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

Sep 8, 2015

Overview

Overview

Problems Arising from High-dimensional Data

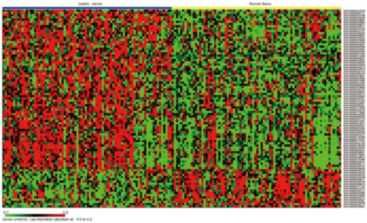


Figure 1: Microarrays

Problems Arising from High-dimensional Data

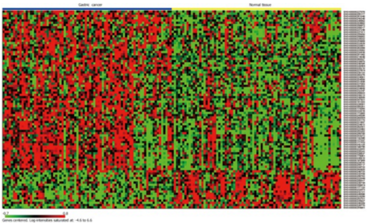


Figure 1: Microarrays

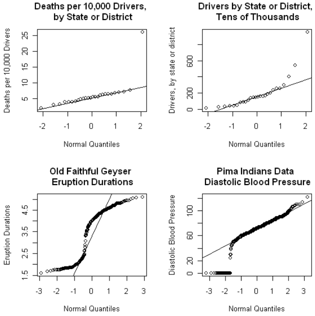


Figure 2: Asymmetric & Heavy-tailed Data

- 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

Model Setup

We consider the linear regression model

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

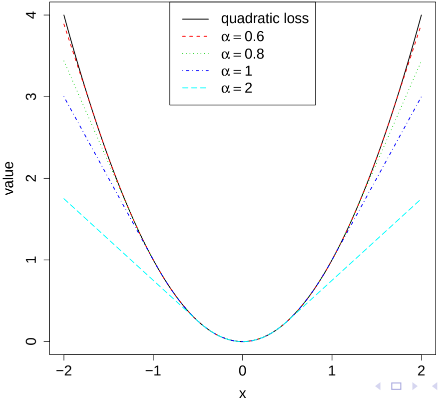
- $\{\boldsymbol{x}_i\}_{i=1}^n$ i.i.d $\mathbb{R}^p$, $\mathrm{E}(\boldsymbol{x}_i) = \boldsymbol{0}$;
- $\{\epsilon_i\}_{i=1}^n$ i.i.d $\mathrm{E}(\epsilon_i) = 0$;
- $p \gg n$, $\log(p) = o(n)$
- $\sum_{j=1}^p \|\beta_j^*\|_1^p \leq R_q, q \in [0, 1)$

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

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:

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

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

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 α.
- We are able to show: $\beta^*_\alpha \to \beta^*$ as $\alpha \to 0$.
- By triangular inequality:

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

RA-Lasso: Approximation Error

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

RA-Lasso: Approximation Error

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

Theorem 1 (Approximation Error)

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

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

RA-Lasso: Estimation Error

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

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

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

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

RA-Lasso: Estimation Error

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

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

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

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

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

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

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

Computational Error

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

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

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

- 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

Robust Estimation of Mean

Under the 1-dim scenario,

$$y_i = \mu + \epsilon_i, \ i = 1, \ldots, n \tag{3}$$

- Estimate μ using the sample mean $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i$? We can do better!

Robust Estimation of Mean

Under the 1-dim scenario,

$$y_i = \mu + \epsilon_i,\ i = 1, \ldots, n \tag{3}$$

- Estimate μ using the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$? We can do better!
- The RA-mean estimator $\hat{\mu}_\alpha$ of μ is the solution of

$$\sum_{i=1}^{n} \psi[\alpha(y_i - \mu)] = 0, \tag{4}$$

where $\psi(x)$ is the **influence function** ("derivative") of Huber loss.

Robust Estimation of Mean

Under the 1-dim scenario,

$$y_i = \mu + \epsilon_i, \; i = 1, \ldots, n \tag{3}$$

- Estimate μ using the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$? We can do better!
- The RA-mean estimator $\hat{\mu}_\alpha$ of μ is the solution of

$$\sum_{i=1}^{n} \psi[\alpha(y_i - \mu)] = 0, \tag{4}$$

  where $\psi(x)$ is the **influence function** ("derivative") of Huber loss.
- We claim: $\hat{\mu}_\alpha$ is **better** than $\hat{\mu}$!

Robust Estimation of Mean

Theorem 5 (Exponential Type of Concentration of $\hat{\mu}_\alpha$)

Assume $var(y_i) = \sigma^2 < \infty$. *Then,*

$$P\left(|\hat{\mu}_\alpha - \mu| \geq t\right) \leq 2\exp(-\frac{nt^2}{16\sigma^2}).$$

Robust Estimation of Mean

Theorem 5 (Exponential Type of Concentration of $\hat{\mu}_\alpha$)

Assume $var(y_i) = \sigma^2 < \infty$. Then,

$$P\left(|\hat{\mu}_\alpha - \mu| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{16\sigma^2}\right).$$

Remark

- $\hat{\mu}_\alpha$: *fast convergence with* ***only*** *2nd moment assumption*
  $\Longrightarrow$ *can deal with heavy-tail and asymmetry;*
- $\hat{\mu}$: *needs sub-Gaussian assumption for fast convergence*
  $\Longrightarrow$ *requires data to be light-tailed*

Robust Estimation of Covariance Matrices

- Observe $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ i.i.d $\sim \boldsymbol{X} \in \mathbb{R}^p, \mathbb{E}(\boldsymbol{X}) = \mathbf{0}$
- Goal: Estimate $\Sigma = \mathrm{cov}(\boldsymbol{X})$
- Sample Cov: $\hat{\Sigma} = (\hat{\sigma}_{ij})$, where $\hat{\sigma}_{ij} = \frac{1}{n}\sum_{k=1}^{n} X_{ki}X_{kj}$

  requires sub-Gaussianity of $\boldsymbol{X}$ for uniform convergence of $\hat{\Sigma}$ to Σ.

Robust Estimation of Covariance Matrices

- Observe $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ i.i.d $\sim \boldsymbol{X} \in \mathbb{R}^p, \mathbb{E}(\boldsymbol{X}) = \mathbf{0}$
- Goal: Estimate $\Sigma = \text{cov}(\boldsymbol{X})$
- Sample Cov: $\hat{\Sigma} = (\hat{\sigma}_{ij})$, where $\hat{\sigma}_{ij} = \frac{1}{n} \sum_{k=1}^{n} X_{ki} X_{kj}$

  requires sub-Gaussianity of $\boldsymbol{X}$ for uniform convergence of $\hat{\Sigma}$ to Σ.
- RA-covariance estimator: $\hat{\Sigma}^{(\alpha)} = (\hat{\sigma}_{ij}^{(\alpha)})$ where $\hat{\sigma}_{ij}^{(\alpha)}$ is the solution of

$$\sum_{k=1}^{n} \psi[\alpha(X_{ki} X_{kj} - \sigma_{ij})] = 0,$$

  only requires $\mathbb{E}(X_j^4) < \infty$.

Overview

Simulation Setup

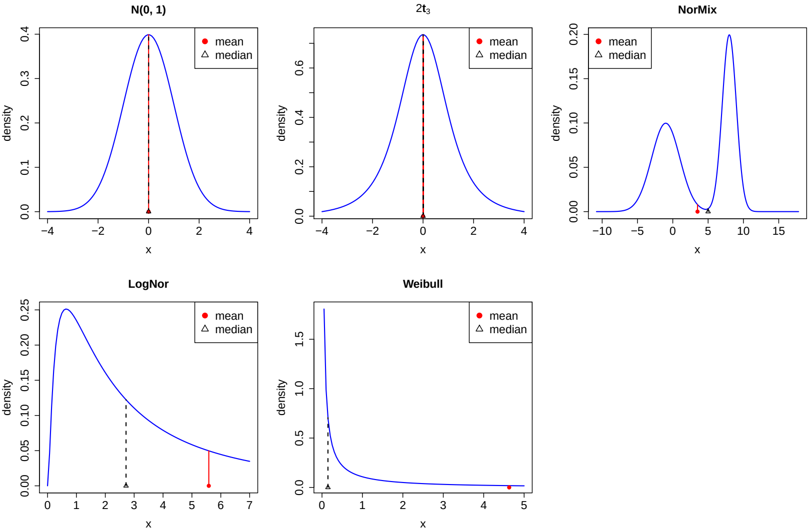
- $y_i = \boldsymbol{x}_i^T\beta^* + \epsilon_i$,
  $\boldsymbol{x}_i \sim N(0, I_p)$, $\epsilon_i = c^{-1}(\boldsymbol{x}_i^T\beta^*)^2\tilde{\epsilon}_i$,, $i = 1, \ldots, n$
- $n = 100$, $p = 400$, $\beta^* = (\underbrace{3, \ldots, 3}_{20}, 0, \ldots, 0)^T$.
- 5 scenarios of noise distributions

	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.

Error Distributions



Simulation Results

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

Table 2: Noise distributions.

		Lasso	LAD	RA-Lasso
N(0, 1)	L_2 loss	4.60	**4.34**	4.60
	L_1 loss	27.16	**27.14**	27.15
2t3	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.

Real data example

- A microarray data for the study of the reaction of innate immune system in face of atherosclerosis (Huang et al., 2011).
- The "TLR8" gene under the Toll-like Receptor (TLR) signaling pathway was found to be a key atherosclerosis-associated gene in the original study.
- We regressed "TLR8" gene on another 464 genes from 12 pathways closely related to TLR pathway.
- $n = 119$ patients were involved.

Real data example

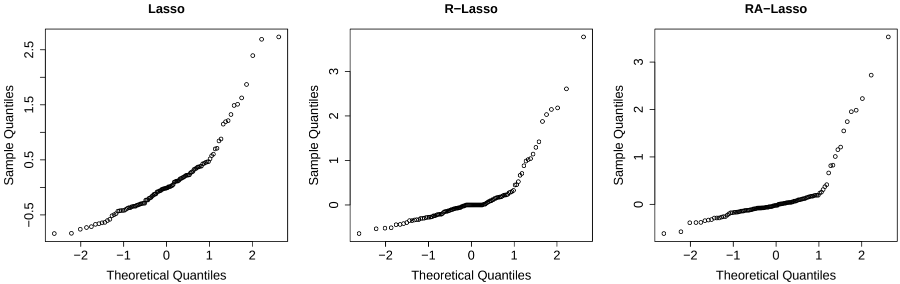


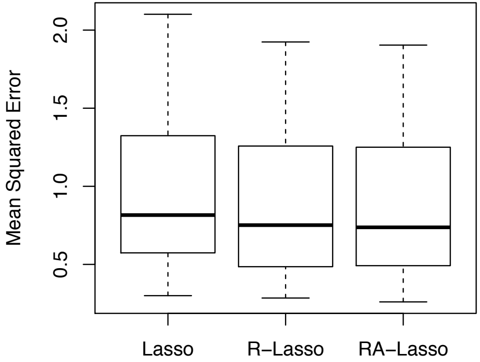
Figure 3: QQ plots of the residuals from three methods.

Real data example

Lasso	CRK						
LAD	CSF3	IL10	AKT1	KPNB1	TLR2	GRB2	MAPK1
	DAPK2	TOLLIP	TLR1	TLR3	SHC1	PSMD1	F12
	EPOR	TJP1	GAB2				
Our	CSF3	CD3E	BTK	CLSPN	RELA	AKT1	IRS2
	IL10	MAP2K4	PMAIP1	BCL2L11	AKT3	DUSP10	IRF4
	IFI6	TLR1	PSMB8	KPNB1	IFNG	FADD	TJP1
	CR2	IL2	PSMC2	HSPA8	SHC1	SPI1	IFNA6
	FYN	EPOR	MASP1	PRKCZ	TOLLIP	BAK1	

Table 4: Selected genes by three methods

Real data example



- Randomly chose 20 subjects as the test set;
- Apply three methods to the rest subjects to obtain the estimated coefficients $\hat{\beta}$;
- Apply $\hat{\beta}$ to the test set to calculate the MSE;
- Repeat random sampling 100 times.

Overview

Discussion

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.
- Robust estimators of mean and covariance matrices.
- Satisfactory finite sample performance of RA-Lasso

Thank you!