Robustness and Tractability for High-Dimensional M-estimators

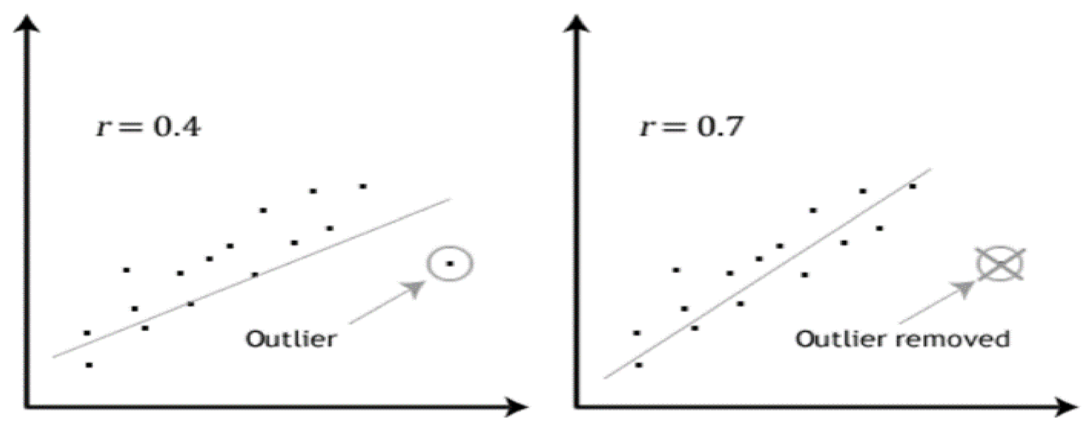
Ruizhi Zhang (rzhang320@gatech.edu), Yajun Mei, Jianjun Shi, Huan Xu
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Abstract

We investigate two important properties of M-estimator, namely robustness and tractability, in linear regression when the data are contaminated by *arbitrary outliers*.
Robustness: the statistical property that the estimator should always be close to the true parameters *regardless of the distribution of the outliers*
Tractability: the computational property that the estimator can be computed efficiently even though the objective function can be *non-convex*.
In this article, by learning the landscape of the empirical risk, we show that under the high-dimensional setting in which $p >> n$, many penalized M-estimators with L_1 regularizer enjoy nice robustness and tractability properties simultaneously when the percentage of outliers is small.

Introduction

Why we need robust regression? Find a good model for majority data, Detect outliers, etc.



Why consider M-estimators?
1. Formulation is simple but general.
2. Statistical properties are well-studied (Consistency and Asymptotic normality [3].)
3. Good robust properties (large breakdown point and bounded influence function [1].)

Our objective: Investigate the ***tractability*** of M-estimators and the relation with ***robustness***.

Model

Assume we have n pairs data $\{(y_i, x_i)\}_{i=1,2,\ldots,n}$, which are generated from the linear model with gross-error [2]:

$$y_i = \langle \theta_0, x_i \rangle + \epsilon_i, \quad \text{where } y_i \in \mathbb{R}, x_i \in \mathbb{R}^p,$$
$$\epsilon_i \sim (1-\delta) f_0 + g, \text{where } f_0 \text{ and } g \text{ denote the density for the idealized noise and outliers.}$$

Remarks:
1. $\delta \in [0,1]$ denotes the percentage of outliers.
2. f_0 has nice idealized properties: symmetric, zero mean, independent to x_i, subgaussian.
3. g may be arbitrary: could be asymmetric, nonzero mean, dependent to x_i.

M-estimators in low-dimensional case

In general, a M-estimator is obtained by solving the optimization problem:

$$\underset{\theta}{\text{Minimize:}} \quad \hat{R}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho(y_i - \langle \theta, x_i \rangle), \qquad (1)$$
$$\text{subject to:} \quad \|\theta\|_2 \le r.$$

Here $\rho : \mathbb{R} \to \mathbb{R}$ is the loss function, and often is *non-convex*.

Table 1: Some well-known loss functions for M-estimators

Type	$\rho(t)$	$\psi(t) = \rho'(t)$
Least Square	$t^2/2$	t
Tukey	$\frac{c^2}{6}\left(1-(1-(t/c)^2)^3\right), \lvert t\rvert \le c$; $c^2/6, \quad \lvert t\rvert \ge c$	$t(1-(t/c)^2)^2, \lvert t\rvert \le c$; $0, \quad \lvert t\rvert > c$
Welsch	$\frac{1-\exp(-\alpha t^2/2)}{\alpha}$	$t\exp(-\alpha t^2/2)$

Theoretical result

We define the score function $\psi(z) := \rho'(z)$.

Assumption 1(a) *The score function $\psi(z)$ is twice differentiable and odd in z with $\psi(z) \ge 0$ for all $z \ge 0$. Moreover, we assume $\max\{\|\psi(z)\|_\infty, \|\psi'(z)\|_\infty, \|\psi''(z)\|_\infty\} \le L_\psi$.*

(b) *The feature vector x_i are i.i.d with zero mean and τ^2-sub-Gaussian, that is $\mathbf{E}[e^{\langle \lambda, x_i \rangle}] \le \exp(\frac{1}{2}\tau^2\|\lambda\|_2^2)$ for all $\lambda \in \mathbb{R}^p$.*

(c) *The feature vector x_i spans all direction in $\mathbb{R}^p$, that is $\mathbf{E}[x_i x_i^T] \succeq \gamma\tau^2 I_{p\times p}$ for some $0 < \gamma < 1$.*

(d) *The idealized noise distribution $f_0(\epsilon)$ is symmetric and decreasing for $\epsilon > 0$.*

Theorem 1
Assume assumption 1 holds and $\|\theta_0\|_2 \le r/3$. There exists constants $\eta_0 = \frac{\delta}{1-\delta}C_1$ and $\eta_1 = C_2 - C_3\delta > 0$, such that for any $\pi > 0$, there exist constant C_π depends on $\pi, \gamma, r, \tau, \psi, f_0$ but independent of n, p, δ and g, such that as $n \ge C_\pi p \log n$, the following statements hold with probability at least $1-\pi$:

(a) *For all $\|\theta - \theta_0\|_2 > 2\eta_0$,*

$$\langle \theta - \theta_0, \nabla \widehat{R}_n(\theta) \rangle > 0. \qquad (2)$$

There is no stationary point of $\widehat{R}_n(\theta)$ outside of the ball $B^p(\theta_0, 2\eta_0)$.

(b) *For all $\|\theta - \theta_0\|_2 \le \eta_1$,*

$$\lambda_{\min}(\nabla^2 \widehat{R}_n(\theta)) > 0. \qquad (3)$$

$\widehat{R}_n(\theta)$ is strong convex in the ball $B^p(\theta, \eta_1)$

Thus, as long as $2\eta_0 < \eta_1$, $\widehat{R}_n(\theta)$ has a unique stationary point, which lies in the ball $B^p(\theta_0, 2\eta_0)$. This is the unique global optimal solution of (1), and denote this unique stationary point by $\widehat{\theta}_n$.

(c) *There exists a positive constant κ that depends on $\pi, \gamma, r, \psi, \delta, f_0$ but independent of n, p and g, such that*

$$\|\widehat{\theta}_n - \theta_0\|_2 \le \eta_0 + \frac{4\tau}{\kappa}\sqrt{\frac{C_\pi p \log n}{n}}. \qquad (4)$$

Penalized M-estimators in high-dimensional case

We consider the case when $p >> n$ and the support of θ_0 is sparse. We consider the penalized M-estimators by solving the optimization problem [4]:

$$\underset{\theta}{\text{Minimize:}} \quad \hat{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho(y_i - \langle \theta, x_i \rangle) + \lambda_n \|\theta\|_1, \qquad (5)$$
$$\text{subject to:} \quad \|\theta\|_2 \le r.$$

Assumption 2
The feature vector x is bounded, i.e., there exists constant $M > 1$ that is independent of dimension p such that $\|x\|_\infty \le M\tau$ almost sure.

Theorem 2
Assume that Assumption 1 and Assumption 2 hold and the true parameter θ_0 satisfies $\|\theta_0\|_2 \le r/3$ and $\|\theta_0\|_0 \le s_0$. Then there exist constants such C, C_0, C_1, C_2 that are dependent on $(\rho(), L_\psi, \tau^2, r, \gamma, \pi)$ but independent on (δ, s_0, n, p, M) such that as $n \ge C s_0 \log p$ and $\lambda_n = C_0 M\sqrt{\frac{\log p}{n}} + \delta\frac{C_1}{\sqrt{s_0}}$, the following hold with probability as least π :

(a) *Any stationary points of problem (5) is in $B_2^p(\theta_0, \eta_0 + \frac{\sqrt{s_0}}{1-\delta}\lambda_n C_2)$*

(b) *As long as n is large enough such that $n \ge C s_0 \log^2 p$ and the contamination ratio δ is smaller such that $(\eta_0 + \frac{1}{1-\delta}\sqrt{s_0}\lambda_n C_2) \le \eta_1$, the problem (5) has a unique local stationary point which is also the global minimizer.*

Remarks:
When $\delta = 0$, we have $\eta_0 = 0$ and $\eta_1 = C > 0$. Thus, by setting $\lambda_n = O(\sqrt{\frac{\log p}{n}})$, if $s_0 = o(\frac{n}{\log p})$, there is a unique stationary point of (5).

Illustration of our theoretical results

Based on our theorems, the two values $\eta_0 = \frac{\delta}{1-\delta}C_1$ and $\eta_1 = C_2 - C_3\delta > 0$ are important. For the penalized M-estimator for the high-dimensional case, we further define a constant r_s and a cone $\mathbb{A}$ by

$$r_s = \eta_0 + \frac{\sqrt{s_0}}{1-\delta}\lambda_n C_2 \qquad (6)$$
$$\mathbb{A} = \{\theta_0 + \Delta : \|\Delta_{S_0^c}\|_1 \le 3\|\Delta_{S_0}\|_1\}. \qquad (7)$$

Then we can illustrate our theoretical results by the following two figures.

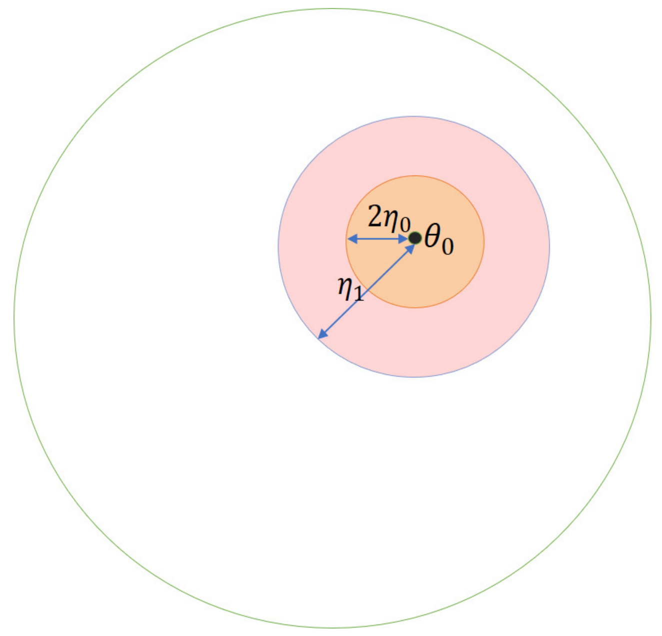


Figure 1: $\widehat{R}_n(\theta)$ in Low-dimensional case

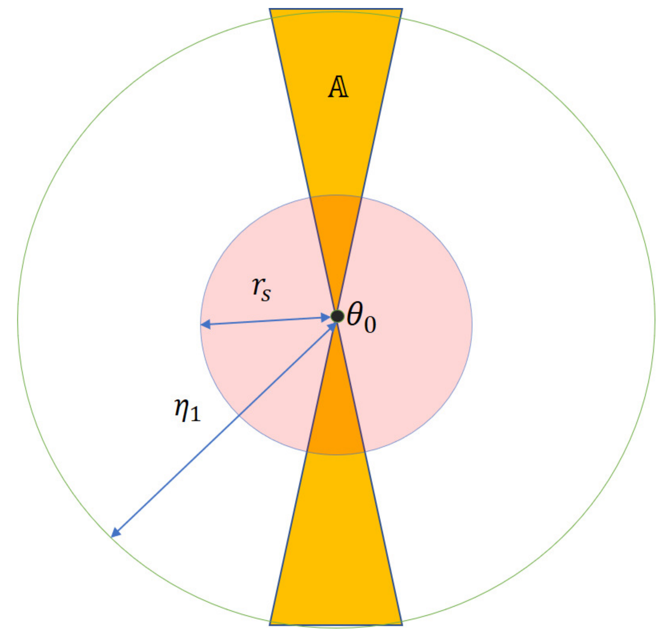


Figure 2: $\widehat{L}_n(\theta)$ in high-dimensional case

Simulation results

Settings:
$x_i \sim N(0, I_{p\times p})$ and responses $y_i = \langle \theta_0, x_i \rangle + \epsilon_i$, where $\|\theta_0\|_2 = 1$.
$\epsilon_i \sim (1-\delta)N(0,1) + \delta N(\|x_i\|_2^2 + 1, 3^2)$.
$r = 10, p = 10, n = 200$
Loss: $\rho_\alpha(t) = \frac{1-\exp(-\alpha t^2/2)}{\alpha}$ (Welsch's)
Algorithm: gradient descent with 20 random initial points.

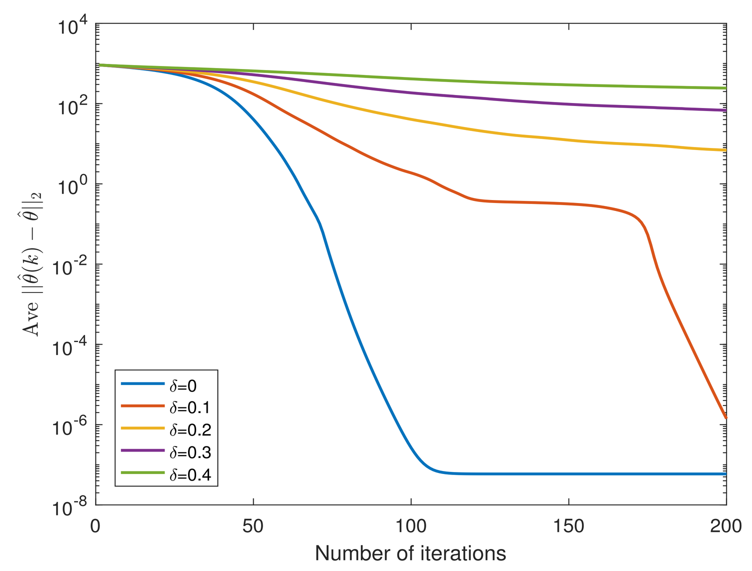


Figure 3: The convergence of gradient descent algorithm for different δ. y-axis is with log scale.

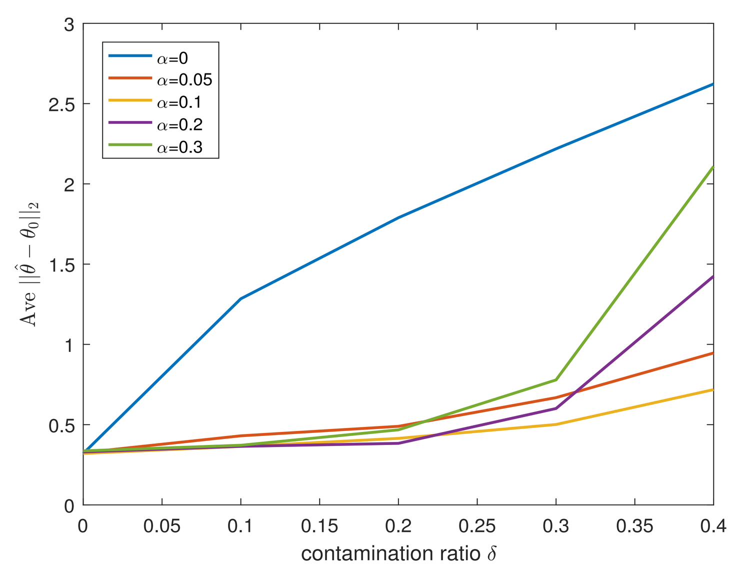


Figure 4: The estimation error for different α and δ

Reference

[1] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust statistics: the approach based on influence functions*, John Wiley & Sons, 2011.

[2] P. J. Huber, *Robust estimation of a location parameter*, The annals of mathematical statistics, 35 (1964), pp. 73–101.

[3] P.-L. Loh, *Statistical consistency and asymptotic normality for high-dimensional robust m-estimators*, The Annals of Statistics, 45 (2017), pp. 866–896.

[4] S. Mei, Y. Bai, and A. Montanari, *The landscape of empirical risk for non-convex losses*, arXiv preprint arXiv:1607.06534, (2016).