Robust penalized least squares of depth trimmed regression

Yijun Zuo[a], Ping-Shou Zhong[b], Haolei Weng[a]

[a] *Department of Statistics and Probability*
Michigan State University, East Lansing, MI 48824, USA
zuo@msu.edu, wenghaol@msu.edu

and

[b] Department of Mathematics, Statistics, and Computer Science
University of Illinois Chicago
pszhong@uic.edu

April 17, 2024

Abstract

Challenges with data in the big-data era include (i) the dimension p is often larger than the sample size n (ii) outliers or contaminated points are frequently hidden and more difficult to detect. Challenge (i) renders most conventional methods inapplicable. It, thus, attracts tremendous attention from statistics, computer science, industry, and bio-medical communities. Numerous penalized regression methods have been introduced as modern methods for analyzing high-dimensional data. Disproportionate attention has been paid to the challenge (ii) though. Penalized regression methods can do their job very well and are expected to handle the challenge (ii) simultaneously. The fact is most of them can break down by a single outlier (or single adversary contaminated point) as revealed in this article. The latter systematically examines a large class of penalized regression methods in terms of their robustness and provides quantitative assessment and reveals that most of leading penalized regression methods can break down by a single outlier. Consequently, a novel robust penalized regression method based on the least squares of depth trimmed residuals is proposed and studied carefully. Experiments with simulated and real data reveal that the newly proposed method can outperform some leading competitors in terms of estimation and prediction accuracy in the cases considered.

AMS 2000 Classification: Primary 62J05, 62G36; Secondary 62J99, 62G99

Key words and phrase: penalized regression, robustness of regularized regression estimators, least squares of depth trimmed residuals, finite sample prediction error bound, approximate computation algorithms.

Running title: robust penalized depth trimmed regression

1 Introduction

Least squares regression, the classic approach Consider the classic linear regression model,

$$y_i = (1, \boldsymbol{x}_i')\boldsymbol{\beta}_0 + e_i := \boldsymbol{w}_i'\boldsymbol{\beta}_0 + e_i, \tag{1}$$

where random variables y_i and $e_i \in \mathbb{R}$, random vector $\boldsymbol{x}_i \in \mathbb{R}^{p-1}$ ($p \geq 2$), and $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown parameter of interest, $'$ stands for the transpose. One wants to estimate the $\boldsymbol{\beta}_0$ based on a given sample $\mathbf{z}^{(n)} := \{(\boldsymbol{x}_i', y_i)', i \in \{1, \cdots, n\}\}$ from a parent model $y = \boldsymbol{w}'\boldsymbol{\beta}_0 + e$.

Call the difference between y_i (observed value) and $\boldsymbol{w}_i'\boldsymbol{\beta}$ (predicted value), the ith residual for a candidate coefficient vector $\boldsymbol{\beta}$ (which is often suppressed). Denote it by r_i, that is,

$$r_i := r_i(\boldsymbol{\beta}) = y_i - (1, \boldsymbol{x}_i')\boldsymbol{\beta} := y_i - \boldsymbol{w}_i'\boldsymbol{\beta}. \tag{2}$$

To estimate $\boldsymbol{\beta}_0$, the classic *least squares* (LS) estimator is the minimizer of the sum of the squared residuals (SSR): $\widehat{\boldsymbol{\beta}}_{ls} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n r_i^2$. Alternatively, one can replace the square by the absolute value to obtain the *least absolute deviations* (lad) estimator (aka, L_1, in contrast to the L_2 (LS), estimator). A straightforward derivation leads to

$$\widehat{\boldsymbol{\beta}}_{ls} = (\boldsymbol{X}_n'\boldsymbol{X}_n)^{-1}\boldsymbol{X}_n'\boldsymbol{Y}_n. \tag{3}$$

where $\boldsymbol{Y}_n = (y_1, \cdots, y_n)'$, $\boldsymbol{X}_n = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)'$ and the columns of $\boldsymbol{X}_n$ are assumed to be linearly independent (i.e. $\boldsymbol{X}_n$ has a full rank p ($\leq n$)).

The LS estimator is popular in practice across a broader spectrum of disciplines due to its (i) great computability (with the computation formula); and (ii) optimal properties (the best linear unbiased estimator (BLUE) and the uniformly minimum variance unbiased estimator (UMVUE), page 186 of [42] when the i.i.d. error e_i follows a normal $\mathcal{N}(0, \sigma^2)$.

It, however, can behave badly when the error distribution is slightly departed from the normal distribution, particularly when the errors are heavy-tailed or contain outliers.

Penalized regression, the state of the art In modern applied data analysis, the number of variables often is even larger than the number of observations. Traditional methods such as the LS can then no longer be applied due to the design matrix $\boldsymbol{X}$ being less than p rank ($n < p$), hence LS estimator is no longer unique and its variance is large if $\boldsymbol{X}$ is close to collinear. Furthermore, models that include the full set of explanatory variables often have poor prediction performance as they tend to have large variance while large models are in general difficult to interpret. *Ridge regression*, minimizing SSR, subject to a constraint $\sum_{i=1}^p |\beta_i|^2 < t$

$$\widehat{\boldsymbol{\beta}}_{ridge}(\lambda) := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \sum_{i=1}^n r_i^2 + \lambda \sum_{i=1}^p |\beta_i|^2 \Big\}, \tag{4}$$

first proposed in [21, 22], is a useful tool for improving prediction in regression situations with highly correlated predictors and tackling the non-inverse issue,

$$\widehat{\boldsymbol{\beta}}_{ridge}(\lambda) = (\boldsymbol{X_n}'\boldsymbol{X_n} + \lambda I_{d\times d})^{-1}\boldsymbol{X_n}'\boldsymbol{Y_n}, \tag{5}$$

its variance is smaller than that of the LS estimator. Therefore, better estimation can be achieved on the average in terms of mean squared error (MSE) with a little sacrifice of bias, and predictions can be improved overall.

Bridge regression, introduced in [14], generalized the ridge regression, minimizes SSR subject to a constraint $\sum_{i=1}^{p} |\beta_i|^\gamma \leq t$ with $\gamma \geq 0$,

$$\widehat{\boldsymbol{\beta}}_{bridge}(\lambda, \gamma) := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \Big\{ \sum_{i=1}^{n} r_i^2 + \lambda \sum_{i=1}^{p} |\beta_j|^\gamma \Big\}. \tag{6}$$

Ridge regression ($\gamma = 2$) and subset selection ($\gamma = 0$) are special cases.

Least absolute shrinkage and selection operator (lasso) was introduced in [47], minimizing SSR subject to a constraint $\sum_{j=1}^{p} |\beta_j| \leq t$, is a special case of the bridge with $\gamma = 1$. As pointed out in [47], the lasso shrinks the LS estimator $\widehat{\boldsymbol{\beta}}_{ls}$ towards 0 and potentially sets $\widehat{\beta}_j = 0$ for some j. That is, it performs as a variable selection operator simultaneously.

Other remarkable approaches of regularized regression for high dimensional sparse data include, among others, **(i)** *elastic nets*, introduced in [55], a generalization of the ridge and lasso models, which combines the two penalties and yields

$$\widehat{\boldsymbol{\beta}}_{enet}(\lambda_1, \lambda_2) := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \Big\{ \sum_{i=1}^{n} r_i^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} \beta_i^2 \Big\}. \tag{7}$$

(ii) *square-root lasso*, introduced in [3], to avoid the pre-estimation of standard deviation σ of the error term in lasso and achieve a better performance, is defined as

$$\widehat{\boldsymbol{\beta}}_{sqrt-lasso} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \Big\{ \big(\sum_{i=1}^{n} r_i^2 \big)^{1/2} + \lambda \sum_{i=1}^{p} |\beta_i| \Big\}. \tag{8}$$

(iii) *slope* (Sorted L-One Penalized Estimation) introduced in [5], aim to control the false discover rate (FDR),

$$\widehat{\boldsymbol{\beta}}_{slope} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \Big\{ \sum_{i=1}^{n} r_i^2 + \sum_{i=1}^{p} \lambda_i |\beta_{(i)}| \Big\}, \tag{9}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and $|\beta_{(1)}| \geq |\beta_{(2)}| \geq \cdots \geq |\beta_{(p)}|$.

Strong connections between some modern methods and a method called *least angle regression* (lar) was revealed in [10] where they developed an algorithmic framework that includes all of these methods (lasso, boosting, forward stagewise regression) and provided a fast implementation, for which they used the term 'lars'. Lars is a promising technique/algorithm for variable selection applications, offering a nice alternative to stepwise regression. For an excellent review on lars and lasso, see [20].

Other outstanding penalized regression estimators include, among others, SCAD [12], [13] and MCP [53]. It is not our goal to review all promising penalized/regularized regression estimators in the literature. For a detailed account about lasso and its variants, refer to Table 6 of [15] or Fig. 1 of [50], and [54] and references therein.

The penalized regression estimators above improve prediction accuracy meanwhile enhance the interpretability of the model. They, however, pay the price of inducing a little bit of bias in addition to the lack of robustness. There are numerous published articles related to

lasso and regularized regression in the literature. However, there are disproportionately few addressing the robustness of the estimators. Are they as robust as supposed (or expected)? Or rather can they resist the influence of just a single contaminated point (or outlier) that is typically buried in high-dimensional data?

Robust versions of the lasso (or ridge) estimators have been sporadically considered in the literature. The LS in lasso (or ridge), is replaced by M-estimators, as in [51] and [30]; or replaced by a Huber-type loss function, as in [35] and [45]; or by lad, as in [49];

$$\widehat{\boldsymbol{\beta}}_{lad-lasso} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \sum_{i=1}^{n} |r_i| + \lambda \sum_{i=1}^{p} |\beta_i| \Big\}, \tag{10}$$

or replacing correlations in lars by a robust type of correlation, as in [25] (Rlars); or by S- ([37]) and MM- ([52]) estimators, as in [32] for ridge regression (Rrr); or by the least trimmed squares (LTS) ([36]), as in [2]. The LTS is defined as

$$\widehat{\boldsymbol{\beta}}^{n}_{lts} := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{h} (r^2)_{i:n}, \tag{11}$$

where $(r^2)_{1:n} \leq (r^2)_{2:n} \leq \cdots, (r^2)_{n:n}$ are the ordered squared residuals, $\lceil n/2 \rceil \leq h \leq n$, and $\lceil \cdot \rceil$ is the ceiling function. The authors of [2] replaced SSR by the loss of LTS and defined

$$\widehat{\boldsymbol{\beta}}_{lts-lasso} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \sum_{i=1}^{h} (r^2)_{i:n} + h\lambda \sum_{i=1}^{p} |\beta_i| \Big\}, \tag{12}$$

The idea of [2] has extended to logistic regression with elastic net penalty in [27], and penalized weighted M-type estimators for the logistic regression have also been studied in [4].

Most estimators above (except Rlars, Rrr, and $\widehat{\boldsymbol{\beta}}_{lts-lasso}$), like both L_1 and L_2 (LS) estimators, unfortunately, have the worst case 0% asymptotic breakdown point (i.e., one bad point can ruin (break down) the estimator), in sharp contrast to the 50% of the least squares of trimmed (LST) estimator (see Section 3.1 of [61] or Section 3 here). Estimators in [27] and [4] are claimed to be robust, but no qualitative robustness assessment of these estimators has been established yet. The same situation with the estimator in [41].

Now let us take a close look at the three exceptions above. The main drawback of the Rlars is the lack of a natural definition or a clear objective function, as commented in [2]. The main focus of [32] is robustifying ridge regression (Rrr).

Only $\widehat{\boldsymbol{\beta}}_{lts-lasso}$ in [2] has an established high finite sample breakdown point (see Section 3 for definition). Their result, though covers the lasso-type estimators, does not cover the elastic nets and other estimators; the authors failed to (i) explain why their estimator can have a breakdown point higher than 50% and (ii) study the properties (such as equivariance) of their estimator. Furthermore, the LTS is notorious for its inefficiency (i.e., it usually has a large variance). On the other hand, the LST introduced in [61] is as robust as LTS and can outperform the LTS (especially in efficiency) as demonstrated in [61].

Open problems and major Contributions Based on observations above, questions we want to address are: **(i)** Can one replace the loss of LS with that of the LST in the penalized regression? How does the resulting estimator perform? Is it more robust, compared with existing ones? **(ii)** Can one provide a more general breakdown robustness assessment that covers a large general class of regularized regression estimators and explain why a breakdown point higher than 50% can be **(iii)** Besides robustness, what are other desirable properties for a regression estimator?

The main contributions of this article include **(i)** it examines the robustness for a large class of (penalized) regression estimators (including all leading ones, e.g. the L_1 and L_2, the ridge, the bridge, the lasso, the elastic net, the slope, the lad-lasso, and the sqrt-lasso) and proves that they can break down by a single adversary contaminating point; **(ii)** it, consequently, introduces a novel and robust penalized least squares of depth trimmed estimator ($\widehat{\boldsymbol{\beta}}_{lst-enet}$) that has a much higher breakdown point robustness and outperforms leading competitors; and **(iii)** it proposes an efficient computational algorithm for the estimator and vets its performance for simulated and real high-dimensional data and reveals that $\widehat{\boldsymbol{\beta}}_{lst-enet}$ can outperform its competitors in most cases considered; **(iv)** it discusses three desired equivariance properties for regression estimators.

Section 2 establishes a robust result for a large class of general regularized regression estimators and reveals that most of leading estimators (including bridge, lasso, slope, and enet) has the worst breakdown point robustness. Section 3 invokes the least squares of depth trimmed (LST) procedure introduced in [61] and reviews its robust property meanwhile introduces a class of penalized regression estimators based on the LST and studies their properties including, existence and uniqueness, robustness, and equivariance. Section 4 addresses the computation issue of $\widehat{\boldsymbol{\beta}}_{lst-enet}$. Section 5 consists of simulation/comparison study and real data application of five competing methods. Section 6 ends the article with some concluding discussions. Proofs of main results are deferred to an Appendix.

2 Robustness of a class of penalized regression estimators

Are the existing numerous penalized regression methods mentioned above as robust as they are expected or believed to be? Or rather can they resist the influence of just a single outlier (or adversary single-point contamination)? We now formally address this issue.

2.1 A robustness measure

In the finite sample practice, the most prevailing quantitative measure of the robustness of any regression or location estimators is the *finite sample breakdown point*, introduced in [9].

Definition [9] The finite sample *replacement breakdown point* (RBP) of a regression estimator $\mathbf{t}$ at a given un-contaminated sample $\mathbf{z}^{(n)} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_n\}$, where $\boldsymbol{z}_i := (\boldsymbol{x}_i', y_i)'$, is

$$\mathrm{RBP}(\mathbf{t}, \mathbf{z}^{(n)}) = \min_{1\le m\le n} \left\{ \frac{m}{n} : \sup_{\mathbf{z}_m^{(n)}} \|\mathbf{t}(\mathbf{z}_m^{(n)}) - \mathbf{t}(\mathbf{z}^{(n)})\|_2 = \infty \right\}, \tag{13}$$

where $\mathbf{z}_m^{(n)}$ stands for an arbitrary contaminated sample by replacing m original sample points in $\mathbf{z}^{(n)}$ with arbitrary points in $\mathbb{R}^p$ and $\|\boldsymbol{x}\|_q = (\sum_{i=1}^n x_i^q)^{1/q}$ is the ℓ_q-norm for vector $\boldsymbol{x} \in \mathbb{R}^n$.

Namely, the RBP of an estimator is the minimum replacement fraction that could drive the estimator beyond any bounds. It turns out that both L_1 (least absolute deviations) and L_2 (least squares) estimators have RBP $1/n$ (or 0%) whereas LST can have $(\lfloor n/2 \rfloor - p + 2)/n$ (or 50%) (see Section 3), the highest possible asymptotic value for any regression equivariant estimators (see [38] or Section 3), where $\lfloor \cdot \rfloor$ is the floor function. We now present a general RBP result for a large class of penalized regression estimators.

2.2 A general result on penalized regression estimators

Theorem 2.1 For any un-contaminated data set $\boldsymbol{z}^{(n)} = \{(\boldsymbol{x}_i', y_i)', i \in \{1, \cdots, n\}\}$ in $\mathbb{R}^p$ ($p \geq 2$), let $\widehat{\boldsymbol{\beta}}^*(\lambda_1, \lambda_2, \gamma, \boldsymbol{z}^{(n)})$ be the penalized regression estimator, which minimizes the objective

$$O(\boldsymbol{\beta}, \lambda_1, \lambda_2, \gamma, \boldsymbol{z}^{(n)}) := f\Big(\frac{1}{n}\sum_{i=1}^n \mathcal{L}(r_i)\Big) + g(\boldsymbol{\beta}, \lambda_1, \lambda_2, \gamma), \tag{14}$$

where $\lambda_i, \gamma \geq 0$, the combined penalty function $g(\boldsymbol{\beta}, \lambda_1, \lambda_2, \gamma) \geq 0$ and $f(x)$ is continuous, strictly increasing over $[0, \infty)$, $f(0) = 0$ and $f(x) \to \infty$ as $x \to \infty$, the loss function $\mathcal{L}(x)$ is non-negative, non-decreasing over $[0, \infty)$, $\mathcal{L}(0) = 0$ and $\mathcal{L}(x) \to \infty$ when $x \to \infty$. Then

$$\mathrm{RBP}(\widehat{\boldsymbol{\beta}}^*(\lambda_1, \lambda_2, \gamma, \boldsymbol{z}^{(n)}), \boldsymbol{z}^{(n)}) = 1/n.$$

Proof: see the Appendix. □

Remarks 2.1

(i) Conditions on loss $\mathcal{L}(x)$ are relative loose, they hold automatically if $\mathcal{L}(x)$ is non-negative, non-decreasing, and convex in x and $\mathcal{L}(0) = 0$. The $\mathcal{L}(x)$ in theorem covers almost all loss functions (with two exceptions out of fifteen) in Table 6 of [15]. The penalty function $g(\boldsymbol{\beta}, \lambda_1, \lambda_2, \gamma)$ covers almost all existing ones including, among others, $\lambda_1\|\boldsymbol{D}_1\boldsymbol{\beta}\|_\gamma^\gamma + \lambda_2\|\boldsymbol{D}_2\boldsymbol{\beta}\|_2^2$, with $\boldsymbol{D}_i$ being penalty matrices. Typical f includes the identity or square-root function (covering the square-root lasso [3]). To the best of our knowledge, a similar general result, judging the robustness of a large class of penalized regression estimators, never appeared in the literature before.

(ii) The RBP result in the theorem is very general since the loss function covers most of the existing loss functions in the machine learning and AI literature, e.g., the most popular ones: negative log-likelihood; the ℓ_1, the ℓ_2, or any ℓ_q ($q \geq 1$) loss, Huber loss, and the loss of the lasso and most of its variants (see Table 6 of [15]). The penalty format covers most of the existing ones in the literature (indeed, it covers all twenty-five penalty functions listed in Table 1 of [50]).

(iii) The great generality of the result in the theorem implies that most of the existing penalized regression (and the classic L_1 and L_2) estimators are not robust. In fact, they can break down with just one single outlier which often buries in high dimensional data. □

Now that most of the existing penalized regression estimators can be broken down by a single outlier (or single-point contamination). Furthermore, existing robust penalized regression estimators are most ad hoc, e.g., Rlars of [25] is for robustifying lars, and Rrr of [32] is for robustifying ridge regression, and [4] is mainly for robustifying the penalized logistic regression estimators. Only $\widehat{\boldsymbol{\beta}}_{lts-lasso}$ of [2] and $\widehat{\boldsymbol{\beta}}_{enetLTS}$ of [27] that employed the loss of LTS to replace the SSR in lasso have really high breakdown robustness meanwhile do the variable selection job. But the major drawback of the LTS is its inefficiency (it has a larger variance) as demonstrated in [61] (also Sections 3 and 6) and Figures 2 and 3 of [27].

A natural question is: can one construct a penalized regression estimator that is robust against the outliers or contamination meanwhile is more efficient (i.e., with a smaller variance than that of the LTS)? In the sequel, we achieve this goal by adopting the loss of least squares of depth trimmed (LST) introduced in [61] and applying it to the penalized regression setting.

3 A class of penalized regression estimators based on the LST

3.1 Least Squares of depth trimmed (LST) estimator

Definition 3.1 To robustify the LS, least trimmed squares (LTS) estimator was introduced in [36]. The procedure orders the squared residuals and then trims the larger ones and keeps at least $h \geq \lceil n/2 \rceil$ squared residuals, the minimizer of the sum of those *trimmed squared residuals* is called an LTS estimator as defined in (11). $\widehat{\boldsymbol{\beta}}_{lts}^{n}$ is highly robust but is not very efficient, as reported in [33] (page 132) or in [44] having just 7% or 8% asymptotic efficiency. A more efficient competitor, least squares of trimmed (LST) estimator, is introduced in [61], overcoming LTS drawback while sharing its high robustness and fast computation advantages.

Depth (or outlyingness)-based trimming scheme trims points that lie on the outskirts (i.e. points that are less deep, or outlying). The outlyingness (or, equivalently, depth) of a point x is defined to be (strictly speaking, depth=1/(1+outlyingness) in [56] and [60])

$$O(x, \boldsymbol{x}^{(n)}) = |x - \text{Med}(\boldsymbol{x}^{(n)})|/\text{MAD}(\boldsymbol{x}^{(n)}), \tag{15}$$

where $\boldsymbol{x}^{(n)} = \{x_1, \cdots, x_n\}$ is a data set in $\mathbb{R}^1$, $\text{Med}(\boldsymbol{x}^{(n)}) = \text{median}(\boldsymbol{x}^{(n)})$ is the median of the data points, and $\text{MAD}(\boldsymbol{x}^{(n)}) = \text{Med}(\{|x_i - \text{Med}(\boldsymbol{x}^{(n)})|,\ i \in \{1, 2, \cdots, n\}\})$ is the median of absolute deviations to the center (median). It is readily seen that $O(x, \boldsymbol{x}^{(n)})$ is a generalized standard deviation, or equivalent to the one-dimensional projection depth (see [60] and [56, 57] for a high dimensional version). For the notion of outlyingness, cf. [43], [7], and [8].

For a given constant α (hereafter assume $\alpha \geq 1$), $\boldsymbol{\beta}$, and $\boldsymbol{z}^{(n)}$, define a set of indexes

$$I(\boldsymbol{\beta}) = \Big\{ i : O(r_i, \boldsymbol{r}^{(n)}) \leq \alpha,\ i \in \{1, \cdots, n\} \Big\}. \tag{16}$$

where $\boldsymbol{r}^{(n)} = \{r_1, r_2, \cdots, r_n\}$ and r_i is defined in (2). Namely, the set of subscripts so that the outlyingness (see (15)) (or depth) of the corresponding residuals are no greater (or less)

than α (or $1/(1+\alpha)$). It depends on $\mathbf{z}^{(n)}$ and α, which are suppressed in the notation. For a fixed constant α in the depth trimming scheme, consider the quantity

$$Q^n(\boldsymbol{\beta}) := Q(\boldsymbol{z}^{(n)}, \boldsymbol{\beta}, \alpha) := \sum_{i=1}^{n} r_i^2 \mathbb{1}\left(O(r_i, \boldsymbol{r}^{(n)}) \leq \alpha\right) = \sum_{i \in I(\boldsymbol{\beta})} r_i^2, \tag{17}$$

where $\mathbb{1}(A)$ is the indicator of A (i.e., it is one if A holds and zero otherwise). Namely, residuals with their outlyingness (or equivalently reciprocal of depth minus one) greater than α will be trimmed. When there is a majority ($\geq \lfloor (n+1)/2 \rfloor$) identical r_is, we define $\mathrm{MAD}(\mathbf{r}^{(n)}) = 1$ (since those r_i lie in the deepest position (or are the least outlying points)).

Minimizing $Q(\boldsymbol{z}^{(n)}, \boldsymbol{\beta}, \alpha)$, one gets the *least* sum of *squares* of *trimmed* (LST) estimator,

$$\widehat{\boldsymbol{\beta}}_{lst}^n := \widehat{\boldsymbol{\beta}}_{lst}(\mathbf{z}^{(n)}, \alpha) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} Q(\boldsymbol{z}^{(n)}, \boldsymbol{\beta}, \alpha). \tag{18}$$

Differences between LST and LTS Compared with the LTS defined in (11), it is readily seen that both estimators trim residuals. However, there are two essential differences: **(i)** the trimming schemes are different. The LTS employs a rank-based trimming scheme that focuses only on the relative position of points (squared residuals) with respect to others and ignores the magnitude of the point and the relative distance between points whereas the LST exactly catches these two important attributes. It orders data (residuals) from a center (the median) outward and trims the points (residuals) that are far away from the center. This is known as depth-based trimming. **(ii)** Besides the trimming scheme difference, there is another difference between the LTS and the LST, that is, the order of trimming and squaring. In the LTS, squaring is first, followed by trimming whereas, in the LST, the order is reversed.

Seemingly negligible differences lead to an unexpected performance difference in the LTS and the LST as demonstrated in [61] and in Figure 1 and Table 1 here.

Existence and uniqueness of $\widehat{\boldsymbol{\beta}}_{lst}^n$ have been addressed in [61], it is also equivariant (see [61]).

Equivariance A regression estimator $\mathbf{t}$ is called *regression, scale, and affine equivariant* if, respectively (see page 116 of [38]) with $N = \{1, 2, \cdots, n\}$

$$\begin{aligned}
\mathbf{t}\left(\{(\boldsymbol{w}_i', y_i + \boldsymbol{w}_i'\mathbf{b})', i \in N\}\right) &= \mathbf{t}\left(\{(\boldsymbol{w}_i', y_i)', i \in N\}\right) + \mathbf{b}, \ \forall\ \mathbf{b} \in \mathbb{R}^p \\
\mathbf{t}\left(\{(\boldsymbol{w}_i', s y_i)', i \in N\}\right) &= s\mathbf{t}\left(\{(\boldsymbol{w}_i', y_i)', i \in N\}\right), \ \forall\ s \in \mathbb{R}^1 \\
\mathbf{t}\left(\{(A'\boldsymbol{w}_i)', y_i)', i \in N\}\right) &= A^{-1}\mathbf{t}\left(\{(\boldsymbol{w}_i', y_i)', i \in N\}\right), \ \forall \text{ nonsingular } A \in \mathbb{R}^{p \times p}.
\end{aligned}$$

Now with the measure of robustness (presented in the last section), naturally one wants to ask the question: is $\widehat{\boldsymbol{\beta}}_{lst}^n$ theoretically more robust than the LS estimator $\widehat{\boldsymbol{\beta}}_{ls}^n$?

Robustness and efficiency of LST

We shall say $\mathbf{z}^{(n)}$ is *in general position* when any p of observations in $\mathbf{z}^{(n)}$ gives a unique determination of $\boldsymbol{\beta}$. In other words, any $(p-1)$ dimensional subspace of the space $(\boldsymbol{x}', y)'$

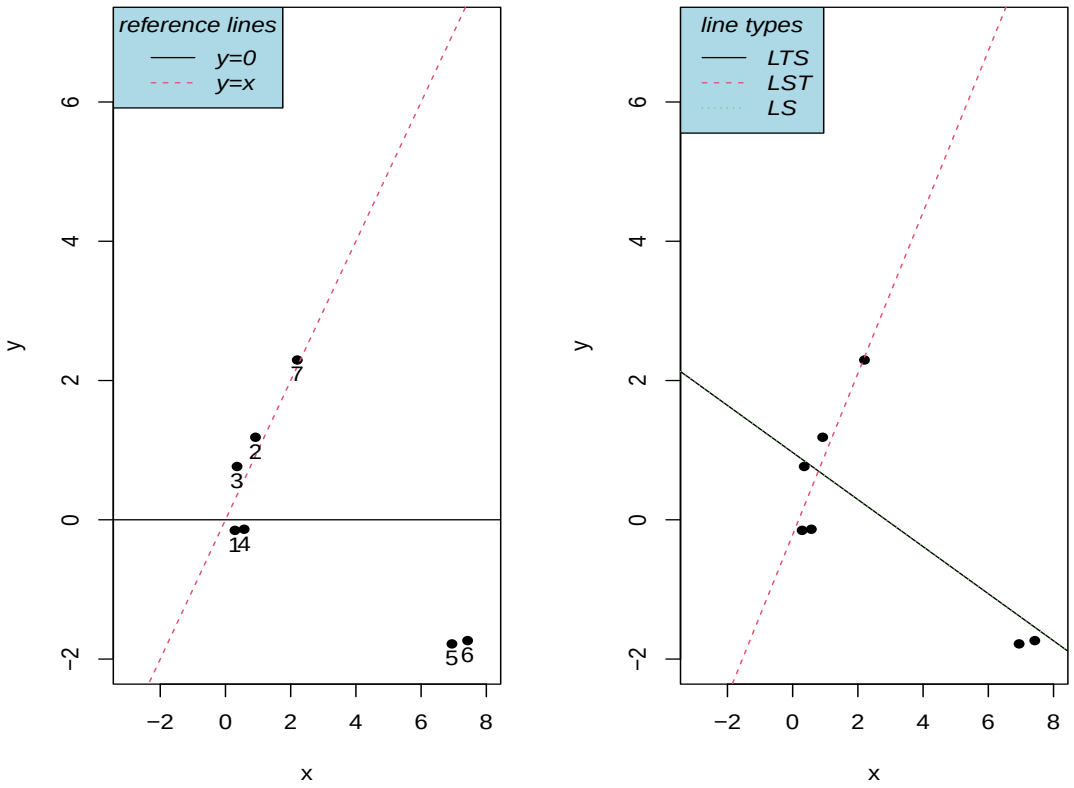

Figure 1: Left panel: plot of seven highly correlated normal points (with two points contaminated) and two reference lines $y = 0$ and $y = x$. Right panel: the same seven points are fitted by LTS, LST, and the LS (benchmark). A solid black line is LTS given by ltsReg which missed overall positive linear pattern. Red dashed line is LST given by lstReg which captured the overall positive linear pattern. Red dotted line is given by the LS which is identical to LTS line in this case, both are sensitive to the two outliers or 30% contamination. note that in theory, LTS can resist up to 50% contamination without breakdown.

contains at most p observations of $\mathbf{z}^{(n)}$. When the observations come from continuous distributions, the event ($\mathbf{z}^{(n)}$ being in general position) happens with probability one.

Lemma 3.1 [61] For $\widehat{\boldsymbol{\beta}}^n_{lst}$ defined in (18) with ($\alpha \geq 1$, $n \geq 2p+1$) and $\mathbf{z}^{(n)}$ in general position, we have

$$\mathrm{RBP}(\widehat{\boldsymbol{\beta}}^n_{lst}, \mathbf{z}^{(n)}) = \begin{cases} \lfloor (n+1)/2 \rfloor / n, & \text{if } p = 1, \\ (\lfloor n/2 \rfloor - p + 2)/n, & \text{if } p > 1. \end{cases} \tag{19}$$

□

Namely, the LST shares the best 50% asymptotic breakdown value of the LTS meanwhile it is much more efficient than the LTS as demonstrated in Table 1 below (also see [61])).

Robustness and efficiency usually do not work in tandem. To investigate the efficiency of LST and LTS, we carry a simulation.

Here we generate $R = 1000$ samples $\{\mathbf{z}_i = (\boldsymbol{x}_i', y_i)', i \in \{1, \cdots, n\}\}$ with various ns from the normal distribution, where mean being a zero-vector in $\mathbb{R}^{20}$, and covariance being a 20 by 20 matrix with diagonal entries being 1 and off-diagonal entries being 0.9. Then $\varepsilon\%$ of each sample are contaminated by $m = \lceil n\varepsilon \rceil$ points, we select m points of $\{\boldsymbol{z}_i, i \in \{1, \cdots, n\}\}$ randomly and replace them by $(7, 7, \cdots, 7, -7)' \in \mathbb{R}^{20}$, then apply the three methods, the LTS, the LST, and the LS, to the contaminated data sets.

We calculate EMSE $:= \sum_{i=1}^{R} \|\mathbf{t}_i - \boldsymbol{\beta}_0\|^2/R$, the empirical mean squared error (EMSE) for estimator $\mathbf{t}$. Here $\mathbf{t}_i$ is the realization of $\mathbf{t}$ obtained from the ith sample with size n and dimension $p = 20$. If $\mathbf{t}$ is regression equivariant (this is true for any of three), then we can assume (w.l.o.g.) that the true parameter $\boldsymbol{\beta}_0 = \mathbf{0} \in \mathbb{R}^p$.

Meanwhile, we also compute, SVAR($\boldsymbol{t}$) $:= \sum_{i=1}^{R} \|\mathbf{t}_i - \bar{\boldsymbol{t}}\|^2/(R-1)$, the sample variance of $\boldsymbol{t}$, where $\bar{\boldsymbol{t}}$ is the mean of all $\{\boldsymbol{t}_i, i \in \{1, 2, \cdots, R\}\}$, and obtain the finite sample relative efficiency (RE) of a procedure (denoted by P) with respect to the LS by the ratio of RE(P) := SVAR($\widehat{\boldsymbol{\beta}}_{ls}$)/SVAR($\widehat{\boldsymbol{\beta}}_P$). At the same time, we record the total time (TT) consumed by different procedures for all replications. The performance of LTS, LST, and LS is assessed by the four criteia: EMSE, SVAR, TT, and RE. Results are listed in Table 1 (note that theoretically both LTS and LST can resist up to 50% contamination without breakdown).

procedure	EMSE	SVAR	TT	RE	EMSE	SVAR	TT	RE
		n=100	$\varepsilon = 0\%$			n=200	$\varepsilon = 0\%$	
LTS	0.5282	0.4764	153.82	0.5164	0.2172	0.1651	265.98	0.6510
LST	0.2969	0.2447	10.911	1.0053	0.1594	0.1073	31.690	1.0019
LS	0.2979	0.2460	1.3254	1.0000	0.1596	0.1075	1.5022	1.0000
		n=100	$\varepsilon = 30\%$			n=200	$\varepsilon = 30\%$	
LTS	142.73	142.54	891.18	0.0864	31.662	31.627	855.82	0.1583
LST	0.4394	0.3875	25.787	31.768	0.2128	0.1606	60.492	31.182
LS	12.410	12.311	1.3358	1.0000	5.1072	5.1072	1.5138	1.0000

Table 1: EMSE, SVAR, TT (seconds), and RE for LTS, LST, and LS based on 1000 Gaussian samples for various ns and contamination rates. LST is computed by lstReg with $\alpha = 3$, LTS was computed by ltsReg.

Inspecting the table reveals some shocking findings (i) in the pure Gaussian sample setting, LST can outperform both LTS and LS in terms of EMSE and SVAR, this is even more striking in the contaminated sample setting; (ii) in terms of speed, no one can run faster than the LS, but LST (R-based) running faster than LTS (Fortran-based) meanwhile with a smaller EMSE and SVAR in all cases considered; (iii) LTS performs unexpectedly disappointing not only always being the slowest but also always having the worst EMSE and SVAR, robust LTS is ironically inferior to non-robust LS for the contaminated samples (RE is just 8.6%).

3.2 A class of penalized regression estimators based on the LST

Definition 3.2 Now that we have a much more robust regression estimator than the LS, which turns out to be more efficient than the LTS. It is quite natural to replace the SSR in (7) by the Q^n defined in (17), and minimize it, subject to two constraints: ℓ_γ-constraint $\sum_{i=1}^p |\beta_i|^\gamma \le t_1$, $t_1 \ge 0$, $\gamma \ge 1$; and ℓ_2-constraint $\sum_{i=1}^p \beta_i^2 \le t_2$, $t_2 \ge 0$, the minimizer is

$$\widehat{\boldsymbol{\beta}}^n_{lst-enet}(\alpha, \lambda_1, \lambda_2, \gamma) := \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Big\{ \frac{1}{n}\sum_{i=1}^n r_i^2 w_i + \lambda_1 \sum_{i=1}^p |\beta_j|^\gamma + \lambda_2 \sum_{i=1}^p \beta_i^2 \Big\}, \tag{20}$$

where $\lambda_i := \lambda(t_i) \ge 0$, $\alpha,\ \gamma \ge 1$, and $w_i := w_i(\boldsymbol{\beta}) := w_i(\boldsymbol{\beta}, \alpha, r_i, \boldsymbol{z}^{(n)}) = \mathbb{1}\left(O(r_i, r^{(n)}) \le \alpha\right)$.

Before studying its robustness, we address existence and uniqueness of $\widehat{\boldsymbol{\beta}}^n_{lst-enet}(\alpha, \lambda_1, \lambda_2, \gamma)$.

Existence and uniqueness

Existence and uniqueness are implicitly assumed for many other penalized regression estimators in the literature. We formally address them for $\widehat{\boldsymbol{\beta}}^n_{lst-enet}(\alpha, \lambda_1, \lambda_2, \gamma)$ below.

Theorem 3.1

(i) $\widehat{\boldsymbol{\beta}}^n_{lst-enet}(\alpha, \lambda_1, \lambda_2, \gamma)$ in (20) always exists;

(ii) $\widehat{\boldsymbol{\beta}}^n_{lst-enet}(\alpha, \lambda_1, \lambda_2, \gamma)$ in (20) is unique provided that (a) $\lambda_1 > 0$ and $\gamma > 1$ or (b) $\lambda_2 > 0$.

Proof: see the Appendix. □

The proof of above theorem needs the following result.

Lemma 3.2 Let $S \subset \mathbb{R}^p$ be an open set and $f(\boldsymbol{x})$: $\mathbb{R}^p \to \mathbb{R}^1$ be strictly convex over S and continuous over $\overline{S}$ (the closure of S). Let $\boldsymbol{x}^*$ be the global minimum of $f(\boldsymbol{x})$ over S and $\boldsymbol{y}$ be a point on the boundary of S, then $f(\boldsymbol{y}) > f(\boldsymbol{x}^*)$.

Proof: see the Appendix. □

Remarks 3.1

(i) Note that $\widehat{\boldsymbol{\beta}}^n_{lst-enet}(\alpha, 0, 0, \gamma) = \widehat{\boldsymbol{\beta}}^n_{lst}$. A sufficient condition for its uniqueness is $\boldsymbol{C}_n := \boldsymbol{X}_n' \mathrm{diag}(w_1, \cdots, w_n)\boldsymbol{X}_n/n$ being invertible. That is, the rank of $\boldsymbol{X}_n$ and the matrix formed by any its $k := \sum_i w_i$ sub-rows is p (see [61]). However, in many applied data set cases, the number of variables p is even larger than the number of observations n, we must have rank $< p$. So it might not be unique. However, if (a) or (b) in (ii) of the theorem holds, then the strictly convexity guarantees the uniqueness of $\widehat{\boldsymbol{\beta}}^n_{lst-enet}(\alpha, \lambda_1, \lambda_2, \gamma)$.

(ii) The uniqueness of $\widehat{\boldsymbol{\beta}}^n_{lst-enet}(\infty, \lambda_1, 0, \gamma)$ (here $\alpha = \infty$ is in the sense that $\alpha \to \infty$), that is, the uniqueness of $\widehat{\boldsymbol{\beta}}_{bridge}(\lambda_1, \gamma)$ has been intensively discussed in the literature, see

e.g., in the Theorems 1 and 2 of [16], it was shown that $\widehat{\boldsymbol{\beta}}_{bridge}(\lambda_1, \gamma)$ is unique if $\lambda_1 > 0$ and $\gamma > 1$ plus some condition on the Hessian matrix of SSR; in their Lemma 2 ($\gamma = 1$) of [55], authors showed that it is not unique when there is repeated row of $\boldsymbol{X}_n$, it is unique with probability one under the some assumption on predictor variables, argued in [48] and in [1] ($\gamma = 1$). Also see section 2.6 of [19]. □

The most relevant question now is: Is $\widehat{\boldsymbol{\beta}}^n_{lst-enet}(\alpha, \lambda_1, \lambda_2, \gamma)$ much more robust than the existing ones? Or rather, what is its RBP? The next result covers both the LST and the LTS based regularized estimators and provides an affirmative answer to the question.

Theorem 3.2 Let $\widehat{\boldsymbol{\beta}}(\lambda_1, \lambda_2, \gamma, \boldsymbol{z}^{(n)})$ be the penalized regression estimator which minimizes the objective function

$$Q(\boldsymbol{\beta}, \lambda_1, \lambda_2, \gamma, z^{(n)}) := \frac{1}{n}\sum_{i=1}^{n} r_i^2 w_i + \lambda_1 \sum_{i=1}^{p} |\beta_i|^\gamma + \lambda_2 \sum_{i=1}^{p} \beta_i^2, \tag{21}$$

where $w_i \in \{0, 1\}$ is an indicator function so that $w_i = \mathbb{1}(r_i^2 \le r^2_{h:n})$ or $w_i = \mathbb{1}(O(r_i, r^{(n)}) \le \alpha)$ and constant $k = \sum_{i=1}^n w_i$ ($\lceil n/2 \rceil \le k \le n$), $\lambda_i \ge 0$, and $\lambda_1 + \lambda_2 > 0$, $1 \le \gamma \le 2$. Then

$$\mathrm{RBP}(\widehat{\boldsymbol{\beta}}(\lambda_1, \lambda_2, \gamma, \boldsymbol{z}^{(n)}), \boldsymbol{z}^{(n)}) = (n - k + 1)/n.$$

Proof: see the Appendix. □

Remarks 3.2

(i) The theorem covers the special case, the main result (Theorem 1) of [2] for the $\widehat{\boldsymbol{\beta}}^n_{lts-lasso}$, where the loss is limited to the sum of smallest h ordered squared residuals and $k = h$, $\gamma = 1$, and $\lambda_2 = 0$. The theorem covers the RBP of $\widehat{\boldsymbol{\beta}}^n_{lst-enet}(\alpha, \lambda_1, \lambda_2, \gamma)$ in (19) for any $\alpha \ge 1$, the RBP reaches its highest value $(\lfloor n/2 \rfloor + 1)/n$ when $\alpha = 1$ ($k = \lfloor (n+1)/2 \rfloor$ in this case). It also for the first time provides the RBP for the reweighted sparse-LTS (or enet-LTS) in [2] (or [27]) with $k = n_w$ there. The theorem tells that ridge, bridge, lasso, and enet all have the lowest RBP $1/n$. The square loss in (21) can be extended to be the more general loss $\mathcal{L}$ in Theorem 2.1.

(ii) Note that the RBP result can be higher than $(\lfloor (n-p)/2 \rfloor + 1)/n$, the latter is the upper bound for any *regression equivariant* estimator (see Theorem 4 on page 125 of [38]). The main reason for this is that the estimator violates the regression equivariance. Without the regression equivariance, any constant vector will have the best possible RBP (100%), but it is not a good estimator at all. Note that the RBP definition in [2] and [26], promoting a constant vector as the most robust estimator, is different from the traditional one. □

Equivariance

Among *regression, scale, and affine equivariance*, the three desired properties (discussed in Sec.3), the regression equivariance is the most fundamental (see [58]), it demands that if one shifts response variable y up and down, then the regression line (or hyperplane) should

shift accordingly up and down. The LS estimator and all its robust alternatives mentioned so far satisfy the three properties. But this is not the case for most of regularized regression estimators. In fact,

Theorem 3.3 Among three equivariant properties, only scale equivariance is posessed by $\widehat{\boldsymbol{\beta}}_{ridge}$ in (4), the $\widehat{\boldsymbol{\beta}}_{sqrt-lasso}$ in (8), the $\widehat{\boldsymbol{\beta}}_{lad-lasso}$ in (10), and $\widehat{\boldsymbol{\beta}}^n_{lts-enet}(\alpha, 0, \lambda_2, \gamma)$ in (20) among all penalized regression estimators discussed previously.

Proof: Scale equivariance of the $\widehat{\boldsymbol{\beta}}_{ridge}$, the $\widehat{\boldsymbol{\beta}}_{sqrt-lasso}$, the $\widehat{\boldsymbol{\beta}}_{lad-lasso}$, the $\widehat{\boldsymbol{\beta}}_{lad-lasso}$, and $\widehat{\boldsymbol{\beta}}^n_{lts-enet}(\alpha, 0, \lambda_2, \gamma)$ is trivial verification in light of (4), (8), (10), and (20). For other properties and penalized estimators, it suffices to show that regression equivariance is violated.

When y_i is shifted to $y_i + \boldsymbol{w}_i'\mathbf{b}$, if the regression coefficients $\boldsymbol{\beta}$ is also shifted to $\boldsymbol{\beta} + \boldsymbol{b}$, then SSR is invariant whereas the constraint or penalty still on $\boldsymbol{\beta}$. □

Remarks 3.3

(i) There has been an abundance of theoretical and computational work on the generalized lasso and its variants and its special cases. Among hundreds, if not thousands, publications on penalized regression in the literature, very few addressed equivariance. Exceptions are [34], and [26]. It is admitted in [34] that their shooting S-estimator fails to meet the regression equivariance. Via transformation and re-transformation [26] asserted that their estimator enjoys the three equivariance properties.

(ii) Standardizing y and $\boldsymbol{x}$ columns are common practice in the literature for many computational algorithms for regularized estimators. This, however, amounts to assuming implicitly that these estimators meet the three equivariance properties. Furthermore, centering the observations of y and $\boldsymbol{x}$ might spread the contamination or outlyingness. □

4 Computation algorithm

Re-parametrizations

(i) Following the notation used in (31), we note that the objective function on the RHS of (20) can be written as (also see the proof of Lemma 4.1)

$$O_n(\boldsymbol{\beta}, \lambda_1, \lambda_2, \gamma) = \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2_{D(\boldsymbol{\beta})} + \lambda_1 \sum_{j=1}^{p} |\beta_j|^\gamma + \lambda_2 \|\boldsymbol{\beta}\|_2^2,$$

where $D(\boldsymbol{\beta}) = \mathrm{diag}(w_1, \cdots, w_n)$ with w_i defined in (20). Now for every $\lambda_2 > 0$, if we write $\boldsymbol{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-1/2}(\boldsymbol{X}'_{n\times p}, \sqrt{\lambda_2}\boldsymbol{I}_{p\times p})'$, $Y^*_{(n+p)\times 1} = (\boldsymbol{Y}'_{n\times 1}, \mathbf{0}'_{p\times 1})'$, $\boldsymbol{\beta}^*_{p\times 1} = (1+\lambda_2)^{1/2}\boldsymbol{\beta}$. Then $D^*(\boldsymbol{\beta}^*)_{(n+p)\times p} := (D(\boldsymbol{\beta^*}), \boldsymbol{I}_{p\times p})' = (D(\boldsymbol{\beta}), \boldsymbol{I}_{p\times p})'$. If let $\lambda_1^* := \lambda_1/(1+\lambda_2)^{1/2}$, we have

$$O_n(\boldsymbol{\beta}^*, \lambda_1, \lambda_2, \gamma) = O_n(\boldsymbol{\beta}, \lambda_1, \lambda_2, \gamma) = \frac{1}{n}\|\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*\|^2_{D^*(\boldsymbol{\beta}^*)} + \lambda_1^* \sum_{j=1}^{p} |\beta_j^*|^\gamma,$$

An ℓ_1-type penalized regression with an objective function much resembling that of a lasso-type problem (especially when in the $\gamma = 1$ case). Denote the minimizer of the objective

function above by $\widehat{\boldsymbol{\beta}}^*$. It can be computed via the approach for lasso such as the lars algorithm of [10].

(ii) Alternatively, if we set $\lambda^* = \lambda_1 + \lambda_2$ and $\eta^* = \lambda_2/(\lambda_1 + \lambda_2)$ (note that $\lambda_1 + \lambda_2 > 0$, otherwise we have a non-penalized problem addressed in [61]), then we have

$$O_n(\boldsymbol{\beta}, \lambda_1, \lambda_2, \gamma) = O_n(\boldsymbol{\beta}, \eta^*, \lambda^*, \gamma) := \frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2_{D(\boldsymbol{\beta})} + \lambda^*\Big((1-\eta^*)\sum_{j=1}^{p}|\beta_j|^\gamma + \eta^*\|\boldsymbol{\beta}\|_2^2\Big). \quad (22)$$

Note that $\eta^* \in [0, 1)$ (a pure ridge regression case is excluded) and $\lambda^* \in (0, \lambda_0]$ for some λ_0 (which is set to be $\max_{1\le j\le p}|2\boldsymbol{Y}'\boldsymbol{x}^{(j)}|/n$ as in the literature, see e.g., [2] and Section 2.12 of [6]). *Boundedness of parameters* is the advantage of this formulation. For a given data set $\boldsymbol{z}^{(n)} = \{(\boldsymbol{x}_i', y_i)', i \in \{1, \cdots, n\}\}$, we now present the outline our approximate algorithm (AA) for $\widehat{\boldsymbol{\beta}}^n_{lst-enet}$.

Pseudocode for computing $\widehat{\boldsymbol{\beta}}^n_{lst-enet}$ (lst-enet)

(1) Sample two indices $\{i, j\}$ (two points) and obtain at least p $\boldsymbol{\beta}$s using the algorithm AA1 in [61]: denoted them by $\boldsymbol{\beta}^k$ ($k \in \{1, \cdots, p\}$) and obtain index sets $I(\boldsymbol{\beta}^k) := \{i : w_i := w_i(\boldsymbol{\beta}^k) = 1\}$.

(2) For each $\boldsymbol{\beta}^k$, employing the strategy below select a pair (η^*, λ^*) with respect to sub-data sets $(D(\boldsymbol{\beta}^k)\boldsymbol{X}, D(\boldsymbol{\beta}^k)\boldsymbol{Y})$.

(3) Based on the sub-data sets $(D(\boldsymbol{\beta}^k)\boldsymbol{X}, D(\boldsymbol{\beta}^k)\boldsymbol{Y})$ obtain solution $\widehat{\boldsymbol{\beta}}^k$ via LARS algorithm (limited the total steps to 900)

(4) Evaluate $O_n(\boldsymbol{\beta}, \eta^*, \lambda^*, \gamma)$ with respect to $\boldsymbol{\beta}^k$ and $\widehat{\boldsymbol{\beta}}^k$ ($k \in \{1, \cdots, p\}$). Update $\widehat{\boldsymbol{\beta}}^n_{lst-enet}$ (initially it is a $\mathbf{0}$ vector) to be the one that minimizes

$$O_n(\boldsymbol{\beta}, \eta^*, \lambda^*, \gamma) = \frac{1}{n}\sum_{i\in I(\boldsymbol{\beta})} w_i r_i^2(\boldsymbol{\beta}) + \lambda^*\Big((1-\eta^*)\sum_{i=1}^{p}|\beta_j|^\gamma + \eta^*\|\boldsymbol{\beta}\|_2^2\Big).$$

(5) Repeat **(1)-(4)** 50 times and output the one that has the minimum objective value.

In algorithm above, (η^*, λ^*) is assumed to be selected. Now we address the issue how to choose this pair. Obviously, we can search among a finite grids over the region $[0, 1) \times (0, \lambda_0]$.

Choice of the penalty/tuning parameters via cross-validation.

We first pick a (relatively small) grid of values for λ^*, say from 0 (excluded) to λ_0 with $\lambda_0/10$ as the step so that there are 10 equal spaced grid points. For the estimation of λ_0, one can see [10] and [2], or **(ii)** of Re-parametrization above.

For each λ^*, we will select an η value among 10 equal spaced grid points over $[0, 1)$ via five-fold cross-validation (CV). In k-fold cross-validation, the data are split randomly in k blocks (folds) of approximately equal size. Each block is left out once to fit the model, and the left-out block is used as test data (see Section 7.10.1 of [18]).

The CV is a popular method for estimating the prediction error and comparing different models (see [55] and [18]). The popular **R** package *glmnet* can be used to select the parameters as did in [27], which automatically checks the model quality for a sequence of values for η, taking the mean squared error as an evaluation criterion.

We use a 5-fold CV via our own developed program to avoid the drawback of the glmnet which often leads to the error message "from glmnet C++ code (error code 7777); All used predictors have zero variance" (this especially is true under the adversary contamination scenario). The latter leads to a problem for evaluating the performance of the procedure enetLTS of [27] when the contamination at 10% level in next section. We have to drop enetLTS in the comparison in that situation. We will run ten times of our 5-fold CV, then the pair (λ^*, η) with the minimum averaged CV error will be the final chosen pair (λ^*, η^*).

The lars algorithm can be used to fit a linear model based on the $k-1$ blocks to obtain a $\widehat{\boldsymbol{\beta}}(\lambda^*, \eta)$. Other algorithms, such as coordinate descent algorithms (including Fu's shooting algorithm ([16]) (see 2,11.1 of [6]) can be employed to speed up the computation.

5 Illustration examples and comparison

5.1 Simulation

All R code for simulation and examples as well as figures in this article (downloadable via https://github.com/left-github-4-codes/lst-enet) were run on a desktop Intel(R)Core(TM) 21 i7-2600 CPU @ 3.40 GHz.

Five regularized regression procedures We like to compare the performance of our procedure lst-enet with leading regularized regression procedures including lasso, lars, enet, and enetLTS. lasso will be computed via **R** package "lars", it can be obtained via "elasticnet". The latter package is mainly for the enet whereas the former mainly focuses on lars. Though lasso could be obtained via "glmnet" but due to the contamination scenario, the glmnet often does not work. Unfortunately, enetLTS employing glmnet in its CV calculation, it can not hand the model $y = \boldsymbol{w}'\boldsymbol{\beta}_0 + e$ appeared in (1) (an error message "glmnet fails at standardization step"). We use an alternative model given below

Simulation designs To copy with the situation above, we simulate data from the true model: $y = \boldsymbol{X}\boldsymbol{\beta}_0 + \sigma e, \quad e \sim N(0,1)$, where the true unknown parameter $\boldsymbol{\beta}_0$ is assumed to be a p-dimensional vector with the first $p_1 := \lceil 6\% * p \rceil$ components are threes and the rest $p_2 := p - p_1$ components are zeros. σ is set to be 0.5 but could be changed to other values (leading to different signal-to-noise ratio).

Design I: take sample of $\boldsymbol{X}$ from $N(\boldsymbol{0}, \sigma \boldsymbol{I}_{p\times p})$ and e from $N(0,1)$. **Design II**: take sample from $\boldsymbol{X} \sim N\big(\boldsymbol{0}, \boldsymbol{\Sigma}\big)$ with $\boldsymbol{\Sigma}(i,j) = \rho_1^{|i-j|}, 1 \le i,j \le p_1$, $\boldsymbol{\Sigma}(i,j) = \rho_2^{|i-j|}, p_1 < i,j \le p$, $\rho_1 = 0.95$, $\rho_2 = 0.05$, all other entries of $\boldsymbol{\Sigma}$ are zeros and $e \sim N(0,1)$. We take $n \in \{50, 100\}$ samples from the $\boldsymbol{X}$ and e above and calculate the response $y_i = \boldsymbol{X}_i\boldsymbol{\beta}_0 + \sigma e_i, i \in \{1, 2, \cdots, n\}$.

Contamination levels and schemes Let ε be the contamination level, when $\varepsilon = 0$ there is no contamination, an ideal situation (and not realistic). Consider the scenario

$\varepsilon \in \{0, 0.05, 0.1, 0.2\}$ (i.e., 0%, 5%, 10%, 20% contamination). Let $m = \lfloor \varepsilon * n \rfloor$, sample m indices from $\{1, \cdots, n\}$.

Contamination **Scheme I**: add 20 to the corresponding m components of $(e_1, \cdots, e_n)$, compute $y_i = \boldsymbol{X}_i \boldsymbol{\beta}_0 + \sigma e_i, i \in \{1, 2, \cdots, n\}$, and add 20 (component-wise) to the corresponding m rows of $(\boldsymbol{X}_1, \cdots, \boldsymbol{X_n})'$. **Scheme II**: add 20 to the corresponding m components of $(e_1, \cdots, e_n)$, compute $y_i = \boldsymbol{X}_i \boldsymbol{\beta}_0 + \sigma e_i, i \in \{1, 2, \cdots, n\}$. Replace the corresponding m rows of $(\boldsymbol{X}_1, \cdots, \boldsymbol{X_n})'$ by a p-vector with its first component being 10^4 and the rest are zeros, do the same for the corresponding m components of $(y_1, \cdots, y_n)$ but with a scalar 10^{10}.

Four performance criteria The first measure is the estimation error, or L2-error/L2-loss between the true parameter $\boldsymbol{\beta}_0$ and the estimator $\widehat{\boldsymbol{\beta}}_P$ via procedure P and is defined as:

$$\text{L2-error}(\boldsymbol{\beta}_0, \widehat{\boldsymbol{\beta}}_P) := \|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_P\|_2^2, \tag{23}$$

where $\|\boldsymbol{a} - \boldsymbol{b}\|_2$ is the ℓ_2-norm between the two p-dimensional vectors.

On the other hand, one has to take the performance measure into the context of the sparsity model consideration. In the following we introduce the *true sparsity discovery rate* (TSDR) and the *false sparsity discovery rate* (FSDR). For notation simplicity, we denote the unknown parameter by $\boldsymbol{\beta}^0$ (assume it has at least one zero coordinate), an estimator by $\widehat{\boldsymbol{\beta}}^P$.

$$\text{TSDR}(\boldsymbol{\beta}^0, \widehat{\boldsymbol{\beta}}^P) := \frac{\sum_{i=1}^p \mathbb{1}(\beta_i^0 = 0, \hat{\beta}_i^P = 0)}{\sum_{i=1}^p \mathbb{1}(\beta_i^0 = 0)}, \tag{24}$$

namely, the fraction of correctly detecting/discovering the zero coordinates of the true parameter $\boldsymbol{\beta}^0$. The higher the TSDR, the better the $\widehat{\boldsymbol{\beta}}^P$.

$$\text{FSDR}(\boldsymbol{\beta}^0, \widehat{\boldsymbol{\beta}}^P) := \frac{\sum_{i=1}^p \mathbb{1}(\beta_i^0 \neq 0, \hat{\beta}_i^P = 0)}{\sum_{i=1}^p \mathbb{1}(\beta_i^0 \neq 0)}, \tag{25}$$

namely, the fraction of falsely detecting/discovering as zero coordinate for the true parameter $\boldsymbol{\beta}^0$ (assume that it is non zero vector). The lower the FSDR, the better the $\widehat{\boldsymbol{\beta}}^P$.

The fourth performance measure is a popular one, it is (square-)root of mean squared (prediction) error (RMSE) on testing data. That is, for a given data set, one first partitions data into training and testing two parts (we take the ratio 7:3 for partition). Then fit the model and get an estimate based on the training data and using the testing data to get the RMSE. Testing data sets are often assumed to be clean (have no contamination or outliers) in the literature. This, however, is not realistic in practice.

Let $\boldsymbol{X}_{test}$, y_{test} be the testing data and $\widehat{\boldsymbol{\beta}}_P$ be the estimator obtained from the training data. Then

$$\text{RMSE}(\widehat{\boldsymbol{\beta}}_P) := \Big(\text{mean}\big((y_{test} - \boldsymbol{X}_{test}\widehat{\boldsymbol{\beta}}_P)^2\big)\Big)^{1/2}. \tag{26}$$

The four performance measures above were discussed in the literature before, all are hoped to be small except the TSDR which is hoped to be as high as possible. All (but RMSE) depend on the unknown parameter $\boldsymbol{\beta}^0$.

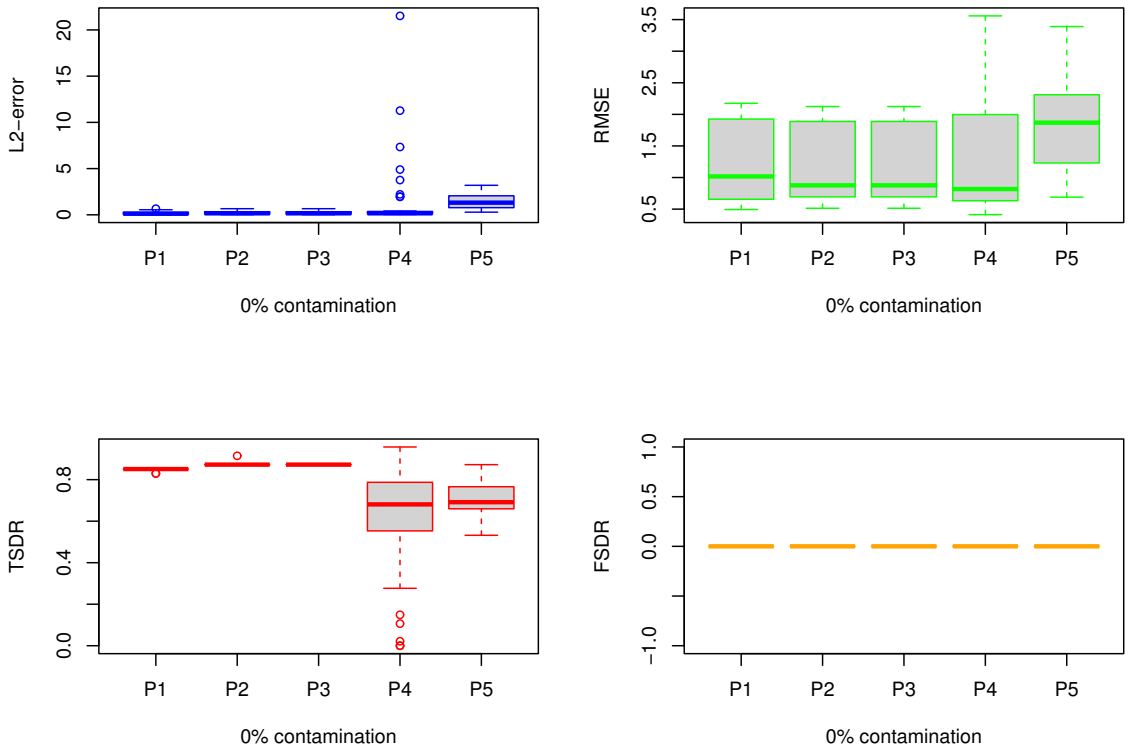

Figure 2: Boxplots for five procedures (P1 stands for lst-enet, P2 for lasso, P3 for lars, P4 for enetLTS, P5 for enet) and 50 samples each with $n = 100$ and $p = 50$ that are generated from design I with 0% contamination rate.

Tuning parameter choices In the following simulation, the tuning parameter α in (17), or (20) is set to be 3 and h in enetLTS is set to be the default value $\lfloor 0.75(n+1) \rfloor$.

Example 5.1 We first consider $\varepsilon = 0$. For simplicity, data are generated according to design **I** and set $n = 100$, $p = 50$ (low dimension case) or $n = 50$, $p = 300$ (high dimension and sparsity case). We generated 50 samples for $\boldsymbol{X}$ and e and obtained corresponding responses y. The simulation results are displayed in Figure 2. For description simplicity, we use hereafter P1 for lst-enet, P2 for lasso, P3 for lars, P4 for enetLTS, P5 for enet in the Figures.

Inspecting Figure 2 reveals that (i) with respect to (w.r.t.) FSDR, all four perform equally well with 0% mis-discovery rate; (ii) w.r.t. TSDR, lst-enet, Lars, and lasso perform stably and at a highest rate while enet with a less stable lower rate but enetLTS performs most unstable with the lowest median rate; (iii)w.r.t. RMSE, lasso and lars are the best followed by lst-enet, enetLST has the median RMSE that is also among the best but with the widest spread of RMSE while enet has the largest (and wider spread of) RMSE; (iv) w.r.t. L2-error, lst-enet, lars, and lasso are among the best while enetLTS has the worst performance followed by enet. Overall, lst-enet, lars, and lasso are among the best whereas enetLTS performs worst overall followed by enet.

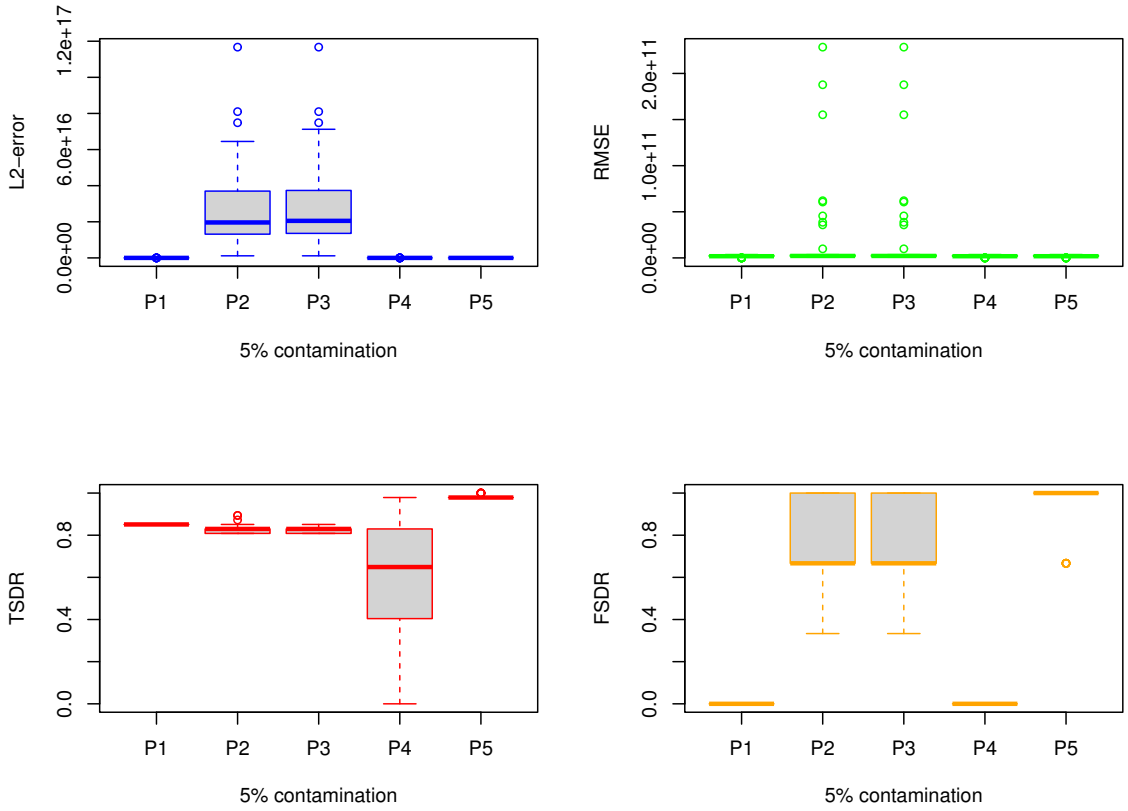

Figure 3: Boxplots for five procedures and 50 samples each with $n = 100$ and $p = 50$ that are generated from design I with 5% contamination rate.

Example 5.2 Perfect normal data are not realistic in practice. We now consider $\varepsilon = 0.05$ (i.e. 5% contamination), all others are the same as Example 6.1 except the contamination scheme II will be adopted (in 6.1 contamination scheme does not matter). We first consider $n = 100$, $p = 50$ (low dimension case) and for simplicity generate data according design **I**. Performance of five procedures in 50 samples is displayed in Figure 3.

Inspecting the Figure reveals that (i) w.r.t. L2-error, lst-enet, enetLTS and enet are the best performers while lasso and lars are equally dissatisfactory; (ii) w.r.t. RMSE, the situation is the same as in the L2-error case; (iii) w.r.t. TSDR, enet is the best performer (this perhaps is false best since it might assign zero to all components of the estimator $\widehat{\boldsymbol{\beta}}$ that could lead to 100% of its FSDR) while enetLTS is the worst and lst-enet is slightly better than lasso and lars; (iv) w.r.t. FSDR, lst-enet, enetLTS are the best performers followed by lasso and lars while enet is the loser. Overall, lst-enet is the only winner.

The simulation study above with 5% contamination is repeated but $n = 50$ and $p = 300$ (high dimensional) and simulation design II is adopted. Results are displayed in Figure 4.

Reviewing the Figure discovered that (i) w.r.t. L2-error, lst-enet, enetLTS, and enet are the best performers while lasso and lars are disappointing; (ii) w.r.t. RMSE; the situation is almost the same as in L2-error case; (iii) w.r.t. TSDR, lst-enet, lasso, lars and enet are the best performers while enetLTS is the loser; (iv) w.r.t. FSDR, enet is the worst performer

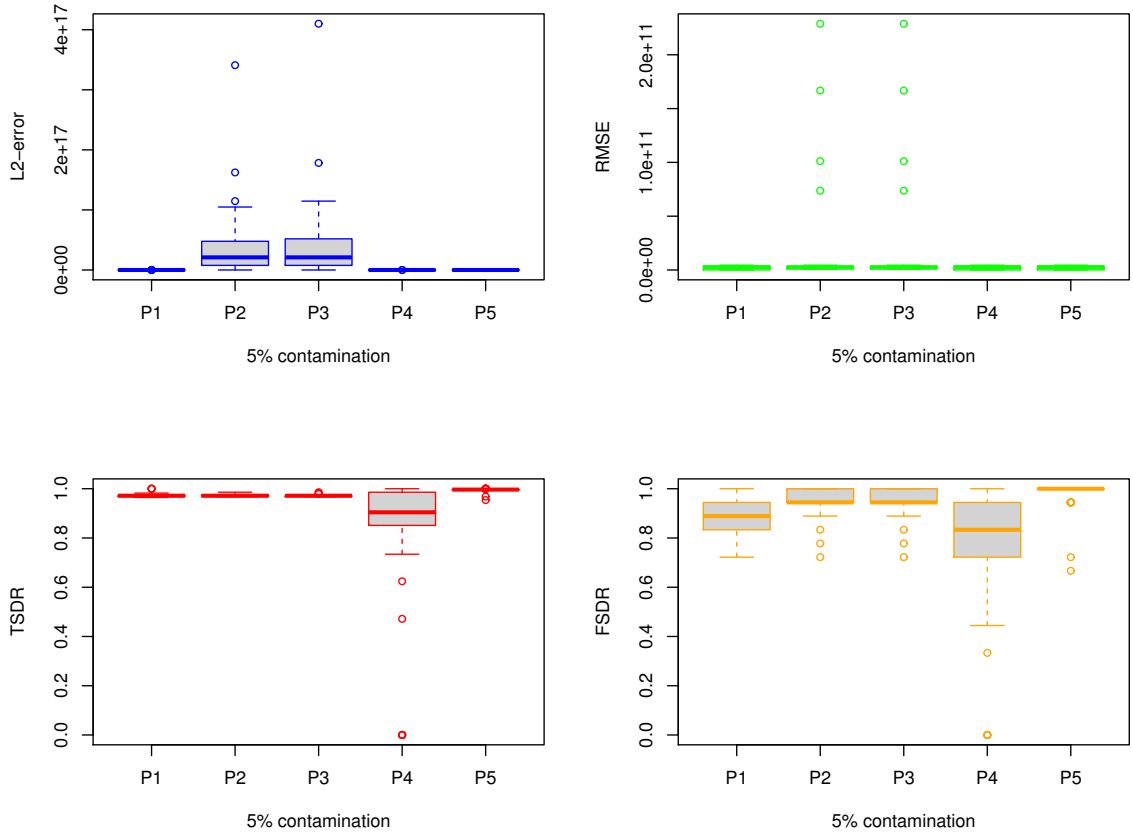

Figure 4: Boxplots for five procedures and 50 samples each with $n = 50$ and $p = 300$ that are generated from design I with 5% contamination rate.

(since its FSDR is almost 100%), enetLTS has the lowest median value while it has the widest spread. lst-enet is the second best performer, lasso and lars are disappointed. Overall, lst-enet is the only winner.

Example 5.3 In practice, 10% (or even 20%) contamination is not rare. Next we consider the case $\varepsilon = 0.1$ (i.e., 10% contamination), contamination scheme II will be adopted. Samples of 50 with $n = 50, p = 300$ are generated with simulation design I. Due to the higher level contamination and the usage of **R** package glmnet in its background CV calculation, enetLTS fails to go through the computation we have to drop it in our comparison. Simulation results are displayed in Figure 5.

Inspecting the Figure reveals that (i) w.r.t. L2-error, lst-enet and enet are the best while lasso and lars are inferior; (ii) w.r.t. RMSE, the situation is the same as in L2-error case; (iii) w.r.t. TSDR, enet is the worst performer (it assigns zeros to almost all components of $\widehat{\boldsymbol{\beta}}$ that will lead to 100% of its FSDR), lst-enet and lasso are the best performers followed by lars; (iv) w.r.t. FSDR, lst-enet is the best performer, enet is the worst one while lasso and lars perform dissatisfactory. Overall, lst-enet is the winner.

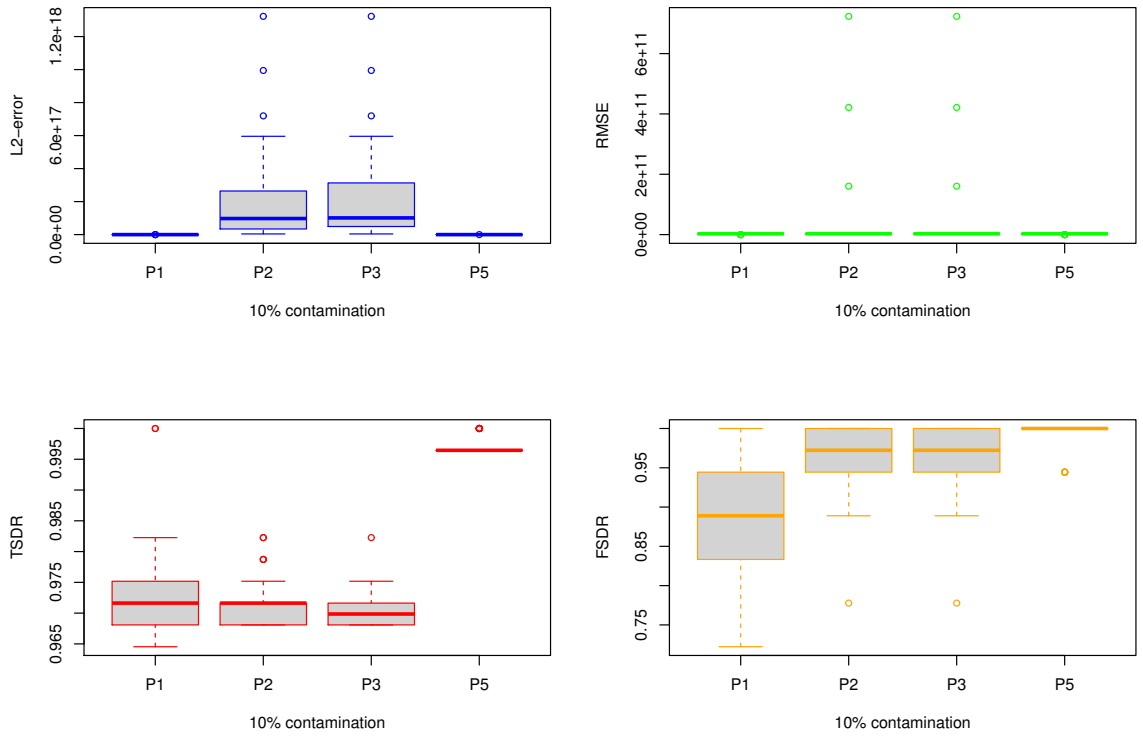

Figure 5: Boxplots for four procedures and 50 samples each with $n = 50$ and $p = 300$ that are generated from design I with 10% contamination rate.

The advantage of lst-enet is even better demonstrated in Figure 6 when $n = 100$ and $p = 50$ and $\varepsilon = 0.2$ (i.e., 20% contamination).

5.2 A real data example

Example 5.4 To analyze a realistic dataset with very large number of variables, we consider the well-known cancer data from the National Cancer Institute (NCI60); see [46] for more detail about this dataset. A total of 59 of the human cancer cell-lines (n= 59) were assayed for gene expression and protein expression. The data set, downloadable from the CellMiner program package, NCI (http://discover.nci.nih.gov/cellminer/) and available from the R package robustHD, has been repeatedly studied in the literature, see e.g., [29].

We process the data set by following the approach in the literature and treat the gene expression microarray data as the predictors $\boldsymbol{X}_{raw}$ (a 59 by 22283 matrix) and the protein expression data as responses variables $\boldsymbol{Y}_{raw}$ (a 59 by 162 matrix). Similar to [29] or [2], we order the protein expression variables according to their scale (employing MAD as a scale estimator instead of the standard deviation) and select the one with median MAD, serving as our dependent variable. It is 75th column of the protein expression data matrix. Denote it by $\boldsymbol{Y}$.

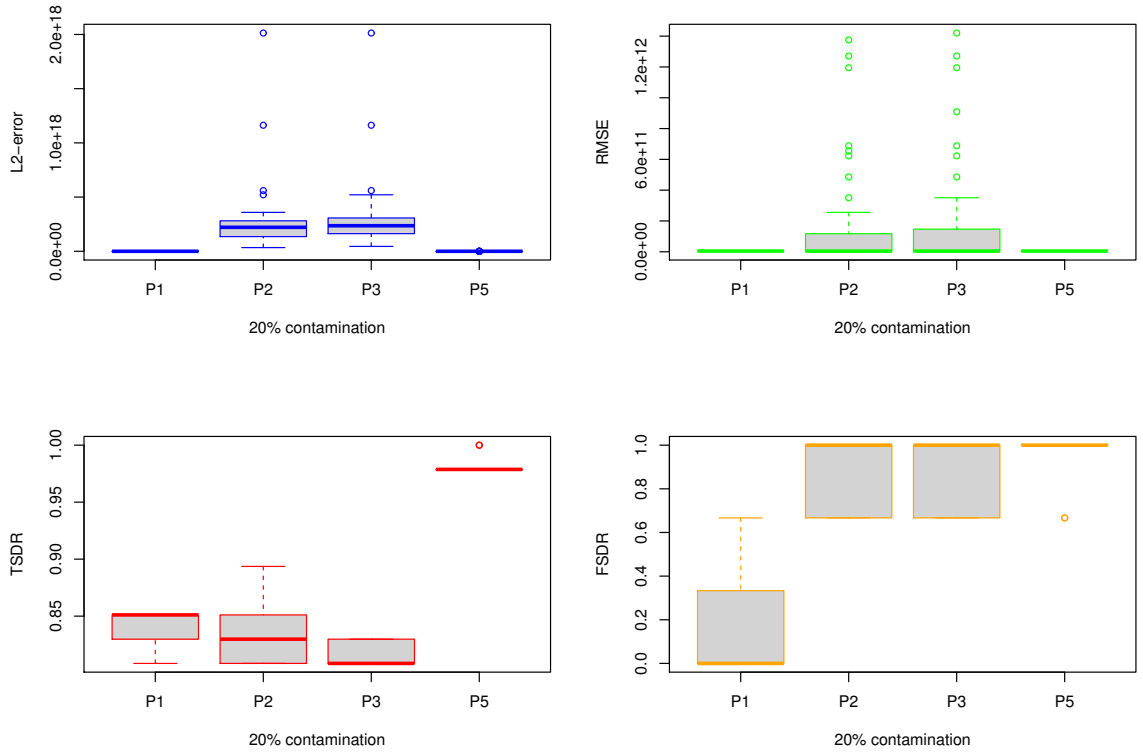

Figure 6: Boxplots for four procedures and 50 samples each with $n = 100$ and $p = 50$ that are generated from design I with 20% contamination rate.

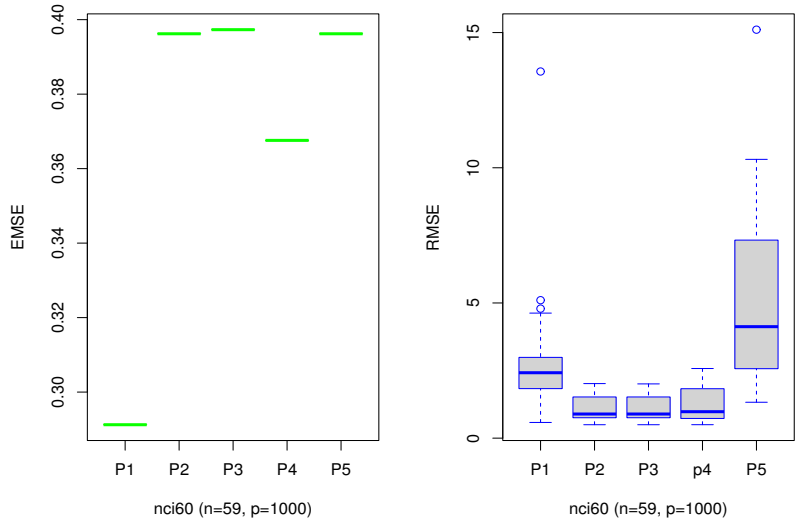

Figure 7: Boxplots for five procedures for real data set NCI60 with 50 times partition of a selected sub-dataset $p = 1000$ and $n = 59$.

Next, we selected out genes using their correlations with $\boldsymbol{Y}$. Here we adopt the robust correlation measure in [25]. We obtain 22283 ordered (decreasing) correlations and select top $k_1 = 100$ corresponding columns of $\boldsymbol{X}_{raw}$ and combined with the bottom $1000 - k_1$ columns as our final $\boldsymbol{X}$, reducing the number of genes from 2,2283 to $p = 1,000$. The number p could easily be changed by adjusting k_1.

We partition (by rows) $\boldsymbol{X}_{59\times 1000}$ and $\boldsymbol{Y}$ into $x.train$, $y.train$ and $x.test$, $y.test$ according the rate 7 : 3. That is 41 rows of $\boldsymbol{X}$ and $\boldsymbol{Y}$ for the training data sets, the rest 18 rows as testing data sets. We do this step 50 times and each time we calculate the RMSE (the only measure that still valid without the given $\boldsymbol{\beta}_0$) for the five procedures. The results are displayed in Figure 7, where the fifth performance measure is introduced, that is, the empirical mean squared error, defined as:

$$\text{EMSE}(\widehat{\boldsymbol{\beta}}_P) := \frac{1}{R}\sum_{i=1}^{R}\big(\widehat{\boldsymbol{\beta}}_P^i - \overline{\widehat{\boldsymbol{\beta}}}_P^i\big)^2, \tag{27}$$

where R is the replication number, $\widehat{\boldsymbol{\beta}}_P^i$ is $\widehat{\boldsymbol{\beta}}_P$ induced from ith sample, and $\overline{\boldsymbol{x}}$ stands for the sample mean of $\boldsymbol{x}_i$. It is readily seen that $\frac{R}{R-1}\text{EMSE}(\widehat{\boldsymbol{\beta}}_P)$ is the sample variance of $\widehat{\boldsymbol{\beta}}_P$.

Inspecting the Figure reveals that (i) lasso, lars, enetLTS (enetLTS has the wider spread) have the smallest RMSE but their sample variances (EMSE) are among the largest; (ii) RMSE of the lst-enet is the second smallest but it is the most stable estimator with the distinguished smallest sample variance (EMSE) which means that with different training and testing data sets obtained by random partitioning, lst-enet produces very closed solutions; (iii) enetLTS has a low median RMSE but its sample variance is the remarkably large; (iv) enet has the categorical largest RMSR while its sample variance is also the largest. Overall, lst-enet is recommended with the rivals enetLTS, lasso and lars.

6 Concluding discussions

The originality of the current work (i) For the first time it discovered that most of the leading penalized regression estimators for high-dimensional sparse data can break down by a single outlier (or contaminating point) (Theorem 2.1). (ii) For the first time it demonstrated that the newly proposed lst-enet estimator not only possesses a high breakdown robustness (Theorem 3.2) but also performs well in simulation studies and a read data example (Section 6), serving as a robust alternative to regularized regression estimators. (iii) the newly proposed lst-enet estimator, its unique existence (Theorem 3.1) is original. (iv) The equivarance discussion and result (Theorem 3.3) are novel and original, and so is the (ii) of Remarks 3.3. (v) Figure 1 and Table 1 provide brand new evidence of the superiority of LST over LTS.

Connection to the existing work (i) It adduced the procedure LST introduced in [61] and cited results from [61] in Section 3.1. (ii) The idea of replacing the square loss of LS in lasso by a loss of robust alternative of LS (e.g. the LTS) appeared in the literature before (see, e.g. [2] and [27]). (iii) A RBP result (Theorem 1, the main result of [2]) intended for $\widehat{\boldsymbol{\beta}}^n_{lts-lasso}$ appeared before. However, our Theorem 3.2 covers their Theorem 1 in the sense: (a) the loss function in Theorem 3.2, being more general, covers the one in Theorem 1 of [2],

it is not vice versa since the loss in [2] involving the sum of h ordered squared residuals and $\lambda_2 = 0$, $k = h$, and $\gamma = 1$; (b) the penalty terms in Theorem 3.2 are more general and cover the one in Theorem 1, it is not vice versa.

Robust measure Finite sample breakdown point has been served as a prevailing quantitative robustness measure in finite sample practice, the main advantage/beauty is its non-randomness and probability-free nature that is exactly why it was enthusiastically welcomed and quick became adopted in a broad spectrum of disciplines after its introduction in 1983.

Critics (e.g., [41]), however, would like to have a more complicated version, a version that includes randomness and Orlicz norm. They argued that worst case performance might not be a good robustness measure. On the other hand, it is common practice to use the worst case performance as in the complexity of an algorithm or the safety of a passenger cars case.

Future possible work (a) Further performance measure could be pursued including finite sample prediction and estimation error bounds, i.e. **(i)** whether $\hat{\boldsymbol{\beta}}$ performs well on future samples (i.e., whether $E\|\boldsymbol{Y} - \boldsymbol{X}'\hat{\boldsymbol{\beta}}\|^2$ is small, or weather $\|\boldsymbol{Y} - \boldsymbol{X}'\hat{\boldsymbol{\beta}}\|^2$ is small with a high probability); **(ii)** whether $\hat{\boldsymbol{\beta}}$ closely approximates the "true" parameter $\boldsymbol{\beta}_0$ (i.e., whether $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|$ is small with a high probability); or **(iii)** whether $\hat{\boldsymbol{\beta}}$ correctly identifies the relevant coordinates of the "true," sparse parameter $\boldsymbol{\beta}_0$ (i.e., whether $(\boldsymbol{\beta}_{0j} = 0) \Leftrightarrow (\hat{\boldsymbol{\beta}}_j = 0)$ with a high probability). (b) Extension of current regression work to a more general setting to cover discriminant analysis, logistic regression, and other topics.

Acknowledgments

The authors thank Prof. Wei Shao for insightful comments and stimulating discussions which significantly improved the manuscript.

References

[1] Ali, A. and Tibshirani, R. J. (2019), "The Generalized Lasso Problem and Uniqueness", *Electron. J. Statist.*, 13(2): 2307-2347.

[2] Alfons, A., Croux, C., and Gelper, S. (2013), "Sparse least trimmed squares regression for analyzing high-dimensional large data sets", *Ann. Appl. Stat.*, 7(1): 226-248.

[3] Belloni, A., Chernozhukov, V., and Wang, L. (2011). "Square-root lasso: Pivotal recovery of sparse signals via conic programming". *Biometrika*, 98(4):791–806.

[4] Bianco, A. M., Boente, G., Chebi, G. (2022), "Penalized robust estimators in sparse logistic regression", *TEST* (2022), 31:563–594 https://doi.org/10.1007/s11749-021-00792-w.

[5] Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). "SLOPE-adaptive variable selection via convex optimization". *Ann. Appl. Stat.* 9(3): 1103-1140, DOI: 10.1214/15-AOAS842.

[6] Bühlmann, P. and Van De Geer, S. (2011), Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer-Verlag Berlin Heidelberg

[7] Donoho, D. L. "Breakdown properties of multivariate location estimators". PhD Qualifying paper, Harvard Univ. (1982).

[8] Donoho, D. L., and Gasko, M. (1992), "Breakdown properties of multivariate location parameters and dispersion matrices", *Ann. Statist.*, **20**, 1803-1827.

[9] Donoho, D. L., and Huber, P. J. (1983), "The notion of breakdown point", in: P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds. *A Festschrift foe Erich L. Lehmann* (Wadsworth, Belmont, CA) pp. 157-184.

[10] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), "Least angle regression". *Ann. Statist.*, 32 407–499

[11] Emmert-Streib, F. and Dehmer, M. (2019), "High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection", *Mach. Learn. Knowl. Extr.*, 1, 359–383; doi:10.3390/make1010021

[12] Fan, J. (1997). Comments on ≪wavelets in statistics: A review≫ by A. Antoniadis. *Journal of the Italian Statistical Society*, 6(2):131

[13] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* 96 1348–1360.

[14] Frank, I.E. and Friedman, J.H. (1993), "A statistical view of some chemometrics regression tools (with discussion)", *Technometrics*, 35: 109–148.

[15] Freijeiro-González, L., Febrero-Bande, M., and González-Manteiga, W. (2022), "A critical review of LASSO and its derivatives for variable selection under dependence among covariates", *International Statistical Review*, 90(1), 118–145 doi: 10.1111/insr.12469

[16] Fu, W.J. (1998), "Penalized regression: The Bridge versus the Lasso", *Journal of Computational and Graphical Statistics*, 7(3): 397–416.

[17] Hastie, T. and Efron, B. (2011), "lars: Least angle regression, lasso and forward stagewise". R package version 0.9-8.

[18] Hastie, T., Tibshirani, R., and Friedman, J. (2017), The Elements of Statistical Learning Data Mining, Inference, and Prediction, Springer Series in Statistics.

[19] Hastie, T., Tibshirani, R., and Wainwright, M. J. (2015), Statistical Learning With Sparsity: The Lasso and Generalizations, Boca Raton, FL: CRC Press. [254]

[20] Hesterberg, T., Choi,N. H., Meier, L., and Fraley, C. (2008), "Least angle and ℓ_1 penalized regression: A review", *Statistics Surveys* Vol. 2 61–93 ISSN: 1935-7516 DOI: 10.1214/08-SS035

[21] Hoerl, A.E., and Kennard, R.W. (1970a), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67.

[22] Hoerl, A.E., and Kennard, R.W. (1970b), "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics*, 12, 69–82.

[23] Huber, P. J. (1981). Robust Statistics. Wiley, New York.

[24] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), An Introduction to Statistical Learning, with Applications in R. Springer New York Heidelberg Dordrecht London

[25] Khan, J. A., Van Aelst, S. and Zamar, R. H. (2007), "Robust linear model selection based on least angle regression". *J. Amer. Statist. Assoc.*, 102 1289-1299.

[26] Kong, D., Bondell, H.D., and Wu, Y. (2018), "Fully efficient robust estimation, outlier detection and variable selection via penalized regression", *Statistica Sinica*, 28 (2018), 1031-1052.

[27] Kurnaz FS, Hoffmann I, Filzmoser P (2018), "Robust and sparse estimation methods for high-dimensional linear and logistic regression", *Chemomet Intell Lab Syst*, 172:211–222,

[28] Laurent, B. and Massart, P. (2000), "Adaptive estimation of a quadratic functional by model selection", *Ann. Statist.*, 28(5): 1302-1338, DOI: 10.1214/aos/1015957395

[29] Lee, D., Lee, W., Lee, Y. and Pawitan, Y. (2011). "Sparse partial least-squares regression and its applications to high-throughput data analysis". *Chemometrics and Intelligent Laboratory Systems*, 109 1–8.

[30] Li, G., Peng, H. and Zhu, L. (2011), "Nonconcave penalized M-estimation with a diverging number of parameters". *Statistica Sinica*, 21 391–419.

[31] Maglott, D., Ostell, J., Pruitt, K. D. and Tatusova, T. (2005), "Entrez Gene: genecentered information at NCBI". *Nucleic Acids Research*, 33 D54–D58.

[32] Maronna, R. A. (2011), "Robust ridge regression for high-dimensional data. *Technometrics*, 53 44–53.",

[33] Maronna, R. A., Martin, R. D., and Yohai, V. J.(2006), " Robust Statistics: Theory and Methods", John Wiley &Sons

[34] Öllerer, V., Alfons A., and Croux, C. (2016), "The shooting S-estimator for robust regression", Comput Stat (2016) 31:829–844.

[35] Rosset, S. and Zhu, J. (2004), "Discussion of "Least angle regression" by Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.", *Ann. Statist.*, 32 469–475.

[36] Rousseeuw, P. J. (1984), "Least median of squares regression". *J. Amer. Statist. Assoc.*, 79, 871-880.

[37] Rousseeuw, P. and Yohai, V. (1984), "Robust regression by means of S-estimators". In Robust and Nonlinear Time Series Analysis (Heidelberg, 1983), vol. 26 of Lecture Notes in Statist. New York: Springer, pp. 256-272.

[38] Rousseeuw, P. J. and Leroy, A. M. (1987). Robust Regression and Outlier Detection. New York, NY, USA: John Wiley & Sons, Inc.

[39] Rousseeuw, P. J. and Van Driessen, K. (2006), "Computing LTS Regression for Large Data Sets", *Data Mining and Knowledge Discovery* 12, 29-45.

[40] Uma T Shankavaram, William C Reinhold, Satoshi Nishizuka, Sylvia Major, Daisaku Morita, Krishna K Chary, Mark A Reimers, Uwe Scherf, Ari Kahn, Douglas Dolginow, Jeffrey Cossman, Eric P Kaldjian, Dominic A Scudiero, Emanuel Petricoin, Lance Liotta, Jae K Lee, John N Weinstein (2007), "Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study", *Mol Cancer Ther* . 2007 Mar; 6(3):820-32.

[41] She, Y., Wang, Z., and Shen, J. (2022), "Gaining Outlier Resistance With Progressive Quantiles: Fast Algorithms andTheoretical Studies", *J. Amer. Statist. Assoc.*, 2022, 117(539), 1282–1295.

[42] Shao, J. (2003), Mathematical Statistics, Springer Texts in Statistics.

[43] Stahel, W. A. (1981), Robuste Schatzungen: Infinitesimale Optimalitiit und Schiitzungen von Kovarianzmatrizen. Ph.D. dissertation, ETH, Zurich.

[44] Stromberg, A. J., Hawkins, D. M., and Hössjer, O. (2000), "The Least Trimmed Differences Regression Estimator and Alternatives", *J. Amer. Statist. Assoc.*, 95, 853-864.

[45] Sun, Q., Zhou, W.X., and Fan, J.Q. (2020), "Adaptive Huber Regression", *J. Amer. Statist. Assoc.*, 115:529, 254-265, DOI: 10.1080/01621459.2018.1543124

[46] Tan, C.S., Salim, A., Ploner, A., Lehtio, J., Chia, K.S., Pawitan, Y., (2009) "Correlating gene and protein expression data using correlated factor analysis", *BMC Bioinformatics* 10 272, doi:10.1186/1471-2105-10-272.

[47] Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society. Series B.*, 58, 267-288.

[48] Tibshirani, R. J. (2013), "The lasso problem and uniqueness", *Electronic Journal of Statistics*, Vol. 7, 1456–1490.

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

[50] Wang, P.; Chen, S.; Yang, S. (2022), "Recent Advances on Penalized Regression Models for Biological Data", *Mathematics*, 2022(10), 3695. https://doi.org/10.3390/math10193695

[51] Van De Geer, S. (2008). "High-dimensional generalized linear models and the lasso". *Ann. Statist.*, 36 614–645.

[52] Yohai, V. J. (1987), "High breakdown-point and high effciency robust estimates for regression". *Ann. Statist.*, 15, 642-656.

[53] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38, 894-942.

[54] Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27, 576-593

[55] Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B*, 67(2), 301–320.

[56] Zuo, Y. (2003) "Projection-based depth functions and associated medians", *Ann. Statist.*, 31, 1460-1490.

[57] Zuo, Y. (2006), "Multi-dimensional trimming based on projection depth", *Ann. Statist.*, 34(5), 2211-2251.

[58] Zuo, Y. (2024), "Non-asymptotic robustness analysis of regression depth median", *Journal of Multivariate Analysis*, (in press), arXiv: 2009.00646

[59] Zuo, Y. (2022), "Non-asymptotic analysis and inference for an outlyingness induced winsorized mean", *Stat. Papers*, https://doi.org/10.1007/s00362-022-01353-5

[60] Zuo, Y., Serfling, R., (2000), "General notions of statistical depth function", *Ann. Statist.*, 28, 461-482.

[61] Zuo, Y. and Zuo, H. (2023), "Least sum of squares of trimmed residuals regression", *Electronic Journal of Statistics* 17(2), 2416-2446. arXiv:2202.10329

Appendix: proofs of main results

Proof of Theorem 2.1

Proof: Clearly, it suffices to show that $\text{RBP}(\widehat{\boldsymbol{\beta}}^*(\lambda_1,\lambda_2,\gamma,\boldsymbol{z}^{(n)}),\boldsymbol{z}^{(n)}) \le 1/n$. Equivalently, to show that one contaminating point can break down the estimator. Assume, otherwise, one point is not enough to break down the estimator. That is, there exists an M such that

$$\sup_{\boldsymbol{z}_1^{(n)}} \|\widehat{\boldsymbol{\beta}}^*(\lambda_1,\lambda_2,\gamma,\boldsymbol{z}_1^{(n)})\|_2 < M < \infty, \tag{28}$$

where $\boldsymbol{z}_1^{(n)}$ stands for any contaminated data set by replacing one point in the original data set $\boldsymbol{z}^{(n)}$ with an arbitrary point in $\mathbb{R}^p$. We seek a contradiction now.

Replace $\boldsymbol{z}_1 = (\boldsymbol{x}_1', y_1)'$ in $\boldsymbol{z}^{(n)} = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n\}$ by $\boldsymbol{z}_1^* = ((\delta, 0, \cdots, 0), \kappa\delta)'$. Denote the contaminated data set by $\boldsymbol{z}_1^{(n)}$ and the estimator based on it as $\widehat{\boldsymbol{\beta}}^* := \widehat{\boldsymbol{\beta}}^*(\lambda_1,\lambda_2,\gamma,\boldsymbol{z}_1^{(n)})$.

Let $M_y = \max_i |y_i|, M_x = \max_i |\boldsymbol{x}_{i1}|$. Let $\boldsymbol{\beta}_\kappa = (0,\kappa,0,\cdots,0)' \in \mathbb{R}^p$ and set $\kappa = (\sqrt{p}+1)M+1$. Select a large δ such that $f\Big(\mathcal{L}(\delta)/n\Big) \ge f\Big(\mathcal{L}(M_y+M_x\kappa)\Big)+g(\boldsymbol{\beta}_\kappa,\lambda_1,\lambda_2,\gamma)+1$, This is possible since $\mathcal{L}(x) \to \infty$ when $|x| \to \infty$ and $\mathcal{L}(x)$ is non-decreasing over $(0,\infty)$, $f(x)$ is strictly increasing over $[0,\infty)$ and $f(x) \to \infty$ as $x \to \infty$. Then

$$\begin{aligned}
O(\widehat{\boldsymbol{\beta}}^*) \le O(\boldsymbol{\beta}_\kappa) &= f\Big(\frac{1}{n}\sum_{i=1}^n \mathcal{L}(r_i(\boldsymbol{\beta}_\kappa))\Big) + g(\boldsymbol{\beta}_\kappa,\lambda_1,\lambda_2,\gamma) \\
&= f\Big(\frac{1}{n}\sum_{i=2}^n \mathcal{L}(r_i(\boldsymbol{\beta}_\kappa))\Big) + g(\boldsymbol{\beta}_\kappa,\lambda_1,\lambda_2,\gamma) \quad (\text{since } \mathcal{L}(r_1)=\mathcal{L}(0)=0) \\
&\le f\Big(\frac{n-1}{n}\mathcal{L}(M_y+M_x\kappa)\Big) + g(\boldsymbol{\beta}_\kappa,\lambda_1,\lambda_2,\gamma) \\
&< f\Big(\frac{1}{n}\mathcal{L}(\delta)\Big) - 1.
\end{aligned} \tag{29}$$

On the other hand, for any $\boldsymbol{\beta} = (\beta_1,\beta_2,\cdots,\beta_p)' \in \mathbb{R}^p$ such that $(\sqrt{p}+1)\|\boldsymbol{\beta}\|_2 \le \kappa - 1$, one has

$$\begin{aligned}
O(\boldsymbol{\beta}) &\ge f\Big(\frac{1}{n}\mathcal{L}(y_1 - \boldsymbol{w}_1\boldsymbol{\beta})\Big) = f\Big(\frac{1}{n}\mathcal{L}(\kappa\delta - (\beta_1+\delta\beta_2))\Big) \\
&\ge f\Big(\frac{1}{n}\mathcal{L}(\kappa\delta - (\delta|\beta_1| + \delta|\beta_2|))\Big) \\
&\ge f\big(\frac{1}{n}\mathcal{L}(\delta(\kappa - (\sqrt{p}+1)\|\boldsymbol{\beta}\|_2))\Big) \ge f\Big(\frac{1}{n}\mathcal{L}(\delta)\Big),
\end{aligned} \tag{30}$$

where the facts: (i) $|\beta_i| \le \|\boldsymbol{\beta}\|_i (i=1,2)$ and (ii) $\|\boldsymbol{\beta}\|_1 \le \sqrt{p}\|\boldsymbol{\beta}\|_2$ are utilized.

Combining (29) and (30), leads to the conclusion that

$$\|\widehat{\boldsymbol{\beta}}^*(\lambda_1,\lambda_2,\gamma,\boldsymbol{z}_1^{(n)})\|_2 > \frac{\kappa-1}{\sqrt{p}+1} = M,$$

which contradicts (28). □

Proof of Theorem 3.1

Proof: $\lambda_1+\lambda_2=0$ case has been treated in [61], we treat $\lambda_1+\lambda_2>0$ case here.

(i) Denote the objective function on the RHS of (20) as

$$O_n(\boldsymbol{\beta},\lambda_1,\lambda_2,\gamma) := O(\boldsymbol{\beta},\lambda_1,\lambda_2,\alpha,\gamma,\boldsymbol{z}^{(n)}) = \frac{1}{n}\sum_{i=1}^{n} r_i^2 w_i + \lambda_1\sum_{j=1}^{p}|\beta_j|^\gamma + \lambda_2\sum_{i=1}^{p}\beta_i^2. \tag{31}$$

Name the three terms on the RHS above as $g(\alpha,\boldsymbol{\beta},\boldsymbol{z}^{(n)})$, $g_1(\lambda_1,\gamma,\boldsymbol{\beta})$, and $g(\lambda_2,\boldsymbol{\beta})$, respectively. It is readily seen that the RHS of (20) is equivalent to minimizing $G(\boldsymbol{\beta},\alpha,\lambda_1,\gamma) := g(\alpha,\boldsymbol{\beta},\boldsymbol{Z}^{(n)}) + g_1(\lambda_1,\gamma,\boldsymbol{\beta})$ subject to $\sum_{i=1}^{p}\beta_i^2 \le t_2$, $t_2 \ge 0$

By Lemma 2.2 of [61], $g(\alpha,\boldsymbol{\beta},\boldsymbol{z}^{(n)})$ is continuous in $\boldsymbol{\beta}$ (this is not as obvious as one believed) while the continuity of $g_1(\lambda_1,\gamma,\boldsymbol{\beta})$ in $\boldsymbol{\beta}$ is obvious. Therefore we have a continuous function of $\boldsymbol{\beta}$, $G(\boldsymbol{\beta},\alpha,\lambda_1,\gamma)$, which obviously has minimum value over the compact set $\|\boldsymbol{\beta}\|_2 \le t_2$.

(ii) Follows the approach originated in [61], we partition the parameter space $\mathbb{R}^p$ of $\boldsymbol{\beta}$ into disjoint open pieces $R_{\boldsymbol{\beta}^k}$, $1 \le k \le L \le \binom{n}{\lfloor (n+1)/2 \rfloor}$ and $\cup_{1\le k\le L}\overline{R}_{\boldsymbol{\beta}^k} = \mathbb{R}^p$, where $\overline{A}$ stands for the closure of the set A, and

$$R_{\boldsymbol{\beta}^k} = \{\boldsymbol{\beta}\in\mathbb{R}^p : I(\boldsymbol{\beta}) = I(\boldsymbol{\beta}^k), O_{i_1}(\boldsymbol{\beta}) < O_{i_2}(\boldsymbol{\beta})\cdots < O_{i_K}(\boldsymbol{\beta})\}, \tag{32}$$

where $O_i(\boldsymbol{\beta}) := O(r_i, r^{(n)})$ for a given $\mathbf{z}^{(n)}$ and $\boldsymbol{\beta}$, $i_1,\cdots,i_K$ in $I(\boldsymbol{\beta})$ and $K = |I(\boldsymbol{\beta})|$ with w_i defined in (20)

$$I(\boldsymbol{\beta}) = \Big\{i : w_i = 1\Big\}. \tag{33}$$

For any $\boldsymbol{\beta}\in\mathbb{R}^p$, either there is $R_{\boldsymbol{\eta}}$ and $\boldsymbol{\beta}\in R_{\boldsymbol{\eta}}$ or there is $R_{\boldsymbol{\xi}}$, such that $\boldsymbol{\beta}\notin R_{\boldsymbol{\eta}}\cup R_{\boldsymbol{\xi}}$ and $\boldsymbol{\beta}\in\overline{R}_{\boldsymbol{\eta}}\cap\overline{R}_{\boldsymbol{\xi}}$. Now we claim that $\widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}^n_{lts-enet}(\alpha,\lambda_1,\lambda_2,\gamma) \in R_{\boldsymbol{\beta}^{k_0}}$ for some $1\le k_0\le L$.

Otherwise, assume that $\widehat{\boldsymbol{\beta}}\in\overline{R}_{\boldsymbol{\beta}^{k_0}}$. By Lemma 2.2 of [61], $g(\alpha,\boldsymbol{\beta},\boldsymbol{Z}^{(n)})$ (denoted by $Q^n(\boldsymbol{\beta})$ there) is convex over $R_{\boldsymbol{\beta}^{k_0}}$. Therefore, $O_n(\boldsymbol{\beta},\lambda_1,\lambda_2,\gamma)$ is strictly convex in $\boldsymbol{\beta}$ over $R_{\boldsymbol{\beta}^{k_0}}$. Assume that $\boldsymbol{\beta}^*$ is the global minimum of $O_n(\boldsymbol{\beta},\lambda_1,\lambda_2,\gamma)$ over $R_{\boldsymbol{\beta}^{k_0}}$. Then it is obviously that $O_n(\widehat{\boldsymbol{\beta}},\lambda_1,\lambda_2,\gamma) \le O_n(\boldsymbol{\beta}^*,\lambda_1,\lambda_2,\gamma)$. But this is impossible in light of Lemma 3.2. The strict convexity of $O_n(\boldsymbol{\beta},\lambda_1,\lambda_2,\gamma)$ over $R_{\boldsymbol{\beta}^{k_0}}$ guarantees the uniqueness. □

Proof of Lemma 3.2

Proof: Let $B(\boldsymbol{x}^*,r)$ be a small ball centered at $\boldsymbol{x}^*$ with a small radius r and $B(\boldsymbol{x}^*,r)\subset S$. Let $B^c := S - B(\boldsymbol{x}^*,r)$ and $\alpha^* = \inf_{\boldsymbol{x}\in B^c} f(\boldsymbol{x})$. Then, $\alpha^* > f(\boldsymbol{x}^*)$ (in light of strict convexity) Since $y\in\overline{S}$, then there is a sequence $\{\boldsymbol{x}_j\}\in B^c$ such that $\boldsymbol{x}_j\to\boldsymbol{y}$ and $f(\boldsymbol{x}_j)\to f(\boldsymbol{y})$ as $j\to\infty$. Hence $f(\boldsymbol{y}) = \lim_{j\to\infty} f(\boldsymbol{x}_j) \ge \alpha^* > f(\boldsymbol{x}^*)$. □

Proof of Theorem 3.2

Proof: We complete it in two steps.

(i) $m \leq n-k$ *contaminating points are not enough to break down the estimator.* Let $M_y = \max_i |y_i|$, denote the minimizer of the Q in (21) for the contaminated sample as $\widehat{\boldsymbol{\beta}}$, Then, it is obviously that

$$Q(\widehat{\boldsymbol{\beta}}) \leq Q(\mathbf{0}, \lambda_1, \lambda_2, \gamma, \boldsymbol{z}_m^{(n)}) = \frac{1}{n}\sum_{i=1}^{n} r_i^2 w_i \leq \frac{k}{n}(M_y^2),$$

where the last inequality deserves further explanations. Note that there are at least k uncontaminated (original) points. Therefore, in the case that $w_i = \mathbb{1}(r_i^2 \leq r_{h:n}^2)$, the RHS of the above display wants to keep the sum of smallest $k = h$ squared residuals (y_i^2), this sum is certainly no greater than that of $k = h$ squared residuals from the k original points.

Likewise, in the case of $w_i = \mathbb{1}(O(r_i, r^{(n)}) \leq \alpha)$, the RHS wants to keep the sum of squared residuals (y_i^2) from $k = |I(\mathbf{0})|$ points that have the smallest outlyingness no greater than α, which is certainly no greater than the sum of k squared residuals from the k original points.

Assume, w.l.o.g. that $\lambda_1 > 0$ ($\lambda_2 > 0$ is even easier). Consider any $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\|_2 \geq M := ((k+1)M_y^2/n\lambda_1)^{1/\gamma}$, then

$$Q(\boldsymbol{\beta}, \lambda_1, \lambda_2, \gamma, \boldsymbol{z}_m^{(n)}) > \lambda_1 \sum_{i=1}^{p} |\beta_i|^\gamma = \lambda_1 \|\boldsymbol{\beta}\|_\gamma^\gamma \geq \lambda_1 \|\boldsymbol{\beta}\|_2^\gamma \geq \frac{k+1}{n} M_y^2,$$

where the fact that $\|x\|_t \leq \|x\|_s$ when $1 \leq s \leq t < \infty$ is invoked.

The two displays above imply that

$$\|\widehat{\boldsymbol{\beta}}(\lambda_1, \lambda_2, \gamma, \boldsymbol{z}_m^{(n)})\|_2 < M.$$

(ii) $m = n-k+1$ *contaminating points are enough to break down the estimator.*

The structure and basic idea of this part is an analogue to that of proof of Theorem 2.1. Assume, otherwise, m points are not enough to break down the estimator. That is, there exists an M such that

$$\sup_{\boldsymbol{z}_m^{(n)}} \|\widehat{\boldsymbol{\beta}}(\lambda_1, \lambda_2, \gamma, \boldsymbol{z}_m^{(n)})\|_2 < M < \infty, \tag{34}$$

where $\boldsymbol{z}_m^{(n)}$ stands for any contaminated data set by replacing m points in the original data set $\boldsymbol{z}^{(n)}$ with m arbitrary points in $\mathbb{R}^p$. We seek a contradiction now.

Replacing m original points $\boldsymbol{z}_i$s with the point $((\delta, 0, \cdots, 0), \delta\kappa)'$. Denote the contaminated data set by $\boldsymbol{z}_m^{(n)}$ and the estimator based on it as $\widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}(\lambda_1, \lambda_2, \gamma, \boldsymbol{z}_m^{(n)})$.

Let $M_y = \max_i |y_i|, M_x = \max_i |\boldsymbol{x}_{i1}|$. Let $\boldsymbol{\beta}_\kappa = (0, \kappa, 0, \cdots, 0)' \in \mathbb{R}^p$ and set $\kappa = M+1$. Select a large δ such that

$$\delta^2 \geq \frac{\max(k-m, k-1)}{n}(M_y + \kappa M_x)^2 + \lambda_1 \kappa^\gamma + \lambda_2 \kappa^2 + 1,$$

This is possible since $x^2 \to \infty$ when $|x| \to \infty$. Note that $k \geq m$.

It is readily seen that all residuals based on $\boldsymbol{\beta}_\kappa$ and m contaminated points are zeros. All non-zero residuals correspond to uncontaminated original points. Then in the case that $w_i = \mathbb{1}(r_i^2 \leq r_{k:n}^2)$

$$\begin{aligned} Q(\widehat{\boldsymbol{\beta}}) \leq Q(\boldsymbol{\beta}_\kappa) &= \frac{1}{n}\sum_{i=1}^{n} r_i^2 w_i + \lambda_1 \kappa^\gamma + \lambda_2 \kappa^2 \\ &= \begin{cases} \frac{1}{n}\sum_{i=1}^{k-m} r_{j_i}^2 + \lambda_1 \kappa^\gamma + \lambda_2 \kappa^2, & \text{if } k > m \\ \lambda_1 \kappa^\gamma + \lambda_2 \kappa^2 & \text{else,} \end{cases} \end{aligned} \tag{35}$$

where the last equality is due that fact that the objective function sums the smallest k squared residuals, but among n squared residuals, m of them are zeros.

For the case $w_i = \mathbb{1}(O(r_i, r^{(n)}) \leq \alpha)$, one has

$$\begin{aligned} Q(\widehat{\boldsymbol{\beta}}) \leq Q(\boldsymbol{\beta}_\kappa) &= \frac{1}{n}\sum_{i=1}^{n} r_i^2 w_i + \lambda_1 \kappa^\gamma + \lambda_2 \kappa^2 \\ &= \frac{1}{n}\sum_{i=1}^{k-1} r_{j_i}^2 + \lambda_1 \kappa^\gamma + \lambda_2 \kappa^2, \end{aligned} \tag{36}$$

where the last equality is due the fact that there is at most $n - m = k - 1$ non-zero residuals. Overall we have

$$Q(\widehat{\boldsymbol{\beta}}) \leq Q(\boldsymbol{\beta}_\kappa) \leq \frac{\max(k-m, k-1)}{n}(M_y + \kappa M_x)^2 + \lambda_1 \kappa^\gamma + \lambda_2 \kappa^2 \leq \delta^2 - 1. \tag{37}$$

On the other hand, for any $\boldsymbol{\beta}$ with $\beta_1 \leq \|\boldsymbol{\beta}\|_2 \leq \kappa - 1$, one has

$$Q(\boldsymbol{\beta}) \geq (\kappa\delta - \delta\beta_1)^2 \geq \delta^2, \tag{38}$$

where we utilize the fact that among the k squared residuals, there is at least one residual that is based on a contaminated point since un-contaminated points are at most $k - 1$.

Combining (37) and (38) leads to

$$\|\widehat{\boldsymbol{\beta}}(\lambda_1, \lambda_2, \gamma, \boldsymbol{z}_m^{(n)})\|_2 > \kappa - 1 = M,$$

which contradicts (34). □