Tutorial:

Quantifying Predictive Uncertainty Without Distributional Assumptions Via Conformal Prediction

Rina Foygel Barber

http://rinafb.github.io/

Distribution-free inference: questions

Why do we want "distribution-free" guarantees?

When we analyze data, we...

- Run a model/algorithm that is valid under certain assumptions
  (parametric model / smoothness conditions / sparsity assumption / ...)
- But if the assumptions don't hold, can we trust the output?
  (parameter estimate / predicted value / error bound / hypothesis test / ...)

Distribution-free inference: questions

Why do we want "distribution-free" guarantees?

When we analyze data, we...

- Run a model/algorithm that is valid under certain assumptions
  (parametric model / smoothness conditions / sparsity assumption / ...)
- But if the assumptions don't hold, can we trust the output?
  (parameter estimate / predicted value / error bound / hypothesis test / ...)
- So, we run a test to check if the assumptions hold
  (goodness of fit / overdispersion / calibration ...)
- But, what if this test is only guaranteed to detect violations,
  under some other assumptions?

Distribution-free inference: questions

Why do we want "distribution-free" guarantees?

When we analyze data, we...

- Run a model/algorithm that is valid under certain assumptions
  (parametric model / smoothness conditions / sparsity assumption / ...)
- But if the assumptions don't hold, can we trust the output?
  (parameter estimate / predicted value / error bound / hypothesis test / ...)
- So, we run a test to check if the assumptions hold
  (goodness of fit / overdispersion / calibration ...)
- But, what if this test is only guaranteed to detect violations,
  under some other assumptions?

The goal of distribution-free inference is to provide guarantees
that are valid universally over all data distributions.

Distribution-free inference: questions

What are inference questions we might want to ask, distribution-free?

- Prediction: the unobserved response Y will lie in [some range]
- Effect size: the dependence between X and Y lies in [some range]
- Independence: test if X & Y independent given [some confounders]
- Regression: the distribution of Y given X satisfies [some property]

Distribution-free inference: questions

What are inference questions we might want to ask, distribution-free?

- Prediction: the unobserved response Y will lie in [some range]

this talk

- Effect size: the dependence between X and Y lies in [some range]
- Independence: test if X & Y independent given [some confounders]
- Regression: the distribution of Y given X satisfies [some property]

The prediction problem

Setting:

- Training data $(X_1, Y_1), \ldots, (X_n, Y_n)$, test point (X_{n+1}, Y_{n+1})

  observed want to predict

- If fitted model $\widehat{\mu}$ overfits to training data,

$$|Y_{n+1} - \widehat{\mu}(X_{n+1})| \gg \frac{1}{n} \sum_{i=1}^{n} |Y_i - \widehat{\mu}(X_i)|$$

  even if training & test data are from the same distribution

The prediction problem

Run algorithm $\mathcal{A}$ on the training data $\rightsquigarrow$ fitted model $\widehat{\mu}$

Prediction interval for Y_{n+1}:

$$\widehat{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \text{(margin of error)}$$

Use training residuals? ("naive")

Use a parametric model?

Use smoothness assumptions?

Use cross-validation?

Using a holdout set

- Using any algorithm, fit model

$$\widehat{\mu} = \mathcal{A}\Big((X_1, Y_1), \ldots, (X_{n/2}, Y_{n/2})\Big)$$

- Compute holdout residuals

$$R_i = |Y_i - \widehat{\mu}(X_i)|, \quad i = n/2 + 1\,, \ \ldots\ , \ n$$

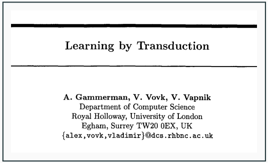
- Prediction interval:

$$\widehat{C}(X_{n+1}) \ = \ \widehat{\mu}(X_{n+1}) \ \pm \ \big(\text{the } (1-\alpha)\text{-quantile of } R_{n/2+1}, \ldots, R_n\big)$$

Conformal prediction framework

Background on the conformal prediction (CP) framework:

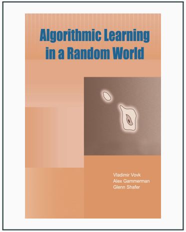
key idea = statistical inference via exchangeability of the data



Learning by Transduction

A. Gammerman, V. Vovk, V. Vapnik
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, UK
{alex,vovk,vladimir}@dcs.rhbnc.ac.uk

Gammerman, Vovk, Vapnik
UAI 1998



Vovk, Gammerman, Shafer
2005 — see `alrw.net`

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 523, 1094–1111, Theory and Methods
https://doi.org/10.1080/01621459.2017.1307116

Taylor & Francis
Taylor & Francis Group

Check for updates

Distribution-Free Predictive Inference for Regression

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA

ABSTRACT
We develop a general framework for distribution-free predictive inference in regression, using conformal inference. The proposed methodology allows for the construction of a prediction band for the response variable using any estimator of the regression function. The resulting prediction band preserves the consistency properties of the original estimator under standard assumptions, while guaranteeing finite-sample marginal coverage even when these assumptions do not hold. We analyze and compare, both empirically and theoretically, the two major variants of our conformal framework: full conformal inference and split conformal inference, along with a related jackknife method. These methods offer different tradeoffs between statistical accuracy (length of resulting prediction intervals) and computational efficiency. As extensions, we develop a method for constructing valid in-sample prediction intervals called *rank-one-out* conformal inference, which has essentially the same computational efficiency as split conformal inference. We also describe an extension of our procedures for producing prediction bands with locally varying length, to adapt to heteroscedasticity in the data. Finally, we propose a model-free notion of variable importance, called *leave-one-covariate-out* or LOCO inference. Accompanying this article is an R package `conformalInference` that implements all of the proposals we have introduced. In the spirit of reproducibility, all of our empirical results can also be easily (re)generated using this package.

ARTICLE HISTORY
Received April 2016
Revised February 2017

KEYWORDS
Distribution-free; Model misspecification; Prediction band; Regression; Variable importance

Lei, G'Sell, Rinaldo,
Tibshirani, Wasserman
JASA 2018

Conformal prediction framework

What is exchangeability?

$(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ & $(X_{\sigma(1)}, Y_{\sigma(1)}), \ldots, (X_{\sigma(n+1)}, Y_{\sigma(n+1)})$
have the same joint distribution for any permutation σ

Equivalently:
Given an *unordered* data set, any ordering is equally likely

Conformal prediction framework

What is exchangeability?

$(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ & $(X_{\sigma(1)}, Y_{\sigma(1)}), \ldots, (X_{\sigma(n+1)}, Y_{\sigma(n+1)})$
have the same joint distribution for any permutation σ

Equivalently:
Given an *unordered* data set, any ordering is equally likely

Examples:

- (X_i, Y_i)'s are i.i.d. from any distribution
- (X_i, Y_i)'s sampled uniformly without replacement from any set
- Not an example: a stationary time series w/ dependence

Split conformal prediction

Split conformal prediction interval (a.k.a. holdout):[1]

$$\widehat{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\Big\{R_{n/2+1}, \ldots, R_n\Big\}$$

the $\lceil (1-\alpha)(n/2+1) \rceil$-th smallest value in the list

Theorem:

If $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exchangeable (e.g., i.i.d.), then for any algorithm $\mathcal{A}$, the split conformal method satisfies

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

[1] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Split conformal prediction

Proof:

After conditioning on $\widehat{\mu}$, holdout + test data is exchangeable

$\Rightarrow$ residuals $R_{n/2+1}, \ldots, R_n, R_{n+1}$ are exchangeable

$\Rightarrow \mathbb{P}\left\{R_{n+1} \leq \left(\text{the } (1-\alpha)\text{-quantile of } R_{n/2+1}, \ldots, R_{n+1}\right)\right\} \geq 1-\alpha$

Split conformal prediction

Proof:

After conditioning on $\widehat{\mu}$, holdout + test data is exchangeable

$\Rightarrow$ residuals $R_{n/2+1}, \ldots, R_n, R_{n+1}$ are exchangeable

$$\Rightarrow \mathbb{P}\left\{R_{n+1} \leq \left(\text{the } (1-\alpha)\text{-quantile of } R_{n/2+1}, \ldots, R_{n+1}\right)\right\} \geq 1-\alpha$$

$$\Updownarrow$$

$$R_{n+1} \leq \mathrm{Q}_{1-\alpha}\{R_{n/2+1}, \ldots, R_n\}$$

$$\Updownarrow$$

$$Y_{n+1} \in \widehat{C}(X_{n+1})$$

The nonconformity score

In the above construction,

$$\widehat{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm [...] \; = \{ \text{ all } y \text{ values with } |y - \widehat{\mu}(X_{n+1})| \leq [...] \; \}$$

[2] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

The nonconformity score

In the above construction,

$$\widehat{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm [...] \ = \{ \text{ all } y \text{ values with } |y - \widehat{\mu}(X_{n+1})| \leq [...] \ \}$$

We can generalize to any score function:[2]

$$\widehat{C}(X_{n+1}) = \{ \text{ all } y \text{ values with } \widehat{S}(X_{n+1}, y) \leq [...] \ \}$$

where $\widehat{S}(x, y)$ measures "nonconformity" of the data point (x, y)

$\widehat{S}$ may be called the "nonconformity score", or the "conformity score"

[2] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

The nonconformity score

Split conformal with an arbitrary nonconformity score:

- Using data $i = 1, \ldots, n/2$, fit nonconformity score function $\widehat{S}$
- Compute $S_i = \widehat{S}(X_i, Y_i)$ for $i = n/2 + 1, \ldots, n$
- Prediction interval:
  $\widehat{C}(X_{n+1}) = \{y : \widehat{S}(X_{n+1}, y) \leq \mathrm{Q}_{1-\alpha}\{S_{n/2+1}, \ldots, S_n\}\}$

The nonconformity score

Split conformal with an arbitrary nonconformity score:

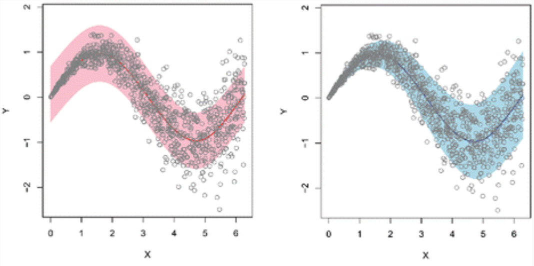
- Using data $i = 1, \ldots, n/2$, fit nonconformity score function $\widehat{S}$
- Compute $S_i = \widehat{S}(X_i, Y_i)$ for $i = n/2 + 1, \ldots, n$
- Prediction interval:
  $\widehat{C}(X_{n+1}) = \{y : \widehat{S}(X_{n+1}, y) \leq \mathrm{Q}_{1-\alpha}\{S_{n/2+1}, \ldots, S_n\}\}$

Choose $\widehat{S}(x, y) = |y - \widehat{\mu}(x)| \rightsquigarrow \widehat{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm [\ldots]$ as before

The nonconformity score: examples

If noise level varies with X, may want varying interval width:[3]

$$\widehat{S}(x,y) = \frac{|y - \widehat{\mu}(x)|}{\widehat{\sigma}(x)} \quad \Rightarrow \quad \widehat{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{\sigma}(X_{n+1}) \cdot \mathrm{Q}_{1-\alpha}\{\dots\}$$



(figure from Lei et al 2018)

[3] Lei et al 2018, *Distribution-Free Predictive Inference for Regression*

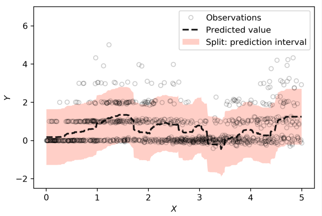
The nonconformity score: examples

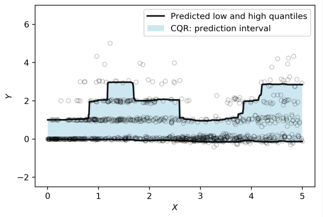
If the shape of distrib. of $Y|X$ varies with X,
centering $\widehat{C}(X_{n+1})$ at $\widehat{\mu}(X_{n+1})$ may not be optimal

Instead, can estimate conditional quantiles directly:[4,5]

- Estimate $\widehat{q}_{\alpha/2}$, $\widehat{q}_{1-\alpha/2}$ on the training set
- Nonconformity score: $\widehat{S}(x,y) = \max\{\widehat{q}_{\alpha/2}(x) - y,\ y - \widehat{q}_{1-\alpha/2}(x)\}$

$$\Rightarrow\ \widehat{C}(X_{n+1}) = \left[\widehat{q}_{\alpha/2}(X_{n+1}) - \mathrm{Q}_{1-\alpha}\{\ldots\},\ \widehat{q}_{1-\alpha/2}(X_{n+1}) + \mathrm{Q}_{1-\alpha}\{\ldots\}\right]$$





(figure from Romano et al 2019)

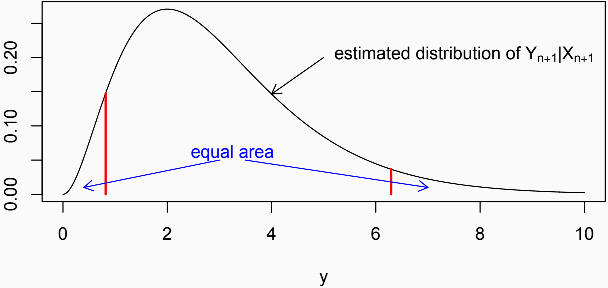
[4] Romano et al 2019, *Conformalized quantile regression*

The nonconformity score: examples

Or, the score can directly use the estimated distribution of $Y|X$:[6]

- Estimate the conditional CDF, $\widehat{F}(y|x)$, on training set
- Nonconformity score: $\widehat{S}(x,y) = \left|\widehat{F}(y|x) - 0.5\right|$

$$\Rightarrow \quad \widehat{C}(X_{n+1}) = \{y : \widehat{F}(y|X_{n+1}) \in 0.5 \pm \mathrm{Q}_{1-\alpha}\{\ldots\}\}$$



[6] Chernozhukov et al 2019, *Distributional conformal prediction*

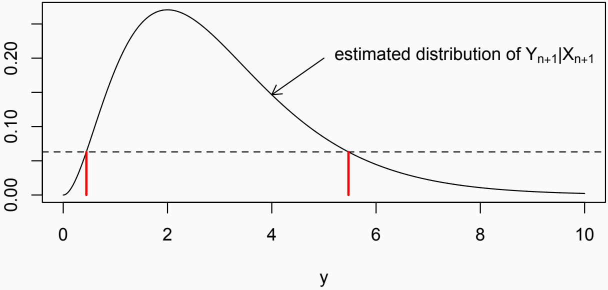
The nonconformity score: examples

An alternative:[7]

- Estimate the conditional density, $\widehat{f}(y|x)$, on training set
- Nonconformity score = two-tailed test:

$$\widehat{S}(x,y) = -\widehat{f}(y|x)$$

$$\Rightarrow \ \widehat{C}(X_{n+1}) = \{y : \widehat{f}(y|X_{n+1}) \geq -Q_{1-\alpha}\{...\}\}$$



[7] Izbicki et al 2020, *Flexible distribution-free conditional predictive bands using density estimators*

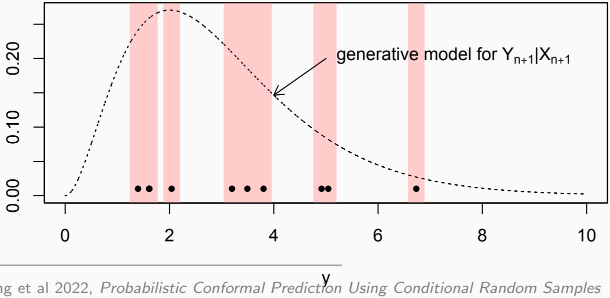
The nonconformity score: examples

Probabilistic conformal prediction—use a generative model:[8]

- Fit a generative model for $Y|X$, on the training data
- Given X_i, draw samples $\hat{Y}_{i,1}, \ldots, \hat{Y}_{i,K}$ from the generative model
- Nonconformity score:

$$\widehat{S}(X_i, y) = \min_{k=1,\ldots,K} \text{distance}(y, \hat{Y}_{ik})$$

$$\Rightarrow \widehat{C}(X_{n+1}) = \{y : \text{distance}(y, \hat{Y}_{n+1,k}) \leq \text{Q}_{1-\alpha}\{\ldots\} \text{ for any } k\}$$



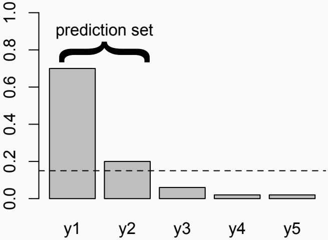
[8] Wang et al 2022, *Probabilistic Conformal Prediction Using Conditional Random Samples*

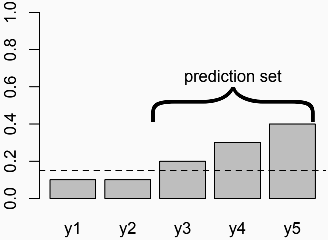
Categorical response variable

If the response Y is categorical, with values $\mathcal{Y} = \{y_1, \ldots, y_K\}$—

- $\hat{p}_k(x)$ estimates $\mathbb{P}\{Y = y_k \mid X = x\}$ using training data
- A natural score function: $\widehat{S}(x, y_k) = -\hat{p}_k(x)$

$\Rightarrow \widehat{C}(X_{n+1}) = \{y : \hat{p}_k(X_{n+1}) \geq -\mathrm{Q}_{1-\alpha}\{...\}$ for any $k\}$



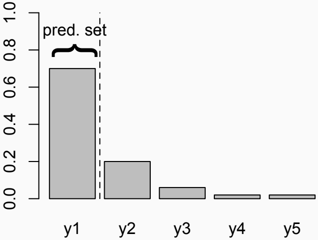


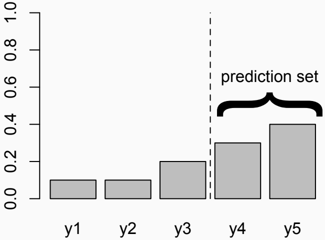
Categorical response variable

A more efficient construction:[9]

- How far into the tail of the distribution, is the label Y?

$$\widehat{S}(x, y_k) = \sum_{k'} \hat{p}_{k'}(x) \cdot \mathbf{1}\{\hat{p}_{k'}(x) \geq \hat{p}_k(x)\}$$





[9] Podkopaev & Ramdas 2021, *Distribution-free uncertainty quantification for classification under label shift*

Split vs full conformal prediction

All methods so far rely on data splitting:

- Training: use $n/2$ data points to develop a score function $\widehat{S}$
- Calibration: use $n/2$ data points to learn the distrib. of $\widehat{S}(X,Y)$
- Then we can predict $\widehat{S}(X_{n+1},Y_{n+1}) \rightsquigarrow$ can predict Y_{n+1}

Split vs full conformal prediction

All methods so far rely on data splitting:

- Training: use $n/2$ data points to develop a score function $\widehat{S}$
- Calibration: use $n/2$ data points to learn the distrib. of $\widehat{S}(X, Y)$
- Then we can predict $\widehat{S}(X_{n+1}, Y_{n+1}) \rightsquigarrow$ can predict Y_{n+1}

The drawback: sample splitting means that we only use $n/2$ data points to fit the model / the score function

Split vs full conformal prediction

Naive method vs Holdout method

$$\underbrace{\widehat{\mu}(X_{n+1})}_{\substack{\text{fitted on } n \text{ points} \\ \Rightarrow \text{ more accurate}}} \pm \underbrace{\substack{\text{training resid.} \\ \text{quantile}}}_{\substack{\text{too small} \\ \text{(overfitted)}}} \qquad \underbrace{\widehat{\mu}(X_{n+1})}_{\substack{\text{fitted on } n/2 \text{ points} \\ \Rightarrow \text{ less accurate}}} \pm \underbrace{\substack{\text{holdout resid.} \\ \text{quantile}}}_{\substack{\text{calibrated} \\ \text{(but wider)}}}$$

[10] Vovk, Gammerman, Shafer 2005

Split vs full conformal prediction

Naive method vs Holdout method

$$\underbrace{\widehat{\mu}(X_{n+1})}_{\substack{\text{fitted on } n \text{ points} \\ \Rightarrow \text{ more accurate}}} \pm \underbrace{\substack{\text{training resid.} \\ \text{quantile}}}_{\substack{\text{too small} \\ \text{(overfitted)}}} \qquad \underbrace{\widehat{\mu}(X_{n+1})}_{\substack{\text{fitted on } n/2 \text{ points} \\ \Rightarrow \text{ less accurate}}} \pm \underbrace{\substack{\text{holdout resid.} \\ \text{quantile}}}_{\substack{\text{calibrated} \\ \text{(but wider)}}}$$

An alternative—the *full conformal* method:[10]

- Models fitted on all n training samples (no data splitting)
- Guaranteed distribution-free predictive coverage
- High computational cost

[10] Vovk, Gammerman, Shafer 2005

Split vs full conformal prediction

$$\text{Terminology: } \begin{cases} \text{holdout method / split conformal / inductive conformal} \\ \text{conformal / full conformal / transductive conformal} \end{cases}$$

Full conformal prediction

- Fit model to training+test data

$$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}))$$

- Compute residuals

$$R_i = |Y_i - \widehat{\mu}(X_i)| \text{ for } i \leq n; \;\; R_{n+1} = |Y_{n+1} - \widehat{\mu}(X_{n+1})|$$

- Check if $R_{n+1} \leq \Big[(1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\Big]$

Full conformal prediction

- Fit model to training+test data

$$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}))$$

- Compute residuals

$$R_i = |Y_i - \widehat{\mu}(X_i)| \text{ for } i \leq n; \;\; R_{n+1} = |Y_{n+1} - \widehat{\mu}(X_{n+1})|$$

- Check if $R_{n+1} \leq \Big[(1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\Big]$

If data points are exchangeable, and $\mathcal{A}$ treats data points symmetrically,
then $R_1, \ldots, R_{n+1}$ are exchangeable
$\Rightarrow$ this event has $\geq 1 - \alpha$ probability

Full conformal prediction

- Fit model to training+test data

$$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, \overset{y}{\cancel{Y_{n+1}}}))$$

- Compute residuals

$$R_i = |Y_i - \widehat{\mu}(X_i)| \text{ for } i \leq n; \;\; R_{n+1} = |\overset{y}{\cancel{Y_{n+1}}} - \widehat{\mu}(X_{n+1})|$$

- Check if $R_{n+1} \leq \Big[(1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\Big]$

If data points are exchangeable, and $\mathcal{A}$ treats data points symmetrically, then $R_1, \ldots, R_{n+1}$ are exchangeable

$\Rightarrow$ this event has $\geq 1 - \alpha$ probability if we plug in $y = Y_{n+1}$

Full conformal prediction

- Fit model to training+test data
$$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$$
- Compute residuals
$$R_i = |Y_i - \widehat{\mu}(X_i)|, \; i = 1, \ldots, n, \quad R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$
- Check if $R_{n+1} \leq \Big[(1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\Big]$

Full conformal prediction

- Fit model to training+test data
$$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$$
- Compute residuals
$$R_i = |Y_i - \widehat{\mu}(X_i)|, \; i = 1, \ldots, n, \quad R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$
- Check if $R_{n+1} \leq \left[(1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\right]$

$y \rightsquigarrow \{\text{Yes}, \text{No}\}$

Full conformal prediction

Test value $y \in \mathbb{R}$

- Fit model to training+test data
  $$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$$
- Compute residuals
  $$R_i = |Y_i - \widehat{\mu}(X_i)|, \; i = 1, \ldots, n, \quad R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$
- Check if $R_{n+1} \leq \left[(1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\right]$

$$y \rightsquigarrow \{\text{Yes}, \text{No}\}$$

if Yes: add $y \in \widehat{C}(X_{n+1})$

if No: discard y

Full conformal prediction

Test value $y \in \mathbb{R}$

- Fit model to training+test data
$$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$$
- Compute residuals
$$R_i = |Y_i - \widehat{\mu}(X_i)|, \; i = 1, \ldots, n, \quad R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$
- Check if $R_{n+1} \leq \left[(1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\right]$

$y \rightsquigarrow \{\text{Yes}, \text{No}\}$

if Yes: add $y \in \widehat{C}(X_{n+1})$

if No: discard y

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1})\right\} = \mathbb{P}\{\text{for test value } y = Y_{n+1}, \text{ answer is Yes}\} \geq 1 - \alpha$$

Full conformal prediction

Validity guarantee for full conformal:[11]

Theorem:

If $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable (e.g., i.i.d.), and the algorithm $\mathcal{A}$ treats data points symmetrically, then full CP satisfies

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

- Split conformal can be viewed as a special case.

[11] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

Full conformal prediction

Full conformal can be run with any score function *on data sets*:

$$\big((x_1, y_1), \dots, (x_{n+1}, y_{n+1})\big) \mapsto (S_1, \dots, S_{n+1})$$

$S_i =$ "nonconformity score" of data point i, relative to rest of the data

Full conformal prediction

Full conformal can be run with any score function *on data sets*:

$$\big((x_1, y_1), \dots, (x_{n+1}, y_{n+1})\big) \mapsto (S_1, \dots, S_{n+1})$$

S_i = "nonconformity score" of data point i, relative to rest of the data

- Regression: $S_i = |y_i - \widehat{\mu}(x_i)|$ or $S_i = |y_i - \widehat{\mu}(x_i)|/\widehat{\sigma}(x_i)|$
- Quantile regr.: S_i compares y_i to $\widehat{q}_{\alpha/2}(x_i)$ & $\widehat{q}_{1-\alpha/2}(x_i)$
- Classification: $S_i = -\widehat{p}(y_i|x_i)$
- & many more

Full conformal: computational challenges

Full conformal prediction requires that the algorithm $\mathcal{A}$ is re-run:

- For each test value X_{n+1} of interest
- For every possible value of Y_{n+1} (e.g, all $y \in \mathbb{R}$)

Approaches:

- In practice — restrict to a grid of y values (but no theory)
- Specialized methods for specific algorithms e.g. Lasso[12]
- Discretized CP — use a discretized version of $\mathcal{A}$ to restore theoretical guarantees[13]

[12] Lei 2017, *Fast Exact Conformalization of Lasso using Piecewise Linear Homotopy*

[13] Chen, Chun, & B. 2017, *Discretized conformal prediction for efficient distribution-free inference*

Full conformal: computational challenges

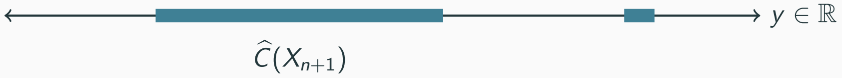
A preliminary observation:[14]
It is valid to run CP on the interval $[\min_{1\leq i\leq n} Y_i,\ \max_{1\leq i\leq n} Y_i]$

- With prob. $\geq 1 - \frac{2}{n+1}$, including Y_{n+1} doesn't change the endpoints
- So, coverage is $\geq 1 - \alpha - \frac{2}{n+1}$

[14] Chen, Wang, Ha, & B. 2016, *Trimmed conformal prediction for high-dimensional models.*

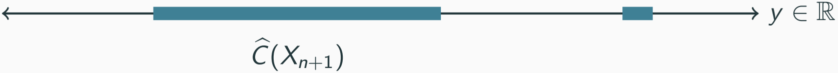
Full conformal: computational challenges

Conformal prediction:



Full conformal: computational challenges

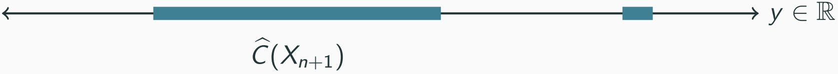
Conformal prediction:



Conformal prediction with rounding (informal version):

Full conformal: computational challenges

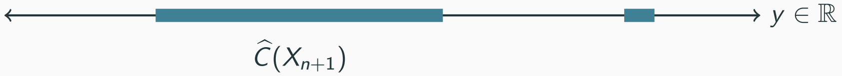
Conformal prediction:



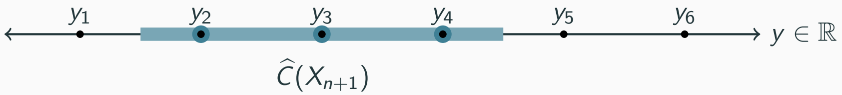
Conformal prediction with rounding (informal version):

Full conformal: computational challenges

Conformal prediction:

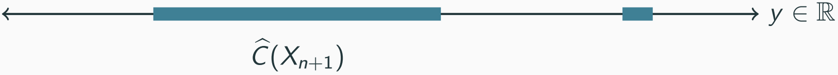


Conformal prediction with rounding (informal version):

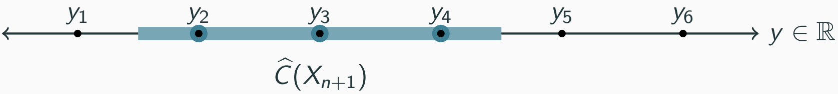


Full conformal: computational challenges

Conformal prediction:



Conformal prediction with rounding (informal version):



Problems to solve:

- Theory to guarantee coverage rate $1 - \alpha$?
- Avoid wider intervals due to discretized grid?

Full conformal: computational challenges

Why do we lose the coverage guarantee?

- If only fit $\widehat{\mu}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$ for y in a grid...
- Equivalent to: fit $\widehat{\mu}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, [Y_{n+1}])$

Y_{n+1} rounded to grid

Full conformal: computational challenges

Why do we lose the coverage guarantee?

- If only fit $\widehat{\mu}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$ for y in a grid...
- Equivalent to: fit $\widehat{\mu}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, [Y_{n+1}])$

Y_{n+1} rounded to grid

To maintain exchangeability:
need to fit $\widehat{\mu}$ on $(X_1, [Y_1]), \ldots, (X_n, [Y_n]), (X_{n+1}, \ y \ )$

Full conformal: computational challenges

Discretizing the model to restore theoretical guarantees:[15]

[15] Chen, Chun, & B. 2017, *Discretized conformal prediction for efficient distribution-free inference*

Full conformal: computational challenges

Discretizing the model to restore theoretical guarantees:[15]

- Run a discretized algorithm for model fitting:

$$(X_1, Y_1), \ldots, (X, y) \xrightarrow{[\ ]} (X_1, [Y_1]), \ldots, (X, [y]) \xrightarrow{\text{fit model } \widehat{\mu}} [\widehat{\mu}]$$

- Calculate residuals

$$R_i = |Y_i - [\widehat{\mu}](X_i)|, \ i = 1, \ldots, n, \quad R_{n+1} = |y - [\widehat{\mu}](X)|$$

- Check if $|R_{n+1}| \leq \Big[(1-\alpha) \text{ quantile of } R_1, \ldots, R_n, R_{n+1}\Big]$

Computational cost: $\mathcal{A}$ only needs to be rerun for each y in the grid

[15]Chen, Chun, & B. 2017, *Discretized conformal prediction for efficient distribution-free inference*

Cross-validation methods

Holdout methods	vs	full conformal
(lose sample size)		(high computational cost)

Cross-validation methods

Holdout methods vs full conformal
(lose sample size) (high computational cost)

To avoid this tradeoff, can we use cross-validation?

Split data into k folds, $\{1, \ldots, n\} = A_1 \cup \cdots \cup A_K$

For $i \in A_k$, $R_i^{\mathrm{CV}} = |Y_i - \widehat{\mu}_{-A_k}(X_i)|$ $\leftarrow \widehat{\mu}_{-A_k}$ is trained on data pts $\{1, \ldots, n\} \backslash A_k$

$$\widehat{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\left\{R_1^{\mathrm{CV}}, \ldots, R_n^{\mathrm{CV}}\right\}$$

the $\lceil (1-\alpha)(n+1) \rceil$-th smallest value in the list

— Computational cost: $K+1$ regressions

— Problem: theory from holdout setting no longer holds

Cross-validation methods

Jackknife a.k.a. leave-one-out cross-validation ($K = n$)

Residuals $R_i^{\text{LOO}} = |Y_i - \widehat{\mu}_{-i}(X_i)|$ $\leftarrow$ $\widehat{\mu}_{-i}$ is trained on data pts $\{1, \ldots, n\} \backslash \{i\}$

$$\widehat{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\left\{R_1^{\text{LOO}}, \ldots, R_n^{\text{LOO}}\right\}$$

[16]Steinberger & Leeb 2018, *Conditional predictive inference for high-dimensional stable algorithms*

Cross-validation methods

Jackknife a.k.a. leave-one-out cross-validation ($K = n$)

Residuals $R_i^{\text{LOO}} = |Y_i - \widehat{\mu}_{-i}(X_i)|$ $\leftarrow$ $\widehat{\mu}_{-i}$ is trained on data pts $\{1, \ldots, n\} \backslash \{i\}$

$$\widehat{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathrm{Q}_{1-\alpha}\left\{R_1^{\text{LOO}}, \ldots, R_n^{\text{LOO}}\right\}$$

— Predictive coverage under algorithmic stability assumption:[16]

$$\mathbb{P}\left\{|\widehat{\mu}(X_{n+1}) - \widehat{\mu}_{-i}(X_{n+1})| \leq \epsilon\right\} \geq 1 - \nu$$

[16] Steinberger & Leeb 2018, *Conditional predictive inference for high-dimensional stable algorithms*

Jackknife+ & CV+

Jackknife+:[17]

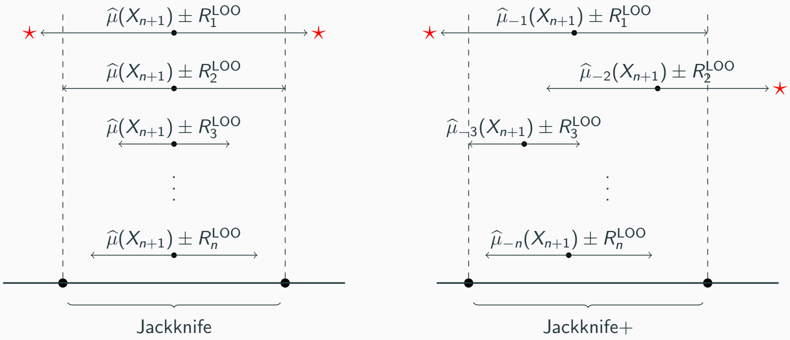
$$\widehat{C}(X_{n+1}) = \left[\mathrm{Q}_{\alpha}\left\{\widehat{\mu}_{-i}(X_{n+1}) - R_i^{\mathrm{LOO}}\right\},\ \mathrm{Q}_{1-\alpha}\left\{\widehat{\mu}_{-i}(X_{n+1}) + R_i^{\mathrm{LOO}}\right\}\right]$$

Compare to jackknife:

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

[17] B., Candès, Ramdas, Tibshirani 2019, *Predictive inference with the jackknife+*

Jackknife+ & CV+



Jackknife+ & CV+

Extension to K-fold CV+:[18]

$$\widehat{C}(X_{n+1}) = \left[\mathrm{Q}_{\alpha}\left\{\widehat{\mu}_{-A_k}(X_{n+1}) - R_i^{\mathrm{CV}}\right\},\ \mathrm{Q}_{1-\alpha}\left\{\widehat{\mu}_{-A_k}(X_{n+1}) + R_i^{\mathrm{CV}}\right\}\right]$$

Closely related to the cross-conformal prediction method[19,20]

[18] B., Candès, Ramdas, Tibshirani 2019, *Predictive inference with the jackknife+*

[19] Vovk 2015, *Cross-conformal predictors*

[20] Vovk et al 2018, *Cross-conformal predictive distributions*

Jackknife+ & CV+

Theorem: For any distrib. P and any $\mathcal{A}$, jackknife+ satisfies

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

(If also assume algorithmic stability, then $\geq 1 - \alpha - o(1)$)

[21] B., Candès, Ramdas, Tibshirani 2019, *Predictive inference with the jackknife+*

[22] Vovk et al 2018, *Cross-conformal predictive distributions*

Jackknife+ & CV+

Theorem: For any distrib. P and any $\mathcal{A}$, jackknife+ satisfies

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

(If also assume algorithmic stability, then $\geq 1 - \alpha - o(1)$)

Theorem: For any distrib. P and any $\mathcal{A}$, K-fold CV+ satisfies

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1})\right\} \geq \begin{cases} 1 - 2\alpha - 1/K & [21] \\ 1 - 2\alpha - 2K/n & [22] \end{cases}$$

$$\rightsquigarrow \quad \geq 1 - 2\alpha - \sqrt{2/n}.$$

[21] B., Candès, Ramdas, Tibshirani 2019, *Predictive inference with the jackknife+*

[22] Vovk et al 2018, *Cross-conformal predictive distributions*

Jackknife+ & CV+

Proof idea: embed jackknife+ into a larger exchangeable problem

Jackknife+ & CV+

Proof idea: embed jackknife+ into a larger exchangeable problem

- Exchangeable data $\{(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})\}$
- $\binom{n+1}{2}$ leave-two-out regressions: $\widetilde{\mu}_{-\{i,j\}}$ for $1 \le i, j \le n+1$
- We can observe n of these, i.e., $\widetilde{\mu}_{-\{i,n+1\}} = \widehat{\mu}_{-i}$ for $1 \le i \le n$

Cross-conformal

For a general score function — can use cross-conformal prediction:[23,24]

- Fit score function $\widehat{S}^{(k)}$ on k-th data set $\{(X_i, Y_i) : i \in \{1, \dots, n\} \backslash A_k\}$
- For $i \in A_k$ define $S_i^{\mathrm{CV}} = \widehat{S}^{(k)}(X_i, Y_i)$
- Prediction set

$$\widehat{C}(X_{n+1}) = \left\{ y : \widehat{S}^{k(i)}(X_{n+1}, y) \leq S_i^{\mathrm{CV}} \text{ for at least } \alpha(n+1) \text{ many } i\text{'s} \right\}$$

- Coverage guarantees as for jackknife+ / CV+

[23] Vovk 2015, *Cross-conformal predictors*

[24] Vovk et al 2018, *Cross-conformal predictive distributions*

Generalizing CP to other definitions of risk

CP methods bound $\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}(X_{n+1})\right\} = \mathbb{E}\left[\underbrace{\mathbf{1}\{Y_{n+1} \notin \widehat{C}(X_{n+1})\}}_{\text{zero/one loss}}\right]$

[25]Angelopolous et al 2021, *Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control*

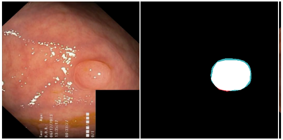
[26]Bates et al 2021, *Distribution-Free, Risk-Controlling Prediction Sets*

Generalizing CP to other definitions of risk

CP methods bound $\mathbb{P}\left\{Y_{n+1} \notin \widehat{C}(X_{n+1})\right\} = \mathbb{E}\left[\underbrace{\mathbf{1}\{Y_{n+1} \notin \widehat{C}(X_{n+1})\}}_{\text{zero/one loss}}\right]$

Idea — use CP-type approach to control other definitions of risk:[25,26]

- Example: FDR for flagging out-of-distribution data points
- Example: false pos./neg. rates if Y = a set of labels
- Example: accuracy rate for selecting pixels within an image



(figures from Bates et al 2021)

[25]Angelopolous et al 2021, *Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control*

[26]Bates et al 2021, *Distribution-Free, Risk-Controlling Prediction Sets*

Limitations of distribution-free prediction

The guarantee for conformal prediction / holdout methods:

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

w.r.t. distribution of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ (assumed to be exchangeable)

Limitations:

- The guarantee is *on average* over the training data
- The guarantee is *on average* over the test point X_{n+1}
- And, what if the data is not exchangeable?

The PAC framework

Limitation 1: The guarantee is *on average* over the training data

All guarantees so far:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1})\right\} = \mathbb{E}\left[\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1}) \;\middle|\; \begin{array}{c}\text{training}\\ \text{data}\end{array}\right\}\right] \geq 1 - \alpha$$

[27] Vovk 2012, *Conditional validity of inductive conformal predictors*

[28] Bian & B. 2021, *Training-conditional coverage for distribution-free predictive inference*

The PAC framework

Limitation 1: The guarantee is *on average* over the training data

All guarantees so far:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1})\right\} = \mathbb{E}\left[\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1}) \;\middle|\; \begin{array}{c}\text{training}\\ \text{data}\end{array}\right\}\right] \geq 1-\alpha$$

The PAC (Probably Approximately Correct) framework:

$$\mathbb{P}\left\{\,\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1}) \;\middle|\; \begin{array}{c}\text{training}\\ \text{data}\end{array}\right\} \geq 1-\alpha\,\right\} \geq 1-\delta$$

[27]Vovk 2012, *Conditional validity of inductive conformal predictors*

[28]Bian & B. 2021, *Training-conditional coverage for distribution-free predictive inference*

The PAC framework

Limitation 1: The guarantee is *on average* over the training data

All guarantees so far:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1})\right\} = \mathbb{E}\left[\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1}) \;\middle|\; \begin{array}{c}\text{training}\\ \text{data}\end{array}\right\}\right] \geq 1-\alpha$$

The PAC (Probably Approximately Correct) framework:

$$\mathbb{P}\left\{\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1}) \;\middle|\; \begin{array}{c}\text{training}\\ \text{data}\end{array}\right\} \geq 1-\alpha\right\} \geq 1-\delta$$

- Split conformal satisfies PAC with no additional assumptions[27]
- No PAC guarantee is possible for full conformal or jackknife+ (unless we make further assumptions)[28]

[27] Vovk 2012, *Conditional validity of inductive conformal predictors*

[28] Bian & B. 2021, *Training-conditional coverage for distribution-free predictive inference*

Conditional prediction

Limitation 2: The guarantee is *on average* over the test point X_{n+1}

Is it possible to provide prediction that's valid conditional on X_{n+1}, i.e.,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1}) \;\middle|\; X_{n+1}\right\} \geq 1 - \alpha \text{ ?}$$

(Motivation—the marginal guarantee doesn't exclude, e.g.,

90% of individuals have 100% coverage / 10% of individuals have 0% coverage)

[29] Vovk 2012, *Conditional validity of inductive conformal predictors*

[30] Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

Conditional prediction

Limitation 2: The guarantee is *on average* over the test point X_{n+1}

Is it possible to provide prediction that's valid conditional on X_{n+1}, i.e.,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1}) \;\middle|\; X_{n+1}\right\} \geq 1-\alpha \text{ ?}$$

(Motivation—the marginal guarantee doesn't exclude, e.g.,

90% of individuals have 100% coverage / 10% of individuals have 0% coverage)

This is impossible for nonatomic X (i.e., $P_X(x) = 0$ for all $x \in \mathcal{X}$):[29,30]

Theorem: If X is nonatomic,
$\mathbb{E}\left[\text{length}(\widehat{C}(X_{n+1}))\right] = \infty$ for any $\widehat{C}$ that's valid distribution-free

[29] Vovk 2012, *Conditional validity of inductive conformal predictors*

[30] Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

Conditional prediction

Limitation 2: The guarantee is *on average* over the test point X_{n+1}

Is it possible to provide prediction that's valid conditional on X_{n+1}, i.e.,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}(X_{n+1}) \;\middle|\; X_{n+1}\right\} \geq 1 - \alpha \text{ ?}$$

(Motivation—the marginal guarantee doesn't exclude, e.g.,

90% of individuals have 100% coverage / 10% of individuals have 0% coverage)

This is impossible for nonatomic X (i.e., $P_X(x) = 0$ for all $x \in \mathcal{X}$):[29,30]

Theorem: If X is nonatomic,

$$\underbrace{\mathbb{E}\left[\text{length}(\widehat{C}(X_{n+1}))\right]}_{\text{expected length when data} \overset{\text{iid}}{\sim} P} = \infty \text{ for any } \underbrace{\widehat{C} \text{ that's valid distribution-free}}_{\text{coverage must hold when data} \overset{\text{iid}}{\sim} \text{any distribution}}$$

[29] Vovk 2012, *Conditional validity of inductive conformal predictors*

[30] Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

Conditional prediction

Can we relax the notion of conditionally valid coverage,
to obtain a nontrivial $\widehat{C}$?

$(1-\alpha,\delta)$-conditional coverage:[31] for any P & any $\mathcal{X}$ with $P_X(\mathcal{X}) \geq \delta$,

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}(X_{n+1}) \;\middle|\; X_{n+1} \in \mathcal{X} \right\} \geq 1-\alpha \text{ w.r.t. data} \stackrel{\text{iid}}{\sim} P.$$

[31] B., Candès, Ramdas, Tibshirani 2019, *The limits of distribution-free conditional predictive inference*

Conditional prediction

Can we relax the notion of conditionally valid coverage,
to obtain a nontrivial $\widehat{C}$?

$(1-\alpha,\delta)$-conditional coverage:[31] for any P & any $\mathcal{X}$ with $P_X(\mathcal{X}) \geq \delta$,

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}(X_{n+1}) \;\middle|\; X_{n+1} \in \mathcal{X} \right\} \geq 1-\alpha \text{ w.r.t. data} \stackrel{\text{iid}}{\sim} P.$$

Theorem: for nonatomic P_X, if $\widehat{C}$ satisfies $(1-\alpha,\delta)$-conditional cov., then

$$\mathbb{E}\left[\text{length}(\widehat{C}(X_{n+1}))\right] \geq \begin{pmatrix} \text{min. length of any oracle method} \\ \text{with } 1-\alpha\delta \text{ coverage for } P \end{pmatrix}$$

trivially achieves $(1-\alpha,\delta)$-conditional cov.
(and must be very wide)

[31] B., Candès, Ramdas, Tibshirani 2019, *The limits of distribution-free conditional predictive inference*

Conditional prediction

Conditional on bins: partition $\mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$,

& require $\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}(X_{n+1}) \;\middle|\; X_{n+1} \in \mathcal{X}_k \right\} \geq 1 - \alpha$ for each k[32,33]

- For each k, data points $\{(X_i, Y_i) : X_i \in \mathcal{X}_k\}$ are exchangeable $\rightsquigarrow$ run CP separately for each k to guarantee bin-conditional cov.
- Note — the model $\widehat{\mu}$ can still be fitted on the entire data set!

An application — fairness with respect to subpopulations[34]

[32]Vovk 2012, *Conditional validity of inductive conformal predictors*

[33]Lei & Wasserman 2014, *Distribution-free prediction bands for nonparametric regression*

[34]Romano, B., Sabatti, Candès 2019, *With malice toward none: assessing uncertainty via equalized coverage*

Conditional prediction

Extension — distributional conformal prediction[35]

- Estimate the conditional distribution of $Y|X \rightsquigarrow \widehat{F}(y|x)$
- Nonconformity score $\widehat{S}(x,y) = |\widehat{F}(y|x) - 0.5|$
  - CP is valid with any score $\Rightarrow$ finite-sample marginal cov.
  - If $\widehat{F}$ satisfies consistency $\Rightarrow$ asymptotic conditional cov.

[35] Chernozhukov et al 2019, *Distributional conformal prediction*

Conditional prediction

Extension — a localized form of the prediction guarantee:[36]

Construct $\widehat{C}(X_{n+1})$ using a kernel around the test point,

e.g., only the nearest neighbors of X_{n+1}

[36] Guan 2020, *Conformal prediction with localization*

Conditional prediction

Extension — a localized form of the prediction guarantee:[36]

Construct $\widehat{C}(X_{n+1})$ using a kernel around the test point,

e.g., only the nearest neighbors of X_{n+1}

Define weights $w_i = w(X_i, X_{n+1})$, then

$$\widehat{C}(X_{n+1}) = \Big\{ y : \widehat{S}(X_{n+1}, y) \leq \mathrm{Q}_{1-\tilde{\alpha}}\{ S_i \text{ with weight } w_i \} \Big\}$$

adjust α to maintain coverage

$\rightsquigarrow$ achieves marginal coverage, and asymptotic conditional coverage

[36] Guan 2020, *Conformal prediction with localization*

Settings beyond exchangeability

Limitation 3: what if data is not exchangeable?

Conformal prediction (or holdout method) assumes:
training & test data are from the same distribution

Settings beyond exchangeability

Limitation 3: what if data is not exchangeable?

Conformal prediction (or holdout method) assumes:
training & test data are from the same distribution

Possible violations:

- Train & test data are from different distributions (transfer learning)
- Data distribution changes over time (drift / changepoints)
- Dependence (over time / spatial location / network / etc)

Weighted conformal prediction

The covariate shift setting:

- Marginal distribution of X is different in training vs. test data
  (e.g., some subpopulations are over- or under-represented in the training data)
- But, distribution of $Y|X$ is the same

Weighted conformal prediction

The covariate shift setting:

- Marginal distribution of X is different in training vs. test data (e.g., some subpopulations are over- or under-represented in the training data)
- But, distribution of $Y|X$ is the same

$$P^{\text{train}} = P_X^{\text{train}} \times P_{Y|X}, \quad P^{\text{test}} = P_X^{\text{test}} \times P_{Y|X}$$

Weighted conformal prediction

Assuming we know the shift (i.e., $\mathrm{d}P_X^{\mathrm{test}}(x) \propto \overbrace{w(x)}^{\text{known fn.}} \cdot \ \mathrm{d}P_X^{\mathrm{train}}(x)$),
conformal can adjust for the shift with *weighted exchangeability*[37,38]

[37] Tibshirani, B., Ramdas, & Candès 2019, *Conformal prediction under covariate shift*

[38] Hu & Lei 2020, *A distribution-free test of covariate shift using conformal prediction*

Weighted conformal prediction

Assuming we know the shift (i.e., $\mathrm{d}P_X^{\text{test}}(x) \propto \overbrace{w(x)}^{\text{known fn.}} \cdot\ \mathrm{d}P_X^{\text{train}}(x)$), conformal can adjust for the shift with *weighted exchangeability*[37,38]

Given $n+1$ data points...

- With (unweighted) exchangeability,
  each one is equally likely to be the test point
  $\rightsquigarrow R_{n+1} \leq \mathrm{Q}_{1-\alpha}\{R_1, \ldots, R_{n+1}\}$ with prob. $1-\alpha$

- With weighted exchangeability, the distribution is nonuniform:

$$\mathbb{P}\left\{(x, y) \text{ is the test point}\right\} \propto w(x)$$

$\rightsquigarrow$ need to compute a weighted quantile: $\mathrm{Q}_{1-\alpha}\{R_i \text{ with weight } w_i\}$

[37] Tibshirani, B., Ramdas, & Candès 2019, *Conformal prediction under covariate shift*

[38] Hu & Lei 2020, *A distribution-free test of covariate shift using conformal prediction*

Weighted conformal prediction

Application: survival analysis & censored data[39]

- "Clean" data $(X_i, Y_i) =$ (features, survival time)
- Censored observations $(X_i, \tilde{Y}_i)$ where $\tilde{Y}_i = \min\{C_i, Y_i\}$
- Main idea: choose a cutoff c_0 so that "usually" $Y_i \leq c_0$,
  & keep only data with $C_i \geq c_0$ (i.e., most Y's are not censored)

[39] Candès, Lei, Ren 2021, *Conformalized survival analysis*

Weighted conformal prediction

Application: survival analysis & censored data[39]

- "Clean" data $(X_i, Y_i) =$ (features, survival time)
- Censored observations $(X_i, \tilde{Y}_i)$ where $\tilde{Y}_i = \min\{C_i, Y_i\}$
- Main idea: choose a cutoff c_0 so that "usually" $Y_i \leq c_0$,
  & keep only data with $C_i \geq c_0$ (i.e., most Y's are not censored)

  — On this data set, can use CP to predict survival time Y
  — But, this may be a different distribution
  (population with $C_i \geq c_0 \quad \neq \quad$ general population)
  — If distrib. of $C|X$ known,
  can use weighted CP to correct for distribution shift

[39] Candès, Lei, Ren 2021, *Conformalized survival analysis*

Weighted conformal prediction

Application: estimating individual treatment effects[40]

- Data $(X_i, T_i, Y_i) =$ (features, treatment group $= 0$ or 1, outcome)
- $\text{ITE}_i =$ (value of Y_i, if $T_i = 1$) − (value of Y_i, if $T_i = 0$)
- Challenge: treatment assignment may depend on X
- Main idea: if propensity score $\mathbb{P}\{T = 1 \mid X = x\}$ is known,
  can use weighted CP to adjust for $X|T = 1$ versus $X|T = 0$

[40] Lei & Candès 2020, *Conformal inference of counterfactuals and individual treatment effects*

Weighted conformal prediction

An extension: the design problem (active learning)[41]

- Training data $(X_i, Y_i) \sim P$
- Test data $X_{n+1} \sim \tilde{P}_X$, where $\tilde{P}_X$ depends on training data
- Can use an extension of weighted CP for valid predictive inference

[41] Fannjiang et al 2022, *Conformal prediction for the design problem*

Weighted conformal prediction

A related problem — label shift (for categorical Y / classification)[42]

- Marginal distribution of Y is different in training vs. test data
  (e.g., some subpopulations are over- or under-represented in the training data)
- But, distribution of $X|Y$ is the same

$$P^{\text{train}} = P_Y^{\text{train}} \times P_{X|Y}, \quad P^{\text{test}} = P_Y^{\text{test}} \times P_{X|Y}$$

If the label shift is known,
can use weighted exchangeability to guarantee coverage

[42] Podkopaev & Ramdas 2021, *Distribution-free uncertainty quantification for classification under label shift*

Weighted conformal prediction

Covariate shift & label shift methods — both assume $\frac{\mathrm{d}P^{\text{test}}}{\mathrm{d}P^{\text{train}}}$ is known

If the distribution shift is unknown, but can be bounded:[43]
construct $\widehat{C}$ that is valid assuming $D_f\left(P^{\text{test}}||P^{\text{train}}\right) \leq \rho$

f-divergence (e.g., KL-divergence)

[43] Cauchois et al 2020, *Robust Validation: Confident Predictions Even When Distributions Shift*

Online / streaming / time series data

Conformal prediction can also be applied to an online setting...

- If data points are iid,
  conformal p-values are valid (and $\perp\!\!\!\perp$) at each time t
  $\Rightarrow$ can use conformal to predict / to test for changepoints[44]
- Can bound cumulative error under arbitrary distribution drift[45,46]
- If data points form a time series,
  CP achieves asymptotic coverage under some assumptions[47]

---

[44] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[45] Gibbs & Candès 2021, *Adaptive conformal inference under distribution shift*

[46] Feldman et al 2022, *Conformalized Online Learning: Online Calibration Without a Holdout Set*

[47] Xu & Xie 2021, *Conformal prediction interval for dynamic time-series*

Robustness & nonsymmetric algorithms

Theory for CP relies on:

1. $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable (e.g., i.i.d.)
2. Regression algorithm $\mathcal{A}$ treats input data points symmetrically

Robustness & nonsymmetric algorithms

Theory for CP relies on:

1. $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are exchangeable (e.g., i.i.d.)
2. Regression algorithm $\mathcal{A}$ treats input data points symmetrically

Challenges in practice:

1. $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ may be nonexchangeable
   (e.g., distribution drift, dependence over time, ...)
2. May want to choose $\mathcal{A}$ that treats data nonsymmetrically
   (e.g., weighted regression, autoregressive model, ...)

Robustness & nonsymmetric algorithms

Nonexchangeable conformal prediction (nexCP):[48]

Draw a random index K with $\mathbb{P}\{K = i\} = w_i$, then:

- Fit model to training+test data

$$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \dots, (X_{n+1}, y), \dots, (X_n, Y_n), (X_K, Y_K))$$

- Compute residuals

$$R_i = |Y_i - \widehat{\mu}(X_i)| \text{ for } i \leq n; \;\; R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$

- Check if $R_{n+1} \;\leq\; \mathrm{Q}_{1-\alpha}\{R_i \text{ with weight } w_i\}$

fixed weights $w_i \geq 0$, e.g., $w_1 \leq w_2 \leq \dots$ for distrib. drift

$\widehat{C}(X_{n+1}) = \{\text{all } y \in \mathbb{R} \text{ for which the above holds}\}$

[48] B., Candès, Ramdas, Tibshirani 2022, *Conformal prediction beyond exchangeability*

Robustness & nonsymmetric algorithms

Nonexchangeable conformal prediction (nexCP):[48]

Draw a random index K with $\mathbb{P}\{K = i\} = w_i$, then:

- Fit model to training+test data
$$\widehat{\mu} = \mathcal{A}((X_1, Y_1), \dots, (X_{n+1}, y), \dots, (X_n, Y_n), (X_K, Y_K))$$
- Compute residuals
$$R_i = |Y_i - \widehat{\mu}(X_i)| \text{ for } i \le n;\ \ R_{n+1} = |y - \widehat{\mu}(X_{n+1})|$$
- Check if $R_{n+1} \le \mathrm{Q}_{1-\alpha}\{R_i \text{ with weight } w_i\}$

fixed weights $w_i \ge 0$, e.g., $w_1 \le w_2 \le \dots$ for distrib. drift

$$\widehat{C}(X_{n+1}) = \{\text{all } y \in \mathbb{R} \text{ for which the above holds}\}$$

- Theory: coverage $\ge 1 - \alpha - \sum_i w_i \cdot \mathrm{d_{TV}}(\text{data}, \text{data}_{\text{swap } i \text{ \& } n+1})$

[48] B., Candès, Ramdas, Tibshirani 2022, *Conformal prediction beyond exchangeability*

Beyond the prediction problem

What are inference questions we might want to ask, distribution-free?

- Prediction: the unobserved response Y_{n+1} will lie in [some range]
- Effect size: the dependence between X and Y lies in [some range]
- Independence: test if X & Y independent given [some confounders]
- Regression: the distribution of Y given X satisfies [some property]

---

[49] Shah & Peters 2018, *The hardness of conditional independence testing and the generalised covariance measure*

[50] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[51] B. 2020, *Is distribution-free inference possible for binary regression?*

[52] Lee & B. 2021, *Distribution-free inference for regression: discrete, continuous, and in between*

Beyond the prediction problem

What are inference questions we might want to ask, distribution-free?

- Prediction: the unobserved response Y_{n+1} will lie in [some range]
- Effect size: the dependence between X and Y lies in [some range]
- Independence: test if X & Y independent given [some confounders]
- Regression: the distribution of Y given X satisfies [some property]

mostly impossible (if X is continuous), or trivial (if X is discrete with bounded # values)

[49] Shah & Peters 2018, *The hardness of conditional independence testing and the generalised covariance measure*

[50] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[51] B. 2020, *Is distribution-free inference possible for binary regression?*

[52] Lee & B. 2021, *Distribution-free inference for regression: discrete, continuous, and in between*

Beyond the prediction problem

What are inference questions we might want to ask, distribution-free?

- Prediction: the unobserved response Y_{n+1} will lie in [some range]
- Effect size: the dependence between X and Y lies in [some range]
- Independence: test if X & Y independent given [some confounders]
- Regression: the distribution of Y given X satisfies [some property]

mostly impossible (if X is continuous), or trivial (if X is discrete with bounded # values)

- Hardness results for testing independence[49]
- Hardness results for inference on $\mathbb{E}[Y \mid X]$[50,51,52]

[49] Shah & Peters 2018, *The hardness of conditional independence testing and the generalised covariance measure*

[50] Vovk, Gammerman, Shafer 2005, *Algorithmic Learning in a Random World*

[51] B. 2020, *Is distribution-free inference possible for binary regression?*

[52] Lee & B. 2021, *Distribution-free inference for regression: discrete, continuous, and in between*

Calibration

Calibration $=$ an alternative definition of validity for a predictor

- Perfect calibration: $\mathbb{E}\left[Y \mid f(X)\right] = f(X)$ almost surely
- Approx. calibration:[53] $\left|\mathbb{E}\left[Y \mid f(X)\right] - f(X)\right| \leq \epsilon$ w/ prob. $\geq 1 - \alpha$

[53] Gupta et al 2020, *Distribution-free binary classification: prediction sets, confidence intervals and calibration*

Calibration

Distribution-free calibration is possible
only if the set of output values is $\leq$ countably infinite:[54]

- Let error level α be fixed, and let sample size $n \to \infty$
- A sequence of functions f_n is asymptotically calibrated if $\epsilon_n = o_P(1)$
- If there exists an asymptotically calibrated sequence f_n, then

$$\limsup_{n\to\infty} \left|\{\text{possible values of } f_n(X)\}\right| \leq \text{countably infinite}$$

[54] Gupta et al 2020, *Distribution-free binary classification: prediction sets, confidence intervals and calibration*

Calibration

Distribution-free calibration is possible
only if the set of output values is $\leq$ countably infinite:[54]

- Let error level α be fixed, and let sample size $n \to \infty$
- A sequence of functions f_n is asymptotically calibrated if $\epsilon_n = o_P(1)$
- If there exists an asymptotically calibrated sequence f_n, then

$$\lim\sup_{n\to\infty} \left|\{\text{possible values of } f_n(X)\}\right| \leq \text{countably infinite}$$

If $f(X)$ takes finitely many values... an example procedure:

- Use data $i = 1, \ldots, \frac{n}{2}$ to partition into bins $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_K$
  (e.g., $\mathcal{X}_k = \{x : \frac{k-1}{N} < \widehat{\mu}(x) \leq \frac{k}{N}\}$)
- Use holdout set $i = \frac{n}{2} + 1, \ldots, n$ to estimate $\mathbb{E}\left[Y \mid X \in \mathcal{X}_k\right]$

[54] Gupta et al 2020, *Distribution-free binary classification: prediction sets, confidence intervals and calibration*

Summary

Distribution-free prediction means that we can:

- Start with any algorithm / modeling procedure...
- ...and then calibrate it to have valid predictive coverage

Summary

Distribution-free prediction means that we can:

- Start with any algorithm / modeling procedure...
- ...and then calibrate it to have valid predictive coverage

The framework relies on assuming:

- Data is exchangeable (e.g., i.i.d. data)
- Or, data has a *bounded* deviation from exchangeability
- Or, a *known* deviation from exchangeability (e.g., covariate shift)

Open questions

- How to detect or adapt to violations of exchangeability?
- Computationally efficient versions of conformal / jackknife+,
  when model alg. is expensive / when Y is multidimensional / etc
- Can we use the data to guide choices (e.g., score function $\widehat{S}(x, y)$),
  without the need for an additional split of the training data?
- Finite-sample guarantees for approximate local/conditional validity?
- Beyond prediction — can we find weaker definitions of validity
  (for testing conditional indep. / for inference on regression / etc)
  for which distribution-free inference is possible?