post-selection inference (Kuchibhotla, 2022)

Setting: $\mathcal{Q} = \{q\}$ all possible selection (model, covariates, transformation, ...)

interested in misspecification-robust target $\theta_q = \underset{\theta \in \Theta_q}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} E[l_q(\theta, Z_i)]$

① It's only a loss function, no need to give a param model

② θ_q can depend on data generating process

stage 1: select $\hat{q}$ (with some data-driven procedure)

stage 2: estimate $\theta_{\hat{q}}$ using $\hat{\theta}_{\hat{q}}$ from data.

$\{Z_i\}_{i=1}^{n}$

$\hat{q} \longrightarrow \hat{\theta}_{\hat{q}}$ $\boxed{?? \ \liminf_{n\to\infty} P(\theta_{\hat{q}} \in \hat{CI}_{\hat{q}}) \geq 1-\alpha, \ \hat{CI}_{\hat{q}} \text{ based on } \hat{\theta}_{\hat{q}}}$ example of VIDE.

$q \longrightarrow \hat{\theta}_q$ easy to find $\hat{CI}_q$ s.t. $\liminf_{n\to\infty} P(\theta_q \in \hat{CI}_q) \geq 1-\alpha$

Further reading: difference with dimension reduction (Berk 2013)

Solution 1: Data Splitting. $\{Z_i\}_{i=1}^{m} \perp \{Z_i\}_{i=m+1}^{n}$ $P(\hat{\theta}_{\hat{q}} - \theta_{\hat{q}} \in A \mid \hat{q} = q) = P(\hat{\theta}_q - \theta_q \in A)$

$\hat{q} \longrightarrow \hat{\theta}_{\hat{q}}$

conformal inference

$X_i \ Y_i \ \hat{Y}_i$

pro: no restriction on selection procedure

cons: ① unacceptable model (?)

② invalid for dependent data.

③ effect brought by splitting (size & randomness).

Note: in this case, the inference being conservative is because: ① it should be (original $\hat{\theta}_{\hat{q}}$ biased)

② smaller sample size.

Solution 2: Simultaneous Inference.

$\forall \hat{q} \in \mathcal{Q}$. $P(\theta_{\hat{q}} \in \hat{CI}_{\hat{q}}) \geq P(\bigcap_{q\in\mathcal{Q}} \{\theta_q \in \hat{CI}_q\})$ Controlling this needs to construct $\hat{CI}_q$ for $\forall q \in \mathcal{Q}$

$\liminf_{n\to\infty} P(\theta_{\hat{q}} \in \hat{CI}_{\hat{q}}) \geq \liminf_{n\to\infty} P(\bigcap_{q\in\mathcal{Q}} \{\theta_q \in \hat{CI}_q\}) \geq 1-\alpha$?? have to ~~use~~ do simultaneous inference for arbitrary selection procedure?

Construction procedure: Assumptions: ① uniform asymptotic linear representation, $\exists \{\psi_q(\cdot), \forall q \in \mathcal{Q}\}$

$\max_{q\in\mathcal{Q}} \left| \Psi_{n,q}^{-\frac{1}{2}} \left(\hat{\theta}_q - \theta_q - \frac{1}{n}\sum_{i=1}^{n} \psi_q(Z_i)\right) \right| = o_P\left(\frac{1}{\sqrt{n}}\right)$ $\frac{\sqrt{n}(\hat{\theta}_q - \theta_q)}{\Psi_{n,q}} \approx \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_q(Z_i)\right)}{\Psi_{n,q}}$

② Some Conditions guarantee $\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_q(Z_i)}{\Psi_{n,q}} \xrightarrow{d} N(0,1)$.

($\{Z_i\}$ needs to be independent / weakly dependent).

$\Rightarrow (\sqrt{n}\Psi_{n,q}^{-\frac{1}{2}}(\hat{\theta}_q - \theta_q) : q \in \mathcal{Q}) \xrightarrow{d} (G_q : q \in \mathcal{Q}) \sim N(0, \underset{\sim}{R})$ $\operatorname{diag}(R) = 1$.

$\max |\cdot| \longrightarrow \max_q |G_q|$ correlation between q unknown.

$\Rightarrow \lim_{n\to\infty} P(\max_{q\in\mathcal{Q}} |\sqrt{n}\,\Psi_{n,q}^{-\frac{1}{2}}(\hat{\theta}_q - \theta_q)| \leq K_\alpha) = 1-\alpha$ K_α: upper α quantile of $\max_q |G_q|$

$\Rightarrow \hat{CI}_q = [\hat{\theta}_q - K_\alpha\sqrt{\Psi_{n,q}/n}, \hat{\theta}_q + K_\alpha\sqrt{\Psi_{n,q}/n}]$ K_α, $\Psi_{n,q}$ unknown.

need to approximate K_α, $\hat{\Upsilon}_{nq}$ using bootstrap.

$\widehat{CI}_q = [\hat{\theta}_q - \hat{K}_\alpha \frac{\hat{\Upsilon}_{nq}^{\frac{1}{2}}}{\sqrt{n}}, \hat{\theta}_q + \hat{K}_\alpha \frac{\hat{\Upsilon}_{nq}^{\frac{1}{2}}}{\sqrt{n}}]$ more conservative $\frac{\hat{K}_\alpha}{z_{\frac{\alpha}{2}}} \geq 1$

$\widehat{CI}_q^{unadj} = [\hat{\theta}_q - z_{\frac{\alpha}{2}} \frac{\hat{\Upsilon}_{nq}^{\frac{1}{2}}}{\sqrt{n}}, \hat{\theta}_q + z_{\frac{\alpha}{2}} \frac{\hat{\Upsilon}_{nq}}{\sqrt{n}}]$ ~~grow~~ ratio grows with dimension. (dim↑ conservative↑)

$\hat{K}_\alpha$ depends on correlation.

pros: ① arbitrary selection procedure (graphical / repeated / multiple model)

② Change to block bootstrap: apply to dependent data (?? will CLT hold)

③ better selection result on larger sample.

Cons: ① might be too conservative ~~espec~~ (but not for arbitrary selection method)

② still only handles weakly dependent data.

③ ~~m~~ need to do calculation on all $q \in \mathcal{Q}$.

④ $\mathcal{Q}$ needs to be decided before data exploration.

⑤ Too conservative if ~~$\$$~~ $P(\hat{q} \in \mathcal{Q}_0 \subset \mathcal{Q})$ is large & $\mathcal{Q}_0$ is small in size.

Assumptions might not be needed ~~so~~ for some Gaussian models.

Solution 3: Conditional Selective Inference.

interested in: $\forall q$, $\liminf_{n\to\infty} P(\theta_q \in \widehat{CI}_q \mid \hat{q} = q) \geq 1-\alpha$

Idea: with ~~Gaussian assumption~~ (CLT), condition on subset of span$(\{z_i\})$ that decides the selection q.

Assumptions: 1. $\exists$ random vector $D_{n,q} \in R^{d_D}$ s.t. it decides selection: $\{\hat{q} = q\} \equiv \{D_{n,q} \preceq 0\}$

2. nonzero denominator: $\liminf_{n\to\infty} P(\hat{q} = q) = \liminf_{n\to\infty} (D_{n,q} \preceq 0) > 0$ (often fails, but can be dropped).

3. $\exists$ vec $\mu_{n,q} \in R^{d_D}$ & cov mat Ω_q s.t.

$$\begin{bmatrix} \sqrt{n}(\hat{\theta}_q - \theta_q) \\ D_{n,q} - \mu_{n,q} \end{bmatrix} \xrightarrow{d} \begin{bmatrix} G_{\theta q} \\ G_{Dq} \end{bmatrix} \sim N(0, \Omega_q) \overset{\Omega_q \text{ P.D.}}{\Longrightarrow} \sup_C \left| P\left(\begin{bmatrix} \sqrt{n}(\hat{\theta}_q - \theta_q) \\ D_{nq} - \mu_{nq} \end{bmatrix} \in C\right) - P\left(\begin{bmatrix} G_{\theta q} \\ G_{Dq} \end{bmatrix} \in C\right) \right| \to 0$$

need this uniformity.

4. Ω_q estimable (for later use)

$$\hat{\Omega}_q = \begin{pmatrix} \hat{W}_q^2 & \hat{\Omega}_{\theta D} \\ \hat{\Omega}_{D\theta} & \hat{\Omega}_{DD} \end{pmatrix} \xrightarrow{P} \Omega_q = \begin{pmatrix} W_q^2 & \Omega_{\theta D} \\ \Omega_{D\theta} & \Omega_{DD} \end{pmatrix}$$

Algorithm: not copied here.

intuitive understanding: use $\hat{\theta}_q$, $\hat{\Omega}_q$ to construct an empirical conditional distribut and build CI on it.

Improvements: ① Data carving: randomized $D_{n,q} \Rightarrow$ bdd expected CI length.

② Combine simul... & selective inference $\Rightarrow$ shorter length.

pros: ① ~~ste~~ selection on full data.

② computationally easier.

③ result can vary in naive CI (short) to data splitting CI (larger).

Cons: ① $\{\hat{q} = q\} \equiv \{D_{n,q} \preceq 0\}$ too strong. (restriction on selection).

② Theoretical analysis ≠ in every new implementation.

③ vanilla version may yields infinite width of CI.

Uniform Validity

$\liminf_{n\to\infty} \inf_{P\in \mathcal{P}^{\otimes n}} P(\theta_{\hat{q}} \in \widehat{CI}_{\hat{q}}) \geq 1-\alpha$

.... $P(\theta_{\hat{q}} \in \widehat{CI}_{\hat{q}} \mid \hat{q}=q)$

uniform for $P \in \mathcal{P}^{\otimes n}$ hold for all 3 methods with condition on $\mathcal{P}^{\otimes n}$

this is done on partial model.

Impossibility result for simultaneous & selective inference.

can't get uniform estimation because the inference is done on β_0 : full model parameter.