Selective inference: a conditional perspective

Xiaoying Tian Harris
Joint work with Jonathan Taylor

September 26, 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), $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$
- model = lm(y ~ X1 + X2 + X3 + X4)

Model selection

- Observe data (y, X), $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$
- model = lm(y ~ X1 + X2 + X3 + X4)
  model = lm(y ~ X1 + X2 + X4)

Model selection

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

Model selection

- Observe data (y, X), $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$
- model $=$ lm(y $\sim$ X1 $+$ X2 $+$ X3 $+$ X4)
  model $=$ lm(y $\sim$ X1 $+$ X2 $+$ X4)
  model $=$ lm(y $\sim$ X1 $+$ X3 $+$ X4)
- Inference after model selection
  1. Use data to select a set of variables E
  2. Normal z-test to get p-values

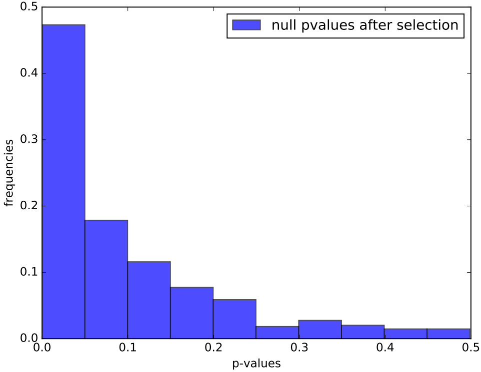
Model selection

- Observe data (y, X), $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$
- model = lm(y $\sim$ X1 + X2 + X3 + X4)
  model = lm(y $\sim$ X1 + X2 + X4)
  model = lm(y $\sim$ X1 + X3 + X4)
- 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
- $y = X\beta + \epsilon$, $\epsilon \sim N(0, I)$
- $\beta = (\underbrace{5, \ldots, 5}_{10}, 0, \ldots, 0)$
- LASSO, nonzero coefficient set E
- z-test, null pvalues 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

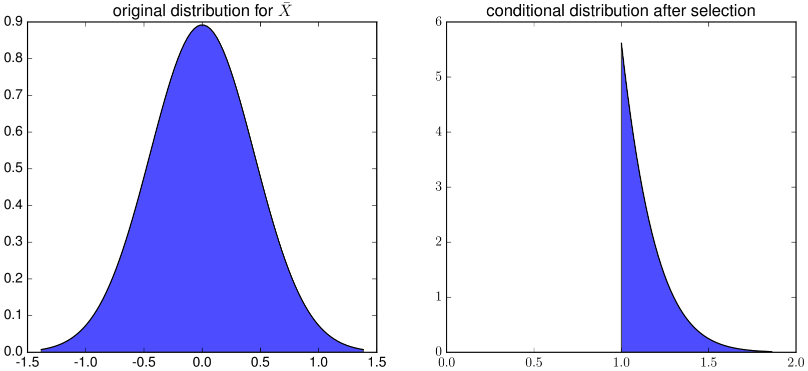
Post-selection inference

- 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":

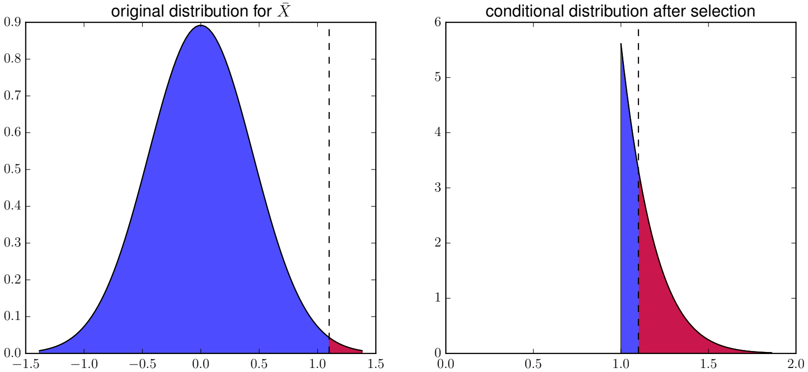
- $X_1, \dots, X_n \overset{i.i.d}{\sim} N(0,1)$, $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$
- Select for "big effects", $\overline{X} > 1$
- Observation: $\overline{X}_{obs} = 1.1$, with $n = 5$
- Normal z-test v.s. selective test for $H_0 : \mu = 0$.



Conditional approach: example

Consider the selection for "big effects":

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



Moral of selective inference

Conditional approach:

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

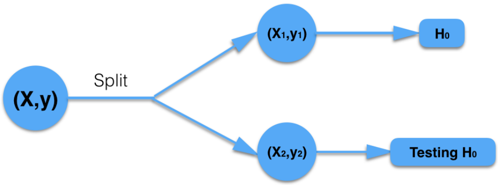
- Selection, e.g. $\overline{X} > 1$.
- Conditional distribution after selection, e.g. $N(\mu, \frac{1}{n})$, truncated at 1.
- Target of inference may (or may not) depend on outcome of the selection.
  1. Not dependent: e.g. $H_0 : \mu = 0$.
  2. Dependent: e.g. two-sample problem, inference for variables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

Random hypothesis

- Replication studies

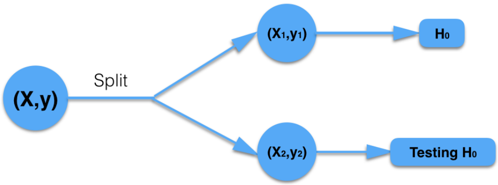
Random hypothesis

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



Random hypothesis

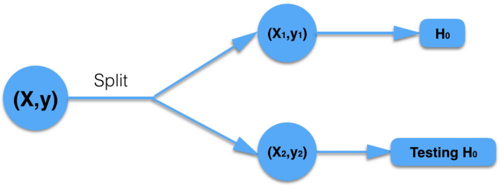
- Replication studies
- Data splitting: observe data (X, y), with X fixed, entries of y are independent (given X)



Random hypothesis selected by the data

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:

$$\mathcal{L}(y | H_0 \text{ selected by } y^*), \qquad y^* = y^*(y, \omega),$$

where ω 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:

$$\mathcal{L}(y | H_0 \text{ selected by } y^*), \qquad y^* = y^*(y, \omega),$$

  where ω is some randomization independent of y.

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

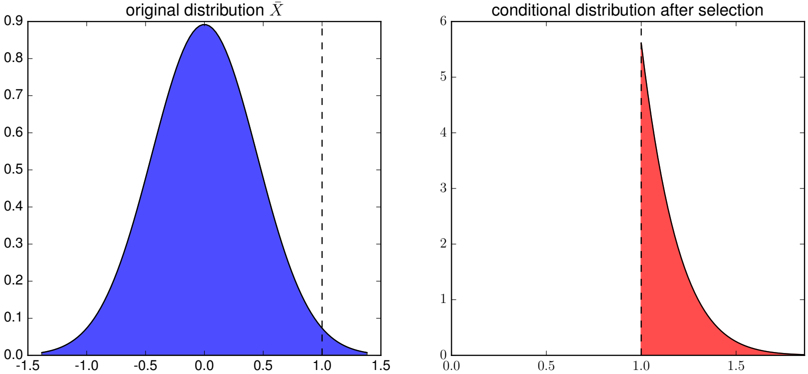
Different y^*

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

	$y^* = y$	$y^* = y_1$	$y^* = y + \omega$	randomized LASSO
y	Lee et al. (2013), Taylor et al.(2014)	Data splitting, Fithian et al.(2014)	T. & Taylor (2015)	T. & Taylor (2015)

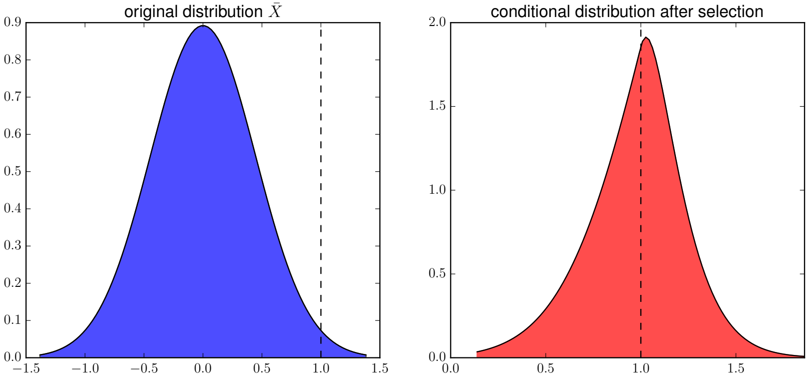
Selective v.s. unselective distributions

Example: $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(0,1)$, $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$, $n = 5$.
Selection: $\overline{X} > 1$.



Selective v.s. unselective distributions

Example: $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(0,1)$, $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$, $n = 5$.
Selection: $\overline{X} + \omega > 1$, 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 $X_i \overset{i.i.d}{\sim} \mathbb{F}$, $X_i \in \mathbb{R}^k$.
- Linearizable statistics: $T = \frac{1}{n}\sum_{i=1}^n \xi_i(X_i) + o_p(n^{-\frac{1}{2}})$, with ξ_i being measurable to X_i's.
- Central limit theorem:

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

where

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

Selective v.s. Unselective distributions

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

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

where

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

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:

$$f(t)\ell(t), \qquad \ell(t) \propto \int \mathbf{1}\left\{\hat{M}\left[T^*(t+\omega)\right] = M\right\} g(\omega)\, d\omega$$

  where g is the density for ω.

- $\ell(t)$ 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*

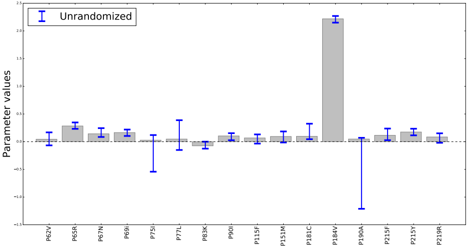
then

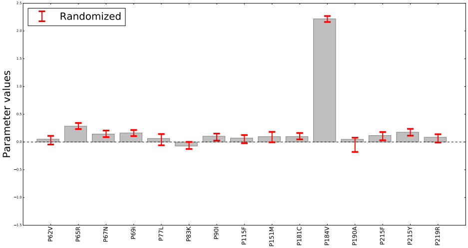
$$\mathcal{L}(T \mid H_0 \text{ selected by } T^*) \Rightarrow \mathcal{L}(N(\mu, \Sigma) \mid H_0 \text{ selected by } T^*),$$

Power comparison

HIVDB http://hivdb.stanford.edu/

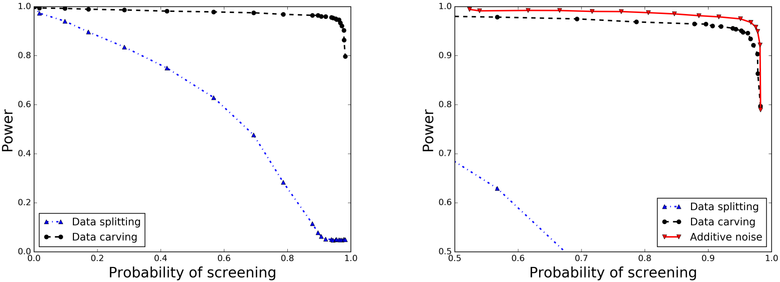
Unrandomized $y^* = y$, randomized $y^* = y + \omega$, $\omega \sim N(0, 0.1\sigma^2)$.





Tradeoff between power and model selection

- Setup $y = X\beta + \epsilon$, $n = 100, p = 200$, $\epsilon \sim N(0, I)$,
  $\beta = (\underbrace{7, \ldots, 7}_{7}, 0, \ldots, 0)$. 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).

Fithian, W., Sun, D. & Taylor, J. (2014), 'Optimal inference after model selection', *arXiv:1410.2597 [math, stat]* . arXiv: 1410.2597.
URL: *http://arxiv.org/abs/1410.2597*