Statistics and Its Interface Volume 11 (2018) 79–89

# Penalized weighted least absolute deviation regression*

Xiaoli Gao† and Yang Feng‡,§

In a linear model where the data is contaminated or the random error is heavy-tailed, least absolute deviation (LAD) regression has been widely used as an alternative approach to least squares (LS) regression. However, it is well known that LAD regression is not robust to outliers in the explanatory variables. When the data includes some leverage points, LAD regression may perform even worse than LS regression. In this manuscript, we propose to improve LAD regression in a penalized weighted least absolute deviation (PWLAD) framework. The main idea is to associate each observation with a weight reflecting the degree of outlying and leverage effect and obtain both the weight and coefficient vector estimation simultaneously and adaptively. The proposed PWLAD is able to provide regression coefficients estimate with strong robustness, and perform outlier detection at the same time, even when the random error does not have finite variances. We provide sufficient conditions under which PWLAD is able to identify true outliers consistently. The performance of the proposed estimator is demonstrated via extensive simulation studies and real examples.

AMS 2000 subject classifications: Primary 62F35, 62F12; secondary 62P35.
Keywords and phrases: Lasso, Leverage points, Outlier detection, Robust regression, Weighted least absolute deviation.

## 1. INTRODUCTION

Given $n$ observation pairs $(\mathbf{x}_i, y_i), i = 1, \cdots, n$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})'$ is a $p$-dimensional predictor and $y_i$ is the response, we consider the following linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}^* + \varepsilon_i, \quad 1 \le i \le n, \tag{1}$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_p^*)'$ is the true coefficients and $\{\varepsilon_i, i = 1, \cdots, n\}$ are independent random errors such that $\varepsilon_i / w_i^{*2}$ are identically distributed with $\max_i w_i^{*2} = 1$. Here, for $1 \le i \le n$, the weights $0 < w_i^{*2} \le 1$ represent the heterogeneity of the errors with $w_i < 1$ representing an outlier and $w_i = 1$ representing a "normal" observation. Let $\mathcal{O} = \{1 \le i \le n : 0 < w_i^* < 1\}$ be the true outlier set and the goal of outlier detection is to recover the set $\mathcal{O}$ accurately. Here, it is assumed the size of $\mathcal{O}$ is smaller than $n/2$, i.e., the majority of the observations are "normal" observations. We set $x_{i1} = 1$ for $1 \le i \le n$ if an intercept is included in the regression model.

Least squares (LS) regression is often used when $\{\varepsilon_i, i = 1, \cdots, n\}$ are well behaved and $w_i^* = 1$ for all $i$. It is well known that LS is lack of robustness and strongly sensitive to outliers. When the data is contaminated or the random error is heavy tailed, least absolute deviation (LAD) regression is considered to be a good alternative to LS regression. An LAD regression estimates the coefficient vector by minimizing the $\ell_1$ loss,

$$\widetilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} |y_i - \mathbf{x}_i'\boldsymbol{\beta}| \right\}, \tag{2}$$

where $\mathbf{y} = (y_1, \cdots, y_n)'$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$. However, it is well known that LAD regression is also lack of robustness when the data include outliers in the explanatory variables [28] (i.e., there exist leverage points). In this scenario, LAD regression may perform even worse than LS regression.

One global measure of an estimator's robustness is the breakdown value [19]. The finite-sample breakdown value is the minimum proportion of observations that, if arbitrarily modified, can cause the estimates to increase above any bound. The range of possible value of breakdown value is between $1/n$ and $1/2$. The higher, the more robust a regression approach can be. The LAD estimator has a smallest breakdown value of $1/n$. This is mainly due to the fact that the LAD estimator is not robust to leverage points. One can refer [14] for a detailed discussion on the breakdown value of the LAD estimator.

There are many robust procedures with high breakdown value. See for example, the least median of squares [29], the least trimmed squares [25], S-estimates [24], Generalized S-estimates [7], MM-estimates [36], the robust and efficient weighted least squares estimators [13], and forward search [2]. One can refer to [19] and [15] for broader reviews of some recent robust regression procedures.

In the framework of LAD regression, the robustness can also be improved by down-weighting those leverage points

*The authors would like to thank the Editor, the AE and two referees for the insightful comments which led to substantial improvement over an earlier version.
†Partially supported by Simons Foundation #359337.
‡Partially supported by NSF CAREER grant DMS-1554804.
§Corresponding author. E-mail address: yang.feng@columbia.edu.

which are detected in advance. For example, [14] proposed a weighted LAD (WLAD) regression to improve the breakdown value of LAD regression. In WLAD regression, each observation is assigned a weight in advance, and the final estimation is expected be robust to the outliers in $x$ direction if those pre-assigned weights correctly reflect the outlying information among all covariates. The idea of including weight under the $\ell_1$ loss is also considered for dealing with possibly infinite variances in several time series models such as [18, 23, 22].

As pointed out by [34] and to be demonstrated by our simulation studies, the robustness of WLAD can be significantly deteriorated by a high percentage of outliers, especially when multiple covariates exist and the outlying percentage is high, even in the case of no leverage points. An alternative approach to robust LAD regression is to adopt the trimming procedure during the LAD regression, such as the Least Trimmed Absolute Deviation (LTAD) estimator studied by [16, 32, 33]. The LTAD was also extended to least trimmed quantile regression [21], where the robust properties of LTAD were discussed as a special case under different trimming percentages. LTAD can also be interpreted as one special type of WLAD by assigning some observations with weight 0 and most others with weight 1.

However, none of the above robust LAD procedures are designed for simultaneous detection of the outliers and robust estimation. On one hand, the WLAD estimator is based upon a pre-assigned weight parameter depending upon a pre-selected clean subset. If a large amount of leverage points exist, the clean subset and the produced weight assignment can be misleading, which cause the WLAD estimates to be severely biased. On the other hand, the performance of LTAD methods also depends closely on the chosen trimming percentage. As pointed in [21], the LTAD can be strongly biased if the assigned trimming rate is less than the true contamination rate. On the contrary, if the trimming percentage is higher than the contamination rate, the estimation can have a large variation even though the bias can be corrected. Thus, having the information on the true trimming rate plays an important role in obtaining an LTAD estimate with strong robustness. An ideal approach should take into account the weight estimation (or outlier detection) simultaneously with the coefficients estimation.

The study of outlier detection has a long history. See [8, 4, 30], among others. In this paper, we propose a new method named penalized weighted least absolute deviation regression (PWLAD) for simultaneous outlier identification and robust LAD estimation. PWLAD borrows ideas from recent work on penalized weighted least squares regression (PWLS) models [11, 12] to deal with the scenarios where the random error may have certain heteroscedasticity or infinite variances.

PWLAD associates each observation with a weight and obtain both the weight and regression coefficient estimates simultaneously using a lasso-type penalty on the weight vector. It leads to estimators with potentially strong robustness, and at the same time, produces the observations' outlying information. Even under the scenario where the data include both contaminated observations (in both $y$ and $x$ directions) and the random errors may not have finite variance, the proposed PWLAD is still able to detect corresponding outliers and provide robust regression coefficient estimates.

The remainder of the paper is organized as follows. In Section 2, we introduce the PWLAD estimator. A corresponding Bayesian interpretation and model implementation are also presented in Section 2. In Section 3, we investigate the theoretical properties of the model regarding outlier detection. In particular, we provide sufficient conditions under which PWLAD estimator is able to separate outliers from normal observations with high probability. In Section 4, we present extensive simulation studies and two real data examples by comparing the proposed approach with some popular methods. We conclude the paper with a discussion in Section 5 and provide all proofs in Section 6.

## 2. PWLAD: METHOD AND IMPLEMENTATION

In the linear model specified in (1), a penalized weighted least absolute deviations (PWLAD) estimator of $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ and $\mathbf{w} = (w_1, \cdots, w_n)'$, is to minimize a penalized objective function consisting of weighted $\ell_1$ loss and a penalty on the weight vector,

(3)
$$(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{w}})(\lambda) = \arg\min_{\boldsymbol{\beta}, \mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^{n} w_i^2 |y_i - \mathbf{x}_i' \boldsymbol{\beta}| + \lambda \sum_{i=1}^{n} \varpi_i |1 - w_i| \right\},$$

where $0 < w_i^2 \leq 1$ represent the weights quantifying the outlying effects for each observation, and $\sum_{i=1}^{n} \lambda \varpi_i |1 - w_i|$ is a penalty shrinking all weight to the direction of 1. Here $\varpi_i$s include some prior information on the outlying status of all observations, and $\lambda$ is a tuning parameter in $(0, \infty)$. Ideally, this penalty term is expected to generate small weights for those leverage points (outliers in the $x$ direction) or $y$-outliers (outliers in the $y$ direction) and large weight for normal observations.

*Remark 1*: The non-differentiability of penalty $\varpi_i |1 - w_i|$ at $w_i = 1$ implies that some of the components of $\widehat{\mathbf{w}}$ may be exactly equal to one, the corresponding observations of which are called "normal" observations. The other observations with $\widehat{w}_i < 1$ are "abnormal", with possible outlying in the $x$ and/or $y$ direction. In this regards, the estimated weights provide an automatic way to conduct outlier detection. When $\lambda$ is sufficiently large, all observations have $\widehat{w}_i = 1$, and no outlier is claimed. When $\lambda$ is sufficiently small, all $\widehat{w}_i$ are close to zero. Therefore, the tuning parameter selection plays an important role in determining the proportion of outlying observations. In Section 2.3, we will

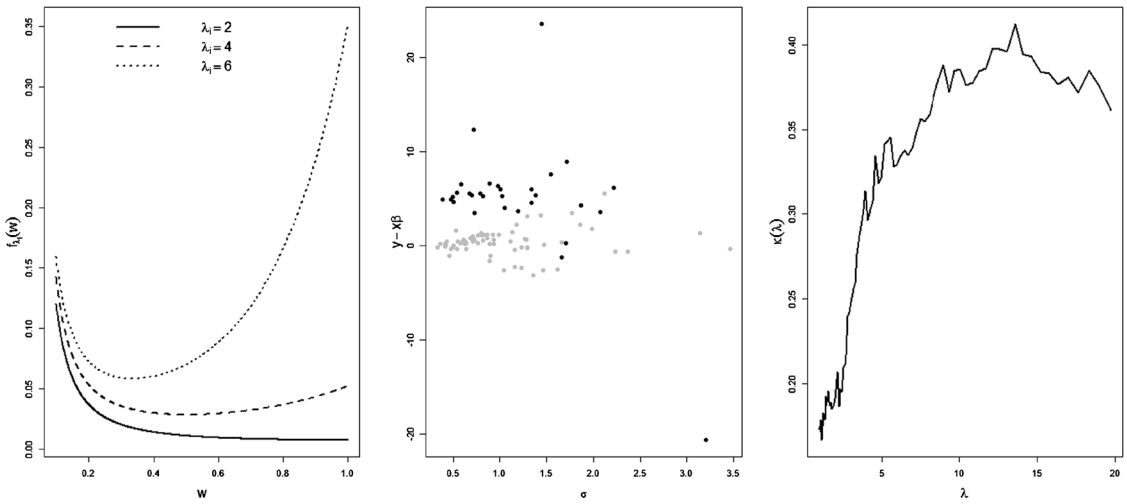


*Figure 1. Left panel: Curves of prior distribution of weight, $f_{\lambda_i}(w_i)$ for $\lambda_i = 2$, 4, and 6; Middle panel: a simulated data with $t(2)$ random error and 30% data contamination under Simulation Example 2 in Section 4.1 (normal observations and outliers are plotted using gray and black dots); Right panel: the $\kappa(\lambda)$ curve (for tuning parameter selection introduced in Section 2.3) produced for the sample data in the middle panel.*

introduce a random weighting technique to select the optimal $\lambda$.

*Remark 2*: If some outlying information is incorporated into $\varpi_i$, the outlier detection accuracy can be significantly improved. For example, suppose an initial value on weight $w_i^{(0)}$ is obtained, we can set $\varpi_i = 1/|\log(w_i^{(0)})|$. A larger $0 < w_i^{(0)} \leq 1$ produces a larger penalty $\lambda\varpi_i$ on $|1 - w_i|$, which pushes $\widehat{w}_i$ more to 1. If $w_i^{(0)} = 1$, then we have $\varpi_i = \infty$, which would leads to $\hat{w}_i = 1$. On the other hand, when $w_i^{(0)} \to 0$, $\hat{w}_i$ is usually much smaller than 1 since $\varpi_i \to 0$ leads to very small penalty being imposed for the $i$-th observation. This resembles the idea of adaptive lasso [38].

We would like to point out a Bayesian interpretation of PWLAD estimator in (3). Suppose $y_i$ follows a Laplace distribution (LD) with mean $\mathbf{x}_i'\boldsymbol{\beta}$ and scale parameter $2/w_i^2$ with $0 < w_i \leq 1$,

$$f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, w_i) = (w_i^2/4)\exp\left\{-(w_i^2/2)|y_i - \mathbf{x}_i'\boldsymbol{\beta}|\right\}.$$

If we assume $w_i$ follows a prior distribution with hyperparameter $\lambda_i \geq 1$,

$$f_{\lambda_i}(w_i) \propto w_i^{-2}e^{-\lambda_i|1-w_i|}, \quad 0 < w_i \leq 1, \tag{4}$$

and $\beta_j \propto 1$, then a posterior distribution of those parameters is

$$f(\boldsymbol{\beta}, \mathbf{w}|y_i, \mathbf{x}_i) \propto \exp[-\sum_{i=1}^{n}(w_i^2/2)|y_i - \mathbf{x}_i'\boldsymbol{\beta}| - \sum_{i=1}^{n}\lambda_i|1 - w_i|].$$

Thus for each $\lambda_i = \lambda\varpi_i$s, PWLAD in (3) is a posterior mode of $\mathbf{w} = (w_1, \cdots, w_n)'$.

It is easy to check $f_{\lambda_i}(w_i)$ in (4) decreases when $w_i < 2/\lambda_i$ and increases when $w_i > 2/\lambda_i$. Three shapes of $f_{\lambda_i}(w_i)$ regarding under $\lambda_i = 2$, 3, and 5 are plotted in the left panel of Figure 1. Those curves show that the larger $\lambda_i$ is, the higher prior probability for $w_i$ being close to 1.

## 2.1 Model implementation

For any tuning parameter $\lambda$ and $\varpi_i$s, the penalized objective function in (3) is convex in $\mathbf{w}$ when $\boldsymbol{\beta}$ is fixed, it is also convex in $\boldsymbol{\beta}$ if $\mathbf{w}$ is fixed. Therefore, we are facing a bi-convex optimization problem in (3). Thus, once an initial $\boldsymbol{\beta}^{(0)}$ and $\mathbf{w}^{(0)}$ are available, $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{w}})$ can be solved alternatively via the following algorithm.

**Algorithm 1** PWLAD Solution for a given $\lambda$

**Given** initial estimates $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\varpi}$ and $\mathbf{w}^{(0)}$
  Let $j = 1$ and $\lambda_i = \lambda\varpi_i$
**While** not converged **do**
  [Update $\boldsymbol{\beta}$]
    $\mathbf{y}^{\text{adj}} = \mathbf{w}^{(j-1)} \cdot \mathbf{w}^{(j-1)} \cdot \mathbf{y}$,
    $\mathbf{X}^{\text{adj}} = \mathbf{w}^{(j-1)} \cdot \mathbf{w}^{(j-1)} \cdot \mathbf{X}$,
    let $\boldsymbol{\beta}^{(j)} = \arg\min_{\boldsymbol{\beta}}\{\|\mathbf{y}^{\text{adj}} - \mathbf{X}^{\text{adj}}\boldsymbol{\beta}\|_1\}$
  [Update $\mathbf{w}$]
    $\mathbf{r}^{(j)} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{(j)}$,
    If $|r_i^{(j)}| > \lambda_i$, let $\mathbf{w}_i^{(j)} \leftarrow \lambda_i/|r_i^{(j)}|$, otherwise $\mathbf{w}_i^{(j)} \leftarrow 1$
    converged $\leftarrow \|\mathbf{w}^{(j)} - \mathbf{w}^{(j-1)}\|_\infty < \epsilon$
  $j \leftarrow j + 1$
**end while**
output $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(j)}$ and $\widehat{\mathbf{w}} = \mathbf{w}^{(j)}$.

Here "$\mathbf{a} \cdot \mathbf{B}$" in Algorithm 1 is a special product between vector $\mathbf{a}$ and matrix $\mathbf{B}$. In particular, if $\mathbf{a} = (a_1, \cdots, a_n)$ is

a vector and $\mathbf{B}$ is a $n \times p$ matrix with $\mathbf{b}_i$ being its $i$th row, then "$\mathbf{a} \cdot \mathbf{B}$" is obtained by multiplying each element of $\mathbf{b}_i$ by $a_i$ for $1 \le i \le n$.

## 2.2 Choice of initial weight

If all $\varpi_i = 1$ in (3), then the penalty in PWLAD becomes a lasso penalty on $1 - \mathbf{w}$. It is well known that the lasso penalty needs stringent conditions to achieve variable selection consistency, which would translate to the difficulty in detecting the outliers consistently. To improve the robustness of PWLAD regarding the leverage points, we suggest to choose $\varpi_i = 1/|\log(w_i^{(0)})|$ in (3), where $w_i^{(0)}$ is an initial estimate of $w_i$ for $1 \le i \le n$ associated with high-breakdown measures of leverage. In particular, we compute the leverage values from a clean data set [3] and the corresponding squared Mahalanobis distance [5] as suggested in [14]. Before the leverage value computation, each predictor is required to be scaled to be between 0 and 1 by subtracting its minimum value and then dividing by its maximum value. After denoting the scaled design matrix as $\breve{\mathbf{X}} = (\breve{\mathbf{x}}_1, \cdots, \breve{\mathbf{x}}_n)'$, and letting $\breve{\mathbf{d}}$ be the vector consists of median value of each of $p$ columns in $\breve{\mathbf{X}}$, the clean subset $S$ consists of all $m$ observations with the smallest distances between $\breve{\mathbf{x}}_i$ and $\breve{\mathbf{d}}$. Thus, the leverage values for observation $i$ relative to the clean subset $S$ is computed from $h_i = \mathbf{x}_i'(\mathbf{X}_S'\mathbf{X}_S)^{-1}\mathbf{x}_i$.

Some other leverage value computation approaches are also provided in [27] and [26]. Once all leverage values are obtained, we quantify the severity of the data contamination (among $\mathbf{x}_i$'s) by a ratio, $L = \dfrac{\max_{1\le i\le n} h_i}{\min_{1\le i\le n} h_i} \ge 1$. If the data produce a large $L$, there is a good chance that the data is contaminated by leverage points. In this case, a pre-screening step based upon $h_i$'s tends to improve the robustness. If $L$ is close to 1, the chance of having the leverage points is small, then the weighting procedure in WLAD regression may cause unnecessary bias during the estimation. Therefore, we only use the high-breakdown leverage value information selectively based upon a cutoff value $L_0 > 1$. If $L > L_0$, we let $w_i^{(0)} = w_0 \ll 1$ for all those $(n-m)$ observations with smallest $h_i$s and $w_i^{(0)} = 1$ for the rest observations. If $L \le L_0$, we compute $w_i^{(0)}$s from a non-adaptive PWLAD obtained from Algorithm 1 in Section 2.1 under $\varpi_i = 1$ and an initial $\boldsymbol{\beta}^{(0)}$ from LAD regression.

*Remark 3*: [14] provided detailed discussions on the choice of clean dataset size $m$ and suggested to use $m = 0.6n$. The clean subset should be large enough to include much of the data, but not too large so that it does not include outlying observations. Our results are not sensitive to the choice of clean dataset size since the procedure is designed to be data-adaptive. The leverage point information is only used if the leverage severity ratio is large enough ($L > L_0$).

*Remark 4*: The cutoff value $L_0 > 1$ is user-defined. It is used to decide whether the above leverage screening approach is worthwhile to be used in the PWLAD approach. A smaller $L_0$ gives more chance on using the measured leverage information. A larger $L_0$ is more conservative. In general, $L_0$ is chosen based upon what ratio is big enough to quantify some observations among all $x_i$'s to be significantly different from others. We suggest to use $\log(n)$ based upon the universal threshold value idea in the lasso variable selection [9, 37]. The PWLAD estimator is robust to the choice of $w_0$ during the pre-screening approach as long as $w_0 \ll 1$. In our numerical results, we set $w_0 = 0.01$.

## 2.3 Tuning parameter selection

The selection of tuning parameter $\lambda$ plays an important role in the performance of both outlier identification and parameter estimation. We propose to use the stability selection method [20, 31] to select an "optimal" $\widehat{\lambda}$ from a fine grid of $\lambda$. This method is referred as the random weighting method in [11]. One important by-product of random weighting method is that it can produce an estimate of the outlying probability of all observations. Here we only introduce random weighting steps briefly with detailed discussions available in [11].

Following [10], let $\omega_1, \cdots, \omega_n$ be some i.i.d. random weights with $E(\omega_i) = Var(\omega_i) = 1$, and $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_n)'$. With these random weights, we obtain the corresponding perturbed estimates,

$$(\widehat{\boldsymbol{\beta}}(\lambda;\boldsymbol{\omega}), \widehat{\mathbf{w}}(\lambda;\boldsymbol{\omega})) = \arg\min_{\boldsymbol{\beta},\mathbf{w}} \left\{ \frac{1}{2}\sum_{i=1}^{n} \omega_i w_i^2 |y_i - \mathbf{x}_i'\boldsymbol{\beta}| + \sum_{i=1}^{n} \lambda \varpi_i |1 - w_i| \right\}. \tag{5}$$

Via (5), any two sets of random weights, $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$, give two perturbed weight estimates $\widehat{\mathbf{w}}(\lambda;\boldsymbol{\omega}_1)$ and $\widehat{\mathbf{w}}(\lambda;\boldsymbol{\omega}_2)$, which lead to two sets of suspected outliers, $\mathcal{O}(\lambda;\boldsymbol{\omega}_1)$ and $\mathcal{O}(\lambda;\boldsymbol{\omega}_2)$. The agreement of these two sets of suspected outliers can be measured by Cohen's kappa coefficient [6], $\kappa(\lambda) \equiv \kappa(\mathcal{O}(\lambda;\boldsymbol{\omega}_1), \mathcal{O}(\lambda;\boldsymbol{\omega}_2))$. A sample $\kappa(\lambda)$ curve (produced for a sample data in Example 2 in Section 4.1) along a sequence of tuning parameter $\lambda$ is plotted in the right panel of Figure 1.

Finally, if we repeatedly generate $B$ pairs of random weights, $\boldsymbol{\omega}_{b1}$ and $\boldsymbol{\omega}_{b2}$, $b = 1, \cdots, B$, we can estimate the stability of the outlier detection by

$$\widehat{S}(\lambda) = \frac{1}{B}\sum_{b=1}^{B} \kappa\left(\mathcal{O}(\lambda;\boldsymbol{\omega}_{b1}), \mathcal{O}(\lambda;\boldsymbol{\omega}_{b2})\right), \tag{6}$$

and then select $\widehat{\lambda}$ that maximizes $\widehat{S}(\lambda)$. In addition, for each observation, the proposed method can provide an estimate for the probability of it being an outlier as $\lambda$ changes,

$$\widehat{P}_i^o(\lambda) = \frac{1}{2B}\sum_{b=1}^{B}\sum_{k=1}^{2} I\left\{i \in \mathcal{O}(\lambda;\boldsymbol{\omega}_{bk})\right\}. \tag{7}$$

## 3. OUTLIER DETECTION CONSISTENCY

In this section, we investigate the outlier detection properties of the PWLAD estimator. In particular, we want to know whether the proposed PWLAD is able to detect all the true outliers with high probability. First, we introduce several notations.

Recall that $\mathcal{O} = \{1 \le i \le n : 0 < w_i^* < 1\}$ is the true outlier set with cardinality $q_n$, and $\widehat{\mathcal{O}} = \{1 \le i \le n : 0 < \widehat{w}_i < 1\}$ be the estimated outlier set. Denote $\bar{a}_n = \max_{i \in \mathcal{O}} w_i^*$. Let $b_n = \max_{1 \le i \le n, 1 \le j \le p} |x_{ij}|$. Intuitively, as the $q_n$ gets larger, the problem is more difficult. In addition, $w_i^*$ in the outlier set should be considerably smaller than 1. We list the precise conditions as follows.

(A1) The random error $\varepsilon_i / w_i^{*2}$ are i.i.d. with cumulative distribution function $F$, where $F$ is twice differentiable and $f(0) = F'(0) > 0$, $0 < w_i^* \le 1$ for $1 \le i \le n$.
(B1) There exists $\underline{\varpi}_n > 0$ and $\bar{\varpi}_n > 0$ satisfying

$$P\left(\{\max_{i \in \mathcal{O}} \varpi_i \le \underline{\varpi}_n\} \cap \{\min_{i \in \mathcal{O}^c} \varpi_i \ge \bar{\varpi}_n\}\right) = 1 - o(1).$$

(B2) (i) $\underline{\varpi}_n = o\left((q_n \lambda \bar{a}_n^2)^{-1}\right)$ and (ii) $1/\bar{\varpi}_n = o\left(\lambda / \sqrt{\log(n)}\right)$.

The error distribution of $\varepsilon_i$ in Condition (A1) is weaker than that assumed under regular LAD regression since some $w_i^*$ are allowed to be very small. Conditions (B1) and (B2) imply that $\varpi_i$s in the true outlier set should be small enough and $\varpi_i$s in normal data set should be large enough. This is reasonable since the penalty in PWLAD is to shrink all $w_i$s from 0 to 1: larger penalties ($\lambda \varpi_i$) on $|1 - w_i|$ for $i \in \mathcal{O}^c$ encourages $\widehat{w}_i$ to be 1, and smaller penalties on $|1 - w_i|$ for $i \in \mathcal{O}$ encourages $\widehat{w}_i$ to be close to 0. (B2) also indicates that $\underline{\varpi}_n / \bar{\varpi}_n = o\left((q_n \bar{a}_n^2 \sqrt{\log(n)})^{-1}\right)$. Thus if $\bar{a}_n^4 = c / \log(n)$ for some constant $c > 0$, then $\underline{\varpi}_n / \bar{\varpi}_n = o(q_n^{-1})$. It provides a rate requirement on $\underline{\varpi}_n / \bar{\varpi}_n$: the faster $q_n$ grows with $n$, the faster $\underline{\varpi}_n / \bar{\varpi}_n \to 0$.

**Theorem 1.** *Suppose Conditions (A1) and (B1-B2) hold. Let* $\widehat{\mathbf{w}}$ *be a PWLAD solution in* (3) *under a given initial estimator* $\widetilde{\boldsymbol{\beta}}$ *satisfying* $P\left(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \lambda / (\sqrt{p} b_n \max\{\bar{\varpi}_n / 2, \underline{\varpi}_n\})\right) = o(1)$. *Then*

$$\lim_{n \to \infty} P(\widehat{\mathcal{O}} = \mathcal{O}) = 1.$$

The proof of Theorem 1 is provided in Section 6. Theorem 1 indicates that the PWLAD estimator with appropriately chosen $\varpi_i$s can detect all outliers with probability going to 1 asymptotically if we start from a well behaved initial estimator, $\widetilde{\boldsymbol{\beta}}$. For example, such a consistent initial estimator can be obtained from the trimmed LAD estimator under regular conditions [21]. In the numerical studies, we choose initial weight $w_i^{(0)}$s as introduced in Section 2.2 and let $\varpi_i = 1/|\log(w_i^{(0)})|$. We obtain $\widetilde{\boldsymbol{\beta}}$ from the WLAD using the above initial weight vector.

*Remark 5:* If $w_i^*$ in the true outlier set $\mathcal{O}$ is small enough such that $\bar{a}_n^2 = O(n^{-\alpha})$ for some $\alpha > 1$, then $\mathcal{O}$ can be identified consistently as long as the initial estimator $|\widetilde{\beta}_j|$ is bounded for $1 \le j \le p$, even when the number of outliers is proportional to the sample size, i.e. $q_n = O(n)$ and the leverage points exist. In particular, if $b_n = o(\log(n))$ and $0 < \underline{\varpi}_n < 1/4 < 1/2 < \bar{\varpi}_n \le 1$, we can choose $c_1 \log(n) < \lambda < c_2 \log(n)$ for some constant $\sqrt{p}/2 < c_1 < c_2$, then both conditions in (B2) are satisfied and the rate conditions on $\widetilde{\boldsymbol{\beta}}$ in Theorem 1 becomes

$$P\left(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \lambda / (\sqrt{p} b_n \max\{\bar{\varpi}_n / 2, \underline{\varpi}_n\})\right) < P\left(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \log(n)/b_n\right) = o(1).$$

## 4. NUMERICAL STUDIES

### 4.1 Simulation experiments

In this section, we conduct extensive simulation studies under different settings to demonstrate the performance of PWLAD in terms of both outlier detection and regression coefficients estimation.

**Example 1.** *[Scale-shifted model] The data is generated from a homogenous linear model* (1) *with* $\boldsymbol{\beta} = \mathbf{0}_5$ *and* $n = 100$. *The covariance matrix* $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)' = \mathbf{U}\boldsymbol{\Sigma}_1^{1/2}$, *where* $\mathbf{U} = (u_{jk})_{n \times p}$ *with* $u_{jk} \overset{i.i.d.}{\sim} \text{Unif}(-5, 5)$ *and* $\boldsymbol{\Sigma}_1$ *has diagonal elements of* 1 *and all non-diagonal elements of* 0.5. *The first* 30% *observations are set as the outlier set* $\mathcal{O}$ *by letting* $w_i^* = 0.05$ *for* $i \in \mathcal{O}$ *and* 1 *otherwise. We generate i.i.d. random error from either* $\varepsilon_i \sim N(0, 1)$ *or* $t(2)$.

We compare the performance of PWLAD with those of three other robust methods: LAD, WLAD and PWLS in terms of both outlier detection and regression coefficients estimation. All simulations are repeated 100 times. To ensure a fair comparison, the initial weighting with 25% prescreening step introduced in Section 2.2 and the random weighting method introduced in Section 2.3 for tuning parameter selection are adopted in both PWLAD and PWLS.

To evaluate the estimation performance of $\widehat{\boldsymbol{\beta}}$, we report the $\ell_1$ estimation error, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1$, from 100 iterations in Figure 1. The left panel in Figure 1 contains the box plots of the $\ell_1$ estimation error for LAD (black), PWLAD (red), PWLS (green) and WLAD (blue). From the graph, we observe that the performance of the four methods are comparable when the error distribution is $N(0, 1)$. When we are in the heavy tail error case ($t(2)$), it appears that PWLAD outperforms the other methods by having the smallest median and the smallest maximum value.

To evaluate the outlier detection performance, we plot the ROC curves in the middle and right panels of Figure 2. Note

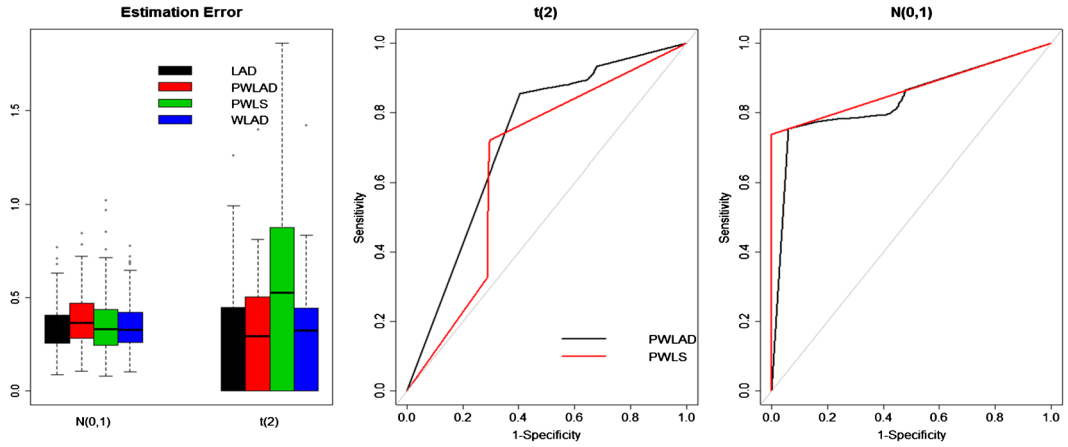


*Figure 2. Estimation and outlier detection results of Example 1. Left panel: the $\ell_1$ estimation error from all 4 methods (PWLAD: red, PWLS: green, WLAD: blue, and LAD: black; left–$N(0,1)$ and right–$t(2)$). Middle panel: ROC curves (PWLAD:Black, PWLS: Red) when random errors follow $t(2)$. Right panel: ROC curves when random errors follow $N(0,1)$.*

that since LAD and WLAD are not designed to detect outliers, we only plot ROC curves generated from PWLAD and PWLS. It is observed from Figure 2 that PWLAD outperforms PWLS regarding the outlier detection accuracy when the random error is heavy tailed. Under the normal random error case, the outlier detection performance of both methods greatly improves and PWLS works slightly better than PWLAD.

**Example 2.** *[Mean-shifted model] We generate the data from a heterogenous mean linear model*

$$\mathbf{y}_i = \gamma_i + \mathbf{x}_i'\boldsymbol{\beta} + \sigma_i\eta_i, \quad 1 \le i \le 100, \tag{8}$$

*where $\boldsymbol{\beta} = \mathbf{0}_5$, $\sigma_i = \exp\{0.055(x_{i1} + x_{i2})\}$, $\eta_i$ are i.i.d. from one of the three types of distributions: $N(0,1)$, $t(2)$ and standard double exponential distribution, $DE(0,1)$. The first $\lfloor 100r \rfloor\%$ ($r = 0.1$, $0.2$ or $0.3$) observations are set as the potential outlier set $\mathcal{O}$ by letting $\gamma_i^* = 5$ for $i \in \mathcal{O}$ and $0$ otherwise. The other settings are the same as Example 1, except that those corresponding outliers are also leverage points by letting $x_{i4} = x_{i5} = 20$ for $i \in \mathcal{O}$.*

In Example 2, we generate outliers using a mean shift model. Other more complicated outlying generation schemes including heteroscedasticity, the coexistence of outliers and leverage points are also considered.

In Figure 3, we compare the ROC curves regarding the outlier detection for PWLAD and PWLS under three different error distributions ($t(2)$, DE(0,1) and N(0,1)) as well as three different outlying proportions (10%, 20% and 30%).

From Figure 3, it is clear that PWLAD performs considerably better than the PWLS among all settings including different types of random errors and various outlier percentages. In the cases where the outlier proportion is 10% and 20%, PWLAD can detect 80% to 90% outliers correctly without any false positives. When the outlier percentage becomes 30%, the outlier detection performance of PWLAD become worse, but can still detect 40% to 60% outliers correctly without any false positives. However, the PWLS loses its outlier detection ability almost completely across all settings, with the possible reason being that the data is generated with heterogeneous random error.

*Table 1. Example 2 – Outlier detection evaluation (M: the mean masking probability; S: the mean swamping probability; JD: the joint outlier detection rate)*

| | | PWLAD(%) | | | PWLS (%) | | |
|---|---|---|---|---|---|---|---|
| | $r$ | JD | M | S | JD | M | S |
| $t(2)$ | 0.1 | 50 | 21 | 9 | 0 | 91 | 11 |
| | 0.2 | 54 | 26 | 4 | 0 | 86 | 16 |
| | 0.3 | 0 | 61 | 0 | 0 | 90 | 9 |
| DB$(0,1)$ | 0.1 | 67 | 8 | 8 | 0 | 84 | 19 |
| | 0.2 | 70 | 13 | 4 | 0 | 83 | 20 |
| | 0.3 | 0 | 64 | 0 | 0 | 84 | 18 |
| $N(0,1)$ | 0.1 | 81 | 7 | 8 | 0 | 78 | 24 |
| | 0.2 | 85 | 6 | 5 | 0 | 78 | 24 |
| | 0.3 | 0 | 44 | 0 | 0 | 81 | 23 |

To further investigate the outlier detection performance of PWLAD, we also compute the mean masking probability (M: fraction of undetected true outliers), the mean swamping probability (S: fraction of non-outliers labeled as outliers), and the joint outlier detection rate (JD: fraction of repetitions with 0 masking) out of all repetitions. The higher JD is, the better; the smaller M and S are, the better. The results are reported in Table 1. It is observed that if the outlier percentage is below 30%, the entire outlier set can

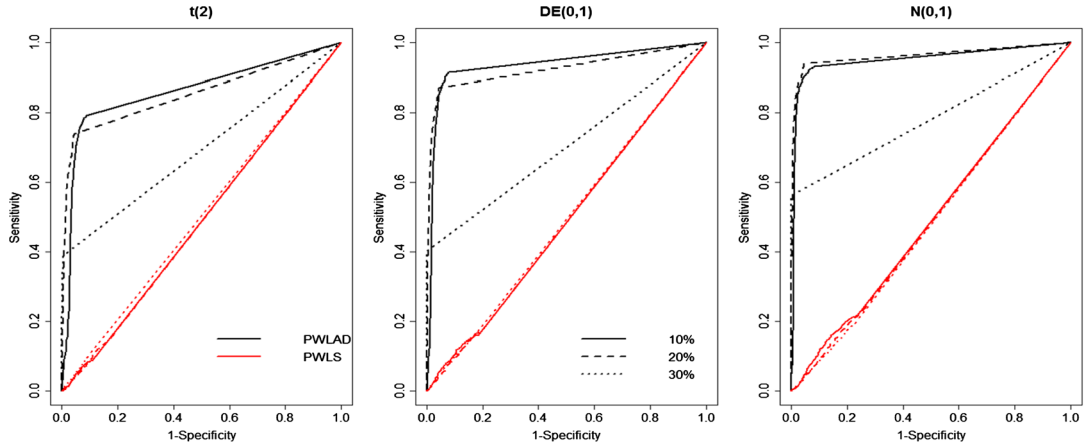


*Figure 3. ROC comparisons of PWLAD (Black) and PWLS (Red) for Example 2. Left panel: $t(2)$; Middle panel: DE(0,1); Right panel: N(0,1).*

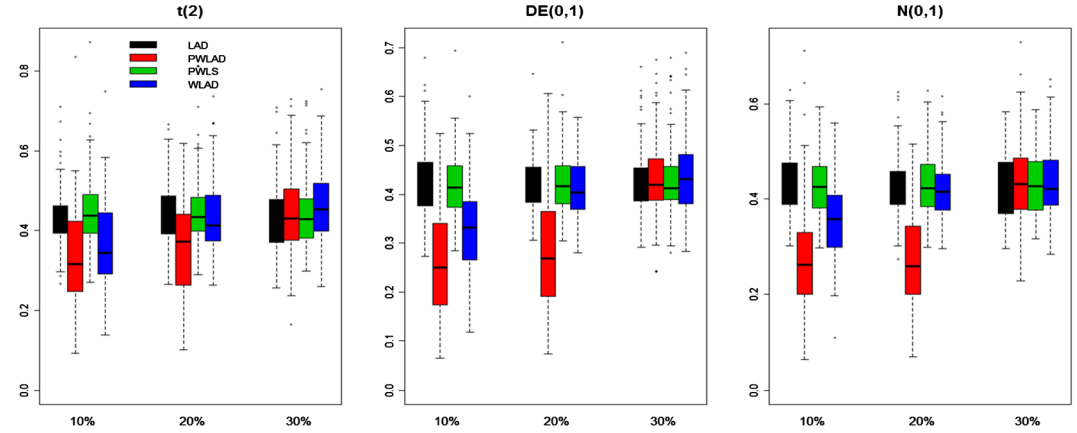


*Figure 4. Regression coefficients estimation comparison for Example 2. The $\ell_1$ estimation error from all 4 methods (PWLAD: red, PWLS: green, WLAD: blue, and LAD: black) are compared under three different outlier proportions (10%, 20%, and 30%). Left panel: $t(2)$. Middle panel: DE$(0,1)$. Right panel: $N(0,1)$.*

be detected by PWLAD as high as 85% of simulation iterations.

The $\ell_1$ estimation error output from all four methods for Example 2 are reported in Figure 4. It is observed that when the outlier percentage is below 30%, PWLAD performs much better than all three other methods: LAD, PWLS, and WLAD. WLAD performs the second best if the outlier percentage is as low as 10%. When the outlier percentage reaches 30%, all four methods produce comparable MAEs.

In summary, PWLAD appears to be the most robust method across all settings: heterogeneous random errors, data contamination at both $x$ and $y$ directions, and heavy tailed distributions.

## 4.2 Real data analysis

Two data sets will be investigated in this section to demonstrate the performance of the PWLAD approach.

The first data set is the Hertzsprung–Russell stars data [28] studied in both [14] and [21]. In this dataset, the logarithm of the light intensity and effective surface temperature were measured for 47 stars. One is interested in fitting the logarithm of light intensity using the logarithm of the surface temperature linearly.

The data is plotted in the left panel of Figure 5. Using PWLAD, 5 observations (7, 11, 20, 30 and 34) are claimed as outliers or leverage points. Among them, weights of observations 11, 20, 30 and 34 (highlighted with black "*") are estimated with 0.006, 0.006, 0.005 and 0.005, while weight of observation 7 (highlighted with black "▲") is estimated to be 0.016. This implies that even though observation 7 is also singled out as a non-normal observation, it plays a more important role during the model fitting process than the other 4 outliers. We also plot four fitted regression lines from LAD (dotted line), WLAD (dash line), PWLAD (solid line), and the re-fitted LAD line (LAD-4, red solid line) after

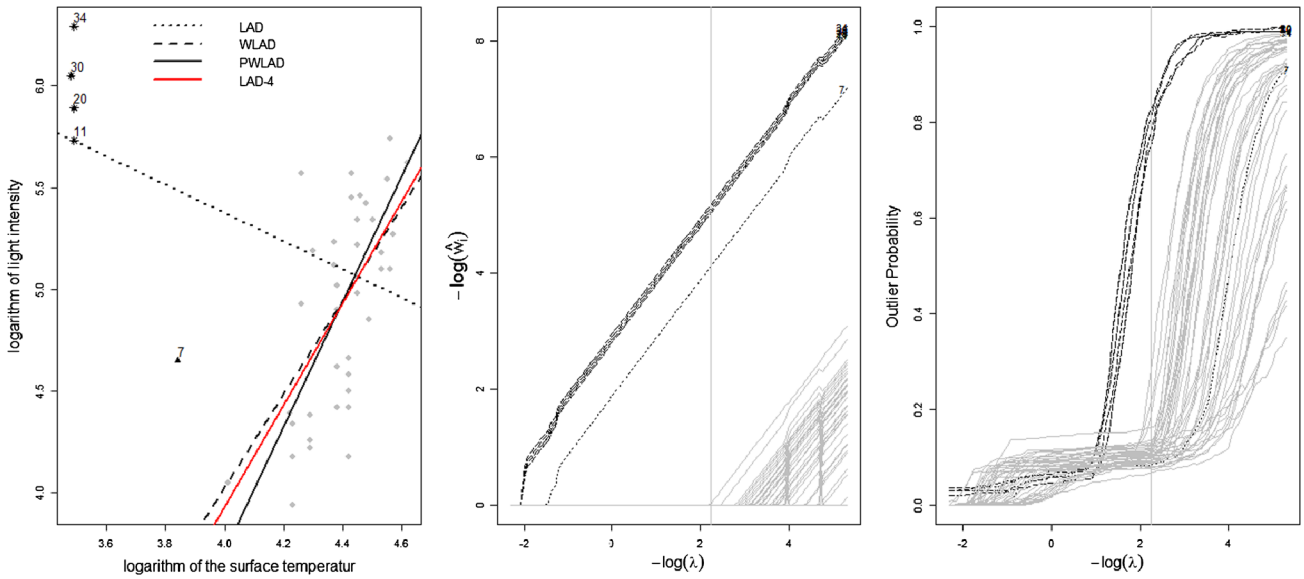


*Figure 5. The Hertzsprung—Russell stars data analysis. The left panel plots the original observations (Outliers: "*" or "▲"; Normal observations: Gray dots) and three fitted regression lines: (LAD: dotted line, WLAD: dash line, PWLAD: solid line. LAD–4 (LAD fit after removing observations 11, 20, 30 and 34): red solid line). The middle panel plots entire solution paths of* $-\log(\widehat{w}_i)$ *versus* $-\log(\lambda)$ *(Outliers: Dark curves, Normal observations: Gray curves) for all* $n$ *observations. The right panel plots all outlying probabilities solution paths versus* $-\log(\lambda)$ *(Outliers: Dark curves, Normal observations: Gray curves). The vertical lines on the last two panels identify the optimal location of the tuning parameter.*

removing 4 observations (11, 20, 30 and 34) in this plot. It is observed that the LAD fitting line is significantly affected by all those 5 observations. The WLAD fits the data much better than the LAD by adjusting those leverage points with different weights. PWLAD calibrates the LAD slightly by data-adaptively adjusting those weights assigned for all observations. The LAD-4 regression line is located between PWLAD and WLAD. This is reasonable since the WLAD only downweights but still uses observations 11, 20, 30 and 34. Compared with the WLAD and LAD-4, PWLAD not only removes (almost) all those four observations, but also downweights an additional observation 17 in the model fitting process.

The solution paths of $-\log(\widehat{w}_i)$ along a sequence of tuning parameter $(-\log(\lambda))$ are reported in the middle panel, where the vertical line identifies the location of the chosen tuning parameter via stability selection. In the right panel, we show the path of outlying probabilities for all observations along a sequence of tuning parameter $(-\log(\lambda))$. Here the outlying probability is computed using the random weighting method introduced in Section 2.3. During the process, the data are reweighted for 100 times. Each time all observations are judged to be an outlier or not. The outlying probability of observation $i$ is the proportion of this observation is identified as an outlier out of 100 reweighted samples. At the chosen tuning parameter, the outlying probabilities of observations 7, 11, 20, 30 and 34 are 11%, 81%, 82%, 85% and 89%, respectively. Thus we found that although observation 7 shows some difference from other normal observations, it is only suggested to be singled out with 11% probability.

The second data set is the modified wood gravity data [25] analyzed using the WLAD method in [14]. The data include 5 covariates and 20 observations, with observations 4, 6, 8 and 19 being modified to be outliers. Those four observations are identified using PWLAD, with weights being estimated as 0.18, 0.15, 0.16 and 0.13, respectively. Compared with the previous star example, none of those outliers has weight close to 0. It means that although those observations are singled out as outliers from others, their information is still used in the regression analysis, but with lower weights. The solution paths of $-\log(\widehat{w}_i)$ are plotted in the middle panel in Figure 6. Apparently, those four observations are separated from others regarding to their weight importance during the regression analysis. It is also interesting to see that the outlying probabilities of those four observations are only around 10%. The outlying probability plot (in the right panel of Figure 6) indicates that simply removing those observations from the regression analysis are not ideal.

In the left panel of Figure 6, we plot the residuals for all 20 observations using LAD (gray circles), WLAD (gray solid dots), and PWLAD (black solid dots for normal observations, black "*" for outliers), respectively. We can see that PWLAD calibrates the WLAD regression by having smaller residuals for normal observations and larger residuals for those outlying observations.

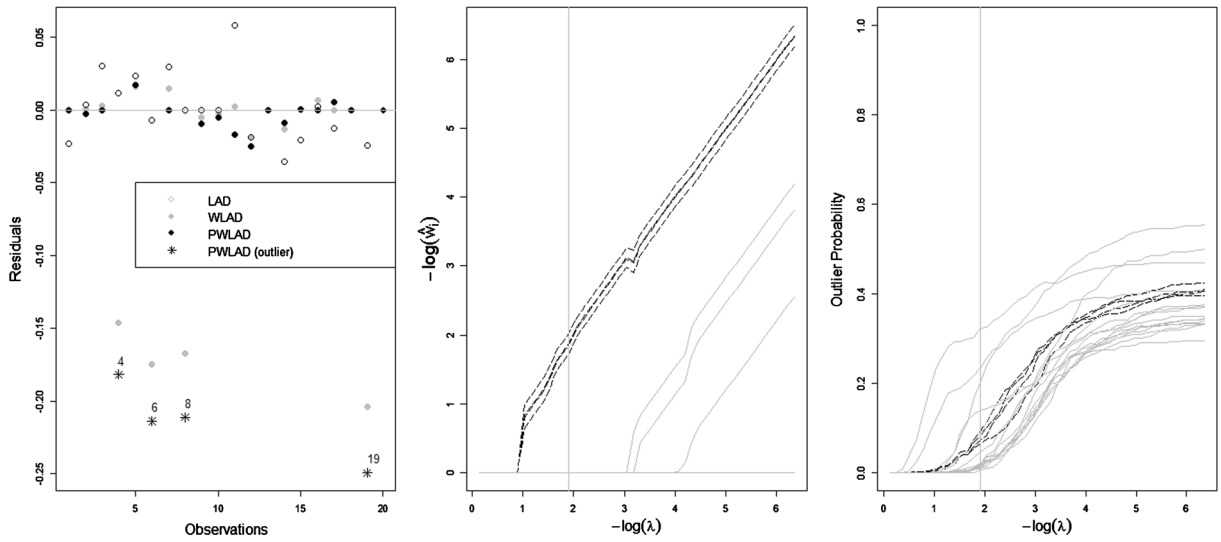


*Figure 6. The modified wood gravity data analysis. The left panel plots the residuals from all three regression analysis: LAD (gray circles), WLAD (gray solid dots) and PWLAD (black solid dots for normal observations, black "*" for those downweighted 4 observations), respectively. The gray horizontal line plots at* 0 *residuals. The middle panel plots entire solution paths of* $-\log(\widehat{w}_i)$ *versus* $-\log(\lambda)$ *(Outliers: Dark curves, Normal observations: Gray curves) for all observations. The right panel plots all outlying probabilities solution paths versus* $-\log(\lambda)$ *(Outliers: Dark curves, Normal observations: Gray curves). The vertical lines on the last two panels identify the optimal location of the tuning parameter.*

## 5. DISCUSSION

In this work, we propose a robust LAD regression method called PWLAD. By assigning each observation an individual weight, $w_i$, and imposing a lasso-type penalty on $1 - w_i$, PWLAD is able to perform simultaneous outlier detection and robust regression, even when the random error is both heterogenous and heavy tailed.

Different from the trimmed LAD regression, PWLAD does not remove detected outliers completely from the LAD regression. All observations contribute in the model fitting process, with observations having large probability being an outlier are used with a considerably smaller weight (close to 0) than the remaining observations. In addition, using PWLAD, there is no need to specify a prior trimmed percentage in the model fitting process, which enhances the robustness of the procedure. PWLAD provides the entire solution path along a sequence of penalty parameters with the final solution chosen using a stability selection approach. As a by-product, an outlying probability is obtained for quantifying each observation's outlying behavior.

Robust regression with variable selection has attracted much attention lately in high-dimensional data analysis. See, for example, the LAD-Lasso in [35] and the least trimmed squares estimator in [1]. An interesting future work is to conduct variable selection and outlier detection simultaneously, e.g., by adding an extra penalty on the regression coefficients, say $\lambda_2 \sum_{j=1}^{p} |\beta_j|$, to the objective function of (3).

Moreover, it is important to point out that quantile regression is also not robust to outliers in $x$ direction. Although the weighted quantile regression (WQR) [17] or trimmed quantile regression (LTQR) were shown [21] to improve the robustness, there are still limitations on the choice of weight parameter and the trimming percentage. Therefore, both LTQR and WQR would lack robustness under large percentage of data contamination. It is also worthwhile to extend the penalized weight idea to the quantile regression framework.

## 6. PROOFS

*Proof of Theorem 1.* The proof of outlier detection consistency is similar to [11]. We only provide the main idea as follows. Under an initial estimator $\widetilde{\boldsymbol{\beta}}$, the objective function of PWLAD on $\mathbf{w}$ becomes

$$L(\mathbf{w}|\widetilde{\boldsymbol{\beta}}, \lambda, \boldsymbol{\varpi}) = \sum_{i=1}^{n} (w_i^2/2)|y_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}| + \lambda\varpi_i|1 - w_i|.$$

Thus the solution of $\widehat{\boldsymbol{w}}$ satisfies,

$$\begin{cases} \widehat{w}_i|y_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}| = \lambda\varpi_i & \text{if } 0 < \widehat{w}_i < 1, \\ \widehat{w}_i|y_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}| < \lambda\varpi_i & \text{if } \widehat{w}_i = 1. \end{cases} \tag{9}$$

Then $\widehat{\mathcal{O}} = \mathcal{O}$ if

$$\begin{cases} |y_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}| > \lambda\varpi_i & \text{if } i \in \mathcal{O}, \\ |y_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}| \le \lambda\varpi_i & \text{if } i \in \mathcal{O}^c. \end{cases} \tag{10}$$

Thus

$$P(\widehat{\mathcal{O}} \neq \mathcal{O}) \leq P\left(\bigcup_{i\in\mathcal{O}}\{|y_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}| < \lambda\varpi_i\}\right) + P\left(\bigcup_{i\in\mathcal{O}^c}\{|y_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}| \geq \lambda\varpi_i\}\right). \tag{11}$$

First,

$$\begin{aligned}
&P\left(\bigcup_{i\in\mathcal{O}}\{|\widetilde{r}_i| < \lambda\varpi_i\}\right)\\
&\leq P\left(\bigcup_{i\in\mathcal{O}}\{|\varepsilon_i/w_i^{*2}| < \lambda\varpi_i + |\mathbf{x}_i'(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|\}\right)\\
&\leq P\left(\bigcup_{i\in\mathcal{O}}\{|\varepsilon_i/w_i^{*2}| < \lambda\underline{\varpi}_n + |\mathbf{x}_i'(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|\}\right)\\
&\quad + P\left(\max_{i\in\mathcal{O}}\varpi_i > \underline{\varpi}_n\right)\\
&\leq q_n P\left(|\varepsilon_i| < 2\lambda\bar{a}_n^2\underline{\varpi}_n\right)\\
&\quad + P\left(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \lambda\underline{\varpi}_n/(b_n\sqrt{p})\right) + o(1)\\
&\leq \frac{2\sqrt{2}\sigma}{\sqrt{\pi}}q_n\bar{a}_n^2\lambda\underline{\varpi}_n\\
&\quad + P\left(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \lambda\underline{\varpi}_n/(b_n\sqrt{p})\right) + o(1)\\
&= P\left(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \lambda\underline{\varpi}_n/(b_n\sqrt{p})\right) + o(1),
\end{aligned}$$

where the third "$\leq$" is from Condition (B1), and the last "$\leq$" is from Conditions (A1) and (B2-i). Second,

$$\begin{aligned}
&P\left(\bigcup_{i\in\mathcal{O}^c}\{|\widetilde{r}_i| \geq \lambda\varpi_i\}\right)\\
&\leq P\left(\bigcup_{i\in\mathcal{O}^c}\{|\varepsilon_i/w_i^{*2}| \geq \lambda\varpi_i - |\mathbf{x}_i'(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)|\}\right)\\
&\leq P\left(\bigcup_{i\in\mathcal{O}^c}\{|\varepsilon_i| \geq \lambda\bar{\varpi}_n - b_n\sqrt{p}\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2\}\right)\\
&\quad + P\left(\min_{i\in\mathcal{O}^c}\varpi_i < \bar{\varpi}_n\right)\\
&\leq P\left(\max_{i\in\mathcal{O}^c}|\varepsilon_i| \geq (\lambda/2)\bar{\varpi}_n\right)\\
&\quad + P\left(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \lambda\bar{\varpi}_n/(2\sqrt{p}b_n)\right) + o(1)\\
&\leq \frac{3\sqrt{1+\log(2n)}\sigma}{\lambda\bar{\varpi}_n} + P\left(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \lambda\bar{\varpi}_n/(2\sqrt{p}b_n)\right)\\
&\quad + o(1)\\
&= P\left(\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 > \lambda\bar{\varpi}_n/(2\sqrt{p}b_n)\right) + o(1),
\end{aligned}$$

where the third "$\leq$" is from Condition (B1), and the last "$\leq$" is from Conditions (A1) and (B2-ii).

The theorem is proved by combining the results regarding the two terms on the right hand side of (11). □

*Received 16 March 2016*

## REFERENCES


[1] Alfons, A., Croux, C., and Gelper, S. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Annals of Applied Statistics*, 7(1):226–248, 2013. MR3086417

[2] Atkinson, A. C., Riani, M., and Cerioli, A. *Exploring Multivariate Data with the Forward Search*. Springer, 2003. MR2055967

[3] Billor, N., Hadi, A. S., and Velleman, P. F. BACON:blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, 34:279–298, 2000.

[4] Chaloner, K. and Brant, R. A bayesian approach to outlier detection and residual analysis. *Biometrika*, 75(4):651–659, 1988.

[5] Chatterjee, S. and Hadi, A. S. *Sensitivity Analysis in Linear Regression*. Wiley, NewYork, 1988. MR0939610

[6] Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measuremen*, 20:37–46, 1960.

[7] Croux, C., Rousseeuw, P. J., and Hössjer, O. Generalized s-estimators. *Journal of the American Statistical Association*, 89(428):1271–1281, 1994.

[8] Hoaglin D. C., Iglewicz B., and Tukey, J. W. Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 81(396):991–999, 1986. MR0867622

[9] Donoho, D. L. and Johnstone, I. M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994. MR1311089

[10] Fang, Y. and Zhao, L. Approximation to the distribution of LAD-estimators for censored regression by random weighting method. *Journal of Statistical Planning and Inference*, 136:1302–1316, 2006.

[11] Gao, X. and Fang, Y. Penalized weighted least squares for outlier detection and robust regression. *Journal of Business Statistics and Economics*, 2016. https://arxiv.org/abs/1603.07427.

[12] Gao, X. Penalized weighted low-rank approximation for robust recovery of recurrent copy number variations. *BMC Bioinformatics*, 16:407, 2015.

[13] Gervini, D. and Yohai, V. J. A class of robust and fully efficient regression estimators. *Annals of Statistics*, 30(2):583–616, 2002.

[14] Gilonia, A., Simonoffb, J. S., and Senguptac, B. Robust weighted lad regression. *Computational Statistics & Data Analysis*, 50:3124–3140, 2006.

[15] Hadi, A. S., Imon, A. H. M. R., and Werner, M. Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):57–70, 2009.

[16] Hawkins, D. M. and Olive, D. Applications and algorithms for least trimmed sum of absolute deviations regression. *Computational Statistics & Data Analysis*, 32:119–134, 1999.

[17] Hubert, M. and Rousseeuw, P. J. The catline for deep regression. *Journal of Multivariate Analysis*, 66:270–296, 1998. MR1642481

[18] Ling, S. Self-weighted least absolute deviation estimation for infinite variance autoregressive models. *Journal of the Royal Statistical Society, Series B*, 67:381–393, 1990.

[19] Maronna, R. A., Martin, D. R., and Yohai, V. J. *Robust Statistics: Theory and Methods*. Wiley, 2006.

[20] Meinshausen, N. and Buhlmann, P. Stability selection (with discussion). *Journal of the Royal Statistical Society, Series B*, 72:417–473, 2010.

[21] Neykov, N. M., Čížek, P., Filzmoserc, P., and Neytcheva, P. N. The least trimmed quantile regression. *Computational Statistics & Data Analysis*, 56(6):1757–1770, 2012.

[22] Pan, B., Chen, M., and Wang, Y. Weighted least absolute deviations estimation for periodic arma models. *Acta Mathematica Sinica, English Series*, 31:1273–1288, 2015. MR3367688
[23] Pan, J., Wang, H., and Yao, Q. Weighted least absolute deviations estimation for arma models with infinite variance. *Econometric Theory*, 23:852–879, 2007. MR2395837
[24] Rousseeuw, P. and Yohai, V. *Robust regression by means of S-estimators.*, volume 26 of *In Robust and nonlinear time series analysis, Lecture Notes in Statistics*. Springer, New York, 1984.
[25] Rousseeuw, P. J. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.
[26] Rousseeuw, P. J. Multivariate estimation with high breakdown point. pages 283–297, 1985. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (Eds.), Mathematical Statistics and Applications, vol. B. Dordrecht, Reidel.
[27] Rousseeuw, P. J. and van Zomeren, B. A comparison of some quick algorithms for robust regression. *Computational Statistics & Data Analysis*, 14:107–116, 1992.
[28] Rousseuw, P. J. and Leroy, A. M. *Robust Regression and Outlier Detection*. New York: Wiley, 1987. MR0914792
[29] Siegel, A. F. Robust regression using repeated medians. *Biometrika*, 69(1):242–244, 1982.
[30] Sim, C. H., Gan, F. F., and Chang, T. C. Outlier labeling with boxplot procedures. *Journal of the American Statistical Association*, 100(470):642–652, 2005.
[31] Sun, W., Wang, J., and Fang, Y. Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research*, 14:3419–3440, 2013.
[32] Tableman, M. The asymptotics of the least trimmed absolute deviations (ltad) estimator. *Statistics and Probability Letters*, 19:387–398, 1994.
[33] Tableman, M. Tthe influence functions for the least trimmed squares and the least trimmed absolute deviations estimator. *Statistics and Probability Letters*, 19:329–337, 1994. MR1278670
[34] Čížek, P. Least trimmed squares in nonlinear regression under dependence. *Journal of Statistical Planning and Inference*, 136(11):3967–3988, 2006.
[35] Wang, H., Li, G., and Jiang, G. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business and Economic Statistics*, 25:347–355, 2007.
[36] Yohai, V. J. High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, 15(2):642–656, 1987.
[37] Zhang, C. H. and Zhang, S. S. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1):217–242, 2014.
[38] Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

Xiaoli Gao
Department of Mathematics and Statistics
University of North Carolina at Greensboro
USA
E-mail address: x_gao2@uncg.edu

Yang Feng
Department of Statistics
Columbia University
USA
E-mail address: yang.feng@columbia.edu