PredSet-1Step: Distribution-Free Prediction Sets Adaptive to Unknown Covariate Shift

Hongxiang Qiu, Edgar Dobriban, Eric Tchetgen Tchetgen

Department of Statistics & Data Science, The Wharton School, University of Pennsylvania

Wharton UNIVERSITY of PENNSYLVANIA

Background

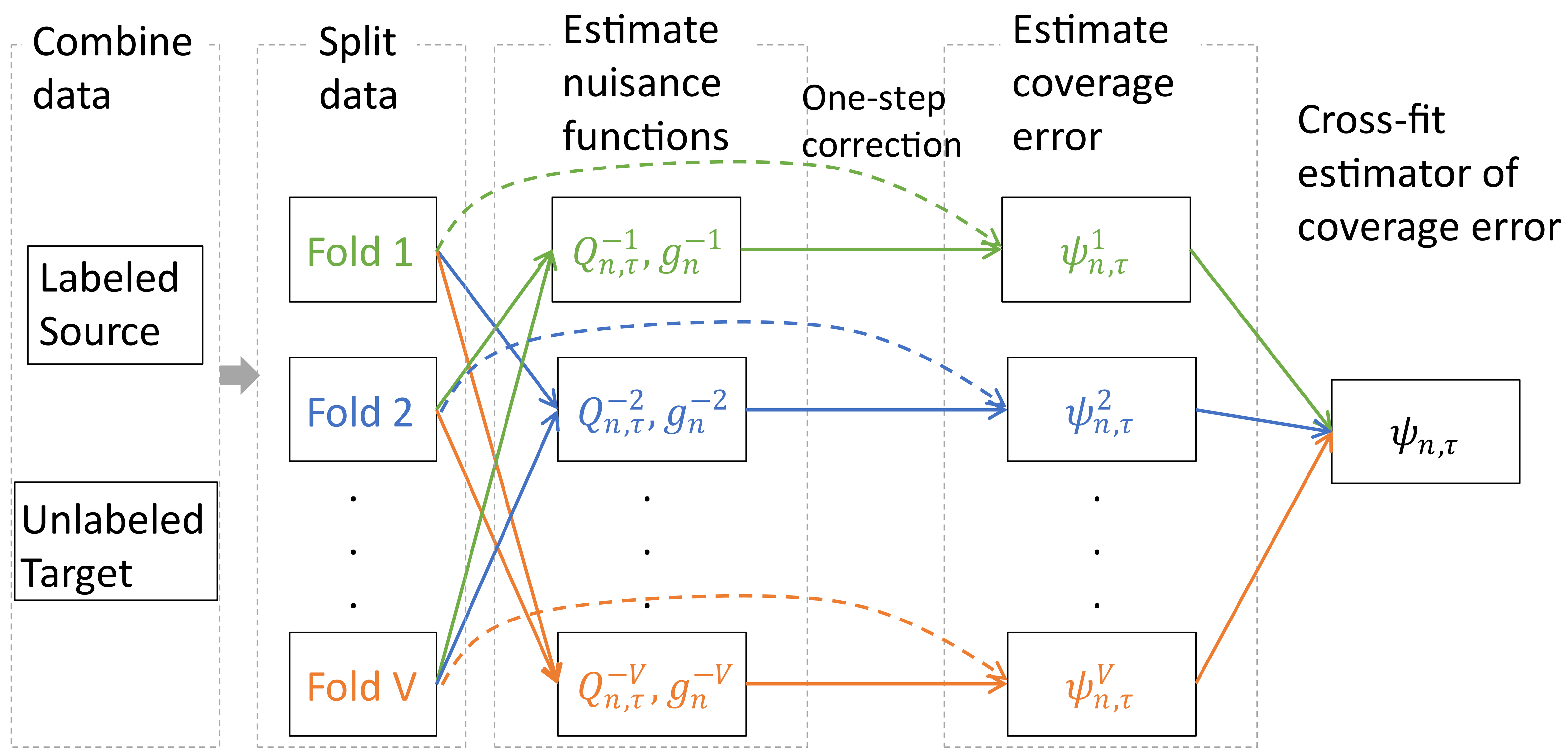
- Uncertainty quantification for prediction is a key issue in modern statistics: prediction sets, tolerance regions, conformal inference, etc
- Data: n observations consisting of Population indicator: source ($A=1$), target ($A=0$); Covariate: X; outcome/label to be predicted: Y
- Goal: *Asymptotically Probably Approximately Correct (APAC)* prediction set $\hat{C}$ in the target population:

$$\Pr\big(\Pr(Y \notin \hat{C}(X) \mid \hat{C}, A=0) \leq \alpha_{\text{error}}\big) \geq 1-\alpha_{\text{conf}} - o(1)$$

- Interpretation: with high confidence approaching $1-\alpha_{\text{conf}}$, the coverage error of $\hat{C}$ is below α_{error}
- Y missing in target population ($A=0$)
- Key assumption: *covariate shift* $p(Y \mid X, A=0) = p(Y \mid X, A=1)$
- Connection with causal inference & missing data: equivalent to missing at random & no unmeasured confounding
- Applications: predict individual treatment effect (ITE) $Y(1)-Y(0)$; predict Y in a different target population
- A key quantity: likelihood ratio

$$w_0(x) := \frac{p(X=x \mid A=0)}{p(X=x \mid A=1)}$$

- Previous PAC prediction set works assume *known* w_0 (ref in preprint); lack methods when w_0 is *unknown*



Procedure of cross-fit one-step estimator of coverage error corresponding to threshold τ

Prediction set coverage error estimator

- Fixed arbitrary scoring function $s(X,Y)$
  E.g., estimated $\Pr(Y \mid X)$ or $p(Y \mid X)$ using separate source population data
- Candidate prediction set: $C_\tau(x) = \{y: s(x,y) \geq \tau\}$ (threshold τ in finite grid $\mathcal{T}_n$)
- Under nonparametric model, coverage error $\psi_{0,\tau} := \Pr(Y \notin C_\tau(X) \mid A=0)$ in target population is pathwise differentiable with gradient (similar to ATT)

$$(a,x,y) \mapsto \frac{a}{1-\gamma_0}\frac{1-g_0(x)}{g_0(x)}\big[I\big(y \notin C_\tau(x)\big) - Q_{0,\tau}(x)\big] + \frac{1-a}{1-\gamma_0}\big[Q_{0,\tau}(x) - \psi_{0,\tau}\big]$$

$Q_{0,\tau}: x \mapsto \Pr(Y \notin C_\tau(X) \mid X=x)$ is the conditional coverage error;
$g_0: x \mapsto \Pr(A=1 \mid X=x)$ is the propensity score; $\gamma_0 := \Pr(A=1)$

- Cross-fit one-step corrected estimator $\psi_{n,\tau}$: split data into V folds, for each fold v,
  1. estimate $(Q_{0,\tau}, g_0)$ with $(Q_{n,\tau}^{-v}, g_n^{-v})$ using data out of fold v
  2. estimator with one-step correction based on gradient:

$$\psi_{n,\tau}^{v} = \frac{\sum_{i \in \text{fold } v}\left\{(1-A_i)Q_{n,\tau}^{-v}(X_i) + A_i\frac{1-g_n^{-v}(X_i)}{g_n^{-v}(X_i)}\left[I\big(Y_i \notin C_\tau(X_i)\big) - Q_{n,\tau}^{-v}(X_i)\right]\right\}}{\sum_{i \in \text{fold } v}(1-A_i)}$$

Average over folds: $\psi_{n,\tau} = \sum_v \psi_{n,\tau}^{v}/V$

- Corrected estimator $\psi_{n,\tau}$ is asymptotically normal: $\sqrt{n}(\psi_{n,\tau} - \psi_{0,\tau}) \xrightarrow{d} N(0, \sigma_\tau^2)$

Threshold selection

- Wald $(1-\alpha_{\text{conf}})$-confidence upper bound for coverage error $\psi_{0,\tau}$

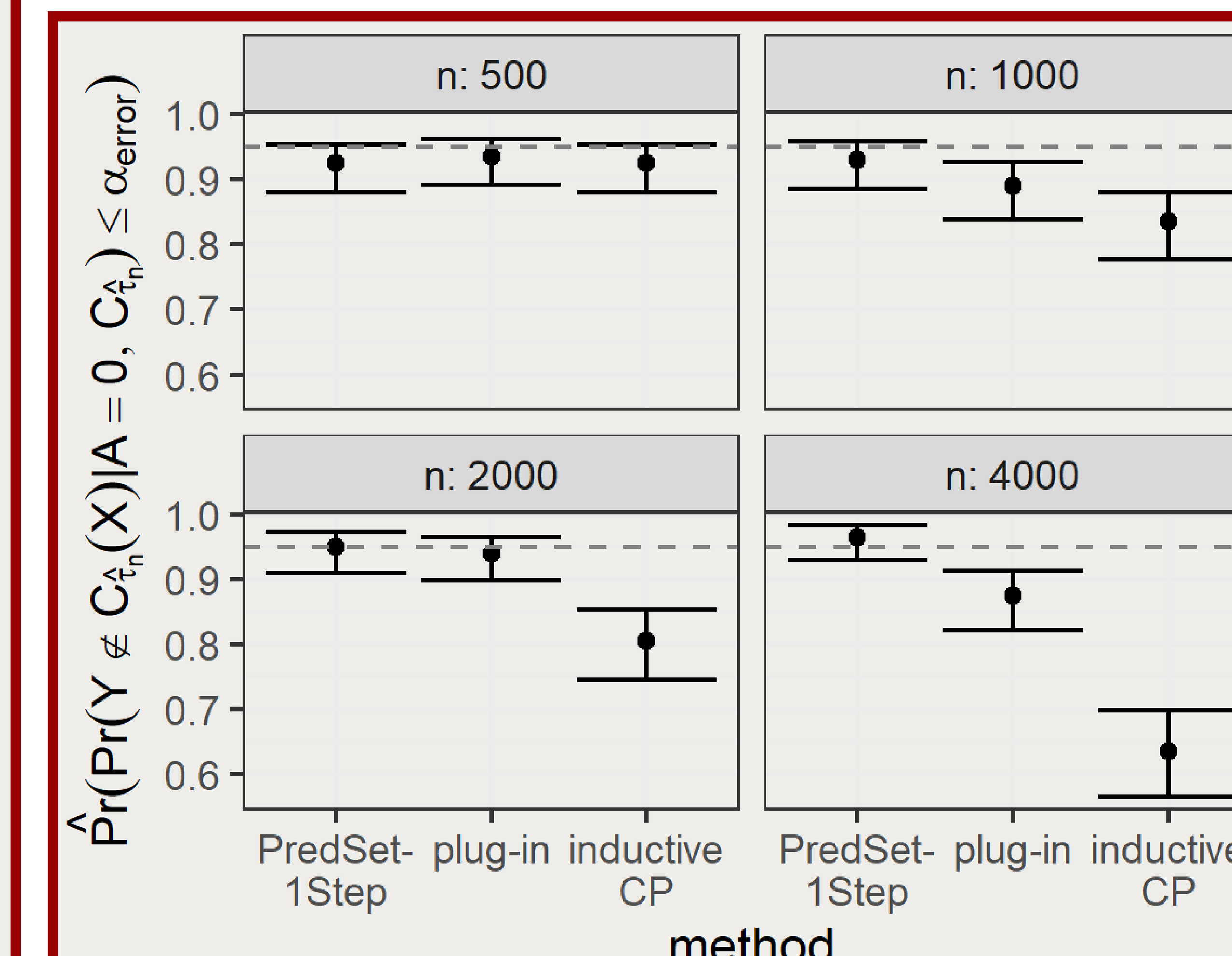
$$\lambda_n(\tau) = \psi_{n,\tau} + z_{\alpha_{\text{conf}}} \times \frac{\sigma_{n,\tau}}{\sqrt{n}}$$

- Select the max threshold in $\mathcal{T}_n$ such that all confidence upper bounds for smaller thresholds are below α_{error}:

$$\hat{\tau}_n := \max\{\tau \in \mathcal{T}_n : \lambda_n(\tau') < \alpha_{\text{error}}, \ \forall \tau' \in \mathcal{T}_n \text{ such that } \tau' \leq \tau\}$$

- $C_{\hat{\tau}_n}$ is APAC in the target population
- The $o(1)$ term is of order

$$n^{1/4}\|g_n^{-v} - g_0\|_2^{1/2}\|Q_{n,\tau}^{-v} - Q_{0,\tau}\|_2^{1/2}$$



Simulation: prob of coverage error $\leq \alpha_{\text{error}}$ (with CI)
Dashed line: target confidence level $1-\alpha_{\text{conf}}$
plug-in: PredSet-1Step without one-step correction
inductive CP: inductive (split) conformal prediction assuming no covariate shift

arXiv preprint:

Supported in part by NSF DMS, NIH, Analytics at Wharton

R package: