Inference For High Dimensional M-estimates: Fixed Design Results

Lihua Lei, Peter Bickel and Noureddine El Karoui

Department of Statistics, UC Berkeley

Berkeley-Stanford Econometrics Jamboree, 2017

Table of Contents

Table of Contents

Setup

Consider a linear Model:

$$Y = X\beta^* + \epsilon.$$

- $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$: response vector;
- $X = (x_1^T, \ldots, x_n^T)^T \in \mathbb{R}^{n\times p}$: design matrix;
- $\beta^* = (\beta_1^*, \ldots, \beta_p^*)^T \in \mathbb{R}^p$: coefficient vector;
- $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n$: random unobserved error with independent entries.

M-Estimator

M-Estimator: Given a convex loss function $\rho(\cdot) : \mathbb{R} \to [0, \infty)$,

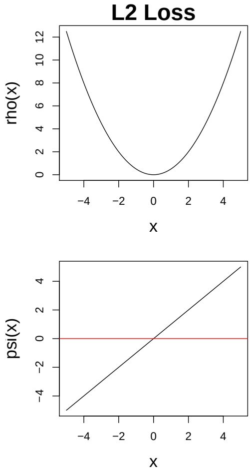
$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - x_i^T \beta).$$

When ρ is differentiable with $\psi = \rho'$, $\hat{\beta}$ can be written as the solution:

$$\frac{1}{n} \sum_{i=1}^{n} \psi(y_i - x_i^T \hat{\beta}) = 0.$$

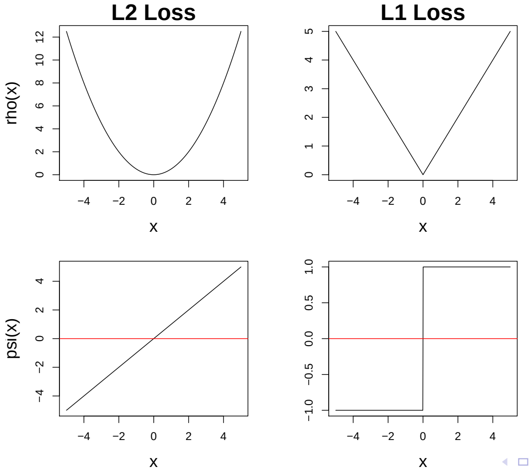
M-Estimator: Examples

- $\rho(x) = x^2/2$ gives the Least-Square estimator;



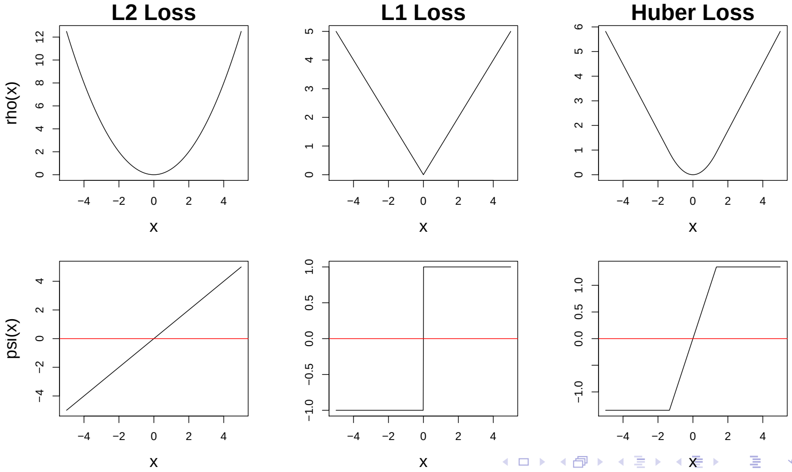
M-Estimator: Examples

- $\rho(x) = x^2/2$ gives the Least-Square estimator;
- $\rho(x) = |x|$ gives the Least-Absolute-Deviation estimator;



M-Estimator: Examples

- $\rho(x) = x^2/2$ gives the Least-Square estimator;
- $\rho(x) = |x|$ gives the Least-Absolute-Deviation estimator;
- $\rho(x) = \begin{cases} x^2/2 & |x| \leq k \\ k(|x| - k/2) & |x| > k \end{cases}$ gives the Huber estimator.



Goals (Informal)

Goal (Informal): Make inference on the **coordinates** of β^* when

- X is treated as **fixed**;
- **no assumption imposed on** β^*;
- and the dimension p is **comparable to** the sample size n.

Goals (Informal)

Goal (Informal): Make inference on the **coordinates** of β^* when

- X is treated as **fixed**;
- **no assumption imposed on** β^*;
- and the dimension p is **comparable to** the sample size n.

- Why coordinates?

Goals (Informal)

Goal (Informal): Make inference on the **coordinates** of β^* when

- X is treated as **fixed**;
- **no assumption imposed on** β^*;
- and the dimension p is **comparable to** the sample size n.

- Why coordinates?
- Why fixed designs?

Goals (Informal)

Goal (Informal): Make inference on the **coordinates** of β^* when

- X is treated as **fixed**;
- **no assumption imposed on** β^*;
- and the dimension p is **comparable to** the sample size n.

- Why coordinates?
- Why fixed designs?
- Why assumption-free β^*?

Goals (Informal)

Goal (Informal): Make inference on the **coordinates** of β^* when

- X is treated as **fixed**;
- **no assumption imposed on** β^*;
- and the dimension p is **comparable to** the sample size n.

- Why coordinates?
- Why fixed designs?
- Why assumption-free β^*?
- Why $p \sim n$?

Asymptotic Arguments: Motivation

- Consider β_1^* WLOG;

Asymptotic Arguments: Motivation

- Consider β_1^* WLOG;
- Ideally, we construct a 95% confidence interval for β_1^* as

$$\left[q_{0.025}\left(\mathcal{L}(\hat{\beta}_1)\right), q_{0.975}\left(\mathcal{L}(\hat{\beta}_1)\right)\right]$$

  where q_α denotes the α-th quantile;

Asymptotic Arguments: Motivation

- Consider β_1^* WLOG;
- Ideally, we construct a 95% confidence interval for β_1^* as

$$\left[q_{0.025}\left(\mathcal{L}(\hat{\beta}_1)\right), q_{0.975}\left(\mathcal{L}(\hat{\beta}_1)\right)\right]$$

  where q_α denotes the α-th quantile;
- Unfortunately, $\mathcal{L}(\hat{\beta}_1)$ is unknown.

Asymptotic Arguments: Motivation

- Consider β_1^* WLOG;
- Ideally, we construct a 95% confidence interval for β_1^* as
$$\left[q_{0.025}\left(\mathcal{L}(\hat{\beta}_1)\right), q_{0.975}\left(\mathcal{L}(\hat{\beta}_1)\right)\right]$$
where q_α denotes the α-th quantile;
- Unfortunately, $\mathcal{L}(\hat{\beta}_1)$ is unknown.
- This motivates the asymptotic arguments, i.e. find a distribution F s.t.
$$\mathcal{L}(\hat{\beta}_1) \approx F.$$

Asymptotic Arguments: Textbook Version

- The limiting behavior of $\hat{\beta}$ when p is fixed, as $n \to \infty$,

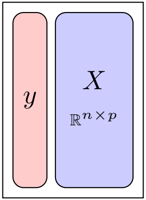
$$\mathcal{L}(\hat{\beta}) \to N\left(\beta^*, (X^T X)^{-1} \frac{\mathbb{E}(\psi^2(\epsilon_1))}{[\mathbb{E}\psi'(\epsilon_1)]^2}\right);$$

- As a consequence, we obtain an approximate 95% confidence interval for β_1^*,

$$\left[\hat{\beta}_1 - 1.96\widehat{\mathrm{sd}}(\hat{\beta}_1), \hat{\beta}_1 + 1.96\widehat{\mathrm{sd}}(\hat{\beta}_1)\right]$$

where $\widehat{\mathrm{sd}}(\hat{\beta}_1)$ could be any consistent estimator of the standard deviation.

Asymptotic Arguments: Hypothetical Problems

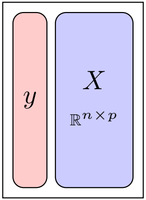


original problem

$(n = 100, p = 30)$

$y \sim X \Rightarrow \hat{\beta}_1$

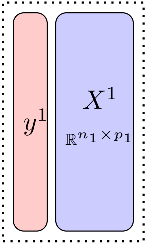
Asymptotic Arguments: Hypothetical Problems



original problem

$(n = 100, p = 30)$

$y \sim X \Rightarrow \hat{\beta}_1$

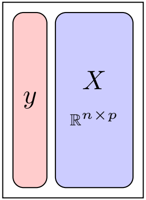


hypothetical problem

$(n_1 = 200, p_1 = 30)$

$y^1 \sim X^1 \Rightarrow \hat{\beta}_1^{(1)}$

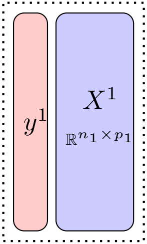
Asymptotic Arguments: Hypothetical Problems



original problem

$(n = 100, p = 30)$

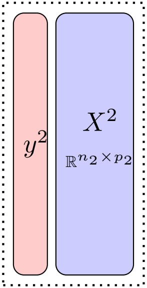
$y \sim X \Rightarrow \hat{\beta}_1$



hypothetical problem

$(n_1 = 200, p_1 = 30)$

$y^1 \sim X^1 \Rightarrow \hat{\beta}_1^{(1)}$

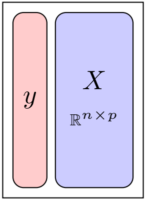


hypothetical problem

$(n_2 = 500, p_2 = 30)$

$y^2 \sim X^2 \Rightarrow \hat{\beta}_1^{(2)}$

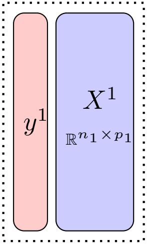
Asymptotic Arguments: Hypothetical Problems



original problem

$(n = 100, p = 30)$

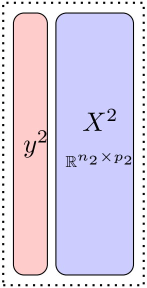
$y \sim X \Rightarrow \hat{\beta}_1$



hypothetical problem

$(n_1 = 200, p_1 = 30)$

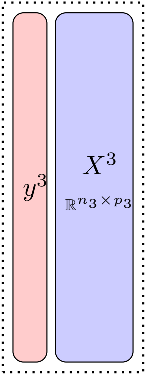
$y^1 \sim X^1 \Rightarrow \hat{\beta}_1^{(1)}$



hypothetical problem

$(n_2 = 500, p_2 = 30)$

$y^2 \sim X^2 \Rightarrow \hat{\beta}_1^{(2)}$

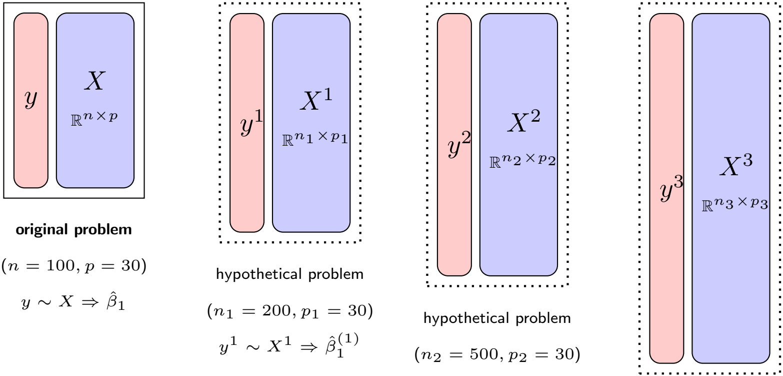


hypothetical problem

$(n_3 = 2000, p_3 = 30)$

$y^3 \sim X^3 \Rightarrow \hat{\beta}_1^{(3)}$

Asymptotic Arguments: Hypothetical Problems



original problem

$(n = 100, p = 30)$

$y \sim X \Rightarrow \hat{\beta}_1$

hypothetical problem

$(n_1 = 200, p_1 = 30)$

$y^1 \sim X^1 \Rightarrow \hat{\beta}_1^{(1)}$

hypothetical problem

$(n_2 = 500, p_2 = 30)$

$y^2 \sim X^2 \Rightarrow \hat{\beta}_1^{(2)}$

hypothetical problem

$(n_3 = 2000, p_3 = 30)$

$y^3 \sim X^3 \Rightarrow \hat{\beta}_1^{(3)}$

Asymptotic argument: use $\lim_{j\to\infty} \mathcal{L}(\hat{\beta}_1^{(j)})$ to approximate $\mathcal{L}(\hat{\beta}_1)$.

Asymptotic Arguments

- Huber [1973] raised the question of understanding the behavior of $\hat{\beta}$ when both n and p tend to infinity;

Asymptotic Arguments

- Huber [1973] raised the question of understanding the behavior of $\hat{\beta}$ when both n and p tend to infinity;
- Huber [1973] showed the L_2 consistency of $\hat{\beta}$:

$$\|\hat{\beta} - \beta^*\|_2^2 \to 0, \quad \text{when } p = o(n^{\frac{1}{3}});$$

Asymptotic Arguments

- Huber [1973] raised the question of understanding the behavior of $\hat{\beta}$ when both n and p tend to infinity;
- Huber [1973] showed the L_2 consistency of $\hat{\beta}$:

$$\|\hat{\beta} - \beta^*\|_2^2 \to 0, \quad \text{when } p = o(n^{\frac{1}{3}});$$

- Portnoy [1984] prove the L_2 consistency of $\hat{\beta}$ when

$$p = o\left(\frac{n}{\log n}\right).$$

Asymptotic Arguments

- Portnoy [1985] and Mammen [1989] showed that $\hat{\beta}$ is **jointly asymptotically normal** when

$$p \ll n^{\frac{2}{3}},$$

Asymptotic Arguments

- Portnoy [1985] and Mammen [1989] showed that $\hat{\beta}$ is **jointly asymptotically normal** when

$$p \ll n^{\frac{2}{3}},$$

in the sense that for any sequence of vectors $a_n \in \mathbb{R}^p$,

$$\mathcal{L}\left(\frac{a_n^T(\hat{\beta} - \beta^*)}{\sqrt{\mathrm{Var}(a_n^T \hat{\beta})}}\right) \to N(0, 1)$$

p/n: A Measure of Difficulty

All of the above works requires

$$p/n \to 0 \quad \text{or} \quad n/p \to \infty.$$

p/n: A Measure of Difficulty

All of the above works requires

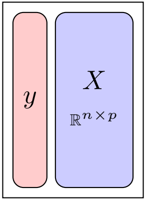
$$p/n \to 0 \text{ or } n/p \to \infty.$$

- n/p is **the number of samples per parameter**;
- Classical rule of thumb: $n/p \geq 5 \sim 10$;
- Heuristically, a larger n/p would give an easier problem;
- Hypothetical problems with $n_j/p_j \to \infty$ are *not appropriate* because they are increasingly easier than the original problem.

Moderate p/n Regime

Formally, we define **Moderate $\mathbf{p/n}$ Regime** as

$$p/n \to \kappa > 0.$$



original problem

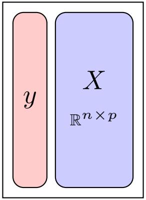
$(n = 100, p = 30)$

$y \sim X \Rightarrow \hat{\beta}_1$

Moderate p/n Regime

Formally, we define **Moderate $\mathbf{p/n}$ Regime** as

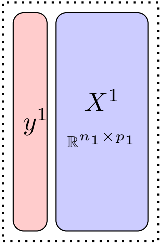
$$p/n \to \kappa > 0.$$



original problem

$(n = 100, p = 30)$

$y \sim X \Rightarrow \hat{\beta}_1$



hypothetical problem

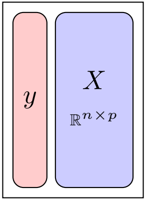
$(n_1 = 200, p_1 = 60)$

$y^1 \sim X^1 \Rightarrow \hat{\beta}_1^{(1)}$

Moderate p/n Regime

Formally, we define **Moderate $\mathbf{p/n}$ Regime** as

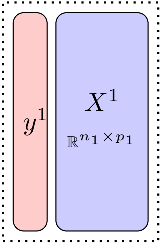
$$p/n \to \kappa > 0.$$



original problem

$(n = 100, p = 30)$

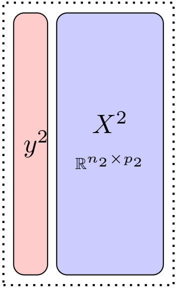
$y \sim X \Rightarrow \hat{\beta}_1$



hypothetical problem

$(n_1 = 200, p_1 = 60)$

$y^1 \sim X^1 \Rightarrow \hat{\beta}_1^{(1)}$



hypothetical problem

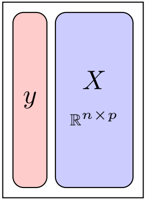
$(n_2 = 500, p_2 = 150)$

$y^2 \sim X^2 \Rightarrow \hat{\beta}_1^{(2)}$

Moderate p/n Regime

Formally, we define **Moderate $\mathbf{p}/\mathbf{n}$ Regime** as

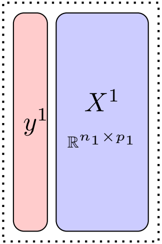
$$p/n \to \kappa > 0.$$



original problem

$(n = 100, p = 30)$

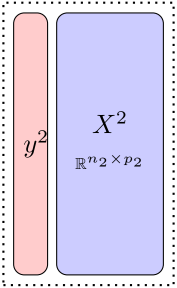
$y \sim X \Rightarrow \hat{\beta}_1$



hypothetical problem

$(n_1 = 200, p_1 = 60)$

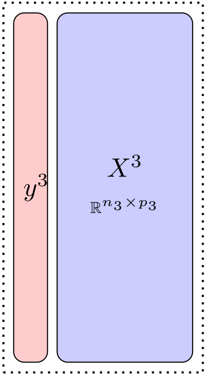
$y^1 \sim X^1 \Rightarrow \hat{\beta}_1^{(1)}$



hypothetical problem

$(n_2 = 500, p_2 = 150)$

$y^2 \sim X^2 \Rightarrow \hat{\beta}_1^{(2)}$



hypothetical problem

$(n_3 = 2000, p_3 = 600)$

$y^3 \sim X^3 \Rightarrow \hat{\beta}_1^{(3)}$

Moderate p/n Regime: More Informative Asymptotics

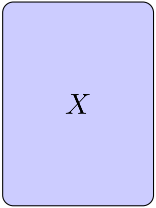
A simulation to compare Fix-p Regime and Moderate p/n Regime:

Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.

Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

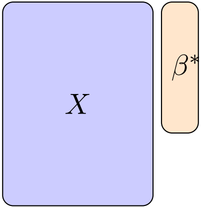
Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.



Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

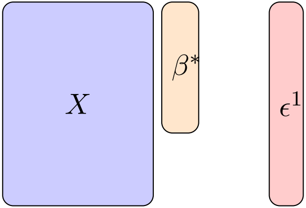
Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.



Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.



Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.

$$y^1 = X \beta^* + \epsilon^1$$

Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.

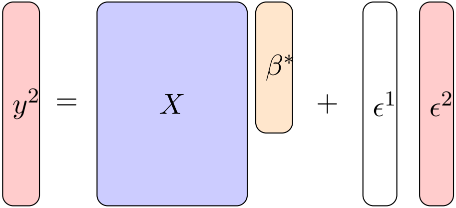
$$y^1 = X \beta^* + \epsilon^1$$

M-Estimates: $\hat{\beta}_1^{(1)}$,

Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.

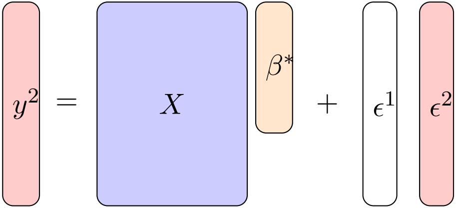


M-Estimates: $\hat{\beta}_1^{(1)}$,

Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.

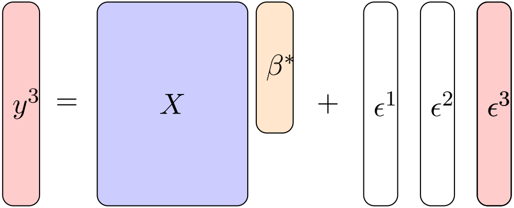


M-Estimates: $\hat{\beta}_1^{(1)}$, $\hat{\beta}_1^{(2)}$,

Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.

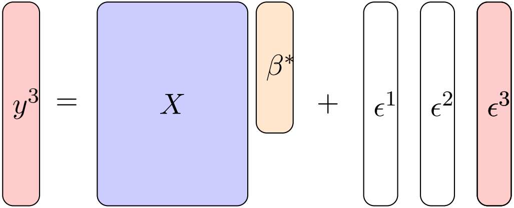


M-Estimates: $\hat{\beta}_1^{(1)}$, $\hat{\beta}_1^{(2)}$,

Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.

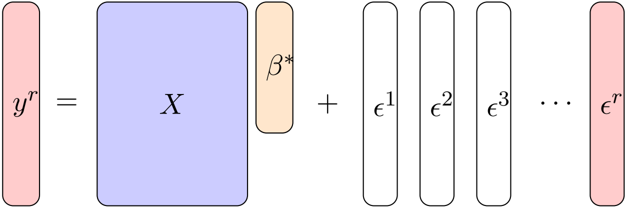


M-Estimates: $\hat{\beta}_1^{(1)}$, $\hat{\beta}_1^{(2)}$, $\hat{\beta}_1^{(3)}$,

Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.



M-Estimates: $\hat{\beta}_1^{(1)}$, $\hat{\beta}_1^{(2)}$, $\hat{\beta}_1^{(3)}$,

Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.

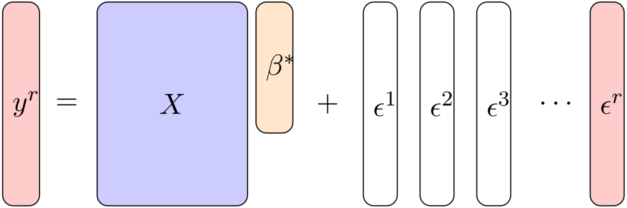
$$y^r = X \beta^* + \quad \epsilon^1 \quad \epsilon^2 \quad \epsilon^3 \quad \cdots \quad \epsilon^r$$

M-Estimates: $\hat{\beta}_1^{(1)}, \hat{\beta}_1^{(2)}, \hat{\beta}_1^{(3)}, \ \ldots \ , \hat{\beta}_1^{(r)}$.

Moderate p/n Regime: More Informative Asymptotics

A simulation to compare Fix-p Regime and Moderate p/n Regime:

Original problem: $n = 50$, $p = 50\kappa$, Huber loss, i.i.d. ϵ_i's.



M-Estimates: $\hat{\beta}_1^{(1)}, \hat{\beta}_1^{(2)}, \hat{\beta}_1^{(3)}, \ \ldots \ , \hat{\beta}_1^{(r)}$.

$$\Longrightarrow \hat{\mathcal{L}}(\hat{\beta}_1; X) = \mathrm{ecdf}(\{\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(r)}\}).$$

Moderate p/n Regime: More Informative Asymptotics

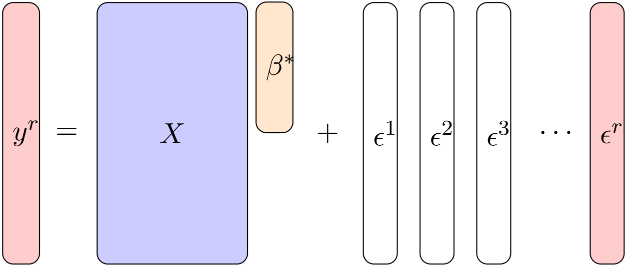
A Simulation to compare Fix-p Regime and Moderate p/n Regime:

Fix-p Approximation: $n = 1000$, $p = 50\kappa$.

Moderate p/n Regime: More Informative Asymptotics

A Simulation to compare Fix-p Regime and Moderate p/n Regime:

Fix-p Approximation: $n = 1000$, $p = 50\kappa$.



M-Estimates: $\hat{\beta}_1^{(F,1)}$, $\hat{\beta}_1^{(F,2)}$, $\hat{\beta}_1^{(F,3)}$, $\ldots$, $\hat{\beta}_1^{(F,r)}$.

$$\Longrightarrow \hat{\mathcal{L}}(\hat{\beta}_1^F; X) = \mathrm{ecdf}(\{\hat{\beta}_1^{(F,1)}, \ldots, \hat{\beta}_1^{(F,r)}\}).$$

Moderate p/n Regime: More Informative Asymptotics

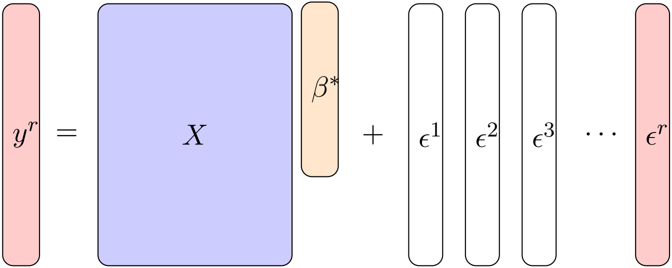
A Simulation to compare Fix-p Regime and Moderate p/n Regime:

Moderate-p/n Approximation: $n = 1000$, $p = 1000\kappa$.

Moderate p/n Regime: More Informative Asymptotics

A Simulation to compare Fix-p Regime and Moderate p/n Regime:

Moderate-p/n Approximation: $n = 1000$, $p = 1000\kappa$.



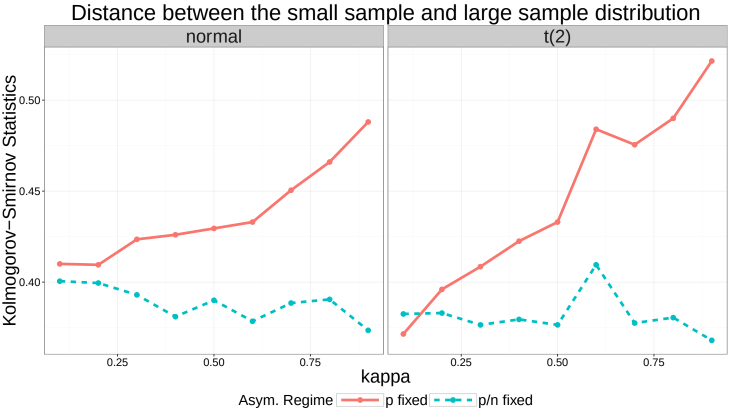
M-Estimates: $\hat{\beta}_1^{(M,1)}$, $\hat{\beta}_1^{(M,2)}$, $\hat{\beta}_1^{(M,3)}$, $\ldots$, $\hat{\beta}_1^{(M,r)}$.

$$\Longrightarrow \hat{\mathcal{L}}(\hat{\beta}_1^M; X) = \mathrm{ecdf}(\{\hat{\beta}_1^{(M,1)}, \ldots, \hat{\beta}_1^{(M,r)}\}).$$

Moderate p/n Regime: More Informative Asymptotics

Measure the accuracy of two approximations by the Kolmogorov-Smirnov statistics

$$d_{KS}\left(\hat{\mathcal{L}}(\hat{\beta}_1), \hat{\mathcal{L}}(\hat{\beta}_1^F)\right) \text{ and } d_{KS}\left(\hat{\mathcal{L}}(\hat{\beta}_1), \hat{\mathcal{L}}(\hat{\beta}_1^M)\right)$$



Moderate p/n Regime: Negative Results

The moderate p/n regime in statistics:

Moderate p/n Regime: Negative Results

The moderate p/n regime in statistics:

- Huber [1973] showed that for least-square estimators there always exists a sequence of vectors $a_n \in \mathbb{R}^p$ such that

$$\mathcal{L}\left(\frac{a_n^T(\hat{\beta}^{LS} - \beta^*)}{\sqrt{\mathrm{Var}(a_n^T \hat{\beta}^{LS})}}\right) \not\to N(0,1).$$

Moderate p/n Regime: Negative Results

The moderate p/n regime in statistics:

- Huber [1973] showed that for least-square estimators there always exists a sequence of vectors $a_n \in \mathbb{R}^p$ such that

$$\mathcal{L}\left(\frac{a_n^T(\hat{\beta}^{LS} - \beta^*)}{\sqrt{\mathrm{Var}(a_n^T \hat{\beta}^{LS})}}\right) \not\to N(0,1).$$

- Bickel and Freedman [1982] showed that the bootstrap fails in the Least-Square case and the usual rescaling does not help;

Moderate p/n Regime: Negative Results

The moderate p/n regime in statistics:

- Huber [1973] showed that for least-square estimators there always exists a sequence of vectors $a_n \in \mathbb{R}^p$ such that

$$\mathcal{L}\left(\frac{a_n^T(\hat{\beta}^{LS} - \beta^*)}{\sqrt{\mathrm{Var}(a_n^T\hat{\beta}^{LS})}}\right) \not\to N(0,1).$$

- Bickel and Freedman [1982] showed that the bootstrap fails in the Least-Square case and the usual rescaling does not help;
- El Karoui et al. [2011] showed that for general loss functions,

$$\|\hat{\beta} - \beta^*\|_2^2 \not\to 0.$$

Moderate p/n Regime: Negative Results

The moderate p/n regime in statistics:

- Huber [1973] showed that for least-square estimators there always exists a sequence of vectors $a_n \in \mathbb{R}^p$ such that

$$\mathcal{L}\left(\frac{a_n^T(\hat{\beta}^{LS} - \beta^*)}{\sqrt{\mathrm{Var}(a_n^T \hat{\beta}^{LS})}}\right) \not\to N(0,1).$$

- Bickel and Freedman [1982] showed that the bootstrap fails in the Least-Square case and the usual rescaling does not help;
- El Karoui et al. [2011] showed that for general loss functions,

$$\|\hat{\beta} - \beta^*\|_2^2 \not\to 0.$$

- El Karoui and Purdom [2015] showed that most widely used resampling schemes give poor inference on β_1^*.

Moderate p/n Regime: Reason of Failure

Qualitatively,

- Influential observation *always* exists [Huber, 1973]: let $H = X(X^T X)^{-1} X^T$ be the hat matrix,

$$\max_i H_{i,i} \geq \frac{1}{n} \operatorname{tr}(H) = \frac{p}{n} \gg 0.$$

Moderate p/n Regime: Reason of Failure

Qualitatively,

- Influential observation *always* exists [Huber, 1973]: let $H = X(X^T X)^{-1} X^T$ be the hat matrix,

$$\max_i H_{i,i} \geq \frac{1}{n} \operatorname{tr}(H) = \frac{p}{n} \gg 0.$$

- Regression residuals fail to mimic true error:

$$R_i \triangleq y_i - x_i^T \hat{\beta} \not\approx \epsilon_i.$$

Moderate p/n Regime: Reason of Failure

Qualitatively,

- Influential observation *always* exists [Huber, 1973]: let $H = X(X^TX)^{-1}X^T$ be the hat matrix,

$$\max_i H_{i,i} \geq \frac{1}{n}\operatorname{tr}(H) = \frac{p}{n} \gg 0.$$

- Regression residuals fail to mimic true error:

$$R_i \triangleq y_i - x_i^T\hat{\beta} \not\approx \epsilon_i.$$

Technically,

- Taylor expansion/Bahadur-type representation fails!

Moderate p/n Regime: Positive Results (Random Designs)

- Bean et al. [2013] showed that when X has i.i.d. Gaussian entries, for any sequence of $a_n \in \mathbb{R}^p$

$$\mathcal{L}_{X,\epsilon}\left(\frac{a_n^T(\hat{\beta} - \beta^*)}{\sqrt{\mathrm{Var}_{X,\epsilon}(a_n^T\hat{\beta})}}\right) \to N(0,1);$$

Moderate p/n Regime: Positive Results (Random Designs)

- Bean et al. [2013] showed that when X has i.i.d. Gaussian entries, for any sequence of $a_n \in \mathbb{R}^p$

$$\mathcal{L}_{X,\epsilon}\left(\frac{a_n^T(\hat{\beta}-\beta^*)}{\sqrt{\mathrm{Var}_{X,\epsilon}(a_n^T\hat{\beta})}}\right) \to N(0,1);$$

- El Karoui [2015] extended it to general random designs.

Moderate p/n Regime: Positive Results (Random Designs)

- Bean et al. [2013] showed that when X has i.i.d. Gaussian entries, for any sequence of $a_n \in \mathbb{R}^p$

$$\mathcal{L}_{X,\epsilon}\left(\frac{a_n^T(\hat{\beta} - \beta^*)}{\sqrt{\mathrm{Var}_{X,\epsilon}(a_n^T\hat{\beta})}}\right) \to N(0,1);$$

- El Karoui [2015] extended it to general random designs.
- The above result does not contradict Huber [1973] in that the randomness comes from both X and ϵ;

Moderate p/n Regime: Positive Results (Random Designs)

- Bean et al. [2013] showed that when X has i.i.d. Gaussian entries, for any sequence of $a_n \in \mathbb{R}^p$

$$\mathcal{L}_{X,\epsilon}\left(\frac{a_n^T(\hat{\beta} - \beta^*)}{\sqrt{\mathrm{Var}_{X,\epsilon}(a_n^T\hat{\beta})}}\right) \to N(0,1);$$

- El Karoui [2015] extended it to general random designs.
- The above result does not contradict Huber [1973] in that the randomness comes from both X and ϵ;
- El Karoui et al. [2011] showed that for general loss functions,

$$\|\hat{\beta} - \beta^*\|_\infty \to 0.$$

Moderate p/n Regime: Summary

- Provides a more accurate approximation of $\mathcal{L}(\hat{\beta}_1)$;

Moderate p/n Regime: Summary

- Provides a more accurate approximation of $\mathcal{L}(\hat{\beta}_1)$;
- Qualitatively different from the classical regimes where $p/n \to 0$;
  - L_2-consistency of $\hat{\beta}$ no longer holds;
  - the residual R_i behaves differently from ϵ_i;
  - fixed design results are different from random design results.

Moderate p/n Regime: Summary

- Provides a more accurate approximation of $\mathcal{L}(\hat{\beta}_1)$;
- Qualitatively different from the classical regimes where $p/n \to 0$;
  - L_2-consistency of $\hat{\beta}$ no longer holds;
  - the residual R_i behaves differently from ϵ_i;
  - fixed design results are different from random design results.
- Inference on the vector $\hat{\beta}$ is hard; but inference on the coordinate / low-dimensional linear contrasts of $\hat{\beta}$ is still possible.

Goals (Formal)

Our Goal (formal): Under the **linear model**

$$Y = X\beta^* + \epsilon,$$

Derive the asymptotic distribution of **coordinates** $\hat{\beta}_j$:

- under the **moderate $\mathbf{p/n}$ regime**, i.e. $p/n \to \kappa \in (0, 1)$;
- with a **fixed design** matrix X;
- **without assumptions on** β^*.

Table of Contents

Main Result (Informal)

Definition 1.

Let P and Q be two distributions on $\mathbb{R}^p$,

$$d_{\mathrm{TV}}(P, Q) = \sup_{A \subset \mathbb{R}^p} |P(A) - Q(A)|.$$

Main Result (Informal)

Definition 1.

Let P and Q be two distributions on $\mathbb{R}^p$,

$$d_{\mathrm{TV}}\left(P, Q\right) = \sup_{A \subset \mathbb{R}^p} |P(A) - Q(A)|.$$

Theorem.

Under appropriate conditions on the design matrix X, the distribution of ϵ and the loss function ρ, as $p/n \to \kappa \in (0,1)$, while $n \to \infty$,

$$\max_j d_{\mathrm{TV}}\left(\mathcal{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = o(1).$$

Main Result (Informal)

If ρ is an even function and $\epsilon \stackrel{d}{=} -\epsilon$, then

$$\hat{\beta} - \beta^* \stackrel{d}{=} \beta^* - \hat{\beta} \Longrightarrow \mathbb{E}\hat{\beta} = \beta^*.$$

Theorem.

Under appropriate conditions on the design matrix X, the distribution of ϵ and the loss function ρ, as $p/n \to \kappa \in (0,1)$, while $n \to \infty$,

$$\max_j d_{\mathrm{TV}} \left(\mathcal{L}\left(\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} \right), N(0,1) \right) = o(1).$$

Why Surprising?

Classical approaches heavily rely on

- L_2 consistency of $\hat{\beta}$, which only holds when $p = o(n)$;
- Bahadur-type representation for $\hat{\beta}$ where

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i + o_p\left(\frac{1}{\sqrt{n}}\right),$$

  for some i.i.d. random variable Z_i's;
- which can be proved only when $p = o\left(n^{2/3}\right)$;

Why Surprising?

Classical approaches heavily rely on

- L_2 consistency of $\hat{\beta}$, which only holds when $p = o(n)$;
- Bahadur-type representation for $\hat{\beta}$ where

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i + o_p\left(\frac{1}{\sqrt{n}}\right),$$

  for some i.i.d. random variable Z_i's;
- which can be proved only when $p = o\left(n^{2/3}\right)$;

Question: What happens when $p \in [O(n^{2/3}), O(n)]$?

Our Contributions and Limitations

Instead, we develops a novel strategy that is built on

- Leave-on-out method [El Karoui et al., 2011];
- and Second-Order Poincaré Inequality [Chatterjee, 2009].

Our Contributions and Limitations

Instead, we develops a novel strategy that is built on

- Leave-on-out method [El Karoui et al., 2011];
- and Second-Order Poincaré Inequality [Chatterjee, 2009].

We prove that

- $\hat{\beta}_1$ is asymptotically normal for all $p \in [O(1), O(n)]$ for *fixed designs* under regularity conditions;
- the conditions are satisfied by "most" design matrices.

Our Contributions and Limitations

Instead, we develops a novel strategy that is built on

- Leave-on-out method [El Karoui et al., 2011];
- and Second-Order Poincaré Inequality [Chatterjee, 2009].

We prove that

- $\hat{\beta}_1$ is asymptotically normal for all $p \in [O(1), O(n)]$ for *fixed designs* under regularity conditions;
- the conditions are satisfied by "most" design matrices.

Limitations:

- we impose strong conditions on ρ and $\mathcal{L}(\epsilon)$;
- we do not know how to estimate $\mathrm{Var}_\epsilon(\hat{\beta}_1)$.

Examples: Realization of i.i.d. Designs

We consider the case where X is **a realization** of a random design Z. The examples below are proved to **satisfy the technical assumptions with high probability** over Z.

Examples: Realization of i.i.d. Designs

We consider the case where X is **a realization** of a random design Z. The examples below are proved to **satisfy the technical assumptions with high probability** over Z.

Example 1 Z has i.i.d. mean-zero sub-gaussian entries with $\mathrm{Var}(Z_{ij}) = \tau^2 > 0$;

Example 2 Z contains an intercept term, i.e. $Z = (\mathbf{1}, \tilde{Z})$ and $\tilde{Z} \in \mathbb{R}^{n\times(p-1)}$ has independent sub-gaussian entries with

$$\tilde{Z}_{ij} - \mu_j \stackrel{d}{=} \mu_j - \tilde{Z}_{ij}, \quad \mathrm{Var}(\tilde{Z}_{ij}) > \tau^2$$

for some arbitrary μ_j's.

A Counter-Example

Consider a one-way ANOVA situation. Each observation i is associated with a label $k_i \in \{1, \ldots, p\}$ and let $X_{i,j} = I(j = k_i)$. This is equivalent to

$$Y_i = \beta^*_{k_i} + \epsilon_i.$$

A Counter-Example

Consider a one-way ANOVA situation. Each observation i is associated with a label $k_i \in \{1, \ldots, p\}$ and let $X_{i,j} = I(j = k_i)$. This is equivalent to

$$Y_i = \beta^*_{k_i} + \epsilon_i.$$

It is easy to see that

$$\hat{\beta}_j = \arg\min_{\beta \in \mathbb{R}} \sum_{i:k_i=j} \rho(y_i - \beta_j).$$

This is a standard location problem.

A Counter-Example

Let $n_j = |\{i : k_i = j\}|$. In the least-square case, i.e. $\rho(x) = x^2/2$,

$$\hat{\beta}_j = \beta_j^* + \frac{1}{n_j} \sum_{i:k_i=j} \epsilon_i.$$

A Counter-Example

Let $n_j = |\{i : k_i = j\}|$. In the least-square case, i.e. $\rho(x) = x^2/2$,

$$\hat{\beta}_j = \beta_j^* + \frac{1}{n_j} \sum_{i:k_i=j} \epsilon_i.$$

Assume a balance design, i.e. $n_j \approx n/p$. Then $n_j \ll \infty$ and

- none of $\hat{\beta}_j$ is normal (unless ϵ_i are normal);
- holds for general loss functions ρ.

A Counter-Example

Let $n_j = |\{i : k_i = j\}|$. In the least-square case, i.e. $\rho(x) = x^2/2$,

$$\hat{\beta}_j = \beta_j^* + \frac{1}{n_j} \sum_{i:k_i=j} \epsilon_i.$$

Assume a balance design, i.e. $n_j \approx n/p$. Then $n_j \ll \infty$ and

- none of $\hat{\beta}_j$ is normal (unless ϵ_i are normal);
- holds for general loss functions ρ.

Conclusion: some "non-standard" assumptions on X are required.

Table of Contents

Least Square Estimator

The L_2 loss, $\rho(x) = x^2/2$, gives the least-square estimator

$$\hat{\beta}^{LS} = (X^T X)^{-1} X^T Y = \beta^* + (X^T X)^{-1} X^T \epsilon.$$

Least Square Estimator

The L_2 loss, $\rho(x) = x^2/2$, gives the least-square estimator

$$\hat{\beta}^{LS} = (X^T X)^{-1} X^T Y = \beta^* + (X^T X)^{-1} X^T \epsilon.$$

Let e_j denote the canonical basis vector in $\mathbb{R}^p$, then

$$\hat{\beta}_j^{LS} - \beta_j^* = e_j^T (X^T X)^{-1} X^T \epsilon \triangleq \alpha_j^T \epsilon.$$

Least Square Estimator

Lindeberg-Feller CLT claims that in order for

$$\mathcal{L}\left(\frac{\hat{\beta}_j^{LS} - \beta_j^*}{\sqrt{\mathrm{Var}(\hat{\beta}_j^{LS})}}\right) \to N(0,1)$$

it is **sufficient and almost necessary** that

$$\frac{\|\alpha_j\|_\infty}{\|\alpha_j\|_2} \to 0. \tag{1}$$

Least Square Estimator

To see the necessity of the condition, recall the one-way ANOVA case. Let $n_j = |\{i : k_i = j\}|$, then

$$X^T X = \mathrm{diag}(n_j)_{j=1}^{p}.$$

Recall that $\alpha_j^T = e_j^T (X^T X)^{-1} X^T$. This gives

$$\alpha_{j,i} = \begin{cases} \frac{1}{n_j} & \text{if } k_i = j \\ 0 & \text{if } k_i \neq j \end{cases}$$

Least Square Estimator

To see the necessity of the condition, recall the one-way ANOVA case. Let $n_j = |\{i : k_i = j\}|$, then

$$X^T X = \mathrm{diag}(n_j)_{j=1}^p.$$

Recall that $\alpha_j^T = e_j^T (X^T X)^{-1} X^T$. This gives

$$\alpha_{j,i} = \begin{cases} \frac{1}{n_j} & \text{if } k_i = j \\ 0 & \text{if } k_i \neq j \end{cases}$$

As a result, $\|\alpha_j\|_\infty = \frac{1}{n_j}, \|\alpha_j\|_2 = \frac{1}{\sqrt{n_j}}$ and hence

$$\frac{\|\alpha_j\|_\infty}{\|\alpha_j\|_2} = \frac{1}{\sqrt{n_j}}$$

However, in moderate p/n regime, there exists j such that $n_j \le 1/\kappa$ and thus $\hat{\beta}_j^{LS}$ is not asymptotically normal.

M-Estimator

The result for LSE is derived from the analytical form of $\hat{\beta}^{LS}$. By contrast, an analytical form is not available for general ρ.

M-Estimator

The result for LSE is derived from the analytical form of $\hat{\beta}^{LS}$. By contrast, an analytical form is not available for general ρ.

Let $\psi = \rho'$, it is the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\psi(y_i - x_i^T\hat{\beta}) = 0 \Longleftrightarrow \frac{1}{n}\sum_{i=1}^{n}\psi(\epsilon_i - x_i^T(\hat{\beta} - \beta^*)) = 0.$$

We show that

- $\hat{\beta}_j$ is a smooth function of ϵ;
- $\frac{\partial\hat{\beta}_j}{\partial\epsilon}$ and $\frac{\partial\hat{\beta}_j}{\partial\epsilon\partial\epsilon^T}$ are computable.

Second-Order Poincaré Inequality

$\hat{\beta}_j$ is a smooth transform of a random vector, ϵ, with independent entries. A powerful CLT for this type of statistics is Second-Order Poincaré Inequality [Chatterjee, 2009].

Second-Order Poincaré Inequality

$\hat{\beta}_j$ is a smooth transform of a random vector, ϵ, with independent entries. A powerful CLT for this type of statistics is Second-Order Poincaré Inequality [Chatterjee, 2009].

Definition 2.

For each $c_1, c_2 > 0$, let $L(c_1, c_2)$ be the class of probability measures on $\mathbb{R}$ that arise as laws of random variables like $u(W)$, where $W \sim N(0, 1)$ and $u \in C^2(\mathbb{R}^n)$ with

$$|u'(x)| \le c_1 \text{ and } |u''(x)| \le c_2.$$

For example, $u = \mathrm{Id}$ gives $N(0, 1)$ and $u = \Phi$ gives $U([0, 1])$.

Second-Order Poincaré Inequality

Proposition 1 (SOPI; Chatterjee [2009]).

Let $\mathscr{W} = (\mathscr{W}_1, \dots, \mathscr{W}_n) \overset{indep.}{\sim} L(c_1, c_2)$. *Take any* $g \in C^2(\mathbb{R}^n)$ *and let* $U = g(\mathscr{W})$,

$$\kappa_1 = (\mathbb{E}\|\nabla g(\mathscr{W})\|_2^4)^{\frac{1}{4}};$$
$$\kappa_2 = (\mathbb{E}\|\nabla^2 g(\mathscr{W})\|_{op}^4)^{\frac{1}{4}};$$
$$\kappa_0 = (\mathbb{E}\sum_{i=1}^n \left|\nabla_i g(\mathscr{W})\right|^4)^{\frac{1}{2}}.$$

If $\mathbb{E}U^4 < \infty$, *then*

$$d_{\mathrm{TV}}\left(\mathcal{L}\left(\frac{U - \mathbb{E}U}{\sqrt{\mathrm{Var}(U)}}\right), N(0,1)\right) \preceq \frac{\kappa_0 + \kappa_1\kappa_2}{\mathrm{Var}(U)}.$$

Assumptions

A1 $\rho(0) = \psi(0) = 0$ and for any $x \in \mathbb{R}$,

$$0 < K_0 \le \psi'(x) \le K_1, \quad |\psi''(x)| \le K_2;$$

A2 ϵ has independent entries with $\epsilon_i \in L(c_1, c_2)$;

A3 Let λ_+ and λ_- be the largest and smallest eigenvalues of $X^T X/n$ and

$$\lambda_+ = O(1), \quad \lambda_- = \Omega(1).$$

A4 "Similar to" the condition for OLS:

$$\max_j \frac{\|e_j^T (X^T X)^{-1} X^T\|_\infty}{\|e_j^T (X^T X)^{-1} X^T\|_2} = o(1)$$

A5 "Similar to" the condition that

$$\min_j \mathrm{Var}(\hat{\beta}_j) = \Omega\left(\frac{1}{n}\right)$$

Main Results

Theorem 3.

Under assumptions $\mathbf{A}1 - \mathbf{A}5$*, as* $p/n \to \kappa$ *for some* $\kappa \in (0,1)$ *while* $n \to \infty$*,*

$$\max_j d_{\mathrm{TV}}\left(\mathcal{L}\left(\frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}\right), N(0,1)\right) = o(1).$$

Table of Contents

Setup

Design matrix $\mathbf{X}$:

- (i.i.d. design): $X_{ij} \overset{i.i.d.}{\sim} F$;
- (partial Hadamard design): a matrix formed by a random set of p columns of a $n \times n$ Hadamard matrix.

Entry Distribution $\mathbf{F}$:

- $F = N(0, 1)$;
- $F = \mathrm{t}_2$.

Error Distribution $\mathcal{L}(\epsilon)$: ϵ_i are i.i.d. with

- $\epsilon_i \sim N(0, 1)$;
- $\epsilon_i \sim \mathrm{t}_2$.

Setup

Sample Size $\mathbf{n}$: $\{100, 200, 400, 800\}$;

$\kappa = \mathbf{p}/\mathbf{n}$: $\{0.5, 0.8\}$;

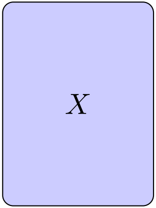
Loss Function ρ: Huber loss with $k = 1.345$,

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & |x| \leq k \\ kx - \frac{k^2}{2} & |x| > k \end{cases};$$

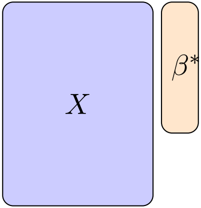
Coefficients: $\beta^* = 0$.

Asymptotic Normality of A Single Coordinate

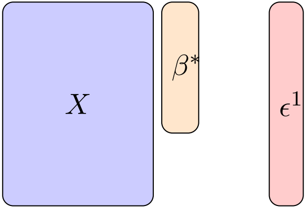
Asymptotic Normality of A Single Coordinate



Asymptotic Normality of A Single Coordinate



Asymptotic Normality of A Single Coordinate



Asymptotic Normality of A Single Coordinate

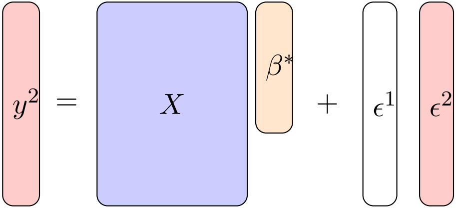
$$y^1 = X \beta^* + \epsilon^1$$

Asymptotic Normality of A Single Coordinate

$$y^1 = X \beta^* + \epsilon^1$$

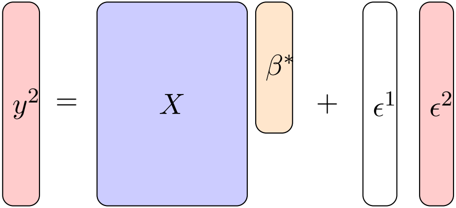
M-Estimates: $\hat{\beta}_1^{(1)}$,

Asymptotic Normality of A Single Coordinate



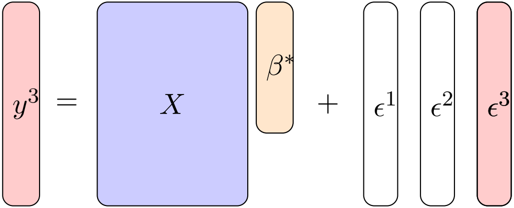
M-Estimates: $\hat{\beta}_1^{(1)}$,

Asymptotic Normality of A Single Coordinate



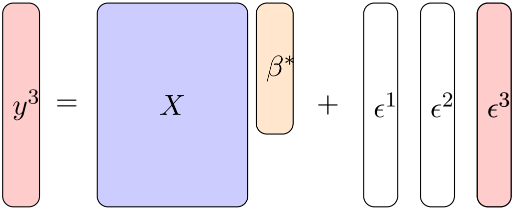
M-Estimates: $\hat{\beta}_1^{(1)}$, $\hat{\beta}_1^{(2)}$,

Asymptotic Normality of A Single Coordinate



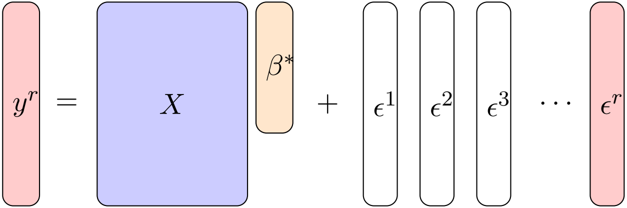
M-Estimates: $\hat{\beta}_1^{(1)}$, $\hat{\beta}_1^{(2)}$,

Asymptotic Normality of A Single Coordinate



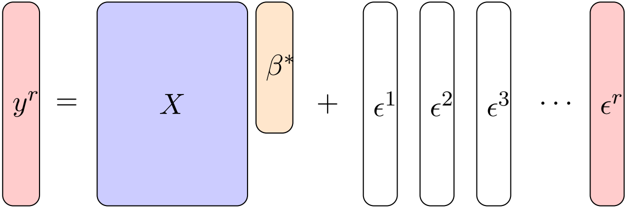
M-Estimates: $\hat{\beta}_1^{(1)}$, $\hat{\beta}_1^{(2)}$, $\hat{\beta}_1^{(3)}$,

Asymptotic Normality of A Single Coordinate



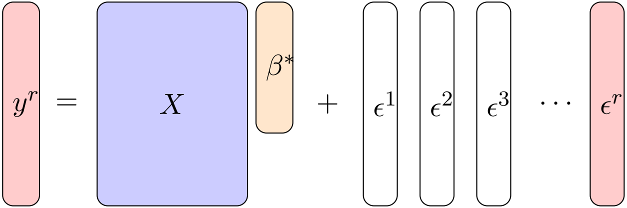
M-Estimates: $\hat{\beta}_1^{(1)}$, $\hat{\beta}_1^{(2)}$, $\hat{\beta}_1^{(3)}$,

Asymptotic Normality of A Single Coordinate



M-Estimates: $\hat{\beta}_1^{(1)}, \hat{\beta}_1^{(2)}, \hat{\beta}_1^{(3)}, \ \ldots \ , \hat{\beta}_1^{(r)}$.

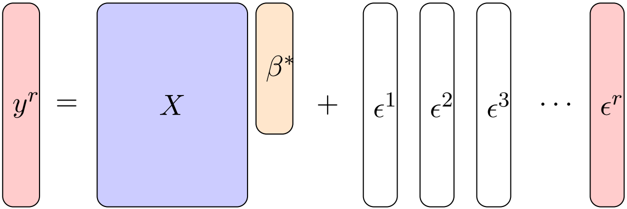
Asymptotic Normality of A Single Coordinate



M-Estimates: $\hat{\beta}_1^{(1)},\ \hat{\beta}_1^{(2)},\ \hat{\beta}_1^{(3)},\ \ \ldots\ \ ,\hat{\beta}_1^{(r)}$.

- $\widehat{sd} \leftarrow \mathrm{se}\left(\{\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(r)}\}\right)$;

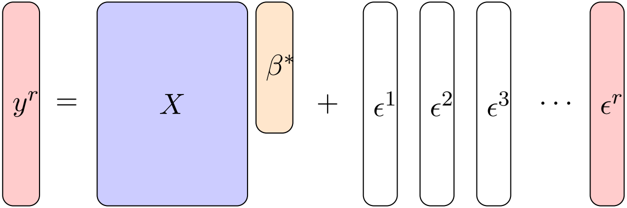
Asymptotic Normality of A Single Coordinate



M-Estimates: $\hat{\beta}_1^{(1)}, \hat{\beta}_1^{(2)}, \hat{\beta}_1^{(3)}, \ \ldots \ , \hat{\beta}_1^{(r)}$.

- $\widehat{sd} \leftarrow \mathrm{se}\left(\{\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(r)}\}\right)$;
- want to compare $\mathcal{L}\left(\hat{\beta}_1/\widehat{sd}\right)$ with $N(0,1)$;

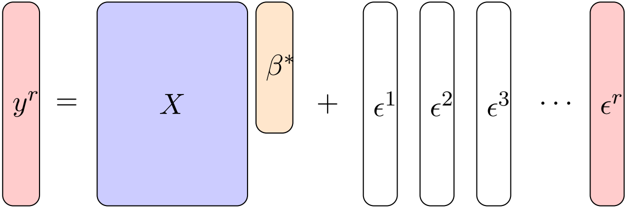
Asymptotic Normality of A Single Coordinate



M-Estimates: $\hat{\beta}_1^{(1)}, \hat{\beta}_1^{(2)}, \hat{\beta}_1^{(3)}, \ \ldots \ , \hat{\beta}_1^{(r)}$.

- $\widehat{sd} \leftarrow \mathrm{se}\left(\{\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(r)}\}\right)$;
- want to compare $\mathcal{L}\left(\hat{\beta}_1/\widehat{sd}\right)$ with $N(0,1)$;
- count the fraction of $\hat{\beta}_1^{(j)} \in [-1.96\widehat{sd}, 1.96\widehat{sd}]$ as the proxy;

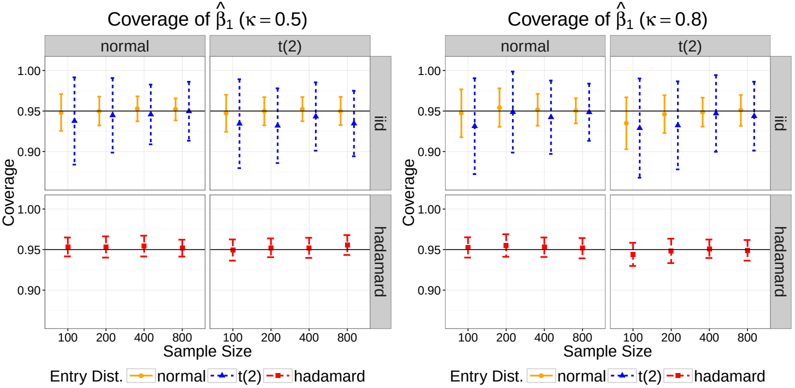
Asymptotic Normality of A Single Coordinate



M-Estimates: $\hat{\beta}_1^{(1)}, \hat{\beta}_1^{(2)}, \hat{\beta}_1^{(3)}, \ \ldots \ , \hat{\beta}_1^{(r)}$.

- $\widehat{sd} \leftarrow \mathrm{se}\left(\{\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(r)}\}\right)$;
- want to compare $\mathcal{L}\left(\hat{\beta}_1 / \widehat{sd}\right)$ with $N(0,1)$;
- count the fraction of $\hat{\beta}_1^{(j)} \in [-1.96\widehat{sd}, 1.96\widehat{sd}]$ as the proxy;
- should be close to 0.95 ideally.

Asymptotic Normality of A Single Coordinate



Conclusion

- We establish the **coordinate-wise asymptotic normality** of the M-estimator for certain **fixed design matrices** under the **moderate $\mathbf{p/n}$ regime** under regularity conditions on $X, \mathcal{L}(\epsilon)$ and ρ but **no condition on** β^*;
- We prove the result by using the novel approach Second-Order Poincaré Inequality [Chatterjee, 2009];
- We show that the regularity conditions are satisfied by a broad class of designs.

Discussion

Discussion

- Inference $\approx$ asym. normality + asym. bias + asym. variance
  - $\text{Var}(\hat{\beta}_1|X) \approx \text{Var}(\hat{\beta}_1)$ when X is indeed a realization of a random design?
  - Resampling method to give conservative variance estimates?
  - More advanced boostrap?

Discussion

- Inference ≈ asym. normality + asym. bias + asym. variance
    - $\mathrm{Var}(\hat{\beta}_1|X) \approx \mathrm{Var}(\hat{\beta}_1)$ when X is indeed a realization of a random design?
    - Resampling method to give conservative variance estimates?
    - More advanced boostrap?
- Relax the regularity conditions:
    - Generalize to non-strongly convex and non-smooth loss functions?
    - Generalize to general error distributions?

Discussion

- Inference $\approx$ asym. normality + asym. bias + asym. variance
  - $\mathrm{Var}(\hat{\beta}_1|X) \approx \mathrm{Var}(\hat{\beta}_1)$ when X is indeed a realization of a random design?
  - Resampling method to give conservative variance estimates?
  - More advanced boostrap?
- Relax the regularity conditions:
  - Generalize to non-strongly convex and non-smooth loss functions?
  - Generalize to general error distributions?
- Get rid of asymptotics:
  - Yes, exact finite-sample guarantee if $n/p > 20$;
  - No assumption on X or β^*;
  - Only exchangeability assumption on ϵ.

Thank You!

References

Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013.

Peter J Bickel and David A Freedman. Bootstrapping regression models with many parameters. *Festschrift for Erich L. Lehmann*, pages 28–48, 1982.

Sourav Chatterjee. Fluctuations of eigenvalues and second order poincaré inequalities. *Probability Theory and Related Fields*, 143(1-2):1–40, 2009.

Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. 2015.

Noureddine El Karoui and Elizabeth Purdom. Can we trust the bootstrap in high-dimension? *UC Berkeley Statistics Department Technical Report*, 2015.

Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2011.

Peter J Huber. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, pages 799–821, 1973.

Enno Mammen. Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *The Annals of Statistics*, pages 382–400, 1989.

Stephen Portnoy. Asymptotic behavior of m-estimators of p regression parameters when p2/n is large. i. consistency. *The Annals of Statistics*, pages 1298–1309, 1984.

Stephen Portnoy. Asymptotic behavior of m estimators of p regression parameters when p2/n is large; ii. normal approximation. *The Annals of Statistics*, pages 1403–1417, 1985.