Inference For High Dimensional M-estimates: Fixed Design Results

Lihua Lei
Advisors: Peter J. Bickel, Michael I. Jordan

joint work with Peter J. Bickel and Noureddine El Karoui

Dec. 8, 2016
Table of Contents

1 Background

2 Main Results and Examples

3 Assumptions and Proof Sketch

4 Numerical Results
Table of Contents

1 Background

2 Main Results and Examples

3 Assumptions and Proof Sketch

4 Numerical Results
Observe \( \{x_1, y_1\}, \{x_2, y_2\}, \ldots, \{x_n, y_n\} \):

- response vector \( Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n \);
- design matrix \( X = (x_1^T, \ldots, x_n^T)^T \in \mathbb{R}^{n \times p} \).
Observe \( \{x_1, y_1\}, \{x_2, y_2\}, \ldots, \{x_n, y_n\} \):

- response vector \( Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n \);
- design matrix \( X = (x_1^T, \ldots, x_n^T)^T \in \mathbb{R}^{n \times p} \).

Model:

- Linear Model: \( Y = X \beta^* + \epsilon \);
- \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \in \mathbb{R}^n \) being a random vector;
M-Estimator: Given a convex loss function $\rho(\cdot) : \mathbb{R} \to [0, \infty)$,

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

$$
\frac{1}{n} \sum_{i=1}^{n} \psi(y_i - x_i^T \hat{\beta}) = 0.
$$
\( \rho(x) = x^2 / 2 \) gives the Least-Square estimator;
\[ \rho(x) = \frac{x^2}{2} \] gives the Least-Square estimator;
M-Estimator: Examples

- \( \rho(x) = \frac{x^2}{2} \) gives the Least-Square estimator;
- \( \rho(x) = |x| \) gives the Least-Absolute-Deviation estimator;

![Graph of L2 Loss](image)

![Graph of Huber Loss](image)
M-Estimator: Examples

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

\[ \rho(x) = \begin{cases} 
\frac{x^2}{2} & \text{if } |x| \leq k \\
\left(|x| - \frac{k}{2}\right)^+ & \text{if } |x| > k
\end{cases} \]
gives the Huber estimator.
M-Estimator: Examples

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

$L_2$ Loss

$L_1$ Loss

Huber Loss
M-Estimator: Examples

- \( \rho(x) = \frac{x^2}{2} \) gives the Least-Square estimator;
- \( \rho(x) = |x| \) gives the Least-Absolute-Deviation estimator;
- \( \rho(x) = \begin{cases} 
  \frac{x^2}{2} & |x| \leq k \\
  k(|x| - k/2) & |x| > k 
\end{cases} \) gives the Huber estimator.
Goal (Informal): Make inference on the coordinates of $\hat{\beta}$ when
- the dimension $p$ is \textit{comparable to} the sample size $n$;
- and $X$ is treated as \textit{fixed};
- \textit{without assumptions on} $\beta^*$.
Goal (Informal): Make inference on the coordinates of $\hat{\beta}$ when
- the dimension $p$ is comparable to the sample size $n$;
- and $X$ is treated as fixed;
- without assumptions on $\beta^*$.

Consider $\beta^*_1$ WLOG;
- Given $X$ and $L(\epsilon)$, $L(\hat{\beta}_1)$ is uniquely determined;
- Ideally, we construct a 95% confidence interval for $\beta^*_1$ as

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

where $q_\alpha$ denotes the $\alpha$-th quantile;
- Unfortunately, $L(\hat{\beta}_1)$ is complicated.
Asymptotic Arguments

Exact finite sample inference is hard. This motivates statisticians to resort to asymptotic arguments, i.e. find a distribution $F$ s.t.

$$\mathcal{L}(\hat{\beta}_1) \approx F.$$
Asymptotic Arguments

Exact finite sample inference is hard. This motivates statisticians to resort to 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 $n \to \infty$,

$$\mathcal{L}(\hat{\beta}) \to N \left( \beta^*, (X^TX)^{-1} \frac{\mathbb{E}(\psi^2(\epsilon_1))}{[\mathbb{E}\psi'(\epsilon_1)]^2} \right);$$

- 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.
In other words, to approximate $\mathcal{L}(\hat{\beta}_1)$, we consider a sequence of hypothetical problems, indexed by $j$, where the $j$-th problem has a sample size $n_j \to \infty$ and a dimension $p_j = p$. For the $j$-th problem, denote by $\hat{\beta}(j)$ the corresponding M-estimator, then the previous slide uses $\lim_{j \to \infty} \mathcal{L}(\hat{\beta}(j))$ to approximate $\mathcal{L}(\hat{\beta}_1)$. In general, $p_j$ is not necessarily fixed and can grow to infinity.
In other words, to approximate $\mathcal{L}(\hat{\beta}_1)$, we consider a sequence of hypothetical problems, indexed by $j$, where the $j$-th problem has a sample size $n_j \to \infty$ and a dimension $p_j = p$.

For $j$-th problem, denote by $\hat{\beta}^{(j)}$ the corresponding M-estimator, then the previous slide uses

$$\lim_{j \to \infty} \mathcal{L}(\hat{\beta}^{(j)})$$

to approximate $\mathcal{L}(\hat{\beta}_1)$. 

In general, $p_j$ is not necessarily fixed and can grow to infinity.
In other words, to approximate $\mathcal{L}(\hat{\beta}_1)$, we consider a sequence of hypothetical problems, indexed by $j$, where the $j$-th problem has a sample size $n_j \to \infty$ and a dimension $p_j = p$.

For $j$-th problem, denote by $\hat{\beta}(j)$ the corresponding M-estimator, then the previous slide uses

$$\lim_{j \to \infty} \mathcal{L}(\hat{\beta}(j))$$


to approximate $\mathcal{L}(\hat{\beta}_1)$.

In general, $p_j$ is not necessarily fixed and can grow to infinity.
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 $\hat{\beta}$:

$$\|\hat{\beta} - \beta^*\|^2 \rightarrow 0$$

under the regime

$$\frac{p^3}{n} \rightarrow 0;$$

Portnoy (1984) prove the $L_2$ consistency of $\hat{\beta}$ under the regime

$$\frac{p \log p}{n} \rightarrow 0;$$
Portnoy (1985) showed that $\hat{\beta}$ is jointly asymptotically normal under the regime

$$\frac{(p \log n)^{\frac{3}{2}}}{n} \rightarrow 0,$$

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

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

\[ \frac{n}{p} \] is the number of samples per parameter. Heuristically, a larger \( \frac{n}{p} \) would give an easier problem.
All of the above works requires

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

\( \frac{n}{p} \) is the number of samples per parameter. Heuristically, a larger \( n/p \) would give an easier problem.
Recall that the approximation can be seen as a sequence of hypothetical problems with sample size $n_j$ and dimension $p_j$. If $n_j/p_j \to \infty$, the problems become increasingly easier as $j$ grows.
Recall that the approximation can be seen as a sequence of hypothetical problems with sample size $n_j$ and dimension $p_j$. If $n_j/p_j \to \infty$, the problems become increasingly easier as $j$ grows.

In other words, the hypothetical problem used for approximation is much easier than the original problem. Then the approximation accuracy might be compromised.
Instead, we can consider a sequence of hypothetical problems with $p_j/n_j$ fixed to be the same as the original problem, i.e.

$$p_j/n_j \equiv p/n.$$
Instead, we can consider a sequence of hypothetical problems with $p_j/n_j$ fixed to be the same as the original problem, i.e.

$$p_j/n_j \equiv p/n.$$ 

In this case, the **difficulty** of the problem is **fixed**.
Formally, we define **Moderate p/n Regime** as

$$p_j/n_j \to \kappa > 0.$$ 

A typical value for $\kappa$ is $p/n$ in the original problem.
Consider a set of small-sample problems where \( n = 50 \) and \( p = n\kappa \) for \( \kappa \in \{0.1, \ldots, 0.9\} \). For each pair \((n, p)\),

**Step 1** Generate \( X \in \mathbb{R}^{n \times p} \) with i.i.d. \( N(0, 1) \) entries;

**Step 2** Fix \( \beta^* = 0 \) and sample \( Y = \epsilon \) with

\[
\epsilon_i \overset{i.i.d.}{\sim} N(0, 1) \quad \text{or} \quad \epsilon_i \overset{i.i.d.}{\sim} t_2;
\]

**Step 3** Estimate \( \beta_1^* \) by \( \hat{\beta}_1 \) with a Huber loss;

**Step 4** Repeat Step 2 - Step 3 for 100 times and estimate \( L(\hat{\beta}_1) \).
Now consider two types of approximations:

- **Fixed-$p$ Approx.**: $N = 1000, P = p$;
- **Moderate-$p/n$ Approx.**: $N = 1000, P = 1000\kappa$;

Repeat Step 1-Step 4 for new pairs $(N, P)$ and estimate

- $\mathcal{L}(\hat{\beta}_1^F)$ (Fixed $p$);
- $\mathcal{L}(\hat{\beta}_1^M)$ (Moderate $p/n$).
Now consider two types of approximations:

- **Fixed-p Approx.**: $N = 1000, \; P = p$;
- **Moderate-p/n Approx.**: $N = 1000, \; P = 1000\kappa$;

Repeat Step 1-Step 4 for new pairs $(N, P)$ and estimate

- $L(\hat{\beta}_1^F)$ (Fixed $p$);
- $L(\hat{\beta}_1^M)$ (Moderate $p/n$).

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

$$d_{KS} \left( L(\hat{\beta}_1), L(\hat{\beta}_1^F) \right) \quad \text{and} \quad d_{KS} \left( L(\hat{\beta}_1), L(\hat{\beta}_1^M) \right)$$
Distance between the small sample and large sample distribution

<table>
<thead>
<tr>
<th></th>
<th>normal</th>
<th>t(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asym. Regime</td>
<td>p fixed</td>
<td>p/n fixed</td>
</tr>
</tbody>
</table>

Kolmogorov–Smirnov Statistics

Distance between the small sample and large sample distribution

Asym. Regime
- p fixed
- p/n fixed
The moderate $p/n$ regime has been widely studied in random matrix theory. 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

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

- 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,

$$\|\hat{\beta} - \beta^*\|_2^2 \not \xrightarrow{a.s.} 0.$$
The moderate $p/n$ regime has been widely studied in random matrix theory. 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;
- El Karoui et al. (2011) showed that for general loss functions,
  \[
  \|\hat{\beta} - \beta^*\|_2^2 \not\rightarrow 0.
  \]

- Main reason: $\hat{F}_n$, the empirical distribution of the residuals, namely $R_i \triangleq y_i - x_i^T \hat{\beta}$, does not converge to $\mathcal{L}(\epsilon_i)$.
Moderate $p/n$ Regime: Positive Results

If $X$ is assumed to be a random matrix under regularity conditions,
Moderate \( p/n \) Regime: Positive Results

If \( X \) is assumed to be a random matrix under regularity conditions,

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

- The above result does not contradict Huber (1973) in that the randomness comes from both \( X \) and \( \varepsilon \);
- El Karoui et al. (2011) showed that for general loss functions,
  \[
  \|\hat{\beta} - \beta^*\|_\infty \to 0.
  \]

- Under weaker assumptions on \( X \), El Karoui (2015) showed
  \[
  \mathcal{L}_{X,\varepsilon} \left( \frac{\hat{\beta}_1(\tau) - \beta^*_1 - \text{bias}(\hat{\beta}_1(\tau))}{\sqrt{\text{Var}_{X,\varepsilon}(\hat{\beta}_1(\tau))}} \right) \to N(0, 1)
  \]
  where \( \hat{\beta}_1(\tau) \) is the ridge-penalized M-estimator.
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 residuals $R_i$ behaves differently from $\epsilon_i$;
  - 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 residuals \( R_i \) 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

1. Background
2. Main Results and Examples
3. Assumptions and Proof Sketch
4. Numerical Results
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 $n \to \infty$,

$$\max_j d_{TV} \left( \mathcal{L} \left( \frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \right), N(0, 1) \right) = o(1).$$
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 $Z$.
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 $\mu_j$. 
Example 3 \( Z \) is matrix-normal with vec\( (Z) \sim \mathcal{N}(0, \Lambda \otimes \Sigma) \) and
\[
\lambda_{\max}(\Lambda), \lambda_{\max}(\Sigma) = O(1), \quad \lambda_{\min}(\Lambda), \lambda_{\min}(\Sigma) = \Omega(1)
\]

Example 4 \( Z \) contains an intercept term, i.e. \( Z = (1, \tilde{Z}) \) and vec\( (\tilde{Z}) \sim \mathcal{N}(0, \Lambda \otimes \Sigma) \) with \( \Lambda \) and \( \Sigma \) satisfy the above condition and
\[
\frac{\max_i |(\Lambda^{-\frac{1}{2}}1)_i|}{\min_j |(\Lambda^{-\frac{1}{2}}1)_j|} = O(1).
\]
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

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

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

2. Main Results and Examples

3. Assumptions and Proof Sketch
   - Least-Square Estimator: A Motivating Example
   - Second-Order Poincaré Inequality
   - Assumptions
   - Main Results

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

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

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

Write $e_j^T (X^T X)^{-1} X$ as $\alpha_j^T$, then

$$\hat{\beta}_j^{LS} - \beta_j^* = \sum_{i=1}^{n} \alpha_{j,i} \epsilon_i.$$
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 \textbf{sufficient and almost necessary} that

\[ \frac{\|\alpha_j\|_\infty}{\|\alpha_j\|_2} \rightarrow 0. \]  \hspace{1cm} (1)
To see the necessity of the condition, recall the one-way ANOVA case. Let $n_j = |\{i : k_i = j\}|$, then

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

This gives

$$\alpha_{j,i} = \begin{cases} \frac{1}{n_j} & \text{if } k_i = j \\ 0 & \text{if } k_i \neq j \end{cases}.$$
To see the necessity of the condition, recall the one-way ANOVA case. Let \( n_j = |\{i : k_i = j\}| \), then

\[
X^T X = \text{diag}(n_j)_j=1^p.
\]

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

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

$$\frac{1}{n} \sum_{i=1}^{n} \psi(y_i - x_i^T \hat{\beta}) = 0$$
The result for LSE is derived from the analytical form of $\hat{\beta}^{LS}$. In contrast, an analytical form is not available for general $\rho$.

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

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

WLOG, assume $\beta^* = 0$, then

$$\frac{1}{n} \sum_{i=1}^{n} \psi(\epsilon_i - x_i^T \hat{\beta}) = 0.$$
Write $R_i$ for $\epsilon_i - x_i^T \hat{\beta}$ and define $D$, $\tilde{D}$ and $G$ as

$$D = \text{diag}(\psi'(R_i)), \quad \tilde{D} = \text{diag}(\psi''(R_i)), \quad G = I - X(X^TDX)^{-1}X^TD.$$
Write $R_i$ for $\epsilon_i - x_i^T \hat{\beta}$ and define $D$, $\tilde{D}$ and $G$ as

$$D = \text{diag}(\psi'(R_i)), \quad \tilde{D} = \text{diag}(\psi''(R_i)), \quad G = I - X(X^T DX)^{-1}X^T D.$$ 

**Lemma 2.**

Suppose $\psi \in C^2(\mathbb{R}^n)$, then

$$\frac{\partial \hat{\beta}_j}{\partial \epsilon} = e_j^T (X^T DX)^{-1}X^T D, \quad (2)$$

$$\frac{\partial \hat{\beta}_j}{\partial \epsilon \partial \epsilon^T} = G^T \text{diag}(e_j^T (X^T DX)^{-1}X^T \tilde{D}) G. \quad (3)$$
\( \hat{\beta}_j \) 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 3.**

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 \( U([0, 1]) \).
Second-Order Poincaré Inequality

**Proposition 1 (SOPI; Chatterjee, 2009).**

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

\[
\kappa_0 = \left( \mathbb{E} \sum_{i=1}^{n} |\nabla_i g(\mathcal{W})|^4 \right)^{\frac{1}{2}};
\]

\[
\kappa_1 = (\mathbb{E} \| \nabla g(\mathcal{W}) \|^4_2)^{\frac{1}{4}};
\]

\[
\kappa_2 = (\mathbb{E} \| \nabla^2 g(\mathcal{W}) \|^4_{op})^{\frac{1}{4}}.
\]

If \( U \) has a finite fourth moment, then

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

Lihua Lei (Qualifying Exam)
Assumptions

Assume that

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

$$\lambda_+ = O(1), \quad \lambda_- = \Omega(1).$$
Apply Second-Order Poincaré Inequality to $\hat{\beta}_j$, we obtain that

**Lemma 4.**

Let $D = \text{diag}(\psi'(\epsilon_i - x_i^T \hat{\beta}))_{i=1}^n$, and

$$M_j = \mathbb{E}\|e_j^T (X^T DX)^{-1} X^T D^{1/2}\|_\infty.$$ 

Then under assumptions A1-A3,

$$\max_j d_{TV} \left( \mathcal{L} \left( \frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \right), N(0, 1) \right) = O_p \left( \frac{\max_j (nM_j^2)^{1/8}}{n \cdot \min_j \text{Var}(\hat{\beta}_j)} \right),$$

The main result is obtained if we prove

$$M_j = o \left( \frac{1}{\sqrt{n}} \right), \quad \text{Var}(\hat{\beta}_j) = \Omega \left( \frac{1}{n} \right).$$
Define the following quantities:

- **leave-one-predictor-out estimate** \( \hat{\beta}_{[j]} \): the M-estimator obtained by removing the \( j \)-th column of \( X \) (El Karoui, 2013);
- **leave-one-predictor-out residuals** \( r_{i,[j]} = \epsilon_i - x_{i,[j]}^T \hat{\beta}_{[j]} \) where \( x_{i,[j]}^T \) is the \( i \)-th row of \( X \) after removing \( j \)-th entry;
- \( h_{j,0} = (\psi(r_{1,[j]}), \ldots, \psi(r_{n,[j]}))^T \);
- \( Q_j = \text{Cov}(h_{j,0}) \) be the covariance matrix of \( \psi(r_{i,[j]}) \).
Besides assumptions $\mathbf{A}1 - \mathbf{A}3$, we assume that

$\mathbf{A}4$ \[ \min_j \frac{X_j^T Q_j X_j}{\text{tr}(Q_j)} = \Omega(1). \]
Besides assumptions A1 - A3, we assume that

\[ A4 \quad \min_j \frac{X_j^T Q_j X_j}{\text{tr}(Q_j)} = \Omega(1). \]

- \( Q_j \) does not involve \( X_j \);
- Assumption A4 guarantees

\[ \text{Var}(\hat{\beta}_j) = \Omega \left( \frac{1}{n} \right). \]
Further Assumptions

If $X_j$ is a realization of a random vector $Z_j$ with i.i.d. entries, then

$$\mathbb{E} Z_j^T Q_j Z_j = \text{tr}(\mathbb{E} Z_j Z_j^T Q_j) = \mathbb{E} Z_{1,j}^2 \cdot \text{tr}(Q_j).$$

If $Z_j^T Q_j Z_j$ concentrates around its mean, then

$$\frac{Z_j^T Q_j Z_j}{\text{tr}(Q_j)} \approx \mathbb{E} Z_{1,j}^2 > 0.$$
If $X_j$ is a realization of a random vector $Z_j$ with i.i.d. entries, then

$$\mathbb{E}Z_j^T Q_j Z_j = \text{tr}(\mathbb{E}Z_j Z_j^T Q_j) = \mathbb{E}Z_{1,j}^2 \cdot \text{tr}(Q_j).$$

If $Z_j^T Q_j Z_j$ concentrates around its mean, then

$$\frac{Z_j^T Q_j Z_j}{\text{tr}(Q_j)} \approx \mathbb{E}Z_{1,j}^2 > 0.$$ For example, when $Z_j$ has i.i.d. sub-gaussian entries, the Hansen-Wright inequality implies the concentration.

$$P(|Z_j^T Q_j Z_j - \mathbb{E}Z_j^T Q_j Z_j| \geq t) \leq 2 \exp \left\{ -c \min \left\{ \frac{t^2}{\|Q_j\|_F^2}, \frac{t}{\|Q_j\|_{op}} \right\} \right\}.$$
To describe the last assumption, we define the following quantities:

- $D_{[j]} = \text{diag}(\psi'(r_{i,[j]}))$: leave-one-predictor-out version of $D$;
- $G_{[j]} = I - X_{[j]}(X_{[j]}^T D_{[j]} X_{[j]})^{-1} X_{[j]}^T D_{[j]}$;
- $h_{j,1,i}^T = e_i^T G_{[j]}$: the $i$-th row of $G_{[j]}$;
- $\Delta_C = \max \left\{ \max_j \frac{|h_{j,0} X_j|}{\|h_{j,0}\|_2}, \max_{i,j} \frac{|h_{j,1,i} X_j|}{\|h_{j,1,i}\|_2} \right\}$. 

Lihua Lei (Qualifying Exam)

Inference For High Dimensional M-estimates
The last assumption:

\[ A_5 \quad \mathbb{E} \Delta^8_C = O(\text{polyLog}(n)). \]
Further Assumptions

The last assumption:

**A5** $\mathbb{E} \Delta^8_C = O\left(\text{polyLog}(n)\right)$.

It turns out that when $\rho(x) = x^2/2$,

$$\Delta_C \approx \max_j \left\| \frac{e_j^T (X^T X)^{-1} X^T}{\|e_j^T (X^T X)^{-1} X^T\|_\infty} \right\|_\infty.$$

Recall that for Least-Squares, $\hat{\beta}_j$ are all asymptotically normal iff the right-handed side tends to 0. This indicates that the assumption **A5** is **not just an artifact of the proof**.
Further Assumptions

Let
\[ \alpha_{j,0} = h_{j,0}/\|h_{j,0}\|_2, \quad \alpha_{j,1,i} = h_{j,1,i}/\|h_{j,1,i}\|_2. \]

Again, if \( X_j \) is a realization of a random vector \( Z_j \) with i.i.d. \( \sigma^2 \)-sub-gaussian entries, then \( \alpha_{j,0}^T Z_j \) and \( \alpha_{j,1,i}^T Z_j \) are all \( \sigma^2 \)-sub-gaussian.
Let

$$\alpha_{j,0} = h_{j,0}/\|h_{j,0}\|_2, \quad \alpha_{j,1,i} = h_{j,1,i}/\|h_{j,1,i}\|_2.$$ 

Again, if $X_j$ is a realization of a random vector $Z_j$ with i.i.d. $\sigma^2$-sub-gaussian entries, then $\alpha_{j,0}^T Z_j$ and $\alpha_{j,1,i}^T Z_j$ are all $\sigma^2$-sub-gaussian.

Then $\Delta_C$ is the maximum of $np + p$ sub-gaussian random variables and hence

$$\mathbb{E}\Delta_C^8 = O(poly\text{Log}(n)).$$
**Review of All 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** \( \min_j \frac{Z_j^T Q_j Z_j}{\text{tr}(Q_j)} = \Omega(1) \).

**A5** \( \mathbb{E} \Delta_C^8 = O(\text{polyLog}(n)) \).
Main Results

Theorem 5.

Under assumptions $\mathbf{A1} - \mathbf{A5}$, as $p/n \to \kappa$ for some $\kappa \in (0, 1)$ while $n \to \infty$,

$$\max \limits_{j} \ d_{TV} \left( \mathcal{L} \left( \frac{\hat{\beta}_j - \mathbb{E}\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \right), N(0, 1) \right) = o(1).$$
A Corollary

If further assume that

\( \textbf{A6} \)  \( \rho \) is an even function and \( \epsilon_i \overset{d}{=} -\epsilon_i \).

Then one can show that \( \hat{\beta} \) is unbiased. As a consequence,
A Corollary

If further assume that

\( A_6 \) \( \rho \) is an even function and \( \epsilon_i \overset{d}{=} -\epsilon_i \).

Then one can show that \( \hat{\beta} \) is unbiased. As a consequence,

**Theorem 6.**

*Under assumptions \( A_1 - A_6 \), as \( p/n \to \kappa \) for some \( \kappa \in (0, 1) \) while \( n \to \infty \),

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

1. Background

2. Main Results and Examples

3. Assumptions and Proof Sketch

4. 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

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

Lihua Lei (Qualifying Exam)  Inference For High Dimensional M-estimates
Sample Size $n$: \{100, 200, 400, 800\};

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

Loss Function $\rho$: Huber loss with $k = 1.345$,

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & |x| \leq k \\ kx - \frac{k^2}{2} & |x| > k \end{cases}$$
For each set of parameters, we run 50 simulations with each consisting of the following steps:

(Step 1) Generate one design matrix $X$;
(Step 2) Generate the 300 error vectors $\epsilon$;
(Step 3) Regress each $Y = \epsilon$ on the design matrix $X$ and end up with 300 random samples of $\hat{\beta}_1$, denoted by $\hat{\beta}_1^{(1)}, \ldots, \hat{\beta}_1^{(300)}$;
(Step 4) Estimate the standard deviation of $\hat{\beta}_1$ by the sample standard error $\hat{sd}$;
(Step 5) Construct a confidence interval $\mathcal{I}^{(k)} = \left[ \hat{\beta}_1^{(k)} - 1.96 \cdot \hat{sd}, \hat{\beta}_1^{(k)} + 1.96 \cdot \hat{sd} \right]$ for each $k = 1, \ldots, 300$;
(Step 6) Calculate the empirical 95% coverage by the proportion of confidence intervals which cover the true $\beta^*_1 = 0$. 

Lihua Lei (Qualifying Exam)

Inference For High Dimensional M-estimates
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

Lihua Lei (Qualifying Exam)
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 $X, \mathcal{L}(\epsilon)$ 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.
Future works for this project:

- Estimate $\text{Var}(\hat{\beta}_j)$
- Relax the assumptions on $L(\epsilon)$
- Relax the strong convexity of $\rho$
- Extend the results to GLM
Future works for this project:

- Estimate $\text{Var}(\hat{\beta}_j)$
- Relax the assumptions on $\mathcal{L}(\epsilon)$
- Relax the strong convexity of $\rho$
- Extend the results to GLM

Future works for my dissertation:

- Distributional properties in high dimensions
- Resampling methods in high dimensions
Thank You!


