STATS 300C: Theory of Statistics **Spring 2022**

Lecture 13 — May 10, 2022

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: Ian Christopher Tanoh*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 and 2022, and scribe notes written by David Fager, Zhimei Ren and Emmanuel Candès.*

13.1 Outline

Agenda: Selective Inference

1. Post-selection inference (POSI)

2. Conditional inference for lasso selection

We live in an era where a multitude of data is available to us, for a variety of topics. This has led to a new paradigm in Statistics: data analysts select models after seeing the data, and not the other way round (e.g using the Akaike Information Criterion - AIC - to select the variables in linear regression). Therefore the classical inference methods cannot provide us with their usual guarantees. Selective inference aims at developing new inference methods that are consistent with this new paradigm.

Through this lecture, we continue to explore selective inference and study this field beyond the scope of hypothesis testing. The results presented today may not stand the test of time. However, the questions discussed are important, and the reader should focus on these questions and how researchers are thinking about them. We are close to the frontier of research, and these questions were at the frontier 5 years ago.

13.2 Post-selection inference (POSI)

The material in this section is from Berk, Brown, Buja, Zhang, and Zhao, 2013.

13.2.1 Problem Formulation

We assume the data follows a linear model as follows:

$$y \sim \mathcal{N}(\mu, \sigma^2 I),$$

with

- $\mu = X\beta$

- X is an $n \times p$ design matrix, and

- σ is known (for convenience).

POSI can be extended to the case where σ is unknown using an independent estimate of σ. Think $p < n$ and $\hat{\sigma}^2 = \text{MSE}_{\text{full model}}$. POSI can also be extended to the case where $\mu \notin span(X)$.

In the classical setting, we specify a model and then fit the model with our data. However, in reality the data analyst selects a model after viewing data, in which case a selection bias is introduced. We hope to develop a scheme so that we can still provide inference about parameters in the selected model.

13.2.2 Classical Inference

In the classical setting, there is a fixed model $M \subset \{1, \ldots, p\}$, and the object of inference is the slopes after adjusting for variables in M only:

$$\beta_M = X_M^\dagger \mu = \mathbb{E}(X_M^\dagger y)$$

with $X_M^\dagger = (X_M' X_M)^{-1} X_M'$. The least-square estimate of β_M is $\hat{\beta}_M = X_M^\dagger y$ and its distribution is:

$$\hat{\beta}_M \sim \mathcal{N}(\beta_M, \sigma^2 (X_M' X_M)^{-1})$$

From this we can get z-scores for testing the hypothesis that the jth regression coefficient in model M is equal to some fixed value $\beta_{j\bullet M}$:

$$z_{j\bullet M} = \frac{\hat{\beta}_{j\bullet M} - \beta_{j\bullet M}}{\sigma \sqrt{(X_M' X_M)^{-1}_{jj}}} = \frac{(y-\mu)' X_{j\bullet M}}{\sigma \|X_{j\bullet M}\|} \sim \mathcal{N}(0,1)$$

where $X_{j\bullet M} = \texttt{lm(X[, j]} \sim \texttt{X[, setdiff(M, j)])\$resid}$ is the residual vector obtained by regressing the jth variable on all the other variables in M.

We observe that this z-score is a linear functional of y. In particular, it is the dot product between a standardized version of y, $\frac{y-\mu}{\sigma} \sim \mathcal{N}(0, I)$, and the unit vector $\frac{X_{j\bullet M}}{\|X_{j\bullet M}\|}$, which we can think of as the direction of the jth variable that is not captured by the other variables in the model or that is orthogonal to them. Also, recall the important following relationship:

$$(X'X)^{-1}_{jj} = \frac{1}{\text{dist}(X_j, X_{-j})^2}$$

We can also construct valid confidence intervals:

$$\hat{\beta}_{j\bullet M} \pm z_{1-\alpha/2} \sigma \|X_{j\bullet M}\|$$

If $\hat{\sigma}^2 = \text{MSE}_{\text{full model}}$, then our confidence intervals are $\hat{\beta}_{j\bullet M} \pm t_{n-p, 1-\alpha/2} \hat{\sigma} \|X_{j\bullet M}\|$.

13.2.3 POSI Setting

For a given variable selection procedure $\hat{M}(y)$, there are a few possible conditions we may want our confidence intervals to satisfy:

- $\mathbb{P}(\beta_{j\bullet\hat{M}} \in C_{j\bullet\hat{M}} | j \in \hat{M}) \geq 1-\alpha$ (inference conditional on being selected)
- $\mathbb{P}(\forall j \in \hat{M}, \beta_{j\bullet\hat{M}} \in C_{j\bullet\hat{M}}) \geq 1-\alpha$ (simultaneous inference over selected variables)

In this setting, the object of inference is random. Further, we do not know $\mathbb{P}(j \in \hat{M})$ because we don't know how the data analyst made their choice. Different selection procedures will lead to different confidence intervals, and it's not obvious how we should construct these confidence intervals.

POSI addresses this problem by constructing confidence intervals that provide simultaneous coverage no matter what the selection procedure was. In other words, they satisfy:

$$\forall \hat{M} \ \ \mathbb{P}(\forall j \in \hat{M}, \beta_{j\bullet\hat{M}} \in C_{j\bullet\hat{M}}) \geq 1-\alpha$$

The pros and cons regarding this approach are:

- **Pros:** This simultaneous inference is the strongest form of protection possible. No matter what the data scientist did, inference is valid. If we can't formalize what the data scientist did, this may be the only valid approach. "The most valuable statistical analyses often arise only after an iterative process involving the data." - Gelman and Loken (2013)
- **Cons:** Confidence intervals can be very wide.
- **Merit:** This work got lots of people thinking.

In principle, POSI is doable using the following observation. For any variable selection procedure $\hat{M}$,

$$\max_{j\in\hat{M}} |z_{j\bullet\hat{M}}| \leq \max_{M} \max_{j\in M} |z_{j\bullet M}|$$

Theorem 1. (Universal Guarantee)

$$\mathbb{P}(\max_{M} \max_{j\in M} |z_{j\bullet M}| \leq K_{1-\alpha/2}) \geq 1-\alpha,$$

where $K_{1-\alpha/2}$ is the POSI constant. Then with $C_{j\bullet\hat{M}} = \hat{\beta}_{j\bullet\hat{M}} \pm K_{1-\alpha/2}\sigma\|X_{j\bullet\hat{M}}\|$,

$$\forall \hat{M} \ \ \mathbb{P}(\forall j \in \hat{M}, \beta_{j\bullet\hat{M}} \in C_{j\bullet\hat{M}}) \geq 1-\alpha.$$

The key to using this universal guarantee theorem is to compute the POSI constant, which is a quantile of the random variable $\max_M \max_{j\in M} |z_{j\bullet M}|$. The difficulty in calculating this quantile is that we have to look at 2^p models. As an alternative, we can try to find asymptotic bounds on this number. It turns out that the POSI constant satisfies:

$$\sqrt{2\log p} \lesssim K_{1-\alpha}(X) \lesssim \sqrt{p}.$$

- The lower bound is achieved for orthogonal designs.
- If we consider a selection process satisfying $\hat{M} = \text{argmax}_M \max_{j \in M} |z_{j \bullet M}|$ ("Single Predictor Adjusted Regression" — SPAR design), the upper bound is achieved as it is a special case of p-hacking.
- The POSI constant can get very large (but necessarily so).

As an aside, note that POSI is similar in spirit to Scheffe's simultaneous confidence intervals for contrasts:

$$c'\beta \quad c \in \mathcal{C} = \left\{ \frac{X_{j\bullet M}}{\|X_{j\bullet M}\|}, j \in M \subset \{1, \ldots, p\} \right\}.$$

Here are some conclusions w.r.t. POSI:

- It provides protection against all kinds of selection.
- It can be very conservative (especially if we don't engage in p-hacking...).
- It is perhaps difficult to implement. It isn't computationally tractable to compute the POSI constant for large p.
- Split sampling is a possible alternative, but it's not always possible because it requires exchangeability, which we don't have, for example, in a designed experiment. (As an aside, note that cross validation, although used in many settings, only works when there is exchangeability…)

A significant impact of POSI is that it asked important questions and stimulated lots of thinking/questioning/research.

13.3 Selective Inference for Lasso

The section relies on the work of Lee, Sun, Sun, and Taylor, 2014.
To get confidence intervals that are shorter than those from POSI, we restrict the analyst's choice. We require that the selection $\hat{M}$ is simply the set of variables selected by Lasso regression for a fixed λ. This is a weakness of this approach, since λ is often chosen using cross-validation. We assume the same setting as before:

$$y \sim \mathcal{N}(\mu, \sigma^2 I) \text{ and } \mu = X\beta.$$

The Lasso selection event is as follows:

$$\hat{\beta} = \underset{b}{\text{argmin}} \frac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1 \Rightarrow \hat{M} = \{j : \hat{\beta}_j \neq 0\}.$$

The objects of inference are the regression coefficients in the reduced model, $\beta_{\hat{M}} := X_{\hat{M}}^{\dagger}\mu$, where $\hat{M}$ is random. The goal is to construct confidence intervals covering the parameters $\beta_{\hat{M}}$.

Figure 13.1. Lasso selection event

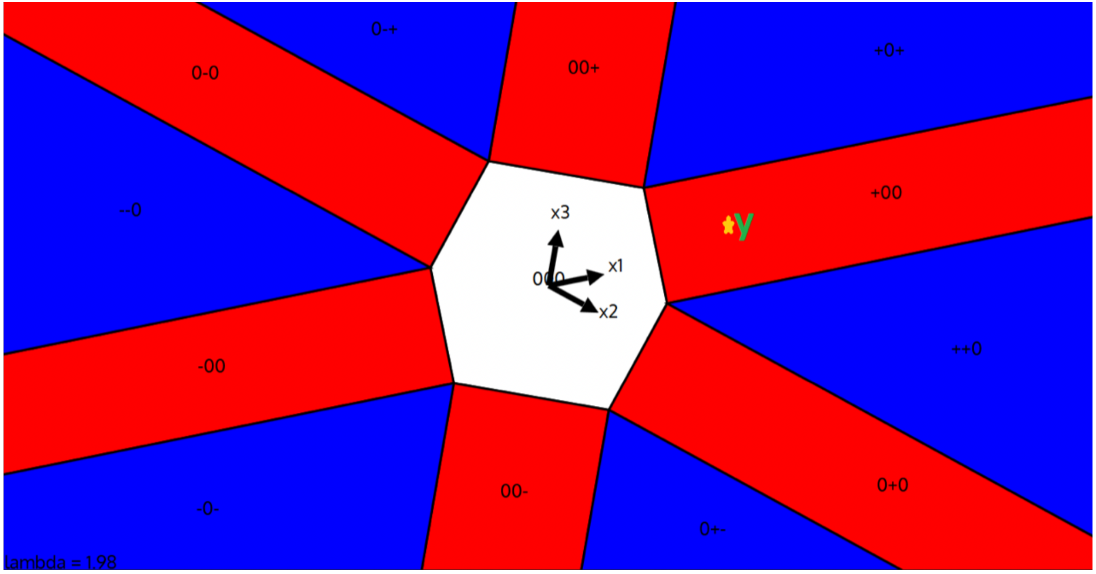


Figure 13.1 is a visualization of the Lasso selection event when $n = 2$ and $p = 3$. The signs in each region represent the signs of the coefficient estimates the Lasso will produce when the data y falls in that region. Each region is a polytope $\{y : Ay \leq b\}$, which can be written as an intersection of half-spaces $\{a_i y \leq bi\}$, where a_i is the ith row of A. These polytopes are easily described via KKT conditions:

$$X_j'(y - X\hat{\beta}) = \lambda \text{sign}(\hat{\beta}_j) \text{ if } \hat{\beta}_j \neq 0,$$

$$|X_j'(y - X\hat{\beta})| \leq \lambda \text{ if } \hat{\beta}_j = 0.$$

Note that a linear equality constraint, as in the case when $\hat{\beta}_j \neq 0$, can also be written as two linear inequality constraints.

The main idea of this work is to condition on the selection event and the signs of the fitted coefficients:

$$y | \{\hat{M} = M, \hat{s} = s\} \sim \mathcal{N}(\mu, \sigma^2 I)\mathbf{1}(Ay \leq b).$$

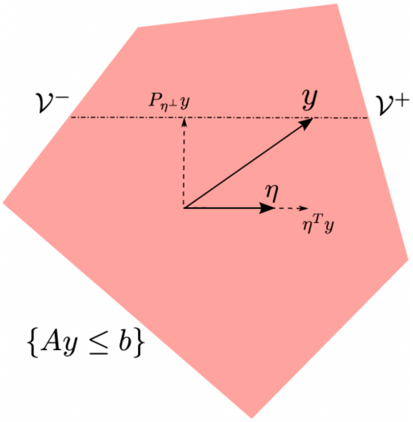
This is a truncated multivariate normal distribution, truncated to a polytope. If we didn't condition on the signs, we would get a multivariate normal truncated to a union of many polytopes, which would not be practical to work with.

We wish to do inference about $\beta_{j \bullet M} = X'_{j \bullet M}\mu := \eta'\mu$. A natural statistic to use is $\eta' y \sim \mathcal{N}(\eta'\mu, \sigma^2\|\eta\|^2)$. In order to do selective inference, we are interested in the distribution of $\eta' y | \{Ay \leq b\}$ which is a complicated mixture of truncated normals that will be computationally expensive to sample from. To make this approach computationally tractable, we also condition on the value of the projection of y onto the space perpendicular to η, $P_{\eta^\perp} y$.

$$\eta' y | \{Ay \leq b, P_{\eta^\perp} y\} \overset{d}{=} \text{TN}(\eta'\mu, \sigma^2\|\eta\|^2, I) \overset{d}{=} \text{TN}(\eta'\mu, \sigma^2\|\eta\|^2, [\mathcal{V}^-(y), \mathcal{V}^+(y)])$$

This is a truncated normal distribution for some truncation interval I. The truncation interval I is the line segment with endpoints $\mathcal{V}^-$ and $\mathcal{V}^+$, which is the intersection of the

Figure 13.2. Conditional sampling distributions on the polytope



polytope $\{Ay \le b\}$ with $P_{\eta^\perp}y$, illustrated in Figure 13.2. The equality in distribution above is not trivial and crucially uses the fact that $\eta' y$ and $P_{\eta^\perp}y$ are independent since they are projections of a Gaussian vector with independent components along orthogonal directions.

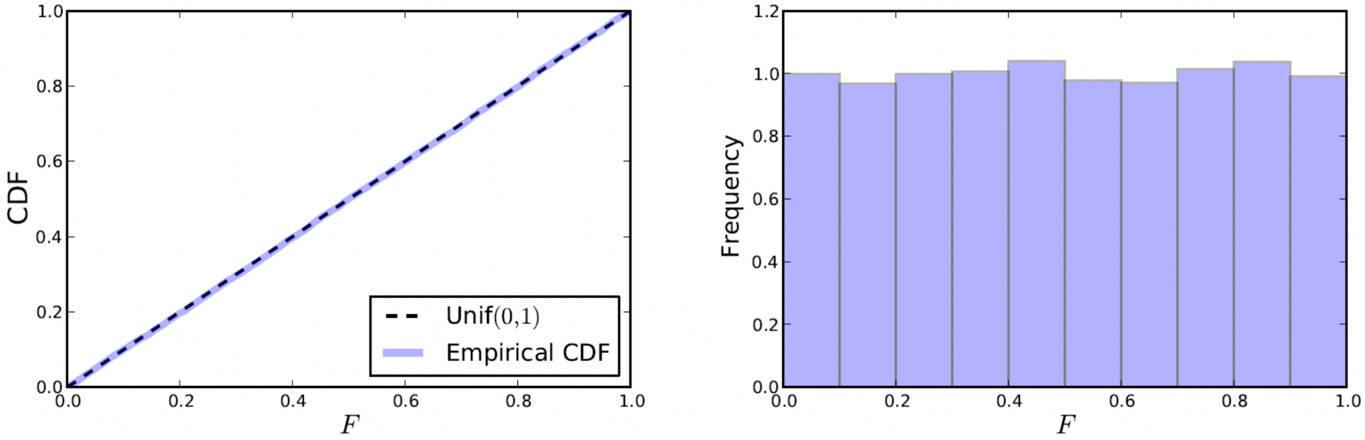
Theorem 2. With $F^{[a,b]}_{\mu,\sigma^2}$ the CDF of $\mathrm{TN}(\mu, \sigma^2, [a,b])$,

$$F^{[\mathcal{V}^-(y),\mathcal{V}^+(y)]}_{\eta'\mu,\sigma^2\|\eta\|^2}(\eta' y)|\{Ay \le b, P_{\eta^\perp}y\} \overset{d}{=} \mathrm{Unif}(0,1).$$

Hence, by integrating over $P_{\eta^\perp}y$, we obtain a pivotal quantity:

$$T := F^{[\mathcal{V}^-(y),\mathcal{V}^+(y)]}_{\eta'\mu,\sigma^2\|\eta\|^2}(\eta' y)|\{Ay \le b\} \sim \mathrm{Unif}(0,1).$$

Figure 13.3. Pivotal quantity is uniform



Using computer simulations, we can see empirically that this pivotal quantity indeed follows a $\mathrm{Unif}(0,1)$ distribution, as shown in Figure 13.3.

We can invert this pivotal quantity to obtain intervals with conditional type-I error control:

$$\begin{aligned} 0.025 \le T \le 0.975 &\Rightarrow a_-(\eta, y) \le \eta' y \le a_+(\eta, y) \\ &\Rightarrow \mathbb{P}(a_-(\eta, y) \le \eta' y \le a_+(\eta, y)|Ay \le b) = 0.95 \end{aligned}$$

We then get conditional coverage

$$\mathbb{P}(\beta_{j\bullet M} \in C_j | \hat{M} = M, \hat{s} = s) = 1 - \alpha,$$

which implies false coverage rate (FCR) control.

$$\mathbb{E}\left[\frac{\#\{j \in \hat{M} : C_j \text{ does not cover } \beta_{j\bullet \hat{M}}\}}{|\hat{M}|}\right] \leq \alpha.$$

Note that we would like to remove the condition $\hat{s} = s$ for more precise inference. However doing so would force us to deal with a huge amount of polytopes (too expensive computationally).

A few closing remarks on the described approach:

- We obtain shorter confidence intervals than with POSI.
- But we have to commit to Lasso with **fixed** λ.
- One other downside of this method is that computation is often numerically unstable because if there are many variables, the polytopes corresponding to specific sign patterns may be very small. This can lead to calculations involving very small probabilities.