An introduction to conformal prediction and distribution-free inference

(Lecture 2)

Rina Foygel Barber (University of Chicago)
Columbia Statistics short course May/June 2024

http://rinafb.github.io/

Topics — Lecture 2

Full conformal prediction

- Motivation
- The full conformal prediction method
- Generalizing to score functions
- Computation
- Comparing conformal and classical methods
- The role of randomization

Summary

- Summary & preview

Full conformal prediction

Motivation

Can we avoid data splitting?

Recall....

- The naive method fits a more accurate $\widehat{\mu}$,
  but the margin of error is too small due to overfitting
- The holdout set method (i.e., split CP) fits a less accurate $\widehat{\mu}$,
  but the margin of error is correctly calibrated
- And, CV type methods tend to work in practice
  but can fail in unstable settings / no distrib.-free theory

Can we avoid data splitting?

Why does distribution-free theory hold for split CP but not for CV?

$$C(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q} \;\rightsquigarrow \text{ coverage if } \underbrace{|Y_{n+1} - \widehat{\mu}(X_{n+1})|}_{=R_{n+1}} \leq \widehat{q}$$

- For split conformal, $\widehat{q}$ is quantile of calibration residuals

  $$R_i = |Y_i - \widehat{\mu}(X_i)|, \;\; i = n_0 + 1, \ldots, n$$

  and $\widehat{\mu}$ is pretrained $\Rightarrow R_{n_0+1}, \ldots, R_n, R_{n+1}$ are exchangeable
- For CV, $\widehat{q}$ is quantile of leave-one-out residuals

  $$R_i = |Y_i - \widehat{\mu}_{-i}(X_i)|, \;\; i = 1, \ldots, n$$

  $\Rightarrow R_1, \ldots, R_n, R_{n+1}$ are *not* exchangeable

Full conformal prediction

The full conformal prediction method

Full conformal prediction

Intuition—

- Split CP fits $\widehat{\mu}$ to part of the data, to ensure R_i's are exch.
- CV uses all the data for fitting $\widehat{\mu}$, but R_i's not exch.
- Full CP: use all the data for fitting $\widehat{\mu}$ *and* ensure R_i's are exch.

An additional assumption:

The symmetric algorithm assumption

For any $Z_1, \ldots, Z_m$ and any $\sigma \in \mathcal{S}_m$,

$$\mathcal{A}(Z_1, \ldots, Z_m) = \mathcal{A}(Z_{\sigma(1)}, \ldots, Z_{\sigma(m)}).$$

Full conformal prediction

Full CP, oracle version: imagine we could observe Y_{n+1}

- 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

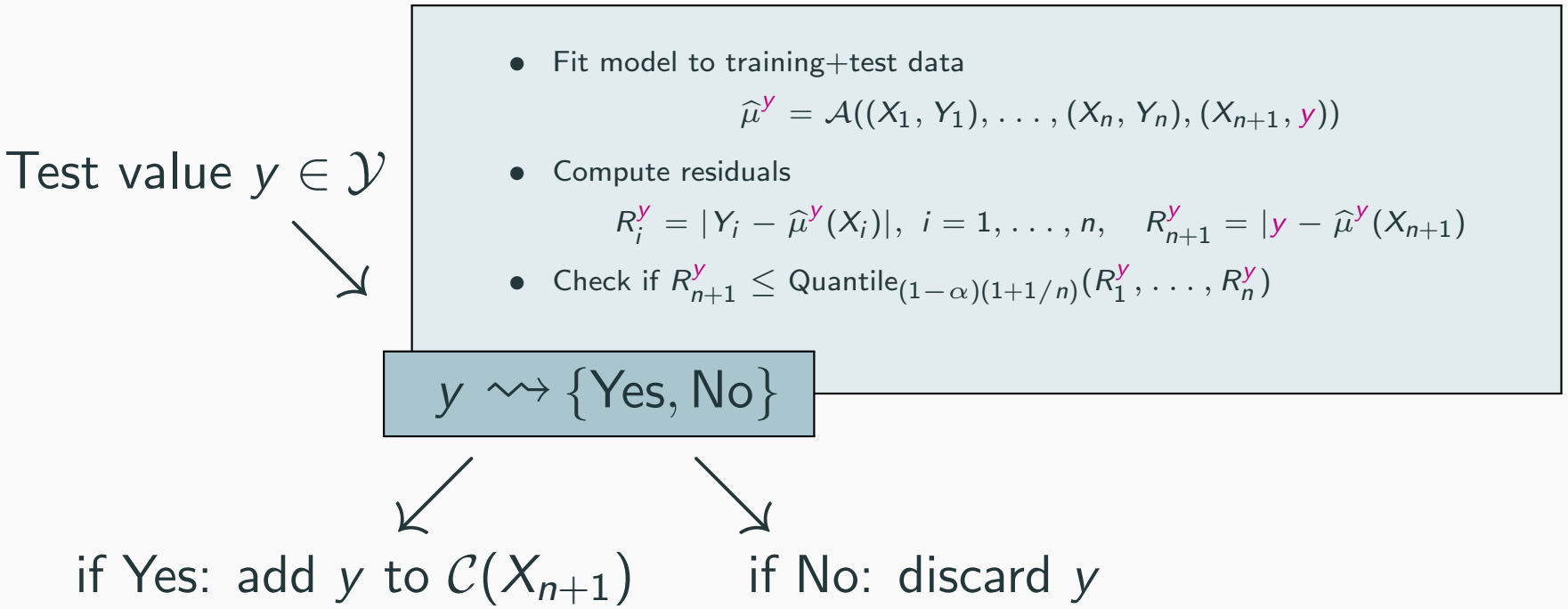
$$R_i = |Y_i - \widehat{\mu}(X_i)|,\ i = 1, \ldots, n;\ \ R_{n+1} = |Y_{n+1} - \widehat{\mu}(X_{n+1})|$$

- Check if $R_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+1/n)}(R_1, \ldots, R_n)$

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

Full conformal prediction

Running full conformal in practice:



Full conformal prediction

Given training data $\{(X_i, Y_i)\}_{i=1,\ldots,n}$, test point X_{n+1}, alg. $\mathcal{A}$:

Full conformal prediction (with residual score)

$$\mathcal{C}(X_{n+1}) = \left\{y \in \mathcal{Y} : R^y_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+1/n)}\left(R^y_1, \ldots, R^y_n\right)\right\}$$

where

$$R^y_i = |Y_i - \widehat{\mu}^y(X_i)|, i = 1, \ldots, n, \quad R^y_{n+1} = |y - \widehat{\mu}^y(X_{n+1})|,$$

for fitted model $\widehat{\mu}^y = \mathcal{A}\big((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)\big)$.

Full conformal prediction

Theorem: full conformal[1]

If $Z_1, \ldots, Z_{n+1}$ are exchangeable, and $\mathcal{A}$ is symmetric, then full conformal prediction satisfies

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

Proof:

$$Y_{n+1} \in \mathcal{C}(X_{n+1}) \iff R_{n+1}^{Y_{n+1}} \leq \text{Quantile}_{(1-\alpha)(1+1/n)}\left(R_1^{Y_{n+1}}, \ldots, R_n^{Y_{n+1}}\right)$$

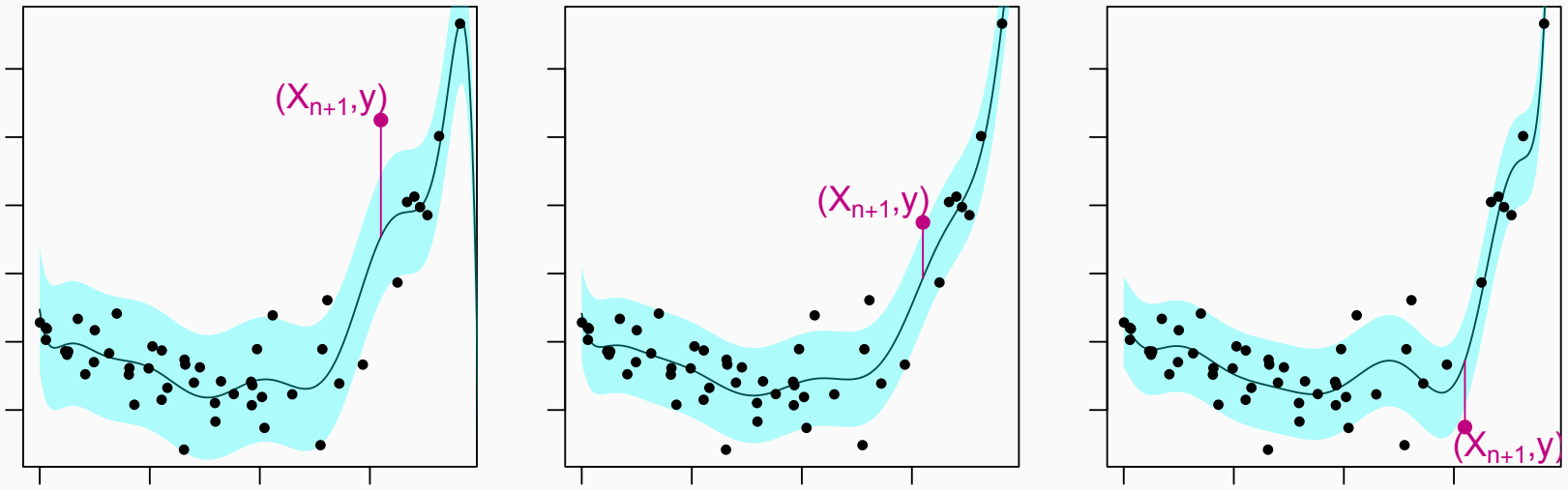
$$\iff R_{n+1}^{Y_{n+1}} \leq \text{Quantile}_{1-\alpha}\left(R_1^{Y_{n+1}}, \ldots, R_n^{Y_{n+1}}, R_{n+1}^{Y_{n+1}}\right)$$

And, this holds with prob. $\geq 1 - \alpha$ (as we proved for oracle version)

[1] Vovk et al 2005, *Algorithmic Learning in a Random World*

Full conformal prediction

How full conformal is run:



- $\widehat{\mu}$ needs to be refitted for each X_{n+1} & each possible y

Full CP versus split CP

Split CP can be viewed as a special case of full CP:

- Let $\widehat{\mu}$ be the pretrained model:

$$\widehat{\mu} = \mathcal{A}(Z_1, \ldots, Z_{n_0})$$

- Define a non-data-dependent algorithm:
  $\mathcal{A}'$ maps *any* data set to the pretrained model $\widehat{\mu}$

- $\big[\text{Full CP with } \{(X_i, Y_i)\}_{n_0 < i \leq n} \;\&\; \mathcal{A}'\big] = \big[\text{Split CP with } \mathcal{A}\big]$

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

Full conformal prediction

Generalizing to score functions

Full conformal prediction: general score

Full CP can be run with any score function (e.g., categorical data).

New definition of an algorithm:

$\mathcal{A}$ maps a data set to a *score function* $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Full conformal prediction (general score)

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S^y_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+1/n)}\left(S^y_1, \ldots, S^y_n\right) \right\}$$

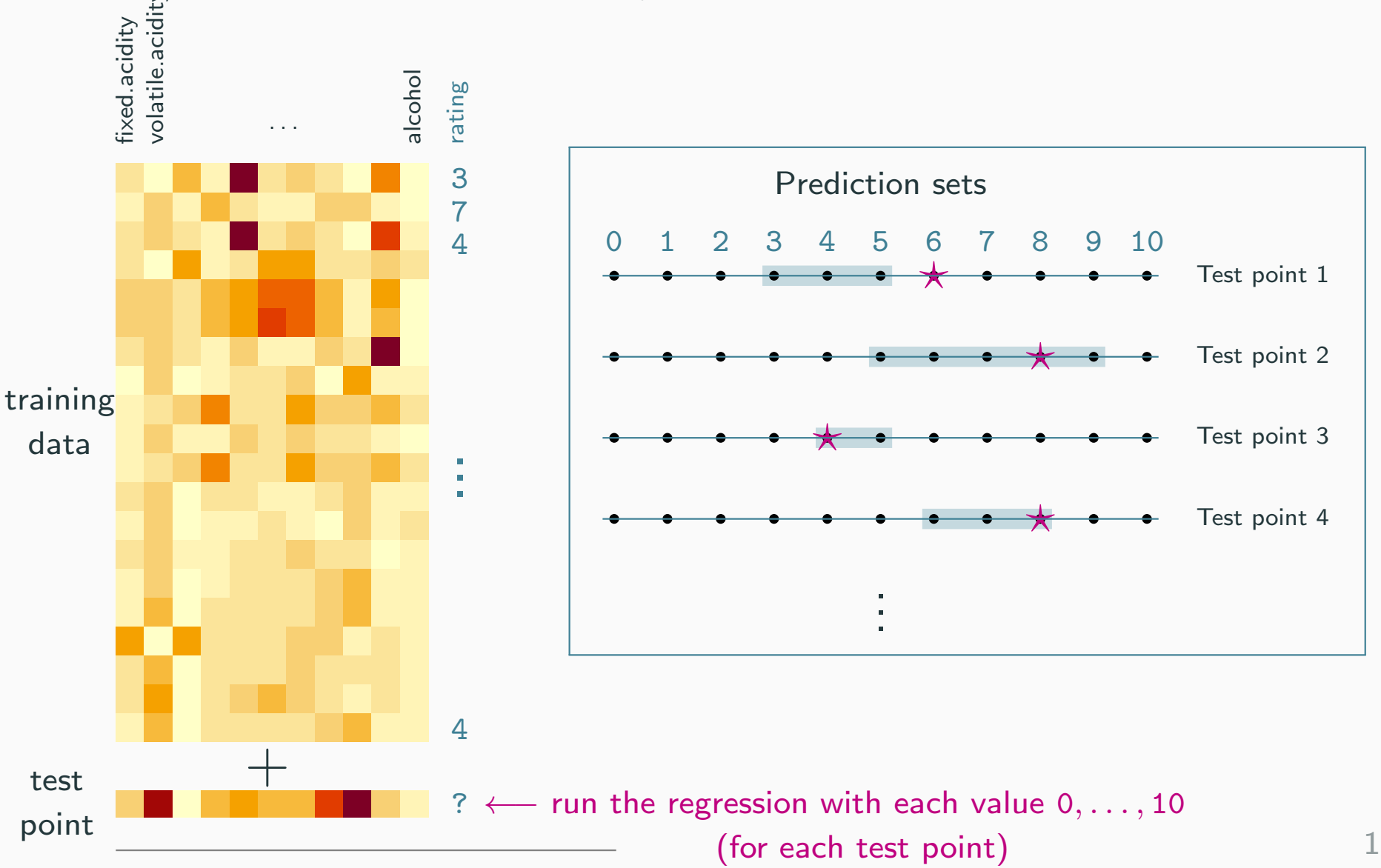
where

$$S^y_i = s^y(X_i, Y_i), i = 1, \ldots, n, \quad S^y_{n+1} = s^y(X_{n+1}, y),$$

for fitted score function $s^y = \mathcal{A}\big((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)\big)$

Full conformal prediction: general score

Example: the `Wine Quality` data set[2]



[2] Cortez et al 2009, `Wine Quality` data set, UCI Machine Learning Repository

Full conformal prediction: general score

An alternative formulation for full CP....

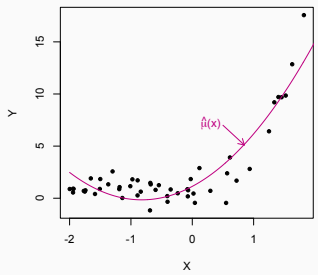
The algorithm $\mathcal{A}$ maps a data set to a vector of scores:

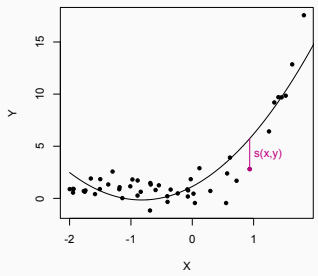
$$(Z_1, \ldots, Z_{n+1}) \mapsto (S_1, \ldots, S_{n+1})$$

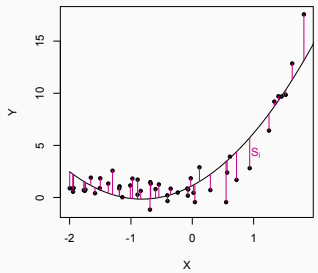
The three versions:

1. $\mathcal{A}$ returns a fitted model:
   $\mathcal{A}(Z_1, \ldots, Z_{n+1}) = \widehat{\mu} \rightsquigarrow$ scores $S_i = R_i = |Y_i - \widehat{\mu}(X_i)|$
2. $\mathcal{A}$ returns a fitted score function:
   $\mathcal{A}(Z_1, \ldots, Z_{n+1}) = s \rightsquigarrow$ scores $S_i = s(X_i, Y_i)$
3. $\mathcal{A}$ returns a vector of scores:
   $\mathcal{A}(Z_1, \ldots, Z_{n+1}) = (S_1, \ldots, S_{n+1})$

Full conformal prediction: general score







$\mathcal{A}$ maps a data set $\{(X_i, Y_i)\}$ to...

1. a fitted model $\widehat{\mu}(x)$
2. a score function $s(x, y)$
3. a vector of scores S_i

Full conformal prediction: general score

For distribution-free coverage to hold,
need (1) exchangeable data, (2) symmetry of $\mathcal{A}$

What does symmetry of $\mathcal{A}$ mean in each setting?

1. Fitted model: $\mathcal{A}$ is invariant to permutations,
$$\widehat{\mu} = \mathcal{A}(Z_1, \ldots, Z_{n+1}) = \mathcal{A}(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)})$$
2. Fitted score function: $\mathcal{A}$ is invariant to permutations,
$$s = \mathcal{A}(Z_1, \ldots, Z_{n+1}) = \mathcal{A}(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)})$$
3. Vector of scores: $\mathcal{A}$ *commutes with* permutations,
$$\mathcal{A}(Z_1, \ldots, Z_{n+1}) = (S_1, \ldots, S_{n+1}) \rightsquigarrow$$
$$\mathcal{A}(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) = (S_{\sigma(1)}, \ldots, S_{\sigma(n+1)})$$

Full conformal prediction

Computation

Full conformal prediction: 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 $y \in \mathbb{R}$)

Full conformal prediction: computational challenges

A preliminary observation:[3]

for $\mathcal{Y} = \mathbb{R}$, we can truncate CP to the empirical range:

$$\mathcal{C}(X_{n+1}) = \left\{ y \in [\min_{1 \le i \le n} Y_i,\ \max_{1 \le i \le n} Y_i] : s(X_{n+1}, y) \le \widehat{q} \right\}$$

split CP version

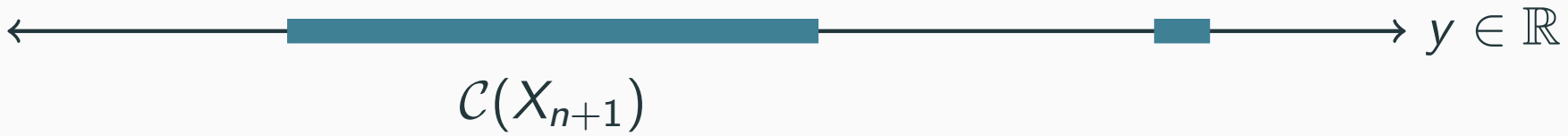
Calculate coverage:

$$\begin{aligned}
&\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \\
&= \mathbb{P}\left\{ s(X_{n+1}, Y_{n+1}) \le \widehat{q} \ \text{ and } \ Y_{n+1} \in [\min_{1 \le i \le n} Y_i,\ \max_{1 \le i \le n} Y_i] \right\} \\
&\ge \mathbb{P}\{s(X_{n+1}, Y_{n+1}) \le \widehat{q}\} - \mathbb{P}\left\{ Y_{n+1} \notin [\min_{1 \le i \le n} Y_i,\ \max_{1 \le i \le n} Y_i] \right\} \\
&\ge 1 - \alpha - \frac{2}{n+1}
\end{aligned}$$

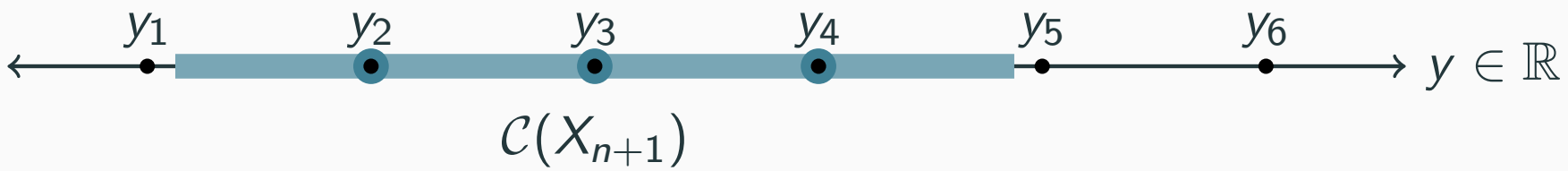
[3] Chen, Wang, Ha, & B. 2016, *Trimmed conformal prediction for high-dimensional models.*

Full conformal prediction: computational challenges

Conformal prediction, in theory:



Conformal prediction, in practice:



Problems to solve:

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

Full conformal prediction: computational challenges

Why do we lose the coverage guarantee?

- If 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

- Essentially, using an algorithm that does not treat data symmetrically

Full conformal prediction: computational challenges

Goal: theoretical validity + computational efficiency.

Discretized conformal prediction:[4]

- Define a new algorithm $\widetilde{\mathcal{A}} = \mathcal{A} \circ [\,\cdot\,]$

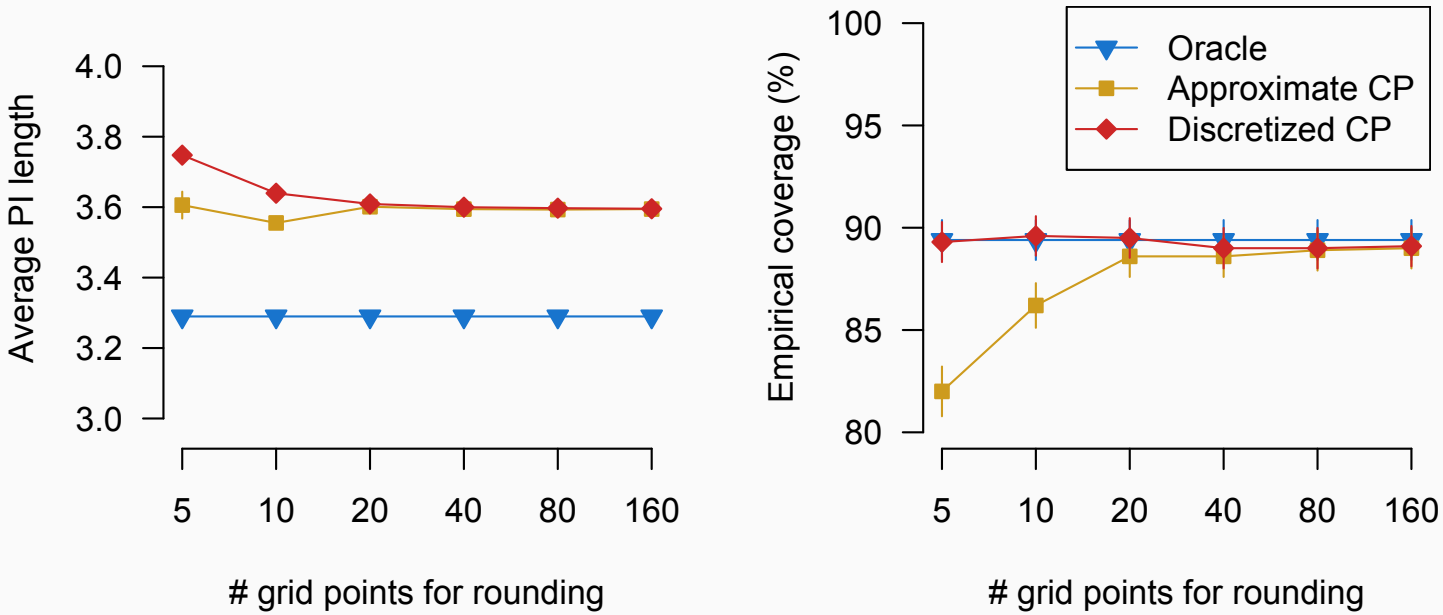
  $$\widetilde{\mathcal{A}}\big((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})\big) = \mathcal{A}\big((X_1, [Y_1]), \ldots, (X_{n+1}, [Y_{n+1}])\big)$$

- Then run full CP with $\widetilde{\mathcal{A}}$ as the regression algorithm

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

Full conformal prediction: computational challenges

Discretized conformal prediction:[5]



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

Full conformal prediction: computational challenges

A special case: methods that are linear in Y

The algorithm $\mathcal{A}$ takes the following form:

For $\widehat{\mu}$ fitted to data points $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$,

$$\begin{pmatrix} \widehat{\mu}(X_1) \\ \vdots \\ \widehat{\mu}(X_{n+1}) \end{pmatrix} = \mathbf{H} \cdot \begin{pmatrix} Y_1 \\ \vdots \\ Y_{n+1} \end{pmatrix}$$

where $\mathbf{H} = \mathbf{H}(X_1, \ldots, X_{n+1})$ depends only on the features.

Examples:

- Linear regression & ridge regression
- Regression with polynomial terms / interaction terms / etc
- Nearest neighbor methods
- Kernel smoothing type methods

Full conformal prediction: computational challenges

How to run full conformal prediction in this special case:[6]

1. Compute $\mathbf{H} = \mathbf{H}(X_1, \ldots, X_{n+1})$
2. Then for $\widehat{\mu}^y = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$,

$$\widehat{\mu}^y(X_i) = \mathbf{H}_{i,*} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ y \end{pmatrix} = \mathbf{H}_{i,*} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ 0 \end{pmatrix} + y \cdot \mathbf{H}_{i,*} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} = a_i + b_i y$$

3. Prediction interval can now be computed in finite time:

$$\mathcal{C}(X_{n+1}) = \left\{ y : |y - (a_{n+1} + b_{n+1} y)| \leq \text{Quantile}_{(1-\alpha)(1+1/n)} \left(|Y_i - (a_i + b_i y)| \right) \right\}$$

[6] Burnaev & Vovk 2014, *Efficiency of conformalized ridge regression*

Full conformal prediction: computational challenges

A special case: stable conformal prediction[7]

Assume stability of $\mathcal{A}$,

$$|\widehat{\mu}^{y}(X_i) - \widehat{\mu}^{y_0}(X_i)| \leq \tau_i, \;\; i = 1, \ldots, n+1$$

- Stability assumption $\Rightarrow$ scores are bounded above & below:

$$\underbrace{|Y_i - \widehat{\mu}(X_i)|}_{=R_i} \in \underbrace{|Y_i - \widehat{\mu}^{y_0}(X_i)|}_{=R_i^{y_0}} \pm \tau_i$$

- Bound prediction set above & below:

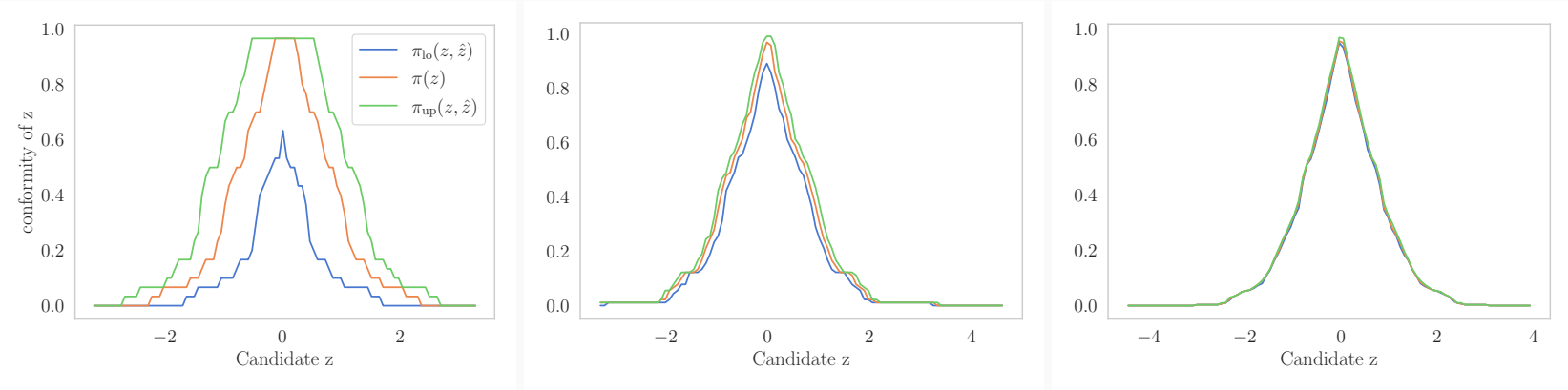
$$\mathcal{C}(X_{n+1}) \subseteq \left\{y : R_{n+1}^{y_0} - \tau_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+1/n)}\left(R_i^{y_0} + \tau_i\right)\right\}$$

$$\mathcal{C}(X_{n+1}) \supseteq \left\{y : R_{n+1}^{y_0} + \tau_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+1/n)}\left(R_i^{y_0} - \tau_i\right)\right\}$$

[7] Ndiaye 2022, *Stable Conformal Prediction Sets*

Full conformal prediction: computational challenges

In general, large $n \rightsquigarrow \tau_i \approx 0$, so the bounds converge:



(a) $n = 30$ (b) $n = 90$ (c) $n = 300$

(figure from Ndiaye 2022)

But, the method requires knowing the stability properties of $\mathcal{A}$.

Full conformal prediction: computational challenges

Summary of approaches:

- In practice — restrict to a grid of y values (but no theory)
- Specialized methods for specific algorithms/settings,
  e.g., Ridge[8], Lasso[9], stable algorithms[10]
- Discretized CP — use a discretized version of $\mathcal{A}$
  to restore theoretical guarantees[11]

[8] Burnaev & Vovk 2014, *Efficiency of conformalized ridge regression*

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

[10] Ndiaye 2022, *Stable Conformal Prediction Sets*

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

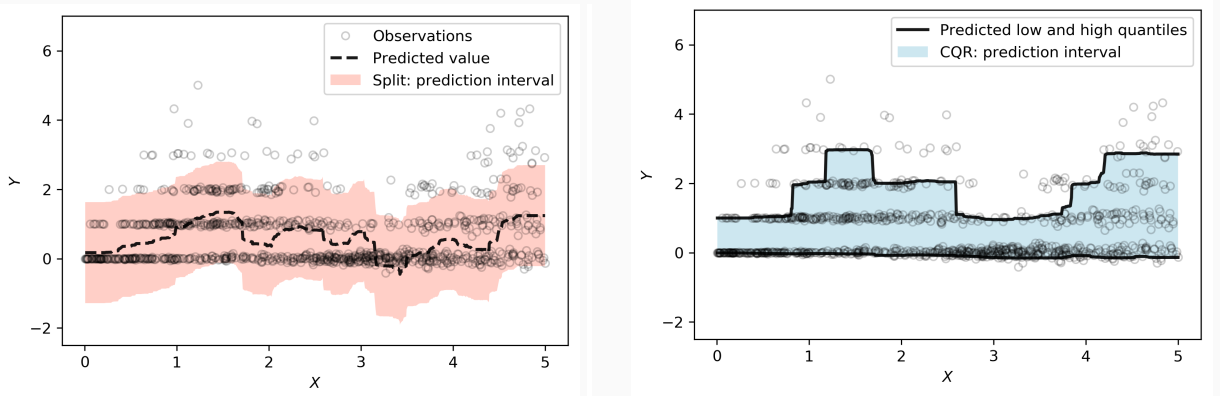
Full conformal prediction

Comparing conformal and classical methods

Comparing conformal & classical methods

For a particular data set, how do we choose between CP versus an existing model/algorithm? What if:

- CP provides a distribution-free coverage guarantee
- But, the existing model/algorithm is a better fit to the data distribution, & provides a narrower interval



(figure from Romano et al 2019)

By choosing a good score function, can have the best of both

Comparing conformal & classical methods

CP is a "wrapper" method that pairs with any model/algorithm,
so we can:

- Choose the model that we believe fits the data
- Use it to guide our definition of the score function
- Apply CP $\rightsquigarrow \mathcal{C}(X_{n+1})$ is an adjusted version of the original model's answer

Drawbacks:
computational cost for full CP / reduced sample size for split CP

Comparing conformal & classical methods

An example strategy...

Construct model-based prediction regions at each confidence level:

$$\dots \subseteq \mathcal{C}^{\text{mod}}_{0.8}(x) \subseteq \mathcal{C}^{\text{mod}}_{0.9}(x) \subseteq \mathcal{C}^{\text{mod}}_{0.95}(x) \subseteq \dots$$

Define a score $s(x, y) = \inf \left\{ t : y \in \mathcal{C}^{\text{mod}}_t(x) \right\}$

Then split CP returns $\mathcal{C}(X_{n+1}) = \mathcal{C}^{\text{mod}}_{\widehat{q}}(X_{n+1})$

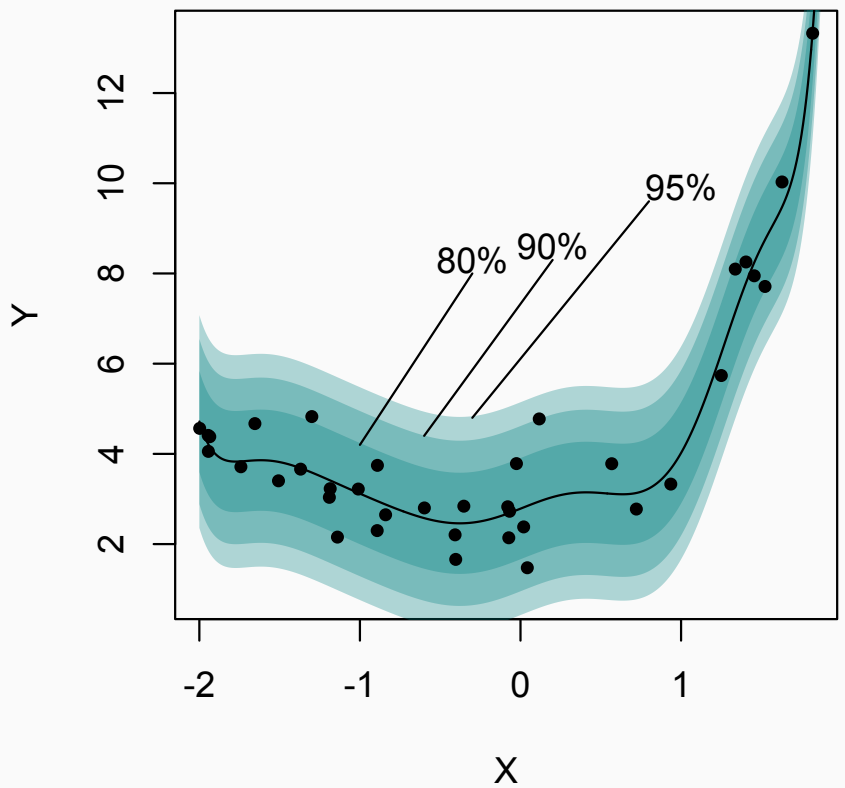
If model is correct, $\widehat{q} \approx 1 - \alpha$

If model is misspecified, $\widehat{q} > 1 - \alpha$

Comparing conformal & classical methods

Interpretation via nested sets:[12]

$$\cdots \subseteq \mathcal{C}^{\text{mod}}_{0.8}(x) \subseteq \mathcal{C}^{\text{mod}}_{0.9}(x) \subseteq \mathcal{C}^{\text{mod}}_{0.95}(x) \subseteq \cdots$$



[12] Gupta et al 2022, *Nested conformal prediction and quantile out-of-bag ensemble methods*

Comparing conformal & classical methods

Theoretical guarantees that CP converges to the "oracle" answer:

Setting 1: estimating the conditional mean[13]

For the residual score, $s(x, y) = |y - \widehat{\mu}(x)|$,
if $\widehat{\mu}(x) \to \mathbb{E}[Y \mid X = x]$ & noise is symmetric,

$$\mathcal{C}(X_{n+1}) \to \mathcal{C}^*(X_{n+1})$$

Setting 2: estimating the conditional density[14]

For the density score, $s(x, y) = -\widehat{f}(y|x)$, if $\widehat{f}(y|x) \to f(y|x)$,

$$\mathcal{C}(X_{n+1}) \to \mathcal{C}^*(X_{n+1})$$

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

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

Full conformal prediction

The role of randomization

Coverage or overcoverage?

Return to the theory:

Theorem: full conformal[15]

If $Z_1, \ldots, Z_{n+1}$ are exchangeable, and $\mathcal{A}$ is symmetric, then full conformal prediction satisfies

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

Proof idea:

$$S_1, \ldots, S_{n+1} \text{ exch.} \;\Rightarrow\; \mathbb{P}\{S_{n+1} \leq \text{Quantile}_{1-\alpha}(S_1, \ldots, S_{n+1})\} \geq 1 - \alpha$$

May be $> 1 - \alpha$ due to:
(1) ties between S_i's
(2) $(n+1)(1-\alpha)$ noninteger

[15] Vovk et al 2005, *Algorithmic Learning in a Random World*

The conformal p-value

A reinterpretation of conformal prediction—

Full conformal prediction (p-value version)[16]

For each $y \in \mathcal{Y}$ define a conformal p-value:

$$p^y = \frac{1 + \sum_{i=1}^{n} \mathbb{1}\left\{S_i^y \geq S_{n+1}^y\right\}}{n+1}$$

where

$$S_i^y = s^y(X_i, Y_i), i = 1, \ldots, n, \quad S_{n+1}^y = s^y(X_{n+1}, y),$$

for fitted score function $s^y = \mathcal{A}\big((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)\big)$

Then define prediction interval: $\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : p^y > \alpha\}$.

[16] Vovk et al 2005, *Algorithmic Learning in a Random World*

The conformal p-value

Interpretation: p^y is a p-value testing the hypothesis $Y_{n+1} = y$, and $\mathcal{C}(X_{n+1})$ is the set of all y's that are *not* rejected:

$$\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : p^y > \alpha\}$$

$S_1^{Y_{n+1}}, \ldots, S_{n+1}^{Y_{n+1}}$ exchangeable $\iff$ $p^{Y_{n+1}}$ is $\underbrace{\text{superuniform}}_{\text{i.e., } \mathbb{P}\left\{p^{Y_{n+1}} \leq \alpha\right\} \leq \alpha}$

The conformal p-value

Use randomization (also called "smoothing")
to make the p-value exactly Unif[0, 1]:

$$p^y = \frac{\sum_{i=1}^{n} \mathbb{1}\left\{S_i^y > S_{n+1}^y\right\} + U \cdot \left[1 + \sum_{i=1}^{n} \mathbb{1}\left\{S_i^y = S_{n+1}^y\right\}\right]}{n+1}$$

where $U \sim \text{Unif}[0, 1]$

- Can also be applied to split CP as a special case
- Note: typically use the *same* draw of U for *all* values $y \in \mathcal{Y}$

Randomized algorithms

Another source of randomness: the algorithm $\mathcal{A}$ may use...

- Stochastic gradient descent (SGD)
- Random initialization
- Resampling / ensembling / CV
- etc.

random seed $\sim$ Unif$[0,1]$

New notation: score function $s = \mathcal{A}(Z_1, \ldots, Z_m; \xi)$

The symmetric algorithm assumption (randomized version)

For any $Z_1, \ldots, Z_m$ and any $\sigma \in \mathcal{S}_m$,

$$\mathcal{A}(Z_1, \ldots, Z_m; \xi) = \mathcal{A}(Z_{\sigma(1)}, \ldots, Z_{\sigma(m)}; \xi')$$

for some coupling of $\xi \sim \text{Unif}[0,1]$ and $\xi' \sim \text{Unif}[0,1]$.

Summary

Summary & preview

Summary: lecture 2

Lecture 2 topics:

- Full conformal prediction: method & theory
- Computational strategies for full CP
- Perspectives on score functions, symmetry, & randomization

Full CP allows us to start with any algorithm/model,
and then calibrate it to have valid predictive coverage
without data splitting (but at a high computational cost)

Preview: lecture 3

Cross-validation based methods offer a compromise between split CP (computational efficiency) & full CP (statistical effic.)

- Cross-conformal prediction
- Jackknife+ and CV+
- Connections to algorithmic stability