Conformal Prediction Under Covariate Shift

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas

Kiho Park

parkkiho@uchicago.edu

1 Introduction

We extend conformal prediction beyond the setting of exchangeable data. In particular, we show that a weighted version of conformal prediction can be used to compute distribution-free prediction intervals in setting of covariate shift. We begin by reviewing the basics of conformal prediction.

1.1 Qunatile lemma

Let Quantile$(\beta; F)$ denote the level β quantile of a distribution F on $\mathbb{R} \cup \{\infty\}$, i.e., for $Z \sim F$,

$$\text{Quantile}(\beta; F) = \inf\{z : \mathbb{P}\{Z \leq z\} \geq \beta\}$$

For values $v_1, \cdots, v_n$, $v_{1:n}$ denotes their multiset. Then, let

$$\text{Quantile}(\beta; v_{1:n}) = \text{Quantile}\left(\beta; \frac{1}{n}\sum_{i=1}^{n} \delta_{v_i}\right)$$

denote quantiles of the empirical distribution of the values $v_1, \cdots, v_n$.

Lemma 1. *If $V_1, \cdots, V_{n+1}$ are exchangeable random variables, then for any $\beta \in (0, 1)$, we have*

$$\mathbb{P}\left\{V_{n+1} \leq \text{Quantile}\left(\beta; V_{1:n} \cup \{\infty\}\right)\right\} \geq \beta$$

Furthermore, if ties between $V_1, \cdots, V_{n+1}$ occur with probability zero, then the above probability is upper bounded by $\beta + 1/(n+1)$.

1.2 Conformal prediction

Denote $Z_i = (X_i, Y_i)$, $i = 1, \cdots, n$, $Z_{1:n} = \{Z_1, \cdots, Z_n\}$, and $Z_{-i} = Z_{1:n} \setminus \{Z_i\}$. We first choose a score function $\mathcal{S}$. For example, we might choose $\mathcal{S}((x, y), Z) = |y - \widehat{\mu}(x)|$ where $\widehat{\mu}$ is a fitted regression function by an algorithm $\mathcal{A}$ on (x, y) and Z. At each $x \in \mathbb{R}^d$, for each $y \in \mathbb{R}$, we calculate the nonconformity scores

$$V_i^{(x,y)} = \mathcal{S}(Z_i, Z_{-i} \cup \{(x, y)\}),\ i = 1, \cdots, n, \quad \text{and} \quad V_{n+1}^{(x,y)} = \mathcal{S}((x, y), Z_{1:n})$$

Theorem 1. *Assume that $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \cdots, n+1$ are exchangeable. For any score function $\mathcal{S}$, and any $\alpha \in (0, 1)$, define the conformal band (based on the first n samples) at $x \in \mathbb{R}^d$ by*

$$\widehat{C}_n(x) = \left\{y \in \mathbb{R} : V_{n+1}^{(x,y)} \leq \text{Quantile}(1 - \alpha; V_{1:n}^{(x,y)} \cup \{\infty\})\right\}$$

where $V_{1:n}^{(x,y)} = \{V_1^{(x,y)}, \cdots, V_n^{(x,y)}\}$. Then, $\widehat{C}_n$ satisfies

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

Furthermore, if ties between $V_1^{(X_{n+1},Y_{n+1})}, \cdots, V_{n+1}^{(X_{n+1},Y_{n+1})}$ occur with probability zero, then this probability is upper bounded by $1 - \alpha + 1/(n+1)$.

2 Covariate shift

Covariate shift is a setting that

$$(X_i, Y_i) \overset{\text{i.i.d}}{\sim} P = P_X \times P_{Y|X},\ i = 1, \cdots, n$$

$$(X_{n+1}, Y_{n+1}) \sim \tilde{P} = \tilde{P}_X \times P_{Y|X},\ \text{independently}$$

Then, the data (X_i, Y_i), $i = 1, \cdots, n+1$ are no longer exchangeable. If we know the ratio of test to training covariate likelihoods, $d\tilde{P}_X/dP_X$, we can still perform a modified version of conformal inference.

2.1 Weighted conformal prediction

Let $w(X_i) = d\tilde{P}_X(X_i)/dP_X(X_i)$ denote the likelihood ratio. We define the weights as

$$p_i^w(x) = \frac{w(X_i)}{\sum_{j=1}^n w(X_j) + w(x)},\ i = 1, \cdots, n, \quad \text{and} \quad p_{n+1}^w(x) = \frac{w(x)}{\sum_{j=1}^n w(X_j) + w(x)}$$

Corollary 1. *Assume that $\tilde{P}_X$ is absolutely continuous with respect to P_X, and denote $w = d\tilde{P}_X/dP_X$. For any score function $\mathcal{S}$, and any $\alpha \in (0,1)$, define for $x \in \mathbb{R}^d$,*

$$\widehat{C}_n(x) = \left\{ y \in \mathbb{R} : V_{n+1}^{(x,y)} \leq \text{Quantile}\left(1-\alpha; \sum_{i=1}^n p_i^w(x)\delta_{V_i^{(x,y)}} + p_{n+1}^w(x)\delta_\infty\right)\right\}$$

Then, $\widehat{C}_n$ satisfies

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

2.2 Weighted split conformal

Constructing a conformal prediction band can be computationally intensive. Split conformal prediction resolves this issue. For example, consider the regression function μ_0 is fitted by a preliminary data set $(X_1^0, Y_1^0), \cdots, (X_{n_0}^0, Y_{n_0}^0)$ and the score function $\mathcal{S}((x,y), Z) = |y - \mu_0(x)|$. Given data $(X_1, Y_1), \cdots, (X_n, Y_n)$, independent of $(X_1^0, Y_1^0), \cdots, (X_{n_0}^0, Y_{n_0}^0)$, we calculate

$$V_i^{(x,y)} = |Y_i - \mu_0(X_i)|\ i = 1, \cdots, n, \quad \text{and} \quad V_{n+1}^{(x,y)} = |y - \mu_0(x)|$$

Then, the conformal prediction interval at a point $x \in \mathbb{R}^d$ is

$$\widehat{C}_n(x) = \mu_0(x) \pm \text{Quantile}\left(1-\alpha; \{|Y_i - \mu_0(X_i)|\}_{i=1}^n \cup \{\infty\}\right)$$

and it has coverage at least $1-\alpha$ conditional on $(X_1^0, Y_1^0), \cdots, (X_{n_0}^0, Y_{n_0}^0)$ by Theorem 1.

Corollary 1 also applies to the split scenario. The band at $x \in \mathbb{R}^d$

$$\widehat{C}_n(x) = \mu_0(x) \pm \text{Quantile}\left(1-\alpha; \sum_{i=1}^n p_i^w(x)\delta_{|Y_i - \mu_0(X_i)|} + p_{n+1}^w(x)\delta_\infty\right)$$

has coverage at least $1-\alpha$ conditional on $(X_1^0, Y_1^0), \cdots, (X_{n_0}^0, Y_{n_0}^0)$.

2.3 Airfoil data example

We demonstrate the use of conformal prediction in the covariate shift setting for an empirical example. We consider the airfoil data set ($N = 1503$) and use ordinary and weighted split conformal prediction methods.

We repeated an experiment for 5000 trials. For each trial, we randomly partitioned the data $\{(X_i, Y_i)\}_{i=1}^N$ into three sets D_{pre} (25% for μ_0), D_{train} (25% for residual quantiles), D_{test} (50% and exchangeable with D_{train}). Also, we constructed a covariate shift test set D_{shift} by sampling 25% of D_{test} with replacement, with probabilities proportional to $w(x) = \exp(x^T\beta)$ where $\beta = (-1, 0, 0, 0, 1)$. Note that $d\tilde{P}_x \propto \exp(x^T\beta)dP_X$ and it is called exponential tilting. This data has heteroskedasticity, and thus we might expect the covariate shift to have some effect on the validity of the ordinary conformal prediction intervals.

First, the ordinary split conformal prediction considerably undercovers for D_{shift} whereas it has empirical coverage at the nominal level for D_{test}. Next, the oracle weighted split conformal prediction covers the points in D_{shift} at the nominal level, but it is more dispersed than it is when there is no covariate shift. This is because we are relying on a reduced "effective sample size":

$$\widehat{n} = \frac{[\sum_{i=1}^n |w(X_i)|]^2}{\sum_{i=1}^n |w(X_i)|^2} = \frac{||w(X_{1:n})||_1^2}{||w(X_{1:n})||_2^2}$$

When we ran unweighted split conformal on D_{test} by using only $\hat{n}$ subsampled points from D_{train}, the histograms are similar. It demonstrates that the overdispersion is generated by the reduced effective sample size.

We can estimate $w = d\tilde{P}_X/dP_X$ by applying logistic regression or random forests to the feature-class pairs (X_i, C_i) , $i = 1, \cdots, n+m$ where $(X_i, C_i) = (X_i, 0)$ for $i = 1, \cdots, n$ correspond to D_{train} and $(X_i, C_i) = (X_i, 1)$ for $i = n+1, \cdots, n+m$ correspond to D_{shift}. If $\widehat{p}(x)$ is an estimate of $\mathbb{P}(C = 1 \mid X = x)$ obtained by fitting a classifier to the pairs, then we can use

$$\widehat{w}(x) = \frac{\mathbb{P}(C=1)}{\mathbb{P}(C=0)}\frac{d\tilde{P}_X}{dP_X}(x) = \frac{\mathbb{P}(C=1 \mid X=x)}{\mathbb{P}(C=0 \mid X=x)} = \frac{\widehat{p}(x)}{1-\widehat{p}(x)}$$

as our estimated weight function (equivalent up to a constant). When we ran weighted split conformal prediction to cover the points in D_{shift} with $\widehat{w}$ using logistic regression and random forests, the coverage histograms have good average coverage. They are only a little more dispersed than that for the oracle weights.

Now we check the median lengths of prediction intervals. There is no differences in the lengths of ordinary split conformal intervals when there is or is not covariate shift. Next, the oracle-weighted split conformal intervals are longer than the ordinary split conformal intervals that use an equivalent effective sample size. This is because μ_0 itself should be adjusted to account for covariate shift. Lastly, the random forests-weighted split conformal intervals are more variable than the logistic regression-weighted split conformal intervals.

Lastly, the empirical coverages and median lengths of split conformal intervals to cover points in D_{test} by the ordinary unweighted approach, logistic regression-weighted approach, and the random forests-weighted approach are very similar. Since random forests are very flexible, the coverages and lengths are a little dispersed.

3 Weighted exchangeability

We develop a general result on conformal prediction for settings in which the data satisfy weighted exchangeability.

3.1 Alternate proof of Lemma 1

Conditioned on the event E_v that $\{V_1, \cdots, V_{n+1}\} = \{v_1, \cdots, v_{n+1}\}$,

$$\mathbb{P}\{V_{n+1} = v_i \mid E_v\} = \frac{\sum_{\sigma:\sigma(n+1)=i} f(v_{\sigma(1)}, \cdots, v_{\sigma(n+1)})}{\sum_{\sigma} f(v_{\sigma(1)}, \cdots, v_{\sigma(n+1)})} = \frac{n!}{(n+1)!} = \frac{1}{n+1}$$

Thus,

$$V_{n+1} \mid E_v \sim \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{v_i}$$

so

$$\mathbb{P}\left\{ V_{n+1} \leq \text{ Quantile}\left(\beta; \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{V_i}\right) \middle| E_V \right\} \geq \beta$$

and we can marginalize it for the result.

3.2 Weighted exchangeability

Definition 1. *Random variables $V_1, \cdots, V_n$ are weighted exchangeable if the density f can be factorized as*

$$f(v_1, \cdots, v_n) = \prod_{i=1}^{n} w_i(v_i) \cdot g(v_1, \cdots, v_n)$$

with weight functions $w_1, \cdots, w_n$, where $g(v_{\sigma(1)}, \cdots, v_{\sigma(n)}) = g(v_1, \cdots, v_n)$ for any permutation σ of $1, \cdots, n$.

Lemma 2. *Let $Z_i \sim P_i$, $i = 1, \cdots, n$ be independent draws, where each P_i is absolutely continuous with respect to P_1, for $i \geq 2$. Then, $Z_1, \cdots, Z_n$ are weighted exchangeable, with weight functions $w_1 \equiv 1$, and $w_i = dP_i/dP_1$, $i \geq 2$.*

3.3 Weighted quantile lemma

We give a weighted generalization of Lemma 1.

Lemma 3. *Let Z_i, $i = 1, \cdots, n+1$ be weighted exchangeable random variables, with weight functions $w_1, \cdots, w_{n+1}$. Let $V_i = \mathcal{S}(Z_i, Z_{-i})$, where $Z_{-i} = Z_{1:(n+1)} \setminus \{Z_i\}$, for $i = 1, \cdots, n+1$, and $\mathcal{S}$ is an arbitrary score function. Define*

$$p_i^w(z_1, \cdots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}{\sum_{\sigma} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}, \ i = 1, \cdots, n+1$$

Then, for any $\beta \in (0, 1)$,

$$\mathbb{P}\left\{ V_{n+1} \leq \text{Quantile}\left(\beta; \sum_{i=1}^{n} p_i^w(Z_1, \cdots, Z_{n+1})\delta_{V_i} + p_{n+1}^w(Z_1, \cdots, Z_{n+1})\delta_{\infty}\right)\right\} \geq \beta$$

The proof is similar to the alternate proof of Lemma 1. Conditioned on the event E_z that $\{Z_1, \cdots, Z_{n+1}\} = \{z_1, \cdots, z_{n+1}\}$, $\mathbb{P}\{V_{n+1} = v_i \mid E_z\} = \mathbb{P}\{Z_{n+1} = z_i \mid E_z\}$. We show

$$V_{n+1} \mid E_z \sim \sum_{i=1}^{n+1} p_i^w(z_1, \cdots, z_{n+1})\delta_{v_i}$$

Then, the lemma is proved by marginalization.

3.4 Weighted conformal prediction

Lemma 3 implies a weighted version of conformal prediction.

Theorem 2. *Assume that $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \cdots, n+1$ are weighted exchangeable with weight functions $w_1, \cdots, w_{n+1}$. For any score function $\mathcal{S}$, and any $\alpha \in (0,1)$, define the weighted conformal band at a point $x \in \mathbb{R}^d$ by*

$$\widehat{C}_n(x) = \left\{ y \in \mathbb{R} : V_{n+1}^{(x,y)} \leq \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} p_i^w(Z_1, \cdots, Z_n, (x,y))\delta_{V_i^{(x,y)}} + p_{n+1}^w(Z_1, \cdots, Z_n, (x,y))\delta_\infty\right)\right\}$$

Then $\widehat{C}_n$ satisfies

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

Note that the split conformal method with $\mathcal{S}((x,y),Z) = |y - \mu_0(x)|$ can be generalized as

$$\widehat{C}_n(x) = \mu_0(x) \pm \text{Quantile}\left(1-\alpha; \sum_{i=1}^{n} p_i^w(Z_1, \cdots, Z_n, (x,y))\delta_{|Y_i - \mu_0(X_i)|} + p_{n+1}^w(Z_1, \cdots, Z_n, (x,y))\delta_\infty\right)$$

This has coverage at least $1-\alpha$, conditional on $(X_1^0, Y_1^0), \cdots, (X_{n_0}^0, Y_{n_0}^0)$.

3.5 Proof of Corollary 1

For the case of covariate shift, the independent draws $Z_i = (X_i, Y_i)$, $i = 1, \cdots, n+1$ are weighted exchangeable with $w_i \equiv 1$ for $i = 1, \cdots, n$ and $w_{n+1}((x,y)) = w(x)$ by Lemma 2. Then, Theorem 2 implies Corollary 1 with $p_i^w(Z_1, \cdots, Z_n, (x,y)) = p_i^w(x)$.

4 Discussion

The weighted conformal methodology can be applied to several other closely related settings.

Graphical models with covariate shift. Assume that the training data $(Z, X, Y) \sim P$ has the Markovian structure $Z \to X \to Y$. Suppose that the distribution of Z has changed on the test data $(Z, X, Y) \sim \tilde{P}$. We can run weighted conformal prediction on only (X, Y), but if X is high-dimensional and Z is low-dimensional, we can just estimate the likelihood ratio of Z between test and training sets for the weighted conformal prediction. This would be more tractable.

Missing covariates with known summaries. Suppose $(Z, X, Y) \sim P^A$ where $Z \in \{0,1\}$ such as sex of the patient. When we know X for each patient and only the fraction of Z values that equal one, instead of collected data $(Z, X) \sim P^B$, we can use weighted conformal prediction by calculating the likelihood ratio of Z between P^B and P^A, assuming that $P^A_{X|Z} = P^B_{X|Z}$.

Towards local conditional coverage? The exact conditional coverage at $x_0 \in \mathbb{R}^d$ is

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(x_0) \mid X_{n+1} = x_0\right\} \geq 1-\alpha$$

This conditional coverage for almost all $x_0 \in \mathbb{R}^d$ and all distributions P is far too strong. Thus we relax it to an approximate conditional coverage:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(x_0) \mid X_{n+1} = x_0 + h\omega\right\} \geq 1-\alpha$$

where the probability is taken over the $n+1$ data points and an independent draw ω from a distribution whose density is proportional to a kernel K. When the bandwidth h is small, this kind of locally-weighted guarantee should be close to a guarantee on conditional coverage.

We can use the weighted conformal prediction for this approximate conditional coverage. It can be written as

$$\mathbb{P}_{x_0}\left\{Y_{n+1} \in \widehat{C}_n(\tilde{X}_{n+1})\right\} \geq 1-\alpha$$

where training points (X_i, Y_i), $i = 1, \cdots, n$, are i.i.d. from $P = P_X \times P_{Y|X}$ and an independent test point $(\tilde{X}_{n+1}, Y_{n+1})$ is from $\tilde{P} = \tilde{P}_X \times P_{Y|X}$ with $d\tilde{P}_X/dP_X \propto K((\cdot - x_0)/h)$. Then, for any score function $\mathcal{S}$, and any $\alpha \in (0,1)$, define

$$\widehat{C}_n(x) = \left\{ y \in \mathbb{R} : V_{n+1}^{(x,y)} \leq \text{Quantile}\left(1-\alpha; \frac{\sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right)\delta_{V_i^{(x,y)}} + K\left(\frac{x-x_0}{h}\right)\delta_\infty}{\sum_{i=1}^n K\left(\frac{X_i-x_0}{h}\right) + K\left(\frac{x-x_0}{h}\right)}\right)\right\}$$

Then by Corollary 1,

$$\mathbb{P}_{x_0}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1}; x_0)\right\} \geq 1-\alpha$$

Note that the prediction band depends on a new point x_0. We have not provided a recipe for the construction of a single band $\widehat{C}_n$ that has the approximate conditional coverage at all $x_0 \in \mathbb{R}^d$. Many important practical and philosophical problems in assumption-lean conditional predictive inference remain open.