# Estimation of High-Dimensional Mean Regression in Absence of Symmetry and Light-tail Assumptions

Yuyan Wang

joint work with Jianqing Fan, Quefeng Li

Princeton University

Aug 10, 2015

# Overview

1. Introduction & Motivation
2. RA-Lasso estimator
   - Optimal Statistical Error
   - Geometric Convergence of Optimization Error
3. Numerical Studies
4. Discussion & Future Work

# Overview


1. Introduction & Motivation
2. RA-Lasso estimator
   - Optimal Statistical Error
   - Geometric Convergence of Optimization Error
3. Numerical Studies
4. Discussion & Future Work

# Problems Arising from High-dimensional Data

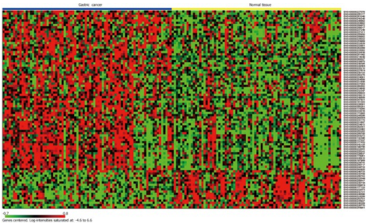


Figure 1 : Microarrays

# Problems Arising from High-dimensional Data

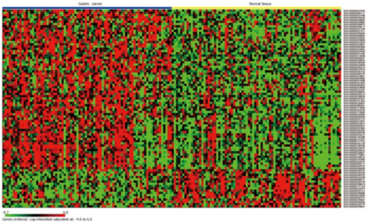


Figure 1 : Microarrays

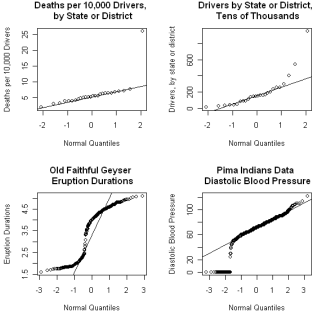



Figure 2 : Asymmetric & Heavy-tailed Data

- High-dimensionality: $p \gg n$
- Abnormal tails: asymmetric and heavy-tailed

# Motivation: Heavy-tailed and asymmetric data

$\mathbf{E}[Y|X]$?

Linear regression in a high-dimensional setting (Large $n$, large $p$, $p \gg n$):

- $L_2$-loss + Penalty: Lasso [Tibshirani, 1996], SCAD [Fan and Li, 2001], MCP [Zhang, 2010]
- need light-tail assumptions

# Motivation: Heavy-tailed and asymmetric data

$\mathbf{E}[Y|X]$?

Linear regression in a high-dimensional setting (Large $n$, large $p$, $p \gg n$):

- $L_2$-loss + Penalty: Lasso [Tibshirani, 1996], SCAD [Fan and Li, 2001], MCP [Zhang, 2010]
- need light-tail assumptions

Robust methods for heavy-tailed data:

- robust loss: $L_1$-loss, Huber loss [Huber, 1964], Catoni loss [Catoni, 2012] etc.
- LAD [Wang, 2013]; AR-Lasso [Fan, Fan and Barut, 2014]
- need symmetry assumptions

# Motivation: Heavy-tailed and asymmetric data

$\mathbf{E}[Y|X]$?

Linear regression in a high-dimensional setting (Large $n$, large $p$, $p \gg n$):

- $L_2$-loss + Penalty: Lasso [Tibshirani, 1996], SCAD [Fan and Li, 2001], MCP [Zhang, 2010]
- need light-tail assumptions

Robust methods for heavy-tailed data:

- robust loss: $L_1$-loss, Huber loss [Huber, 1964], Catoni loss [Catoni, 2012] etc.
- LAD [Wang, 2013]; AR-Lasso [Fan, Fan and Barut, 2014]
- need symmetry assumptions

Heavy-tailed **and** asymmetric? Robustly estimate **mean**?

# Overview

1. Introduction & Motivation
2. RA-Lasso estimator
   - Optimal Statistical Error
   - Geometric Convergence of Optimization Error
3. Numerical Studies
4. Discussion & Future Work

# Model Setup

We consider the linear regression model

$$y_i = \boldsymbol{x}_i \beta^* + \epsilon_i, \ i = 1, \ldots, n \tag{1}$$

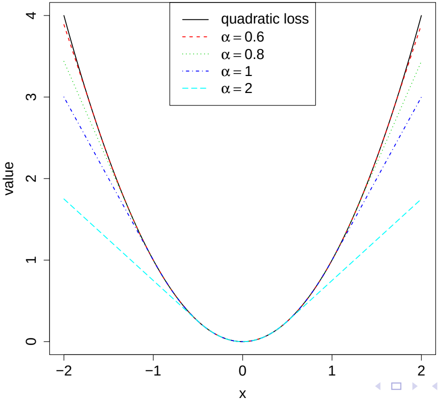
- $\{\boldsymbol{x}_i\}_{i=1}^n$ i.i.d $\mathbb{R}^p$, $\mathrm{E}(\boldsymbol{x}_i) = \mathbf{0}$;
- $\{\epsilon_i\}_{i=1}^n$ i.i.d $\mathrm{E}(\epsilon_i) = 0$;
- $p \gg n$, $\log(p) = o(n)$
- $\sum_{j=1}^p \|\beta_j^*\|_1^p \leq R_q, q \in [0, 1)$

Goal: Estimate the mean effect of $y$ conditioning on $\boldsymbol{x}$, which is $\beta^*$.

# Robust Surrogate Loss: Huber Loss with varying parameter

$$\ell_\alpha(x) = \begin{cases} 2\alpha^{-1}|x| - \alpha^{-2} & \text{if } |x| > \alpha^{-1}; \\ x^2 & \text{if } |x| \le \alpha^{-1}. \end{cases}$$

$$\ell_\alpha(x) \to x^2 \text{ as } \alpha \to 0 \text{ and } \ell_\alpha(x) \to |x| \text{ as } \alpha \to \infty.$$



# Our proposed robust estimator: RA-Lasso

We propose the **RA-Lasso** estimator:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \; \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_\alpha(y_i - \boldsymbol{x}_i^T\beta)}_{\text{Huber loss}} + \lambda_n \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{penalty}}. \tag{2}$$

- $\hat{\beta}$ is an estimator of $\beta_\alpha^* = \text{argmin}_{\beta}\, \mathrm{E}\ell_\alpha(y - \boldsymbol{x}^T\beta)$ for any fixed $\alpha$.

# Our proposed robust estimator: RA-Lasso

We propose the **RA-Lasso** estimator:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \; \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_\alpha(y_i - \mathbf{x}_i^T \beta)}_{\text{Huber loss}} + \lambda_n \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{penalty}}. \tag{2}$$

- $\hat{\beta}$ is an estimator of $\beta_\alpha^* = \operatorname{argmin}_\beta \mathrm{E}\ell_\alpha(y - \mathbf{x}^T\beta)$ for any fixed $\alpha$.
- We are able to show: $\beta_\alpha^* \to \beta^*$ as $\alpha \to 0$.
- By triangular inequality:

$$\underbrace{\|\hat{\beta} - \beta^*\|_2}_{\text{statistical error}} \leq \underbrace{\|\beta_\alpha^* - \beta^*\|_2}_{\text{approximation error}} + \underbrace{\|\hat{\beta} - \beta_\alpha^*\|_2}_{\text{estimation error}}.$$

# RA-Lasso: Approximation Error

$$\underbrace{\|\hat{\beta} - \beta^*\|_2}_{\text{statistical error}} \leq \underbrace{\|\beta^*_\alpha - \beta^*\|_2}_{\text{approximation error}} + \underbrace{\|\hat{\beta} - \beta^*_\alpha\|_2}_{\text{estimation error}}.$$

# RA-Lasso: Approximation Error

$$\underbrace{\|\hat{\beta} - \beta^*\|_2}_{\text{statistical error}} \leq \underbrace{\|\beta^*_\alpha - \beta^*\|_2}_{\text{approximation error}} + \underbrace{\|\hat{\beta} - \beta^*_\alpha\|_2}_{\text{estimation error}}.$$

## Theorem 1 (Approximation Error)

*Suppose*
*(C1)* $\mathrm{E}[\mathrm{E}(|\epsilon|^k|\boldsymbol{x})] \leq M_k < \infty$, *for some* $k \geq 2$,
*it holds that*

$$\|\beta^*_\alpha - \beta^*\|_2 = O(\alpha^{k-1}).$$

# RA-Lasso: Estimation Error

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \underbrace{\|\boldsymbol{\beta}_\alpha^* - \boldsymbol{\beta}^*\|_2}_{\text{approximation error}} + \underbrace{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2}_{\text{estimation error}},$$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \quad \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta})}_{\mathcal{L}_n(\boldsymbol{\beta})} + \lambda_n \|\boldsymbol{\beta}\|_1,$$

$$\boldsymbol{\beta}_\alpha^* = \text{argmin}\, \mathrm{E}\ell_\alpha(y - \mathbf{x}'\boldsymbol{\beta}).$$

- Estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_\alpha^*\|_2$:
  $L_2$-error of a high-dim regularized convex $M$-estimator

# RA-Lasso: Estimation Error

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le \underbrace{\|\boldsymbol{\beta}^*_\alpha - \boldsymbol{\beta}^*\|_2}_{\text{approximation error}} + \underbrace{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha\|_2}_{\text{estimation error}},$$

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta})}_{\mathcal{L}_n(\boldsymbol{\beta})} + \lambda_n\|\boldsymbol{\beta}\|_1,$$

$$\boldsymbol{\beta}^*_\alpha = \operatorname{argmin} \mathrm{E}\ell_\alpha(y - \mathbf{x}'\boldsymbol{\beta}).$$

- Estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_\alpha\|_2$:
  $L_2$-error of a high-dim regularized convex $M$-estimator
- Restricted Strong Convexity (RSC) [Negahban, et al., 2012]:

$$\delta\mathcal{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}^*_\alpha) \ge \kappa_{\mathcal{L}}\|\boldsymbol{\Delta}\|_2^2 - \tau_{\mathcal{L}}^2, \text{ for all } \boldsymbol{\Delta} \in \mathbb{C}_\alpha.$$

where $\delta\mathcal{L}_n(\boldsymbol{\Delta}, \boldsymbol{\beta}^*_\alpha) = \mathcal{L}_n(\boldsymbol{\beta}^*_\alpha + \boldsymbol{\Delta}) - \mathcal{L}_n(\boldsymbol{\beta}^*_\alpha) - [\nabla\mathcal{L}_n(\boldsymbol{\beta}^*_\alpha)]^T\boldsymbol{\Delta}$.

# Main Result

**Theorem 2 (Estimation Error)**

*By choosing* $\lambda_n = O(\sqrt{\frac{\log p}{n}})$ *and* $\alpha \geq c\lambda_n$,

$$\|\hat{\beta} - \beta^*_\alpha\|_2 = O(\sqrt{R_q}[(\log p)/n]^{1/2-q/4}).$$

$$\underbrace{\|\hat{\beta} - \beta^*\|_2}_{\text{statistical error}} \leq \underbrace{\|\beta^*_\alpha - \beta^*\|_2}_{\text{approximation error}} + \underbrace{\|\hat{\beta} - \beta^*_\alpha\|_2}_{\text{estimation error}}.$$

**Theorem 3 (Statistical Error)**

$$\|\hat{\beta} - \beta^*\|_2 = O(\alpha^{k-1}) + O(\sqrt{R_q}[(\log p)/n]^{1/2-q/4}).$$

# Overview


1. Introduction & Motivation
2. RA-Lasso estimator
   - Optimal Statistical Error
   - Geometric Convergence of Optimization Error
3. Numerical Studies
4. Discussion & Future Work

# Computational Error

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \; \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_\alpha(y_i - \boldsymbol{x}_i^T \beta)}_{\mathcal{L}_n(\beta)} + \lambda_n \|\beta\|_1.$$

The gradient descent algorithm to solve the problem: At the $t$-th iteration,

$$\hat{\beta}^{t+1} = \underset{\|\beta\|_1 \le \rho}{\text{argmin}} \; \underbrace{\mathcal{L}_n(\hat{\beta}^t) + [\nabla \mathcal{L}_n(\hat{\beta}^t)]^T(\beta - \hat{\beta}^t) + \frac{\gamma_u}{2}\|\beta - \hat{\beta}^t\|_2^2}_{\text{local quadratic approximation}} + \lambda_n \|\beta\|_1,$$

- Optimization error: $\hat{\beta}^t - \hat{\beta}$

# Geometric convergence of $\hat{\beta}^t - \hat{\beta}$

**Theorem 4**

*We have*

$$\|\hat{\beta}^t - \hat{\beta}\|_2^2 = O\left( \underbrace{R_q \left(\frac{\log p}{n}\right)^{1-(q/2)}}_{o(1)} \left[ \|\hat{\beta} - \beta^*_\alpha\|_2^2 + R_q \left(\frac{\log p}{n}\right)^{1-(q/2)} \right] \right),$$

*w.h.p. after sufficient iterations.*

$$\|\hat{\beta}^t - \beta^*\|_2 \leq \underbrace{\|\hat{\beta}^t - \hat{\beta}\|_2}_{\text{computational error}} + \underbrace{\|\hat{\beta} - \beta^*_\alpha\|_2}_{\text{estimation error}} + \underbrace{\|\beta^*_\alpha - \beta^*\|_2}_{\text{approximation error}}$$

$$= O(\sqrt{R_q}[(\log p)/n]^{1/2-q/4}) \Rightarrow \hat{\beta}^t \text{ is as good as } \hat{\beta}.$$

# Overview


1. Introduction & Motivation
2. RA-Lasso estimator
   - Optimal Statistical Error
   - Geometric Convergence of Optimization Error
3. Numerical Studies
4. Discussion & Future Work

# Simulation Setup

- $y_i = \boldsymbol{x}_i^T \beta^* + \epsilon_i$,
  $\boldsymbol{x}_i \sim N(0, I_p)$, $\epsilon_i = c^{-1}(\boldsymbol{x}_i^T \beta^*)^2 \tilde{\epsilon}_i,$, $i = 1, \ldots, n$
- $n = 100$, $p = 400$, $\beta^* = (\underbrace{3, \ldots, 3}_{20}, 0, \ldots, 0)^T$.
- 5 scenarios of noise distributions

| | **Light Tail** | **Heavy Tail** |
|---|---|---|
| **Symmetric** | $N(0, 1)$ | $2t_3$ |
| **Asymmetric** | MixN | LogNor, Weibull |

Table 1 : categorical summary of the 5 scenarios

- Performance measures: $\|\beta^* - \hat{\beta}\|_2$, $\|\beta^* - \hat{\beta}\|_1$
- Compared with: (1) Lasso: $L_2$-loss + $L_1$-pen;
  (2) LAD: $L_1$-loss + $L_1$-pen.

# Simulation Results

| | Light Tail | Heavy Tail |
|---|---|---|
| **Symmetric** | $N(0,1)$ | $2t_3$ |
| **Asymmetric** | MixN | LogNor, Weibull |

Table 2 : Noise distributions.

| | | Lasso | LAD | RA-Lasso |
|---|---|---|---|---|
| $\mathbf{N(0,1)}$ | $L_2$ loss | 4.60 | **4.34** | 4.60 |
| | $L_1$ loss | 27.16 | **27.14** | 27.15 |
| **$2t_3$** | $L_2$ loss | 8.08 | 6.71 | **6.70** |
| | $L_1$ loss | 41.16 | 42.76 | **38.52** |
| **MixN** | $L_2$ loss | 6.26 | 6.54 | **6.25** |
| | $L_1$ loss | 41.26 | 46.95 | **39.25** |
| **LogNor** | $L_2$ loss | 10.86 | 9.19 | **8.48** |
| | $L_1$ loss | 57.52 | 57.18 | **53.20** |
| **Weibull** | $L_2$ loss | 7.40 | 8.81 | **5.53** |
| | $L_1$ loss | 40.95 | 47.82 | **34.65** |

Table 3 : Simulation results.

# Overview


1 Introduction & Motivation

2 RA-Lasso estimator
- Optimal Statistical Error
- Geometric Convergence of Optimization Error

3 Numerical Studies

4 Discussion & Future Work

# Discussion & Future Work

Our achievements:

- RA-Lasso which estimates *mean* and allows *asymmetry and heavy-tails*;
- Optimal rate of RA-Lasso;
- A computational solution of RA-Lasso that achieves the same optimal rate.

Future work:

- A family of robust loss functions
- RA-Lasso in the estimation of large covariance matrices

***Thank you!***