Exact Post-selection Inference, With Application to The LASSO (Lee. 2016)

1. Justify for building conditional CI.

① unselected parameters should not be considered (β_{uns} doesn't exist, and shouldn't be set 0) (Berk 2013 suppl)

② relation to data splitting : ~~similar to~~ valid only conditional on selected model (Fithian, Sun & Taylor 2014)

③ simultaneous coverage is too stringent, and computationally heavy:

Berk 2013 controls $FWER = P(\beta_j^{\hat{M}} \notin C_j^{\hat{M}}, \forall j \in \hat{M})$ also marginal for all $\hat{M} = M$

④ controlling conditional coverage implies controlling some "familywise properties".

$$FCR = E\left[\frac{|\{j \in \hat{M} : \beta_j^{\hat{M}} \notin C_j^{\hat{M}}\}|}{|\hat{M}|} ; |\hat{M}| > 0\right]$$

PFCR = ↗ but condition on $|\hat{M}| > 0$.

Lemma 2.1 $\Rightarrow P(\beta_j^{\hat{M}} \notin C_j^{\hat{M}} | \hat{M} = M) \le \alpha,\ \forall M, \forall j \in M \Rightarrow FCR \le PFCR \le \alpha.$

2. Setting & Major idea

$y \sim N(\mu, \sigma^2 I)$. $\beta_j^M = e_j^T X_M^+ \mu$ is the inference target (doesn't assume a true model).

use the distribution of $\eta_M^T y | \{\hat{M} = M\}$ to draw inference on $\eta_M^T \mu$. (if $\eta_M = e_j^T X_M^+$, then β_j^M).

model selection procedure : LASSO. $\hat{\beta} \in \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$

3. Major Results:

① selection procedure $\{\hat{M} = M, \hat{s} = s\}$ can be expressed as a polyhedra $\left\{ \begin{pmatrix} A_0(M,s) \\ A_1(M,s) \end{pmatrix} y \le \begin{pmatrix} b_0(M,s) \\ b_1(M,s) \end{pmatrix} \right\}$ (↗ signs)

$\Downarrow$ A_0, b_1 depend on λ.

$$\{\hat{M} = M\} = \bigcup_{s \in \{-1,1\}^{|M|}} \{A(M,s) y \le b(M,s)\}$$

② $P(\beta_j^M \in C_j^M | \hat{M} = M, \hat{s} = s) \ge 1-\alpha \underset{\forall s}{\Rightarrow} P(\beta_j^M \in C_j^M | \hat{M} = M) \ge 1-\alpha.$ condition on either would work. $\{\hat{M} = M\}$ gives shorter interval. ? intuition

③ build statistic & CI conditional on single polehedra $\{Ay \le b\}$

$\{Ay \le b\} = \{\mathcal{V}^-(z) \le \eta^T y \le \mathcal{V}^+(z),\ \mathcal{V}^0(z) \ge 0\}$ $\mathcal{V}$s are functions of z, depend on A, b, η.

$z = (I_n - c\eta^T) y \quad c = \Sigma\eta(\eta^T\Sigma\eta)^{-1} \quad z \perp \eta^T y$

$$\eta^T y | \underset{z = z_0}{\{Ay \le b\}} \overset{d}{=} \eta^T y | \underset{z = z_0}{\{\mathcal{V}^-(z) \le \eta^T y \le \mathcal{V}^+(z), \mathcal{V}^0(z) \ge 0\}} \sim TN(\eta^T\mu, \sigma^2\eta^T\Sigma\eta, \mathcal{V}^-(z_0), \mathcal{V}^+(z_0))$$

$\Rightarrow$ statistic $F_{\eta^T\mu,\, \eta^T\Sigma\eta}^{[\mathcal{V}^-(z_0),\, \mathcal{V}^+(z_0)]}(\eta^T y) | \{Ay \le b, z = z_0\} \sim Unif(0,1)$

$\Rightarrow$ marginalize over z $\quad F_{\eta^T\mu,\, \eta^T\Sigma\eta}^{z}(\eta^T y) | \{Ay \le b\} \sim Unif(0,1)$

$\Rightarrow$ CI: $\eta^T\mu \in [L, R]$ s.t. $F_{L,\, \eta^T\Sigma\eta}^{z}(\eta^T y) = \frac{\alpha}{2}$, $F_{R,\, \eta^T\Sigma\eta}(\eta^T y) = 1 - \frac{\alpha}{2}$

④ ... union of polyhedra $\bigcup_s \{A_s y \le b_s\}$

$$F_{\eta^T\mu,\, \eta^T\Sigma\eta}^{\bigcup_s [\mathcal{V}_s^-(z),\, \mathcal{V}_s^+(z)]}(\eta^T y) \mid \bigcup_s \{A_s y \le b_s\} \sim \mathrm{Unif}(0,1).$$

similar way in ③ to build CI (generally shorter for wider ~~tu~~ truncation)

(but computationally more heavy, for A_s, b_s ...)

4. things I don't (intuitively) understand

long interval ~~of CI~~ for truncated Gaussian near boundary ?

unbiased interval ? $P_\theta(\theta' \in C) \le 1-\alpha \quad \forall\, \theta,\ \theta' \ne \theta$

5. Simulation & example result.

strong signal: recovers OLS interval

Shorter than data splitting & simultaneous interval. suprising.

maybe more info on model selection used.

weak signal: near boundary of truncation (why?)

$\Rightarrow$ longer interval

but similar to data splitting