Distribution-Free Prediction Sets
Adaptive to Unknown Covariate Shift

Hongxiang Qiu, Edgar Dobriban, Eric Tchetgen Tchetgen

Department of Statistics, The Wharton School, University of Pennsylvania
Motivation

- Great advances in prediction using machine learning
- **Prediction sets with coverage guarantees** are useful to quantify uncertainty of prediction
- One useful guarantee is *Probably Approximately Correct* (PAC):
  \[
  \Pr \left( \Pr \left( Y \notin \hat{C}(X) \mid \text{training data} \right) \leq \alpha_{\text{error}} \right) \geq 1 - \alpha_{\text{conf}}
  \]
  
- Interpretation: with high confidence level \(1 - \alpha_{\text{conf}}\) (probably), the coverage error rate of \(\hat{C}\) is below \(\alpha_{\text{error}}\) (approximately correct)
- Also termed “training-set conditional validity”
- Inductive conformal prediction outputs PAC prediction sets if all data come from the same population [Papadopoulos et al., 2002, Vovk, 2013, Park et al., 2020]
Motivation

- Challenge: in many applications, labeled training data are drawn from a different population from the target population.
- For example, labeled data from Africa but want to predict in USA.
- Common assumption: covariate shift (covariate distribution shifts; distribution of label/outcome given covariate remains same).
- Under covariate shift, we learn $Y \mid X$ using labeled data from source population and can extrapolate to target population.

Hongxiang Qiu, Edgar Dobriban, Eric Tchetgen

PAC prediction sets under covariate shift
Can we construct PAC prediction sets adaptive to covariate shift based on an arbitrary predictor under weak assumptions?
Motivation

- Can we construct PAC prediction sets adaptive to covariate shift based on an arbitrary predictor under weak assumptions?
- Previous literature: Sort of: require knowing exactly the covariate distribution shift [Park et al., 2021]
Motivation

- Can we construct PAC prediction sets adaptive to covariate shift based on an arbitrary predictor under weak assumptions?
- Previous literature: Sort of: require knowing exactly the covariate distribution shift [Park et al., 2021]
- What if this shift is unknown?
- Available data: i.i.d. from $P^0$
  - labeled data $(X, Y)$ from source population ($A = 1$), and
  - unlabeled data $(X, \cdot)$ from target population ($A = 0$)
Lemma

Suppose that \( X \) and \( Y \) are continuous. Under unknown covariate shift, if \( \hat{\mathcal{C}} \) is PAC, then under any data-generating distribution \( P^0 \) and for almost every \( y \),

\[
\Pr(y \notin \hat{\mathcal{C}}(X) \mid A = 0) \leq \alpha_{\text{error}} + \alpha_{\text{conf}}.
\]

Any PAC prediction set \( \hat{\mathcal{C}} \) is generally uninformative

- Consider \( X \perp \perp Y \) and \( Y \in \mathbb{R} \): might wish
  \( \hat{\mathcal{C}}(x) = (\hat{q}_{\alpha_{\text{error}}/2}, \hat{q}_{1-\alpha_{\text{error}}/2}) \), but it is impossible to be PAC

- The following \( \hat{\mathcal{C}} \) is PAC but useless

\[
\hat{\mathcal{C}}(x) = \begin{cases} \mathbb{R} & \text{with probability } 1 - \alpha_{\text{error}} \\ \emptyset & \text{with probability } \alpha_{\text{error}} \end{cases}
\]
Resort to asymptotic coverage guarantee

- **Asymptotically Probably Approximately Correct (APAC) guarantee** for prediction set $\hat{C}_n$:

$$\Pr\left(\Pr\left(Y \notin \hat{C}_n(X) \mid \text{training data}\right) \leq \alpha_{\text{error}}\right) \geq 1 - \alpha_{\text{conf}} - o(1)$$

as sample size $n \to \infty$.

- Interpretation: with high confidence level approaching $1 - \alpha_{\text{conf}}$, the coverage error rate of $\hat{C}_n$ is below $\alpha_{\text{error}}$.
Proposed method: PredSet-1Step

- Given an arbitrary scoring function $s$, consider candidate prediction sets $C_{\tau} : x \mapsto \{ y : s(x,y) \geq \tau \}$

- Examples of $s(x,y)$: estimated $\Pr(Y = y \mid X = x)$ or $f(Y = y \mid X = x)$ from held-out labeled data; $-|y - \hat{y}(x)|$ for a predictor $\hat{y}$ trained from held-out labeled data

- Using semiparametric efficiency theory, we construct an asymptotically efficient estimator (cross-fit one-step corrected estimator) $\hat{\psi}_{n, \tau}$ of the coverage error of $C_{\tau}$ in the target population:

\[ \Pr(Y \notin C_{\tau}(X) \mid A = 0) =: \Psi_{\tau}(P^0) \]

- Construct a $(1 - \alpha_{\text{conf}})$-confidence upper bound $\lambda_n(\tau)$ for $\Psi_{\tau}(P^0)$

- Select a threshold $\hat{\tau}_n$ from a grid $\mathcal{T}_n$ based on $\lambda_n(\tau)$
Flowchart of cross-fit one-step corrected estimator

Combine data

Labeled Source

Unlabeled Target

Split data

Fold 1

Fold 2

Fold V

Estimate nuisance functions

\( \hat{\mathcal{E}}_{n,\tau}^{-1}, \hat{g}_{n}^{-1} \)

\( \hat{\mathcal{E}}_{n,\tau}^{-2}, \hat{g}_{n}^{-2} \)

\( \hat{\mathcal{E}}_{n,\tau}^{-V}, \hat{g}_{n}^{-V} \)

One-step correction

Estimate coverage error

\( \hat{\psi}_{n,\tau}^{1} \)

\( \hat{\psi}_{n,\tau}^{2} \)

\( \hat{\psi}_{n,\tau}^{V} \)

Cross-fit estimator of coverage error

\( \hat{\psi}_{n,\tau} \)
Cross-fit one-step corrected estimator

1. Randomly split entire data set into $V$ folds with index sets $I_v$ ($v = 1, \ldots, V$)
2. For each fold $v$, estimate nuisance functions $(\mathcal{E}_{0,\tau}, g_0)$ with $(\hat{\mathcal{E}}_{n,\tau}, \hat{g}_n)$ using data out of fold $v$

$$
\mathcal{E}_{0,\tau}(x) := \Pr(Y \notin C_{\tau}(X) \mid X = x, A = 1)
$$

$$
g_0(x) := \Pr(A = 1 \mid X = x)
$$

3. Let $\hat{\gamma}_n^v$ be the empirical proportion of $A = 1$ in fold $v$ (estimator of $\Pr(A = 1)$)
4. For each fold $v$, compute one-step corrected estimator

$$\hat{\psi}_{n,\tau} := \frac{\sum_{i \in I_v} (1 - A_i) \hat{C}_{n,\tau}(X_i)}{\sum_{i \in I_v} (1 - A_i)}$$

sample analogue of $\Psi_\tau(P^0)$

$$+ \frac{1}{|I_v|} \sum_{i \in I_v} \frac{A_i}{1 - \hat{\gamma}_n^V} \frac{1 - \hat{g}_{n,\tau}^V(X_i)}{\hat{g}_{n,\tau}^V(X_i)} \left[ \mathbb{1}(Y_i \notin C_\tau(X_i)) - \hat{C}_{n,\tau}(X_i) \right].$$

one-step correction

5. Average over folds: $\hat{\psi}_{n,\tau} := \frac{1}{n} \sum_{V=1}^V |I_v| \hat{\psi}_{n,\tau}^V$. 
Asymptotic efficiency

**Theorem (Informal)**

Under conditions, \( \hat{\psi}_{n,\tau} \) is an asymptotically efficient estimator of \( \Psi_{\tau}(P^0) \) and

\[
\sqrt{n}(\hat{\psi}_{n,\tau} - \Psi_{\tau}(P^0)) \xrightarrow{d} N\left(0, \sigma_{0,\tau}^2\right)
\]

with \( \sigma_{0,\tau}^2 = \mathbb{E}_{P^0}[D_{\tau}(P^0)(O)^2] \) where \( D_{\tau}(P^0) \) is the efficient influence function.

\((1 - \alpha_{\text{conf}})\)-Wald confidence upper bound \( \lambda_n(\tau) \) for \( \Psi_{\tau}(P^0) \):

\[
\lambda_n(\tau) = \hat{\psi}_{n,\tau} + z_{\alpha_{\text{conf}}} \frac{\hat{\sigma}_{n,\tau}}{\sqrt{n}}
\]
Selection of threshold

Select the threshold

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

The largest candidate threshold such that all \( \lambda_n \) on the left hand side are below \( \alpha_{\text{error}} \). (Similar to Bates et al. [2021])

**Theorem (Informal)**

*Under conditions, \( C_{\hat{\tau}_n} \) is APAC.*
Illustration of threshold selection

\begin{align*}
(1 - \alpha_{\text{conf}}) - \text{level confidence upper bound} \\
\alpha_{\text{error}}
\end{align*}

Candidate threshold $\tau$

Hongxiang Qiu, Edgar Dobriban, Eric Tchetgen

PAC prediction sets under covariate shift
Simulation result

\[ \hat{P}(Y \notin C_{\hat{A}, n}) \leq \alpha_{\text{error}} \]

Method

- PredSet-1Step
- plug-in
- Inductive CP

\[ \hat{P}(Y \notin C_{\hat{A}, n}(X)|A = 0, C_{\hat{A}, n}) \]

Data sizes: n: 500, n: 1000, n: 2000, n: 4000

Hongxiang Qiu, Edgar Dobriban, Eric Tchetgen (Statistics, UPenn)
Analysis of HIV risk prediction data in South Africa

- **Y**: HIV infection
- Source population: urban and rural communities
- Target population: peri-urban communities with community HIV treatment coverage $\leq 15\%$
- Target coverage error $\alpha_{\text{error}} = 5\%$ (coverage $\geq 95\%$)
- Target confidence level $1 - \alpha_{\text{conf}} = 95\%$

| Method          | Empirical coverage | 95% CI of coverage  |
|-----------------|--------------------|---------------------|
| PredSet-1Step   | 95.98%             | 94.83%–96.89%       |
| Inductive CP    | 91.89%             | 90.35%–93.20%       |
Conclusion

- Prediction sets are useful to quantify uncertainty of prediction
- Unknown covariate shift is a common challenge
- We propose a method, PredSet-1step, to construct APAC prediction sets adaptive to unknown covariate shift
Acknowledgment

Collaborators:

Edgar Dobriban
Eric Tchetgen Tchetgen

Funding supported by NSF, NIH and Analytics at Wharton.

arXiv preprint: https://arxiv.org/abs/2203.06126 (will update soon)

R package available on Github.
Thank you!
S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*, 2021.

L. Lei and E. J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 83(5):911–938, 2021. ISSN 14679868. doi: 10.1111/rssb.12445. URL http://arxiv.org/abs/2006.06138.

H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

S. Park, S. Li, I. Lee, and O. Bastani. Pac confidence predictions for deep neural network classifiers. *arXiv preprint arXiv:2011.00716*, 2020.

S. Park, E. Dobriban, I. Lee, and O. Bastani. Pac prediction sets under covariate shift, 2021.

R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32, 2019. ISSN 10495258.
V. Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, volume 25, pages 475–490. PMLR, 2013. doi: 10.1007/s10994-013-5355-6.
Without one-step correction, the naïve estimator $\Psi_\tau(P_{n,\tau}^v)$ is generally asymptotically inefficient.
More technical results

Key condition for asymptotic efficiency of $\hat{\psi}_{n,\tau}$:

$$\|\hat{\mathcal{C}}_{n,\tau} - \mathcal{C}_{0,\tau}\| \|\hat{\mathcal{G}}_{n} - g_0\| = o_p(n^{-1/2})$$

Rate of $o(1)$ term:

**Theorem (Informal)**

*If the asymptotic variance is nonzero, the coverage probability $\Pr(\Psi_{\tau}(P^0) \leq \lambda_n(\tau))$ equals*

$$1 - \alpha_{\text{conf}} - O\left(n^{1/4} \mathbb{E}_{P^0}[\|\hat{\mathcal{C}}_{n,\tau} - \mathcal{C}_{0,\tau}\|\|\hat{\mathcal{G}}_{n} - g_0\|]^{1/2}\right)$$

The rate of the $o(1)$ term is mainly determined by the product of convergence rates of the two nuisance function estimators.