The Conditional Breakdown Properties of LAD-LASSO Regression

by

Boning (Bernice) Feng

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Business and Economics Honors Program

NYU Shanghai

May 2023

Professor Marti G. Subrahmanyam Professor Christina Wang Professor Wendy Jin	Professor Avi Giloni
Faculty Advisers	Thesis Adviser

Boning (Bernice) Feng, New York University, bf1318@nyu.edu

Avi Giloni, Yeshiva University, agiloni@yu.edu
Jeffrey S. Simonoff, New York University, jss2@stern.nyu.edu

The Conditional Breakdown Properties of LAD-LASSO Regression

Abstract

The breakdown value is a measure of the worst-case robustness properties of an estimator. It represents the smallest number of observations that can be sent to arbitrary values that will result in a parameter estimate becoming infinitely large in absolute value. In regression modeling, the meaningful measure is the conditional breakdown, in which predictor values are taken as given and fixed, and response values are sent to infinity. Least squares (LS)-based methods generally have a breakdown of 1 observation, the smallest possible value, reflecting their lack of robustness.

It is known that regression based on least absolute deviations (LAD) has higher conditional breakdown than do LS-based methods. In this paper, we examine the conditional breakdown properties of the robust regression method LAD-LASSO, a regularization method that effectively performs variable selection by setting specific slopes to zero based on a specified regularization parameter λ while also attempting to be resistant to unusual observations. By formulating the LAD-LASSO problem as a linear program, we are able to use an enumerative algorithm to calculate the conditional breakdown of LAD-LASSO for a given data set. We find that the breakdown depends on several things, including λ and the specific values of the predictors (as would be expected), but also on how and whether variables are centered and scaled. We also find that using LAD-LASSO to choose the predictors with nonzero slopes, and then fitting LAD on those predictors, can improve the breakdown considerably.

0.1 Introduction

It is well-known that many estimators, particularly ones based on least squares, are not robust, being sensitive to the effects of unusual observations (outliers and leverage points). One measure of robustness is the breakdown value, the smallest number of observations that can be sent to arbitrary values that will result in a parameter estimate becoming infinitely large in absolute value (a closely-related measure is the breakdown point, which is the breakdown value divided by the sample size, and thus refers to the smallest proportion of observations that can be sent to arbitrary values that will result in a parameter estimate becoming infinitely large in absolute value). This is a worst-case performance measure, as it is based on observations being set to the most disadvantageous values possible.

Consider a standard linear regression situation, in which the data

constitute a sample of n observations $\{\mathbf{x}^i, y_i\}$, with $\mathbf{x}^i$ being a p-dimensional vector. The model assumes

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i,$$

where $\boldsymbol{\varepsilon}$ is an n-dimensional error vector. The predictors $\mathbf{x}_i$ are taken as given and fixed, so the appropriate breakdown value is the conditional breakdown value, in which only response values are allowed to become infinite. It is well-known that the least squares estimator $\hat{\boldsymbol{\beta}}^{OLS}$, which minimizes

$$\sum_{i=1}^{n}(y_i - [\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}])^2,$$

has a breakdown value of 1 observation, the smallest possible value.

Least absolute deviations (LAD) regression is an alternative regression method that is more resistant to outliers. The LAD estimator $\hat{\boldsymbol{\beta}}^{LAD}$ minimizes

$$\sum_{i=1}^{n}|y_i - [\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}]|. \tag{1}$$

[8] showed that the breakdown value of $\hat{\boldsymbol{\beta}}^{LAD}$ can be determined via mixed-integer programming, and showed that, depending on the configuration of predictor values, it can be considerably larger than 1.

In recent years, the problem of fitting regression models with large numbers of potential predictors has become increasingly important. A popular approach to this problem has been through the use of regularization methods, in which the usual estimation criterion (sum of squared or sum of absolute residuals) is penalized with a term that forces estimated regression slopes to 0. The LASSO estimator, introduced by [17], minimizes

$$\sum_{i=1}^{n}(y_i - [\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}])^2 + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Depending on the choice of λ, the LASSO slope estimates are either forced to equal 0 (effectively acting as a variable selection method) or shrunk towards zero. Unfortunately, since the LASSO criterion is still based on least squares, the estimator is not robust, with a breakdown value of 1.

In order to overcome this weakness, [18] proposed applying regularization to LAD regression, resulting in the LAD-LASSO, which minimizes

$$\sum_{i=1}^{n}|y_i - [\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}]| + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Various authors have examined the asymptotic properties of the LAD-LASSO and generalizations (including allowing for different values of λ for different predictors), including [18], [6], [4], [20], and [19].

Despite its motivation as a method resistant to unusual observations, there has been relatively little study of the robustness properties of LAD-LASSO (notable exceptions include [15], [1], and [3]). In this chapter, we study the breakdown properties of the estimator, demonstrating that it can achieve better (sometimes considerably better) breakdown value than that of LASSO. In Section 0.2 we describe the estimator in more detail. In Section 0.3 we discuss how determining the breakdown value of the method can be formulated as a mixed-integer programming problem. Section 0.4 addresses different implementations of the LAD-LASSO, including versions based on centering and/or scaling the predictors. In Section 0.5 we illustrate the calculations of breakdown values of LAD-LASSO methods on several well-known data sets, showing how the breakdown value depends on the choice of λ and how and whether variables are centered and scaled. We further show how using LAD-LASSO as a variable selection method to choose the predictors with nonzero slopes, and then fitting LAD on those predictors (a method closely related to the relaxed LAD-LASSO proposed in [12]), can improve the breakdown considerably. We conclude the chapter with discussion of potential future work, including the incorporation of weights to improve the breakdown value further.

0.2 LAD regression, LAD-LASSO, and linear programming

The LAD regression problem described in equation (1) can be formulated and solved as a linear program (c.f. [7]). Specifically, let

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & \dots & x_{p1} \\ \vdots & & & & \vdots \\ 1 & x_{1n} & \dots & \dots & x_{pn} \end{pmatrix} = \begin{pmatrix} \mathbf{x}^1 \\ \vdots \\ \mathbf{x}^n \end{pmatrix}.$$

Next, let $r_i = y_i - [\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}]$ and since the sum of the absolute values of each of the residuals is to be minimized, let $r_i^+ - r_i^- = r_i$, where $r_i^+ \geq 0$ and $r_i^- \geq 0$. The LAD regression problem as an optimization problem is to determine the regression coefficients that minimize $\sum |r_i|$, which is equivalent to minimizing $\sum r_i^+ + r_i^-$, since for any i, setting both $r_i^+ > 0$ and $r_i^- > 0$ will necessarily increase the value of the objective function. Thus, the LAD regression problem can

be formulated as

$$\begin{array}{ll} \min & \mathbf{e}_n^T\mathbf{r}^+ + \mathbf{e}_n^T\mathbf{r}^- \\ \text{s.t.} & \mathbf{X}\boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y} \\ & \boldsymbol{\beta} \text{ free},\ \mathbf{r}^+ \geq \mathbf{0},\ \mathbf{r}^- \geq \mathbf{0} \end{array} \tag{2}$$

where the vector of residuals is $\mathbf{r} = \mathbf{r}^+ - \mathbf{r}^-$, and $\mathbf{e}_n^T$ is a vector with all n components equal to 1. The LAD-LASSO problem can be formulated as a nonlinear optimization problem by changing the objective function in the linear program (2) from $\mathbf{e}_n^T\mathbf{r}^+ + \mathbf{e}_n^T\mathbf{r}^-$ to $\mathbf{e}_n^T\mathbf{r}^+ + \mathbf{e}_n^T\mathbf{r}^- + \lambda\mathbf{e}_p^T|\boldsymbol{\beta}|$. However, by setting $\boldsymbol{\beta} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^-$, the objective function of the LAD-LASSO problem can be made into the linear function $\mathbf{e}_n^T\mathbf{r}^+ + \mathbf{e}_n^T\mathbf{r}^- + \lambda\mathbf{e}_p^T\boldsymbol{\beta}^+ + \lambda\mathbf{e}_p^T\boldsymbol{\beta}^-$, since if for some j $\beta_j^+ > 0$ and $\beta_j^- > 0$, the LAD-LASSO objective function would be increased. Thus, the LAD-LASSO problem can be formulated as the following linear program:

$$\begin{array}{ll} \min & \mathbf{e}_n^T\mathbf{r}^+ + \mathbf{e}_n^T\mathbf{r}^- + \lambda\mathbf{e}_p^T\boldsymbol{\beta}^+ + \lambda\mathbf{e}_p^T\boldsymbol{\beta}^- \\ \text{s.t.} & \mathbf{X}\boldsymbol{\beta}^+ - \mathbf{X}\boldsymbol{\beta}^- + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y} \\ & \boldsymbol{\beta}^+ \geq \mathbf{0},\ \boldsymbol{\beta}^- \geq \mathbf{0},\ \mathbf{r}^+ \geq \mathbf{0},\ \mathbf{r}^- \geq \mathbf{0}. \end{array} \tag{3}$$

However, as will be discussed further below, we would like to take a different approach, by formulating the LAD-LASSO problem as an LAD regression problem with an augmented design matrix and response vector. To do so we first reformulate the linear program in equation (3) as

$$\begin{array}{ll} \min & \mathbf{e}_n^T\mathbf{r}^+ + \mathbf{e}_n^T\mathbf{r}^- + \mathbf{e}_p^T\mathbf{c}^+ + \mathbf{e}_p^T\mathbf{c}^- \\ \text{s.t.} & \mathbf{X}\boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y} \\ & -\lambda\boldsymbol{\beta} + \mathbf{c}^+ - \mathbf{c}^- = 0 \\ & \boldsymbol{\beta} \text{ free},\ \mathbf{c}^+ \geq \mathbf{0},\ \mathbf{c}^- \geq \mathbf{0},\ \mathbf{r}^+ \geq \mathbf{0},\ \mathbf{r}^- \geq \mathbf{0}, \end{array} \tag{4}$$

where $\lambda\mathbf{e}_p^T\boldsymbol{\beta}^+ + \lambda\mathbf{e}_p^T\boldsymbol{\beta}^- = \mathbf{e}_p^T\mathbf{c}^+ + \mathbf{e}_p^T\mathbf{c}^-$ since $\lambda\boldsymbol{\beta} = \mathbf{c}^+ - \mathbf{c}^-$. Next, we note that by defining an augmented design matrix $\mathbf{X}^*$ and augmented response vector $\mathbf{y}^*$, the LAD-LASSO linear program in equation (4) can be formulated as an LAD regression problem with design matrix $\mathbf{X}^*$ and response vector $\mathbf{y}^*$. Specifically,

$$\mathbf{y}^* = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ 0 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}, \quad \mathbf{X}^* = \begin{pmatrix} 1 & x_{11} & \dots & \dots & x_{p1} \\ \vdots & & & & \vdots \\ 1 & x_{1n} & \dots & \dots & x_{pn} \\ 0 & -\lambda & 0 & \dots & 0 \\ 0 & 0 & -\lambda & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 0 & -\lambda \end{pmatrix} = \begin{pmatrix} \mathbf{x}^{*1} \\ \vdots \\ \mathbf{x}^{*n+p} \end{pmatrix}, \tag{5}$$

with the LAD-LASSO problem formulated as the LAD regression problem in linear programming form as

$$\begin{array}{ll} \min & \mathbf{e}_n^T\mathbf{r}^+ + \mathbf{e}_n^T\mathbf{r}^- \\ \text{s.t.} & \mathbf{X}^*\boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y}^* \\ & \boldsymbol{\beta} \text{ free, } \mathbf{r}^+ \geq \mathbf{0}, \ \mathbf{r}^- \geq \mathbf{0}. \end{array} \tag{6}$$

The fact that the LAD-LASSO problem can be formulated as a linear program solving an LAD regression problem with an augmented design matrix and augmented response vector is important for two reasons. First, as we demonstrate in the next section, the computational methods of the breakdown value of LAD regression are heavily based upon the LAD regression problem being a linear program. Hence, the LAD-LASSO breakdown value can be computed in a similar manner to that of computing the breakdown value of LAD regression, since it is LAD regression with different data. Second, there is an important property of LAD-LASSO that can be easily seen from its linear programming formulation and solution. Recognizing that the design matrix and y vector for LAD-LASSO in equation (5) are $(n+p)\times(p+1)$ and $(n+p)\times 1$ respectively, Proposition 2 in [7] describes that an optimal solution to the LAD-LASSO regression problem is of the form

$$\hat{\boldsymbol{\beta}}^{LAD} = {\mathbf{X}^*}_B^{-1}{\mathbf{y}^*}_B,$$

where ${\mathbf{X}^*}_B$ is some nonsingular $(p+1)\times(p+1)$ submatrix of $\mathbf{X}^*$ and ${\mathbf{y}^*}_B$ the associated rows of the response vector. In the event that the rows in ${\mathbf{X}^*}_B$ are only among the first n rows, then $\hat{\boldsymbol{\beta}}^{LAD}$ will have $p+1$ nonzero values. On the other hand, any row of ${\mathbf{X}^*}_B$ that comes from the last p rows of $\mathbf{X}^*$ will have associated $\hat{\beta}^{LAD}$ values set to 0.

0.3 Determining the breakdown value of LAD-LASSO

Consider the LAD estimated regression parameters $\hat{\boldsymbol{\beta}}^{LAD}$ based on data $(\mathbf{X}, \mathbf{y})$. If we contaminate m $(1 \leq m < n)$ values of the response vector $\mathbf{y}$ in a way so that row i is replaced by $(\mathbf{x}^i, \tilde{\mathbf{y}}_i)$, we obtain some new data $(\mathbf{X}, \tilde{\mathbf{y}})$. The LAD estimated regression parameters applied to $(\mathbf{X}, \tilde{\mathbf{y}})$ are different from the original ones. We can use any norm $\|\cdot\|$ on $\mathbb{R}^p$ to measure the distance $\|\hat{\boldsymbol{\beta}}^{LAD}(\mathbf{X}, \tilde{\mathbf{y}}) - \hat{\boldsymbol{\beta}}^{LAD}\|$ of the respective estimates. If we vary over all possible choices, then this distance remains either bounded or not bounded. Let

$$b(m, \mathbf{y}|\mathbf{X}) = \sup_{\tilde{y}} \|\hat{\boldsymbol{\beta}}^{LAD}(\mathbf{X}, \tilde{\mathbf{y}}) - \hat{\boldsymbol{\beta}}^{LAD}(\mathbf{X}, \mathbf{y})\|$$

be the maximum bias that results when we replace at most m of the original values of the dependent variable y_i with arbitrary new data.

The conditional breakdown value of LAD regression is

$$a(\mathbf{y}|\mathbf{X}) = \min_{1\leq m<n} \Big\{ m : b(m, \mathbf{y}|\mathbf{X}) \text{ is infinite} \Big\};$$

i.e., it is the minimum number of values of $\mathbf{y}$ that, if replaced with arbitrary new data, make the LAD regression technique break down.

[8] show that the conditional breakdown value of LAD regression can be computed by solving the following mixed-integer program when the design matrix $\mathbf{X}$ is in general position:

$$\begin{aligned}
\min \quad & \sum_{i=1}^{n} u_i + l_i = a(\mathbf{y}|\mathbf{X}) && \text{(7a)}\\
\text{s.t.} \quad & \mathbf{x}^i\boldsymbol{\xi} + \eta^+ - \eta^- + s_i - t_i = 0 \qquad \text{for } i = 1, \ldots, n, && \text{(7b)}\\
& s_i - Mu_i \leq 0, \quad t_i - Ml_i \leq 0 \qquad \text{for } i = 1, \ldots, n, && \text{(7c)}\\
& \eta_i^+ + \eta_i^- + Mu_i + Ml_i \leq M \qquad \text{for } i = 1, \ldots, n, && \text{(7d)}\\
& u_i + l_i \leq 1 \qquad \text{for } i = 1, \ldots, n, && \text{(7e)}\\
& \sum_{i=1}^{n} \eta_i^+ + \eta_i^- - s_i - t_i \leq 0, \quad \sum_{i=1}^{n} s_i + t_i \geq \varepsilon, && \text{(7f)}\\
& \boldsymbol{\xi} \text{ free}, \ \boldsymbol{\eta}^+ \geq \mathbf{0}, \ \boldsymbol{\eta}^- \geq \mathbf{0}, \ \boldsymbol{s} \geq \mathbf{0}, \ \boldsymbol{t} \geq \mathbf{0}, && \text{(7g)}\\
& u_i, l_i \in \{0, 1\} \text{ for } i = 1, \ldots, n && \text{(7h)}
\end{aligned}$$

where we assume that M is a suitably chosen large number and ε a small number so that constraints (7c) and (7d) are nonbinding for the solution that results if we set u_i or l_i equal to 1 or $u_i = l_i = 0$.

[13] and [14] provide an enumerative approach to computing the conditional breakdown of LAD regression. [8] show that this approach provides the same value as the mixed-integer program (when the design matrix is in general position) and [9] describe the enumerative approach through a linear programming framework. Let $N = \{1, \ldots, n\}$ and $E \subseteq N\}$. Then $a(\mathbf{y}|\mathbf{X}) = |E|$ where $|E|$ is the smallest integer such that

$$\max \frac{\sum_{i\in E} |\mathbf{x}^i\boldsymbol{\xi}|}{\sum_{i\in \mathbb{N}} |\mathbf{x}^i\boldsymbol{\xi}|} \geq \frac{1}{2},$$

where

$$\boldsymbol{\xi} = \begin{pmatrix} 1 & \mathbf{0} \\ & \mathbf{X}_B \end{pmatrix}^{-1} \begin{pmatrix} \gamma \\ \mathbf{0} \end{pmatrix},$$

γ is a positive constant and the algorithm enumerates through all $(p) \times (p+1)$ submatrices $\mathbf{X}_B$ of $\mathbf{X}$.

To compute the breakdown value of the LAD-LASSO regression problem described in equation (6), we use the enumerative approach described above with $\mathbf{X}$ substituted with $\mathbf{X}^*$ and $\mathbf{y}$ substituted with $\mathbf{y}^*$. The reason for this is the mixed-integer programming approach is based upon a design matrix that is in general position and the design matrix $\mathbf{X}^*$ is not necessarily in general position, resulting in incorrect breakdown values.

In the next section, we describe the numerical experiments that we conducted in order to explore the breakdown values of different versions of the LAD-LASSO algorithm. In particular, one of the approaches considered is where the original design matrix and response vectors are centered. This has been a popular approach to constructing LASSO-type estimators, because it avoids the need to differentiate in the objective function the intercept coefficient (which is not shrunk to zero) and the slope coefficients (which are). Instead, the intercept term is removed through centering the response and predictor variables (since the OLS fit based on predictors at their sample mean values equals the sample mean of the response), the LASSO is fit based on a model without an intercept, and the intercept is then added back in at the end. In such a case, the LAD-LASSO regression estimates are determined by first solving an LAD regression problem where there is no intercept or constant term. In such a case, the augmented design matrix used will be $p \times (n+p)$ as opposed to $(p+1) \times (n+p)$ and will be of the form (with centering and possibly scaling)

$$\mathbf{y}' = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ 0 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}, \quad \mathbf{X}' = \begin{pmatrix} x_{11} & \dots & \dots & x_{p1} \\ \vdots & & & \vdots \\ x_{1n} & \dots & \dots & x_{pn} \\ -\lambda & 0 & \dots & 0 \\ 0 & -\lambda & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & -\lambda \end{pmatrix} = \begin{pmatrix} \mathbf{x}'^1 \\ \vdots \\ \mathbf{x}'^{n+p} \end{pmatrix}. \tag{8}$$

We note that the conditional breakdown value of the formulation of the LAD-LASSO regression problem is a lower bound on the LAD-LASSO method. This is because the algorithm used to compute the breakdown value permits potential contamination in the response vector described in equation (5) and equation (8). In other words, the breakdown calculation assumes that contamination can take place in the entire response vector, including the 0 values corresponding to the LASSO constraints, where in reality such contamination would not take

place (since those entries do not correspond to real data values). Hence, we also compute the breakdown value of the LAD regression problem on the columns that have nonzero estimated LAD-LASSO parameters. In such a case, if any centering had been performed on the design matrix (as described in the next section), when computing the breakdown value of the LAD regression problem on the columns that have nonzero estimated LAD-LASSO parameters, the original un-centered columns were used including a constant/intercept. Since this breakdown computation is for a traditional LAD regression, either the enumerative approach or the mixed-integer programming approach can be used, and the MIP optimizer Gurobi was used to solve any mixed-integer programs.

0.4 Numerical experiments

Besides the two aforementioned approaches for computing the breakdown value (LAD-LASSO problem vs. remaining columns approach), we now describe the seven approaches considered for computing the breakdown values. The reason for these different approaches is due to the discussions in the literature regarding (i) whether a constant term should be explicitly estimated by the LAD-LASSO or if the design matrix and response vector should be centered, and (ii) whether the data should be scaled before estimating LAD-LASSO parameters, in order to make all of the slopes on the same scale, thereby avoiding the problem of the penalty term in the LASSO objective function potentially being dominated by individual predictor(s) simply because of their scale. In each of the seven approaches below, we compute the breakdown value of both the LAD-LASSO problem and the resulting remaining columns after LAD-LASSO is performed (a method closely related to the relaxed LAD-LASSO proposed in [12]). Specifically, the seven approaches in terms of the design matrix $\mathbf{X}^*$ or $\mathbf{X}'$ are as follows:

1. Estimate the breakdown value of the direct LAD-LASSO regression problem. In other words, compute the breakdown value of the LAD regression with augmented design matrix and response

vector described in equation (5),

$$\mathbf{y}^* = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ 0 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}, \quad \mathbf{X}^* = \begin{pmatrix} 1 & x_{11} & \dots & \dots & x_{p1} \\ \vdots & & & & \vdots \\ 1 & x_{1n} & \dots & \dots & x_{pn} \\ 0 & -\lambda & 0 & \dots & 0 \\ 0 & 0 & -\lambda & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 0 & -\lambda \end{pmatrix}.$$

2. Center the response vector and columns of the design matrix by using their respective sample means. We note that the constant term or intercept will no longer be directly computed by the LAD-LASSO linear program. This results in an augmented design matrix and response vector as described in equation (8),

$$\mathbf{y}^{'} = \begin{pmatrix} y_1^{'} \\ \vdots \\ y_n^{'} \\ 0 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}, \quad \mathbf{X}^{'} = \begin{pmatrix} x_{11}^{'} & \dots & \dots & x_{p1}^{'} \\ \vdots & & & \vdots \\ x_{1n}^{'} & \dots & \dots & x_{pn}^{'} \\ -\lambda & 0 & \dots & 0 \\ 0 & -\lambda & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & -\lambda \end{pmatrix},$$

where

$$y_i^{'} = y_i - \hat{\mu}(\mathbf{y}), \text{ where } \hat{\mu}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$x_{ji}^{'} = x_{ji} - \hat{\mu}(\mathbf{x}_j), \text{ where } \hat{\mu}(\mathbf{x}_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ji}$$

3. Center the response vector and columns of the design matrix by using their respective sample medians. We note that the constant term or intercept will no longer be directly computed by the LAD-LASSO linear program. This results in an augmented

design matrix and response vector as described in equation (8),

$$\mathbf{y}^{'} = \begin{pmatrix} y_1^{'} \\ \vdots \\ y_n^{'} \\ 0 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}, \quad \mathbf{X}^{'} = \begin{pmatrix} x_{11}^{'} & \dots & \dots & x_{p1}^{'} \\ \vdots & & & \vdots \\ x_{1n}^{'} & \dots & \dots & x_{pn}^{'} \\ -\lambda & 0 & \dots & 0 \\ 0 & -\lambda & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & -\lambda \end{pmatrix},$$

where

$$y_i^{'} = y_i - m(\mathbf{y}), \text{ where } m(\mathbf{y}) \text{ is the median of } \mathbf{y}$$

$$x_{ji}^{'} = x_{ji} - m(\mathbf{x}_j), \text{ where } m(\mathbf{x}_j) \text{ is the median of } \mathbf{x}_j.$$

4. Scale the columns of the design matrix by their respective standard deviations and estimate the breakdown value of the direct LAD-LASSO regression problem. In other words, compute the breakdown value of the LAD regression with augmented design matrix and response vector described in equation (5),

$$\mathbf{y}^{*} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ 0 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}, \quad \mathbf{X}^{*} = \begin{pmatrix} 1 & x_{11}^{'} & \dots & \dots & x_{p1}^{'} \\ \vdots & & & & \vdots \\ 1 & x_{1n}^{'} & \dots & \dots & x_{pn}^{'} \\ 0 & -\lambda & 0 & \dots & 0 \\ 0 & 0 & -\lambda & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 0 & -\lambda \end{pmatrix},$$

where the scaling of the predictor variables is computed as

$$x_{ji}^{'} = \frac{x_{ji}}{\hat{\sigma}(\mathbf{x}_j)}, \text{ where } \hat{\sigma}(\mathbf{x}_j) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ji} - \hat{\mu}(\mathbf{x}_j)\right)^2}$$

5. Scale the columns of the design matrix by their respective mean absolute deviation from the sample median and estimate the breakdown value of the direct LAD-LASSO regression problem.

In other words, compute the breakdown value of the LAD regression with augmented design matrix and response vector described in equation (5),

$$\mathbf{y}^* = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ 0 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}, \quad \mathbf{X}^* = \begin{pmatrix} 1 & x_{11}^{'} & \dots & \dots & x_{p1}^{'} \\ \vdots & & & & \vdots \\ 1 & x_{1n}^{'} & \dots & \dots & x_{pn}^{'} \\ 0 & -\lambda & 0 & \dots & 0 \\ 0 & 0 & -\lambda & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 0 & -\lambda \end{pmatrix},$$

where the scaling of the predictor variables is computed as

$$x_{ji}^{'} = \frac{x_{ji}}{\hat{\sigma}(\mathbf{x}_j)}, \text{ where } \hat{\sigma}(\mathbf{x}_j) = \frac{1}{n}\sum_{i=1}^{n} \left| x_{ji} - m(\mathbf{x}_j) \right|$$

6. Center the response vector and columns of the design matrix by using their respective sample means, and scale the columns of the design matrix by their respective standard deviations. We note that the constant term or intercept will no longer be directly computed by the LAD-LASSO linear program. This results in an augmented design matrix and response vector as described in equation (8),

$$\mathbf{y}^{'} = \begin{pmatrix} y_1^{'} \\ \vdots \\ y_n^{'} \\ 0 \\ 0 \\ \vdots \\ \\ 0 \end{pmatrix}, \quad \mathbf{X}^{'} = \begin{pmatrix} x_{11}^{'} & \dots & \dots & x_{p1}^{'} \\ \vdots & & & \vdots \\ x_{1n}^{'} & \dots & \dots & x_{pn}^{'} \\ -\lambda & 0 & \dots & 0 \\ 0 & -\lambda & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & -\lambda \end{pmatrix},$$

where

$$y_i^{'} = y_i - \hat{\mu}(\mathbf{y}), \text{ where } \hat{\mu}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$x_{ji}^{'} = \frac{x_{ji} - \hat{\mu}(\mathbf{x}_j)}{\hat{\sigma}(\mathbf{x}_j)}, \text{ where } \hat{\mu}(\mathbf{x}_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ji}$$

and

$$\hat{\sigma}(\mathbf{x}_j) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ji} - \hat{\mu}(\mathbf{x}_j)\right)^2}.$$

7. Center the response vector and columns of the design matrix by using their respective sample medians, and scale the columns of the design matrix by their respective mean absolute deviations. We note that the constant term or intercept will no longer be directly computed by the LAD-LASSO linear program. This results in an augmented design matrix and response vector as described in equation (8),

$$\mathbf{y}^{'} = \begin{pmatrix} y_1^{'} \\ \vdots \\ y_n^{'} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{X}^{'} = \begin{pmatrix} x_{11}^{'} & \dots & \dots & x_{p1}^{'} \\ \vdots & & & \vdots \\ x_{1n}^{'} & \dots & \dots & x_{pn}^{'} \\ -\lambda & 0 & \dots & 0 \\ 0 & -\lambda & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & -\lambda \end{pmatrix},$$

where

$$y_i^{'} = y_i - m(\mathbf{y}), \text{ where } m(\mathbf{y}) \text{ is the median of } \mathbf{y}$$

$$x_{ji}^{'} = \frac{x_{ji} - m(\mathbf{x}_j)}{\hat{\sigma}(\mathbf{x}_j)}, \text{ where } m(\mathbf{x}_j) \text{ is the median of } \mathbf{x}_j$$

and

$$\hat{\sigma}(\mathbf{x}_j) = \frac{1}{n}\sum_{i=1}^{n}\left|x_{ji} - m(\mathbf{x}_j)\right|.$$

0.5 Examples of breakdown values of LAD-LASSO-based methods

In this section we illustrate the calculation and interpretation of breakdown values for various data sets that have appeared in the robustness literature. The data sets used are the aircraft data, coleman data, Hawkins-Bradu-Kass (HBK) data, salinity data, and the modified wood gravity data. Each of these data sets can be found and is discussed in [16].

For each data set, we produce a three-panel figure of breakdown values (Figures 1 – 5). The top panel plots breakdown values versus the number of active (nonzero slope) predictors over a range of λ values, for the direct LAD-LASSO (L-L), the relaxed LAD-LASSO (RL-L), the versions obtained by estimating the intercept indirectly based on centering variables with sample means (L-L (mean) and RL-L (mean), respectively), and the versions obtained by estimating the intercept indirectly based on centering variables with sample medians (L-L (median) and RL-L (median), respectively). The breakdown value of LAD (which corresponds to that of L-L with $\lambda = 0$) is given by an asterisk. The middle panel presents breakdown values when scaling predictors nonrobustly using sample standard deviations (and centering using means when estimating the intercept indirectly), and the bottom panel presents breakdown values when scaling predictors robustly using sample mean absolute deviations (and centering using medians when estimating the intercept indirectly).

The aircraft data (Figure 1) illustrate the general patterns. From the top panel we see that LAD-LASSO tends to have similar breakdown value to LAD itself when the chosen λ leads to all predictors being active, but as more slopes are driven to zero, its breakdown value stays the same or decreases. Distinctions between direct determination of L-L and indirectly fitting the intercept are minor, and perhaps surprisingly, centering robustly using medians isn't any more effective than doing so using means.

On the other hand, using the LAD-LASSO to determine the number of active predictors, but then using LAD on those predictors to estimate the coefficients (the relaxed LAD-LASSO) can make a difference, and the fewer the number of active predictors, the higher the breakdown value can go. Presumably this is reflecting the fact that if a variable is inactive, changes in its values can no longer affect the estimated coefficients, thereby making it easier to avoid breakdown.

The breakdown properties related to scaling are less clear, but once again we see that relaxed LAD-LASSO outperforms LAD-LASSO. When scaling nonrobustly using standard deviations, the RL-L methods again tend to gain effectiveness as the number of active predictors decreases. Perhaps more surprisingly, the RL-L results when scaling robustly using mean absolute deviations are much more unstable, with breakdown values sometimes higher and sometimes lower than those using nonrobust scaling, and breakdown values counterintuitively sometimes being lower when there are fewer active predictors.

The patterns are similar for the coleman data (Figure 2) and the wood data (Figure 3). The other two data sets exhibit somewhat different behavior. When applied to the HBK data, the relaxed LASSOs

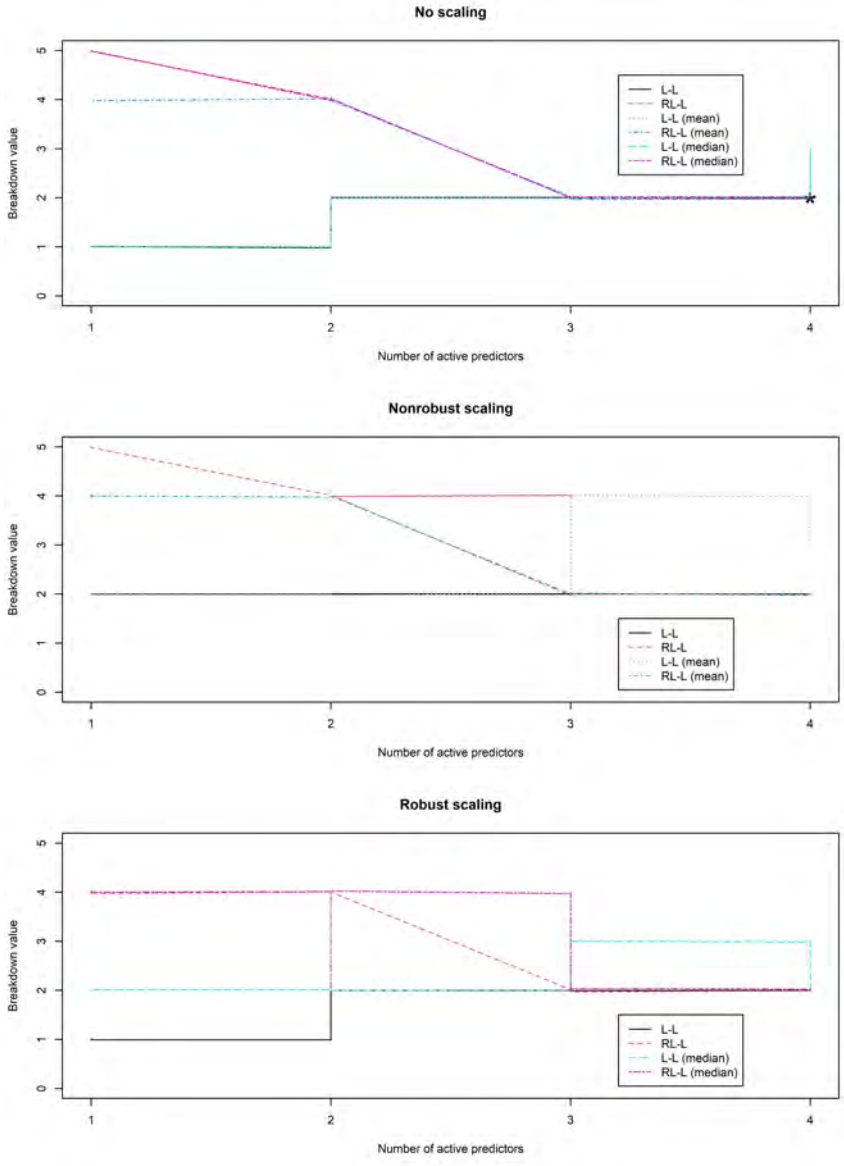

Figure 1: Breakdown values for aircraft data.

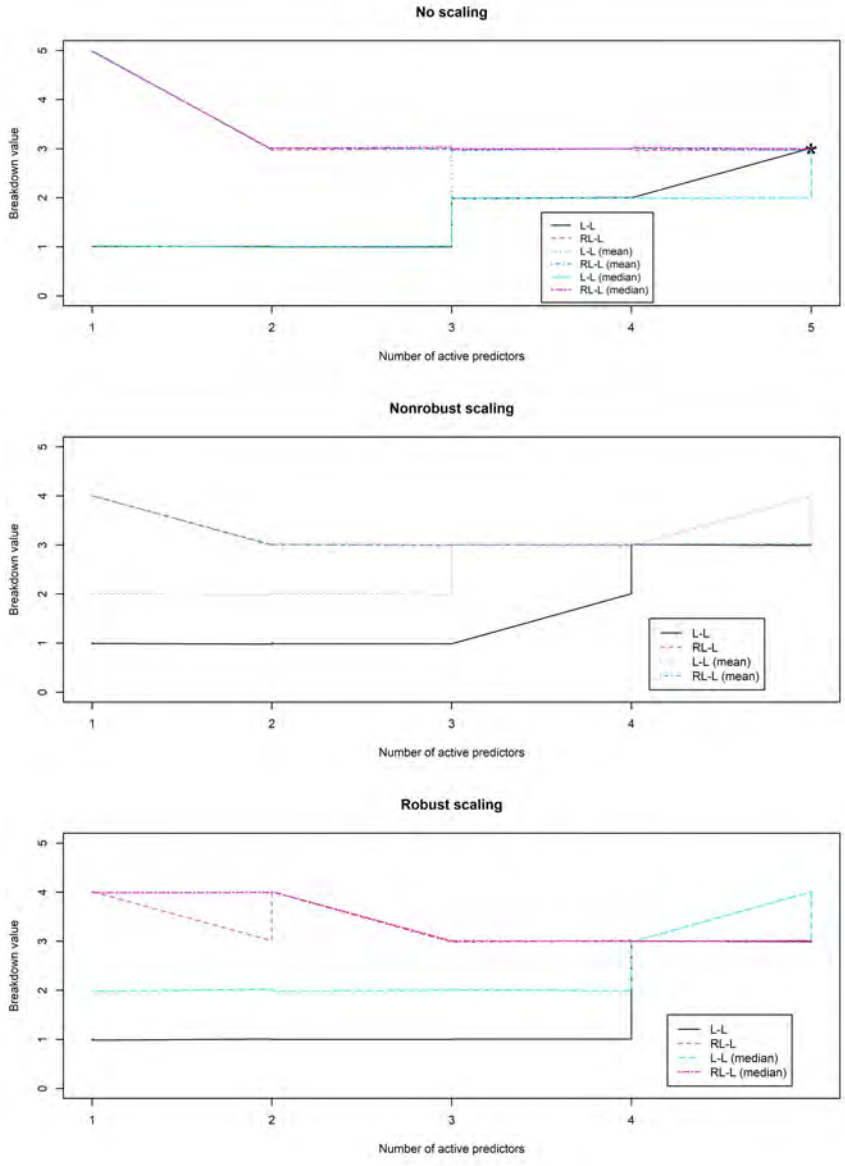

Figure 2: Breakdown values for coleman data.

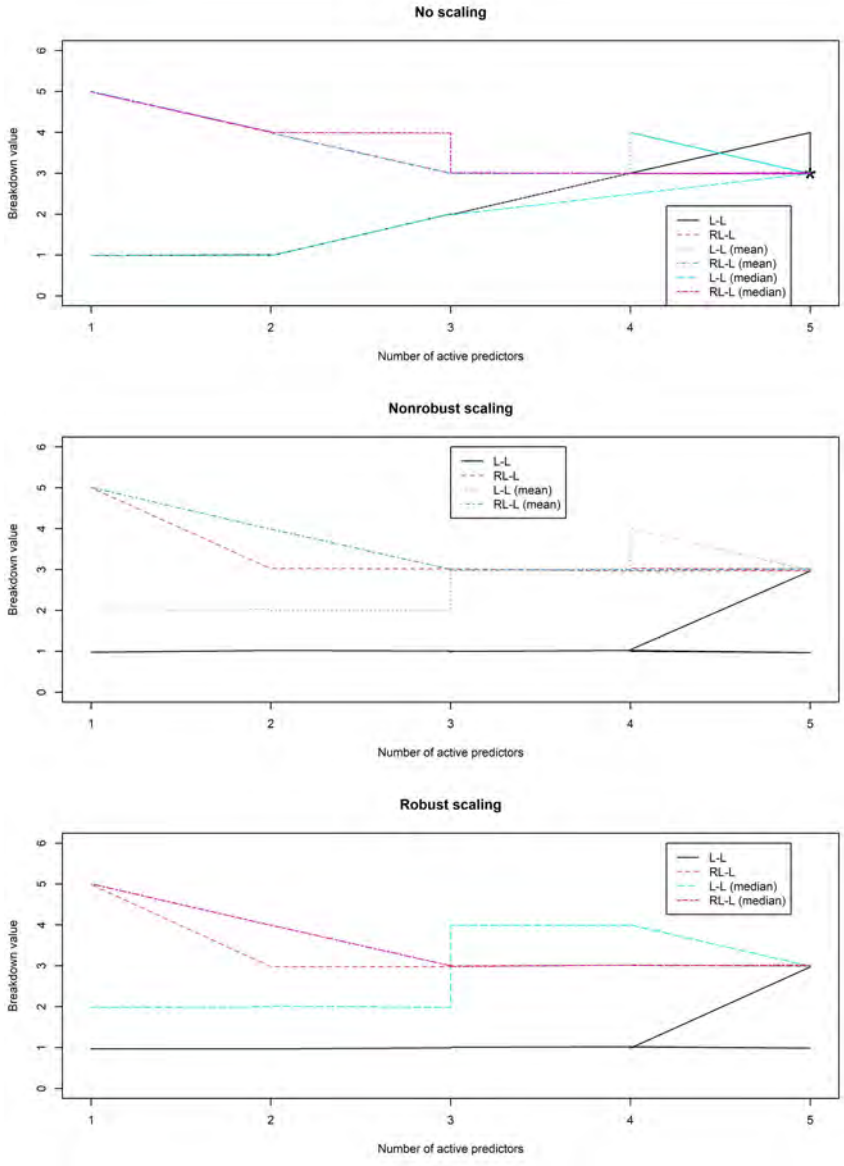

Figure 3: Breakdown values for wood data.

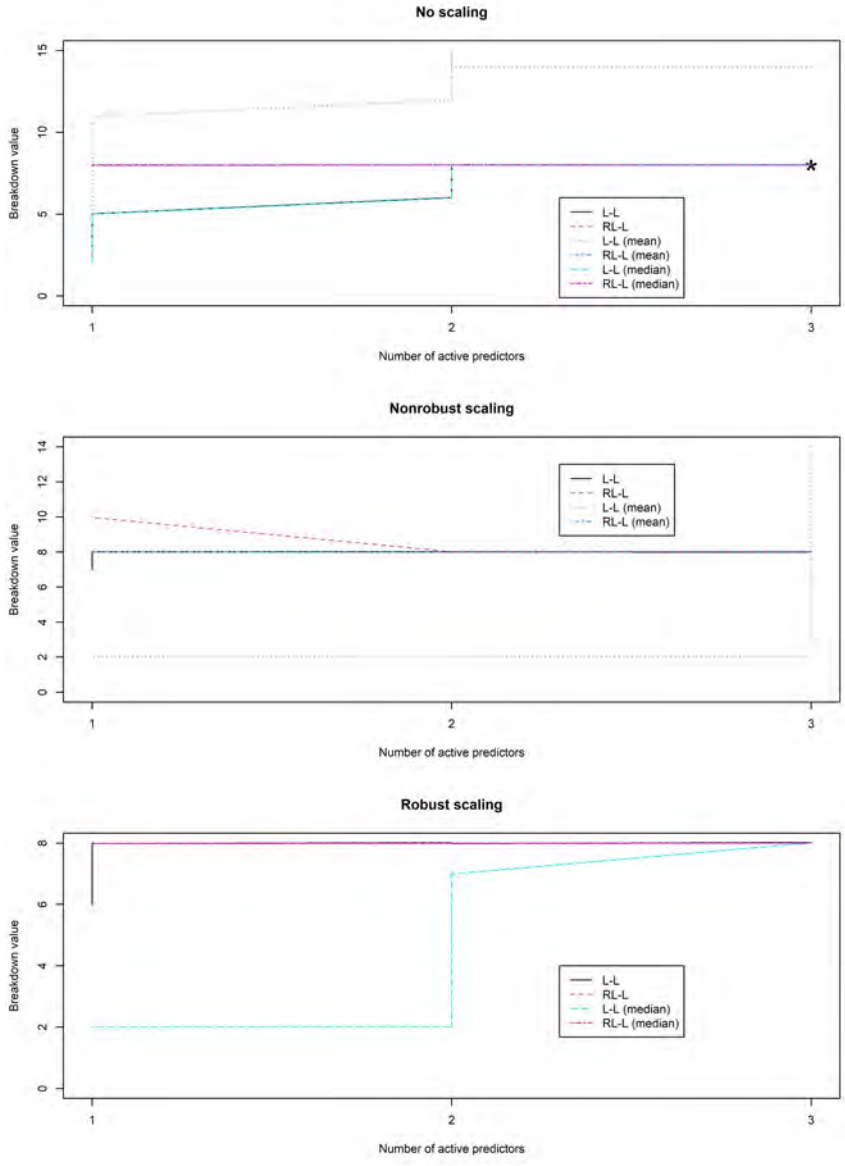

Figure 4: Breakdown values for HBK data.

don't improve on the breakdown from their original L-L method. Further, indirectly estimating the intercept by centering using the mean results in far higher breakdown values for L-L than any other method. We speculate that this could be because fourteen observations in the data set were artificially created to be leverage points that mask each other, with very unusual predictor values relative to the others. When centering using the median, these points are still recognized as unusual, and LAD is affected by them, but when centering by the mean, they are no longer unusual, and their effects on LAD are accordingly less. This notion is only of limited applicability, however, because scaling using the nonrobust standard deviation (while still centering using the mean) results in the lowest breakdown values of any method.

The salinity data occupies a somewhat middle position compared to the others. The RL-L methods still generally improve on breakdown as the number of active predictors becomes fewer, but centering can result in higher breakdown values for L-L methods for more active predictors. Further, while RL-L methods can also improve breakdown when scaling variables, scaling L-L methods (either robust or nonrobust) sometimes have higher breakdown and sometimes have lower breakdown.

It is important to remember what these results mean. The calculated breakdown values are based on an already-specified value of λ. In practice, this value would be chosen in a data-dependent way, and unusual observations could affect that choice in unanticipated ways. Thus, we cannot say that a particular method, applied to data sets with specific values in a data-dependent way, would lead to better performance. What we can say, however, is that apparently the breakdown properties of LAD-LASSO tend to mirror those of LAD itself, and become worse for choices of λ that lead to fewer active predictors. This tendency, however, can be mitigated by applying LAD to those data sets with fewer predictors (i.e., constructed RL-L estimates), which can result in often noticeable improvements in breakdown.

0.6 Conclusion and future work

In this chapter we have illustrated how the worst-case robustness of the LAD-LASSO method can be examined through determination of the breakdown value for a given data set and specified choice of the regularization parameter λ, and how using a version of the relaxed LAD-LASSO can sometimes improve the breakdown considerably. A natural next question is to wonder if the (relaxed) LAD-LASSO method can be modified to improve the breakdown further. [9] and [10] showed how weights can be chosen to produce a weighted LAD (WLAD) estima-

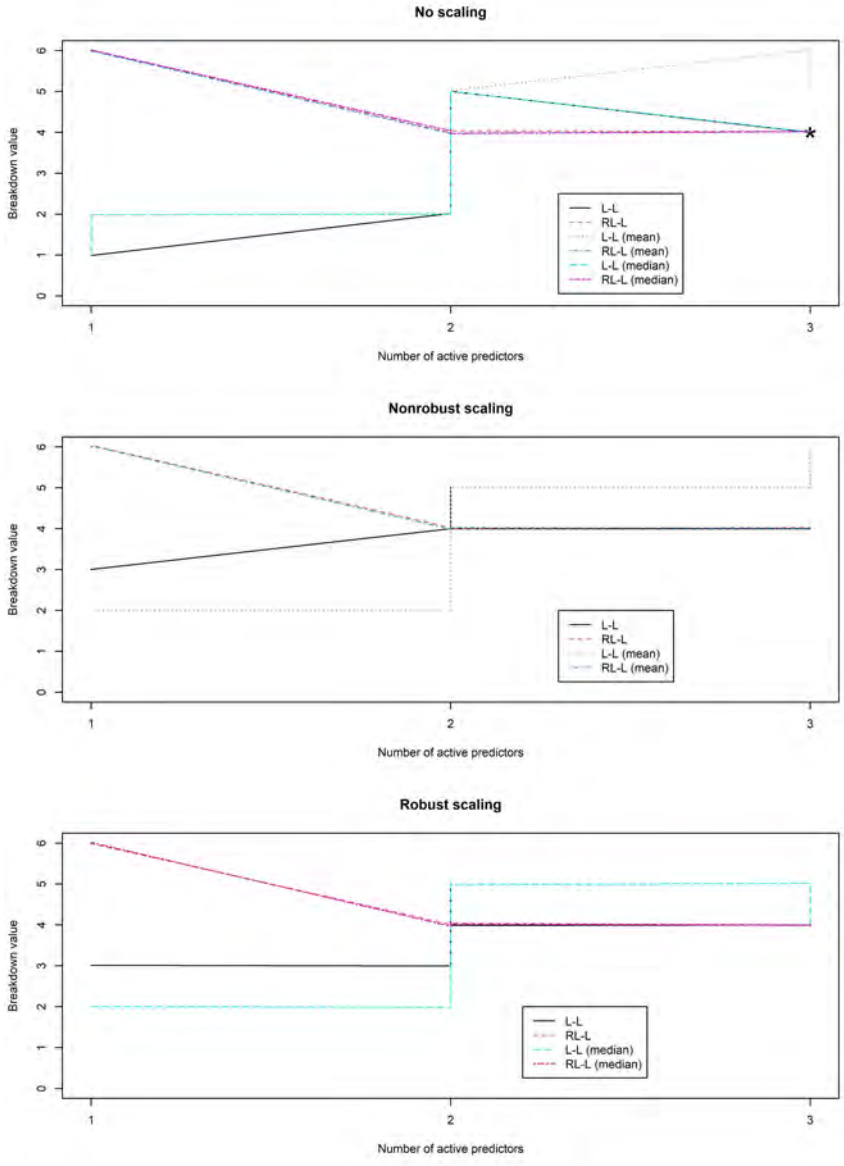

Figure 5: Breakdown values for salinity data.

tor that can increase the breakdown of LAD regression; it would be interesting to see if a similar improvement can be made in constructing WLAD-LASSO methods. Several authors have proposed versions of WLAD-LASSO (for example, [2], [5], and [11]), and it would be interesting to explore the breakdown properties of those methods as well.

References

[1] S.M. Ajeel and H.A. Hashem. Comparison some robust regularization methods in linear regression via simulation study. *Academic Journal of Nawroz University*, 9(2):244–252, 2020. https://doi.org/10.25007/ajnu.v9n2a818.

[2] O. Arslan. Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. *Computational Statistics and Data Analysis*, 56:1952–1965, 2012.

[3] S.D. Cahya, B. Sartono, Indahwati, and E. Purnaningrum. Performance of LAD-LASSO and WLAD-LASSO on high dimensional regression in handling data containing outliers. *Jurnal Teori dan Aplikasi Matematika*, 6:844–856, 2022. https://journal.ummat.ac.id/index.php/jtam/article/view/8968.

[4] X. Chen, Z. Wang, and M. McKeown. Asymptotic analysis of robust LASSOs in the presence of noise with large variance. *IEEE Transactions on Information Theory*, 56:5131–5149, 2010.

[5] X. Gao and Y. Feng. Penalized weighted least absolute deviation regression. *Statistics and Its Interface*, 11:79–89, 2018.

[6] X. Gao and J. Huang. Asymptotic analysis of high-dimensional LAD regression with LASSO. *Statistica Sinica*, 20:1485–1506, 2010.

[7] A. Giloni and M. Padberg. Alternative methods of linear regression. *Mathematical and Computer Modelling*, 35:361–374, 2002.

[8] A. Giloni and M. Padberg. The finite sample breakdown point of ℓ_1-regression. *SIAM Journal of Optimization*, 14:1028–1042, 2004.

[9] A. Giloni, B. Sengupta, and J.S. Simonoff. A mathematical programming approach for improving the robustness of least sum of absolute deviations regression. *Naval Research Logistics*, 53:261–271, 2006.

[10] A. Giloni, J.S. Simonoff, and B. Sengupta. Robust weighted LAD regression. *Computational Statistics and Data Analysis*, 50:3124–3140, 2006.

[11] Y. Jiang, Y. Wang, J. Zhang, B. Xie, J. Liao, and W. Liao. Outlier detection and robust variable selection via the penalized weighted LAD-LASSO method. *Journal of Applied Statistics*, 48:234–246, 2021.

[12] H. Li, X. Xu, Y. Lu, X. Yu, T. Zhao, and R. Zhang. Robust variable selection based on relaxed lad lasso. *Symmetry*, 14:2161, 2022. `https://doi.org/10.3390/sym14102161`.

[13] I. Mizera and C.H. Müller. Breakdown points and variation exponents of robust M-estimators in linear models. *Annals of Statistics*, 27:1164–1177, 1999.

[14] I. Mizera and C.H. Müller. The influence of the design on the breakdown points of ℓ_1-type M-estimators. In A. Atkinson and W. Müller, editors, *MODA6 — Advances in model-oriented design and analysis*, pages 193–200. Physica-Verlag, Heidelberg, 2001.

[15] S. Rahardiantoro and A. Kurnia. LAD-LASSO: Simulation study of robust regression in high dimensional data. *Indonesian Journal of Statistics*, 18:105–107, 2015. `https://journal.ipb.ac.id/index.php/statistika/article/view/16775`.

[16] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., New York, NY, 1987.

[17] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[18] H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business and Economic Statistics*, 25:347–355, 2007.

[19] L. Wang. The L_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

[20] J. Xu and Z. Ying. Simultaneous estimation and variable selection in median regression using Lasso-type penalty. *Annals of the Institute of Statistical Mathematics*, 62:487–514, 2010.