Communications for Statistical Applications and Methods
2023, Vol. 30, No. 2, 135–147

https://doi.org/10.29220/CSAM.2023.30.2.135
Print ISSN 2287-7843 / Online ISSN 2383-4757

# Least clipped absolute deviation for robust regression using skipped median

Hao Li[a], Seokho Lee[1,a]

[a]Department of Statistics, Hankuk University of Foreign Studies, Korea

**Abstract**

Skipped median is more robust than median when outliers are not symmetrically distributed. In this work, we propose a novel algorithm to estimate the skipped median. The idea of skipped median and the new algorithm are extended to regression problem, which is called least clipped absolute deviation (LCAD). Since our proposed algorithm for nonconvex LCAD optimization makes use of convex least absolute deviation (LAD) procedure as a subroutine, regularizations developed for LAD can be directly applied, without modification, to LCAD as well. Numerical studies demonstrate that skipped median and LCAD are useful and outperform their counterparts, median and LAD, when outliers intervene asymmetrically. Some extensions of the idea for skipped median and LCAD are discussed.

Keywords: convex relaxation, least clipped absolute deviation, robust statistics, skipped median

## 1. Motivation

Median is one of most popular robust measures for location. If distribution is skewed or heavy-tailed, median is better than mean as location representative. In case of symmetric distribution, whose mean (if exists) and median are identical, sample median is robust against outliers with high breakdown point while it is less efficient than sample mean. These estimation properties of median are carried over to regression problem as well. Least absolute deviation (LAD) finds the conditional median response, which is known as a robust alternative to the conditional mean response from least squares (LS) in the presence of outliers. Since it is still less efficient, LAD estimate is often used as an initialization for more advanced robust regression methods, such as MM-estimation (Yohai, 1987). However, LAD itself is still practically useful in that it provides the median of response variable conditional to a given covariate.

Even having high breakdown point of 0.5, sample median relying on sample ranks is still vulnerable to the adverse effect from outliers. Figure 1 shows a simple example of location mixture model, where outliers are distributed to the right side far from the population. Sample mean is not a good estimate of the population median as we expected. Sample median is not good enough either because the sample is contaminated asymmetrically by outliers and the mid-ranked observation is not guaranteed to be close to the population median. A better choice for this example is to use skipped median (Hampel *et al.*, 1986). The skipped median is the median of sub-sample belonging to an interval $[\theta - a, \theta + a]$ with a positive constant $a > 0$. A suitably constructed interval can completely

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C1003956).

[1] Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Yongin, Gyeonggi-do 17035, Korea. E-mail: lees@hufs.ac.kr

Published 31 March 2023 / journal homepage: http://csam.or.kr

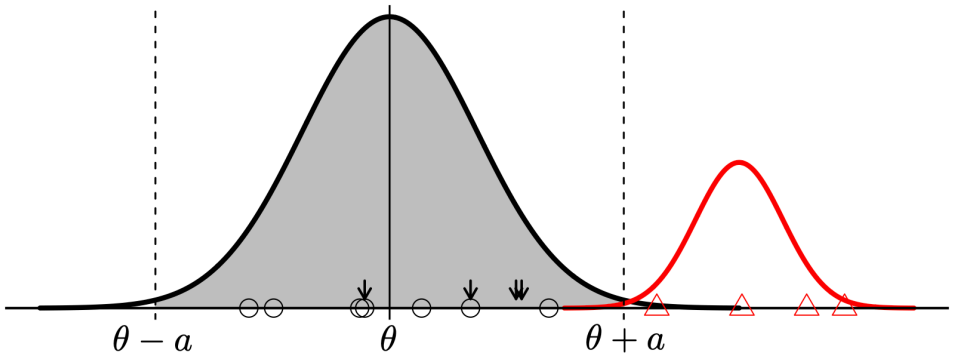


Figure 1: *Location mixture example is presented. Circles are samples from population distribution whose median is $\theta$. Some outliers from their distribution having different location are denoted by triangles. The arrow closest to $\theta$ is the skipped median estimate using the samples belonging to the interval $[\theta - a, \theta + a]$. Other arrows show sample median, sample trimmed mean, and sample mean, indiscriminately.*

remove outliers from the sample so that the sample median of selected observations is not affected at all by outliers. Figure 1 illustrates the usefulness of the skipped median which is denoted by the arrow closest to the population median. Note that skipping procedure is not same as trimming procedure, such as trimmed mean. Trimmed estimator is constructed after removing the equal proportions of the data placed on both tails. Thus, trimming strategy is not successful for estimating the median in the case that outliers are distribution asymmetrically because the legitimate observations in the left tail are also discarded as in Figure 1.

To make this comparison in a concrete way, we demonstrate the performance of estimates under normal mixture error model. Random samples are drawn from $(1 - \pi)N(0, 1) + \pi N(4, 1/2^2)$ with outlier rate of $\pi = 0.1$. This mixture model assumes that legitimate samples are from $N(0, 1)$ so that the purpose of location estimation is to correctly finds the true center of symmetry $\theta = 0$. We generate the noisy data of size 1000 from the mixture model (900 legitimates, 100 outliers) and, then, estimated mean, median, 10% trimmed mean, and skipped median (which will be described in Section 2 with an estimating algorithm). This procedure is repeated 100 times and we provide the boxplot of 100 estimates in Figure 2. As we discussed, skipped median outperforms others, all of which are biased upward due to outlier distribution.

For skipped median estimation, choosing an appropriate interval $[\theta - a, \theta + a]$ for finding $\theta$ seems a difficult task because location ($\theta$) and width ($a$) of the interval should be determined. $a$ controls the amount of samples to be used for median computation. With a too large $a$, the interval becomes wider so that outliers may be included in the interval. A small $a$ gives a narrow interval, causing efficiency loss in estimation. Later, we will discuss how to choose $a$ considering the efficiency of the skipped median estimator under normality. Given $a$, it is known that an estimator of $\theta$, which becomes the sample skipped median, is obtained under risk minimization. This is described in Section 2, where an accompanying efficient algorithm is provided.

In regression problem, least absolute deviation (LAD) estimates the conditional median response, which is resistant against outliers. However, as in the location estimation example, LAD is not robust enough when outlier distribution is asymmetric. We will show that conditional skipped median response is more robust than conditional median response in Section 3, where the skipped median approach is generalized and applied to regression problem. The performance of the proposed method

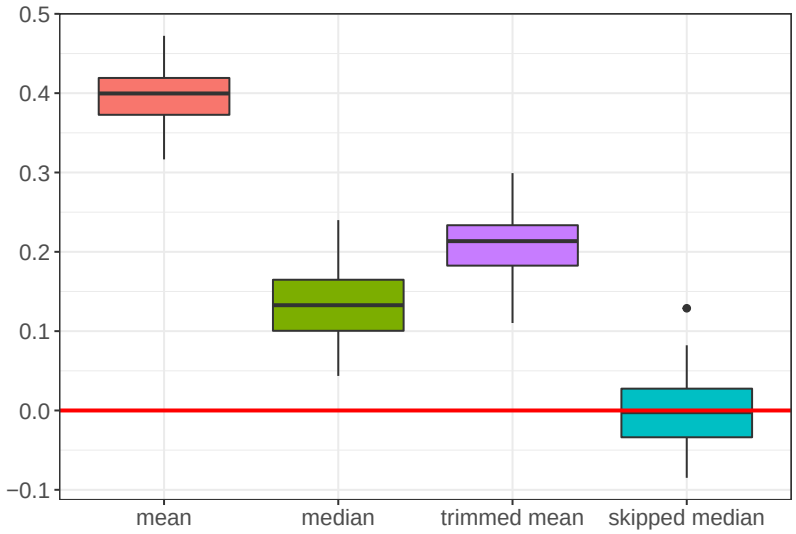


Figure 2: *Boxplot of location estimates under the normal mixture* $(1-\pi)N(0,1)+\pi N(4,1/4)$ *with* $\pi=0.1$. *The horizontal line denotes the center of symmetry for the legitimate distribution* $N(0,1)$.

is demonstrated under simulation in Section 4. This paper ends in Section 5 with concluding remarks.

## 2. Skipped median

In this section, we introduce skipped median briefly and propose a new algorithm for its estimation. Suppose $X\sim F$ with a known scale parameter $\sigma$. Consider the below risk minimization problem and its solution

$$\theta_a=\arg\min_{\theta} E_F\rho_a\left(\frac{X-\theta}{\sigma}\right) \tag{2.1}$$

with a loss function

$$\rho_a(u)=\min\{|u|,a\}=\begin{cases}|u|, & |u|\le a,\\ a, & |u|\ge a.\end{cases} \tag{2.2}$$

We call $\rho_a(u)$ the clipped absolute deviation loss. The positive constant $a$ regulates robustness and efficiency of estimation. With the derivative of the loss function

$$\psi_a(u)=\rho_a'(u)=\begin{cases}I(0<u<a)-I(-a<u<0), & |u|\le a,\\ 0, & |u|\ge a,\end{cases} \tag{2.3}$$

a minimizer of (2.1) is obtained by solving the below estimating equation

$$E_F\psi_a\left(\frac{X-\theta}{\sigma}\right)=0. \tag{2.4}$$

Since $E_F\psi_a((X-\theta)/\sigma)=P(\theta<X<\theta+a\sigma)-P(\theta-a\sigma<X<\theta)$, a solution of (2.4) satisfies

$$F(\theta+a\sigma)-F(\theta)=F(\theta)-F(\theta-a\sigma). \tag{2.5}$$

The solution $\theta_a$ is called the skipped median, which is the median of the distribution confined on the interval $[\theta - a\sigma, \theta + a\sigma]$. Note that, if $a = \infty$, then the clipped absolute deviate loss becomes the absolute value loss, $\rho_\infty(u) = |u|$, and its solution $\theta_\infty$ becomes the median of $F$. In the case that the distribution $F$ is symmetric about $\theta$ satisfying $F(\theta - x) = 1 - F(\theta + x)$ for all $x$, the Equation (2.5) becomes $2F(\theta) = 1$ for any $a \in (0, +\infty)$. The center for symmetry of $F$, that is the median, satisfies this equation. This implies that minimization of (2.1) finds the median of $X$ even if there exist outliers in tail.

Consider that $X_1, \ldots, X_n$ are iid from $F$. The sample version of the optimization solves the below problem

$$\hat{\theta}_a = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \rho_a \left( \frac{X_i - \theta}{\sigma} \right), \tag{2.6}$$

equivalently,

$$0 = \frac{1}{n} \sum_{i=1}^{n} \psi_a \left( \frac{X_i - \hat{\theta}_a}{\sigma} \right) = \# \left\{ X_i : \hat{\theta}_a \le X_i \le \hat{\theta}_a + a\sigma \right\} - \# \left\{ X_i : \hat{\theta}_a - a\sigma \le X_i \le \hat{\theta}_a \right\}. \tag{2.7}$$

Thus, $\hat{\theta}_a$ is a sample median after skipping the samples outside the interval $[\hat{\theta}_a - a\sigma, \hat{\theta}_a + a\sigma]$. The solution in (2.6) is an M-estimator with the bounded loss $\rho_a$. From the theory of M-estimation, we can easily deduce the influence function (IF) becomes

$$\mathrm{IF}(x) = \frac{\sigma \psi_a \left( (x - \theta)/\sigma \right)}{2f(\theta) - f(\theta - a\sigma) - f(\theta + a\sigma)} \tag{2.8}$$

with $f = F'$, the density of $X$. IF is bounded since $\psi_a$ is so. Let us define $\gamma = \min_\theta E_F \rho_a((X - \theta)/\sigma)$. Its breakdown point (BP) is $\epsilon^* = (1 - \gamma/a)/(2 - \gamma/a)$, which is strictly smaller than 0.5. However, the BP is still very close to 0.5 for normal distribution with similar bounded loss (Huber, 1984; Maronna *et al.*, 2019).

## 2.1. The choice of $a$ for skipped median

The additional parameter $a > 0$ takes the balance between robustness and efficiency. If $a$ gets closer to zero, robustness of estimator gets maximized but its efficiency gets lower since the available samples is reduced. If $a$ is too large, an estimator loses robustness since outliers are involved in estimation. A way of choosing $a$ is to ensure the level of efficiency under a certain distribution.

Note that the choice of $a$ must depend on the scale of distribution because $a$ determines the range of available of data. We set $\sigma$ by a previously estimated scale parameter, which should be also a robust scale estimate, such as MAD (Maronna *et al.*, 2019). Here, $\sigma = 1$ is assumed without loss of generality. From M-estimation theory, we can easily show that $\hat{\theta}_a \to \theta_a$ in probability as $n \to \infty$. And the sampling distribution of $\hat{\theta}_a$ is approximately normal with mean $\theta_a$ and variance $v_a/n$ with

$$v_a = \frac{E_F \psi_a (X - \theta_a)^2}{\left( E_F \psi_a' (X - \theta_a) \right)^2} = \frac{F(\theta_a + a) - F(\theta_a - a)}{\left( f(\theta_a + a) + f(\theta_a - a) - 2f(\theta_a) \right)^2}, \tag{2.9}$$

where $f = F'$ is the density function of the random variable $X$. If the distribution is symmetric about $\theta_a$, i.e., $F(\theta_a - a) = 1 - F(\theta_a - a)$ and $f(\theta_a - a) = f(\theta_a + a)$, then we have

$$v_a = \frac{F(\theta_a + a) - 1/2}{2\left\{ f(\theta_a + a) - f(\theta_a) \right\}^2}. \tag{2.10}$$

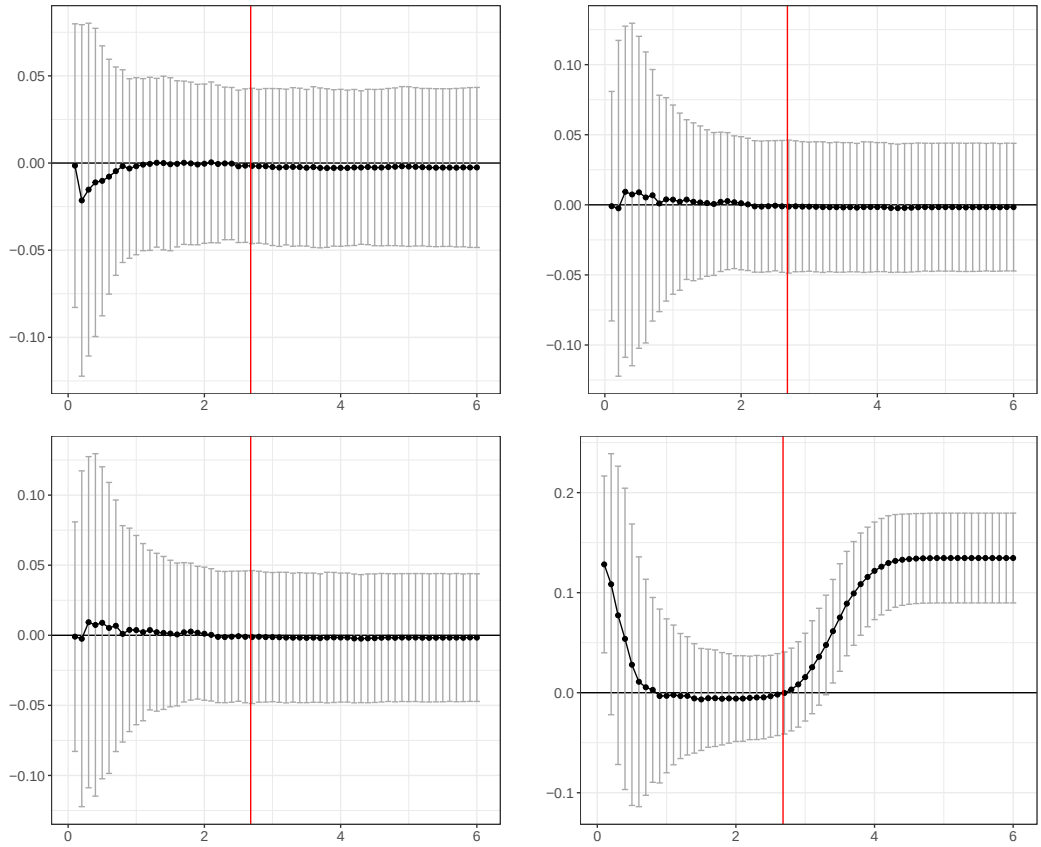


Figure 3: *Averaged location estimates ± standard deviation are displayed along $a$ values. Population median is 0 denoted by the horizontal line. The red vertical line is drawn at $a = 2.68$. The underlying distributions are standard normal (topleft), $t$-distribution with degrees of freedom 3 (topright) and 1 (bottomleft), and contaminated normal (bottomright) used in Figure 2.*

For sample median, $\hat{\theta}_\infty$ approximately follows the distribution $N(\theta_\infty, v_\infty/n)$ with $v_\infty = 1/4f(\theta_\infty)^2$. We can verify that $v_a \geq v_\infty$ holds for any $a > 0$, implying that the sample skipped median is less efficient than the sample median. For the choice of $a$, we consider the relative efficiency of $\hat{\theta}_a$ with respect to the sample median under the standard normal distribution. If $F = N(0, 1)$, we have $v_\infty = 2\pi/4 \approx 1.571$. We propose to use $a = 2.68$, which provides $v_{2.68}/v_\infty \approx 1.05$. The variance of the skipped median is 5% larger than that of the sample median with our proposal. As a different way, it is also possible to select $a$ based on quantiles of distribution in order to remove excessive tail observation for robustness. For example, $a = z_{\alpha/2}$ ignores $100 \times \alpha\%$ observations in tail for estimation. Note that it is different from trimmed estimation since trimming is not symmetric to the distribution when outliers exist. With the purpose of demonstration, we use $a = 2.68$ for subsequent analysis.

Figure 3 shows sensitivity analysis to see the performance change due to the choice of $a$. For the case of no contamination of outlier, bias of estimation quickly diminishes as $a$ gets larger than 1. We can observe that the bias gets worse as $a$ becomes larger when there are outliers in the tail, as shown in the bottom right panel of Figure 3. It is important in estimation performance to select $a$ wisely.

## 2.2. Estimating algorithm for skipped median

Since the loss $\rho_a(u)$ is not convex, direct minimization in (2.6) is not straightforward. We propose a novel iterative algorithm to compute the skipped median. The proposed algorithm is developed for location estimation, but it can be tailored to regression problem with penalization, which will be given in Section 3. Define a function $g(u, r)$ having an additional parameter $r$

$$g(u, r) = |u - r| + \rho_a(r) = \begin{cases} |u - r| + |r|, & |r| \le a, \\ |u - r| + a, & |r| \ge a. \end{cases} \tag{2.11}$$

The below theorem shows some properties of $g(u, r)$ related to $\rho_a(u)$, providing a useful algorithm to minimize (2.6) using convex relaxation.

**Theorem 1.** *For any given u,*

*(a) A hard-thresholding of u is a minimizer of $g(u, r) = |u - r| + \rho_a(r)$, i.e.,*

$$\hat{r} = \arg\min_r g(u, r) = uI(|u| > a). \tag{2.12}$$

*(b) $g(u, r) \ge \rho_a(u)$ for all r.*

*(c) $g(u, \hat{r}) = \rho_a(u)$.*

**Proof**: Since $g(u, r)$ in (2.11) is continuous and piecewise linear in $r$, a minimizer can be easily found by simply observing its functional shape over $r$ for a given $u$. In the case of $-a \le u \le 0$, we see that (i) $g(u, r) = u - r + a$ is decreasing over $r \le -a$, (ii) $g(u, r) = u - 2r$ is decreasing over $-a \le r \le u$, (iii) $g(u, r) = u$ is constant over $u \le r \le 0$, (iv) $g(u, r) = -u + 2r$ is increasing over $0 \le r \le a$, and (v) $g(u, r) = -u + r + a$ is increasing over $a \le r$. From (i)∼(v), any $r^* \in [u, 0]$ is a minimizer with the minimum $g(u, r^*) = u = \rho_a(u)$ if $-a \le u \le 0$. In the similar way, we can find that any $r^* \in [0, u]$ is a minimizer with the minimum $g(u, r^*) = u = \rho_a(u)$ if $0 \le u \le a$, and $r^* = u$ is the minimizer with the minimum $g(u, r^*) = a = \rho_a(u)$ if $u \le -a$ or $u \ge a$. This completes the proof of parts (a) and (c) since $\hat{r}$ defined in (3.4) is in the solution set $\mathcal{S} = \{r^* : r^* = \arg\min_r g(u, r)\}$ and, thus, $g(u, \hat{r}) = g(u, r^*) = \rho_a(u)$. Note that $g(u, r) \ge g(u, \hat{r}) = \rho_a(u)$ holds because $\hat{r}$ is a minimizer of $g(u, r)$, which completes the proof of part (b). □

***Remark 1.*** From the proof of **Theorem 1**, we find that $\tilde{r} = u$ is also a minimizer of $g(u, r)$ for any $u$. However, this trivial solution is not constructive at all for optimization. $\hat{r}$ in **Theorem 1** is chosen by having the smallest absolute value among the solution set $\mathcal{S}$.

***Remark 2.*** One may consider $g(u, r) = |u - r| + h(r)$ with $h(r) = aI(r \ne 0)$ to obtain the hard thresholding solution $\hat{r}$ as in **Theorem 1**. In this case, $g(u, r)$ is not continuous in $r$. However, $\hat{r}$ is uniquely determined, in contrast to **Theorem 1**. Having the same $\hat{r}$, this will eventually lead to the same optimization results.

Using the results of **Theorem 1**, minimization of (2.6) can be performed by iteratively minimizing the surrogate function in below:

$$G(\theta, r) = \frac{1}{n}\sum_{i=1}^{n} g\left(\frac{x_i - \theta}{\sigma}, r_i\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left|\frac{x_i - \theta}{\sigma} - r_i\right| + \rho_a(r_i)\right\}. \tag{2.13}$$

Table 1: Frequency table for iteration number until convergence for the example used in Figure 2

| Number of iterations | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Frequency | 6 | 85 | 8 | 1 |

This surrogate function is minimized alternatively over $\theta$ and $r = (r_1, \ldots, r_n)^T$. If $\hat{\theta}$ is given, minimization of $G(\hat{\theta}, r)$ over $r_1, \ldots, r_n$ is separable so that each $r_i$ is individually obtained by minimizing $g((x_i - \hat{\theta})/\sigma, r_i)$ for $i = 1, \ldots, n$. From **Theorem 1**, we get $\hat{r}_i = z_i I(|z_i| \geq a)$ with $z_i = (x_i - \hat{\theta})/\sigma$. When $\hat{r}_i$ are given, $\theta$ is obtained by minimizing $\sum_{i=1}^n |t_i - \theta|$ with $t_i = x_i - \sigma\hat{r}_i$, resulting in $\hat{\theta} = \text{Med}(t)$, which is the median of $\{t_1, \ldots, t_n\}$. Note that, if $\hat{r}_i \neq 0$ then $\hat{r}_i = (x_i - \hat{\theta}^{\text{old}})/\sigma$ resulting in $t_i = \hat{\theta}^{\text{old}}$. Thus, $t_i$'s having $\hat{r}_i \neq 0$ accumulate into the previous estimate of $\theta$, so that the updated value of $\theta$ severely depends on the previous estimate. With this consideration, we update $\theta$ by $\hat{\theta} = \text{Med}\{t_i : \hat{r}_i = 0\}$. Because $t_i = x_i$ for $\hat{r}_i = 0$, the updated estimate simply becomes $\hat{\theta} = \text{Med}\{x_i : \hat{r}_i = 0\}$, which is a median of $x_i$ inside the interval $[\hat{\theta}^{\text{old}} - a, \hat{\theta}^{\text{old}} + a]$.

This iterative algorithm under alternating scheme turns out to quickly converges to a limit point within a few iterations from our experience. As an illustration, we provide the number of iterations required to convergence in Table 1 for the example used in Figure 2. This estimating algorithm for location of a univariate variable gives us a hint to generalize the skipped median estimation to regression problem to obtain the conditional median response estimate robust to outliers.

## 3. Least clipped absolute deviation

We apply the skipped median approach to regression problem to obtain the conditional skipped median response. Consider the regression model

$$y_i = x_i^T \beta + u_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where $u_i$'s are independent random errors from an error distribution $F$, which is typically assumed to be symmetric about 0. Then, the conditional mean and the conditional median coincide as $E(y|x) = \text{Med}(y|x) = x^T\beta$. If $F = N(0, \sigma^2)$, the least squares (LS) estimation efficiently estimates $\beta$ and provides the conditional mean response. When the presence of outlier is suspicious, LAD serves a better alternative to LS. LAD provides the conditional median response estimate which is a robust estimate of $x^T\beta$. Suppose $u \sim (1 - \pi)F + \pi G$, where $G$ is a distribution of outlying observations. If $G$ is not symmetrically distributed about the center for symmetry of $F$, then LAD still suffers from the presence of outlier, as we have seen in Section 1 for univariate location problem.

We propose to minimize the below:

$$L(\beta) = \frac{1}{n} \sum_{i=1}^n \rho_a \left( \frac{y_i - x_i^T \beta}{\sigma} \right). \tag{3.2}$$

We call this optimization the least clipped absolute deviation (LCAD). If $a = \infty$, then $\rho_\infty(u) = |u|$ so that LCAD becomes LAD. Since $\rho_\infty(u)$ is convex, LAD optimization is well studied and implemented under linear programming (Koenker, 2005). With $0 < a < \infty$, $\rho_a(u)$ is not convex and its optimization is not straightforward. We apply the algorithm for skipped median described in Section 2.2 to LCAD optimization. Instead of minimizing $L(\beta)$ directly, we iteratively minimize the below criterion

$$G(\beta, r) = \frac{1}{n} \sum_{i=1}^n g\left( \frac{y_i - x_i^T \beta}{\sigma}, r_i \right) = \frac{1}{n} \sum_{i=1}^n \left\{ \left| \frac{y_i - x_i^T \beta}{\sigma} - r_i \right| + \rho_a(r_i) \right\}. \tag{3.3}$$

Minimizing $G(\beta, r)$ over $\beta$ and $r$ is performed in alternating fashion. If $\hat{\beta}$ is given, minimization of $G(\hat{\beta}, r)$ over $r = (r_1, \ldots, r_n)^T$ is separable so that each $r_i$ is obtained, from the results of **Theorem 1**, by

$$\hat{r}_i = \arg\min_r \left| \frac{y_i - x_i^T \hat{\beta}}{\sigma} - r \right| + \rho_a(r) = z_i I(|z_i| \geq a) \tag{3.4}$$

with $z_i = (y_i - x_i^T \hat{\beta})/\sigma$ for $i = 1, \ldots, n$. If $\hat{r} = (\hat{r}_1, \ldots, \hat{r}_n)^T$ is given as the estimates in the previous iteration step, then minimizing $G(\beta, \hat{r})$ over $\beta$ is equivalent to

$$\hat{\beta} = \arg\min_\beta \sum_{i=1}^n \left| \frac{y_i - x_i^T \beta}{\sigma} - \hat{r}_i \right| = \arg\min_\beta \sum_{i=1}^n \left| t_i - x_i^T \beta \right| \tag{3.5}$$

with $t_i = y_i - \sigma \hat{r}_i$ for $i = 1, \ldots, n$. As discussed in Section 2, if $\hat{r}_i \neq 0$ then $t_i = x_i^T \hat{\beta}^{\text{old}}$ so that the previous estimate $\hat{\beta}^{\text{old}}$ is preferred for the next update. Thus, $\beta$ is updated using observations having $\hat{r}_i = 0$. In this case,

$$\hat{\beta} = \arg\min_\beta \sum_{i:\hat{r}_i = 0} \left| y_i - x_i^T \beta \right| \tag{3.6}$$

since $t_i = y_i$ with $\hat{r}_i = 0$. (3.6) is LAD problem with the data $\{(x_i, y_i) : \hat{r}_i = 0\}$. LAD is efficiently solved in linear programming and its solvers are available in most standard statistical softwares (for example, `quanreg` or `L1pack` packages in R).

**Algorithm 1** : LCAD algorithm

Step 1 : Initialize $\hat{\beta}$ as LAD estimate. And compute the residual vector $e = y - X^T \hat{\beta}$ and set $\hat{\sigma} = \text{Med}(|e - \text{SkippedMed}(e)|)$.

Step 2 : Repeat until convergence is met

(a) Compute $z_i = (y_i - x_i^T \hat{\beta})/\hat{\sigma}$ and set $\hat{r}_i = z_i I(|z_i| \geq a)$ for $i = 1, \ldots, n$.

(b) Update $\hat{\beta}$ as a LAD solution (3.6) with data $\{(x_i, y_i) : \hat{r}_i = 0\}$.

Before invoking the iterative algorithm, we need initialization for $\beta$ and the scale parameter $\sigma$. We adopt the initialization scheme in robust regression (Maronna *et al.*, 2019). The initial value $\hat{\beta}$ is set by LAD estimate, which has high breakdown point. Then, $\hat{\sigma}$ is set by the modified MAD of the residuals, computed as $\hat{\sigma} = \text{Med}(|e_i - \text{SkippedMed}(e_i)|)$ with $e_i = y_i - x_i^T \hat{\beta}$, where $\text{SkippedMed}(e_i)$ is the skipped median of $e_1, \ldots, e_n$. Since LAD solution is vulnerable to outliers, we change the center of residuals by its robust version. This previously estimated scale parameter $\hat{\sigma}$ is computed in the initial stage and is not changed during later iterations. We summarize the algorithm in **Algorithm 1**.

This alternate fitting algorithm is similar to the mean-shift model used in McCann and Welsh (2007) and She and Owen (2011). These works deal with robust regression using squared loss, different from the absolute loss in this work. In our approach, minimizing the objective function (3.3) is equivalent to fitting the center-shift model $y_i = x_i^T \beta + r_i + u_i$ by introducing shift parameter $r_i$. As in She and Owen (2011), shift parameter $r_i$ can be regarded as an indicator for outlier; $\hat{r}_i \neq 0$ indicates

Table 2: Average of 100 test MSEs and its standard error (in parenthesis) are presented for the case of $b = 3$

| $n$ | $p$ | $\pi$ | TRUE | LS | LAD | Huber | Tukey | She & Owen | LCAD |
|---|---|---|---|---|---|---|---|---|---|
| 200 | 1 | 0.0 | 0.997 (0.004) | 1.007 (0.005) | 1.013 (0.005) | 1.008 (0.005) | 1.008 (0.005) | 1.004 (0.004) | 1.020 (0.005) |
| | | 0.1 | 0.997 (0.004) | 1.102 (0.007) | 1.036 (0.006) | 1.047 (0.006) | 1.039 (0.006) | 1.010 (0.005) | 1.021 (0.005) |
| | | 0.2 | 0.997 (0.004) | 1.375 (0.011) | 1.124 (0.009) | 1.232 (0.009) | 1.214 (0.009) | 1.012 (0.005) | 1.042 (0.007) |
| | | 0.3 | 0.997 (0.004) | 1.829 (0.016) | 1.335 (0.014) | 1.725 (0.015) | 1.670 (0.015) | 1.019 (0.005) | 1.160 (0.014) |
| | 5 | 0.0 | 0.995 (0.004) | 1.025 (0.005) | 1.045 (0.006) | 1.027 (0.005) | 1.028 (0.005) | 1.027 (0.005) | 1.069 (0.006) |
| | | 0.1 | 0.995 (0.004) | 1.131 (0.008) | 1.073 (0.007) | 1.072 (0.007) | 1.065 (0.007) | 1.048 (0.006) | 1.075 (0.007) |
| | | 0.2 | 0.995 (0.004) | 1.406 (0.012) | 1.171 (0.010) | 1.270 (0.011) | 1.254 (0.011) | 1.063 (0.006) | 1.104 (0.010) |
| | | 0.3 | 0.995 (0.004) | 1.867 (0.017) | 1.434 (0.018) | 1.763 (0.017) | 1.721 (0.017) | 1.088 (0.008) | 1.278 (0.020) |
| | 10 | 0.0 | 0.999 (0.004) | 1.059 (0.005) | 1.093 (0.006) | 1.062 (0.005) | 1.063 (0.005) | 1.063 (0.005) | 1.140 (0.009) |
| | | 0.1 | 0.999 (0.004) | 1.198 (0.008) | 1.142 (0.007) | 1.132 (0.007) | 1.128 (0.007) | 1.102 (0.007) | 1.158 (0.009) |
| | | 0.2 | 0.999 (0.004) | 1.505 (0.012) | 1.288 (0.012) | 1.367 (0.012) | 1.355 (0.012) | 1.140 (0.007) | 1.235 (0.013) |
| | | 0.3 | 0.999 (0.004) | 1.995 (0.018) | 1.650 (0.024) | 1.900 (0.019) | 1.872 (0.019) | 1.191 (0.010) | 1.561 (0.029) |
| 400 | 1 | 0.0 | 0.997 (0.004) | 1.002 (0.004) | 1.004 (0.004) | 1.002 (0.004) | 1.002 (0.004) | 1.001 (0.004) | 1.007 (0.005) |
| | | 0.1 | 0.997 (0.004) | 1.094 (0.005) | 1.026 (0.005) | 1.040 (0.004) | 1.032 (0.004) | 1.003 (0.004) | 1.008 (0.004) |
| | | 0.2 | 0.997 (0.004) | 1.365 (0.007) | 1.108 (0.006) | 1.223 (0.007) | 1.206 (0.007) | 1.006 (0.005) | 1.021 (0.005) |
| | | 0.3 | 0.997 (0.004) | 1.816 (0.010) | 1.319 (0.010) | 1.703 (0.010) | 1.654 (0.010) | 1.008 (0.005) | 1.132 (0.009) |
| | 5 | 0.0 | 0.995 (0.004) | 1.009 (0.005) | 1.017 (0.005) | 1.010 (0.005) | 1.010 (0.005) | 1.012 (0.005) | 1.024 (0.005) |
| | | 0.1 | 0.995 (0.004) | 1.115 (0.006) | 1.046 (0.005) | 1.057 (0.005) | 1.049 (0.005) | 1.021 (0.005) | 1.032 (0.005) |
| | | 0.2 | 0.995 (0.004) | 1.400 (0.009) | 1.150 (0.008) | 1.259 (0.008) | 1.244 (0.008) | 1.031 (0.005) | 1.061 (0.006) |
| | | 0.3 | 0.995 (0.004) | 1.861 (0.012) | 1.400 (0.014) | 1.752 (0.012) | 1.710 (0.012) | 1.043 (0.005) | 1.218 (0.014) |
| | 10 | 0.0 | 0.999 (0.004) | 1.029 (0.005) | 1.048 (0.005) | 1.031 (0.005) | 1.031 (0.005) | 1.032 (0.005) | 1.067 (0.005) |
| | | 0.1 | 0.999 (0.004) | 1.142 (0.006) | 1.081 (0.006) | 1.082 (0.005) | 1.075 (0.005) | 1.053 (0.005) | 1.074 (0.006) |
| | | 0.2 | 0.999 (0.004) | 1.430 (0.008) | 1.186 (0.008) | 1.293 (0.008) | 1.281 (0.008) | 1.077 (0.006) | 1.112 (0.007) |
| | | 0.3 | 0.999 (0.004) | 1.891 (0.011) | 1.456 (0.013) | 1.789 (0.012) | 1.752 (0.012) | 1.101 (0.007) | 1.286 (0.014) |

that the $i^{th}$ observation is an outlier because its center must be shifted to adjust its location to the assumed model.

This algorithm turns a nonconvex LCAD problem into a convex LAD problem as a subroutine. This is beneficial since we can use any well-developed LAD solver for LCAD problem. Furthermore, the extension of LAD can be generalized to LCAD problem. For example, we can consider regularization for estimation of regression coefficient $\beta$ by imposing penalty function $p_\lambda(\beta)$ to the objective function $L(\beta)$ in (3.2). Since the subproblem for $\beta$ estimation in LCAD becomes LAD problem, it is straightforward to implement the regularized LCAD by simply replacing (3.6) in Step 2(b) of **Algorithm 1** by the penalized LAD problem

$$\hat{\beta} = \arg\min_{\beta} \sum_{i:\hat{r}_i=0} \left| y_i - x_i^T \beta \right| + p_\lambda(\beta). \tag{3.7}$$

Note that the penalization does not affect the estimation for $r_i$. Some penalized LAD methods have been studied and implemented already. For example, LAD has been combined with $L_1$ penalization (Gao and Huang, 2010), sparse estimation (Wang and Zhu, 2017), fused lasso (Liu *et al.*, 2018), functional regression (Cardot *et al.*, 2005) and many others. Thus, LCAD embraces these works for its regularization without any additional implementation tailored to the nonconvex $\rho_a(u)$.

## 4. Simulation

We consider the standard linear model (3.1) with normal error distribution $u_i \sim F(u) \equiv N(0,1)$. The covariates $x_{ij}$ ($i = 1,\ldots,n$, $j = 1,\ldots,p$) are independently generated from uniform distribution on the interval $(-3,3)$. Since the error distribution is symmetric about mean 0, $f(x) = x^T\beta$ is the conditional median as well as conditional mean of the response given covariate $x$. We demonstrate the performance of LCAD and compare it with LS and LAD in the standard linear regression with contaminated normal noise. To simulate the situation where outliers are involved, we consider a contaminated error distribution with a mixture normal $u \sim (1-\pi)N(0,1) + \pi N(b,1)$.

Table 3: Average of 100 test MSEs and its standard error (in parenthesis) are presented for the case of $b = 9$

| $n$ | $p$ | $\pi$ | TRUE | LS | LAD | Huber | Tukey | She & Owen | LCAD |
|---|---|---|---|---|---|---|---|---|---|
| 200 | 1 | 0.0 | 0.997 (0.004) | 1.007 (0.005) | 1.013 (0.005) | 1.008 (0.005) | 1.008 (0.005) | 1.004 (0.004) | 1.020 (0.005) |
| | | 0.1 | 0.997 (0.004) | 1.857 (0.017) | 1.035 (0.006) | 1.050 (0.006) | 1.009 (0.005) | 1.040 (0.007) | 1.020 (0.005) |
| | | 0.2 | 0.997 (0.004) | 4.313 (0.032) | 1.125 (0.009) | 1.306 (0.012) | 1.011 (0.005) | 1.076 (0.012) | 1.020 (0.005) |
| | | 0.3 | 0.997 (0.004) | 8.401 (0.051) | 1.349 (0.015) | 7.568 (0.055) | 1.015 (0.005) | 1.118 (0.017) | 1.021 (0.005) |
| | 5 | 0.0 | 0.995 (0.004) | 1.025 (0.005) | 1.045 (0.006) | 1.027 (0.005) | 1.028 (0.005) | 1.027 (0.005) | 1.069 (0.006) |
| | | 0.1 | 0.995 (0.004) | 2.013 (0.023) | 1.073 (0.007) | 1.077 (0.007) | 1.029 (0.005) | 1.236 (0.016) | 1.066 (0.006) |
| | | 0.2 | 0.995 (0.004) | 4.543 (0.039) | 1.173 (0.010) | 1.383 (0.015) | 1.033 (0.005) | 1.393 (0.022) | 1.063 (0.006) |
| | | 0.3 | 0.995 (0.004) | 8.740 (0.056) | 1.459 (0.019) | 8.031 (0.062) | 1.048 (0.006) | 1.616 (0.036) | 1.070 (0.007) |
| | 10 | 0.0 | 0.999 (0.004) | 1.059 (0.005) | 1.093 (0.006) | 1.062 (0.005) | 1.063 (0.005) | 1.063 (0.005) | 1.140 (0.009) |
| | | 0.1 | 0.999 (0.004) | 2.283 (0.028) | 1.143 (0.008) | 1.142 (0.007) | 1.070 (0.005) | 1.483 (0.026) | 1.138 (0.008) |
| | | 0.2 | 0.999 (0.004) | 5.016 (0.049) | 1.298 (0.013) | 1.558 (0.019) | 1.082 (0.005) | 1.840 (0.037) | 1.144 (0.008) |
| | | 0.3 | 0.999 (0.004) | 9.378 (0.080) | 1.744 (0.031) | 8.797 (0.085) | 1.170 (0.022) | 2.319 (0.058) | 1.152 (0.009) |
| 400 | 1 | 0.0 | 0.997 (0.004) | 1.002 (0.004) | 1.004 (0.004) | 1.002 (0.004) | 1.002 (0.004) | 1.001 (0.004) | 1.007 (0.005) |
| | | 0.1 | 0.997 (0.004) | 1.829 (0.011) | 1.026 (0.005) | 1.042 (0.005) | 1.003 (0.004) | 1.020 (0.005) | 1.005 (0.004) |
| | | 0.2 | 0.997 (0.004) | 4.270 (0.021) | 1.110 (0.006) | 1.293 (0.008) | 1.003 (0.004) | 1.040 (0.008) | 1.006 (0.004) |
| | | 0.3 | 0.997 (0.004) | 8.325 (0.029) | 1.330 (0.010) | 7.461 (0.037) | 1.006 (0.004) | 1.050 (0.010) | 1.006 (0.004) |
| | 5 | 0.0 | 0.995 (0.004) | 1.009 (0.005) | 1.017 (0.005) | 1.010 (0.005) | 1.010 (0.005) | 1.012 (0.005) | 1.024 (0.005) |
| | | 0.1 | 0.995 (0.004) | 1.930 (0.014) | 1.046 (0.005) | 1.061 (0.005) | 1.012 (0.005) | 1.099 (0.008) | 1.026 (0.005) |
| | | 0.2 | 0.995 (0.004) | 4.461 (0.028) | 1.153 (0.008) | 1.352 (0.011) | 1.014 (0.005) | 1.206 (0.012) | 1.029 (0.005) |
| | | 0.3 | 0.995 (0.004) | 8.576 (0.042) | 1.419 (0.015) | 7.842 (0.052) | 1.023 (0.005) | 1.306 (0.018) | 1.029 (0.005) |
| | 10 | 0.0 | 0.999 (0.004) | 1.029 (0.005) | 1.048 (0.005) | 1.031 (0.005) | 1.031 (0.005) | 1.032 (0.005) | 1.067 (0.005) |
| | | 0.1 | 0.999 (0.004) | 2.028 (0.015) | 1.081 (0.006) | 1.087 (0.005) | 1.034 (0.005) | 1.227 (0.011) | 1.066 (0.005) |
| | | 0.2 | 0.999 (0.004) | 4.623 (0.030) | 1.190 (0.008) | 1.408 (0.011) | 1.038 (0.005) | 1.423 (0.017) | 1.066 (0.006) |
| | | 0.3 | 0.999 (0.004) | 8.768 (0.044) | 1.489 (0.015) | 8.128 (0.053) | 1.051 (0.005) | 1.633 (0.028) | 1.069 (0.006) |

From this mixture normal error model, we generate training data of size $n$. Since the error distribution is symmetric about 0, the conditional median response for legitimate observations are $x_i^T\beta$, whose coefficient $\beta$ is expected to be estimated correctly even under the existence of outlier. In simulation, we set the sample size $n = (200, 400)$ and the number of variables $p = (1, 5, 10)$. True coefficients are set by $\beta_j = 2$ for $j = 1, \ldots, p$, and no intercept is involved in the regression model. We consider $\pi = (0.1, 0.2, 0.3)$ and $b = (3, 9)$ for outlier distribution specification. In order to evaluate the performance, we generate a clean test data of size 1000 from the same model without outliers and compute the test mean squared error (MSE) for prediction. This procedure is repeated 100 times, and the resulting 100 test MSE's are summarized in average (standard error in parenthesis) in Table 2 for $b = 3$ and Table 3 for $b = 9$. 'TRUE' denotes for test MSE using true coefficient of $\beta_j$'s and it is provided as a benchmark for comparison. We apply least clipped absolute deviation ('LCAD') to the noisy training data, as well as several benchmark methods including least squares ('LS'), least absolute deviation ('LAD'), two types of M-estimation ('Huber' and 'Tukey'), and the proposal by She and Owen (2011) ('She & Owen'). Test MSE is computed based on the estimates from these methods. We use `lad()` function provided by `L1pack` package in R for LAD implementation. The same solver is also used for a LAD subroutine in LCAD optimization to make a fair comparison.

From Table 2 and Table 3, we see that all other methods are worse than 'TRUE', which implies that their performance is affected from outliers. As the proportion of outliers increases, their MSEs increase as well. LS is the worst performer as we expected. Although LAD, which is known as a robust estimation, improves LS, LCAD performs better than LAD. Note that, comparing to the results in Table 3, the improvement of LCAD over LAD seems marginal in Table 2. This is because the outlier distribution with a low $b = 3$ is overlapped with the legitimate distribution so that the outliers in the data are not clearly treated by the procedure. However, LCAD improves much better than LAD in Table 3, representing the situation where outliers are well separated from the legitimate distribution with a high $b = 9$. The performance of 'Tukey' and 'She & Owen' is comparable to or even better than LCAD. This result is expected since the bisquare loss for 'Tukey' and the clipped squared loss for 'She & Owen' are noncovex as the clipped absolute loss for LCAD is. 'Tukey' is the best for

Table 4: Average of 100 false positive rates and its standard error (in parenthesis) are presented

| $p$ | $b$ | $\pi$ | LASSO | L1-LAD | L1-LCAD |
|---|---|---|---|---|---|
| 10 | 3 | 0.0 | 0.0480 (0.0099) | 0.0180 (0.0058) | 0.1020 (0.0144) |
| | | 0.1 | 0.1120 (0.0169) | 0.0140 (0.0051) | 0.1160 (0.0151) |
| | | 0.2 | 0.1360 (0.0188) | 0.0260 (0.0073) | 0.1660 (0.0184) |
| | 9 | 0.0 | 0.0480 (0.0099) | 0.0180 (0.0058) | 0.1020 (0.0144) |
| | | 0.1 | 0.3300 (0.0258) | 0.0060 (0.0034) | 0.0800 (0.0124) |
| | | 0.2 | 0.3800 (0.0291) | 0.0100 (0.0044) | 0.1100 (0.0140) |
| 20 | 3 | 0.0 | 0.0127 (0.0028) | 0.0087 (0.0028) | 0.0707 (0.0080) |
| | | 0.1 | 0.0433 (0.0065) | 0.0127 (0.0030) | 0.0727 (0.0083) |
| | | 0.2 | 0.0753 (0.0089) | 0.0200 (0.0037) | 0.1040 (0.0099) |
| | 9 | 0.0 | 0.0127 (0.0028) | 0.0087 (0.0028) | 0.0707 (0.0080) |
| | | 0.1 | 0.1673 (0.0127) | 0.0040 (0.0016) | 0.0480 (0.0070) |
| | | 0.2 | 0.2293 (0.0158) | 0.0113 (0.0032) | 0.0660 (0.0078) |
| 50 | 3 | 0.0 | 0.0093 (0.0019) | 0.0056 (0.0016) | 0.0318 (0.0038) |
| | | 0.1 | 0.0280 (0.0037) | 0.0080 (0.0017) | 0.0349 (0.0040) |
| | | 0.2 | 0.0416 (0.0047) | 0.0167 (0.0021) | 0.0644 (0.0052) |
| | 9 | 0.0 | 0.0093 (0.0019) | 0.0056 (0.0016) | 0.0318 (0.0038) |
| | | 0.1 | 0.0929 (0.0071) | 0.0029 (0.0008) | 0.0273 (0.0036) |
| | | 0.2 | 0.1282 (0.0092) | 0.0129 (0.0020) | 0.0329 (0.0027) |
| 100 | 3 | 0.0 | 0.0049 (0.0011) | 0.0040 (0.0010) | 0.0199 (0.0022) |
| | | 0.1 | 0.0145 (0.0019) | 0.0061 (0.0015) | 0.0259 (0.0022) |
| | | 0.2 | 0.0202 (0.0022) | 0.0119 (0.0016) | 0.0392 (0.0031) |
| | 9 | 0.0 | 0.0049 (0.0011) | 0.0040 (0.0010) | 0.0199 (0.0022) |
| | | 0.1 | 0.0533 (0.0048) | 0.0033 (0.0006) | 0.0178 (0.0021) |
| | | 0.2 | 0.0699 (0.0051) | 0.0116 (0.0013) | 0.0207 (0.0022) |
| 200 | 3 | 0.0 | 0.0026 (0.0005) | 0.0027 (0.0005) | 0.0111 (0.0013) |
| | | 0.1 | 0.0088 (0.0011) | 0.0041 (0.0005) | 0.0168 (0.0013) |
| | | 0.2 | 0.0124 (0.0015) | 0.0076 (0.0010) | 0.0253 (0.0017) |
| | 9 | 0.0 | 0.0026 (0.0005) | 0.0027 (0.0005) | 0.0111 (0.0013) |
| | | 0.1 | 0.0337 (0.0027) | 0.0026 (0.0005) | 0.0094 (0.0008) |
| | | 0.2 | 0.0454 (0.0032) | 0.0095 (0.0010) | 0.0149 (0.0023) |

moderate outliers ($b = 3$ in Table 2) and 'She & Owen' outperforms for severe outliers ($b = 9$ in Table 3). However, in both cases, LCAD performs comparable to these best performers. Note that LCAD is to estimate the conditional median line of the regression and 'She & Owen' produces the conditional mean line, while 'Tukey' is not obvious for its interpretation. Since the mean and median are identical under the normal model as in this simulation, these three estimates are expected to be similar. If, however, the underlying model is not symmetric, then those methods would yield the different results according to their purpose.

In the following simulation, we demonstrate how well LCAD performs in variable selection when outliers exist in training data. We generate, same as in the previous simulation, noisy training data of size $n = 100$ from the linear regression model (3.1) without intercept under the mixture normal error distribution. We consider $p = (10, 20, 50.100.200)$, $b = (3, 9)$, and $\pi = (0.0, 0.1, 0.2)$. For variable selection, we assume that a part of explanatory variables are significantly associated with the response and the remaining variables are redundant. To mimic this situation, we set $\beta_1 = \cdots = \beta_5 = 2$ and $\beta_6 = \cdots = \beta_p = 0$. This implies that the first 5 explanatory variables are important for prediction in response while the others are not. Thus, a successful method should be able to select the first 5 explanatory variables in the final model. In fitting algorithm, we apply $L_1$ penalty $p_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|$ on regression coefficient to promote a sparse solution. We use `hqreg` package in R for the penalized LAD ('L1-LAD') and the subroutine for penalized LCAD ('L1-LCAD'). The penalty parameter $\lambda$ is tuned by cross validation.

We compute false positive rate (proportion of $\hat{\beta}_j \neq 0$ among $\beta_6, \ldots, \beta_p$) through 100 repetitions and present their results in Table 4. LASSO, based on squared loss, is less successful to find the correct variables by showing large false positive rates, including nonsignificant variables in the final

Table 5: Median of 100 test MSEs and its MAD (in parenthesis) are presented

| $p$ | $b$ | $\pi$ | LASSO | L1-LAD | L1-LCAD |
|---|---|---|---|---|---|
| 10 | 3 | 0.0 | 1.0969 (0.0513) | 1.2550 (0.0859) | 1.2160 (0.0882) |
| | | 0.1 | 1.2565 (0.0737) | 1.3975 (0.1530) | 1.2615 (0.0945) |
| | | 0.2 | 1.5863 (0.1059) | 1.6264 (0.1946) | 1.3752 (0.1540) |
| | 9 | 0.0 | 1.0969 (0.0513) | 1.2550 (0.0859) | 1.2160 (0.0882) |
| | | 0.1 | 2.5827 (0.3561) | 1.8775 (0.2474) | 1.1982 (0.0743) |
| | | 0.2 | 5.5191 (0.5663) | 2.2480 (0.3672) | 1.1836 (0.0654) |
| 20 | 3 | 0.0 | 1.1503 (0.0552) | 1.3524 (0.1151) | 1.3507 (0.1372) |
| | | 0.1 | 1.3529 (0.1082) | 1.5363 (0.1917) | 1.4091 (0.1374) |
| | | 0.2 | 1.6727 (0.1559) | 1.8328 (0.2911) | 1.6326 (0.2112) |
| | 9 | 0.0 | 1.1503 (0.0552) | 1.3524 (0.1151) | 1.3507 (0.1372) |
| | | 0.1 | 2.9702 (0.3597) | 1.9629 (0.2134) | 1.2952 (0.1117) |
| | | 0.2 | 6.0470 (0.6938) | 2.5148 (0.4371) | 1.3316 (0.1439) |
| 50 | 3 | 0.0 | 1.2171 (0.0972) | 1.4182 (0.1631) | 1.5157 (0.2098) |
| | | 0.1 | 1.4745 (0.1159) | 1.6646 (0.1919) | 1.5948 (0.1992) |
| | | 0.2 | 1.8740 (0.2035) | 2.0005 (0.3759) | 1.9982 (0.4126) |
| | 9 | 0.0 | 1.2171 (0.0972) | 1.4182 (0.1631) | 1.5157 (0.2098) |
| | | 0.1 | 3.5408 (0.5043) | 2.0519 (0.3061) | 1.4341 (0.1489) |
| | | 0.2 | 7.2305 (1.0202) | 2.5977 (0.6521) | 1.4403 (0.1359) |
| 100 | 3 | 0.0 | 1.2583 (0.0843) | 1.4965 (0.1511) | 1.6724 (0.2025) |
| | | 0.1 | 1.6015 (0.1957) | 1.7381 (0.2752) | 1.9242 (0.3751) |
| | | 0.2 | 2.0254 (0.2369) | 2.1728 (0.4318) | 2.3803 (0.5616) |
| | 9 | 0.0 | 1.2583 (0.0843) | 1.4965 (0.1511) | 1.6724 (0.2025) |
| | | 0.1 | 3.9285 (0.6592) | 2.2364 (0.4416) | 1.5494 (0.2073) |
| | | 0.2 | 8.1748 (1.0662) | 3.0423 (0.9332) | 1.4883 (0.1904) |
| 200 | 3 | 0.0 | 1.3317 (0.0902) | 1.6137 (0.1590) | 1.9211 (0.3127) |
| | | 0.1 | 1.6794 (0.1668) | 1.8606 (0.2773) | 2.1526 (0.5002) |
| | | 0.2 | 2.1615 (0.2360) | 2.5240 (0.4885) | 2.9911 (0.8061) |
| | 9 | 0.0 | 1.3317 (0.0902) | 1.6137 (0.1590) | 1.9211 (0.3127) |
| | | 0.1 | 4.6366 (0.7889) | 2.3286 (0.4044) | 1.7382 (0.2372) |
| | | 0.2 | 8.9148 (1.2367) | 3.3533 (0.9577) | 1.7410 (0.2593) |

model. L1-LAD is the best performer and is even better than L1-LCAD in terms of false positive rate. We do not include true positive rate here since all methods show 100% true positive rates for most cases. Variable selection performance influences prediction. To see this, we provide test MSE in Table 5, showing the clear outperformance of L1-LCAD among candidates when the proportion of outliers increases.

## 5. Concluding remarks

From this work, we demonstrate the usefulness of skipped median, not only for robust location estimation, but also for robust estimation of conditional median response in regression problem. We develop a novel estimating algorithm for skipped median in the univariate case. The proposed iterative algorithm makes use of convex subproblem to the nonconvex optimization for skipped median estimation. This convex relaxation enables us to generalize it to regression problem and use LAD optimization as a subproblem for LCAD. Furthermore, we are able to apply various regularizations to LCAD whenever such regularizations are developed and implemented for LAD.

Skipped median is useful in asymmetric outlier distribution. However, skipped-median based approaches are not popularly studied or used in practice. This is partly because the corresponding loss $\rho_a(u)$ is not continuously differentiable while there are continuously differentiable loss functions in M-estimation (Tukey's biweight as a typical loss) which provide stable estimation. However, skipped median and LCAD provide a robust median estimate. Median is a quantile statistic that is more informative than a center of symmetry.

In this work we consider a symmetric distribution for legitimate errors since the symmetric error is often assumed in regression. For an asymmetric distribution, we can extend the skipping procedure to

find the median. In this case, asymmetric skipping is required so that the corresponding loss, instead of (2.2), becomes an asymmetrically clipped absolute deviation loss

$$\rho_{a_1,a_2}(u) = \begin{cases} a_1, & u \le -a_1, \\ |u|, & -a_1 \le u \le a_2, \\ a_2, & u \ge a_2 \end{cases} \tag{5.1}$$

with $a_1 = (F^{-1}(0.5) - F^{-1}(\epsilon))/\sigma$ and $a_2 = (F^{-1}(1-\epsilon) - F^{-1}(0.5))/\sigma$, where $F^{-1}(\cdot)$ is a quantile function and $0 < \epsilon < 0.5$ is a given skipping rate. With (5.1), the skipped median can be deduced from asymmetric distribution $F$. As another extension of this work, skipping procedure can be applied to robust quantile regression when outliers exist. The corresponding loss function can be obtained by clipping $\rho_\tau(u) = u(\tau - I(u < 0))$ at asymmetric points. These extensions are left to our future works.

## References


Cardot H, Crambes C, and Sarda P (2005). Quantile regression when the covariates are functions, *Journal of Nonparametric Statistics*, **17**, 841–856.

Gao X and Huang J (2010). Asymptotic analysis of high-dimensional LAD regression with lasso, *Statistica Sinica*, **20**, 1485–1506.

Hampel FR, Ronchetti EM, Rousseeuw PJ, and Stahel WA (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, Toronto.

Huber PJ (1984). Finite sample breakdown of M- and P-estimators. *The Annals of Statistics*, **12**, 119–126.

Koenker R (2005). *Quantile Regression*, Cambridge University Press, Cambridge.

Liu Y, Tao J, Zhang H, Xiu X, and Kong L (2018). Fused lasso penalized least absolute deviation estimator for high dimensional linear regression, *Numerical Algebra, Control and Optimization*, **8**, 97–117.

Maronna RA, Martin RD, Yohai VJ, and Salibian-Barrera M (2019). *Robust Statistics: Theory and Methods* (2nd ed), John Wiley & Sons, NewJersy.

McCann L and Welsch RE (2007). Robust variable selection using least angle regression and elemental set sampling, *Computational Statistics and Data Analysis*, **52**, 249–257.

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

Wang Y and Zhu L (2017). Variable selection and parameter estimation via WLAD-SCAD with a diverging number of parameters, *Journal of the Korean Statistical Society*, **46**, 390–403.

Yohai VJ (1987). High breakdown-point and high efficiency estimates for regression, *The Annals of Statistics*, **15**, 642–665.


*Received July 28, 2022; Revised September 19, 2022; Accepted September 26, 2022*