ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 1, *pp.* 247–265

Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions

Jianqing Fan,

Princeton University, USA, and Academy of Mathematics and Systems Science, Beijing, People's Republic of China

Quefeng Li

University of North Carolina at Chapel Hill, USA

and Yuyan Wang

Princeton University, USA

[Received December 2014. Final revision November 2015]

Summary. Data subject to heavy-tailed errors are commonly encountered in various scientific fields. To address this problem, procedures based on quantile regression and least absolute deviation regression have been developed in recent years. These methods essentially estimate the conditional median (or quantile) function. They can be very different from the conditional mean functions, especially when distributions are asymmetric and heteroscedastic. How can we efficiently estimate the mean regression functions in ultrahigh dimensional settings with existence of only the second moment? To solve this problem, we propose a penalized Huber loss with diverging parameter to reduce biases created by the traditional Huber loss. Such a penalized robust approximate (RA) quadratic loss will be called the RA lasso. In the ultrahigh dimensional setting, where the dimensionality can grow exponentially with the sample size, our results reveal that the RA lasso estimator produces a consistent estimator at the same rate as the optimal rate under the light tail situation. We further study the computational convergence of the RA lasso and show that the composite gradient descent algorithm indeed produces a solution that admits the same optimal rate after sufficient iterations. As a by-product, we also establish the concentration inequality for estimating the population mean when there is only the second moment. We compare the RA lasso with other regularized robust estimators based on quantile regression and least absolute deviation regression. Extensive simulation studies demonstrate the satisfactory finite sample performance of the RA lasso.

Keywords: High dimension; Huber loss; *M*-estimator; Optimal rate; Robust regularization

1. Introduction

Our era has witnessed the massive explosion of data and a dramatic improvement of technology in collecting and processing large data sets. We often encounter huge data sets in which the number of features greatly surpasses the number of observations. It makes many traditional statistical analysis tools infeasible and poses great challenge on developing new tools. Regularization methods have been widely used for the analysis of high dimensional data. These methods penalize the least squares or the likelihood function with the L_1-penalty on the unknown par-

Address for correspondence: Jianqing Fan, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA.
E-mail: jqfan@princeton.edu

 1369–7412/17/79247

ameters (the lasso; Tibshirani (1996)), or a folded concave penalty function such as smoothly clipped absolute deviation (Fan and Li, 2001) and the minimum convex penalty (Zhang, 2010). However, these penalized least squares methods are sensitive to the tails of the error distributions, particularly for ultrahigh dimensional covariates, as the maximum spurious correlation between the covariates and the realized noise can be large in those cases. As a result, theoretical properties are often obtained under light-tailed error distributions (Bickel *et al*., 2009; Fan and Lv, 2011). Besides regularization methods, traditional stagewise selection methods (e.g. forward selection) have also been extended to the high dimensional setting. For instance, Fan and Lv (2008) proposed a sure independence screening method and Wang (2009) studied stagewise selection methods in high dimension settings. These methods are usually built on marginal correlations between the response and covariates; hence they also need light tail assumptions on the errors.

To tackle the problem of heavy-tailed errors, robust regularization methods have been extensively studied. Li and Zhu (2008), Wu and Liu (2009) and Zou and Yuan (2008) developed robust regularized estimators based on quantile regression for the case of fixed dimensionality. Belloni and Chernozhukov (2011) studied L_1-penalized quantile regression in high dimensional sparse models. Fan *et al*. (2014) further considered an adaptively weighted L_1-penalty to alleviate the bias problem and established the oracle property and asymptotic normality of the corresponding estimator. Other robust estimators were developed based on least absolute deviation (LAD) regression. Wang (2013) studied L_1-penalized LAD regression and showed that the estimator achieves near oracle risk performance under the high dimensional setting.

These methods essentially estimate the conditional *median (or quantile)* regression, instead of the conditional *mean* regression, function. In applications where mean regression is of interest, these methods are not feasible unless a strong assumption is made that the distribution of errors is symmetric around zero. A simple example is the heteroscedastic linear model with asymmetric noise distribution. Another example is to estimate the conditional variance function such as the auto-regressive conditional heteroscedasticity model (Engle, 1982). In these cases, the conditional mean and conditional median are very different. Another important example is to estimate large covariance matrices without assuming light tails. We shall explain this more in detail in Section 5. In addition, LAD-based methods tend to penalize strongly on small errors. If only a small proportion of samples are outliers, they are expected to be less efficient than the least-squares-based method.

A natural question is then how to conduct ultrahigh dimensional mean regression when the tails of errors are not light and how to estimate the sample mean with very fast concentration when the distribution has only bounded second moment. These simple questions have not been carefully studied. LAD-based methods do not intend to answer these questions as they alter the problems of the study. This leads us to consider Huber loss as another way of robustification. The Huber loss (Huber, 1964) is a hybrid of squared loss for relatively small errors and absolute loss for relatively large errors, where the degree of hybridization is controlled by one tuning parameter. Lambert-Lacroix and Zwald (2011) proposed to use the Huber loss together with the adaptive lasso penalty for robust estimation. However, they needed the strong assumption that the distribution of errors is symmetric around zero. Unlike their method, we waive the symmetry requirement by allowing the regularization parameter to diverge (or to converge if its reciprocal is used) to reduce the bias that is induced by the Huber loss when the distribution is asymmetric. In this paper, we consider the regularized approximate (RA) quadratic estimator (the RA lasso) with an L_1-penalty and show that it admits the same L_2-error rate as the optimal error rate in the light tail situation. In particular, if the distribution of errors is indeed symmetric around zero (where the median and mean agree), this rate is the same as the regularized LAD

estimator that was obtained in Wang (2013). Therefore, the RA lasso estimator does not lose efficiency in this special case. In practice, since the distribution of errors is unknown, the RA lasso is more flexible than the existing methods in terms of estimating the conditional mean regression function.

A by-product of our method is that the RA lasso estimator of the population mean has the exponential type of concentration even in the presence of the finite second moment. Catoni (2012) studied this type of problem and proposed a class of losses to result in a robust M-estimator of mean with exponential type of concentration. We further extend his idea to the sparse linear regression setting and show that Catoni loss is another choice to reach the optimal rate.

As in many other references, estimators with nice sampling properties are typically defined through the optimization of a target function such as penalized least squares. The properties that are established are not necessarily the same as those that are computed. Following the framework of Agarwal *et al.* (2012), we propose the composite gradient descent algorithm for solving the RA lasso estimator and develop the sampling properties by taking computational error into consideration. We show that the algorithm indeed produces a solution that admits the same optimal L_2-error rate as the theoretical estimator after a sufficient number of iterations.

This paper is organized as follows. First, in Section 2, we introduce the RA lasso estimator and give the non-asymptotic upper bound for its L_2-error. We show that it has the same rate as the optimal rate under light tails. In Section 3, we study the property of the composite gradient descent algorithm for solving our problem and show that the algorithm produces a solution that performs as well as the theoretical solution. In Section 4, we apply the idea to robust estimation of the mean and large covariance matrix. In Section 5, we show similar results for Catoni loss in robust sparse regression. Section 6 gives estimation of residual variance. Numerical studies are given in Sections 7 and 8 to compare our method with two competitors. Proofs of theorems 1 and 2 are given in Appendix A, which together imply the main result (theorem 3). A proof of theorem 5 regarding the concentration of the robust mean estimator is also given in Appendix A. Proofs of supporting lemmas and remaining theorems are given in an on-line supplementary file. The relevant MATLAB code is available from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Robust approximate lasso estimator

We consider the linear regression model

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}^* + \epsilon_i, \tag{2.1}$$

where $\{\mathbf{x}_i\}_{i=1}^n$ are independent and identically distributed p-dimensional covariate vectors, $\{\epsilon_i\}_{i=1}^n$ are independent and identically distributed errors and $\boldsymbol{\beta}^*$ is a p-dimensional regression coefficient vector. The assumption of independent and identically distributed errors indeed allows conditional heteroscedastic models, where ϵ_i can depend on $\mathbf{x}_i$. For example, it can be $\epsilon_i = \sigma(\mathbf{x}_i)\tilde{\epsilon}_i$, where $\sigma(\mathbf{x}_i)$ is a function of $\mathbf{x}_i$ and $\tilde{\epsilon}_i$ is independent of $\mathbf{x}_i$. We consider the high dimensional setting, where $\log(p) = O(n^b)$ for some constant $0 < b < 1$. The distributions of $\mathbf{x}$ and $\epsilon|\mathbf{x}$ are both assumed to have mean 0. Under this assumption, $\boldsymbol{\beta}^*$ is related to the mean effect of y conditioning on $\mathbf{x}$, which is assumed to be of interest. $\boldsymbol{\beta}^*$ differs from the median effect of y conditioning on $\mathbf{x}$, especially under the heteroscedastic models or more general models. Therefore, the LAD-based methods are not applicable.

To adapt for different magnitudes of errors and to robustify the estimation, we propose to use the Huber loss (Huber, 1964):

$$l_\alpha(x) = \begin{cases} 2\alpha^{-1}|x| - \alpha^{-2} & \text{if } |x| > \alpha^{-1}; \\ x^2 & \text{if } |x| \leqslant \alpha^{-1}. \end{cases} \tag{2.2}$$

The Huber loss is quadratic for small values of x and linear for large values of x. The parameter α controls the blending of quadratic and linear penalization. Least squares and LAD can be regarded as two extremes of the Huber loss for $\alpha = 0$ and $\alpha = \infty$ respectively. Deviating from the traditional Huber estimator, the parameter α converges to 0 to reduce the biases of estimating the mean regression function when the conditional distribution of ϵ_i is not symmetric. However, α cannot shrink too fast in order to maintain the robustness. In this paper, we regard α as a tuning parameter, whose optimal value will be discussed later in this section. In practice, α needs to be tuned by some data-driven method. By letting α vary, we call $l_\alpha(x)$ the RA quadratic loss.

To estimate β^*, we propose to solve the following convex optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} l_\alpha(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p} |\beta_j|. \tag{2.3}$$

To assess the performance of $\hat{\boldsymbol{\beta}}$, we study the property of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, where $\|\cdot\|_2$ is the Euclidean norm of a vector. When λ_n converges to 0 sufficiently fast, $\hat{\boldsymbol{\beta}}$ is a natural M-estimator of $\boldsymbol{\beta}_\alpha^* = \arg\min_\beta E\{l_\alpha(y - \mathbf{x}'\boldsymbol{\beta})\}$, which is the population minimizer under the RA quadratic loss and varies by α. In general, $\boldsymbol{\beta}_\alpha^*$ differs from $\boldsymbol{\beta}^*$. But, since the RA quadratic loss approximates the quadratic loss as $\alpha \to 0$, $\boldsymbol{\beta}_\alpha^*$ is expected to converge to $\boldsymbol{\beta}^*$. This property will be established in theorem 1. Therefore, we decompose the statistical error $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ into the approximation error $\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*$ and the estimation error $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*$. The statistical error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ is then bounded by

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leqslant \underbrace{\|\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*\|_2}_{\text{approximation error}} + \underbrace{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2}_{\text{estimation error}}.$$

In what follows, we give upper bounds of the approximation and estimation error. We show that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ is upper bounded by the same rate as the optimal rate under light tails, as long as the two tuning parameters α and λ_n are properly chosen. We first give the upper bound of the approximation error under some moment conditions on $\mathbf{x}$ and $\epsilon|\mathbf{x}$. We assume that $\|\boldsymbol{\beta}^*\|_2 \leqslant \rho_2$, where the radius ρ_2 is a sufficiently large constant. This is a mild assumption, which is implied by condition 2 below and a reasonable assumption that $\mathrm{var}(y) < \infty$, since $\mathrm{var}(y) \geqslant \boldsymbol{\beta}^{*\mathrm{T}} E(\mathbf{x}\mathbf{x}^{\mathrm{T}})\boldsymbol{\beta}^* \geqslant \kappa_1 \|\boldsymbol{\beta}^*\|_2^2$.

Theorem 1. First we state the following conditions.

Condition 1. $E\{E(|\epsilon|^k|\mathbf{x})\}^2 \leqslant M_k < \infty$, for some $k \geqslant 2$.

Condition 2. $0 < \kappa_{\mathrm{l}} \leqslant \lambda_{\min}\{E(\mathbf{x}\mathbf{x}^{\mathrm{T}})\} \leqslant \lambda_{\max}\{E(\mathbf{x}\mathbf{x}^{\mathrm{T}})\} \leqslant \kappa_{\mathrm{u}} < \infty$.

Condition 3. For any $\boldsymbol{\nu} \in \mathbb{R}^p$, $\mathbf{x}^{\mathrm{T}}\boldsymbol{\nu}$ is sub-Gaussian with parameter at most $\kappa_0^2\|\boldsymbol{\nu}\|_2^2$, i.e. $E\{\exp(t\mathbf{x}^{\mathrm{T}}\boldsymbol{\nu})\} \leqslant \exp(t^2\kappa_0^2\|\boldsymbol{\nu}\|_2^2/2)$, for any $t \in \mathbb{R}$.

Under these conditions there is a universal positive constant C_1, such that $\|\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*\|_2 \leqslant C_1\sqrt{\kappa_{\mathrm{u}}}\kappa_{\mathrm{l}}^{-1}(\kappa_0^k + \sqrt{M_k})\alpha^{k-1}$.

Theorem 1 reveals that the approximation error vanishes faster if higher moments of $\epsilon|\mathbf{x}$ exist. We next give the non-asymptotic upper bound of the estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2$. This part differs from the existing work regarding the estimation error of high dimensional regularized M-estimators (Negahban *et al.*, 2012; Agarwal *et al.*, 2012) as the population minimizer $\boldsymbol{\beta}_\alpha^*$ now varies with α. However, we shall show that the upper bound of the estimation error does not depend on α, given a uniform sparsity condition.

To be solvable in the high dimensional setting, $\boldsymbol{\beta}^*$ is usually assumed to be sparse or weakly sparse, i.e. many elements of $\boldsymbol{\beta}^*$ are 0 or small. By theorem 1, $\boldsymbol{\beta}^*_\alpha$ converges to $\boldsymbol{\beta}^*$ as $\alpha \to 0$. In view of this fact, we assume that $\boldsymbol{\beta}^*_\alpha$ is uniformly weakly sparse when α is sufficiently small. In particular, we assume that there is a small constant $r > 0$, such that $\boldsymbol{\beta}^*_\alpha$ belongs to an L_q-ball with a uniform radius R_q, i.e.

$$\sum_{j=1}^{p} |\beta^*_{\alpha,j}|^q \leqslant R_q, \tag{2.4}$$

for all $\alpha \in (0, r]$ and some $q \in (0, 1]$. When the conditional distribution of ϵ_i is symmetric, $\beta^*_{\alpha,j} = \beta^*_j$ for all α and j. Therefore the condition reduces to that $\boldsymbol{\beta}^*$ is in the L_q-ball. When the conditional distribution of ϵ_i is asymmetric, we give a sufficient condition showing that, if $\boldsymbol{\beta}^*$ belongs to an L_q-ball with radius $R_q/2$, inequality (2.4) holds for all $\alpha \leqslant c\{R_q p^{-(2-q)/2}\}^{1/\{q(k-1)\}}$, where c is a positive constant. In fact, for any $q \in (0, 1]$, $|r_1|^q + |r_2|^q \geqslant (|r_1| + |r_2|)^q \geqslant |r_1 + r_2|^q$. Using this,

$$\sum_{j=1}^{p} |\beta^*_{\alpha,j}|^q \leqslant \sum_{j=1}^{p} |\beta^*_{\alpha,j} - \beta^*_j|^q + \sum_{j=1}^{p} |\beta^*_j|^q \leqslant p^{(2-q)/2} \left(\sum_{j=1}^{p} |\beta^*_{\alpha,j} - \beta^*_j|^2 \right)^{q/2} + \sum_{j=1}^{p} |\beta^*_j|^q.$$

By theorem 1, $\Sigma_{j=1}^{p} |\beta^*_{\alpha,j} - \beta^*_j|^2 = O(\alpha^{2(k-1)})$. Hence, if $\Sigma_{j=1}^{p} |\beta^*_j|^q \leqslant R_q/2$, we have $\Sigma_{j=1}^{p} |\beta^*_{\alpha,j}|^q \leqslant R_q$ for all $\alpha \leqslant c\{R_q p^{-(2-q)/2}\}^{1/\{q(k-1)\}}$.

Since the RA quadratic loss is convex, we show that with high probability the estimation error $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha$ belongs to a star-shaped set, which depends on α and the threshold level η of signals.

Lemma 1. Under conditions 1 and 3, with the choice of $\lambda_n = \kappa_\lambda \sqrt{\{\log(p)/n\}}$ and $\alpha \geqslant L\lambda_n/(4v)$, where v and L are positive constants depending on M_2 and κ_0, and κ_λ is a sufficiently large constant such that $\kappa_\lambda^2 > 32v$, it holds with probability greater than $1 - 2\exp(-c_0 n)$ that

$$\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha \in \mathbb{C}_{\alpha\eta} := \{\boldsymbol{\Delta} \in \mathbb{R}^p : \|\boldsymbol{\Delta}_{S^c_{\alpha\eta}}\|_1 \leqslant 3\|\boldsymbol{\Delta}_{S_{\alpha\eta}}\|_1 + 4\|\boldsymbol{\beta}^*_{\alpha, S^c_{\alpha\eta}}\|_1\},$$

where $c_0 = \kappa_\lambda^2/(32v) - 1$, η is a positive constant, $S_{\alpha\eta} = \{j : |\beta^*_{\alpha,j}| > \eta\}$ and $\boldsymbol{\Delta}_{S_{\alpha\eta}}$ denotes the subvector of $\boldsymbol{\Delta}$ with indices in set $S_{\alpha\eta}$.

We further verify a restricted strong convexity (RSC) condition, which has been shown to be critical in the study of high dimensional regularized M-estimators (Negahban *et al.*, 2012; Agarwal *et al.*, 2012). Let

$$\delta\mathcal{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}) = \mathcal{L}_n(\boldsymbol{\beta} + \boldsymbol{\Delta}) - \mathcal{L}_n(\boldsymbol{\beta}) - \nabla\mathcal{L}_n(\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Delta}, \tag{2.5}$$

where $\mathcal{L}_n(\boldsymbol{\beta}) = (1/n)\Sigma_{i=1}^{n} l_\alpha(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, $\boldsymbol{\Delta}$ is a p-dimensional vector and $\nabla\mathcal{L}_n(\boldsymbol{\beta})$ is the gradient of $\mathcal{L}_n$ at the point of $\boldsymbol{\beta}$.

Definition 1. The loss function $\mathcal{L}_n$ satisfies the RSC condition on a set S with curvature $\kappa_{\mathcal{L}} > 0$ and tolerance $\tau_{\mathcal{L}}$ if

$$\delta\mathcal{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}) \geqslant \kappa_{\mathcal{L}}\|\boldsymbol{\Delta}\|_2^2 - \tau_{\mathcal{L}}^2, \qquad \text{for all } \boldsymbol{\Delta} \in S.$$

Next, we show that with high probability the RA quadratic loss (2.2) satisfies RSC for $\boldsymbol{\beta} = \boldsymbol{\beta}^*_\alpha$ and all $\boldsymbol{\Delta} \in \mathbb{C}_{\alpha\eta} \cap \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_2 \leqslant 1\}$ with uniform constants $\kappa_{\mathcal{L}}$ and $\tau_{\mathcal{L}}$ that do not depend on α. To prove RSC at $\boldsymbol{\beta}^*_\alpha$ and a stronger version in lemma 4, we first give a uniform lower bound of $\delta\mathcal{L}_n(\boldsymbol{\Delta}, \beta)$ for all $\|\boldsymbol{\beta}\|_2 \leqslant 4\rho_2$, $\|\boldsymbol{\Delta}\|_2 \leqslant 8\rho_2$ and $\alpha \leqslant c_{\mathrm{u}}\rho_2^{-1}$, where c_{u} is a positive constant, depending on M_k, κ_{l}, κ_{u} and κ_0.

Lemma 2. Under conditions 1–3, for all $\|\boldsymbol{\beta}\|_2 \leqslant 4\rho_2$, $\|\boldsymbol{\Delta}\|_2 \leqslant 8\rho_2$ and $\alpha \leqslant c_{\mathrm{u}}\rho_2^{-1}$, there are uniform positive constants κ_1, κ_2, c_1' and c_2' such that, with probability at least $1 - c_1'\exp(-c_2'n)$,

$$\delta\mathcal{L}_n(\boldsymbol{\Delta}, \beta) \geqslant \kappa_1\|\boldsymbol{\Delta}\|_2[\|\boldsymbol{\Delta}\|_2 - \kappa_2\sqrt{\{\log(p)/n\}}\|\boldsymbol{\Delta}\|_1]. \tag{2.6}$$

Lemma 3. Suppose that conditions 1–3 hold and assume that

$$8\kappa_2\kappa_\lambda^{-q/2}\sqrt{R_q}\left\{\frac{\log(p)}{n}\right\}^{(1-q)/2} \leqslant 1, \tag{2.7}$$

by choosing $\eta = \lambda_n$, with probability at least $1 - c_1'\exp(-c_2'n)$, the RSC condition holds for $\delta\mathcal{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}_\alpha^*)$ for any $\boldsymbol{\Delta} \in \mathbb{C}_{\alpha\eta} \cap \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_2 \leqslant 1\}$ with $\kappa_{\mathcal{L}} = \kappa_1/2$ and $\tau_{\mathcal{L}}^2 = 4R_q\kappa_1\kappa_2\kappa_\lambda^{1-q} \times \{n^{-1}\log(p)\}^{1-q/2}$.

Lemma 3 shows that, even though $\boldsymbol{\beta}_\alpha^*$ is unknown and the set $\mathbb{C}_{\alpha\eta}$ depends on α, RSC holds with uniform constants that do not depend on α. This further gives the following upper bound of the estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2$, which also does not depend on α.

Theorem 2. Under the conditions of lemmas 1 and 3, there are positive constants c_1, c_2 and C_2 such that, with probability at least $1 - c_1\exp(-c_2n)$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2 \leqslant C_2k_1^{-2}\kappa_\lambda^{2-q}R_q\{n^{-1}\log(p)\}^{1-q/2}.$$

Finally, theorems 1 and 2 together lead to the following main result, which gives the non-asymptotic upper bound of the statistical error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$.

Theorem 3. Under the conditions of lemmas 1 and 3, with probability at least $1 - c_1\exp(-c_2n)$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leqslant d_1\alpha^{k-1} + d_2\sqrt{R_q}\{\log(p)/n\}^{1/2-q/4}, \tag{2.8}$$

where the constants $d_1 = C_1\sqrt{\kappa_{\mathrm{u}}}\kappa_1^{-1}(\kappa_0^k + \sqrt{M_k})$ and $d_2 = C_2k_1^{-2}\kappa_\lambda^{2-q}$.

Next, we compare our result with the existing results regarding the robust estimation of high dimensional linear regression models.

(a) When the conditional distribution of ϵ is symmetric around 0, then $\boldsymbol{\beta}_\alpha^* = \boldsymbol{\beta}^*$ for any α, which has no approximation error. If ϵ has heavy tails in addition to being symmetric, we would like to choose α sufficiently large to robustify the estimation. Theorem 2 implies that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ has a convergence rate of $\sqrt{R_q}\{\log(p)/n\}^{1/2-q/4}$, where $R_q = \Sigma_{j=1}^p|\beta_j^*|^q$. The rate is the same as the minimax rate (Raskutti *et al.*, 2011) for weakly sparse models under light tails. In a special case that $q = 0$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ converges at a rate of $\sqrt{\{s\log(p)/n\}}$, where s is the number of non-zero elements in $\boldsymbol{\beta}^*$. This is the same rate as the regularized LAD estimator in Wang (2013) and the regularized quantile estimator in Belloni and Chernozhukov (2011). It suggests that our method does not lose efficiency for symmetric heavy-tailed errors.

(b) If the conditional distribution of ϵ is asymmetric around 0, the quantile and LAD-based methods are inconsistent, since they estimate the median instead of the mean. Theorem 3 shows that our estimator still achieves the optimal rate as long as $\alpha \leqslant \{d_1^{-1}d_2R_q\{\log(p)/n\}^{1-q/2}\}^{1/\{2(k-1)\}}$. Recall from the conditions in lemmas 1 and 3 that we also need to choose α, such that $c_{\mathrm{l}}\sqrt{\{\log(p)/n\}} \leqslant \alpha \leqslant c_{\mathrm{u}}\rho_2^{-1}$ for some constants c_{l} and c_{u}. Given the sparsity condition (2.7), α can be chosen to meet the above three requirements. In terms of estimating the conditional mean effect, errors with heavy but asymmetric tails give the case where the RA lasso has the biggest advantage over existing estimators.

In practice, the distribution of errors is unknown. Our method is more flexible than existing methods as it does not require symmetry and light tail assumptions. The tuning parameter α plays a key role by adapting to errors with different shapes and tails. In reality, the optimal values of tuning parameters α and λ_n can be chosen by a two-dimensional grid search using cross-validation or an information-based criterion, e.g. the Akaike information criterion or Bayesian information criterion. More specifically, the search grid is formed by partitioning a rectangle in the scale of $(\log(\alpha), \log(\lambda_n))$. The optimal values are then found by the combination that minimizes the Akaike information criterion, Bayesian information criterion or the cross-validated measurement (such as mean-squared error).

3. Geometric convergence of computational error

The gradient descent algorithm (Nesterov, 2007; Agarwal *et al.*, 2012) is usually applied to solve the convex problem (2.3). For example, we can replace the RA quadratic loss with its local isotropic quadratic approximation and iteratively solve the following optimization problem:

$$\hat{\boldsymbol{\beta}}^{t+1} = \underset{\|\boldsymbol{\beta}\|_1 \leqslant \rho}{\arg\min} \{\mathcal{L}_n(\hat{\boldsymbol{\beta}}^t) + \nabla \mathcal{L}_n(\hat{\boldsymbol{\beta}}^t)^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t) + \frac{\gamma_{\mathrm{u}}}{2}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t\|_2^2 + \lambda_n \|\boldsymbol{\beta}\|_1\}, \tag{3.1}$$

where γ_{u} is a sufficiently large fixed constant whose condition is specified in expression (3.3) below and the side constraint '$\|\boldsymbol{\beta}\|_1 \leqslant \rho$' is introduced to guarantee good performance in the first few iterations and ρ is allowed to be sufficiently large such that $\boldsymbol{\beta}^*$ is feasible. The isotropic local quadratic approximation allows an expedient computation. To solve problem (3.1), the update can be computed by a two-step procedure. We first solve problem (3.1) without the norm constraint, which is the soft threshold of the vector $\hat{\boldsymbol{\beta}}^t - (1/\gamma_{\mathrm{u}})\nabla \mathcal{L}_n(\hat{\boldsymbol{\beta}}^t)$ at level λ_n, and call the solution $\check{\boldsymbol{\beta}}$. If $\|\check{\boldsymbol{\beta}}\|_1 \leqslant \rho$, set $\hat{\boldsymbol{\beta}}^{t+1} = \check{\boldsymbol{\beta}}$. Otherwise, $\hat{\boldsymbol{\beta}}^{t+1}$ is obtained by further projecting $\check{\boldsymbol{\beta}}$ onto the L_1-ball $\{\boldsymbol{\beta}: \|\boldsymbol{\beta}\|_1 \leqslant \rho\}$. The projection can be done (Duchi *et al.*, 2008) by soft thresholding $\check{\boldsymbol{\beta}}$ at level π_n, where π_n is given by the following procedure:

(a) sort $\{|\check{\beta}_j|\}_{j=1}^p$ into $b_1 \geqslant b_2 \geqslant \ldots \geqslant b_p$;
(b) find $J = \max\{1 \leqslant j \leqslant p : b_j - (\Sigma_{r=1}^j b_r - \rho)/j > 0\}$ and let $\pi_n = (\Sigma_{r=1}^J b_j - \rho)/J$.

Agarwal *et al.* (2012) considered the computational error of such a first-order gradient descent method. They showed that, for a convex and differentiable loss functions $l(x)$ and decomposable penalty function $p(\boldsymbol{\beta})$, the error $\|\hat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^*\|_2$ has the same rate as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$ for all sufficiently large t, where $\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} E\{l(\mathbf{x}, y; \boldsymbol{\beta})\}$, and $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}(1/n)\Sigma_{i=1}^n l(\mathbf{x}_i, y_i, \beta) + p(\boldsymbol{\beta})$. Differently from their set-up, our population minimizer $\boldsymbol{\beta}_\alpha^*$ varies by α. Nevertheless, as $\boldsymbol{\beta}_\alpha^*$ converges to the true effect $\boldsymbol{\beta}^*$, by a careful control of α, we can still show that $\|\hat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^*\|_2$ has the same rate as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, where $\hat{\boldsymbol{\beta}}$ is the theoretical solution of expression (2.3) and $\hat{\boldsymbol{\beta}}^t$ is as defined in problem (3.1).

The key is that the RA quadratic loss function $\mathcal{L}_n$ satisfies the RSC condition and the restricted smoothness condition with some uniform constants, namely $\delta\mathcal{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta})$ as defined in expression (2.5) satisfies the following conditions: RSC,

$$\delta\mathcal{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}) \geqslant \frac{\gamma_{\mathrm{l}}}{2}\|\boldsymbol{\Delta}\|_2^2 - \tau_{\mathrm{l}}\|\boldsymbol{\Delta}\|_1^2; \tag{3.2}$$

restricted smoothness,

$$\delta\mathcal{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}) \leqslant \frac{\gamma_{\mathrm{u}}}{2}\|\boldsymbol{\Delta}\|_2^2 + \tau_{\mathrm{u}}\|\boldsymbol{\Delta}\|_1^2, \tag{3.3}$$

for all $\boldsymbol{\beta}$ and $\boldsymbol{\Delta}$ in some set of interest, with parameters γ_{l}, τ_{l}, γ_{u} and τ_{u} that do not depend on α. We show that such conditions hold with high probability.

Lemma 4. Under conditions 1–3, for all $\|\boldsymbol{\beta}\|_2 \leqslant 4\rho_2$, $\|\boldsymbol{\Delta}\|_2 \leqslant 8\rho_2$ and $\alpha \leqslant c_{\mathrm{u}}\rho_2^{-1}$, with probability greater than $1 - c_1\exp(-c_2 n)$, conditions (3.2) and (3.3) hold with $\gamma_{\mathrm{l}} = \kappa_1$, $\tau_{\mathrm{l}} = \kappa_1\kappa_2^2\log(p)/(2n)$, $\gamma_{\mathrm{u}} = 3\kappa_{\mathrm{u}}$ and $\tau_{\mathrm{u}} = \kappa_{\mathrm{u}}\log(p)/n$.

We further give an upper bound of computational error $\|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}\|_2$ in theorem 4. It shows that, with high probability, $\|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}\|_2$ is dominated by $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha\|_2$ after sufficient iterations, as long as $R_q\{\log(p)/n\}^{1-q/2} = o(1)$, which is required for consistency of *any method* over the weak sparse L_q-ball by the known minimax results (Raskutti *et al.*, 2011). Denote $r_n^2 = R_q\{\log(p)/n\}^{1-q/2}$. Theorem 3 and theorem 4 below imply that, with high probability,

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^*\|_2 &\leqslant \|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}\|_2 + \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha\|_2 + \|\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*\|_2\\ &\leqslant \sqrt{d_3 r_n}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha\|_2^2 + r_n^2)^{1/2} + d_2 r_n + d_1\alpha^{k-1}\\ &\leqslant \{d_3(d_2^2+1)\}^{1/2} r_n^2 + d_2 r_n + d_1\alpha^{k-1}\\ &\leqslant 2d_2 r_n + d_1\alpha^{k-1},\end{aligned}$$

when the sample size is sufficiently large to ensure that $r_n \leqslant d_2\{d_3(d_2^2+1)\}^{-1/2}$. Therefore, $\|\hat{\boldsymbol{\beta}}^t - \boldsymbol{\beta}^*\|_2$ has the same rate as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$. Hence, from a statistical point of view, there is no need to iterate beyond t steps.

Theorem 4. Under the conditions of theorem 3, suppose that we choose λ_n as in lemma 1 and also satisfying

$$\lambda_n \geqslant \frac{32\rho}{1-\kappa}\left\{1 - \frac{64\kappa_{\mathrm{u}}|S_{\alpha\eta}|\log(p)}{n\bar{\gamma}_{\mathrm{l}}}\right\}^{-1}\left[1 + \kappa_1\kappa_2^2\left\{\frac{\bar{\gamma}_{\mathrm{l}}}{12\kappa_{\mathrm{u}}} + \frac{128\kappa_{\mathrm{u}}|S_{\alpha\eta}|\log(p)}{n\bar{\gamma}_{\mathrm{l}}}\right\} + 8\kappa_{\mathrm{u}}\right]\frac{\log(p)}{n},$$

where $|S_{\alpha\eta}|$ denotes the cardinality of set $S_{\alpha\eta}$ and $\bar{\gamma}_{\mathrm{l}} = \gamma_{\mathrm{l}} - 64\tau_{\mathrm{l}}|S_{\alpha\eta}|$; then, with probability at least $1 - c_1\exp(-c_2 n)$, there is a generic positive constant d_3 such that

$$\|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}\|_2^2 \leqslant d_3 R_q\left\{\frac{\log(p)}{n}\right\}^{1-q/2}\left[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha\|_2^2 + R_q\left\{\frac{\log(p)}{n}\right\}^{1-q/2}\right], \tag{3.4}$$

for all iterations

$$t \geqslant \frac{2\log[\{\phi_n(\hat{\boldsymbol{\beta}}^0) - \phi_n(\hat{\boldsymbol{\beta}})\}/\delta^2]}{\log(1/\kappa)} + \log_2\left\{\log_2\left(\frac{\rho\lambda_n}{\delta^2}\right)\right\}\left\{1 + \frac{\log(2)}{\log(1/\kappa)}\right\},$$

where $\phi_n(\boldsymbol{\beta}) = \mathcal{L}_n(\boldsymbol{\beta}) + \lambda_n\|\boldsymbol{\beta}\|_1$ and $\hat{\boldsymbol{\beta}}^0$ is the initial value satisfying $\|\hat{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}^*\|_2 \leqslant \rho_2$, $\delta = \varepsilon^2/(1-\kappa)$ is the tolerance level and κ and ε are some constants as will be defined in expressions (19) and (20) in the on-line supplementary file respectively.

4. Robust estimation of mean and covariance matrix

Estimation of the mean can be regarded as a univariate linear regression where the covariate equals 1. In that special case, we have a more explicit concentration result for the RA mean estimator, which is the estimator that minimizes the RA quadratic loss. Let $\{y_i\}_{i=1}^n$ be an independent and identically distributed sample from some unknown distribution with $E(y_i) = \mu$ and $\mathrm{var}(y_i) = \sigma^2$. The RA mean estimator $\hat{\mu}_\alpha$ of μ is the solution of

$$\sum_{i=1}^n \psi\{\alpha(y_i - \mu)\} = 0,$$

for parameter $\alpha \to 0$, where the influence function $\psi(x) = x$ if $|x| \leqslant 1, \psi(x) = 1$ if $x > 1$ and $\psi(x) = -1$ if $x < -1$. The following theorem gives the exponential type of concentration of $\hat{\mu}_\alpha$ around μ.

Theorem 5. Assume that $\log(1/\delta)/n \leqslant \frac{1}{8}$ and let $\alpha = \sqrt{\{\log(1/\delta)/(nv^2)\}}$ where $v \geqslant \sigma$. Then,

$$P\left[|\hat{\mu}_\alpha - \mu| \geqslant 4v\sqrt{\left\{\frac{\log(1/\delta)}{n}\right\}}\right] \leqslant 2\delta.$$

This result provides fast concentration of the mean estimation with only two moments assumption. This is very useful for large-scale hypothesis testing (Efron, 2010; Fan *et al.*, 2012) and covariance matrix estimation (Bickel and Levina, 2008; Fan *et al.*, 2013), where uniform convergence is required. Taking the estimation of a large covariance matrix as an example, for the elements of the sample covariance matrix to converge uniformly, Bickel and Levina (2008) and Fan *et al.* (2013) required the underlying multivariate distribution to be sub-Gaussian. This restrictive assumption can be removed if we apply the robust estimation with concentration bound. Regarding $\sigma_{ij} = E(X_i X_j)$ as the expected value of the random variable $X_i X_j$ (it is typically not the same as the median of $X_i X_j$), it can be estimated with accuracy

$$P\left[|\hat{\sigma}_{ij} - \sigma_{ij}| \geqslant 4v\sqrt{\left\{\frac{\log(1/\delta)}{n}\right\}}\right] \leqslant 2\delta,$$

where $v \geqslant \max_{i,j \leqslant p}\sqrt{\mathrm{var}(X_i X_j)}$ and $\hat{\sigma}_{ij}$ is the RA mean estimator using data $\{X_{ik}X_{jk}\}_{k=1}^n$. Since there are only $O(p^2)$ elements, by taking $\delta = p^{-a}$ for some $a > 2$ and the union bound, we have

$$P\left[\max_{i,j \leqslant p}|\hat{\sigma}_{ij} - \sigma_{ij}| \geqslant 4v\sqrt{\left\{\frac{a\log(p)}{n}\right\}}\right] \leqslant 2p^{2-a},$$

when $\max_{i \leqslant p} E(X_i^4)$ is bounded. This robustified covariance estimator requires a much weaker condition than the sample covariance and has far wider applicability than the sample covariance. It can be regularized further in the same way as the sample covariance matrix.

5. Connection with Catoni loss

Catoni (2012) considered estimation of the mean of heavy-tailed distributions with fast concentration. He proposed an M-estimator by solving

$$\sum_{i=1}^n \psi_c\{\alpha(y_i - \theta)\} = 0,$$

where the influence function $\psi_c(x)$ is chosen such that $-\log(1 - x + x^2/2) \leqslant \psi_c(x) \leqslant \log(1 + x + x^2/2)$. He showed that this M-estimator has the exponential type of concentration by only requiring the existence of the variance. It performed as well as the sample mean under the light tail case.

Catoni's idea can also be extended to the linear regression setting. Suppose that we replace the RA quadratic loss $l_\alpha(x)$ in expression (2.3) with Catoni loss,

$$l_\alpha^c(x) = \frac{2}{\alpha}\int_0^x \psi_c(\alpha t)\mathrm{d}t,$$

where the influence function $\psi_c(t)$ is given by

$$\psi_c(t) = \mathrm{sgn}(t)\{-\log(1 - |t| + t^2/2)I(|t| < 1) + \log(2)I(|t| \geqslant 1)\}.$$

Let $\hat{\boldsymbol{\beta}}^{\mathrm{c}}$ be the orresponding solution. Then, $\hat{\boldsymbol{\beta}}^{\mathrm{c}}$ has the same non-asymptotic upper bound as the RA lasso, which is stated as follows.

Theorem 6. Suppose that condition 1 holds for $k=2$ or $k=3$, and conditions 2, 3 and (2.7) hold. Then there are generic positive constants c_1, c_2, d_4 and d_5, depending on M_k, κ_0, κ_{l}, κ_{u} and κ_λ, such that, with probability at least $1-c_1\exp(-c_2 n)$,

$$\|\hat{\boldsymbol{\beta}}^{\mathrm{c}}-\boldsymbol{\beta}^*\|_2 \leqslant d_4\alpha^{k-1}+d_5\sqrt{R_q}\{\log(p)/n\}^{1/2-q/4}.$$

Unlike the RA lasso, the order of bias of $\hat{\boldsymbol{\beta}}^{\mathrm{c}}$ cannot be further improved, even when higher conditional moments of errors exist beyond the third order. The reason is that the Catoni loss is not exactly the quadratic loss over any finite intervals. Similar results regarding the computational error of $\hat{\boldsymbol{\beta}}^{\mathrm{c}}$ can also be established as in theorem 4, since the RSC and restricted smoothness conditions also hold for Catoni loss with uniform constants.

6. Variance estimation

We estimate the unconditional variance $\sigma^2=E(\epsilon^2)$ based on the RA lasso estimator and a cross-validation scheme. To ease the presentation, we assume that the data set can be evenly divided into J folds with m observations in each fold. Then, we estimate σ^2 by

$$\hat{\sigma}^2=\frac{1}{J}\sum_{j=1}^{J}\frac{1}{m}\sum_{i\in\mathrm{fold}\,j}(y_i-\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(-j)})^2,$$

where $\hat{\boldsymbol{\beta}}^{(-j)}$ is the RA lasso estimator obtained by using data points outside the jth fold. We show that $\hat{\sigma}^2$ is asymptotically efficiently. Differently from the existing cross-validation-based method (Fan *et al.*, 2012), a light tail assumption is not needed because of the utilization of the RA lasso estimator.

Theorem 7. Under the conditions of theorem 3, if $R_q\log(p)^{1-q/2}/n^{(1-q)/2}\to 0$ for $q\in(0,1)$, and $\alpha=O([R_q\{\log(p)/n\}^{1-q/2}]^{1/\{2(k-1)\}})$, then

$$\sqrt{n}(\hat{\sigma}^2-\sigma^2)\xrightarrow{\mathcal{D}} N\{0,E(\epsilon^4)-\sigma^4\}.$$

7. Simulation studies

In this section, we assess the finite sample performance of the RA lasso and compare it with other methods through various models. We simulated data from the high dimensional model

$$y_i=\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*+\epsilon_i,\qquad \mathbf{x}_i\sim N(0,I_p),\tag{7.1}$$

where we generated $n=100$ observations and the number of parameters was chosen to be $p=400$. We chose the true regression coefficient vector as

$$\boldsymbol{\beta}^*=(3,\ldots,3,0,\ldots,0)^{\mathrm{T}},$$

where the first 20 elements are all equal to 3 and the rest are all equal to 0. To involve various shapes of error distributions, we considered the following five scenarios:

(a) normal errors with mean 0 and variance 4 ($N(0,4)$);
(b) two times the t-distribution with degrees of freedom 3 ($2t_3$);
(c) a mixture of normal distributions, MixN, $0.5N(-1,4)+0.5N(8,1)$;

Table 1. Summary of the shapes and tails of the five error distributions

	Light tail	*Heavy tail*
Symmetric	$N(0,4)$	$2t_3$
Asymmetric	MixN	LogNormal, Weibull

(d) a log-normal distribution, LogNormal, $\epsilon = \exp(1 + 1.2Z)$, where Z is the standard normal distribution;
(e) a Weibull distribution with shape parameter 0.3 and scale parameter 0.5.

To meet the model assumptions, the errors were standardized to have mean 0. Table 1 categorizes the five scenarios according to the shapes and tails of the error distributions.

To obtain our estimator, we iteratively applied the gradient descent algorithm. We compared the RA lasso with two other methods in a high dimensional setting:

(a) lasso, the penalized least squares estimator with L_1-penalty as in Tibshirani (1996);
(b) R-Lasso, the R-Lasso estimator in Fan *et al*. (2014), which is the same as the regularized LAD estimator with L_1-penalty as in Wang (2013).

Their performance under the five scenarios was evaluated by the following four measurements:

(a) L_2-error, which is defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$;
(b) L_1-error, which is defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$;
(c) the number of false positive results, FP, which is the number of noise covariates that are selected;
(d) the number of false negative results, FN, which is the number of signal covariates that are not selected.

We also measured the relative gain of the RA lasso with respect to R-Lasso and the lasso, in terms of the difference from the oracle estimator. The oracle estimator $\hat{\boldsymbol{\beta}}_{\text{oracle}}$ is defined to be the least square estimator by using the first 20 covariates only. Then, the relative gains of the RA lasso with respect to the lasso, $\text{RG}_{\text{A,L}}$, in the L_2- and L_1-norm are defined as

$$\frac{\|\hat{\boldsymbol{\beta}}_{\text{lasso}} - \boldsymbol{\beta}^*\|_2 - \|\hat{\boldsymbol{\beta}}_{\text{oracle}} - \boldsymbol{\beta}^*\|_2}{\|\hat{\boldsymbol{\beta}}_{\text{RA lasso}} - \boldsymbol{\beta}^*\|_2 - \|\hat{\boldsymbol{\beta}}_{\text{oracle}} - \boldsymbol{\beta}^*\|_2},$$
$$\frac{\|\hat{\boldsymbol{\beta}}_{\text{lasso}} - \boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}_{\text{oracle}} - \boldsymbol{\beta}^*\|_1}{\|\hat{\boldsymbol{\beta}}_{\text{RA lasso}} - \boldsymbol{\beta}^*\|_1 - \|\hat{\boldsymbol{\beta}}_{\text{oracle}} - \boldsymbol{\beta}^*\|_1}.$$

The relative gain of the RA lasso with respect to R-Lasso, $\text{RG}_{\text{A,R}}$, is defined similarly.

For the RA lasso, the tuning parameters λ_n and α were chosen optimally on the basis of 100 independent validation data sets. We ran a two-dimensional grid search to find the best (λ_n, α) pair that minimizes the mean L_2-loss of the 100 validation data sets. Such an optimal pair was then used in the simulations. A similar method was applied in choosing the tuning parameters in the lasso and R-Lasso.

The above simulation model is based on the additive model (7.1), in which the error distribution is independent of covariates. However, this homoscedastic model makes the conditional

Table 2. Simulation results for the lasso, R-lasso and RA lasso under homoscedastic model (7.1)

Scenario		*Results for the following methods:*				
		Lasso	*R-Lasso*	*RA lasso*	$RG_{\mathrm{A,L}}$	$RG_{\mathrm{A,R}}$
$N(0,4)$	L_2-loss	4.54	4.40	4.53	1.00	0.96
	L_1-loss	27.21	29.11	27.21	1.00	1.08
	FP, FN	52.10, 0.09	66.36, 0.17	52.10, 0.09		
$2t_3$	L_2-loss	6.04	5.10	5.47	1.14	0.91
	L_1-loss	35.22	33.07	30.42	1.19	1.10
	FP, FN	47.13, 0.34	65.84, 0.22	41.34, 0.28		
MixN	L_2-loss	6.14	6.44	6.13	1.00	1.06
	L_1-loss	40.46	46.18	38.48	1.06	1.23
	FP, FN	65.99, 0.34	80.31, 0.33	58.05, 0.39		
LogNormal	L_2-loss	11.08	12.16	10.10	1.14	1.30
	L_1-loss	53.17	57.18	51.58	1.04	1.14
	FP, FN	26.5, 15.00	27.20, 6.90	37.20, 3.90		
Weibull	L_2-loss	7.77	7.11	6.62	1.23	1.10
	L_1-loss	55.65	50.49	42.93	1.34	1.20
	FP, FN	78.70, 0.71	77.13, 0.56	62.27, 0.52		

Table 3. Simulation results of the lasso, R-lasso and RA lasso under heteroscedastic model (7.2)

Scenario		*Results for the following methods:*				
		Lasso	*R-Lasso*	*RA lasso*	$RG_{\mathrm{A,L}}$	$RG_{\mathrm{A,R}}$
$N(0,4)$	L_2-loss	4.60	4.34	4.60	1.00	0.93
	L_1-loss	27.16	27.14	27.15	1.00	1.00
	FP, FN	48.78, 0.10	58.25, 0.27	48.78, 0.10		
$2t_3$	L_2-loss	8.08	6.71	6.70	1.26	1.01
	L_1-loss	41.16	42.76	38.52	1.08	1.12
	FP, FN	55.33, 0.67	71.67, 0.33	45.33, 0.33		
MixN	L_2-loss	6.26	6.54	6.25	1.00	1.06
	L_1-loss	41.26	46.95	39.25	1.06	1.23
	FP, FN	65.98, 0.34	80.30, 0.32	58.80, 0.34		
LogNormal	L_2-loss	10.86	9.19	8.48	1.43	1.13
	L_1-loss	57.52	57.18	53.20	1.10	1.09
	FP, FN	29.70, 5.70	54.10, 2.00	54.30, 1.50		
Weibull	L_2-loss	7.40	8.81	5.53	1.53	1.92
	L_1-loss	40.95	47.82	34.65	1.23	1.48
	FP, FN	38.87, 0.96	35.31, 2.90	58.15, 0.39		

mean and the conditional median differ by only a constant. To examine the deviations between the mean regression and median regression further, we also simulated the data from the heteroscedastic model

$$y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + c^{-1}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*)^2\epsilon_i, \qquad \mathbf{x}_i \sim N(0, I_p), \tag{7.2}$$

where the constant $c = \sqrt{3}\|\boldsymbol{\beta}^*\|^2$ makes $E\{c^{-1}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^*)^2\}^2 = 1$. $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* \sim N(0, \|\boldsymbol{\beta}^*\|^2)$ and therefore c is chosen so that the average noise level is the same as that of ϵ_i. For both the homoscedastic

Table 4. Genes selected by the lasso, R-Lasso and RA lasso

Genes selected by the following methods:								
Lasso	*R-Lasso*			*RA Lasso*				
CRK	CSF3	DAPK2	EPOR	CSF3	IL10	IFI6	CR2	FYN
0.23	−2.46	0.7	−0.17	−2.95	1.52	0.86	0.57	−0.24
	IL10	TOLLIP	TJP1	CD3E	MAP2K4	TLR1	IL2	EPOR
	2.24	−0.68	−0.12	2.67	1.17	0.82	−0.47	0.24
	AKT1	TLR1	GAB2	BTK	PMAIP1	PSMB8	PSMC2	MASP1
	1.68	0.52	−0.01	2.37	−1.14	0.79	0.38	−0.24
	KPNB1	TLR3		CLSPN	BCL2L11	KPNB1	HSPA8	PRKCZ
	1.49	0.33		1.93	−1.13	0.77	−0.35	0.24
	TLR2	SHC1		RELA	AKT3	IFNG	SHC1	TOLLIP
	1.41	−0.28		1.88	−1.01	−0.74	−0.33	−0.19
	GRB2	PSMD1		AKT1	DUSP10	FADD	SPI1	BAK1
	−1.06	0.27		1.61	0.97	0.65	−0.28	0.14
	MAPK1	F12		IRS2	IRF4	TJP1	IFNA6	
	0.98	0.24		1.55	−0.95	−0.57	0.28	

and the heteroscedastic models, we ran 100 simulations for each scenario. The mean of each performance measurement is reported in Table 2 and Table 3 respectively.

Tables 2 and 3 indicate that our method had the biggest advantage when the errors were asymmetric and heavy tailed (the LogNormal and Weibull methods). In this case, R-Lasso had larger L_1- and L_2-errors owing to the bias for estimating the conditional median instead of the mean. Even though the lasso did not have bias in the loss component (quadratic loss), it did not perform well owing to its sensitivity to outliers. The advantage of our method is more pronounced in the heteroscedastic model than in the homoscedastic model. Both of them clearly indicate that, if the errors come from asymmetric and heavy-tailed distributions, our method is better than both the lasso and R-Lasso. When the errors were symmetric and heavy tailed (method $2t_3$), our estimator performed closely to the R-Lasso method, and both outperformed the lasso. These two cases evidently showed that the RA lasso was robust to outliers and did not lose efficiency when the errors were indeed symmetric. Under the light-tailed scenario, if the errors were asymmetric (method MixN), our method performed similarly to the lasso. The R-Lasso method performed worse, since it had bias. For the regular setting ($N(0, 4)$), where the errors were light tailed and symmetric, the three methods were comparable with each other.

In conclusion, the RA lasso is more flexible than the lasso and R-Lasso. The tuning parameter α automatically adapts to errors with different shapes and tails. It enables the RA lasso to render consistently satisfactory results under all scenarios.

8. Real data example

In this section, we use a microarray data set to illustrate the performance of the lasso, R-Lasso and RA lasso. Huang *et al*. (2011) studied the role of the innate immune system on the development of atherosclerosis by examining gene profiles from peripheral blood of 119 patients. The data were collected by using an Illumina HumanRef8 V2.0 Bead Chip and are available from the gene expression omnibus. The original study showed that the toll-like receptor (TLR) signalling pathway plays an important role in triggering the innate immune system in the face

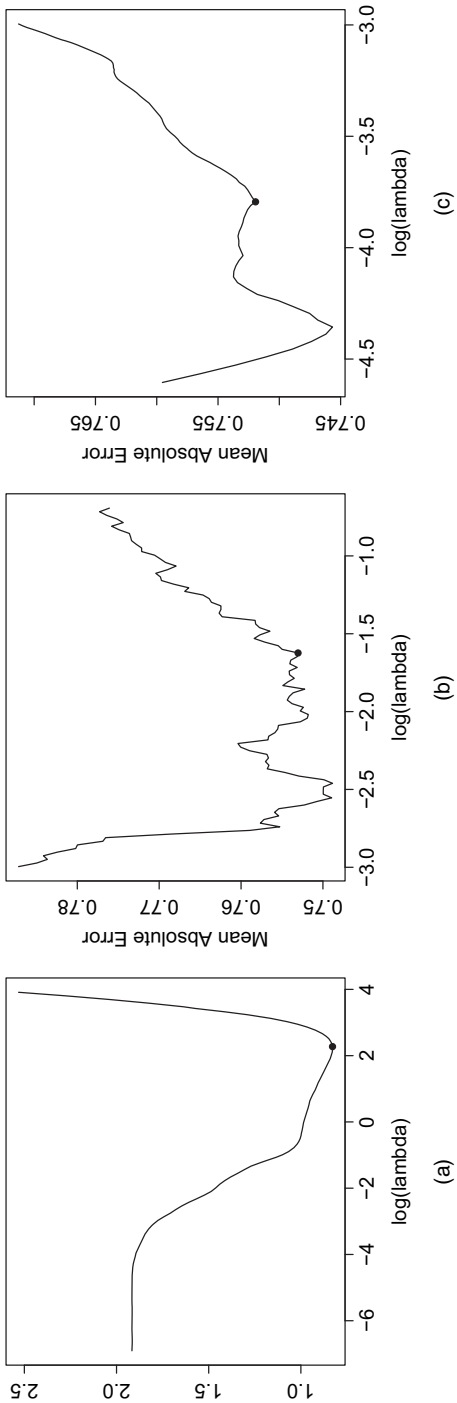

Fig. 1. Fivefold cross-validation results (•, choice of the penalization parameter): (a) lasso; (b) R-Lasso; (c) RA lasso

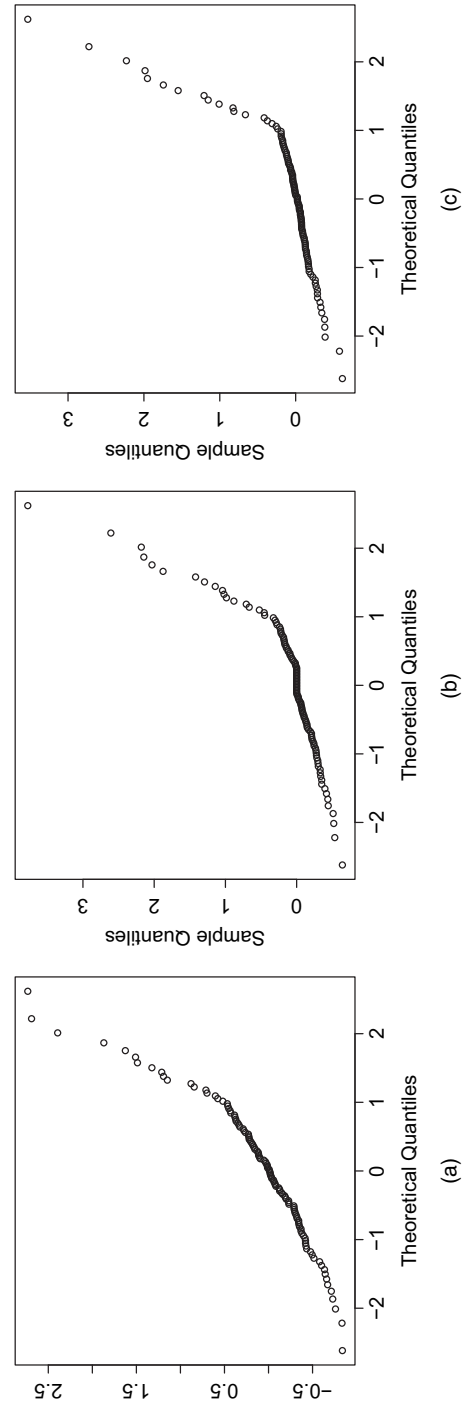

Fig. 2. *QQ*-plots of the residuals from the three methods: (a) lasso; (b) R-Lasso; (c) RA lasso

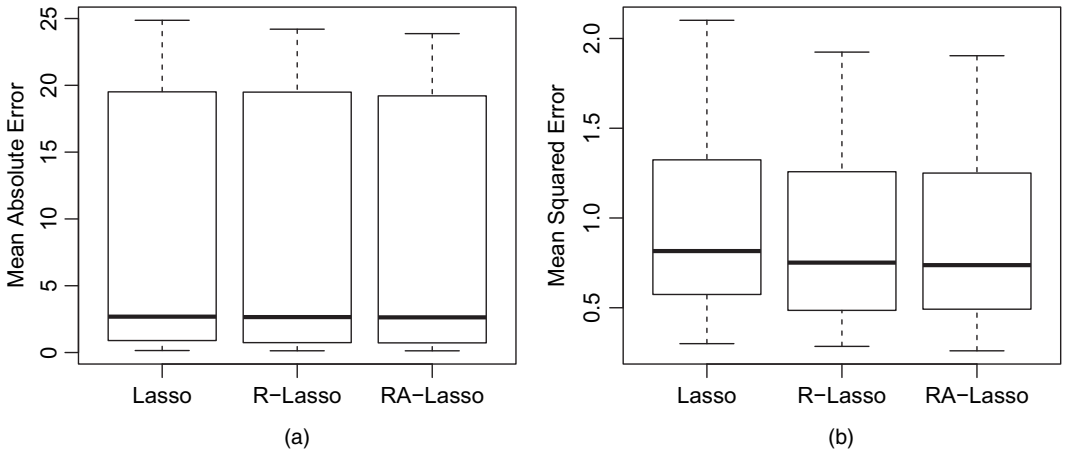

Fig. 3. Boxplots of (a) mean absolute errors of prediction and (b) mean-squared errors of prediction

of atherosclerosis. Under this pathway, the 'TLR8' gene was found to be a key atherosclerosis-associated gene. To study further the relationship between this key gene and the other genes, we regressed it on another 464 genes from 12 different pathways (TLR, CCC, CIR, IFNG, MAPK, RAPO, EXAPO, INAPO, DRS, NOD, EPO and CTR) that are related to the TLR pathway. We applied the lasso, R-Lasso and RA lasso to these data. The tuning parameters for all the methods were chosen by using fivefold cross-validation. Fig. 1 shows our choice of the penalization parameter based on the cross-validation results. For the RA lasso, the choice of α was insensitive to the results and was fixed at 5. We then applied the three methods with the above choice of tuning parameters to select significant genes. The *QQ*-plots of the residuals from the three methods are shown in Fig. 2. The genes selected by the three methods are reported in Table 4. After the selection, we regressed the expression of the TLR8 gene on the selected genes; the t-values from the refittings are also reported in Table 4.

Table 4 shows that the lasso selected only one gene. The R-Lasso method selected 17 genes. Our proposed RA lasso selected 34 genes. Eight genes (CSF3, IL10, AKT1, TOLLIP, TLR1, SHC1, EPOR and TJP1) that were found by R-Lasso were also selected by the RA lasso. Compared with the lasso and R-Lasso, our method selected more genes, which could be useful for a second-stage confirmatory study. It is clearly seen from Fig. 2 that the residuals from the fitted regressions had a heavy right-hand tail and a skewed distribution. We learn from the simulation studies in Section 7 that the RA lasso tends to perform better than the lasso and R-Lasso in this situation. For further investigation, we randomly chose 24 patients as the test set; we applied three methods to the rest of the patients to obtain the estimated coefficients, which in return were used to predict the responses of 24 patients. We repeated the random splitting 100 times; the boxplots of the mean absolute and mean-squared error of predictions are shown in Fig. 3. The RA lasso made better predictions than the lasso and R-Lasso.

Acknowledgements

The authors thank the Joint Editor, the Associate Editor and two referees for their valuable comments, which led to great improvement of the paper.

This work was supported in part by National Science Foundation grants DMS-1206464 and DMS-1406266 and National Institutes of Health grants R01-GM072611-9 and R01-GM100474-4.

Appendix A: Proofs of theorems 1, 2 and 5

A.1. Proof of theorem 1

Let $l(x) = x^2$. Since $\boldsymbol{\beta}^*$ minimizes $E\{l(y - \mathbf{x}^{\mathrm{T}}\beta)\}$, it follows from condition 2 that

$$E\{l(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\alpha^*) - l(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*)\} = (\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)^{\mathrm{T}} E(\mathbf{x}\mathbf{x}^{\mathrm{T}})(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*) \geqslant \kappa_1 \|\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*\|_2^2. \tag{A.1}$$

Let $g_\alpha(x) = l(x) - l_\alpha(x) = (|x| - \alpha^{-1})^2 I(|x| > \alpha^{-1})$. Then, since $\boldsymbol{\beta}_\alpha^*$ is the minimizer of $E\{l_\alpha(y - \mathbf{x}^{\mathrm{T}}\beta)\}$, we have

$$\begin{aligned} E\{l(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\alpha^*) - l(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*)\} &= E\{l(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\alpha^*) - l_\alpha(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\alpha^*)\} + E\{l_\alpha(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\alpha^*) - l_\alpha(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*)\} \\ &\quad + E\{l_\alpha(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*) - l(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*)\} \\ &\leqslant E\{g_\alpha(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\alpha^*)\} - E\{g_\alpha(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*)\}. \end{aligned}$$

By Taylor's expansion, we have

$$E\{l(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\alpha^*) - l_\alpha(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\alpha^*)\} \leqslant 2E\{(z - \alpha^{-1}) I(z > \alpha^{-1}) |\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)|\}, \tag{A.2}$$

where $\tilde{\boldsymbol{\beta}}$ is a vector lying between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}_\alpha^*$ and $z = |y - \mathbf{x}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}|$. With P_ϵ denoting the distribution of ϵ conditioning on $\mathbf{x}$ and E_ϵ the corresponding expectation, we have

$$\begin{aligned} E_\epsilon\{(z - \alpha^{-1}) I(z > \alpha^{-1})\} &= \int_0^\infty P_\epsilon\{z I(z > \alpha^{-1}) > t\}\, \mathrm{d}t - \alpha^{-1} P_\epsilon(z > \alpha^{-1}) \\ &= \int_0^\infty P_\epsilon(z > t \text{ and } z > \alpha^{-1})\, \mathrm{d}t - \alpha^{-1} P_\epsilon(z > \alpha^{-1}) \\ &= \int_{\alpha^{-1}}^\infty P_\epsilon(z > t)\, \mathrm{d}t + \int_0^{\alpha^{-1}} P_\epsilon(z > \alpha^{-1})\, \mathrm{d}t - \alpha^{-1} P_\epsilon(z > \alpha^{-1}) \\ &\leqslant \int_{\alpha^{-1}}^\infty \frac{E_\epsilon(z^k)}{t^k}\, \mathrm{d}t \leqslant \alpha^{k-1} E_\epsilon(z^k). \end{aligned}$$

Therefore, $E\{l(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_\alpha^*) - l(y - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}^*)\}$ is further bounded by

$$\begin{aligned} 2\alpha^{k-1} E\{|y - \mathbf{x}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}|^k |\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)|\} &= 2\alpha^{k-1} E\{|\epsilon + \mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}})|^k |\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)|\} \\ &= 2(2\alpha)^{k-1} [E\{|\epsilon|^k |\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)|\} + E\{|\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}})|^k |\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)|\}]. \end{aligned} \tag{A.3}$$

Note that

$$\begin{aligned} E\{|\epsilon|^k |\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)|\} &= E\{E(|\epsilon|^k |\mathbf{x}) |\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)|\} \leqslant [E\{E(|\epsilon|^k |\mathbf{x})\}^2]^{1/2} [E|\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)|^2]^{1/2} \\ &\leqslant \sqrt{(M_k \kappa_u)} \|\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*\|_2, \end{aligned}$$

where the last inequality follows from conditions 1 and 2. However, by condition 3, $\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}})$ is sub-Gaussian; hence its $2k$th moment is bounded by $c^2 \kappa_0^{2k}$, for a universal positive constant c depending on k only. Then,

$$\begin{aligned} E\{|\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}})|^k |\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)|\} &\leqslant \{E|\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}})|^{2k}\}^{1/2} \{E|\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)|^2\}^{1/2} \\ &\leqslant c \kappa_0^k \sqrt{\kappa_u} \|\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*\|_2. \end{aligned}$$

These results together with equation (A.1) and (A.3) complete the proof.

A.2. Proof of theorem 2

Let A_1 and A_2 denote the events that lemma 1 and lemma 3 hold respectively. By theorem 1 of Negahban

et al. (2012), within $A_1 \cap A_2$, it holds that

$$\begin{aligned}\|\boldsymbol{\Delta}\|_2^2 &\leqslant 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}^2}|S_{\alpha\eta}| + \frac{\lambda_n}{\kappa_{\mathcal{L}}^2}(2\tau_{\mathcal{L}}^2 + 4\|\boldsymbol{\beta}^*_{S^c_{\alpha\eta}}\|_1)\\ &\leqslant \frac{36\lambda_n^2 R_q}{\kappa_1^2\eta^q} + \frac{4\lambda_n}{\kappa_1^2}\left[8R_q\kappa_1\kappa_2{\kappa_\lambda}^{1-q}\left\{\frac{\log(p)}{n}\right\}^{1-q/2} + 4R_q\eta^{1-q}\right]\\ &\overset{\text{(i)}}{=} \frac{36}{\kappa_1^2}R_q\lambda_n^{2-q} + \frac{16}{\kappa_1^2}R_q\lambda_n^{2-q}\left[2\kappa_1\kappa_2\left\{\frac{\log(p)}{n}\right\}^{1/2} + 1\right]\\ &\overset{\text{(ii)}}{\leqslant} C_2 k_1^{-2}\kappa_\lambda^{2-q}R_q\{n^{-1}\log(p)\}^{1-q/2},\end{aligned}$$

where equation (i) follows from the choice of $\eta = \lambda_n$ and in inequality (ii) we assume that the sample size n is sufficiently large that $2\kappa_1\kappa_2\{n^{-1}\log(p)\}^{1/2} \leqslant 1$ and observe that $\kappa_1 = \kappa_{\mathrm{l}}/4$. In contrast, by lemmas 1 and 3, $P(A_1 \cap A_2) \geqslant 1 - c_1\exp(-c_2 n)$, where $c_1 = \max\{2, c_1'\}$ and $c_2 = \min\{c_0, c_2'\}$.

A.3. Proof of theorem 5

The proof of theorem 5 follows the same spirit as the proof of proposition 2.4 in Catoni (2012). The influence function $\psi(x)$ of RA quadratic loss satisfies

$$-\log(1 - x + x^2) \leqslant \psi(x) \leqslant \log(1 + x + x^2).$$

Using this and independence, with $r(\theta) = \{1/(\alpha n)\}\,\Sigma_{i=1}^n \psi\{\alpha(Y_i - \theta)\}$, we have

$$\begin{aligned}E[\exp\{\alpha n r(\theta)\}] &\leqslant E(\exp[\psi\{\alpha(Y_i - \theta)\}])^n\\ &\leqslant [1 + \alpha(\mu - \theta) + \alpha^2\{\sigma^2 + (\mu - \theta)^2\}]^n\\ &\leqslant \exp[n\alpha(\mu - \theta) + n\alpha^2\{v^2 + (\mu - \theta)^2\}].\end{aligned}$$

Similarly, $E[\exp\{-\alpha n r(\theta)\}] \leqslant \exp[-n\alpha(\mu - \theta) + n\alpha^2\{v^2 + (\mu - \theta)^2\}]$. Define

$$\begin{aligned}B_+(\theta) &= \mu - \theta + \alpha\{v^2 + (\mu - \theta)^2\} + \frac{\log(1/\delta)}{n\alpha},\\ B_-(\theta) &= \mu - \theta - \alpha\{v^2 + (\mu - \theta)^2\} - \frac{\log(1/\delta)}{n\alpha}.\end{aligned}$$

By the Chebyshev inequality,

$$P\{r(\theta) > B_+(\theta)\} \leqslant \frac{E[\exp\{\alpha n r(\theta)\}]}{\exp[\alpha n(\mu - \theta) + n\alpha^2\{v^2 + (\mu - \theta)^2\} + \log(1/\delta)]} \leqslant \delta.$$

Similarly, $P\{r(\theta) < B_-(\theta)\} \leqslant \delta$.

Let θ_+ be the smallest solution of the quadratic equation $B_+(\theta_+) = 0$ and θ_- be the largest solution of the equation $B_-(\theta_-) = 0$. Under the assumption that $\log(1/\delta)/n \leqslant \frac{1}{8}$ and the choice of $\alpha = \sqrt{\{\log(1/\delta)/nv^2\}}$, we have $\alpha^2 v^2 + \log(1/\delta)/n \leqslant \frac{1}{4}$. Therefore,

$$\begin{aligned}\theta_+ &= \mu + 2\left\{\alpha v^2 + \frac{\log(1/\delta)}{\alpha n}\right\}\left(1 + \sqrt{\left[1 - 4\left\{\alpha^2 v^2 + \frac{\log(1/\delta)}{n}\right\}\right]}\right)^{-1}\\ &\leqslant \mu + 2\left\{\alpha v^2 + \frac{\log(1/\delta)}{\alpha n}\right\}.\end{aligned}$$

Similarly,

$$\theta_- = \mu - 2\left\{\alpha v^2 + \frac{\log(1/\delta)}{\alpha n}\right\}\left(1 + \sqrt{\left[1 - 4\left\{\alpha^2 v^2 + \frac{\log(1/\delta)}{n}\right\}\right]}\right)^{-1}$$

$$\geqslant \mu - 2\left\{\alpha v^2 + \frac{\log(1/\delta)}{\alpha n}\right\}.$$

With $\alpha = \sqrt{\{\log(1/\delta)/(nv^2)\}}$, $\theta_+ \leqslant \mu + 4v\sqrt{\{\log(1/\delta)/n\}}$, and $\theta_- \geqslant \mu - 4v\sqrt{\{\log(1/\delta)/n\}}$. Since the map $\theta \mapsto r(\theta)$ is non-increasing, under event $\{B_-(\theta) \leqslant r(\theta) \leqslant B_+(\theta)\}$

$$\mu - 4v\sqrt{\left\{\frac{\log(1/\delta)}{n}\right\}} \leqslant \theta_- \leqslant \hat{\mu}_\alpha \leqslant \theta_+ \leqslant \mu + 4v\sqrt{\left\{\frac{\log(1/\delta)}{n}\right\}},$$

i.e. $|\hat{\mu}_\alpha - \mu| \leqslant 4v\sqrt{\{\log(1/\delta)/n\}}$. Meanwhile, $P\{B_-(\theta) \leqslant r(\theta) \leqslant B_+(\theta)\} > 1 - 2\delta$.

References

Agarwal, A., Negahban, S. and Wainwright, M. J. (2012) Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, **40**, 2452–2482.

Belloni, A. and Chernozhukov, V. (2011) L_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.*, **39**, 82–130.

Bickel, P. J. and Levina, E. (2008) Covariance regularization by thresholding. *Ann. Statist.*, **36**, 2577–2604.

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

Catoni, O. (2012) Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. H. Poincaré. Probab. Statist.*, **48**, 1148–1185.

Duchi, J., Shalev-Shwartz, S., Singer, Y. and Chandra, T. (2008) Efficient projections onto the L_1-ball for learning in high dimensions. In *Proc. 25th Int. Conf. Machine Learning*, pp. 272–279. New York: Association for Computing Machinery.

Efron, B. (2010) Correlated z-values and the accuracy of large-scale statistical estimates. *J. Am. Statist. Ass.*, **105**, 1042–1055.

Engle, R. F. (1982) Autoregressive conditional heteroscedasticy with estimates of the variance of U.K. inflation. *Econometrica*, **50**, 987–1008.

Fan, J., Fan, Y. and Barut, E. (2014) Adaptive robust variable selection. *Ann. Statist.*, **42**, 324–351.

Fan, J., Guo, S. and Hao, N. (2012) Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Statist. Soc.* B, **74**, 37–65.

Fan, J., Han, X. and Gu, W. (2012) Estimating the false discovery proportion under arbitrary covariance dependence (with discussion). *J. Am. Statist. Ass.*, **107**, 1019–1035.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fan, J., Liao, Y. and Mincheva, M. (2013) Large covariance estimation by thresholding principal orthogonal complements (with discussion). *J. R. Statist. Soc.* B, **75**, 603–680.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc.* B, **70**, 849–911.

Fan, J. and Lv, J. (2011) Non-concave penalized likelihood with NP-Dimensionality. *IEEE Trans. Inform. Theor.*, **57**, 5467–5484.

Huang, C. C., Liu, K., Pope, R. M., Du, P., Lin, S., Rajamannan, N. M., Huang, Q. Q., Jafar, N., Burke, G. L., Post, W., Watson, K. E., Johnson, C., Daviglus, M. L. and Lloyd-Jones, D. M. (2011) Activated TLR signaling in atherosclerosis among women with lower Framingham risk score: the multi-ethnic study of atherosclerosis. *PLOS ONE*, **6**, article e21067.

Huber, P. J. (1964) Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.

Lambert-Lacroix, S. and Zwald, L. (2011) Robust regression through the Huber's criterion and adaptive lasso penalty. *Electron. J. Statist.*, **5**, 1015–1053.

Li, Y. and Zhu, J. (2008) L_1-norm quantile regression. *J. Computnl Graph. Statist.*, **17**, 163–185.

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

Nesterov, Y. (2007) Gradient methods for minimizing composite objective function. *Technical Report 76*. Center for Operations Research and Econometrics, Catholic University of Louvain, Louvain-la-Neuve.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011) Minimax rates of estimation for high-dimensional linear regression over L_q-balls. *IEEE Trans. Inform. Theor.*, **57**, 6976–6994.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Wang, H. (2009) Forward regression for ultra-high dimensional variable screening. *J. Am. Statist. Ass.*, **104**, 1512–1524.

Wang, L. (2013) The L_1 penalized LAD estimator for high dimensional linear regression. *J. Multiv. Anal.*, **120**, 135–151.

Wu, Y. and Liu, Y. (2009) Variable selection in quantile regression. *Statist. Sin.*, **19**, 801–817.
Zhang, C.-H. (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.
Zou, H. and Yuan, M. (2008) Composite quantile regression and the oracle model selection theory. *Ann. Statist.*, **36**, 1108–1126.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'A supplementary file for "Estimation of high-dimensional mean regression in absence of symmetry and light-tail assumptions"'.