# Inference for High Dimensional Robust Regression

Peter Bickel, Noureddine El Karoui, Lihua Lei

Department of Statistics
UC Berkeley

Stanford-Berkeley Joint Colloquium, 2015

# Table of Contents

1. Background
2. Main Results
3. OLS: A Motivating Example

# Table of Contents

1 Background

2 Main Results

3 OLS: A Motivating Example

# Setup

Consider a linear regression model:

$$Y_i = X_i^T \beta_0 + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Here $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}^p$ and $\epsilon_i \in \mathbb{R}$.

- OLS Estimator ($p < n$):

$$\hat{\beta}^{OLS} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2;$$

- M Estimator ($p < n$):

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

# Background

- The limiting behavior for $\hat{\beta}$ when $p$ is fixed

$$\mathcal{L}(\hat{\beta}) \approx N\left(\beta_0, (X^T X)^{-1} \frac{\mathbb{E}(\psi^2(\epsilon))}{[\mathbb{E}\psi'(\epsilon)]^2}\right);$$

# Background

- The limiting behavior for $\hat{\beta}$ when $p$ is fixed

$$\mathcal{L}(\hat{\beta}) \approx N\left(\beta_0, (X^T X)^{-1} \frac{\mathbb{E}(\psi^2(\epsilon))}{[\mathbb{E}\psi'(\epsilon)]^2}\right);$$

- Huber (1973) raised the question of understanding the behavior of $\hat{\beta}$ when $p \to \infty$;

# Background

- The limiting behavior for $\hat{\beta}$ when $p$ is fixed

$$\mathcal{L}(\hat{\beta}) \approx N\left(\beta_0, (X^T X)^{-1} \frac{\mathbb{E}(\psi^2(\epsilon))}{[\mathbb{E}\psi'(\epsilon)]^2}\right);$$

- Huber (1973) raised the question of understanding the behavior of $\hat{\beta}$ when $p \to \infty$;
- Huber (1973) proved that $\hat{\beta}^{OLS}$ is **jointly asymptotically normal** iff

$$\kappa = \max_i (X(X^T X)^{-1} X^T)_{i,i} \to 0$$

which requires

$$\frac{p}{n} \to 0.$$

# Background

- Portnoy (1984, 1985, 1986, 1987) proved the **joint asymptotic normality** of $\hat{\beta}$ in the case

$$\frac{(p \log n)^{\frac{3}{2}}}{n} \to 0;$$

# Background

- Portnoy (1984, 1985, 1986, 1987) proved the **joint asymptotic normality** of $\hat{\beta}$ in the case

$$\frac{(p \log n)^{\frac{3}{2}}}{n} \to 0;$$

- Mammen (1989) provided an expansion for $\hat{\beta}$ (which leads to **joint asymptotic normality**) by assuming

$$\kappa n^{\frac{1}{3}} (\log n)^{\frac{2}{3}} \to 0.$$

# Background

- Portnoy (1984, 1985, 1986, 1987) proved the **joint asymptotic normality** of $\hat{\beta}$ in the case

$$\frac{(p \log n)^{\frac{3}{2}}}{n} \to 0;$$

- Mammen (1989) provided an expansion for $\hat{\beta}$ (which leads to **joint asymptotic normality**) by assuming

$$\kappa n^{\frac{1}{3}} (\log n)^{\frac{2}{3}} \to 0.$$

- All works are based on **fixed designs** but requires

$$\frac{p}{n} \to 0.$$

# Background

- El Karoui et al. (2011, 2013), Bean et al. (2013) established the **joint asymptotic normality** of $\hat{\beta}$ in the regime

$$\frac{p}{n} \to \kappa \in (0, 1),$$

by assuming a **random design** $X$, which has i.i.d. Gaussian entries;

# Background

- El Karoui et al. (2011, 2013), Bean et al. (2013) established the **joint asymptotic normality** of $\hat{\beta}$ in the regime
$$\frac{p}{n} \to \kappa \in (0, 1),$$
by assuming a **random design** $X$, which has i.i.d. Gaussian entries;
- Zhang and Zhang (2014), Van de Geer et al. (2014) proved the **partial asymptotic normality** of LASSO estimator by assuming a **fixed design** and imposing a **sparsity** condition on $\beta_0$:
$$\frac{||\beta_0||_0 \log p}{\sqrt{n}} \to 0.$$

# Main Research Question and Our Contributions

Suppose $\frac{p}{n} \to \kappa \in (0,1)$ and the design matrix $X$ is fixed, can we make inference on the coordinate (or lower dimensional linear contrast) of $\hat{\beta}$?

# Main Research Question and Our Contributions

Suppose $\frac{p}{n} \to \kappa \in (0,1)$ and the design matrix $X$ is fixed, can we make inference on the coordinate (or lower dimensional linear contrast) of $\hat{\beta}$?

**YES**! In this work, we prove

- the **coordinatewise asymptotic normality** of $\hat{\beta}$
- in the regime $\mathbf{\frac{p}{n} \to \kappa \in (0,1)}$
- for **fixed designs**;
- show that the conditions for fixed design matrix is satisfied by a broad class of random designs.

# Table of Contents

1 Background

2 Main Results

3 OLS: A Motivating Example

Consider the ridge-regularized M estimator

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - x_i^T \beta) + \frac{\tau}{2} ||\beta||^2.$$

Assume that $\rho \in \mathcal{C}^2$ is a convex function with $\psi = \rho'$ and $\beta_0 = 0$, then the first order condition implies that

$$\sum_{i=1}^{n} x_i \psi(\epsilon_i - x_i^T \hat{\beta}) = n\tau\hat{\beta}.$$

In most cases, there is no closed form solution and $\hat{\beta}$ is an implicit function of $\epsilon$.

# Final Conclusions

**Theorem 1.**

*Under Assumptions* **A**1-**A**4, $\hat{\beta}$ *is coordinatewise asymptotically normal in the sense that*

$$\max_j d_{TV}\left(\frac{\hat{\beta}_j - \mathbb{E}_\epsilon \hat{\beta}_j}{\sqrt{Var_\epsilon(\hat{\beta}_j)}}, N(0,1)\right) = O\left(\frac{PolyLog(n)}{\sqrt{n}}\right).$$

# Assumptions: Loss Function

Assumption **A**1: Let $\psi = \rho'$, for any $x$,

- $0 < D_0 \le \psi'(x) \le D_1(|x| \vee 1)^{m_1}$;
- $|\psi''(x)| \le D_2(|x| \vee 1)^{m_2}$;
- $|\psi'''(x)| \le D_3(|x| \vee 1)^{m_3}$;
- $\max\{D_0^{-1}, D_1, D_2, D_3\} = O(PolyLog(n))$;
- $\max\{m_1, m_2, m_3\} = O(1)$.

# Assumptions: Error Distribution

Assumption **A**2: $\epsilon$ are transformations of i.i.d. Gaussian random variables, i.e. $\epsilon_i = u_i(\nu_i)$, where

- $\nu_i \overset{i.i.d.}{\sim} N(0, 1)$;
- $||u_i'||_\infty \le c_1, \ ||u_i''||_\infty \le c_2$;
- $\max\{c_1, c_2\} = O(PolyLog(n))$.

# Assumptions: Design Matrix

Assumption **A**3: for design matrix $X$,

- $\max_{i,j} |X_{ij}| = O(PolyLog(n))$;
- $\lambda_{max}\left(\frac{X^T X}{n}\right) = O(PolyLog(n))$;
- $\left\|\frac{1}{n}\sum_{i=1}^{n} x_i\right\| = O(PolyLog(n))$, where $x_i$ is the $i$-th row.

Assumption **A**4:

Let $x_i$ be the $i$-th row of $X$ and $X_j$ be the $j$-th column of $X$.
$\{\alpha_{k,i} \in \mathbb{R}^p : k = 1, \ldots, N_n^{(1)}; i = 1, \ldots, n\}$ and
$\{\gamma_{r,j} \in \mathbb{R}^n : r = 1, \ldots, N_n^{(2)}; j = 1, \ldots, p\}$ are two sequences of **unit vectors** (with explicit forms but omitted here for concision)

- $\max\{N_n^{(1)}, N_n^{(2)}\} = O(n^2)$.
- $\alpha_{k,i}$ only relies on $\epsilon$ and $x_{i'}$ for $i' \neq i$;
- $\gamma_{r,j}$ only relies on $\epsilon$ and $X_{j'}$ for $j' \neq j$;
- $\mathbb{E}_\epsilon \max_{k,i} |\alpha_{k,i}^T x_i| = O(PolyLog(n))$;
- $\mathbb{E}_\epsilon \max_{r,j} |\gamma_{r,j}^T X_j| = O(PolyLog(n))$;

Consider i.i.d. standard gaussian designs

$$X_{ij} \overset{i.i.d.}{\sim} N(0,1), \quad X \perp \epsilon.$$

For given $k$ and $i$, $\alpha_{k,i} \perp x_i$ and

$$\alpha_{k,i}^T x_i \sim N(0,1).$$

Then $\mathbb{E}_{\epsilon,X} \max_{k,i} |\alpha_{k,i}^T x_i|$ is the expectation of $N_n^{(1)}$ standard gaussian random variables and hence

$$\mathbb{E}_{\epsilon,X} \max_{k,i} |\alpha_{k,i}^T x_i| \preceq \sqrt{\log nN_n^{(1)}} = O(PolyLog(n)).$$

By Markov Inequality,

$$\mathbb{E}_{\epsilon} \max_{k,i} |\alpha_{k,i}^T x_i| = O_p\left(\mathbb{E}_{\epsilon,X} \max_{k,i} |\alpha_{k,i}^T x_i|\right) = O_p(PolyLog(n)).$$

# Table of Contents

1 Background

2 Main Results

3 OLS: A Motivating Example

# Lindeberg-Feller Condition

Assume that $p/n \to \kappa \in (0,1)$, $p < n$ and $\beta_0 = 0$, denote

$$\hat{\beta}_p^{OLS} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = (X^T X)^{-1} X^T \epsilon;$$

then each coordinate is a **linear constrast** of $\hat{\beta}$.

**Proposition 1 (Lindeberg-Feller Condition).**

*Suppose $\epsilon_n = (\epsilon_1, \ldots, \epsilon_n)^T$ has i.i.d. zero mean components with variance $\sigma^2$. If $||c_n||_\infty / ||c_n||_2 \to 0$ where $c_n = (c_{n,1}, \ldots, c_{n,k_n})$, then*

$$\frac{c_n^T \epsilon_n}{||c_n||_2} \xrightarrow{d} N(0, \sigma^2).$$

# Lindeberg-Feller Condition

Note that

$$\hat{\beta}^{OLS}_{p,j_p} = e_{j_p}^T (X^T X)^{-1} X^T \epsilon \triangleq c_{p,j_p}^T \epsilon.$$

For given matrix $X \in \mathbb{R}^{n \times p}$, define

$$H(X) \triangleq \max_{j=1,\dots,p} \frac{||e_j^T (X^T X)^{-1} X^T||_\infty}{||e_j^T (X^T X)^{-1} X^T||_2},$$

then for any $j_p \in \{1, \dots, p\}$,

$$\frac{||c_{p,j_p}^T||_\infty}{||c_{p,j_p}^T||_2} \leq H(X_p)$$

and this leads to

**Theorem 2.**

*$\hat{\beta}^{OLS}_p$ is c.a.s.n. if $H(X_p) \to 0$.*

We prove that $H(X_p) \to 0$ for a broad class of random designs.

**Theorem 3.**

*Let $X \in \mathbb{R}^{n\times p}$ be a random matrix with independent zero mean entries, such that $\sup_{i,j} ||X_{ij}||_{8+\delta} \le M$ for some constant $M$ and $\delta > 0$. Assume that $X$ has full column rank almost surely and $Var(X_{ij}) > \tau^2$ for some $\tau > 0$. Then*

$$H(X) = O_p(n^{-\frac{1}{4}})$$

*provided* $\limsup p/n < 1$.

# Future Works

- Extend to heavy-tailed errors, e.g. $\epsilon_i \sim Cauchy$;
- Explore more general random designs that satisfy **A**3 and **A**4;
- Calculate the bias $\mathbb{E}_{\epsilon}\hat{\beta}_j$ and variance $Var_{\epsilon}(\hat{\beta}_j)$;
- Prove the result for unregularized M estimator, i.e. $\tau = 0$;
- Extend to low dimensional linear contrasts of $\hat{\beta}$, i.e. $\alpha^T\hat{\beta}$ with $||\alpha||_0 = o(n)$.

# Thank You!