Finite-sample Efficient Conformal Prediction

http://arxiv.org/abs/2104.13871

Yachong Elsa Yang, University of Pennsylvania
Joint work with Arun Kumar Kuchibhotla, Carnegie Mellon University

Table of contents

1. An Introduction to Conformal Prediction
2. Problem Formulation
3. Validity First Conformal Prediction
4. Efficiency First Conformal Prediction
5. Simulations
6. Conclusions

An Introduction to Conformal Prediction

Goal

Given i.i.d. pairs $(X_i, Y_i) \sim P, i = 1, \ldots, N$, for a distribution P on

$$\mathcal{X} \times \mathbb{R} \quad \left(\text{e.g., } \mathcal{X} = \mathbb{R}^d\right)$$

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).

Can we do so without any assumptions on the distribution P, and hope for something nontrivial?

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.
- Define

  $$\tilde{C}_N = \{(x, y) : \ y \in [\widehat{\mu}_m(x) - \tilde{q}_{n,\alpha}, \widehat{\mu}_m(x) + \tilde{q}_{n,\alpha}]\}$$

  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 $1 - \alpha$.

Example

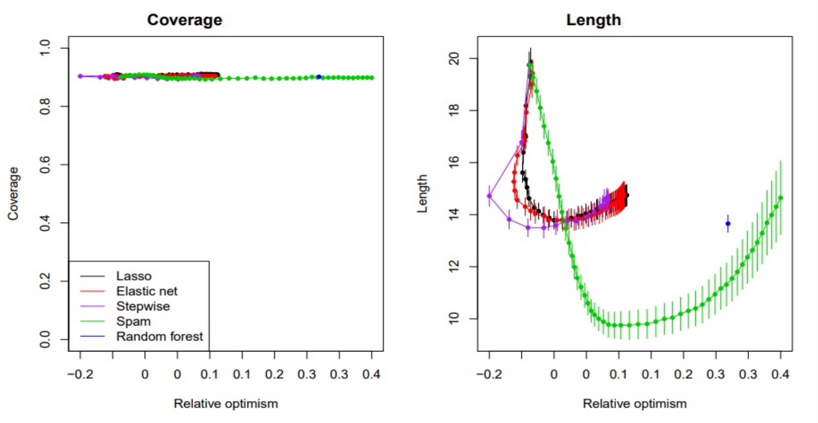


Figure 1: Conformal prediction intervals across several base estimators.[1]

[1] Figure 1 of Lei et al. JASA 2018.

Problem Formulation

Formalizing the problem

- 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.

Validity First Conformal Prediction

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 $\mu(\cdot)$.

$$n := N/3.$$

1. Randomly split the data into three parts each with n observations.
2. Fit the estimators $\widehat{\mu}_1(\cdot), \ldots, \widehat{\mu}_K(\cdot)$ on the first split of the data.
3. Define $T_{\alpha,k}$ as the $(1-\alpha)(1+1/n)$-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. Define $T^*_{\alpha,\widehat{k}}$ as the $(1-\alpha)(1+1/n)$-th quantile of $|Y_i - \widehat{\mu}_{\widehat{k}}(X_i)|$, residuals in the third split of the data.
6. **Output:** $\widehat{C}_\alpha^{\texttt{VFCP}} := \left\{(x,y) :\ y \in \left[\widehat{\mu}_{\widehat{k}}(x) \pm T^*_{\alpha,\widehat{k}}\right]\right\}$.

Validity and Efficiency of VFCP

Theorem (Validity of VFCP)

The prediction interval $\widehat{C}_\alpha^{\mathrm{VFCP}}$ *satisfies*

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

whenever $(X_i, Y_i), 1 \leq i \leq N+1$ *form an exchangeable sequence of random variables.*

Theorem (Efficiency of VFCP)

Suppose $(X_i, Y_i), 1 \leq i \leq N$ *are independent and identically distributed. Fix* $\delta \in [0, 1]$*, under some assumptions, with probability at least* $1 - \delta$*,*

$$\mathrm{Width}(\widehat{C}_\alpha^{\mathrm{VFCP}}) \ \leq \ \min_{1 \leq k \leq K} \mathrm{Width}(\widehat{C}_k) + \mathfrak{C}\sqrt{\frac{\log(4K/\delta)}{N}}.$$

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 $\mu(\cdot)$.

$$n := N/3.$$

1. Randomly split the data into three parts each with n observations.
2. Fit the estimators $\widehat{\mu}_1(\cdot), \ldots, \widehat{\mu}_K(\cdot)$ on the first split of the data.
3. Define $T_{\alpha,k}$ as the $(1-\alpha)(1+1/n)$-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. Define $T^*_{\alpha,\widehat{k}}$ as the $(1-\alpha)(1+1/n)$-th quantile of $|Y_i - \widehat{\mu}_{\widehat{k}}(X_i)|$, residuals in the third split of the data.
6. **Output:** $\widehat{C}_\alpha^{\texttt{VFCP}} := \left\{(x,y) : \, y \in \left[\widehat{\mu}_{\widehat{k}}(x) \pm T^*_{\alpha,\widehat{k}}\right]\right\}.$

Efficiency First Conformal Prediction

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 $\mu(\cdot)$.

$$n := N/2.$$

1. Randomly split the data into two parts each with n observations.
2. Fit the estimators $\widehat{\mu}_1(\cdot), \ldots, \widehat{\mu}_K(\cdot)$ on the first split of the data.
3. Define $T_{\alpha,k}$ as the $(1-\alpha)(1+1/n)$-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:** $\widehat{C}_\alpha^{\mathtt{EFCP}} := \widehat{C}_{\widehat{k}} = \{(x, y) : \ y \in [\widehat{\mu}_{\widehat{k}}(x) \pm T_{\alpha,\widehat{k}}]\}$.

Validity and Efficiency of EFCP

EFCP is basically data snooping where we are using the same data to both tune the parameter and report inference. But surprisingly,

Theorem (Validity of EFCP)

If $(X_i, Y_i), 1 \leq i \leq N+1$ *are independent and identically distributed, then the efficiency first conformal prediction region* $\widehat{C}_\alpha^{\mathrm{EFCP}}$ *satisfies*

$$\mathbb{P}\left((X_{N+1}, Y_{N+1}) \in \widehat{C}_\alpha^{\mathrm{VFCP}}\right) \;\geq\; 1 - \alpha - \frac{\sqrt{\log(2K)} + 2/3}{\sqrt{N}}.$$

Theorem (Efficiency of EFCP)
From the definition of $\widehat{k}$*, we have*

$$\mathrm{Width}(\widehat{C}_\alpha^{\mathrm{EFCP}}) \;\leq\; \min_{1\leq k\leq K} \mathrm{Width}(\widehat{C}_k) \;+\; 0.$$

Comparison of VFCP and EFCP

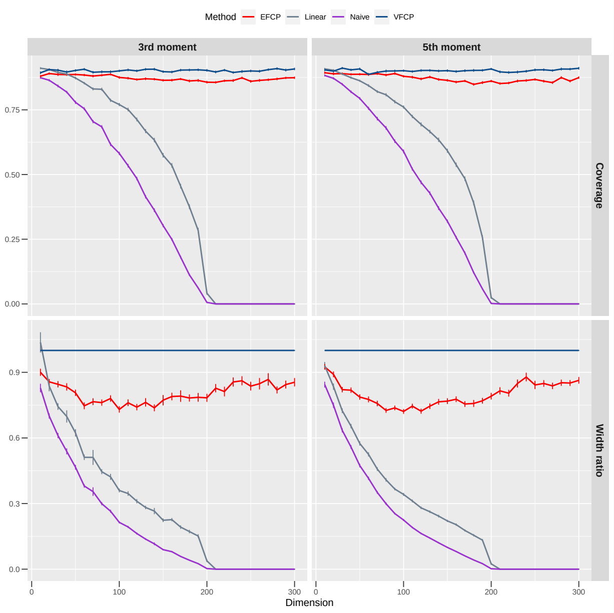
Table 1: Comparison of validity first conformal prediction (VFCP) and efficiency first conformal prediction (EFCP) in terms of coverage and width.

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

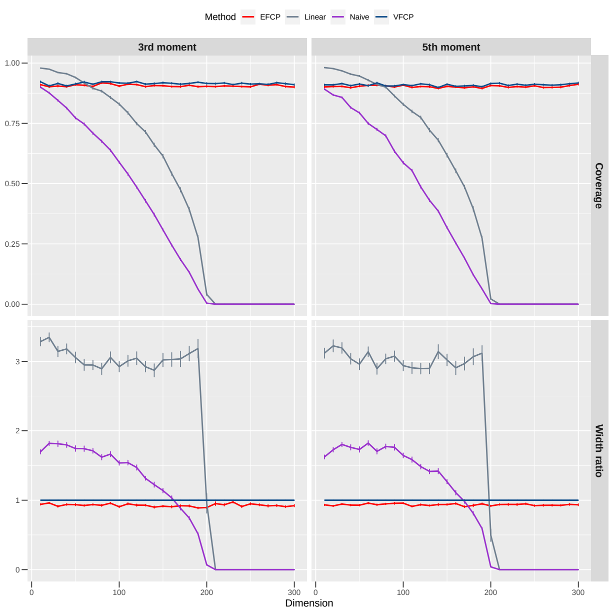
1. VFCP requires three splits of the data while EFCP only requires two splits of the data.
2. Validity and efficiency of VFCP requires some continuity assumptions on the distributions, while EFCP only requires i.i.d. data.

Simulations

Simulation 1: Ridge Regression with a Linear Model



Simulation 2: Ridge Regression with a Non-linear Model



Results

- VFCP has valid coverage while EFCP undercovers by around 2 percent across all dimensions.
- Compared with VFCP, EFCP has an improved efficiency ranging from 10 percent to 30 percent across all dimensions, with an average of over 20 percent improvement.
- The standard errors of the coverage for the methods are similar.
- Before computing the ratio of the widths, we find that the standard error of the widths of EFCP is, on average, over 30 percent smaller than that of VFCP over 100 repetitions.
- Neither the *Linear* method nor the *Naive* method gives valid coverage, this is expected because the two settings are heteroskedastic models.

Conclusions

Take home message

1. We have proposed two algorithms to find conformal prediction regions with the smallest widths among a class of methods;
2. The first method (VFCP) provides finite sample validity and approximate efficiency;
3. The second method (EFCP) gives finite sample efficiency and approximate validity;
4. Surprisingly we are not losing much by doing data snooping to tune the parameter and do inference on the same data with EFCP.
5. The work is done with split conformal method. It would be interesting to see if it can be applied to Jackknife and other conformal methods.

http://arxiv.org/abs/2104.13871

Thank You!

References

Lei, Robins, and Wasseerman (2013) Distribution-free prediction sets. *JASA.*

Lei and Wasseerman (2014) Distribution-free prediction bands for non-parametric regression. *JRSSB.*

Lei, G'Sell, Rinaldo, Tibshirani, and Wasserman (2018) Distribution-free predictive inference for regression. *JASA.*