Does Compressed Sensing have applications in Robust Statistics?

Salvador Flores

December 1, 2014

Abstract

The connections between robust linear regression and sparse reconstruction are brought to light. We show that in the context of fixed design, the notion of breakdown point coincides with exact recovery of sparse signals from highly incomplete information. The main consequence of this connection in robust regression is that there exists, for any dimension, "many" designs on which the ℓ_1 estimator has a positive breakdown point. This result clarifies a common misunderstanding on the robustness of M-estimators.

1 Introduction

In recent years there has been a lot of excitement about the advances in the reconstruction of sparse signals by ℓ_1-norm minimization and its applications to compressed sensing. The basic question in compressed sensing (Donoho, 2006) is the following: if a vector $e \in \mathbb{R}^n$ has a sparse representation in some basis, can we reconstruct e from less than n linear combinations of its components? A first theoretical answer is positive; if F is a $p \times n$ matrix $(p < n)$ with columns in general position and e has at most $(n-p-1)/2$ nonzero components, then e is the unique solution to

$$\begin{array}{ll} \min\limits_{s \in \mathbb{R}^n} & \|s\|_0 \\ s.t & Fs = \tilde{y}. \end{array} \tag{1}$$

where $\tilde{y} = Fe$ are the measurements and $\|e\|_0$ denotes the "ℓ_0 norm" of e, defined as the number of nonzero components. However, the difficulty of solving Problem (1) makes it impractical. Instead, one solves its closest convex problem

$$\begin{array}{ll} \min\limits_{s \in \mathbb{R}^n} & \|s\|_1 \\ s.t & Fs = \tilde{y}. \end{array} \tag{2}$$

Conditions on p and F ensuring the equivalence of problems (1) and (2) has been the subject of extensive research with many impressive results.

The study of under-determined problems, such as (2) has concentrated most of this research effort. However, in some circumstances Problem (2) is equivalent to the following overdetermined least absolute deviations problem (Candes and Tao, 2005; Zhang, 2005)

$$\min_{g \in \mathbb{R}^p} \|y - Xg\|_1, \tag{3}$$

provided that

$$FX = 0 \quad \text{and} \quad \tilde{y} = Fy.$$

Built on this equivalence, we explore the striking connections between the sparse reconstruction problem and robust linear regression, with ℓ_1-norm minimization as common theme.

Let us consider the following problem; we send a vector $f \in \mathbb{R}^p$ (the signal) encoded by a $n \times p$ matrix X. When delivered to its recipients, the encoded information is corrupted by arbitrary and unknown errors. In its noiseless version, the problem is to recover f from the corrupted measurements $y = Xf + e$, where e is a sparse vector of errors. Candes and Tao (2005) show that the information f can be *exactly* recovered as the unique solution to Problem (3), provided that X satisfies some conditions and the number of nonzero components of e is small enough. Viewed from a statistical perspective, the result seems overly strong; recovering *exactly* the signal is not the kind of result one expects from a regression procedure, even less if the errors are of arbitrary magnitude. Before looking at the rather involved conditions on X, we point out an unusual hypothesis: sparsity. In simple words, the sparsity hypothesis reads " a small fraction of the observations can be subject to arbitrary errors, but the large majority of them is completely free of errors". The study of the effects of arbitrary contamination on statistical analysis is not new and there exists a whole branch of statistics devoted to the subject. On the contrary, that the majority of the observations are completely free of errors is a situation unlikely to have been considered before in the statistics literature. In robust statistics, a central role is played by the notion of breakdown point, which mesures the resistance level of an estimator when some observations are replaced by arbitrary ones. Though, the part of the observations considered as "clean" follows a linear model and is by no means supposed to be completely free of errors. One may think that the results in robust regression dealing with the more general noisy version of the problem should recover Candes and Tao ones under the same hypothesis, but it is not so. In fact, a major drawback of the breakdown point notion is that by looking at the limit as contamination diverges it negliges a lot of useful information. It is regrettable that the dichotomic character of these results (bounded/unbounded) in some cases overshadow elaborated developments. This is case of He et al. (1990), where it was derived for the first time the regression breakdown point of the ℓ_1 estimator. Recently, Flores (2014) extended those results and gave sharp error bounds for the ℓ_1 estimator. We shall show that using Flores bounds we are able to recover all of the most recent results on sparse recovery, establishing an unexpected link between a 25 years old statistical theory with one of the most dynamic research fields in the last decade.

2 Connections with sparse reconstruction

Let us consider the problem of recovering an input f from corrupted measurements

$$y = Xf + e, \tag{4}$$

when the error term e is sparse. This problem can solved by exploiting recent advances in the study of Sparse Reconstruction Problems. If we consider a matrix F such that $\ker(F) = \operatorname{ran}(X)$, from (4) we obtain $Fy = Fe$. Thus, if additionaly e is the *unique* solution to the convex problem

$$\begin{array}{cc} \min\limits_{s \in \mathbb{R}^n} & \|s\|_1 \\ & F(s - y) = 0, \end{array} \tag{5}$$

then it is possible to recover the signal f from e by solving the system

$$Xf = y - e. \tag{6}$$

Hereafter we say that F has the exact recovery property of order k if, whenever $\|e\|_0 \leq k$,

$$\{e\} = \operatorname{argmin}\{\|s\|_1 : \ Fs = Fe\}. \tag{7}$$

Candes and Tao (2005) provide sufficient conditions for exact recovery for matrices F satisfying a *restricted isometry property.* Zhang (2005) gives an account of necessary and sufficient conditions for exact recovery. An important concept there is k- balancedness, defined below

Definition 1 *The vector subspace V is strictly k-balanced if for any set $M \subseteq \{1, ..., n\}$ of cardinality less or equal to k it holds*

$$\sum_{i \in M} |z_i| < \sum_{i \in N \setminus M} |z_i| \quad \textit{for any } z \in V$$

Zhang (2013, 2005) show that F has the exact recovery property of order k if and only if $\ker(F)$ is strictly k-balanced. In a recent result Juditsky and Nemirovski (2011, Theorem 1) show that F has the exact recovery property of order k if and only if $\hat{\gamma}_k(F) < 1/2$, where $\hat{\gamma}_k(F)$ is defined by

$$\hat{\gamma}_k(F) = \max_{\substack{s \in \ker F \\ \|s\|_1 \leq 1}} \max_{\substack{M \subset N \\ |M| = k}} \sum_{i \in M} |s_i|. \tag{8}$$

Now we would like to compare this exact recovery result with statistical ones. As already said there are not results in statistics considering analysis of noiseless data

In practice one expects that all observations carry some noise. A more realistic model is

$$y = Xf + z + e, \tag{9}$$

where z is a dense, presumably small, vector of noise and e is an arbitrary sparse vector. Under this model, exact recovery is not longer possible. The goodness of an estimator is measured by its distance to some reference point, which can be f or some estimator of it. If there is a bound on that distance which is finite for any e such that $\|e\|_0 \leq k$, then the RBP of the estimator is at least k. Moreover, fine bounds on the recovery error provides not only an asymptotic description of an estimator, as RBP does, but also a more accurate study of it in front of finite errors.

We will rather focus on the bounds by Juditsky and Nemirovski (2011), which are based on necessary and sufficient conditions, for the noisy recovery problem

$$\begin{array}{cc} \min\limits_{s \in \mathbb{R}^n} & \|s\|_1 \\ & \|Fs - \tilde{y}\| \leq \sigma. \end{array} \tag{10}$$

In (10) it is supposed that the information Fe is not longer available with infinite precision, but instead we receive a noisy version of it $\tilde{y} = Fe + \xi$, for some $\|\xi\| \leq \varepsilon$. This framework is adapted to the error model (9) for bounded b, since

$$\tilde{y} = Fy = Fe + Fz = Fe + \bar{b},$$

where $\bar{b}$ is the projection of z onto $\operatorname{ran}(X)^{\perp}$.

In order to compare the results of Juditsky and Nemirovski (2011) with those of Flores (2014), we suppose that provided an optimal solution d to (10), we obtain g by solving $Xg = y - d$ in the least squares sense. Then

$$\|H(e-d)\| = \|X(f_n - g)\|. \tag{11}$$

where $H = X(X^\top X)^{-1}X^\top$ is the hat matrix.

In this way, we can bring back the errors bounds obtained for under-determined problems to compare with bounds available in the overdetermined case. Juditsky and Nemirovski (2011) show the following

Theorem 1 *Let M be a subset of $N := \{1, ..., n\}$ with cardinality $k \leq n$ and F a $p \times n$ matrix such that $\hat{\gamma}_k(F) < 1/2$.*

(i) If $\tilde{y} = Fe$ and d is a solution to Problem (5)*, then*

$$\|d - e\|_1 \leq \frac{2}{1 - 2\hat{\gamma}_k(F)} \sum_{i \in N \setminus M} |e_i|$$

(ii) If $\tilde{y} = Fe + \xi$, $\|\xi\| \leq \varepsilon$ and d is a solution to Problem (10)*, then*

$$\|d - e\|_1 \leq \frac{2}{1 - 2\hat{\gamma}_k(F)} (\beta(\sigma + \varepsilon) + \sum_{i \in N \setminus M} |e_i|)$$

for some constant $\beta > 0$ large enough to satisfy certain conditions.

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

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

and

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

The quantity $m(X)$ is an alternative representation of the breakdown point of the ℓ_1 estimator. Note that the condition $c_k > 1/2$ is equivalent to say that $\mathrm{ran}(X)$ is strictly k-balanced. The leverage constants are closely related to the constants $\hat{\gamma}_k$ and $s_*(F)$ as follows

Lemma 1 *Let F be such that $\ker(F) = \mathrm{ran}(X)$ as in Section 2. Let $\hat{\gamma}_k(F)$ be defined in (8) and $s_*(F) = \max\left\{k \in N \;\middle|\; \hat{\gamma}_k(F) < \frac{1}{2}\right\}$. These constants are related to $c_K(X)$ and $m(X)$ via*

$$c_K(X) = 1 - \hat{\gamma}_k(F) \tag{14}$$

and $m(X) = s_*(F)$.

Now we are in position to compare with the bounds in Theorem 1 with the bounds available for robust regression

Theorem 2 *Let $y = Xf + z + e$ and $M \subseteq N$ 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 above. Then the following hold for the ℓ_1 estimator f_1.*

$$\|X(f_1 - f_n)\|_1 \leq \frac{1}{2c_k - 1} \left(\sum_{i \in N \setminus M} |\bar{b}_i + e_i| + \frac{\sum\limits_{i \in N \setminus M} |\bar{b}_i + e_i|^2}{\max\limits_{i \in N \setminus M} |\bar{b}_i + e_i|} \right). \tag{15}$$

Note that, by Hölder inequality,

$$\frac{\sum\limits_{i \in N \setminus M} |\bar{b}_i + e_i|^2}{\max\limits_{i \in N \setminus M} |\bar{b}_i + e_i|} \leq \sum_{i \in N \setminus M} |\bar{b}_i + e_i| \tag{16}$$

then (15) can be simplified to yield

$$\|X(f_1 - f_n)\|_1 \leq \frac{1}{c_k - 1/2} \sum_{i \in N \setminus M} |\bar{b}_i + e_i|. \tag{17}$$

From here, if $b = 0$ and $|M| = |\operatorname{supp}(e)| \leq m(X) = k$ we rediscover that exact recovery occurs for noiseless data if and only if $\operatorname{ran}(X)$ is strictly k-balanced.

Altogether, (15), (14), (11) and (16) show that neither of the bounds in Theorem 1 improve the existing results in robust statistics. In particular, the bounds for the noisy decode problem (10) do not improve the bounds for the pure ℓ_1 estimator, contrarily to Huber M-estimator, which improves the bound (15) by reducing the noise (Flores, 2014, Theorem 3). To be fair we should mention that Problem (10) is intended to solve the under determined problem while the bounds in Flores (2014) are specialized to the over determined case. Some cruder bounds based on restricted isometry constants for the underdetermined case have been derived for the Dantzig selector, a variant of (10) (Candes and Tao, 2007), and for the regularized least squares problem (Zhang, 2009).

3 The breakdown point of ℓ_1 regression for large p

Until now, we have seen that all the results obtained in compressed sensing have a counterpart, often sharper, in robust regression. However, there is one result related to the Kolmogorov diameter of balls dig out by compressed sensing researchers (cf. Yin and Zhang, 2008; Zhang, 2005; Candes et al., 2006) which has deep consequences when applied to robust regression.

Theorem 3 *Let p and n be any natural numbers with $p < n$. There exists a set Ξ of $n \times p$ matrices with positive Grassmanian measure such that any design matrix in Ξ has breakdown order k whenever*

$$\frac{k}{n} < \alpha\phi\left(\frac{n-p}{n}\right)$$

for an absolute constant $\alpha > 0$ independent of n and p, where $\phi(t) = t/(1 - \log(t))$.

Remark 1 *The function ϕ is monotone and satisfies $0 < \phi(t) \leq 1$ for $0 < t \leq 1$, $\lim_{t\to 0}\phi(t) = 0$ and $\phi(1) = 1$. In particular $\liminf\limits_{n\to\infty, p\to\infty} \phi((n-p)/n) > 0$ whenever $\liminf\limits_{n\to\infty, p\to\infty} p/n < 1$.*

Theorem 3 disproves a common belief in robust regression that M-estimators and ℓ_1 estimation in particular have a breakdown point going to 0 as p increases. This belief originated in a result by Maronna et al. (1979) showing that if the rows of X are sampled from a spherically symmetric distribution, then the breakdown point of X for ℓ_1 estimation behaves like $(2p)^{-1/2}$ for large p. Concrete examples of matrices achieving the abstract result of Theorem 3 has concentrated the essential of the research effort in compressed sensing, and it has been well stablished that properly scaled gaussian matrices do so with high probability.

4 Conclusions

We have explored the connection between the problems of robust regression and sparse recovery by ℓ_1 norm minimization. Both theories have independently arrived to the common fundamental results. However, the treatment of the problem in both cases differ greatly. In robust regression, most of the results on ℓ_1 minimization have been presented in a negative way, in contrast to high-breakdown point estimators, despite the fact that the later are not computable in practice, except for very small datasets. For this reason, we can hardly find in the robust regression literature examples where the ℓ_1 estimator gives right answers in contaminated observations. On the contrary, compressed sensing theory focused from the very beginning in identifying the most favourable cases and prving positive results. This explains in our opinion the fact that Theorem 3, which a cornerstone in compressed sensing, was still unknown in robust regression.

References

Candes, E., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509.

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

Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35:2313–2351.

Donoho, D. (2006). Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289 –1306.

Flores, S. (2014). Sharp non-asymptotic performance bounds for ℓ_1 and Huber robust regression estimators. *Test*, xx:x–xx.

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

Juditsky, A. and Nemirovski, A. (2011). On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *Math. Program.*, 127(1).

Maronna, R., Bustos, O., and Yohai, V. (1979). Bias- and efficiency-robustness of general m-estimators for regression with random carriers. In Gasser, T. and Rosenblatt, M., editors, *Smoothing Techniques for Curve Estimation*, volume 757 of *Lecture Notes in Mathematics*, pages 91–116. Springer Berlin / Heidelberg.

Yin, W. and Zhang, Y. (2008). Extracting salient features from less data via ℓ_1-minimization. *SIAG/OPT Newsletter: Views & News*, 19(1):11–19.

Zhang, T. (2009). some sharp performance bounds for least squares regression with l_1 regularization. *Ann. Statist.*, 37:2109–2144.

Zhang, Y. (2005). A simple proof for recoverability of ℓ_1 minimization: Go over or under? *Rice University CAAM technical report*, TR05-09.

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.