ESTIMATION OF HIGH-DIMENSIONAL MEAN REGRESSION IN ABSENCE OF SYMMETRY AND LIGHT-TAIL ASSUMPTIONS

PRINCETON UNIVERSITY

{ JIANQING FAN, QUEFENG LI, YUYAN WANG }

PRINCETON UNIVERSITY, DEPARTMENT OF OPERATIONS RESEARCH & FINANCIAL ENGINEERING

PRINCETON UNIVERSITY

OBJECTIVES

Data subject to heavy-tailed and/or asymmetric errors are encountered in various scientific fields. In the pursuit of estimating mean regression function under this setting and assuming high-dimensionality, our goal is to:

1. Propose a penalized method for the estimation of mean regression function under minimal assumptions on the data.
2. Develop efficient algorithms to solve for the estimator.

OPTIMAL STATISTICAL ERROR

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

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

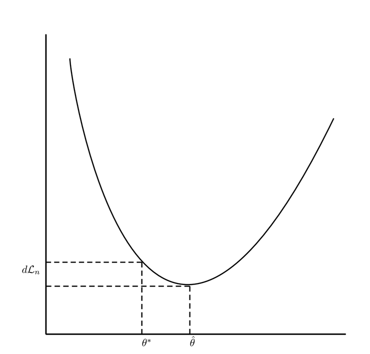


Fig. 2: Large Curvature

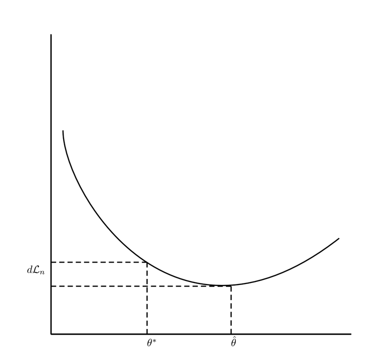


Fig. 3: Large Curvature

Restricted Strong Convexity (RSC):

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

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

Thm 2. $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*_\alpha\|_2 = O(\sqrt{R_q}[(\log p)/n]^{1/2-q/4})$.
By choosing proper α, one gets:

$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 = O(\sqrt{R_q}[(\log p)/n]^{1/2-q/4})$. $\Rightarrow$ Only requires existence of 2nd moment; Rate is the same as under the light tails!

INTRODUCTION

The existing techniques for estimating mean regression function for high-dimensional data with possibly heavy OR asymmetric tails:

- Lasso: Light & Asymmetric Tails
- LAD: Heavy & Symmetric Tails

We propose a penalized robust approximate quadratic (RA-quadratic) loss and the **RA-Lasso** estimator, which deals with heavy & asymmetric tails.

NUMERICAL STUDIES

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta}^* + c^{-1}(\mathbf{x}_i^T\boldsymbol{\beta}^*)^2\epsilon_i,$$

$$\mathbf{x}_i \sim N(0, I_p),\ c = \sqrt{3}||\boldsymbol{\beta}^*||^2.$$

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

Relative gain (RG_L, RG_{LAD}):

$$\frac{\|\hat{\boldsymbol{\beta}}_{\text{Lasso/LAD}}-\boldsymbol{\beta}^*\|_2 - \|\hat{\boldsymbol{\beta}}_{\text{oracle}}-\boldsymbol{\beta}^*\|_2}{\|\hat{\boldsymbol{\beta}}_{\text{our}}-\boldsymbol{\beta}^*\|_2 - \|\hat{\boldsymbol{\beta}}_{\text{oracle}}-\boldsymbol{\beta}^*\|_2}$$

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

		RG_L	RG_{LAD}
N(0, 1)	L_2 loss	1.00	0.93
	L_1 loss	1.00	1.00
2t₃	L_2 loss	1.26	1.01
	L_1 loss	1.08	1.12
MixN	L_2 loss	1.00	1.06
	L_1 loss	1.06	1.23
LogNor	L_2 loss	1.43	1.13
	L_1 loss	1.10	1.09
Weibull	L_2 loss	1.53	1.92
	L_1 loss	1.23	1.48

MODEL & METHODS: RA-LASSO

Consider the linear regression model:

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta}^* + \epsilon_i.$$

- $\{\mathbf{x}_i\}_{i=1}^n$ are i.i.d p-dim with $\mathrm{E}(\mathbf{x}_i) = \mathbf{0}$;
- $\{\epsilon_i\}_{i=1}^n$ are i.i.d errors with $\mathrm{E}(\epsilon_i) = 0$;
- $p \gg n$, $\log(p) = o(n^b)$ for some $0 < b < 1$.

Build a sequence of surrogate losses to approximate the least squares *at a good rate*, in order to balance bias and robustification.

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

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

We propose **RA-Lasso**:

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

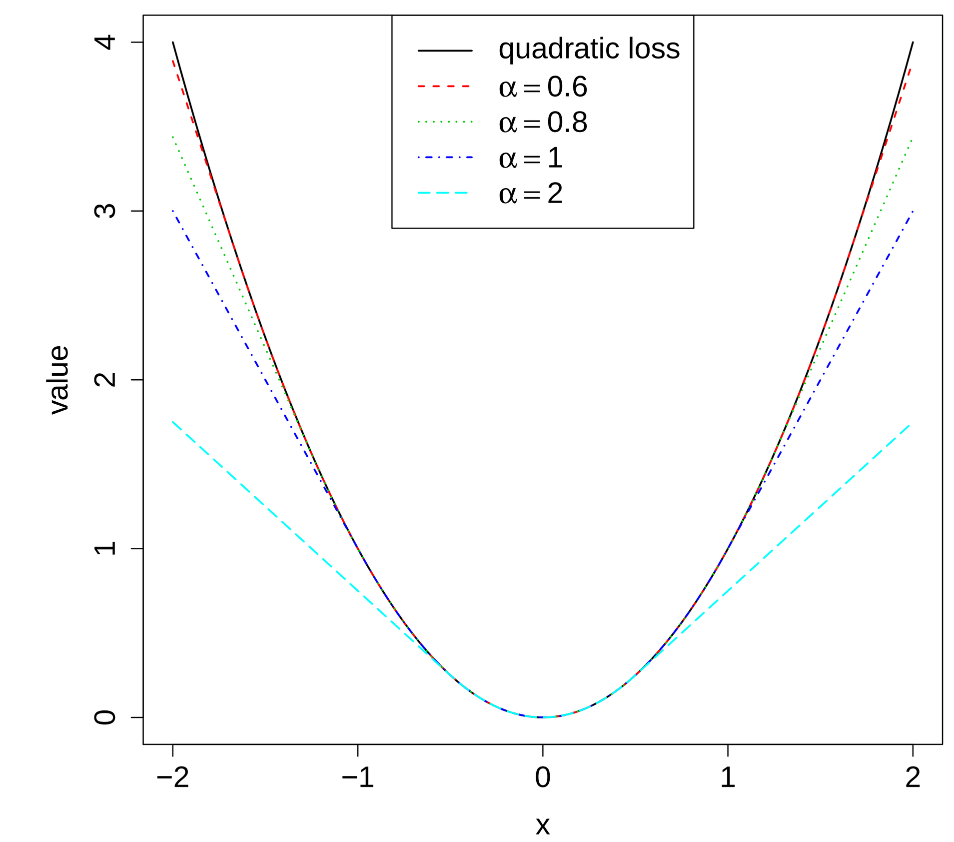


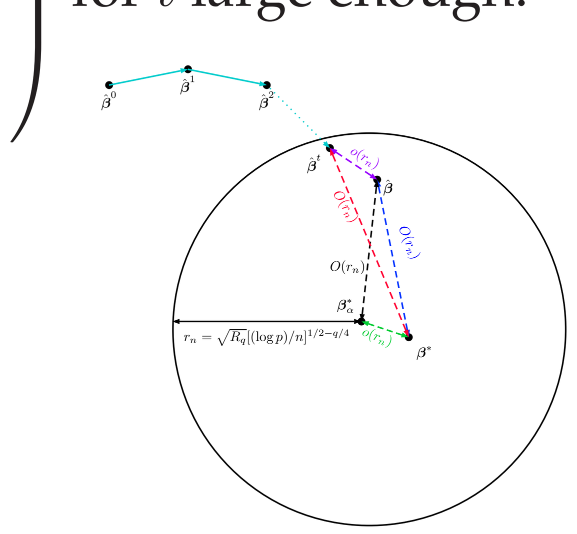
Fig. 1: Huber Loss with Varying Parameter

- For a fixed α, $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}^*_\alpha = \operatorname{argmin}_{\boldsymbol{\beta}} \mathrm{E}\,\ell_\alpha(y - \mathbf{x}^T\boldsymbol{\beta})$.
- In general, $\boldsymbol{\beta}^*_\alpha \to \boldsymbol{\beta}^*$ as $\alpha \to 0$.
- We recommend letting $\alpha \to 0$ *at a good rate*.

GEOMETRIC CONVERGENCE OF COMPUTATIONAL ERROR

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

Thm 3. $\|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}\|_2^2 = O\left(\underbrace{R_q\left(\frac{\log p}{n}\right)^{1-(q/2)}}_{o(1)}\left[\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*_\alpha\|_2^2 + R_q\left(\frac{\log p}{n}\right)^{1-(q/2)}\right]\right)$ for t large enough.



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

$$= o(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*_\alpha\|_2) + \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*_\alpha\|_2 + o(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*_\alpha\|_2)$$

$$= O(\sqrt{R_q}[(\log p)/n]^{1/2-q/4})$$

$\Rightarrow \hat{\boldsymbol{\beta}}^t$ is as good as $\hat{\boldsymbol{\beta}}$.

FUTURE RESEARCH

General theory for the robust estimation of the mean regression function in high dimension:
- A proper sequence of surrogate losses?
– should be close to quadratic loss for small errors.
– should penalize less severely for large errors.

SUMMARY

- Robust estimation of the mean function in HD.
- Approximate L_2-losswith a sequence of surrogate loss functions to balance bias and robustification.
- Only requires the existence of 2nd moment.
- Attain optimality for a *wide* range of distributions.
- No efficiency loss under special cases, compared with existing methods (LAD, quantile regression).
- A computable solution that attains the optimal error rate after sufficient iterations.

CONTACT INFORMATION

Address ORFE, Sherrerd Hall, Princeton University, Princeton, NJ, 08540
Email yuyan@princeton.edu
Phone +1(609)815-5855