Robust Approximate Lasso for High-Dimensional Regression

Yuyan Wang

Joint work with Jianqing Fan, Quefeng Li
Princeton University

Outline

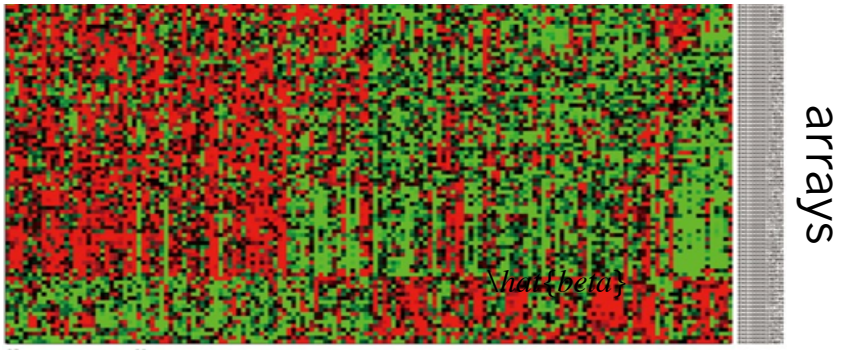
- Motivation: regression in high-D with outliers
- Existing efforts: Lasso, LAD
- RA-Lasso
- Numerical results
- Theoretical justification
- My other projects

Outline

- Motivation: regression in high-D with outliers
- Existing efforts: Lasso, LAD
- RA-Lasso
- Numerical results
- Theoretical justification
- My other projects

Challenges arising from Big Data

- High-dimensionality: p >> n

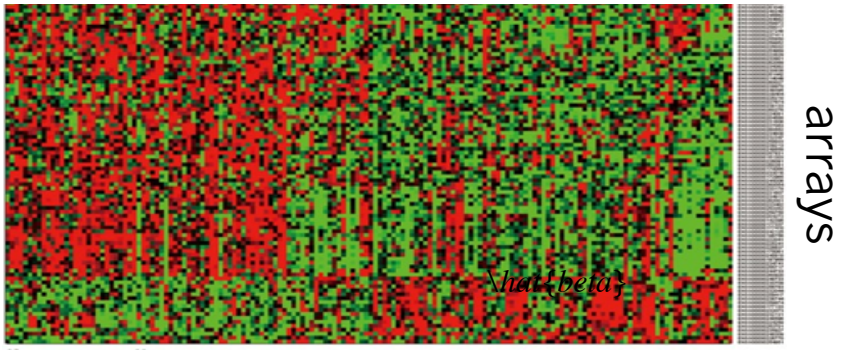


- Measure gene expression. Which genes are significant?

Often tens of thousands of genes (features); only tens of hundreds of samples.

Challenges arising from Big Data

- High-dimensionality: p >> n



- Measure gene expression. Which genes are significant?

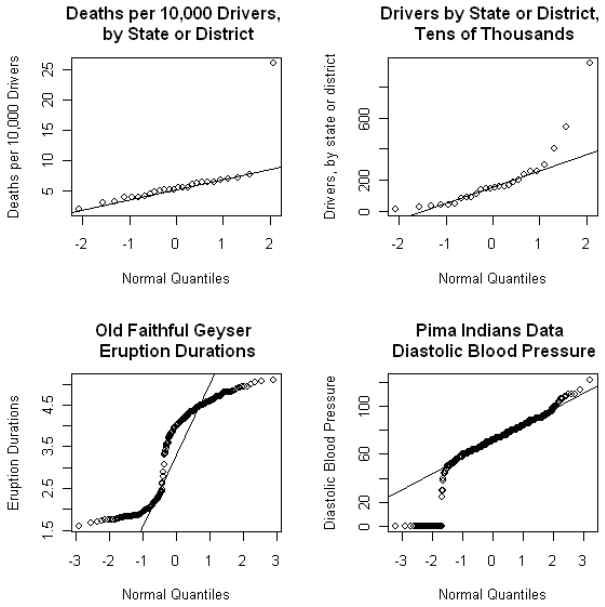
Often tens of thousands of genes (features); only tens of hundreds of samples.

- Traditional Linear Regression (OLS): $\hat{\beta} = (X^T X)^{-1} X^T Y$

where $X \in \mathbb{R}^{n \times p}$ and $X^T X \in \mathbb{R}^{p \times p}$

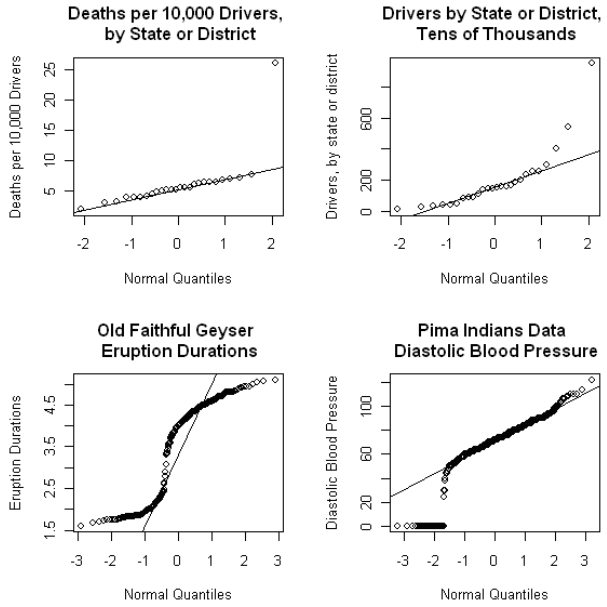
Challenges arising from Big Data

- Asymmetric and Heavy-tailed data



Challenges arising from Big Data

- Asymmetric and Heavy-tailed data



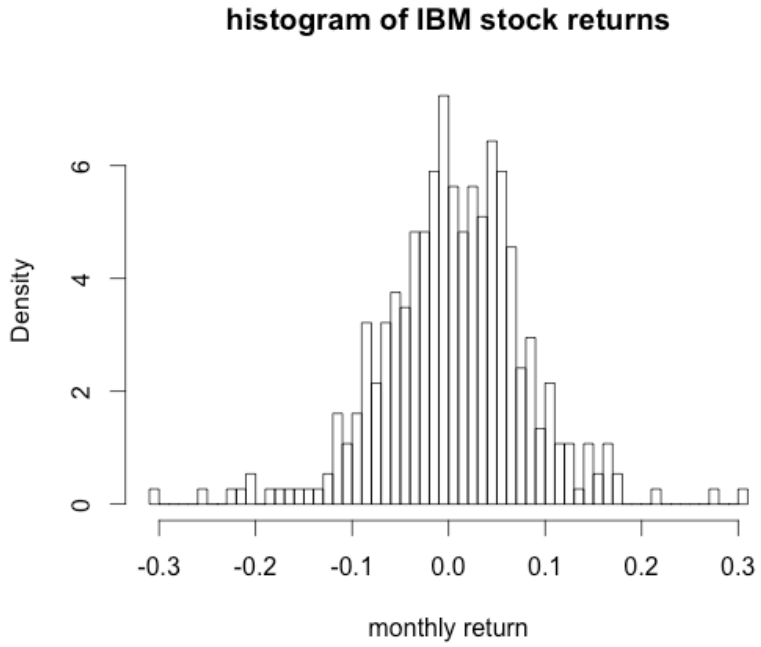
- The distribution of real data deviates from normal/light tail assumptions.

Why these are problems?

- n << p: OLS no longer applies. Do not have info to recover the true linear relation.

Why these are problems?

- n << p: OLS no longer applies. Do not have info to recover the true linear relation.
- Abnormal tails & outliers: violating the assumptions of existing methods.



Outline

- Motivation: regression in high-D with outliers
- Existing efforts: Lasso, LAD
- RA-Lasso
- Numerical results
- Theoretical justification
- My other projects

Linear regression in High-dimensional setting

- Linear model in high-D:

$$y_i = x_i^T \beta^* + \epsilon_i,\ i = 1, \ldots, n$$

- L-2 loss + penalty

e.g. Lasso [Tibshirani, 1996], SCAD [Fan & Li, 2001] and MCP [Zhang, 2010]

$$\hat{\beta}^{\text{Lasso}} = arg\,min_\beta \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|$$

Linear regression in High-dimensional setting

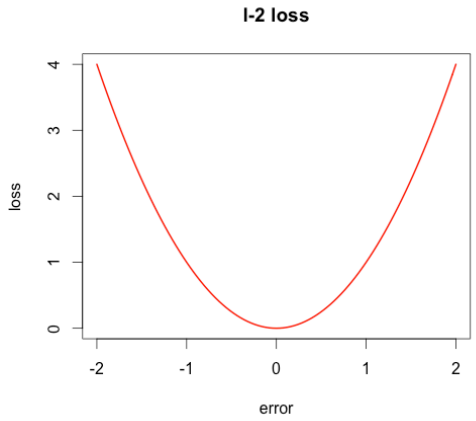
- Linear model in high-D:

$$y_i = x_i^T \beta^* + \epsilon_i, \ i = 1, \ldots, n$$

- L-2 loss + penalty

e.g. Lasso [Tibshirani, 1996], SCAD [Fan & Li, 2001] and MCP [Zhang, 2010]

$$\hat{\beta}^{\text{Lasso}} = arg\,min_\beta \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|$$



Needs light-tail assumption on the data

Linear regression in High-dimensional setting

- Linear model in high-D:

$$y_i = x_i^T \beta^* + \epsilon_i,\ i = 1, \ldots, n$$

- L-1 loss + penalty

$$\hat{\beta}^{\mathrm{LAD}} = arg\, min_\beta \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i^T \beta| + \lambda_n \sum_{j=1}^{p} |\beta_j|$$

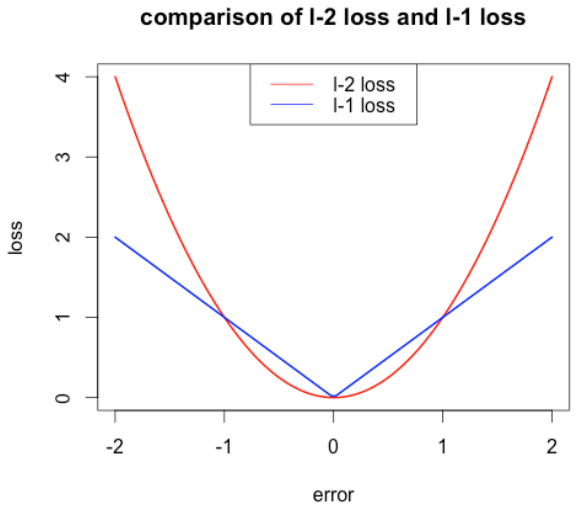
Linear regression in High-dimensional setting

- Linear model in high-D:

$$y_i = x_i^T \beta^* + \epsilon_i, \ i = 1, \ldots, n$$

- L-1 loss + penalty

$$\hat{\beta}^{\mathrm{LAD}} = arg\, min_\beta \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i^T \beta| + \lambda_n \sum_{j=1}^{p} |\beta_j|$$



Needs symmetry assumption on the data

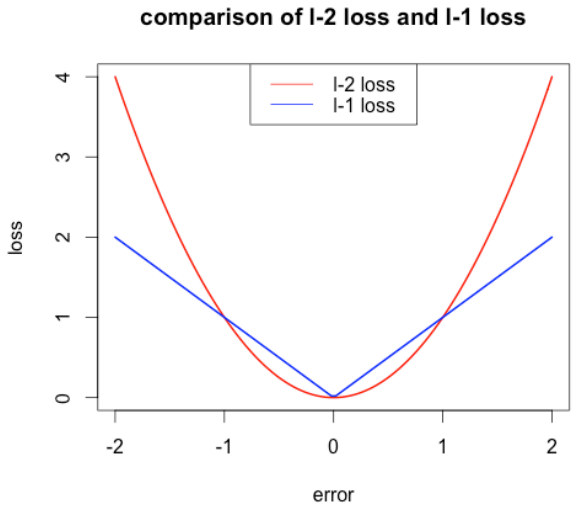
Linear regression in High-dimensional setting

- Linear model in high-D:

$$y_i = x_i^T \beta^* + \epsilon_i, \ i = 1, \ldots, n$$

- L-1 loss + penalty

$$\hat{\beta}^{\mathrm{LAD}} = arg\, min_\beta \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i^T \beta| + \lambda_n \sum_{j=1}^{p} |\beta_j|$$



Needs symmetry assumption on the data

Linear regression in High-dimensional setting

- To sum up:
- L-2 loss is **unbiased** but not robust;
- L-1 loss is **robust** but biased.

Linear regression in High-dimensional setting

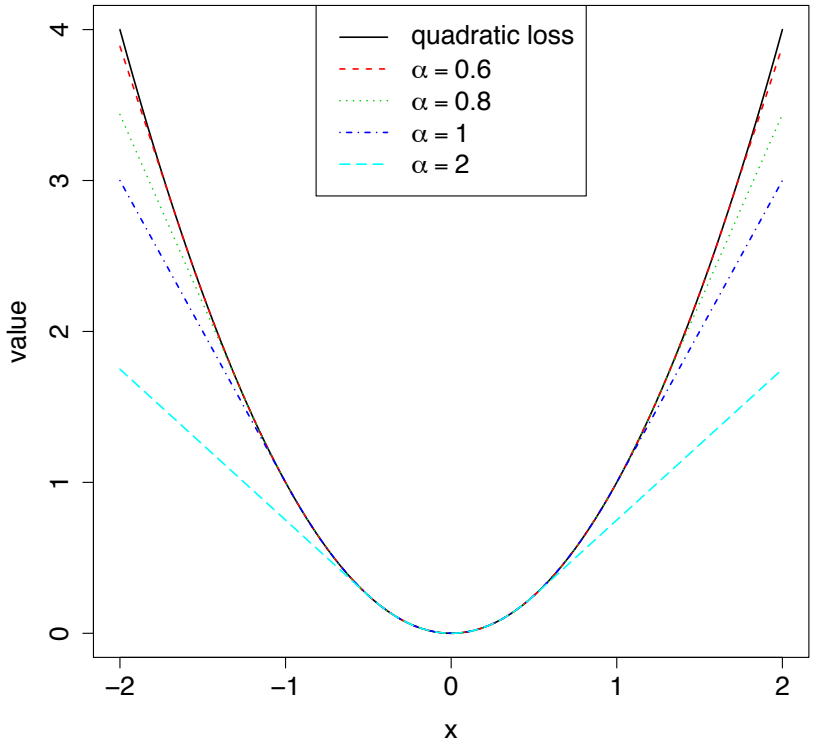
- To sum up:
- L-2 loss is **unbiased** but not robust;
- L-1 loss is **robust** but biased.
- Of interest: Robust linear regression in high-d

Outline

- Motivation: regression in high-D with outliers
- Existing efforts: Lasso, LAD
- RA-Lasso
- Numerical results
- Theoretical justification
- My other projects

Robust surrogate loss

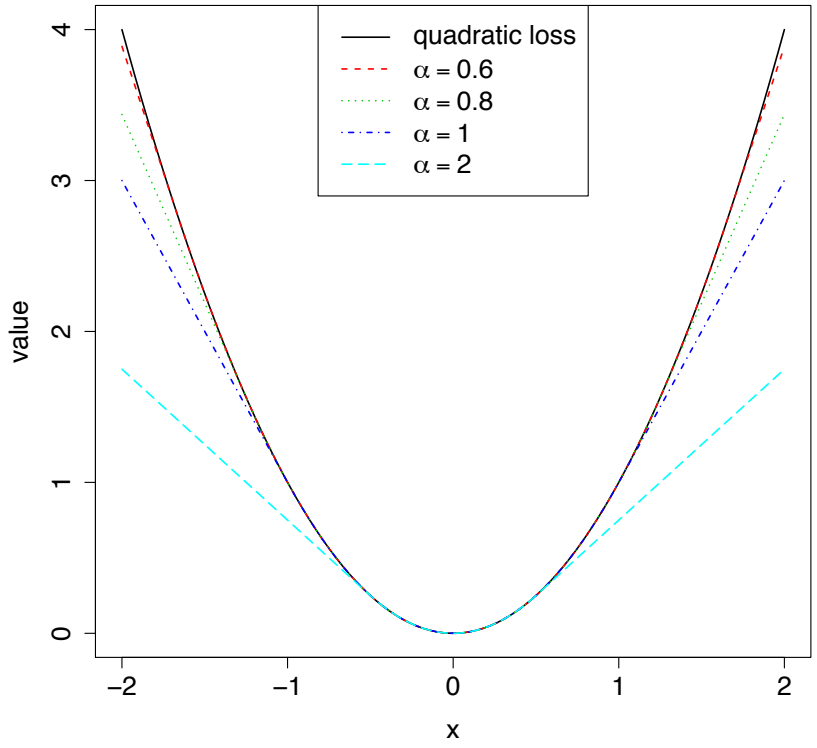
- Huber loss with varying parameter



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

Robust surrogate loss

- Huber loss with varying parameter



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

A compromise between l-2 loss and l-1 loss

Our proposed robust estimator: RA-Lasso

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

$$\hat{\beta} = arg\,min_{\beta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_{\alpha}(y_i - x_i^T\beta)}_{\text{Huber loss}} + \lambda_n \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{penalty}}$$

Our proposed robust estimator: RA-Lasso

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

$$\hat{\beta} = arg\,min_{\beta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_{\alpha}(y_i - x_i^T\beta)}_{\text{Huber loss}} + \lambda_n \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{penalty}}$$

- We will show that RA-Lasso:

(i) preserves the **optimal** convergence rate as Lasso and LAD;

(ii) requires **only** existence of 2nd moment (covers a large class of heavy-tailed and asymmetric data)

Outline

- Motivation: regression in high-D with outliers
- Existing efforts: Lasso, LAD
- RA-Lasso
- Numerical results
- Theoretical justification
- My other projects

Simulation Studies

- Homoscedastic Model:

$$y_i = x_i^T \beta + \epsilon_i, \quad x_i \sim N(0, I_p)$$

- High-D: $n = 100, p = 400, \beta^* = (\underbrace{3, \ldots, 3}_{20}, 0, \ldots, 0)^T$

Simulation Studies

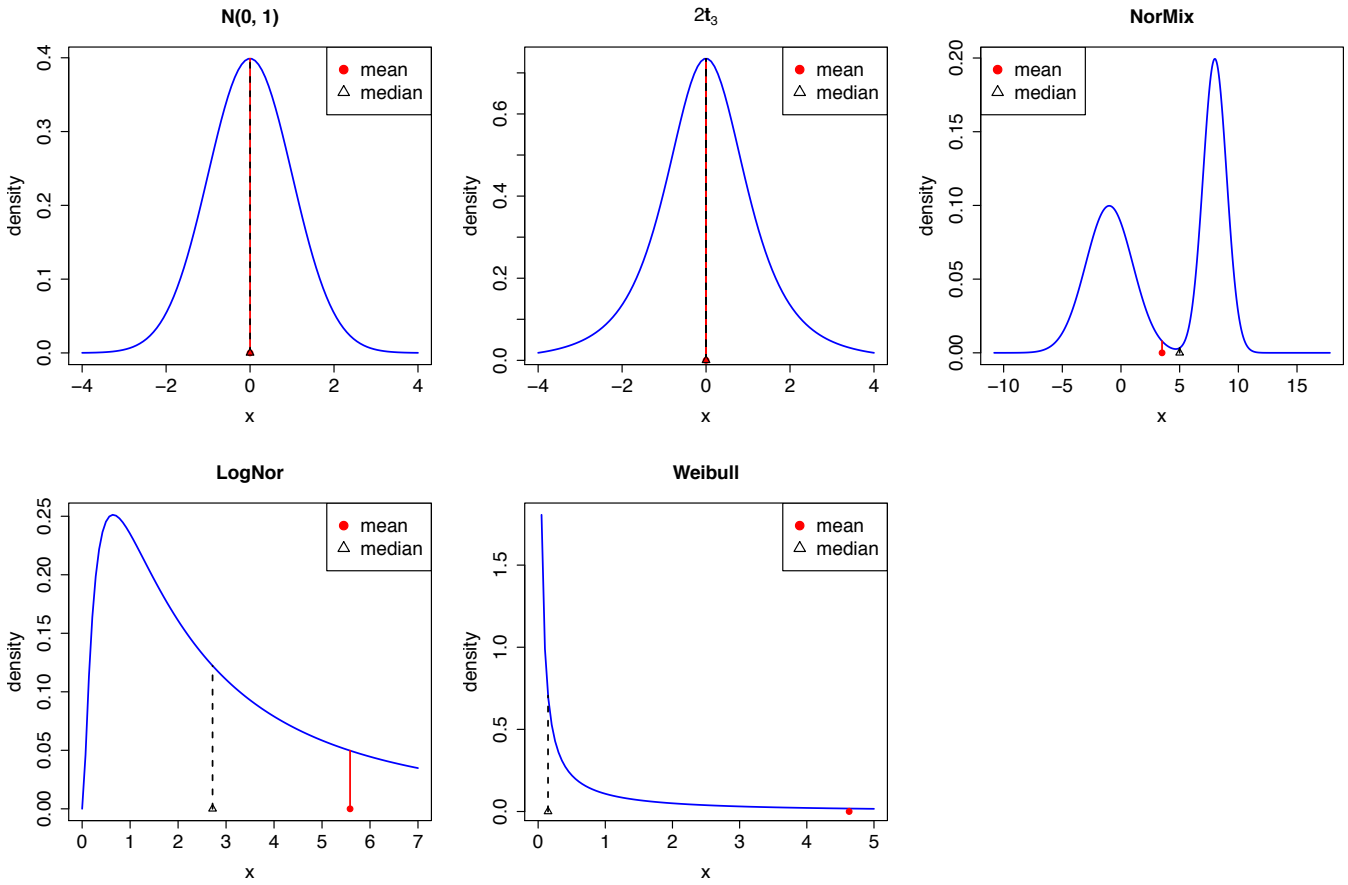
- Homoscedastic Model:

$$y_i = x_i^T \beta + \epsilon_i, \quad x_i \sim N(0, I_p)$$

- High-D: $n = 100, p = 400, \beta^* = (\underbrace{3, \ldots, 3}_{20}, 0, \ldots, 0)^T$
- Five scenarios of error distribution:
  1. $N(0, 1)$
  2. $2t_3$
  3. $\text{NorMix} : 0.5N(-1, 4) + 0.5N(8, 1)$
  4. $\text{LogNor} : \epsilon = e^{1+1.2Z}, \text{where} Z \sim N(0, 1)$
  5. $Weibull(0.3, 0.5)$

Simulation Studies

- Error distributions:



Simulation Studies

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

Table 1 : Shapes and tails of five error distributions

Simulation Studies

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

Table 1 : Shapes and tails of five error distributions

- Heteroscedastic Model:

$$y_i = x_i^T\beta + c^{-1}(x_i^T\beta)^2\epsilon_i, \quad x_i \sim N(0, I_p),\ c = \sqrt{3}||\beta^*||^2$$

Simulation Studies

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

Table 1 : Shapes and tails of five error distributions

- Heteroscedastic Model:

$$y_i = x_i^T\beta + c^{-1}(x_i^T\beta)^2\epsilon_i, \quad x_i \sim N(0, I_p),\ c = \sqrt{3}||\beta^*||^2$$

- Choice of tuning parameters: grid search based on 100 independent validation datasets.

Simulation Studies

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

Table 1 : Shapes and tails of five error distributions

- Heteroscedastic Model:

$$y_i = x_i^T\beta + c^{-1}(x_i^T\beta)^2\epsilon_i, \quad x_i \sim N(0, I_p),\ c = \sqrt{3}||\beta^*||^2$$

- Choice of tuning parameters: grid search based on 100 independent validation datasets.
- Three competitors: Lasso, LAD and ours(RA-Lasso).

Simulation Studies

- Measurements of performance:
  - L_2 error: $\|\hat{\beta} - \beta^*\|_2$;
  - L_1 error: $\|\hat{\beta} - \beta^*\|_1$;
  - Relative gain of our method against Lasso and LAD *with respect to the oracle estimator* (RG_L, RG_{LAD}):

$$\frac{\|\hat{\beta}_{\text{Lasso}} - \beta^*\|_2 - \|\hat{\beta}_{\text{oracle}} - \beta^*\|_2}{\|\hat{\beta}_{\text{our}} - \beta^*\|_2 - \|\hat{\beta}_{\text{oracle}} - \beta^*\|_2} \quad \text{and} \quad \frac{\|\hat{\beta}_{\text{LAD}} - \beta^*\|_2 - \|\hat{\beta}_{\text{oracle}} - \beta^*\|_2}{\|\hat{\beta}_{\text{our}} - \beta^*\|_2 - \|\hat{\beta}_{\text{oracle}} - \beta^*\|_2}.$$

$RG_L > 1 \Rightarrow$ ours is better than LASSO.
$RG_{LAD} > 1 \Rightarrow$ ours is better than LAD.

Simulation Studies

$RG_L > 1 \Rightarrow$ ours is better than LASSO.
$RG_{LAD} > 1 \Rightarrow$ ours is better than LAD.

		Lasso	LAD	Our	RG_L	RG_{LAD}
$\mathbf{N(0,1)}$	L_2 loss	4.54	4.40	4.53	1.00	0.96
	L_1 loss	27.21	29.11	27.21	1.00	1.08
$\mathbf{2t_3}$	L_2 loss	6.04	5.10	5.47	1.14	0.91
	L_1 loss	35.22	33.07	30.42	1.19	1.10
MixN	L_2 loss	6.14	6.44	6.13	1.00	1.06
	L_1 loss	40.46	46.18	38.48	1.06	1.23
LogNor	L_2 loss	11.08	12.16	10.10	1.14	1.30
	L_1 loss	53.17	57.18	51.58	1.04	1.14
Weibull	L_2 loss	7.77	7.11	6.62	1.23	1.10
	L_1 loss	55.65	50.49	42.93	1.34	1.20

Table 2 : Simulation results of homoscedastic models.

Simulation Studies

		Lasso	LAD	Our	RG_L	RG_{LAD}
$\mathbf{N(0,1)}$	L_2 loss	4.60	4.34	4.60	1.00	0.93
	L_1 loss	27.16	27.14	27.15	1.00	1.00
$\mathbf{2t_3}$	L_2 loss	8.08	6.71	6.70	1.26	1.01
	L_1 loss	41.16	42.76	38.52	1.08	1.12
MixN	L_2 loss	6.26	6.54	6.25	1.00	1.06
	L_1 loss	41.26	46.95	39.25	1.06	1.23
LogNor	L_2 loss	10.86	9.19	8.48	1.43	1.13
	L_1 loss	57.52	57.18	53.20	1.10	1.09
Weibull	L_2 loss	7.40	8.81	5.53	1.53	1.92
	L_1 loss	40.95	47.82	34.65	1.23	1.48

Table 3 : Simulation results of heteroscedastic models.

Microarray Data

- 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 p = 464 genes from 12 pathways closely related to TLR pathway.
- n=119 patients were involved.

Microarray Data

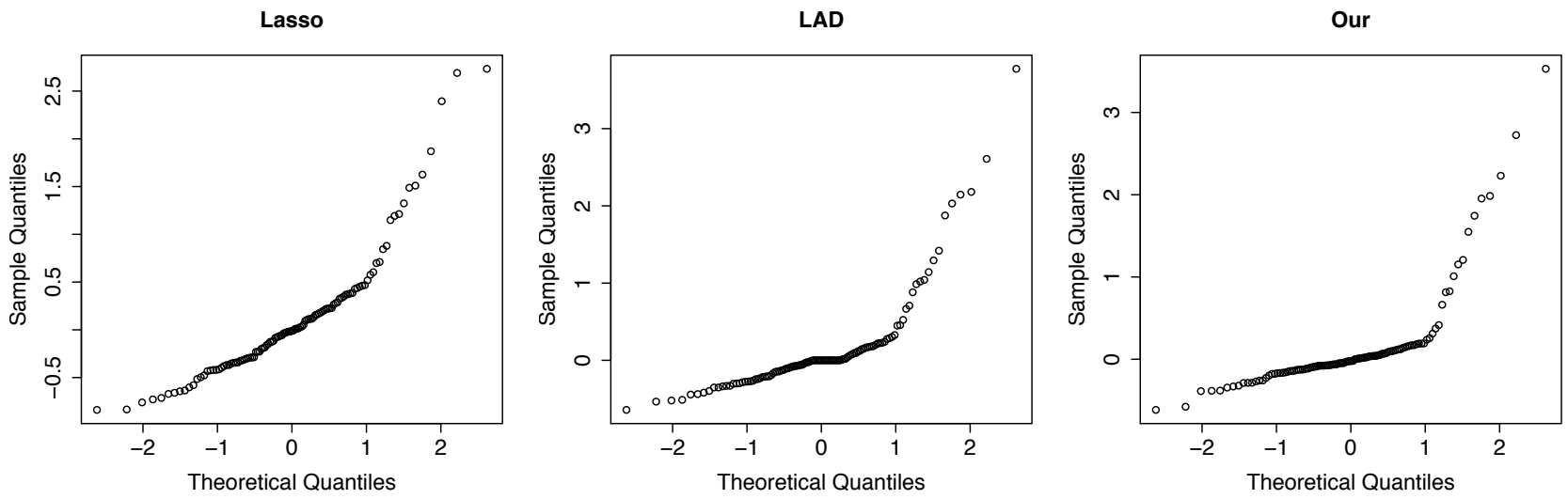


Figure 3 : QQ plots of the residuals from three methods.

Microarray Data

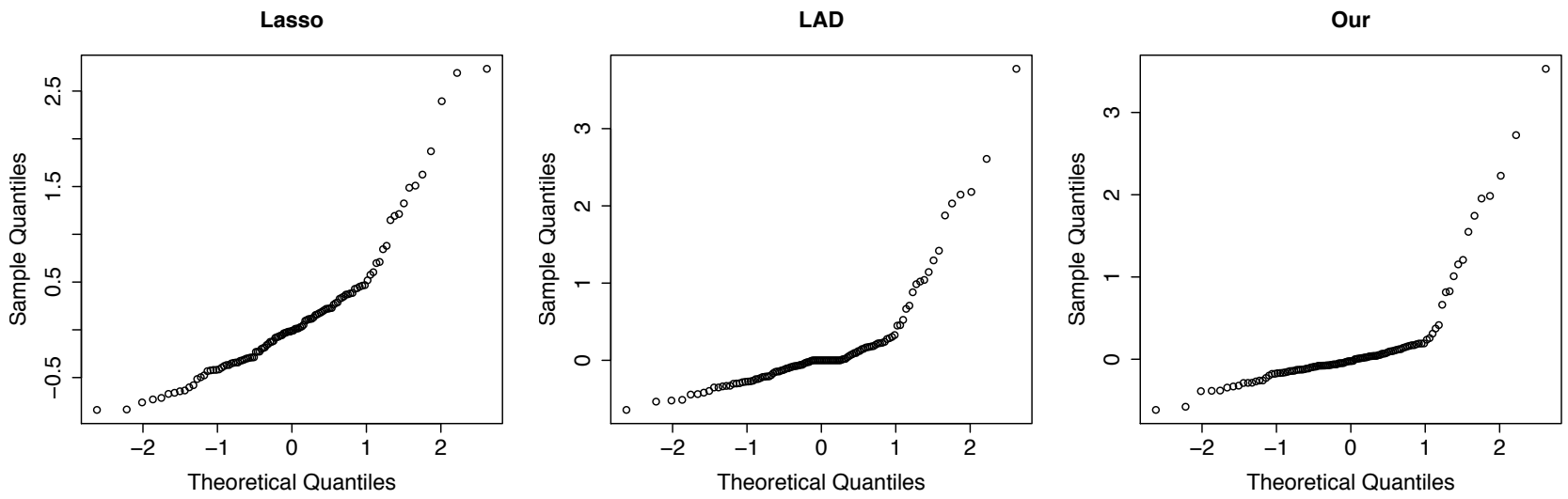


Figure 3 : QQ plots of the residuals from three methods.

Data is to the skewed right and our method captures the skewness the best.

Microarray Data

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

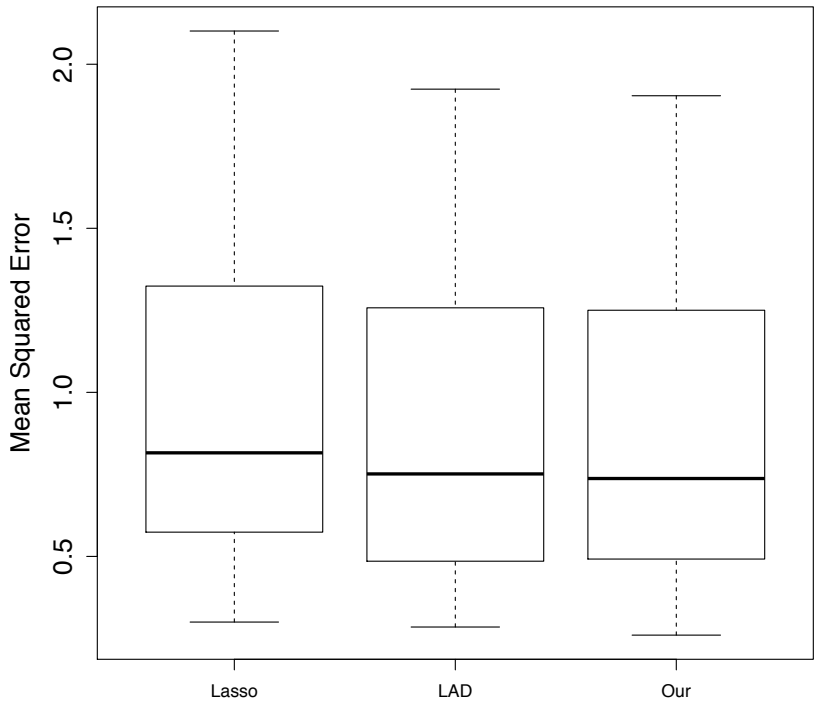
Microarray Data

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

Our method selected most genes, which is useful for a second-stage confirmatory study.

Microarray Data



- 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.

Outline

- Motivation: regression in high-D with outliers
- Existing efforts: Lasso, LAD
- RA-Lasso
- Numerical results
- Theoretical justification
- My other projects

Robust estimator: RA-Lasso

- Our proposed **RA-Lasso** estimator:

$$\hat{\beta} = arg\,min_{\beta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_{\alpha}(y_i - x_i^T\beta)}_{\text{Huber loss}} + \lambda_n \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{penalty}}$$

Robust estimator: RA-Lasso

- Our proposed **RA-Lasso** estimator:

$$\hat{\beta} = arg\,min_{\beta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_{\alpha}(y_i - x_i^T\beta)}_{\text{Huber loss}} + \lambda_n \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{penalty}}$$

- $\hat{\beta}$, for any $\hat{\alpha}$, is an estimator of:

$$\beta_{\alpha}^{*} = arg\,min_{\beta} E\ell_{\alpha}(y - x^T\beta)$$

Robust estimator: RA-Lasso

- Our proposed **RA-Lasso** estimator:

$$\hat{\beta} = arg\,min_{\beta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_{\alpha}(y_i - x_i^T\beta)}_{\text{Huber loss}} + \lambda_n \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{penalty}}$$

- $\hat{\beta}$, for any $\hat{\alpha}$, is an estimator of:

$$\beta_{\alpha}^{*} = arg\,min_{\beta} E\ell_{\alpha}(y - x^T\beta)$$

- Our goal: estimate $\beta^{*} = arg\,min_{\beta} E(y - x^T\beta)^2$

Robust estimator: RA-Lasso

- Our proposed **RA-Lasso** estimator:

$$\hat{\beta} = arg\,min_{\beta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell_{\alpha}(y_i - x_i^T\beta)}_{\text{Huber loss}} + \lambda_n \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\text{penalty}}$$

- $\hat{\beta}$, for any $\hat{\alpha}$, is an estimator of:

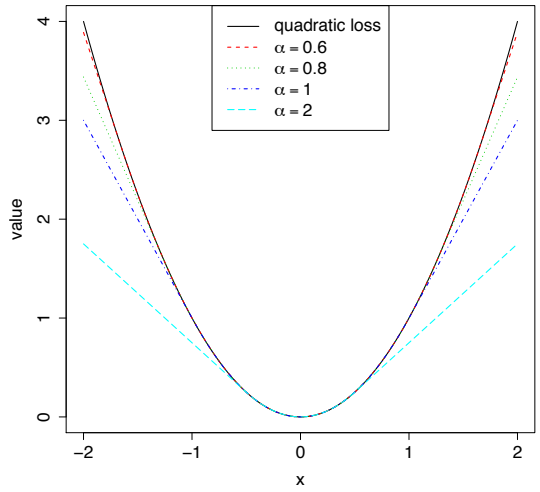
$$\beta_{\alpha}^* = arg\,min_{\beta} E\ell_{\alpha}(y - x^T\beta)$$

- Our goal: estimate $\beta^* = arg\,min_{\beta} E(y - x^T\beta)^2$
- By triangular inequality,

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

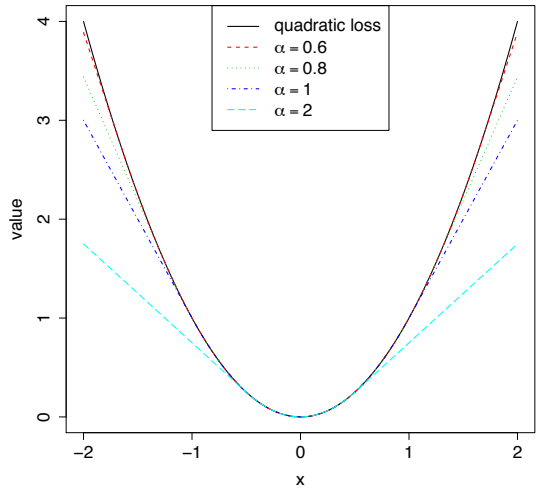
RA-Lasso: Approximation Error

- Observe that $\ell_\alpha(\cdot) \to \ell_2$-loss as $\alpha \to 0$.
- Intuitively $\beta^*_\alpha \to \beta^*$ as $\alpha \to 0$



RA-Lasso: Approximation Error

- Observe that $\ell_\alpha(\cdot) \to \ell_2$-loss as $\alpha \to 0$.
- Intuitively $\beta^*_\alpha \to \beta^*$ as $\alpha \to 0$



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

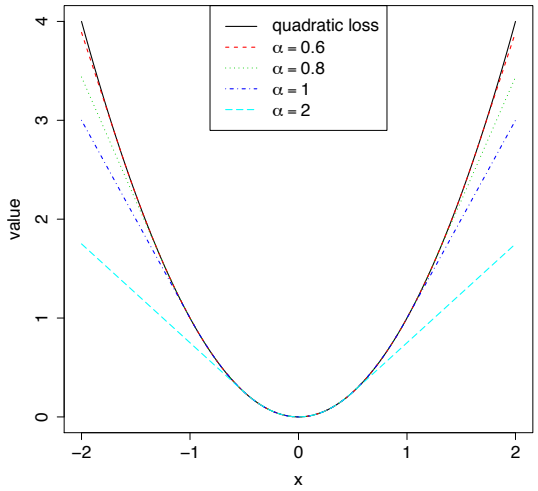
Theorem 1 (Approximation Error)

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: Approximation Error

- Observe that $\ell_\alpha(\cdot) \to \ell_2$-loss as $\alpha \to 0$.
- Intuitively $\beta^*_\alpha \to \beta^*$ as $\alpha \to 0$



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

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

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

We only need existence of 2nd moment!

RA-Lasso: Estimation Error

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

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

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

- L-2 error of a high-dim regularized convex M-estimator
- Restricted Strong Convexity (RSC)

RA-Lasso: Statistical 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 3 (Statistical Error)

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

RA-Lasso: Statistical 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 3 (Statistical Error)

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

Minimax optimal even under light tails!

Robust Estimation of Mean

Under the 1-dim scenario,

$$y_i = \mu + \epsilon_i,\ i = 1, \ldots, n$$

- 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,

$$y_i = \mu + \epsilon_i,\ i = 1, \ldots, n$$

- 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

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

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

Robust Estimation of Mean

Under the 1-dim scenario,

$$y_i = \mu + \epsilon_i,\ i = 1, \ldots, n$$

- 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

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

  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*

Our achievements

- RA-Lasso which robustly estimates mean with minimal assumption on the data;
- Optimal rate of RA-Lasso;
- Robust estimator of mean;
- Satisfactory finite sample performance.

Outline

- Motivation: regression in high-D with outliers
- Existing efforts: Lasso, LAD
- RA-Lasso
- Numerical results
- Theoretical justification
- My other projects

My other projects

- **Factor model**: the blessing of dimensionality in high-dimensional factor models.

My other projects

- **Factor model**: the blessing of dimensionality in high-dimensional factor models.
- **Statistical Modeling of Hurricane**: a mixture of sparse generalized additive model for modeling hurricane intensity, which captures sparsity, heterogeneity and nonlinearity.

My other projects

- **Factor model**: the blessing of dimensionality in high-dimensional factor models.
- **Statistical Modeling of Hurricane**: a mixture of sparse generalized additive model for modeling hurricane intensity, which captures sparsity, heterogeneity and nonlinearity.
- **Modeling effect of annotation**: a modified varying coefficient model with fused-Lasso penalty to demonstrate the effect of prior knowledge of coefficients on prediction.

My other projects

- **Factor model**: the blessing of dimensionality in high-dimensional factor models.
- **Statistical Modeling of Hurricane**: a mixture of sparse generalized additive model for modeling hurricane intensity, which captures sparsity, heterogeneity and nonlinearity.
- **Modeling effect of annotation**: a modified varying coefficient model with fused-Lasso penalty to demonstrate the effect of prior knowledge of coefficients on prediction.
- **Online search query analysis**: a novel "sample series" approach that tracks the change of online query frequencies 30% more responsively than existing time series approaches.

Thank you!

Appendix

Computational Error

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

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

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

- 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.

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

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