HHS Public Access

Author manuscript

J R Stat Soc Series B Stat Methodol. Author manuscript; available in PMC 2018 January 01.

Published in final edited form as:

J R Stat Soc Series B Stat Methodol. 2017 January ; 79(1): 247–265. doi:10.1111/rssb.12166.

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

Jianqing Fan, **Quefeng Li**, and **Yuyan Wang**

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544

Abstract

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 (LAD) 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 ultra-high dimensional setting 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 quadratic (RA-quadratic) loss will be called RA-Lasso. In the ultra-high 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 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 byproduct, we also establish the concentration inequality for estimating population mean when there exists only the second moment. We compare RA-Lasso with other regularized robust estimators based on quantile regression and LAD regression. Extensive simulation studies demonstrate the satisfactory finite-sample performance of 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 that 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

*Supported in part by NSF Grants DMS-1206464 and DMS-1406266 and NIH grants R01-GM072611-9 and NIH R01-GM100474-4.

on the unknown parameters (Lasso, Tibshirani (1996)), or a folded concave penalty function such as the SCAD (Fan and Li, 2001) and the MCP(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 noises can be large in those cases. As a result, theoretical properties are often obtained under light-tailed error distributions (Bickel, Ritov and Tsybakov, 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 the stagewise selection methods in high-dimension setting. 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 the L_1-penalized quantile regression in high dimensional sparse models. Fan, Fan, and Barut (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 the L_1-penalized LAD regression and showed that the estimator achieves near oracle risk performance under the high dimensional setting.

The above methods essentially estimate the conditional *median (or quantile)* regression, instead of the conditional *mean* regression function. In the applications where the 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 ARCH model (Engle, 1982). In these cases, the conditional mean and conditional median are very different. Another important example is to estimate large covariance matrix without assuming light-tails. We will explain this more in details 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? 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 the 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 converge if its reciprocal is used) in order to reduce the bias induced by the Huber loss when the distribution is asymmetric. In this paper, we consider the regularized approximate quadratic (RA-Lasso) estimator 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 0 (where the median and mean agree), this rate is the same as the regularized LAD estimator 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, 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 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 in order to reach the optimal rate.

As done in many other papers, estimators with nice sampling properties are typically defined through the optimization of a target function such as the penalized least-squares. The properties that are established are not necessarily the same as the ones that are computed. Following the framework of Agarwal, Negahban, and Wainwright (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 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 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 Section 7 and 8 to compare our method with two competitors. Proofs of Theorems 1 and 2 are given in the appendix, which together imply the main result (Theorem 3). Proof of Theorem 5 regarding the concentration of the robust mean estimator is also given in the appendix. Proofs of supporting lemmas and remaining theorems are given in an on-line supplementary file. The relevant matlab code is available at the site: http://orfe.princeton.edu/~jqfan/papers/15/RA-Lasso.zip.

2 RA-Lasso estimator

We consider the linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + \epsilon_i, \quad (2.1)$$

where $\{\mathbf{x}_i\}_{i=1}^n$ are independent and identically distributed (i.i.d) p-dimensional covariate vectors, $\{\epsilon_i\}_{i=1}^n$ are i.i.d errors, and $\boldsymbol{\beta}^*$ is a p-dimensional regression coefficient vector. The i.i.d assumption on $\boldsymbol{\varepsilon}_i$ indeed allows conditional heteroscedastic models, where $\boldsymbol{\varepsilon}_i$ can depend on $\mathbf{x}_i$. For example, it can be $\epsilon_i = \sigma(\mathbf{x}_i)\tilde{\epsilon}_i$, where $\boldsymbol{\sigma}(\mathbf{x}_i)$ 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 $\boldsymbol{\varepsilon}|\mathbf{x}$ are assumed to both 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 magnitude of errors and robustify the estimation, we propose to use the Huber loss (Huber, 1964):

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

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

To estimate $\boldsymbol{\beta}^*$, we propose to solve the following convex optimization problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{argmin} \frac{1}{n} \sum_{i=1}^n \ell_\alpha\left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right) + \lambda_n \sum_{j=1}^p |\beta_j|. \quad (2.3)$$

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 zero sufficiently fast, $\hat{\boldsymbol{\beta}}$ is a natural M-estimator of $\boldsymbol{\beta}_\alpha^* = \underset{\beta}{argmin}\, \mathrm{E}\, \ell_\alpha\left(y - \mathbf{x}'\boldsymbol{\beta}\right)$, which is the population minimizer under the RA-quadratic loss and varies by $\boldsymbol{\alpha}$. In general, $\boldsymbol{\beta}_\alpha^*$ differs from $\boldsymbol{\beta}^*$. But, since the RA-quadratic loss approximates the quadratic loss as $\boldsymbol{\alpha}$ tends to 0, $\boldsymbol{\beta}_\alpha^*$ is expected to converge to

β^*. This property will be established in Theorem 1. Therefore, we decompose the statistical error $\hat{\beta} - \beta^*$ into the approximation error $\beta^*_\alpha - \beta^*$ and the estimation error $\hat{\beta} - \beta^*_\alpha$. The statistical error $\|\hat{\beta} - \beta^*\|_2$ is then bounded by

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

In the following, we give upper bounds of the approximation and estimation error, respectively. We show that $\|\hat{\beta} - \beta^*\|_2$ 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 $\varepsilon|\mathbf{x}$. We assume that $\|\beta^*\|_2 \leq \rho_2$, where the radius ρ_2 is a sufficiently large constant. This is a mild assumption, which is implied by (C2) and a reasonable assumption that var(y) > ∞, since $var(y) \geq (\beta^*)^T \mathrm{E}\left(\mathbf{x}\mathbf{x}^T\right)\beta^* \geq \kappa_l \|\beta^*\|_2^2$.

Theorem 1

Under the following conditions:

(C1) $\mathrm{E}\{\mathrm{E}(|\varepsilon|^k|\mathbf{x})\}^2 \leq M_k < \infty$, *for some* $k \geq 2$.

(C2) $0 < \kappa_l \leq \lambda_{\min}(\mathrm{E}[\mathbf{x}\mathbf{x}^T]) \leq \lambda_{\max}(\mathrm{E}[\mathbf{x}\mathbf{x}^T]) \leq \kappa_u < \infty$,

(C3) *For any* $\nu \in$ P, $\mathbf{x}^T\nu$ *is sub-Gaussian with parameter at most* $\kappa_0^2\|\nu\|_2^2$, *i.e.* $\mathrm{E}\, exp\left(t\mathbf{x}^T\nu\right) \leq exp\left(t^2\kappa_0^2\|\nu\|_2^2/2\right)$, *for any* $t \in$,

there exosts a universal positive constant C_1, *such that*

$$\|\beta^*_\alpha - \beta^*\|_2 \leq C_1 \sqrt{\kappa_u}\kappa_l^{-1}\left(\kappa_0^k + \sqrt{M_k}\right)\alpha^{k-1}.$$

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

In order to be solvable in the high-dimensional setting, β^* is usually assumed to be sparse or weakly sparse, i.e. many elements of β^* are zero or small. By Theorem 1, β^*_α converges to β^* as α goes to 0. In view of this fact, we assume that β^*_α is uniformly weakly sparse when α is sufficiently small. In particular, we assume that there exists a small constant $r > 0$, such that β^*_α belongs to an L_q-ball with a uniform radius R_q that

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

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 β^* is in the L_q-ball. When the conditional distribution of ε_i is asymmetric, we give a sufficient condition showing that if β^* belongs to an L_q-ball with radius $R_q/2$, (2.4) holds for all $\alpha \le c\left\{R_q p^{-(2-q)/2}\right\}^{\frac{1}{q(k-1)}}$, where c is a positive constant. In fact, for any $q \in (0,1]$, $|r_1|^q+|r_2|^q \ge (|r_1|+|r_2|)^q \ge |r_1+r_2|^q$. Using this,

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

By Theorem 1, $\sum_{j=1}^{p}|\beta^*_{\alpha,j}-\beta^*_j|^2=O\left(\alpha^{2(k-1)}\right)$. Hence, if $\sum_{j=1}^{p}|\beta^*_j|^q \le R_q/2$, we have $\sum_{j=1}^{p}|\beta^*_{\alpha,j}|^q \le R_q$ for all $\alpha \le c\left\{R_q p^{-(2-q)/2}\right\}^{\frac{1}{q(k-1)}}$.

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

Lemma 1

Under Conditions (C1) and (C3), with the choice of $\lambda_n=\kappa_\lambda\sqrt{(\log p)/n}$ and $\alpha \ge L\lambda_n/(4v)$, where v and L are positive constants depending on M_2 *and* κ_0, *and* κ_λ *is a sufficiently large constant such that* $\kappa^2_\lambda>32v$, *it holds with probability greater than* $1-2\exp(-c_0 n)$ *that,*

$$\hat{\Delta}=\hat{\beta}-\beta^*_\alpha \in \mathbb{C}_{\alpha\eta} := \left\{\Delta \in {}^{\mathrm{p}}:\|\Delta_{\mathrm{S}^{\mathrm{c}}_{\alpha\eta}}\|_1 \le 3\|\Delta_{\mathrm{S}_{\alpha\eta}}\|_1+4\|\beta^*_{\alpha,\mathrm{S}^{\mathrm{c}}_{\alpha\eta}}\|_1\right\},$$

where $c_0=\kappa^2_\lambda/(32v)-1$, η is a positive constant, $S_{\alpha\eta}=\left\{j:|\beta^*_{\alpha,j}|>\eta\right\}$ *and* $\Delta S_{\alpha\eta}$ *denotes the subvector of* Δ *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*-estimator (Negahban, *et al.*, 2012; Agarwal, Negahban, and Wainwright, 2012). Let

$$\delta\mathscr{L}_n(\Delta,\beta)=\mathscr{L}_n(\beta+\Delta)-\mathscr{L}_n(\beta)-[\nabla\mathscr{L}_n(\beta)]^T\Delta, \tag{2.5}$$

where $\mathscr{L}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \ell_\alpha\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)$, $\boldsymbol{\Delta}$ is a p-dimensional vector and $\nabla\mathscr{L}_n(\boldsymbol{\beta})$ is the gradient of $\mathscr{L}_n$ at the point of $\boldsymbol{\beta}$.

Definition 1

The loss function $\mathscr{L}_n$ satisfies RSC condition on a set $\mathbf{S}$ with curvature $\kappa_{\mathscr{L}} > 0$ and tolerance $\tau_{\mathscr{L}}$ *if*

$$\delta\mathscr{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}) \geq \kappa_{\mathscr{L}} \|\boldsymbol{\Delta}\|_2^2 - \tau_{\mathscr{L}}^2, \quad for \quad all \quad \boldsymbol{\Delta} \in \mathbf{S}.$$

Next, we show that with high probability the RA-quadratic loss (2.2) satisfies RSC for $\boldsymbol{\beta}=\boldsymbol{\beta}_\alpha^*$ and all $\boldsymbol{\Delta} \in \mathbb{C}_{\alpha\eta} \cap \{\boldsymbol{\Delta}: \|\boldsymbol{\Delta}\|_2 \leq 1\}$ with uniform constants $\kappa_{\mathscr{L}}$ and $\tau_{\mathscr{L}}$ that do not depend on α. To prove the RSC at $\boldsymbol{\beta}_\alpha^*$ and a stronger version in Lemma 4, we first give a uniform lower bound of $\delta\mathscr{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta})$ for all $\|\boldsymbol{\beta}\|_2 \leq 4\rho_2$, $\|\boldsymbol{\Delta}\|_2 \leq 8\rho_2$ and $\alpha \leq c_u\rho_2^{-1}$, where c_u is a positive constant, depending on M_k, κ_l, κ_u and $\kappa 0$.

Lemma 2

Under conditions (C1)-(C3), for all $\|\boldsymbol{\beta}\|_2 \leq 4\rho_2$, $\|\boldsymbol{\Delta}\|_2 \leq 8\rho_2$ and $\alpha \leq c_u\rho_2^{-1}$, there exist uniform positive constants c_1' and c_2' such that, with probability at least $1 - c_1' \exp\left(-c_2' n\right)$,

$$\delta\mathscr{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}) \geq \kappa_1 \|\boldsymbol{\Delta}\|_2 \left(\|\boldsymbol{\Delta}\|_2 - \kappa_2 \sqrt{(\log p)/n} \|\boldsymbol{\Delta}\|_1\right). \quad (2.6)$$

Lemma 3

Suppose conditions (C1)-(C3) hold and assume that

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

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

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 conditions of Lemma 1 and 3, there exist positive constants c_1, c_2, *and C_2 such that with probability at least* $1 - c_1 \exp(-c_2 n)$,

$$\|\hat{\beta} - \beta^*_\alpha\|_2 \leq C_2 k_l^{-2} \kappa_\lambda^{2-q} R_q \left(n^{-1} \log p\right)^{1-(q/2)}.$$

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

Theorem 3

Under conditions of Lemmas 1 and 3, with probability at least $1 - c_1 \exp(-c_2 n)$,

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

where the constants $d_1 = C_1 \sqrt{\kappa_u} \kappa_l^{-1} \left(\kappa_0^k + \sqrt{M_k}\right)$ and $d_2 = C_2 \kappa_l^{-2} \kappa_\lambda^{2-q}$.

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

1. When the conditional distribution of ε is symmetric around 0, then $\beta^*_\alpha = \beta^*$ 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 $\|\hat{\beta} - \beta^*\|_2$ has a convergence rate of $\sqrt{R_q}[(\log p)/n]^{1/2-q/4}$, where $R_q = \sum_{j=1}^p |\beta_j^*|^q$. The rate is the same as the minimax rate (Raskutti, Wainwright, and Yu, 2011) for weakly sparse model under the light tails. In a special case that $q = 0$, $\|\hat{\beta} - \beta^*\|_2$ converges at a rate of $\sqrt{s(\log p)/n}$, where s is the number of nonzero elements in β^*. 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.
2. 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 $\alpha \leq \left\{d_1^{-1} d_2 R_q [(\log p)/n]^{1-\frac{q}{2}}\right\}^{\frac{1}{2(k-1)}}$. Recall from conditions in Lemmas 1 and 3 that we also need to choose α, such that $c_l \sqrt{(\log p)/n} \leq \alpha \leq c_u \rho_2^{-1}$ 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 the existing estimators.

In practice, the distribution of errors is unknown. Our method is more flexible than the 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 a and λ_n can be chosen by a two-dimensional grid search using cross-validation or information-based criterion, for example, Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). More specifically, the searching grid is formed by partitioning a rectangle in the scale of (log(a), log(λ_n)). The optimal values are then found by the combination that minimizing AIC, BIC or the cross-validated measurement (such as Mean Squared Error).

3 Geometric convergence of computational error

The gradient descent algorithm (Nesterov, 2007; Agarwal, Negahban, and Wainwright, 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 (LQA) and iteratively solve the following optimization problem:

$$\hat{\boldsymbol{\beta}}^{t+1}=\underset{\|\boldsymbol{\beta}\|_1\le\rho}{argmin}\left\{\mathscr{L}_n\left(\hat{\boldsymbol{\beta}}^t\right)+\left[\nabla\mathscr{L}_n\left(\hat{\boldsymbol{\beta}}^t\right)\right]^T\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{\mathbf{t}}\right)+\frac{\gamma_u}{2}\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{\mathbf{t}}\|_2^2+\lambda_n\|\boldsymbol{\beta}\|_1\right\}, \tag{3.1}$$

where γ_u is a sufficiently large fixed constant whose condition is specified in (3.3) and the side constraint "$\|\boldsymbol{\beta}\|_1 \le \rho$" 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 (3.1), the update can be computed by a two-step procedure. We first solve (3.1) without the norm constraint, which is the soft-threshold of the vector $\hat{\boldsymbol{\beta}}^t-\frac{1}{\gamma_u}\nabla\mathscr{L}_n\left(\hat{\boldsymbol{\beta}}^t\right)$ at level λ_n, and call the solution $\check{\boldsymbol{\beta}}$. If $\|\check{\boldsymbol{\beta}}\|_1 \le \rho$, 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 $\{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|_1 \le \rho\}$. The projection can be done (Duchi, *et al.*, 2008) by soft-thresholding $\check{\beta}$ at level π_n, where π_n is given by the following procedure: (1) sort $\left\{|\check{\beta}_j|\right\}_{j=1}^p$ into $b_1 \ge b_2 \ge \ldots \ge b_p$; (2) find $J=max\left\{1\le j\le p: b_j-\left(\sum_{r=1}^j b_r-\rho\right)/j>0\right\}$ and $\left(\sum_{r=1}^J b_j-\rho\right)/J$.

Agarwal, Negahban, and Wainwright (2012) considered the computational error of such firstorder gradient descent method. They showed that, for a convex and differentiable loss functions $\ell(x)$ 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 $\boldsymbol{\beta}^*=argmin_{\boldsymbol{\beta}}\,\mathrm{E}\,\ell(\mathbf{x},y;\boldsymbol{\beta})$, and $\hat{\boldsymbol{\beta}}=argmin_{\boldsymbol{\beta}}\frac{1}{n}\sum_{i=1}^n\ell(\mathbf{x}_i,y_i,\boldsymbol{\beta})+p(\boldsymbol{\beta})$. Different from their setup, our population minimizer $\boldsymbol{\beta}_\alpha^*$ varies by a. Nevertheless, as $\boldsymbol{\beta}_\alpha^*$ converges to the true effect $\boldsymbol{\beta}^*$, by a careful control of a, 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 (2.3) and $\hat{\boldsymbol{\beta}}^t$ is as defined in (3.1).

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

$$RSC{:}\delta\mathscr{L}_n(\boldsymbol{\Delta},\boldsymbol{\beta}) \geq \frac{\gamma_l}{2}\|\boldsymbol{\Delta}\|_2^2 - \tau_l\|\boldsymbol{\Delta}\|_1^2, \quad (3.2)$$

$$RSM{:}\delta\mathscr{L}_n(\boldsymbol{\Delta},\boldsymbol{\beta}) \leq \frac{\gamma_u}{2}\|\boldsymbol{\Delta}\|_2^2 + \tau_u\|\boldsymbol{\Delta}\|_1^2, \quad (3.3)$$

for all $\boldsymbol{\beta}$ and $\boldsymbol{\Delta}$ in some set of interest, with parameters $\boldsymbol{\gamma_l}$, $\boldsymbol{\tau_l}$, $\boldsymbol{\gamma_u}$ and $\boldsymbol{\tau_u}$ that do not depend on $\boldsymbol{\alpha}$. We show that such conditions hold with high probability.

Lemma 4

Under condition (C1)-(C3), for all $\|\boldsymbol{\beta}\|_2 \leq 4\rho_2$, $\|\boldsymbol{\Delta}\|_2 \leq 8\rho_2$ *and* $\alpha \leq c_u\rho_2^{-1}$, *with probability greater than* $1 - c_1\exp(-c_2 n)$, *(3.2) and (3.3) hold with* $\boldsymbol{\gamma_l} = \boldsymbol{\kappa}1$, $\tau_l = \kappa_1\kappa_2^2(\log p)/(2n)$, $\boldsymbol{\gamma_u} = 3\boldsymbol{\kappa_u}$, $\boldsymbol{\tau_u} = \boldsymbol{\kappa_u}(\log p)/n$.

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 $R_q\left(\frac{\log p}{n}\right)^{1-(q/2)} = o(1)$, which is required for consistency of *any method* over the weak sparse L_q ball by the known minimax results (Raskutti, Wainwright, and Yu, 2011). Denote $r_n^2 = R_q\left(\frac{\log p}{n}\right)^{1-(q/2)}$. Theorems 3 and 4 below imply that, with high probability,

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

when the sample size is large enough to ensure $r_n \leq d_2\left\{d_3\left(d_2^2+1\right)\right\}^{-1/2}$. 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 conditions of Theorem 3, suppose we choose λ_n *as in Lemma 1 and also satisfying*

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

where $|S_{\alpha\eta}|$ *denotes the cardinality of set* $S_{\alpha\eta}$ *and* $\bar{\gamma}_l = \gamma_l - 64\tau_l|S_{\alpha\eta}|$, *then with probability at least* $1 - c_1 \exp(-c_2 n)$, *there exists a generic positive constant* d_3 *such that*

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

for all iterations

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

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

4 Robust estimation of mean and covariance matrix

The estimation of mean can be regarded as a univariate linear regression where the covariate equals to 1. In that special case, we have 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 i.i.d sample from some unknown distribution with $E(y_i) = \mu$ and $\text{var}(y_i) = \sigma^2$. 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 $|x| \le 1$, $\psi(x) = 1$, 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 $\frac{\log(1/\delta)}{n} \le 1/8$ and let $\alpha = \sqrt{\frac{\log(1/\delta)}{nv^2}}$ where $v \ge \sigma$. *Then,*

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

The above 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, Han, and Gu, 2012) and covariance matrix estimation (Bickel and Levina, 2008; Fan, Liao and Mincheva, 2013), where uniform convergence is required. Taking the estimation of large covariance matrix as an example, in order for the elements of the sample covariance matrix

to converge uniformly, the aforementioned authors require the underlying multivariate distribution be sub-Gaussian. This restrictive assumptions can be removed if we apply the robust estimation with concentration bound. Regarding $\sigma_{ij} = \mathrm{E}\, X_iX_j$ as the expected value of the random variable X_iX_j (it is typically not the same as the median of X_iX_j), it can be estimated with accuracy

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

where $v \geq max_{i,j\leq p}\sqrt{var\,(X_iX_j)}$ and $\hat{\sigma}_{ij}$ is 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 $a > 2$ and the union bound, we have

$$P\left\{\underset{i,j\leq p}{max}|\hat{\sigma}_{ij} - \sigma_{ij}| \geq 4v\sqrt{\frac{a\,log\,p}{n}}\right\} \leq 2p^{2-a},$$

when $max_{i\leq p}\,\mathrm{E}\,X_i^4$ is bounded. This robustified covariance estimator requires much weaker condition than the sample covariance and has far wide 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 the estimation of the mean of heavy-tailed distribution with fast concentration. He proposed an *M*-estimator by solving

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

where the influence function $\psi_C(x)$ is chosen such that $-\log(1-x+x^2/2) \leq \psi_C(x) \leq \log(1+x+x^2/2)$. 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 we replace the RA-quadratic loss $\ell_\alpha\,(x)$ in (2.3) with Catoni loss

$$\ell_\alpha^c\,(x) = \frac{2}{\alpha}\int_0^x\psi_c\,(\alpha t)\,dt,$$

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

$$\psi_c\,(t) = sgn\,(t)\left\{-log\left(1 - |t| + t^2/2\right)I\,(|t|<1) + log\,(2)\,I\,(|t|\geq 1)\right\}.$$

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

Theorem 6

Suppose condition (C1) holds for $k = 2$ or 3, (C2), (C3) and (2.7) hold. Then there exist 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* $1 - c_1 \exp(-c_2 n)$,

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

Unlike the RA-lasso, the order of bias of $\hat{\beta}^c$ 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 $\hat{\beta}^c$ could also be established as in Theorem 4, since the RSC/RSM conditions also hold for Catoni loss with uniform constants.

6 Variance Estimation

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

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

where $\hat{\beta}^{(-j)}$ is the RA-Lasso estimator obtained by using data points outside the j-th fold. We show that $\hat{\sigma}^2$ is asymptotically efficient. Different from the existing cross-validation based method (Fan, Guo, and Hao, 2012), light-tail assumption is not needed due to the utilization of the RA-Lasso estimator.

Theorem 7

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

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

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 following high dimensional model

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

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

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

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:

1. Normal with mean 0 and variance 4 (N(0,4));
2. Two times the t-distribution with degrees of freedom 3 ($2t_3$);
3. Mixture of Normal distribution(MixN): $0.5N(-1, 4) + 0.5N(8, 1)$;
4. Log-normal distribution (LogNormal): $\varepsilon = e^{1+1.2Z}$, where Z is standard normal.
5. Weibull distribution with shape parameter = 0.3 and scale parameter = 0.5 (Weibull).

In order to meet the model assumption, 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 RA-Lasso with two other methods in high-dimensional setting: (a) Lasso: the penalized least-squares estimator with L_1-penalty as in Tibshirani (1996); and (b) R-Lasso: the R-Lasso estimator in Fan, Fan, and Barut (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:

(1) L_2 error, which is defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$.

(2) L_1 error, which is defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$.

(3) Number of false positives (FP), which is number of noise covariates that are selected.

(4) Number of false negatives (FN), which is number of signal covariates that are not selected.

We also measured the relative gain of RA-Lasso with respect to R-Lasso and Lasso, in terms of the difference to 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 gain of RA-Lasso with respect to Lasso ($\text{RG}_{\text{A,L}}$) in L_2 and L_1 norm are defined as

$$\frac{\|\hat{\beta}_{\text{Lasso}}-\beta^*\|_2-\|\hat{\beta}_{\text{oracle}}-\beta^*\|_2}{\|\hat{\beta}_{\text{RA- Lasso}}-\beta^*\|_2-\|\hat{\beta}_{\text{oracle}}-\beta^*\|_2} \quad \text{and} \quad \frac{\|\hat{\beta}_{\text{Lasso}}-\beta^*\|_1-\|\hat{\beta}_{\text{oracle}}-\beta^*\|_1}{\|\hat{\beta}_{\text{RA- Lasso}}-\beta^*\|_1-\|\hat{\beta}_{\text{oracle}}-\beta^*\|_1}.$$

The relative gain of RA-Lasso with respect to R-Lasso ($RG_{A,R}$) is defined similarly.

For RA-Lasso, the tuning parameters λ_n and a were chosen optimally based on 100 independent validation datasets. We ran a 2-dimensional grid search to find the best (λ_n, a) pair that minimizes the mean L_2-loss of the 100 validation datasets. Such an optimal pair was then used in the simulations. Similar method was applied in choosing the tuning parameters in Lasso and R-Lasso.

The above simulation model is based on the additive model (7.1), in which error distribution is independent of covariates. However, this homoscedastic model makes the conditional mean and the conditional median differ only by a constant. To further examine the deviations between the mean regression and median regression, we also simulated the data from the heteroscedastic model

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

where the constant $c=\sqrt{3}\|\beta^*\|^2$ makes $\mathrm{E}\left[c^{-1}\left(\mathbf{x}_i^T\beta^*\right)^2\right]^2=1$. Note that $\mathbf{x}_i^T\beta^*\sim N\left(0,\|\beta^*\|^2\right)$ and therefore c is chosen so that the average noise levels is the same as that of ε_i. For both the homoscedastic and the heteroscedastic models, we ran 100 simulations for each scenario. The mean of each performance measurement is reported in Table 2 and 3, respectively.

Tables 2 and 3 indicate that our method had the biggest advantage when the errors were asymmetric and heavy-tailed (LogNormal and Weibull). In this case, R-Lasso had larger L_1 and L_2 errors due to the bias for estimating the conditional median instead of the mean. Even though Lasso did not have bias in the loss component (quadratic loss), it did not perform well due 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 Lasso and R-Lasso. When the errors were symmetric and heavy-tailed ($2t_3$), our estimator performed closely as R-Lasso, both of which outperformed Lasso. The above two cases evidently showed that RA-Lasso was robust to the outliers and did not lose efficiency when the errors were indeed symmetric. Under the light-tailed scenario, if the errors were asymmetric (MixN), our method performed similarly as Lasso. R-Lasso 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, RA-Lasso is more flexible than Lasso and R-Lasso. The tuning parameter a automatically adapts to errors with different shapes and tails. It enables RA-Lasso to render consistently satisfactory results under all scenarios.

8 Real Data Example

In this section, we use a microarray data to illustrate the performance of Lasso, R-Lasso and RA-Lasso. Huang, *et al.* (2011) studied the role of innate immune system on the development of atherosclerosis by examining gene profiles from peripheral blood of 119 patients. The data were collected using Illumina HumanRef8 V2.0 Bead Chip and are available on Gene Expression Omnibus. The original study showed that the toll-like receptors (TLR) signaling pathway plays an important role on triggering the innate immune system in face of atherosclerosis. Under this pathway, the "TLR8" gene was found to be a key atherosclerosis-associated gene. To further study 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, CTR) that are related to the TLR pathway. We applied Lasso, R-Lasso and RA-Lasso to this data. The tuning parameters for all methods were chosen by using five-fold cross validation. Figure 1 shows our choice of the penalization parameter based on the cross validation results. For RA-Lasso, the choice of a 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 Figure 2. The selected genes by the three methods are reported in Table 4. After the selection, we regressed the expression of TLR8 gene on the selected genes, the t-values from the refittings are also reported in Table 4.

Table 4 shows that Lasso only selected one gene. R-Lasso selected 17 genes. Our proposed RA-Lasso selected 34 genes. Eight genes (CSF3, IL10, AKT1, TOLLIP, TLR1, SHC1, EPOR, and TJP1) found by R-Lasso were also selected by RA-Lasso. Compared with Lasso and R-Lasso, our method selected more genes, which could be useful for a second-stage confirmatory study. It is clearly seen from Figure 2 that the residuals from the fitted regressions had heavy right tail and skewed distribution. We learn from the simulation studies in Section 7 that RA-Lasso tends to perform better than Lasso and R-Lasso in this situation. For further investigation, we randomly chose 24 patients as the test set; applied three methods to the rest 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/Squared Error of predictions are shown in Figure 3. RA-Lasso has better predictions than Lasso and R-Lasso.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

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

Appendix: Proofs of Theorem 1, 2 and 5

Proof of Theorem 1

Let $\ell(x) = x^2$. Since $\boldsymbol{\beta}^*$ minimizes $\mathrm{E}\,\ell\left(y - \mathbf{x}^T\boldsymbol{\beta}\right)$, it follows from condition(C2) that

$$\mathrm{E}\left[\ell\left(y - \mathbf{x}^T\boldsymbol{\beta}_\alpha^*\right) - \ell\left(y - \mathbf{x}^T\boldsymbol{\beta}^*\right)\right] = (\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)^T \mathrm{E}\left(\mathbf{x}\mathbf{x}^T\right)(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*) \geq \kappa_l \|\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*\|_2^2. \quad \text{(A.1)}$$

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

$$\begin{aligned} & \mathrm{E}\left[\ell\left(y - \mathbf{x}^T\boldsymbol{\beta}_\alpha^*\right) - \ell\left(y - \mathbf{x}^T\boldsymbol{\beta}^*\right)\right] \\ = \; & \mathrm{E}\left[\ell\left(y - \mathbf{x}^T\boldsymbol{\beta}_\alpha^*\right) - \ell_\alpha\left(y - \mathbf{x}^T\boldsymbol{\beta}_\alpha^*\right)\right] + \mathrm{E}\left[\ell_\alpha\left(y - \mathbf{x}^T\boldsymbol{\beta}_\alpha^*\right) - \ell_\alpha\left(y - \mathbf{x}^T\boldsymbol{\beta}^*\right)\right] + \mathrm{E}\left[\ell_\alpha\left(y - \mathbf{x}^T\boldsymbol{\beta}^*\right) - \ell\left(y - \mathbf{x}^T\boldsymbol{\beta}^*\right)\right] \\ \leq \; & \mathrm{E}\left[g_\alpha\left(y - \mathbf{x}^T\boldsymbol{\beta}_\alpha^*\right)\right] - \mathrm{E}\left[g_\alpha\left(y - \mathbf{x}^T\boldsymbol{\beta}^*\right)\right]. \end{aligned}$$

By Taylor's expansion, we have

$$\mathrm{E}\left[\ell\left(y - \mathbf{x}^T\boldsymbol{\beta}_\alpha^*\right) - \ell_\alpha\left(y - \mathbf{x}^T\boldsymbol{\beta}_\alpha^*\right)\right] \leq 2\,\mathrm{E}\left[\left(z - \alpha^{-1}\right) I\left(z > \alpha^{-1}\right) |\mathbf{x}^T(\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*)|\right], \quad \text{(A.2)}$$

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

$$\begin{aligned} \mathrm{E}_\epsilon\left[\left(z - \alpha^{-1}\right) I\left(z > \alpha^{-1}\right)\right] = & \int_0^\infty P_\epsilon\left(zI\left(z > \alpha^{-1}\right) > t\right) dt - \alpha^{-1} P_\epsilon\left(z > \alpha^{-1}\right) \\ = & \int_0^\infty P_\epsilon\left(z > t \quad \text{and} \quad z > \alpha^{-1}\right) dt - \alpha^{-1} P_\epsilon\left(z > \alpha^{-1}\right) \\ = & \int_{\alpha^{-1}}^\infty P_\epsilon(z > t)\, dt + \int_0^{\alpha^{-1}} P_\epsilon\left(z > \alpha^{-1}\right) dt - \alpha^{-1} P_\epsilon\left(z > \alpha^{-1}\right) \\ \leq & \int_{\alpha^{-1}}^\infty \frac{\mathrm{E}_\epsilon\left(z^k\right)}{t^k} dt \leq \alpha^{k-1}\,\mathrm{E}_\epsilon\left(z^k\right). \end{aligned}$$

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

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

Note that,

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

where the last inequality follows from (C1) and (C2). On the other hand, by (C3), $\mathbf{x}^T\left(\boldsymbol{\beta}^*-\tilde{\boldsymbol{\beta}}\right)$ 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,

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

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

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

$$\begin{aligned}\|\boldsymbol{\Delta}\|_2^2 &\le 9\tfrac{\lambda_n^2}{\kappa_{\mathscr{L}}^2}|S_{\alpha\eta}|+\tfrac{\lambda_n}{\kappa_{\mathscr{L}}^2}\left\{2\tau_{\mathscr{L}}^2+4\|\boldsymbol{\beta}^*_{S^c_{\alpha\eta}}\|_1\right\} \\ &\le \tfrac{36\lambda_n^2R_q}{\kappa_1^2\eta^q}+\tfrac{4\lambda_n}{\kappa_1^2}\left\{8R_q\kappa_1\kappa_2\kappa_\lambda{}^{1-q}\left(\tfrac{log\,p}{n}\right)^{1-(q/2)}+4R_q\eta^{1-q}\right\} \\ &\overset{(ii)}{=} \tfrac{36}{\kappa_1^2}R_q\lambda_n^{2-q}+\tfrac{16}{\kappa_1^2}R_q\lambda_n^{2-q}\left\{2\kappa_1\kappa_2\left(\tfrac{log\,p}{n}\right)^{\frac{1}{2}}+1\right\} \\ &\overset{(ii)}{\le} C_2k_l^{-2}\kappa_\lambda^{2-q}R_q(n^{-1}\,log\,p)^{1-(q/2)},\end{aligned}$$

where (i) follows from the choice of $\eta = \lambda_n$, in (ii) we assume that the sample size n is large enough such that $2\kappa_1\kappa_2(n^{-1}\log p)^{1/2} \le 1$ and observe that $\kappa_1 = \kappa_l/4$. On the other hand, by Lemma 1 and 3, $P(A_1\cap A_2) \ge 1-c_1\,exp(-c_2n)$, where $c_1{=}max\left\{2, c_1^{'}\right\}$ and $c_2{=}min\left\{c_0, c_2^{'}\right\}$.

Proof of Theorem 5

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

$$-log\left(1-x+x^2\right) \le \psi(x) \le log\left(1+x+x^2\right).$$

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

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

Similarly, $\mathrm{E}\left\{exp\left[-\alpha nr\left(\theta\right)\right]\right\} \leq exp\left\{-n\alpha\left(\mu-\theta\right)+n\alpha^2\left[v^2+\left(\mu-\theta\right)^2\right]\right\}$. Define

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

By Chebyshev inequality,

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

Similarly, $P(r(\theta) < B_-(\theta)) \leq \delta$.

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 $\frac{log(1/\delta)}{n} \leq 1/8$ and the choice of $\alpha = \sqrt{\frac{log(1/\delta)}{nv^2}}$, we have $\alpha^2v^2+\frac{log(1/\delta)}{n} \leq 1/4$. Therefore,

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

Similarly,

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

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

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

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

References

Agarwal A, Negahban S, Wainwright MJ. Fast global convergence of gradient methods for high-dimensional statistical recovery. The Annals of Statistics. 2012; 40:2452–2482.

Alexander KS. Rates of growth and sample moduli for weighted empirical processes indexed by sets. Probability Theory and Related Fields. 1987; 75:379–423.

Belloni A, Chernozhukov V. L_1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics. 2011; 39:82–130.

Bickel PJ, Levina E. Covariance regularization by thresholding. The Annals of Statistics. 2008; 36:2577–2604.

Bickel PJ, Ritov Y, Tsybakov A. Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics. 2009; 37:1705–1732.

Bühlmann, P., Van De Geer, S. Statistics for high-dimensional data: methods, theory and applications. Springer; 2011.

Catoni O. Challenging the empirical mean and empirical variance: a deviation study. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques. 2012; 48:1148–1185.

Duchi J, Shalev-Shwartz S, Singer Y, Chandra T. Efficient projections onto the L_1-ball for learning in high dimensions. Proceedings of the 25th international conference on Machine learning. 2008:272–279.

Efron B. Correlated z-values and the accuracy of large-scale statistical estimates. Journal of the American Statistical Association. 2010; 105:1042–1055. [PubMed: 21052523]

Engle RF. Autoregressive conditional heteroscedasticy with estimates of the variance of U.K. inflation. Econometrica. 1982; 50:987–1008.

Fan J, Fan Y, Barut E. Adaptive robust variable selection. The Annals of Statistics. 2014; 42:324–351. [PubMed: 25580039]

Fan J, Guo S, Hao N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. The Annals of Statistics. 2014; 42:324–351. [PubMed: 25580039]

Fan J, Han X, Gu W. Estimating false discovery proportion under arbitrary covariance dependence (with discussion). Jour. Roy. Statist. Soc. B. 2012; 74:37–65.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

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

Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. Jour. Roy. Statist. Soc. B. 2008; 70:849–911.

Fan J, Lv J. Non-concave penalized likelihood with NP-Dimensionality. IEEE – Information Theory. 2011; 57:5467–5484. [PubMed: 22287795]

Huang CC, Liu K, Pope RM, Du P, Lin S, Rajamannan NM, et al. Activated TLR signaling in atherosclerosis among women with lower Framingham risk score: the multi-ethnic study of atherosclerosis. PLoS ONE. 2011; 6:e21067. [PubMed: 21698167]

Huber PJ. Robust estimation of a location parameter. The Annals of Mathematical Statistics. 1964; 35:73–101.

Lambert-Lacroix S, Zwald L. Robust regression through the Hubers criterion and adaptive lasso penalty. Electronic Journal of Statistics. 2011; 5:1015–1053.

Ledoux, M., Talagrand, M. Probability in Banach Spaces: isoperimetry and processes. Springer; 1991.

Li Y, Zhu J. L_1-norm quantile regression. Journal of Computational and Graphical Statistics. 2008; 17:163–185.

Loh P-L, Wainwright MJ. Regularized *M*-estimators with nonconvexity: Statistical and algorithmic theory for local optima. Advances in Neural Information Processing Systems. 2013:476–484.

Massart, P., Picard, J. Concentration inequalities and model selection. Springer; 2007.

Negahban SN, Ravikumar P, Wainwright MJ, Yu B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statistical Science. 2012; 27:538–557.

Nesterov, Y. Technical Report 76, Center for Operations Research and Econometrics (CORE). Catholic Univ. Louvain (UCL); 2007. Gradient methods for minimizing composite objective function..

Raskutti G, Wainwright MJ, Yu B. Minimax rates of estimation for high-dimensional linear regression over L_q-balls. Information Theory, IEEE Transactions on Information Theory. 2011; 57:6976–6994.

Rivasplata O. Subgaussian random variables: an expository note. 2012

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, B. 1996; 58:267–288.

Van de Geer, S. Empirical Processes in M-estimation. Cambridge university press Cam bridge; 2000.

Wang H. Forward regression for ultra-high dimensional variable screening. Journal of American Statistical Association. 2009; 104:1512–1524.

Wang L. The L_1 penalized LAD estimator for high dimensional linear regression. Journal of Multivariate Analysis. 2013; 120:135–151.

Wu Y, Liu Y. Variable selection in quantile regression. Statistica Sinica. 2009; 19:801.

Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics. 2010; 38:894–942.

Zou H, Yuan M. Composite quantile regression and the oracle model selection theory. The Annals of Statistics. 2008; 36:1108–1126.

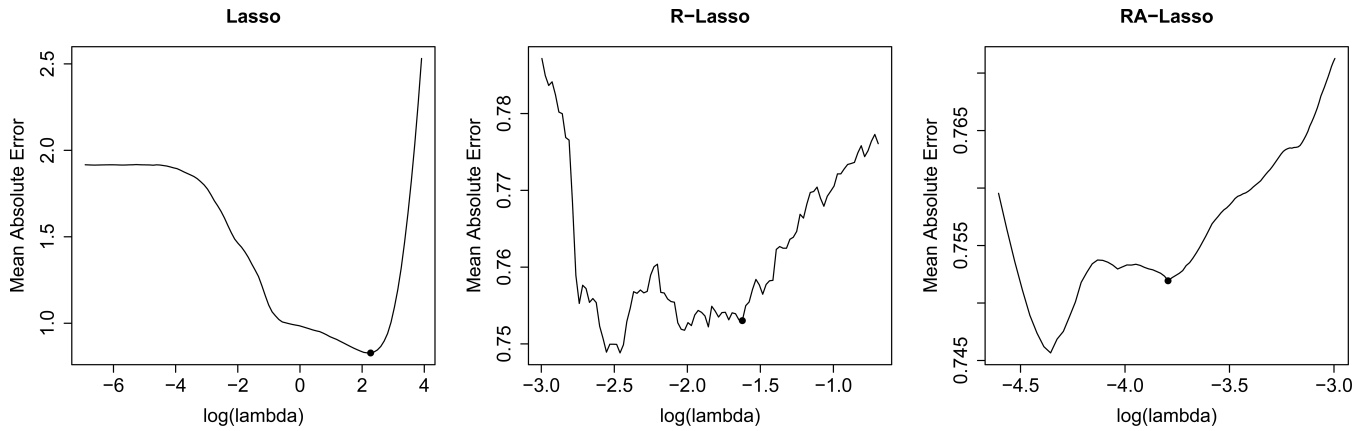

Figure 1.
Five-fold cross validation results: black dot marks the choice of the penalization parameter.

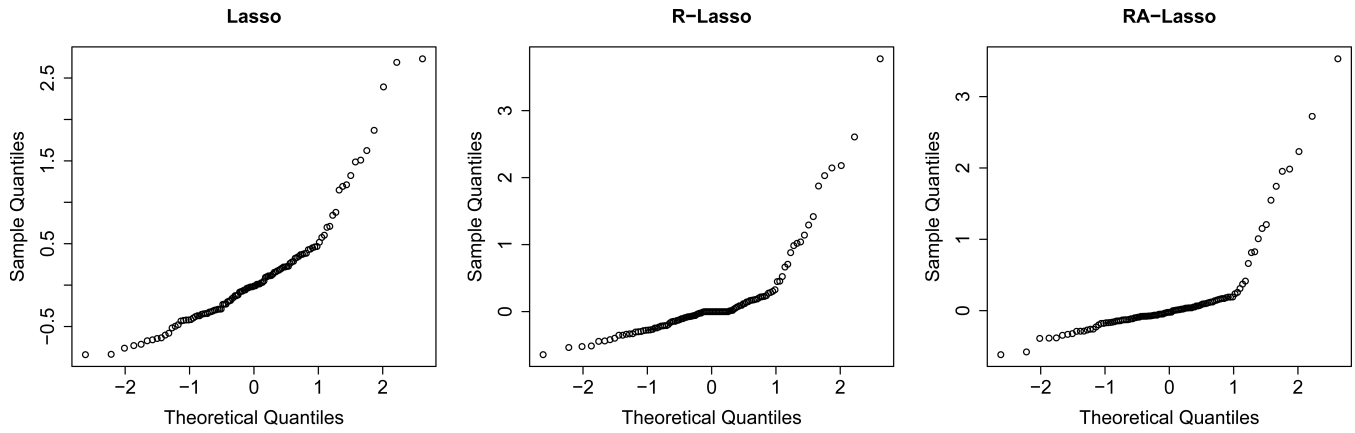

Figure 2.
QQ plots of the residuals from three methods.

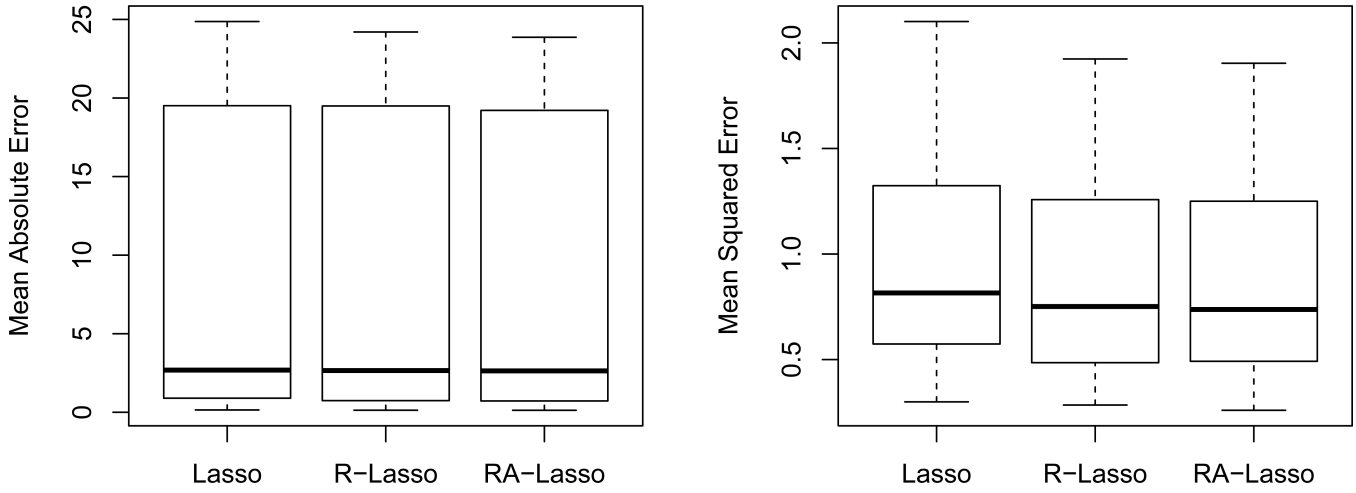

Figure 3.
Boxplot of the Mean Absolute/Squared Error of predictions.

Table 1

Summary of the shapes and tails of five error distributions

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

Table 2

Simulation results of Lasso, R-Lasso and RA-Lasso under homoscedastic model. (7.1)

		Lasso	R-Lasso	RA-Lasso	$RG_{A,L}$	$RG_{A,R}$
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 Lasso, R-Lasso and RA-Lasso under heteroscedastic model (7.2).

		Lasso	R-Lasso	RA-Lasso	$RG_{A,L}$	$RG_{A,R}$
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		

Table 4

Selected genes by Lasso, R-Lasso and RA-Lasso.

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