Review of Conformal+methodology

by Weihao LI

Feb 11th 2023

Outline

Conformal Inference: Lei's work

- Consider i.i.d. regression data

$$Z_1, \ldots, Z_n \sim P,$$

where each $Z_i = (X_i, Y_i)$ is a random variable in $\mathbb{R}^d \times \mathbb{R}$, comprised of a response variable Y_i and a d-dimensional vector of features (or predictors, or covariates) $X_i = (X_i(1), \ldots, X_i(d))$.

Conformal Inference: Lei's work

- Consider i.i.d. regression data

$$Z_1, \ldots, Z_n \sim P,$$

  where each $Z_i = (X_i, Y_i)$ is a random variable in $\mathbb{R}^d \times \mathbb{R}$, comprised of a response variable Y_i and a d-dimensional vector of features (or predictors, or covariates) $X_i = (X_i(1), \ldots, X_i(d))$.
- Constructing a prediction band $C \subseteq \mathbb{R}^d \times \mathbb{R}$ based on $Z_1, \ldots, Z_n$ with the property that

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

Objective of conformal inference

make CI without assumption

A simple exercise

Exercise:

Suppose we have positive i.i.d random variables $R_1, \cdots, R_n, R_{n+1}$. Let $Q_{1-\alpha}$ denote the empirical $1-\alpha$ quantile for $\{R_1, \cdots, R_n\}$, what is approximate value of $\mathbf{P}(R_{n+1} \in [0, Q_{1-\alpha}])$?

A simple exercise

Exercise:

Suppose we have positive i.i.d random variables $R_1, \cdots, R_n, R_{n+1}$. Let $Q_{1-\alpha}$ denote the empirical $1-\alpha$ quantile for $\{R_1, \cdots, R_n\}$, what is approximate value of $\mathbf{P}(R_{n+1} \in [0, Q_{1-\alpha}])$?

Answer: Consider ordered statisitics for $n+1$ R_i

$$R_{(1)}, R_{(2)}, R_{(3)}, \cdots, R_{(n)}, R_{(n+1)}$$

$$\boldsymbol{P}(R_{n+1} \in [0, Q_{1-\alpha}]) \approx \boldsymbol{P}(R_{n+1} \text{ rank lower than } (1-\alpha)(n+1)) \approx 1-\alpha \quad (2)$$

A simple exercise

Exercise:

Suppose we have positive i.i.d random variables $R_1, \cdots, R_n, R_{n+1}$. Let $Q_{1-\alpha}$ denote the empirical $1-\alpha$ quantile for $\{R_1, \cdots, R_n\}$, what is approximate value of $\mathbf{P}(R_{n+1} \in [0, Q_{1-\alpha}])$?

Answer: Consider ordered statisitics for $n+1$ R_i

$$R_{(1)}, R_{(2)}, R_{(3)}, \cdots, R_{(n)}, R_{(n+1)}$$

$$\boldsymbol{P}(R_{n+1} \in [0, Q_{1-\alpha}]) \approx \boldsymbol{P}(R_{n+1} \text{ rank lower than } (1-\alpha)(n+1)) \approx 1-\alpha \tag{2}$$

Remark: relax i.i.d to exchangeable.

Construction of full conformal prediction set

- For each value $y \in \mathbb{R}$, we construct an augmented regression estimator $\widehat{\mu}_y$, which is trained on the augmented data set $Z_1, \ldots, Z_n, (X_{n+1}, y)$. Now we define

$$R_{y,i} = |Y_i - \widehat{\mu}_y(X_i)|, i = 1, \ldots, n \quad \text{and} \quad R_{y,n+1} = |y - \widehat{\mu}_y(X_{n+1})|,$$

Construction of full conformal prediction set

- For each value $y \in \mathbb{R}$, we construct an augmented regression estimator $\widehat{\mu}_y$, which is trained on the augmented data set $Z_1, \ldots, Z_n, (X_{n+1}, y)$. Now we define

$$R_{y,i} = \left|Y_i - \widehat{\mu}_y\left(X_i\right)\right|, i = 1, \ldots, n \quad \text{and} \quad R_{y,n+1} = \left|y - \widehat{\mu}_y\left(X_{n+1}\right)\right|,$$

- Hypothesis test: $H_0 : Y_{n+1} = y$, define $1-$"p-value"

$$\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\left\{R_{y,i} \leq R_{y,n+1}\right\} \tag{3}$$

Construction of full conformal prediction set

- For each value $y \in \mathbb{R}$, we construct an augmented regression estimator $\widehat{\mu}_y$, which is trained on the augmented data set $Z_1, \ldots, Z_n, (X_{n+1}, y)$. Now we define

$$R_{y,i} = \left| Y_i - \widehat{\mu}_y\left(X_i\right) \right|, i = 1, \ldots, n \quad \text{and} \quad R_{y,n+1} = \left| y - \widehat{\mu}_y\left(X_{n+1}\right) \right|,$$

- Hypothesis test: $H_0 : Y_{n+1} = y$, define $1-$"p-value"

$$\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\left\{R_{y,i} \leq R_{y,n+1}\right\} \tag{3}$$

we have $\mathbb{P}\left(\pi\left(Y_{n+1}\right) \leq 1-\alpha\right) \geq 1-\alpha$

$$C_{\text{conf}}\left(X_{n+1}\right) = \{y \in \mathbb{R} : \pi(y) \leq 1-\alpha\}. \tag{4}$$

Remark on full conformal prediction set

- If (X_i, Y_i), $i = 1, \ldots, n$ are i.i.d., then for an new i.i.d. pair (X_{n+1}, Y_{n+1}),

$$\mathbb{P}\left(Y_{n+1} \in C_{\text{conf}}\left(X_{n+1}\right)\right) \geq 1 - \alpha,$$

$$\mathbb{P}\left(Y_{n+1} \in C_{\text{conf}}\left(X_{n+1}\right)\right) \leq 1 - \alpha + \frac{1}{n+1}$$

Remark on full conformal prediction set

- If $(X_i, Y_i), i = 1, \ldots, n$ are i.i.d., then for an new i.i.d. pair (X_{n+1}, Y_{n+1}),

$$\mathbb{P}\left(Y_{n+1} \in C_{\text{conf}}\left(X_{n+1}\right)\right) \geq 1 - \alpha,$$

$$\mathbb{P}\left(Y_{n+1} \in C_{\text{conf}}\left(X_{n+1}\right)\right) \leq 1 - \alpha + \frac{1}{n+1}$$

- Computationally intensive: For any X_{n+1} and y, in order to tell if y is to be included in $C_{\text{conf}}\left(X_{n+1}\right)$, we **retrain** the model on the augmented data set (which includes the new point (X_{n+1}, y))
$\Rightarrow$ split conformal prediction set

Construction of split conformal prediction set

Algorithm Split Conformal Prediction

Input: Data $(X_i, Y_i), i = 1, \ldots, n$, miscoverage level $\alpha \in (0, 1)$,
regression algorithm $\mathcal{A}$

Output: Prediction band, over $x \in \mathbb{R}^d$

Randomly split $\{1, \ldots, n\}$ into two equal-sized subsets $\mathcal{I}_1, \mathcal{I}_2$

$\widehat{\mu} = \mathcal{A}(\{(X_i, Y_i) : i \in \mathcal{I}_1\})$

$R_i = |Y_i - \widehat{\mu}(X_i)|, i \in \mathcal{I}_2$

$d =$ the k th smallest value in $\{R_i : i \in \mathcal{I}_2\}$, where $k = (n/2 + 1)(1 - \alpha)$

Return $C_{\text{split}}(x) = [\widehat{\mu}(x) - d, \widehat{\mu}(x) + d]$, for all $x \in \mathbb{R}^d$

- If $(X_i, Y_i), i = 1, \ldots, n$ are i.i.d., then for an new i.i.d. draw (X_{n+1}, Y_{n+1}),

$$\mathbb{P}(Y_{n+1} \in C_{\text{split}}(X_{n+1})) \geq 1 - \alpha$$

$$\mathbb{P}(Y_{n+1} \in C_{\text{split}}(X_{n+1})) \leq 1 - \alpha + \frac{2}{n+2}$$

Problem of conformal prediction band

- For split conformal, the width is exactly constant over x.
  For full conformal, the width can vary slightly as x varies
  Solution:
  $$R_i = \frac{|Y_i - \widehat{\mu}(X_i)|}{\widehat{\rho}(X_i)}, i \in \mathcal{I}_2$$
  where the conditional mean $\widehat{\mu}$ and $\widehat{\rho}(x)$ denotes an estimate of the conditional conditional mean absolute deviation (MAD) of $(Y - \mu(X)) \mid X = x$, are fit on the samples in $\mathcal{I}_1$

Problem of conformal prediction band

- For split conformal, the width is exactly constant over x.
  For full conformal, the width can vary slightly as x varies
  Solution:
  $$R_i = \frac{|Y_i - \widehat{\mu}(X_i)|}{\widehat{\rho}(X_i)}, i \in \mathcal{I}_2$$
  where the conditional mean $\widehat{\mu}$ and $\widehat{\rho}(x)$ denotes an estimate of the conditional conditional mean absolute deviation (MAD) of $(Y - \mu(X)) \mid X = x$, are fit on the samples in $\mathcal{I}_1$
- Marginal coverage: $\mathbb{P}(Y_{n+1} \in C_{\text{conf}}(X_{n+1})) \geq 1 - \alpha$
  Much stronger property
  $$\mathbb{P}(Y_{n+1} \in C(x) \mid X_{n+1} = x) \geq 1 - \alpha \text{ for all } x \in \mathbb{R}^d \tag{5}$$

Statistical accuracy

- Base estimator is accurate $\Rightarrow$ the conformal prediction band is near-optimal;
  Base estimator is bad, then we still have valid marginal coverage.
- **Assumption A0:** i.i.d. data (X_i, Y_i) with mean function $\mu(x) = \mathbb{E}(Y \mid X = x), x \in \mathbb{R}^d$.
  Assumption A1: (Independent and symmetric noise) $\epsilon = Y - \mu(X)$

Statistical accuracy

- Base estimator is accurate ⇒ the conformal prediction band is near-optimal;
  Base estimator is bad, then we still have valid marginal coverage.
- **Assumption A0:** i.i.d. data (X_i, Y_i) with mean function $\mu(x) = \mathbb{E}(Y \mid X = x), x \in \mathbb{R}^d$.
  Assumption A1: (Independent and symmetric noise) $\epsilon = Y - \mu(X)$
- Two oracle bands:
  1. Super oracle band:
     $$C_s^*(x) = [\mu(x) - q_\alpha, \mu(x) + q_\alpha]$$
     where q_α is the α upper quantile of $\mathcal{L}(|\epsilon|)$.
     (i) valid conditional coverage (ii) shortest length for both conditional/marginal coverage
  2. Regular oracle band:
     $$C_o^*(x) = [\widehat{\mu}_n(x) - q_{n,\alpha}, \widehat{\mu}_n(x) + q_{n,\alpha}]$$
     where $q_{n,\alpha}$ is the α upper quantile of $\mathcal{L}\left(|Y - \widehat{\mu}_n(X)|\right)$.
     Only offers marginal coverage.

Property: Split conformal prediction bands

- Compare two oracle: estimation error is $\Delta_n(x) = \widehat{\mu}_n(x) - \mu(x)$

$$|q_{n,\alpha} - q_\alpha| \lesssim \mathbb{E}\Delta_n^2(X)$$

- **Assumption A2:** (Sampling stability). For large enough n,

$$\mathbb{P}\left(\|\widehat{\mu}_n - \widetilde{\mu}\|_\infty \geq \eta_n\right) \leq \rho_n,$$

  for some sequences satisfying $\eta_n = o(1), \rho_n = o(1)$ as $n \to \infty$, and some function $\widetilde{\mu}$.

- Theorem(Split conformal approximation of regular oracle): under A0,A1,A2, and the density of $|Y - \widetilde{\mu}(X)|$, is lower bounded away from zero $\nu_{n,\text{ split}}$ denote the split conformal interval's width

$$\nu_{n,\text{split}} - 2q_{n,\alpha} = O_{\mathbb{P}}\left(\rho_n + \eta_n + n^{-1/2}\right)$$

Super Oracle Approximation Under Consistency Assumptions

- Weaker condition than $\mathbb{E}\Delta_n^2(X) = o(1)$
  Assumption A4:(Consistency of base estimator). For n large enough,

$$\mathbb{P}\left(\mathbb{E}_X\left[(\widehat{\mu}_n(X) - \mu(X))^2 \mid \widehat{\mu}_n\right] \geq \eta_n\right) \leq \rho_n,$$

  for some sequences satisfying $\eta_n = o(1), \rho_n = o(1)$ as $n \to \infty$
- Theorem (Split conformal approximation of super oracle): under A0, A1, A4 and $|Y - \mu(X)|$ has density bounded away from zero

$$L\left(C_{n,\text{ split}}(X) \Delta C_s^*(X)\right) = o_{\mathbb{P}}(1)$$

  where $L(A)$ denotes the Lebesgue measure of a set A, and $A \triangle B$ the symmetric difference between sets A, B. Thus, $C_{n,\text{ split}}$ has asymptotic conditional coverage at the level $1 - \alpha$.

Limitation of conformal prediction

- **Assumption A1** require homoscedastic noise.
- Without modeling assumption, it is known to be impossible to construct non-trival prediction intervals with guaranteed conditional coverage.

Theory of CQR

Conformity score: $E_i^{\mathrm{CQR}} = \max\left\{\hat{q}_{\alpha/2}(X_i) - Y_i, Y_i - \hat{q}_{1-\alpha/2}(X_i)\right\}$

$$\hat{C}_\alpha^{\mathrm{CQR}}(X_{n+1}) = \left[\hat{q}_{\alpha/2}(X_{n+1}) - \hat{Q}_{1-\alpha}\left(E^{\mathrm{CQR}}; \mathcal{I}_2\right), \hat{q}_{1-\alpha/2}(X_{n+1}) + \hat{Q}_{1-\alpha}\left(E^{\mathrm{CQR}}; \mathcal{I}_2\right)\right]$$

- i.i.d + regularity + consistency

$$C_\alpha^{\text{oracle}}(X_{n+1}) = \left[q_{\alpha/2}(X_{n+1}), q_{1-\alpha/2}(X_{n+1})\right]$$

$$\mathbb{P}\left[\mathbb{E}\left[\left(\hat{q}_{\alpha/2}(X) - q_{\alpha/2}(X)\right)^2 \mid \hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}\right] \leq \eta_n\right] \geq 1 - \rho_n$$
$$\mathbb{P}\left[\mathbb{E}\left[\left(\hat{q}_{1-\alpha/2}(X) - q_{1-\alpha/2}(X)\right)^2 \mid \hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}\right] \leq \eta_n\right] \geq 1 - \rho_n,$$

for some sequences $\eta_n = o(1)$ and $\rho_n = o(1)$, as $n \to \infty$.

$$L\left(\hat{C}_\alpha(X_{n+1}) \triangle C_\alpha^{\text{oracle}}(X_{n+1})\right) = o_{\mathbb{P}}(1)$$

Candes work—own perspective

1. holdout method
2. conformal under covariate shift
3. conformal hypothesis testing

Jackknife and jackknife+

$$\widehat{q}^{+}_{n,\alpha}\{v_i\} = \text{ the } \lceil(1-\alpha)(n+1)\rceil\text{-th smallest value of } v_1,\ldots,v_n,$$

$$\widehat{q}^{-}_{n,\alpha}\{v_i\} = \text{ the } \lfloor\alpha(n+1)\rfloor\text{-th smallest value of } v_1,\ldots,v_n = -\widehat{q}^{+}_{n,\alpha}\{-v_i\}.$$

- Jackknife (no holdout), no theoretical guarantee

$$\widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}^{+}_{n,\alpha}\{|Y_i - \widehat{\mu}_{-i}(X_i)|\}$$

- Split conformal: train and holdout

$$\widehat{C}^{\text{split,conf}}_{n,\alpha}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}^{+}_{n,\alpha}\{|Y_i - \widehat{\mu}(X_i)|, i \in holdout\}$$

Jackknife+

$$\widehat{C}_{n,\alpha}^{\text{jackknife }+}(X_{n+1}) = \left[\widehat{q}_{n,\alpha}^{-}\left\{\widehat{\mu}_{-i}(X_{n+1}) - R_i^{\mathrm{LOO}}\right\}, \widehat{q}_{n,\alpha}^{+}\left\{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{\mathrm{LOO}}\right\}\right]$$
$$\widehat{C}_{n,\alpha}^{\text{jackknife}}(X_{n+1}) = \left[\widehat{q}_{n,\alpha}^{-}\left\{\widehat{\mu}(X_{n+1}) - R_i^{\mathrm{LOO}}\right\}, \widehat{q}_{n,\alpha}^{+}\left\{\widehat{\mu}(X_{n+1}) + R_i^{\mathrm{LOO}}\right\}\right]$$

- Theorem:

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

- To ease computation, K-fold CV+

$$R_i^{\mathrm{CV}} = \left|Y_i - \widehat{\mu}_{-S_{k(i)}}(X_i)\right|, i = 1, \ldots, n$$

$$\widehat{C}_{n,K,\alpha}^{\mathrm{CV}+}(X_{n+1}) = \left[\widehat{q}_{n,\alpha}^{-}\left\{\widehat{\mu}_{-S_{k(i)}}(X_{n+1}) - R_i^{\mathrm{CV}}\right\}, \widehat{q}_{n,\alpha}^{+}\left\{\widehat{\mu}_{-S_{k(i)}}(X_{n+1}) + R_i^{\mathrm{CV}}\right\}\right]$$

Theorem:

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

Some comparison: theoretical and computational

Method	Assumption-free theory
Split conf. (holdout)	$\geq 1-\alpha$ coverage
Jackknife	No guarantee
Jackknife +	$\geq 1-2\alpha$ coverage
Full conformal	$\geq 1-\alpha$ coverage
K-fold CV+	$\geq 1-2\alpha$ coverage

Method	Model training cost
Split conf. (holdout)	1
Jackknife	n
Jackknife +	n
K-fold CV+	K
Full conformal	$n_{\text{test}} \cdot n_{\text{grid}}$

Conformal prediction under covariate shift: weighted conformal inference [TFBCR19]

$$(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P = P_X \times P_{Y|X}, i = 1, \ldots, n,$$
$$(X_{n+1}, Y_{n+1}) \sim \widetilde{P} = \widetilde{P}_X \times P_{Y|X}, \text{ independently.}$$

- Assume $w(X_i) = \mathrm{d}\widetilde{P}_X(X_i) / \mathrm{d}P_X(X_i)$ is known
- no covariate shift: $\frac{1}{n+1}\sum_{i=1}^{n} \delta_{V_i^{(x,y)}} + \frac{1}{n+1}\delta_{\infty}$
- covariate shift: $\sum_{i=1}^{n} p_i^w(x)\delta_{V_i^{(x,y)}} + p_{n+1}^w(x)\delta_{\infty}$

$$p_i^w(x) = \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j) + w(x)}, i = 1, \ldots, n$$

$$p_{n+1}^w(x) = \frac{w(x)}{\sum_{j=1}^{n} w(X_j) + w(x)}$$

"look exchangable"

Conformal prediction under covariate shift

Distribution-free but with problem-specific assumption, naive bands is not good since we do not take advantage of those assumption.

1. Adaptive conformal inference under distribution shift: online setting [GC22]
2. Conformalized survival analysis [CLR21]: lower predictive bounds on survival times, censoring matter

$$\mathbb{P}\left(T \wedge c_0 \geq \hat{L}(X)\right) \geq 1-\alpha$$

3. Conformal casual inference [LC20]

$$\mathbb{P}\left(Y(1) - Y(0) \in \hat{C}_{\mathrm{ITE}}(X)\right) \geq 1-\alpha$$

Conformal casual inference [LC20]

Testing for outliers with conformal p-value [BCL+21]

- Clean i.i.d training data, given many testing data want to test $\mathcal{H}_{0,i} : X_i \sim P_X$, for any $X_i \in \mathcal{D}^{\text{test}}$

- marginally superuniform (conservative) p-values $\hat{u}^{(\text{marg})}(X_{n+1})$:

$$\mathbb{P}\left[\hat{u}^{(\text{marg })}(X_{n+1}) \leq t\right] \leq t,$$

  under $\mathcal{H}_{0,i}$

- Calibration-conditional conformal p-value, $\mathcal{D} = \mathcal{D}^{\text{train }} \cup \mathcal{D}^{\text{cal}}$

$$\mathbb{P}\left[\mathbb{P}\left[\hat{u}^{(\mathrm{ccv})}(X_{n+1}) \leq t \mid \mathcal{D}\right] \leq t \text{ for all } t \in (0,1)\right] \geq 1-\delta$$

  under $\mathcal{H}_{0,i}$

- Significance: (1)leverage any black-box machine-learning tool (2) control FDR via multiple testing procedure

Functional data–conformal

- The sequence $z_1(\cdot), \ldots, z_n(\cdot)$ consists now of $L^2[0,1]$ functions. The definition of validity for a confidence predictor γ^α is:

$$\mathbb{P}\left(z_{n+1}(t) \in \gamma^\alpha\left(z_1, \ldots, z_n\right)(t) \forall t\right) \geq 1-\alpha \quad \text{for all } P.$$

  To apply conformal prediction, a nonconformity measure is needed. A choice might be:

$$R_i = \int \left(z_i(t) - \bar{z}(t)\right)^2 dt$$

  where $\bar{z}(t)$ is the average of the augmented data set.

Functional data–conformal

- The sequence $z_1(\cdot), \ldots, z_n(\cdot)$ consists now of $L^2[0,1]$ functions. The definition of validity for a confidence predictor γ^α is:

$$\mathbb{P}\left(z_{n+1}(t) \in \gamma^\alpha\left(z_1, \ldots, z_n\right)(t) \forall t\right) \geq 1-\alpha \quad \text{for all } P.$$

To apply conformal prediction, a nonconformity measure is needed. A choice might be:

$$R_i = \int \left(z_i(t) - \bar{z}(t)\right)^2 dt$$

where $\bar{z}(t)$ is the average of the augmented data set. Then, one more step is mandatory. Given a conformal prediction set γ^α, the inherent prediction bands are defined in terms of lower and upper bounds:

$$l(t) = \inf_{z \in \gamma^\alpha} z(t) \quad \text{and} \quad u(t) = \sup_{z \in \gamma^\alpha} z(t).$$

Consequently, thanks to provable conformal properties,

$$\mathbb{P}\left(l(t) \leq z_{n+1}(t) \leq u(t), \forall t\right) \geq 1-\alpha$$

Reference I

- Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia, *Testing for outliers with conformal p-values*, arXiv preprint arXiv:2104.08279 (2021).
- Emmanuel J Candès, Lihua Lei, and Zhimei Ren, *Conformalized survival analysis*, arXiv preprint arXiv:2103.09763 (2021).
- Isaac Gibbs and Emmanuel Candès, *Conformal inference for online prediction with arbitrary distribution shifts*, arXiv preprint arXiv:2208.08401 (2022).
- Lihua Lei and Emmanuel J Candès, *Conformal inference of counterfactuals and individual treatment effects*, arXiv preprint arXiv:2006.06138 (2020).

Reference II

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas, *Conformal prediction under covariate shift*, Advances in neural information processing systems **32** (2019).

Testing for outliers with conformal p-value [BCL+21]

- Clean i.i.d observed data from P_X, given many testing data want to test $\mathcal{H}_{0,i} : X_i \sim P_X$, for any $X_i \in \mathcal{D}^{\text{test}}$

- marginally superuniform (conservative) p-values $\hat{u}^{(\text{marg})}(X_{n+1})$:

$$\mathbb{P}\left[\hat{u}^{(\text{marg })}(X_{n+1}) \le t\right] \le t,$$

  under $\mathcal{H}_{0,i}$

- Calibration-conditional conformal p-value, $\mathcal{D} = \mathcal{D}^{\text{train}} \cup \mathcal{D}^{\text{cal}}$

$$\mathbb{P}\left[\mathbb{P}\left[\hat{u}^{(\text{ccv})}(X_{n+1}) \le t \mid \mathcal{D}\right] \le t \text{ for all } t \in (0,1)\right] \ge 1 - \delta$$

  under $\mathcal{H}_{0,i}$

- Significance: (1)leverage any black-box machine-learning tool (2) control FDR via multiple testing procedure

Functional data–conformal

- The sequence $z_1(\cdot), \ldots, z_n(\cdot)$ consists now of $L^2[0,1]$ functions. The definition of validity for a confidence predictor γ^α is:

$$\mathbb{P}\left(z_{n+1}(t) \in \gamma^\alpha\left(z_1, \ldots, z_n\right)(t) \forall t\right) \geq 1-\alpha \quad \text{for all } P.$$

  To apply conformal prediction, a nonconformity measure is needed. A choice might be:

$$R_i = \int \left(z_i(t) - \bar{z}(t)\right)^2 dt$$

  where $\bar{z}(t)$ is the average of the augmented data set.

Functional data–conformal

- The sequence $z_1(\cdot), \ldots, z_n(\cdot)$ consists now of $L^2[0,1]$ functions. The definition of validity for a confidence predictor γ^α is:

$$\mathbb{P}\left(z_{n+1}(t) \in \gamma^\alpha\left(z_1, \ldots, z_n\right)(t) \forall t\right) \geq 1-\alpha \quad \text{for all } P.$$

To apply conformal prediction, a nonconformity measure is needed. A choice might be:

$$R_i = \int \left(z_i(t) - \bar{z}(t)\right)^2 dt$$

where $\bar{z}(t)$ is the average of the augmented data set. Then, one more step is mandatory. Given a conformal prediction set γ^α, the inherent prediction bands are defined in terms of lower and upper bounds:

$$l(t) = \inf_{z \in \gamma^\alpha} z(t) \quad \text{and} \quad u(t) = \sup_{z \in \gamma^\alpha} z(t).$$

Consequently, thanks to provable conformal properties,

$$\mathbb{P}\left(l(t) \leq z_{n+1}(t) \leq u(t), \forall t\right) \geq 1-\alpha$$

Statistical accuracy

- Base estimator is accurate $\Rightarrow$ the conformal prediction band is near-optimal;
  Base estimator is bad, then we still have valid marginal coverage.
- **Assumption A0:** i.i.d. data (X_i, Y_i) with mean function $\mu(x) = \mathbb{E}(Y \mid X = x), x \in \mathbb{R}^d$.
  Assumption A1: (Independent and symmetric noise) $\epsilon = Y - \mu(X)$

Statistical accuracy

- Base estimator is accurate $\Rightarrow$ the conformal prediction band is near-optimal;
  Base estimator is bad, then we still have valid marginal coverage.
- **Assumption A0:** i.i.d. data (X_i, Y_i) with mean function $\mu(x) = \mathbb{E}(Y \mid X = x), x \in \mathbb{R}^d$.
  Assumption A1: (Independent and symmetric noise) $\epsilon = Y - \mu(X)$
- Two oracle bands:
  1. Super oracle band:
  $$C_s^*(x) = [\mu(x) - q_\alpha, \mu(x) + q_\alpha]$$
  where q_α is the α upper quantile of $\mathcal{L}(|\epsilon|)$.
  (i) valid conditional coverage (ii) shortest length for both conditional/marginal coverage
  2. Regular oracle band:
  $$C_o^*(x) = [\widehat{\mu}_n(x) - q_{n,\alpha}, \widehat{\mu}_n(x) + q_{n,\alpha}]$$
  where $q_{n,\alpha}$ is the α upper quantile of $\mathcal{L}\left(|Y - \widehat{\mu}_n(X)|\right)$.
  Only offers marginal coverage.

Property: Split conformal prediction bands

- Compare two oracle: estimation error is $\Delta_n(x) = \widehat{\mu}_n(x) - \mu(x)$

$$|q_{n,\alpha} - q_\alpha| \lesssim \mathbb{E}\Delta_n^2(X)$$

- **Assumption A2:** (Sampling stability). For large enough n,

$$\mathbb{P}\left(\|\widehat{\mu}_n - \widetilde{\mu}\|_\infty \geq \eta_n\right) \leq \rho_n,$$

  for some sequences satisfying $\eta_n = o(1), \rho_n = o(1)$ as $n \to \infty$, and some function $\widetilde{\mu}$.

- Theorem(Split conformal approximation of regular oracle): under A0,A1,A2, and the density of $|Y - \widetilde{\mu}(X)|$, is lower bounded away from zero $\nu_{n,\text{ split}}$ denote the split conformal interval's width

$$\nu_{n,\text{split}} - 2q_{n,\alpha} = O_{\mathbb{P}}\left(\rho_n + \eta_n + n^{-1/2}\right)$$

Super Oracle Approximation Under Consistency Assumptions

- Weaker condition than $\mathbb{E}\Delta_n^2(X) = o(1)$
  Assumption A4:(Consistency of base estimator). For n large enough,

$$\mathbb{P}\left(\mathbb{E}_X\left[(\widehat{\mu}_n(X) - \mu(X))^2 \mid \widehat{\mu}_n\right] \geq \eta_n\right) \leq \rho_n,$$

  for some sequences satisfying $\eta_n = o(1), \rho_n = o(1)$ as $n \to \infty$
- Theorem (Split conformal approximation of super oracle): under A0, A1, A4 and $|Y - \mu(X)|$ has density bounded away from zero

$$L\left(C_{n,\text{ split}}\,(X) \Delta C_s^*(X)\right) = o_{\mathbb{P}}(1)$$

  where $L(A)$ denotes the Lebesgue measure of a set A, and $A \triangle B$ the symmetric difference between sets A, B. Thus, $C_{n,\text{ split}}$ has asymptotic conditional coverage at the level $1 - \alpha$.

Limitation of conformal prediction

- Without modeling assumption, it is known to be impossible to construct non-trival prediction intervals with guaranteed conditional coverage.

Theory of CQR

Conformity score: $E_i^{\mathrm{CQR}} = \max\left\{\hat{q}_{\alpha/2}\left(X_i\right) - Y_i, Y_i - \hat{q}_{1-\alpha/2}\left(X_i\right)\right\}$

$$\hat{C}_\alpha^{\mathrm{CQR}}\left(X_{n+1}\right) = \left[\hat{q}_{\alpha/2}\left(X_{n+1}\right) - \hat{Q}_{1-\alpha}\left(E^{\mathrm{CQR}}; \mathcal{I}_2\right), \hat{q}_{1-\alpha/2}\left(X_{n+1}\right) + \hat{Q}_{1-\alpha}\left(E^{\mathrm{CQR}}; \mathcal{I}_2\right)\right]$$

- i.i.d + regularity + consistency

$$C_\alpha^{\text{oracle}}\left(X_{n+1}\right) = \left[q_{\alpha/2}\left(X_{n+1}\right), q_{1-\alpha/2}\left(X_{n+1}\right)\right]$$

$$\mathbb{P}\left[\mathbb{E}\left[\left(\hat{q}_{\alpha/2}(X) - q_{\alpha/2}(X)\right)^2 \mid \hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}\right] \leq \eta_n\right] \geq 1 - \rho_n$$

$$\mathbb{P}\left[\mathbb{E}\left[\left(\hat{q}_{1-\alpha/2}(X) - q_{1-\alpha/2}(X)\right)^2 \mid \hat{q}_{\alpha/2}, \hat{q}_{1-\alpha/2}\right] \leq \eta_n\right] \geq 1 - \rho_n,$$

for some sequences $\eta_n = o(1)$ and $\rho_n = o(1)$, as $n \to \infty$.

$$L\left(\hat{C}_\alpha\left(X_{n+1}\right) \triangle C_\alpha^{\text{oracle}}\left(X_{n+1}\right)\right) = o_{\mathbb{P}}(1)$$