Statistica Sinica **27** (2017), 1243-1264
doi:https://doi.org/10.5705/ss.202015.0179

SPARSE AND ROBUST LINEAR REGRESSION: AN OPTIMIZATION ALGORITHM AND ITS STATISTICAL PROPERTIES

Shota Katayama and Hironori Fujisawa

Tokyo Institute of Technology and The Institute of Statistical Mathematics

Abstract: This paper studies sparse linear regression analysis with outliers in the responses. A parameter vector for modeling outliers is added to the standard linear regression model and then the sparse estimation problem for both coefficients and outliers is considered. The ℓ_1 penalty is imposed for the coefficients, while various penalties including redescending type penalties are for the outliers. To solve the sparse estimation problem, we introduce an optimization algorithm. Under some conditions, we show the algorithmic and statistical convergence property for the coefficients obtained by the algorithm. Moreover, it is shown that the algorithm can recover the true support of the coefficients with probability going to one.

Key words and phrases: Algorithmic and statistical convergence, robust estimation, sparse linear regression, support recovery.

1. Introduction

Linear regression with a large number of covariates is a general and fundamental problem in recent data analysis. A standard method to overcome this problem is the least absolute shrinkage and selection operator (Lasso) proposed by Tibshirani (1996). For a large number of covariates, it is natural to assume sparsity, which means that many of the covariates are not relevant to the responses. The Lasso can draw relevant covariates automatically and simultaneously estimate the remaining coefficients to be zero. The Lasso has been studied by, for instance, Bickel, Ritov and Tsybakov (2009), Meinshausen and Yu (2009) and Wainwright (2009) with giving the convergence rate under several norms and showing that the Lasso estimates can recover the true support of the coefficients. See also Efron et al. (2004), Zhao and Yu (2006), Zou (2006), and van de Geer and Bühlmann (2009).

Recent linear regression analysis requires some complex structures in addition to a large number of covariates. One of them is the outlier structure. That appears in such applications as signal detection, image and speech processing,

communication networks, and so on. A popular way for robustifying against outliers is to use the M-estimation procedure that replaces the ℓ_2 loss by an other loss with a bounded influence function; e.g., Huber's, the skipped-mean, and Hampel's robust loss (see, for instance, Huber and Ronchetti (2009) for more details). However, optimizing such a robust loss function with a sparse penalty requires much computational cost. Recently, She and Owen (2011) proposed a novel approach for robust parameter estimation, adding a parameter vector for modeling outliers to the standard linear regression model. The corresponding ℓ_2 loss function with a sparse penalty for the outlier parameter vector is optimized.

This paper studies sparse and robust linear regression based on their outlier model from a theoretical point of view, which was not treated in She and Owen (2011). Some theoretical analyses were discussed by Nguyen and Tran (2013), but only the ℓ_1 penalty was used for the outlier parameters. Their results were derived from the fact that the resulting estimate is a global optimum. As She and Owen (2011) showed, many penalties having good robustness are non-convex and the resulting estimate is often a local optimum. A general theory including non-convex penalties is thus of interest.

We consider a larger class of penalties for outlier parameters including non-convex penalties and derive some statistical properties. To avoid the problem of local optima, we directly analyze the estimated coefficients that an optimization algorithm outputs. We provide an upper bound for its ℓ_2 error, and show that the algorithm can recover the true support of the coefficients.

The remainder of this paper is organized as follows. We introduce the model and the optimization algorithm in Section 2. In Section 3, theoretical analyses for the estimated coefficients which the algorithm outputs are provided. We report numerical performances in Section 4. The proofs are in Section 5.

Throughout the paper, a bold symbol denotes a matrix or a vector, e.g., $\boldsymbol{A} = (a_{ij})$ for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{a} = (a_1, \ldots, a_m)^T$ for $\boldsymbol{a} \in \mathbb{R}^m$. For any vector $\boldsymbol{a} \in \mathbb{R}^m$ and $1 \leq q < \infty$, $\|\boldsymbol{a}\|_q = (\sum_{i=1}^m |a_i|^q)^{1/q}$, $\|\boldsymbol{a}\|_0 = |\{i|\, a_i \neq 0\}|$, and $\|\boldsymbol{a}\|_\infty = \max_{1\leq i\leq m} |a_i|$. Given a set $S \subset \{1, \ldots, m\}$, $\boldsymbol{a}_S = \{a_i|\ i \in S\}$. For any two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, $\langle \boldsymbol{a}, \boldsymbol{b}\rangle$ denotes the standard inner product. The matrix ℓ_1 norm is $\|\boldsymbol{A}\|_{\ell_1} = \max_{1\leq j\leq n} \sum_{i=1}^m |a_{ij}|$ and, for any symmetric matrix $\boldsymbol{B} \in \mathbb{R}^{m\times m}$, the largest eigenvalue is $\xi_{max}(\boldsymbol{B})$. For two positive sequences a_n, b_n depending on n, the notation $a_n = O(b_n)$ means that there exists a finite constant $C > 0$ such that $a_n \leq Cb_n$ for a sufficiently large n, while $a_n = \Omega(b_n)$ means that $a_n \geq Cb_n$. The notation $a_n = o(b_n)$ means that $a_n/b_n \to 0$ as n goes to infinity.

2. Sparse and Robust Linear Regression

2.1. Model and parameter estimation

Consider the linear regression model with outliers

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \sqrt{n}\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is an n dimensional response vector, $\boldsymbol{X} = (x_{ij}) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is an $n \times p$ covariate matrix, $\boldsymbol{\beta}^*$ is a p dimensional unknown coefficient vector, $\boldsymbol{\gamma}^*$ is an n dimensional unknown vector whose nonzero elements correspond to outliers and $\boldsymbol{\varepsilon}$ is an n dimensional random error vector. At (2.1), we assume that the ℓ_2 norms of columns of $\boldsymbol{X}$ are $\sqrt{n}$. Correspondingly, the coefficient of $\boldsymbol{\gamma}^*$ is assumed to be $\sqrt{n}$ to match its scale with the columns of $\boldsymbol{X}$. The model (2.1) can be found also in She and Owen (2011) and Nguyen and Tran (2013). Our purpose is to estimate $\boldsymbol{\beta}^*$ accurately even when the number of covariates is large. We suppose that $\boldsymbol{\beta}^*$ has many zero elements (sparse), and also that $\boldsymbol{\gamma}^*$ is sparse since the number of outliers is usually not large. For this setup, we introduce sparse penalties for coefficients and outliers. The parameters are estimated by solving the optimization problem

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\gamma}\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sqrt{n}\boldsymbol{\gamma}\|_2^2 + \lambda_\beta \sum_{j=1}^{p} w_{\beta,j}|\beta_j| + \sum_{i=1}^{n} P\left(\gamma_i; \frac{\lambda_\gamma w_{\gamma,i}}{\sqrt{n}}\right), \tag{2.2}$$

where $\lambda_\beta > 0$ and $\lambda_\gamma > 0$ are tuning parameters for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively, and P is a penalty function that encourages sparsity. We often use a redescending P, since it can yield a small bias against strong outliers. We consider the adaptive Lasso (Zou (2006)) type optimization problem. In (2.2), $w_{\beta,i}$ and $w_{\gamma,j}$ are known weights. Suppose that we have the preliminary estimators $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$ and $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_n)^T$. The weights are given, with constant $R_w > 0$, by

$$w_{\beta,j} = \max\left(\frac{1}{|\tilde{\beta}_j|}, \frac{1}{R_w}\right), \quad w_{\gamma,i} = \min\left(\frac{1}{|\tilde{\gamma}_i|}, R_w\right) \tag{2.3}$$

for $j \in \tilde{S}$ and $i \in \tilde{G}$, where $\tilde{S}$ and $\tilde{G}$ are the support of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$, respectively, and by $w_{\beta,j} = 1/|\tilde{\beta}_j|$ and $w_{\gamma,i} = 1/|\tilde{\gamma}_i|$ for the rest. These restrictions have $\min_{j\in\tilde{S}} w_{\beta,j} \geq R_w^{-1}$ and $\max_{i\in\tilde{G}} w_{\gamma,i} \leq R_w$. The same results hold if different constants $R_{w,\beta}$ and $R_{w,\gamma}$ are used in (2.3). For the details of the preliminary estimators used in this paper, see Section 3.3.

2.2. Optimization algorithm

Let $L(\boldsymbol{\beta}, \boldsymbol{w})$ be the objective function in (2.2). To solve (2.2), we introduce Algorithm 1. Here $L(\boldsymbol{\beta}, \boldsymbol{w})$ depends on the choice of preliminary estimators $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$

Algorithm 1

Step 1. Initialize $k \leftarrow 0$, $\boldsymbol{\beta}^k \leftarrow \boldsymbol{\beta}^{init}$ and $\boldsymbol{\gamma}^k \leftarrow \mathrm{argmin}_{\boldsymbol{\gamma}} L(\boldsymbol{\beta}^0, \boldsymbol{\gamma})$.
Step 2. Update $k \leftarrow k+1$,

$$\boldsymbol{\beta}^k \leftarrow \mathrm{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, \boldsymbol{\gamma}^{k-1}), \tag{2.4}$$

$$\boldsymbol{\gamma}^k \leftarrow \mathrm{argmin}_{\boldsymbol{\gamma}} L(\boldsymbol{\beta}^k, \boldsymbol{\gamma}). \tag{2.5}$$

Step 3. If they converge, then output current $(\boldsymbol{\beta}^k, \boldsymbol{\gamma}^k)$ and stop the algorithm, otherwise return to Step 2.

and Algorithm 1 requires an initial value $\boldsymbol{\beta}^{init}$. A typical example of $\boldsymbol{\beta}^{init}$ is $\tilde{\boldsymbol{\beta}}$. The preliminary estimators and the initial value must satisfy certain properties, later in Section 3, to ensure good behavior of the algorithm. The algorithm converges since

$$L(\boldsymbol{\beta}^0, \boldsymbol{\gamma}^0) \geq L(\boldsymbol{\beta}^1, \boldsymbol{\gamma}^0) \geq L(\boldsymbol{\beta}^1, \boldsymbol{\gamma}^1) \geq \cdots \geq 0$$

from its construction. The optimization problem at (2.4) can be solved by such as the coordinate descent algorithm (Friedman, Hastie and Tibshirani (2010)). The optimization problem in (2.5) can be rewritten as

$$\underset{\boldsymbol{\gamma} \in \mathbb{R}^n}{\mathrm{argmin}} \sum_{i=1}^{n} \left\{ \frac{1}{2} \left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}^k}{\sqrt{n}} - \gamma_i \right)^2 + P\left(\gamma_i; \frac{\lambda_\gamma w_{\gamma,i}}{\sqrt{n}} \right) \right\}.$$

If the problem $\mathrm{argmin}_x (z-x)^2/2 + P(x;\lambda)$ has the explicit solution $\Theta(z;\lambda)$ satisfying $c\Theta(z;\lambda) = \Theta(cz; c\lambda)$ for $c > 0$, then (2.5) can be written as

$$\gamma_i^k \leftarrow \frac{1}{\sqrt{n}} \Theta(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}^k; \lambda_\gamma w_{\gamma,i}), \quad i = 1, \ldots, n,$$

where γ_i^k is the ith element of $\boldsymbol{\gamma}^k$. Let this expression be denoted by $\boldsymbol{\gamma}^k \leftarrow h(\boldsymbol{\beta}^k)$. Then, step 2 can be expressed with only the update of $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}^k \leftarrow \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\mathrm{argmin}} L\big\{ \boldsymbol{\beta}, h(\boldsymbol{\beta}^{k-1}) \big\}.$$

The function $\Theta(z;\lambda)$ is often called the thresholding function. As seen in Antoniadis and Fan (2001), many sparse penalties, including the ℓ_1, ℓ_0, and smoothly clipped absolute deviation (SCAD; Fan (1997)) penalties, have an explicit solution $\Theta(z;\lambda)$. The ℓ_1 penalty leads to the soft thresholding function $\Theta(z;\lambda) = \mathrm{sgn}(z)\max(|z| - \lambda, 0)$ where $\mathrm{sgn}(\cdot)$ denotes the sign function, and the ℓ_0 penalty leads to the hard thresholding function $\Theta(z;\lambda) = zI(|z| > \lambda)$, where $I(A)$ denotes the indicator function on the event A. For the SCAD penalty, the corresponding thresholding function is given by

$$\Theta(z;\lambda)=\begin{cases}\operatorname{sgn}(z)(|z|-\lambda) & \text{if } |z|\le 2\lambda,\\ \dfrac{(a-1)z-a\lambda\operatorname{sgn}(z)}{a-2} & \text{if } 2\lambda<|z|\le a\lambda,\\ z & \text{if } |z|>a\lambda,\end{cases}$$

where $a=3.7$ is recommended by Fan and Li (2001).

2.3. Connection to robust M-estimators

As in She and Owen (2011), our procedure has a close connection to robust M-estimators.

Proposition 1. *Let $\hat{\boldsymbol{\beta}}$ be the coefficient output of Algorithm 1 and let $\psi(z;\lambda)=z-\Theta(z;\lambda)$, where $\Theta(\cdot;\lambda)$ is a thresholding function yielded from a penalty function $P(\cdot)$. Then, for $j=1,\dots,p$,*

$$\frac{1}{n}\sum_{i=1}^{n}x_{ij}\psi(y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}};\lambda_\gamma w_{\gamma,i})+\lambda_\beta w_{\beta,j}\partial|\hat{\beta}_j|=0, \tag{2.6}$$

where $\partial|\hat{\beta}_j|=\operatorname{sgn}(\hat{\beta}_j)$ if $\hat{\beta}_j\neq 0$, $\partial|\hat{\beta}_j|\in[-1,1]$ otherwise.

Proposition 1 shows that the output $\hat{\boldsymbol{\beta}}$ satisfies the estimating equations (2.6) with the ψ function and the ℓ_1 penalty. Thus, our algorithm is closely related to the optimization problem

$$\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\frac{1}{2n}\sum_{i=1}^{n}\Psi(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta};\lambda_\gamma w_{\gamma,i})+\lambda_\beta\sum_{j=1}^{p}w_{\beta,j}|\beta_j|, \tag{2.7}$$

where $(d\Psi)/(dt)(t;\lambda)=\psi(t;\lambda)$. The output $\hat{\beta}$ and any local solution to (2.7) share the same first-order KKT condition. It may take much computation to directly solve (2.7), particularly when $\Psi(\cdot;\lambda)$ is non-convex. For fast computation, we can use the first-order approximation of $\Psi(\cdot;\lambda)$, but it loses the information of the original function. Thus, we use Algorithm 1 instead of solving (2.7).

The relationship between sparse penalties and robust loss functions is stated through the equation $\psi(z;\lambda)=z-\Theta(z;\lambda)$. For instance, the ℓ_1 penalty corresponds to the Huber's loss, the ℓ_0 penalty to the skipped-mean loss, and the SCAD penalty to the Hampel's loss. She and Owen (2011) gave illustrations and more details. One measure to characterize a robust loss function is the redescending property; $|z-\Theta(z;\lambda)|$ goes to 0 as $|z|$ goes infinity. The soft thresholding function that corresponds to the ℓ_1 penalty does not have the redescending property, but the hard and SCAD thresholding functions that correspond to non-convex penalties have it. Some other non-convex penalties, including the non-negative garrote penalty (Garrote; Gao (1998)) and the minimax concave penalty (MCP;

Zhang (2010)), also lead to that property.

3. Theoretical Analysis

The optimization problem (2.2) with a non-convex penalty P suffers from local minima, so we directly analyze the output of Algorithm 1. Some analyses of computable solutions in linear regression model without outliers were provided by Zhang and Zhang (2012), Fan and Lv (2013), and Fan and Lv (2014), but our analysis is essentially different from theirs. They first derived some properties for a global minimum and then showed that a computable solution shared those properties under additional conditions on the solution.

3.1. Notations

Let $S^* = \mathrm{supp}(\boldsymbol{\beta}^*) = \{i | \, \beta_i^* \neq 0\} \subset \{1, \ldots, p\}$ and $G^* = \mathrm{supp}(\boldsymbol{\gamma}^*) \subset \{1, \ldots, n\}$. These support sizes are written $s^* = |S^*|$ and $g^* = |G^*|$. For preliminary estimators, let $\tilde{S} = \mathrm{supp}(\tilde{\boldsymbol{\beta}})$ and $\tilde{G} = \mathrm{supp}(\tilde{\boldsymbol{\gamma}})$ with sizes $\tilde{s} = |\tilde{S}|$ and $\tilde{g} = |\tilde{G}|$. We define a restricted smallest eigenvalue δ_{min} as

$$\delta_{min}(u) = \inf_{\|\boldsymbol{\delta}\|_0 \leq u} \frac{\|\boldsymbol{X}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}\|_2^2} \tag{3.1}$$

and a (doubly) restricted largest eigenvalue $\delta_{\max}$ as

$$\delta_{max}(u, u') = \sup_{\|\boldsymbol{\delta}\|_0 \leq u} \sup_{|G| \leq u'} \frac{\|\boldsymbol{X}_{(G)}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}\|_2^2}, \tag{3.2}$$

where $\boldsymbol{X}_{(G)} = \{x_{ij} | \; i \in G, 1 \leq j \leq p\}$. The restricted eigenvalue is popular in the analysis of the Lasso (see, e.g., Bickel, Ritov and Tsybakov (2009)), but our δ_{max} is slightly different from the existing one. The restriction is imposed on rows of $\boldsymbol{X}$ in addition to columns, corresponding to the outliers. We provide an asymptotic analysis as n goes to infinity. In our theory, p, s^* and g^* may go to infinity depending on n. Notice that the restricted eigenvalues also depend on n.

3.2. Properties of Algorithm 1 output

To derive a convergence property of Algorithm 1, we require conditions on the random error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, the thresholding function $\Theta(\cdot; \lambda)$, and the preliminary estimator $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$.

Condition 1. The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independently and identically distributed as a zero mean sub-Gaussian distribution with a parameter $\sigma > 0$; $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}\{\exp(t\varepsilon_i)\} \leq \exp(t^2\sigma^2/2)$ for all $t \in \mathbb{R}$.

Condition 2. The thresholding function $\Theta(\cdot;\lambda)$ satisfies $\Theta(x;\lambda)=0$ if $|x|\le\lambda$ and $|\Theta(x;\lambda)-x|\le\lambda$ for all $x\in\mathbb{R}$.

Condition 3. There exist a sequence $a_{n,1}\to 0$ and a constant $\kappa>0$ such that $\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2+\|\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2\le \tilde{C}a_{n,1}$ and $\delta_{min}(\tilde{s})\ge\kappa$ with probability going to one, where $\tilde{C}$ is some positive constant.

The sub-Gaussian distribution family, see Vershynin (2012), covers the Gaussian, Bernoulli, and any distribution with a bounded support. Condition 2 decides a class of thresholding functions that include the soft, hard, non-negative garrote, SCAD, and MCP thresholding functions. The first part of Condition 3 implies consistent preliminary estimators at the rate $a_{n,1}$. We find $a_{n,1}^2=\{(s^*+g^*)\log\max(n,p)\}/n$ when we use the preliminary estimators in Section 3.3. Then we have $|\tilde{\beta}_j|\ge|\beta_j^*|-|\tilde{\beta}_j-\beta_j^*|\ge|\beta_j^*|-\tilde{C}a_{n,1}$ for $j\in S^*$ by the triangle inequality. Hence, if $\min_{j\in S^*}|\beta_j^*|>\tilde{C}a_{n,1}$, then $|\tilde{\beta}_j|>0$ for $j\in S^*$. Similarly, we can show that if $\min_{i\in G^*}|\gamma_i^*|>\tilde{C}a_{n,1}$ then $|\tilde{\gamma}_i|>0$ for $i\in G^*$. Thus, Condition 3 leads to the screening property that $S^*\subset\tilde{S}$ and $G^*\subset\tilde{G}$ if

$$\min\Big\{\min_{j\in S^*}|\beta_j^*|,\min_{i\in G^*}|\gamma_i^*|\Big\}>\tilde{C}a_{n,1}. \tag{3.3}$$

The second part of Condition 3 is the sparse Riesz condition (Zhang and Huang (2008)) if $\tilde{s}$ is non-random. When we use a preliminary estimator that has a non-random upper bound s_u, it reduces to $\delta_{min}(s_u)\ge\kappa$. In Section 3.3, we clarify the order of $a_{n,1}$ and the non-random upper bound of $\tilde{s}$ for some preliminary estimators.

Theorem 1. *Assume Conditions* 1–3 *and* (3.3). *Let*

$$\lambda_\beta\ge 2CR_w\sqrt{\frac{\sigma^2\log p}{n}},\quad \lambda_\gamma\le\frac{Cn}{R_w\max_{j\in\tilde{S}}\sum_{i\in\tilde{G}}|x_{ij}|}\sqrt{\frac{\sigma^2\log p}{n}}$$

for a constant $C>\sqrt{2}$ *and* $R_w>0$ *used in* (2.3). *Then, for any iteration* $k\ge 1$ *and any initial value* $\boldsymbol{\beta}^{init}$ *in Algorithm 1,*

$$\|\boldsymbol{\beta}^k-\boldsymbol{\beta}^*\|_2\le\rho^k\|\boldsymbol{\beta}^{init}-\boldsymbol{\beta}^*\|_2+2\kappa^{-1}\sqrt{s^*}\lambda_\beta\big(R_w^{-1}+\max_{j\in S^*}w_{\beta,j}\big)\sum_{i=0}^{k-1}\rho^i, \tag{3.4}$$

with probability going to one, where $\rho=2\kappa^{-1}\delta_{\max}(\tilde{s},\tilde{g})$.

The first term of the right side of (3.4) can be regarded as the algorithmic error. It represents the effect of an initial value $\boldsymbol{\beta}^{init}$ in Algorithm 1. This shows that if $\rho<1$ then the effect vanishes from the bound exponentially as the number of iterations increases. Since $|\xi_{max}(\boldsymbol{M})|\le\|\boldsymbol{M}\|_{\ell_1}$ for any symmetric matrix $\boldsymbol{M}$,

we have

$$\delta_{max}(\tilde{s},\tilde{g}) \leq \max_{1\leq i\leq n; 1\leq j\leq p} x_{ij}^2 \frac{\tilde{s}\tilde{g}}{n}. \tag{3.5}$$

Thus, if $\max_{ij} x_{ij}^2 a_{n,2}^2 = o(n)$, we have $\rho < 1$ for sufficiently large n with probability going to one, where the rate $a_{n,2}$ is to be found in Condition 4. When we use the preliminary estimators defined in Section 3.3, $a_{n,2} = \xi_{max}(\boldsymbol{Z}^T\boldsymbol{Z}/n)(s^* + g^*)$ or $s^* + g^*$ where $\boldsymbol{Z} = (\boldsymbol{X}, \sqrt{n}\boldsymbol{I}_n)$. Thus, Algorithm 1 can remove the effect from $\boldsymbol{\beta}^{init}$ if $\max_{ij} x_{ij}^2[\xi_{max}(\boldsymbol{Z}^T\boldsymbol{Z}/n)(s^* + g^*)]^2 = o(n)$ or $\max_{ij} x_{ij}^2(s^* + g^*)^2 = o(n)$. Thus, we need an n sufficiently larger than $(s^* + g^*)^2$ to hold $\rho < 1$. The second term can be regarded as the statistical error. The rate $a_{n,1}$ in Condition 3 affects the rate of $\max_{j\in S^*} w_{\beta,j}$ and $\max_{j\in S^*} w_{\beta,j} = O(1)$ with probability going to one if $a_{n,1} = o(1)$ and $\min_{j\in S^*} |\beta_j^*| = \Omega(1)$. In this case, the second term has $O(\sqrt{s^*}\lambda_\beta)$, which is equivalent to the order of the standard Lasso excluding the term $\sqrt{n}\boldsymbol{\gamma}^*$ from (2.1) in advance.

Theorem 1 shows only the convergence result for the output. Next, we shall show that the output can recover the true support. We need an extra condition and a corollary.

Condition 4. There exist a constant $C > 0$ and a sequence $a_{n,2}$, which may diverge, such that $\max(\tilde{s},\tilde{g}) \leq Ca_{n,2}$ with probability going to one.

Corollary 1. Assume Conditions 1–4, (3.3), $\max_{ij} |x_{ij}| = O(1)$, $\min_{j\in S^*} |\beta_j^*| = \Omega(1)$, and $a_{n,2}^2 = o(n)$. Let $\lambda_\beta \geq C_\beta\{(\log p)/n)\}^{1/2}$ and $\lambda_\gamma \leq C_\gamma(n/a_{n,2})\{(\log p)/n\}^{1/2}$ with some constants $C_\beta > 0$ and $C_\gamma > 0$. If there exist some $k_0 \geq 1$ and $C_0 > 0$ such that $\mathbb{P}(\|\boldsymbol{\beta}^{init} - \boldsymbol{\beta}^*\|_2 \leq C_0\rho^{-k_0}\sqrt{s^*}\lambda_\beta) \to 1$, then it follows that $\mathbb{P}(\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^*\|_2 \leq C\sqrt{s^*}\lambda_\beta) \to 1$ for any $k \geq k_0$ for some constant $C > 0$.

When $\max_{i,j} |x_{ij}| = O(1)$ and $a_{n,2}^2 = o(n)$, we have $\rho = o(1)$. Consequently, Corollary 1 can be obtained from Theorem 1. The requirement $\mathbb{P}(\|\boldsymbol{\beta}^{init} - \boldsymbol{\beta}^*\|_2 \leq C_0\rho^{-k_0}\sqrt{s^*}\lambda_\beta) \to 1$ is not so restrictive since $\rho = o(1)$. If we use $\tilde{\boldsymbol{\beta}}$ for the initial value $\boldsymbol{\beta}^{init}$, then the condition $(a_{n,2}^2/n)^{k_0}a_{n,1} = O(\sqrt{s^*}\lambda_\beta)$ is required. From Corollary 1, we can show a theorem about support recovery. For this we only need to analyze the elements of $\boldsymbol{\beta}^k$ on $\tilde{S}\cap S^{*c}$ since the screening property $S^* \subset \tilde{S}$ holds.

Theorem 2. *Assume the conditions of Corollary* 1. *If* $\sqrt{\log n}$=$o(\sqrt{n}\min_{i\in G^*} |\gamma_i^*|)$, $a_{n,2}s^*\log p = o(n\log n)$, *and* $a_{n,1}\max\left(\sqrt{a_{n,2}s^*}, g^*/\sqrt{n}\right) = o(1)$, *then we have* $\mathbb{P}\left(\boldsymbol{\beta}_{\tilde{S}\cap S^{*c}}^k = \boldsymbol{0}\right) \to 1$ *for any* $k \geq k_0 + 1$, *where* k_0 *is as in Corollary* 1.

The proof is given in Section 5.3. As seen in the proof, an explicit convergence rate for $\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^*\|_2$ is required to derive the conditions involving (s^*, g^*) in

Theorem 2. This means that the rate need not be $\sqrt{s^*}\lambda_\beta$, and hence the condition $\min_{j\in S^*}|\beta_j^*| = \Omega(1)$ can be weakened. Instead, the conditions involving (s^*, g^*) become stronger. There exists a trade-off between the conditions for $\min_{j\in S^*}|\beta_j^*|$ and (s^*, g^*). To keep conditions simple, we only report the case where the rate $\sqrt{s^*}\lambda_\beta$ is attained.

3.3. Preliminary estimator and its properties

In this section, we give an example, and specify the orders $a_{n,1}$ and $a_{n,2}$ in Conditions 3 and 4. Consider the Lasso type preliminary estimators of $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}^T, \tilde{\boldsymbol{\gamma}}^T)^T$, given by

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^{np}}{\operatorname{argmin}} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta}\|_2^2 + \lambda_\theta\|\boldsymbol{\theta}\|_1, \tag{3.6}$$

where $\lambda_\theta > 0$ is the tuning parameter and $\boldsymbol{Z} = (\boldsymbol{X}, \sqrt{n}\boldsymbol{I}_n)$. The estimator $\tilde{\boldsymbol{\theta}}$ is a simple variant of that proposed in Nguyen and Tran (2013). Using different tuning parameters for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ may improve the accuracy of estimates. However, even if the preliminary estimates do not have the high accuracy, we can improve accuracy by calculating (2.2) with different tuning parameters. For this reason, it would be enough to use the single tuning parameter in (3.6).

After Bickel, Ritov and Tsybakov (2009) and Wainwright (2009), for example, the following can be shown.

Proposition 2. *Assume Condition 1 and that there exists a constant $\tilde{\kappa} > 0$ such that*

$$\min_{\boldsymbol{\theta}\neq\boldsymbol{0};\|\boldsymbol{\theta}_{U^{*c}}\|_1\leq 3\|\boldsymbol{\theta}_{U^*}\|_1} \frac{\|\boldsymbol{Z}\boldsymbol{\theta}\|_2^2}{n\|\boldsymbol{\theta}\|_2^2} \geq \tilde{\kappa}, \tag{3.7}$$

where $U^* = \operatorname{supp}(\boldsymbol{\theta}^*)$ with $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*T}, \boldsymbol{\gamma}^{*T})^T$. Let $\lambda_\theta = C_\theta\{(\log\max(n,p))/n\}^{1/2}$ for sufficiently large $C_\theta > 0$. Then, it follows that

$$\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq C\sqrt{\frac{(s^*+g^*)\log\max(n,p)}{n}}, \tag{3.8}$$

$$|\operatorname{supp}(\tilde{\boldsymbol{\theta}})| \leq C\xi_{max}(\frac{\boldsymbol{Z}^T\boldsymbol{Z}}{n})(s^*+g^*), \tag{3.9}$$

with probability going to one, where $C > 0$ is some constant.

The condition (3.7) is slightly different from that commonly used in Lasso analysis. It is involved in the extended matrix $\boldsymbol{Z}$ not in $\boldsymbol{X}$. For more details, see Nguyen and Tran (2013). The bounds (3.8) and (3.9) correspond to the orders $a_{n,1}$ and $a_{n,2}$ of Condition 3 and 4, respectively. However, the term $\xi_{max}(\boldsymbol{Z}^T\boldsymbol{Z}/n)$ is troublesome since it may diverge as n or p increases. To exclude it, after $\tilde{\boldsymbol{\theta}}$ is

obtained with λ_θ, we consider the threshold version of (3.6) with an additional tuning parameter $\tau_\theta > 0$ as

$$\tilde{\theta}_j^{th} = \tilde{\theta}_j I(|\tilde{\theta}_j| > \tau_\theta \lambda_\theta), \quad j = 1, \ldots, n+p. \tag{3.10}$$

Now, let $\tilde{\boldsymbol{\theta}}^{th} = (\tilde{\theta}_1^{th}, \ldots, \tilde{\theta}_{n+p}^{th})^T$ and $a_{n,1} = \{(s^*+g^*)\log\max(n,p)/n)\}^{1/2}$. Under (3.3), in which $\tilde{C}$ is replaced by $2\tilde{C}$, we have $|\tilde{\theta}_j| > \tilde{C}a_{n,1}$ for any $j \in \mathrm{supp}(\boldsymbol{\theta}^*)$. Thus, if we select τ_θ such that $\tau_\theta\lambda_\theta \leq \tilde{C}a_{n,1}$, then $\tilde{\theta}_j^{th} = \tilde{\theta}_j$ for any $j \in \mathrm{supp}(\boldsymbol{\theta}^*)$. Hence, the threshold version also satisfies (3.8) with the same order as the original. Meanwhile, note that

$$\begin{aligned}|\mathrm{supp}(\tilde{\boldsymbol{\theta}}^{th})\backslash\mathrm{supp}(\boldsymbol{\theta}^*)| &= \sum_{j\in\mathrm{supp}(\tilde{\boldsymbol{\theta}}^{th})\backslash\mathrm{supp}(\boldsymbol{\theta}^*)} 1 \leq \sum_{j\notin\mathrm{supp}(\boldsymbol{\theta}^*)} \frac{\tilde{\theta}_j^2}{\tau_\theta^2\lambda_\theta^2} \\ &\leq \frac{\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2}{\tau_\theta^2\lambda_\theta^2} \leq C(s^*+g^*),\end{aligned}$$

if we select $\tau_\theta \geq C_\tau$ for sufficiently small $C_\tau > 0$, which implies that $|\mathrm{supp}(\tilde{\boldsymbol{\theta}}^{th})| \leq (1+C)(s^*+g^*)$. Here, the term $\xi_{max}(\boldsymbol{Z}^T\boldsymbol{Z}/n)$ is excluded. The conditions $\tau_\theta\lambda_\theta \leq \tilde{C}a_{n,1}$ and $\tau_\theta \geq C_\tau$ are compatible since the order of λ_θ is smaller than or equal to that of $a_{n,1}$.

If we further assume the mutual incoherence in Nguyen and Tran (2013), the support recovery for (3.6) and (3.10) can be obtained. However, we do not require it when we use them as preliminary estimators for the proposed method. We only need the convergence rate and the upper bound of the support size. Moreover, we can show the support recovery of the proposed method without the mutual incoherence as seen in Theorem 2.

4. Numerical Performance

We examined numerical performances of our procedure based on 100 Monte Carlo simulations. The tuning parameters λ_β, λ_γ, λ_θ, and τ_θ were selected by the Bayesian information criteria (BIC; Schwarz (1978)). For instance, consider the selection of $(\lambda_\beta, \lambda_\gamma)$ by BIC. Let $\hat{\boldsymbol{\beta}}(\lambda_\beta, \lambda_\gamma)$ and $\hat{\boldsymbol{\gamma}}(\lambda_\beta, \lambda_\gamma)$ be the outputs of Algorithm 1 with the tuning parameters $(\lambda_\beta, \lambda_\gamma)$. Then, the optimal tuning parameters $(\hat{\lambda}_\beta, \hat{\lambda}_\gamma)$ are given by

$$\begin{aligned}(\hat{\lambda}_\beta, \hat{\lambda}_\gamma) = \underset{\lambda_\beta>0;\lambda_\gamma>0}{\mathrm{argmin}} &\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_\beta,\lambda_\gamma) - \sqrt{n}\hat{\boldsymbol{\gamma}}(\lambda_\beta,\lambda_\gamma)\|_2^2 \\ &+ \frac{\log n}{n}\big\{|\mathrm{supp}(\hat{\boldsymbol{\beta}}(\lambda_\beta,\lambda_\gamma))| + |\mathrm{supp}(\hat{\boldsymbol{\gamma}}(\lambda_\beta,\lambda_\gamma))|\big\}.\end{aligned}$$

Practically, since it is impossible to search all possible tuning parameters, we

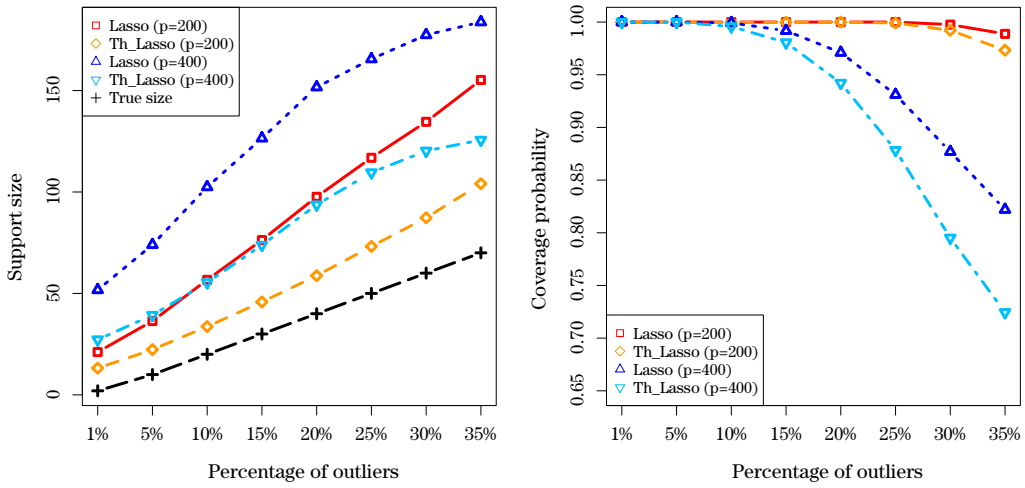

Figure 1. The support size (left) and the coverage probability (right) for the two preliminary estimators. "Lasso" means (3.6) and "Th_Lasso" means (3.10). Each point on the curve shows the mean based on 100 Monte Carlo simulations.

searched them over some candidate values which were generated as in Friedman, Hastie and Tibshirani (2010).

The first simulations were designed to see the impact of the number of outliers. The scenarios $(n, p, s^*) = (200, 200, 10)$ (moderate dimension) and $(n, p, s^*) = (200, 400, 20)$ (high dimension) with various g^* were considered. The covariates $\boldsymbol{x}_i$'s were independently drawn from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $(\boldsymbol{\Sigma})_{ij} = 0.3^{|i-j|}$, the true coefficients were given by $\beta_j^* = \operatorname{sgn}(u_j)$, and the true outliers were given by $\sqrt{n}\gamma_i^* = 8$. The positions of the non-zero coefficients and outliers were uniformly drawn from $\{1, \ldots, p\}$ and $\{1, \ldots, n\}$, respectively. The ε_i's and u_j's were independently drawn from $N(0, 1)$. We used $R_w = 100$ in (2.3) and the preliminary estimator for $\boldsymbol{\beta}^{init}$ in Algorithm 1. The algorithm stopped when $\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}\|_1/\tilde{s} \le 10^{-3}$ was satisfied at the iteration k.

Figure 1 shows the support size $\tilde{s} + \tilde{g}$ (left) and the coverage probability $\mathbb{P}(S^* \subset \tilde{S}, G^* \subset \tilde{G})$ (right) for the two preliminary estimators (3.6) and (3.10) when the percentage of outliers increased from 1% to 35%. It can be seen that the two preliminary estimators performed well if the percentage of outliers was lower than around 20%. The threshold version (3.10) had a smaller support size than (3.6), but its coverage probability was worse. We also notice that the coverage probability tended to be low as the percentage of outliers increased. It would come from the violation of the condition (3.3). As the number of outliers increases, the order $a_{n,1}$ increases.

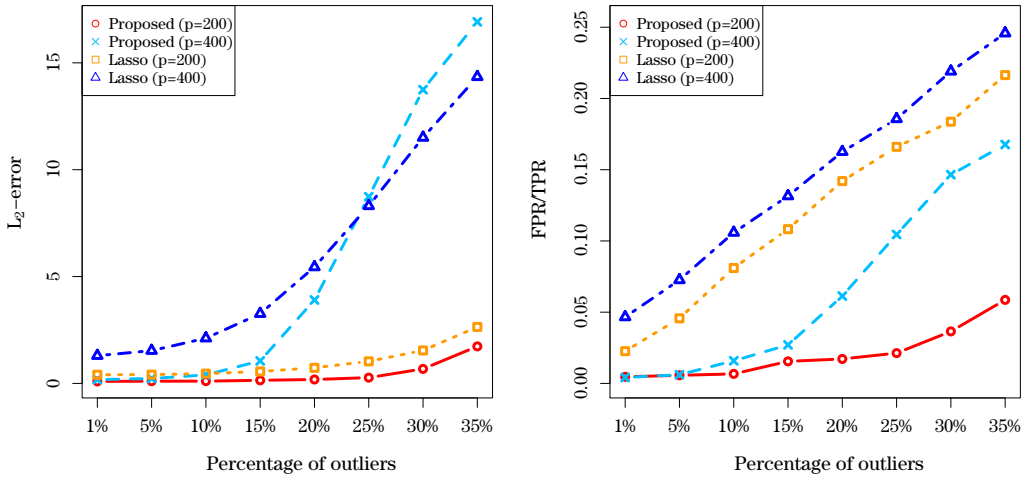

Figure 2. The squared ℓ_2-error (left) and the false positive rate divided by the true positive rate (right) for the proposed estimator with (3.6) and $P(\cdot) = |\cdot|$. The performances of (3.6) are provided for comparison.

Figure 2 shows the performance of the proposed method with the preliminary estimator (3.6) and the ℓ_1 penalty $P(x;\lambda) = \lambda|x|$ as the number of outliers increases. We omitted the other cases from the figure since they had similar behaviors. Their concrete values for more precise investigation were reported in Tables 1–4. The left figure shows the squared ℓ_2-error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$ and the right shows the false positive rate divided by the true positive rate (FPR/TPR) given by $(|\{j|\beta_j^* = 0, \hat{\beta}_j \neq 0\}|/|\{j|\beta_j^* = 0\}|)/(|\{j|\beta_j^* \neq 0, \hat{\beta}_j \neq 0\}|/\{j|\beta_j^* \neq 0\})$. This index is in $[0,1]$ and 0 is the best. In Tables 1–4, we show the squared ℓ_2-error, the number of the false positives $|\{j|\beta_j^* = 0, \hat{\beta}_j \neq 0\}|$ (FP), and the number of the true positives $|\{j|\beta_j^* \neq 0, \hat{\beta}_j \neq 0\}|$ (TP) for various penalties for outlier parameters. We reported the performance with the "Soft", "Hard", "SCAD", and "Garotte" thresholding functions. Only the Soft does not have the redescending property. The Garrote has a different behavior from the Hard and the SCAD, in fact, its $\psi(z;\lambda)$ function never vanishes if z is finite. For comparison, we also investigated the performances of the standard "Lasso" and its "Oracle" version where the true outliers are excluded in advance. The symbol "-" means that the standard Lasso does not depend on preliminary estimators and its oracle version does not so on outliers.

From Figure 2 and Tables 1–2, for moderate dimensions, our procedure provided quite good estimates and recovered the true support well, although the performance at 30% and more outliers was marginal. Interestingly, the perfor-

Table 1. Numerical performances of the proposed procedure for various outlier percentages when $(n, p, s^*) = (200, 200, 10)$ and (3.6) is used for the preliminary estimators. Each value shows the mean (standard deviation) based on 100 Monte Carlo simulations.

Outlier	Criterion	Soft	Hard	SCAD	Garrote	Lasso	Oracle
5%	ℓ_2-error	0.1036	0.1008	0.1026	0.1021	1.7400	0.3514
		(0.0504)	(0.0471)	(0.0483)	(0.0491)	(0.7256)	(0.1198)
	FP	1.08	0.88	0.97	0.95	66.86	5.17
		(1.55)	(1.50)	(1.58)	(1.54)	(18.58)	(2.74)
	TP	10.00	10.00	10.00	10.00	10.00	10.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
10%	ℓ_2-error	0.1059	0.1058	0.0992	0.1014	9.4954	-
		(0.0456)	(0.0480)	(0.0417)	(0.0424)	(3.9652)	
	FP	1.27	1.09	1.07	1.11	124.53	-
		(1.50)	(1.39)	(1.24)	(1.26)	(20.47)	
	TP	10.00	10.00	10.00	10.00	9.97	-
		(0.00)	(0.00)	(0.00)	(0.00)	(0.17)	
20%	ℓ_2-error	0.1828	0.1436	0.1377	0.1485	32.0149	-
		(0.1192)	(0.0752)	(0.0693)	(0.0807)	(6.5770)	
	FP	3.26	2.38	2.62	2.74	153.42	-
		(1.50)	(1.39)	(1.24)	(1.26)	(20.47)	
	TP	10.00	10.00	10.00	10.00	9.76	-
		(0.00)	(0.00)	(0.00)	(0.00)	(0.45)	
30%	ℓ_2-error	0.6788	0.4886	0.4683	0.5107	52.2651	-
		(0.8848)	(0.7465)	(0.7398)	(0.7833)	(9.1842)	
	FP	6.77	5.04	5.33	5.68	157.06	-
		(7.07)	(6.12)	(6.38)	(6.36)	(5.84)	
	TP	9.87	9.89	9.89	9.89	9.53	-
		(0.39)	(0.35)	(0.35)	(0.35)	(0.69)	

mance was better than the Oracle. This would be because the error $\boldsymbol{\varepsilon}$ can yield extreme values in simulations and the Oracle is not robust against these values. As seen in Figure 2 and Tables 3–4, however, for high dimensions and large outlier percentages, our procedure did poorly. For high-dimensional data, our procedure would perform well only when the outlier percentage is low (up to 15% in this case). In particular, Figure 2 shows that, for $p = 400$, the proposed method was poor at around 20% outliers, and at around 30% for $p = 200$. As the number of outliers increases, the value of $s^* + g^*$ increases. This result would come from violating the conditions in Section 3 relating to $s^* + g^*$. Comparing preliminary estimators, (3.10) performed better than (3.6) for the true support recovery, but the opposite was true for the ℓ_2-error. The Soft performed worse than the other thresholding functions. This would be explained by the redescending property.

The final simulations were designed to investigate the impact of the magni-

Table 2. The same simulations as those of Table 1 when (3.10) is used for the preliminary estimators. The results of the standard Lasso and its oracle version are omitted since they are equivalent to Table 1.

Pre.	Outlier	Criterion	Soft	Hard	SCAD	Garrote
(3.10)	5%	ℓ_2-error	0.1067	0.1060	0.1059	0.1061
			(0.0473)	(0.0474)	(0.0474)	(0.0477)
		FP	1.05	1.07	1.07	1.07
			(1.26)	(1.26)	(1.26)	(1.26)
		TP	10.00	10.00	10.00	10.00
			(0.00)	(0.00)	(0.00)	(0.00)
	10%	ℓ_2-error	0.1097	0.1080	0.1084	0.1083
			(0.0466)	(0.0476)	(0.0451)	(0.0447)
		FP	1.18	1.09	1.10	1.10
			(1.37)	(1.33)	(1.36)	(1.36)
		TP	10.00	10.00	10.00	10.00
			(0.00)	(0.00)	(0.00)	(0.00)
	20%	ℓ_2-error	0.1654	0.1501	0.1487	0.1541
			(0.0800)	(0.0675)	(0.0646)	(0.0721)
		FP	1.14	1.11	1.06	1.11
			(1.40)	(1.66)	(1.47)	(1.57)
		TP	10.00	10.00	10.00	10.00
			(0.00)	(0.00)	(0.00)	(0.00)
	30%	ℓ_2-error	0.6583	0.5377	0.5502	0.5788
			(0.9306)	(0.8802)	(0.9146)	(0.9216)
		FP	3.66	2.91	2.87	3.20
			(5.21)	(4.39)	(4.38)	(4.83)
		TP	9.84	9.84	9.84	9.84
			(0.42)	(0.42)	(0.42)	(0.42)

tude of outliers. We used $(n, p, s^*) = (200, 400, 20)$ with $g^* = 20$ (10% outliers) and various magnitudes of $\sqrt{n}\boldsymbol{\gamma}^*$. We considered the situations $\sqrt{n}\boldsymbol{\gamma}^* = \gamma^* \mathbf{1}_n$ with $\gamma^* \in \{2, 4, 6, \ldots, 14\}$. The Monte Carlo samples were generated as before. In Figure 3, only the performances of the Soft and Hard are shown. The SCAD and Garotte performed similarly to the Hard. The true positives are also omitted since they were around 20 for all the cases considered. Our procedure performed well with low and high magnitudes, but it did not do so with a moderate magnitude. This also would be come from the violation of the condition (3.3). For a low magnitude, the outliers would be hidden by the random errors. When $\boldsymbol{\varepsilon}$ is drawn from $N_n(\mathbf{0}, \boldsymbol{I}_n)$, the maximum magnitude of ε_i's is less than $\sqrt{2 \log n}$ (it is around 3.3 in this case). We also note that the performances tended to be stable as the magnitude increased.

Table 3. Numerical performances of the proposed procedure for various outlier percentages when $(n, p, s^*) = (200, 400, 20)$ and (3.6) is used for the preliminary estimators. Each value shows the mean (standard deviation) based on 100 Monte Carlo simulations.

Outlier	Criterion	Soft	Hard	SCAD	Garrote	Lasso	Oracle
5%	ℓ_2-error	0.2287	0.2139	0.2041	0.2087	3.8340	1.0337
		(0.1129)	(0.0874)	(0.0870)	(0.0944)	(0.8725)	(0.3036)
	FP	2.32	2.05	2.04	2.09	97.95	21.69
		(2.82)	(2.81)	(2.82)	(2.76)	(16.26)	(8.09)
	TP	20.00	20.00	20.00	20.00	19.98	20.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.00)
10%	ℓ_2-error	0.3987	0.3096	0.3039	0.3218	8.5953	-
		(0.3029)	(0.1970)	(0.1978)	(0.2071)	(1.5790)	
	FP	6.03	5.02	5.33	5.30	134.80	-
		(7.61)	(7.06)	(7.14)	(7.20)	(13.70)	
	TP	20.00	20.00	20.00	20.00	19.87	-
		(0.00)	(0.00)	(0.00)	(0.00)	(0.34)	
20%	ℓ_2-error	3.9041	3.5611	3.5540	3.6990	19.5640	-
		(3.4908)	(3.5130)	(3.5647)	(3.5669)	(2.9656)	
	FP	21.51	18.74	19.77	20.44	158.11	-
		(13.56)	(13.69)	(14.38)	(14.60)	(11.62)	
	TP	19.01	18.99	19.00	19.09	18.79	-
		(1.31)	(1.37)	(1.35)	(1.22)	(1.01)	
30%	ℓ_2-error	13.7451	13.0448	13.7788	13.9348	30.6065	-
		(3.9886)	(3.6637)	(4.1995)	(3.9989)	(4.2319)	
	FP	43.13	40.86	42.78	43.41	169.16	-
		(11.82)	(10.72)	(11.44)	(11.00)	(11.72)	
	TP	15.85	15.81	15.90	15.82	17.52	-
		(1.94)	(1.98)	(1.95)	(1.96)	(1.31)	

5. Proofs

5.1. Proof of Proposition 1

Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ be the output of Algorithm 1. Then, from (2.4), we have for $j = 1, \ldots, p$,

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \sqrt{n}\hat{\gamma}_i) + \lambda_{\beta,j}\partial|\hat{\beta}_j| = 0.$$

Since $\psi(z;\lambda) = z - \Theta(z;\lambda)$, it follows from (2.5) that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \sqrt{n}\hat{\gamma}_i) &= \frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \Theta(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}; \lambda_\gamma w_{\gamma,i})) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_{ij}\psi(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}; \lambda_\gamma w_{\gamma,i}),\end{aligned}$$

Table 4. The same simulations as those of Table 3 when (3.10) is used for the preliminary estimators. The results of the standard Lasso and its oracle version are omitted since they are equivalent to Table 3.

Outlier	Criterion	Soft	Hard	SCAD	Garrote
5%	ℓ_2-error	0.2502	0.2280	0.2232	0.2293
		(0.1184)	(0.1079)	(0.1027)	(0.1061)
	FP	2.17	2.06	1.98	2.10
		(2.21)	(2.26)	(2.26)	(2.33)
	TP	20.00	20.00	20.00	20.00
		(0.00)	(0.00)	(0.00)	(0.00)
10%	ℓ_2-error	0.4013	0.3253	0.3137	0.3345
		(0.3243)	(0.2805)	(0.2741)	(0.2899)
	FP	2.91	2.43	2.57	2.56
		(3.28)	(2.72)	(2.81)	(2.96)
	TP	19.99	19.99	19.99	19.99
		(0.10)	(0.10)	(0.10)	(0.10)
20%	ℓ_2-error	4.1649	4.0372	3.9793	4.0969
		(3.6045)	(3.7870)	(3.7536)	(3.7765)
	FP	19.78	19.04	18.81	19.04
		(13.83)	(14.01)	(13.69)	(13.58)
	TP	18.99	19.01	19.00	19.00
		(1.33)	(1.28)	(1.33)	(1.23)
30%	ℓ_2-error	13.2793	13.9873	13.7711	14.3158
		(3.6615)	(3.8248)	(3.8830)	(3.7659)
	FP	41.10	41.02	41.00	41.39
		(7.71)	(7.99)	(7.47)	(7.37)
	TP	16.02	16.03	15.99	16.00
		(1.89)	(1.90)	(1.88)	(1.92)

which concludes the proof.

5.2. Proof of Theorem 1

To prove the theorem we need a lemma. The proof is omitted since it is now standard in Lasso analysis. See, e.g., Negahban et al. (2012).

Lemma 1. *Let $Z_1, \ldots, Z_n$ be independently identically distributed as a zero mean sub-Gaussian distribution with a parameter $\sigma > 0$. Then, for any vector $\boldsymbol{a} \in \mathbb{R}^n$ and any $t \geq 0$,*

$$\mathbb{P}\Big(\Big| \sum_{i=1}^{n} a_i Z_i \Big| > t \Big) \leq 2 \exp\Big(- \frac{t^2}{2\sigma^2 \|\boldsymbol{a}\|_2^2} \Big).$$

Since we consider the adaptive Lasso type estimator and the screening property that $S^* \subset \tilde{S}$ and $G^* \subset \tilde{G}$ is satisfied under Condition 3 and (3.3), it suffices

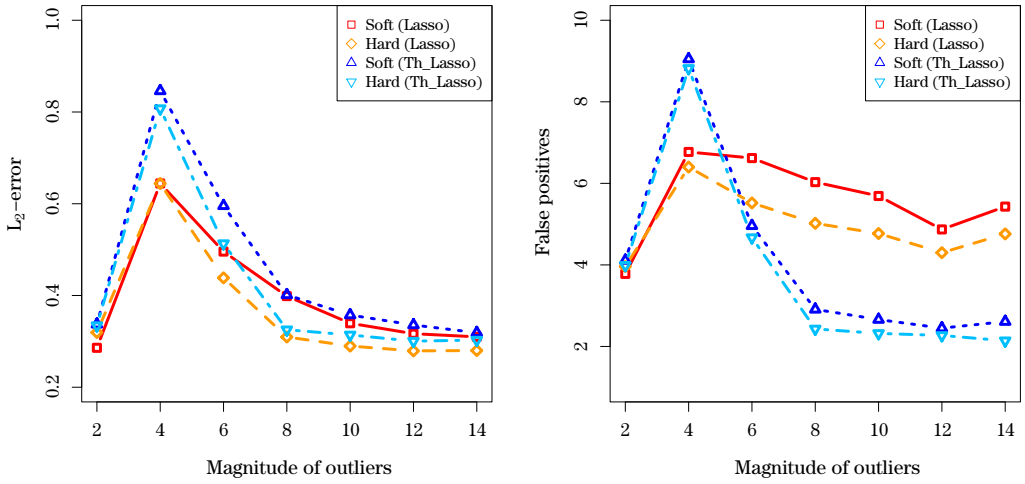

Figure 3. The squared ℓ_2-error (left) and the false positives (right) for various magnitudes of outliers. Each point on the curve shows the mean based on 100 Monte Carlo simulations.

to focus only on the covariates selected by the preliminary estimator $\tilde{\boldsymbol{\beta}}$, the submatrix $\boldsymbol{X}_{\tilde{S}} = \{x_{ij}|\ 1 \le i \le n; j \in \tilde{S}\}$. Going forward we omit the subscript $\tilde{S}$ for simplicity, keeping in mind that $\boldsymbol{X}$, $\boldsymbol{\beta}^k$, and $\boldsymbol{\beta}^*$ have dimension $\tilde{s}$.

We will show the bound (3.4) on the event that Condition 3 is satisfied and on

$$E = \left\{ \left\| \frac{1}{n} \boldsymbol{X}^T_{(\tilde{G}^c)} \boldsymbol{\varepsilon}_{\tilde{G}^c} \right\|_\infty \le C \sqrt{\frac{\sigma^2 \log p}{n}} \right\}, \tag{5.1}$$

both of which have probabilities going to one, where $\boldsymbol{X}_{(\tilde{G}^c)} = \{x_{ij}|\ i \in \tilde{G}^c, j \in \tilde{S}\}$, $\boldsymbol{\varepsilon}_{\tilde{G}^c} = \{\varepsilon_i|\ i \in \tilde{G}^c\}$, and $C > \sqrt{2}$. In fact, from Lemma 1 and Condition 1, we have for given preliminary estimators $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$,

$$\begin{aligned}
\mathbb{P}(E) &= 1 - \mathbb{P}(E^c) = 1 - \mathbb{P}\left\{ \left\| \frac{1}{n} \boldsymbol{X}^T_{(\tilde{G}^c)} \boldsymbol{\varepsilon}_{\tilde{G}^c} \right\|_\infty > C \sqrt{\frac{\sigma^2 \log p}{n}} \right\} \\
&\ge 1 - \sum_{j \in \tilde{S}} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i \in \tilde{G}^c} x_{ij} \varepsilon_i \right| > C \sqrt{\frac{\sigma^2 \log p}{n}} \right) \\
&\ge 1 - 2 \sum_{j \in \tilde{S}} \exp(-2^{-1} C^2 \log p) \\
&\ge 1 - 2 \exp(\log p - 2^{-1} C^2 \log p) = 1 - o(1).
\end{aligned}$$

The lower bound here does not depend on the preliminary estimators, and hence the probability of E goes to one.

Since $\boldsymbol{\beta}^k = \mathrm{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, \boldsymbol{\gamma}^{k-1})$, it follows that $L(\boldsymbol{\beta}^k, \boldsymbol{\gamma}^{k-1}) \leq L(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^{k-1})$. Hence, we have

$$\begin{aligned}\frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\Delta}^k\|_2^2 &\leq \frac{1}{n}\langle \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^* - \sqrt{n}\boldsymbol{\gamma}^{k-1}), \boldsymbol{\Delta}^k\rangle + \lambda_\beta \sum_{j\in\tilde{S}} w_{\beta,j}(|\beta_j^*| - |\beta_j^k|) \\ &= A_1 + A_2, \quad \text{say},\end{aligned}$$

where $\boldsymbol{\Delta}^k = \boldsymbol{\beta}^k - \boldsymbol{\beta}^*$. First, we evaluate the term A_1. Since $\sqrt{n}\gamma_i^{k-1} = \Theta(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{k-1}; \lambda_\gamma w_{\gamma,i})$ and $\sqrt{n}\gamma_i^{k-1} = 0$ when $i \in \tilde{G}^c$, the j-th element of $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^* - \sqrt{n}\boldsymbol{\gamma}^{k-1})$ is given by

$$\begin{aligned}&\sum_{i\in\tilde{G}} x_{ij}\{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^* - \Theta(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{k-1}; \lambda_\gamma w_{\gamma,i})\} + \sum_{i\in\tilde{G}^c} x_{ij}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*) \\ &= \sum_{i\in\tilde{G}} x_{ij}\big\{\boldsymbol{x}_i^T\boldsymbol{\Delta}^{k-1} + (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{k-1}) - \Theta(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{k-1}; \lambda_\gamma w_{\gamma,i})\big\} \\ &\quad + \sum_{i\in\tilde{G}^c} x_{ij}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*).\end{aligned}$$

Thus, we obtain

$$\begin{aligned}A_1 &= \frac{1}{n}(\boldsymbol{X}_{(\tilde{G})}\boldsymbol{\Delta}^{k-1})^T(\boldsymbol{X}_{(\tilde{G})}\boldsymbol{\Delta}^k) \\ &\quad + \frac{1}{n}\sum_{j\in\tilde{S}} \Delta_j^k \sum_{i\in\tilde{G}} x_{ij}\{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{k-1}) - \Theta(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{k-1}; \lambda_\gamma w_{\gamma,i})\} \\ &\quad + \frac{1}{n}\sum_{j\in\tilde{S}} \Delta_j^k \sum_{i\in\tilde{G}^c} x_{ij}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^*) = A_{11} + A_{12} + A_{13}, \quad \text{say}.\end{aligned}$$

Since $\|\boldsymbol{\Delta}^k\|_0 \leq \tilde{s}$ for any $k \geq 1$, the definition of the (doubly) restricted largest eigenvalue in (3.2) implies that

$$|A_{11}| \leq \frac{1}{n}\|\boldsymbol{X}_{(\tilde{G})}\boldsymbol{\Delta}^{k-1}\|_2\|\boldsymbol{X}_{(\tilde{G})}\boldsymbol{\Delta}^k\|_2 \leq \delta_{max}(\tilde{s}, \tilde{g})\|\boldsymbol{\Delta}^{k-1}\|_2\|\boldsymbol{\Delta}^k\|_2. \tag{5.2}$$

Note that $|\Theta(x;\lambda) - x| \leq \lambda$ under Condition 2. Then,

$$|A_{12}| \leq \frac{R_w\lambda_\gamma}{n}\|\boldsymbol{X}_{(\tilde{G})}\|_{\ell_1}\|\boldsymbol{\Delta}^k\|_1 \leq C\sqrt{\frac{\sigma^2\log p}{n}}\|\boldsymbol{\Delta}^k\|_1, \tag{5.3}$$

where $R_w > 0$ is defined in (2.3). Since $G^* \subset \tilde{G}$, we have $\gamma_i^* = 0$ for $i \in \tilde{G}^c$. Then, from (5.1), we have

$$|A_{13}| \leq \frac{1}{n}\|\boldsymbol{X}_{(\tilde{G}^c)}\boldsymbol{\varepsilon}_{\tilde{G}^c}\|_\infty\|\boldsymbol{\Delta}^k\|_1 \leq C\sqrt{\frac{\sigma^2\log p}{n}}\|\boldsymbol{\Delta}^k\|_1. \tag{5.4}$$

For the term A_2, since $S^* \subset \tilde{S}$,

$$
\begin{aligned}
A_2 &\le \lambda_\beta \sum_{j\in S^*} w_{\beta,j}|\beta_j^*| - \lambda_\beta \sum_{j\in S^*} w_{\beta,j}|\beta_j^* + \Delta_j^k| - \lambda_\beta \sum_{j\in S^{*c}} w_{\beta,j}|\beta_j^* + \Delta_j^k| \\
&\le \lambda_\beta \sum_{j\in S^*} w_{\beta,j}|\Delta_j^k| - \lambda_\beta \sum_{j\in S^{*c}} w_{\beta,j}|\Delta_j^k| \\
&\le \lambda_\beta \max_{j\in S^*} w_{\beta,j} \|\boldsymbol{\Delta}_{S^*}^k\|_1 - \frac{\lambda_\beta}{R_w}\|\boldsymbol{\Delta}_{S^{*c}}^k\|_1.
\end{aligned}
$$

Combined with (5.2)–(5.4), it follows from $\|\boldsymbol{\Delta}^k\|_1 = \|\boldsymbol{\Delta}_{S^*}^k\|_1 + \|\boldsymbol{\Delta}_{S^{*c}}^k\|_1$ that

$$
\begin{aligned}
\frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\Delta}^k\|_2^2 &\le \delta_{max}(\tilde{s},\tilde{g})\|\boldsymbol{\Delta}^{k-1}\|_2\|\boldsymbol{\Delta}^k\|_2 + 2C\sqrt{\frac{\sigma^2\log p}{n}}\|\boldsymbol{\Delta}^k\|_1 \\
&\quad + \lambda_\beta \max_{j\in S^*} w_{\beta,j}\|\boldsymbol{\Delta}_{S^*}^k\|_1 - \frac{\lambda_\beta}{R_w}\|\boldsymbol{\Delta}_{S^{*c}}^k\|_1 \\
&= \delta_{max}(\tilde{s},\tilde{g})\|\boldsymbol{\Delta}^{k-1}\|_2\|\boldsymbol{\Delta}^k\|_2 + \left(\lambda_\beta \max_{j\in S^*} w_{\beta,j} + 2C\sqrt{\frac{\sigma^2\log p}{n}}\right)\|\boldsymbol{\Delta}_{S^*}^k\|_1 \\
&\quad + \left(2C\sqrt{\frac{\sigma^2\log p}{n}} - \frac{\lambda_\beta}{R_w}\right)\|\boldsymbol{\Delta}_{S^{*c}}^k\|_1 \\
&\le \delta_{max}(\tilde{s},\tilde{g})\|\boldsymbol{\Delta}^{k-1}\|_2\|\boldsymbol{\Delta}^k\|_2 + \lambda_\beta\big(R_w^{-1} + \max_{j\in S^*} w_{\beta,j}\big)\sqrt{s^*}\|\boldsymbol{\Delta}^k\|_2.
\end{aligned}
$$

From Condition 3 and (3.1), we have $(1/n)\|\boldsymbol{X}\boldsymbol{\Delta}^k\|_2^2 \ge \kappa\|\boldsymbol{\Delta}^k\|_2^2$. Thus, for $k \ge 1$,

$$
\|\boldsymbol{\Delta}^k\|_2 \le \rho\|\boldsymbol{\Delta}^{k-1}\|_2 + 2\kappa^{-1}\sqrt{s^*}\lambda_\beta\big(R_w^{-1} + \max_{j\in S^*} w_{\beta,j}\big).
$$

Let $\boldsymbol{\Delta}^0 = \boldsymbol{\beta}^{init} - \boldsymbol{\beta}^*$. Then, the bound (3.4) is derived by solving the above recurrence relation for $k = 1, 2, \ldots$, which concludes the proof.

5.3. Proof of Theorem 2

We denote positive constants by C_i $(i \ge 1)$ which may be different from each other. Suppose that for some $\ell \in \tilde{S} \cap S^{*^c}$, $\beta_\ell^k \ne 0$. Without loss of generality, we can assume $\beta_\ell^k > 0$. Then, from the first order condition for $\boldsymbol{\beta}^k$, the value

$$
\frac{1}{n}\sum_{j\in\tilde{S}}\sum_{i=1}^n x_{ij}x_{i\ell}\beta_j^k - \frac{1}{n}\sum_{i=1}^n x_{i\ell}\left\{y_i - \Theta\big(y_i - \sum_{j\in\tilde{S}} x_{ij}\beta_j^{k-1}; \lambda_\gamma w_{\gamma,i}\big)\right\} + \lambda_\beta w_{\beta,\ell} \tag{5.5}
$$

should be zero. But, if we can show the first two terms are dominated by the third term $\lambda_\beta w_{\beta,\ell}$ for any $\ell \in \tilde{S} \cap S^{*^c}$, we have a contradiction.

First, we evaluate the middle term of (5.5). From the definition, the inside of Θ is represented as

$$y_i - \sum_{j\in\tilde{S}} x_{ij}\beta_j^{k-1} = \begin{cases} \sqrt{n}\gamma_i^* + \varepsilon_i - \sum_{j\in\tilde{S}} x_{ij}(\beta_j^{k-1} - \beta_j^*), & i \in G^* \cap \tilde{G} = G^*, \\ \varepsilon_i - \sum_{j\in\tilde{S}} x_{ij}(\beta_j^{k-1} - \beta_j^*), & i \in G^{*c} \cap \tilde{G}. \end{cases}$$

By Lemma 1, we can show that $\max_{1\le i\le n}|\varepsilon_i| \le C_1\sqrt{\log n}$ with probability going to one. Let $\lambda_\beta = C_2\{(\log p)/n\}^{1/2}$, then it follows from Corollary 1 and $a_{n,2}s^*\log p = o(n\log n)$ that

$$\begin{aligned}\max_{1\le i\le n}\left|\varepsilon_i - \sum_{j\in\tilde{S}} x_{ij}(\beta_j^{k-1} - \beta_j^*)\right| &\le \max_{1\le i\le n}|\varepsilon_i| + C_3\sqrt{\tilde{s}}\|\boldsymbol{\beta}^{k-1} - \boldsymbol{\beta}^*\|_2 \\ &\le C_4\sqrt{\log n}(1 + o(1)),\end{aligned}$$

which implies that if $\lambda_\gamma \min_{i\in G^{*c}\cap\tilde{G}} w_{\gamma,i} \ge C_4\sqrt{\log n}$, then $y_i - \sum_{j\in\tilde{S}} x_{ij}\beta_j^{k-1} = 0$ for $i \in G^{*c} \cap \tilde{G}$. Note that $\min_{i\in G^{*c}\cap\tilde{G}} w_{\gamma,i} = R_w$ for sufficiently large n since $\min_{i\in G^{*c}\cap\tilde{G}}(1/|\tilde{\gamma}_i|) \ge 1/(Ca_{n,1}) \to \infty$. Hence it suffices to put $\lambda_\gamma = C_5\sqrt{\log n}$ for a sufficiently large $C_5 > 0$. Such a λ_γ satisfying the condition of Corollary 1 can be selected since $a_{n,2}^2 = o(n)$. Meanwhile, since $\sqrt{\log n} = o(\sqrt{n}\min_{i\in G^*}|\gamma_i^*|)$, we have $y_i - \sum_{j\in\tilde{S}} x_{ij}\beta_j^{k-1} = \sqrt{n}\gamma_i^*(1 + o(1))$ for $i \in G^*$. Thus, for sufficiently large n, it holds that $\min_{i\in G^*}|y_i - \sum_{j\in\tilde{S}} x_{ij}\beta_j^{k-1}| \ge \lambda_\gamma R_w \ge \lambda_\gamma \max_{i\in G^*} w_{\gamma,i}$. Therefore, under Condition 2,

$$\Theta\big(y_i - \sum_{j\in\tilde{S}} x_{ij}\beta_j^{k-1}; \lambda_\gamma w_{\gamma,i}\big) = \begin{cases} y_i - \sum_{j\in\tilde{S}} x_{ij}\beta_j^{k-1} + O(\sqrt{\log n}), & i \in G^*, \\ 0, & i \in G^{*c} \cap \tilde{G}. \end{cases}$$

Then, it follows that

$$\begin{aligned}&\sum_{i=1}^n x_{i\ell}\Big\{y_i - \Theta\big(y_i - \sum_{j\in\tilde{S}} x_{ij}\beta_j^{k-1}; \lambda_\gamma w_{\gamma,i}\big)\Big\} \\ &= \sum_{i\in\tilde{G}} x_{i\ell}\Big\{y_i - \Theta\big(y_i - \sum_{j\in\tilde{S}} x_{ij}\beta_j^{k-1}; \lambda_\gamma w_{\gamma,i}\big)\Big\} + \sum_{i\in\tilde{G}^c} x_{i\ell}y_i \\ &= \sum_{i\in G^*} x_{i\ell}\sum_{j\in\tilde{S}} x_{ij}\beta_j^{k-1} + O(\sqrt{\log n})\sum_{i\in G^*} x_{i\ell} + \sum_{i\in G^{*c}\cap\tilde{G}} x_{i\ell}y_i + \sum_{i\in\tilde{G}^c} x_{i\ell}y_i \\ &= \sum_{i\in G^*} x_{i\ell}\sum_{j\in\tilde{S}} x_{ij}\big\{\beta_j^* + (\beta_j^{k-1} - \beta_j^*)\big\} + \sum_{i\in G^{*c}} x_{i\ell}y_i + O(g^*\sqrt{\log n}) \\ &= \sum_{i=1}^n\sum_{j\in\tilde{S}} x_{i\ell}x_{ij}\beta_j^* + \sum_{i\in G^*}\sum_{j\in\tilde{S}} x_{i\ell}x_{ij}(\beta_j^{k-1} - \beta_j^*) + \sum_{i\in G^{*c}} x_{i\ell}\varepsilon_i + O(g^*\sqrt{\log n}).\end{aligned}$$

Thus, (5.5) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\tilde{S}} x_{ij}x_{i\ell}(\beta_j^k-\beta_j^*) - \frac{1}{n}\sum_{i\in G^*}\sum_{j\in\tilde{S}} x_{ij}x_{i\ell}(\beta_j^{k-1}-\beta_j^*) - \frac{1}{n}\sum_{i\in G^{*c}} x_{i\ell}\varepsilon_i + O\left(\frac{g^*}{n}\sqrt{\log p}\right) + \lambda_\beta w_{\beta,\ell}.$$

Clearly the fist and second terms have the order $\sqrt{a_{n,2}s^*}\lambda_\beta$. From the proof of Theorem 1, the third term is of order $\{(\log p)/n\}^{1/2}$. By $a_{n,1}\max\left(\sqrt{a_{n,2}s^*}, g^*/\sqrt{n}\right) = o(1)$, the first four terms of (5.5) are dominated by $\lambda_\beta \min_{j\in\tilde{S}\cap S^{*c}} w_{\beta,j}$ since $\min_{j\in\tilde{S}\cap S^{*c}} w_{\beta,j} \geq 1/(Ca_{n,1})$.

Acknowledgements

We thank the associate editor and the referee for their careful reading of the original article and for their valuable comments that greatly helped improve the article. This work was supported by the System Genetics Project of the Research Organization of Information and Systems and by JSPS KAKENHI Grant Number 15K15946.

References

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association* **96**, 939-967.

Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37**, 1705-1732.

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

Fan, J. (1997). Comments on 'wavelets in statistics: A review' by A. Antoniadis. *Journal of the American Statistical Association* **6**, 131-139.

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

Fan, Y. and Lv, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *Journal of the American Statistical Association* **108**, 1044-1061.

Fan, Y. and Lv, J. (2014). Asymptotic properties for combined L_1 and concave regularization. *Biometrika* **101**, 57-70.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1-22.

Gao, H. (1998). Wavelet shrinkage denoising using the non-negative garrote. *Journal of Computational and Graphical Statistics* **7**, 469-488.

Huber, P. and Ronchetti, E. (2009). *Robust Statistics, 2nd edition.* Wiley, New York.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics* **37**, 246-270.

Negahban, S., Ravikumar, P., Wainwright, M. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27**, 538-557.

Nguyen, N. and Tran, T. (2013). Robust lasso with missing and grossly corrupted observations. *IEEE Transactions Information Theory* **59**, 2036-2058.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461-464.

She, Y. and Owen, A. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association* **106**, 626-639.

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

Van De Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* **3**, 1360-1392.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications* **5**, 210-268.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions Information Theory* **55**, 2183-2202.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894-942.

Zhang, C. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *The Annals of Statistics* **36**, 1567-1594.

Zhang, C. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science* **27**, 576-593.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7**, 2541-2563.

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

Department of Industrial Engineering and Economics, Tokyo Institute of Technology, 2-12-1, Ookayama, Meguro-ku Tokyo 1528550 Japan.

E-mail: katayama.s.ad@m.titech.ac.jp

The Institute of Statistical Mathematics; Department of Statistical Science, The Graduate University for Advanced Studies; Department of Mathematical Statistics, Nagoya University Graduate School of Medicine, Nagoya, 464-8601, Japan.

E-mail: fujisawa@ism.ac.jp

(Received May 2015; accepted June 2016)