Finite-sample Efficient Conformal Prediction

Yachong Yang and Arun Kuchibhotla
University of Pennsylvania Carnegie Melon University

What is conformal prediction?

- Conformal prediction is a generic method for finite-sample valid distribution-free prediction and can be applied with most machine learning algorithms to yield valid prediction regions.
- Given i.i.d. pairs $(X_i, Y_i) \sim P, i = 1, \ldots, N$, for a distribution P on $\mathcal{X} \times \mathbb{R}$ (e.g., $\mathcal{X} = \mathbb{R}^d$)

  Goal. Build a prediction set $\widehat{C}_N : \mathcal{X} \to \mathcal{P}(\mathbb{R})$, such that for new i.i.d. pair (X_{N+1}, Y_{N+1}) :

$$\mathbb{P}\left(Y_{N+1} \in \widehat{C}_N\left(X_{N+1}\right)\right) \geq 1 - \alpha,$$

  (where the probability is over all $N + 1$ pairs).

Split Conformal Prediction

- Split the sample $(X_i, Y_i), 1 \leq i \leq N$ into two parts each with $n = N/2$ observations.
- Compute the estimator $\widehat{\mu}_m(\cdot)$ based on the first split.
- Let $\tilde{q}_{n,\alpha}$ denote the $(1-\alpha)(1+1/n)$-th quantile of the residuals $|Y_i - \widehat{\mu}_m(X_i)|$, on the second split.
- $\tilde{C}_N := \{(x, y) : y \in [\widehat{\mu}_m(x) - \tilde{q}_{n,\alpha}, \widehat{\mu}_m(x) + \tilde{q}_{n,\alpha}]\}$

By exchangeability, we have finite-sample coverage:

$$\mathbb{P}\left((X_{N+1}, Y_{N+1}) \in \tilde{C}_N\right) \geq 1 - \alpha.$$

This is valid regardless of whether $\widehat{\mu}_m$ is a consistent estimator. In practice, the coverage is close to $1 - \alpha$.

Problem Formulation

- Suppose $\widehat{\mu}_m^1(\cdot), \ldots, \widehat{\mu}_m^K(\cdot)$ are K different estimators with different tuning parameters. They can be from different estimation methods such as LASSO, random forest.
- ⋆ The question we discuss is how to select $\widehat{k} \in \{1, 2, \ldots, K\}$ and construct a valid prediction region with width close to the smallest.
- We will call coverage guarantee as validity and smallest width property as efficiency.

Efficient First Conformal Prediction (EFCP)

Input: Data $(X_i, Y_i), 1 \leq i \leq N$, coverage probability $1 - \alpha$, and K estimation methods for $\mu(\cdot)$.

1. Randomly split the data into three parts each with $n := N/3$ observations.
2. Fit the estimators $\widehat{\mu}_1(\cdot), \ldots, \widehat{\mu}_K(\cdot)$ on the first split of the data.
3. $T_{\alpha,k} := (1 - \alpha)(1 + 1/n)$-th quantile of $|Y_i - \widehat{\mu}_k(X_i)|$, residuals in the second split of the data. The corresponding conformal prediction region is

$$\widehat{C}_k := \{(x, y) : y \in [\widehat{\mu}_k(x) \pm T_{\alpha,k}]\}.$$

4. Set $\widehat{k}$ as the minimizer of $T_{\alpha,k}$ over $1 \leq k \leq K$.
5. $T^*_{\alpha,\widehat{k}} := (1 - \alpha)(1 + 1/n)$-th quantile of $|Y_i - \widehat{\mu}_{\widehat{k}}(X_i)|$, residuals in the third split of the data.
6. **Output:**

$$\widehat{C}^{\texttt{VFCP}}_\alpha := \left\{(x, y) : y \in \left[\widehat{\mu}_{\widehat{k}}(x) \pm T^*_{\alpha,\widehat{k}}\right]\right\}.$$

Validity First Conformal Prediction (VFCP)

Input: Data $(X_i, Y_i), 1 \leq i \leq N$, coverage probability $1 - \alpha$, and K estimation methods for $\mu(\cdot)$.

1. Randomly split the data into two parts each with $n := N/2$ observations.
2. Fit the estimators $\widehat{\mu}_1(\cdot), \ldots, \widehat{\mu}_K(\cdot)$ on the first split.
3. $T_{\alpha,k} := (1 - \alpha)(1 + 1/n)$-th quantile of $|Y_i - \widehat{\mu}_k(X_i)|$, residuals in the second split of the data. The corresponding conformal prediction region is

$$\widehat{C}_k := \{(x, y) : y \in [\widehat{\mu}_k(x) \pm T_{\alpha,k}]\}.$$

4. Set $\widehat{k}$ as the minimizer of $T_{\alpha,k}$ over $1 \leq k \leq K$.
5. **Output:**

$$\widehat{C}^{\texttt{EFCP}}_\alpha := \widehat{C}_{\widehat{k}} = \{(x, y) : y \in [\widehat{\mu}_{\widehat{k}}(x) \pm T_{\alpha,\widehat{k}}]\}.$$

Results and comparison of VFCP and EFCP

	Coverage	Width
VFCP	$1 - \alpha - 0$	min-width $+ \mathfrak{C}\sqrt{\frac{\log(K/\delta)}{N}}$
EFCP	$1 - \alpha - \mathfrak{C}\sqrt{\frac{\log(K)}{N}}$	min-width $+ 0$

1. VFCP requires three splits of the data while EFCP only requires two splits of the data.
2. Validity and efficiency of VFCP requires some continuity assumptions on the distributions, while EFCP only requires i.i.d. data.

Application: tuning-free ridge regression

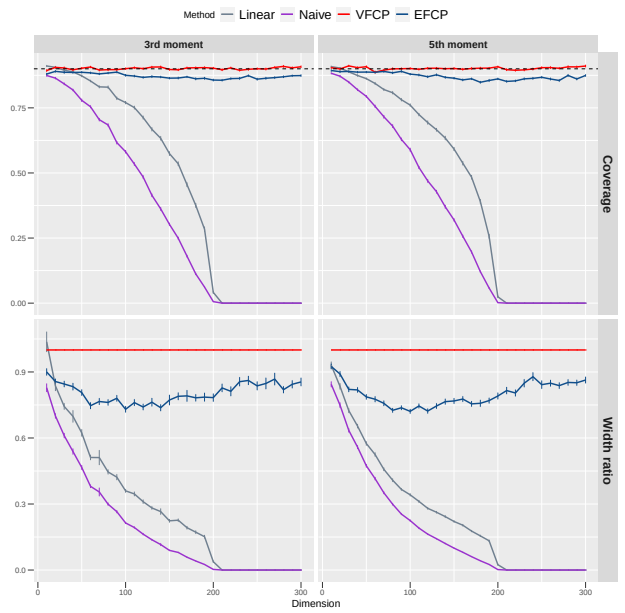

Figure 1: Ridge Regression with a Linear Model

Full paper and supplement (ArXiv):