A modern take on Huber regression

Po-Ling Loh

University of Cambridge
Department of Pure Mathematics and Mathematical Statistics

Data Science Seminar
LSE
15 November 2021

Introduction

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)

Introduction

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)
- **Goals:**
  1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
  2. Quantify performance with respect to deviations

Introduction

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)
- **Goals:**
  1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
  2. Quantify performance with respect to deviations
- Local stability captured by *influence function*

$$IF(x; T, F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\Delta_x) - T(F)}{\epsilon}$$

Introduction

- Robust statistics introduced in 1960s (Huber, Tukey, Hampel, et al.)
- **Goals:**
  1. Develop estimators $T(\cdot)$ that are reliable under deviations from model assumptions
  2. Quantify performance with respect to deviations
- Local stability captured by *influence function*

$$IF(x; T, F) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon \Delta_x) - T(F)}{\epsilon}$$

- Global stability captured by *breakdown point*

$$\epsilon^*(T; X_1, \dots, X_n) = \min \left\{ \frac{m}{n} : \sup_{X^m} \| T(X^m) - T(X) \| = \infty \right\}$$

Regression M-estimators

- Linear model:
$$y_i = x_i^T \beta^* + \epsilon_i, \qquad i = 1, \dots, n$$
- Assume $\epsilon_i \perp\!\!\!\perp x_i$ and $\mathbb{E}(\epsilon_i) = 0$

Regression M-estimators

- Linear model:

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

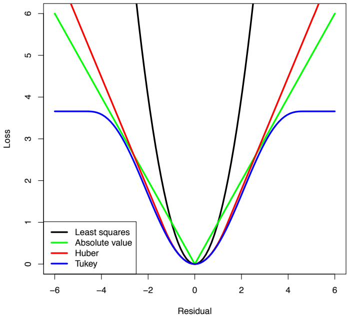
- Assume $\epsilon_i \perp\!\!\!\perp x_i$ and $\mathbb{E}(\epsilon_i) = 0$
- Generalization of OLS suitable for heavy-tailed/contaminated errors:

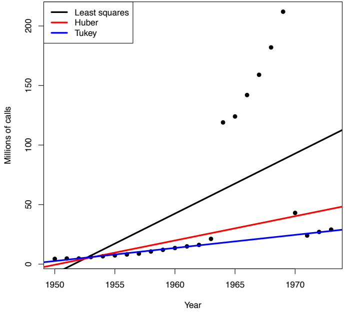
$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) \right\}$$

Regression M-estimators

- Bounded ℓ' limits influence of outliers:

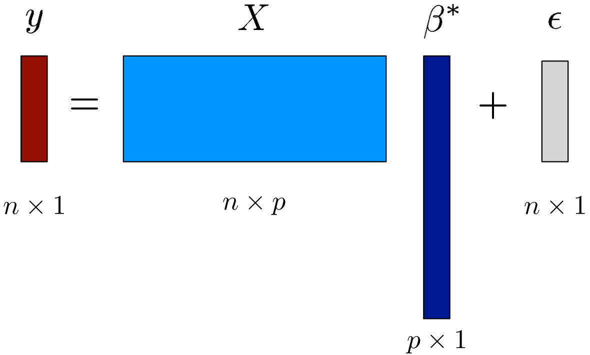
$$IF((x,y);T,F_\beta) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\Delta_{(x,y)}) - T(F)}{\epsilon} \propto \ell'(x^T\beta - y)x$$





Huber regression with scale calibration

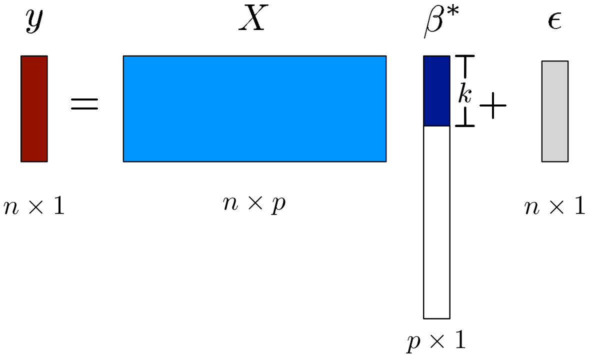
High-dimensional linear regression



- Linear model:

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

High-dimensional linear regression



- Linear model:

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

- When $p \gg n$, assume sparsity: $\|\beta^*\|_0 \leq k$

High-dimensional M-estimators

- **Natural idea:** For $p > n$, use regularized version:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \lambda \|\beta\|_1 \right\}$$

High-dimensional M-estimators

- **Natural idea:** For $p > n$, use regularized version:

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) + \lambda \|\beta\|_1 \right\}$$

Complications:

- Optimization for nonconvex ℓ?
- Statistical theory? Are certain losses provably better than others?

Motivating calculation

- Lasso analysis (e.g., van de Geer (2007), Bickel et al. (2008)):

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \underbrace{\frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1}_{\mathcal{L}_n(\beta)} \right\}$$

Motivating calculation

- Lasso analysis (e.g., van de Geer (2007), Bickel et al. (2008)):

$$\widehat{\beta} \in \arg\min_{\beta} \Bigg\{ \underbrace{\frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1}_{\mathcal{L}_n(\beta)} \Bigg\}$$

- Rearranging *basic inequality* $\mathcal{L}_n(\widehat{\beta}) \leq \mathcal{L}_n(\beta^*)$ and assuming $\lambda \geq 2\left\|\frac{X^T\epsilon}{n}\right\|_\infty$, obtain

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

Motivating calculation

- Lasso analysis (e.g., van de Geer (2007), Bickel et al. (2008)):

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \underbrace{\frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1}_{\mathcal{L}_n(\beta)} \right\}$$

- Rearranging *basic inequality* $\mathcal{L}_n(\widehat{\beta}) \leq \mathcal{L}_n(\beta^*)$ and assuming $\lambda \geq 2\left\|\frac{X^T\epsilon}{n}\right\|_\infty$, obtain

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

- Sub-Gaussian assumptions on x_i's and ϵ_i's provide $\mathcal{O}\left(\sqrt{\frac{k \log p}{n}}\right)$ bounds, minimax optimal

Motivating calculation

- **Key observation:** For general loss function, if $\lambda \geq 2 \left\| \frac{X^T \ell'(\epsilon)}{n} \right\|_\infty$, obtain
$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

Motivating calculation

- **Key observation:** For general loss function, if $\lambda \geq 2 \left\| \frac{X^T \ell'(\epsilon)}{n} \right\|_\infty$, obtain

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

- $\ell'(\epsilon)$ sub-Gaussian whenever ℓ' bounded

Motivating calculation

- **Key observation:** For general loss function, if $\lambda \geq 2 \left\| \frac{X^T \ell'(\epsilon)}{n} \right\|_\infty$, obtain
$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

- $\ell'(\epsilon)$ sub-Gaussian whenever ℓ' bounded
  $\implies$ can achieve estimation error
$$\|\widehat{\beta} - \beta^*\|_2 \leq c\sqrt{\frac{k \log p}{n}},$$
  without assuming ϵ_i is sub-Gaussian

Motivating calculation

- **Key observation:** For general loss function, if $\lambda \geq 2 \left\| \frac{X^T \ell'(\epsilon)}{n} \right\|_\infty$, obtain

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\lambda\sqrt{k}$$

- $\ell'(\epsilon)$ sub-Gaussian whenever ℓ' bounded
  $\implies$ can achieve estimation error

$$\|\widehat{\beta} - \beta^*\|_2 \leq c\sqrt{\frac{k \log p}{n}},$$

  without assuming ϵ_i is sub-Gaussian

- Also require verifying RE/RSC condition, derived from local strong convexity of ℓ near 0

The problem of scale ...

- However, hidden condition that $\mathrm{Var}(\epsilon_i) < c\gamma^2$, where γ corresponds to radius of robust loss function

The problem of scale . . .

- However, hidden condition that $\text{Var}(\epsilon_i) < c\gamma^2$, where γ corresponds to radius of robust loss function
- For non-OLS regression, "optimal" loss function should depend on scale of ϵ_i's

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \beta - y_i) \right\}$$

Some proposals

- *MM*-estimator

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\left(\frac{y_i - x_i^T \beta}{\widehat{\sigma}_0} \right) \right\},$$

using robust estimate of scale $\widehat{\sigma}_0$ based on preliminary estimate $\widehat{\beta}_0$

Some proposals

- *MM*-estimator

$$\widehat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\left(\frac{y_i - x_i^T \beta}{\widehat{\sigma}_0} \right) \right\},$$

  using robust estimate of scale $\widehat{\sigma}_0$ based on preliminary estimate $\widehat{\beta}_0$
- How to obtain $(\widehat{\beta}_0, \widehat{\sigma}_0)$?
  - *S*-estimators/LMS:

$$\widehat{\beta}_0 \in \arg\min_{\beta} \left\{ \widehat{\sigma}(r(\beta)) \right\},$$

    where $\widehat{\sigma}(r) = r_{(n-\lfloor n\delta \rfloor)}$
  - Least trimmed squares:

$$\widehat{\beta}_0 \in \arg\min_{\beta} \left\{ \sum_{i=1}^{n-\lfloor n\alpha \rfloor} (y_i - x_i^T \beta)^2_{(i)} \right\}$$

Scale calibration (L. '18)

- **Lepski's method** originally proposed for adaptive bandwidth selection in nonparametric regression

Scale calibration (L. '18)

- **Lepski's method** originally proposed for adaptive bandwidth selection in nonparametric regression
- Can be used to select σ in location/scale problem:

$$\widehat{\beta}_\sigma \in \arg\min_\beta \left\{ \frac{1}{n} \sum_{i=1}^n \ell\left(\frac{y_i - x_i^T \beta}{\sigma} \right) + \lambda\sigma \|\beta\|_1 \right\},$$

where ℓ is Huber loss with parameter 1

Lepski's method

- Preceding theory implies

$$\|\widehat{\beta}_\sigma - \beta^*\|_2 \leq C\sigma\sqrt{\frac{k \log p}{n}},$$

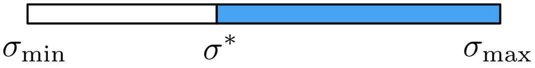
w.h.p., assuming $\sigma \geq \sqrt{\mathrm{Var}(\epsilon_i)} := \sigma^*$

Lepski's method

- Preceding theory implies

$$\|\widehat{\beta}_\sigma - \beta^*\|_2 \leq C\sigma\sqrt{\frac{k \log p}{n}},$$

  w.h.p., assuming $\sigma \geq \sqrt{\mathrm{Var}(\epsilon_i)} := \sigma^*$
- Basic idea of Lepski's method: Compute $\widehat{\beta}_\sigma$ on gridding $\{\sigma_1, \dots, \sigma_M\}$ of interval $[\sigma_{\min}, \sigma_{\max}] \ni \sigma^*$



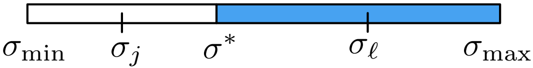
Lepski's method

- Preceding theory implies

$$\|\widehat{\beta}_\sigma - \beta^*\|_2 \leq C\sigma\sqrt{\frac{k \log p}{n}},$$

w.h.p., assuming $\sigma \geq \sqrt{\text{Var}(\epsilon_i)} := \sigma^*$

- Basic idea of Lepski's method: Compute $\widehat{\beta}_\sigma$ on gridding $\{\sigma_1, \dots, \sigma_M\}$ of interval $[\sigma_{\min}, \sigma_{\max}] \ni \sigma^*$
- For each σ_j, check if $\|\widehat{\beta}_{\sigma_j} - \widehat{\beta}_{\sigma_\ell}\|_2 \leq 2C\sigma_\ell\sqrt{\frac{k \log p}{n}}$ for all $\ell > j$, and let $\widehat{\sigma}$ be argmin in this set



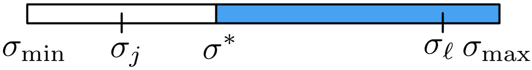
Lepski's method

- Preceding theory implies

$$\|\widehat{\beta}_\sigma - \beta^*\|_2 \leq C\sigma\sqrt{\frac{k \log p}{n}},$$

  w.h.p., assuming $\sigma \geq \sqrt{\mathrm{Var}(\epsilon_i)} := \sigma^*$

- Basic idea of Lepski's method: Compute $\widehat{\beta}_\sigma$ on gridding $\{\sigma_1, \dots, \sigma_M\}$ of interval $[\sigma_{\min}, \sigma_{\max}] \ni \sigma^*$
- For each σ_j, check if $\|\widehat{\beta}_{\sigma_j} - \widehat{\beta}_{\sigma_\ell}\|_2 \leq 2C\sigma_\ell\sqrt{\frac{k \log p}{n}}$ for all $\ell > j$, and let $\widehat{\sigma}$ be argmin in this set



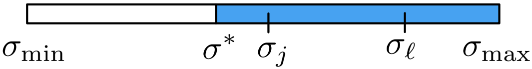
Lepski's method

- Preceding theory implies

$$\|\widehat{\beta}_\sigma - \beta^*\|_2 \le C\sigma\sqrt{\frac{k \log p}{n}},$$

  w.h.p., assuming $\sigma \ge \sqrt{\mathrm{Var}(\epsilon_i)} := \sigma^*$

- Basic idea of Lepski's method: Compute $\widehat{\beta}_\sigma$ on gridding $\{\sigma_1, \dots, \sigma_M\}$ of interval $[\sigma_{\min}, \sigma_{\max}] \ni \sigma^*$
- For each σ_j, check if $\|\widehat{\beta}_{\sigma_j} - \widehat{\beta}_{\sigma_\ell}\|_2 \le 2C\sigma_\ell\sqrt{\frac{k \log p}{n}}$ for all $\ell > j$, and let $\widehat{\sigma}$ be argmin in this set



Lepski's method

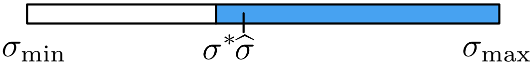
- Preceding theory implies

$$\|\widehat{\beta}_\sigma - \beta^*\|_2 \leq C\sigma\sqrt{\frac{k \log p}{n}},$$

  w.h.p., assuming $\sigma \geq \sqrt{\text{Var}(\epsilon_i)} := \sigma^*$

- Basic idea of Lepski's method: Compute $\widehat{\beta}_\sigma$ on gridding $\{\sigma_1, \dots, \sigma_M\}$ of interval $[\sigma_{\min}, \sigma_{\max}] \ni \sigma^*$

- For each σ_j, check if $\|\widehat{\beta}_{\sigma_j} - \widehat{\beta}_{\sigma_\ell}\|_2 \leq 2C\sigma_\ell\sqrt{\frac{k \log p}{n}}$ for all $\ell > j$, and let $\widehat{\sigma}$ be argmin in this set



Statistical guarantee

Theorem (L. '18)

With high probability, output of Lepski's method satisfies

$$\|\widehat{\beta}_{\widehat{\sigma}} - \beta^*\|_2 \leq C'\sigma^* \sqrt{\frac{k \log p}{n}},$$

- Method does **not** require prior knowledge of scale σ^*

Statistical guarantee

Theorem (L. '18)

With high probability, output of Lepski's method satisfies

$$\|\widehat{\beta}_{\widehat{\sigma}} - \beta^*\|_2 \leq C'\sigma^* \sqrt{\frac{k \log p}{n}},$$

- Method does **not** require prior knowledge of scale σ^*
- Constant C' still depends on properties of design matrix (RE constant)
- Choice of λ depends only on $\sqrt{\frac{\log p}{n}}$ and universal constants

Contributions

- New theory for **robust high-dimensional M-estimators** implies $\mathcal{O}\left(\sqrt{\frac{k \log p}{n}}\right)$ error rates when $\|\ell'\|_\infty \leq C$ based on local RSC
- **Lepski's method** proposed to avoid joint scale parameter estimation

Huber regression with covariate filtering

Joint work with Ankit Pensia (UW-Madison) and Varun Jog (Cambridge)

Adversarial contamination

- Instead of drawing i.i.d. data from an ϵ-contaminated mixture, draw i.i.d. data points $\{z_i\}_{i=1}^n$ and arbitrarily contaminate ϵ-fraction $\rightarrow$ observations $\{x_i\}_{i=1}^n$

Adversarial contamination

- Instead of drawing i.i.d. data from an ϵ-contaminated mixture, draw i.i.d. data points $\{z_i\}_{i=1}^n$ and arbitrarily contaminate ϵ-fraction $\rightarrow$ observations $\{x_i\}_{i=1}^n$
- Seminal papers by Diakonikolas et al. and Lai et al. on mean estimation for adversarially contaminated data (2016) for contaminated Gaussian data with $\tilde{O}(\epsilon)$ error

Adversarial contamination

- Instead of drawing i.i.d. data from an ϵ-contaminated mixture, draw i.i.d. data points $\{z_i\}_{i=1}^n$ and arbitrarily contaminate ϵ-fraction $\rightarrow$ observations $\{x_i\}_{i=1}^n$
- Seminal papers by Diakonikolas et al. and Lai et al. on mean estimation for adversarially contaminated data (2016) for contaminated Gaussian data with $\tilde{O}(\epsilon)$ error
- In our model, assume both covariates and responses may be ϵ-contaminated

Filtering algorithm

- Algorithm of Diakonikolas et al. iteratively computes weights of (remaining) data points according to projection onto top eigenvector of sample covariance matrix

Filtering algorithm

- Algorithm of Diakonikolas et al. iteratively computes weights of (remaining) data points according to projection onto top eigenvector of sample covariance matrix
- Use weights to probabilistically remove data points at each iteration

Filtering algorithm

- Success of algorithm is based on stability condition

Definition

Observations $\{x_i\}_{i=1}^n$ satisfy (ϵ, δ)-*stability* w.r.t. (μ, σ) if

$$\left\| \frac{1}{|S'|} \sum_{i \in S'} x_i - \mu \right\|_2 \le \sigma\delta, \qquad \text{and}$$
$$\left\| \frac{1}{|S'|} \sum_{i \in S'} (x_i - \mu)(x_i - \mu)^T - \sigma^2 I \right\|_2 \le \frac{\sigma^2\delta^2}{\epsilon},$$

whenever $|S'| \ge (1 - \epsilon)n$

Filtering algorithm

- Success of algorithm is based on stability condition

Definition

Observations $\{x_i\}_{i=1}^n$ satisfy (ϵ, δ)-*stability* w.r.t. (μ, σ) if

$$\left\| \frac{1}{|S'|} \sum_{i \in S'} x_i - \mu \right\|_2 \leq \sigma\delta, \qquad \text{and}$$

$$\left\| \frac{1}{|S'|} \sum_{i \in S'} (x_i - \mu)(x_i - \mu)^T - \sigma^2 I \right\|_2 \leq \frac{\sigma^2 \delta^2}{\epsilon},$$

whenever $|S'| \geq (1 - \epsilon)n$

- Filtering algorithm identifies large stable set, w.h.p., when data are ϵ-corrupted and/or heavy-tailed

Linear model assumptions

- Linear model:
$$y_i = x_i^T \beta^* + z_i, \qquad i = 1, \dots, n$$
- Distributional assumptions:
  - Covariates: $\mathbb{E}(x_i) = 0$, $\mathbb{E}(x_i x_i^T) = I$, and $\mathbb{E}[(v^T x_i)^4]^{1/4} \le C\mathbb{E}[(v^T x_i)^2]^{1/2}$ for all $\|v\|_2 = 1$
  - Noise: $z_i \perp\!\!\!\perp x_i$ and $\mathbb{E}(z_i) = 0$ (moment assumptions specified later)

Linear model assumptions

- Linear model:

$$y_i = x_i^T \beta^* + z_i, \qquad i = 1, \dots, n$$

- Distributional assumptions:
  - Covariates: $\mathbb{E}(x_i) = 0$, $\mathbb{E}(x_i x_i^T) = I$, and $\mathbb{E}[(v^T x_i)^4]^{1/4} \leq C\mathbb{E}[(v^T x_i)^2]^{1/2}$ for all $\|v\|_2 = 1$
  - Noise: $z_i \perp\!\!\!\perp x_i$ and $\mathbb{E}(z_i) = 0$ (moment assumptions specified later)
- Low-dimensional setting, $n \geq p$

Linear model assumptions

- Linear model:
$$y_i = x_i^T \beta^* + z_i, \qquad i = 1, \dots, n$$
- Distributional assumptions:
  - Covariates: $\mathbb{E}(x_i) = 0$, $\mathbb{E}(x_i x_i^T) = I$, and $\mathbb{E}[(v^T x_i)^4]^{1/4} \le C\mathbb{E}[(v^T x_i)^2]^{1/2}$ for all $\|v\|_2 = 1$
  - Noise: $z_i \perp\!\!\!\perp x_i$ and $\mathbb{E}(z_i) = 0$ (moment assumptions specified later)
- Low-dimensional setting, $n \ge p$
- After seeing i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$, adversary can contaminate ϵn data points to obtain $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n$

Huber regression

- Huber loss:
$$\ell_\gamma(x) = \begin{cases} \frac{x^2}{2}, & |x| \le \gamma, \\ \gamma|x| - \frac{\gamma^2}{2}, & |x| > \gamma \end{cases}$$
- Huber estimator: $\widehat{\beta}_{Hub} \in \arg\min_\beta \left\{\sum_{i=1}^n \ell_\gamma(y_i - x_i^T \beta)\right\}$

Huber regression

- Huber loss:

$$\ell_\gamma(x) = \begin{cases} \frac{x^2}{2}, & |x| \le \gamma, \\ \gamma|x| - \frac{\gamma^2}{2}, & |x| > \gamma \end{cases}$$

- Huber estimator: $\widehat{\beta}_{Hub} \in \arg\min_\beta \left\{\sum_{i=1}^n \ell_\gamma(y_i - x_i^T \beta)\right\}$
- Existing analysis for sub-Gaussian/uncontaminated covariates:
  - Sun et al. (2020) derived theory for $\widehat{\beta}_{Hub}$ for fixed design, heavy-tailed errors
  - Sasai and Fujisawa (2020) derived theory for $\widehat{\beta}_{Hub}$ under adversarially contaminated responses

Huber regression

- Huber loss:

$$\ell_\gamma(x) = \begin{cases} \frac{x^2}{2}, & |x| \leq \gamma, \\ \gamma|x| - \frac{\gamma^2}{2}, & |x| > \gamma \end{cases}$$

- Huber estimator: $\widehat{\beta}_{Hub} \in \arg\min_\beta \left\{\sum_{i=1}^n \ell_\gamma(y_i - x_i^T \beta)\right\}$
- Existing analysis for sub-Gaussian/uncontaminated covariates:
  - Sun et al. (2020) derived theory for $\widehat{\beta}_{Hub}$ for fixed design, heavy-tailed errors
  - Sasai and Fujisawa (2020) derived theory for $\widehat{\beta}_{Hub}$ under adversarially contaminated responses
- **Our idea:** Apply filtering algorithm with parameter ϵ' on x_i's, then run Huber regression on remaining data points

Filtered Huber regression

Theorem

Suppose $\mathbb{E}[z_i^2] = \sigma^2$, *and suppose* $n = \Omega(p \log p + \log(1/\tau))$. *Then the filtered Huber regression algorithm with* $\epsilon' = \Theta(\epsilon)$ *and* $\gamma = \Omega(\sigma)$ *satisfies*

$$\|\widehat{\beta} - \beta^*\|_2 \lesssim \gamma \left(\sqrt{\frac{p \log p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \epsilon^{3/4} \right),$$

with probability at least $1 - \tau$.

Filtered Huber regression

Theorem

Suppose $\mathbb{E}[z_i^2] = \sigma^2$, *and suppose* $n = \Omega(p \log p + \log(1/\tau))$. *Then the filtered Huber regression algorithm with* $\epsilon' = \Theta(\epsilon)$ *and* $\gamma = \Omega(\sigma)$ *satisfies*

$$\|\widehat{\beta} - \beta^*\|_2 \precsim \gamma \left(\sqrt{\frac{p \log p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \epsilon^{3/4} \right),$$

with probability at least $1 - \tau$.

- Assuming k^{th}-moment condition on covariates, can improve rate to $O(\epsilon^{1-1/k})$

Filtered Huber regression

Theorem

Suppose $\mathbb{E}[z_i^2] = \sigma^2$*, and suppose* $n = \Omega(p \log p + \log(1/\tau))$*. Then the filtered Huber regression algorithm with* $\epsilon' = \Theta(\epsilon)$ *and* $\gamma = \Omega(\sigma)$ *satisfies*

$$\|\widehat{\beta} - \beta^*\|_2 \lesssim \gamma \left(\sqrt{\frac{p \log p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \epsilon^{3/4} \right),$$

with probability at least $1 - \tau$.

- Assuming k^{th}-moment condition on covariates, can improve rate to $O(\epsilon^{1-1/k})$
- Rate-optimal for linear regression under adversarial contamination

Filtered Huber regression

Theorem

Suppose $\mathbb{E}[z_i^2] = \sigma^2$*, and suppose* $n = \Omega(p \log p + \log(1/\tau))$*. Then the filtered Huber regression algorithm with* $\epsilon' = \Theta(\epsilon)$ *and* $\gamma = \Omega(\sigma)$ *satisfies*

$$\|\widehat{\beta} - \beta^*\|_2 \lesssim \gamma \left(\sqrt{\frac{p \log p}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \epsilon^{3/4} \right),$$

with probability at least $1 - \tau$*.*

- Assuming k^{th}-moment condition on covariates, can improve rate to $O(\epsilon^{1-1/k})$
- Rate-optimal for linear regression under adversarial contamination
- Huber parameter can again be calibrated using Lepski-type procedure

Proof idea

- Filtered covariates satisfy weak stability, w.h.p.:

$$L \leq \lambda_{\min}\left(\frac{1}{n}\sum_{i \in S} \tilde{x}_i \tilde{x}_i^T\right) \leq \lambda_{\max}\left(\frac{1}{n}\sum_{i \in S} \tilde{x}_i \tilde{x}_i^T\right) \leq U,$$

whenever $|S| \geq (1-\epsilon)n$

Proof idea

- Filtered covariates satisfy weak stability, w.h.p.:

$$L \leq \lambda_{\min}\left(\frac{1}{n}\sum_{i \in S} \tilde{x}_i \tilde{x}_i^T\right) \leq \lambda_{\max}\left(\frac{1}{n}\sum_{i \in S} \tilde{x}_i \tilde{x}_i^T\right) \leq U,$$

  whenever $|S| \geq (1-\epsilon)n$
- Also need to establish deviation bound on gradient of loss:

$$\|\nabla \mathcal{L}_\gamma(\beta^*)\|_2 \precsim \gamma\left(\sqrt{\frac{p \log p}{n}} + \epsilon^{1-1/k} + \sqrt{\frac{\log(1/\tau)}{n}}\right)$$

  and local strong convexity of $\mathcal{L}_\gamma$ around β^*

Related work

- Relatively little work for adversarial contamination in both covariates and responses

Related work

- Relatively little work for adversarial contamination in both covariates and responses
  - General framework for robust ERM by Diakonikolas et al. (2019) and Prasad et al. (2020) does not achieve optimal rates for linear regression
  - Diakonikolas et al. (2019) analyzed contaminated model for Gaussian setting
  - Recent works by Zhu et al. (2020), Bakshi and Prasad (2020), Cherapanamjeri et al. (2020), Depersin (2020) analyzed slightly different assumptions on covariate/noise distributions, but algorithms are somewhat different and sometimes rather complicated (e.g., sum-of-squares procedure)

Related work

- Relatively little work for adversarial contamination in both covariates and responses
  - General framework for robust ERM by Diakonikolas et al. (2019) and Prasad et al. (2020) does not achieve optimal rates for linear regression
  - Diakonikolas et al. (2019) analyzed contaminated model for Gaussian setting
  - Recent works by Zhu et al. (2020), Bakshi and Prasad (2020), Cherapanamjeri et al. (2020), Depersin (2020) analyzed slightly different assumptions on covariate/noise distributions, but algorithms are somewhat different and sometimes rather complicated (e.g., sum-of-squares procedure)
- **Note:** Several connections between optimal estimators for heavy-tailed/adversarially contaminated data have appeared in past few years

LTS regression

- Least trimmed squares (LTS):

$$\widehat{\beta}_{LTS} \in \arg\min_{\beta} \left\{ \sum_{i=1}^{n-m} (y_i - x_i^T \beta)^2_{(i)} \right\}$$

LTS regression

- Least trimmed squares (LTS):

$$\widehat{\beta}_{LTS} \in \arg\min_{\beta} \left\{ \sum_{i=1}^{n-m} (y_i - x_i^T \beta)_{(i)}^2 \right\}$$

- Bhatia et al. (2015) established error bound for LTS with adversarially contaminated responses, when covariates satisfy subset strong convexity/smoothness (SSC/S) condition:

$$\lambda_m \leq \min_{|S|=m} \lambda_{\min} \left(\sum_{i \in S} x_i x_i^T \right) \leq \max_{|S|=m} \lambda_{\max} \left(\sum_{i \in S} x_i x_i^T \right) \leq \Lambda_m,$$

  with $\frac{\Lambda_{2m}}{\lambda_n} < \frac{1}{4}$ and $\Lambda_n = O(\lambda_n)$
- Condition holds w.h.p. for i.i.d. Gaussian covariates

Alternating minimization algorithm

- Recast LTS problem as

$$\min_{\beta \in \mathbb{R}^p, \|b\|_0 \le m} \|X\beta - (y - b)\|_2^2$$

Alternating minimization algorithm

- Recast LTS problem as

$$\min_{\beta \in \mathbb{R}^p, \|b\|_0 \le m} \|X\beta - (y - b)\|_2^2$$

- Alternately minimize over β and b:

$$\beta^j = (X^T X)^{-1} X^T (y - b^{j-1}),$$
$$b^j = HT_m(y - X\beta^j)$$

Alternating minimization algorithm

- Recast LTS problem as

$$\min_{\beta \in \mathbb{R}^p, \|b\|_0 \le m} \|X\beta - (y - b)\|_2^2$$

- Alternately minimize over β and b:

$$\beta^j = (X^T X)^{-1} X^T (y - b^{j-1}),$$
$$b^j = HT_m(y - X\beta^j)$$

- May converge to local optimum, but proved statistical error bound on output

Filtered LTS

Theorem

Suppose $\mathbb{E}[z_i^2] = \sigma^2$ *and* $\mathbb{E}[z_i^{k'}]^{1/k'} \le C$ *for* $k' \ge 2$*, and suppose* $n = \Omega(p \log p + \log(1/\tau))$*. Then the filtered LTS regression algorithm with* $m = \Theta(p \log p + \epsilon n + \log(1/\tau))$ *and* $\epsilon' = \Theta(\frac{m}{n})$ *satisfies*

$$\|\widehat{\beta} - \beta^*\|_2 \precsim \sigma \left(\frac{p \log p}{n} + \frac{\log(1/\tau)}{n} + \epsilon \right)^{1/2 - 1/k'},$$

with probability at least $1 - \tau$*.*

- Suboptimal error rate can be improved via postprocessing step (later)

LAD regression

- Least absolute deviation (LAD):

$$\widehat{\beta}_{LAD} \in \arg\min_{\beta} \left\{ \sum_{i=1}^{n} |y_i - x_i^T \beta| \right\}$$

LAD regression

- Least absolute deviation (LAD):

$$\widehat{\beta}_{LAD} \in \arg\min_{\beta} \left\{ \sum_{i=1}^{n} |y_i - x_i^T \beta| \right\}$$

- Karmalkar and Price (2019) established error bound for LAD when covariates satisfy ℓ_1-*stability*:

$$\frac{1}{n} \sum_{i \in S} |x_i^T v| \geq M, \quad \text{and} \quad \frac{1}{n} \sum_{i \notin S} |x_i^T v| \leq m,$$

for all $|S| \geq (1 - \epsilon)n$ and unit vectors v

LAD regression

- Least absolute deviation (LAD):

$$\widehat{\beta}_{LAD} \in \arg\min_{\beta} \left\{ \sum_{i=1}^{n} |y_i - x_i^T \beta| \right\}$$

- Karmalkar and Price (2019) established error bound for LAD when covariates satisfy ℓ_1-*stability*:

$$\frac{1}{n} \sum_{i \in S} |x_i^T v| \geq M, \quad \text{and} \quad \frac{1}{n} \sum_{i \notin S} |x_i^T v| \leq m,$$

  for all $|S| \geq (1 - \epsilon)n$ and unit vectors v
- Responses may be adversarially contaminated, but again, covariates are i.i.d. Gaussian
- Focus of that paper was ℓ_1-penalized LAD

Filtered LAD

Theorem

Suppose $\mathbb{E}|z_i| = \kappa$ *and* $n = \Omega(p \log p + \log(1/\tau))$. *Then the filtered LAD regression algorithm with* $\epsilon' = \Theta(1)$ *satisfies*

$$\|\widehat{\beta} - \beta^*\|_2 \precsim \kappa,$$

with probability at least $1 - \tau$.

- Suboptimal error rate can also be improved via postprocessing

Filtered LAD

Theorem

Suppose $\mathbb{E}|z_i| = \kappa$ *and* $n = \Omega(p \log p + \log(1/\tau))$. *Then the filtered LAD regression algorithm with* $\epsilon' = \Theta(1)$ *satisfies*

$$\|\widehat{\beta} - \beta^*\|_2 \lesssim \kappa,$$

with probability at least $1 - \tau$.

- Suboptimal error rate can also be improved via postprocessing
- Benefits of LAD estimator: no tuning parameter, only requires bounded first moment of error distribution (and does not even require $z_i \perp\!\!\!\perp x_i$ or $\mathbb{E}(z_i) = 0$)

Postprocessing

- Suppose $\mathbb{E}[z_i^2] = \sigma^2$ and initial estimator $\widehat{\beta}_1$ satisfies

$$\|\widehat{\beta}_1 - \beta^*\|_2 = O(\sigma)$$

Postprocessing

- Suppose $\mathbb{E}[z_i^2] = \sigma^2$ and initial estimator $\widehat{\beta}_1$ satisfies

$$\|\widehat{\beta}_1 - \beta^*\|_2 = O(\sigma)$$

- Apply filtering (mean estimation) to vectors $\left\{\widehat{\beta}_1 + (y_i - x_i^T \widehat{\beta}_1) x_i\right\}_{i=1}^n$

Postprocessing

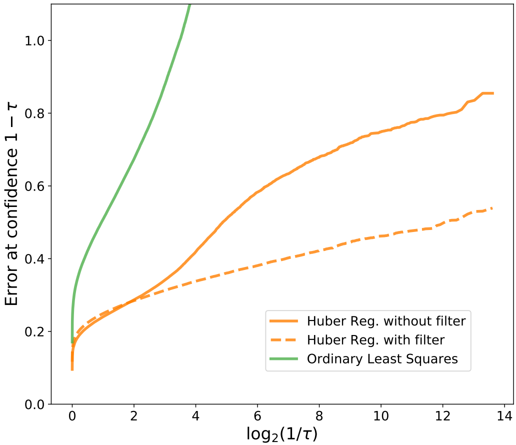
- Suppose $\mathbb{E}[z_i^2] = \sigma^2$ and initial estimator $\widehat{\beta}_1$ satisfies

$$\|\widehat{\beta}_1 - \beta^*\|_2 = O(\sigma)$$

- Apply filtering (mean estimation) to vectors $\left\{\widehat{\beta}_1 + (y_i - x_i^T \widehat{\beta}_1) x_i\right\}_{i=1}^n$
- Output $\widehat{\beta}$ has near-optimal error rates:

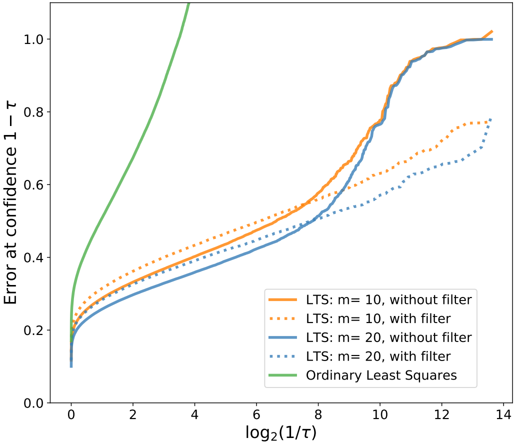
$$\|\widehat{\beta} - \beta^*\|_2 \precsim \sigma \left(\sqrt{\frac{p \log(pn)}{n}} + \sqrt{\frac{\log(1/\tau)}{n}} + \sqrt{\epsilon}\right)$$

Simulations: Huber + heavy-tailed data



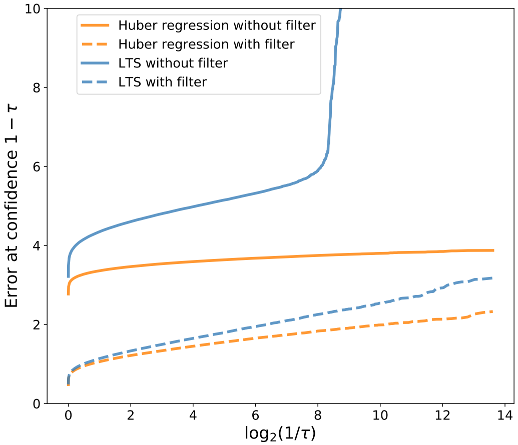
- x_i's and z_i's sampled from Pareto distribution, $f(u) \propto \left(\frac{1}{|u|+1}\right)^{1+\alpha}$
- $n = 200$, $p = 40$, Huber parameter $\gamma = 0.5$
- Filter removes 10 points

Simulations: LTS + heavy-tailed data



- LTS parameter $m \in \{10, 20\}$

Simulations: Adversarially contaminated, heavy-tailed data



- 20 points set to deterministic (large) outlying values
- Filter removes 30 points
- Huber parameter $\gamma = 0.5$, LTS parameter $m = 30$

Contributions

- Showed that various classical robust regression estimators (Huber, LTS, LAD) can be made robust to heavy tails and adversarial contamination by **simple covariate filtering** step
- Filtered Huber regression leads to near-optimal rates in ϵ, p, τ, n
- Filtered LTS and LAD can be made near-optimal after **additional postprocessing** step

Open questions

- Extension of filtering method to high-dimensional linear regression
- Unknown covariance Σ_x, relaxing independence assumption $x_i \perp\!\!\!\perp z_i$

References

- Loh (2021). Scale calibration for high-dimensional robust regression. *To appear in Electronic Journal of Statistics*.
- Pensia, Jog & Loh (2020). Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint*.

Thank you!!