Training-conditional coverage for distribution-free predictive inference

Michael Bian and Rina Foygel Barber

Kiho Park
parkkiho@uchicago.edu

1 Introduction

We say that $\widehat{C}_n$ is a distribution-free predictive interval if, for every distribution P on $\mathcal{X} \times \mathbb{R}$ it holds that

$$\mathbb{P}_{P^{n+1}}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha \tag{1}$$

In practice, we are often interested in the coverage rate for test points once we fit a regression algorithm to a particular training set. However, the guarantee in (1) does not directly address this. Let $\mathcal{D}_n = ((X_1, Y_1), \cdots, (X_n, Y_n))$ be the training data set. Then, define the miscoverage rate as a function of the training data:

$$\alpha_P(\mathcal{D}_n) = \mathbb{P}_P\left\{Y_{n+1} \notin \widehat{C}_n(X_{n+1}) \mid \mathcal{D}_n\right\}$$

Then, the guarantee in (1) can be re-written as

$$\mathbb{E}_{P^n}[\alpha_P(\mathcal{D}_n)] \leq \alpha$$

While this expectation is bounded, $\alpha_P(\mathcal{D}_n)$ may have high variance over the training data. In this paper, we examine the variability of $\alpha_P(\mathcal{D}_n)$ for popular distribution-free methods for arbitrary P and A. In particular, we seek to provide guarantees of the form

$$\mathbb{P}_{P^n}\{\alpha_P(\mathcal{D}_n) > \alpha + o(1)\} \leq o(1) \tag{2}$$

also known as a "Probably Approximately Correct" (PAC) predictive interval.

2 Background

Consider an algorithm $\mathcal{A}$ that maps datasets $\{(X_i, Y_i)\}$ to fitted regression functions $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$. Throughout, we will assume implicitly that $\mathcal{A}$ treats input data points symmetrically. In many practical settings, the distribution of the data is likely unknown, and the regression algorithm $\mathcal{A}$ may be a complex "black box" methods, and therefore it may be challenging to guarantee a particular error bound for $\widehat{\mu}(X_{n+1})$ as an estimator of the unseen response Y_{n+1}.

2.1 Distribution-free methods

2.1.1 Conformal prediction

The conformal prediction framework provides a mechanism for constructing prediction intervals in this challenging setting, with distribution-free coverage guarantees. While split conformal offers both computational efficiency and distribution-free coverage, its precision may suffer from the loss of sample size incurred by splitting the data set. In contrast, full conformal uses all the available training data for model fitting, but comes at a high computational cost.

2.1.2 Jackknife+ and CV+

The jackknife+ and CV+ methods proposed by Barber et al. (2021) offer a compromise between the computational efficiency of split conformal and the statistical efficiency of full conformal. The jackkife+ prediction interval satisfies

$$\mathbb{P}_{P^{n+1}}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - 2\alpha$$

for every distribution P on $\mathcal{X} \times \mathbb{R}$. The CV+ prediction interval satisfies

$$\mathbb{P}_{P^{n+1}}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - 2\alpha - \sqrt{2/n}$$

for every distribution P on $\mathcal{X} \times \mathbb{R}$.

2.2 Marginal or conditional validity

The predictive coverage bound achieved by split and full conformal, or the bounds for the jackknife+ and CV+ methods, are all marginal guarantees. This means that the probability is calculated over a random draw of both the training and test data. However, this may be unsatisfactory for practical purposes, in several ways.

Training-conditional coverage We may be interested in training-conditional coverage. For the split conformal method, Vovk (2012) [Proposition 2a] establishes training-conditional coverage through a Hoeffding bound:

Theorem 1 (Vovk (2012) Proposition 2a)**.** *Consider the split conformal method with sample size $n = n_0+n_1$. Then, for any distribution P and any $\delta \in (0, 0.5]$,*

$$\mathbb{P}_{P^n}\left\{\alpha_P(\mathcal{D}_n) \leq \alpha + \sqrt{\frac{\log(1/\delta)}{2n_1}}\right\} \geq 1 - \delta$$

In other words, the probability that a training set results in a significantly higher training-conditional miscoverage rate than the nominal rate, is vanishingly small under the split conformal method. No analogous finite-sample results are known for distribution-free prediction methods beyond split conformal. In this work, our goal will be to examine the finite-sample training-conditional coverage properties of distribution-free methods beyond split conformal.

Object-conditional or label-conditional coverage We may also be interested in coverage at a particular new test feature vector X_{n+1}. That is, we would like to show that the conditional coverage probability $\mathbb{P}_{P^{n+1}}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \mid X_{n+1} = x\right\}$ is $\geq 1-\alpha$, at least approximately. However, Vovk (2012) shows that this type of guarantee is impossible under any distribution P for which X is nonatomic. A third type of conditional guarantee is that of label-conditional coverage for the setting where the response Y is categorical, requiring $\mathbb{P}_{P^{n+1}}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}) \mid Y_{n+1} = y\right\} \geq 1 - \alpha$ for each category y. Both of these type of conditional guarantees are fundamentally very different from training-conditional coverage, and we will not address these further in this work.

3 Theoretical results

We find that a guarantee of the form (2) can also be shown for the K-fold CV+ method, but no such guarantees are possible for the full conformal or jackknife+ methods.

First, we consider the full conformal prediction method. In contrast to split conformal, it is impossible to guarantee training-conditional coverage for the full conformal method without further assumptions.

Theorem 2. *For any sample size $n \geq 2$ and any distribution P for which the marginal P_X is nonatomic, there exists a regression algorithm $\mathcal{A}$ such that the full conformal prediction method satisfies*

$$\mathbb{P}_{P^n}\left\{\alpha_P(\mathcal{D}_n) \geq 1 - n^{-2}\right\} \geq \alpha - 6\sqrt{\frac{\log n}{n}}$$

In other words, without placing assumptions on the distribution P and/or the algorithm $\mathcal{A}$, we cannot avoid the worst-case scenario. Next, for thejackknife+, the same worst-case result holds.

Theorem 3. *For any sample size $n \geq 2$ and any distribution P for which the marginal P_X is nonatomic, there exists a regression algorithm $\mathcal{A}$ such that the jackknife+ prediction interval satisfies*

$$\mathbb{P}_{P^n}\left\{\alpha_P(\mathcal{D}_n) \geq 1 - n^{-2}\right\} \geq \alpha - 6\sqrt{\frac{\log n}{n}}$$

In contrast, for CV+, we will now see that the lower bound on marginal coverage, which is $\gtrapprox 1-2\alpha$, can also be obtained as a training-conditional guarantee.

Theorem 4. *For any integers $K \geq 2$ and $m \geq 1$, and let $n = Km$. Suppose CV+ is run with K folds each of size m. Then, for any regression algorithm $\mathcal{A}$ and any distribution P, the K-fold CV+ method satisfies*

$$\mathbb{P}_{P^n}\left\{\alpha_P(\mathcal{D}_n) \leq 2\alpha + \sqrt{\frac{2\log(K/\delta)}{m}}\right\} \geq 1 - \delta$$

for any $\delta > 0$.

As long as the size of each fold m is large, the bound on $\alpha_P(\mathcal{D}_n)$ is approximately 2α. Comparing to the marginal result for the CV+ method, we see that the conditional coverage guarantee essentially matches the marginal coverage guarantee, and thus could not be improved.

4 Proofs

4.1 Proofs for split conformal and CV+

Both Theorem 1 and Theorem 4 can be proved as consequences of the following lemma.

Lemma 1. *Let $n \geq 2$ and choose a holdout set A with $\emptyset \subsetneq A \subsetneq [n]$. Let $\widehat{\mu}_{[n]\setminus A} = \mathcal{A}\left((X_i, Y_i) : i \in [n] \setminus A\right)$. Define*

$$p_A(x,y) = \frac{1}{|A|}\sum_{i\in A} \mathbb{1}\{|Y_i - \widehat{\mu}_{[n]\setminus A}(X_i)| \geq |y - \widehat{\mu}_{[n]\setminus A}(x)|\}$$

and

$$p_A^*(x,y) = \mathbb{P}_P\left\{|Y - \widehat{\mu}_{[n]\setminus A}(X)| \geq |y - \widehat{\mu}_{[n]\setminus A}(x)| \mid \widehat{\mu}_{[n]\setminus A}\right\}$$

Then $p_A^*(X_{n+1}, Y_{n+1})$ is a valid p-value conditional on the training data, i.e., $\mathbb{P}_P\{p_A^*(X_{n+1}, Y_{n+1}) \leq a \mid \mathcal{D}_n\} \leq a$ for all $a \in [0,1]$, almost surely over $\mathcal{D}_n$. Moreover, for any $\Delta \geq \sqrt{\frac{\log 2}{2|A|}}$,

$$\mathbb{P}_{P^n}\left\{\sup_{(x,y)\in\mathcal{X}\times\mathbb{R}} (p_A^*(x,y) - p_A(x,y)) > \Delta \mid \mathcal{D}_n\right\} \leq e^{-2|A|\Delta^2}$$

For the split conformal method, we use the event that implies

$$\begin{aligned}\alpha_P(\mathcal{D}_n) &\le \mathbb{P}_P\{p_{[n]\setminus[n_0]}(X_{n+1}, Y_{n+1}) \le \alpha \mid \mathcal{D}_n\} \\ &\le \mathbb{P}_P\{p^*_{[n]\setminus[n_0]}(X_{n+1}, Y_{n+1}) \le \alpha + \Delta \mid \mathcal{D}_n\} \le \alpha + \Delta\end{aligned}$$

For CV+ method, we use the event that implies

$$\begin{aligned}\alpha_P(\mathcal{D}_n) &\le \mathbb{P}_P\left\{\frac{1}{K}\sum_{k=1}^{K} p_{S_k}(X_{n+1}, Y_{n+1}) \le \alpha \mid \mathcal{D}_n\right\} \\ &\le \mathbb{P}_P\left\{\frac{1}{K}\sum_{k=1}^{K} p^*_{S_k}(X_{n+1}, Y_{n+1}) \le \alpha + \Delta \mid \mathcal{D}_n\right\} \le 2\alpha + 2\Delta\end{aligned}$$

4.2 Proofs for full conformal and jackknife+

Fix some large integer M, and partition $\mathcal{X} = A_0 \cup A_1 \cup \cdots \cup A_{M-1}$ where $\mathbb{P}_P\{X \in A_m\} = \frac{1}{M}$ for each $m = 0, \cdots, M-1$. Define a map $a(x) = m$ if $x \in A_m$ for $m = 0, \cdots, M-1$. Then,

$$a(X) \sim \mathrm{Unif}\{0, \cdots, M-1\}$$

under the distribution P. By extension, $\mathrm{mod}(\sum_{i=1}^n a(X_i), M) \sim \mathrm{Unif}\{0, \cdots, M-1\}$.

Define y_* as the $(1-n^{-2})$-quantile of Y under distribution P, and let

$$M_1 = \left\lfloor M\left(\alpha - \sqrt{\frac{2\log n}{n}} - \frac{2}{n}\right)\right\rfloor \approx \alpha M$$

Let $\mathcal{E}_{\max}$ be the event that $\max_{i\in[n]}|Y_i| < y_*$. Let $\mathcal{E}_{\mathrm{mod}}$ be the event that $\mathrm{mod}(\sum_{i=1}^n a(X_i), M) < M_1$. Let $\mathcal{E}_{\mathrm{unif}}$ be the event that $\sum_{i=1}^n \mathbb{1}\{\mathrm{mod}(a(X_i)+m, M) < M - M_1\} \ge \lceil(1-\alpha)(n+1)\rceil$ for all integers m.

Lemma 2. *Under the definitions and notation above, for $(X_1, Y_1), \cdots, (X_n, Y_n) \overset{iid}{\sim} P$, we have*

$$\mathbb{P}_{P^n}\{\mathcal{E}_{\mathrm{mod}}\} \ge \alpha - \sqrt{\frac{2\log n}{n}} - \frac{2}{n} - \frac{1}{M}, \quad \mathbb{P}_{P^n}\{\mathcal{E}_{\max}\} \ge 1 - \frac{1}{n}, \quad \mathbb{P}_{P^n}\{\mathcal{E}_{\mathrm{unif}}\} \ge 1 - \frac{2}{n}$$

and therefore,

$$\mathbb{P}_{P^n}\{\mathcal{E}_{\mathrm{mod}} \cap \mathcal{E}_{\max} \cap \mathcal{E}_{\mathrm{unif}}\} \ge \alpha - \sqrt{\frac{2\log n}{n}} - \frac{1}{M} - \frac{5}{n}$$

By choosing M to be sufficiently large, we obtain

$$\mathbb{P}_{P^n}\{\mathcal{E}_{\mathrm{mod}} \cap \mathcal{E}_{\max} \cap \mathcal{E}_{\mathrm{unif}}\} \ge \alpha - 6\sqrt{\frac{\log n}{n}}$$

We define regression algorithms $\mathcal{A}$ that map a data set $\{(x_1, y_1), \cdots, (x_{n+1}, y_{n+1})\}$ to the fitted function

$$\widehat{\mu}(x) = 2y_*\mathbb{1}\left\{\mathrm{mod}\left(-a(x) + \sum_{i=1}^{n+1} a(x_i), M\right) < M_1\right\}$$

for the full conformal prediction, and

$$\widehat{\mu}(x) = 2y_*\mathbb{1}\left\{\mathrm{mod}\left(a(x) + \sum_{i=1}^{n+1} a(x_i), M\right) \ge M_1\right\}$$

for the jackknife+. Then, we prove Theorem 2 and 3 by showing that if $\mathcal{E}_{\mathrm{mod}} \cap \mathcal{E}_{\max} \cap \mathcal{E}_{\mathrm{unif}}$ holds, $\widehat{C}_n(X_{n+1}) \subseteq (y_*, \infty)$ almost surely over X_{n+1}. Note that we use Dvoretzky-Kiefer-Wolfowitz inequality in the proof of Lemma 1 and 2.

5 Empirical results

The algorithms used as counterexamples in the proof of the Theroem 2 and 3 are extremely unrealistic. Therefore, a natural question is how large the variability of $\alpha_P(\mathcal{D}_n)$ is in practice, particularly in unstable environments. In our simulation, we examine the empirical performance of $\alpha_P(\mathcal{D}_n)$ for the four distribution-free predictive inference tools studied in this work.

5.1 Setting

We choose $\alpha = 0.1$, and compare the performance of split conformal with $n_0 = n_1 = n/2$, full conformal, jackknife+, and CV+ with $K = 20$. We use training sample size $n = 500$ and $n_{\text{test}} = 1000$. For each trial, we generate i.i.d. (X_i, Y_i), $i = 1, \cdots, n+n_{\text{test}}$ from $X_i \sim \mathcal{N}(0, I_d)$ and $Y_i \mid X_i \sim (X_i^T\beta, 1)$ with 200 independent trials for each $d = 125, 250, 500, 1000$. Our algorithm $\mathcal{A}$ is given by ridge regression with $\lambda = 0.0001$. Finally, we estimate the training-conditional miscoverage rate $\alpha_P(\mathcal{D}_n) \approx \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{1}\left\{Y_{n+i} \in \widehat{C}_n(X_{n+i})\right\}$.

Instability of the algorithm The ridge regression algorithm is highly unstable if the number of training points is $\approx d$. However, instability will not occur if the training set size is substantially smaller than d or larger than d. For split conformal, this instability will be high for $d = 250$. For the other three methods, we expect instability to be high for $d = 500$.

5.2 Results

For both split conformal and CV+, as the theory suggests, $\alpha_P(\mathcal{D}_n)$ is consistently near or below the nominal level $\alpha = 0.1$, for all d. Especially, CV+ appears to be highly conservative when $d = 500$. In contrast, for full conformal and for jackknife+, we see that at $d = 500$, $\alpha_P(\mathcal{D}_n)$ is highly variable. On the other hand, it concentrates near $\alpha = 0.1$ for both methods, for all other d.

6 Conclusion

In this paper, we examine one form of conditional validity for methods of distribution free predictive inference: training-conditional validity. We showed that training conditional coverage guarantees can be ensured for the split conformal and CV+ method, but are not possible for either the full conformal or jackknife+ methods without additional assumptions. In addition, we demonstrated empirically that training-conditional miscoverage rates far above the nominal level α can occur in realistic data sets with the latter two methods.

6.1 The role of algorithmic stability

An interesting open question is whether there are any mild assumptions that would ensure training-conditional coverage for full conformal and/or for jackknife+. One possibility is to consider algorithmic stability assumptions. In particular, our empirical results show that poor training-conditional coverage for these two methods is observed exactly in those settings where the behavior of the regression algorithm $\mathcal{A}$ is highly unstable. This suggests that assuming stability of $\mathcal{A}$ could potentially be sufficient to ensure training-conditional coverage for these methods.

References

Rina Foygel Barber, Emmanuel J Candés, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486-507, 2021.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475-490. PMLR, 2012.

7 Question

1. Theorem 4 also holds when $K = n$. Then, it is meaningless because $m = 1$? Then, we can find m and K using similar algorithm in proofs of Theorem 3. large m and small K can make it impossible.

2. why conservative?