STAT 689: Large-scale and high-dimensional statistical inference **Fall 2024**

Lecture 11

1 Conformal prediction

Conformal prediction is a relatively new framework for quantifying uncertainty in the predictions made by arbitrary prediction algorithms. Fundamentally, it does so by converting an algorithm's predictions into prediction sets, which have strong finite-sample coverage properties.

2 Distribution-free predictive inference

Conformal inference is motivated by the problem of distribution-free predictive inference, which can be described as follows. Consider a data set $D_n := \{(X_i, Y_i)\}_{i=1}^n$ drawn independently from the distribution $\mathbb{P}_{XY} = \mathbb{P}_X \times \mathbb{P}_{Y|X}$ on $\mathcal{X} \times \mathcal{Y}$. Here, we can think of $\{X_i\}$ as a set of covariates and $\{Y_i\}$ as the responses. Let X_{n+1} be generated from $\mathbb{P}_X$. Our goal is to construct a prediction set $C(X_{n+1}) := C(D_n, \alpha, X_{n+1})$ for Y_{n+1} which satisfies that

$$\mathbb{P}(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha,$$

where the probability is taken over all $n+1$ points and $\alpha \in (0, 1)$ is a predefined confidence level. A trivial way to achieve this is by setting

$$C(X_{n+1}) = \begin{cases} \mathcal{Y} & \text{with probability } 1 - \alpha, \\ \emptyset & \text{with probability } \alpha. \end{cases}$$

So the real question is: can we achieve the desired coverage in finite samples, without any assumptions on $\mathbb{P}$, by doing something "nontrivial"? In particular, we would like our strategy to adapt to the hardness of the problem in the following sense: the more easily we can predict Y_{n+1} from X_{n+1}, the smaller we would like our set $C(X_{n+1})$ to be.

2.1 Basic idea

The basic idea behind conformal prediction is two-fold. The first key idea can actually be explained in a simpler context, where there are no covariates at all, and we have a sequence of responses $Y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$. Our goal here is to find q_n such that

$$\mathbb{P}(Y_{n+1} \leq q_n) \geq 1 - \alpha. \tag{1}$$

A nature idea would be to set q_n as the $1 - \alpha$ sample quantile of $Y_1, \ldots, Y_n$, which would lead to

$$\mathbb{P}(Y_{n+1} \leq q_n) \approx 1 - \alpha.$$

This would become exact as $n \to +\infty$. But the question here is whether (1) can be fulfilled exactly in finite sample.

2.2 Exchangeability

The key to achieve (1) is to explore the exchangeability among $(Y_1, \ldots, Y_{n+1})$. As $Y_1, \ldots, Y_{n+1}$ are i.i.d, $\{Y_1, \ldots, Y_{n+1}\}$ are exchangeable, meaning that for any permutation π of $\{1, 2, \ldots, n\}$, $(Y_1, \ldots, Y_{n+1})$ and $(Y_{\pi(1)}, \ldots, Y_{\pi(n+1)})$ have the same joint distribution. Therefore, the rank of Y_{n+1} among $\{Y_1, \ldots, Y_{n+1}\}$ is uniformly distributed over $\{1, \ldots, n+1\}$.

Proposition. We have

$$\mathbb{P}(Y_{n+1} \text{ is among the } \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } Y_1, \ldots, Y_{n+1}) \geq 1-\alpha.$$

Set $k = \lceil (1-\alpha)(n+1) \rceil$ for the ease of notation. Let A be the event that Y_{n+1} is among the $\lceil (1-\alpha)(n+1) \rceil$ smallest of $Y_1, \ldots, Y_{n+1}$. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots Y_{(n)}$ be the order statistics and $q_n = Y_{(k)}$.

Exercise 11.1: Prove that A is equivalent to the event that

$$Y_{n+1} \leq q_n = Y_{(k)}.$$

Hint: show that A^c is equivalent to $Y_{n+1} > q_n$.

With the result of Exercise 11.1, the proposition would imply that

$$\mathbb{P}(Y_{n+1} \leq q_n) = \mathbb{P}(A) \geq 1-\alpha.$$

Proof of Proposition. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots Y_{(n+1)}$ and note that Y_{n+1} is equally likely to take any of the values in $\{Y_{(j)} : j = 1, 2, \ldots, n+1\}$. We have

$$\mathbb{P}(A) \geq \frac{k}{n+1} = \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \geq 1-\alpha,$$

where the inequality accounts for the fact that there could be ties among $Y_1, \ldots, Y_{n+1}$, e.g., there are multiple Y_i's that are equal to $Y_{(k)}$.

Remark. Equivalently, q_n is the $k/n = \lceil (1-\alpha)(n+1) \rceil / n$ quantile of the empirical cdf based on $\{Y_1, \ldots, Y_n\}$. In this way, we view q_n as the sample quantile at an adjusted level: we use $\lceil (1-\alpha)(n+1) \rceil / n$, instead of $1-\alpha$, which is a sort of finite-sample correction.

Remark. The proof essentially relies on the fact the rank of Y_{n+1} among $\{Y_1, Y_2, \ldots, Y_{n+1}\}$ is uniformly distributed over $\{1, 2, \ldots, n+1\}$. This follows from the assumption that $Y_1, \ldots, Y_{n+1}$ are i.i.d. However, the i.i.d assumption is not necessary and can be replaced by exchangeability, i.e., $(Y_1, \ldots, Y_{n+1})$ and $(Y_{\pi(1)}, \ldots, Y_{\pi(n+1)})$ have the same joint distribution for any permutation $\pi(\cdot)$.

2.3 Coverage upper bound

We can also derive a coverage upper bound when there are almost surely no ties among $Y_1, \ldots, Y_{n+1}$ (e.g., they are generated from continuous distributions). Let R_{n+1} be the rank of Y_{n+1} among $\{Y_1, Y_2, \ldots, Y_{n+1}\}$. Then R_{n+1} is uniformly distributed over $\{1, 2, \ldots, n+1\}$. Thus we have

$$\mathbb{P}(Y_{n+1} \leq q_n) = \mathbb{P}(A) = \mathbb{P}(R_{n+1} \leq k) = \frac{k}{n+1} = \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} < \frac{(1-\alpha)(n+1)+1}{n+1} = 1-\alpha+\frac{1}{n+1},$$

where we have used the fact that $\lceil k \rceil < k+1$. Therefore,

$$1-\alpha \leq \mathbb{P}(Y_{n+1} \leq q_n) < 1-\alpha+\frac{1}{n+1}.$$

2.4 Regression problems

We now apply the above idea to the regression problem where we observe a set of covariates $X_i \in \mathcal{X}$ and a response $Y_i \in \mathcal{Y} = \mathbb{R}$ for $i = 1, 2, \ldots, n$. We aim to construct a prediction set for Y_{n+1} based on X_{n+1}. Suppose $\hat{f}$ is a point predictor trained based on (X_i, Y_i) such that $\hat{f}(x)$ predicts the value of y that we expect to see at x.

We define

$$R_i = |\hat{f}(X_i) - Y_i|, \quad i = 1, 2, \ldots, n,$$

as the absolute residuals. Following the above idea, we let q_n be the $\lceil (1-\alpha)(n+1) \rceil$ smallest value of $R_1, \ldots, R_n$. We could then define the prediction set

$$C_n(x) = [\hat{f}(x) - q_n, \hat{f}(x) + q_n].$$

We hope to show that $Y_{n+1} \in C_n(X_{n+1})$ with probability at least $1-\alpha$. We note that $Y_{n+1} \in C_n(X_{n+1})$ is equivalent to $R_{n+1} \leq q_n$. Using the same argument as in Exercise 11.1, $R_{n+1} \leq q_n$ is equivalent to that R_{n+1} is among the $\lceil (1-\alpha)(n+1) \rceil$ smallest of $R_1, \ldots, R_{n+1}$. However, we do not have the exchangeability in this case as $\hat{f}$ depends on $(X_1, Y_1), \ldots, (X_n, Y_n)$ which will make $\{R_1, \ldots, R_n\}$ smaller than R_{n+1} in general. To address this issue, we will consider a sample-splitting strategy.

3 Sample-splitting conformal prediction

Another key idea in conformal prediction is to build residuals in a way that treats all of the data, including the test data, in a symmetric fashion. This will ensure that the residuals obey the exchangeability condition we require to get coverage. Sample-splitting is a computationally efficient approach to achieve this goal.

Concretely, we do the following in sample-splitting conformal prediction (split CP). We first divide the training set into two sets:

- D_1, the proper training set;
- D_2, the calibration set.

We can think of D_i as the set of indices so that $D_1 \cup D_2 = \{1, 2, \ldots, n\}$ and $D_1 \cap D_2 = \emptyset$. Suppose the size of D_i is n_i for $n_1 + n_2 = n$.

We train a predictor $\hat{f}_{n_1}$ based on the proper training set $\{(X_i, Y_i) : i \in D_1\}$. Then, we define the absolute residuals based on the calibration set

$$R_i = |\hat{f}_{n_1}(X_i) - Y_i|, \quad i \in D_2.$$

Now let q_{n_2} be the $\lceil (1-\alpha)(n_2+1) \rceil$ smallest value of $R_i, i \in D_2$ and define

$$C_n(x) = [\hat{f}_{n_1}(x) - q_{n_2}, \hat{f}_{n_1}(x) + q_{n_2}].$$

As $\hat{f}_{n_1}(\cdot)$ only depends on the proper training set, $\{R_i : i \in D_2\}$ and R_{n+1} are exchangeable. Using similar arguments as in the previous section, we can show that

$$\begin{aligned}
&\mathbb{P}(Y_{n+1} \in C_n(X_{n+1})) \\
=&\mathbb{P}\left(R_{n+1} \leq q_{n_2}\right) \\
=&\mathbb{P}\left(R_{n+1} \text{ is among the } \lceil (1-\alpha)(n_2+1) \rceil \text{ smallest of } \{R_i, i \in D_2\} \cup \{R_{n+1}\}\right) \\
\geq&\frac{\lceil (1-\alpha)(n_2+1) \rceil}{n_2+1} \geq 1-\alpha.
\end{aligned}$$

Furthermore, if there are no ties among the absolute residuals almost surely, we can show that

$$\mathbb{P}(Y_{n+1} \in C_n(X_{n+1})) < 1 - \alpha + \frac{1}{n_2+1}.$$

3.1 General score functions

Above, we utilized absolute residuals as a negatively oriented score function, where lower values are preferable. However, any negatively oriented score function will suffice, and the argument holds just as before. More precisely, we consider

$$S(x, y) = S(x, y, \hat{f}_{n_1}),$$

which assigns a conformity score to the point (x, y) based on $\hat{f}_{n_1}$. Define the calibration set scores

$$R_i = S(X_i, Y_i), \quad i \in D_2,$$

and a conformal set

$$C_n(x) = \{y \in \mathcal{Y} : S(x, y) \leq \lceil (1-\alpha)(n_2+1) \rceil \text{ smallest of } R_i, i \in D_2\}$$

Then, we can get the same guarantee as before.

3.2 Randomization to get exact coverage

We can always use auxiliary randomization to get the exact coverage $1-\alpha$ in our prediction sets. To illustrate the construction, we begin with the following facts.

Exercise 11.2: Let W be a random variable with the cdf F. Define the inverse cdf $F^{-1}(x) = \inf\{u : F(u) \geq x\}$. First show that

$$\mathbb{P}(F(W) \leq t) \leq t.$$

If u is a discontinuous point of F such that $F(u) - F(u-) > 0$, then for $F(u-) < t < F(u)$, we have $\mathbb{P}(F(W) \leq t) < t$. Let U be a uniform random variable over $[0, 1]$. Define

$$F^*(u) = F(u-) + U(F(u) - F(u-)).$$

Note that if u is a continuous point of F, $F^*(u) = F(u)$. Show that

$$\mathbb{P}(F^*(W) \leq t) = t.$$

Let $R_{n+1} = S(X_{n+1}, Y_{n+1})$ and F_{n_2+1} be the empirical cdf based on $\{R_i : i \in D_2\} \cup \{R_{n+1}\}$. Applying the result from Exercise 11.2 to F_{n_2+1}, we can construct $F^*_{n_2+1}$. Now define the conformal set

$$C^*_n(x) = \left\{ y \in \mathcal{Y} : \frac{1}{n_2+1} \sum_{i \in D_2} \mathbf{1}\{S(X_i, Y_i) < S(x, y)\} + \frac{U}{n_2+1} \left(\sum_{i \in D_2} \mathbf{1}\{R_i = S(x, y)\} + 1 \right) \leq 1 - \alpha \right\},$$

where $U \sim \text{Unif}(0, 1)$ is independent of the data. We have $Y_{n+1} \in C^*_n(X_{n+1})$ if and only if $F^*_{n_2+1}(R_{n+1}) \leq 1 - \alpha$. Therefore, we get

$$\mathbb{P}(Y_{n+1} \in C^*_n(X_{n+1}) | (X_i, Y_i) : i \in D_1) = \mathbb{P}(F^*_{n_2+1}(R_{n+1}) \leq 1 - \alpha | (X_i, Y_i) : i \in D_1) = 1 - \alpha. \tag{2}$$

Exercise 11.3: Suppose there are no ties almost surely among $\{R_1, \ldots, R_{n+1}\}$. Prove (2).

4 Conformal p-value and PRDS property

Consider a score function S and suppose $(Z_1, \ldots, Z_{n+1})$ are drawn independently from a distribution P. We define a conformal p-value as

$$p(z) = \frac{1 + |1 \leq i \leq n : S(Z_i) \leq S(z)|}{n+1},$$

where $|A|$ denotes the cardinality of a set A. The function S here measures how much a new observation conforms to previous data. A high score means high conformity and a low score means low conformity (different from the definition above). The goal of defining the conformal p-value is that we now use it to test if new data points are from the same distribution as $Z_1, \ldots, Z_n \sim P$. If a new data point Z_{n+1} conforms much more poorly than $Z_1, \ldots, Z_n$ to previous data, that counts as evidence against it being from P. Thus,

conformal p-values can be used to test for equality of distributions without any knowledge of the actual underlying distributions.

Exercise 11.4: Assume that $S(Z)$ is continuously distributed for $Z \sim P$. Then $p(Z_{n+1})$ is uniformly distributed over the set $\{1/(n+1), 2/(n+1), \ldots, 1\}$.

Given a test set $\{Z_{n+1}, \ldots, Z_{n+m}\}$ that is independent of $(Z_1, \ldots, Z_n)$, we define the conformal p-values

$$p_i = p(Z_{n+i}), \quad i = 1, 2, \ldots, m.$$

Here the p-value p_i can be used to test $H_{0,i} : Z_{n+i} \sim P$, e.g., testing for outliers.

Example. Let $S(x, y) = -|y - \hat{\mu}(x)|$ as our conformity score, where $\hat{\mu}$ is obtained through a separate data set. Let $Z_{n+1} = (X, Y)$. Recall that in this case

$$p(z) = \frac{1 + |\{1 \le i \le n : |Y - \hat{\mu}(X)| \le |Y_i - \hat{\mu}(X_i)|\}|}{n+1}.$$

We have

$$\begin{aligned} Y \notin C_n(X) &\iff |Y - \hat{\mu}(X)| > \lceil (1-\alpha)(n+1) \rceil \text{ smallest of } |Y_i - \hat{\mu}(X_i)| \text{ for } 1 \le i \le n, \\ &\iff p(Z_{n+1}) \le \frac{1 + n - \lceil (1-\alpha)(n+1) \rceil}{n+1} \\ &\implies p(Z_{n+1}) \le \alpha. \end{aligned}$$

Theorem. If the distribution of $S(Z)$ is continuous for $Z \sim P$, then the conformal p-values are PRDS on the set of true nulls.

Proof. Without loss of generality, assume that $H_1, \ldots, H_{n_0}$ are true nulls and let $S_j = S(Z_j)$ for $1 \le j \le n+m$. For any increasing set D, we aim to show that

$$P((p_1, \ldots, p_m) \in D | p_1 = x)$$

is an increasing function of x.

We note that $(p_2, \ldots, p_m)$ is a deterministic function of p_1 and $W_1 := (S_{(1)}, \ldots, S_{(n+1)}, S_{n+2} \ldots, S_{n+m})$, where $S_{(1)} \le S_{(2)} \le \cdots S_{(n+1)}$ is the order statistics of $S_1, \ldots, S_{n+1}$. Given p_1, we can figure out the rank of S_{n+1} among $S_1, \ldots, S_{n+1}$ and hence we know the order statistics of $S_1, \ldots, S_n$. The p-values $p_2, \ldots, p_m$ can be computed based on $S_{n+2}, \ldots, S_{n+m}$ and the order statistics of $S_1, \ldots, S_n$. Because of this, we write $(p_1, p_2, \ldots, p_m) = G(p_1, W_1)$. We now show that the function $G(p_1, W_1)$ is increasing in p_1. We prove that increasing p_1 and keeping W_1 fixed does not decrease any p-values. Note that

$$\begin{aligned} p_1 = \frac{k}{n+1} &\implies \{S_1, \ldots, S_n\} = \{S_{(1)}, \ldots, S_{(k-1)}, S_{(k+1)}, \ldots, S_{(n+1)}\} \\ &\implies p_j = \frac{1}{n+1} \left\{ 1 + \sum_{i \ne k, i \le n+1} \mathbf{1}\{S_{n+j} \ge S_{(i)}\} \right\} \\ &\iff p_j = \frac{1}{n+1} \left\{ 1 - \mathbf{1}\{S_{n+j} \ge S_{(k)}\} + \sum_{1 \le i \le n+1} \mathbf{1}\{S_{n+j} \ge S_{(i)}\} \right\}. \end{aligned}$$

The term $1 - \mathbf{1}\{S_{n+j} \ge S_{(k)}\} + \sum_{1 \le i \le n+1} \mathbf{1}\{S_{n+j} \ge S_{(i)}\}$ is increasing in k. Thus, $G(p_1, W_1)$ is increasing in p_1.

Finally, we note that p_1 is independent of W_1 under the null. The null here means that $Z_{n+1} \sim P$. In that case, we know that p_1 is uniform over $\{1/(n+1), \ldots, 1\}$. Also, given $S_{(1)}, ..., S_{(n+1)}$, the value of S_{n+1} is uniform over this set since the random variables are i.i.d under the null and hence, p_1 has the same

distribution even after conditioning which shows that $p_1 \perp (S_{(1)}, ..., S_{(n+1)})$. Since $Z_{n+1} \perp Z_{n+2}, ..., Z_{n+m}$ by assumption, we have that $p_1 \perp W_1$. This enables the following calculation. For any increasing set D,

$$\begin{aligned} P((p_1, p_2, \ldots, p_m) \in D \mid p_1 = x) &= P(G(p_1, W_1) \in D \mid p_1 = x) \\ &= \mathbb{E}_{W_1}[P(G(p_1, W_1) \in D \mid p_1 = x, W_1)] \\ &= \mathbb{E}_{W_1}[\mathbf{1}(G(x, W_1) \in D)] \qquad (p_1 \perp W_1) \end{aligned}$$

Since $G(p_1, W_1)$ is increasing in p_1 and D is an increasing set, $\mathbf{1}(G(x, W_1) \in D)$ is increasing in x which implies the PRDS property.

5 Conformalized quantile regression

The materials in this section are from Romano et al. (2019). One limitation of the sample-splitting conformal prediction is that the length of the prediction interval is equal to $2q_{n_2}$, which is independent of X_{n+1}. In other words, the interval does not reflect the heterogeneity of the observations. In this section, we will learn a new method fully adaptive to heteroscedasticity. It combines conformal prediction with classical quantile regression, inheriting the advantages of both.

5.1 Quantile regression

The conditional distribution of Y given $X = x$ is

$$F(y|X = x) = P(Y \leq y|X = x)$$

and the αth conditional quantile function is

$$q_\alpha(x) = \inf\{y \in \mathbb{R} : F(y|X = x) \geq \alpha\}.$$

We can obtain a conditional prediction interval for Y given $X = x$ with mis-coverage rate α as

$$C(x) = [q_{\alpha/2}(x), q_{1-\alpha/2}(x)].$$

By construction, this interval satisfies that

$$P(Y \in C(X)|X = x) \geq 1 - \alpha.$$

Notice that the length of the interval $C(X)$ can vary greatly depending on the value of X. The uncertainty in the prediction of Y is naturally reflected in the length of the interval. In practice, we cannot know this ideal prediction interval, but we can try to estimate it from the data.

Classical regression estimates the conditional mean of Y given $X = x$ by $\mu(x; \hat{\theta})$ with

$$\hat{\theta} = \text{argmin}_\theta \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu(X_i, \theta))^2 + R(\theta).$$

Here θ are the parameters associated with the regression model, $\mu(x; \theta)$ is the regression function (for modeling the conditional mean), and $R(\theta)$ is a penalty/regularizer.

Quantile regression estimates a conditional quantile function $q_\alpha(x)$ of Y given $X = x$ by $\hat{q}_\alpha(x) = f(x; \hat{\theta})$, with

$$\hat{\theta} = \text{argmin}_\theta \frac{1}{n} \sum_{i=1}^{n} \rho_\alpha(Y_i, f(X_i; \theta)) + R(\theta).$$

Here $f(x; \theta)$ is the quantile function and ρ_α is the so-called check function or pinball loss defined by

$$\rho_\alpha(y, b) = \begin{cases} \alpha(y - b) & \text{if } y > b, \\ (1 - \alpha)(b - y) & \text{if } y \leq b. \end{cases}$$

The simplicity and generality of this formulation make quantile regression widely applicable. As in classical regression, one can leverage the great variety of machine learning methods to design and learn $q_\alpha(x)$. An obvious strategy to construct a prediction band with the nominal mis-coverage rate α: estimate $q_\alpha(x)$ by $\hat{q}_\alpha(x)$ using quantile regression and construct the interval:

$$\hat{C}(x) = [\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)].$$

However, it is not guaranteed to satisfy

$$P(Y \in \hat{C}(X)|X = x) \geq 1 - \alpha$$

when $C(X)$ is replaced by the estimated interval $\hat{C}(X)$. To address this issue, we shall use the idea of conformal prediction.

5.2 Conformalized quantile regression

Similar to split conformal prediction, we split the data into a proper training set indexed by D_1 and a calibration set indexed by D_2.

1. Given any quantile regression algorithm $\mathcal{A}$, we fit two quantile regression functions $\hat{q}_{\alpha/2}(x)$ and $\hat{q}_{1-\alpha/2}(x)$ on the proper training set.

2. We compute conformity scores that quantify the error made by the plug-in prediction interval $\hat{C}(x) = [\hat{q}_{\alpha/2}(x), \hat{q}_{1-\alpha/2}(x)]$. The scores are evaluated on the calibration set as

   $$E_i = \max\{\hat{q}_{\alpha/2}(X_i) - Y_i, Y_i - \hat{q}_{1-\alpha/2}(X_i)\}$$

   for each $i \in D_2$. If $Y_i < \hat{q}_{\alpha/2}(X_i)$, $E_i = |Y_i - \hat{q}_{\alpha/2}(X_i)|$. Similarly, if $Y_i > \hat{q}_{1-\alpha/2}(X_i)$, $E_i = |Y_i - \hat{q}_{1-\alpha/2}(X_i)|$. If $\hat{q}_{\alpha/2}(X_i) < Y_i < \hat{q}_{1-\alpha/2}(X_i)$, E_i is non-positive. The conformity score thus accounts for both undercoverage and overcoverage.

3. Given a new data point X_{n+1}, we construct a prediction interval

   $$\widetilde{C}(X_{n+1}) = [\hat{q}_{\alpha/2}(X_i) - Q_{1-\alpha}(E, D_2), \hat{q}_{1-\alpha/2}(X_i) + Q_{1-\alpha}(E, D_2)].$$

   where $Q_{1-\alpha}(E, D_2)$ is the $\lceil (1-\alpha)(n_2+1) \rceil / n_2$th empirical quantile of $\{E_i : i \in D_2\}$ with $n_2 = |D_2|$.

Theorem. Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable, then the prediction interval $\widetilde{C}(X_{n+1})$ satisfies that

$$P(Y_{n+1} \in \widetilde{C}(X_{n+1})) \geq 1 - \alpha.$$

When E_is have no ties almost surely, we have

$$P(Y_{n+1} \in \widetilde{C}(X_{n+1})) \leq 1 - \alpha + \frac{1}{|D_2| + 1}.$$

Proof. By the construction of $\widetilde{C}$, we have $Y_{n+1} \in \widetilde{C}(X_{n+1})$ if and only if

$$E_{n+1} \leq Q_{1-\alpha}(E, D_2).$$

As $E_i, i \in D_2$ and E_{n+1} are exchangeable, using the same argument as before, we have

$$P(Y_{n+1} \in \widetilde{C}(X_{n+1})|(X_i, Y_i), i \in D_1) \geq 1 - \alpha.$$

and when there is no tie,

$$P(Y_{n+1} \in \widetilde{C}(X_{n+1})|(X_i, Y_i), i \in D_1) \leq 1 - \alpha + \frac{1}{|D_2| + 1}.$$

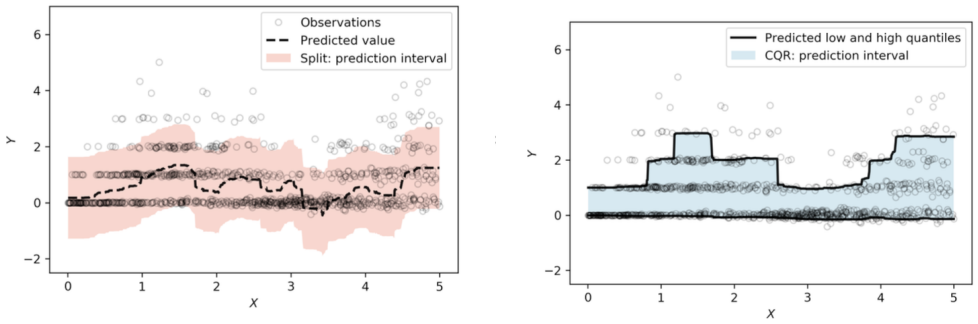

Figure 1: Left panel: standard split conformal prediction; Right panel: conformal quantile regression.

6 The jackknife and jackknife+

Sample-splitting conformal prediction is computationally very cheap and has the desired coverage under exchangeability (no other assumption on the predictive algorithm is needed). However, these benefits come at a statistical cost. If n_1 is small, the fitted model $\hat{f}_{n_1}$ can be very poor, leading to a wide interval. If n_2 is small, the interval can again be wide. The jackknife and jackknife+ are alternative approaches that can use the data more efficiently but are computationally expensive.

To describe the jackknife prediction interval, we let $\hat{f}_{-i}$ be a predictor based on the observations $(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)$. Denote by

$$R_i^{\text{LOO}} = |Y_i - \hat{f}_{-i}(X_i)|$$

the ith leave-one-out residual. Further, let

$$q_n = \text{the } \lceil (1-\alpha)(n+1) \rceil \text{th smallest value of } R_1^{\text{LOO}}, \ldots, R_n^{\text{LOO}},$$

the $1-\alpha$ quantile of the empirical distribution of these values. Then, the jackknife prediction interval is given by

$$[\hat{f}(X_{n+1}) - q_n, \hat{f}(X_{n+1}) + q_n],$$

where $\hat{f}$ is the predictor based on $(X_1, Y_1), \ldots, (X_n, Y_n)$. However, the jackknife procedure does not have any universal theoretical guarantees. In particular, it may lose predictive coverage when $\hat{f}$ is unstable. For example, the jackknife can have extremely poor coverage using least squares regression when the sample size n is close to the dimension p; Figure 2.

To overcome this issue, Barber et al. (2020, arXiv:1905.02928) introduce the jackknife+, which comes with universal theoretical guarantees. Let

$$q_{n,\text{low}} = \text{the } \lfloor \alpha(n+1) \rfloor \text{th smallest value of } \hat{f}_{-1}(X_{n+1}) - R_1^{\text{LOO}}, \ldots, \hat{f}_{-n}(X_{n+1}) - R_n^{\text{LOO}},$$

and

$$q_{n,\text{up}} = \text{the } \lceil (1-\alpha)(n+1) \rceil \text{th smallest value of } \hat{f}_{-1}(X_{n+1}) + R_1^{\text{LOO}}, \ldots, \hat{f}_{-n}(X_{n+1}) + R_n^{\text{LOO}}.$$

The jackknife+ prediction interval is given by

$$[q_{n,\text{low}}, q_{n,\text{up}}].$$

While both versions of jackknife use the leave-one-out residuals, the difference is that for jackknife, we center our interval on the predicted value $\hat{f}(X_{n+1})$ fitted on the full training data, while for jackknife+, we use the leave-one-out predictions $\hat{f}_{-i}(X_{n+1})$ for the test point.

Theorem. The jackknife+ prediction interval satisfies

$$P(Y_{n+1} \in [q_{n,\text{low}}, q_{n,\text{up}}]) \geq 1 - 2\alpha.$$

See Barber et al. (2020) for a proof of this result.

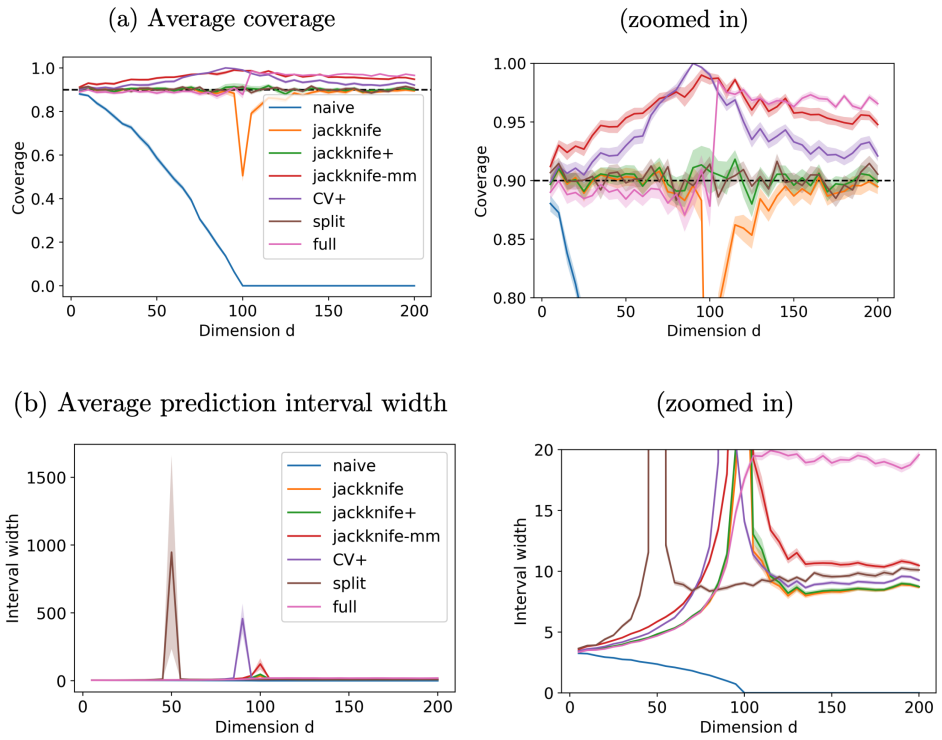

Figure 2: Average coverage and prediction interval width for different methods. The plots are from Barber et al. (2020).

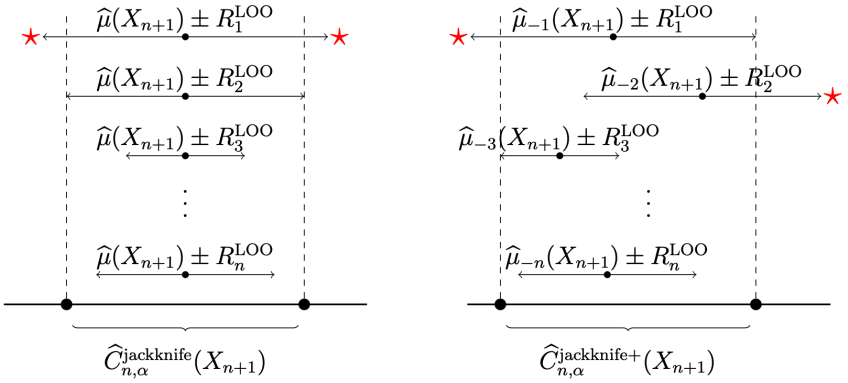

Figure 3: Illustration of the usual jackknife and the new jackknife+. The resulting prediction intervals are chosen so that, on either side, the boundary is exceeded by a sufficiently small proportion of the two-sided arrows—above, these are marked with a star.