De-biasing the Lasso: Optimal Sample Size for Gaussian Designs

Adel Javanmard

USC Marshall School of Business
Data Science and Operations department

Based on joint work with

Andrea Montanari

Oct 2015

An example

Kaggle challenge: Identify patients diagnosed with type-2 diabetes

Statistical model

Data $(Y_1, X_1), \ldots, (Y_n, X_n)$:

- $Y_i =$ Patient i gets type-2 diabetes $\quad \in \{0, 1\}$
- $X_i =$ Features of patient i $\quad \in \mathbb{R}^p$

$$Y_i \sim f_{\theta_0}(\cdot | X_i) \qquad \theta_0 \in \mathbb{R}^p$$

$\theta_{0,j} =$ contribution of feature j

Statistical model

Data $(Y_1, X_1), \ldots, (Y_n, X_n)$:

- $Y_i =$ Patient i gets type-2 diabetes $\in \{0, 1\}$
- $X_i =$ Features of patient i $\in \mathbb{R}^p$

$$Y_i \sim f_{\theta_0}(\,\cdot\,|X_i) \qquad \theta_0 \in \mathbb{R}^p$$

$\theta_{0,j} =$ contribution of feature j

Statistical model

Data $(Y_1, X_1), \dots, (Y_n, X_n)$:

- $Y_i =$ Patient i gets type-2 diabetes $\in \{0, 1\}$
- $X_i =$ Features of patient i $\in \mathbb{R}^p$

$$Y_i \sim f_{\theta_0}(\,\cdot\,|X_i) \qquad \theta_0 \in \mathbb{R}^p$$

$\theta_{0,j} =$ contribution of feature j

Regularized estimator

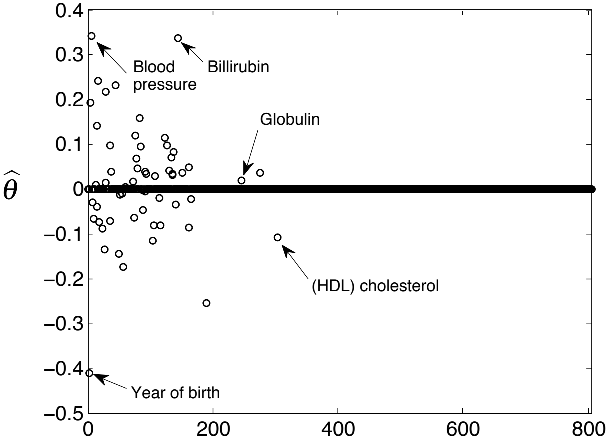
$$\widehat{\theta} \equiv \arg\min_{\theta \in \mathbb{R}^p} \Big(\underbrace{\mathscr{L}(\theta)}_{\text{logistic loss}} + \lambda \underbrace{\|\theta\|_1}_{\text{regularizer}} \Big).$$

- Convex optimization
- Variable selection

Practice fusion data set (Kaggle)

Database

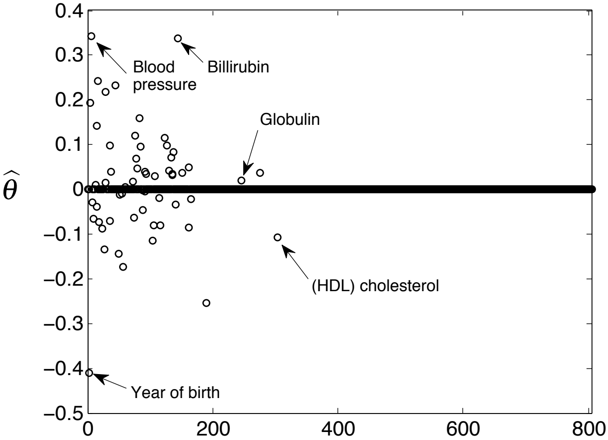
- $n = 500$: patients
- $p = 805$: medical information
  (meds, lab results, diagnosis, ...)



Regularized logreg selects 62 features

(λ chosen via cross validation resulting AUC $= 0.75$)

Shall we trust our findings?



Regularized logreg selects 62 features
(λ chosen via cross validation resulting AUC $= 0.75$)

Shall we trust our findings?

In summary

- Will focus on linear model and Lasso

- Compute confidence intervals/p-values

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 $i \in \{1, \dots, p\}$, $\underline{\theta}_i, \overline{\theta}_i \in \mathbb{R}$ 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:**

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

LASSO

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

[Tibshirani 1996, Chen, Donoho 1996]

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

LASSO

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

[Tibshirani 1996, Chen, Donoho 1996]

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

LASSO

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

[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:

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

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:

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

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

- Confidence intervals:

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

High-dimensional setting ($n < p$)

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

Problem in high dimension:

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

High-dimensional setting ($n < p$)

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

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

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

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

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

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

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

Debiased estimator: Choosing M?

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

- Gaussian design ($x_i \sim \mathrm{N}(0, \Sigma)$)
  - Assume known Σ (relevant in semi-supervised learning)
  - $M = \Sigma^{-1}$

[Javanmard, Montanari 2012]

Debiased estimator: Choosing M?

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

- Gaussian design ($x_i \sim \mathrm{N}(0, \Sigma)$)
  - Assume known Σ (relevant in semi-supervised learning)
  - $M = \Sigma^{-1}$

[Javanmard, Montanari 2012]

Does this remind you anything?

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

Debiased estimator: Choosing M?

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

- Gaussian design ($x_i \sim \mathrm{N}(0, \Sigma)$)
  - Assume known Σ (relevant in semi-supervised learning)
  - $M = \Sigma^{-1}$

[Javanmard, Montanari 2012]

Does this remind you anything?

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

(pseudo-) Newton method

Debiased estimator: Choosing M?

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

- Gaussian design ($x_i \sim \mathrm{N}(0, \Sigma)$)
  - Assume known Σ (relevant in semi-supervised learning)
  - $M = \Sigma^{-1}$

[Javanmard, Montanari 2012]

- 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 2014]

Debiased estimator: Choosing M?

Our approach:

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

[A. Javanmard, A. Montanari 2014]

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

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

Debiased estimator: Choosing M?

Our approach:

- Find M by solving an optimization problem:

[A. Javanmard, A. Montanari]

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

Debiased estimator: Choosing M?

Our approach:

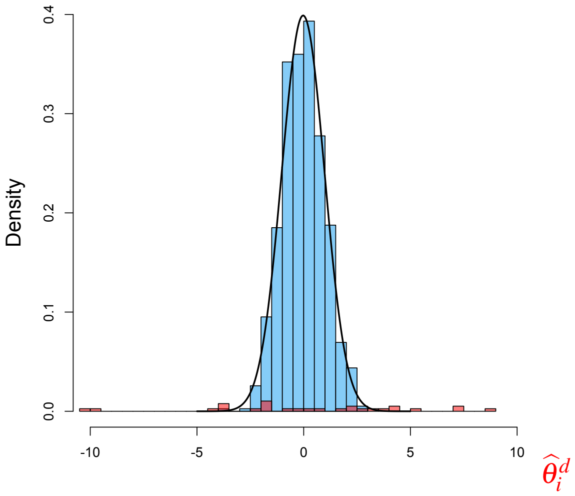
- Find M by solving an optimization problem:

[A. Javanmard, A. Montanari]

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

The optimization can be decoupled and solved in parallel.

What does it look like?



Can estimate σ

'Ground truth' from $n_{\text{tot}} = 10,000$ records.

Confidence intervals

Neglecting the bias ($\widehat{\sigma}$ estimator of σ)

$$\widehat{\theta}_i^d \approx \mathrm{N}(\theta_{0,i}, \Delta_i^2)\,, \qquad \Delta_i^2 \equiv \frac{\widehat{\sigma}^2}{n}(M\widehat{\Sigma}M^{\mathsf{T}})_{ii}$$

$$[\underline{\theta}_i, \overline{\theta}_i] = [\widehat{\theta}_i^d - c_\alpha \Delta_i, \widehat{\theta}_i^d + c_\alpha \Delta_i]$$

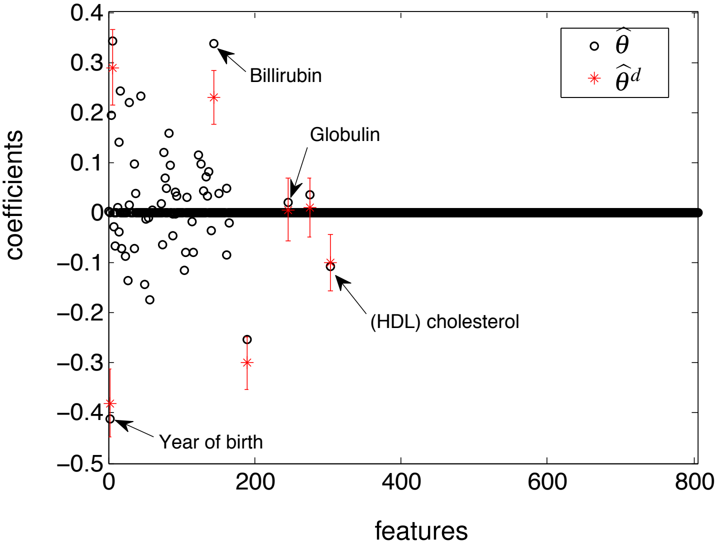
Confidence intervals

Neglecting the bias ($\widehat{\sigma}$ estimator of σ)

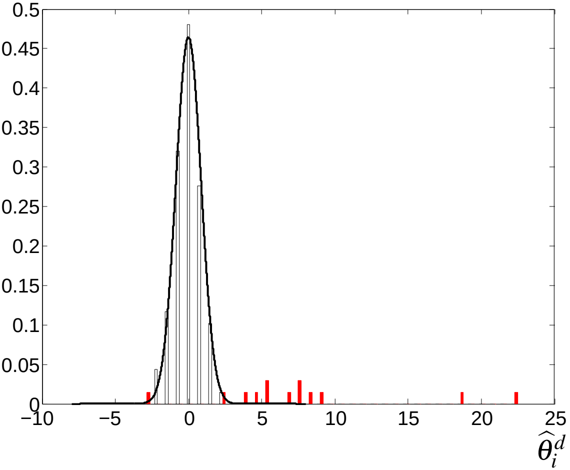
$$\widehat{\theta}_i^d \approx \mathrm{N}(\theta_{0,i}, \Delta_i^2)\,, \qquad \Delta_i^2 \equiv \frac{\widehat{\sigma}^2}{n}(M\widehat{\Sigma}M^{\mathsf{T}})_{ii}$$

$$[\underline{\theta}_i, \overline{\theta}_i] = [\widehat{\theta}_i^d - c_\alpha \Delta_i, \widehat{\theta}_i^d + c_\alpha \Delta_i]$$

What does it look like?



UCI crime dataset



$n = 84$, $p = 102$, $n_{\text{tot}} = 1994$.

A theorem

Theorem [Javanmard, Montanari 2013] (Deterministic designs)

Let $\mathbf{X}$ be any deterministic design that satisfies compatibility condition. Define the coherence parameter

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

Let $s_0 = |\mathrm{supp}(\theta_0)|$. Then

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

$$\|\Delta\|_\infty \quad \le \quad c\mu_* \sigma s_0 \sqrt{\log p}\,, \quad \text{w.h.p.}$$

A theorem

Theorem [Javanmard, Montanari 2013] (Deterministic designs)

Let $\mathbf{X}$ be any deterministic design that satisfies compatibility condition. Define the coherence parameter

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

Let $s_0 = |\mathrm{supp}(\theta_0)|$. Then

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

$$\|\Delta\|_\infty \;\le\; c\mu_*\sigma s_0\sqrt{\log p}, \quad \text{w.h.p.}$$

Remark:

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

A theorem

Theorem [Javanmard, Montanari 2013] (Random designs)

Consider population covariance Σ with bounded eigenvalues and assume assume $\mathbf{X}\Sigma^{-1}$ has independent subgaussian rows. Then

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

$$\|\Delta\|_\infty \le c\sigma \frac{s_0 \log p}{\sqrt{n}}, \quad \text{w.h.p.}$$

A theorem

Theorem [Javanmard, Montanari 2013] (Random designs)

Consider population covariance Σ with bounded eigenvalues and assume assume $\mathbf{X}\Sigma^{-1}$ has independent subgaussian rows. Then

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

$$\|\Delta\|_\infty \leq c\sigma \frac{s_0 \log p}{\sqrt{n}}, \quad \text{w.h.p.}$$

Remark on sample size:

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

Consequences

- Confidence intervals for single parameters:

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

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

(n<p)

Consequences

- Confidence intervals for single parameters:

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

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

(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:

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

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

(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)

Hypothesis testing (based on de-biased estimator)

- Null/alternative hypothesis:

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

- 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.

- We provide precise characterization of type I and type II error.

- Test (using de-biased estimator) has minimax optimal statistical power.

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.

Further related work

- Debiasing:
  - Group sparsity [R. Mitra & C.H.Zhang 2014]
  - Confidence interval for inverse covariance estimation [J. Jankova, S.v.d. Geer 2015]
  - Genomics [Q.Zhao et. al. 2015, B. Rakitsch 2015]
  - Econometrics [A. Belloni & V. Chernozhukov 2014, D. Kozbur 2015]
- Other methods for uncertainty assessment
  - Uncertainty quantification under group sparsity [Q.Zhou 2015]
  - Post double selection [Belloni et. al. 2014]

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, $n \gtrsim s_0 \log p$?

Where is the bottleneck?

Where is the bottleneck?

The bias is given by

$$\Delta = \sqrt{n}(\Omega\widehat{\Sigma} - \mathbf{I})(\theta_0 - \widehat{\theta}^{\mathrm{Lasso}}) \,.$$

Earlier work bound bias a simple $\ell_1 - \ell_\infty$ inequality:

$$\begin{aligned}
\|\Delta\|_\infty &\leq \sqrt{n}|M\widehat{\Sigma} - \mathbf{I}|_\infty \|\theta^* - \widehat{\theta}^{\mathrm{Lasso}}\|_1 \\
&\leq \sqrt{n} \times C\sqrt{\frac{\log p}{n}} \times Cs_0\sigma\sqrt{\frac{\log p}{n}} \\
&\leq C^2\sigma\frac{s_0 \log p}{\sqrt{n}} \,.
\end{aligned}$$

Plan for this part

- Focus on Gaussian design: $x_i \sim \mathrm{N}(0, \Sigma)$
- Assume that Σ is known. (See paper for unknown covariance.)
- We show that the required sample rate is indeed artifact of the argument!

Plan for this part

- Focus on Gaussian design: $x_i \sim \mathrm{N}(0, \Sigma)$
- Assume that Σ is known. (See paper for unknown covariance.)
- We show that the required sample rate is indeed artifact of the argument!

De-biased estimator is asymptotically Gaussian under condition $n \gtrsim s_0 (\log p)^2$.

'Leave-one-out' technique

Fix coordinate i.

- Define

$$\begin{aligned} \widehat{\theta}^{\rm p} \quad &\equiv \quad \arg\min_{\theta} \quad \frac{1}{2n}\|y - X\theta\|^2 + \lambda\|\theta\|_1 \\ &\quad\quad \text{subject to} \quad \widehat{\theta}_i^{\rm p} = \theta_{0,i} \end{aligned}$$

'Leave-one-out' technique

Fix coordinate i.

- Define

$$\begin{aligned} \widehat{\theta}^{\rm p} \quad &\equiv \quad \arg\min_{\theta} \quad \frac{1}{2n}\|y - X\theta\|^2 + \lambda\|\theta\|_1 \\ &\quad \text{subject to} \quad \widehat{\theta}^{\rm p}_i = \theta_{0,i} \end{aligned}$$

  We then have

$$y - X\widehat{\theta}^{\rm p} = w + \cancel{\tilde{x}_i(\theta_{0,i} - \widehat{\theta}^{\rm p}_i)} + X_{\sim i}(\theta_{0,\sim i} - \widehat{\theta}^{\rm p}_{\sim i})$$

'Leave-one-out' technique

Fix coordinate i.

- Define

$$\begin{aligned}\widehat{\theta}^{\mathrm{p}} \quad &\equiv \quad \arg\min_{\theta} \quad \frac{1}{2n}\|y - X\theta\|^2 + \lambda\|\theta\|_1 \\ &\quad\quad \text{subject to} \quad \widehat{\theta}^{\mathrm{p}}_i = \theta_{0,i}\end{aligned}$$

We then have

$$y - X\widehat{\theta}^{\mathrm{p}} = w + \cancel{\tilde{x}_i(\theta_{0,i} - \widehat{\theta}^{\mathrm{p}}_i)} + X_{\sim i}(\theta_{0,\sim i} - \widehat{\theta}^{\mathrm{p}}_{\sim i})$$

$\widehat{\theta}^{\mathrm{p}}$ is the Lasso estimator when $\tilde{x}_i$ is left out!

'Leave-one-out' technique

Let v be the ith column of $X\Sigma^{-1}$.
The bias is given by

$$\Delta_i = R_1 + R_2 + R_3$$

‘Leave-one-out’ technique

Let v be the ith column of $X\Sigma^{-1}$.
The bias is given by

$$\Delta_i = R_1 + R_2 + R_3$$

$$\begin{aligned}
R_1 &= \sqrt{n}\Big(1 - \frac{\langle v, \widetilde{x}_i\rangle}{n}\Big)(\widehat{\theta}_i^{\rm Lasso} - \theta_i^*)\\
R_2 &= \frac{v^{\sf T}}{\sqrt{n}} X_{\sim i}(\theta_{0,\sim i} - \widehat{\theta}^{\rm p}_{\sim i})\\
R_3 &= \frac{v^{\sf T}}{\sqrt{n}} X_{\sim i}(\widehat{\theta}^{\rm p}_{\sim i} - \widehat{\theta}^{\rm Lasso}_{\sim i})
\end{aligned}$$

‘Leave-one-out’ technique

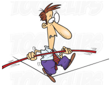
Let v be the ith column of $X\Sigma^{-1}$.
The bias is given by

$$\Delta_i = R_1 + R_2 + R_3$$

$$R_1 = \sqrt{n}\left(1 - \frac{\langle v, \widetilde{x}_i\rangle}{n}\right)(\widehat{\theta}_i^{\text{Lasso}} - \theta_i^*) \xrightarrow{\text{concentration}}$$

$$R_2 = \frac{v^{\mathsf{T}}}{\sqrt{n}} X_{\sim i}(\theta_{0,\sim i} - \widehat{\theta}^{\text{p}}_{\sim i}) \xrightarrow{\text{independence}}$$

$$R_3 = \frac{v^{\mathsf{T}}}{\sqrt{n}} X_{\sim i}(\widehat{\theta}^{\text{p}}_{\sim i} - \widehat{\theta}^{\text{Lasso}}_{\sim i}) \xrightarrow{\text{perturbation}}$$



Summary

Combining the bounds on R_1, R_2, R_3, we obtain

$$\|\Delta\|_\infty \le C\sqrt{\frac{s_0}{n}\log p}\,, \quad \text{w.h.p}$$

Summary

Combining the bounds on R_1, R_2, R_3, we obtain

$$\|\Delta\|_\infty \le C\sqrt{\frac{s_0}{n}}\log p\,, \quad \text{w.h.p}$$

Therefore,

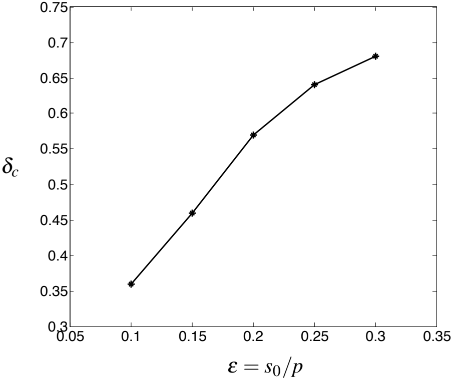
$$\|\Delta\|_\infty \to 0 \quad \text{provided that} \quad n \ge s_0(\log p)^2\,.$$

Numerical illustration

- Fix $p = 3000$
- Design matrix X with rows i.i.d. from $\mathrm{N}(0, \Sigma)$
- $\Sigma_{ij} = 0.8^{|i-j|}$
- Define $\delta = n/p$ (undersampling rate) and $\varepsilon = s_0/p$ (sparsity proportion)
- δ_c: Critical value above which the de-biased estimator is Gaussian.

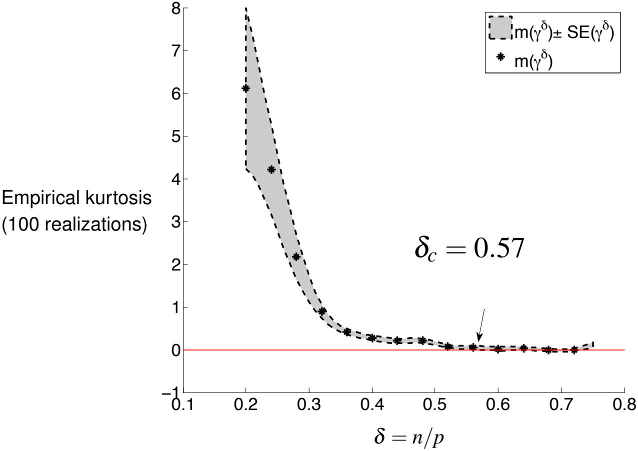
Numerical illustration

- Fix $p = 3000$
- Design matrix X with rows i.i.d. from $\mathrm{N}(0, \Sigma)$
- $\Sigma_{ij} = 0.8^{|i-j|}$
- Define $\delta = n/p$ (undersampling rate) and $\varepsilon = s_0/p$ (sparsity proportion)
- δ_c: Critical value above which the de-biased estimator is Gaussian.



How to define δ_c?

- Fix ε and change $\delta = n/p$.



$(\varepsilon = 0.2)$

Conclusion

- De-biasing regularized estimators
- Compute confidence intervals/p-values for high dimensional models
- Optimal sample size for Gaussian designs

Thanks!