Accuracy Assessment for High-dimensional Linear Regression

Zijian Guo

Department of Statistics
University of Pennsylvania
Philadelphia, PA 19104

November 3, 2016

Joint work with Professor T. Tony Cai.

High-dimensional linear regression

The linear regression model

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad n \ll p,$$

where $\|\beta\|_0 \leq k$.

Motivating applications: Genomics study; Compressed sensing.

High-dimensional linear regression

The linear regression model

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad n \ll p,$$

where $\|\beta\|_0 \leq k$.

Motivating applications: Genomics study; Compressed sensing.

Methods: Basis Pursuit (Chen & Donoho, 1994), Lasso (Tibshirani, 1996), SCAD (Fan & Li, 2001), Dantzig Selector (Candès & Tao, 2007), square-root Lasso (Belloni, et. al., 2011) and scaled Lasso (Sun & Zhang, 2010).

Accuracy Assessment

Not enough to just provide a good estimator.

Accuracy Assessment

Not enough to just provide a good estimator.
Need to know the accuracy of the estimator.

Accuracy Assessment

Not enough to just provide a good estimator.
Need to know the accuracy of the estimator.

Accuracy assessment

- Margin of error $\rightarrow$ inference for binomial proportion.
- Width of confidence interval $\rightarrow$ inference for one-dimensional parameter.
- Stein's Unbiased Risk Estimate $\rightarrow$ empirical selection of tuning parameter.

Accuracy Assessment

Not enough to just provide a good estimator.
Need to know the accuracy of the estimator.

Accuracy assessment

- Margin of error $\rightarrow$ inference for binomial proportion.
- Width of confidence interval $\rightarrow$ inference for one-dimensional parameter.
- Stein's Unbiased Risk Estimate $\rightarrow$ empirical selection of tuning parameter.
- A doctor needs to know the accuracy of reconstructed image based on MRI. (Janson et. al., 2015)
- Choose the best estimator among the proposed estimators.

Research Problem

How to assess the accuracy of these proposed estimators?

Research Problem

How to assess the accuracy of these proposed estimators?

1. Confidence intervals for the accuracy $\|\widehat{\beta} - \beta\|_2^2$.

Research Problem

How to assess the accuracy of these proposed estimators?

1. Confidence intervals for the accuracy $\|\widehat{\beta} - \beta\|_2^2$.
2. Is it possible to construct confidence intervals for $\|\widehat{\beta} - \beta\|_2^2$
    - Minimax rate-optimal
    - Adaptive to the sparsity.

Adaptive and rate-optimal estimators

Lasso, Dantzig Selector and scaled Lasso satisfy, for β being sparse,

$$\mathbb{P}\left(\|\widehat{\beta}-\beta\|_2^2 \leq C\frac{\|\beta\|_0 \log p}{n}\right) \geq 1-o(1). \quad (1)$$

See Candès and Tao (2007); Bickel, Ritov and Tsybakov(2009); Sun and Zhang (2010).

Adaptive and rate-optimal estimators

Lasso, Dantzig Selector and scaled Lasso satisfy, for β being sparse,

$$\mathbb{P}\left(\|\widehat{\beta}-\beta\|_2^2 \leq C\frac{\|\beta\|_0 \log p}{n}\right) \geq 1 - o(1). \quad (1)$$

See Candès and Tao (2007); Bickel, Ritov and Tsybakov(2009); Sun and Zhang (2010).

Adaptive to sparsity!

Adaptive and rate-optimal estimators

Lasso, Dantzig Selector and scaled Lasso satisfy, for β being sparse,

$$\mathbb{P}\left(\|\widehat{\beta}-\beta\|_2^2 \leq C\frac{\|\beta\|_0 \log p}{n}\right) \geq 1-o(1). \tag{1}$$

See Candès and Tao (2007); Bickel, Ritov and Tsybakov(2009); Sun and Zhang (2010).

Adaptive to sparsity!

Focus on adaptive and rate-optimal estimators satisfying (1).

Adaptive and rate-optimal estimators

Lasso, Dantzig Selector and scaled Lasso satisfy, for β being sparse,

$$\mathbb{P}\left(\|\widehat{\beta}-\beta\|_2^2 \leq C\frac{\|\beta\|_0 \log p}{n}\right) \geq 1-o(1). \tag{1}$$

See Candès and Tao (2007); Bickel, Ritov and Tsybakov(2009); Sun and Zhang (2010).

Adaptive to sparsity!

Focus on adaptive and rate-optimal estimators satisfying (1).

Let $\widehat{\beta}^{L}$ and $\widehat{\beta}^{SL}$ denote the Lasso or scaled Lasso estimator with a proper chosen tuning parameter.

Two parameter spaces

Recall the high-dimensional linear model with random design,

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N_n(0, \sigma^2 \mathrm{I}).$$

where $X_{i\cdot} \overset{iid}{\sim} N(0, \Sigma)$ and $X_{i\cdot}$ and ϵ are independent.

Two parameter spaces

Recall the high-dimensional linear model with random design,

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N_n(0, \sigma^2 \mathrm{I}).$$

where $X_{i\cdot} \overset{iid}{\sim} N(0, \Sigma)$ and $X_{i\cdot}$ and ϵ are independent.

Two parameter spaces for (β, Σ, σ)

1. Known $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$

$$\Theta_0(k) = \{(\beta, \mathrm{I}, \sigma_0) : \|\beta\|_0 \le k\}.$$

2. Unknown Σ and σ

$$\Theta(k) = \left\{(\beta, \Sigma, \sigma) : \|\beta\|_0 \le k,\ \frac{1}{M_1} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le M_1,\ 0 < \sigma \le M_2\right\}.$$

Framework for minimaxity and adaptivity

Two levels of sparsity $k_1 \leq k_2$

- $\|\beta\|_0 = k_1$ – precise knowledge of sparsity.
- $\|\beta\|_0 \leq k_2$ – rough knowledge of sparsity.

Framework for minimaxity and adaptivity

Two levels of sparsity $k_1 \leq k_2$

- $\|\beta\|_0 = k_1$ – precise knowledge of sparsity.
- $\|\beta\|_0 \leq k_2$ – rough knowledge of sparsity.

Adaptive estimation of β

- Implementation does not require prior knowledge of k_1.
- The convergence rate is $k_1 \log p/n$.

Framework for minimaxity and adaptivity

Two levels of sparsity $k_1 \leq k_2$

- $\|\beta\|_0 = k_1$ – precise knowledge of sparsity.
- $\|\beta\|_0 \leq k_2$ – rough knowledge of sparsity.

Adaptive estimation of β

- Implementation does not require prior knowledge of k_1.
- The convergence rate is $k_1 \log p/n$.

Two aspects of confidence intervals

- Coverage: Guaranteed coverage probability.
- Precision: As short as possible.

Framework for minimaxity and adaptivity

Confidence intervals for $\|\widehat{\beta} - \beta\|_2^2$

What if we only know k_2?

- Coverage: Guaranteed coverage probability over $\Theta(k_2)$.
- Precision: Evaluate the length over $\Theta(k_1) \subset \Theta(k_2)$.

Framework for minimaxity and adaptivity

Confidence intervals for $\|\widehat{\beta} - \beta\|_2^2$

What if we only know k_2?

- Coverage: Guaranteed coverage probability over $\Theta(k_2)$.
- Precision: Evaluate the length over $\Theta(k_1) \subset \Theta(k_2)$.

Define benchmark for adaptivity between $\Theta(k_1) \subset \Theta(k_2)$ as

$$\mathbf{L}_{\alpha}^{*}\left(\Theta(k_1), \Theta(k_2), \widehat{\beta}\right) = \inf_{\substack{\text{CI has guaranteed} \\ \text{coverage over } \Theta(k_2)}} \sup_{\theta \in \Theta(k_1)} \mathbf{E}_{\theta} \mathbf{L}\left(\mathrm{CI}\right).$$

Framework for minimaxity and adaptivity

Confidence intervals for $\|\widehat{\beta}-\beta\|_2^2$

What if we only know k_2?

- Coverage: Guaranteed coverage probability over $\Theta(k_2)$.
- Precision: Evaluate the length over $\Theta(k_1) \subset \Theta(k_2)$.

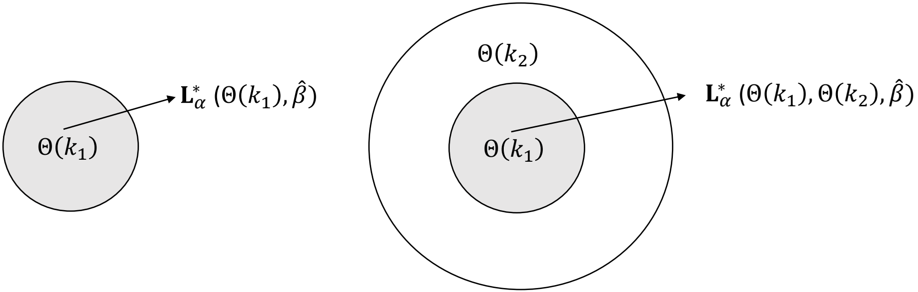
Define benchmark for adaptivity between $\Theta(k_1) \subset \Theta(k_2)$ as

$$\mathbf{L}_{\alpha}^{*}\left(\Theta(k_1), \Theta(k_2), \widehat{\beta}\right) = \inf_{\substack{\text{CI has guaranteed} \\ \text{coverage over } \Theta(k_2)}} \sup_{\theta \in \Theta(k_1)} \mathbf{E}_{\theta} \mathbf{L}\left(\mathrm{CI}\right).$$

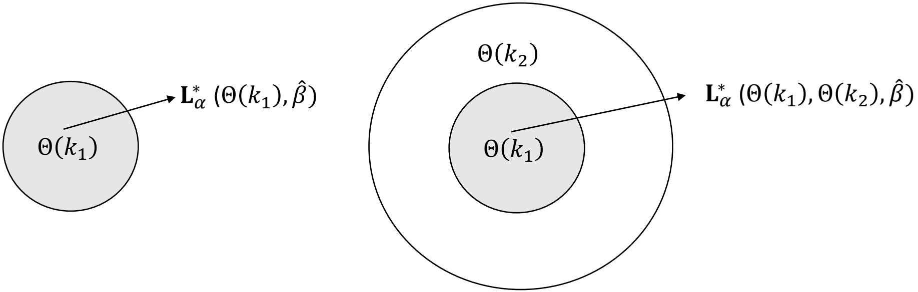
Define benchmark for minimaxity as

$$\mathbf{L}_{\alpha}^{*}\left(\Theta(k_1), \widehat{\beta}\right) = \inf_{\substack{\text{CI has guaranteed} \\ \text{coverage over } \Theta(k_1)}} \sup_{\theta \in \Theta(k_1)} \mathbf{E}_{\theta} \mathbf{L}\left(\mathrm{CI}\right).$$

Framework for minimaxity and adaptivity



Framework for minimaxity and adaptivity



Impossibility of adaptivity

$$\mathbf{L}_{\alpha}^{*}\left(\Theta(k_1), \Theta(k_2), \widehat{\beta}\right) \gg \mathbf{L}_{\alpha}^{*}\left(\Theta(k_1), \widehat{\beta}\right). \tag{2}$$

Confidence intervals for $\|\widehat{\beta} - \beta\|_2^2$ over $\Theta_0(k)$

Theorem

For any adaptive and rate-optimal estimator $\widehat{\beta}$, then there is some constant $c > 0$ such that

$$\mathbf{L}_{\alpha}^{*}\left(\Theta_0(k_1), \widehat{\beta}\right) \geq c \min\left\{\frac{k_1 \log p}{n}, \frac{1}{\sqrt{n}}\right\} \sigma_0^2. \tag{3}$$

$$\mathbf{L}_{\alpha}^{*}\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}\right) \geq c \min\left\{\frac{k_2 \log p}{n}, \frac{1}{\sqrt{n}}\right\} \sigma_0^2. \tag{4}$$

Confidence intervals for $\|\widehat{\beta} - \beta\|_2^2$ over $\Theta_0(k)$

Theorem

For any adaptive and rate-optimal estimator $\widehat{\beta}$, then there is some constant $c > 0$ such that

$$\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \widehat{\beta}\right) \geq c \min\left\{\frac{k_1 \log p}{n}, \frac{1}{\sqrt{n}}\right\} \sigma_0^2. \tag{3}$$

$$\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}\right) \geq c \min\left\{\frac{k_2 \log p}{n}, \frac{1}{\sqrt{n}}\right\} \sigma_0^2. \tag{4}$$

The lower bounds can be achieved for confidence intervals for $\|\widehat{\beta}^L - \beta\|_2^2$.

Case 1: $k_1 \le k_2 \lesssim \frac{\sqrt{n}}{\log p}$

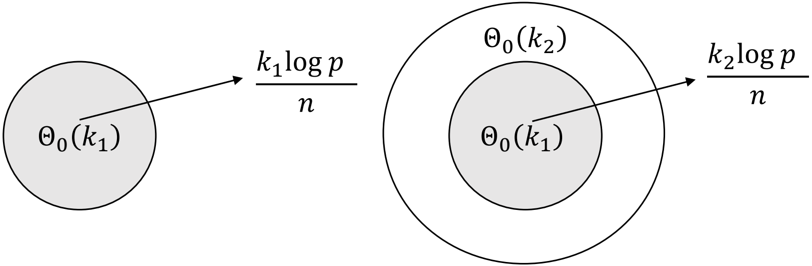


Figure: $\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \widehat{\beta}^L\right)$ v.s. $\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^L\right)$

Impossible to construct adaptive CI for $\|\widehat{\beta}^L - \beta\|_2^2$.

Case 2: $k_1 \lesssim \frac{\sqrt{n}}{\log p} \ll k_2 \lesssim \frac{n}{\log p}$

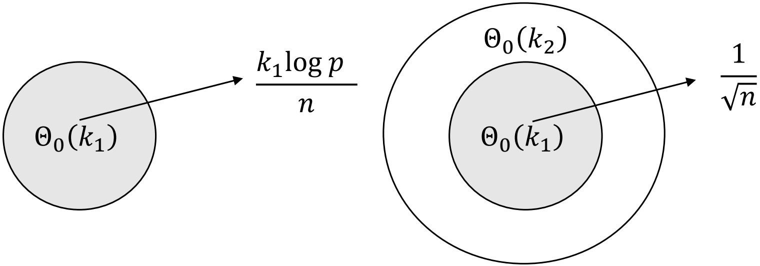


Figure: $\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \widehat{\beta}^L\right)$ v.s. $\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^L\right)$

Impossible to construct adaptive CI for $\|\widehat{\beta}^L - \beta\|_2^2$.

Case 3: $\frac{\sqrt{n}}{\log p} \ll k_1 \leq k_2 \lesssim \frac{n}{\log p}$

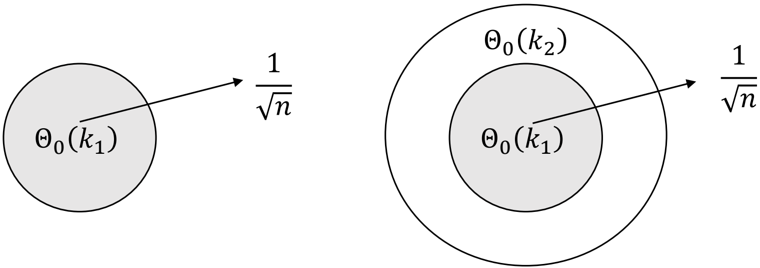


Figure: $\mathbf{L}_{\alpha}^{*}\left(\Theta_0(k_1), \widehat{\beta}^L\right)$ v.s. $\mathbf{L}_{\alpha}^{*}\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^L\right)$

Possible to construct adaptive CI for $\|\widehat{\beta}^L - \beta\|_2^2$.

Confidence intervals for $\|\widehat{\beta}^L - \beta\|_2^2$ over $\Theta_0(k)$

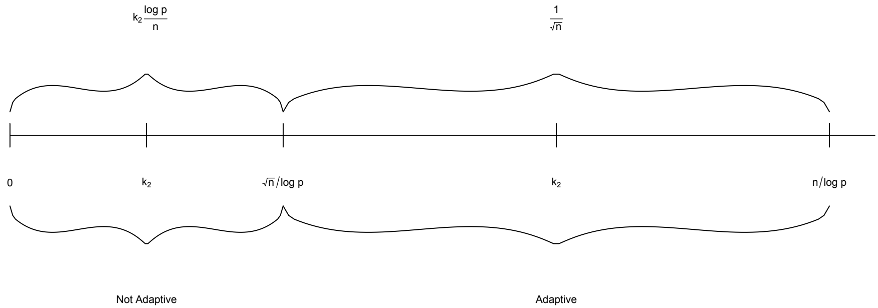


Figure: Summary of $\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^L\right)$

Confidence intervals for $\|\widehat{\beta} - \beta\|_2^2$ over $\Theta(k)$

Theorem

For any adaptive and rate-optimal estimator $\widehat{\beta}$, then there is some constant $c > 0$ such that

$$\mathbf{L}_{\alpha}^{*}\left(\Theta\left(k_1\right), \widehat{\beta}\right) \geq c k_1 \frac{\log p}{n}; \tag{5}$$

and

$$\mathbf{L}_{\alpha}^{*}\left(\Theta\left(k_1\right), \Theta\left(k_2\right), \widehat{\beta}\right) \geq c k_2 \frac{\log p}{n}. \tag{6}$$

Confidence intervals for $\|\widehat{\beta} - \beta\|_2^2$ over $\Theta(k)$

Theorem

For any adaptive and rate-optimal estimator $\widehat{\beta}$, then there is some constant $c > 0$ such that

$$\mathbf{L}_{\alpha}^{*}\left(\Theta\left(k_1\right), \widehat{\beta}\right) \geq c k_1 \frac{\log p}{n}; \tag{5}$$

and

$$\mathbf{L}_{\alpha}^{*}\left(\Theta\left(k_1\right), \Theta\left(k_2\right), \widehat{\beta}\right) \geq c k_2 \frac{\log p}{n}. \tag{6}$$

The lower bounds can be achieved for confidence intervals for $\|\widehat{\beta}^{SL} - \beta\|_2^2$.

Confidence intervals for $\|\widehat{\beta}^{SL}-\beta\|_2^2$ over $\Theta(k)$

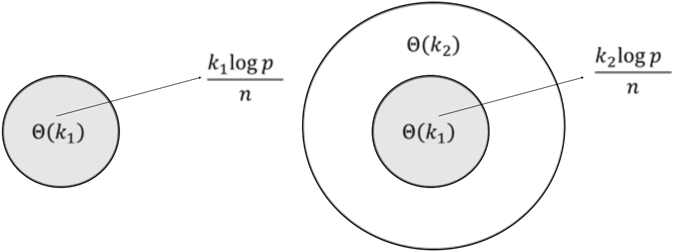


Figure: $\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \widehat{\beta}^{SL}\right)$ v.s. $\mathbf{L}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^{SL}\right)$

Impossible to construct adaptive CI for $\|\widehat{\beta}^{SL}-\beta\|_2^2$.

Confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ with $1 \leq q < 2$

1. There is fundamental difference between $q = 2$ and $1 \leq q < 2$.

Confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ with $1 \leq q < 2$

1. There is fundamental difference between $q = 2$ and $1 \leq q < 2$.
2. No adaptive regime for both $\Theta_0(k)$ and $\Theta(k)$.

Conclusion and Discussion

1. For any adaptive rate-optimal estimator, accuracy assessment is hard in high dimension linear regression.
2. Adaptive confidence interval for the accuracy $\|\widehat{\beta} - \beta\|_2^2$ is only possible
   - With the prior information $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$;
   - Over the regime $\frac{\sqrt{n}}{\log p} \leq k \leq \frac{n}{\log p}$.

Conclusion and Discussion

1. For any adaptive rate-optimal estimator, accuracy assessment is hard in high dimension linear regression.
2. Adaptive confidence interval for the accuracy $\|\widehat{\beta}-\beta\|_2^2$ is only possible
   - With the prior information $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$;
   - Over the regime $\frac{\sqrt{n}}{\log p} \leq k \leq \frac{n}{\log p}$.
3. Significant differences between
   - $\|\widehat{\beta}-\beta\|_2^2$and the $\|\widehat{\beta}-\beta\|_q^2$ loss with $1 \leq q < 2$;
   - the two parameter spaces $\Theta(k)$ and $\Theta_0(k)$.
4. In the paper, we have developed a general tool for establishing minimax lower bounds for accuracy assessment.

Conclusion and Discussion

1. For any adaptive rate-optimal estimator, accuracy assessment is hard in high dimension linear regression.
2. Adaptive confidence interval for the accuracy $\|\widehat{\beta} - \beta\|_2^2$ is only possible
   - With the prior information $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$;
   - Over the regime $\frac{\sqrt{n}}{\log p} \leq k \leq \frac{n}{\log p}$.
3. Significant differences between
   - $\|\widehat{\beta} - \beta\|_2^2$and the $\|\widehat{\beta} - \beta\|_q^2$ loss with $1 \leq q < 2$;
   - the two parameter spaces $\Theta(k)$ and $\Theta_0(k)$.
4. In the paper, we have developed a general tool for establishing minimax lower bounds for accuracy assessment.
5. It is interesting to investigate the estimation of loss for more general estimators that are not adaptive and rate-optimal estimators.