Selective inference: a conditional perspective

Xiaoying Tian Harris
Joint work with Jonathan Taylor

September 30, 2016
Model selection

- Observe data \((y, X)\), \(X \in \mathbb{R}^{n \times p}\), \(y \in \mathbb{R}^n\)
Model selection

- Observe data \((y, X), X \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n\)
- \[\text{model} = \text{lm}(y \sim X1 + X2 + X3 + X4)\]
Model selection

- Observe data \((y, X), X \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n\)

- model = \text{lm}(y \sim X1 + X2 + X3 + X4)
  
- model = \text{lm}(y \sim X1 + X2 + X4)
  
- Inference after model selection
  1. Use data to select a set of variables
  2. Normal z-test to get p-values

- Problem: inflated significance
  1. Normal z-tests need adjustment
  2. Selection is biased towards "significance"
Model selection

- Observe data \((y, X), X \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n\)

- model = \text{lm}(y \sim X1 + X2 + X3 + X4)
  model = \text{lm}(y \sim X1 + X2 + X4)
  model = \text{lm}(y \sim X1 + X3 + X4)
Model selection

- Observe data \((y, X), X \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n\)

- \(\text{model} = \text{lm}(y \sim X_1 + X_2 + X_3 + X_4)\)
  \(\text{model} = \text{lm}(y \sim X_1 + X_2 + X_4)\)
  \(\text{model} = \text{lm}(y \sim X_1 + X_3 + X_4)\)

- Inference after model selection
  1. Use data to select a set of variables \(E\)
  2. Normal z-test to get p-values

Problem: inflated significance

- Normal z-tests need adjustment
- Selection is biased towards "significance"
Model selection

- Observe data \((y, X), X \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n\)
- \(\text{model} = \text{lm}(y \sim X_1 + X_2 + X_3 + X_4)\)
- \(\text{model} = \text{lm}(y \sim X_1 + X_2 + X_4)\)
- \(\text{model} = \text{lm}(y \sim X_1 + X_3 + X_4)\)
- Inference after model selection
  1. Use data to select a set of variables \(E\)
  2. Normal z-test to get p-values
- Problem: inflated significance
  1. Normal z-tests need adjustment
  2. Selection is biased towards “significance”
Inflated Significance

Setup:

- $X \in \mathbb{R}^{100 \times 200}$ has i.i.d normal entries
- $y = X\beta + \epsilon$, $\epsilon \sim N(0, I)$
- $\beta = (5, \ldots, 5, 0, \ldots, 0)$
- LASSO, nonzero coefficient set $E$
- z-test, null p-values for $i \in E$, $i \notin \{1, \ldots, 10\}$
Post-selection inference

- PoSI approach:
  1. Reduce to simultaneous inference
  2. Protects against any selection procedure
  3. Conservative and computationally expensive
Post-selection inference

- PoSI approach:
  1. Reduce to simultaneous inference
  2. Protects against any selection procedure
  3. Conservative and computationally expensive

- Selective inference approach:
  1. Conditional approach
  2. Specific to particular selection procedures
  3. More powerful tests
Conditional approach: example

Consider the selection for “big effects”:

- \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(0, 1), \quad \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \)
- Select for “big effects”, \( \bar{X} > 1 \)
- Observation: \( \bar{X}_{obs} = 1.1 \), with \( n = 5 \)
- Normal z-test v.s. selective test for \( H_0 : \mu = 0 \).
Conditional approach: example

Consider the selection for “big effects”:

- $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(0, 1), \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
- Select for “big effects”, $\bar{X} > 1$
- Observation: $\bar{X}_{\text{obs}} = 1.1$, with $n = 5$
- Normal $z$-test v.s. selective test for $H_0 : \mu = 0$. 

![Graph of original distribution for $\bar{X}$](attachment:image1.png)

![Graph of conditional distribution after selection](attachment:image2.png)
Moral of selective inference

Conditional approach:

- Selection, e.g. $\bar{X} > 1$.

- Conditional distribution after selection, e.g. $N(\mu, \frac{1}{n})$, truncated at 1.

- Target of inference may (or may not) depend on outcome of the selection.
  1. Not dependent: e.g. $H_0 : \mu = 0$.
  2. Dependent: e.g. two-sample problem, inference for variables selected by LASSO
Moral of selective inference

Conditional approach:

- Selection, e.g. $\bar{X} > 1$.
- Conditional distribution after selection, e.g. $N(\mu, \frac{1}{n})$, truncated at 1.
- Target of inference may (or may not) depend on outcome of the selection.
  1. Not dependent: e.g. $H_0 : \mu = 0$.
  2. Dependent: e.g. two-sample problem, inference for variables selected by LASSO
- Random hypothesis?
Random hypothesis

- Replication studies
Random hypothesis

- Replication studies
- Data splitting: observe data \((X, y)\), with \(X\) fixed, entries of \(y\) are independent (given \(X\))
Random hypothesis

- Replication studies
- Data splitting: observe data \((X, y)\), with \(X\) fixed, entries of \(y\) are independent (given \(X\))

Random hypothesis selected by the data

\[
L(y_2) = L(y_2 | H_0 \text{ selected by } y_1)
\]
Random hypothesis

- Replication studies
- Data splitting: observe data \((X, y)\), with \(X\) fixed, entries of \(y\) are independent (given \(X\))

Random hypothesis selected by the data

- Data splitting as a conditional approach:

\[
\mathcal{L}(y_2) = \mathcal{L}(y_2 | H_0 \text{ selected by } y_1).
\]
Selective inference: a conditional approach

- Data splitting as a conditional approach:
  \[ \mathcal{L}(y_2) = \mathcal{L}(y_2|H_0 \text{ selected by } y_1). \]

- Inference based on the conditional law:
  \[ \mathcal{L}(y|H_0 \text{ selected by } y^*), \quad y^* = y^*(y, \omega), \]
  where \( \omega \) is some randomization independent of \( y \).
Selective inference: a conditional approach

- Data splitting as a conditional approach:

\[ \mathcal{L}(y_2) = \mathcal{L}(y_2|H_0 \text{ selected by } y_1). \]

- Inference based on the conditional law:

\[ \mathcal{L}(y|H_0 \text{ selected by } y^*), \quad y^* = y^*(y, \omega), \]

where \( \omega \) is some randomization independent of \( y \).

- Examples of \( y^* \):
  1. \( y^* = y_1 \), where \( \omega \) is a random split
  2. \( y^* = y \), \( \omega \) is void
  3. \( y^* = y + \omega \), where \( \omega \sim N(0, \gamma^2) \), additive noise
Different $y^*$

- Much more powerful tests.
- Randomization transfers the properties of unselective distributions to selective counterparts.

<table>
<thead>
<tr>
<th>$y^* = y$</th>
<th>$y^* = y_1$</th>
<th>$y^* = y + \omega$</th>
<th>randomized LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>Lee et al. (2013),</td>
<td>Data splitting,</td>
<td>T. &amp; Taylor (2015)</td>
</tr>
</tbody>
</table>
Selective v.s. unselective distributions

Example: $X_1, \ldots, X_n \sim i.i.d. N(0, 1)$, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, $n = 5$.
Selection: $\bar{X} > 1$.
Selective v.s. unselective distributions

Example: \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(0, 1), \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, n = 5. \)
Selection: \( \bar{X} + \omega > 1, \) where \( \omega \sim \text{Laplace}(0.15) \)

Explicit formulas for the densities of the selective distribution.

The selective distribution is much better behaved after randomization.
Selective v.s. Unselective distributions

- Suppose $X_i \overset{i.i.d.}{\sim} F$, $X_i \in \mathbb{R}^k$.
- Linearizable statistics: $T = \frac{1}{n} \sum_{i=1}^{n} \xi_i(X_i) + o_p(n^{-\frac{1}{2}})$, with $\xi_i$ being measurable to $X_i$'s.
- Central limit theorem:

$$T \Rightarrow N\left(\mu, \frac{\Sigma}{n}\right),$$

where

$$\mathbb{E}[\xi_i] = \mu \in \mathbb{R}^P, \quad \text{Var}(\xi_i) = \Sigma.$$
Selective v.s. Unselective distributions

- Suppose $X_i \overset{i.i.d.}{\sim} F$, $X_i \in \mathbb{R}^k$.
- Linearizable statistics: $T = \frac{1}{n} \sum_{i=1}^{n} \xi_i(X_i) + o_p(n^{-\frac{1}{2}})$, with $\xi_i$ being measurable to $X_i$'s.
- Central limit theorem:

$$T \Rightarrow N \left( \mu, \frac{\Sigma}{n} \right),$$

where

$$\mathbb{E}[\xi_i] = \mu \in \mathbb{R}^p, \quad \text{Var}(\xi_i) = \Sigma.$$

Would this still hold under the selective distribution?
Selective distributions

Randomized selection with $T^* = T^*(T, \omega), \hat{M} : T^* \mapsto M$,

- Original distribution of $T$ (with density $f$):
  $$f(t)$$

- Selective distribution:
  $$f(t)\ell(t), \quad \ell(t) \propto \int 1 \left\{ \hat{M} [T^*(t, \omega)] = M \right\} g(\omega) \, d\omega$$

where $g$ is the density for $\omega$. Special case, when $T^* = T + \omega$.

- $\ell(t)$ is also called the selective likelihood.
Selective central limit theorem

Theorem (Selective CLT, T. and Taylor (2015))

If

1. Model selection is made with $T^* = T^*(T, \omega)$,
2. Selective likelihood $\ell(t)$ satisfies some regularity conditions,
3. $T$ has moment generating function in a neighbourhood of the origin,

$$L(T \mid H_0 \text{ selected by } T^*) \Rightarrow L(N(\mu, \Sigma) \mid H_0 \text{ selected by } T^*),$$
Power comparison

HIVDB http://hivdb.stanford.edu/

Unrandomized $y^* = y$, randomized $y^* = y + \omega$, $\omega \sim N(0, 0.1\sigma^2)$. 

![Parameter values](Unrandomized) 

![Parameter values](Randomized)
Tradeoff between power and model selection

▶ Setup $y = X\beta + \epsilon$, $n = 100$, $p = 200$, $\epsilon \sim N(0, I)$, $\beta = (7, \ldots, 7, 0, \ldots, 0)$. $X$ is equicorrelated with $\rho = 0.3$.

▶ Use randomized $y^*$ to fit Lasso, active set $E$:
   1. Data splitting / Data carving: $y^* = y_1$ random subset of $y$,
   2. Additive randomization: $y^* = y + \omega$, $\omega \sim N(0, \gamma^2 I)$.

Data carving picture credit Fithian et al. (2014).
Thank you!

URL: http://arxiv.org/abs/1410.2597