Simultaneous Confidence Intervals Using Entire Solution Paths

Xiaorui Zhu

Operations, Business Analytics & Information Systems Department
Lindner College of Business
University of Cincinnati

Co-Authors:
Yichen Qin, University of Cincinnati
Peng Wang, University of Cincinnati

July 28, 2019

Outline

- Motivation for the study
- Existing Methods and Preliminaries
- General approach of constructing simultaneous confidence intervals
- Simulation studies
- Real Examples

1. The high-dimensional problems are prevalent

- Document classification: bag-of-words(similarity) can result in $p = 20K$
- Genomics: say $p = 20K$ genes for each subject

2. Two objectives in the high-dimensional sparse linear models:

- Sparse estimation
- **Statistical inference** (our focus)

We focus on linear model as follow:

$$\mathbf{y} = \mathbf{X}\beta^* + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n), \tag{1}$$

- $\mathbf{y}$ is the response vector
- $\mathbf{X}_{n\times p} \in \mathbb{R}^p$ is the fixed design matrix containing p dimensional covariates.
- The parameter vector $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_p^*)' \in \mathbb{R}^p$ is assumed to be sparse.
- $S = \{j : \beta_j^* \neq 0, j = 1, \cdots, p\} \subset \{j : j = 1, \cdots, p\}$, we assume that $|S| = s < p$. The set of the truly zero coefficients is $S^c = \{j : \beta_j^* = 0\}$.

Motivation: Ideal simultaneous confidence intervals

An ideal simultaneous confidence intervals should:

1. Provide *simultaneous confidence intervals* with the nominal confidence level (can be shown by the coverage probability);

2. Have *tight intervals for all coefficients* at a given level of confidence (can be shown by the width of nonzero and zero coefficients);

3. Be able to reveal the *variable selection results* in a way that the truly irrelevant coefficients have zero width intervals.

Motivation: Drawbacks of Existing Methods

The ideal simultaneous confidence intervals **require** the variable selection method to have:

- Unbiasedness of estimation (But, Lasso estimator is biased)
- High selection accuracy (But, the selection accuracy of Lasso and Adaptive Lasso is highly unstable due to a single tuning parameter)

Missing of selection information

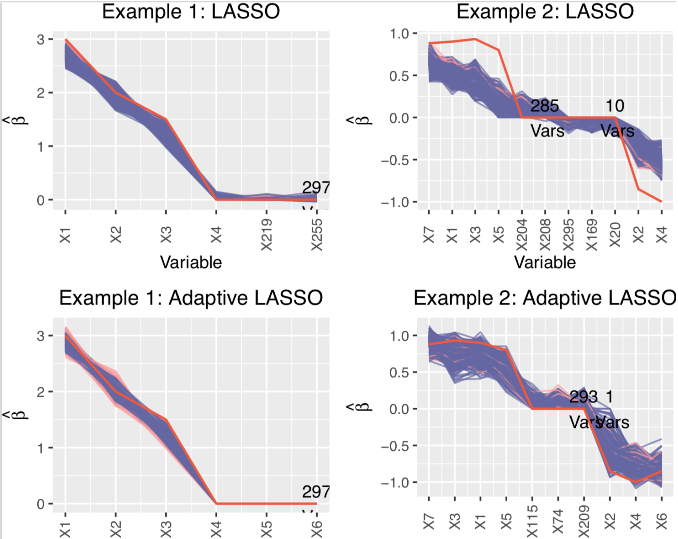
- Main stream: "Debiased" estimator hide the variable selection information (S. van de Geer et al. (2014), Javanmard and Montanari (2014), Dezeure, Bühlmann, and Zhang (2017), X. Zhang and Cheng (2017))

- *Example 1* (Moderate Correlation, $p > n$, Tibshirani (1996)). $\beta_i^* = (3, 2, 1.5), i = 1, 2, 3,\ \beta_i^* = 0, i = 4, \ldots, 300$, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The correlation between x_{j_1} and x_{j_2} is $0.5^{|j_1 - j_2|}$.

- *Example 2*: ($p > n$, positive and negative coefficients). Assume $\boldsymbol{\beta}^* = (0.9, -0.85, 0.93, -1, 0.8, -0.85, 0.88)$, and the remaining coefficients equal zero. The correlation between x_{j_1} and x_{j_2} is $0.5^{|j_1 - j_2|}$.

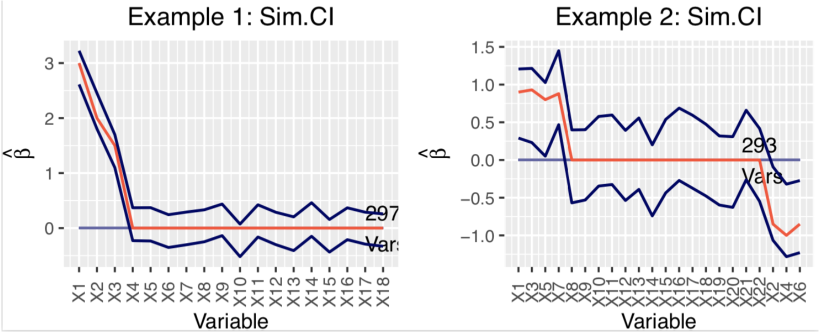
- For both examples, $n = 200$, $p = 300$, and $\sigma = 1$.

1. *Biased estimators*
2. *Poor selection accuracy*

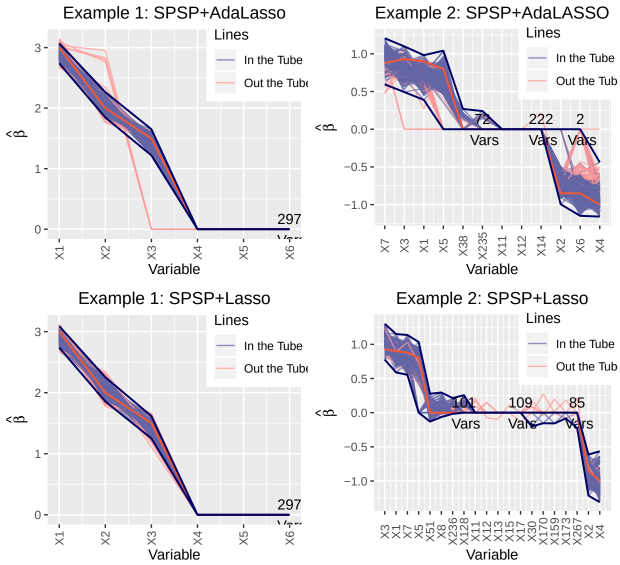


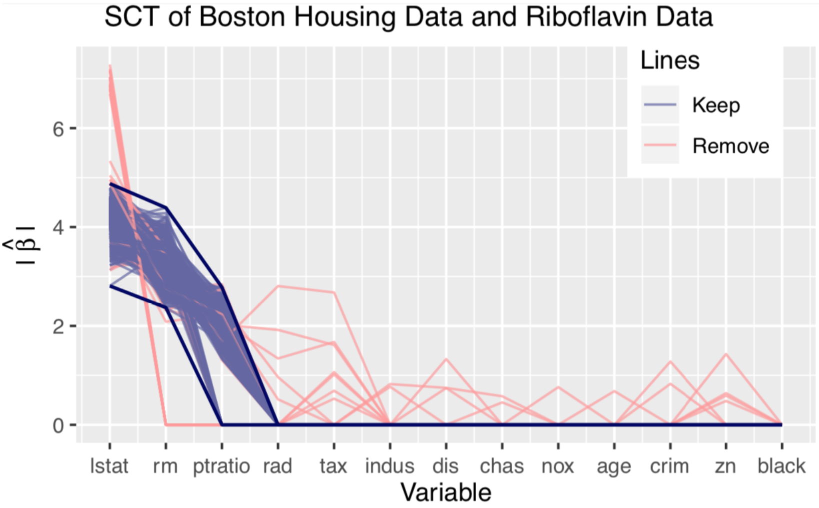
3. *Missing of selection information*

The simultaneous confidence intervals method by X. Zhang and Cheng (2017) (named as “Sim.CI”):



How about this type of SCI




SCT of Boston Housing Data and Riboflavin Data
Lines
Keep
Remove
6
4
2
0
|β̂|
lstat
rm
ptratio
rad
tax
indus
dis
chas
nox
age
crim
zn
black
Variable

Preliminaries

Lasso (Tibshirani (1996)):

$$\hat{\beta}^{\text{Lasso}} = \underset{\beta}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.1}$$

Adaptive Lasso (Zou (2006)):

$$\hat{\beta}^{\text{AdaLasso}} = \underset{\beta}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|, \tag{2.2}$$

Selection by Partitioning the Solution Paths (SPSP)

Idea: Using the whole solution paths of all coefficients and applying the clustering approach.

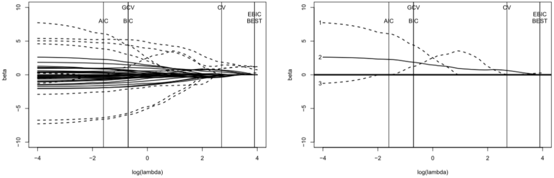


FIG 1. *Left: The lasso solution paths for the simulated example. The dashed lines are the paths of the* 10 *non-zero coefficients, while the black lines are the paths of the* 30 *zero coefficients The vertical lines represent the tuning parameters selected by different criteria. Right: The lasso solution paths for the non-zero coefficients, 1 and 3, and the zero coefficient, 2. Here CV is cross-validation, GCV is generalized cross-validation and EBIC is extended BIC.*

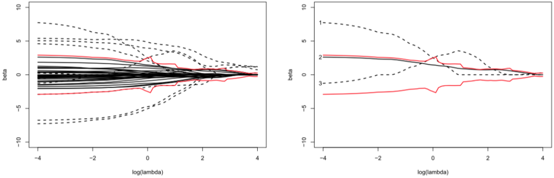


FIG 2. *Left: Partitions of the lasso solution paths of the same simulated example. Right: Partitions of the lasso solution paths for the non-zero coefficients, 1 and 3, and the zero coefficient, 2.*

Assumption 2.1: Compatibility Condition (Bühlmann and Geer (2011); S. van de Geer (2007)). For some constant $\phi > 0$ and for any vector ζ satisfying $\|\zeta\|_1 \leq 3\|\zeta_S\|_1$, the following compatibility condition holds:

$$\|\zeta_S\|_1^2 \leq (\zeta^T \hat{\Sigma} \zeta) s / \phi^2,$$

where $s = |S|$ is the dimension of β_S.

Assumption 2.2: Weak Identifiability Condition Let $\eta > 0$ be some constant. For any $\bar{\beta} = (\bar{\beta}_S, \bar{\beta}_{S^C})$, then for $k = \frac{2}{2s+Rs(s+1)}$ and some κ that satisfies

$$D_{\max} > \lambda_0 \frac{4s(1+R)}{\phi^2}\left\{\frac{Rs^2 + (2+R)S + 2}{\eta} - 1 + \kappa\right\},$$

then the **WIC**,

$$\|\mathbf{X}\beta^* - \mathbf{X}_S\bar{\boldsymbol{\beta}}_S - \mathbf{X}_{S^C}\bar{\boldsymbol{\beta}}_{S^C}\|^2 \geq \min_{\beta \in \Theta(\|\bar{\boldsymbol{\beta}}_S\|_1, \|\bar{\boldsymbol{\beta}}_{S^C}\|_1)} \|\mathbf{X}\beta^* - \mathbf{X}\beta\|^2 - \kappa\eta\|\bar{\boldsymbol{\beta}}_{S^C}\|_1,$$

holds. The $\Theta(\|\bar{\boldsymbol{\beta}}_S\|_1, \|\bar{\boldsymbol{\beta}}_{S^C}\|_1) = \{\boldsymbol{\beta} = (\boldsymbol{\beta}_S, \boldsymbol{\beta}_{S^C}) : \|\boldsymbol{\beta}\|_1 \leq \|\bar{\boldsymbol{\beta}}_S\|_1 + (1-\eta)\|\bar{\boldsymbol{\beta}}_{S^C}\|_1, \|\boldsymbol{\beta}_{S^C}\|_1 \leq k\|\boldsymbol{\beta}_S\|_1\}$.

Apply the residual bootstrap method to obtain SPSP+AdaLasso (SPSP+Lasso) bootstrap estimators (Efron (1979), Freedman (1981), Knight and Fu (2000), Chatterjee and Lahiri (2011))

	Residual Bootstrap for SPSP
(1)	apply SPSP+Lasso or SPSP+AdaLasso to get: $\tilde{\boldsymbol{\beta}}$ and $\tilde{S}$;
(2)	compute residuals: $\tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$;
(3)	center residuals: $\tilde{\varepsilon}_{\text{cent},i} = \tilde{\varepsilon}_i - \bar{\tilde{\varepsilon}}\ (i = 1, \ldots, n), \bar{\tilde{\varepsilon}} = n^{-1} \sum \tilde{\varepsilon}_i$;
(4)	i.i.d resample B copies of $\tilde{\boldsymbol{\varepsilon}}^{(b)} = (\varepsilon_1^{(b)}, \ldots, \varepsilon_n^{(b)})$ from $\tilde{\varepsilon}_{\text{cent},i}$;
(5)	construct bootstrapped response as: $\mathbf{y}^{(b)} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}}^{(b)}$;
	then, the B bootstrap samples are: $\{(\mathbf{y}^{(b)}, \mathbf{X}, \tilde{\boldsymbol{\varepsilon}}^{(b)})\}_{b=1}^{B}$;
(6)	apply SPSP methods for B times to get: $\{\hat{\boldsymbol{\beta}}^{(b)} = (\hat{\beta}_1^{(b)}, \ldots, \hat{\beta}_p^{(b)})\}$

Simultaneous Confidence Intervals

A general approach for the constructing of simultaneous confidence intervals:

We define outlyingness score as follow:

$$O^{(b)} = g(\hat{\beta}^{(b)}) = (o_1^{(b)}, \ldots, o_d^{(b)}) \in \mathbb{R}^{+d},\ b \in 1, \ldots, B.$$

Procedure:	Simultaneous Confidence Rigion
Step 1 :	Apply residual bootstrap for SPSP to obtain: $\{\hat{\beta}^{(b)}\}_{b=1}^{B}$;
Step 2 :	Construct outlyingness score: $O^{(b)} = (o_1, o_2, \ldots, o_d) = g(\hat{\beta}^{(b)}) \in \mathbb{R}^{+d}$;
Step 3 :	Calculate the $q_i(1-\frac{\nu}{d})$ is $(1-\frac{\nu}{d})$ quintile of o_i;
Step 4 :	Construct a set $\mathcal{A}_\nu \subset \{1, \ldots, B\}$: $\mathcal{A}_\nu = \{b \in (1, \ldots, B);\ o_i^{(b)} \le q_i(1-\frac{\nu}{d}), i = 1, \ldots, d\}$;
Step 5 :	Construct the SCI as: $\mathrm{SCI}_{(1-\alpha)} = \left\{\beta \in \mathbb{R}^p;\ \min_{b \in \mathcal{A}_{\nu^*}} \beta_j^{(b)} \le \beta_j \le \max_{b \in \mathcal{A}_{\nu^*}} \beta_j^{(b)}, j = 1, \ldots, p\right\}$, where the $\nu^* = \underset{\nu}{\mathrm{argmax}} \lvert\mathcal{A}_\nu\rvert$, s.t. $\lvert\mathcal{A}_\nu\rvert \le (1-\alpha)B$.

1. $O^{\mathsf{F},(b)} = (o^{\mathsf{F},(b)}) = g^{\mathsf{F}}(\hat{\beta}^{(b)}) = \hat{F}(\gamma_b, \gamma_f) = \frac{(\mathrm{RSS}_{\gamma_b} - \mathrm{RSS}_{\gamma_f})/(df_{\gamma_b} - df_{\gamma_f})}{\mathrm{RSS}_{\gamma_f}/df_{\gamma_f}}$

   - It is based on the residual sum of squares of the bootstrap model.
   - This outlyingness score can rule out too simple models.

We can obtain the set

$$\mathcal{A}^{\mathsf{F}} = \{b \in (1, \ldots, B);\ o^{\mathsf{F},(b)} \leq q_F(1-\alpha)\} \subset (1, \ldots, B),$$

where the $q^{\mathsf{F}}(1-\alpha)$ is $(1-\alpha)$-quantile of bootstrap distribution of o^{F}

In the end, the

$$\mathrm{SCI}^{\mathsf{F}}(1-\alpha) = \Big\{\boldsymbol{\beta} \in \mathbb{R}^p;\ \min_{b \in \mathcal{A}^{\mathsf{F}}} \beta_j^{(b)} \leq \beta_j \leq \max_{b \in \mathcal{A}^{\mathsf{F}}} \beta_j^{(b)}, j = 1, \ldots, p\Big\}.$$

Outlyingness Score: Standardized Maximum-Minimum

$$
\textbf{2.}\quad O^{\text{MaxMin},(b)} = (o_{\text{max}}^{(b)}, o_{\text{min}}^{(b)}) = g^{\text{MaxMin}}(\hat{\beta}^{(b)})
$$
$$
= \left(\max_{j \in \{1,\dots,p\}} \left(\frac{\hat{\beta}_j^{(b)} - \bar{\hat{\beta}}_j}{s.e._{\hat{\beta}_j}} \right), \min_{j \in \{1,\dots,p\}} \left(\frac{\hat{\beta}_j^{(b)} - \bar{\hat{\beta}}_j}{s.e._{\hat{\beta}_j}} \right) \right).
$$

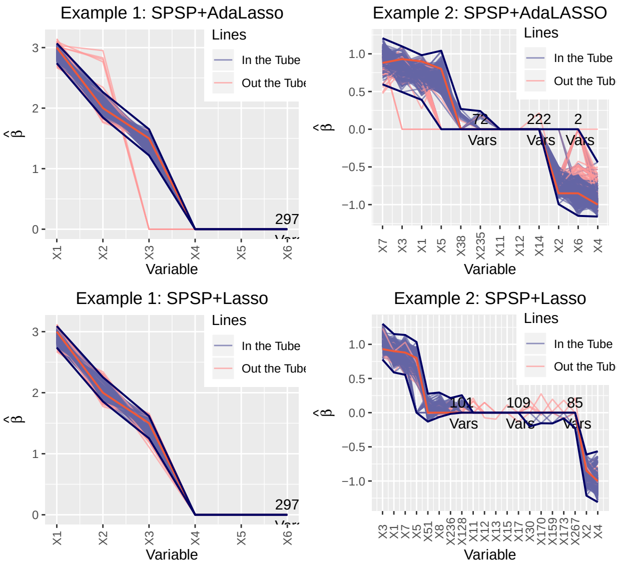
- It is designed for SCI only rely on the empirical bootstrapping distribution of coefficients
- Ruling out tails: those bootstrap estimators with either very large maximum or very small minimum among all bootstrap samples

Outlyingness Score: Standardized Maximum-Minimum

$$\mathcal{A}_{\nu^*}^{\text{MaxMin}} = \{b \in (1, \dots, B);\ o_{\max}^{(b)} \le q_{\max}(1-\frac{\nu^*}{d}),\ o_{\min}^{(b)} \le q_{\min}(1-\frac{\nu^*}{d})\}.$$

$$\text{SCI}_{(1-\alpha)}^{\text{MaxMin}} = \left\{\boldsymbol{\beta} \in \mathbb{R}^p;\ \min_{b \in \mathcal{A}^{\text{MaxMin}}} \beta_j^{(b)} \le \beta_j \le \max_{b \in \mathcal{A}^{\text{MaxMin}}} \beta_j^{(b)}, j = 1, \dots, p\right\}$$

Simultaneous Confidence Tube



- **Example 1**: (Tibshirani, 1996) $\beta_i^* = (3, 2, 1.5), i = 1, 2, 3$, the remaining coefficients equal zero. The correlation between x_{j_1} and x_{j_2} is $0.5^{|j_1 - j_2|}$.

Table 1: The Comparison of SCIs in Example 1

SCI	W.Nzero	W.Zero	Cover Pr	Avg Card	Med Card	Std Card
SPSP+AdaLasso(MaxMin)	0.66	0.00	97.50	1.30	1.00	0.67
SPSP+AdaLasso(F)	0.80	0.00	100.00			
SPSP+Lasso(MaxMin)	0.40	0.00	94.50	1.00	1.00	0.00
SPSP+Lasso(F)	0.40	0.00	100.00			
AdaLasso(MaxMin)	0.42	0.00	60.50	1.00	1.00	0.00
AdaLasso(F)	0.43	0.00	82.00			
Lasso(MaxMin)	0.54	0.17	56.00	898.23	896.00	17.58
Lasso(F)	0.54	0.17	58.50			
True model(MaxMin)	0.39	0.00	96.00	1.00	1.00	0.00
True model(F)	0.40	0.00	100.00			

- **Example 2**: Let $\beta^* = (0.9, -0.85, 0.93, -1, 0.8, -0.85, 0.88)$, and let the remaining coefficients equal zero. The correlation between x_{j_1} and x_{j_2} is $0.5^{|j_1 - j_2|}$. We set $n = 200$, $p = 300$, and $\sigma = 1$ of error.

Table 2: The Comparison of SCIs in Example 2.

SCI	W.Nzero	W.Zero	Cover Pr	Avg Card	Med Card	Std Card
SPSP+AdaLasso(MaxMin)	0.60	0.04	96.50	68.31	59.00	51.66
SPSP+AdaLasso(F)	0.61	0.06	98.50			
SPSP+Lasso(MaxMin)	0.92	0.19	96.50	734.19	770.50	150.75
SPSP+Lasso(F)	0.92	0.19	96.50			
AdaLasso(MaxMin)	0.64	0.21	66.00	949.24	950.00	1.56
AdaLasso(F)	0.64	0.21	65.50			
Lasso(MaxMin)	0.54	0.25	0.00	950.00	950.00	0.00
Lasso(F)	0.54	0.25	0.00			
True model(MaxMin)	0.45	0.00	92.50	1.00	1.00	0.00
True model(F)	0.46	0.00	99.50			

- **Example 3**: Let $\beta^* = (1, -1.25, 0.75, -0.95, 1.5)$, and let the remaining coefficients equal zero. The correlation between x_{j_1} and x_{j_2} is $0.5^{|j_1 - j_2|}$.

Table 3: The Comparison of SCIs in Example 3.

SCI	W.Nzero	W.Zero	Cover Pr	Avg Card	Med Card	Std Card
SPSP+AdaLasso(MaxMin)	0.74	0.01	88.00	15.92	3.00	74.82
SPSP+AdaLasso(F)	0.82	0.01	89.50			
SPSP+Lasso(MaxMin)	1.07	0.08	79.50	239.66	219.50	160.10
SPSP+Lasso(F)	1.07	0.09	79.50			
AdaLasso(MaxMin)	0.65	0.13	68.00	895.24	914.00	55.85
AdaLasso(F)	0.65	0.13	68.50			
Lasso(MaxMin)	0.54	0.23	0.00	950.00	950.00	0.00
Lasso(F)	0.54	0.23	0.00			
True model(MaxMin)	0.43	0.00	92.50	1.00	1.00	0.00
True model(F)	0.44	0.00	98.50			

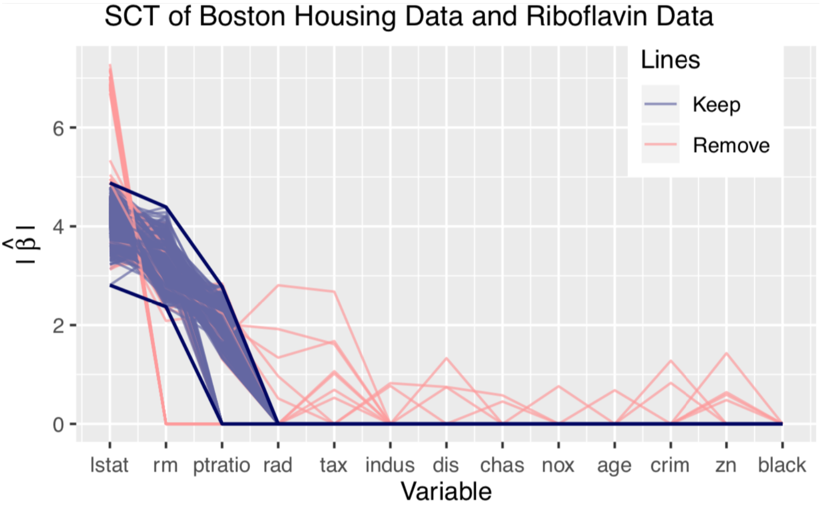
- **Example 4**: (Independent, $p > n$) Let $\beta^* = (4, 3.5, 3, 2.5, 2)$, and let the remaining coefficients equal zero. Covariates are independent.

Table 4: The Comparison of SCIs in Example 4.

SCI	W.Nzero	W.Zero	Cover Pr	Avg Card	Med Card	Std Card
SPSP+AdaLasso(MaxMin)	0.35	0.00	94.50	1.00	1.00	0.00
SPSP+AdaLasso(F)	0.35	0.00	97.50			
SPSP+Lasso(MaxMin)	1.07	0.08	95.00	1.00	1.00	0.00
SPSP+Lasso(F)	1.07	0.09	98.00			
AdaLasso(MaxMin)	0.36	0.00	22.50	1.00	1.00	0.00
AdaLasso(F)	0.36	0.00	56.00			
Lasso(MaxMin)	0.45	0.20	2.50	949.98	950.00	0.17
Lasso(F)	0.45	0.20	2.50			
True model(MaxMin)	0.35	0.00	93.50	1.00	1.00	0.00
True model(F)	0.35	0.00	98.50			

Real Data Examples

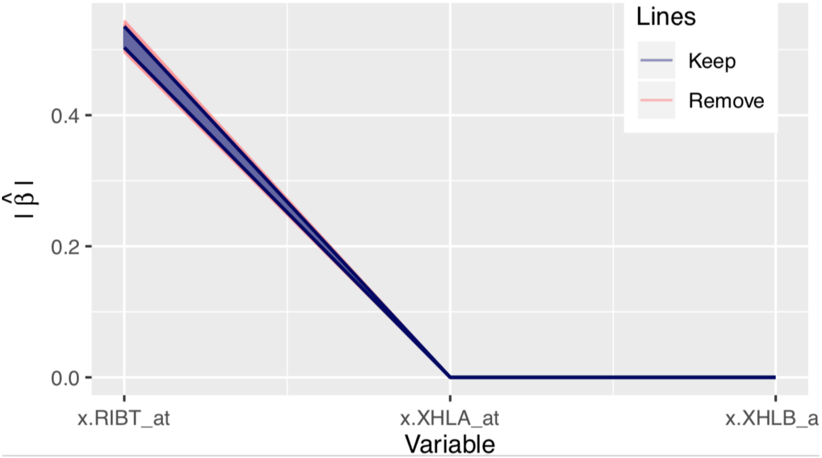
Real Data Example: Boston house pricing



- **LSTAT**, **RM**, **PTRATIO** are the only three plausibly relevant factors
- *PTRATIO* is not significantly relevant at 95% level

Real Data Example: riboflavin (vitamin B_2) production

This dataset contains only 71 (n) observations, but it has 4088 covariates representing the logarithm of the expression level of genes.



- Only gene **ribT** (Reductase) has nonzero confidence interval

Summary

Summary

Our proposed approach can construct the ideal simultaneous confidence intervals with triplefold advantages:

1. They can achieve the *nominal confidence level*;
2. They have *tight intervals for all coefficients* at a given level of confidence;
3. They have the *variable selection results* embedded (the truly irrelevant coefficients have zero width intervals).

Thank you!