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

Background

Main Results

Heuristics and Proof Techniques

Numerical Results
Table of Contents

Background

Main Results

Heuristics and Proof Techniques

Numerical Results
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: Given a convex loss function \( \rho(\cdot) : \mathbb{R} \rightarrow [0, \infty) \),

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

When \( \rho \) 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

- $\rho(x) = \frac{x^2}{2}$ 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;

![Graphs of L2 Loss, L1 Loss, and Huber Loss](image-url)
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.
Goal (Informal): Make inference on the coordinates of $\beta^*$ when

- $X$ is treated as fixed;
- no assumption imposed on $\beta^*$;
- and the dimension $p$ is comparable to the sample size $n$. 
Goals (Informal)

Goal (Informal): Make inference on the coordinates of $\beta^*$ when

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

- Why coordinates?
Goal (Informal): Make inference on the coordinates of $\beta^*$ when

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

- Why coordinates?
- Why fixed designs?
Goal (Informal): Make inference on the coordinates of $\beta^*$ when

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

- Why coordinates?
- Why fixed designs?
- Why assumption-free $\beta^*$?
Goals (Informal)

**Goal (Informal):** Make inference on the coordinates of $\beta^*$ when

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

- Why coordinates?
- Why fixed designs?
- Why assumption-free $\beta^*$?
- Why $p \sim n$?
Asymptotic Arguments: Motivation

▶ Consider $\beta_1^*$ WLOG;

Ideally, we construct a 95% confidence interval for $\beta_1^*$ as $[q_{0.025}(L(\hat{\beta}_1)), q_{0.975}(L(\hat{\beta}_1))]$ where $q_\alpha$ denotes the $\alpha$-th quantile;

Unfortunately, $L(\hat{\beta}_1)$ is unknown. This motivates the asymptotic arguments, i.e. find a distribution $F$ s.t. $L(\hat{\beta}_1) \approx F$. 
Asymptotic Arguments: Motivation

- Consider $\beta_1^*$ WLOG;

- Ideally, we construct a 95% confidence interval for $\beta_1^*$ as

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

where $q_\alpha$ denotes the $\alpha$-th quantile;
Asymptotic Arguments: Motivation

- Consider $\beta^*_1$ WLOG;

- Ideally, we construct a 95% confidence interval for $\beta^*_1$ as
  
  \[
  \left[ q_{0.025} \left( \mathcal{L}(\hat{\beta}_1) \right), q_{0.975} \left( \mathcal{L}(\hat{\beta}_1) \right) \right]
  \]

  where $q_\alpha$ denotes the $\alpha$-th quantile;

- Unfortunately, $\mathcal{L}(\hat{\beta}_1)$ is unknown.
Asymptotic Arguments: Motivation

- Consider $\beta_1^*$ WLOG;

- Ideally, we construct a 95% confidence interval for $\beta_1^*$ as

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

  where $q_\alpha$ denotes the $\alpha$-th quantile;

- Unfortunately, $\mathcal{L}(\hat{\beta}_1)$ is unknown.

- This motivates the asymptotic arguments, i.e. find a distribution $F$ s.t.

  $$\mathcal{L}(\hat{\beta}_1) \approx F.$$
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 \( \beta_1^* \),

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

where \( \hat{sd}(\hat{\beta}_1) \) 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
\[(n_2 = 500, p_2 = 30)\]
\[y^2 \sim X^2 \Rightarrow \hat{\beta}^{(2)}_1\]

hypothetical problem
\[(n_3 = 2000, p_3 = 30)\]
\[y^3 \sim X^3 \Rightarrow \hat{\beta}^{(3)}_1\]
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
\[(n_2 = 500, p_2 = 30)\]
\[y^2 \sim X^2 \Rightarrow \hat{\beta}^{(2)}_1\]

hypothetical problem
\[(n_3 = 2000, p_3 = 30)\]
\[y^3 \sim X^3 \Rightarrow \hat{\beta}^{(3)}_1\]

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

\[ \| \hat{\beta} - \beta^* \|_2 \to 0, \quad \text{when } p = o\left(\frac{n^{1/3}}{\log n}\right). \]
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 $\hat{\beta}$:
  \[ \| \hat{\beta} - \beta^* \|_2^2 \to 0, \quad \text{when } p = o(n^{1/3}); \]

- Portnoy [1984] prove the $L_2$ consistency of $\hat{\beta}$ when
  \[ p = o \left( \frac{n}{\log n} \right). \]
Portnoy [1985] and Mammen [1989] showed that $\hat{\beta}$ is \textbf{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 $a_n \in \mathbb{R}^p$,

\[ \mathcal{L} \left( \frac{a_n^T (\hat{\beta} - \beta^*)}{\sqrt{\text{Var}(a_n^T \hat{\beta})}} \right) \rightarrow N(0, 1) \]
All of the above works requires

\[ p/n \rightarrow 0 \text{ or } n/p \rightarrow \infty. \]
All of the above works requires

\[ \frac{p}{n} \rightarrow 0 \quad \text{or} \quad \frac{n}{p} \rightarrow \infty. \]

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

Formally, we define Moderate $p/n$ Regime as

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

Formally, we define **Moderate p/n Regime** as

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

Formally, we define **Moderate $p/n$ Regime** as

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

![Diagram](image)

- **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 p/n Regime** as

\[ p/n \rightarrow \kappa > 0. \]

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. $\epsilon_i$'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. $\epsilon_i$’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. $\epsilon_i$’s.

\[ y_1 = X \beta^* + \epsilon_1 \\
 y_2 = \epsilon_1 + \epsilon_2 \\
 y_3 = \epsilon_2 + \epsilon_3 \\
 \vdots \\
 y_r = \epsilon_{r-1} + \epsilon_r \\
 \]

M-Estimates:

\[ \hat{\beta}^{(1)}_1, \hat{\beta}^{(2)}_1, \hat{\beta}^{(3)}_1, \ldots, \hat{\beta}^{(r)}_1. \]

\[ \Rightarrow \hat{L}(\hat{\beta}^{(1)}_1; X) = \text{ecdf}\{\hat{\beta}^{(1)}_1, \ldots, \hat{\beta}^{(r)}_1\}. \]
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. $\epsilon_i$’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. $\epsilon_i$’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. $\epsilon_i$'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. $\epsilon_i$'s.

\[
y^2 = X \beta^* + \epsilon^1 + \epsilon^2
\]

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

\[
\hat{L}(\hat{\beta}_1^{(1)}; X) = \text{ecdf}\left\{\hat{\beta}_1^{(1)},\ldots,\hat{\beta}_1^{(r)}\right\}.
\]
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. $\epsilon_i$’s.

\[
y^2 = X \beta^* + \epsilon^1 + \epsilon^2
\]

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

\[
y^3 = X \beta^* + \epsilon^1 + \epsilon^2 + \epsilon^3
\]

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. $\epsilon_i$'s.

$$y^3 = X + \epsilon^1 + \epsilon^2 + \epsilon^3$$

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

\[ y_1 = X \beta^* + \epsilon_1 \]

\[ y_2 = X \beta^* + \epsilon_1 + \epsilon_2 \]

\[ y_3 = X \beta^* + \epsilon_1 + \epsilon_2 + \epsilon_3 \]
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. $\epsilon_i$'s.

\[
y^r = X \beta^* + \epsilon^1 \epsilon^2 \epsilon^3 \cdots \epsilon^r
\]

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

\[
y^r = X \beta^* + \epsilon^1 + \epsilon^2 + \epsilon^3 + \cdots + \epsilon^r
\]

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

\[
y^{(r)} = \begin{bmatrix} X \\ \beta^* + \epsilon_1, \epsilon_2, \epsilon_3, \ldots, \epsilon_r \end{bmatrix}
\]

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

\[
\implies \hat{L}(\hat{\beta}_1; X) = \text{ecdf}(\{\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(r)}\}).
\]
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:** \( n = 1000, \ p = 50\kappa. \)

\[ y^r = X \beta^* + \epsilon^1 + \epsilon^2 + \epsilon^3 + \cdots + \epsilon^r \]

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

\[ \Rightarrow \hat{\mathcal{L}}(\hat{\beta}_1^F; X) = \text{ecdf}(\{\hat{\beta}_{1}^{(F,1)}, \ldots, \hat{\beta}_{1}^{(F,r)}\}). \]
Moderate $p/n$ Regime: More Informative Asymptotics

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

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

\[
X = \beta^* + \epsilon_1 + \epsilon_2 + \epsilon_3 + \cdots + \epsilon_r.
\]

\[
\hat{\beta}_m = \Rightarrow \hat{L}(\hat{\beta}_m) = \text{ecdf}\left\{\hat{\beta}_m, \ldots, \hat{\beta}_r\right\}.
\]
Moderate $p/n$ Regime: More Informative Asymptotics

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

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

\[
y^r = X \beta^* + \epsilon^1 + \epsilon^2 + \epsilon^3 + \cdots + \epsilon^r
\]

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

\[
\Longrightarrow \hat{\mathcal{L}}(\hat{\beta}_1^M; X) = \text{ecdf}(\{\hat{\beta}_1^{(M,1)}, \ldots, \hat{\beta}_1^{(M,r)}\})
\]
Moderate $p/n$ Regime: More Informative Asymptotics

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

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

Distance between the small sample and large sample distribution

![Graph showing the Kolmogorov-Smirnov Statistics for different regimes and sample types.](image)
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

$$\frac{L}{\sqrt{\text{Var}(a^T_n \hat{\beta}_{LS} - \beta^*)}} \not\to N(0, 1).$$

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

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

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

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

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

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

- 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,

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

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

- 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 $\beta_1^*$. 
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,

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

- 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,

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

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

$$\mathcal{L}_{X,\epsilon} \left( \frac{a_n^T (\hat{\beta} - \beta^*)}{\sqrt{\text{Var}_{X,\epsilon} (a_n^T \hat{\beta})}} \right) \rightarrow N(0, 1);$$
Bean et al. [2013] showed that when $X$ has i.i.d. Gaussian entries, for any sequence of $a_n \in \mathbb{R}^p$

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

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 $a_n \in \mathbb{R}^p$

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

- 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 $\epsilon$;
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 $a_n \in \mathbb{R}^p$

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

- 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 $\epsilon$;
- El Karoui et al. [2011] showed that for general loss functions,

$$
\|\hat{\beta} - \beta^*\|_\infty \rightarrow 0.
$$
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 $\epsilon_i$;
  ▶ 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.
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 $\epsilon_i$;
  - 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.
Our Goal (formal): Under the linear model

\[ Y = X \beta^* + \epsilon, \]

Derive the asymptotic distribution of coordinates \( \hat{\beta}_j \):

- under the moderate \( p/n \) regime, i.e. \( p/n \to \kappa \in (0, 1) \);
- with a fixed design matrix \( X \);
- without assumptions on \( \beta^* \).
Table of Contents

Background

Main Results

Heuristics and Proof Techniques

Numerical Results
Main Result (Informal)

**Definition 1.**

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

\[
d_{TV}(P, Q) = \sup_{A \subset \mathbb{R}^p} |P(A) - Q(A)|.
\]
Main Result (Informal)

**Definition 1.**

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

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

**Theorem.**

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

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

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

$$\hat{\beta} - \beta^* \overset{d}{=} \beta^* - \hat{\beta} \implies \mathbb{E}\hat{\beta} = \beta^*.$$ 

**Theorem.**

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

$$\max_j d_{TV} \left( \mathcal{L} \left( \frac{\hat{\beta}_j - \beta^*_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \right), N(0, 1) \right) = o(1).$$
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 develop a novel strategy that is built on

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

Limitations:

- we impose strong conditions on \( \rho \) and \( L(\epsilon) \);
- we do not know how to estimate \( \text{Var}(\hat{\beta}_1) \).
Our Contributions and Limitations

Instead, we develop 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 develop 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 \( \rho \) and \( \mathcal{L}(\epsilon) \);
▶ we do not know how to estimate \( \text{Var}_\epsilon(\hat{\beta}_1) \).
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 $Z$. 

Example 1

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

Example 2

$Z$ contains an intercept term, i.e. $Z = (1, \tilde{Z})$ and $\tilde{Z} \in \mathbb{R}^{n \times (p-1)}$ has independent sub-gaussian entries with $\tilde{Z}_{ij} - \mu_j d = \mu_j - \tilde{Z}_{ij}$, $\text{Var}(\tilde{Z}_{ij}) > \tau^2$ for some arbitrary $\mu_j$'s.
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

$$\text{Var}(Z_{ij}) = \tau^2 > 0;$$

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

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

for some arbitrary $\mu_j$'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

\[
Y_i = \beta^{\star}_{k_i} + \epsilon_i.
\]
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

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

This is a standard location problem.
A Counter-Example

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

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

Assume a balance design, i.e. \( n_j \approx \frac{n}{p} \). Then \( n_j < \infty \) and none of \( \hat{\beta}_j \) is normal (unless \( \epsilon_i \) are normal); holds for general loss functions \( \rho \).
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 $\epsilon_i$ are normal);
- holds for general loss functions $\rho$. 
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 $\epsilon_i$ are normal);
- holds for general loss functions $\rho$.

**Conclusion:** some “non-standard” assumptions on $X$ are required.
Table of Contents

Background

Main Results

Heuristics and Proof Techniques
  Least-Square Estimator: A Motivating Example
  Second-Order Poincaré Inequality
  Assumptions
  Main Results

Numerical Results
Least Square Estimator

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

$$\hat{\beta}^{LS} = (X^TX)^{-1}XTY = \beta^* + (X^TX)^{-1}XT\epsilon.$$
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

$$\hat{\beta}_{LS}^j - \beta_j^* = e_j^T (X^T X)^{-1} X^T \epsilon \triangleq \alpha_j^T \epsilon.$$
Least Square Estimator

Lindeberg-Feller CLT claims that in order for

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

it is **sufficient and almost necessary** that

$$\frac{\|\alpha_j\|_\infty}{\|\alpha^*_j\|_2} \rightarrow 0.$$  \hspace{1cm} (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

$$X^TX = \text{diag}(n_j)^p_{j=1}.$$ 

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

$$X^T X = \text{diag}(n_j)^p_{j=1}.$$ 

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 $n_j \leq 1/\kappa$ 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 $\rho$. 
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 $\rho$.

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

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

We show that

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

\[
|u'(x)| \leq c_1 \text{ and } |u''(x)| \leq c_2.
\]

For example, \( u = \text{Id} \) gives \( N(0, 1) \) and \( u = \Phi \) gives \( U([0, 1]) \).
Proposition 1 (SOPI; Chatterjee [2009]).

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

\[
\kappa_1 = (\mathbb{E} \| \nabla g(\mathcal{W}) \|_2^4)^{\frac{1}{4}}; \\
\kappa_2 = (\mathbb{E} \| \nabla^2 g(\mathcal{W}) \|_{4 \text{ op}}^4)^{\frac{1}{4}}; \\
\kappa_0 = (\mathbb{E} \sum_{i=1}^n \| \nabla_i g(\mathcal{W}) \|_4^4)^{\frac{1}{2}}.
\]

If \( \mathbb{E} U^4 < \infty \), then

\[
d_{TV} \left( \mathcal{L} \left( \frac{U - \mathbb{E} U}{\sqrt{\text{Var}(U)}} \right), N(0, 1) \right) \leq \frac{\kappa_0 + \kappa_1 \kappa_2}{\text{Var}(U)}.
\]
Assumptions

A1  \( \rho(0) = \psi(0) = 0 \) and for any \( x \in \mathbb{R} \),
\[
0 < K_0 \leq \psi'(x) \leq K_1, \quad |\psi''(x)| \leq K_2;
\]

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

A3  Let \( \lambda_+ \) and \( \lambda_- \) 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
\[
\min_j \text{Var}(\hat{\beta}_j) = \Omega \left( \frac{1}{n} \right)
\]
Main Results

**Theorem 3.**

_Under assumptions A1 – A5, as \( p/n \to \kappa \) for some \( \kappa \in (0, 1) \) while \( n \to \infty \),

\[
\max_j d_{TV} \left( \mathcal{L} \left( \frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \right) , N(0, 1) \right) = o(1).
\]
Table of Contents

Background

Main Results

Heuristics and Proof Techniques

Numerical Results
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 $F$:

- $F = N(0, 1)$;
- $F = t_2$.

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

- $\epsilon_i \sim N(0, 1)$;
- $\epsilon_i \sim t_2$. 
Setup

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

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

Loss Function $\rho$: 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

\[ y_1 = X\beta^* + \epsilon_1 \]

\[ y_2 = \epsilon_1 + \epsilon_2 \]

\[ y_3 = \epsilon_2 + \epsilon_3 \]

\[ \cdots \]

\[ y_r = \epsilon_{r-1} + \epsilon_r \]

M-Estimates:

\[ \hat{\beta}_1^{(1)}, \hat{\beta}_1^{(2)}, \hat{\beta}_1^{(3)}, \ldots, \hat{\beta}_1^{(r)} \]

\[ \hat{s}_d \leftarrow se(\{\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(r)}\}) ; \]

\[ \text{want to compare} \ L(\hat{\beta}_1 / \hat{s}_d) \text{ with } N(0, 1) ; \]

\[ \text{count the fraction of} \ \hat{\beta}_1^{(j)}_1 \in \left[ -1.96 \hat{s}_d, 1.96 \hat{s}_d \right] \text{ as the proxy;} \]

\[ \text{should be close to 0.95 ideally.} \]
Asymptotic Normality of A Single Coordinate

\[ y_1 = X \beta^* + \epsilon_1 \]

\[ y_2 = \epsilon_1 \epsilon_2 \]

\[ y_3 = \epsilon_2 \epsilon_3 \]

\[ \vdots \]

\[ y_r = \epsilon_{r-1} \epsilon_r \]

M-Estimates:

\[ \hat{\beta}(1)_1, \hat{\beta}(2)_1, \hat{\beta}(3)_1, \ldots, \hat{\beta}(r)_1. \]

\[ \hat{sd} \left\langle \text{se} \left\{ \hat{\beta}(1)_1, \ldots, \hat{\beta}(r)_1 \right\} \right\rangle; \]

want to compare \( L(\hat{\beta}_1/\hat{sd}) \) with \( N(0, 1) \);

count the fraction of \( \hat{\beta}(j)_1 \in [-1.96 \hat{sd}, 1.96 \hat{sd}] \) as the proxy;

should be close to 0.95 ideally.
Asymptotic Normality of A Single Coordinate

\[ y_1 = X \beta^* + \epsilon_1 \]

\[ y_2 = \epsilon_1 \epsilon_2 \]

\[ y_3 = \epsilon_2 \epsilon_3 \]

\[ \vdots \]

\[ y_r = \epsilon_3 \epsilon_4 \epsilon_5 \ldots \epsilon_r \]

M-Estimates:
\[ \hat{\beta}_{(1)}^1, \hat{\beta}_{(2)}^1, \hat{\beta}_{(3)}^1, \ldots, \hat{\beta}_{(r)}^1. \]

\[ \hat{sd} \leftarrow se(\{\hat{\beta}_{(1)}^1, \ldots, \hat{\beta}_{(r)}^1\}); \]

Want to compare \( L(\hat{\beta}_1 / \hat{sd}) \) with \( N(0, 1) \);

Count the fraction of \( \hat{\beta}_{(j)}^1 \in [-1.96 \hat{sd}, 1.96 \hat{sd}] \) as the proxy;

Should be close to 0.95 ideally.
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

\[ y^2 = X \beta^* + \epsilon^1 + \epsilon^2 \]

M-Estimates: \( \hat{\beta}_1^{(1)} \),
Asymptotic Normality of A Single Coordinate

\[ y^2 = X \beta^* + \epsilon^1 + \epsilon^2 \]

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

\[ \hat{\beta}_1 \]
Asymptotic Normality of A Single Coordinate

\[ y^3 = X \beta^* + \epsilon^1 + \epsilon^2 + \epsilon^3 \]

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

\[ \hat{\beta}_1^{(1)} \]

\[ \hat{\beta}_1^{(2)} \]
Asymptotic Normality of A Single Coordinate

\[ y^3 = X \beta^* + \epsilon^1 + \epsilon^2 + \epsilon^3 \]

M-Estimates: \( \hat{\beta}_1^{(1)}, \hat{\beta}_1^{(2)}, \hat{\beta}_1^{(3)}, \)
Asymptotic Normality of A Single Coordinate

\[ y^r = X \beta^* + \epsilon^1 + \epsilon^2 + \epsilon^3 + \cdots + \epsilon^r \]

M-Estimates: \( \hat{\beta}_1^{(1)}, \hat{\beta}_1^{(2)}, \hat{\beta}_1^{(3)} \),
Asymptotic Normality of A Single Coordinate

\[ y^r = X \beta^* + \epsilon^1 + \epsilon^2 + \epsilon^3 + \cdots + \epsilon^r \]

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

\[ y^r = X \beta^* + \epsilon_1 + \epsilon_2 + \epsilon_3 + \cdots + \epsilon_r \]

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

\( \hat{sd} \leftarrow \text{se} \left( \{ \hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(r)} \} \right) \);
Asymptotic Normality of A Single Coordinate

\[ y^r = X_\beta^* + \epsilon_1 + \epsilon_2 + \epsilon_3 + \cdots + \epsilon_r \]

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

\[ \hat{s}d \leftarrow \text{se} \left( \{\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(r)}\} \right); \]

\( \text{want to compare } L \left( \frac{\hat{\beta}_1}{\hat{s}d} \right) \text{ with } N(0, 1); \)
Asymptotic Normality of A Single Coordinate

\[ y^r = X \beta^* + \epsilon_1 + \epsilon_2 + \epsilon_3 + \ldots + \epsilon_r \]

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

- \( \hat{s}d \leftarrow \text{se} \left( \{\hat{\beta}^{(1)}_1, \ldots, \hat{\beta}^{(r)}_1\} \right) \);

- want to compare \( \mathcal{L} \left( \hat{\beta}_1/\hat{s}d \right) \) with \( N(0, 1) \);

- count the fraction of \( \hat{\beta}_1^{(j)} \in [-1.96\hat{s}d, 1.96\hat{s}d] \) as the proxy;
Asymptotic Normality of A Single Coordinate

\[ y^r = X \beta^* + \epsilon_1 + \epsilon_2 + \epsilon_3 + \cdots + \epsilon_r \]

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

\( \hat{s}d \leftarrow \text{se}\left( \{ \hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(r)} \} \right) ; \)

want to compare \( \mathcal{L}\left( \hat{\beta}_1 / \hat{s}d \right) \) with \( N(0, 1) \);

count the fraction of \( \hat{\beta}_1^{(j)} \in [-1.96\hat{s}d, 1.96\hat{s}d] \) as the proxy;

should be close to 0.95 ideally.
Asymptotic Normality of A Single Coordinate

Coverage of $\hat{\beta}_1$ ($\kappa = 0.5$)

Coverage of $\hat{\beta}_1$ ($\kappa = 0.8$)

Entry Dist. normal $t(2)$ hadamard

Sample Size

Coverage

Sample Size

Coverage
Conclusion

- We establish the **coordinate-wise asymptotic normality** of the M-estimator for certain **fixed design matrices** under the **moderate p/n regime** under regularity conditions on $X$, $L(\epsilon)$ and $\rho$ but **no condition on $\beta^*$**;

- 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

Inference

\[ \text{Var}(\hat{\beta}_1 | X) \approx \text{Var}(\hat{\beta}_1) \text{ when } X \text{ is indeed a realization of a random design?} \]

Resampling method to give conservative variance estimates?

More advanced bootstrap?

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 \( \beta^* \);

Only exchangeability assumption on \( \epsilon \).
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 $\approx$ 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 bootstrap?

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

- 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 $\beta^*$;
  - Only exchangeability assumption on $\epsilon$. 
Thank You!
References


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


