Unpublished manuscript No.
(will be inserted by the editor)

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

Salvador Flores

October 22, 2014

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 usign 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. A Montecarlo simulation illustrates the ideas previously developed.

Keywords ℓ_1 norm minimization, Huber M-estimator, leverage constants, sparse outliers, breakdown point, robustness

Mathematics Subject Classification (2000) 62J05, 62F35, 90C31

1 Introduction

In this paper we address the problem of estimating a vector $f \in \mathbb{R}^p$ from a set of n measurements ($n > p$),

$$y = Xf + \delta, \tag{1}$$

where $y \in \mathbb{R}^n$ is the vector of measurements or observations, X is an $n \times p$ real matrix, whose rows are realizations of the explicative variables, and $\delta \in \mathbb{R}^n$ is an error term. For simplicity we suppose that the matrix X has full rank.

S. Flores
Centro de Modelamiento Matemático (CNRS UMI 2807) – Universidad de Chile
Av. Blanco Encalada 2120, Santiago, Chile.
Tel.: +56-22-97 80611
Fax: +56-22-68 89705
E-mail: sflores@dim.uchile.cl

This problem has special interest in many fields, including linear regression and signal recovery.

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 $x_1, ..., x_n \in \mathbb{R}^p$. We assume that the random variables $x_1, ..., x_n$, and y are related through a linear model, which implies the existence of a vector $f \in \mathbb{R}^p$ such that

$$(\forall i \in \{1, ..., n\}) \quad y_i = x_i^\top f + \delta_i, \tag{2}$$

where $(\delta_i)_{1 \leq i \leq n}$ are i.i.d. random variables with zero mean and finite variance. The objective in linear regression is to estimate f. The Least Squares Estimator (LSE) of f is defined as the solution to

$$\begin{array}{cc} \min\limits_{g \in \mathbb{R}^p, r \in \mathbb{R}^n} & \sum\limits_{i=1}^{n} r_i^2 \\ s.t & r = y - Xg, \end{array} \tag{3}$$

where r denotes the vector of residuals. Under the usual assumption that the errors δ_i are Gaussian, the LSE is the best linear unbiased estimator of f (Shao 2003). However, the LSE is very sensitive to deviations from normality, even moderate ones. As the hypothesis of normality is often violated in practice, there is a great interest in developing statistical procedures that are robust face to different error distributions.

In robust regression, model (2) is enlarged by considering that errors come from *contaminated distributions* (Tukey 1960) in a neighbourhood of an assumed distribution F

$$\mathcal{F}_\varepsilon = \{(1 - \varepsilon)F + \varepsilon G : G \in \mathcal{G}\}, \tag{4}$$

where F is usually the normal distribution, $\mathcal{G}$ is a family of contaminating distribution, supposed to model outliers, and $0 < \varepsilon < 1$ represents the fraction of contamination. The ability of an estimation method to give reasonable results under model (4) 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 in order 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. The M-estimators (Huber 1973, 1981) aim to perform robust and computationally efficient estimation. They are a generalization of (3), defined as a solution to

$$\begin{array}{cc} \min\limits_{g \in \mathbb{R}^p, r \in \mathbb{R}^n} & \sum\limits_{i=1}^{n} \rho(r_i/\tau) \\ s.t & r = y - Xg, \end{array} \tag{5}$$

for some differentiable pair function $\rho : \mathbb{R} \to \mathbb{R}_+$ which is non-decreasing in $\mathbb{R}_+$ and a scale of residuals $\tau > 0$. The scale τ is supposed to be known, and can be

taken as one without loss of generality. The first order optimality conditions of problem (5) yields

$$\sum_{i=1}^{n} w_i x_i = 0, \tag{6}$$

where $w_i := \rho'(r_i)$ acts as a weight of the influence of each observation on the fit. Hence, if the function ρ is additionally convex the observations with large residuals have a higher weight, which implies that the M-estimator is sensitive to outliers in this case. In the opposite case, if the function ρ has non-increasing derivative, we face a nonconvex optimization problem, which are beyond the capabilities of the state-of-the-art of global optimization methods, even for problems of modest size.

The border case is the ℓ_1 estimator, also called Least Absolute Deviations, which is defined as a solution to

$$\begin{array}{cc} \min\limits_{g\in\mathbb{R}^p, r\in\mathbb{R}^n} & \sum\limits_{i=1}^{n} |r_i| \\ s.t & r = y - Xg. \end{array} \tag{7}$$

It does not fit in the framework of (5) since the function to minimize is not differentiable. Nonetheless, it satisfies system (6) for w_i equal to one if $r_i > 0$, equal to minus one if $r_i < 0$, and between -1 and 1 for null residuals. Therefore, the ℓ_1 estimator gives a bounded weight to each observation while keeping the estimation problem convex. The increased robustness of the ℓ_1 estimator with respect to the LSE comes at the cost of a lower statistical efficiency or, equivalently, an increased asymptotic variance. Huber (1981) proposed to choose the function ρ so as to minimize the maximum asymptotic variance over the neighbourhood $\mathcal{F}_\varepsilon$ in (4), yielding to

$$\rho_\sigma(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \le \sigma \\ \sigma|r| - \frac{1}{2}\sigma^2 & \text{if } |r| > \sigma \end{cases} \tag{8}$$

The value of σ can be chosen to obtain a given asymptotic variance at the normal distribution. The Huber estimate gives a bounded weight equal to $\pm\sigma$ to large residuals and a least squares weight to smaller ones, combining in this way statistical efficiency and robustness. Besides, the functions ρ_σ are convex and differentiable. The Huber M-estimator has been successfully applied to numerous problems such as earthquake hypocenter location (Anderson 1982), GPS position estimation (Chang and Guo 2005), robust face recognition (Naseem et al 2012) and non-invasive measurement of the cerebral blood flow by MRI (Maumet et al 2014), just to cite a few. Algorithms for computing this estimator can be found in Huber (1981); Madsen and Nielsen (1990); Michelot and Bougeard (1994); Chen and Pinar (1998).

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). Their work lies in the fixed design framework and they also suppose that contamination is restricted to the dependent variable y. Moreover, 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. This hypothesis, that would horrify any statistician, permits to solve the problem via the successful theory of sparse solutions to linear systems. It provides 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 in order 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. Think for instance in the earthquake hypocenter location problem to realize that the asymptotic study of the hypocenter location estimation with infinite isotropically distributed sensing stations is of little practical interest. Likewise, the notion of breakdown point gives information on the behaviour of an estimator when data is replaced by divergent observations; nonetheless, it is preferable 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 aim at filling this gap by providing non-asymptotic error bounds in finite samples for two of the most widespread convex robust estimators.

1.1 Outline of the paper

In Section 2 we introduce the leverage constants and derive the fundamental ℓ_1 error estimate, which is useful to examine any estimation technique related to ℓ_1 norm minimization. Then, in Section 3, we consider the following model for the errors in (1):

$$\delta = z + e, \tag{9}$$

where z is a dense, presumably small, vector of noise and e is an arbitrary sparse vector. We perform a detailed quantitative error analysis of the ℓ_1 estimator, obtaining a sharp error bound taking advantage of the ℓ_1 error estimate of Section 2 and the dual problem. In particular, we show that the RBP of the ℓ_1 estimator characterizes the critical sparsity level of outliers in order to obtain exact recovery by ℓ_1 minimization in the noiseless case. In Section 4 we derive from (9) an alternative formulation of the Huber M-estimation problem permiting to extend the analysis done for the ℓ_1 estimator. In Section 5 we show some bias curves obtained by a Montecarlo simulation confirming the behaviour predicted by our analytical results. We conclude the article with a summary and a discussion, presented in Section 6.

1.2 Notation and preliminaries

We shall use the notation $N = \{1, ..., n\}$ 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 nonzero 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 of N and $p \in [1, +\infty[$, we define

$$\|\cdot\|_{p,M} \colon x \mapsto \left(\sum_{i \in M} |x_i|^p \right)^{1/p}$$

and

$$\|\cdot\|_{\infty,M} \colon x \mapsto \max_{i \in M} |x_i|.$$

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

Let $\phi \colon \mathbb{R}^n \to]-\infty, +\infty]$ 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

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

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

$$x \in \operatorname*{Argmin}_{y \in \mathbb{R}^n} \phi(x) \quad \Leftrightarrow \quad 0 \in \partial\phi(x). \tag{10}$$

The proximal mapping associated with ϕ is defined by

$$\operatorname{prox}_\phi \colon \mathbb{R}^n \to \mathbb{R}^n \colon x \mapsto \operatorname*{argmin}_{u \in \mathbb{R}^n} \left(\phi(u) + \frac{1}{2} \|u - x\|_2^2 \right). \tag{11}$$

From (10) 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 $x \in \mathbb{R}^n$.

The following lemma will be useful throughout the paper.

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

(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). □

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 influencial 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

$$c_k(X) = \min_{\substack{M \subset N \\ |M| = k}} \min_{\substack{g \in \mathbb{R}^p \\ g \neq 0}} \frac{\sum_{i \in N \setminus M} |x_i^\top g|}{\sum_{i \in N} |x_i^\top g|} = \min_{\substack{M \subset N \\ |M| = k}} \min_{\substack{g \in \mathbb{R}^p \\ \|g\|_2 = 1}} \frac{\sum_{i \in N \setminus M} |x_i^\top g|}{\sum_{i \in N} |x_i^\top g|} \tag{12}$$

and

$$m(X) = \max\Big\{k \in N \;\Big|\; c_k(X) > \frac{1}{2}\Big\}. \tag{13}$$

Note that the two minima in (12) are achieved since the feasible set in both cases is compact and the objective function is continuous.

Lemma 2 *We have $c_0 = 1$, $c_n = 0$ and, for every $k \in \{1, \ldots, n\}$, $c_k \le c_{k-1}$.*

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

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

Now let $i_0 \in N \setminus M$ and let $\widetilde{M} = M \cup \{i_0\}$. We have $|\widetilde{M}| = k$ and, from (12) we obtain

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

which yields the result. □

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. We shall see in the sequel that 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).

Many of the 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 3 (ℓ_1 error estimate) *Let X be a $n \times p$ real matrix, and let $(c_k)_{1 \leq k \leq n}$ and $m(X)$ be defined as in* (12) *and* (13)*, respectively. In addition, let $M \subset N$, and let $y, b^* \in \mathbb{R}^n$ and $g^*, g \in \mathbb{R}^p$ be arbitrary. Then the following hold.*

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

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

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

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

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 $|u-v| \geq ||u|-|v|| \geq |u|-|v|$ we obtain

$$\|y - Xg - b^*\|_1 \geq 2 \sum_{i \in N \setminus M} |x_i^\top g - x_i^\top g^*| - \sum_{i \in N} |x_i^\top g - x_i^\top g^*| + \sum_{i \in N} |y_i - x_i^\top g^* - b_i^*| - 2 \sum_{i \in N \setminus M} |y_i - x_i^\top g^* - b_i^*|. \tag{14}$$

It follows from (13) and (12) that $c_k > 1/2$ and there exists $g_k \neq g^*$ such that

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

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 (14) 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

$$y = Xf + z + e, \tag{15}$$

where z is a dense vector of noise and e is an arbitrary sparse vector modelling outliers. Since the least squares estimator 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 least squares estimator in this case. More precisely, if $y_n := y - e = Xf + z$ is the noisy part of the data, devoid of outliers, and $z = X\overline{g} + \overline{b}$, with $\overline{b} \in KerX^\top$ is the orthogonal decomposition of the noise, the LSE on the data y_n is $f_n = (X^\top X)^{-1} X^\top y_n = f + \overline{g}$.

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 $\overline{g} \in \mathbb{R}^p$ and $\overline{b} \in KerX^\top$, and let $f_n = f + \overline{g}$ as discused above. Then the following hold for the ℓ_1 estimator f_1.*

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

$$\|X(f_1 - f_n)\|_1 \leq \frac{1}{2c_k - 1}\left(\|\bar{b}\|_{1,N\setminus M} + \frac{\|\bar{b}\|^2_{2,N\setminus M}}{\|\bar{b}\|}\right). \tag{16}$$

Proof Using Lemma 3(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

$$(2c_k - 1)\|X(f_1 - f_n)\|_1 \leq 2\|\bar{b}\|_{1,N\setminus M} + \|y - Xf_1\|_1 - \|y - Xf_n\|_1. \tag{17}$$

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 (17), the full rank of X, and $c_k > 1/2$ that $f_1 = f_n$. Now suppose that $\|\bar{b}\|_{\infty,N\setminus M} > 0$. Problem (7) can be formulated as a linear program; using linear programming duality we have (Giloni and Padberg 2004, p. 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, by using Lemma 5, we obtain

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

which, altogether with (17), yields (16). □

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.

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, 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 $\|e\|_0 \leq m(X)$, by 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 definiton 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

$$\sum_{i \in N \setminus M} |x_i^\top g_k| \leq \sum_{i \in M} |x_i^\top g_k|. \tag{18}$$

Now define, for $\alpha > 0$,

$$(\forall i \in N) \quad \overline{e}_i = \begin{cases} \alpha x_i^\top g_k, & \text{if } i \in M; \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

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 (18) 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 (19) 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 (15). 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

$$\begin{array}{c} \underset{(g,b,s) \in \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^n}{\text{minimize}} \quad \sigma \|s\|_1 + \frac{1}{2}\|b\|_2^2 \\ s.t \quad y = Xg + b + s, \end{array} \tag{20}$$

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:

$$\underset{(g,b)\in\mathbb{R}^p\times\mathbb{R}^n}{\text{minimize}} \quad \psi(g,b) := \sigma\|y - Xg - b\|_1 + \frac{1}{2}\|b\|_2^2, \tag{21}$$

which can be set in the M-estimator form (5) with

$$\rho\colon r \mapsto \rho(r) = \inf_{b\in\mathbb{R}^n} \sigma\|r - b\|_1 + \frac{1}{2}\|b\|_2^2.$$

The function ρ above equals the Huber's criterion (8) for any $r \in \mathbb{R}^n$ (Michelot and Bougeard 1994). The alternative formulation (20) of Huber's estimation problem based on the error model (15) provides an interpretation of the estimator on finite samples.

Problem (21) can be studied using the ℓ_1 error estimate and duality, as we did in Section 3 for the ℓ_1 estimator. To this end we need to study the optimality conditions and the dual problem. This is done in the following Lemma.

Lemma 4 *The following hold.*

(i) $(\hat{g}, \hat{b})$ *is a solution to* (21) *if and only if* $X^\top \hat{b} = 0$ *and*

$$(\forall i \in \{1,\dots,n\}) \quad \hat{b}_i = \begin{cases} \sigma, & \text{if } y_i - x_i^\top \hat{g} > \sigma; \\ y_i - x_i^\top \hat{g}, & \text{if } y_i - x_i^\top \hat{g} \in [-\sigma, \sigma]\,; \\ -\sigma, & \text{if } y_i - x_i^\top \hat{g} < -\sigma. \end{cases} \tag{22}$$

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

(ii) *A dual of* (21) *is*

$$\gamma := \max_{u\in\sigma P^*} u^\top y - \frac{1}{2}\|u\|_2^2, \tag{23}$$

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

$$\psi(g,b) = \sigma\|y - \begin{bmatrix} X & I_n \end{bmatrix} \begin{pmatrix} g \\ b \end{pmatrix}\|_1 + \frac{1}{2}\|\begin{bmatrix} 0_p & I_n \end{bmatrix} \begin{pmatrix} g \\ b \end{pmatrix}\|_2^2 \tag{24}$$

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 conditions for a solution $(\hat{g}, \hat{b})$ to Problem (21) is

$$0 \in \partial\psi(\hat{g}, \hat{b}). \tag{25}$$

Hence, by using (Hiriart-Urruty and Lemaréchal 1993, Theorem 4.2.1) in (24), (25) is equivalent to

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

Therefore, there exists $u \in \partial\sigma\|\cdot\|_1(y - X\hat{g} - \hat{b})$ such that

$$\begin{cases} X^T u = 0, \\ \hat{b} = u \end{cases}$$

or, equivalently,

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

Hence $y - X\hat{g} - \hat{b} = \operatorname{prox}_{\sigma\|\cdot\|_1}(y - X\hat{g})$, and the result follows from Lemma 1(ii).

(ii): Problem (21) is equivalent to (20) 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,

$$\max_{u\in\mathbb{R}^p} \Big(u^\top y + \big(\min_{b\in\mathbb{R}^n} \frac{1}{2}\|b\|_2^2 - u^\top b\big) + \big(\min_{s\in\mathbb{R}^n} \sigma\|s\|_1 - u^\top s\big) - \max_{g\in\mathbb{R}^p} g^\top(X^\top u)\Big). \quad (26)$$

The optimality conditions associated to the convex optimization problem

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

yields $b = u$, 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 (26) that the dual to (21) is given by (23) and the absence of duality gap follows from the Slater qualification condition and the existence of multipliers (Hiriart-Urruty and Lemaréchal 1993, section 4). □

We pursue the study of the Huber estimator by showing that the additional term b in (21), which makes the difference with respect to the ℓ_1 estimator, improves its response to noisy observations. The numerical simulations performed in Section 5 confirm that the additional term actually plays an important role reducing the bias induced by noise.

Theorem 3 *Let $y = Xf + z + e$, let $M = \mathrm{supp}(e)$, and suppose that $|M| = k \leq m(X)$. Consider the unique decomposition of z as $z = X\overline{g} + \overline{b}$, where $\overline{g} \in \mathbb{R}^p$ and $\overline{b} \in KerX^\top$. Then any solution $(\hat{g}, \hat{b})$ to (21) satisfies*

$$\|X(\hat{g} - f_n)\|_1 \leq \frac{1}{2c_k - 1}\left(\|\overline{b} - \hat{b}\|_{1,N\setminus M} + \frac{\|\overline{b} - \hat{b}\|^2_{2,N\setminus M}}{\|\overline{b} - \hat{b}\|_{\infty,N\setminus M}}\right), \tag{27}$$

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

Proof From Lemma 3(i) and (21) we deduce

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

Hence, it follows from $f_n = f + \bar{g}$ that, for every $i \in \{1, \ldots, n\}$, $y_i - x_i^\top f_n = e_i + \bar{b}_i$ 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 $i \in N \setminus M$,

$$\sigma(2c_k - 1)\|X(\hat{g} - f_n)\|_1 \leq 2\sigma\|\bar{b} - \hat{b}\|_{1,N\setminus M} - \sigma\|e + \bar{b} - \hat{b}\|_1 + \psi(\hat{g}, \hat{b}) - \frac{1}{2}\|\hat{b}\|_2^2. \tag{28}$$

From Lemma 4(ii), the dual problem to (21) is

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

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

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

Hence, it follows from Lemma 5 that

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

which, combined with (28), yields

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

as claimed. □

5 Numerical Illustrations

In this Section we report on a Monte Carlo study intended to illustrate the results of this paper. 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 $f_n = (X^\top X)^{-1} X^\top z$.

In Figure 1 the bias for data with gaussian noise and sparse contamination is plotted. On the left the bias is plotted for different levels of contamination with light-tailed outliers. The LSE outperforms Huber's estimator when the vector of outliers is very sparse (less than 5% of contamination) and, hence, the gaussian noise predominates. However, Huber's estimator has a lower bias in general. The difference of the bias between LSE and ℓ_1 estimator decreases as the percentage of contamination raises. In the figure on the right the bias is plotted for different levels of contamination with heavy-tailed outliers. Both ℓ_1 norm based estimators perform much better than the LSE even for very low levels of contamination. Observing the curves for the ℓ_1 and Huber estimators on both plots, we notice that the behaviour is similar, and consistent with (16) and (27).

In Figure 2 we plot the bias under gaussian noise and very large adversarial sparse errors. On the left, Huber's estimator outperforms dramatically LSE for

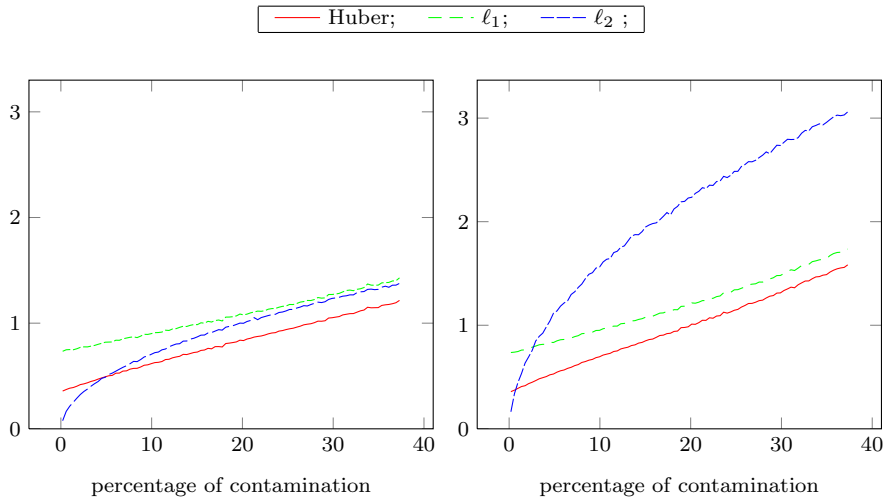

Fig. 1 Relative error $\|\hat{f} - f_n\|/\|f_n\|$ for different percentage of outliers with gaussian noise. On the left, the contamination is drawn from a $N(0, 5)$ distribution and on the right from a Laplace $(0, 5)$ distribution.

any level of contamination; on the right we focus on the low contamination zone to show the better performance of the Huber's estimator with respect to the ℓ_1 estimator. In addition, we appreciate a clear breakdown phenomenon when the level of contamination exceeds the 30% approximately.

In summary, we confirm the high sensitivity of LSE with respect to the percentage of outliers and, in special, with respect to heavy-tailed and adversarial ones. In every examined case Huber's estimator has a better performance compared to the ℓ_1 estimator, and the curves have similar shapes, as expected in view of Theorem 3.

6 Conclusions

We have filled an existing gap in the literature by performing a detailed non-asymptotic study of the robustness of estimators involving ℓ_1 norm minimization. The importance of these estimators stems from the fact that they permit to perform robust regression on very large datasets. The main results are the sharp performance bounds for the ℓ_1 and Huber estimators. Nonetheless, other results such as the ℓ_1 error estimate and the equivalent formulation of Huber's problem bear some interest by their own. The results presented in this article are quantitative, in contrast with the qualitative (bounded/unbounded) character of results related uniquely to the notion of breakdown point, which are particular cases of our analysis. Also, our results are valid for general data, in contrast with previous works based on the restricted isometry property.

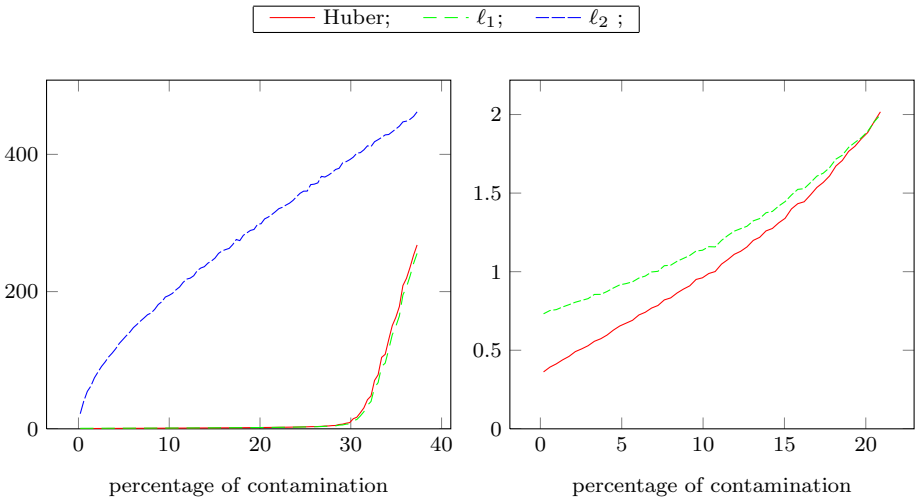

Fig. 2 Plot of relative error $\|\hat{f} - f_n\|/\|f_n\|$, with z standard gaussian, for different fractions of gross errors; at left, with adversarial contamination in the order of 50; at right a closeup comparing Huber's and ℓ_1 on the zone of low contamination.

Appendix

Lemma 5 *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 $\sigma > 0$,*

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

Proof Let

$$\tilde{b}_i = \begin{cases} 0, & \text{if } i \in M; \\ b_i, & \text{otherwise,} \end{cases} \quad \text{and} \quad \tilde{e}_i = \begin{cases} b_i + e_i, & \text{if } i \in M; \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

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

$$\max_{d \in \sigma P^*} d^\top (e + b) = \max_{d \in \sigma P^*} d^\top (\tilde{e} + \tilde{b}) \leq \max_{d \in \sigma P^*} d^\top \tilde{e} + \max_{d \in \sigma P^*} d^\top \tilde{b}. \tag{30}$$

On one hand, it follows from Lemma 3(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, $(\exists u \in P^*) \quad u^\top \tilde{e} = \|\tilde{e}\|_1$. 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

$$\max_{d \in \sigma P^*} d^\top \tilde{e} = \sigma \cdot \max_{u \in P^*} u^\top \tilde{e} = \sigma \|\tilde{e}\|_1. \tag{31}$$

On the other hand,

$$\max_{d\in\sigma P^*} d^\top\tilde{b} \leq \max_{\|d\|_\infty\leq\sigma} d^\top\tilde{b} = \frac{\sigma}{\|\tilde{b}\|_\infty}\tilde{b}^\top\tilde{b} = \frac{\sigma}{\|\tilde{b}\|_\infty}\|\tilde{b}\|_2^2. \tag{32}$$

Therefore, by replacing (31) and (32) in (30), the result follows from (29). □

Acknowledgements The author is grateful to Luis Briceño-Arias for his careful reading and 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). This work was supported by Comisión Nacional de Investigación Científica y Tecnológica through FONDECYT program, grant 3120166, and FONDAP-BASAL program.

References

Anderson KR (1982) Robust earthquake location using m-estimates. Physics of the Earth and Planetary Interiors 30(23):119 – 130

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

Chang XW, Guo Y (2005) Hubers m-estimation in relative gps positioning: computational aspects. Journal of Geodesy 79(6-7):351–362

Chen B, Pinar M (1998) On newton's method for huber's robust m-estimation problems in linear regression. BIT Numerical Mathematics 38(4):674–684

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

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

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

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, Grundlehren der mathematischen Wissenschaften, vol 305. Springer-Verlag

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

Huber PJ (1981) Robust statistics. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons Inc., New York

Madsen K, Nielsen H (1990) Finite alogorithms for robust linear regression. BIT Numerical Mathematics 30(4):682–699

Maumet C, Maurel P, Ferr JC, Barillot C (2014) Robust estimation of the cerebral blood flow in arterial spin labelling. Magnetic Resonance Imaging 32(5):497 – 504

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

Naseem I, Togneri R, Bennamoun M (2012) Robust regression for face recognition. Pattern Recognition 45(1):104 – 118

Shao J (2003) Mathematical statistics. Springer Texts in Statistics, Springer-Verlag, New York

Tukey JW (1960) A survey of sampling from contaminated distributions. In: Contributions to probability and statistics, Stanford Univ. Press, Stanford, Calif., pp 448–485

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