TEST (2015) 24:796–812
DOI 10.1007/s11749-015-0435-5

ORIGINAL PAPER

Sharp non-asymptotic performance bounds for ℓ_1 and Huber robust regression estimators

Salvador Flores[1]

Received: 24 July 2014 / Accepted: 5 March 2015 / Published online: 14 March 2015

Abstract A quantitative study of the robustness properties of the ℓ_1 and the Huber M-estimator on finite samples is presented. The focus is on the linear model involving a fixed design matrix and additive errors restricted to the dependent variables consisting of noise and sparse outliers. We derive sharp error bounds for the ℓ_1 estimator in terms of the leverage constants of a design matrix introduced here. A similar analysis is performed for Huber's estimator using an equivalent problem formulation of independent interest. Our analysis considers outliers of arbitrary magnitude, and we recover breakdown point results as particular cases when outliers diverge. The practical implications of the theoretical analysis are discussed on two real datasets.

Keywords ℓ_1 norm minimization · Huber M-estimator · Leverage constants · Sparse outliers · Breakdown point · Leverage plot

Mathematics Subject Classification 62J05 · 62F35 · 90C31

1 Introduction

In classical linear regression, a vector of responses or dependent variables $y \in \mathbb{R}^n$ is given along with the same number of explanatory variables or carriers $x_1, \ldots, x_n \in \mathbb{R}^p$. We assume that the random variables $x_1, \ldots, x_n$, and y are related through a linear model, which implies the existence of a vector $f \in \mathbb{R}^p$ such that

✉ Salvador Flores
sflores@dim.uchile.cl

[1] Centro de Modelamiento Matemático (CNRS UMI 2807), Universidad de Chile, Beauchef 851, Santiago, Chile

$$(\forall i \in \{1, \ldots, n\}) \quad y_i = x_i^\top f + \delta_i, \tag{1}$$

where $(\delta_i)_{1 \le i \le n}$ are i.i.d. random variables with zero mean and finite variance. The objective in linear regression is to estimate f.

Under the usual assumption that the errors δ_i are Gaussian, the least squares estimator (LSE) is the best linear unbiased estimator of f. However, the LSE is very sensitive to deviations from normality, even moderate ones. The ability of an estimation method to give reasonable results on contaminated samples is measured by the regression breakdown point (RBP), defined in the fixed design context as the minimum fraction of the components of δ that must diverge in an arbitrary way to take the estimator out of any bound (see, e.g. He et al. 1990; Giloni and Padberg 2004). For example, the LSE has an asymptotic RBP of 0 %, since a single divergent observation can completely mislead the fit, independently of the sample size. M-estimators (Huber 1973; Rousseeuw and Leroy 1987) aim to perform robust and computationally efficient estimation. The quantitative study of the robustness properties of M-estimators for non-random carriers (also called *fixed design*) was started by He et al. (1990). In that work, the authors introduce a finite-sample measure of performance for regression estimators based on tail behaviour. For the ℓ_1-estimator as well as for a class of M-estimators, their tail performance measure equals the RBP; a simple characterization of the RBP in terms of the design configuration is provided. In particular, they show that the RBP of the ℓ_1 estimator can be positive if the matrix X is not subject to contamination, closing a long-standing discussion about the robustness of the ℓ_1 estimator. The same expression for the RBP is obtained by Ellis and Morgenthaler (1992), where its role as a leverage measure is studied as well. Giloni and Padberg (2004) obtain an alternative characterization of the RBP using mixed-integer programming.

In the context of signal processing, the estimation problem is considered by Candes and Tao (2005). They assume that the vector δ in (1) is *sparse*, i.e., only a small fraction of the observations is contaminated and the rest is completely free of errors. They provide sufficient conditions for *exact recovery* of a signal from corrupted measurements. The sufficient condition is known as the restricted isometry property (RIP) and it is verified with high probability for random normal matrices X when n and p go to infinity in a proper ratio. Later, in Candes and Randall (2008), a modification of ℓ_1 minimization for linear regression is proposed to deal with outliers and noise. The sufficient conditions for the noiseless case are adapted to this more realistic context. However, the analysis is restricted to the particular instance when X is normal random and has orthonormal columns. Leaving aside the drawbacks of the RIP (c.f. Zhang 2013, Sect. 1.3), any error analysis taking the design matrix X as a degree of freedom rather than as part of the data of the problem is unsatisfactory, because in many applications the design is fixed and non-scalable. Likewise, the notion of breakdown point gives information on the behaviour of an estimator when data are replaced by divergent observations; nonetheless, it is more informative to have a quantitative measure of the prediction error when some observations are affected by finite errors of any magnitude that cannot be reasonably considered as noise.

We fill this gap by providing non-asymptotic error bounds in finite samples for two of the most widespread convex robust estimators.

1.1 Notation and preliminaries

We shall use the notation $N = \{1, \ldots, n\}$ for the index set of all the observations. For a set of indexes M, $|M|$ denotes its cardinality. For a vector $x \in \mathbb{R}^n$, we denote by $\operatorname{supp}(x)$ its support, i.e., the index set of non-zero components, $\operatorname{supp}(x) = \{i \in N \mid x_i \neq 0\}$. The cardinality of the support of a vector, often called the "ℓ_0-norm" or "cardinality norm", is denoted by $\|x\|_0$; thus $\|x\|_0 = |\{i \in N \mid x_i \neq 0\}|$. For a subset $M \subseteq N$ and $p \in [1, +\infty[$, we define

$$\|\cdot\|_{p,M} : x \mapsto \left(\sum_{i \in M} |x_i|^p\right)^{1/p} \quad \text{and} \quad \|\cdot\|_{\infty,M} : x \mapsto \max_{i \in M} |x_i|.$$

Moreover, for every $x \in \mathbb{R}^n$ and $p \in [1, +\infty[$, we denote $\|x\|_p = \|x\|_{p,N}$ and $\|x\|_\infty = \|x\|_{\infty,N}$.

For the sake of readability, we postpone some lemmas and proofs to Appendix A.

2 Range conditions on the design matrix

We carry out a non-asymptotic analysis of two estimation techniques which are valid for any sample size, *ergo* for an arbitrary design matrix X. To this end we introduce the *leverage constants* of a matrix, measuring the relative weight of the most influential observations on the fit.

For a $n \times p$ matrix X, define for every $k \in \{1, \ldots, n\}$ the *leverage constants* c_k of X as

$$c_k(X) = \min_{\substack{M \subset N \\ |M| = k}} \min_{\substack{g \in \mathbb{R}^p \\ g \neq 0}} \frac{\sum_{i \in N \setminus M} |x_i^\top g|}{\sum_{i \in N} |x_i^\top g|} = \min_{\substack{M \subset N \\ |M| = k}} \min_{\substack{g \in \mathbb{R}^p \\ \|g\|_2 = 1}} \frac{\sum_{i \in N \setminus M} |x_i^\top g|}{\sum_{i \in N} |x_i^\top g|} \tag{2}$$

and

$$m(X) = \max\left\{k \in N \;\middle|\; c_k(X) > \frac{1}{2}\right\}. \tag{3}$$

Note that the two minima in (2) are achieved since the feasible set in both cases is compact and the objective function is continuous. However, the minimum may not be unique as there could be two or more groups of points with equal leverage.

Proposition 1 *We have $c_0 = 1$, $c_n = 0$ and, for every $k \in \{1, \ldots, n\}$, $c_k \leq c_{k-1}$.*

The quantity $m(X)$ defined above is already known to characterize, up to a constant, the RBP of the ℓ_1 and Huber's estimators. The leverage constants of a matrix provide the essential information for describing the response of a class of estimates to groups of influential observations (see also Ellis and Morgenthaler 1992, for a related discussion).

The main results in this article rely on the following fundamental ℓ_1 error estimate, which is inspired on (He et al. 1990, Lemma 5.2). When there is no place for confusion, we shall omit the dependency of the constants c_k on X.

Lemma 1 (ℓ_1 error estimate) *Let X be a $n \times p$ real matrix, and $(c_k)_{1\le k\le n}$ and $m(X)$ be defined as in (2) and (3), respectively. In addition, let $M \subset N$, and $y, b^* \in \mathbb{R}^n$ as well as $g^*, g \in \mathbb{R}^p$ be arbitrary. The following holds.*

(i) *Suppose that $|M| = k < m(X)$. Then*

$$\begin{aligned}&\|y - Xg - b^*\|_1 - \|y - Xg^* - b^*\|_1\\ &\quad\ge (2c_k - 1)\|X(g - g^*)\|_1 - 2\sum_{i\in N\setminus M} |y_i - x_i^\top g^* - b_i^*|.\end{aligned}$$

(ii) *Suppose that $|M| = 0$. Then, for every $b \in \mathbb{R}^n$,*

$$\begin{aligned}&\|y - Xg - b\|_1 - \|y - Xg^* - b^*\|_1\\ &\quad\ge \|X(g - g^*) + b - b^*\|_1 - 2\sum_{i\in N} |y_i - b_i^* - x_i^\top g^*|.\end{aligned}$$

The inequalities (i) and (ii) are sharp for there exist values of y, b^*, g^* and g for which equality holds.

Proof (i): Let $y, b^* \in \mathbb{R}^n$ and $g^*, g \in \mathbb{R}^p$. We have

$$\begin{aligned}\|y - Xg - b^*\|_1 &= \sum_{i\in N} |y_i - x_i^\top g - b_i^*|\\ &= \sum_{i\in N} |(y_i - x_i^\top g^* - b_i^*) - (x_i^\top g - x_i^\top g^*)|\\ &= \sum_{i\in N\setminus M} |(x_i^\top g - x_i^\top g^*) - (y_i - x_i^\top g^* - b_i^*)|\\ &\quad + \sum_{i\in M} |(y_i - x_i^\top g^* - b_i^*) - (x_i^\top g - x_i^\top g^*)|\end{aligned}$$

and using the reverse triangle inequality $|u - v| \ge ||u| - |v|| \ge |u| - |v|$ we obtain

$$\begin{aligned}\|y - Xg - b^*\|_1 \ge\ & 2\sum_{i\in N\setminus M} |x_i^\top g - x_i^\top g^*| - \sum_{i\in N} |x_i^\top g - x_i^\top g^*|\\ &+ \sum_{i\in N} |y_i - x_i^\top g^* - b_i^*| - 2\sum_{i\in N\setminus M} |y_i - x_i^\top g^* - b_i^*|. \qquad (4)\end{aligned}$$

It follows from (3) and (2) that $c_k > 1/2$ and there exists $g_k \ne g^*$ such that

$$(\forall g, g^* \in \mathbb{R}^p) \quad \text{s.t.} \quad g \ne g^* \quad \frac{\sum_{i\in N\setminus M} |x_i^\top (g - g^*)|}{\sum_{i\in N} |x_i^\top (g - g^*)|} \ge \frac{\sum_{i\in N\setminus M} |x_i^\top (g_k - g^*)|}{\sum_{i\in N} |x_i^\top (g_k - g^*)|} = c_k,$$

Thus,

$$\sum_{i\in N\setminus M} |x_i^\top (g-g^*)| \geq c_k \sum_{i\in N} |x_i^\top (g-g^*)|.$$

By replacing in (4) we obtain

$$\|y-Xg-b^*\|_1 - \|y-Xg^*-b^*\|_1 \geq (2c_k-1)\|X(g-g^*)\|_1 - 2\sum_{i\in N\setminus M} |y_i - x_i^\top g^* - b_i^*|$$

and the result holds.

(ii): The result is a direct consequence of the triangle inequality for the ℓ_1 norm. □

3 Characterization of the behaviour of the ℓ_1-estimator

In this section, we study the problem of estimating by ℓ_1 minimization the vector f from observations of the form

$$y = Xf + z + e, \tag{5}$$

where z is a dense vector of noise and e is an arbitrary sparse vector modelling outliers. Since the LSE is optimal in the absence of outliers, we measure the reconstruction error by comparing the ℓ_1 estimator f_1 with f_n, which is the LSE applied to the noisy part of the data, devoid of outliers.

Theorem 1 *Let $y = Xf + z + e$ and $M = \mathrm{supp}(e)$ satisfying $|M| = k \leq m(X)$. Consider the unique decomposition of z as $z = X\overline{g} + \overline{b}$, where $\overline{b} \in Ker X^\top$, and let $f_n = f + \bar{g}$. Then, the following holds for the ℓ_1 estimator f_1.*

(i) *If $\|\overline{b}\|_{\infty, N\setminus M} = 0$, then $f_1 = f_n$.*
(ii) *If $\|\overline{b}\|_{\infty, N\setminus M} > 0$, then*

$$\|X(f_1 - f_n)\|_1 \leq \frac{1}{2c_k - 1}\left(\|\bar{b}\|_{1,N\setminus M} + \frac{\|\bar{b}\|^2_{2,N\setminus M}}{\|\bar{b}\|_{\infty,N\setminus M}}\right). \tag{6}$$

There exist data for which inequality (6) holds with equality; therefore, the estimate cannot be further improved.

The estimates of Theorem 1 can be easily extended to the case when the number of outliers exceeds $m(X)$ by taking M to be the set of indices of the components of e with the m largest absolute values. Thus, the ℓ_1 estimator mitigates the effect of the largest outliers. They can also be generalized to weighted ℓ_1 regression (Giloni et al. 2006a, b) by adding weights to the absolute values in the definition of the leverage constants. The use of the leverage constants to improve the RBP of weighted ℓ_1 regression is the subject of current research.

Giloni and Padberg (2002, Prop. 1) proved, without assuming model (5), that $\|y - Xf_1\|_1 \geq \|\bar{y}\|_2^2/\|\bar{y}\|_\infty$, where $\bar{y} = y - Xf_{LS}$. The ideas behind that result are generalized in Lemma 4.

Proof of Theorem 1. Using Lemma 1(i) with $b^* = 0$, $g = f_1$, and $g^* = f_n$ we obtain

$$\|y - Xf_1\|_1 - \|y - Xf_n\|_1 \geq (2c_k - 1)\|X(f_1 - f_n)\|_1 - 2\sum_{i \in N\setminus M} |y_i - x_i^\top f_n|.$$

Since, by hypothesis, $y_i = x_i^\top(f + \bar{g}) + \bar{b}_i = x_i^\top f_n + \bar{b}_i$ for $i \in N\setminus M$ we have

$$(2c_k - 1)\|X(f_1 - f_n)\|_1 \leq 2\|\bar{b}\|_{1,N\setminus M} + \|y - Xf_1\|_1 - \|y - Xf_n\|_1. \tag{7}$$

First note that since f_1 is a minimizer, $\|y - Xf_1\|_1 - \|y - Xf_n\|_1 \leq 0$. Thus if $\|\bar{b}\|_{\infty,N\setminus M} = 0$ it follows from (7), the full rank of X, and $c_k > 1/2$ that $f_1 = f_n$. Now suppose that $\|\bar{b}\|_{\infty,N\setminus M} > 0$. The ℓ_1 minimization problem can be formulated as a linear program; using linear programming duality, we have (Giloni and Padberg 2004, pp. 1031–1032)

$$\|y - Xf_1\|_1 = \min_{g\in\mathbb{R}^p} \|y - Xg\|_1 = \max_{d\in P^*} d^\top y = \max_{d\in P^*} d^\top(e + \bar{b}),$$

where $P^* = \left\{d \in \ker X^\top \;\middle|\; \|d\|_\infty \leq 1\right\}$. Thus,

$$\|y - Xf_1\|_1 - \|y - Xf_n\|_1 = \max_{d\in P^*} d^\top(e + \bar{b}) - \|e + \bar{b}\|_1.$$

Hence, using Lemma 4, we obtain

$$\begin{aligned}\|y - Xf_1\|_1 - \|y - Xf_n\|_1 &\leq \|e + \bar{b}\|_{1,M} + \frac{\|\bar{b}\|_{2,N\setminus M}^2}{\|\bar{b}\|_{\infty,N\setminus M}} - \|e + \bar{b}\|_1 \\ &= -\|\bar{b}\|_{1,N\setminus M} + \frac{\|\bar{b}\|_{2,N\setminus M}^2}{\|\bar{b}\|_{\infty,N\setminus M}}\end{aligned}$$

which, altogether with (7), yields (6). □

In the particular case when only sparse errors are present ($z = 0$), the following result is a characterization of the exact recovery property (see also Zhang 2013; Giloni and Padberg 2002, for related results).

Theorem 2 *Let $f \in \mathbb{R}^p$, $e \in \mathbb{R}^n$, and set $y = Xf + e$. Then, f is the unique solution of the problem*

$$\min_{g\in\mathbb{R}^d} \|y - Xg\|_1.$$

for any $\|e\|_0 \leq k$ if and only if $k \leq m(X)$.

Proof First note that, in this case, $f_n = f$. If $\|e\|_0 \leq m(X)$, using Theorem 1 with $z = 0$, we obtain that $X(f_1 - f_n) = X(f_1 - f) = 0$, and since X has full rank, we conclude that $f_1 = f$. Now let us show that for $k = \|e\|_0 > m(X)$ we can find an instance of the problem for which f, whether is not a solution, or it is not the unique solution. Let $f \in \mathbb{R}^p$ be arbitrary. From the definition of c_k, there exists $g_k \in \mathbb{R}^p$ such that $\|g_k\|_2 = 1$ and $M \subseteq N$, $|M| = k$ such that

$$\sum_{i \in N \setminus M} |x_i^\top g_k| \leq \sum_{i \in M} |x_i^\top g_k|. \tag{8}$$

Now define, for $\alpha > 0$,

$$(\forall i \in N) \quad \overline{e}_i = \begin{cases} \alpha x_i^\top g_k, & \text{if } i \in M; \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

and $\overline{y} = Xf + \overline{e}$. Then,

$$\|\overline{y} - Xf\|_1 = \alpha \sum_{i \in M} |x_i^\top g_k|$$
$$\|\overline{y} - X(f + \alpha g_k)\|_1 = \alpha \sum_{i \in N \setminus M} |x_i^\top g_k|.$$

Hence, it follows from (8) that $\|\overline{y} - X(f + \alpha g_k)\|_1 \leq \|\overline{y} - Xf\|_1$, then $f + \alpha g_k$ is a minimizer. □

The proof of Theorem 2 shows that if $k > m(X)$, then, for any $\alpha > 0$, we can find a vector e such that $\|e\|_0 = k$ and the ℓ_1 estimator f_1 on the data $y = Xf + \alpha e$ satisfies $\|f_1 - f\|_2 = \alpha$. Combined with Theorem 1 this shows that the RBP of the ℓ_1 estimator equals $m(X) + 1$, recovering results of Giloni and Padberg (2004), Mizera and Müller (1999).

Also, we can see from (9) that the existence of an unexpected sub-population following a linear model with a different slope is the most troublesome scenario for ℓ_1 estimation.

4 Error bounds for Huber M-estimator face to sparse outliers and noise

In this section, we study the performance of Huber's M-estimator at model (5). The derivation of error bounds for Huber's estimator relies on an alternative formulation of the minimization problem, the ℓ_1 error estimate and duality theory.

Let $\sigma > 0$, let $y \in \mathbb{R}^n$, and let X be a $n \times p$ real matrix with full rank. Consider the problem

$$\begin{array}{c} \underset{(g,b,s) \in \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^n}{\text{minimize}} \quad \sigma \|s\|_1 + \frac{1}{2}\|b\|_2^2 \\ s.t \quad y = Xg + b + s, \end{array} \tag{10}$$

where g, b and s are optimization variables estimating f, the dense error term z and the sparse errors e, respectively, and σ is an estimate of the magnitude of the noise. Isolating b from the linear constraint brings up the following equivalent problem:

$$\underset{(g,b)\in\mathbb{R}^p\times\mathbb{R}^n}{\text{minimize}} \quad \psi(g,b) := \sigma\|y - Xg - b\|_1 + \frac{1}{2}\|b\|_2^2. \tag{11}$$

Theorem 3 *Let $y = Xf+z+e$, let $M = \operatorname{supp}(e)$, and suppose that $|M| = k \le m(X)$. Then any solution $(\hat{g}, \hat{b})$ to (11) satisfies*

$$\|X(\hat{g} - f_n)\|_1 \le \frac{1}{2c_k - 1}\left(\|\bar{b} - \hat{b}\|_{1,N\setminus M} + \frac{\|\bar{b} - \hat{b}\|_{2,N\setminus M}^2}{\|\bar{b} - \hat{b}\|_{\infty,N\setminus M}}\right), \tag{12}$$

where $f_n = f + \bar{g}$ is the LSE on $y_n = Xf + z$.

Theorem 3 provides an error bound for the Huber estimator g_H since in any solution pair $(\hat{g}, \hat{b})$ to (11) the first component $\hat{g}$ coincides with g_H. To see this notice that the minimization in (11) can be written as $\min_{g\in\mathbb{R}^p}\rho(y - Xg)$ for $\rho\colon r \mapsto \rho(r) = \inf_{b\in\mathbb{R}^n}\sigma\|r - b\|_1 + \frac{1}{2}\|b\|_2^2$; the function ρ above equals Huber's criterion

$$\rho_H(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \le \sigma \\ \sigma|r| - \frac{1}{2}\sigma^2 & \text{if } |r| > \sigma \end{cases} \tag{13}$$

for any $r \in \mathbb{R}^n$ (Michelot and Bougeard 1994).

The alternative formulation (10) of Huber's estimation problem based on the error model (5) provides an interpretation of the estimator on finite samples. The additional term b in (11), which makes the difference with respect to the ℓ_1 estimator, improves its response to noisy observations.

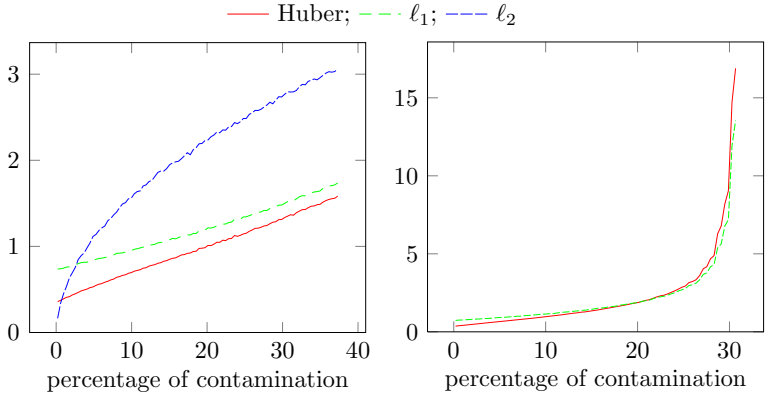

Fig. 1 Relative error $\|\hat{f} - f_n\|/\|f_n\|$ with gaussian noise and different percentage of outliers. On the *left*, the contamination is drawn from a Laplace $(0, 5)$ distribution and on the *right* contamination consists of large grouped outliers

In Fig. 1 two bias curves illustrate Theorems 1 and 3 in different situations. At the left outliers are drawn from a heavy-tail distribution, the denoising effect of $\hat{b}$ is clearly perceived. At the right outliers are very large and predominate, there is not significative difference between Huber's and ℓ_1 estimators as both breakdown at the same point as expected. The methodology used in the simulation is standard and can be found in Appendix B.

5 Leverage in real datasets

The importance of quantifying the concept of leverage becomes apparent by analyzing some examples. For our analysis, the quantities

$$\gamma^i = \max_{\substack{g\in\mathbb{R}^p \\ \|g\|_2=1}} \frac{|x_i^\top g|}{\sum_{i\in N} |x_i^\top g|}, \qquad i = 1, \ldots, n; \tag{14}$$

measuring the leverage of a particular observation will be particularly useful. They correspond to taking $k = |M| = 1$ in the notation of Theorems 1 and 3, in particular $c_1 = 1 - \max_{1\le i\le n} \gamma^i$.

The *aircraft* model (Rousseeuw and Leroy 1987, pp. 154) intends to explain the cost of 23 single-engine aircrafts in terms of their aspect ratio, lift-to-drag ratio, weight and thrust. The ℓ_1 estimator has a breakdown point of $2/23 = 8.69\,\%$ on these data, which amounts to have $m(X) = 1$. The data contain one outlier in the y-direction (observation 22) and a *good leverage point* at observation 14.

Recall that a high leverage point is not always influential. It has the potential to influence the fit, but it does not necessarily do so. Indeed, the leverage depends only on the x_i, while the fit depends on the pairs (x_i, y_i). For this reason, it is convenient to visualize the leverage constants γ^i jointly with the responses y_i. After all, whether a leverage point will exert its influence depends on its associated response. In Fig. 2, at the left, we show a scatterplot of the pairs (γ^i, y_i) for the aircraft dataset. There are

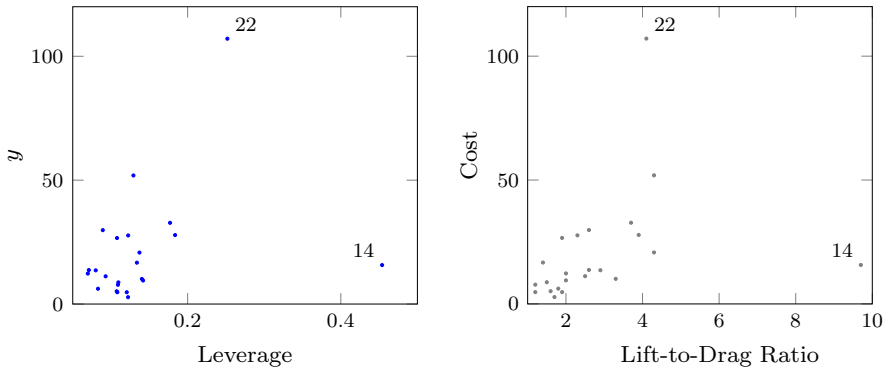

Fig. 2 Aircraft dataset. At the *left*, a scatterplot of the pairs (leverage, response). At the *right*, the response is plotted against the lift-to-drag ratio

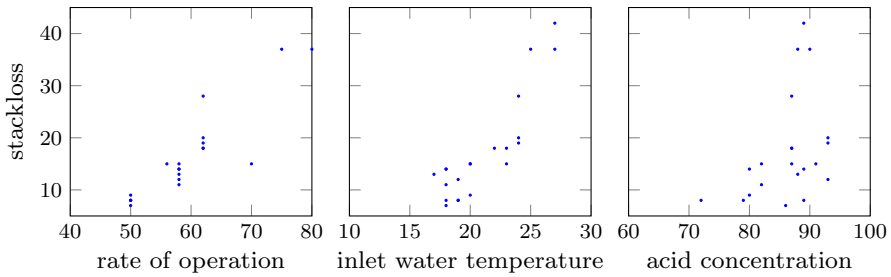

Fig. 3 Plot of stackloss versus each of the explanatory variables

two points (14 and 22) that escape from the main cloud. However, they seem to be of different nature; point 14 has the greatest leverage, but its response is definitively normal. On the other hand, point 14 has a leverage barely above that of the main group, but a clearly outlying response, thus a high influence. This is a benign combination of leverage points and outliers. In fact the structure of the data can be explained by one of the explanatory variables, the lift-to-drag ratio, as shown at the right in Fig. 2.

Another interesting fact in these data is that observation 14 has a leverage of 0.4539, which is quite close to the critical value of 0.5 above which the BDP falls down to 0 %. If we perturb row 14 of X by adding the vector $(0.6876, -0.1692, 0.5337, 0, 0)$, which changes its norm by 0.0000012 %, the leverage of observation 14 increases to 0.4982. The estimator obtained from the perturbed data is $(-0.9656, -2.6731, 1.4467, 0.0023, -0.0010)$, which is quite apart from the vector $(1.9511, -3.0466, 1.4736, 0.0022, -0.0010)$ obtained from the original data. The same would happen, to a minor extent, even if observation 22 was not contaminated. This phenomenon, called *instability* by Ellis (1998), is just another aspect of leverage, such as breakdown, although much less obvious.

A look at the leverage is therefore a must when analyzing data using ℓ_1 regression. Another advantage of the leverage constants is that they synthesize multi-dimensional data in such a way that it can be plotted, as in Fig. 2. The leverage plot can also spot non-trivial aspects of the data, as the following example shows.

The *stackloss* dataset (Rousseeuw and Leroy 1987, pp. 76) is well known in the robustness literature. It describes an oxidation process; the stackloss is to be explained by the rate of operation, the inlet water temperature and the acid concentration. Observations 1, 2, 3, 4 and 21 are outliers. However, this is not apparent by looking at the plots of each explanatory variable, shown in Fig. 3, as it was the case for the aircraft data.

In the leverage plot of Fig. 4 (left) observations 1, 2, 3 and 4 stand out. Observation 21 blends into the bulk of the data, and observation 17 seems to be a 'good' leverage point. For this datum $m(X) = 3$, thus the BDP of ℓ_1 regression is $4/21 = 19.04\,\%$. Computing the leverage constant $c_4 = 0.4144$ gives the minima $M = \{1, 2, 3, 21\}$ and $g_4 = (0.9953, -0.0855, 0.0165, 0.0412)$. At the right in Fig. 4 we show a scatterplot of the projections of the data onto the "outlying direction" g_4, along with y_i. The structure of the data is clearly depicted. Note in particular that the roles of observations 17 and 21 are clarified. Point 17 is absolved, since it forms undeniably part of the main group of points, and point 21 is exposed. Over the x-axis we plot again the projections

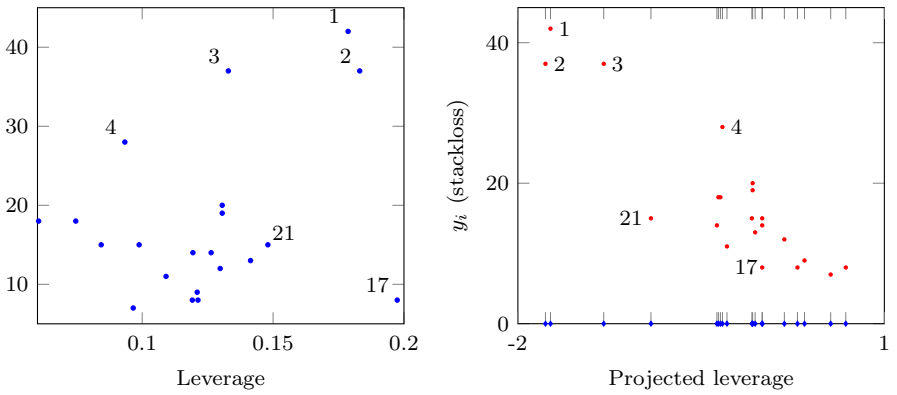

Fig. 4 Scatterplots of the pairs (leverage, response) for each point of the stackloss dataset at the *left*, and of the pairs $(x_i^\top g_4, y_i)$ at the *right*

$x_i^\top g_4$ to notice the separation of the points in M from the rest. This plot also confirms that point 4 is a vertical outlier.

Acknowledgments The author is grateful to Luis Briceño-Arias for his thoughtful comments on an earlier version of this paper, and to Jean-Baptiste Hiriart-Urruty for bringing to my attention the work of Candes and Tao (2005). The author also thanks the Editor and the anonymous referees for their valuable suggestions which have improved the presentation of the paper. This work was supported by Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) through FONDECYT program, Grant 3120166, and FONDAP-BASAL program.

Appendix A: Proofs and Lemmas

Let $\phi\colon \mathbb{R}^n \to \left]-\infty,+\infty\right]$ be a lower semicontinuous convex function which is proper in the sense that $\operatorname{dom}\phi = \left\{x \in \mathbb{R}^n \;\middle|\; \phi(x) < +\infty\right\} \neq \varnothing$. The subdifferential operator of ϕ is

$$\partial\phi\colon \mathbb{R}^n \to 2^{\mathbb{R}^n} : x \mapsto \left\{u \in \mathbb{R}^n \;\middle|\; (\forall y \in \mathbb{R}^n)\; u^\top(y-x) + \phi(x) \leq \phi(y)\right\}$$

and we have (Hiriart-Urruty and Lemaréchal 1993, Theorem 2.2.1)

$$x \in \operatorname*{Argmin}_{y\in\mathbb{R}^n} \phi(x) \quad \Leftrightarrow \quad 0 \in \partial\phi(x). \tag{15}$$

The proximal mapping associated with ϕ is defined by

$$\operatorname{prox}_\phi\colon \mathbb{R}^n \to \mathbb{R}^n : x \mapsto \operatorname{argmin}_{u\in\mathbb{R}^n}\left(\phi(u) + \frac{1}{2}\|u-x\|_2^2\right). \tag{16}$$

From (15) we obtain, for every $x \in \mathbb{R}^n$ and $p \in \mathbb{R}^n$,

$$p = \operatorname{prox}_\phi x \quad \Leftrightarrow \quad x - p \in \partial\phi(p),$$

and since $\phi + \|\cdot - x\|^2/2$ is strongly convex, $\operatorname{prox}_\phi(x)$ exists and is unique for all $x \in \mathbb{R}^n$.

Lemma 2 *Let $\gamma \in \left]0, +\infty\right[$ and $\phi\colon \mathbb{R}^n \to \mathbb{R}\colon x \mapsto \gamma\|x\|_1 = \gamma \cdot \sum_{i=1}^n |x_i|$. Then the following holds.*

(i) *For every $x \in \mathbb{R}^n$, $\partial\phi(x) = \times_{i=1}^n \partial\gamma|\cdot|(x_i)$, where*

$$(\forall \xi \in \mathbb{R}) \quad \partial\gamma|\cdot|(\xi) = \begin{cases} \gamma, & \text{if } \xi > 0; \\ [-\gamma, \gamma], & \text{if } \xi = 0; \\ -\gamma, & \text{if } \xi < 0. \end{cases}$$

(ii) *For every $x \in \mathbb{R}^n$, $\operatorname{prox}_{\gamma\phi}\, x = (\operatorname{prox}_{\gamma|\cdot|}(x_i))_{1\le i\le n}$, where*

$$(\forall \xi \in \mathbb{R}) \quad \operatorname{prox}_{\gamma|\cdot|}(\xi) = \begin{cases} \xi - \gamma & \text{if } \xi > \gamma; \\ 0, & \text{if } \xi \in [-\gamma, \gamma]; \\ \xi + \gamma, & \text{if } \xi < -\gamma. \end{cases}$$

Proof The results follow from Combettes and Wajs (2005, Lemma 2.1, Lemma 2.9, and Example 2.16). □

Proof of Proposition 1

Proof It is clear that $c_0 = 1$ and that $c_n = 0$. Let $k \in \{1, \dots, n\}$, let $g \in \mathbb{R}^p \backslash \{0\}$, and let M with $|M| = k - 1$ such that

$$c_{k-1} = \frac{\sum\limits_{i \in N\backslash M} |x_i^\top g|}{\sum\limits_{i \in N} |x_i^\top g|}.$$

Now let $i_0 \in N\backslash M$ and $\widetilde{M} = M \cup \{i_0\}$. We have $|\widetilde{M}| = k$ and from (2), we obtain

$$c_{k-1} = \frac{\sum\limits_{i \in N\backslash M} |x_i^\top g|}{\sum\limits_{i \in N} |x_i^\top g|} = \frac{\sum\limits_{i \in N\backslash \widetilde{M}} |x_i^\top g| + |x_{i_0} g|}{\sum\limits_{i \in N} |x_i^\top g|} \ge \frac{\sum\limits_{i \in N\backslash \widetilde{M}} |x_i^\top g|}{\sum\limits_{i \in N} |x_i^\top g|} \ge c_k,$$

which yields the result. □

Lemma 3 *The following holds.*

(i) *$(\hat{g}, \hat{b})$ is a solution to (11) if and only if $X^\top \hat{b} = 0$ and*

$$(\forall i \in \{1, \dots, n\}) \quad \hat{b}_i = \begin{cases} \sigma, & \text{if } y_i - x_i^\top \hat{g} > \sigma; \\ y_i - x_i^\top \hat{g}, & \text{if } y_i - x_i^\top \hat{g} \in [-\sigma, \sigma]; \\ -\sigma, & \text{if } y_i - x_i^\top \hat{g} < -\sigma. \end{cases} \tag{17}$$

In particular, $\|\hat{b}\|_\infty \le \sigma$.

(ii) *A dual of* (11) *is*

$$\gamma := \max_{u \in \sigma P^*} u^\top y - \frac{1}{2}\|u\|_2^2, \tag{18}$$

where $P^* = \left\{u \in \ker X^\top \;\middle|\; \|u\|_\infty \leq 1\right\}$ *and*

$$\min_{(g,b)\in\mathbb{R}^p\times\mathbb{R}^n} \psi(g,b) = \gamma.$$

Proof Note that $\psi(g,b)$ can be equivalently written as

$$\psi(g,b) = \sigma\|y - \begin{bmatrix} X & I_n \end{bmatrix}\begin{pmatrix} g \\ b \end{pmatrix}\|_1 + \frac{1}{2}\|\begin{bmatrix} 0_p & I_n \end{bmatrix}\begin{pmatrix} g \\ b \end{pmatrix}\|_2^2 \tag{19}$$

where I_n denotes the identity matrix of size $n \times n$ and 0_p the zero matrix of size $p \times p$.

(i): Since the function $\psi(g,b)$ is convex, a necessary and sufficient condition for a solution $(\hat{g}, \hat{b})$ to Problem (11) is

$$0 \in \partial\psi(\hat{g}, \hat{b}). \tag{20}$$

Hence, using (Hiriart-Urruty and Lemaréchal 1993, Theorem 4.2.1) in (19), (20) is equivalent to

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in -\begin{bmatrix} X^T \\ I_n \end{bmatrix} \partial\sigma\|\cdot\|_1(y - X\hat{g} - \hat{b}) + \begin{pmatrix} 0 \\ \hat{b} \end{pmatrix}.$$

Therefore, there exists $u \in \partial\sigma\|\cdot\|_1(y - X\hat{g} - \hat{b})$ such that $X^\top u = 0$ and $b = u$, or equivalently,

$$\begin{cases} \hat{b} \in \partial\sigma\|\cdot\|_1(y - X\hat{g} - \hat{b}), \\ X^\top\hat{b} = 0. \end{cases}$$

Hence $y - X\hat{g} - \hat{b} = \text{prox}_{\sigma\|\cdot\|_1}(y - X\hat{g})$, and the result follows from Lemma 2(ii).

(ii): Problem (11) is equivalent to (10), and applying Lagrangian duality, the dual is

$$\max_{u\in\mathbb{R}^p} \min_{(g,b,s)\in\mathbb{R}^d\times\mathbb{R}^n\times\mathbb{R}^n} \sigma\|s\|_1 + \frac{1}{2}\|b\|_2^2 + u^\top(y - Xg - b - s),$$

or equivalently,

$$\max_{u\in\mathbb{R}^p}\left(u^\top y + \left(\min_{b\in\mathbb{R}^n}\frac{1}{2}\|b\|_2^2 - u^\top b\right) + \left(\min_{s\in\mathbb{R}^n}\sigma\|s\|_1 - u^\top s\right) - \max_{g\in\mathbb{R}^p} g^\top(X^\top u)\right). \tag{21}$$

The optimality conditions associated to the convex optimization problem

$$\min_{b\in\mathbb{R}^n}\frac{1}{2}\|b\|_2^2 - u^\top b$$

yield $b = u$, hence $\min_{b\in\mathbb{R}^n} \frac{1}{2}\|b\|_2^2 - u^\top b = -\frac{1}{2}\|u\|_2^2$. The second minimization problem can be written as

$$\min_{s\in\mathbb{R}^n} \sigma\|s\|_1 - u^\top s = \sum_{i=1}^n \min_{s_i\in\mathbb{R}} \sigma|s_i| - u_i s_i = \begin{cases} -\infty, & \text{if } \|u\|_\infty > \sigma; \\ 0, & \text{if } \|u\|_\infty \le \sigma. \end{cases}$$

Finally, we have

$$\max_{g\in\mathbb{R}^p} g^\top (X^\top u) = \begin{cases} +\infty, & \text{if } u \notin \ker X^\top; \\ 0, & \text{if } u \in \ker X^\top. \end{cases}$$

Altogether, it follows from (21) that the dual to (11) is given by (18) and the absence of duality gap follows from the Slater qualification condition and the existence of multipliers (Hiriart-Urruty and Lemaréchal 1993, Sect. 4). □

Proof of Theorem 3

Proof From Lemma 1(i) and (11) we deduce

$$\psi(\hat{g},\hat{b}) - \psi(f_n,\hat{b}) \ge \sigma(2c_k - 1)\|X(\hat{g} - f_n)\|_1 - 2\sigma\|y - Xf_n - \hat{b}\|_{1,N\setminus M}.$$

Hence, it follows from $f_n = f + \bar{g}$ that, for every $i \in \{1,\dots,n\}$, $y_i - x_i^\top f_n = e_i + \bar{b}_i$ and thus, $\psi(f_n,\hat{b}) = \sigma\|e + \bar{b} - \hat{b}\|_1 + \|\hat{b}\|_2^2/2$. Therefore, since $e_i = 0$ for any $i \in N\setminus M$,

$$\sigma(2c_k - 1)\|X(\hat{g} - f_n)\|_1 \le 2\sigma\|\bar{b} - \hat{b}\|_{1,N\setminus M} - \sigma\|e + \bar{b} - \hat{b}\|_1 + \psi(\hat{g},\hat{b}) - \frac{1}{2}\|\hat{b}\|_2^2. \quad (22)$$

From Lemma 3(ii), the dual problem to (11) is

$$\max_{u\in\sigma P^*} u^\top(e + \bar{b}) - \frac{1}{2}\|u\|_2^2$$

and $\psi(\hat{g},\hat{b}) = \max_{u\in\sigma P^*} u^\top(e + \bar{b}) - \frac{1}{2}\|u\|_2^2$. Therefore,

$$\begin{aligned} \psi(\hat{g},\hat{b}) - \frac{1}{2}\|\hat{b}\|_2^2 &= \max_{u\in\sigma P^*} u^\top(e + \bar{b}) - \frac{1}{2}\|u\|_2^2 - \frac{1}{2}\|\hat{b}\|_2^2 \\ &= \max_{u\in\sigma P^*} u^\top(e + \bar{b} - \hat{b}) - \frac{1}{2}\|u - \hat{b}\|_2^2 \\ &\le \max_{u\in\sigma P^*} u^\top(e + \bar{b} - \hat{b}). \end{aligned}$$

Hence, it follows from Lemma 4 that

$$\psi(\hat{g},\hat{b}) - \frac{1}{2}\|\hat{b}\|_2^2 \le \sigma\|e + \bar{b} - \hat{b}\|_{1,M} + \frac{\sigma}{\|\bar{b} - \hat{b}\|_{\infty,N\setminus M}}\|\bar{b} - \hat{b}\|_{2,N\setminus M}^2,$$

which, combined with (22), yields

$$\begin{aligned}(2c_k-1)\|X(\hat{g}-f_n)\|_1 &\leq 2\|\bar{b}-\hat{b}\|_{1,N\setminus M}-\|e+\bar{b}-\hat{b}\|_1+\|e+\bar{b}-\hat{b}\|_{1,M}\\ &\quad+\frac{1}{\|\bar{b}-\hat{b}\|_{\infty,N\setminus M}}\|\bar{b}-\hat{b}\|_{2,N\setminus M}^2\\ &=\|\bar{b}-\hat{b}\|_{1,N\setminus M}+\frac{1}{\|\bar{b}-\hat{b}\|_{\infty,N\setminus M}}\|\bar{b}-\hat{b}\|_{2,N\setminus M}^2\end{aligned}$$

as claimed. □

Lemma 4 *Let $b\in\mathbb{R}^n$, $e\in\mathbb{R}^n$ and let $M=\operatorname{supp}(e)$. Suppose that $|M|\leq m(X)$ and $\max_{i\in N\setminus M}|b_i|>0$. Let us define $P^*=\left\{d\in\ker X^\top \;\middle|\; \|d\|_\infty\leq 1\right\}$. Then, for every $\sigma>0$,*

$$\max_{d\in\sigma P^*} d^\top(e+b)\leq\sigma\|e+b\|_{1,M}+\frac{\sigma}{\|b\|_{\infty,N\setminus M}}\|b\|_{2,N\setminus M}^2.$$

Proof Let

$$\tilde{b}_i=\begin{cases}0, & \text{if } i\in M;\\ b_i, & \text{otherwise,}\end{cases}\quad\text{and}\quad \tilde{e}_i=\begin{cases}b_i+e_i, & \text{if } i\in M;\\ 0, & \text{otherwise.}\end{cases}\tag{23}$$

Then $\operatorname{supp}(\tilde{e})=M$, $b+e=\tilde{b}+\tilde{e}$, $\|b+e\|_1=\|\tilde{b}\|_1+\|\tilde{e}\|_1$, and

$$\max_{d\in\sigma P^*} d^\top(e+b)=\max_{d\in\sigma P^*} d^\top(\tilde{e}+\tilde{b})\leq\max_{d\in\sigma P^*} d^\top\tilde{e}+\max_{d\in\sigma P^*} d^\top\tilde{b}.\tag{24}$$

On one hand, it follows from Lemma 1(i) with $y=\tilde{e}$, $g^*=0$, and $b^*=0$ that, for every $g\in\mathbb{R}^p$, $\|\tilde{e}\|_1\leq\|\tilde{e}-Xg\|_1$, hence $0\in\operatorname{argmin}_{g\in\mathbb{R}^p}\|\tilde{e}-Xg\|_1$ and from the first-order optimality condition $0\in X^\top\partial\|\cdot\|_1(\tilde{e})$, or equivalently, $(\exists u\in P^*)\;\; u^\top\tilde{e}=\|\tilde{e}\|_1$. Since, for every $u\in P^*$, $u^\top e\leq\|e\|_1$ we hence deduce that $\max_{u\in P^*}u^\top\tilde{e}=\|\tilde{e}\|_1$. Therefore, by considering the change of variables $u=d/\sigma$, we obtain

$$\max_{d\in\sigma P^*} d^\top\tilde{e}=\sigma\cdot\max_{u\in P^*} u^\top\tilde{e}=\sigma\|\tilde{e}\|_1.\tag{25}$$

On the other hand,

$$\max_{d\in\sigma P^*} d^\top\tilde{b}\leq\max_{\|d\|_\infty\leq\sigma} d^\top\tilde{b}=\frac{\sigma}{\|\tilde{b}\|_\infty}\tilde{b}^\top\tilde{b}=\frac{\sigma}{\|\tilde{b}\|_\infty}\|\tilde{b}\|_2^2.\tag{26}$$

Therefore, by replacing (25) and (26) in (24), the result follows from (23). □

Appendix B: Additional information on curves in Sect. 4, Fig. 1

The experimental setup is the following. The matrix X is generated randomly with independent entries drawn from a standard normal distribution. Its size is $n\times p=512\times 128$. The vector of data is generated according to

$$y = Xf + z + e,$$

with $f = 0$ and z standard normal, for different types and levels of contamination.

We estimate f by three different methods: LSE, ℓ_1, and Huber's with $\sigma = \sqrt{\chi_1^2(.95)}$. The size of the support of e ranges from 1 to $(n - p - 1)/2$, which means that the maximum fraction of contamination is close to 40 %. We consider three types of sparse contamination. In the first and second types, each non-zero component of e is drawn i.i.d. from a Normal (light-tailed) and Laplace (heavy-tailed) distribution with mean 0 and standard deviation 5, respectively. The last type of sparse error is considered to be very large and adversarial, inspired from the proof of Theorem 2. For generating the adversarial contamination we first create the vector $\tilde{e} = X\mathbb{1}_p$, where $\mathbb{1}_p$ is the vector of ones of size $p \times 1$. Then, the sparse errors are obtained by selecting some components of $\tilde{e}$ randomly and by multiplying them by 50.

For each type of contamination, for every $k \in \{1, \ldots, (n - p - 1)/2\}$, we repeat 1000 times the following:

1. Choose randomly a subset M of N of size k.
2. Construct the sparse vector e by filling the entries indexed by M with the corresponding type of large errors.
3. Generate z with independent $N(0, 1)$ entries.
4. Set $y = z + e$ and estimate $f = 0$ by LSE, ℓ_1, and Huber's methods.

For each percentage of outliers, the bias is quantified by the mean of the quotients $\|\hat{f} - f_n\|_2 / \|f_n\|_2$, where $\hat{f}$ is the estimation of f obtained by each of the three methods and $f_n = (X^\top X)^{-1} X^\top z$.

References

Candes E, Randall P (2008) Highly robust error correction by convex programming. IEEE Trans Inform Theory 54(7):2829–2840

Candes E, Tao T (2005) Decoding by linear programming. IEEE Trans Inform Theory 51(12):4203–4215

Combettes PL, Wajs VR (2005) Signal recovery by proximal forward–backward splitting. Multiscale Model Simul 4(4):1168–1200

Ellis SP (1998) Instability of least squares, least absolute deviation and least median of squares linear regression, with a comment by stephen portnoy and ivan mizera and a rejoinder by the author. Stat Sci 13(4):337–350

Ellis SP, Morgenthaler S (1992) Leverage and breakdown in L_1 regression. J Am Stat Assoc 87(417):143–148

Giloni A, Padberg M (2002) Alternative methods of linear regression. Math Comput Model 35:361–374

Giloni A, Padberg M (2004) The finite sample breakdown point of ℓ_1-regression. SIAM J Optim 14:1028–1042

Giloni A, Sengupta B, Simonoff JS (2006a) A mathematical programming approach for improving the robustness of least sum of absolute deviations regression. Naval Res Logist 53(4):261–271

Giloni A, Simonoff JS, Sengupta B (2006b) Robust weighted LAD regression. Comput Stat Data Anal 50(11):3124–3140

He X, Jurečková J, Koenker R, Portnoy S (1990) Tail behavior of regression estimators and their breakdown points. Econometrica 58(5):1195–1214

Hiriart-Urruty JB, Lemaréchal C (1993) Convex analysis and minimization algorithms I: fundamentals. In: Grundlehren der mathematischen Wissenschaften, vol 305. Springer, Berlin

Huber PJ (1973) Robust regression: asymptotics, conjectures and Monte Carlo. Ann Stat 1:799–821

Michelot C, Bougeard ML (1994) Duality results and proximal solutions of the Huber M-estimator problem. Appl Math Optim 30(2):203–221

Mizera I, Müller CH (1999) Breakdown points and variation exponents of robust M-estimators in linear models. Ann Statist 27(4):1164–1177

Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, New York

Zhang Y (2013) Theory of compressive sensing via ℓ_1-minimization: a non-rip analysis and extensions. J Oper Res Soc China 1(1):79–105