Hypothesis Testing for High-Dimensional Regression: Nearly Optimal Sample Size

Adel Javanmard

Stanford University- UC Berkeley

Based on joint work with

Andrea Montanari

January 2015

Outline

Problem definition

Linear model

We focus on linear models:

$$Y = \mathbf{X}\theta_0 + W$$

- $Y \in \mathbb{R}^n$ (response), $\mathbf{X} \in \mathbb{R}^{n \times p}$ (design matrix), $\theta_0 \in \mathbb{R}^p$ (parameters)
- Noise vector has independent entries with

$$\mathbb{E}(W_i) = 0, \qquad E(W_i^2) = \sigma^2\,,$$

$$\mathbb{E}(|W_i|^{2+\kappa}) < \infty\,, \text{ for some } \kappa > 0\,.$$

Problem

- **Confidence intervals:** For each $i \in \{1, \ldots, p\}$, $\underline{\theta}_i, \overline{\theta}_i \in \mathbb{R}$ such that

$$\mathbb{P}\Big(\theta_{0,i} \in [\underline{\theta}_i, \overline{\theta}_i]\Big) \geq 1 - \alpha$$

  We would like $|\underline{\theta}_i - \overline{\theta}_i|$ as small as possible.

- **Hypothesis testing:**

$$H_{0,i} : \theta_{0,i} = 0\,, \qquad H_{A,i} : \theta_{0,i} \neq 0$$

LASSO

$$\widehat{\theta} \equiv \underset{\theta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 + \lambda \|\theta\|_1 \right\}.$$

[Tibshirani 1996, Chen, Donoho 1996]

- Distribution of $\widehat{\theta}$?

LASSO

$$\widehat{\theta} \equiv \underset{\theta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 + \lambda \|\theta\|_1 \right\}.$$

[Tibshirani 1996, Chen, Donoho 1996]

- Distribution of $\widehat{\theta}$?
- Debiasing approach:
  (LASSO is biased towards small ℓ_1 norm.)

LASSO

$$\widehat{\theta} \equiv \underset{\theta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 + \lambda \|\theta\|_1 \right\}.$$

[Tibshirani 1996, Chen, Donoho 1996]

- Distribution of $\widehat{\theta}$?
- Debiasing approach:
  (LASSO is biased towards small ℓ_1 norm.)

$$\widehat{\theta} \xrightarrow{\textit{debiasing}} \widehat{\theta}^d$$

We characterize distribution of $\widehat{\theta}^d$.

Debiasing approach

Classical setting $(n \gg p)$

We know everything about the least-square estimator:

$$\widehat{\theta}^{\mathrm{LS}} = \frac{1}{n}\widehat{\Sigma}^{-1}\mathbf{X}^{\mathsf{T}}Y\,,$$

where $\widehat{\Sigma} \equiv (\mathbf{X}^{\mathsf{T}}\mathbf{X})/n$ is empirical covariance.

Classical setting $(n \gg p)$

We know everything about the least-square estimator:

$$\widehat{\theta}^{\rm LS} = \frac{1}{n}\widehat{\Sigma}^{-1}\mathbf{X}^{\mathsf{T}}Y\,,$$

where $\widehat{\Sigma} \equiv (\mathbf{X}^{\mathsf{T}}\mathbf{X})/n$ is empirical covariance.

- Confidence intervals:

$$[\underline{\theta}_i, \overline{\theta}_i] = [\widehat{\theta}_i^{\rm LS} - c_\alpha \Delta_i, \widehat{\theta}_i^{\rm LS} + c_\alpha \Delta_i]\,, \qquad \Delta_i \equiv \sigma\sqrt{\frac{(\widehat{\Sigma}^{-1})_{ii}}{n}}$$

High-dimensional setting ($n < p$)

$$\widehat{\theta}^{\mathrm{LS}} = \frac{1}{n}\widehat{\Sigma}^{-1}\mathbf{X}^{\mathsf{T}}Y$$

Problem in high dimension:

$\widehat{\Sigma}$ is not invertible!

High-dimensional setting ($n < p$)

$$\widehat{\theta}^{\mathrm{LS}} = \frac{1}{n}\widehat{\Sigma}^{-1}\mathbf{X}^{\mathsf{T}}Y$$

Take your favorite $M \in \mathbb{R}^{p\times p}$:

$$\begin{aligned}
\widehat{\theta}^{*} &= \frac{1}{n}M\mathbf{X}^{\mathsf{T}}Y \\
&= \frac{1}{n}M\mathbf{X}^{\mathsf{T}}\mathbf{X}\theta_0 + \frac{1}{n}M\mathbf{X}^{\mathsf{T}}W \\
&= \theta_0 + \underbrace{(M\widehat{\Sigma} - \mathbf{I})\theta_0}_{\text{bias}} + \underbrace{\frac{1}{n}M\mathbf{X}^{\mathsf{T}}W}_{\text{Gaussian error}}
\end{aligned}$$

Debiased estimator

$$\widehat{\theta}^* = \theta_0 + \underbrace{(M\widehat{\Sigma} - \mathbf{I})\theta_0}_{\text{bias}} + \underbrace{\frac{1}{n} M\mathbf{X}^\mathsf{T} W}_{\text{Gaussian error}}$$

Debiased estimator

$$\widehat{\theta}^* = \theta_0 + \underbrace{(M\widehat{\Sigma} - \mathbf{I})\theta_0}_{\text{bias}} + \underbrace{\frac{1}{n} M\mathbf{X}^{\mathsf{T}} W}_{\text{Gaussian error}}$$

Let us (try to) subtract the bias

$$\widehat{\theta}^u = \widehat{\theta}^* - (M\widehat{\Sigma} - \mathbf{I})\widehat{\theta}^{\text{Lasso}}$$

Debiased estimator

$$\widehat{\theta}^* = \theta_0 + \underbrace{(M\widehat{\Sigma} - \mathbf{I})\theta_0}_{\text{bias}} + \underbrace{\frac{1}{n} M\mathbf{X}^\mathsf{T} W}_{\text{Gaussian error}}$$

Let us (try to) subtract the bias

$$\widehat{\theta}^u = \widehat{\theta}^* - (M\widehat{\Sigma} - \mathbf{I})\widehat{\theta}^{\text{Lasso}}$$

Debiased estimator ($\widehat{\theta} = \widehat{\theta}^{\text{Lasso}}$)

$$\widehat{\theta}^d \equiv \widehat{\theta} + \frac{1}{n} M\mathbf{X}^\mathsf{T}(Y - \mathbf{X}\widehat{\theta})$$

Debiased estimator: Choosing M?

$$\widehat{\theta}^d \equiv \widehat{\theta} + \frac{1}{n} M \mathbf{X}^{\mathsf{T}} (y - \mathbf{X}\widehat{\theta})$$

- Low dimensional projection estimator (LDPE)
  - Start with a linear estimator, debias by a nonlinear estimator
  - M constructed via nodewise LASSO on $\mathbf{X}$

[C-H. Zhang, S. S. Zhang]

Debiased estimator: Choosing M?

$$\widehat{\theta}^d \equiv \widehat{\theta} + \frac{1}{n} M \mathbf{X}^{\mathsf{T}} (y - \mathbf{X}\widehat{\theta})$$

- Low dimensional projection estimator (LDPE)
  - Start with a linear estimator, debias by a nonlinear estimator
  - M constructed via nodewise LASSO on $\mathbf{X}$

[C-H. Zhang, S. S. Zhang]

- Approximate inverse of $\widehat{\Sigma}$: nodewise LASSO on $\mathbf{X}$
  (under row-sparsity assumption on Σ^{-1})

[S. van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure]

Debiased estimator: Choosing M?

Our approach:

- Optimizing two objectives (bias and variance of $\widehat{\theta}^d$)

[A. Javanmard, A. Montanari]

$$\sqrt{n}(\widehat{\theta}^d - \theta_0) = \underbrace{\sqrt{n}(M\widehat{\Sigma} - \mathbf{I})(\theta_0 - \widehat{\theta})}_{bias\downarrow} + Z$$

$$Z|\mathbf{X} \sim \mathrm{N}(0, \underbrace{\sigma^2 M\widehat{\Sigma}M^{\mathsf{T}}}_{\substack{noise \\ covariance\downarrow}}), \qquad \widehat{\Sigma} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}$$

Debiased estimator: Choosing M?

Our approach:

- Find M by solving an optimization problem:

[A. Javanmard, A. Montanari]

$$\begin{aligned} \underset{M}{\text{minimize}} \quad & \max_{1\le i\le p} \left(M\widehat{\Sigma}M^{\mathsf{T}}\right)_{i,i} \\ \text{subject to} \quad & \left|M\widehat{\Sigma}-\mathbf{I}\right|_{\infty} \le \xi \end{aligned}$$

Debiased estimator: Choosing M?

Our approach:

- Find M by solving an optimization problem:

[A. Javanmard, A. Montanari]

$$\begin{aligned} \underset{m_i}{\text{minimize}} \quad & m_i^{\mathsf{T}} \widehat{\Sigma} m_i \\ \text{subject to} \quad & \|\widehat{\Sigma} m_i - e_i\|_\infty \leq \xi \end{aligned}$$

The optimization can be decoupled and solved in parallel.

Main theorems

Theorem [Javanmard, Montanari 2013] (Deterministic designs)

Let $\mathbf{X}$ be any deterministic design that satisfies compatibility condition for the set $S = \mathrm{supp}(\theta_0)$, ($|S| \le s_0$), with constant ϕ_0 . Further define the coherence parameter

$$\mu_* \equiv \min_{M \in \mathbb{R}^{p\times p}} |M\widehat{\Sigma} - \mathbf{I}|_\infty .$$

Let $K \equiv \max_{i\in[p]} \widehat{\Sigma}_{ii}$. Then, letting $\lambda = c\sigma\sqrt{\log p/n}$, we have

$$\sqrt{n}(\widehat{\theta}^d - \theta_0) = Z + \Delta, \qquad Z \sim \mathrm{N}(0, \sigma^2 M\widehat{\Sigma}M^{\mathsf{T}})$$

$$\mathbb{P}\left(\|\Delta\|_\infty \ge \frac{4c\mu_*\sigma s_0}{\phi_0^2}\sqrt{\log p}\right) \le 2p^{-c_0}, \quad c_0 = \frac{c^2}{32K} - 1$$

Main theorems

Theorem [Javanmard, Montanari 2013] (Deterministic designs)

Let $\mathbf{X}$ be any deterministic design that satisfies compatibility condition for the set $S = \mathrm{supp}(\theta_0)$, ($|S| \le s_0$), with constant ϕ_0 . Further define the coherence parameter

$$\mu_* \equiv \min_{M \in \mathbb{R}^{p\times p}} |M\widehat{\Sigma} - \mathbf{I}|_\infty .$$

Let $K \equiv \max_{i\in[p]} \widehat{\Sigma}_{ii}$. Then, letting $\lambda = c\sigma\sqrt{\log p/n}$, we have

$$\sqrt{n}(\widehat{\theta}^d - \theta_0) = Z + \Delta, \qquad Z \sim \mathrm{N}(0, \sigma^2 M\widehat{\Sigma}M^{\mathsf{T}})$$

$$\mathbb{P}\left(\|\Delta\|_\infty \ge \frac{4c\mu_*\sigma s_0}{\phi_0^2}\sqrt{\log p} \right) \le 2p^{-c_0}, \quad c_0 = \frac{c^2}{32K} - 1$$

Remark:

$$\mu_* \le \frac{1}{n}\max_{i\neq j} |\langle \mathbf{X}e_i, \mathbf{X}e_j\rangle| .$$

Main theorems

Theorem [Javanmard, Montanari 2013] (Random designs)

Let Σ be such that $\sigma_{\min}(\Sigma) \geq C_{\min} > 0$ and $\sigma_{\max}(\Sigma) \leq C_{\max} < \infty$ and $\max_{i\in[p]} \Sigma_{ii} \leq 1$. Assume $\mathbf{X}\Sigma^{-1}$ has independent subgaussian rows with mean zero and subgaussian norm K. Letting $\lambda = c\sigma\sqrt{\log p/n}$, we have

$$\sqrt{n}(\widehat{\theta}^d - \theta_0) = Z + \Delta, \qquad Z|\mathbf{X} \sim \mathrm{N}(0, \sigma^2 M\widehat{\Sigma}M^{\mathsf{T}}),$$

$$\mathbb{P}\Big\{ \|\Delta\|_\infty \geq (\frac{16c\sigma}{C_{\min}})\frac{s_0 \log p}{\sqrt{n}} \Big\} \leq 4e^{-c_1 n} + 4p^{-c_2},$$

for some explicit constants $c_1 = C(K)$, $c_2 = C(c, K, C_{\min}, C_{\max})$.

Main theorems

Theorem [Javanmard, Montanari 2013] (Random designs)

Let Σ be such that $\sigma_{\min}(\Sigma) \geq C_{\min} > 0$ and $\sigma_{\max}(\Sigma) \leq C_{\max} < \infty$ and $\max_{i\in[p]} \Sigma_{ii} \leq 1$. Assume $\mathbf{X}\Sigma^{-1}$ has independent subgaussian rows with mean zero and subgaussian norm K. Letting $\lambda = c\sigma\sqrt{\log p/n}$, we have

$$\sqrt{n}(\widehat{\theta}^d - \theta_0) = Z + \Delta, \qquad Z|\mathbf{X} \sim \mathrm{N}(0, \sigma^2 M\widehat{\Sigma}M^{\mathsf{T}}),$$

$$\mathbb{P}\Big\{ \|\Delta\|_\infty \geq (\frac{16c\sigma}{C_{\min}})\frac{s_0 \log p}{\sqrt{n}} \Big\} \leq 4e^{-c_1 n} + 4p^{-c_2},$$

for some explicit constants $c_1 = C(K)$, $c_2 = C(c, K, C_{\min}, C_{\max})$.

Remark on sample size:

$$\text{If } \frac{n}{(s_0 \log p)^2} \to \infty \text{ then } \|\Delta\|_\infty = o_p(1).$$

Consequences

- Confidence intervals for single parameters:

$$\lim_{n\to\infty} \mathbb{P}\Big(\theta_{0,i} \in [\underline{\theta}_i, \overline{\theta}_i]\Big) \geq 1-\alpha$$

$$|\underline{\theta}_i - \overline{\theta}_i| \leq (2+o(1))c_\alpha \sqrt{\frac{\sigma^2}{n}(\Sigma^{-1})_{ii}}$$

(n<p)

Consequences

- Confidence intervals for single parameters:

$$\lim_{n\to\infty} \mathbb{P}\Big(\theta_{0,i} \in [\underline{\theta}_i, \overline{\theta}_i]\Big) \geq 1-\alpha$$

$$|\underline{\theta}_i - \overline{\theta}_i| \leq (2+o(1))c_\alpha \sqrt{\frac{\sigma^2}{n}(\Sigma^{-1})_{ii}}$$

(n<p)

$$|\underline{\theta}_i - \overline{\theta}_i| \leq 2c_\alpha \sqrt{\frac{\sigma^2}{n}(\widehat{\Sigma}^{-1})_{ii}}$$

Least square (n>p)

Consequences

- Confidence intervals for single parameters:

$$\lim_{n\to\infty} \mathbb{P}\Big(\theta_{0,i} \in [\underline{\theta}_i, \overline{\theta}_i]\Big) \geq 1-\alpha$$

$$|\underline{\theta}_i - \overline{\theta}_i| \leq (2+o(1))c_\alpha \sqrt{\frac{\sigma^2}{n}(\Sigma^{-1})_{ii}}$$

(n<p)

$$|\underline{\theta}_i - \overline{\theta}_i| \leq 2c_\alpha \sqrt{\frac{\sigma^2}{n}(\widehat{\Sigma}^{-1})_{ii}}$$

Least square (n>p)

Remark:

No need for irrepresentability / $\theta_{\min}$ condition (common assumptions for support recovery)

Consequences

- Confidence intervals for single parameters:

$$\lim_{n\to\infty} \mathbb{P}\Big(\theta_{0,i} \in [\underline{\theta}_i, \overline{\theta}_i]\Big) \geq 1-\alpha$$

$$|\underline{\theta}_i - \overline{\theta}_i| \leq (2+o(1))c_\alpha \sqrt{\frac{\sigma^2}{n}(\Sigma^{-1})_{ii}}$$

(n<p)

- Hypothesis testing: minimax optimal statistical power

Framework

- Hypothesis testing

$$H_{0,i} : \theta_{0,i} = 0\,, \qquad H_{A,i} : \theta_{0,i} \neq 0\,.$$

- Two-sided p-values:

$$P_i = 2\Big(1 - \Phi\big(\frac{|\widehat{\theta}_i^d|}{\tau}\big)\Big)$$

  with $\Phi(\cdot)$ cdf of standard normal.

- Decision rule:

$$T_{i,\mathbf{X}}(y) = \begin{cases} 1 & \text{if } P_i \leq \alpha \quad \text{(reject the null hypothesis } H_{0,i}\text{)}\,, \\ 0 & \text{otherwise} \quad \text{(accept the null hypothesis).} \end{cases}$$

Theorem [Javanmard, Montanari, 2013]

Consider designs with subgaussian rows and let $S \equiv \mathrm{supp}(\theta_0)$.
Assume that $s_0 \equiv |S| = o(\sqrt{n}/\log p)$.
Then, for any fixed sequence of integers $i = i(n)$, we have that for $i \notin S$

$$\lim_{n\to\infty} \mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y) = 1) \le \alpha\,.$$

Further, assuming that for all $i \in S$, $|\theta_{0,i}| \ge \mu$, we have

$$\liminf_{p\to\infty} \frac{1}{1-\beta_{i,n}(\alpha,\mu)} \mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y) = 1) \ge 1\,,$$

$$1-\beta_{i,n}(\alpha,\mu) \equiv G\Big(\alpha, \frac{\sqrt{n}\mu}{\sigma\sqrt{(\Sigma^{-1})_{ii}}}\Big)\,.$$

with $G(\alpha,u)$ given by ...

Theorem [Javanmard, Montanari, 2013]

Consider designs with subgaussian rows and let $S \equiv \mathrm{supp}(\theta_0)$.
Assume that $s_0 \equiv |S| = o(\sqrt{n}/\log p)$.
Then, for any fixed sequence of integers $i = i(n)$, we have that for $i \notin S$

$$\lim_{n\to\infty} \mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y) = 1) \le \alpha\,.$$

Further, assuming that for all $i \in S$, $|\theta_{0,i}| \ge \mu$, we have

$$\liminf_{p\to\infty} \frac{1}{1-\beta_{i,n}(\alpha,\mu)} \mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y) = 1) \ge 1\,,$$

$$1-\beta_{i,n}(\alpha,\mu) \equiv G\Big(\alpha, \frac{\sqrt{n}\mu}{\sigma\sqrt{(\Sigma^{-1})_{ii}}}\Big)\,.$$

with $G(\alpha,u)$ given by ...

Minimax optimal power over the family of s_0-sparse vectors θ_0.

Related work on bias-correction

- Ridge projection and bias correction [P. Bühlmann]
  - (Remaining) bias is not negligible.
  - Conservative tests

- Low dimensional projection estimator (LDPE) [C-H. Zhang, S. S. Zhang]
  - Initial projection based on nodewise LASSO on $\mathbf{X}$.
  - Bias correction via LASSO.

Hypothesis testing under nearly optimal sample size

Smaller sample size

- Estimation, prediction: $n \gtrsim s_0 \log p$.

  [Candés, Tao 2007, Bickel et al. 2009]

- Hypothesis testing, confidence intervals: $n \gtrsim (s_0 \log p)^2$.

  [This talk]

  - Bias corrected ridge regression [P. Bühlmann]
  - LDPE [C-H. Zhang, S. S. Zhang]
  - Desparsified LASSO [S. van de Geer et. al.]

Smaller sample size

- Estimation, prediction: $n \gtrsim s_0 \log p$.

[Candés, Tao 2007, Bickel et al. 2009]

- Hypothesis testing, confidence intervals: $n \gtrsim (s_0 \log p)^2$.

[This talk]

  - Bias corrected ridge regression [P. Bühlmann]
  - LDPE [C-H. Zhang, S. S. Zhang]
  - Desparsified LASSO [S. van de Geer et. al.]

Can we match the optimal sample size, $n \gtrsim s_0 \log p$?

Theorem [Javanmard, Montanari 2013]

Consider designs with subgassian rows and assume $n \gtrsim s_0(\log p)^2$. Then

$$\limsup_{p\to\infty} \frac{1}{p-s_0} \sum_{i\in S^c} \mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y)=1) \le \alpha\,.$$

Further, assuming that for all $i \in S$, $|\theta_{0,i}| \ge \mu$, we have

$$\liminf_{p\to\infty} \frac{1}{1-\beta_n^*(\alpha,\theta_0)} \Big\{ \frac{1}{s_0} \sum_{i\in S} \mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y)=1) \Big\} \ge 1\,,$$

where

$$1-\beta_n^*(\alpha,\theta_0) \equiv \frac{1}{s_0} \sum_{i\in S} G\Big(\alpha, \frac{\sqrt{n}|\theta_{0,i}|}{\sigma\sqrt{(\Sigma^{-1})_{ii}}}\Big)\,.$$

Theorem [Javanmard, Montanari 2013]

Consider designs with subgassian rows and assume $n \gtrsim s_0(\log p)^2$. Then

$$\limsup_{p\to\infty} \frac{1}{p-s_0} \sum_{i\in S^c} \mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y) = 1) \le \alpha\,.$$

Further, assuming that for all $i \in S$, $|\theta_{0,i}| \ge \mu$, we have

$$\liminf_{p\to\infty} \frac{1}{1-\beta_n^*(\alpha,\theta_0)} \Big\{ \frac{1}{s_0} \sum_{i\in S} \mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y) = 1) \Big\} \ge 1\,,$$

where

$$1-\beta_n^*(\alpha,\theta_0) \equiv \frac{1}{s_0} \sum_{i\in S} G\Big(\alpha, \frac{\sqrt{n}|\theta_{0,i}|}{\sigma\sqrt{(\Sigma^{-1})_{ii}}}\Big)\,.$$

- Controls average type I error
- Minimax optimal average power

High-level idea of the proof

Recall

$$\sqrt{n}(\widehat{\theta}^d - \theta_0) = Z + \underbrace{\Delta}_{\text{bias}}.$$

$$\|\Delta\|_\infty = O_p\left(\frac{s_0 \log p}{\sqrt{n}}\right) \longrightarrow n \gtrsim (s_0 \log p)^2$$

High-level idea of the proof

Recall

$$\sqrt{n}(\widehat{\theta}^d - \theta_0) = Z + \underbrace{\Delta}_{\text{bias}}.$$

$$\|\Delta\|_\infty = O_p\left(\frac{s_0 \log p}{\sqrt{n}}\right) \longrightarrow n \gtrsim (s_0 \log p)^2$$

To ensure average performance, we do not need to control $\|\Delta\|_\infty$.

High-level idea of the proof

A new norm:

$$\|\Delta\|_{(\infty,k)} \equiv \max_{A\subset[p],|A|\geq k} \frac{\|\Delta_A\|_2}{\sqrt{|A|}}\,.$$

High-level idea of the proof

A new norm:

$$\|\Delta\|_{(\infty,k)} \equiv \max_{A\subset[p],|A|\geq k} \frac{\|\Delta_A\|_2}{\sqrt{|A|}}\,.$$

Properties of $\|\cdot\|_{\infty,k}$:

- Non-increasing in k
- As k gets smaller, it gives tighter control on the individual entries of Δ.

High-level idea of the proof

A new norm:

$$\|\Delta\|_{(\infty,k)} \equiv \max_{A\subset[p],|A|\geq k} \frac{\|\Delta_A\|_2}{\sqrt{|A|}}\,.$$

Properties of $\|\cdot\|_{\infty,k}$:

- Non-increasing in k
- As k gets smaller, it gives tighter control on the individual entries of Δ.

Lemma

$$\|\Delta\|_{(\infty,cs_0)} = O\Big(\sqrt{\frac{s_0}{n}}\log p\Big)\,.$$

A few steps of the proof

- Any set $|A| \geq cs_0$ can be partitioned as

$$A = A_1 \cup A_2 \cup \ldots \cup A_L$$

with A_i disjoint and $cs_0 \leq |A_i| \leq 2cs_0$.

A few steps of the proof

- Any set $|A| \geq cs_0$ can be partitioned as

$$A = A_1 \cup A_2 \cup \ldots \cup A_L$$

with A_i disjoint and $cs_0 \leq |A_i| \leq 2cs_0$.
(WLOG, we can assume $cs_0 \leq |A| \leq 2cs_0$.)

A few steps of the proof

- Any set $|A| \geq cs_0$ can be partitioned as

$$A = A_1 \cup A_2 \cup \ldots \cup A_L$$

with A_i disjoint and $cs_0 \leq |A_i| \leq 2cs_0$.
(WLOG, we can assume $cs_0 \leq |A| \leq 2cs_0$.)

- $$\Delta \equiv \sqrt{n}(M\widehat{\Sigma} - \mathbf{I})(\widehat{\theta} - \theta_0)\,.$$

A few steps of the proof

- Any set $|A| \geq cs_0$ can be partitioned as

$$A = A_1 \cup A_2 \cup \ldots \cup A_L$$

with A_i disjoint and $cs_0 \leq |A_i| \leq 2cs_0$.
(WLOG, we can assume $cs_0 \leq |A| \leq 2cs_0$.)

- $$\Delta \equiv \sqrt{n}(\Sigma^{-1}\widehat{\Sigma} - \mathbf{I})(\widehat{\theta} - \theta_0)\,.$$

A few steps of the proof

- Any set $|A| \geq cs_0$ can be partitioned as

$$A = A_1 \cup A_2 \cup \ldots \cup A_L$$

with A_i disjoint and $cs_0 \leq |A_i| \leq 2cs_0$.
(WLOG, we can assume $cs_0 \leq |A| \leq 2cs_0$.)

-

$$\Delta \equiv \sqrt{n}(\Sigma^{-1}\widehat{\Sigma} - \mathbf{I})(\widehat{\theta} - \theta_0)\,.$$

Let $T = \mathrm{supp}(\theta_0) \cup \mathrm{supp}(\widehat{\theta})$. We have

$$\|\Delta_A\|_2 \leq \sqrt{n}\|(\Sigma^{-1}\widehat{\Sigma} - \mathbf{I})_{A,T}\|_2 \|(\widehat{\theta} - \theta_0)_T\|_2\,.$$

A few steps of the proof (cont'd)

- Applying tail bounds we get

$$\sup_{\substack{|A|\le cs_0\\ |T|\le c's_0}} \|(\Sigma^{-1}\widehat{\Sigma}-\mathbf{I})_{A,T}\|_2 = O\Big(\sqrt{\frac{s_0\log p}{n}}\Big).$$

  We also know that

$$\|\widehat{\theta}-\theta_0\|_2 = O\Big(\sqrt{\frac{s_0\log p}{n}}\Big).$$

- Combining the bounds, we get

$$\frac{\|\Delta_A\|_2}{|A|} \le O\Big(\sqrt{\frac{s_0}{n}}\log p\Big).$$

A few steps of the proof (cont'd)

- Applying tail bounds we get

$$\sup_{\substack{|A|\le cs_0\\ |T|\le c's_0}} \|(\Sigma^{-1}\widehat{\Sigma}-\mathbf{I})_{A,T}\|_2 = O\Big(\sqrt{\frac{s_0\log p}{n}}\Big).$$

  We also know that

$$\|\widehat{\theta}-\theta_0\|_2 = O\Big(\sqrt{\frac{s_0\log p}{n}}\Big).$$

- Combining the bounds, we get

$$\frac{\|\Delta_A\|_2}{|A|} \le O\Big(\sqrt{\frac{s_0}{n}}\log p\Big).$$

$$n \gtrsim s_0(\log p)^2$$

Standard gaussian design

Suppose $X_{ij} \sim \mathrm{N}(0,1)$ independently.

$$\widehat{\theta}^d = \widehat{\theta} + \frac{1}{n}\mathbf{X}^\mathsf{T}(y - \mathbf{X}\widehat{\theta})$$

Standard gaussian design

Suppose $X_{ij} \sim \mathrm{N}(0,1)$ independently.

$$\widehat{\theta}^d = \widehat{\theta} + \frac{1}{n}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta})$$

- SDL test [J., Montanari 2013]

$$\widehat{\theta}^d = \widehat{\theta} + \frac{\mathrm{d}}{n}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta})$$

Standard gaussian design

Suppose $X_{ij} \sim \mathrm{N}(0,1)$ independently.

$$\widehat{\theta}^d = \widehat{\theta} + \frac{1}{n}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta})$$

- SDL test [J., Montanari 2013]

$$\begin{aligned} \widehat{\theta}^d &= \widehat{\theta} + \frac{\mathrm{d}}{n}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta}) \\ \mathrm{d} &= \left(1 - \frac{1}{n}\|\widehat{\theta}\|_0\right)^{-1} \end{aligned}$$

Based on the analysis of Approximate Message Passing (AMP).

[M. Bayati, D. Donoho, A. Maleki, A. Montanari]

Exact asymptotic characterization

$$\widehat{\theta}^d \equiv \widehat{\theta} + \frac{\mathsf{d}}{n}\mathbf{X}^{\mathsf{T}}(y - \mathbf{X}\widehat{\theta})$$

Theorem [M. Bayati, A. Montanari 2012]

Consider the standard gaussian setting where $n/p \to \delta$, $s_0/p \to \varepsilon$ and $n\sigma^2 \to \sigma_\infty^2$. If $\delta \geq \varepsilon \log(1/\varepsilon)$, then on finite-dimensional marginals

$$\widehat{\theta}^d = \theta_0 + \tau Z\,, \qquad Z \sim \mathrm{N}(0, \mathbf{I}_{p\times p})\,,$$

with τ, d given by ...

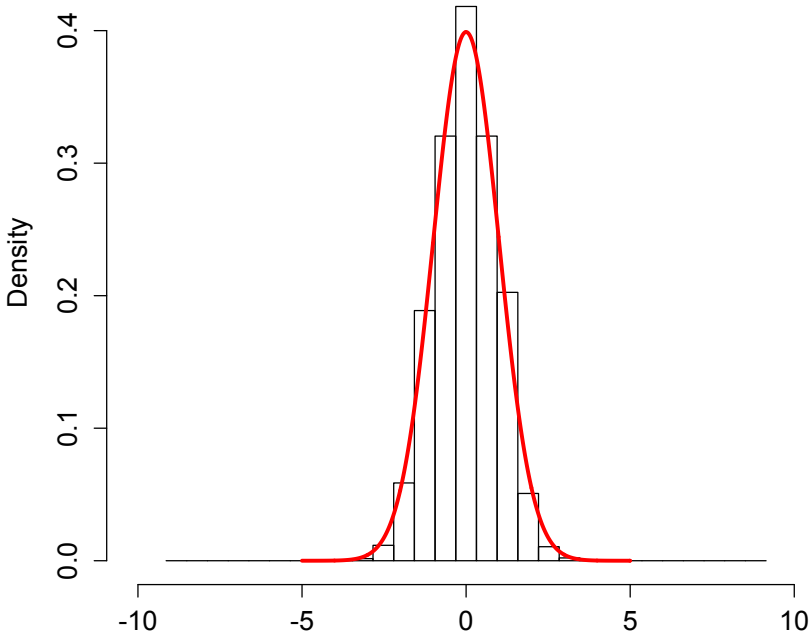
$$\delta \geq \varepsilon \log(1/\varepsilon) \longrightarrow n \geq s_0 \log(p/s_0)$$

Effect of factor d

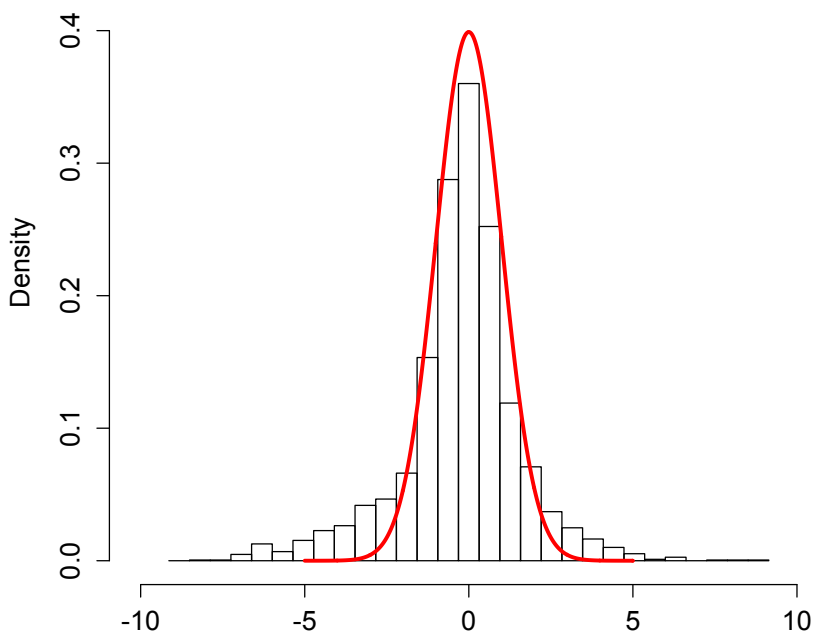
$$\mathsf{d} = \left(1 - \frac{\|\widehat{\theta}\|_0}{n}\right)^{-1} = 1 + O\left(\frac{s_0}{n}\right)$$

Effect of factor d

$$\mathsf{d} = \left(1 - \frac{\|\widehat{\theta}\|_0}{n}\right)^{-1} = 1 + O\left(\frac{s_0}{n}\right)$$



with factor d

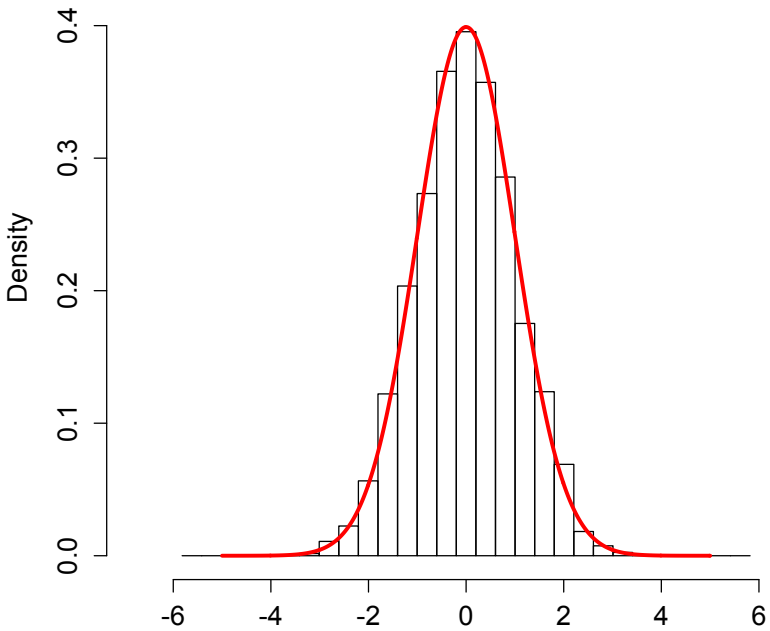


without factor d

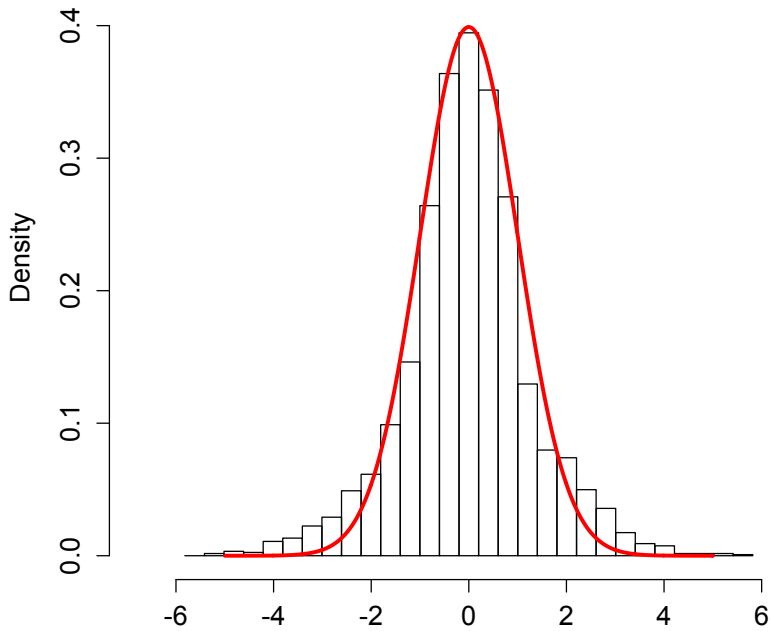
Figure: Histogram of $v = (\widehat{\theta}^d - \theta_0)/\tau$ for $n = 3\,s_0$ ($\varepsilon = 0.2$, $\delta = 0.6$) and $p = 3000$.

Effect of factor d

$$\mathsf{d} = \Big(1 - \frac{\|\widehat{\theta}\|_0}{n}\Big)^{-1} = 1 + O\Big(\frac{s_0}{n}\Big)$$



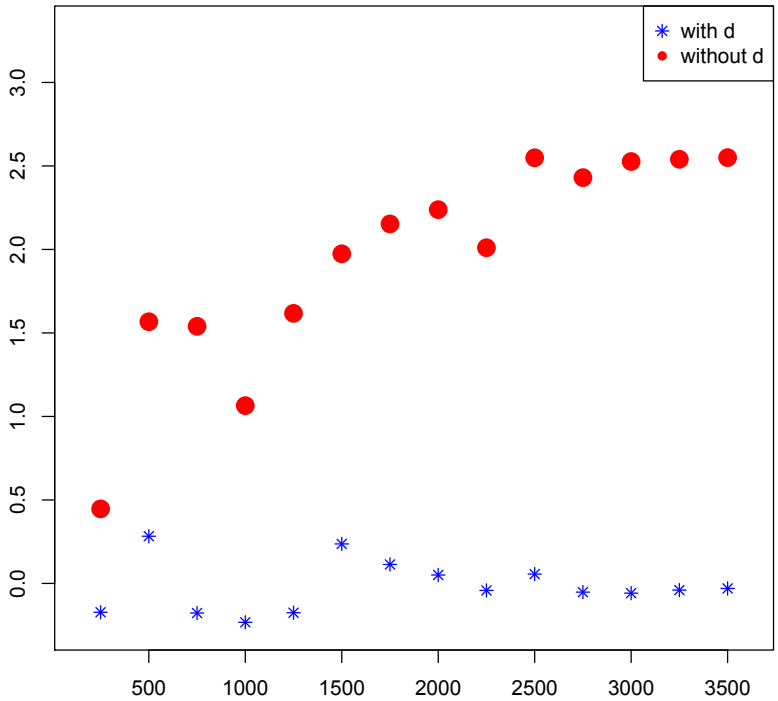
with factor d



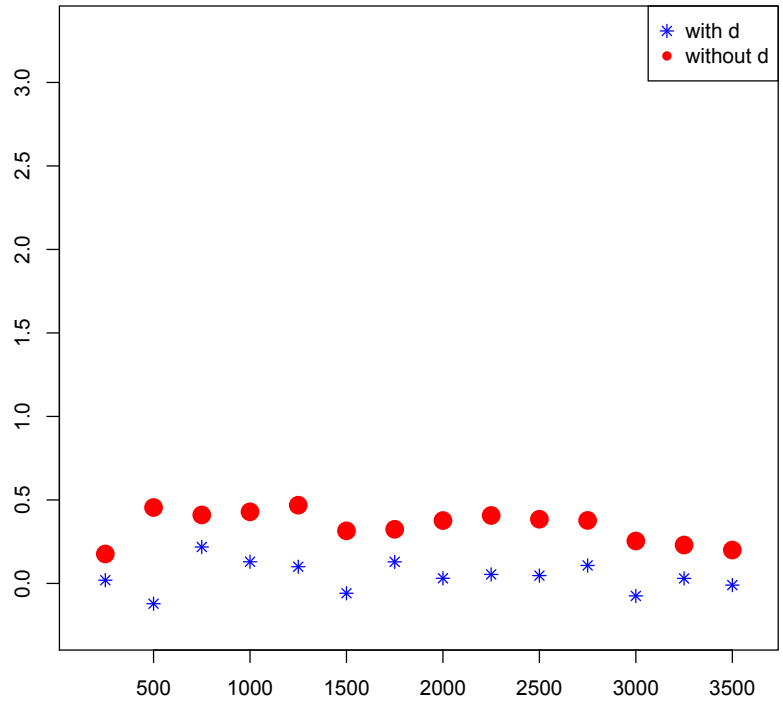
without factor d

Figure: Histogram of $v = (\widehat{\theta}^d - \theta_0)/\tau$ for $n = 30\, s_0$ ($\varepsilon = 0.02$, $\delta = 0.6$) and $p = 3000$.

Exact asymptotic characterization



$n = 3s_0$



$n = 30s_0$

Figure: Empirical kurtosis of $v = (\widehat{\theta}^d - \theta_0)/\tau$ with and without normalization factor d.

Theorem [Javanmard, Montanari, 2013]

Consider the setting where $n/p \to \delta$, $s_0/p \to \varepsilon$ and $\delta \geq \varepsilon \log(1/\varepsilon)$. Then, for $i \notin S$ we have

$$\lim_{p\to\infty} \mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y) = 1) = \alpha\,.$$

Further, assuming that for all $i \in S$, $|\theta_{0,i}| \geq \mu$, we have

$$\lim_{p\to\infty} \mathbb{P}_{\theta_0}(T_{i,\mathbf{X}}(y) = 1) \geq G\Big(\alpha, \frac{\mu}{\tau}\Big)\,,$$

with

$$G(\alpha,u) \equiv 2 - \Phi\Big(\Phi^{-1}(1-\frac{\alpha}{2}) + u\Big) - \Phi\Big(\Phi^{-1}(1-\frac{\alpha}{2}) - u\Big)\,.$$

Summary

- Random designs
  - $n \gtrsim s_0(\log p)^2$ ✓
    - guarantee on average type I error and power
    - requires good estimate of precision matrix
      (can be done e.g., under sparsity assumption)
  - $n \gtrsim s_0 \log p$?
- Standard gaussian designs
  - $n \gtrsim s_0 \log(p/s_0)$ ✓


[1] A. Javanmard and A. Montanari, *Confidence Intervals and Hypothesis Testing for High-Dimensional Regression*. JMLR, 2014


[2] A. Javanmard and A. Montanari, *Nearly Optimal Sample Size in Hypothesis Testing for High-Dimensional Regression*. Allerton, 2013.


[3] A. Javanmard and A. Montanari, *Hypothesis Testing in High-Dimensional Regression under the Gaussian Random Design Model: Asymptotic Theory*. IEEE transaction on Info. Theory, 2013.

Thanks!