High Dimensional Robust Sparse Regression

Liu Liu  Yanyao Shen  Tianyang Li  Constantine Caramanis

The University of Texas at Austin

December, 2020
Introduction

- Large-scale statistical problems: both the dimension $d$ and the sample size $n$ may be large (possibly $n \ll d$).

- Low dimensional structures in the high dimensional setting.
Large-scale statistical problems: both the dimension $d$ and the sample size $n$ may be large (possibly $n \ll d$).

Low dimensional structures in the high dimensional setting.

Many examples of this:
  - Sparse regression.
  - Low rank matrix completion.
  - Low rank + sparse matrix decomposition.
  - etc...
Motivation

Well known that most state of the art approaches for these problems are fragile.

- Typically need very light tails.

- Data must be pristine: A single corrupted sample can arbitrarily corrupt the original maximum likelihood estimation.
Problem setup: robust estimation for sparse regression

Sparse regression model:
- dimensions: $n \ll d$.
- iid Gaussian $X$.
- $y_i = x_i^T \beta^* + \xi_i$.
- noise: $\xi_i \sim \mathcal{N}(0, \sigma^2)$.
- $\beta^* \in \mathbb{R}^d$ is $k$-sparse.

Contamination model:
- we observe $z_i = (y_i, x_i)$.
- $\{z_1, \cdots, z_n\} \sim (1 - \epsilon)P + \epsilon Q$.
- $P$: sparse regression model.
- $Q$: arbitrary distribution.
- $\epsilon$: const fraction of outliers.
Problem setup: robust estimation for sparse regression

### Sparse regression model:
- dimensions: $n \ll d$.
- iid Gaussian $X$.
- $y_i = x_i^T \beta^* + \xi_i$.
- noise: $\xi_i \sim \mathcal{N}(0, \sigma^2)$.
- $\beta^* \in \mathbb{R}^d$ is $k$-sparse.

### Contamination model:
- we observe $z_i = (y_i, x_i)$.
- $\{z_1, \cdots, z_n\} \sim (1 - \epsilon)P + \epsilon Q$.
- $P$: sparse regression model.
- $Q$: arbitrary distribution.
- $\epsilon$: const fraction of outliers.
### Related work

#### Robust regression
- [Li13][BJK15][DT19]: robust regression with corruptions only in $y$.
- [KKM18] [PSBR18] [DKK$^+$18] [DKS19]: low dimensional linear regression with corruptions in $x$ and $y$, $n = \Omega(d)$ and $\epsilon = \text{const}$.  
- [CCM13] [LLC19]: robust sparse regression resilient to corruptions in $x$ and $y$, with $\epsilon = O(1/\sqrt{k})$.

#### Robust mean estimation
- [LRV16] [DKK$^+$16]: robust mean estimation with $\epsilon = \text{const}$, $n = \Omega(d)$.
- [BDLS17]: robust sparse mean estimation with $\epsilon = \text{const}$, $n = \Omega(k^2 \log(d))$. $^a$ This is based on the ellipsoid algorithm in [DKK$^+$16].

---

$^a$[DKS16]: statistical query-based l.b. of $\Omega(k^2)$ on rob. sparse mean estimation.
Estimation tasks for robust sparse regression

Problem: $\epsilon$-corrupted samples from robust sparse regression model, can we recover $\beta^*$?

- [CCM13]: corruptions in $x$ and $y$, but cannot deal with constant $\epsilon$.
- [Gao17, LM16, LL$^+$20] show the minimax rate $O(\epsilon \sigma)$, but only provides exponential-time algorithm.
- [BDLS17] has sub-optimal rates depending on $\|\beta^*\|_2$.
- [KKM18] [PSBR18] [DKK$^+$18] [DKS19]: recent advances in robust regression, but require at least $n = \Omega(d)$.
- [Li13][BJK15][DT19]: corruptions only in $y$. 
Our approach

Algorithmic idea:

Iterative Hard Thresholding

+ 

Robust Sparse Mean Estimation on gradients
Our approach

Algorithmic idea:

Iterative Hard Thresholding

+  

Robust Sparse Mean Estimation on gradients

Required ingredients:

- Robust Sparse Mean Estimation
- Stability of IHT
This work

- Meta-Theorem: stability of IHT. Given any Robust Sparse Mean Estimation sub-procedure, IHT has controlled error.

- We provide order-wise faster Robust Sparse Mean Estimation algorithm based on filtering, which is scalable and practical.

- With the ellipsoid algorithm, we have optimal rate of convergence.

- With the faster filtering algorithm, we can deal with unknown but sparse covariance matrix. Exact recovery when $\epsilon$ or $\sigma$ goes to zero.
Iterative Hard Thresholding

We look at the gradient part of uncorrupted IHT:

$$\beta^{t+1} = P_k(\beta^t - \frac{1}{n} \sum_{i=1}^{n} g_i^t),$$

where $g_i^t = x_i(x_i^T \beta^t - y_i)$ is gradient of the $i^{th}$ sample $(y_i, x_i)$. 
Iterative Hard Thresholding

We look at the gradient part of uncorrupted IHT:

\[ \beta^{t+1} = P_k(\beta^t - \frac{1}{n} \sum_{i=1}^{n} g_i^t), \]

where \( g_i^t = x_i(x_i^T \beta^t - y_i) \) is gradient of the \( i \)th sample \((y_i, x_i)\).

If \( \Sigma = I_d \), \( \mathbb{E}(g_i) \) is guaranteed to be \( 2k \)-sparse:

\[ \mathbb{E}_P(g_i^t) = \mathbb{E}_P(x_i x_i^T (\beta^t - \beta^*)) = \beta^t - \beta^* = G^t. \]
Iterative Hard Thresholding

We look at the gradient part of uncorrupted IHT:

$$\beta^{t+1} = P_k(\beta^t - \frac{1}{n} \sum_{i=1}^{n} g^t_i),$$

where $g^t_i = x_i(x^T_i \beta^t - y_i)$ is gradient of the $i^{th}$ sample $(y_i, x_i)$.

If $\Sigma = I_d$, $E(g_i)$ is guaranteed to be $2k$-sparse:

$$E_P(g^t_i) = E_P(x_i x^T_i (\beta^t - \beta^*)) = \beta^t - \beta^* = G^t.$$  

When $\{y_i, x_i\}_{i=1}^{n}$ come from $(1 - \epsilon)P + \epsilon Q$, we can use robust sparse mean estimation on $G^t$, and then use inexact IHT.
**Robust Sparse Gradient Estimator (RSGE)**

### Definition 1 (RSGE)

We call $\hat{G}(\beta)$ a $\psi(\epsilon)$-RSGE, if given $\{g_i\}_{i=1}^n$, $\hat{G}(\beta)$ guarantees

$$\| \hat{G}(\beta) - G(\beta) \|_2^2 \leq \alpha \| G(\beta) \|_2^2 + \psi(\epsilon),$$

with high probability, where $\alpha \in (0, 0.1)$ is a constant.
Robust Sparse Gradient Estimator (RSGE)

**Definition 1 (RSGE)**

We call $\hat{G}(\beta)$ a $\psi(\epsilon)$-RSGE, if given $\{g_i\}_{i=1}^n$, $\hat{G}(\beta)$ guarantees

$$\|\hat{G}(\beta) - G(\beta)\|_2^2 \leq \alpha \|G(\beta)\|_2^2 + \psi(\epsilon),$$

with high probability, where $\alpha \in (0, 0.1)$ is a constant.

**Theorem 1 (Thm. 2.1 in our paper)**

*IHT is stable. In particular, using an $\psi(\epsilon)$-RSGE as defined in Definition 1, IHT outputs $\hat{\beta}$, such that*

$$\|\hat{\beta} - \beta^*\|_2 = O\left(\sqrt{\psi(\epsilon)}\right),$$

*with high probability.*
Robust sparse regression with corrupted gradients

Algorithm 1: Robust sparse regression by RSGE

1: **Input:** Data samples \( \{y_i, x_i\}_{i=1}^{N} \), RSGE subroutine.
2: **Output:** The estimation \( \hat{\beta} \)
3: Split samples into \( T \) subsets of size \( n \).
4: Initialize with \( \beta^0 = 0 \).
5: **for** \( t = 0 \) to \( T - 1 \), **do**
6: At current \( \beta^t \), calculate all gradients:
\[
g_i^t = x_i \left( x_i^\top \beta^t - y_i \right), \ i \in [n].
\]
7: We use a RSGE to get \( \hat{G}^t \).
8: \[
\beta^{t+1} = P_k \left( \beta^t - \hat{G}^t \right).
\]
9: **end for**
10: Output the estimation \( \hat{\beta} = \beta^T \).
Theorem 2 (RSGE by ellipsoid algorithm in [BDLS17], Cor. 3.1 in our paper)

With \( n \geq \Omega\left(\frac{k^2 \log d}{\epsilon^2}\right) \), we can guarantee

\[
\| \hat{G}^t - G^t \|_2^2 = O(\epsilon^2 \| G^t \|_2^2 + \epsilon^2 \sigma^2).
\]

Theorem 3

Combining Theorem 2 with Theorem 1, we have \( \| \hat{\beta} - \beta^* \|_2 = O(\epsilon \sigma) \).
Theorem 2 (RSGE by ellipsoid algorithm in [BDLS17], Cor. 3.1 in our paper)

With $n \geq \Omega\left(\frac{k^2 \log d}{\epsilon^2}\right)$, we can guarantee

$$\|\hat{G}^t - G^t\|_2^2 = O(\epsilon^2 \|G^t\|_2^2 + \epsilon^2 \sigma^2).$$

Theorem 3

Combining Theorem 2 with Theorem 1, we have $\|\hat{\beta} - \beta^*\|_2 = O(\epsilon \sigma)$.  

- This algorithm’s time complexity is polynomial.  
- However, it cannot handle unknown covariance.
How to design RSGE?

We provide a new, faster filtering algorithm.

**Theorem 4 (RSGE by filtering algorithm, Cor. 4.1 in our paper)**

*With* \( n \geq \Omega\left(\frac{k^2 \log d}{\epsilon}\right) \), *we can guarantee*

\[
\|\hat{G}^t - G^t\|_2^2 = O(\epsilon \|G^t\|_2^2 + \epsilon \sigma^2).
\]

**Theorem 5**

*Combining Theorem 1 and Theorem 4, we have* \( \|\hat{\beta} - \beta^*\|_2 = O(\sqrt{\epsilon} \sigma) \).

- The new filtering algorithm is orderwise faster, at the expense of \( \sqrt{\epsilon} \) rather than \( \epsilon \) in the guarantee.

- This new filtering algorithm also works for unknown yet sparse covariance matrix.
Experimental results I: robust sparse mean estimation

We generate authentic samples through $g_i = x_i x_i^\top G$, where $G$ is $k$-sparse. The rescaled relative MSE: $\| \hat{G} - G \|_2^2 / (\epsilon \| G \|_2^2)$ should be independent of the parameters $\{\epsilon, k, d\}$.

(a) Rescaled relative MSE vs. sparsity. (b) Rescaled relative MSE vs. dimension.

Figure: Sample complexity $n \propto k^2 \log(d) / \epsilon$. Different curves for $\epsilon \in \{0.1, 0.15, 0.2\}$ are the average of 15 trials.
Experimental results II: robust sparse regression

We use filtering algorithm as our RSGE, and generate authentic samples $y_i = x_i^\top \beta^* + \xi_i$. As expected, the convergence is linear, and flattens out at the level of the final error.

(a) $\log(\|\beta^t - \beta^*\|_2^2)$ vs. iterates.  

(b) $\log(\|\beta^t - \beta^*\|_2^2)$ vs. iterates.

Figure: In all cases, we fix $k = 5$, $d = 500$, and choose the sample complexity $n \propto 1/\epsilon$. (2a) has fixed $\sigma^2 = 0.1$. (2b) has fixed $\epsilon = 0.1$. 

\[ \]
Experimental results III: Large scale experiments

The wall clock time vs. the sample size or the dimensionality.

![Graphs showing time vs. sample size and dimensionality]

**Figure:** In both plots, we use $\epsilon = 0.1$. In the left plot, we fix $d = 500$ and in the right plot, we fix $n = 1000$. 
Other Important Directions

- What if the gradients are not sparse? For example: for general (non-sparse, non-identity) covariance.

- Then RSGE cannot be used! It is too much to ask for

\[ \| \hat{G}(\beta) - G(\beta) \|^2_2 \]

- to be small.

- Different tools/ideas are needed. For some results along these lines, see: https://arxiv.org/abs/1901.08237
Our contribution

- Sparse regression algorithm that is resilient to a constant fraction of arbitrary outliers. Our algorithm requires $n = \Omega(k^2 \log d)$ samples.

* [Gao17]: this error rate is minimax optimal under the $\epsilon$-contamination model.
Our contribution

- Sparse regression algorithm that is resilient to a constant fraction of arbitrary outliers. Our algorithm requires $n = \Omega(k^2 \log d)$ samples.

- Meta-theorem which allows the use of any robust sparse mean estimation subroutine:
  - By ellipsoid algorithm in [BDLS17], we can recover $\beta^*$ within additive error $O(\epsilon \sigma)$.

---

*[Gao17]: this error rate is minimax optimal under the $\epsilon$-contamination model.*
Our contribution

- Sparse regression algorithm that is resilient to a constant fraction of arbitrary outliers. Our algorithm requires $n = \Omega(k^2 \log d)$ samples.

- Meta-theorem which allows the use of any robust sparse mean estimation subroutine:
  - By ellipsoid algorithm in [BDLS17], we can recover $\beta^*$ within additive error $O(\epsilon \sigma)$.

- Efficient filtering algorithm for robust sparse mean estimation.
  - By this algorithm, we can recover $\beta^*$ within additive error $O(\sqrt{\epsilon \sigma})$.
  - The filtering algorithm is practical and faster by at least $d^2$.

- In particular: exact recovery as $\sigma \to 0$.

*[Gao17]: this error rate is minimax optimal under the $\epsilon$-contamination model.*
For more information please refer to our paper

Liu Liu, Yanyao Shen, Tianyang Li, Constantine Caramanis. High Dimensional Robust Sparse Regression. 
https://arxiv.org/abs/1805.11643
References I


