Electronic Journal of Statistics
Vol. x (202x) 1–50
ISSN: 1935-7524

Least sum of squares of trimmed residuals regression

Hanwen Zuo and Yijun Zuo

Department of Computer Science and Department of Statistics and Probability
Michigan State University, East Lansing, MI 48824, USA
e-mail: zuohanwe@msu.edu; zuo@msu.edu

Abstract: In the famous least sum of trimmed squares (LTS) of residuals estimator [21], residuals are first squared and then trimmed. In this article, we first trim residuals - using a depth trimming scheme - and then square the rest of residuals. The estimator that can minimize the sum of squares of the trimmed residuals, is called an LST estimator.

It turns out that the LST is a robust alternative to the classic least sum of squares (LS) estimator. Indeed, it has a very high finite sample breakdown point, and can resist, asymptotically, up to 50% contamination without breakdown - in sharp contrast to the 0% of the LS estimator.

The population version of the LST is Fisher consistent, and the sample version is strong and root-n consistent and asymptotically normal. Approximate algorithms for computing the LST are proposed and tested in synthetic and real data examples. These experiments indicate that one of the algorithms can compute the LST estimator very fast and with relatively smaller variances, compared with that of the famous LTS estimator. All the evidence suggests that the LST deserves to be a robust alternative to the LS estimator and is feasible in practice for high dimensional data sets (with possible contamination and outliers).

MSC2020 subject classifications: Primary 62J05, 62G36; secondary 62J99, 62G99.
Keywords and phrases: trimmed residuals, robust regression, finite sample breakdown point, consistency, approximate computation algorithm..

Contents

arXiv: 2202.10329

1. Introduction

In the classical regression analysis, we assume that there is a relationship for a given data set $\{(\boldsymbol{x}_i', y_i)', i \in \{1, 2, \cdots, n\}\}$:

$$y_i = (1, \boldsymbol{x}_i')\boldsymbol{\beta}_0 + e_i, \quad i \in \{1, \cdots, n\} \tag{1.1}$$

where $y_i \in \mathbb{R}^1$, $'$ stands for the transpose, $\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0p})'$ (the true unknown parameter) in $\mathbb{R}^p$ and $\boldsymbol{x_i} = (x_{i1}, \cdots, x_{i(p-1)})'$ in $\mathbb{R}^{p-1}$, $e_i \in \mathbb{R}^1$ is called an error term (or random fluctuation/disturbances, which is usually assumed to have zero mean and variance σ^2 in classic regression theory). That is, β_{01} is the intercept term of the model. Write $\boldsymbol{w}_i = (1, \boldsymbol{x}_i')'$, then one has $y_i = \boldsymbol{w}_i'\boldsymbol{\beta}_0 + e_i$, which will be used interchangeably with model (1.1).

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 the model $y = (1, \boldsymbol{x}')\boldsymbol{\beta}_0 + e$. Call the difference between y_i and $\boldsymbol{w_i'\beta}$ the ith residual, $r_i(\boldsymbol{\beta})$, for a candidate coefficient vector $\boldsymbol{\beta}$ (which is often suppressed). That is,

$$r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{w_i'\beta}. \tag{1.2}$$

To estimate $\boldsymbol{\beta}_0$, the classic *least squares* (LS) minimizes the sum of squares of residuals,

$$\widehat{\boldsymbol{\beta}}_{ls} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} r_i^2.$$

Alternatively, one can replace the square above by absolute value to obtain the least absolute deviations estimator (aka, L_1 estimator, in contrast to the L_2 (LS) estimator).

The LS estimator is very popular in practice across a broader spectrum of disciplines due to its great computability and optimal properties when the error e_i follows a normal $\mathcal{N}(\mathbf{0}, \sigma^2)$ distribution. 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.

Robust alternatives to the $\widehat{\boldsymbol{\beta}}_{ls}$ abound in the literature for a long time. The most popular ones are, among others, M-estimators [14], least median squares (LMS) and least trimmed squares (LTS) estimators [21], S-estimators [27], MM-estimators [46], τ-estimators [47], and maximum depth estimators ([22], [52], and [53]). For more related discussions, please see, Sections 1.2 and 4.4 of [23], and Section 5.14 of [17].

Among all robust alternatives, in practice, the LTS is one of the most prevailing crossing multiple disciplines. Its idea is simple, ordering the squared residuals and then trimming the larger ones and keeping at least $\lceil n/2 \rceil$ squared residuals, where $\lceil\ \rceil$ is the ceiling function, the minimizer of the sum of those *trimmed squared residuals* is called an LTS estimator:

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

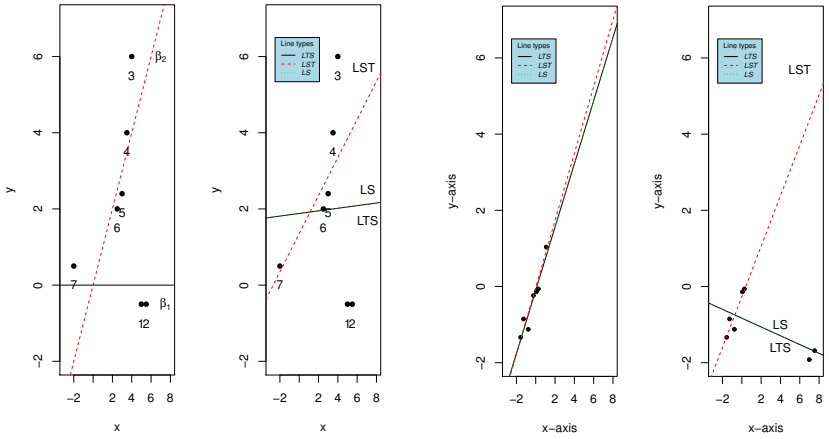
where $(r^2)_{1:n} \leq (r^2)_{2:n} \leq \cdots, (r^2)_{n:n}$ are the ordered squared residuals and constant h satisfies $\lceil n/2 \rceil \leq h \leq n$.

One naturally wonders, what if one first trims (employing the scheme given in Section 2) the residuals and then minimizes the sum of *squares of trimmed residuals* (the minimizer will be called an LST)? Is there any difference between the two procedures? Outlying (extremely large or small) original residuals are trimmed after squaring in the LTS - those residuals certainly are trimmed in the LST. But the outlying residuals which have a small squared magnitude will not be trimmed in the LTS and are trimmed in the LST (see (a) of Figure 1). Before formally introducing the LST in Section 2, let us first appreciate the difference between the two procedures.

Example 1.1 We constructed a small data set in $\mathbb{R}^2$ with $\boldsymbol{x} = (5, 5.5, 4, 3.5, 3, 2.5, -2)$ and $\boldsymbol{y} = (-.5, -.5, 6, 4, 2.4, 2, .5)$, they are plotted in the left panel of the (a) of Figure 1. We also provide two candidate regression lines $\boldsymbol{\beta}_1$ ($y = 0$) and $\boldsymbol{\beta}_2$ ($y = x$). Which one would you pick to represent the overall pattern of the data set?

If one uses the number $h = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ given on page 132 of [23] to achieve the maximum possible breakdown point (see Section 3 for definition) for the LTS estimator, that is, employing four smallest squared residuals, then the LTS prefers $\boldsymbol{\beta}_1$ (using residuals from points 1, 2, 6, and 7) to $\boldsymbol{\beta}_2$ (using points $4, 5, 6, 7$), whereas for the LST, $\boldsymbol{\beta}_2$ (using residuals from points $4, 5, 6, 7$) is the preferred. One might immediately argue that this is not representative since the LTS searches all possible (not just two) lines and outputs the best one.

If one utilized the R function ltsReg, then it produced the solid (black) line whereas the line based on algorithms (see Section 5) for the LST is the dashed (red) one in the right panel of the (a) of Figure 1. For benchmark purposes, the LS line dotted (green) is also given, which is overlapping with the LTS line. From this instance, One can appreciate the difference between trimming schemes of



(a) Left panel: plot of seven artificial points and two candidate lines ($\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$), which line would you pick? Sheerly based on the trimming scheme and objective function value, if one uses the number $h = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ given on page 132 of [23], that is, employing four squared residuals, then the LTS prefers $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_2$ whereas the LST reverses the preference.

Right panel: the same seven points are fitted by the LTS, the LST, and the LS (benchmark). A solid black line is the LTS given by ltsReg. Red dashed line is given by the LST, and green dotted line is given by the LS - which is identical to the LTS line in this case.

(b) Left panel: plot of seven highly correlated normal points (with mean being the zero vector and covariance matrix with diagonal entries being one and off-diagonal entries being 0.88) and three lines given by the LST, the LTS, and the LS. The LS line is identical to the LTS line again.

Right panel: The LTS line (solid black) and the LST line (dashed red), and the LS (dotted green) for the same seven highly correlated normal points but with two points contaminated nevertheless. The LS line is identical to the LTS line due to the attributes in the R function ltsReg that is based on [26]).

Fig 1: (a) Difference between the two procedures: the LST and the LTS. (b) Performance difference between the LST and the LTS when there are contaminated points (x-axis leverage points).

the LTS and the LST. Of course, one might argue that the data set in the (a) is purely synthetic and fixed.

So, in the (b) of Figure 1, we generated seven highly correlated normal points (with correlation 0.88 between x and y), when there is no contamination the LTS (identical to the LS again) and the LST pick perfectly the linear pattern whereas if there are two contaminated points (note that the LTS allows $m := \lfloor (n-p)/2 \rfloor = 2$ contaminated points in this case in light of Theorem 6 on page

132 of [23]), the line from the LTS drastically changes in this particular instance, which again is identical to the LS.

For examples with an increased sample size, see Section 6. Incidentally, the instability of the LMS (not the LTS) was already documented in [13]. □

The idea of trimming residuals and then doing regression has appeared in the literature for quite some time. The trimming idea was first introduced in location setting and later extended to regression, see, [15], [2], [28], [44], and [23], among others. Trimmed mean has been used in practice for more than two centuries (see [8], page 34, and is attributed to "Anonymous" (1821)([1]) (Gergonne, see [33]), or [18]. Tukey ([37], [4]) is one of the outstanding advocators for the trimmed mean in the last century.

However, trimming residuals based on depth or outlyingness employed in this article (see Section 2) is novel and has never been utilized before. A more recent study on the topic is given by Johansen and Nielsen (2013), where the authors used an iterated one-step approximation to Huber-skip estimator to detect outliers in regression and theoretical justification for the approximation is provided. Their Huber-skip estimator defined on page 56 is closely related to our LST, but there are two essential differences (i) their estimator more resembles the least winsorized squares regression (see page 135 of [23]), (ii) residuals in their estimator are not centered by the median of residuals.

In light of [52], both the LTS and the LST could be regarded as the deepest estimator (aka regression median) with respect to the corresponding objective function type of regression depth (see Section 2.3.1 of [52] and Section 4).

The rest of the article is organized as follows. Section 2 introduces trimming schemes and the least sum of squares of trimmed (LST) residuals estimator and establishes the existence and equivariance properties. Section 3 investigates the robustness of the LST in terms of its finite sample breakdown point and its influence function. Section 4 establishes the Fisher as well as the strong and the root-n consistency. The asymptotic normality is derived from stochastic equicontinuity in Section 5. Section 6 is devoted to the computation algorithms of the LST where two approximate algorithms are proposed. Section 7 presents examples of simulated and real data and carries out the comparison with the leading regression estimators, the LTS and the LMS. Section 8 consists of some concluding discussions. Long proofs are deferred to the Appendix.

2. Least sum of squares of trimmed residuals estimator

2.1. Trimming schemes

Rank based trimming This scheme is based on the ranks of data points, usually trimming an equal number of points at both tails of a data set (that is, lower or higher rank points are trimmed) and also can trim points one-sided

if needed (such as when all data points lie on the positive (or negative) side of number axis).

This scheme is closed related to the trimmed mean, which can keep a good balance between robustness and efficiency, alleviating the extreme sensitivity of sample mean and enhancing the efficiency of the sample median.

Rank-based trimming focuses only on the relative position of points with respect to others and ignores the magnitude of the point and the relative distance between points. [49] and [45] discussed an alternative trimming scheme, which exactly catches these two important attributes (magnitude and relative distance). It orders data from a center (the median) outward and trims the points that are far away from the center. This is known as depth-based trimming.

Depth (or outlyingness) based trimming In other words, the depth-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 [48])

$$D(x, X^{(n)}) = |x - \mathrm{Med}(X^{(n)})|/\mathrm{MAD}(X^{(n)}), \tag{2.1}$$

where $X^{(n)} = \{x_1, \cdots, x_n\}$ is a data set in $\mathbb{R}^1$, $\mathrm{Med}(X^{(n)}) = \mathrm{median}(X^{(n)})$ is the median of the data points, and $\mathrm{MAD}(X^{(n)}) = \mathrm{Med}(\{|x_i - \mathrm{Med}(X^{(n)})|,\ i \in \{1, 2, \cdots, n\}\})$ is the median of absolute deviations to the center (median). It is readily seen that $D(x, X^{(n)})$ is a generalized standard deviation, or equivalent to the one-dimensional projection depth/outlyingness (see [55] and [48, 49] for a high dimensional version). For notion of outlyingness, cf. [32], [5], and [6].

The LTS essentially employs one-sided rank based trimming scheme (w.r.t. squared residuals), whereas depth based trimming is utilized in the LST which is introduced next.

2.2. Definition and properties of the LST

Definition For a given sample $\mathbf{Z}^{(n)} = \{(\boldsymbol{x}_i', y_i)', i \in \{1, 2, \cdots, n\}\}$ in $\mathbb{R}^p$ from $y = \boldsymbol{w}'\boldsymbol{\beta}_0 + e$ and a $\boldsymbol{\beta} \in \mathbb{R}^p$, define

$$m_n(\boldsymbol{\beta}) := m(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta}) = \mathrm{Med}_i\{r_i\}, \tag{2.2}$$

$$\sigma_n(\boldsymbol{\beta}) := \sigma(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta}) = \mathrm{MAD}_i\{r_i\}, \tag{2.3}$$

where operators Med and MAD are used for discrete data sets (and distributions as well) and r_i defined in (1.2). For a constant α in the depth trimming scheme, consider the quantity

$$Q(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta}, \alpha) := \sum_{i=1}^{n} r_i^2 \mathbb{1}\left(\frac{|r_i - m(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})} \le \alpha\right), \tag{2.4}$$

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 $\sigma(\mathbf{Z}^{(n)}, \boldsymbol{\beta}) = 1$ (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) residuals estimator,

$$\widehat{\boldsymbol{\beta}}_{lst}^{n} := \widehat{\boldsymbol{\beta}}_{lst}(\mathbf{Z}^{(n)}, \alpha) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} Q(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta}, \alpha). \tag{2.5}$$

One might take it for granted that the minimizer of $Q(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta}, \alpha)$ always exists. Does the right-hand side (RHS) of (2.5) always have a minimizer? If it exists, is it unique? We treat this problem formally next. Assume $\mathbf{X}_n = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_n)'$ has a full rank p ($p < n$) throughout.

Hereafter we will assume that $\alpha \geq 1$. That is, we will keep the residuals that are no greater than one MAD away from the center (the median of residuals) untrimmed. For a given α, $\boldsymbol{\beta}$, and $\boldsymbol{Z}^{(n)}$, define a set of indexes for $1 \leq i \leq n$

$$I(\boldsymbol{\beta}) = \Big\{ i : \frac{|r_i - m(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})} \leq \alpha \Big\}. \tag{2.6}$$

Namely, the set of subscripts so that the outlyingness (see (2.1)) of the corresponding residuals are no greater than α. It depends on $\mathbf{Z}^{(n)}$ and α, which are suppressed in the notation. Following the convention, we denote the cardinality of set A by $|A|$. We have

Lemma 2.1 For any $\boldsymbol{\beta} \in \mathbb{R}^p$ and the given $\mathbf{Z}^{(n)}$ and α, $|I(\boldsymbol{\beta})| \geq \lfloor (n+1)/2 \rfloor$.

Proof: By the definition of MAD (the median of the absolute deviations to the center (median)), it is readily seen that

$$\begin{aligned} |I(\boldsymbol{\beta})| &= \sum_{i=1}^{n} \mathbb{1}\left(\frac{|r_i - m(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})} \leq \alpha \right) \\ &\geq \sum_{i=1}^{n} \mathbb{1}\left(\frac{|r_i - m(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})} \leq 1 \right) = \lfloor (n+1)/2 \rfloor, \end{aligned}$$

This completes the proof. □

The lemma implies that the RHS of (2.4) sums a majority of squared residuals.

Properties of the objective function

Write $D_i := D(r_i, \boldsymbol{\beta}) = |r_i - m(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})| / \sigma(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})$ for a given $\mathbf{Z}^{(n)}$ and $\boldsymbol{\beta}$. Let $i_1, \cdots, i_K$ be in $I(\boldsymbol{\beta})$ such that $D_{i_1} \leq D_{i_2} \cdots \leq D_{i_K}$ (i.e. ordered depth values of residuals), $K := |I(\boldsymbol{\beta})|$. Both i_j and D_{i_j} clearly depend on $\boldsymbol{\beta}$ and $\mathbf{Z}^{(n)}$.

Generally, the inequalities between the D_i's cannot be strict unless we assume that $r := y - \boldsymbol{w}'\boldsymbol{\beta}$ has a density for any $\boldsymbol{\beta} \in \mathbb{R}^p$. In the latter case, the strict inequalities hold almost surely (a.s.), i.e., $D_{i_1} < D_{i_2} \cdots < D_{i_K}$ (a.s.). Define for any $\boldsymbol{\beta}^1 \in \mathbb{R}^p$ and a given $\mathbf{Z}^{(n)}$

$$R_{\boldsymbol{\beta}^1} = \{\boldsymbol{\beta} \in \mathbb{R}^p : I(\boldsymbol{\beta}) = I(\boldsymbol{\beta}^1), D_{i_1}(\boldsymbol{\beta}) < D_{i_2}(\boldsymbol{\beta}) \cdots < D_{i_K}(\boldsymbol{\beta})\}. \tag{2.7}$$

That is, the set of all $\boldsymbol{\beta}$s that share the same index set $I(\boldsymbol{\beta}^1)$ of $\boldsymbol{\beta}^1$. If $y - \boldsymbol{w}'\boldsymbol{\beta}$ has a density at $\boldsymbol{\beta}^1 \in \mathbb{R}^p$, then $R_{\boldsymbol{\beta}^1} \neq \emptyset$ (a.s.). There are at most finitely many $R_{\boldsymbol{\beta}^k}$s, $\boldsymbol{\beta}^k \in \mathbb{R}^p$, $1 \leq k \leq L := \binom{n}{\lfloor (n+1)/2 \rfloor}$ such that $\cup_{k=1}^{L} \overline{R}_{\boldsymbol{\beta}^k} = \mathbb{R}^p$, where $R_{\boldsymbol{\beta}^k}$ is defined similarly to (2.7) and $\overline{A}$ stands for the closure of the set A. 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 $\boldsymbol{\beta} \in \overline{R}_{\boldsymbol{\eta}} \cap \overline{R}_{\boldsymbol{\xi}}$. In the latter case, there are $i_k, i_l \in I(\boldsymbol{\beta})$ $i_k \neq i_l$, such that $D_{i_k} = D_{i_l}$.

For a given sample $\boldsymbol{Z}^{(n)}$, write $Q^n(\boldsymbol{\beta})$ for $Q(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta}, \alpha)$, $B(\boldsymbol{\eta}, \delta)$ for an open ball in $\mathbb{R}^p$ centered at $\boldsymbol{\eta}$ with a radius $\delta > 0$, and $\mathbb{1}_i$, which depends on $\boldsymbol{\beta}$, for $\mathbb{1}\left(|y_i - \boldsymbol{w}'_i\boldsymbol{\beta} - m_n(\boldsymbol{\beta})|/\sigma_n(\boldsymbol{\beta}) \leq \alpha\right)$. Let $\mathbf{Y}_n = (y_1, \cdots, y_n)'$ and $\boldsymbol{M}_n := \boldsymbol{M}(\mathbf{Y}_n, \mathbf{X}_n, \boldsymbol{\beta}, \alpha) = \sum_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{w}'_i \mathbb{1}_i = \sum_{i \in I(\boldsymbol{\beta})} \boldsymbol{w}_i \boldsymbol{w}'_i$. We have

Lemma 2.2

(i) For a given $\mathbf{Z}^{(n)}$ and α, for any $1 \leq k \leq L$ and any $\boldsymbol{\eta} \in R_{\boldsymbol{\beta}^k}$, there exists a $B(\boldsymbol{\eta}, \delta)$ such that for any $\boldsymbol{\beta} \in B(\boldsymbol{\eta}, \delta)$, $\boldsymbol{\beta} \in R_{\boldsymbol{\beta}^k}$, i.e.,

$$Q^n(\boldsymbol{\beta}) = \sum_{i \in I(\boldsymbol{\beta}^k)} r_i^2,$$

(ii) For any $1 \leq k \leq L$, $R_{\boldsymbol{\beta}^k}$ is open,

(iii) $Q^n(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta} \in \mathbb{R}^p$,

(iv) Over each $R_{\boldsymbol{\beta}^k}$, $1 \leq k \leq L$, $Q^n(\boldsymbol{\beta})$ is twice differentiable and convex, and strictly convex if the rank of $\boldsymbol{X}_n$ is p.

Proof: See the Appendix. □

Remark 2.1

(i) By discussions above and Lemma 2.2, we see that the domain of $Q^n(\boldsymbol{\beta})$ (the parameter space) is partitioned into at most L pieces and over each piece the graph of $Q^n(\boldsymbol{\beta})$ is that of the quadratic function of the sum of squared residuals. Hence the graph of $Q^n(\boldsymbol{\beta})$ is composed of at most L those components.

(ii) The continuity deduced from $Q^n(\boldsymbol{\beta})$ being the sum of some squared residuals without (i) of Lemma 2.2 might not be flawless. The unified expression for $Q^n(\boldsymbol{\beta})$ around the small neighborhood of $\boldsymbol{\beta}$ such as the one given in (i) of the Lemma 2.2 is indispensable. □

2.3. Existence, uniqueness and equivariance

Theorem 2.1

(i) $\widehat{\boldsymbol{\beta}}_{lst}^n$ exists and is the unique local minima of $Q^n(\boldsymbol{\beta})$ over $R_{\boldsymbol{\beta}^{k_0}}$ for some k_0 $(1 \le k_0 \le L)$.

(ii) Over $R_{\boldsymbol{\beta}^{k_0}}$, $\widehat{\boldsymbol{\beta}}_{lst}^n$ is the solution of the system of equations

$$\sum_{i=1}^{n}(y_i - \boldsymbol{w}_i'\boldsymbol{\beta})\boldsymbol{w}_i \mathbb{1}_i = \mathbf{0}, \tag{2.8}$$

(iii) Over $R_{\boldsymbol{\beta}^{k_0}}$, the unique solution is (assume that $\boldsymbol{X}_n$ has a full rank)

$$\widehat{\boldsymbol{\beta}}_{lst}^n = \boldsymbol{M}_n(\mathbf{Y}_n, \mathbf{X}_n, \widehat{\boldsymbol{\beta}}_{lst}^n, \alpha)^{-1} \sum_{i \in I(\boldsymbol{\beta}^{k_0})} y_i \boldsymbol{w}_i \tag{2.9}$$

Proof: See the Appendix. □

Note that $\boldsymbol{X}_n$ having a full rank is sufficient for the matrix in the theorem to be invertible. The existence could also be established as follows. In the sequel, we will assume that

(A0) there is no vertical hyperplane which contains at least $\lfloor (n+1)/2 \rfloor$ points of $\mathbf{Z}^{(n)}$.

This holds true with probability one if $(\boldsymbol{x}', y)'$ has a joint density or holds if $\mathbf{Z}^{(n)}$ is *in general position* (see Section 3 for definition) (assume that $n > 2p+1$ hereafter).

Theorem 2.2 The minimizer $\widehat{\boldsymbol{\beta}}_{lst}^n$ of $Q(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta}, \alpha)$ defined in (2.4) over $\boldsymbol{\beta} \in \mathbb{R}^p$ always exists for a given $\boldsymbol{Z}^{(n)}$ and an α provided that **(A0)** holds.

Proof: See the Appendix. □

Equivariance A regression estimator $\mathbf{T}$ is called *regression, scale, and affine equivariant* if, respectively (see page 116 of [23]) with $i \in \mathbb{N} := \{1, 2, \cdots, n\}$

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

Theorem 2.3 $\widehat{\boldsymbol{\beta}}_{lst}^n$ is regression, scale, and affine equivariant.

Proof: We have the identities

$$\begin{aligned} y_i + \boldsymbol{w}_i'\mathbf{b} - \boldsymbol{w}_i'(\boldsymbol{\beta}+\mathbf{b}) &= y_i - \boldsymbol{w}_i'\boldsymbol{\beta},\ \forall\ \mathbf{b}\in\mathbb{R}^p \\ sy_i - \boldsymbol{w}_i'(s\boldsymbol{\beta}) &= s(y_i - \boldsymbol{w}_i'\boldsymbol{\beta}),\ \forall\ s\in\mathbb{R}^1 \\ y_i - (A'\boldsymbol{w}_i)'A^{-1}\boldsymbol{\beta} &= y_i - \boldsymbol{w}_i'\boldsymbol{\beta},\ \forall \text{ nonsingular } A\in\mathbb{R}^{p\times p}. \end{aligned}$$

The desired result follows by these identities and the (regression, scale, and affine) invariance (see page 148 of [52] for definition) of $\frac{|r_i - m(\boldsymbol{Z}^{(n)},\ \boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}^{(n)},\ \boldsymbol{\beta})}$. □

3. Robustness of LST

3.1. Finite sample breakdown point

As an alternative to the least-squares, is the LST estimator more robust? The most prevailing quantitative measure of global robustness of any location or regression estimators in the finite sample practice is the *finite sample breakdown point* (FSBP), introduced by [7].

Roughly speaking, the FSBP is the minimum fraction of 'bad' (or contaminated) data points that can force the estimator beyond any bound (becoming useless). For example, in the context of estimating the center of a data set, the sample mean has a breakdown point of $1/n$ (or 0%), because even one bad observation can change the mean by an arbitrary amount; in contrast, the sample median has a breakdown point of $\lfloor (n+1)/2 \rfloor / n$ (or 50%).

Definition 3.1 [7] The finite sample *replacement breakdown point* (RBP) of a regression estimator $\mathbf{T}$ at the given sample $\mathbf{Z}^{(n)} = \{Z_1, Z_2, \cdots, Z_n\}$, where $Z_i := (\boldsymbol{x}_i', y_i)'$, is defined as

$$\text{RBP}(\mathbf{T}, \mathbf{Z}^{(n)}) = \min_{1\le m\le n, m\in\mathbb{N}} \left\{ \frac{m}{n} : \sup_{\mathbf{Z}_m^{(n)}} \|\mathbf{T}(\mathbf{Z}_m^{(n)}) - \mathbf{T}(\mathbf{Z}^{(n)})\| = \infty \right\}, \tag{3.1}$$

where $\mathbf{Z}_m^{(n)}$ denotes an arbitrary contaminated sample by replacing m original sample points in $\mathbf{Z}^{(n)}$ with arbitrary points in $\mathbb{R}^p$. Namely, the RBP of an estimator is the minimum replacement fraction that could drive the estimator beyond any bound. It turns out that both the L_1 (least absolute deviations) and the L_2 (least squares) estimators have RBP $1/n$ (or 0%), the lowest possible value whereas the LTS can have $(\lfloor (n-p)/2 \rfloor + 1)/n$ (or 50%), the highest possible value for any regression equivariant estimators (see pages 124-125 of [23]).

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 $(\boldsymbol{x}', y)'$ 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.

Theorem 3.1 For $\widehat{\boldsymbol{\beta}}_{lst}^n$ defined in (2.5) and $\mathbf{Z}^{(n)}$ in general position, we have

$$\mathrm{RBP}(\widehat{\boldsymbol{\beta}}_{lst}^n, \mathbf{Z}^{(n)}) = \begin{cases} \lfloor (n+1)/2 \rfloor / n, & \text{if } p = 1, \\ (\lfloor n/2 \rfloor - p + 2)/n, & \text{if } p > 1. \end{cases} \tag{3.2}$$

Proof: See the Appendix. □

Remark 3.1

(I) The assumption that $\mathbf{Z}^{(n)}$ is in general position seems to play a central role in the proof. But actually, one can drop it and introduce an index: $c(\mathbf{Z}^{(n)})$ (which is the maximum number of observations from $\mathbf{Z}^{(n)}$ contained in any $(p-1)$ dimensional subspace/hyperplane) to replace p in the derivation of the proof and the final RBP result (when $p > 1$).

(II) Asymptotically speaking (i.e. as $n \to \infty$), $\widehat{\boldsymbol{\beta}}_{lst}^n$ has the best possible asymptotic breakdown point (ABP) 50%, the same as that of the LTS. The RBP of $\widehat{\boldsymbol{\beta}}_{lst}^n$, albeit very high (indeed as high as that of the LMS), is slightly less than that of the LTS (with the best choice of h). However, it can be improved to attain the best possible value if one modifies α so that it is the hth quantile of the n outlyingness of residuals with $h = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ to include exact h squares of residuals in the sum of the RHS of (2.4). □

3.2. Influence function

Throughout $F_{\mathbf{z}}$ stands for the distribution of random vector $\mathbf{z}$ unless otherwise stated. Write $F_{(\boldsymbol{x}',y)}$ for the joint distribution of $\boldsymbol{x}'$ and y in (1.1), $r := r(F_{(\boldsymbol{x}',y)}, \boldsymbol{\beta}) = y - (1, \boldsymbol{x}')\boldsymbol{\beta} := y - \boldsymbol{w}'\boldsymbol{\beta}$.

$$\begin{aligned} m :=& m(F_{(\boldsymbol{x}',y)}, \boldsymbol{\beta}) = \mathrm{Med}(F_r), \\ \sigma :=& \sigma(F_{(\boldsymbol{x}',y)}, \boldsymbol{\beta}) = \mathrm{MAD}(F_r), \end{aligned}$$

hereafter we assume that m and σ exist uniquely. The population counterparts of (2.4) and (2.5) are respectively:

$$Q(F_{(\boldsymbol{x}',y)}, \boldsymbol{\beta}, \alpha) := \int (y - \boldsymbol{w}'\boldsymbol{\beta})^2 \mathbb{1}\left(\frac{|y - \boldsymbol{w}'\boldsymbol{\beta} - m|}{\sigma} \le \alpha \right) dF_{(\boldsymbol{x}',y)}, \tag{3.3}$$

$$\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}',y)}, \alpha) := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} Q(F_{(\boldsymbol{x}',y)}, \boldsymbol{\beta}, \alpha). \tag{3.4}$$

The RBP gauges the global robustness of an estimator at finite sample practice. To assess the local robustness at the population setting, one can use the influence function approach (see [8]), which depicts the local robustness of a functional with an infinitesimal point-mass contamination at a single point $\boldsymbol{z} \in \mathbb{R}^p$.

For a given distribution F defined on $\mathbb{R}^p$ and an $\varepsilon > 0$, the version of F contaminated by an ε amount of an *arbitrary distribution* G on $\mathbb{R}^p$ is denoted by $F(\varepsilon, G) = (1-\varepsilon)F + \varepsilon G$ (ε amount deviation from the assumed F). Hereafter it is assumed that $\varepsilon < 1/2$, otherwise $F(\varepsilon, G) = G((1-\varepsilon), F)$, and one can't distinguish which one is contaminated by which one.

Definition 3.2 [8] The *influence function* (IF) of a functional T at a given point $\boldsymbol{z} \in \mathbb{R}^p$ for a given F is defined as

$$\mathrm{IF}(\boldsymbol{z}; \boldsymbol{T}, F) = \lim_{\varepsilon \to 0^+} \frac{\boldsymbol{T}(F(\varepsilon, \delta_{\boldsymbol{z}})) - \boldsymbol{T}(F)}{\varepsilon}, \tag{3.5}$$

where $\delta_{\boldsymbol{z}}$ is the point-mass probability measure at $\boldsymbol{z} \in \mathbb{R}^p$.

The function $\mathrm{IF}(\boldsymbol{z}; \boldsymbol{T}, F)$ describes the relative influence on $\boldsymbol{T}$ of an infinitesimal point-mass contamination at $\boldsymbol{z}$ and gauges the local robustness of $\boldsymbol{T}$.

It is desirable that a regression estimating functional has a bounded influence function. This, however, does not hold for an arbitrary regression estimating functional (such as the classical least squares functional). Now we investigate this for the functional of the least sum of squares of trimmed residuals, $\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}', y)}, \alpha)$. Put

$$\begin{aligned} F_\varepsilon(\mathbf{z}) :=& F(\varepsilon, \delta_{\mathbf{z}}) = (1-\varepsilon)F_{(\boldsymbol{x}', y)} + \varepsilon \delta_{\mathbf{z}}, \\ m_\varepsilon(\mathbf{z}) :=& m(F_\varepsilon(\mathbf{z}), \boldsymbol{\beta}) = \mathrm{Med}(F_{R_\varepsilon(\mathbf{z})}), \\ \sigma_\varepsilon(\mathbf{z}) :=& \sigma(F_\varepsilon(\mathbf{z}), \boldsymbol{\beta}) = \mathrm{MAD}(F_{R_\varepsilon(\mathbf{z})}), \end{aligned}$$

where $R_\varepsilon(\mathbf{z}) = r(F_\varepsilon(\mathbf{z}), \boldsymbol{\beta}) = t - (1, \mathbf{s}')\boldsymbol{\beta}$, and $F_\varepsilon(\mathbf{z}) =: F_{\boldsymbol{u}}(\mathbf{z})$ with a random vector $\mathbf{u} = (\mathbf{s}', t)' \in \mathbb{R}^p$, $\mathbf{s} \in \mathbb{R}^{p-1}$, and $t \in \mathbb{R}^1$ (i.e., $\boldsymbol{u}$ is the random vector that has the CDF $F_\varepsilon(\boldsymbol{z})$). Hereafter we assume that $m_\varepsilon(\mathbf{z})$ and $\sigma_\varepsilon(\mathbf{z})$ uniquely exist. The versions of (3.3) and (3.4) at the contaminated distribution $F_\varepsilon(\mathbf{z})$ are respectively

$$Q(F_\varepsilon(\mathbf{z}), \boldsymbol{\beta}, \alpha) := \int (t - (1, \mathbf{s}')\boldsymbol{\beta})^2 \mathbb{1}\left(\frac{|(t - (1, \mathbf{s}')\boldsymbol{\beta}) - m_\varepsilon(\mathbf{z})|}{\sigma_\varepsilon(\mathbf{z})} \le \alpha \right) dF_{\boldsymbol{u}}(\mathbf{s}', t), \tag{3.6}$$

$$\boldsymbol{\beta}_{lst}(F_\varepsilon(\mathbf{z}), \alpha) := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} Q(F_\varepsilon(\mathbf{z}), \boldsymbol{\beta}, \alpha). \tag{3.7}$$

Lemma 3.1 $\boldsymbol{\beta}_{lst} := \boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}', y)}, \alpha)$ is regression, scale, and affine equivariant (see [52] for definition).

Proof: It is trivial (analogous to that of Theorem 2.3). □

To investigate the influence function of $\boldsymbol{\beta}_{lst}$ especially the consistency of its sample version in the next section, we first need to establish its existence and uniqueness. We need assumptions: **(A1)** y has a density, and **(A2)** the distribution F_r with $r = y - \boldsymbol{w}'\boldsymbol{\beta}$ is non-flat around $m = \mathrm{Med}(F_r)$ and $\sigma = \mathrm{MAD}(F_r)$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$.

Write $Q(\boldsymbol{\beta})$ for $Q(F_{(\boldsymbol{x}',y)},\boldsymbol{\beta},\alpha)$ in (3.3). We have a population counterpart of Lemma 2.2.

Lemma 3.2 Assume **(A1)-(A2)** hold. Then $Q(\boldsymbol{\beta})$

(i) is continuous in $\boldsymbol{\beta}\in\mathbb{R}^p$;

(ii) is twice differentiable in $\boldsymbol{\beta}\in\mathbb{R}^p$ with (assume that $E(\boldsymbol{x}\boldsymbol{x}')$ exists)

$$\partial^2 Q(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^2 = 2E\boldsymbol{w}\boldsymbol{w}'\mathbb{1}\left(|y-\boldsymbol{w}'\boldsymbol{\beta}-m|/\sigma\le\alpha\right);$$

(iii) is convex in $\boldsymbol{\beta}\in\mathbb{R}^p$ and strictly convex if $E\boldsymbol{w}\boldsymbol{w}'\mathbb{1}\left(|y-\boldsymbol{w}'\boldsymbol{\beta}-m|/\sigma\le\alpha\right)$ is invertible .

Proof: See the Appendix. □

Theorem 3.2 Assume that **(A1)-(A2)** hold and $m(F_\varepsilon(\mathbf{z}),\boldsymbol{\beta})$ and $\sigma(F_\varepsilon(\mathbf{z}),\boldsymbol{\beta})$ are continuous in $\boldsymbol{\beta}$ around a small neighborhood of $\boldsymbol{\beta}_{lst}((F_\varepsilon(\mathbf{z}),\alpha)$. Write $\boldsymbol{v}'=(1,\boldsymbol{s}')$ and let $\boldsymbol{u}$ be the random variable with CDF $F_\varepsilon(\boldsymbol{z})$. We have

(i) $\boldsymbol{\beta}_{lts}(F_{(\boldsymbol{x}',y)},\alpha)$ and $\boldsymbol{\beta}_{lts}(F_\varepsilon(\mathbf{z}),\alpha)$ exist.

(ii) Furthermore, they are the solution of system of equations, respectively

$$\int(y-\boldsymbol{w}'\boldsymbol{\beta})\boldsymbol{w}\mathbb{1}\left(|y-\boldsymbol{w}'\boldsymbol{\beta}-m|/\sigma\le\alpha\right)dF_{(\boldsymbol{x}',y)}(\boldsymbol{x},y)=\mathbf{0}, \tag{3.8}$$

$$\int(t-\boldsymbol{v}'\boldsymbol{\beta})\boldsymbol{v}\mathbb{1}\left(|(t-\boldsymbol{v}'\boldsymbol{\beta})-m_\varepsilon(\mathbf{z})|/\sigma_\varepsilon(\mathbf{z})\le\alpha\right)dF_{\mathbf{u}}(\mathbf{s},t)=\mathbf{0}. \tag{3.9}$$

(iii) $\boldsymbol{\beta}_{lts}(F_{(\boldsymbol{x}',y)},\alpha)$ and $\boldsymbol{\beta}_{lts}(F_\varepsilon(\mathbf{z}),\alpha)$ are unique provided that

$$\int\boldsymbol{w}\boldsymbol{w}'\mathbb{1}\left(|y-\boldsymbol{w}'\boldsymbol{\beta}-m|/\sigma\le\alpha\right)dF_{(\boldsymbol{x}',y)}(\boldsymbol{x},y), \tag{3.10}$$

$$\int\boldsymbol{v}\boldsymbol{v}'\mathbb{1}\left(|(t-\boldsymbol{v}')\boldsymbol{\beta})-m_\varepsilon(\mathbf{z})|/\sigma_\varepsilon(\mathbf{z})\le\alpha\right)dF_{\mathbf{u}}(\mathbf{s},t) \tag{3.11}$$

are respectively invertible.

Proof: See the Appendix. □

Theorem 3.3 If assumptions in theorem 3.2 hold, then for any $\mathbf{z}_0:=(\mathbf{s}'_0,t_0)'\in\mathbb{R}^p$, we have that

$$\dot{\boldsymbol{\beta}}_{lst}(\mathbf{z}_0,F_{(\boldsymbol{x}',y)})=\begin{cases}\mathbf{0}, & \text{if } t_0-(1,\mathbf{s}'_0)\boldsymbol{\beta}_{lst}\notin[m(\boldsymbol{\beta}_{lst})\pm\alpha\sigma(\boldsymbol{\beta}_{lst})]\\ (t_0-(1,\mathbf{s}'_0)\boldsymbol{\beta}_{lst})M^{-1}(1,\mathbf{s}'_0)', & \text{otherwise,}\end{cases}$$

where $\dot{\boldsymbol{\beta}}_{lst}(\mathbf{z}_0,F_{(\boldsymbol{x}',y)})$ stands for the $\mathrm{IF}(\boldsymbol{z}_0;\boldsymbol{\beta}_{lst},F_{(\boldsymbol{x}',y)})$, M^{-1} stands for the inverse of the matrix $E\left(\boldsymbol{w}\boldsymbol{w}'\mathbb{1}\left(|r(\boldsymbol{\beta})-m(F_{r(\boldsymbol{\beta})})|/\sigma(F_{r(\boldsymbol{\beta})})\le\alpha\right)\right)$ with $\boldsymbol{\beta}=\boldsymbol{\beta}_{lst}$, and $[a\pm b]$ stands for $[a-b,a+b]$.

Proof: See the Appendix. □

Remark 3.2 see the Appendix. □

Overall, we see that LST is globally robust with the best possible ABP of 50% and robust locally against point-mass contamination when there are vertical and bad leverage outliers.

Besides robustness, one wonders: does the $\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}',y)}, \alpha)$ really catch the true parameter (i.e. is it Fisher consistent)? And how fast does the sample $\boldsymbol{\beta}_{lst}(Z^{(n)})$ converge to $\boldsymbol{\beta}_{lst}$ (or the true parameter $\boldsymbol{\beta}_0$) (i.e. strong or root-n consistency)? We answer these questions next.

4. Consistency

4.1. Fisher Consistency

Before establishing the strong or root-n consistency, we like to first show that the population version of the LST, $\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}',y)}, \alpha)$, is consistent with (or rather identical to) the true unknown parameter $\boldsymbol{\beta}_0$ under some assumptions - which is called Fisher consistency of the estimation functional. To that end, let us first recall our general model:

$$y = (1, \boldsymbol{x}')\boldsymbol{\beta}_0 + e, \tag{4.1}$$

with its sample version given in model (1.1). In addition to the assumptions given in Theorem 3.2 for the existence and uniqueness of $\boldsymbol{\beta}_{lst}$, we need one more assumption:

(A3) $\boldsymbol{x}$ and e are independent and $E_{(\boldsymbol{x}',y)}\left(e\mathbb{1}\left(|e - m(F_e)|/\sigma(F_e) \le \alpha\right)\right) = 0$. Hereafter we assume that $m(F_e)$ and $\sigma(F_e)$ exist uniquely.

The independence assumption between $\boldsymbol{x}$ and e is typical in the traditional regression analysis. However, one can drop it here by modifying the integration appropriately (see the proof below), and it is unnecessary if $\boldsymbol{x}$ is a non-random covariate (carrier). The assumption that integration equals to zero is very mild, and it automatically holds under the common assumption that the e is symmetric with respect to 0 (that is, $e \overset{d}{=} -e$). We have

Theorem 4.1 Under assumptions **(A1)-(A3)**, $\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}',y)}, \alpha) = \boldsymbol{\beta}_0$ (i.e. it is Fisher consistent) provided that $E\boldsymbol{w}\boldsymbol{w}'\mathbb{1}\left(|e - m(F_e)|/\sigma(F_e) \le \alpha\right)$ is invertible.

Proof: Notice that $y - \boldsymbol{w}'\boldsymbol{\beta} = \boldsymbol{w}'(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + e$. This in conjunction with equation (3.8) yields,

$$\int (\boldsymbol{w}'(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + e)\boldsymbol{w}\mathbb{1}\left(|(\boldsymbol{w}'(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + e) - m|/\sigma \le \alpha\right) dF_{(\boldsymbol{x}',y)} = \mathbf{0},$$

one sees that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ indeed is one solution of the equation system by virtue of **(A3)**. In light of Theorem 3.2 and the uniqueness of the solution, the desired result follows. □

4.2. Strong consistency

To establish the strong consistency of $\widehat{\boldsymbol{\beta}}_{lst}(\mathbf{Z}^{(n)}, \alpha)$ for the $\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}',y)}, \alpha)$, write $\widehat{\boldsymbol{\beta}}_{lst}(F^n_{\mathbf{Z}}) := \widehat{\boldsymbol{\beta}}_{lst}(\mathbf{Z}^{(n)}, \alpha)$, $\boldsymbol{\beta}_{lst}(F_{\mathbf{Z}}) := \boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}',y)}, \alpha)$, $Q(F^n_{\mathbf{Z}}, \boldsymbol{\beta}) := Q(\mathbf{Z}^{(n)}, \boldsymbol{\beta}, \alpha)$, and $Q(F_{\mathbf{Z}}, \boldsymbol{\beta}) := Q(F_{(\boldsymbol{x}',y)}, \boldsymbol{\beta}, \alpha)$, for notation simplicity. where $F^n_{\mathbf{Z}}$ is the sample version of $F_{\mathbf{Z}} := F_{(\boldsymbol{x}',y)}$, corresponding to $\mathbf{Z}^{(n)}$ and α are suppressed.

We will follow the approach in [51] and treat the problem in a more general setting. To that end, we introduce the regression depth functions $D(F^n_{\mathbf{Z}}, \boldsymbol{\beta}) = (1 + Q(F^n_{\mathbf{Z}}, \boldsymbol{\beta}))^{-1}$ and $D(F_{\mathbf{Z}}, \boldsymbol{\beta}) = (1 + Q(F_{\mathbf{Z}}, \boldsymbol{\beta}))^{-1}$ (see page 144 of [52] for the objective function approach). The original minimization issue becomes a maximization problem.

Let M_n be stochastic processes indexed by a metric space Θ of $\boldsymbol{\theta}$, and M: $\Theta \to \mathbb{R}$ be a deterministic function of $\boldsymbol{\theta}$ which has its maximum at a point $\boldsymbol{\theta}_0$.

The sufficient conditions for the consistency of this type of problem were given in [38] and [39], they are:

C1: $\sup_{\boldsymbol{\theta}\in\Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| = o_p(1)$;

C2: $\sup_{\{\boldsymbol{\theta}:\ d(\boldsymbol{\theta},\boldsymbol{\theta_0})\geq\delta\}} M(\boldsymbol{\theta}) < M(\boldsymbol{\theta_0})$, for any $\delta > 0$ and the metric d on Θ;

Then any sequence $\boldsymbol{\theta}_n$ is consistent for $\boldsymbol{\theta}_0$ providing that it satisfies

C3: $M_n(\boldsymbol{\theta_n}) \geq M_n(\boldsymbol{\theta_0}) - o_p(1)$.

Lemma 4.1 [38] If **C1** and **C2** hold, then any $\boldsymbol{\theta}_n$ satisfying **C3** is consistent for $\boldsymbol{\theta}_0$. □

Remark 4.1

(I) C1 requires that the $M_n(\boldsymbol{\theta})$ converges to $M(\boldsymbol{\theta})$ in probability uniformly in $\boldsymbol{\theta}$. For the depth process $\mathrm{D}(F^n_{\mathbf{Z}}, \boldsymbol{\beta})$ and $\mathrm{D}(F_{\mathbf{Z}}, \boldsymbol{\beta})$, it holds true (the convergence here is almost surely (a.s.) and uniformly in $\boldsymbol{\beta}$ as shown in Lemma 4.2 below).

(II) C2 essentially demands that the unique maximizer $\boldsymbol{\theta_0}$ is well separated. This holds true for $\mathrm{D}(F_{\mathbf{Z}}, \boldsymbol{\beta})$ as shown in Lemma 4.3 below.

(III) C3 asks that $\boldsymbol{\theta_n}$ is very close to $\boldsymbol{\theta}_0$ in the sense that the difference of images of the two at M_n is within $o_p(1)$. In [10] and [39] a stronger version of **C3** is required:

$$\mathbf{C3}^*: \quad M_n(\boldsymbol{\theta_n}) \geq \sup_{\theta\in\Theta} M_n(\boldsymbol{\theta}) - o_p(1),$$

which implies **C3**. This strong version mandates that $\boldsymbol{\theta_n}$ *nearly* maximizes $M_n(\boldsymbol{\theta})$. Our maximum regression depth estimator $\widehat{\boldsymbol{\beta}}_{lst}(F_Z^n, \alpha)(:= \boldsymbol{\theta_n})$ is defined to be the maximizer of $M_n(\boldsymbol{\theta}) := D(F_{\mathbf{Z}}^n, \boldsymbol{\beta})$, hence $\mathbf{C3}^*$ (and thus **C3**) holds automatically. □

In light of above, we have

Corollary 4.1 $\widehat{\boldsymbol{\beta}}_{lst}(F_{\mathbf{Z}}^n)$ induced from $\mathrm{D}(F_{\mathbf{Z}}^n, \boldsymbol{\beta})$ (or $Q(F_{\mathbf{Z}}^n, \boldsymbol{\beta})$) is consistent for $\boldsymbol{\beta}_{lst}(F_{\mathbf{Z}})$. □

But, we can have more.

Based on the proofs of Theorems 2.2 and 3.2 and in light of Theorem 4.1, under assumptions **(A0)-(A3)**, we assume without loss of generality (w.l.o.g.) that $\widehat{\boldsymbol{\beta}}_{lst}(F_{\mathbf{Z}}^n) \in B(\boldsymbol{\beta}_0, r)$ and $\boldsymbol{\beta}_{lst}(F_{\mathbf{Z}}) \in B(\boldsymbol{\beta}_0, r)$, where $B(\boldsymbol{\beta}_0, r)$ is a ball centered at $\boldsymbol{\beta}_0$ with radius r which is large enough. Now $B(\boldsymbol{\beta}_0, r)$ can serve, w.l.o.g., as out parameter space Θ of $\boldsymbol{\beta}$ in the sequel.

Lemma 4.2 Under assumption **(A2)**, (a) $\sup_{\boldsymbol{\beta}\in\Theta} |Q(F_{\mathbf{Z}}^n, \boldsymbol{\beta}) - Q(F_{\mathbf{Z}}, \boldsymbol{\beta})| = o(1)$, a.s. and (b) $\sup_{\boldsymbol{\beta}\in\Theta} |D(F_{\mathbf{Z}}^n, \boldsymbol{\beta}) - D(F_{\mathbf{Z}}, \boldsymbol{\beta})| = o(1)$, a.s..

Proof: See the Appendix. □

Lemma 4.3 Assume that a regression (or location) depth function $D(\boldsymbol{\beta}; F_{\mathbf{Z}})$ is continuous in $\boldsymbol{\beta}$ and $\boldsymbol{\beta} \in \Theta$ is bounded. Let $\boldsymbol{\eta} \in \Theta$ be the unique point with $\boldsymbol{\eta} = \arg\max_{\boldsymbol{\beta}\in\Theta} D(\boldsymbol{\beta}; F_{\mathbf{Z}})$ and $D(\boldsymbol{\eta}; F_{\mathbf{Z}}) > 0$. Then $\sup_{\boldsymbol{\beta}\in N_\varepsilon^c(\boldsymbol{\eta})} D(\boldsymbol{\beta}; F_{\mathbf{Z}}) < D(\boldsymbol{\eta}; F_{\mathbf{Z}})$, for any $\varepsilon > 0$, where $N_\varepsilon^c(\boldsymbol{\eta}) = \{\boldsymbol{\beta} \in \Theta : \|\boldsymbol{\beta} - \boldsymbol{\eta}\| \geq \varepsilon\}$ and "A^c" stands for "complement" of the set A.

Proof: See the Appendix. □

Theorem 4.2 Under assumptions **(A1) -(A3)**, $\widehat{\boldsymbol{\beta}}_{lst}(F_{\mathbf{Z}}^n)$ is *strongly* consistent for $\boldsymbol{\beta}_{lst}(F_{\mathbf{Z}})$ (i.e., $\widehat{\boldsymbol{\beta}}_{lst}^n - \boldsymbol{\beta}_{lst} = o(1)$ a.s.).

Proof: The proof for the consistency of Lemma 4.1 could be easily extended to the strong consistency with a strengthened version of **C1**

$$\mathbf{C1}^*: \quad \sup_{\boldsymbol{\theta}\in\Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| = o(1), \text{ a.s.},$$

In the light of the proof of Lemma 4.1, we need only verify the sufficient conditions $\mathbf{C1}^*$ and **C2-C3**. By (III) of Remark 4.1, **C3** holds automatically, so we need to verify $\mathbf{C1}^*$ and **C2**. $\mathbf{C1}^*$ follows from Lemma 4.2. So the only item left is to verify **C2** for $\mathrm{D}(F_Z, \boldsymbol{\beta})$ which is guaranteed by Lemma 4.3. □

Remark 4.2

(I) The approach utilizing a generalized Glivenko-Cantelli theorem over a

class of functions with polynomial discrimination in the proof of lemma 4.2 is very powerful and applicable to many regression estimators to obtain the strong consistency result. It is certainly applicable to the least trimmed squares (LTS).

(II) The consistency (not the strong version in Theorem 4.2) of the LTS has been obtained in [40] using standard analysis (under many assumptions on non-random $\boldsymbol{x}_i$ and on the distribution of e) which, of course, is difficult, lengthy (consumed an entire article), and tedious. The approach here is different, concise and the estimator (LST) is, of course, different to the LTS. □

Consistency does not reveal the speed of convergence of sample $\widehat{\boldsymbol{\beta}}_{lst}(F_{\mathbf{Z}}^n)$ to its population counterpart $\boldsymbol{\beta}_{lst}(F_{\mathbf{Z}})$. Standard speed of $O_p(1/\sqrt{n})$ is desirable and expected for $\widehat{\boldsymbol{\beta}}_{lst}(F_{\mathbf{Z}}^n)$. We investigate this issue next.

4.3. $\sqrt{n}$- consistency

To establish the root-n consistency we need one more assumption:

(A4) $\mathrm{E}(e) = 0$ and $\mathrm{E}(\boldsymbol{x}\boldsymbol{x}')$ exists.

$\mathrm{E}(e) = 0$ is commonly required in the traditional regression analysis. The existence of covariance (and the mean) of $\boldsymbol{x}$ is sufficient for the existence of $\mathrm{E}(\boldsymbol{x}\boldsymbol{x}')$.

In the following, we will employ big O and little o notations for the vectors or matrices.

Definition 4.1 For a sequence of random vectors or matrices $\boldsymbol{X}_n$, we say
$\boldsymbol{X}_n = o_p(1)$ means $\|\boldsymbol{X}_n\| \xrightarrow{p} 0$;
$\boldsymbol{X}_n = O_p(1)$ means $\|\boldsymbol{X}_n\| = O_p(1)$,
where norm of a matrix $A_{m\times n}$ is defined as $\|A\| := \sup_{\boldsymbol{x}\neq 0\in\mathbb{R}^n} \|A\boldsymbol{x}\|_p/\|\boldsymbol{x}\|_p$, p could be $1, 2,$ or ∞ (see page 82 of [3]). □

Theorem 4.3 Assume that assumptions in Theorem 4.1 and **(A4)** hold, then $\widehat{\boldsymbol{\beta}}_{lst}^n - \boldsymbol{\beta}_{lst} = \widehat{\boldsymbol{\beta}}_{lst}^n - \boldsymbol{\beta}_0 = O_p(1/\sqrt{n})$.

Proof: See the Appendix. □

Remark 4.4

(I) The root-n consistency of an arg max estimator could be established by a general approach given in [30, 31] Theorem 1. With the depth process introduced in the section 4.2, we are unable to verify the second requirement in that theorem though.

(II) The approach here for the root-n consistency of the LST is analogous to what is given in [41] for the LTS. However, the latter is lengthy and costs a

twenty-two pages article. □

5. Asymptotic normality

The root-n consistency above could be obtained as a by-product of the asymptotic normality which will be established in the following via stochastic equicontinuity (see page 139 of [20], or the supplementary of [51]).

Stochastic equicontinuity refers to a sequence of stochastic processes $\{Z_n(t) : t \in T\}$ whose shared index set T comes equipped with a semi metric $d(\cdot,\cdot)$.

Definition 5.1 (IIV. 1, Def. 2 of [20]). Call Z_n stochastically equicontinuous at t_0 if for each $\eta > 0$ and $\epsilon > 0$ there exists a neighborhood U of t_0 for which

$$\limsup P\left(\sup_U |Z_n(t) - Z_n(t_0)| > \eta\right) < \epsilon. \tag{5.1}$$

□

If τ_n is a sequence of random elements of T that converges in probability to t_0, then

$$Z_n(\tau_n) - Z_n(t_0) \to 0 \text{ in probability}, \tag{5.2}$$

because, with probability tending to one, τ_n will belong to each U. The form above will be easier to apply, especially when behavior of a particular τ_n sequence is under investigation.

Suppose $\mathscr{F} = \{f(\cdot,t) : t \in T\}$, with T a subset of $\mathbb{R}^k$, is a collection of real, P-integrable functions on the set S where P (probability measure) lives. Denote by P_n the empirical measure formed from n independent observations on P, and define the empirical process E_n as the signed measure $n^{1/2}(P_n - P)$. Define

$$F(t) = Pf(\cdot,t),$$
$$F_n(t) = P_n f(\cdot,t).$$

Suppose $f(\cdot,t)$ has a linear approximation near the t_0 at which $F(\cdot)$ takes on its minimum value:

$$f(\cdot,t) = f(\cdot,t_0) + (t - t_0)'\nabla(\cdot) + |t - t_0| r(\cdot,t). \tag{5.3}$$

For completeness set $r(\cdot,t_0) = 0$, where ∇ (differential operator) is a vector of k real functions on S. We cite theorem 5 of IIV.1 of [20] (page 141) for the asymptotic normality of τ_n.

Lemma 5.1 . Suppose $\{\tau_n\}$ is a sequence of random vectors converging in probability to the value t_0 at which $F(\cdot)$ has its minimum. Define $r(\cdot,t)$ and the vector of functions $\nabla(\cdot)$ by (5.3). If

(i) t_0 is an interior point of the parameter set T;

(ii) $F(\cdot)$ has a non-singular second derivative matrix V at t_0;
(iii) $F_n(\tau_n) = o_p(n^{-1}) + \inf_t F_n(t)$;
(iv) the components of $\nabla(\cdot)$ all belong to $\mathscr{L}^2(P)$;
(v) the sequence $\{E_n r(\cdot, t)\}$ is stochastically equicontinuous at t_0 ;

then

$$n^{1/2}(\tau_n - t_0) \xrightarrow{d} \mathcal{N}(O, V^{-1}[P(\nabla\nabla') - (P\nabla)(P\nabla)']V^{-1}).$$

□

Theorem 5.1 Assume that

(i) the uniqueness assumptions for $\widehat{\boldsymbol{\beta}}_{lst}^n$ and $\boldsymbol{\beta}_{lst}$ in theorems 2.1 and 3.2 hold respectively;
(ii) $P(x_i^2)$ exists;

then

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_{lst}^n - \boldsymbol{\beta}_{lst}) \xrightarrow{d} \mathcal{N}(O, V^{-1}[P(\nabla\nabla') - (P\nabla)(P\nabla)']V^{-1}),$$

where $\boldsymbol{\beta}$ in V and ∇ is replaced by $\boldsymbol{\beta}_{lst}$ (which could be assumed to be zero).

Proof: See the Appendix. □

Assume that $\boldsymbol{z} = (\boldsymbol{x}', y)'$ follows elliptical distributions $E(g; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with density

$$f_{\boldsymbol{z}}(\boldsymbol{x}', y) = \frac{g(((\boldsymbol{x}', y)' - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}((\boldsymbol{x}', y)' - \boldsymbol{\mu}))}{\sqrt{\det(\boldsymbol{\Sigma})}}, \tag{5.4}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}$ a positive definite matrix of size p which is proportional to the covariance matrix if the latter exists. We assume the function g to have a strictly negative derivative, so that the $f_{\boldsymbol{z}}$ is unimodal.

In light of Lemma 3.1 and under some transformations (see the Appendix in the supplementary material), we can assume, w.l.o.g. that $(\boldsymbol{x}', y)$ follows an $E(g; \boldsymbol{0}, \boldsymbol{I}_{p\times p})$ distribution and $\boldsymbol{I}_{p\times p}$ is the covariance matrix of $(\boldsymbol{x}', y)$ hereafter.

Corollary 5.1 Assume that

(i) assumptions of Theorem 5.1 hold;
(ii) $e \sim \mathcal{N}(0, \sigma^2)$ and $\boldsymbol{x}$ are independent.

Then

(1) $P\nabla = \boldsymbol{0}$ and $P(\nabla\nabla') = 8\sigma^2 C \boldsymbol{I}_{p\times p}$,
with $C = \Gamma(1/2, 1)(\alpha c/\sigma)$, where $c = \sigma\Phi^{-1}(3/4)$, $\Gamma(1/2, 1)(x)$ is the cumulative distribution function (CDF) of random variable $\Gamma(a, b)$ which has a pdf: $\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}$, and $\Phi(x)$ is the CDF of $\mathcal{N}(0, 1)$.

(2) $\mathbf{V} = 2C_1\boldsymbol{I}_{p\times p}$ with $C_1 = 2 * \Phi(\alpha c/\sigma) - 1$.

(3) $n^{1/2}(\widehat{\boldsymbol{\beta}}_{lst}^n - \boldsymbol{\beta}_{lst}) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \frac{2C\sigma^2}{C_1^2}\boldsymbol{I}_{p\times p})$.

Proof: By Theorem 4.1 and Lemma 3.1, we can assume, w.l.o.g., that $\boldsymbol{\beta}_{lst} = \boldsymbol{\beta}_0 = \mathbf{0}$. Utilizing the independence between e and $\boldsymbol{x}$ and Theorem 5.1, a straightforward calculation leads to the results. □

6. Computation

Now we address one of the most important topics on robust regression estimation, that is, the computation of the estimator. Unlike the LS estimator, which has an analytical formula for the computation, for the LST estimator, we do not have such a formula. The formula given in (2.9) can not serve our purpose (due to the circular dependency: the RHS depends on the LHS). For small sample size n and dimension p, one can compute the LST exactly (the L in Theorem 2.1 is not a big number), but that is not affordable for moderate n and p. That is, generally, we have to appeal to approximate algorithms (AAs).

6.1. A procedure based Theorem 2.1

In light of Theorem 2.1, if one discovers all $R_{\boldsymbol{\beta}^k}$s for $1 \le k \le L$, then one can get the exact result. But in practice for some cases, this might not be computationally affordable. However, one can simply search as many $R_{\boldsymbol{\beta}^k}$s as possible to get a good approximation of the estimate $\widehat{\boldsymbol{\beta}}^n_{lst}$.

To identify $R_{\boldsymbol{\beta}^k}$ is equivalent to identifying $i_1, \cdots, i_K$ so that $D_{i_1} < D_{i_2} < \cdots, D_{i_K}$ in light to (2.7), where $K = |I(\boldsymbol{\beta}^k)|$. The latter is equivalent to finding a $\boldsymbol{\beta} \in R_{\boldsymbol{\beta}^k}$, then one gets the desired $i_1, \cdots, i_K$. To find the desired $\boldsymbol{\beta}$, one way is to find a $\overline{\boldsymbol{\beta}}$ on the common boundary of $R_{\boldsymbol{\beta}^k}$ and $R_{\boldsymbol{\beta}^l}$ so that there are $i \neq j$, $D_i = D_j$ for some $1 \le l \neq k \le L$ and $1 \le i, j \le n$. Small perturbation of the coordinates of the $\overline{\boldsymbol{\beta}} = (\beta_1, \cdots, \beta_p)'$ leads to more than one $\boldsymbol{\beta}$s ($\boldsymbol{\beta} = (\beta_1, \cdots, \beta_j \pm \delta, \cdots, \beta_p)'$ (for some $1 \le j \le p$ and $\delta > 0$) that belong to $R_{\boldsymbol{\beta}^k}$ or $R_{\boldsymbol{\beta}^l}$.

Now we address the way to find out the $\overline{\boldsymbol{\beta}}$. In light of (2.7), there are $i \neq j$, $D_i = D_j$ for some $1 \le l \neq k \le L$ and $1 \le i, j \le n$. The equality $D_i = D_j$ implies that (i) $r_i = r_j$ or (ii) $(r_i + r_j)/2 = m_n(\boldsymbol{\beta})$. Both equalities could lead to some $\overline{\boldsymbol{\beta}}$s, but the first one $r_i = r_j$ is more convenient.

We now focus on the first one which amounts to $y_i - y_j = (\boldsymbol{w}_i - \boldsymbol{w}_j)'\boldsymbol{\beta} = (\boldsymbol{x}_i - \boldsymbol{x}_j)'(\beta_2, \cdots, \beta_p)'$, where $\boldsymbol{w}' = (1, \boldsymbol{x}')$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$. Assume that $\boldsymbol{x}_i \neq \boldsymbol{x}_j$ for $i \neq j$. If $y_i = y_j$, then, $\boldsymbol{\beta} = (\beta_1, \mathbf{0}'_{p-1})'$ is one of solutions, otherwise, from this equation, we see that (i) β_1 could be any number in $\mathbb{R}^1$, (ii) the equation defines a $(p-1)$-dimensional hyperplane. Consequently, all $\boldsymbol{\beta} = (\beta_1, 0, \cdots, 0, \frac{y_i - y_j}{x_{ik} - x_{jk}}, 0, \cdots, 0) \in \mathbb{R}^p$ are solutions, where $\beta_1 \in \mathbb{R}^1$ and $x_{ik} \neq x_{jk}$, $1 \le k \le (p-1)$. Simple choices for β_1 could be 0 and 1 or any constant. From here we obtain at least two $\boldsymbol{\beta}$s that lie on the common boundary.

With the small perturbation ($\pm\delta$) to the ith coordinate of the $\boldsymbol{\beta}$s above we could obtain $4p$ new $\boldsymbol{\beta}$s. For each such $\boldsymbol{\beta}$, we first obtain $i_1, \cdots, i_K$ with $K = |I(\boldsymbol{\beta})|$ and then check if the strict inequalities in (2.7) hold.

If they do not hold, then move to the next $\boldsymbol{\beta}$. Otherwise, check if the K indices already appear before, if it has, then do nothing, else update the data structure that stores the indices, and obtain the least square solution $\boldsymbol{\beta}_{ls}$-new based on the sub-data set with the K subscripts ($I(\boldsymbol{\beta})$) and the sum of squared residuals. If the latter is smaller than SS-min, then set it to be the SS-min and update $\widehat{\boldsymbol{\beta}}^n_{lst}$ with $\boldsymbol{\beta}_{ls}$-new. Increase T_{ls}, which is the counter for the number of LS calculations, by one. Move to the next $\boldsymbol{\beta}$ until all $4p$ $\boldsymbol{\beta}$s are exhausted. Then repeat the entire process with a new pair (i, j). Summarizing discussions so far, we have

AA1– pseudocode for computing the LST based on Theorem 2.1

Input: A data set $\boldsymbol{Z}^{(n)} = \{(\boldsymbol{x}_i', y_i)', i = 1, 2, \cdots, n\}$, a fixed α. Assume that $\boldsymbol{x}_i \neq \boldsymbol{x}_j$ if $i \neq j$.

(1) Sample two indices i and j from $\{1, \cdots, n\}$, assume that $x_{ik} \neq x_{jk}$, $1 \leq k \leq (p-1)$ (i.e. the kth coordinates of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ do not equal). Consider

$$\boldsymbol{\beta}^0 = (0, 0, \cdots, 0, b_{k+1}, 0, \cdots, 0)', \boldsymbol{\beta}^1 = (1, 0, \cdots, 0, b_{k+1}, 0, \cdots, 0)' \text{ in } \mathbb{R}^p$$

Both have the same $(k+1)$th coordinate, $b_{k+1} := (y_i - y_j)/(x_{ik} - x_{jk})$.

(2) Write $\boldsymbol{\beta}^j(l, \pm\delta)$ for the perturbed $\boldsymbol{\beta}^j$ with its lth coordinate adding or subtracting a $\delta > 0$. Define a set

$$S(\boldsymbol{\beta}) = \cup_{l=1}^p \{\boldsymbol{\beta}^0(l, \pm\delta)\} \cup_{l=1}^p \{\boldsymbol{\beta}^1(l, \pm\delta)\}.$$

(3) For each $\boldsymbol{\beta}$ of $4p$ $\boldsymbol{\beta}$s in the set $S(\boldsymbol{\beta})$,

(a) obtain $i_1, \cdots, i_K$ with $K = |I(\boldsymbol{\beta})|$ and check to see if the strict inequalities in (2.7) hold.

(a1) If not, move to the next $\boldsymbol{\beta}$; else

(a2) check if the K indices already appear in a structure S_{ind}

(i) if yes, then move to the next $\boldsymbol{\beta}$; else

(ii) update S_{ind} by storing the K indices in the structure S_{ind} and calculate the LS estimate $\boldsymbol{\beta}_{ls}$-new based on the sub-data set with index in $I(\boldsymbol{\beta})$ and obtain the sum of $|I(\boldsymbol{\beta})|$ squared residuals, SS($\boldsymbol{\beta}_{ls}$-new).

(iii) Update SS_{min} if it is greater than SS($\boldsymbol{\beta}_{ls}$-new) and update $\widehat{\boldsymbol{\beta}}^n_{lst}$ with $\boldsymbol{\beta}_{ls}$-new. Update the counter for the total number T_{ls} of LS calculations, if the latter is less than N (the total

number of LS calculations decided to perform), then continue the loop (go to **(3)**), else stop.

(b) If $T_{ls} < N$, then go to **(1)**, else break the loop.

Output: $\widehat{\boldsymbol{\beta}}^n_{lst}$

Remark 6.1 see the Appendix. □

6.2. A subsampling procedure

Subsampling procedures are prevailing in practice for most robust regression estimators (see [23], [11], [12], [24], [43], [25, 26], [50, 54], among others).

The basic idea is straightforward: (1) draw a sub-sample of size m from data set $\mathbf{Z}^{(n)} = \{(\boldsymbol{x}'_i, y_i)' \in \mathbb{R}^p, \boldsymbol{x}_i \in \mathbb{R}^{p-1}, i = 1, 2, \cdots, n\}$. (2) compute an estimate based on the sub-sample and obtain the objective function value. (3) if the objective function value can be further improved (reduced), then go to (1), otherwise, stop and output the final step estimate.

Natural questions for the above procedure include (1) how to guarantee the convergence of the procedure and the final answer is the global minimum? (2) what is the exact size m and what is the relationship with n and dimension p? To better address these matters, we first propose the corresponding procedure for our LST.

AA2 pseudocode for a sub-sampling procedure for LST

Input: A data set $\mathbf{Z}^{(n)} = \{\mathbf{Z}_1, \cdots, \mathbf{Z}_n\} = \{(\boldsymbol{x}'_i, y_i)', i = 1, 2, \cdots, n\} \in \mathbb{R}^p$ (assume that $p \geq 2$) and an $\alpha \geq 1$ (default is one).

(a) **Initialization**: N=min$\{\binom{n}{p}, 300(p-1)\}$, R=0, $Q_{old} = 10^8$, $\boldsymbol{\beta}_{old} = \mathbf{0}$ (or a LS (or LTS) estimate).

(b) **Iteration**: while $(R \leq N)$

keep sampling p indices $\{i_1, \cdots, i_p\}$ from $\{1, 2, \cdots, n\}$ (without replacement) until $M'_{\boldsymbol{x}} := (\boldsymbol{w}_{i_1}, \cdots, \boldsymbol{w}_{i_p})$ being invertible. Let $\boldsymbol{\beta}_{new} = (M_{\boldsymbol{x}})^{-1}(y_{i_1}, \cdots, y_{i_p})'$.

(1) Calculate $I(\boldsymbol{\beta}_{new})$ (based on (2.6)) and $Q_{new} := Q^n(\boldsymbol{\beta}_{new})$ (based on (2.4)).

(2)
- If $Q_{new} < Q_{old}$, then $Q_{old} = Q_{new}$, $\boldsymbol{\beta}_{old} = \boldsymbol{\beta}_{new}$. Get an LS estimator $\boldsymbol{\beta}_{ls}$ based on the data points of $\mathbf{Z}^{(n)}$ with subscripts from $I(\boldsymbol{\beta}_{new})$. Go to (1) with $\boldsymbol{\beta}_{new} = \boldsymbol{\beta}_{ls}$.
- Else if $Q_{new} = Q_{old}$ break
  else R=R+1, go to (b)

Output: $\boldsymbol{\beta}_{new}$. □

Remark 6.2 see the Appendix. □

Table entries (a, b) are: a:=empirical mean squared error, b:=total time consumed

n	p	AA1	AA2
50	3	(0.3499, 566.49)	(0.5290, 651.25)
	5	(0.5817, 457.49)	(0.7645, 861.75)
	10	(0.5390, 682.41)	(1.7177, 1016.6)
100	3	(0.1755, 573.07)	(0.3619, 879.01)
	5	(0.2023, 638.76)	(0.4528, 1042.6)
	10	(0.2576, 702.02)	(0.7000, 1071.5)
200	3	(0.0825, 619.75)	(0.3025, 1309.7)
	5	(0.1055, 676.63)	(0.3501, 1285.6)
	10	(0.1283, 698.14)	(0.4178, 1310.2)

TABLE 1
Total computation time for all 1000 *samples (seconds) and empirical mean squared error (EMSE) of different AAs for various ns and ps.*

7. Examples and comparison

This section investigates the performance of AAs and compares it with that of the benchmark LTS. First, we like to give some guidance for selection among the two AAs.

Example 7.1 Performance of the two AAs There are two AAs and which of them should be recommended for users? This example tries to achieve this by examining the speed and accuracy of the two AAs.

We generate 1000 samples $\mathbf{Z}^{(n)} = \{(\boldsymbol{x}_i', y_i)', i \in \{1, \cdots, n\}, \boldsymbol{x}_i \in \mathbb{R}^{p-1}\}$ from the standard Gaussian distribution for various sample size n and dimension p. For the speed, we calculate the *total time* consumed for all 1000 samples (dividing it by 1000, one gets the average time consumed per sample) by different AAs. For accuracy (or variance, or efficiency), we will compute their empirical mean squared error (EMSE).

For a general estimator $\mathbf{T}$, if it is regression equivariant, then we can assume (w.l.o.g.) that the true parameter $\boldsymbol{\beta}_0 = \mathbf{0} \in \mathbb{R}^p$. We calculate EMSE $:= \sum_{i=1}^{R} \|\mathbf{T}_i - \boldsymbol{\beta}_0\|^2/R$, the empirical mean squared error (EMSE) for $\mathbf{T}$, where $R = 1000$, $\boldsymbol{\beta}_0 = (0, \cdots, 0)' \in \mathbb{R}^p$, and $\mathbf{T}_i$ is the realization of $\mathbf{T}$ obtained from the ith sample with size n and dimension p. The EMSE and the total time consumed (in seconds) by different AAs are listed in Table 1.

Inspecting Table 1 immediately reveals that (i) AA2 is not only the slowest but is the most inaccurate (with the largest EMSEs) in all cases considered. (ii) AA1 has both speed and accuracy advantages for all cases considered.

Overall, we recommend AA1 for users. That does not exclude the potential of improvement of AA2 via the idea in [26]. □

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

The data points in the example above are perfect standard normal and hence are not practically realistic. In the following, we will investigate the performance of AA1 versus the LTS for contaminated standard normal data sets and for moderate as well as large ns and ps.

Example 7.2 Multiple regression with contaminated normal data sets. Now we consider data with contamination, which is typical for big data sets in the "big-data era".

We consider the contaminated highly correlated normal data points scheme. We generate 1000 samples $\mathbf{Z}_i = (\boldsymbol{x_i}', y_i)'$ with various ns from the normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is a zero-vector in $\mathbb{R}^p$, and $\boldsymbol{\Sigma}$ is a p by p matrix with diagonal entries being 1 and off-diagonal entries being 0.9. Then ε% of them are contaminated by normal points with $\boldsymbol{\mu}$ being the p-vector with all elements being 7 except the last one being -2 and the covariance matrix being diagonal with diagonal being 0.1 and off-diagonal being zero. The results are listed in Table 2 .

Normal data sets, each with ε% contamination
Table entries (a, b) are: a:=empirical mean squared error, b:=total time consumed

		$\varepsilon = 5\%$		$\varepsilon = 10\%$	
p	n	AA1	ltsReg	AA1	ltsReg
	100	(0.2971, 9.6581)	(0.3010, 22.867)	(0.2843, 494.01)	(0.2942, 25.289)
5	200	(0.2503, 26.045)	(0.2650, 41.861)	(0.2517, 26.629)	(0.2630, 43.504)
	300	(0.2396, 54.100)	(0.2551, 63.639)	(0.2366, 54.885)	(0.2534, 63.522)
	400	(0.1335, 1085.6)	(0.1394, 181.18)	(0.1340, 1056.2)	(0.1382, 175.92)
10	500	(0.1280, 1207.7)	(0.1321, 222.81)	(0.1289, 1178.5)	(0.1321, 218.94)
	600	(0.1247, 1308.4)	(0.1285, 152.47)	(0.1253, 1273.6)	(0.1276, 149.99)
	700	(0.0815, 2044.9)	(0.0885, 549.61)	(0.0838, 1994.0)	(0.0882, 547.53)
20	800	(0.0776, 2261.7)	(0.0837, 620.63)	(0.0796, 2177.0)	(0.0837, 616.87)
	900	(0.0748, 2436.1)	(0.0804, 541.20)	(0.0761, 2353.7)	(0.0795, 538.43)
		$\varepsilon = 30\%$		$\varepsilon = 40\%$	
	300	(0.4347, 53.248)	(1.9236, 1635.1)	(0.4352, 56.430)	(1.3517, 1712.8)
40	400	(0.3362, 100.04)	(1.2604, 2401.5)	(0.3314, 102.81)	(0.8995, 2399.5)
	500	(0.2594, 147.66)	(0.9514, 2963.4)	(0.2873, 146.67)	(0.6851, 2787.7)
	300	(0.5242, 58.736)	(2.7826, 2861.8)	(0.5700, 59.903)	(1.9808, 2896.3)
50	400	(0.4085, 89.897)	(1.7562, 3292.0)	(0.4539, 108.88)	(1.2547, 3925.5)
	500	(0.3107, 145.84)	(1.2870, 4510.5)	(0.3406, 145.75)	(0.9086, 4419.6)

TABLE 2

Total computation time for all 1000 *samples (seconds) and empirical mean squared error (EMSE) of the LST(AA1) versus the LTS(ltsReg) for various ns, ps, and contaminations.*

Inspecting the table reveals that (i) in terms of EMSE, the AA1 is the overall winner (with the smallest EMSE in all cases considered), the LTS has the largest EMSE in all the cases; (ii) in terms of speed, the LTS (or rather ltsReg) is the winner when $p = 10$ or 20. The AA1 is the winner for all other p's, except when $p = 5$, $n = 100$ and $\varepsilon = 10\%$. For the latter case, the AA1 can still be faster if tuning T_{ls} to be 1, then one gets $(0.2986, 10.396)$ for AA1 versus $(0.2948, 23.133)$ for ltsReg (suffering a slight increase in EMSE). □

The LTS (or lstReg) demonstrates its well-known speedy advantage, which is partially due to its background computation via Fortran subroutine and the computation scheme proposed in [26]. The AA1 (a pure R programming procedure), on the other hand, has the potential to speed up via Rcpp or even via Fortran in one or more order of magnitude.

Remark 7.1

(I) Parameters tuning Two parameters in AA1 that can be tuned. The T_{ls} is set to be 300 for better EMSE (as in the $p = 5$, $n = 100$, and $\varepsilon = 10\%$ case). If tuning it to be 1, one gets a much faster AA1 (as in the cases $p = 30$ 40, and $p = 5$, except when $n = 100$, and $\varepsilon = 10\%$). For the α in the definition of the LST, it is set to be 1 (default value) in Table 1, it is set to be 3 as in Table 2 when there are contaminations (or outliers). Note that theoretically speaking, both the LST and the LTS can resist 50% contamination without breakdown. So 40% contamination rate in Table 2 is relevant which is also employed in [26].

(II) The LTS estimate is obtained via R package ltsReg, h is the default value $\lfloor (n+p+1)/2 \rfloor$, one might tune this h to get better performance from the LTS. But this will decrease LTS's finite sample breakdown value. This is not the case for the LST with the α (see Theorem 3.1). □

So far we have assumed that the true $\boldsymbol{\beta}_0$ is the zero vector based on the regression equivariance. One might not be used to this assumption.

Example 7.3 Performance of the LST and the LTS with respect to a given $\boldsymbol{\beta}_0$. Now we examine the performance of the three regression estimators the LST, the LTS, and LMS in a slightly different setting. We generate 1000 samples $\{(\boldsymbol{x}_i', y_i)' \in \mathbb{R}^p\}$ with a fixed sample size 100 from an assumed model: $y_i = \boldsymbol{\beta_0}'\boldsymbol{x}_i + e_i$, where $\boldsymbol{x}_i = (1, x_{i1}, \cdots, x_{ip-1})'$ and $\boldsymbol{\beta_0} = (\beta_0, \cdots, \beta_{p-1})'$ are in $\mathbb{R}^p$ and x_{ij} and ε_i are independently from either the Cauchy or $\mathcal{N}(0,1)$ distribution.

We list the total time consumed (in seconds) and the EMSE (the same formula as before but the true $\boldsymbol{\beta}_0$ is the given one no longer being the zero vector) for the three methods with respect to different $\boldsymbol{\beta_0}$'s in Table 3. **Case I** $\boldsymbol{\beta_0} = (-2, 0.1, 1)'$, all x_{ij} and e_i are from $\mathcal{N}(0,1)$ distribution. **Case II** $\boldsymbol{\beta_0} = (-2, 0.1, 1, 5)'$, x_{i1}, x_{i2}, and e_i are from $\mathcal{N}(0,1)$ and x_{i3} is from Cauchy distribution. **Case III** $\boldsymbol{\beta_0} = (50, 0.1, -2, 15, 100)'$, all x_{ij} and e_i are from $\mathcal{N}(0,1)$.

Replication 1000 times, $n = 100$

Performance criteria	LST(AA1)	LMS(lmsreg)	LTS(ltsReg)
	Case I	$p = 3$	
EMSE	3.525451	4.204053	3.806951
Total time consumed	11.53858	10.49865	17.81713
	Case II	$p = 4$	
EMSE	29.91539	30.23814	29.97682
Total time consumed	9.919189	6.087584	10.31606
	Case III	$p = 5$	
EMSE	12724.32	12726.87	12724.74
Total time consumed	14.54680	17.42145	22.08751

TABLE 3
Performance of the LST, the LTS, and the LMS for three true $\boldsymbol{\beta}_0$'s.

Inspecting the Table reveals that (i) in terms of EMSE, the LST (AA1) is the overall winner (has the smallest EMSE in all cases) whereas the LMS is the loser; (ii) in terms of speed, there is no overall winner. In two respective cases, the LMS is the fastest whereas the LST is fastest in $p = 5$ case and the LTS is the slowest in all cases.

Up to this point, we have dealt with synthetic data sets. Next we examine the performance of the LST, the LTS and the LMS with respect to real data sets in high dimensions.

Example 7.4 Textbook size real data sets We first look at real data sets with relatively small sample size n and moderate dimension p. For a description of data sets, see [23], all are studied there. Since some of methods might depend on randomness, So we run the computation $R = 1000$ times to alleviate the randomness. We then calculate the *total* time consumed (in seconds) by different methods for all replications, and the EMSE (with true $\boldsymbol{\beta}_0$ being replaced by the sample mean of 1000 $\widehat{\boldsymbol{\beta}}$s), which is the sample variance of all $\widehat{\boldsymbol{\beta}}$s up to a factor 1000/999. The results are reported in Table 4, where the parameters α and T_{ls} in AA1 are tuned.

Inspecting the Table reveals that (i) in terms of the EMSE (or rather empirical variance), AA1 and ltsReg are the overall winners for all cases considered (no randomness) and the LMS has the largest sample variance. (ii) in terms of computation speed, there is no overall winner, but AA1 is faster than ltsReg in three out of four cases. The LMS is the fastest in one case. □

The limitation of this example is that the data sets are still relatively small and not in very high dimensions. We examine a high dimension and large sample dataset next.

Table entries (a, b) are: a:=empirical variance of $\widehat{\boldsymbol{\beta}}$s, b:=total time consumed

data set	(n, p)	AA1	ltsReg	lmsreg
salinity	(28, 4)	(0.0, 2.3290)	(0.0, 8.8385)	(1.3719, 4.9425)
phosphor	(18, 3)	(0.0, 4.9218)	(0.0, 8.3902)	(0.0000, 1.5153)
wood	(20, 6)	(0.0, 4.8013)	(0.0, 10.343)	(2.6470, 8.3714)
coleman	(20, 6)	(0.0, 14.585)	(0.0, 10.159)	(243.11, 8.3560)

TABLE 4

Total time consumed (in seconds) and sample variance in 1000 *replications by the LST (AA1), the LTS (ltsReg), and the LMS (lmsreg) for various real data sets.*

Example 7.5 A large real data set Boston housing is a famous data set ([9]) and studied by many authors with different emphasizes (transformation, quantile, nonparametric regression, etc.) in the literature. For a more detailed description of the data set, see http://lib.stat.cmu.edu/datasets/.

The analysis reported here did not include any of the previous results, but consisted of just a straight linear regression of the dependent variable (median price of a house) on the thirteen explanatory variables as might be used in an initial exploratory analysis of a new data set. We have sample size $n = 506$ and dimension $p = 14$.

We assess the performance of the LST, the LTS, and the LMS as follows: (i) we sample m points (without replacement) ($m = 506$, entire data set, or $m = 200, 250, 300, 350$) from the entire data set, and compute the $\widehat{\boldsymbol{\beta}}$s with different methods, we do this RepN times, where replication number RepN varies with respect to different ms. (ii) we calculate the total time consumed (in seconds) by different methods for all replications, and the EMSE (with true $\boldsymbol{\beta}_0$ being replaced by the sample mean of RepN $\widehat{\boldsymbol{\beta}}$s from (i)), which is the sample variance of all $\widehat{\boldsymbol{\beta}}$s up to a factor $RepN/(RepN-1)$. The results are reported in Table 5.

Inspecting the Table reveals that (i) the LMS has the largest EMSEs while it is faster than the LTS in all cases; (ii) the LST has smallest EMSE in three cases among the five (in those cases it is slower than the LTS) (in the other two cases the LTS takes its turn); (iii) in the entire data-set case, the LST returned the same estimate every replication whereas it is not the case for the LTS and the LMS. □

8. Final discussions

The difference between the LTS and the LST The least sum of squares of trimmed (LST) residuals estimator, which is proven to have the best 50% asymptotic breakdown point, is another robust alternative to the classical least sum of squares (LS) of residuals estimator. The latter keeps all squared residuals

Table entries (a, b) are: a:=empirical variance of $\widehat{\boldsymbol{\beta}}$s, b:=total time consumed

p	m	RepN	LST(AA1)	LTS(ltsReg)	LMS(lmsreg)
	200	10^4	(195.3379, 595.7677)	(220.8644, 480.0612)	(847.2457, 472.4671)
	250	10^4	(164.4042, 723.5861)	(169.5725, 597.2802)	(791.2557, 555.3318)
14	300	10^4	(461.5653, 514.8522)	(126.7703, 683.3362)	(754.2416, 623.5828)
	350	10^4	(453.3266, 695.9286)	(97.86377, 821.1486)	(724.2104, 732.2517)
	506	10^3	(0.000000, 142.4225)	(42.58697, 116.5830)	(703.7999, 101.0454)

TABLE 5

Total time consumed (in seconds) and sample variance in RepN replications by the LTS (ltsReg), the LST (AA1), and the LMS(lmsreg) for real data sets with various sample size m's and $p = 14$.

whereas the former trims some residuals and then squares the left. Trimming is also utilized in the prevailing least sum of trimmed squares (LTS) of the residuals estimator. However, the two trimming schemes are quite different, the one used in the LTS is a one-sided trimming (only large squared residuals are trimmed, of course, it also might be regarded as a two-sided trimming with respect to the un-squared residuals) whereas the one utilized in the LST is a depth-based (or outlyingness-based) trimming (see [49] and [45] for more discussions on trimming schemes) which can trim both ends of un-squared residuals and trim not a fixed number of residuals.

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. All the difference leads to an unexpected performance difference in the LTS and the LST as demonstrated in the last section.

Fairness of performance criteria For comparison of the performance of the LST and the LTS, we have focused on the variance (accuracy, efficiency, or EMSE) and the computation speed of the algorithms for the estimators. The asymptotic efficiency (AE) of the LTS has been reported to be just 7% in [34] or 8% in [17] (page 132), the AE of the LST is yet to be discovered, which however is expected to be better than 8%. This assentation is verified and supported by the experimental results in the last section (Tables 2, and 3 indicate that the LST is much more efficient than the LTS). Furthermore, it was also supported by the results of [45] for various trimming schemes in the case of $p = 1$.

The computation speed comparison of the LTS versus the LST in the last section is somewhat not based on a level ground. It is essentially a speed comparison of pure R verse R plus Fortran since the Fortran subroutine (rfltsreg) is called in ltsReg (similarly lmsreg also calling a Fortran subroutine). Even with that, ltsReg does not have an overwhelming advantage in speed over AA1. For the latter, however, there is still room for improvement by utilizing Fortran or

even better Rcpp to speed up by at least one order of magnitude.

Parameters tuning and finite sample breakdown point There are two parameters h in the LTS and α in the LST which can be tuned in the program for computation. Their values have a connection with the finite sample breakdown point. For example, when h takes its default value $\lfloor (n+p+1)/2 \rfloor$, then the FSBP of the LTS is $(n-h+1)/n$ which will decrease from the best FSBP result $(\lfloor (n-p)/2 \rfloor + 1)/n$ (see pages 125, 132 of [23]) when h increases. For the parameter α in LST, as long as $\alpha \geq 1$ then the high FSBP in theorem 3.1 remains valid. This is due to the difference in the trimming schemes (see [45]).

Open and future problems By simply switching the order of trimming and squaring and adopting a depth-based trimming scheme, the LTS and the LST can have different performances. One naturally wonders what if one does the same thing with respect to the famous the LMS introduced also by [21] (i.e. the least square of the median (LSM) of residuals estimator). It turns out, this is not a good idea since there is a universal solution, it is $\widehat{\boldsymbol{\beta}} = (\text{Med}\{y_i\}, 0, \cdots, 0) \in R^p$.

One interesting problem that remains is to investigate the least sum of squares of trimmed residuals with yet another trimming scheme such as the winsorized version given in [45], that is, replacing the residuals beyond the cut-off values at the two ends with just the cutoff values or even a more generalized weighted (trimming) scheme which includes the hard 0 and 1 trimming scheme. Other challenging open topics that deserve to be pursued independently elsewhere include (i) providing a finite sample estimation error analysis (non-asymptotic analysis) (ii) regularized regression based on the LST to handle variable selection and model interpretation issues when dimension p is much larger than sample size n.

Acknowledgments

Authors thank Denis Selyuzhitsky, Nadav Langberg, and Profs. Wei Shao and Yimin Xiao for their stimulating discussions and two anonymous referees for their insightful comments and helpful suggestions. All of which significantly improved the manuscript.

Declarations

Funding

Authors declare that there is no funding received for this study.

Conflicts of interests/Competing interests

Authors declare that there is no conflict of interests/Competing interests.

References

[1] Anonymous (1821). Dissertation sur la recherche du milieu le plus probable, entre les rbsultats de plusieurs observations ou experiences. *Ann. Math. Pures Appl.* 12, 181-204.

[2] Bickel, P.J. (1975), "One-step Huber estimates in the linear model". *J. Am. Statist. Assoc.*, 70, 428-434.

[3] Boyd, S. and Vandenberghe, L. (2004), *Convex Optimization*. Cambridge University Press.

[4] Dixon, W.J. and Tukey, J.W. (1968), "Approximate Behavior of the Distribution of Winsorized t (Trimming/Winsorization 2)", *Technometrics*, 10(1), pp. 83-98.

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

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

[7] 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.

[8] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, John Wiley & Sons, New York.

[9] Harrison, D. and Rubinfeld, D.L. (1987), "Hedonic prices and the demand for clean air", *J. Environ. Economics and Management*, vol.5, 81-102.

[10] Kim, J. and Pollard, D. (1990), "Cube root asymptotics". *Ann. Statist.*, 18 191-219.

[11] Hawkins, D. M. (1994), "The feasible solution algorithm for least trimmed squares regression", *Computational Statistics & Data Analysis*, 17, 185-196.

[12] Hawkins, D. M. and Olive, D. J. (1999), "Improved feasible solution algorithms for high breakdown estimation " *Computational Statistics & Data Analysis*, 30(1), 1-11.

[13] Hettmansperger, T.P. and Sheather, S. J. (1992), "A Cautionary Note on the Method of Least Median Squares", The American Statistician, 46:2, 79-83.

[14] Huber, P. J. (1964), "Robust estimation of a location parameter", *Ann. Math. Statist.*, 35 73-101.

[15] Huber, P. J. (1973), "Robust Regression," *Ann. Statist.*, 1, 799-821.

[16] Johansen, S., and Nielsen, B., (2013), "Outlier Detection in Regression Using an Iterated One-Step Approximation to the Huber-Skip Estimator", *Econometrics*, 1, 53-70.

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

[18] Mendeleev, D. I. (1895), "Course of work on the renewal of prototypes or standard measures of lengths and weights (Russian)". *Vremennik Glavnoi Paluty Mer i Vesw* 2, 157-185. Reprinted 1950: *Collected Writings (Soch-*

eneniya), Izdat. Akad. Nauk, SSSR, Leningrad-Moscow, Vol. 22, pp. 175-213.

[19] Öllerer, V., Croux, C., and Alfons, A. (2015) "The influence function of penalized regression estimators", *Statistics*, 49:4, 741-765

[20] Pollard, D. (1984), *Convergence of Stochastic Processes*, Springer, Berlin.

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

[22] Rousseeuw, P. J., and Hubert, M. (1999), "Regression depth (with discussion)", *J. Amer. Statist. Assoc.*, 94, 388–433.

[23] Rousseeuw, P.J., and Leroy, A. (1987), *Robust regression and outlier detection*. Wiley New York.

[24] Rousseeuw, P. J., Struyf, A. (1998), "Computing location depth and regression depth in higher dimensions", *Statistics and Computing*, 8:193-203.

[25] Rousseeuw, P. J. and Van Driessen, K. (1999), "A fast algorithm for the minimum covariance determinant estimator", *Technometrics*, 41(3), 212-223.

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

[27] Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In Robust and Nonlinear Time Series Analysis. Lecture Notes in Statist. Springer, New York. 26 256-272

[28] Ruppert, D. and Carroll, R. J. (1980), "Trimmed Least Squares Estimation in the Linear Model", *J. Amer. Statist. Assoc.*, 75, 828-838.

[29] Serfling, R. J. (1980), '*Approximation Theorems of Mathematical Statistics*". New York: Wiley.

[30] Sherman, R. P. (1993), "The Limiting Distribution of the Maximum Rank Correlation Estimator", *Econometrica* ,61(1), pp. 123-137.

[31] Sherman, R. P. (1994), ""Maximal Inequalities for Degenerate U-Processes with Applications to Optimization Estimators", *Ann. Statist.* 22(1): 439-459.

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

[33] Stigler, S.M., (1976), "The anonymous Professor Gergonne", *Hist. Math.* 3, 71-74.

[34] 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.

[35] Transtrum, M. K., Machta, B. B., and Sethna, J. P. (2011), "Geometry of nonlinear least squares with applications to sloppy models and optimization", *Phys. Rev. E* 83, 036701

[36] Tableman, M. (1994), "The influence functions for the least trimmed squares and the least trimmed absolute deviations estimators", *Statistics & Probability Letters* 19 (1994) 329-337.

[37] Tukey, J.W. and McLaughlin, D.H. (1963), " Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization 1", *Sankhyā: The Indian Journal of Statistics, Series*

A , 25(3), pp. 331-352.

[38] Van Der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press.

[39] Van Der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer, New York.

[40] Víšek, J. Á. (2006a), The least trimmed squares. Part I: Consistency. *Kybernetika*, 42, 1-36.

[41] Víšek, J. Á. (2006b) The least trimmed squares. Part II: $\sqrt{n}$-consistency. *Kybernetika*, 42, 181-202.

[42] Víšek, J. Á. (2006c), The least trimmed squares. Part III: Asymptotic normality. *Kybernetika*, 42, 203-224.

[43] Víšek, J. Á. (2001), "Regression with high breakdown point", ROBUST'2000, 324 – 356.

[44] Welsh, A. H. (1987), "The Trimmed Mean in the Linear Model", Ann. Statist. 15(1): 20-36.

[45] Wu, M., and Zuo, Y. (2009), "Trimmed and Winsorized means based on a scaled deviation", *J. Statist. Plann. Inference*, 139(2), 350-365.

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

[47] Yohai, V.J. and Zamar, R.H. (1988), "High breakdown estimates of regression by means of the minimization of an efficient scale", *J. Amer. Statist. Assoc.*, 83, 406–413.

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

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

[50] Zuo, Y. (2018), "A new approach for the computation of halfspace depth in high dimensions". *Communications in Statistics - Simulation and Computation*, 48(3): 900-921.

[51] Zuo, Y. (2020), "Large sample properties of the regression depth induced median", *Statistics and Probability Letters*, November 2020 166, arXiv1809.09896.

[52] Zuo, Y. (2021a), "On general notions of depth for regression" *Statistical Science* 2021, Vol. 36, No. 1, 142–157, arXiv:1805.02046.

[53] Zuo, Y. (2021b), "Robustness of the deepest projection regression depth functional", *Statistical Papers*, vol. 62(3), pages 1167-1193.

[54] Zuo, Y. (2021c), "Computation of projection regression depth and its induced median", *Computational statistics and data analysis*, Vol. 158, 107184.

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

Supplementary Material

R code downloadable at https://github.com/left-github-4-codes/LST

Appendix: main proofs and remarks

Proof of Lemma 2.2

(i) For $\boldsymbol{\eta} \in R_{\boldsymbol{\beta}^k}$, we have $I(\boldsymbol{\eta}) = I(\boldsymbol{\beta}^k)$. Let $J = |I(\boldsymbol{\beta}^k)|$, then $D_{i_{j+1}}(\boldsymbol{\eta}) > D_{i_j}(\boldsymbol{\eta})$ for $1 \le j \le (J-1)$. Let $\gamma := \min_{1\le j\le(J-1)} |D_{i_{j+1}}(\boldsymbol{\eta}) - D_{i_j}(\boldsymbol{\eta})|\sigma_n(\boldsymbol{\eta})$, then by (2.7) we have $\gamma > 0$.

Due to the continuity of residuals in $\boldsymbol{\beta}$, we can choose a small radius δ such that for any $\boldsymbol{\beta} \in B(\boldsymbol{\eta}, \delta)$, $|r_i(\boldsymbol{\eta}) - r_i(\boldsymbol{\beta})| < \gamma/4$ for any i. After a straightforward derivation, one gets $|m_n(\boldsymbol{\eta}) - m_n(\boldsymbol{\beta})| \le \gamma/4$. In light of these two inequalities and the definition of γ, one obtains

$$\begin{aligned} |r_{i_{j+1}}(\boldsymbol{\beta}) - m_n(\boldsymbol{\beta})| &\ge \big|r_{i_{j+1}}(\boldsymbol{\eta}) - \gamma/4 - [m_n(\boldsymbol{\eta}) + \gamma/4]\big| \\ &= \big|r_{i_{j+1}}(\boldsymbol{\eta}) - m_n(\boldsymbol{\eta}) - \gamma/2\big| \\ &\ge |r_{i_j}(\boldsymbol{\eta}) - m_n(\boldsymbol{\eta})| + \gamma/2, \end{aligned}$$

for any $\boldsymbol{\beta} \in B(\boldsymbol{\eta}, \delta)$ and any $1 \le j \le (J-1)$, and

$$\begin{aligned} |r_{i_j}(\boldsymbol{\beta}) - m_n(\boldsymbol{\beta})| &\le |r_{i_j}(\boldsymbol{\eta}) + \gamma/4 - [m_n(\boldsymbol{\eta}) - \gamma/4]| \\ &= |r_{i_j}(\boldsymbol{\eta}) - m_n(\boldsymbol{\eta}) + \gamma/2| \\ &\le |r_{i_j}(\boldsymbol{\eta}) - m_n(\boldsymbol{\eta})| + \gamma/2 \end{aligned}$$

The last two displays imply that $D_{i_{j+1}}(\boldsymbol{\beta}) > D_{i_j}(\boldsymbol{\beta})$ for any $1 \le j \le (J-1)$. That is, for any $\boldsymbol{\beta} \in B(\boldsymbol{\eta}, \delta)$, $\boldsymbol{\beta} \in R_{\boldsymbol{\beta}^k}$. Consequently, $Q^n(\boldsymbol{\beta}) = \sum_{i\in I(\boldsymbol{\beta}^k)} r_i^2$.

(ii) The openness of $R_{\boldsymbol{\beta}^k}$ follows from the proof (i) above straightforwardly.

(iii) For any $\boldsymbol{\beta} \in \mathbb{R}^p$, (i) either $\boldsymbol{\beta} \in R_{\boldsymbol{\beta}^k}$ for some $0 \le k \le L$ and $Q^n(\boldsymbol{\beta}) = \sum_{i\in I(\boldsymbol{\beta})} r_i^2$, or (ii) $\boldsymbol{\beta}$ lies on the common boundary of $R_{\boldsymbol{\beta}^s}$ and $R_{\boldsymbol{\beta}^t}$ for some $1 \le s \ne t \le L$ such that there are $i \ne j$ $D_i(\boldsymbol{\beta}) = D_j(\boldsymbol{\beta})$, and $D_i(\boldsymbol{\eta}) > D_j(\boldsymbol{\eta})$ if $\boldsymbol{\eta} \in R_{\boldsymbol{\beta}^s}$ and $D_i(\boldsymbol{\eta}) < D_j(\boldsymbol{\eta})$ if $\boldsymbol{\eta} \in R_{\boldsymbol{\beta}^t}$, and $Q^n(\boldsymbol{\beta}) = \sum_{i\in I(\boldsymbol{\beta})} r_i^2$ for $\boldsymbol{\beta} \in \overline{R}_{\boldsymbol{\beta}^s} \cap \overline{R}_{\boldsymbol{\beta}^t}$.

The continuity of $Q^n(\boldsymbol{\beta})$ over $R_{\boldsymbol{\beta}^k}$ is obvious. We show that is true at any $\boldsymbol{\eta} \in \overline{R}_{\boldsymbol{\beta}^s} \cap \overline{R}_{\boldsymbol{\beta}^t}$. Let $\{\boldsymbol{\beta}_j\}$ be a sequence approaching to $\boldsymbol{\eta}$, where $\boldsymbol{\beta}_j$ could be in $\overline{S}_{\boldsymbol{\beta}^s}$ or in $\overline{S}_{\boldsymbol{\beta}^t}$. We show that $Q^n(\boldsymbol{\beta}_j)$ approaches to $Q^n(\boldsymbol{\eta})$. Note that $Q^n(\boldsymbol{\eta}) = \sum_{i\in I(\boldsymbol{\eta})} r_i^2$ for $\boldsymbol{\eta} \in \overline{R}_{\boldsymbol{\beta}^s} \cup \overline{R}_{\boldsymbol{\beta}^t}$. Partition $\{\boldsymbol{\beta}_j\}$ into $\{\boldsymbol{\beta}_{j_s}\}$ and $\{\boldsymbol{\beta}_{j_t}\}$, and all members of the former belong to $\overline{R}_{\boldsymbol{\beta}^s}$ where the latter are all within $\overline{R}_{\boldsymbol{\beta}^t}$. By continuity of the sum of squared residuals in

$\boldsymbol{\beta}$, both $Q^n(\boldsymbol{\beta}_{j_s}))$ and $O^n(\boldsymbol{\beta}_{j_t}))$ approach to $Q^n(\boldsymbol{\eta})$ since both $\{\boldsymbol{\beta}_{j_s}\}$ and $\{\boldsymbol{\beta}_{j_t}\}$ approach η as $\min\{j_s, j_t\} \to \infty$.

(iv) Over each $R_{\boldsymbol{\beta}^k}$, $1 \le k \le L$, $Q^n(\boldsymbol{\beta}) = \sum_{i\in I(\boldsymbol{\beta})} r_i^2$ which is clearly twice differentiable and convex since

$$\frac{\partial}{\partial \boldsymbol{\beta}} Q^n(\boldsymbol{\beta}) = -2\sum_{i=1}^{n} r_i \mathbb{1}_i \boldsymbol{w}_i = -2\boldsymbol{R}'\boldsymbol{D}\boldsymbol{W}'_n,$$

$$\frac{\partial^2}{\partial \boldsymbol{\beta}^2} O^n(\boldsymbol{\beta}) = 2\boldsymbol{W}_n \boldsymbol{D}\boldsymbol{W}'_n,$$

where $\boldsymbol{R} = (r_1, r_2, \cdots, r_n)'$, $\boldsymbol{D} = \mathrm{diag}(\mathbb{1}_i)$, and $\boldsymbol{W}_n = (\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_n)'$. Strict convexity follows from the positive definite of Hessian matrix: $2\boldsymbol{W}_n\boldsymbol{D}\boldsymbol{W}'_n$. □

Proof of Theorem 2.1

(i) Over each $S_{\boldsymbol{\beta}^k}$, $Q^n(\boldsymbol{\beta})$ is twice differentiable and strictly convex in light of given condition, hence it has a unique minimizer. Since there are only finitely many $R_{\boldsymbol{\beta}^k}$, the assertion follows if we can prove that the minimum does not reach at a boundary point of some $R_{\boldsymbol{\beta}^k}$.

Assume it is otherwise. That is, $Q^n(\boldsymbol{\beta})$ reaches its minimum at point $\boldsymbol{\beta}_1$ which is a boundary point of $R_{\boldsymbol{\beta}^k}$ for some k. Assume that over $R_{\boldsymbol{\beta}^k}$, $Q^n(\boldsymbol{\beta})$ attains its minimum value at the unique point $\boldsymbol{\beta}_2$. Then, $Q^n(\boldsymbol{\beta}_1) \le Q^n(\boldsymbol{\beta}_2)$, If equality holds then we already have the desired result, otherwise, there is a point $\boldsymbol{\beta}_3$ in the small neighborhood of $\boldsymbol{\beta}_1$ so that $Q^n(\boldsymbol{\beta}_3) \le Q^n(\boldsymbol{\beta}_1) + (Q^n(\boldsymbol{\beta}_2) - Q^n(\boldsymbol{\beta}_1))/2 < Q^n(\boldsymbol{\beta}_2)$. A contradiction is obtained.

(ii) It is seen from **(i)** that $Q^n(\boldsymbol{\beta})$ is twice continuously differentiable, hence its first derivative evaluated at the global minimum must be zero. By **(i)**, we have (2.8).

(iii) This part directly follows from **(ii)** and the invertibility of $\boldsymbol{M}_n$ that follows from the full rank of $\boldsymbol{X}_n$. □

Proof of Theorem 2.2

For the given $\boldsymbol{Z}^{(n)}$ and α, write $M = Q(\boldsymbol{Z}^{(n)}, \boldsymbol{0}, \alpha) = \sum_{i\in I(\boldsymbol{0})} y_i^2$. For a given $\boldsymbol{\beta} \in \mathbb{R}^p$, assume that $H_{\boldsymbol{\beta}}$ is the hyperplane determined by $y = \boldsymbol{w}'\boldsymbol{\beta}$ and let H_h being the horizontal hyperplane (i.e. $y = 0$, the $\boldsymbol{w}$-space). Partition the space of $\boldsymbol{\beta}$s into two parts:S_1 and S_2, with S_1 containing all $\boldsymbol{\beta}$s such that $H_{\boldsymbol{\beta}}$ and H_h are parallel and S_2 consisting of the rest of $\boldsymbol{\beta}$s so that $H_{\boldsymbol{\beta}}$ and H_h are not parallel.

If one can show that there are minimizers of $Q(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta}, \alpha)$ over S_i $i =$

$1, 2$ respectively, then one can have an overall minimizer. Over S_1, the minimizer is $\widehat{\boldsymbol{\beta}} = (\overline{y}, \mathbf{0}'_{(p-1)\times 1})'$ and the minimum value of $Q(\boldsymbol{Z}^{(n)}, \widehat{\boldsymbol{\beta}}, \alpha)$ is $M - \overline{y}^2$, where $\overline{y}$ is the average of y_i over all $i \in I(\mathbf{0})$.

Over S_2, denote by $l_{\boldsymbol{\beta}}$ the intersection part of $H_{\boldsymbol{\beta}}$ with the horizontal hyperplane H_h (we call it a hyperline, though it is $p-1$-dimensional). Let $\theta_{\boldsymbol{\beta}} \in (-\pi/2, \pi/2)$ be the angle between the $H_{\boldsymbol{\beta}}$ and H_h (and $\theta_{\boldsymbol{\beta}} \neq 0$). Consider two cases.

Case I. All $\boldsymbol{w}_i$, $i \in I(\boldsymbol{\beta})$ on the hyperline $l_{\boldsymbol{\beta}}$. Then we have a vertical hyperplane that is perpendicular to the horizontal hyperplane H_h ($y = 0$) and intersect H_h at $l_{\boldsymbol{\beta}}$, which contains, in light of lemma 2.1, at least $\lfloor (n+1)/2 \rfloor$ points of $\boldsymbol{Z}^{(n)}$. But this contradicts the assumption just before the theorem. We only need to consider the other case.

Case II. Otherwise, define

$$\delta = \frac{1}{2} \inf\{\tau, \text{such that } N(l_{\boldsymbol{\beta}}, \tau) \text{ contains all } \boldsymbol{w}_i \text{ with } i \in I(\boldsymbol{\beta})\},$$

where $N(l_{\boldsymbol{\beta}}, \tau)$ is the set of points in $\boldsymbol{w}$-space such that each distance to the $l_{\boldsymbol{\beta}}$ is no greater than τ. Clearly, $0 < \delta < \infty$ (since $\delta = 0$ has been covered in **Case I** and $2\delta \leq \max_i\{\|\boldsymbol{w}_i\|\} < \infty$, where the first inequality follows from the fact that hypotenuse is always longer than any legs).

We now show that when $\|\boldsymbol{\beta}\| > (1+\eta)\sqrt{M}/\delta$, where $\eta > 1$ is a fixed number, then

$$\sum_{i \in I(\boldsymbol{\beta})} r_i^2(\boldsymbol{\beta}) > M = \sum_{i \in I(\mathbf{0})} r_i^2(\mathbf{0}). \tag{8.1}$$

That is, for the solution of minimization of (2.4), one only needs to search over the ball $\|\boldsymbol{\beta}\| \leq (1+\eta)\sqrt{M}/\delta$, a compact set. Note that $Q(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta}, \alpha)$ is continuous in $\boldsymbol{\beta}$ by Lemma 2.2. Then the minimization problem certainly has a solution over the compact set.

The proof is complete if we can show (8.1) when $\|\boldsymbol{\beta}\| > (1+\eta)\sqrt{M}/\delta$. It is not difficult to see that there is at least one $i \in I(\boldsymbol{\beta})$ such that $\boldsymbol{w}_i \notin N(l_{\boldsymbol{\beta}}, \delta)$ since otherwise it contradicts the definition of δ above. Note that $\theta_{\boldsymbol{\beta}}$ is the angle between the normal vectors $(-\boldsymbol{\beta}', 1)'$ and $(\mathbf{0}', 1)'$ of hyperplanes $H_{\boldsymbol{\beta}}$ and H_h, respectively. Then $|\tan\theta_{\boldsymbol{\beta}}| = \|\boldsymbol{\beta}\|$ and (see Figure 2)

$$|\boldsymbol{w}_i'\boldsymbol{\beta}| > \delta |\tan\theta_{\boldsymbol{\beta}}| = \delta\|\boldsymbol{\beta}\| > (1+\eta)\sqrt{M}.$$

Now we have

$$|r_i(\boldsymbol{\beta})| = |\boldsymbol{w}_i'\boldsymbol{\beta} - y_i| \geq \big||\boldsymbol{w}_i'\boldsymbol{\beta}| - |y_i|\big| > (1+\eta)\sqrt{M} - |y_i|. \tag{8.2}$$

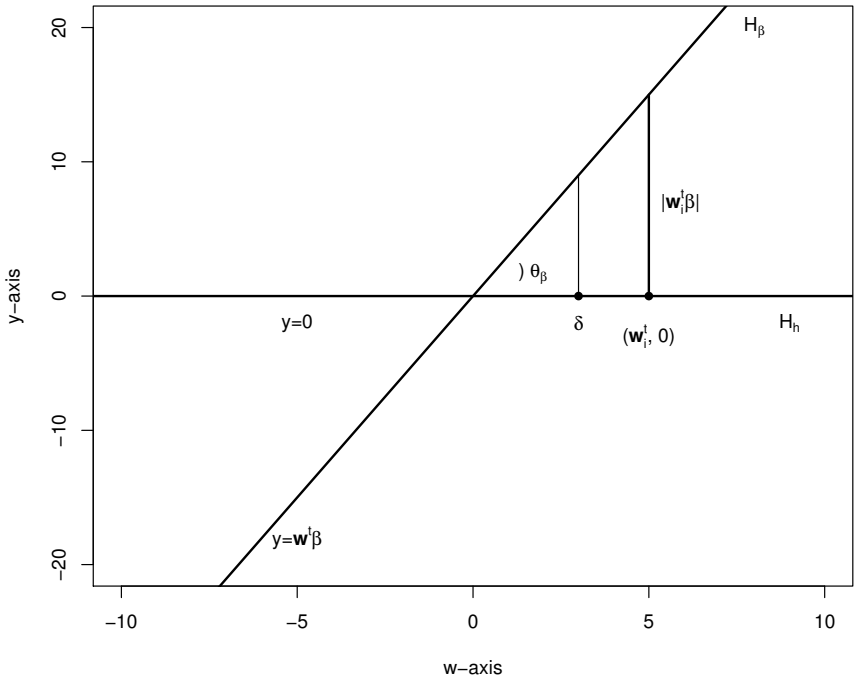


Fig 2: A two-dimensional vertical cross-section (that goes through points $(\boldsymbol{w}_i^t, 0)$ and $(\boldsymbol{w}_i^t, \boldsymbol{w}_i^t\boldsymbol{\beta})$) of a figure in $\mathbb{R}^p$ ($\boldsymbol{w}_i^t = \boldsymbol{w}_i'$). Hyperplanes H_h and $H_{\boldsymbol{\beta}}$ intersect at hyperline $l_{\boldsymbol{\beta}}$ (which does not necessarily pass through $(\mathbf{0}, 0)$, here just for illustration). The vertical distance from point $(\boldsymbol{w}_i^t, 0)$ to the hyperplane $H_{\boldsymbol{\beta}}$, $|\boldsymbol{w}_i^t\boldsymbol{\beta}|$, is greater than $\delta|\tan(\theta_{\boldsymbol{\beta}})|$.

Therefore,

$$\sum_{j\in I(\boldsymbol{\beta})} r_j^2(\boldsymbol{\beta}) \geq r_i^2(\boldsymbol{\beta}) > \Big((1+\eta)\sqrt{M} - |y_i|\Big)^2 \geq \Big((1+\eta)\sqrt{M} - \sqrt{M}\Big)^2$$
$$= \eta^2 M > M = \sum_{j\in I(\mathbf{0})} r_j^2(\mathbf{0}).$$

That is, we have certified (8.1). □

Proof of theorem 3.1

Case A: $p = 1$. The problem becomes an estimation of a location parameter β_1 (the intercept term in the model $y_i = \beta_1 + e_i$). The solution is the

depth trimmed mean based on $y_i, i \in N$, which has the RBP as claimed (see [45]).

Case B: $p > 1$.

(i) **First**, *we show that* $m = \lfloor n/2 \rfloor - p + 2$ *points are enough to breakdown* $\widehat{\boldsymbol{\beta}}_{lst}^n$. Recall the definition of $\widehat{\boldsymbol{\beta}}_{lst}^n$. One has

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{lst}(\mathbf{Z}^{(n)}, \alpha) &= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} Q(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta}, \alpha) \\ &= \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} r_i^2 \mathbb{1}\left(\frac{|r_i - m(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}^{(n)}, \boldsymbol{\beta})} \leq \alpha \right). \quad (8.3) \end{aligned}$$

Select $p-1$ points from $\boldsymbol{Z}^{(n)} = \{(\boldsymbol{x_i'}, y_i)'\}$. $(\boldsymbol{w}_i', y_i)$, together with the origin, form a $(p-1)$-dimensional subspace (hyperline) L_h in the $(p+1)$-dimensional space of $(\boldsymbol{w}', y)'$.

Construct a non-vertical hyperplane H through L_h (that is, it is not perpendicular to the horizontal hyperplane $y = 0$). Let $\boldsymbol{\beta}$ be determined by the hyperplane H through $y = \boldsymbol{w'\beta}$.

We can tilt the hyperplane H so that it approaches its ultimate vertical position. Meanwhile, we put all the m contaminating points onto this hyperplane H so that it contains no less than $m + (p-1) = \lfloor n/2 \rfloor + 1$ observations. Call the resulting contaminated sample by $\boldsymbol{Z}_m^{(n)}$. Therefore the majority of $r_i = y_i - \boldsymbol{w}_i'\boldsymbol{\beta}$ will now be zero. Therefore, $\sigma(\mathbf{Z}^{(n)}, \boldsymbol{\beta})$, in this case, is defined to be one.

When H approaches its ultimate vertical position, $\|\boldsymbol{\beta}\| \to \infty$ (for the reasoning, see the **case (II)** of the proof of Theorem 2.2) and r_i for points $(\boldsymbol{w}_i', y_i))'$ not on the H will also approach ∞. This implies that this $\boldsymbol{\beta}$ is the solution for $\widehat{\boldsymbol{\beta}}_{lst}^n$ at this contaminated data $\boldsymbol{Z}_m^{(n)}$ since it attains the minimum possible value (zero) on the RHS of (2.5). That is, $m = \lfloor n/2 \rfloor - p + 2$ contaminating points are enough to break down $\widehat{\boldsymbol{\beta}}_{lst}^n$.

(ii) **Second**, *we now show that* $m = \lfloor n/2 \rfloor - p + 1$ *points are not enough to break down* $\widehat{\boldsymbol{\beta}}_{lst}^n$. Let $\boldsymbol{Z}_m^{(n)}$ be an arbitrary contaminated sample and $\boldsymbol{\beta_c} := \widehat{\boldsymbol{\beta}}_{lst}(\boldsymbol{Z}_m^{(n)}, \alpha)$ and $\boldsymbol{\beta_o} = \widehat{\boldsymbol{\beta}}_{lst}(\boldsymbol{Z}^{(n)}, \alpha)$, where $\boldsymbol{Z}^{(n)} = \{\boldsymbol{Z}_i\} = \{(\boldsymbol{x_i'}, y_i)'\}$ are uncontaminated original points. Assume that $\boldsymbol{\beta_c} \neq \boldsymbol{\beta_o}$ (Otherwise, we are done). It suffices to show that $\|\boldsymbol{\beta_c} - \boldsymbol{\beta_o}\|$ is bounded.

Note that since $n - m = \lfloor (n+1)/2 \rfloor + p - 1$, then both m and σ in respective (2.2) and (2.3) are bounded for both contaminated $\boldsymbol{Z}_m^{(n)}$ and

$\boldsymbol{\beta}_c$ and original $\boldsymbol{Z}^{(n)}$ and $\boldsymbol{\beta}_o$. Define

$$\delta = \frac{1}{2}\inf\ \{\tau>0;\ \exists \text{ a } (p-1)\text{-dimensional subspace } L \text{ of } (y=0) \text{ such that } L^\tau \text{ contains at least } p \text{ of uncontaminated } (1,\boldsymbol{x_i'}) \text{ from } \boldsymbol{Z}^{(n)}\},$$

where L^τ is the set of all points $\boldsymbol{w'}$ that have the distance to L no greater than τ. Since $\boldsymbol{Z}^{(n)}$ is in general position, $\delta>0$.

Let H_o and H_c be the hyperplanes determined by $y=\boldsymbol{w'\beta_o}$ and $y=\boldsymbol{w'\beta_c}$, respectively, and $M=\max_i\{|y_i-\boldsymbol{w_i'\beta}_o|\}$ for all original y_i and $\boldsymbol{x_i}$ in $Z^{(n)}$. Since $\boldsymbol{\beta_o}\neq\boldsymbol{\beta_c}$, then $H_o\neq H_c$.

(I) Assume that H_o and H_c are not parallel. Denote the vertical projection of the intersection $H_o\cap H_c$ to the horizontal hyperplane $y=0$ by $L_{vp}(H_o\cap H_c)$, then it is $(p-1)$-dimensional. By the definition of δ, there are at most $p-1$ of uncontaminated points of $\boldsymbol{w}_i=(1,\boldsymbol{x}_i')'$ from the original $\{\boldsymbol{Z}_i, i=1,\cdots,n\}$ within $L^\delta_{vp}(H_o\cap H_c)$. Denote the set of all these possible $\boldsymbol{w}_i$ (at most $p-1$) by S_{cap} and $|S_{cap}|=n_{cap}\leq(p-1)$. Denote the set of all remaining uncontaminated $\boldsymbol{Z}_i$ from the original $\{\boldsymbol{Z}_i, i=1,\cdots,n\}$ by S_r and the set of all such i as J, then there are at least $n-m-n_{cap}\geq n-\lfloor n/2\rfloor=\lfloor(n+1)/2\rfloor$ such $\boldsymbol{Z}_i$ in S_r.

For each $(\boldsymbol{w_i'},y_i)'$ with $i\in J$, construct a two-dimensional vertical plane P_i that goes through $(\boldsymbol{w_i'},y_i)'$ and $(\boldsymbol{w_i'},y_i+1)'$ and is perpendicular to $L_{vp}(H_o\cap H_c)$ (see Figure 2 and/or Figure 16 of [23]). Denote the angle formed by H_o and the horizontal line in P_i by $\alpha_o\in(-\pi/2,\pi/2)$, similarly by α_c for H_c and P_i. They are essentially the angles formed between H_o and H_c with the horizontal hyperplane $y=0$, respectively.

We see that for $i\in J$ and each $(\boldsymbol{w_i'},y_i)'$, $|\boldsymbol{w_i'\beta_o}|>\delta|\tan(\alpha_o)|$ and $|\boldsymbol{w_i'\beta_c}|>\delta|\tan(\alpha_c)|$ (see Figure 2 or Figure 16 of [23] of a geographical illustration for better understanding) and $\|\boldsymbol{\beta_o}\|=|\tan(\alpha_o)|$ and $\|\boldsymbol{\beta_c}\|=|\tan(\alpha_c)|$.

Now for each $i\in J$, denote $r_i^o:=(y_i-\boldsymbol{w_i'\beta_o})$ and $r_i^c:=(y_i-\boldsymbol{w_i'\beta_c})$. For any $i\in J$, it follows that (see Figure 2 or Figure 16 of [23])

$$\begin{aligned}|r_i^o-r_i^c| &= \left|\boldsymbol{w_i'\beta_o}-\boldsymbol{w_i'\beta_c}\right|>\delta|\tan(\alpha_o)-\tan(\alpha_c)|\\ &\geq \delta\big||\tan(\alpha_o)|-|\tan(\alpha_c)|\big|=\delta\big|\|\boldsymbol{\beta_o}\|-\|\boldsymbol{\beta_c}\|\big|\\ &\geq \delta\big|\|\boldsymbol{\beta_o}-\boldsymbol{\beta_c}\|-2\|\boldsymbol{\beta_o}\|\big|\end{aligned}$$

Let $M_1:=|m(\mathbf{Z}_m^{(n)},\boldsymbol{\beta}_c)|+\alpha\sigma(\mathbf{Z}_m^{(n)},\boldsymbol{\beta}_c)$, which is obviously bounded. Then it is obvious that

$$Q(\mathbf{Z}_m^{(n)},\boldsymbol{\beta}_c,\alpha)=\sum_{i\in I(\boldsymbol{\beta}_c)}(r_i^c)^2\mathbb{1}\left(\frac{|r_i^c-m(\boldsymbol{Z}_m^{(n)},\boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}_m^{(n)},\boldsymbol{\beta})}\leq\alpha\right)\leq I(\boldsymbol{\beta}_c)M_1^2, \tag{8.4}$$

If we assume that $\|\boldsymbol{\beta_o}-\boldsymbol{\beta_c}\| \geq 2\|\boldsymbol{\beta_o}\| + (M_1\sqrt{I(\boldsymbol{\beta}_c)}+M)/\delta$, then by the inequality above we have for $i \in J$

$$|r_i^o - r_i^c| > \delta\big|\|\boldsymbol{\beta_o}-\boldsymbol{\beta_c}\| - 2\|\boldsymbol{\beta_o}\|\big| \geq M_1\sqrt{I(\boldsymbol{\beta}_c)}+M,$$

which implies that for any $i \in J$,

$$|r_i^c| \geq |r_i^o - r_i^c| - |r_i^o| > M_1\sqrt{I(\boldsymbol{\beta}_c)}+M-M = M_1\sqrt{I(\boldsymbol{\beta}_c)}.$$

Notice that $|J| \geq \lfloor (n+1)/2 \rfloor$ which implies that there is at least one $i_0 \in J$ that belongs to $I(\boldsymbol{\beta}_c)$ in light of Lemma 2.1. Therefore

$$\begin{aligned} Q(\mathbf{Z}_m^{(n)}, \boldsymbol{\beta}_c, \alpha) &= \sum_{i\in I(\boldsymbol{\beta}_c)} (r_i^c)^2 \mathbb{1}\left(\frac{|r_i^c - m(\boldsymbol{Z}_m^{(n)}, \boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}_m^{(n)}, \boldsymbol{\beta})} \leq \alpha\right) \\ &\geq (r_{i_0}^c)^2 > I(\boldsymbol{\beta}_c)M_1^2, \end{aligned}$$

which contradicts (8.4). Thus, $\|\boldsymbol{\beta_o}-\boldsymbol{\beta_c}\|$ $\Big(< 2\|\boldsymbol{\beta_o}\| + (M_1\sqrt{I(\boldsymbol{\beta}_c)}+M)/\delta\Big)$ is bounded.

(II) Assume that H_o and H_c are parallel. That is, $\boldsymbol{\beta_c} = \rho\boldsymbol{\beta_o}$. We claim that $\|\boldsymbol{\beta_c}-\boldsymbol{\beta_o}\|$ is bounded. If ρ is finite or $\|\boldsymbol{\beta}_o\| = 0$, then $\|\boldsymbol{\beta_c}-\boldsymbol{\beta_o}\|$ is automatically bounded. We are done. Otherwise, consider the case that $\boldsymbol{\beta}_o \neq 0$ and $|\rho| \to \infty$.

(A) Assume that H_o is not parallel to $y = 0$.

The proof is very similar to part **(I)**. Denote the intersection of H_c and the horizontal hyperplane $y = 0$: $H_c \cap \{y = 0\}$ by L_c. Then L_c^δ contains at most $p-1$ uncontaminated points from $\{\boldsymbol{Z}^{(n)}\}$. Denote the set of all the remaining uncontaminated points in $\{\boldsymbol{Z}^{(n)}\}$ as S_r. Hence $|S_r| \geq n - m - (p-1) \geq \lfloor (n+1/2 \rfloor$. Denote again by J the set of all i such that $\boldsymbol{Z}_i \in S_r$. Again let the angle between H_c and $y = 0$ be α_c, then it is seen that $\|\boldsymbol{\beta_c}\| = |\tan(\alpha_c)|$ and $|\boldsymbol{w_i'\beta_c}| > \delta|\tan(\alpha_c)|$ for any $i \in J$.

Note that for $i \in J$, $r_i^c = (y_i - \boldsymbol{w_i'\beta_c})$. Write $M_y = \max_{i\in J}|y_i|$. It follows that for $i \in J$

$$\big|r_i^c\big| \geq \big||\mathbf{w_i'\beta_c}| - |y_i|\big| \geq |\ \delta|\tan(\alpha_c)| - M_y|.$$

Since $|S_r| \geq \lfloor (n+1/2 \rfloor$, then $M_1 := |m(\mathbf{Z}_m^{(n)}, \boldsymbol{\beta}_c)| + \alpha\sigma(\mathbf{Z}_m^{(n)}, \boldsymbol{\beta}_c)$ is obviously bounded (see reasing in (I) above) and

$$Q(\mathbf{Z}_m^{(n)}, \boldsymbol{\beta}_c, \alpha) = \sum_{i\in I(\boldsymbol{\beta}_c)} (r_i^c)^2 \mathbb{1}\left(\frac{|r_i^c - m(\boldsymbol{Z}_m^{(n)}, \boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}_m^{(n)}, \boldsymbol{\beta})} \leq \alpha\right) \leq I(\boldsymbol{\beta}_c)M_1^2, \tag{8.5}$$

Notice that $|J| \geq \lfloor (n+1)/2 \rfloor$ which implies that there is at least one $i_0 \in J$ that belongs to $I(\boldsymbol{\beta}_c)$ in light of Lemma 2.1. Therefore

$$\begin{aligned} Q(\mathbf{Z}_m^{(n)}, \boldsymbol{\beta}_c, \alpha) &= \sum_{i \in I(\boldsymbol{\beta}_c)} (r_i^c)^2 \mathbb{1}\left(\frac{|r_i^c - m(\boldsymbol{Z}_m^{(n)}, \boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}_m^{(n)}, \boldsymbol{\beta})} \leq \alpha \right) \\ &\geq (r_{i_0}^c)^2 > (\delta |\tan(\alpha_c)| - M_y)^2 = (\delta |\rho| \|\boldsymbol{\beta_o}\| - M_y)^2 \end{aligned}$$

Since $|\rho|$ could be arbitrarily large, then the above inequality contradicts (8.5).

(B) Assume that H_o is parallel to $y = 0$. Then, it means that $\boldsymbol{\beta_c} = \rho \boldsymbol{\beta_o} = (\rho \beta_{o1}, 0, \cdots, 0)$. Assume that $\beta_{o1} \neq 0$. Otherwise, we are done. Now we can repeat the argument above since $n - m = (p-1) + \lfloor (n+1)/2 \rfloor$. Let A be the set of all uncontaminated points from $\mathbf{Z}^{(n)}$, then $|A| = n - m = (p-1) + \lfloor (n+1)/2 \rfloor$. Let J be the set of all i such that $\boldsymbol{Z}_i \in A$ and $M_y = \max_{i \in J} |y_i|$, then $M_1 := |m(\mathbf{Z}_m^{(n)}, \boldsymbol{\beta}_c)| + \alpha \sigma(\mathbf{Z}_m^{(n)}, \boldsymbol{\beta}_c)$ is obvious bounded. We still have

$$Q(\mathbf{Z}_m^{(n)}, \boldsymbol{\beta}_c, \alpha) = \sum_{i \in I(\boldsymbol{\beta}_c)} (r_i^c)^2 \mathbb{1}\left(\frac{|r_i^c - m(\boldsymbol{Z}^{(n)_m}, \boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}^{(n)_m}, \boldsymbol{\beta})} \leq \alpha \right) \leq I(\boldsymbol{\beta}_c) M_1^2, \tag{8.6}$$

On the one hand we have that for $i \in J$

$$|r_i^c| = |\boldsymbol{w}_i' \boldsymbol{\beta}_c - y_i| \geq \big| |\boldsymbol{w}_i' \boldsymbol{\beta}_c| - |y_i| \big| \geq \big| |\rho| |\beta_{o1}| - M_y \big|,$$

which implies that $(r_i^c)^2$ becomes unbounded when $\rho \to \infty$. Since there is at least one $i_0 \in J$ that belongs to $I(\boldsymbol{\beta}_c)$ in light of Lemma 2.1, now we have

$$\begin{aligned} Q(\mathbf{Z}_m^{(n)}, \boldsymbol{\beta}_c, \alpha) &= \sum_{i \in I(\boldsymbol{\beta}_c)} (r_i^c)^2 \mathbb{1}\left(\frac{|r_i^c - m(\boldsymbol{Z}_m^{(n)}, \boldsymbol{\beta})|}{\sigma(\boldsymbol{Z}_m^{(n)}, \boldsymbol{\beta})} \leq \alpha \right) \\ &\geq (r_{i_0}^c)^2 \geq (|\rho| |\beta_{o1}| - M_y)^2 \to \infty \quad (\text{as } \rho \to \infty), \end{aligned}$$

which contradicts to (8.6).
That is, m contaminating points are not enough to breakdown $\widehat{\boldsymbol{\beta}}_{lst}^n$ since $\|\boldsymbol{\beta}_o - \boldsymbol{\beta}_c\|$ remains bounded. □

Remark A.1

Parallel cases considered in the proofs of Theorems 2.2 and 3.1 (often missed the related discussions in the literature) are important. This is especially true in the latter case since one can not afford to miss the parallel cases when considering the all possibilities of contamination. □

Proof of Lemma 3.2

Denote the integrand in (3.3) as $G(\beta) := (y-\boldsymbol{w}'\boldsymbol{\beta})^2\mathbb{1}\left(\frac{|y-\boldsymbol{w}'\boldsymbol{\beta}-m|}{\sigma} \leq \alpha\right)$ for a given point $(\boldsymbol{x}', y) \in \mathbb{R}^p$. Put $G(\beta) := (y-\boldsymbol{w}'\boldsymbol{\beta})^2\big(1-\mathbb{1}\left(\frac{|y-\boldsymbol{w}'\boldsymbol{\beta}-m|}{\sigma} > \alpha\right)\big)$.

(i) By the strictly non-flatness of F_r around m and σ, we have the continuity of the $m(\boldsymbol{\beta}$ and $\sigma(\boldsymbol{\beta})$. Consequently, $G(\boldsymbol{\beta})$ is obvious continuous in $\boldsymbol{\beta} \in \mathbb{R}^p$. Hence, $Q(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta} \in \mathbb{R}^p$.

(ii) For arbitrary points $(\boldsymbol{x}', y)$ and $\boldsymbol{\beta}$ in $\mathbb{R}^p$ and fixed distribution F_r, there are three cases for consideration: (a) $|y - \boldsymbol{w}'\boldsymbol{\beta} - m|/\sigma < \alpha$ (b) $|y-\boldsymbol{w}'\boldsymbol{\beta}-m|/\sigma > \alpha$ and (c) $|y-\boldsymbol{w}'\boldsymbol{\beta}-m|/\sigma = \alpha$. Case (c) happens with probability zero, we thus skip this case and treat (a) and (b) only. By the continuity in $\boldsymbol{\beta}$, there is a small neighborhood of $\boldsymbol{\beta}$: $B(\boldsymbol{\beta}, \delta)$, centered at $\boldsymbol{\beta}$ with radius $\delta > 0$ such that (a) (or (b)) holds for all $\boldsymbol{\beta} \in B(\boldsymbol{\beta}, \delta)$. This implies that

$$\frac{\partial}{\partial \boldsymbol{\beta}}\mathbb{1}\left(\frac{|y - \boldsymbol{w}'\boldsymbol{\beta} - m|}{\sigma} \leq \alpha\right) = \mathbf{0},$$

and

$$\frac{\partial}{\partial \boldsymbol{\beta}}G(\boldsymbol{\beta}) = -2(y - \boldsymbol{w}'\boldsymbol{\beta})\boldsymbol{w}\mathbb{1}\left(\frac{|y - \boldsymbol{w}'\boldsymbol{\beta} - m|}{\sigma} \leq \alpha)\right),$$

Hence, we have that

$$\frac{\partial^2}{\partial \boldsymbol{\beta}^2}G(\boldsymbol{\beta}) = 2\boldsymbol{w}\boldsymbol{w}'\mathbb{1}\left(\frac{|y - \boldsymbol{w}'\boldsymbol{\beta} - m|}{\sigma} \leq \alpha)\right),$$

Note that $G(\boldsymbol{\beta})$ is uniformly bounded over $\boldsymbol{\beta} \in \mathbb{R}^p$, then by the Lebesgue dominated convergence theorem, the desired result follows.

(iii) The convexity follows from the twice differentiability and the positive semidefinte of the second order derivative of $Q(\boldsymbol{\beta})$ and the strict convexity follows from the invertibility of Hessian matrix. □

Proof of Theorem 3.2

We will treat $\boldsymbol{\beta}_{lts}(F_{(\boldsymbol{x};,y)}, \alpha)$, the counterpart for $\boldsymbol{\beta}_{lts}(F_\varepsilon(\boldsymbol{z}), \alpha)$ can be treated analogously.

(i) Existence follows from the positive smidefinite of the Hessian matrix (see proof of (ii) of Lemma 3.2) and the convexity of $Q(\boldsymbol{\beta})$.

(ii) The equation follows from the Lebesgue dominated convergence theorem, the differentiability and the first order derivative of $Q(\boldsymbol{\beta})$ given in the proof (ii) of Lemma 3.2.

(iii) The uniqueness follows from the Lebesgue dominated convergence theorem, the positive definite of the Hessian matrix based on the given condition (invertibility). □

Remark 3.2

(I) Generally, the influence function for a regression estimator when $p > 1$ is not often provided in the literature (exceptions including [53]) for the projection regression median, and [19] for the penalized regression estimators. In the latter case for the spare LTS, it is still restricted to $p = 1$ and x and e are independent and normally distributed, though). In the location setting ($p = 1$) the IF of the LTS estimator has been given in [36]. In this special case ($p = 1$) in our model (1.1), we have a location problem for the β_{01} and the IF was given in [45] and is bounded.

(II) If setting $\alpha \to \infty$, then one immediately obtains the influence function for LS estimating functional, $\boldsymbol{\beta}_{ls}$, which is with $\mathbf{z}_0 = (\mathbf{s}'_0, t_0)' \in \mathbb{R}^p$

$$\mathrm{IF}(\mathbf{z}_0; \boldsymbol{\beta}_{ls}, F_{(\boldsymbol{x}',y)}) = (E(\boldsymbol{w}\boldsymbol{w}'))^{-1}(1, \mathbf{s}'_0)'(t_0 - (1, \mathbf{s}'_0)\boldsymbol{\beta}_{ls}).$$

Of course, assuming that the inverse exists. Obviously, one can follow the approach in the theorem to obtain the IF for LTS in the case $p > 1$.

(III) When the depth of the residual of the contaminating point $\mathbf{z}'_0 = (\mathbf{s}'_0, t_0)$ with respect to the $\boldsymbol{\beta}_{lst}$ is larger than α, then the point mass contamination does not affect at all the functional $\boldsymbol{\beta}_{lst}$ with its influence function remaining bounded. It, unfortunately, might be unbounded (in $p > 1$ case), sharing the same drawback of that of LTS (in the $p = 1$ case). The latter was shown in [19] even in the simple regression case with x and e are independent and normally distributed. □

Proof of theorem 3.3

Insert $\boldsymbol{\beta}^{\varepsilon}_{lst}(\mathbf{z}_0) := \boldsymbol{\beta}_{lst}(F_{\varepsilon}(\mathbf{z}_0), \alpha)$ for $\boldsymbol{\beta}$ in (3.9) and take derivative with respect to ε and let $\varepsilon \to 0$, we obtain (in light of dominated convergence theorem)

$$\left(\int \frac{\partial}{\partial \boldsymbol{\beta}^{\varepsilon}_{lst}(\mathbf{z}_0)} (r(\boldsymbol{\beta}^{\varepsilon}_{lst}(\mathbf{z}_0))\boldsymbol{v}\mathbb{1}(\boldsymbol{\beta}^{\varepsilon}_{lst}(\mathbf{z}_0), F_{\varepsilon}(\mathbf{z}_0))\Big|_{\varepsilon=0} dF_{(\boldsymbol{x}',y)}\right)\dot{\boldsymbol{\beta}}_{lst}(\mathbf{z}_0, F_{(\boldsymbol{x}',y)})$$

$$+I_2 - I_3 = \mathbf{0}, \tag{8.7}$$

where $r(\boldsymbol{\beta}) = y - \boldsymbol{w}'\boldsymbol{\beta}$, $\mathbb{1}(\boldsymbol{\beta}, G) = \mathbb{1}\left(|(y - \boldsymbol{w}'\boldsymbol{\beta}) - m(G)|/\sigma(G) \le \alpha\right)$, and

$$I_2 = \int (r(\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}',y)}, \alpha))\boldsymbol{v}\mathbb{1}(\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}',y)}, \alpha), F_{(\boldsymbol{x}',y)}),$$

$$I_3 = \int (r(\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}',y)}, \alpha))\boldsymbol{w}\mathbb{1}(\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}',y)}, \alpha), F_{(\boldsymbol{x}',y)})dF_{(\boldsymbol{x}',y)}.$$

Denote by I_1 for the first term on the LHS of the above first equation. We

have $I_1 + I_2 - I_3 = 0$, and

$$I_2 - I_3 = (t_0 - (1, \mathbf{s}_0')\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}', y)}, \alpha))(1, \mathbf{s}_0')'\mathbb{1}\left(\frac{|(t_0 - (1, \mathbf{s}_0')\boldsymbol{\beta}_{lst}(F_{(\boldsymbol{x}', y)}, \alpha) - m|}{\sigma} \leq \alpha\right),$$

where the equality follows from (3.8) (i.e. $I_3 = \mathbf{0}$). The RHS of the last display is:

$$= \begin{cases} \mathbf{0}, & \text{if } t_0 - (1, \mathbf{s}_0')\boldsymbol{\beta}_{lst} \notin [m(\boldsymbol{\beta}_{lst}) \pm \alpha\sigma(\boldsymbol{\beta}_{lst})], \\ (t_0 - (1, \mathbf{s}_0')\boldsymbol{\beta}_{lst})(1, \mathbf{s}_0')', & \text{otherwise,} \end{cases}$$

Now we focus on the I_1 and especially its integrand. Denote the latter by I_4. We have

$$\begin{aligned} I_4 &= \frac{\partial}{\partial \boldsymbol{\beta}_{lst}^{\varepsilon}(\mathbf{z}_0)}\left((y - \boldsymbol{w}'\boldsymbol{\beta}_{lst}^{\varepsilon}(\mathbf{z}_0))\boldsymbol{w}\mathbb{1}\left(\frac{|(y - \boldsymbol{w}'\boldsymbol{\beta}_{lst}^{\varepsilon}(\mathbf{z}_0)) - m_{\varepsilon}(\mathbf{z}_0)|}{\sigma_{\varepsilon}(\mathbf{z}_0)} \leq \alpha\right)\right)\Bigg|_{\varepsilon=0} \\ &= \left(-\boldsymbol{w}\boldsymbol{w}'\mathbb{1}\left(\frac{|(y - \boldsymbol{w}'\boldsymbol{\beta}_{lst}^{\varepsilon}(\mathbf{z}_0)) - m_{\varepsilon}(\mathbf{z}_0)|}{\sigma_{\varepsilon}(\mathbf{z}_0)} \leq \alpha\right)\right)\Bigg|_{\varepsilon=0} \\ &+ \left((y - \boldsymbol{w}'\boldsymbol{\beta}_{lst}^{\varepsilon}(\mathbf{z}_0))\boldsymbol{w}\frac{\partial}{\partial \boldsymbol{\beta}_{lst}^{\varepsilon}(\mathbf{z}_0)}\mathbb{1}\left(\frac{|(y - \boldsymbol{w}'\boldsymbol{\beta}_{lst}^{\varepsilon}(\mathbf{z}_0)) - m_{\varepsilon}(\mathbf{z}_0)|}{\sigma_{\varepsilon}(\mathbf{z}_0)} \leq \alpha\right)\right)\Bigg|_{\varepsilon=0}. \end{aligned}$$

Hence

$$\begin{aligned} I_4 &= -\boldsymbol{w}\boldsymbol{w}'\mathbb{1}\left(\frac{|(y - \boldsymbol{w}'\boldsymbol{\beta}_{lst}) - m(\boldsymbol{\beta}_{lst})|}{\sigma(\boldsymbol{\beta}_{lst})} \leq \alpha\right) \\ &\quad + (y - \boldsymbol{w}'\boldsymbol{\beta}_{lst})\boldsymbol{w}\frac{\partial}{\partial \boldsymbol{\beta}}\mathbb{1}\left(\frac{|(y - \boldsymbol{w}'\boldsymbol{\beta}) - m(\boldsymbol{\beta})|}{\sigma(\boldsymbol{\beta})} \leq \alpha\right)\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{lst}} \\ &= -\boldsymbol{w}\boldsymbol{w}'\mathbb{1}\left(\frac{|(y - \boldsymbol{w}'\boldsymbol{\beta}_{lst}) - m(\boldsymbol{\beta}_{lst})|}{\sigma(\boldsymbol{\beta}_{lst})} \leq \alpha\right), \end{aligned}$$

where the last step follows from the proof of Lemma 3.2.

Now we have in light of (8.7)

$$\left(\int (-I_4)dF_{(\boldsymbol{x}', y)}\right)\dot{\boldsymbol{\beta}}_{lst}(\mathbf{z}_0, F_{(\boldsymbol{x}', y)}) = I_2.$$

The desired result follows. □

Proof of lemma 4.2

It suffices to establish (a), (b) follows. Put $m_{\sup} = \sup_{\boldsymbol{\beta}\in\Theta} m(F_{y-\boldsymbol{w}'\boldsymbol{\beta}})$, $m_{\inf} = \inf_{\boldsymbol{\beta}\in\Theta} m(F_{y-\boldsymbol{w}'\boldsymbol{\beta}})$, and $\sigma_{\sup} = \sup_{\boldsymbol{\beta}\in\Theta} \sigma(F_{y-\boldsymbol{w}'\boldsymbol{\beta}})$, by continuity in $\boldsymbol{\beta}$ and boundedness of Θ, all are finite numbers. Define two classes of functions for a fixed α, $m_{\sup}$, $m_{\inf}$, and $\sigma_{\sup}$ with $r(\boldsymbol{\beta}) = y - \boldsymbol{w}'\boldsymbol{\beta}$

$$\mathscr{F}_1(\boldsymbol{\beta}) := \left\{ f(\boldsymbol{x}, y, \boldsymbol{\beta}) = (r(\boldsymbol{\beta}))^2 \mathbb{1}\left(\frac{|r(\boldsymbol{\beta}) - m(F_R)|}{\sigma(F_R)} \leq \alpha \right), \boldsymbol{\beta} \in \Theta \right\},$$

$$\mathscr{F}_2(\boldsymbol{\beta}) :=$$

$$\left\{ f(\boldsymbol{x}, y, \boldsymbol{\beta}) = (r(\boldsymbol{\beta}))^2 \mathbb{1}\left(m_{\inf} - \alpha\sigma_{\sup} \leq r(\boldsymbol{\beta}) \leq m_{\sup} + \alpha\sigma_{\sup} \right), \boldsymbol{\beta} \in \Theta \right\}.$$

Obviously, $\mathscr{F}_1(\boldsymbol{\beta}) \subset \mathscr{F}_2(\boldsymbol{\beta})$. Following the notation of [20], we have for any $\boldsymbol{\beta} \in \Theta$,

$$Q(F_{\mathbf{Z}}^n, \boldsymbol{\beta}) - Q(F_{\mathbf{Z}}, \boldsymbol{\beta}) = P_n f(\boldsymbol{x}, y, \boldsymbol{\beta}) - P f(\boldsymbol{x}, y, \boldsymbol{\beta}) := P_n f - P f,$$

where $f := f(\boldsymbol{x}, y, \boldsymbol{\beta}) \in \mathscr{F}_1(\boldsymbol{\beta})$ (hereafter for consistency we assume that there is a factor $\frac{1}{n}$ in the RHS of (2.4). This will not affect the minimization or all previous discussions). And

$$\sup_{\boldsymbol{\beta}\in\Theta} |Q(F_{\mathbf{Z}}^n, \boldsymbol{\beta}) - Q(F_{\mathbf{Z}}, \boldsymbol{\beta})| = \sup_{f\in\mathscr{F}_1(\boldsymbol{\beta})} |P_n f - P f| \leq \sup_{f\in\mathscr{F}_2(\boldsymbol{\beta})} |P_n f - P f|. \tag{8.8}$$

It suffices to show the most right hand side equals to $o(1)$ a.s. (cf, supplement of [51]) for this part of proof).

To achieve that, we invoke Theorem 24 of II.5 of [20]. First $\mathscr{F}_2(\boldsymbol{\beta})$ is a permissible class of functions with an envelop $F = (m_{\sup} + \alpha\sigma_{\sup})^2$. Second, to verify the logarithm of the covering number is $o_p(n)$, by Theorem 25 of II.5 of [20], it suffices to show that the graphs of functions in $\mathscr{F}_2(\boldsymbol{\beta})$ have only polynomial discrimination (for related concepts, cf [20]), also see Example 26 of II.5 of [20] (page 29) and Example 18 of VII.4 of [20] (page 153).

The graph of a real-valued function f on a set S is defined as the subset (see page 27 of [20])

$$G_f = \{(s,t) : 0 \leq t \leq f(s) \text{ or } f(s) \leq t \leq 0, s \in S\}.$$

The graph of a function in $\mathscr{F}_2(\boldsymbol{\beta})$ contains a point $(\mathbf{x}(\omega), y(\omega), t)$ if and only if $0 \leq t \leq f(\boldsymbol{x}, y, \boldsymbol{\beta})$ or $f(\boldsymbol{x}, y, \boldsymbol{\beta}) \leq t \leq 0$. The latter case could be excluded since the function is always nonnegative (and equals 0 case covered by the former case). The former case happens if and only if $0 \leq \sqrt{t} \leq y - \boldsymbol{w}'\boldsymbol{\beta}$.

Given a collection of n points, the graph of a function in $\mathscr{F}_2(\boldsymbol{\beta})$ picks out only points that belong to $\{\sqrt{t} \geq 0\} \cap \{y - \boldsymbol{\beta}'\boldsymbol{w} - \sqrt{t} \geq 0\}$. Given n points

$(\boldsymbol{x}_i, y_i, t_i)$ $(t_i \geq 0)$, introduce n new points $(\boldsymbol{x}_i, y_i, z_i) := (\boldsymbol{x}_i, y_i, \sqrt{t_i})$ in $\mathbb{R}^{p+1}$. On $\mathbb{R}^{p+1}$ define a vector space $\mathscr{G}$ of functions

$$g_{a,b,c}(\boldsymbol{x}, y, z) = \mathbf{a}'\boldsymbol{x} + by + cz,$$

where $a \in \mathbb{R}^p$, $b \in \mathbb{R}^1$, and $c \in \mathbb{R}^1$ and $\mathscr{G} := \{g_{a,b,c}(\boldsymbol{x}, y, z) = \mathbf{a}'\boldsymbol{x} + by + cz, a \in \mathbb{R}^p, b \in \mathbb{R}^1, \text{ and } c \in \mathbb{R}^1\}$ which is $\mathbb{R}^{p+1}$-dimensional vector space.

It is clear now that the graph of a function in $\mathscr{F}_2(\boldsymbol{\beta})$ picks out only points that belong to the sets of $\{g \geq 0\}$ for $g \in \mathscr{G}$. By Lemma 18 of II.4 of [20] (page 20), the graphs of functions in $\mathscr{F}_2(\boldsymbol{\beta})$ pick only polynomial numbers of subsets of $\{w_i := (\boldsymbol{x}_i, y_i, z_i), i = 1, \cdots, n\}$; those sets corresponding to $g \in \mathscr{G}$ with $a \in \{\mathbf{0}, -\boldsymbol{\beta}\}$, $b \in \{0, 1\}$, and $c \in \{1, -1\}$ pick up even few subsets from $\{w_i, i = 1, \cdots, n\}$. This in conjunction with Lemma 15 of II.4 of [20] (page 18), yields that the graphs of functions in $\mathscr{F}_2(\boldsymbol{\beta})$ have only polynomial discrimination.

By Theorem 24 of II.5 of [20] we have completed the proof. □

Proof of lemma 4.3

Assume conversely that $\sup_{\boldsymbol{\beta} \in N_\varepsilon^c(\boldsymbol{\eta})} D(\boldsymbol{\beta}; F_{\mathbf{Z}}) = D(\boldsymbol{\eta}; F_{\mathbf{Z}})$. Then by the given conditions, there is a sequence of *bounded* $\boldsymbol{\beta_j}$ $(j = 0, 1, \cdots)$ in $N_\varepsilon^c(\boldsymbol{\eta})$ such that $\boldsymbol{\beta_j} \to \boldsymbol{\beta_0} \in N_\varepsilon^c(\boldsymbol{\eta})$ and $D(\boldsymbol{\beta_j}; F_{\mathbf{Z}}) \to D(\boldsymbol{\eta}; F_{\mathbf{Z}})$ as $j \to \infty$. Note that $D(\boldsymbol{\eta}; F_{\mathbf{Z}}) > D(\boldsymbol{\beta_0}; F_{\mathbf{Z}})$. The continuity of $D(\cdot; F_{\mathbf{Z}})$ now leads to a contradiction: for sufficiently large j, $D(\boldsymbol{\beta_j}; F_{\mathbf{Z}}) \leq (D(\boldsymbol{\eta}; F_{\mathbf{Z}}) + D(\boldsymbol{\beta_0}; F_{\mathbf{Z}}))/2 < D(\boldsymbol{\eta}; F_{\mathbf{Z}})$. This completes the proof. □

Proof of theorem 4.3

For convenience of description, we write

$$\mathbb{1}(\boldsymbol{\beta}, F_{r(\boldsymbol{\beta})}) := \mathbb{1}\left(\frac{|y - \boldsymbol{w'}\boldsymbol{\beta} - m(F_{r(\boldsymbol{\beta})})|}{\sigma(F_{r(\boldsymbol{\beta})})} \leq \alpha\right), \tag{8.9}$$

where $r(\boldsymbol{\beta}) = y - \boldsymbol{w'}\boldsymbol{\beta}$ and $m(F_{r(\boldsymbol{\beta})})$ and $\sigma(F_{r(\boldsymbol{\beta})})$ are the median and MAD of the distribution of $r(\boldsymbol{\beta})$.

Adding the derivative of $Q(\mathbf{Z}^{(n)}, \boldsymbol{\beta}, \alpha)$ with respect to $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ to the both sides of equation (2.8) and multiplying $1/(2\sqrt{n})$ we obtain

$$\frac{1}{\sqrt{n}} \sum_i (y_i - \boldsymbol{w}_i'\boldsymbol{\beta}_0)\boldsymbol{w}_i \mathbb{1}(\boldsymbol{\beta}_0, F^n_{r(\boldsymbol{\beta}_0)}) =$$

$$\frac{1}{\sqrt{n}} \sum_i \boldsymbol{w}_i \boldsymbol{w'}_i (\widehat{\boldsymbol{\beta}}^n_{lst} - \boldsymbol{\beta}_0) \mathbb{1}(\boldsymbol{\beta}_0, F^n_{r(\boldsymbol{\beta}_0)})$$

$$-\frac{1}{\sqrt{n}} \sum_i r_i(\widehat{\boldsymbol{\beta}}^n_{lst}) \boldsymbol{w}_i \Big[\mathbb{1}(\widehat{\boldsymbol{\beta}}^n_{lst}, F^n_{r(\widehat{\boldsymbol{\beta}}^n_{lst})}) - \mathbb{1}(\boldsymbol{\beta}_0, F^n_{r(\boldsymbol{\beta}_0)})\Big],$$

where $\mathbb{1}(\boldsymbol{\beta}, F^n_{r(\boldsymbol{\beta})})$ has the same meaning as in (8.9) except that the median and MAD are the sample version, respectively based on $\{y_i - \boldsymbol{w'}_i\boldsymbol{\beta}\}$. For further simplicity, we write $\mathbb{1}(\boldsymbol{\beta}, n)$ for $\mathbb{1}(\boldsymbol{\beta}, F^n_{r(\boldsymbol{\beta})})$, and I_0 for the LHS of the equation above. Rewrite the RHS of the equation above, we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_i (y_i - \boldsymbol{w}_i'\boldsymbol{\beta}_0)\boldsymbol{w}_i\mathbb{1}(\boldsymbol{\beta}_0, F^n_{r(\boldsymbol{\beta}_0)}) = \frac{1}{n}\sum_i \boldsymbol{w}_i\boldsymbol{w}_i'\mathbb{1}(\boldsymbol{\beta}_0, n)\sqrt{n}(\widehat{\boldsymbol{\beta}}^n_{lst} - \boldsymbol{\beta}_0)\\ + \frac{1}{n}\sum_i \boldsymbol{w}_i\boldsymbol{w}_i'\Big[\mathbb{1}(\widehat{\boldsymbol{\beta}}^n_{lst}, n) - \mathbb{1}(\boldsymbol{\beta}_0, n)\Big]\sqrt{n}(\widehat{\boldsymbol{\beta}}^n_{lst} - \boldsymbol{\beta}_0)\\ - \frac{1}{\sqrt{n}}\sum_i e_i\boldsymbol{w}_i\Big[\mathbb{1}(\widehat{\boldsymbol{\beta}}^n_{lst}, n) - \mathbb{1}(\boldsymbol{\beta}_0, n)\Big]\end{aligned}$$

Denote the three terms on the RHS above as I_1, I_2, and I_3, respectively. Now we have, based on the short notations,

$$I_0 = I_1 + I_2 + I_3.$$

If we can show that $I_0 = O_p(1)$, $I_1 = (O_p(1) + o_p(1))\sqrt{n}(\widehat{\boldsymbol{\beta}}^n_{lst} - \boldsymbol{\beta}_0)$, $I_2 = o_p(1)\sqrt{n}(\widehat{\boldsymbol{\beta}}^n_{lst} - \boldsymbol{\beta}_0)$, and $I_3 = o_p(1)$, then the desired result follows immediately. On the other hand, these results are established in Lemmas 4.4 and 4.5. This completes the proof. □

Lemma 4.4 With the assumptions **(A3)-(A4)**, we have

$$\frac{1}{\sqrt{n}}\sum_i (y_i - \boldsymbol{w}_i'\boldsymbol{\beta}_0)\boldsymbol{w}_i\mathbb{1}(\boldsymbol{\beta}_0, F^n_{R(\boldsymbol{\beta}_0)}) = O_p(1).$$

Proof: Notice that $y_i - \boldsymbol{w}_i'\boldsymbol{\beta}_0 = e_i$. It suffices to show that

$$\frac{1}{\sqrt{n}}\sum_i e_i\boldsymbol{w}_i = O_p(1).$$

This however follows straightforwardly from the CLT and $\mathrm{E}(e_i\boldsymbol{w}_i) = 0$. □

Lemma 4.5 With the assumptions **(A0)-(A4)**, we have

$$\frac{1}{n}\sum_i \boldsymbol{w}_i\boldsymbol{w}_i'\mathbb{1}(\boldsymbol{\beta}_0, n)\sqrt{n}(\widehat{\boldsymbol{\beta}}^n_{lst} - \boldsymbol{\beta}_0) = (O_p(1) + o_p(1))\sqrt{n}(\widehat{\boldsymbol{\beta}}^n_{lst} - \boldsymbol{\beta}_0), \tag{8.10}$$

$$\frac{1}{n}\sum_i \boldsymbol{w}_i\boldsymbol{w}_i'\Big[\mathbb{1}(\widehat{\boldsymbol{\beta}}^n_{lst}, n) - \mathbb{1}(\boldsymbol{\beta}_0, n)\Big]\sqrt{n}(\widehat{\boldsymbol{\beta}}^n_{lst} - \boldsymbol{\beta}_0) = o_p(1)\sqrt{n}(\widehat{\boldsymbol{\beta}}^n_{lst} - \boldsymbol{\beta}_0), \tag{8.11}$$

$$\frac{1}{\sqrt{n}}\sum_i e_i\boldsymbol{w}_i\Big[\mathbb{1}(\widehat{\boldsymbol{\beta}}^n_{lst}, n) - \mathbb{1}(\boldsymbol{\beta}_0, n)\Big] = o_p(1). \tag{8.12}$$

Proof: By theorems 4.1 and 4.2, we have that $\widehat{\boldsymbol{\beta}}_{lst}^n - \boldsymbol{\beta}_0 = o(1)$ a.s. Furthermore, sample median $m(F^n_{r(\boldsymbol{\beta}_0)})$ converges to its popular version $m(F_{r(\boldsymbol{\beta}_0)})$ a.s. by Glivenko-Cantelli theorem, the continuity of the median functional (see page 7 of [20]), and Theorem 2.3.1 of [29]), hence we have

$$\mathbb{1}(\boldsymbol{\beta}_0, n) = \mathbb{1}(\boldsymbol{\beta}_0, F_{r(\boldsymbol{\beta}_0)}) + o(1), a.s. \text{ and } \mathbb{1}(\widehat{\boldsymbol{\beta}}_{lst}^n, n) - \mathbb{1}(\boldsymbol{\beta}_0, n) = o(1), \ a.s.$$

In light of the CLT and by **(A3)** and **(A4)**, we have that

$$\frac{1}{\sqrt{n}} \sum_i e_i \boldsymbol{w}_i = \sqrt{n} E(e\boldsymbol{w}) + O_p(1) = O_p(1).$$

Now in virtue of the LLN, we have that

$$\frac{1}{n} \sum_i \boldsymbol{w}_i \boldsymbol{w}_i' = E(\boldsymbol{w}\boldsymbol{w}') + o_p(1).$$

The last three displays lead to the desired results. □

Proof of Theorem 5.1

In order to apply the Lemma 5.1, we first realize that in our case, $\widehat{\boldsymbol{\beta}}_{lst}^n$ and $\boldsymbol{\beta}_{lst}$ correspond to τ_n and t_0 (assume, w.l.o.g. that $\boldsymbol{\beta}_{lts} = \mathbf{0}$ in light of regression equivariance); $\boldsymbol{\beta}$ and Θ correspond to t and T; $f(\cdot, t) := f(\cdot, \cdot, \boldsymbol{\beta}, \alpha)$ and α is a fixed constant, where $f(\boldsymbol{x}, y, \boldsymbol{\beta}, \alpha) = r^2 \mathbb{1}(F_{(\boldsymbol{x}', y)}, \boldsymbol{\beta})$ and $\mathbb{1}(F_{(\boldsymbol{x}', y)}, \boldsymbol{\beta}) := \mathbb{1}\left(\frac{|y - \boldsymbol{w}'\boldsymbol{\beta} - \mu(F_r)|}{\sigma(F_r)} \le \alpha\right)$, $r := r(\boldsymbol{\beta}) = y - \boldsymbol{w}'\boldsymbol{\beta}$. In our case,

$$\nabla(\boldsymbol{x}, y, \boldsymbol{\beta}, \alpha) = \frac{\partial}{\partial \boldsymbol{\beta}} f(\boldsymbol{x}, y, \boldsymbol{\beta}, \alpha) = 2(y - \boldsymbol{w}'\boldsymbol{\beta}) \boldsymbol{w} \mathbb{1}(F_{(\boldsymbol{x}', y)}, \boldsymbol{\beta}, \alpha).$$

We will have to assume that $P(\nabla_i^2) = P(4(y - \boldsymbol{w}'\boldsymbol{\beta})^2 w_i^2 \mathbb{1}(F_{(\boldsymbol{x}', y)}, \boldsymbol{\beta}, \alpha)$ exists to meet (iv) of the lemma, where $i \in \{1, \cdots, p\}$ and $\boldsymbol{w}' = (w_1, \cdots, w_p) = (1, \boldsymbol{x}')$. It is readily seen that a sufficient condition for this assumption to hold is the existence of $P(x_i^2)$. In our case, $V = 2P(\boldsymbol{w}\boldsymbol{w}' \mathbb{1}(F_{(\boldsymbol{x}', y)}, \boldsymbol{\beta}, \alpha)$, we will have to assume that it is invertible when $\boldsymbol{\beta}$ is replaced by $\boldsymbol{\beta}_{lst}$ (it is covered by the assumption in Theorem 3.2) to meet (ii) of the lemma. In our case,

$$r(\cdot, t) = \left(\frac{\boldsymbol{\beta}'}{\|\boldsymbol{\beta}\|} V / 2 \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \right) \|\boldsymbol{\beta}\|.$$

We will assume that λ_{min} and λ_{max} are the minimum and maximum eigenvalues of positive semidefinite matrix V overall $\boldsymbol{\beta} \in \Theta$ and a fixed $\alpha \ge 1$.

Now to apply Lemma 5.1, we need to verify the five conditions, among them only (iii) and (v) need to be addressed, all others are satisfied trivially. For (iii), it holds automatically since our $\tau_n = \widehat{\boldsymbol{\beta}}_{lst}^n$ is defined to be the minimizer of $F_n(t)$ over $t \in T (= \Theta)$.

So the only condition that needs to be verified is the (v), the stochastic equicontinuity of $\{E_n r(\cdot,t)\}$ at t_0. For that, we will appeal to the Equicontinuity Lemma (VII.4 of [20], page 150). To apply the Lemma, we will verify that the condition for the random covering numbers satisfy the uniformity condition. To that end, we look at the class of functions for a fixed $\alpha \geq 1$

$$\mathscr{R}(\boldsymbol{\beta}) = \left\{ r(\cdot,\cdot,\alpha,\boldsymbol{\beta}) = \left(\frac{\boldsymbol{\beta}'}{\|\boldsymbol{\beta}\|} V/2 \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \right) \|\boldsymbol{\beta}\| : \ \boldsymbol{\beta} \in \Theta \right\}.$$

Obviously, $\lambda_{max} r_0/2$ is an envelope for the class $\mathscr{R}$ in $\mathscr{L}^2(P)$, where r_0 is the radius of the ball $\Theta = B(\boldsymbol{\beta}_{lts}, r_0)$. We now show that the covering numbers of $\mathscr{R}$ are uniformly bounded, which amply suffices for the Equicontinuity Lemma. For this, we will invoke Lemmas II.25 and II.36 of [20]. To apply Lemma II.25, we need to show that the graphs of functions in $\mathscr{R}$ have only polynomial discrimination.

The graph of a real-valued function f on a set S is defined as the subset (see page 27 of [20])

$$G_f = \{(s,t) : 0 \leq t \leq f(s) \text{ or } f(s) \leq t \leq 0, s \in S\}.$$

The graph of $r(\boldsymbol{x}, y, \alpha, \boldsymbol{\beta})$ contains a point $(\boldsymbol{x}, y, t)$, $t \geq 0$ iff $\left(\frac{\boldsymbol{\beta}'}{\|\boldsymbol{\beta}\|} V/2 \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} \right)$ $\|\boldsymbol{\beta}\| \geq t$ for all $\boldsymbol{\beta} \in \Theta$. Equivalently, the graph of $r(\boldsymbol{x}, y, \alpha, \boldsymbol{\beta})$ contains a point $(\boldsymbol{x}, y, t)$, $t \geq 0$ iff $\lambda_{min}/2\|\boldsymbol{\beta}\| \geq t$. For a collection of n points $(\boldsymbol{x}_i', y_i, t_i)$ with $t_i \geq 0$, the graph picks out those points satisfying $\lambda_{min}/2\|\boldsymbol{\beta}\| - t_i \geq 0$. Construct from $(\boldsymbol{x}_i, y_i, t_i)$ a point $z_i = t_i$ in $\mathbb{R}$. On $\mathbb{R}$ define a vector space $\mathscr{G}$ of functions

$$g_{a,b}(x) = ax + b, \quad a, \ b \in \mathbb{R}.$$

By Lemma 18 of [20], the sets $\{g \geq 0\}$, for $g \in \mathscr{G}$, pick out only a polynomial number of subsets from $\{z_i\}$; those sets corresponding to functions in $\mathscr{G}$ with $a = -1$ and $b = \lambda_{min}/2\|\boldsymbol{\beta}\|$ pick out even fewer subsets from $\{z_i\}$. Thus the graphs of functions in $\mathscr{R}$ have only polynomial discrimination. □

Transformation in Section 5 before Corollary 5.1 Assume the Cholesky decomposition of $\boldsymbol{\Sigma}$ in (5.4) yields a nonsingular lower triangular matrix $\boldsymbol{L}$ of the form

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{v}' & c \end{pmatrix}$$

with $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'$. Hence $\det(\boldsymbol{A}) \neq 0 \neq c$. Now transfer $(\boldsymbol{x}', y)$ to $(\boldsymbol{s}', t)$ with $(\boldsymbol{s}', t)' = \boldsymbol{L}^{-1}((\boldsymbol{x}', y)' - \boldsymbol{\mu})$. It is readily seen that the distribution of $(\boldsymbol{s}', t)'$ follows $E(g; \boldsymbol{0}, \boldsymbol{I}_{\boldsymbol{p}\times\boldsymbol{p}})$.

Note that $(\boldsymbol{x}', y)' = \boldsymbol{L}(\boldsymbol{s}', t)' + (\boldsymbol{\mu}_1', \mu_2)'$ with $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \mu_2)'$. That is,

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s} + \boldsymbol{\mu}_1, \tag{8.13}$$

$$y = \boldsymbol{v}'\boldsymbol{s} + ct + \mu_2. \tag{8.14}$$

Equivalently,

$$(1, \boldsymbol{s}')' = \boldsymbol{B}^{-1}(1, \boldsymbol{x}')', \tag{8.15}$$

$$t = \frac{y - (1, \boldsymbol{s}')(\mu_2, \boldsymbol{v}')'}{c}, \tag{8.16}$$

where

$$\boldsymbol{B} = \begin{pmatrix} 1 & \boldsymbol{0}' \\ \boldsymbol{\mu}_1 & \boldsymbol{A} \end{pmatrix}, \quad \boldsymbol{B}^{-1} = \begin{pmatrix} 1 & \boldsymbol{0}' \\ -\boldsymbol{A}^{-1}\boldsymbol{\mu}_1 & \boldsymbol{A}^{-1} \end{pmatrix},$$

It is readily seen that (8.15) is an affine transformation on $\boldsymbol{w}$ and (8.16) is first an affine transformation on $\boldsymbol{w}$ then a regression transformation on y followed by a scale transformation on y. In light of Theorem 2.4, we can assume hereafter, w.l.o.g. that $(\boldsymbol{x}', y)$ follows an $E(g; \boldsymbol{0}, \boldsymbol{I}_{p\times p})$ (spherical) distribution and $\boldsymbol{I}_{p\times p}$ is the covariance matrix of $(\boldsymbol{x}', y)$.

Remark 6.1
(I) Stopping criteria for the algorithm include (i) the total number of the LS estimation decided to perform (ii) the total number of two indices sampled from $\{1, 2, \cdots, n\}$ or (iii) the total number of distinct index sequences $i_1, \cdots, i_K$ in the step (a2) of (3).

(II) There are $O(n^2)$ two-point pairs, all other operations cost at most $O(np^2 + p^3)$, theoretically, overall the worst time complexity is $O(pn^3 + n^2p^3)$. However, in the program, N is the minimum of $\{1000, \binom{n}{\lfloor (n+1)/2 \rfloor}, T_{ls}\}$, where T_{ls} is a turning parameter, the total number of the LS estimation decided to perform, which usually set to be $100 \sim 500$, so in practice the real time complexity is $O(np^2 + p^3)$ (see Section 7).

(III) When $\boldsymbol{x}_i = \boldsymbol{x}_j$ for some $i \neq j$, one can add a small ε say, to $\boldsymbol{x}_i$, to force them are not identical. So that one can still apply the AA1. □

Remark 6.2
(I) It is readily seen that the worst case time complexity of algorithm AA2 is $O(N(p^2n + p^3))$ where p^3 comes from finding the inverse of p by p matrix and from $p \times p$ matrix multiply a p vector and the most costly step is (1) to compute the $I(\boldsymbol{\beta}_{new})$ which, however, can achieve in $O(np^2)$. When n and p are small (say $n \leq 50$, $p \leq 3$), then N might just be $\binom{n}{p}$, otherwise it will be $300(p-1)$. Here 300 could be tuned to a larger number - such as 500 - or even larger. It is readily seen that the AA2 produces a non-negative and non-increasing sequence: $Q_1 > Q_2 \cdots > Q_k > \cdots$. So the convergence of AA2 is always achievable.

(II) For large n, say $n \geq 200$, we suggest that one first partitions the data set into disjoint (say five) subsets, then applies the AA2 to each subset to obtain $\boldsymbol{\beta}$ from each subset. Finally, one carries out step (1) above with respect to the entire data set and selects the $\boldsymbol{\beta}$ which produces the smallest objective function value $Q(\boldsymbol{\beta})$.

(III) In the algorithm AA2, the sub-sample size m is p. Other choices include $\lfloor (n+1)/2 \rfloor$ (corresponding to $\alpha = 1$) and $I(\boldsymbol{\beta}_{new})$ (which requires an initial $\boldsymbol{\beta}_{new}$). The latter however is generally not recommended. □