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

Two objectives in the high-dimensional sparse linear models:

- Sparse estimation
- **Statistical inference** (our focus)
We focus on linear model as follow:

\[ y = X\beta^* + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \]  

(1)

- \( y \) is the response vector
- \( X_{n \times p} \in \mathbb{R}^p \) is the fixed design matrix containing \( p \) dimensional covariates.
- The parameter vector \( \beta^* = (\beta_1^*, \cdots, \beta_p^*)' \in \mathbb{R}^p \) 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 \( S^c = \{j : \beta_j^* = 0\} \).
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

Illustrative Examples

- **Example 1** (Moderate Correlation, $p > n$, Tibshirani (1996)).
  $\beta^*_i = (3, 2, 1.5), i = 1, 2, 3, \beta^*_i = 0, i = 4, \ldots, 300,$
  $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$. 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
  $\beta^* = (0.9, -0.85, 0.93, -1, 0.8, -0.85, 0.88)$, 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 $\sigma = 1$. 
Illustrative Examples of Drawbacks

1. Biased estimators
2. Poor selection accuracy
Illustrative Examples of Drawbacks

3 Missing of selection information

The simultaneous confidence intervals method by X. Zhang and Cheng (2017) (named as “Sim.CI”):
Example 1: SPSP+AdaLasso

- Vars: 297
- Examples:
  - In the Tube
  - Out the Tube

Example 2: SPSP+AdaLASSO

- Vars: 297
- Examples:
  - In the Tube
  - Out the Tube

Example 1: SPSP+Lasso

- Vars: 297
- Examples:
  - In the Tube
  - Out the Tube

Example 2: SPSP+Lasso

- Vars: 297
- Examples:
  - In the Tube
  - Out the Tube
Preliminaries
Selection by Partitioning the Solution Paths (SPSP)

Idea: Using the whole solution paths of all coefficients and applying the clustering approach (can be applied to Lasso or Adaptive Lasso)

**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 $\zeta$ 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 $\beta_S$. 
Assumption 2.2: Weak Identifiability Condition

Let $\eta > 0$ be some constant. For any $\bar{\beta} = (\bar{\beta}_S, \bar{\beta}_SC)$, then for $k = \frac{2}{2s + Rs(s+1)}$ and some $\kappa$ that satisfies

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

then the WIC,

$$\|X\beta^* - X_S\bar{\beta}_S - X_SC\bar{\beta}_SC\|_2^2 \geq \min_{\beta \in \Theta(\|\bar{\beta}_S\|_1,\|\bar{\beta}_SC\|_1)} \|X\beta^* - X\beta\|_2^2 - \kappa \eta \|\bar{\beta}_SC\|_1,$$

holds. The $\Theta(\|\bar{\beta}_S\|_1,\|\bar{\beta}_SC\|_1) = \{\beta = (\beta_S, \beta_SC) : \|\beta\|_1 \leq \|\bar{\beta}_S\|_1 + (1 - \eta)\|\bar{\beta}_SC\|_1, \|\beta_SC\|_1 \leq k\|\beta_S\|_1\}$. 
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: $\tilde{\beta}$ and $\tilde{S}$;
2. compute residuals: $\tilde{\varepsilon} = y - X\tilde{\beta}$;
3. center residuals: $\tilde{\varepsilon}_{\text{cent},i} = \tilde{\varepsilon}_i - \bar{\tilde{\varepsilon}}$ (i = 1, ..., n), $\bar{\tilde{\varepsilon}} = n^{-1} \sum \tilde{\varepsilon}_i$;
4. i.i.d resample B copies of $\tilde{\varepsilon}^{(b)} = (\varepsilon_{1}^{(b)}, \ldots, \varepsilon_{n}^{(b)})$ from $\tilde{\varepsilon}_{\text{cent},i}$;
5. construct bootstrapped response as: $y^{(b)} = X\tilde{\beta} + \tilde{\varepsilon}^{(b)}$; then, the B bootstrap samples are: $\{(y^{(b)}, X, \tilde{\varepsilon}^{(b)})\}_{b=1}^{B}$;
6. apply SPSP methods for B times to get: $\hat{\beta}^{(b)} = (\hat{\beta}_1^{(b)}, \ldots, \hat{\beta}_p^{(b)})$.
Simultaneous Confidence Intervals
Suppose $\beta_1 = 4, \beta_2 = 0.2, \beta_3 = 0$. Bootstrap times is 1000.

Red dots are the 5% outlying bootstrap estimators.

Geometrical Differences:
- SCI based on debiased lasso estimator is a ellipsoid
- Ours is a rectangle in this example in two dimension, since $\beta_3$ is always estimated as 0
We propose a general approach for the constructing of simultaneous confidence intervals. It relies on outlyingness score as following form:

\[ O^{(b)} = g(\hat{\beta}) = (o^{(b)}_1, \ldots, o^{(b)}_d) \in \mathbb{R}^+^d, \ b \in 1, \ldots, B. \]

It measures the relative location of a bootstrap estimator among all B bootstrap estimators.

Then, we can rule out \( \alpha \) percent of outlying bootstrap estimators among all to construct the simultaneous confidence intervals with confidence level \( 1 - \alpha \).
Simultaneous Confidence Intervals

**Procedure:** Simultaneous Confidence Region

**Step 1:** Apply residual bootstrap for SPSP to obtain:
\[
\{ \hat{\beta}^{(b)} \}_{b=1}^B
\]

**Step 2:** Construct outlyingness score:
\[
O^{(b)} = (o_1, o_2, \ldots, o_d) = g(\hat{\beta}) \in \mathbb{R}^+^d;
\]

**Step 3:** Calculate the \(q_i(1 - \frac{\alpha}{d})\) is \((1 - \frac{\alpha}{d})\) quintile of \(o_i\);

**Step 4:** Construct a set \(A_\alpha \subset \{1, \ldots, B\}\):
\[
A_\alpha = \{ b \in (1, \ldots, B); \quad o_i^{(b)} \leq q_i(1 - \frac{\alpha}{d}), \quad i = 1, \ldots, d \};
\]

**Step 5:** Construct the SCI as:
\[
SCI_{(1-\alpha)} = \left\{ \beta \in \mathbb{R}^p; \quad \min_{b \in A_\alpha} \beta_j^{(b)} \leq \beta_j \leq \max_{b \in A_\alpha} \beta_j^{(b)}, j = 1, \ldots, p \right\},
\]
Outlyingness Score: F-stat

1. \( O^{F,(b)} = (o^{F,(b)}) = g^F(\hat{\beta}) = \hat{F}(\gamma_b, \gamma_f) = \frac{(RSS_{\gamma_b} - RSS_{\gamma_f})/(df_{\gamma_b} - df_{\gamma_f})}{RSS_{\gamma_f}/df_{\gamma_f}}. \)

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

\[ A^F = \{ b \in (1, \ldots, B); \quad o^{F,(b)} \leq q_F(1 - \alpha) \} \subset (1, \ldots, B). \]

\[ SCI^F(1 - \alpha) = \left\{ \beta \in \mathbb{R}^p; \quad \min_{b \in A^F} \beta_j^{(b)} \leq \beta_j \leq \max_{b \in A^F} \beta_j^{(b)}, j = 1, \ldots, p \right\}. \]
Outlyingness Score: Standardized Maximum-Minimum

2. $O^{\text{MaxMin},(b)} = (o^{(b)}_{\text{max}}, o^{(b)}_{\text{min}}) = g^{\text{MaxMin}}(\hat{\beta})$

$$= \left( \max_{j \in \{1, \ldots, p\}} \left( \frac{\hat{\beta}_j^{(b)} - \bar{\hat{\beta}}_j}{\text{s.e.} \hat{\beta}_j} \right), \left| \min_{j \in \{1, \ldots, p\}} \left( \frac{\hat{\beta}_j^{(b)} - \bar{\hat{\beta}}_j}{\text{s.e.} \hat{\beta}_j} \right) \right) \right).$$

- 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

\[ A_{\alpha}^{\text{MaxMin}} = \{ b \in (1, \ldots, B); \ o_{\text{max}}^{(b)} \leq q_{\text{max}}(1-\frac{\alpha}{d}), \ o_{\text{min}}^{(b)} \leq q_{\text{min}}(1-\frac{\alpha}{d}) \}. \]

\[ \text{SCI}_{\text{MaxMin}}^{(1-\alpha)} = \left\{ \beta \in \mathbb{R}^p; \ \min_{b \in A_{\text{MaxMin}}} \beta_j^{(b)} \leq \beta_j \leq \max_{b \in A_{\text{MaxMin}}} \beta_j^{(b)}, j = 1, \ldots, p \right\} \]
**Theorem**: Under the assumptions (1, 2.1, and 2.2), for $\alpha \in (0, 1)$ and all $\beta \in \mathbb{R}^p$, we have

$$P(\beta \in SCI_{n,(1-\alpha)}) \to 1 - \alpha \text{ as } n \to \infty.$$
We design a graphical tool to display the resulting simultaneous confidence intervals:

Example 1: SPSP+AdaLasso

Example 2: SPSP+AdaLASSO

Example 1: SPSP+Lasso

Example 2: SPSP+Lasso
**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>
<thead>
<tr>
<th>SCI</th>
<th>W.Nzero</th>
<th>W.Zero</th>
<th>Cover Pr</th>
<th>Avg Card</th>
<th>Med Card</th>
<th>Std Card</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPSP+AdaLasso(MaxMin)</td>
<td>0.66</td>
<td>0.00</td>
<td>97.50</td>
<td>1.30</td>
<td>1.00</td>
<td>0.67</td>
</tr>
<tr>
<td>SPSP+AdaLasso(F)</td>
<td>0.80</td>
<td>0.00</td>
<td>100.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPSP+Lasso(MaxMin)</td>
<td>0.40</td>
<td>0.00</td>
<td>94.50</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>SPSP+Lasso(F)</td>
<td>0.40</td>
<td>0.00</td>
<td>100.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AdaLasso(MaxMin)</td>
<td>0.42</td>
<td>0.00</td>
<td>60.50</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>AdaLasso(F)</td>
<td>0.43</td>
<td>0.00</td>
<td>82.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso(MaxMin)</td>
<td>0.54</td>
<td>0.17</td>
<td>56.00</td>
<td>898.23</td>
<td>896.00</td>
<td>17.58</td>
</tr>
<tr>
<td>Lasso(F)</td>
<td>0.54</td>
<td>0.17</td>
<td>58.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True model(MaxMin)</td>
<td>0.39</td>
<td>0.00</td>
<td>96.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>True model(F)</td>
<td>0.40</td>
<td>0.00</td>
<td>100.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
- **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>
<thead>
<tr>
<th>SCI</th>
<th>W.Nzero</th>
<th>W.Zero</th>
<th>Cover Pr</th>
<th>Avg Card</th>
<th>Med Card</th>
<th>Std Card</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPSP+AdaLasso(MaxMin)</td>
<td>0.60</td>
<td>0.04</td>
<td>96.50</td>
<td>68.31</td>
<td>59.00</td>
<td>51.66</td>
</tr>
<tr>
<td>SPSP+AdaLasso(F)</td>
<td>0.61</td>
<td>0.06</td>
<td>98.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPSP+Lasso(MaxMin)</td>
<td>0.92</td>
<td>0.19</td>
<td>96.50</td>
<td>734.19</td>
<td>770.50</td>
<td>150.75</td>
</tr>
<tr>
<td>SPSP+Lasso(F)</td>
<td>0.92</td>
<td>0.19</td>
<td>96.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AdaLasso(MaxMin)</td>
<td>0.64</td>
<td>0.21</td>
<td>66.00</td>
<td>949.24</td>
<td>950.00</td>
<td>1.56</td>
</tr>
<tr>
<td>AdaLasso(F)</td>
<td>0.64</td>
<td>0.21</td>
<td>66.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso(MaxMin)</td>
<td>0.54</td>
<td>0.25</td>
<td>0.00</td>
<td>950.00</td>
<td>950.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Lasso(F)</td>
<td>0.54</td>
<td>0.25</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True model(MaxMin)</td>
<td>0.45</td>
<td>0.00</td>
<td>92.50</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>True model(F)</td>
<td>0.46</td>
<td>0.00</td>
<td>99.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCI(Debiased)</td>
<td>0.97</td>
<td>0.97</td>
<td>98.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 $0.5|j_1 - j_2|$.
**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>
<thead>
<tr>
<th>SCI</th>
<th>W.Nzero</th>
<th>W.Zero</th>
<th>Cover Pr</th>
<th>Avg Card</th>
<th>Med Card</th>
<th>Std Card</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPSP+AdaLasso(MaxMin)</td>
<td>0.35</td>
<td>0.00</td>
<td>94.50</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>SPSP+AdaLasso(F)</td>
<td>0.35</td>
<td>0.00</td>
<td>97.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPSP+Lasso(MaxMin)</td>
<td>1.07</td>
<td>0.08</td>
<td>95.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>SPSP+Lasso(F)</td>
<td>1.07</td>
<td>0.09</td>
<td>98.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AdaLasso(MaxMin)</td>
<td>0.36</td>
<td>0.00</td>
<td>22.50</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>AdaLasso(F)</td>
<td>0.36</td>
<td>0.00</td>
<td>56.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso(MaxMin)</td>
<td>0.45</td>
<td>0.20</td>
<td>2.50</td>
<td>949.98</td>
<td>950.00</td>
<td>0.17</td>
</tr>
<tr>
<td>Lasso(F)</td>
<td>0.45</td>
<td>0.20</td>
<td>2.50</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True model(MaxMin)</td>
<td>0.35</td>
<td>0.00</td>
<td>93.50</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>True model(F)</td>
<td>0.35</td>
<td>0.00</td>
<td>98.50</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
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!