Consistent Robust Regression

Kush Bhatia kushbhatia@berkeley.edu
University of California, Berkeley
Prateek Jain prajain@microsoft.com
Microsoft Research, India
Parameswaran Kamalaruban kamalaruban.parameswaran@epfl.ch
EPFL, Switzerland
Purushottam Kar purushot@cse.iitk.ac.in
Indian Institute of Technology, Kanpur

Abstract

We present the first efficient and provably consistent estimator for the robust regression problem. The area of robust learning and optimization has generated a significant amount of interest in the learning and statistics communities in recent years owing to its applicability in scenarios with corrupted data. In particular, special interest has been devoted to the problem of robust linear regression where estimators that can tolerate corruption in up to a constant fraction of the response variables are widely studied. Surprisingly however, to this date, we are not aware of a polynomial time estimator that offers a consistent estimate in the presence of dense, unbounded corruptions. In this work we present such an estimator, called CRR. This solves an open problem put forward in the work of [2]. Our consistency analysis requires a novel two-stage proof technique involving a careful analysis of the stability of ordered lists which may be of independent interest.

1 Introduction

The problem of robust learning involves designing and analyzing learning algorithms that can extract the underlying model despite dense, possibly malicious, corruptions in the training data provided to the algorithm. The problem has been studied in a dizzying variety of models and settings ranging from regression [12], classification [7], dimensionality reduction [3] and matrix completion [6].

In this paper we are interested in the Robust Least Squares Regression (RLSR) problem that finds several applications to robust methods in face recognition and vision [15, 14], and economics [12]. In this problem, we are given a set of n covariates in d dimensions, arranged as a data matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, and a response vector $\mathbf{y} \in \mathbb{R}^n$. However, it is known apriori that a certain number k of these responses cannot be trusted since they are corrupted. These may correspond to corrupted pixels in visual recognition tasks or untrustworthy measurements in general sensing tasks.

Using these corrupted data points in any standard least-squares solver, especially when $k = \mathcal{O}(n)$, is likely to yield a poor model with little predictive power. A solution to this is to exclude corrupted points from consideration. The RLSR problem formalizes this requirement as follows:

$$(\widehat{\mathbf{w}}, \widehat{S}) = \underset{\substack{\mathbf{w} \in \mathbb{R}^p, S \subset [n] \\ |S| = n-k}}{\arg\min} \sum_{i \in S} (y_i - \mathbf{x}_i^T \mathbf{w})^2, \tag{1}$$

This formulation seeks to simultaneously extract the set of uncorrupted points and estimate the least-squares solutions over those uncorrupted points. Due to the combinatorial nature of the RLSR formulation (1), solving it directly is challenging and in fact, NP-hard in general [2, 13].

Literature in robust statistics suggests several techniques to solve (1). The most common model assumes a realizable setting wherein there exists a gold model $\mathbf{w}^*$ that generates the non-corrupted responses. A vector of *corruptions* is then introduced to model the corrupted responses i.e.

$$\mathbf{y} = X^T \mathbf{w}^* + \mathbf{b}^*. \tag{2}$$

The goal of RLSR is to recover $\mathbf{w}^* \in \mathbb{R}^d$, the *true* model. The vector $\mathbf{b}^* \in \mathbb{R}^n$ is a k-sparse vector which takes non-zero values on at most k corrupted samples out of the n total samples, and a zero value elsewhere.

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

Paper	Breakdown Point	Adversary	Consistent	Technique
Wright & Ma, 2010 [14]	$\alpha \to 1$	Oblivious	No	L_1 regularization
Chen & Dalalyan, 2010 [4]	$\alpha \geq \Omega(1)$	Adaptive	No	Second-order cone program
Chen et al., 2013 [5]	$\alpha \geq \Omega\left(\frac{1}{\sqrt{d}}\right)$	Adaptive	No	Robust thresholding
Nguyen & Tran, 2013 [11]	$\alpha \to 1$	Oblivious	No	L_1 regularization
Bhatia et al., 2015 [2]	$\alpha \geq \Omega(1)$	Adaptive	No	Iterative hard thresholding
This paper	$\alpha \geq \Omega(1)$	**Oblivious**	**Yes**	**Iterative hard thresholding**

Table 1: A comparison of different RLSR algorithms and their properties. CRR is the first efficient RLSR algorithm to guarantee consistency in the presence of a constant fraction of corruptions.

A more useful, but challenging model is one in which (mostly heteroscedastic and i.i.d.) Gaussian noise is injected into the responses in addition to the corruptions.

$$\mathbf{y} = X^T \mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon}. \tag{3}$$

2 Related Works

A string of recent works have looked at the RLSR problem in various settings. To facilitate a comparison among these, we set the following benchmarks for RLSR algorithms

1. (Breakdown Point) This is the number of corruptions k that an RLSR algorithm can tolerate is a direct measure of its robustness. The breakdown point k is frequently represented as a fraction α of the total number of data points i.e. $k = \alpha \cdot n$.
2. (Adversary Model) RLSR algorithms frequently resort to an adversary model to specify how are the corruptions introduced into the regression problem. The strictest is the *adaptive adversarial* model wherein the adversary is able to view X and $\mathbf{w}^*$ (as well as $\boldsymbol{\epsilon}$ if Gaussian noise is present) before deciding upon $\mathbf{b}^*$. In contrast, in an *oblivious adversarial* model the adversary generates a k-sparse vector in complete ignorance of X and $\mathbf{w}^*$ (and $\boldsymbol{\epsilon}$).
3. (Consistency) RLSR algorithms that are able to operate in the *hybrid* noise model with sparse adversarial corruptions as well as dense Gaussian noise are more valuable. An RLSR algorithms is said to be consistent if, when invoked in the hybrid noise model, the algorithm returns an estimate $\widehat{\mathbf{w}}_n$ such that $\lim_{n\to\infty} \mathbb{E}\left[\widehat{\mathbf{w}}_n - \mathbf{w}^*\right]_2 = 0$.

In Table 1, we present a summarized view of existing RLSR techniques and their performance vis-a-vis the benchmarks discussed above. Past work has seen the application of a wide variety of algorithmic techniques to solve this problem, inclding more expensive methods involving L_1 regularization (for example $\min_{\mathbf{w},\mathbf{b}} \lambda_w \|\mathbf{w}\|_1 + \lambda_b \|\mathbf{b}\|_1 + \|X^\top \mathbf{w} + \mathbf{b} - \mathbf{y}\|_2^2$) and second-order cone programs such as [14, 4, 10, 11], as well as more scalable methods such as the robust thresholding and iterative hard thresholding [5, 2].

An important point of consideration is the breakdown point of these methods. Among those cite in Table 1, the works of [14] and [10] obtain the best breakdown points that allow a fraction of points to be corrupted that is arbitrarily close to 1. They require the data to be generated from either an isotropic Gaussian ensemble or be row-sampled from an incoherent orthogonal matrix. Most results mentioned in the table allow a constant fraction of points to be corrupted i.e. allow $k = \alpha \cdot n$ corruptions for some fixed constant $\alpha > 0$. This is still impressive since it allows a dense subset of data points to be corrupted and yet guarantee recovery. However, as we shall see below, these results cannot guarantee consistency while allowing allow $k = \alpha \cdot n$ corruptions.

We note that we use the term *dense* to refer to the corruptions in our model since they are a constant fraction of the total available data. Moreover, as we shall see, this constant shall be universal and independent of the ambient dimensionality d. This terminology is used to contrast against some other works which can tolerate only $o(n)$ corruptions which is arguably much sparser. For instance, as we shall see below, the work of [11] can tolerate only $o(n/\log n)$ corruptions if a consistent estimate is expected. The work of [5] also offers a weak guarantee wherein they are only able to tolerate a $1/\sqrt{d}$ fraction of corruptions. However, [5] allow corruptions in covariates as well.

However, we note that *none* of the algorithms listed here are able to guarantee a consistent solution, irrespective of assumptions on the adversary model. At best, they guarantee $\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \mathcal{O}(\sigma)$ when $k = \Omega(n)$ where σ is the standard deviation of the white noise (see Equation 3). Thus, their estimation error is of the order of the white noise in the system, even if the algorithm is supplied with an infinite amount of data.

Algorithm 1 CRR: Consistent Robust Regression

Input: Covariates $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, responses $\mathbf{y} = [y_1, \ldots, y_n]^\top$, corruption index k, tolerance ϵ

$\mathbf{b}^0 \leftarrow \mathbf{0}, t \leftarrow 0,$
$P_X \leftarrow X^\top (XX^\top)^{-1} X$
while $\left\|\mathbf{b}^t - \mathbf{b}^{t-1}\right\|_2 > \epsilon$ **do**
  $\mathbf{b}^{t+1} \leftarrow \mathrm{HT}_k(P_X \mathbf{b}^t + (I - P_X)\mathbf{y})$
  $t \leftarrow t + 1$
end while
return $\mathbf{w}^t \leftarrow (XX^\top)^{-1} X(\mathbf{y} - \mathbf{b}^t)$

3 Our Contributions

In this paper, we remedy the above problem by using a simple and scalable iterative hard-thresholding algorithm called CRR along with a novel two-stage proof technique. Given n covariates that form a Gaussian ensemble, our method in time poly(n, d), outputs an estimate $\widehat{\mathbf{w}}_n$ s.t. $\|\widehat{\mathbf{w}}_n - \mathbf{w}^*\|_2 \to 0$ as $n \to \infty$ (see Theorem 1 for a precise statement). In fact, our method guarantees a nearly optimal error rate of $\|\widehat{\mathbf{w}}_n - \mathbf{w}^*\|_2 \leq \sigma\sqrt{\frac{d}{n}}$. It is noteworthy that CRR can tolerate a constant fraction of corruptions i.e. tolerate $k = \alpha \cdot n$ corruptions for some fixed $\alpha > 0$.

We note that although hard thresholding techniques have been applied to the RLSR problem earlier [2, 5], none of those methods are able to guarantee a consistent solution to the problem. Our results hold in the setting where a *constant fraction* of the responses are corrupted by an *oblivious adversary* (i.e. the one which corrupts observations without information of the data points themselves). Our algorithm runs in time $\widetilde{\mathcal{O}}\left(d^3 + nd\right)$, where d is the dimensionality of the data. Moreover, as we shall see, our technique makes more efficient use of data than previous hard thresholding methods such as TORRENT [2].

To the best of our knowledge, this is the *first* efficient and consistent estimator for the RLSR problem in the challenging setting where a *constant* fraction of the responses may be corrupted in the presence of dense noise. We would like to note that the problem of consistent robust regression is especially challenging because without the assumption of an *oblivious* adversary, consistent estimation with a constant fraction of corruptions (even for an arbitrarily small constant) is *impossible* [1].

However, by crucially using the restriction of obliviousness on the adversary along with a novel proof technique, we are able to provide a consistent estimator for RLSR with optimal (up to constants) statistical and computational complexity.

4 Problem Formulation

We are given n data points $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where $\mathbf{x}_i \in \mathbb{R}^d$ are the *covariates* and, for some *true* model $\mathbf{w}^* \in \mathbb{R}^d$, the vector of *responses* $\mathbf{y} \in \mathbb{R}^n$ is generated

$$\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b}^* + \boldsymbol{\epsilon}, \tag{4}$$

The responses suffer two kinds of perturbations – *dense white noise* $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ that is chosen in an i.i.d. fashion independently of the data X and the model $\mathbf{w}^*$, and *adversarial corruptions* in the form of $\mathbf{b}^*$. We assume that $\mathbf{b}^*$ is a k^*-sparse vector albeit one with potentially unbounded entries. The constant k^* will be called the *corruption index* of the problem. We assume the *oblivious adversary* model where $\mathbf{b}^*$ is chosen independently of $X, \mathbf{w}^*$ and $\boldsymbol{\epsilon}$.

Although there exist works that operate under a fully adaptive adversary [2, 4], none of these works are able to give a consistent estimate. We also note that existing works are unable to give consistent estimates even in the oblivious adversary model.

5 CRR: A Hard Thresholding Approach to Consistent Robust Regression

We now present a consistent method CRR for the RLSR problem. CRR takes a different approach to the problem than previous works. Instead of attempting to exclude data points deemed unclean (as done by the TORRENT algorithm of [2]), CRR focuses on correcting the errors.

To motivate the CRR algorithm, we start with the RLSR formulation $\min_{\mathbf{w}\in\mathbb{R}^p, \|\mathbf{b}\|_0 \le k^*} \frac{1}{2}\left\|X^\top \mathbf{w} - (\mathbf{y}-\mathbf{b})\right\|_2^2$, and realize that given any estimate $\widehat{\mathbf{b}}$ of the corruption vector, the optimal model with respect to this estimate is given by the expression $\widehat{\mathbf{w}} = (XX^\top)^{-1}X(\mathbf{y}-\widehat{\mathbf{b}})$. Plugging this expression for $\widehat{\mathbf{w}}$ into the formulation allows us to reformulate the RLSR problem.

$$\min_{\|\mathbf{b}\|_0 \le k^*} f(\mathbf{b}) = \frac{1}{2}\|(I - P_X)(\mathbf{y}-\mathbf{b})\|_2^2 \tag{5}$$

where $P_X = X^\top(XX^\top)^{-1}X$. This greatly simplifies the problem by casting it as a *sparse parameter estimation* problem instead of a data subset selection problem (as done by TORRENT). CRR directly optimizes (5) by using a form of iterative hard thresholding. At each step, CRR performs the following update: $\mathbf{b}^{t+1} = \mathrm{HT}_k(\mathbf{b}^t - \nabla f(\mathbf{b}^t))$, where k is a parameter for CRR. We note that CRR functions with a fixed, unit step length, which is convenient in practice as it avoids step length tuning, something most IHT algorithms [8, 9] require.

6 Consistency Guarantees for CRR

Theorem 1. *Let $x_i \in \mathbb{R}^d, 1 \le i \le n$ be generated i.i.d. from a Gaussian distribution, let y_i's be generated using (4) for a fixed $\mathbf{w}^*$, and let σ^2 be the noise variance. Also let the number of corruptions k^* be s.t. $2k^* \le k \le n/10000$. Then, with probability at least $1-\delta$, after $\mathcal{O}\left(\log \frac{\|\mathbf{b}^*\|_2}{\sigma k + \epsilon} + \log\frac{n}{d}\right)$ steps,* CRR *ensures that $\|\mathbf{w}^t - \mathbf{w}^*\|_2 \le \epsilon + \mathcal{O}\left(\frac{\sigma}{\sqrt{\lambda_{\min}(\Sigma)}}\sqrt{\frac{d}{n}\log\frac{nd}{\delta}}\right)$.*

The above result establishes consistency of the CRR method with an error rate of $\tilde{\mathcal{O}}(\sigma\sqrt{d/n})$ that is known to be statistically optimal. It is notable that this optimal rate is being ensured in the presence of gross and unbounded outliers.

For our analysis, we will divide CRR's execution into two phases – a *coarse convergence* phase and a *fine convergence* phase. CRR will enjoy a linear rate of convergence in both phases. However, the coarse convergence analysis will only ensure $\|\mathbf{w}^t - \mathbf{w}^*\|_2 = \mathcal{O}(\sigma)$. The fine convergence phase will then use a much more careful analysis of the algorithm to show that in at most $\mathcal{O}(\log n)$ more iterations, CRR ensures $\|\mathbf{w}^t - \mathbf{w}^*\|_2 = \tilde{\mathcal{O}}(\sigma\sqrt{d/n})$, thus establishing consistency of the method.

Coarse convergence: Here we establish a result that guarantees that after a certain number of steps T_0, CRR identifies the corruption vector with a relatively high accuracy and consequently ensure that $\left\|\mathbf{w}^{T_0} - \mathbf{w}^*\right\|_2 \le \mathcal{O}(\sigma)$. We utilize the notions of *Subset Strong Convexity* and *Subset Strong Smoothness* in our proofs similar to [2].

Lemma 2. *For any data matrix X that satisfies the SSC and SSS properties such that $\frac{2\Lambda_{k+k^*}}{\lambda_n} < 1$,* CRR*, when executed with a parameter $k \ge k^*$, ensures that after $T_0 = \mathcal{O}\left(\log\frac{\|\mathbf{b}^*\|_2}{e_0+\epsilon}\right)$ steps, $\left\|\mathbf{b}^{T_0} - \mathbf{b}^*\right\|_2 \le 3e_0 + \epsilon$, where $e_0 = \mathcal{O}\left(\sigma\sqrt{(k+k^*)\log\frac{n}{\delta(k+k^*)}}\right)$ for standard Gaussian designs.*

Fine convergence: We now show that CRR progresses further at a linear rate to achieve a consistent solution. In Lemma 3, we show that $\|X(\mathbf{b}^t - \mathbf{b}^*)\|$ has a linear decrease for every iteration $t > T_0$ along with a term which is $\tilde{\mathcal{O}}(\sqrt{dn})$.

Lemma 3. *Let $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ be a data matrix consisting of i.i.d. standard normal vectors i.e $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, I_{d\times d})$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \cdot I_{n\times n})$ be a standard normal vector of white noise values drawn independently of X. For any $\boldsymbol{\lambda} \in \mathbb{R}^d$ such that $\|\boldsymbol{\lambda}\|_2 \le \frac{\sigma}{100}$, define $\mathbf{b}^{new} = HT_k(X^\top\boldsymbol{\lambda} + \boldsymbol{\epsilon} + \mathbf{b}^*)$, $\mathbf{z}^{new} = \mathbf{b}^{new} - \mathbf{b}^*$ and $\boldsymbol{\lambda}^{new} = (XX^T)^{-1}X\mathbf{z}^{new}$, where $k \ge 2k^*$, $|supp(\mathbf{b}^*)| \le k^*$, $k^* \le n/1000$, and $d \le n/1000$. Then, with probability at least $1 - 1/n^5$, for every $\boldsymbol{\lambda}$ s.t. $\|\boldsymbol{\lambda}\|_2 \le \frac{\sigma}{100}$, we have*

$$\|X\mathbf{z}^{new}\|_2 \le .9n\|\boldsymbol{\lambda}\| + 100\sigma\sqrt{d\cdot n}\log^2 n,$$

$$\|\boldsymbol{\lambda}^{new}\|_2 \le .91\|\boldsymbol{\lambda}\| + 110\sigma\sqrt{\frac{d}{n}}\log^2 n.$$

Putting all these results together establishes Theorem 1.

For full proofs, convergence analyses of the other variants, and experimental results, see the long version of the paper (Full Version).

References

[1] Anonymous. Lower Bounds on Consistent Robust Regression. *Personal communication*, 2016.

[2] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust Regression via Hard Thresholding. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS)*, 2015.

[3] Emmanuel J. Candès, Xiaodong Li, and John Wright. Robust Principal Component Analysis? *Journal of the ACM*, 58(1):1–37, 2009.

[4] Yin Chen and Arnak S. Dalalyan. Fused sparsity and robust estimation for linear models with unknown variance. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.

[5] Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust Sparse Regression under Adversarial Corruption. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

[6] Yeshwanth Cherapanamjeri, Kartik Gupta, and Prateek Jain. Nearly-optimal Robust Matrix Completion. arXiv:1606.07315 [cs.LG], 2016.

[7] Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Robust Logistic Regression and Classification. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

[8] Rahul Garg and Rohit Khandekar. Gradient Descent with Sparsification: An Iterative Algorithm for Sparse Recovery with Restricted Isometry Property. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009.

[9] Prateek Jain, Ambuj Tewari, and Purushottam Kar. On Iterative Hard Thresholding Methods for High-dimensional M-estimation. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS)*, 2014.

[10] Nam H Nguyen and Trac D Tran. Exact recoverability from dense corrupted observations via ℓ_1-minimization. *IEEE transactions on information theory*, 59(4):2017–2035, 2013.

[11] Nam H Nguyen and Trac D Tran. Robust Lasso With Missing and Grossly Corrupted Observations. *IEEE Transaction on Information Theory*, 59(4):2036–2058, 2013.

[12] Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. John Wiley and Sons, 1987.

[13] Christoph Studer, Patrick Kuppinger, Graeme Pope, and Helmut Bölcskei. Recovery of Sparsely Corrupted Signals. *IEEE Transaction on Information Theory*, 58(5):3115–3130, 2012.

[14] John Wright and Yi Ma. Dense Error Correction via ℓ^1 Minimization. *IEEE Transactions on Information Theory*, 56(7):3540–3560, 2010.

[15] John Wright, Alan Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.