) (MathWorks, Natick, MA).

For the implementation, we now address the initialization and the termination criteria for the method. The initial iterate is taken as \((\beta^0, \gamma^0, \tau^0, \alpha^0, \theta^0) = (1, 1, 1, 1, 1)\). For the test algorithms, the maximum iteration number is set as 1000 and the stopping criterion is set as \( \text{RelErr} < 2.5e^{-3} \), which is defined as
\[
\text{RelErr} := \frac{\|\beta_k - \beta_{k-1}\|_2}{\max\{\|\beta_k\|_2, 1\}}.
\]

We first show how to generate the synthetic data. The Gaussian matrix \( X \in \mathbb{R}^{n \times p} \) with unit column norm is generated randomly. To generate a sparse solution with close relationship among the successive coefficient, we divide \( p \) into 80 groups and randomly select 10 groups denoted as a sample set \( G \), whose cardinality is \( g \). Then, the coefficient vector \( \beta^* \) is generated by \( \beta_i^* = U([-3, 3]) \), if \( i \in G \), otherwise
\[ \beta_i^* = 0, \] where \( U[-3, 3] \) represents the uniform distribution on the interval \([-3, 3]\). Finally, we can get the observation data \( y \) by \( y = X\beta^* + \varepsilon \) with \( \varepsilon \sim n(0, \sigma^2 I) \).

We test the cases \( \sigma = 0.001, 0.005 \) and \( 0.01 \), with \((n,p,g) = (360i, 1280i, 160i)\) for \( i = 1, \cdots, 4 \). As suggested by the SLEP, the tuning parameters are chosen as \( \lambda_1 = 0.1 \) and \( \lambda_2 = 0.1 \) in (3). For the parameters \( \mu \) and \( \nu \), we take them as 0.5.

For the \( \eta > \rho(\mu X^T X + \nu A^T A) \), we report the values of \( \rho(\mu X^T X + \nu A^T A) \) for some cases of \( \sigma \) and \( i \) in Table 2. As suggested in [13], the value of \( \eta \) is determined by the following rule

\[
\eta^{k+1} = \begin{cases} 
\eta^k, & \text{if } \text{mod}(k, 5) \neq 0; \\
\max\{\eta^k \times 0.78, 0.1\}, & \text{if } \text{mod}(k, 5) = 0 \text{ and } k \leq 20; \\
\min\{\eta^k \times 1.01, 50\}, & \text{if } \text{mod}(k, 5) = 0 \text{ and } k > 20,
\end{cases}
\]

starting from \( \eta = 5.5 \).

**Table 2.** \( \rho(0.5X^T X + 0.5A^T A) \) for design matrix with unit column norms.

| \( i \) | 1   | 2   | 3   | 4   |
|-------|-----|-----|-----|-----|
| \( \sigma = 0.001 \) | 5.246 | 5.289 | 5.324 | 5.344 |
| \( \sigma = 0.005 \) | 5.265 | 5.296 | 5.333 | 5.350 |
| \( \sigma = 0.01 \) | 5.273 | 5.301 | 5.342 | 5.356 |

Table 3 reports the results of our algorithm for (3) in terms of the number of iterations (“Iter”), the computing time in seconds (“Time”), the objective function (“Obj”) and the estimation error (“Error”) measured by \( \|\beta - \beta^*\|_1 \). We can see that the computing time of the algorithm is increasing when the value of \( i \) is increasing.

**Table 3.** The results for RFLASSO with size \((n,p,g) = (360i, 1280i, 160i)\).

| \( i \) | \( \sigma \) | Iter | CPU  | Obj  | Error  |
|-------|-------------|------|------|------|--------|
| \( \sigma = 0.001 \) | 165 | 0.424 | 31.052 | 20.3164 |
| \( \sigma = 0.001 \) | 150 | 1.783 | 48.669 | 24.9877 |
| \( \sigma = 0.01 \) | 168 | 1.953 | 59.415 | 25.5110 |
| \( \sigma = 0.01 \) | 189 | 2.113 | 67.510 | 31.8536 |
| \( \sigma = 0.001 \) | 167 | 4.109 | 91.263 | 34.8521 |
| \( \sigma = 0.001 \) | 166 | 7.525 | 123.087 | 40.0416 |
| \( \sigma = 0.001 \) | 157 | 6.833 | 137.114 | 44.1079 |
Figures 1-2 visualize the results for RFLASSO and FLASSO where \( i = 2 \) and \( \sigma = 0.001 \), respectively. It is easy to derive that both the estimators which are generated by RFLASSO and FLASSO are very closely with the true value. Furthermore, we summarize the results when \( i = 2 \) and \( \sigma = 0.001 \) in Table 4. In this table, we use “\( \#$\)” to denote the number of the elements, which is a more specific measurement to the proximity of the approximate solution \( \beta \) to the true solution \( \beta^* \) for the RFLASSO model. This table illustrates that our proposed RFLASSO model can compare with FLASSO model under Gaussian noise.

Next, we will compare RFLASSO with FLASSO under non-Gaussian noise, i.e., Cauchy distribution. In Figures 3-4, we visualize the results of RFLASSO and
Table 4. Selected results for RFLASSO and FLASSO.

|                      | RFLASSO | FLASSO |
|----------------------|---------|--------|
| $\sharp\{ |\hat{\beta}_i| < 0.1, \ i \in G^c \}$ | 2240    | 2240   |
| $\max_{i \in G^c} |\hat{\beta}_i|$ | 0.0147  | 0.0153 |
| $\sharp\{ |\hat{\beta}_i - \beta^*_i| < 0.1, \ i \in G \}$ | 320     | 320    |
| $\max_{i \in G} |\hat{\beta}_i - \beta^*_i|$ | 0.1447  | 0.1365 |

FLASSO for the Cauchy noise case where $i = 2$ and $\sigma = 0.001$. We further summarize the results in Table 5. This table shows that RFLASSO achieve the solutions with higher quality to the true solution $\beta^*$, that is to say, RFLASSO is more robust than FLASSO when the noise is heavy-tailed distribution.

Figure 3. Results for RFLASSO when $i = 2, \sigma = 0.001$.

Figure 4. Results for FLASSO when $i = 2, \sigma = 0.001$. 
Table 5. Selected results for RFLASSO and FLASSO.

|                  | RFLASSO | FLASSO |
|------------------|---------|--------|
| $\sharp \{ |\hat{\beta}_i | < 0.1, i \in G^c \}$ | 2240    | 2372   |
| $\max_{i \in G^c} |\hat{\beta}_i |$         | 0.0442  | 0.0997 |
| $\sharp \{ |\hat{\beta}_i - \beta^*_i | < 0.1, i \in G \}$ | 320     | 188    |
| $\max_{i \in G} |\hat{\beta}_i - \beta^*_i |$         | 0.1329  | 1.7218 |

5. Conclusions. In this article we deal with Robust Fused LASSO under the assumption that both the unknown vector and its successive differences are sparse. The main contribution of this paper is to prove the theoretical results on the Robust Fused LASSO estimator. For a wide range of noise distributions included heavy-tailed or the Cauchy distribution, we show that under some mild condition the Robust Fused LASSO estimator possesses near oracle performance, i.e., with high probability, the $\ell_2$ norm of the estimation error is of order $O(\sqrt{k \log p/n})$. Moreover, we apply the linearized alternating direction method of multipliers to find the Robust Fused LASSO estimator, which possesses the global convergence. We compare Robust Fused LASSO and Fused LASSO under different noise distributions by the numerical experiments, which demonstrates the efficiency of our proposed method. In the future study, we will try to apply our model to real data.

Acknowledgments. The authors would like to thank the anonymous referees for their numerous insightful comments and suggestions, which have greatly improved the paper.

REFERENCES

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends® in Machine Learning, 3 (2010), 1–122.
[2] E. Candès, Compressive sampling, In Proceedings of the International Congress of Mathematicians, 3 (2006), 1433–1452.
[3] E. Candès and T. Tao, Decoding by linear programming, IEEE Transactions on Information Theory, 51 (2005), 4203–4215.
[4] T. Cai, L. Wang and G. Xu, Shifting inequality and recovery of sparse signals, IEEE Transactions on Signal Processing, 58 (2010), 1300–1308.
[5] H.B. Chen, Y.J. Wang, G. Wang, Strong convergence of extragradient method for generalized variational inequalities in Hilbert space, Journal of Inequalities and Applications, 2014 (2014), 223, 11 pages.
[6] W. Deng and W. Yin, On the global and linear convergence of the generalized alternating direction method of multipliers, Journal of Scientific Computing, 66 (2016), 889–916.
[7] J. Friedman, Multivariate adaptive regression splines, The Annals of Statistics, (1991), 1–67.
[8] J. Friedman, T. Hastie, H. H"ofling and R. Tibshirani, Pathwise coordinate optimization, Annals of Applied Statistics, 1 (2007), 302–332.
[9] M. Grant, S. Boyd and Y. Ye, CVX: Matlab Software for Disiplined Convex Programming, 2009.
[10] D. Han, X. Yuan and W. Zhang, An augmented Lagrangian based parallel splitting method for separable convex minimization with applications to image processing, Mathematics of Computation, 83 (2014), 2263–2291.
[11] H. H"ofling, A path algorithm for the fused lasso signal approximator, Journal of Computational and Graphical Statistics, 19 (2009), 984–1006.
[12] E. Koc and H. Bozdogan, Model selection in multivariate adaptive regression splines (MARS) using information complexity as the fitness function, Machine Learning, 101 (2015), 35–58.
[13] X. Li, L. Mo, X. Yuan and J. Zhang, Linearized alternating direction method of multipliers for sparse group and fused-LASSO models, *Computational Statistics & Data Analysis*, 79 (2013), 203–221.

[14] J. Liu, S. Ji and J. Ye, SLEP: sparse learning with efficient projections, http://www.public.asu.edu/~jye02/Software/SLEP, 2009.

[15] H. Ma and Y. Jia, Stability Analysis For Stochastic Differential Equations With Infinite Markovian Switchings, *Journal of Mathematical Analysis and Applications*, 435 (2016), 593–605.

[16] S. Ma, L. Xue and H. Zou, Alternating direction methods for latent variable Gaussian graphical model selection, *Neural Computation*, 25 (2013), 2172–2198.

[17] D. Martinez, D. Shih, V. Chen and S. Kim, A convex version of multivariate adaptive regression splines, *Computational Statistics & Data Analysis*, 81 (2015), 89–106.

[18] B. Pete and V. Sara, *Statistics for High-dimensional Data: Methods, Theory and Applications*, Springer, 2011.

[19] A. Rinaldo, Properties and refinements of the fused Lasso, *Annals of Statistics*, 37 (2009), 2922–2952.

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

[21] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu and K. Knight, Sparsity and smoothness via the fused Lasso, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2005), 91–108.

[22] R. Tibshirani and P. Wang, Spatial smoothing and hot pot detection for CGH Data using the fused Lasso, *Biostatistics*, 9 (2008), 18–29.

[23] T. Wu and K. Lange, Coordinate decent algorithms for Lasso penalized regression, *Annals of Applied Statistics*, (2008), 224–244.

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

[25] L. Wang, L1 penalized LAD estimator for high dimensional linear regression, *Journal of Multivariate Analysis*, 120 (2013), 135–151.

[26] M. Wang, L. Song and G. Tian, SCAD-penalized least absolute deviation regression in high dimensional models, *Communications in Statistics-Theory and Methods*, 44 (2015), 2452–2472.

[27] X. Wang and X. Yuan, The linearized alternating direction method of multipliers for Dantzig Selector, *Siam Journal on Scientific Computing*, 34 (2012), A2793–A2811.

[28] G. Weber, I. Batmaz, G. Köksal, P. Taylan and F. Yerlikaya-özkurt, CMARS: a new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization, *Inverse Problems in Science and Engineering*, 20 (2012), 371–400.

[29] J. Yang and X. Yuan, Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization, *Mathematics of Computation*, 82 (2013), 301–329.

[30] X. Zhang, M. Burger and S. Osher, A unified primal-dual algorithm framwork based on Bregman iteration, *Journal of Scientific Computing*, 46 (2011), 20–46.

[31] H. Zou, G. Wahba, Y. Lin, M. Voelker, M. Ferris, R. Klein and B. Klein, Variable selection and model building via likelihood basis pursuit, *Journal of the American Statistical Association*, 96 (2011), 659–672.

[32] H. Zou, T. Hastie and R. Tibshirani, Sparse principle component analysis, *Journal of Computational and Graphical Statistics*, 15 (2006), 265–286.

Received January 2017; 1st revision April 2017; final revision December 2017.

E-mail address: liuyanqing6780126.com
E-mail address: 13269338163@163.com
E-mail address: xcxiu@bjtu.edu.cn
E-mail address: jtao@loyola.edu
E-mail address: lchkong@bjtu.edu.cn