STAT 689: Large-scale and high-dimensional statistical inference **Fall 2024**

Lecture 14

1 Background

We are living in an era of unprecedented data abundance, with vast troves of information available across a multitude of subjects. This data-rich landscape has ushered in a new paradigm in statistics, where data analysts now opt to select their models after carefully examining the available data, rather than making those choices upfront (for instance, using tools like the Akaike Information Criterion to choose variables in linear regression). As a result, traditional statistical inference methods may no longer be able to reliably provide their usual guarantees. Selective inference is an active area of research that aims to develop new inferential techniques aligned with this updated data-driven approach to modeling. In this lecture, we will delve deeper into the concepts and challenges of selective inference, extending our exploration beyond just hypothesis testing. The specific results presented here may not stand the test of time, but the underlying questions being grappled with are of great significance. Readers should focus on understanding these key questions and consider how researchers are working to address them.

2 POSI Setting

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

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

In this setting, the object of inference is random. Further, we do not know $P(j \in \widehat{M})$ 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 (POst Selection Inference) addresses this problem by constructing confidence intervals that provide simultaneous coverage no matter what the selection procedure is. In other words, they satisfy:

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

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, the 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 $\widehat{M}$,

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

where

$$z_{j\bullet\widehat{M}} = \frac{\hat{\beta}_{j\bullet\widehat{M}} - \beta_{j\bullet\widehat{M}}}{\widehat{\mathrm{sd}}(\hat{\beta}_{j\bullet\widehat{M}})}$$

is the z-score for testing the significance of the jth covariate in the model $\widehat{M}$.

Theorem.

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

where $K_{1-\alpha/2}$ is the POSI constant. Then with $C_{j \bullet \widehat{M}} = \hat{\beta}_{j \bullet \widehat{M}} \pm K_{1-\alpha/2} \widehat{\mathrm{sd}}(\hat{\beta}_{j \bullet \widehat{M}})$,

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

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 $\widehat{M} = \mathrm{argmax}_M \max_{j \in M} |z_{j \bullet M}|$ ("Single Predictor Adjusted Regression" — SPAR design), the upper bound is achieved as it is the "significance hunting" procedure that selects the model containing the most significant "effect". See Section 6.2 of Berk et al. (2013) for an example.
- The POSI constant can get very large (but not necessarily so).

Here are some conclusions with respect to 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. For more details, please check Berk et al. (2013).

3 Selective inference for clustering

The materials in this section are from Gao et al. (2022, JASA). Suppose we have a set of n observations $X_1, \ldots, X_n \in \mathbb{R}^q$, where

$$X_i \sim N(\mu_i, \sigma^2 \mathbf{I}_q), \quad i = 1, 2, \ldots, n,$$

where $\mu_i \in \mathbb{R}^q$ is the mean vector for the ith observation and σ^2 is assumed to be known. For any $G \subseteq \{1, 2, \ldots, n\}$, we define

$$\bar{\mu}_G = \frac{1}{|G|} \sum_{i \in G} \mu_i, \quad \bar{X}_G = \frac{1}{|G|} \sum_{i \in G} X_i.$$

Given a realization $\mathbf{x} = [x_1, \ldots, x_n]^\top$ of $\mathbf{X} = [X_1, \ldots, X_n]^\top \in \mathbb{R}^{n \times q}$, we first apply a clustering algorithm $\mathcal{C}$ to obtain $\mathcal{C}(\mathbf{x})$ which is a partition of $\{1, 2, \ldots, n\}$. We then use $\mathbf{x}$ to test, for a pair of clusters $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{x})$,

$$H_{0, \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2} : \bar{\mu}_{\hat{\mathcal{C}}_1} = \bar{\mu}_{\hat{\mathcal{C}}_2} \quad \text{versus} \quad H_{a, \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2} : \bar{\mu}_{\hat{\mathcal{C}}_1} \neq \bar{\mu}_{\hat{\mathcal{C}}_2}.$$

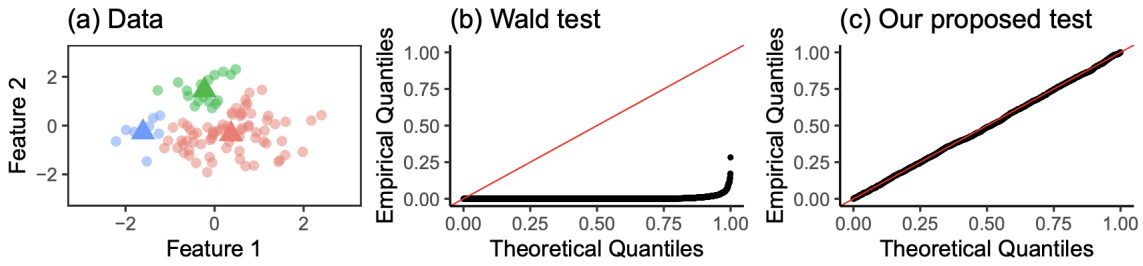

Figure 1: (a) a simulated data set with $\mu_i = \mathbf{0} \in \mathbb{R}^2$ for all i and $\sigma^2 = 1$; QQ-plots of the Unif$(0,1)$ distribution against the p-values from (b) the Wald test and (c) the test based on selective inference.

As σ^2 is known, it is tempting to simply apply a Wald test, with a p-value given by

$$P_{H_{0,\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2}}\left(\|\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}\|_2 \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2\right)$$

where

$$\|\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}\|_2 \sim \sigma\sqrt{\frac{1}{|\hat{\mathcal{C}}_1|} + \frac{1}{|\hat{\mathcal{C}}_2|}}\chi_q.$$

However, since we clustered $\mathbf{x}$ to get $\mathcal{C}(\mathbf{x}) = \{\hat{\mathcal{C}}_k : k = 1, 2, \ldots, K\}$, we shall observe substantial differences between $\{\bar{x}_{\hat{\mathcal{C}}_k}\}_{k=1}^K$ even when there is no signal in the data. In other words, the scaled chi distribution does not provide the correct characterization of the behavior of $\|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2^2$; see Figure 1. To overcome this issue, a natural idea is to define the p-value as

$$P_{H_{0,\hat{\mathcal{C}}_1,\hat{\mathcal{C}}_2}}\left(\|\bar{X}_{\hat{\mathcal{C}}_1} - \bar{X}_{\hat{\mathcal{C}}_2}\|_2 \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2 \middle| \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2 \in \mathcal{C}(\mathbf{X})\right),$$

which accounts for the fact that $\hat{C}_1, \hat{C}_2$ are chosen based on the data. However, it is difficult to characterize this p-value as it depends on some unknown nuisance quantities.

To construct a valid p-value, the core idea would be to condition $\mathbf{X}$ on additional events so that the conditional probability is tractable. To illustrate the idea, let us consider two groups $\mathcal{C}_1, \mathcal{C}_2 \subseteq \{1, 2, \ldots, n\}$ such that $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$ (later on $\mathcal{C}_1$ and $\mathcal{C}_2$ will be specified to be $\hat{\mathcal{C}}_1$ and $\hat{\mathcal{C}}_2$). We define $v(\mathcal{C}_1, \mathcal{C}_2) = (v_1, \ldots, v_n)^\top$ with

$$v_i = \frac{1}{|\mathcal{C}_1|}\mathbf{1}\{i \in \mathcal{C}_1\} - \frac{1}{|\mathcal{C}_2|}\mathbf{1}\{i \in \mathcal{C}_2\}.$$

Let

$$P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)} = \mathbf{I}_n - \frac{v(\mathcal{C}_1, \mathcal{C}_2)v(\mathcal{C}_1, \mathcal{C}_2)^\top}{\|v(\mathcal{C}_1, \mathcal{C}_2)\|_2^2}$$

be the projection matrix associated with the orthogonal complement of the vector $v(\mathcal{C}_1, \mathcal{C}_2)$. For a vector $u \in \mathbb{R}^q$, we define $\mathrm{dir}(u) = u/\|u\|_2$ as the direction of u. We first notice some facts for multivariate normal distribution:

- Fact 1: $\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2} = \mathbf{X}^\top v(\mathcal{C}_1, \mathcal{C}_2)$ and $P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{X}$ are independent.
- Fact 2: $\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2$ and $\mathrm{dir}(\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2})$ are independent under the null that $\bar{\mu}_{\mathcal{C}_1} = \bar{\mu}_{\mathcal{C}_2}$.

Now we define the p-value

$$p(\mathbf{x}, \mathcal{C}_1, \mathcal{C}_2) = P_{H_{0,\mathcal{C}_1,\mathcal{C}_2}}\Bigg(\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2 \geq \|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2 \Big| \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathbf{X}), P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{X} = P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{x},$$

$$\mathrm{dir}(\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}) = \mathrm{dir}(\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2})\Bigg),$$

which will be shown to be valid. Conditional on $P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{X} = P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{x}$ and $\text{dir}(\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}) = \text{dir}(\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2})$, we have

$$\begin{aligned}
\mathbf{X} =& P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{X} + \frac{v(\mathcal{C}_1,\mathcal{C}_2)v(\mathcal{C}_1,\mathcal{C}_2)^\top}{\|v(\mathcal{C}_1,\mathcal{C}_2)\|_2^2}\mathbf{X} \\
=& P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{X} + \frac{v(\mathcal{C}_1,\mathcal{C}_2)(\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2})^\top}{\|v(\mathcal{C}_1,\mathcal{C}_2)\|_2^2} \\
=& P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{X} + \frac{\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2}{|\mathcal{C}_1|^{-1} + |\mathcal{C}_2|^{-1}} v(\mathcal{C}_1,\mathcal{C}_2)\text{dir}(\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2})^\top \\
=& P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{x} + \frac{\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2}{|\mathcal{C}_1|^{-1} + |\mathcal{C}_2|^{-1}} v(\mathcal{C}_1,\mathcal{C}_2)\text{dir}(\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2})^\top,
\end{aligned}$$

where $\|v(\mathcal{C}_1,\mathcal{C}_2)\|_2^2 = |\mathcal{C}_1|^{-1} + |\mathcal{C}_2|^{-1}$. In this case, the event $\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathbf{X})$ is equivalent to $\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2 \in S(\mathbf{x}, \mathcal{C}_1, \mathcal{C}_2)$ where

$$S(\mathbf{x}, \mathcal{C}_1, \mathcal{C}_2) = \left\{ \phi \geq 0 : \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}\left(P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{x} + \frac{\phi}{|\mathcal{C}_1|^{-1} + |\mathcal{C}_2|^{-1}} v(\mathcal{C}_1,\mathcal{C}_2)\text{dir}(\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2})^\top \right) \right\}$$

Using Facts 1-2, the p-value can be rewritten as

$$\begin{aligned}
p(\mathbf{x}, \mathcal{C}_1, \mathcal{C}_2) =& P_{H_{0,\mathcal{C}_1,\mathcal{C}_2}}\Bigg(\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2 \geq \|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2 \,\Big|\, \|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2 \in S(\mathbf{x}, \mathcal{C}_1, \mathcal{C}_2), P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{X} = P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{x}, \\
& \text{dir}(\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}) = \text{dir}(\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}) \Bigg) \\
=& P_{H_{0,\mathcal{C}_1,\mathcal{C}_2}}\left(\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2 \geq \|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2 \,\Big|\, \|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2 \in S(\mathbf{x}, \mathcal{C}_1, \mathcal{C}_2) \right) \\
=& 1 - F\left(\|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2; \sigma(|\mathcal{C}_1|^{-1} + |\mathcal{C}_2|^{-1}), S(\mathbf{x}, \mathcal{C}_1, \mathcal{C}_2) \right),
\end{aligned}$$

where $F(\cdot; c, \mathcal{S})$ denotes the cdf of a $c \cdot \chi_q$ random variable truncated to the set $\mathcal{S}$. Essentially, in the above derivation, we use the fact that conditional on $\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathbf{X}), P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{X} = P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{x}$ and $\text{dir}(\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}) = \text{dir}(\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2})$, $\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2$ follows the truncated $\sigma(|\mathcal{C}_1|^{-1} + |\mathcal{C}_2|^{-1}) \cdot \chi_q$ distribution.

Under $H_{0,\mathcal{C}_1,\mathcal{C}_2}$, we claim that

$$P_{H_{0,\mathcal{C}_1,\mathcal{C}_2}}(p(\mathbf{X}, \mathcal{C}_1, \mathcal{C}_2) \leq \alpha | \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathbf{X})) \leq \alpha.$$

Proof. For the ease of notation, we write $F(x) = F\left(x; \sigma(|\mathcal{C}_1|^{-1} + |\mathcal{C}_2|^{-1}), S(\mathbf{x}, \mathcal{C}_1, \mathcal{C}_2)\right)$. From the above discussion, we know that

$$\begin{aligned}
& P_{H_{0,\mathcal{C}_1,\mathcal{C}_2}}\left(p(\mathbf{X}, \mathcal{C}_1, \mathcal{C}_2) \leq \alpha \Big| \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathbf{X}), P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{X} = P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{x}, \text{dir}(\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}) = \text{dir}(\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}) \right) \\
=& P_{H_{0,\mathcal{C}_1,\mathcal{C}_2}}\left(1 - F(\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2) \leq \alpha \Big| \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathbf{X}), P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{X} = P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{x}, \text{dir}(\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}) = \text{dir}(\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}) \right) \\
=& P_{H_{0,\mathcal{C}_1,\mathcal{C}_2}}\left(\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2 \geq F^{-1}(1-\alpha) \Big| \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathbf{X}), P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{X} = P^{\perp}_{v(\mathcal{C}_1,\mathcal{C}_2)}\mathbf{x}, \text{dir}(\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}) = \text{dir}(\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}) \right) \\
=& 1 - F(F^{-1}(1-\alpha)) = \alpha.
\end{aligned}$$

Therefore,

$$P_{H_{0,\mathcal{C}_1,\mathcal{C}_2}}\left(p(\mathbf{X}, \mathcal{C}_1, \mathcal{C}_2) \leq \alpha \Big| \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathbf{X}) \right) = \alpha.$$

In practice, the p-value can be computed using the Monte Carlo method. Recall that

$$p(\mathbf{x}, \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2) = \frac{P\left(\phi \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2, \phi \in S(\mathbf{x}, \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2)\right)}{P\left(\phi \in S(\mathbf{x}, \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2)\right)},$$

where $\phi \sim \sigma\sqrt{|\hat{\mathcal{C}}_1|^{-1} + |\hat{\mathcal{C}}_2|^{-1}}\chi_q$. We can approximate the numerator by

$$\frac{1}{B}\sum_{i=1}^{B} \mathbf{1}\left\{\phi_i \geq \|\bar{x}_{\hat{\mathcal{C}}_1} - \bar{x}_{\hat{\mathcal{C}}_2}\|_2, \phi_i \in S(\mathbf{x}, \hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2)\right\},$$

where $\phi_1, \ldots, \phi_B$ are independent $\sigma\sqrt{|\hat{\mathcal{C}}_1|^{-1} + |\hat{\mathcal{C}}_2|^{-1}}\chi_q$ random variables.

High-level idea. The key difference with the classical inference setting is that the hypothesis to be tested (say $H_0(\mathbf{X})$) depends on the data (instead of being pre-determined before the data collection process). The hypothesis to be tested can be described by an event of the data $\mathbf{X} \in A$ for some set A. In our case, given $\mathcal{C}_1, \mathcal{C}_2$,

$$\{\mathbf{X} \in A\} = \{\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(\mathbf{X})\}.$$

Suppose we use a test statistic $T(\mathbf{X})$ to test the hypothesis $H_0(\mathbf{X})$. Let $t = T(\mathbf{x})$ be the realization of $T(\mathbf{X})$ based on $\mathbf{x}$ (a realization of $\mathbf{X}$). In selective inference, a valid p-value can be defined

$$P(T(\mathbf{X}) \geq t | \mathbf{X} \in A).$$

However, it may be difficult to evaluate this probability. The essential idea is to express the event $\{\mathbf{X} \in A\}$ as

$$\{(T(\mathbf{X}), S(\mathbf{X})) \in B\}$$

where $S(\mathbf{X})$ is independent of $T(\mathbf{X})$. Conditional on $S(\mathbf{X}) = S(\mathbf{x})$, we define the p-value to be

$$\begin{aligned}
&P(T(\mathbf{X}) \geq t | \mathbf{X} \in A, S(\mathbf{X}) = S(\mathbf{x})) \\
=&P(T(\mathbf{X}) \geq t | (T(\mathbf{X}), S(\mathbf{x})) \in B, S(\mathbf{X}) = S(\mathbf{x})) \\
=&P(T(\mathbf{X}) \geq t | (T(\mathbf{X}), S(\mathbf{x})) \in B),
\end{aligned}$$

which is related to a truncated distribution.

4 Selective Inference for Lasso

The section relies on the work of Lee et al. (2016). To get confidence intervals that are shorter than those from POSI, we restrict the analyst's choice. We require that the selection $\widehat{M}$ 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 setting:

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

The Lasso selection event is as follows:

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

The objects of inference are the regression coefficients in the reduced model,

$$\beta_{\widehat{M}} := (X_{\widehat{M}}^\top X_{\widehat{M}})^{-1} X_{\widehat{M}}^\top \mu = \text{argmin}_b \mathbb{E}[\|y - X_{\widehat{M}} b\|^2 | X],$$

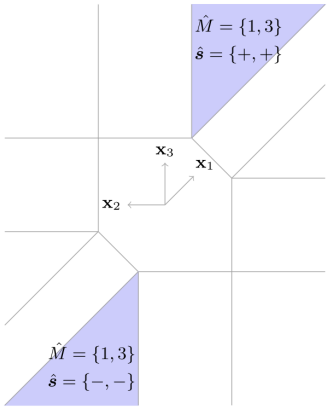

Figure 2: A geometric picture illustrating the Lasso selection events.

where $\widehat{M}$ is random. The goal is to construct confidence intervals covering the parameters $\beta_{\widehat{M}}$.

Figure 2 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 $\{a_i y \leq b_i\}$, where a_i is the ith row of A. These polytopes are easily described via KKT conditions:

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

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. Thus, we can write the KKT condition as

$$Ay \leq b,$$

where it can be shown that A and b only depends on the selected model $\widehat{M}$ and the signs of the non-zero coefficients $\hat{s}$; see Lemma 4.1 of Lee et al. (2016).

The main idea of this work is to condition on the selection event and the signs of the fitted coefficients (i.e., $\hat{s} = s$):

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

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 $\beta_{j \bullet \widehat{M}} = \eta^\top \mu$ for $\eta^\top$ being the jth row of $(X_{\widehat{M}}^\top X_{\widehat{M}})^{-1} X_{\widehat{M}}^\top$. A natural statistic to use is $\eta^\top y \sim N(\eta^\top \mu, \sigma^2 \|\eta\|^2)$. In order to do selective inference, we are interested in the distribution of $\eta^\top y|\{Ay \leq b\}$, 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, $P_{\eta^\perp} y$.

$$\eta^\top y|\{Ay \leq b, P_{\eta^\perp} y = z\} \stackrel{d}{=} \text{TN}(\eta^\top \mu, \sigma^2 \|\eta\|^2, I) \stackrel{d}{=} \text{TN}(\eta^\top \mu, \sigma^2 \|\eta\|^2, [V_-(z), V_+(z)])$$

This is a truncated normal distribution for some truncation interval I. The truncation interval I is the line segment with endpoints V_- and V_+, which is the intersection of the polytope $\{Ay \leq b\}$ with $P_{\eta^\perp} y$, illustrated in Figure 3.

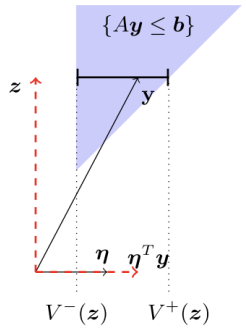

Figure 3: Intersection of the polytope $\{Ay \leq b\}$ with $P_{\eta^\perp} y = z$

The equality in distribution above is not trivial and crucially uses the fact that $\eta^\top y$ and $P_{\eta^\perp} y$ are independent since they are projections of a Gaussian vector with independent components along orthogonal directions.

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

$$F^{\eta^\top \mu, \sigma^2 \|\eta\|^2}_{[V_-(z), V_+(z)]}(\eta^\top y) | \{Ay \leq b, P_{\eta^\perp} y = z\} \overset{d}{=} \mathrm{Unif}(0, 1),$$

where z is computed from the observed dataset. Hence, by integrating over $P_{\eta^\perp} y$, we obtain a pivotal quantity:

$$T := F^{\eta^\top \mu, \sigma^2 \|\eta\|^2}_{[V_-(z), V_+(z)]}(\eta^\top y) | \{Ay \leq b\} \sim \mathrm{Unif}(0, 1).$$

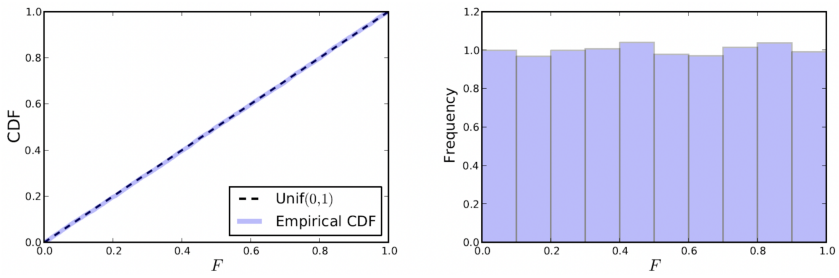

Figure 4: Uniformly distributed p-values.

Using computer simulations, we can see empirically that this pivotal quantity indeed follows a $\mathrm{Unif}(0, 1)$ distribution, as shown in Figure 4.

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

$$C_j = \left\{ \beta : \frac{\alpha}{2} \leq F^{\beta, \sigma^2 \|\eta\|^2}_{[V_-(z), V_+(z)]}(\eta^\top y) \leq 1 - \frac{\alpha}{2} \right\}.$$

We then get the conditional coverage

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

which implies

$$P(\beta_{j \bullet M} \in C_j | \widehat{M} = M) \geq 1 - \alpha,$$

and the false coverage rate (FCR) control:

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

We want to remove the condition $\hat{s} = s$ for more precise inference. However, doing so would force us to deal with many 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.