From selective inference to adaptive data analysis

Xiaoying Tian Harris

December 9, 2016

Acknowledgement

My advisor:

- Jonathan Taylor

Other coauthors:

- Snigdha Panigrahi
- Jelena Markovic
- Nan Bi

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 $\sim$ 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 $\sim$ X1 + X2 + X3 + X4)
  model = lm(y $\sim$ 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

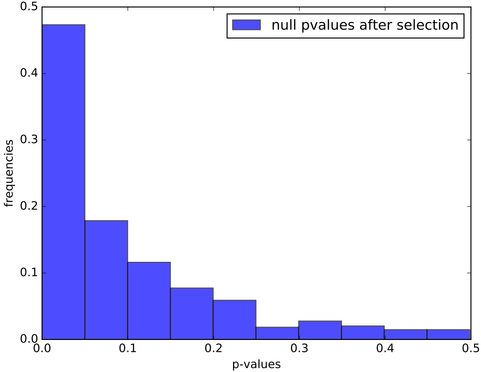
Model selection

- 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:

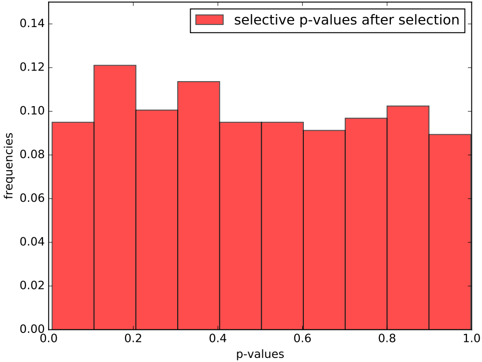
- $X \in \mathbb{R}^{100\times 200}$ has i.i.d normal entries
- $y = X\beta + \epsilon$, $\epsilon \sim N(0, I)$
- $\beta = (\underbrace{5, \ldots, 5}_{10}, 0, \ldots, 0)$
- LASSO, nonzero coefficient set E
- z-test, null pvalues for $i \in E,\ i \notin \{1, \ldots, 10\}$



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 $i \in E$, $i \notin \{1, \ldots, 10\}$



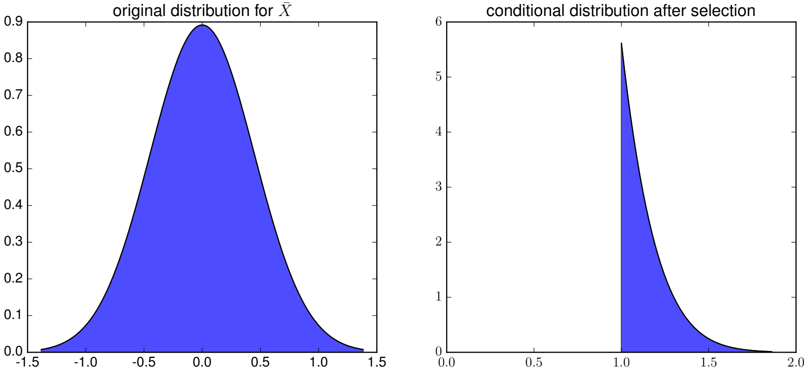
Selective inference: features and caveat

- Specific to particular selection procedures
- Exact post-selection test
- More powerful test

Selective inference: popping the hood

Consider the selection for "big effects":

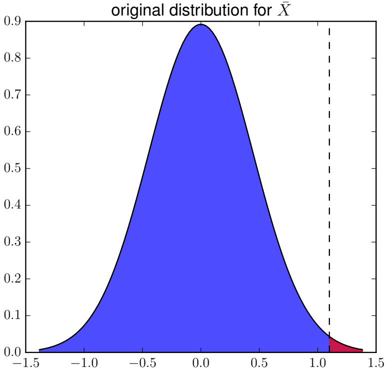
- $X_1, \dots, X_n \overset{i.i.d}{\sim} N(0,1)$, $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$
- Select for "big effects", $\overline{X} > 1$
- Observation: $\overline{X}_{obs} = 1.1$, with $n = 5$
- Normal z-test v.s. selective test for $H_0 : \mu = 0$.

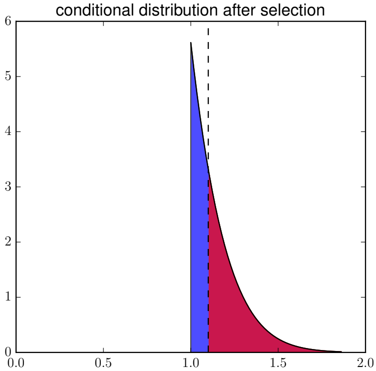


Selective inference: popping the hood

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 $H_0 : \mu = 0$.





Selective inference: in a nutshell

- Selection, e.g. $\overline{X} > 1$.
- Change of the reference measure
  - the conditional distribution, e.g. $N(\mu, \frac{1}{n})$, truncated at 1.
- Target of inference may depend on the outcome of selection
  - Example: selection by LASSO

What is the "selected" model?

Suppose a set of variables E are suggested by the data for further investigation.

- Selected model by Fithian et al. (2014):

$$\mathcal{M}_E = \{N(X_E\beta_E, \sigma_E^2 I), \beta_E \in \mathbb{R}^{|E|}, \sigma_E^2 > 0\}.$$

  Target is β_E.

- Full model by Lee et al. (2016), Berk et al. (2013):

$$\mathcal{M} = \{N(\mu, \sigma^2 I), \mu \in \mathbb{R}^n\}.$$

  Target is $\beta_E(\mu) = X_E^\dagger \mu$.

- Nonparametric model:

$$\mathcal{M} = \{\otimes^n F : (X, Y) \sim F\}.$$

  Target is $\beta_E(F) = \mathbb{E}_F[X_E^T X_E]^{-1} \mathbb{E}_F[X_E \cdot Y]$.

What is the "selected" model?

Suppose a set of variables E are suggested by the data for further investigation.

- Selected model by Fithian et al. (2014):

$$\mathcal{M}_E = \{N(X_E\beta_E, \sigma_E^2 I), \beta_E \in \mathbb{R}^{|E|}, \sigma_E^2 > 0\}.$$

  Target is β_E.

- Full model by Lee et al. (2016), Berk et al. (2013):

$$\mathcal{M} = \{N(\mu, \sigma^2 I), \mu \in \mathbb{R}^n\}.$$

  Target is $\beta_E(\mu) = X_E^\dagger \mu$.

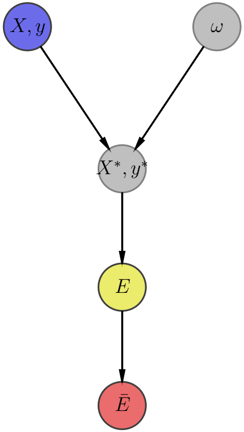
- Nonparametric model:

$$\mathcal{M} = \{\otimes^n F : (X, Y) \sim F\}.$$

  Target is $\beta_E(F) = \mathbb{E}_F[X_E^T X_E]^{-1}\mathbb{E}_F[X_E \cdot Y]$.

A tool for valid inference after exploratory data analysis.

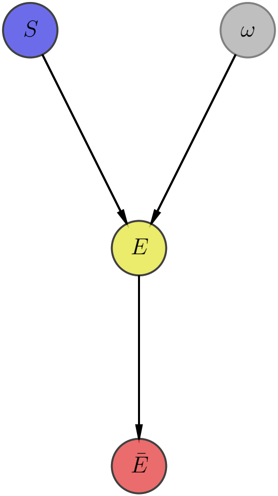
Selective inference on a DAG



- Incoporate randomness through ω
  1. $(X^*, y^*) = (X, y)$
  2. $(X^*, y^*) = (X_1, y_1)$
  3. $(X^*, y^*) = (X, y + \omega)$
- Reference measure conditioning on E, the yellow node.
- Target of inference can be $\overline{E}$
  1. Not E, but depends on the data through E
  2. "Liberating" target of inference from selection
  3. $\overline{E}$ incorporate knowledge from previous literature.

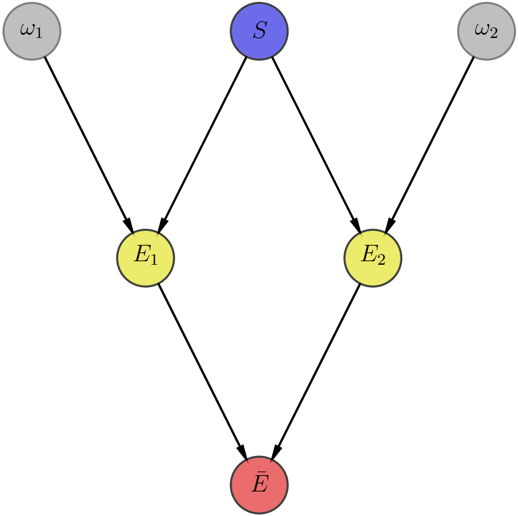
From selective inference to adaptive data analysis

Denote the data by S



From selective inference to adaptive data analysis

Denote the data by S



Reference measure after selection

- Given any point null F_0, use the conditional distribution F_0^* as reference measure,

$$\frac{dF_0^*}{dF_0}(S) = \ell_F(S).$$

- ℓ_F is called the **selective likelihood ratio**. Depends on the selection algorithm and the randomization distribution $\omega \sim G$.
- Tests of the form $H_0 : \theta(F) = \theta_0$ can be reduced to testing point nulls, e.g.
  - Score test
  - Conditioning in exponential families

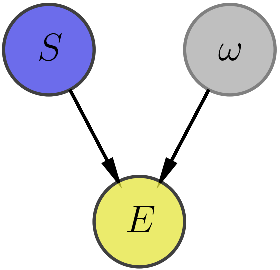
Computing the reference measure after selection

- **Selection map** $\hat{\mathcal{Q}}$ results from an optimization problem,

$$\hat{\beta}(S,\omega) = \arg\min_{\beta} \ell(S;\beta) + \mathcal{P}(\beta) + \omega^T \beta.$$

  E is the active set of $\hat{\beta}$.
- Selection region $A(S) = \{\omega : \hat{\mathcal{Q}}(S,\omega) = E\}$, $\omega \sim G$

$$\frac{dF_0^*}{dF_0}(S) = \int_{A(S)} dG(\omega).$$



$\{\hat{\mathcal{Q}}(S,\omega) = E\}$ is difficult to describe.

Computing the reference measure after selection

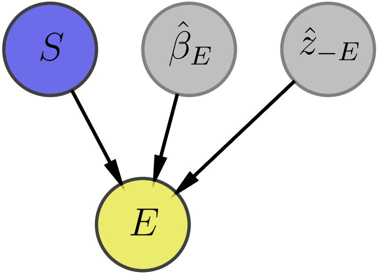
- **Selection map** $\hat{\mathcal{Q}}$ results from an optimization problem,

$$\hat{\beta}(S,\omega) = \arg\min_{\beta} \ell(S;\beta) + \mathcal{P}(\beta) + \omega^T\beta.$$

  E is the active set of $\hat{\beta}$.
- Selection region $A(S) = \{\omega : \hat{\mathcal{Q}}(S,\omega) = E\}$, $\omega \sim G$

$$\frac{dF_0^*}{dF_0}(S) = \int_{A(S)} dG(\omega).$$

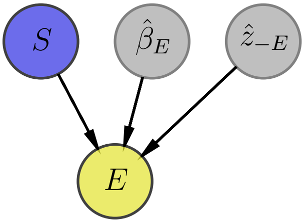
  Let $\hat{z}(S,\omega)$ be the subgradient of the optimization problem.



$\{(\hat{\beta}_E, \hat{z}_{-E}) \in \mathcal{B}\}$, $\mathcal{B}$ depends only on the penalty $\mathcal{P}$.

Monte-Carlo sampler for the conditional distribution

Suppose F_0 has density f_0 and G has density g,



$$\frac{dF_0^*}{dF_0}(S) = \int_{\mathcal{B}} g(\psi(S, \hat{\beta}_E, \hat{z}_{-E}))d\hat{\beta}_E d\hat{z}_{-E},$$

where $\omega = \psi(S, \hat{\beta}_E, \hat{z}_{-E})$.

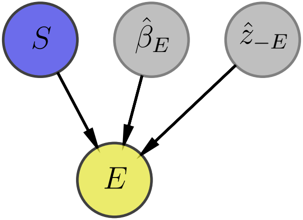
- The reparametrization map ψ is easy to compute, Harris et al. (2016)
- In simulation, we jointly sample $(S, \hat{\beta}_E, \hat{z}_{-E})$ from the density below,

$$f_0(S)g(\psi(S, \hat{\beta}_E, \hat{z}_{-E}))\mathbf{1}_{\mathcal{B}}.$$

Samples of S can be used as reference measure for selective inference.

Interactive Data Analysis

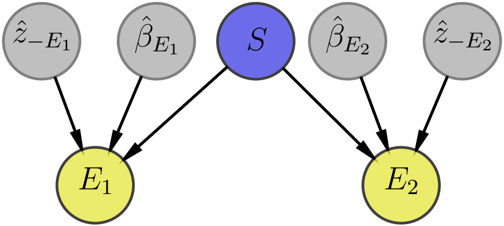
Easily generalizable in a sequential/interactive fashion.



$$f_0(S)g(\psi(S, \hat{\beta}_E, \hat{z}_{-E}))\mathbf{1}_{\mathcal{B}}.$$

Interactive Data Analysis

Easily generalizable in a sequential/interactive fashion.



$$f_0(S)g(\psi_1(S,\hat{\beta}_{E_1},\hat{z}_{-E_1}))\mathbf{1}_{\mathcal{B}_1}\cdot g(\psi_2(S,\hat{\beta}_{E_2},\hat{z}_{-E_2}))\mathbf{1}_{\mathcal{B}_2}.$$

- Flexible framework. Any selection procedure resulting from a "Loss + Penalty" convex problem.
- Examples such as Lasso, logistic Lasso, marginal screening, forward stepwise, graphical Lasso, group Lasso, are considered in Harris et al. (2016).
- Many more is possible.

Summary

- Selective inference on a DAG
- Selection: more than one shot
- Feasible implementation of the selective tests
  https://github.com/selective-inference/Python-software

Thank you!

Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. (2013), 'Valid post-selection inference', *The Annals of Statistics* **41**(2), 802–837.
URL: *http://projecteuclid.org/euclid.aos/1369836961*

Fithian, W., Sun, D. & Taylor, J. (2014), 'Optimal Inference After Model Selection', *arXiv preprint arXiv:1410.2597* . arXiv: 1410.2597.
URL: *http://arxiv.org/abs/1410.2597*

Harris, X. T., Panigrahi, S., Markovic, J., Bi, N. & Taylor, J. (2016), 'Selective sampling after solving a convex problem', *arXiv preprint arXiv:1609.05609* .

Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016), 'Exact post-selection inference with the lasso', *The Annals of Statistics* **44**(3), 907–927.
URL: *http://projecteuclid.org/euclid.aos/1460381681*