# An introduction to conformal prediction and distribution-free inference

## CIRM tutorial (part 1)

Rina Foygel Barber (University of Chicago)

CIRM December 2024

http://rinafb.github.io/

# Introduction

# Regression & prediction

Supervised learning setting:

$$\text{Training data } (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$$

Goals:

- Inference on the regression — model distribution of $Y$ given $X$
- Predictive inference — predict value of $Y$ given $X$
  for test points $(X_{n+1}, Y_{n+1}), (X_{n+2}, Y_{n+2}), \ldots$

## Regression & prediction — classical approach

- We assume a parametric model on $(X, Y)$ or on $Y \mid X$

- We perform estimation & inference on the parameters....

- ....& then we can provide prediction intervals:

# Regression & prediction — classical approach

- We assume a parametric model on $(X, Y)$ or on $Y \mid X$

  e.g., for linear regression, $Y = X^\top \beta + \mathcal{N}(0, \sigma^2)$

- We perform estimation & inference on the parameters....

  e.g., for linear regression, distribution of $\widehat{\beta}$ and $\widehat{\sigma}^2$

- ....& then we can provide prediction intervals:

  e.g., for linear regression, $Y_{n+1} \in X_{n+1}^\top \widehat{\beta} \pm \ldots\ldots.$

## Regression & prediction — nonparametric approach

- We allow a nonparametric model for $(X, Y)$ or $Y \mid X$,
  with assumptions/constraints

- We perform estimation & inference on the model....

- ....& then we can provide prediction intervals:

## Regression & prediction — nonparametric approach

- We allow a nonparametric model for $(X, Y)$ or $Y \mid X$,
  with assumptions/constraints

  e.g., assume $\mathbb{E}[Y \mid X]$ is smooth

- We perform estimation & inference on the model....

  e.g., $\widehat{\mu}(x) =$ estimate of $\mathbb{E}[Y \mid X = x]$, via a Gaussian kernel

- ....& then we can provide prediction intervals:

  e.g., $Y_{n+1} \in \widehat{\mu}(X_{n+1}) \pm \ldots\ldots$

## Regression & prediction — ML approach

- Train an overparametrized model for $Y \mid X$

- Provide predictions for new feature vectors

- Use a data-driven strategy for uncertainty quantification

## Regression & prediction — ML approach

- Train an overparametrized model for $Y \mid X$

  e.g., train a neural net on $\{(X_i, Y_i)\}$

- Provide predictions for new feature vectors

  e.g., $\widehat{Y}_{n+i}$ = neural net's prediction for feature $X_{n+i}$

- Use a data-driven strategy for uncertainty quantification

  e.g., holdout data / cross-validation / bootstrapping / etc

# Regression & prediction — challenges

What can go wrong?

- For the parametric approach — our model may be wrong
- For the nonparametric approach — our assumptions (e.g., smoothness) may not hold
- For the ML approach — is data-driven inference guaranteed to give valid answers?

## Regression & prediction — challenges

Our choices:

- Rely on assumptions being correct
- Or, test empirically whether our assumptions hold
- Or, use inference methods that don't rely on assumptions (or, only rely on weaker assumptions)

# Regression & prediction — data-driven predictive inference

Setting:

- Features $X \in \mathcal{X}$, response $Y \in \mathbb{R}$ (or $Y \in \mathcal{Y}$)
- Available training data $(X_1, Y_1), \ldots, (X_n, Y_n) \rightsquigarrow$ fit model $\widehat{\mu}$
- Goal: given $X_{n+1}, X_{n+2}, \ldots$, predict $Y_{n+1}, Y_{n+2}, \ldots$

## Regression & prediction — data-driven predictive inference

Setting:

- Features $X \in \mathcal{X}$, response $Y \in \mathbb{R}$ (or $Y \in \mathcal{Y}$)
- Available training data $(X_1, Y_1), \dots, (X_n, Y_n) \rightsquigarrow$ fit model $\widehat{\mu}$
- Goal: given $X_{n+1}, X_{n+2}, \dots$, predict $Y_{n+1}, Y_{n+2}, \dots$

| Prediction? | or predictive inference? |
|---|---|
| $\widehat{Y}_{n+i} = \widehat{\mu}(X_{n+i})$ | $Y_{n+i} \in \widehat{\mu}(X_{n+i}) \pm \underbrace{\text{(margin of error)}}_{\text{how to calculate?}}$ |

# Using the training set for inference

Using the training loss:

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

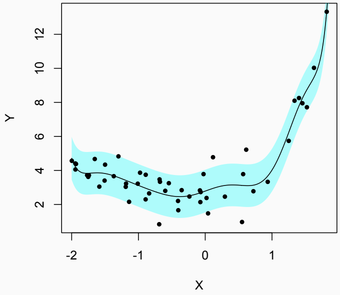
$$\underbrace{|Y_{n+i} - \widehat{\mu}(X_{n+i})|}_{\text{test error}} \gg \underbrace{\frac{1}{n}\sum_{i=1}^{n} |Y_i - \widehat{\mu}(X_i)|}_{\text{avg. training error}}$$

even if training & test data are from the same distribution

# Regression & prediction — data-driven predictive inference

Simulation: suppose we construct prediction intervals as

$$\mathcal{C}(X_{n+i}) = \widehat{\mu}(X_{n+i}) \pm \mathrm{Quantile}_{1-\alpha}\big(\underbrace{|Y_1 - \widehat{\mu}(X_1)|, \ldots, |Y_n - \widehat{\mu}(X_n)|}_{\text{residuals on training data}}\big)$$

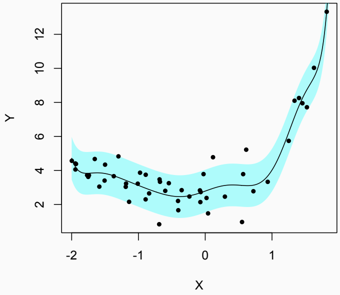



Train: 90% coverage

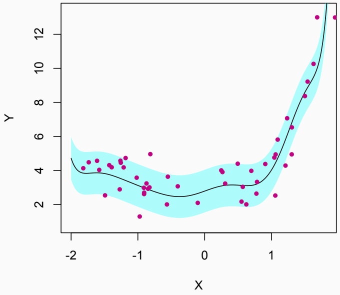
# Regression & prediction — data-driven predictive inference

Simulation: suppose we construct prediction intervals as

$$\mathcal{C}(X_{n+i}) = \widehat{\mu}(X_{n+i}) \pm \text{Quantile}_{1-\alpha}(\underbrace{|Y_1 - \widehat{\mu}(X_1)|, \ldots, |Y_n - \widehat{\mu}(X_n)|}_{\text{residuals on training data}})$$




Train: 90% coverage




Test: 78% coverage

# Using a holdout set for inference

To avoid overfitting — use a holdout set ("calibration set")

- Split the training data, $n = n_0 + n_1$
- Fit model $\widehat{\mu}$ on pretraining set $\{(X_i, Y_i)\}_{1 \leq i \leq n_0}$
- Compute residuals on calibration set, $\{|Y_i - \widehat{\mu}(X_i)|\}_{n_0 < i \leq n}$
- Prediction interval:

$$\mathcal{C}(X_{n+i}) = \widehat{\mu}(X_{n+i}) \pm \mathrm{Quantile}_{1-\alpha}\big(\{|Y_i - \widehat{\mu}(X_i)|\}_{n_0 < i \leq n}\big)$$

# Using a holdout set for inference

To avoid overfitting — use a holdout set ("calibration set")

- Split the training data, $n = n_0 + n_1$
- Fit model $\widehat{\mu}$ on pretraining set $\{(X_i, Y_i)\}_{1 \le i \le n_0}$
- Compute residuals on calibration set, $\{|Y_i - \widehat{\mu}(X_i)|\}_{n_0 < i \le n}$
- Prediction interval:

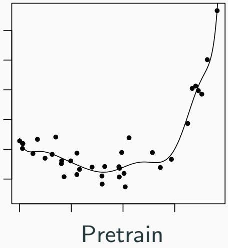
$$\mathcal{C}(X_{n+i}) = \widehat{\mu}(X_{n+i}) \pm \mathrm{Quantile}_{1-\alpha}\big(\{|Y_i - \widehat{\mu}(X_i)|\}_{n_0 < i \le n}\big)$$

fitted on pretraining data

computed on calibration data

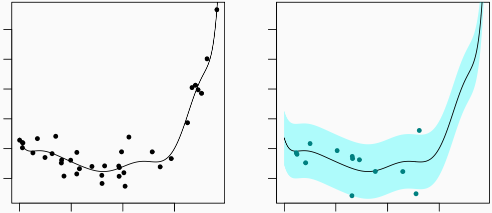
# Using a holdout set for inference

Simulation:



# Using a holdout set for inference

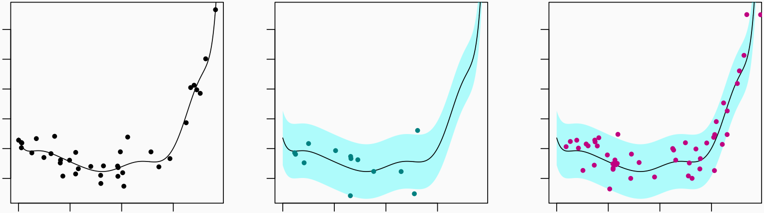
Simulation:



Pretrain　　Calibration: 90% coverage

# Using a holdout set for inference

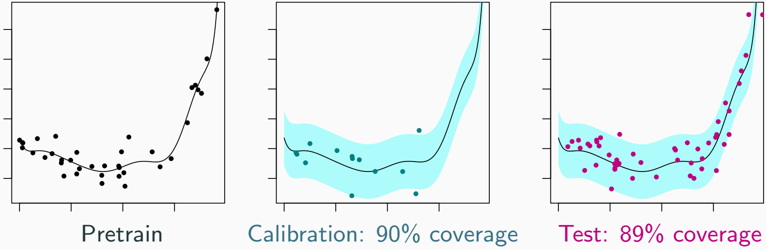
Simulation:



Pretrain    Calibration: 90% coverage    Test: 89% coverage

# Using a holdout set for inference

Simulation:




Note: lower sample size $\rightsquigarrow$ $\widehat{\mu}$ is less accurate $\rightsquigarrow$ intervals are wider

## Using a holdout set for inference

- The naive method fits a more accurate $\widehat{\mu}$,
  but the margin of error is too small due to overfitting
- A holdout set method fits a less accurate $\widehat{\mu}$,
  but the margin of error is correctly calibrated
- Can we use cross-validation (CV) to get the best of both?
  Will return to this!

# Distribution-free prediction: aims

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

---

For the problem of predictive inference...

- Can we construct a prediction interval $\mathcal{C}(X_{n+i}) \subseteq \mathcal{Y}$ such that

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

# Distribution-free prediction: aims

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

---

For the problem of predictive inference...

- Can we construct a prediction interval $\mathcal{C}(X_{n+i}) \subseteq \mathcal{Y}$ such that

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

- Want to avoid overly conservative solutions ($\mathcal{C}(X_{n+1}) = \mathcal{Y}$)
- Want to be able to use *any* regression method to construct $\mathcal{C}$ (classical or ML methods)

# Intro to exchangeability

# Introduction to exchangeability

For the rest of this talk: let $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$

**The i.i.d. data setting**

Assume $\underbrace{Z_1, \ldots, Z_n}_{\text{training}}, \underbrace{Z_{n+1}, Z_{n+2}, \ldots}_{\text{test}}$ are i.i.d. from some distrib. $P$

## Introduction to exchangeability

For the rest of this talk: let $Z_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$

**The i.i.d. data setting**

Assume $\underbrace{Z_1, \ldots, Z_n}_{\text{training}}, \underbrace{Z_{n+1}, Z_{n+2}, \ldots}_{\text{test}}$ are i.i.d. from some distrib. $P$

Can we call this "distribution-free"?

- No assumptions on $P$ (e.g., $P$ does not need to be smooth)
- But, this does not allow for dependence across time / distribution shift / etc
- We will return to these settings later

# Introduction to exchangeability

**The exchangeable data setting**

Assume that the data points

$$\underbrace{Z_1, \ldots, Z_n}_{\text{training}}, \underbrace{Z_{n+1}, Z_{n+2}, \ldots}_{\text{test}}$$

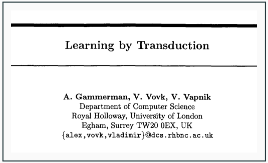
are exchangeable, i.e., $(Z_1, \ldots, Z_m) \stackrel{d}{=} (Z_{\sigma(1)}, \ldots, Z_{\sigma(m)})$ for every $m$ and every permutation $\sigma$.

- The i.i.d. data setting is a special case
- Conditionally i.i.d. data is another special case
- Note: finite sequences can be exchangeable but not i.i.d. (de Finetti's theorem does not apply)

# Background on conformal prediction

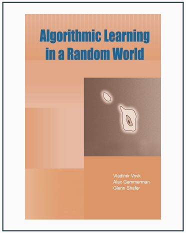
# Conformal prediction: background

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.

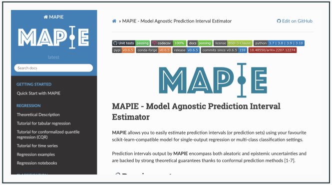
ARTICLE HISTORY
Received April 2016
Revised February 2017

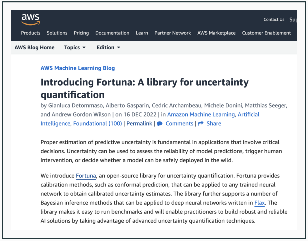
KEYWORDS
Distribution-free; Model misspecification; Prediction band; Regression; Variable importance

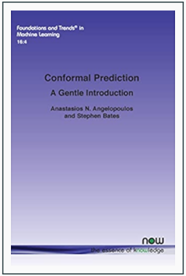
Lei, G'Sell, Rinaldo,
Tibshirani, Wasserman
JASA 2018

# Conformal prediction: background

Recent developments — software packages & user-friendly tutorials

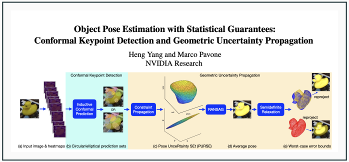


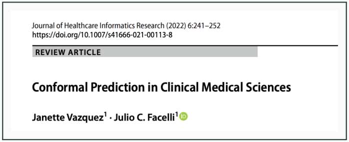

& a new theory textbook (75% on arXiv, forthcoming from CUP)

Recent developments — successful applications in biological sciences, machine learning, & many more domains


**Object Pose Estimation with Statistical Guarantees: Conformal Keypoint Detection and Geometric Uncertainty Propagation**

Heng Yang and Marco Pavone
NVIDIA Research

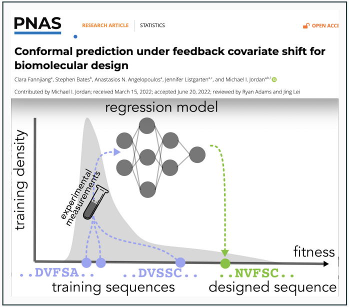



Journal of Healthcare Informatics Research (2022) 6:241–252
https://doi.org/10.1007/s41666-021-00113-8

REVIEW ARTICLE

**Conformal Prediction in Clinical Medical Sciences**

Janette Vazquez[1] · Julio C. Facelli[1]


PNAS RESEARCH ARTICLE | STATISTICS OPEN ACCE

**Conformal prediction under feedback covariate shift for biomolecular design**

Clara Fannjiang[a], Stephen Bates[b], Anastasios N. Angelopoulos[a], Jennifer Listgarten[a,c], and Michael I. Jordan[a,b,1]

Contributed by Michael I. Jordan; received March 15, 2022; accepted June 20, 2022; reviewed by Ryan Adams and Jing Lei

# Split conformal prediction

# Split conformal prediction

## The split conformal prediction method

1. Using pretraining data $Z_1, \ldots, Z_{n_0}$,
   construct fitted model $\widehat{\mu}$ using any regression algorithm:

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

2. Compute quantile $\widehat{q}$ of calibration set residuals:

$$\widehat{q} = \text{Quantile}_{(1-\alpha)(1+1/n_1)}\Big(\{|Y_i - \widehat{\mu}(X_i)|\}_{n_0 < i \leq n}\Big)$$

3. For test point $n+1$ return prediction interval

$$\mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}$$

# Split conformal prediction

**Theorem**[1]

If $Z_1, \ldots, Z_{n+1}$ are exchangeable, then split conformal satisfies:

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

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

## Split conformal prediction

**Theorem**[1]

If $Z_1, \ldots, Z_{n+1}$ are exchangeable, then split conformal satisfies:

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

**Proof:**

Define $S_i = |Y_i - \widehat{\mu}(X_i)|$ for $i = \underbrace{n_0 + 1, \ldots, n}_{\text{calibration}}, \underbrace{n+1}_{\text{test}}$

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

# Split conformal prediction

**Theorem**[1]

If $Z_1, \ldots, Z_{n+1}$ are exchangeable, then split conformal satisfies:

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

**Proof:**

Define $S_i = |Y_i - \widehat{\mu}(X_i)|$ for $i = \underbrace{n_0 + 1, \ldots, n}_{\text{calibration}}, \underbrace{n + 1}_{\text{test}}$

$$Y_{n+1} \in \mathcal{C}(X_{n+1}) \iff S_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+1/n_1)}(S_{n_0+1}, \ldots, S_n)$$

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

# Split conformal prediction

**Theorem**[1]

If $Z_1, \ldots, Z_{n+1}$ are exchangeable, then split conformal satisfies:

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

**Proof:**

Define $S_i = |Y_i - \widehat{\mu}(X_i)|$ for $i = \underbrace{n_0 + 1, \ldots, n}_{\text{calibration}}, \underbrace{n+1}_{\text{test}}$

$$\begin{aligned} Y_{n+1} \in \mathcal{C}(X_{n+1}) &\iff S_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+1/n_1)}(S_{n_0+1}, \ldots, S_n) \\ &\iff S_{n+1} \leq \text{Quantile}_{1-\alpha}(S_{n_0+1}, \ldots, S_n, S_{n+1}) \end{aligned}$$

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

# Split conformal prediction

**Exchangeability for holdout set methods**

If $\widehat{\mu} = \mathcal{A}(Z_1, \ldots, Z_{n_0})$, & the data points are exchangeable, then

$$\underbrace{|Y_{n_0+1} - \widehat{\mu}(X_{n_0+1})|, \ldots, |Y_n - \widehat{\mu}(X_n)|}_{\text{calibration residuals}}, \underbrace{|Y_{n+1} - \widehat{\mu}(X_{n+1})|}_{\text{test residual}}$$

are exchangeable.

# Split conformal prediction

**Exchangeability for holdout set methods**

If $\widehat{\mu} = \mathcal{A}(Z_1, \ldots, Z_{n_0})$, & the data points are exchangeable, then

$$\underbrace{|Y_{n_0+1} - \widehat{\mu}(X_{n_0+1})|, \ldots, |Y_n - \widehat{\mu}(X_n)|}_{\text{calibration residuals}}, \underbrace{|Y_{n+1} - \widehat{\mu}(X_{n+1})|}_{\text{test residual}}$$

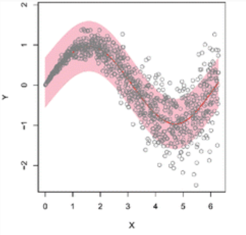
are exchangeable.

$$\Longrightarrow \mathbb{P}\left\{S_{n+1} \leq \text{Quantile}_{1-\alpha}(S_{n_0+1}, \ldots, S_n, S_{n+1})\right\} \geq 1 - \alpha$$

$\square$

# The conformal score

Due to the construction of the split conformal method,
$\mathcal{C}(X_{n+1})$ has the same width regardless of the value of $X_{n+1}$




(figure from Lei et al 2018)

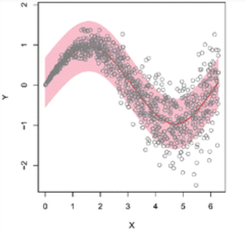
Why?

- $\mathcal{C}(X_{n+1}) = [\ \widehat{\mu}(X_{n+1}) \pm \widehat{q}\ ] = \{y \in \mathbb{R} : |y - \widehat{\mu}(X_{n+1})| \leq \widehat{q}\}$

# The conformal score

Due to the construction of the split conformal method,

$\mathcal{C}(X_{n+1})$ has the same width regardless of the value of $X_{n+1}$




(figure from Lei et al 2018)

Why?

- $\mathcal{C}(X_{n+1}) = \left[\ \widehat{\mu}(X_{n+1}) \pm \widehat{q}\ \right] = \{y \in \mathbb{R} : |y - \widehat{\mu}(X_{n+1})| \leq \widehat{q}\}$
- Equivalently: we are using $|y - \widehat{\mu}(X_{n+1})|$ as a *score* to determine whether $y$ is contained in $\mathcal{C}(X_{n+1})$ or not

# The conformal score

**The split conformal prediction method[2] (general score)**

1. Using pretraining data $Z_1, \ldots, Z_{n_0}$,
   construct score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ using any algorithm
2. Compute quantile $\widehat{q}$ of calibration set scores:

$$\widehat{q} = \mathrm{Quantile}_{(1-\alpha)(1+1/n_1)}\left(S_{n_0+1}, \ldots, S_n\right)$$

   where $S_i = s(X_i, Y_i)$

3. For test point $n+1$ return prediction interval

$$\mathcal{C}(X_{n+1}) = \{y \in \mathcal{Y} : s(X_{n+1}, y) \le \widehat{q}\}$$

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

# The conformal score

The residual score:

$$s(x, y) = |y - \widehat{\mu}(x)|, \text{ where } \widehat{\mu} \text{ fitted on pretraining data}$$

$$\implies \mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q}$$

# The conformal score

An alternative score function — (will see more examples later on)

The scaled residual score:[3]

$$s(x,y) = \frac{|y - \widehat{\mu}(x)|}{\widehat{\sigma}(x)}, \text{ where } \widehat{\mu}, \widehat{\sigma} \text{ fitted on pretraining data}$$

$$\implies \mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q} \cdot \widehat{\sigma}(X_{n+1})$$



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

## The conformal score

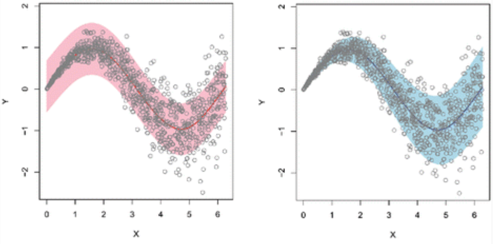
An alternative score function — (will see more examples later on)

The scaled residual score:[3]

$$s(x,y) = \frac{|y - \widehat{\mu}(x)|}{\widehat{\sigma}(x)}, \text{ where } \widehat{\mu}, \widehat{\sigma} \text{ fitted on pretraining data}$$

$$\implies \mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \widehat{q} \cdot \widehat{\sigma}(X_{n+1})$$

Compare to residual score:




(figure from Lei et al 2018)

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

## Split conformal: summary

Split CP allows us to start with any pretrained model/score,
and then calibrate it to have valid predictive coverage
(as long as we can assume exchangeability!)

Drawback: model $\widehat{\mu}$ (or score $s$) less accurate due to data splitting

# Full conformal prediction

# Full conformal prediction

Intuition—

- Split CP fits $\widehat{\mu}$ to part of the data, to ensure $S_i$'s are exch.
- Full CP: use all the data for fitting $\widehat{\mu}$ *and* ensure $S_i$'s are exch.

# Full conformal prediction

Intuition—

- Split CP fits $\widehat{\mu}$ to part of the data, to ensure $S_i$'s are exch.
- Full CP: use all the data for fitting $\widehat{\mu}$ *and* ensure $S_i$'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

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

- Check if $S_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+1/n)}(S_1, \ldots, S_n)$

# 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

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

- Check if $S_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+1/n)}(S_1, \ldots, S_n)$

$\nwarrow$

If data points are exchangeable, and $\mathcal{A}$ is symmetric,
then $S_1, \ldots, S_{n+1}$ are exchangeable
$\Rightarrow$ this event has $\geq 1 - \alpha$ probability

# Full conformal prediction

Running full conformal in practice:

- Fit model to training+test data
$$\widehat{\mu}^{y} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$$
- Compute residuals
$$S_i^y = |Y_i - \widehat{\mu}^y(X_i)|, \ i = 1, \ldots, n, \quad S_{n+1}^y = |y - \widehat{\mu}^y(X_{n+1})$$
- Check if $S_{n+1}^y \leq \text{Quantile}_{(1-\alpha)(1+1/n)}(S_1^y, \ldots, S_n^y)$

# Full conformal prediction

Running full conformal in practice:

- Fit model to training+test data
$$\widehat{\mu}^{y} = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$$
- Compute residuals
$$S_i^{y} = |Y_i - \widehat{\mu}^{y}(X_i)|, \ i = 1, \ldots, n, \quad S_{n+1}^{y} = |y - \widehat{\mu}^{y}(X_{n+1})$$
- Check if $S_{n+1}^{y} \leq \mathrm{Quantile}_{(1-\alpha)(1+1/n)}(S_1^{y}, \ldots, S_n^{y})$

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

# Full conformal prediction

Running full conformal in practice:

Test value $y \in \mathcal{Y}$

- Fit model to training+test data
$$\widehat{\mu}^y = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$$
- Compute residuals
$$S_i^y = |Y_i - \widehat{\mu}^y(X_i)|, \ i = 1, \ldots, n, \quad S_{n+1}^y = |y - \widehat{\mu}^y(X_{n+1})$$
- Check if $S_{n+1}^y \leq \text{Quantile}_{(1-\alpha)(1+1/n)}(S_1^y, \ldots, S_n^y)$

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

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

if No: discard $y$

# Full conformal prediction

Running full conformal in practice:

Test value $y \in \mathcal{Y}$

- Fit model to training+test data
$$\widehat{\mu}^y = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$$
- Compute residuals
$$S_i^y = |Y_i - \widehat{\mu}^y(X_i)|, \ i = 1, \ldots, n, \quad S_{n+1}^y = |y - \widehat{\mu}^y(X_{n+1})|$$
- Check if $S_{n+1}^y \leq \text{Quantile}_{(1-\alpha)(1+1/n)}(S_1^y, \ldots, S_n^y)$

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

if Yes: add $y$ to $\mathcal{C}(X_{n+1})$ if No: discard $y$

Note: split CP can be viewed as a special case of full CP:

$\mathcal{A}$ returns a *pretrained* model $\widehat{\mu}$ — doesn't depend on data

# Full conformal prediction

**Theorem: full conformal**[4]

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**:

1. Need to verify
   $Y_{n+1} \in \mathcal{C}(X_{n+1}) \iff S_{n+1} \leq \text{Quantile}_{1-\alpha}(S_1, \ldots, S_n, S_{n+1})$
2. Need to verify that $S_1, \ldots, S_n, S_{n+1}$ are exchangeable

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

# Full conformal prediction

❶ $Y_{n+1} \in \mathcal{C}(X_{n+1}) \iff S_{n+1} \leq \text{Quantile}_{1-\alpha}(S_1, \ldots, S_n, S_{n+1})$

---

By construction,

$$\text{If } y = Y_{n+1} \rightsquigarrow \underbrace{S_i^y}_{\text{train } \mathcal{A} \text{ on } (X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)} = \underbrace{S_i}_{\text{train } \mathcal{A} \text{ on } (X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})} \text{ for all } i = 1, \ldots, n+1$$

# Full conformal prediction

❶ $Y_{n+1} \in \mathcal{C}(X_{n+1}) \iff S_{n+1} \leq \text{Quantile}_{1-\alpha}(S_1, \ldots, S_n, S_{n+1})$

---

By construction,

$$\text{If } y = Y_{n+1} \ \rightsquigarrow \ S_i^y = S_i \text{ for all } i = 1, \ldots, n+1$$

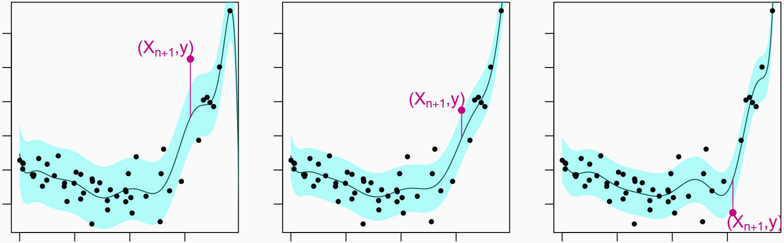
train $\mathcal{A}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)$

train $\mathcal{A}$ on $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$

$$\begin{aligned} Y_{n+1} \in \mathcal{C}(X_{n+1}) &\iff S_{n+1}^{Y_{n+1}} \leq \text{Quantile}_{(1-\alpha)(1+1/n)}(S_1^{Y_{n+1}}, \ldots, S_n^{Y_{n+1}}) \\ &\iff S_{n+1} \leq \text{Quantile}_{(1-\alpha)(1+1/n)}(S_1, \ldots, S_n) \\ &\iff S_{n+1} \leq \text{Quantile}_{1-\alpha}(S_1, \ldots, S_n, S_{n+1}) \end{aligned}$$

# Full conformal prediction

How full conformal is run:




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

# Full conformal prediction: general score

Full CP can be run with any conformal score function

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)**

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

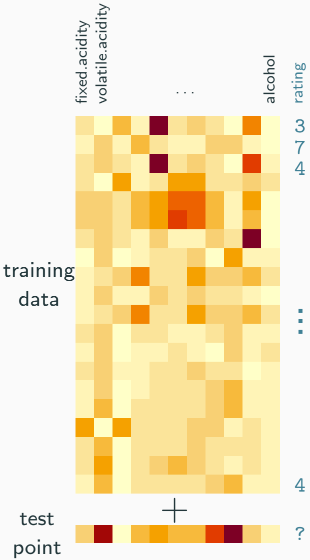
where

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

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

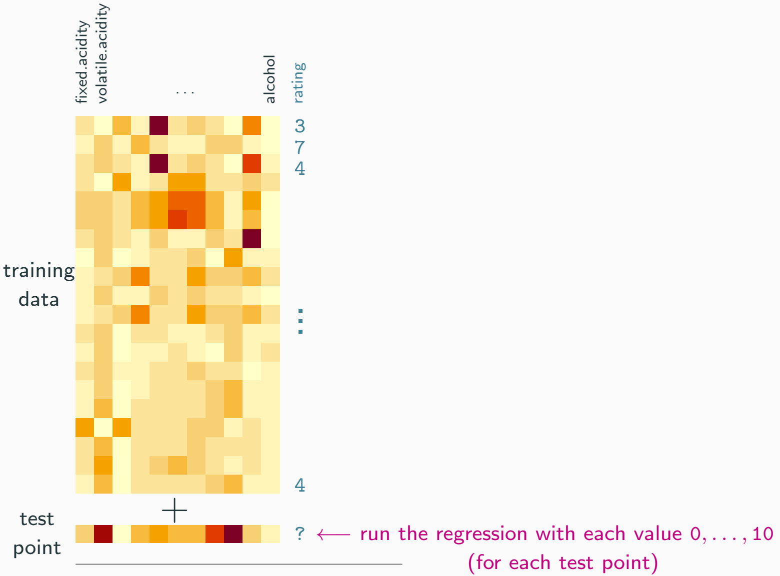
# Full conformal prediction: general score

Example: the `Wine Quality` data set[5]




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

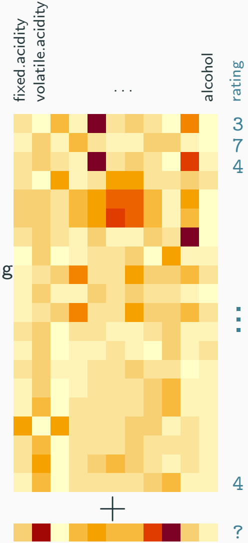
Example: the `Wine Quality` data set[5]



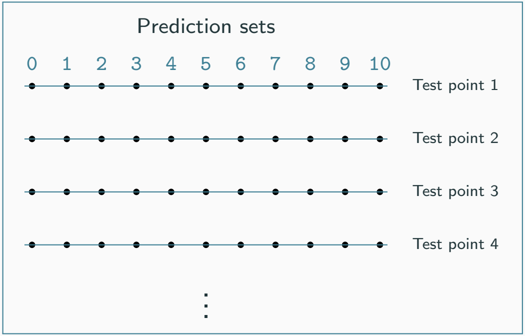

[5] Cortez et al 2009, `Wine Quality` data set, UCI Machine Learning Repository

Example: the `Wine Quality` data set[5]

training data




test point




? ⟵ run the regression with each value $0, \ldots, 10$
(for each test point)

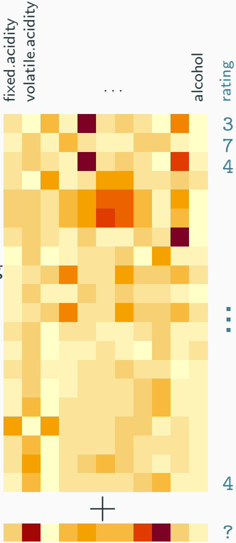
[5] Cortez et al 2009, `Wine Quality` data set, UCI Machine Learning Repository

# Full conformal prediction: general score

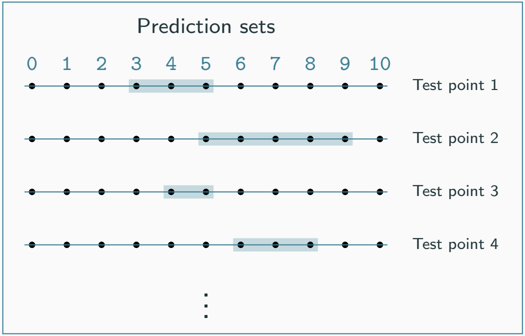
Example: the `Wine Quality` data set[5]

training data

test point



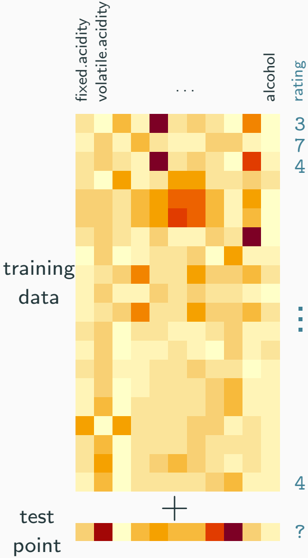

$\longleftarrow$ run the regression with each value $0, \ldots, 10$
(for each test point)

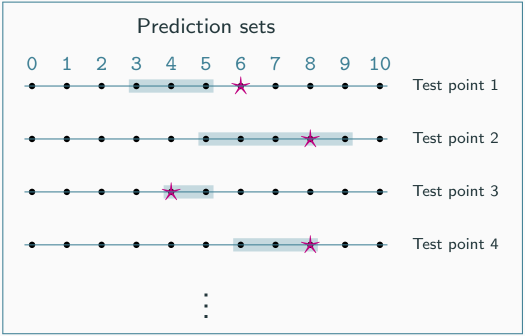



[5] Cortez et al 2009, `Wine Quality` data set, UCI Machine Learning Repository

# Full conformal prediction: general score

Example: the `Wine Quality` data set[5]




⟵ run the regression with each value $0, \ldots, 10$
(for each test point)

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

## Full conformal prediction: computational challenges

For a real-valued response...

- Running full CP requires refitting model for *every* value $y \in \mathbb{R}$

# Full conformal prediction: computational challenges

For a real-valued response...

- Running full CP requires refitting model for *every* value $y \in \mathbb{R}$

Summary of approaches used in practice:

- Most common — restrict to a grid of $y$ values (but no theory)
- Can use a discretized version of $\mathcal{A}$ to restore theory[6]
- Specialized methods for specific algorithms/settings,
  e.g., Ridge,[7] Lasso,[8] stable algorithms[9]

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

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

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

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

# Another look at the theory

Recall theoretical guarantee for split CP & full CP:

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

Limitations:

- Coverage is *marginal*, may not hold conditional on $X_{n+1}$ — what if we undercover for certain subpopulations?
- Requires exchangeability — what if there is distribution drift?
- Full CP requires symmetric $\mathcal{A}$

See Part 2 for some methods to address these limitations

# Conformal + CV

# Using cross-validation for inference

Summarizing different methods:

- Split CP fits $\widehat{\mu}$ to part of the data $\rightsquigarrow$ distrib.-free theory
- Full CP: use all the data for $\widehat{\mu}$ *and* achieves distrib.-free theory, but computationally very expensive
- Can cross-validation based methods offer a compromise?

# Using cross-validation for inference

Leave-one-out CV (the "jackknife"):

$$\mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \text{Quantile}_{1-\alpha}\big(\{|Y_i - \widehat{\mu}_{-i}(X_i)|\}_{i=1,\dots,n}\big)$$

fitted on all data

$\widehat{\mu}_{-i}$ fitted on $\{(X_\ell, Y_\ell)\}_{\ell \neq i}$
(the leave-one-out model)

# Using cross-validation for inference

Leave-one-out CV (the "jackknife"):

$$\mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \text{Quantile}_{1-\alpha}\big(\{|Y_i - \widehat{\mu}_{-i}(X_i)|\}_{i=1,\dots,n}\big)$$

fitted on all data

$\widehat{\mu}_{-i}$ fitted on $\{(X_\ell, Y_\ell)\}_{\ell \neq i}$
(the leave-one-out model)

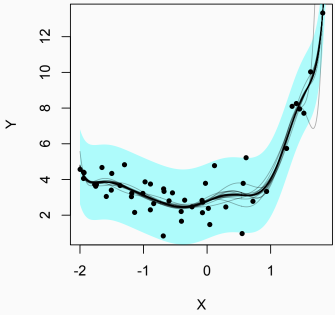
More computationally efficient: $K$-fold CV

- Partition $\{1,\dots,n\} = A_1 \cup \dots \cup A_K$, with $|A_k| = n/K$
- Fit models $\widehat{\mu}_{-A_k}$ to data $\{(X_i, Y_i)\}_{i \notin A_k}$
- Compute the margin of error using residuals

$$\big\{|Y_i - \widehat{\mu}_{-A_k}(X_i)|\big\}_{k=1,\dots,K; i \in A_k}$$

# Using cross-validation for inference

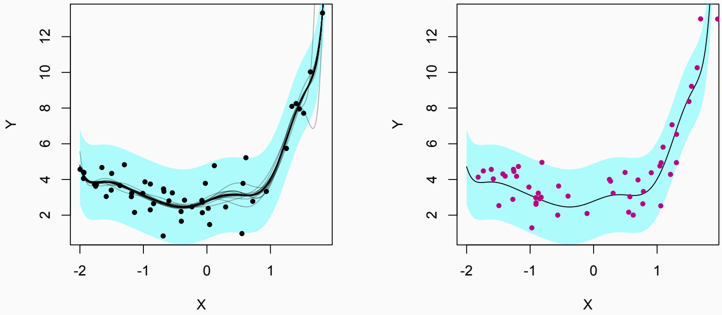
Leave-one-out CV: simulation




Train: 90% coverage

# Using cross-validation for inference

Leave-one-out CV: simulation



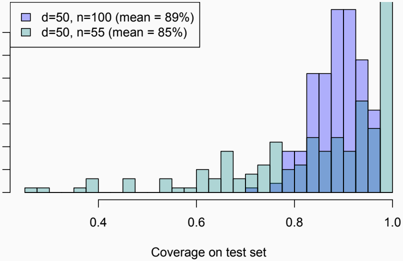

Train: 90% coverage

Test: 92% coverage

# Using cross-validation for inference

However, no assumption-free theory for CV...

Example: least squares regression + jackknife




- Theoretical guarantees under asymptotic settings
- In practice, generally we see $\approx 1 - \alpha$ coverage,
  but unstable models may lead to undercoverage

# Challenges for cross-validation

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

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

# Challenges for cross-validation

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

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

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

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

  and $\widehat{\mu}$ is pretrained $\Rightarrow S_{n_0+1}, \ldots, S_n, S_{n+1}$ are exchangeable

# Challenges for cross-validation

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

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

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

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

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

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

  $\Rightarrow S_1, \ldots, S_n, S_{n+1}$ are *not* exchangeable

# Jackknife & jackknife+

$$\text{Jackknife:} \quad \mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \text{Quantile}_{1-\alpha}\big(S_i\big)$$

Jackknife can equivalently be defined as:

$$\mathcal{C}(X_{n+1}) = \Big[\text{Quantile}_{\alpha}(\widehat{\mu}(X_{n+1}) - S_i), \text{Quantile}_{1-\alpha}(\widehat{\mu}(X_{n+1}) + S_i)\Big]$$

# Jackknife & jackknife+

$$\text{Jackknife:} \quad \mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \text{Quantile}_{1-\alpha}(S_i)$$

Jackknife can equivalently be defined as:

$$\mathcal{C}(X_{n+1}) = \left[\cancel{\text{Quantile}_{\alpha}(\widehat{\mu}(X_{n+1}) - S_i)}, \text{Quantile}_{1-\alpha}(\widehat{\mu}(X_{n+1}) + S_i)\right]$$
$$-\text{Quantile}_{1-\alpha}(-\widehat{\mu}(X_{n+1}) + S_i)$$

# Jackknife & jackknife+

$$\text{Jackknife:} \quad \mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \text{Quantile}_{1-\alpha}\big(S_i\big)$$

Jackknife can equivalently be defined as:

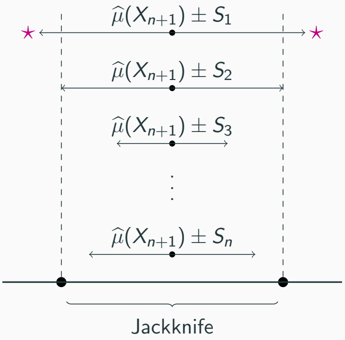
$$\begin{aligned}\mathcal{C}(X_{n+1}) = \Big[ & \cancel{\text{Quantile}_{\alpha}(\widehat{\mu}(X_{n+1}) - S_i)}, \text{Quantile}_{1-\alpha}(\widehat{\mu}(X_{n+1}) + S_i)\Big] \\ & -\text{Quantile}_{1-\alpha}(-\widehat{\mu}(X_{n+1}) + S_i)\end{aligned}$$

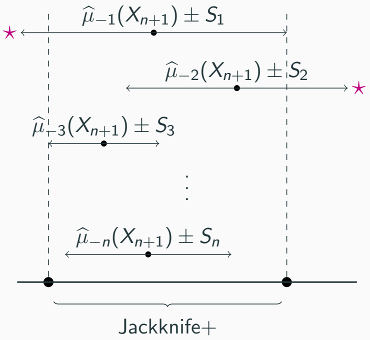
A modified version of the method: the jackknife+.[10]

$$\begin{aligned}\mathcal{C}(X_{n+1}) = \Big[ & -\text{Quantile}_{(1-\alpha)(1+1/n)}\big(-\widehat{\mu}_{-i}(X_{n+1}) + S_i\big), \\ & \text{Quantile}_{(1-\alpha)(1+1/n)}\big(\widehat{\mu}_{-i}(X_{n+1}) + S_i\big)\Big]\end{aligned}$$

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

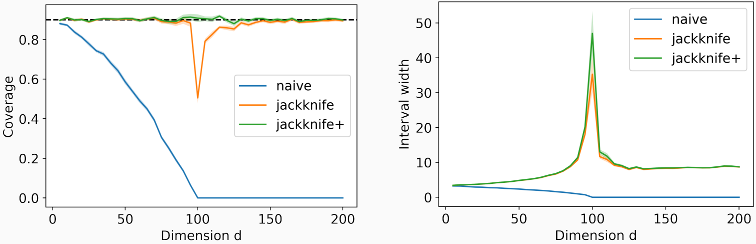
# Jackknife & jackknife+



# Jackknife & jackknife+

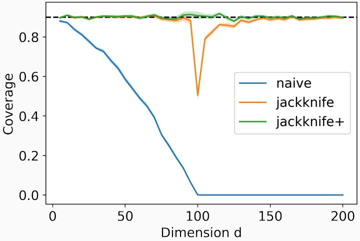
Empirical comparison (linear regression with $n = 100$):

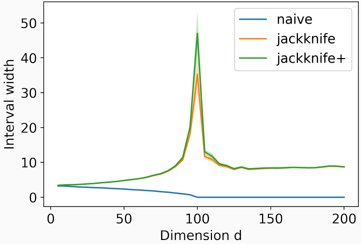



- “Ridgeless” regression — minimum-$\ell_2$-norm solution, if $d > n$

# Jackknife & jackknife+

Empirical comparison (linear regression with $n = 100$):




- "Ridgeless" regression — minimum-$\ell_2$-norm solution, if $d > n$
- Note: ridgeless regression is stable except the $d \approx n$ regime[11]

[11] Hastie et al 2022, *Surprises in High-Dimensional Ridgeless Least Squares Interpolation*

# Jackknife+ coverage guarantee

**Theorem: coverage for jackknife+**[12]

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

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - 2\alpha$$

(In contrast, jackknife may have zero coverage, in the worst case)

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

# From leave-one-out to $K$-fold

To avoid computational cost of leave-one-out CV —

$K$-fold CV (e.g., $K = 5$ or $K = 10$)

# From leave-one-out to $K$-fold

To avoid computational cost of leave-one-out CV —
$K$-fold CV (e.g., $K = 5$ or $K = 10$)

- Partition $\{1, \ldots, n\}$ into $K$ folds $A_1 \cup \cdots \cup A_K$
- Fit model $\widehat{\mu}_{-A_k} = \mathcal{A}\Big(\{(X_i, Y_i) : i \in \{1, \ldots, n\} \backslash A_k\}\Big)$
- For $i \in A_k$ define $S_i = |Y_i - \widehat{\mu}_{-A_k}(X_i)|$

$$\mathcal{C}(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \text{Quantile}_{1-\alpha}(S_1, \ldots, S_n)$$

# From leave-one-out to $K$-fold

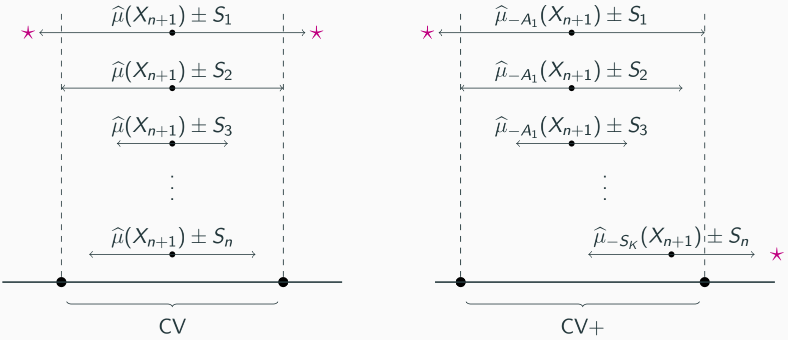
Generalize jackknife+ to the $K$-fold setting $\rightsquigarrow$ CV+

**K-fold CV+**

- Partition $\{1, \dots, n\}$ into $K$ folds $A_1 \cup \dots \cup A_K$
- Fit model $\widehat{\mu}_{-A_k} = \mathcal{A}\Big(\{(X_i, Y_i) : i \in \{1, \dots, n\} \backslash A_k\}\Big)$
- For $i \in A_k$ define $S_i = |Y_i - \widehat{\mu}_{-A_k}(X_i)|$
- Prediction set

$$\mathcal{C}(X_{n+1}) = \Big[ -\mathrm{Quantile}_{(1-\alpha)(1+1/n)}\big(\{-\widehat{\mu}_{-A_k}(X_{n+1}) + S_i\}\big), \mathrm{Quantile}_{(1-\alpha)(1+1/n)}\big(\{\widehat{\mu}_{-A_k}(X_{n+1}) + S_i\}\big) \Big]$$

# From leave-one-out to $K$-fold



# Cross-conformal prediction

CV+ is related to a more general method:

**Cross-conformal prediction**[13,14]

- Partition $\{1, \ldots, n\}$ into $K$ folds $A_1 \cup \cdots \cup A_K$
- Fit score function $s^{(k)} = \mathcal{A}\Big(\{(X_i, Y_i) : i \in \{1, \ldots, n\} \backslash A_k\}\Big)$
- For $i \in A_k$ define $S_i = s^{(k)}(X_i, Y_i)$
- Prediction set

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \sum_{k=1}^{K} \sum_{i \in A_k} \mathbb{1}\{S_i \geq s^{(k)}(X_{n+1}, y)\} \geq \alpha(n+1) \right\}$$

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

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

# Relating cross-conformal & CV+

For the residual score function $s(x, y) = |y - \widehat{\mu}(x)|$,

$$\mathcal{C}_{\text{cross-conf.}}(X_{n+1}) \subseteq \mathcal{C}_{\text{CV+}}(X_{n+1})$$

Comparison:

- Cross-conformal is more flexible (can use any score function)
- CV+ always returns an interval (by construction)

# Theory for cross-conformal & CV+

**Theorem: coverage for CV+ and cross-conformal**

If $Z_1, \ldots, Z_{n+1}$ are i.i.d., and $\mathcal{A}$ is symmetric, then $K$-fold cross-conformal satisfies

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq \begin{cases} 1 - 2\alpha - 2/K & {}^{15} \\ 1 - 2\alpha - 2K/n & {}^{16} \end{cases}$$

As a special case, the same is true for $K$-fold CV+.

$\Longrightarrow$ For any $K$,

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - 2\alpha - \frac{2}{\sqrt{n}}.$$

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

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

# Conformal methods vs model-based methods

## Conformal or classical?

In a practical application....

Should we use a model?

- Model is probably a good approximation
- Obtain more precise answers
- But, may lose coverage if assumptions don't hold

Should we use conformal prediction?

- Coverage doesn't depend on assumptions
- But, coverage guarantee is only marginal
- Would we get wider intervals (less informative)?

# Conformal or classical?

Answer: use both, & get the best of both worlds!

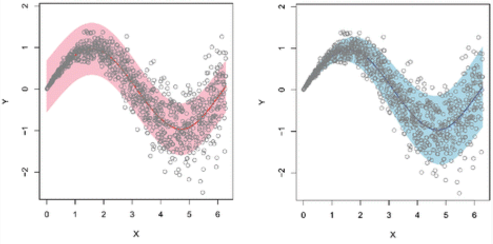
Conformal prediction is a *family* of methods

- Choosing a score specifies a particular method
- Can incorporate models / assumptions into the score

# The conformal score: examples

Recall....

1. The residual score: $s(x, y) = |y - \widehat{\mu}(x)|$
2. The scaled residual score:[17]
$$s(x, y) = \frac{|y - \widehat{\mu}(x)|}{\widehat{\sigma}(x)},$$ where $\widehat{\mu}$, $\widehat{\sigma}$ fitted on pretraining data




(figure from Lei et al 2018)

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

# The conformal score: examples

❸ Conformalized quantile regression:[18]

$$s(x, y) = \max\{y - \widehat{\gamma}_{\text{hi}}(x), \widehat{\gamma}_{\text{lo}}(x) - y\}$$

where $\underbrace{\widehat{\gamma}_{\text{lo}},\ \widehat{\gamma}_{\text{hi}}}_{\text{estimated quantiles of } Y|X}$ fitted on pretraining data

$$\implies \mathcal{C}(X_{n+1}) = [\widehat{\gamma}_{\text{lo}}(X_{n+1}) - \widehat{q},\ \widehat{\gamma}_{\text{hi}}(X_{n+1}) + \widehat{q}]$$



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

# The conformal score: examples

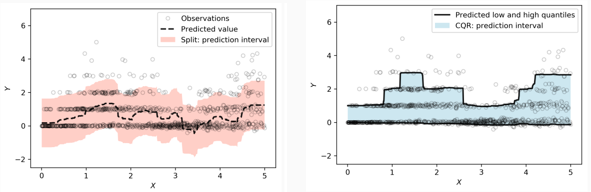
❸ Conformalized quantile regression:[18]

$$s(x, y) = \max\{y - \widehat{\gamma}_{\mathrm{hi}}(x), \widehat{\gamma}_{\mathrm{lo}}(x) - y\}$$

where $\underbrace{\widehat{\gamma}_{\mathrm{lo}}, \widehat{\gamma}_{\mathrm{hi}}}_{\text{estimated quantiles of } Y|X}$ fitted on pretraining data

$$\implies \mathcal{C}(X_{n+1}) = [\widehat{\gamma}_{\mathrm{lo}}(X_{n+1}) - \widehat{q},\ \widehat{\gamma}_{\mathrm{hi}}(X_{n+1}) + \widehat{q}]$$

Compare to residual score:




(figure from Romano et al 2019)



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

# The conformal score: examples

4. Distributional conformal prediction:[19]

$$s(x,y) = \left|\widehat{F}(y|x) - 0.5\right| \text{ where } \underbrace{\widehat{F}(\cdot|x)}_{\substack{\text{estimated conditional CDF} \\ \text{of } Y \text{ given } X = x}} \text{ is fitted on pretraining data}$$

$$\Longrightarrow \mathcal{C}(X_{n+1}) = \left[\widehat{F}^{-1}(0.5 - \widehat{q} \,|X_{n+1}),\ \widehat{F}^{-1}(0.5 + \widehat{q} \,|X_{n+1})\right]$$

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

# The conformal score: examples

❺ The high-density score:[20]

$$s(x,y) = -\widehat{f}(y|x) \text{ where } \underbrace{\widehat{f}(\cdot|x)}_{\substack{\text{estimated conditional density}\\ \text{of } Y \text{ given } X=x}} \text{ is fitted on pretraining data}$$

$$\Longrightarrow \mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \widehat{f}(y|X_{n+1}) \geq -\widehat{q} \right\}$$

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

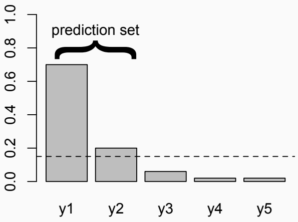
# The conformal score: examples

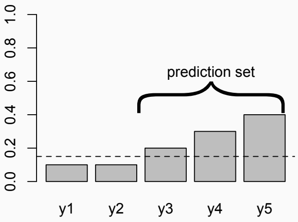
If the response $Y$ is categorical, with values $\mathcal{Y} = \{y_1, \dots, y_K\}$—

❻ The high-probability score:

$$s(x, y_k) = -\widehat{p}_k(x) \text{ where } \underbrace{\widehat{p}_k(x)}_{\text{estimate of } \mathbb{P}\{Y = y_k \mid X = x\}} \text{ is fitted on pretraining data}$$

$$\implies \mathcal{C}(X_{n+1}) = \left\{y_k : \hat{p}_k(X_{n+1}) \geq -\widehat{q}\right\}$$



# Model-based theory for conformal

A general recipe:

- Suppose that if our model is correct,
  the "oracle" answer can be written in the form

$$\mathcal{C}^*(X_{n+1}) = \{y : s^*(X_{n+1}, y) \le q^*\}$$

# Model-based theory for conformal

A general recipe:

- Suppose that if our model is correct,

  the "oracle" answer can be written in the form

$$\mathcal{C}^*(X_{n+1}) = \{y : s^*(X_{n+1}, y) \leq q^*\}$$

- Compare to the split conformal prediction set:

$$\mathcal{C}(X_{n+1}) = \{y : s(X_{n+1}, y) \leq \widehat{q}\}$$

$\rightsquigarrow$ for conformal to approximate the oracle, need

$$s \approx s^* \text{ and } \widehat{q} \approx q^*$$

# Model-based theory for conformal

A general recipe:

- Suppose that if our model is correct,

  the "oracle" answer can be written in the form

$$\mathcal{C}^*(X_{n+1}) = \{y : s^*(X_{n+1}, y) \leq q^*\}$$

- Compare to the split conformal prediction set:

$$\mathcal{C}(X_{n+1}) = \{y : s(X_{n+1}, y) \leq \widehat{q}\}$$

$\rightsquigarrow$ for conformal to approximate the oracle, need

$$s \approx s^* \text{ and } \widehat{q} \approx q^*$$

relies on fitting a good model
using the pretraining data

relies on concentration of quantiles
for calibration set scores

# Model-based theory for conformal: examples

❶ If the true model is $Y = \mu(X) + \epsilon$ where $\epsilon \perp\!\!\!\perp X$ is symmetric & unimodal noise,

$$\mathcal{C}^*(X_{n+1}) = \mu(X_{n+1}) \pm \text{Quantile}_{1-\alpha}(|\epsilon|)$$
$$= \{y : s^*(X_{n+1}, y) \leq \text{Quantile}_{1-\alpha}(|\epsilon|)\}$$

for $s^*(x, y) = |y - \mu(x)|$

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

## Model-based theory for conformal: examples

1. If the true model is $Y = \mu(X) + \epsilon$ where $\epsilon \perp\!\!\!\perp X$ is symmetric & unimodal noise,

$$\begin{aligned} \mathcal{C}^*(X_{n+1}) &= \mu(X_{n+1}) \pm \mathrm{Quantile}_{1-\alpha}(|\epsilon|) \\ &= \{y : s^*(X_{n+1}, y) \leq \mathrm{Quantile}_{1-\alpha}(|\epsilon|)\} \end{aligned}$$

for $s^*(x, y) = |y - \mu(x)|$

$\implies$ if we use the residual score, and if $\widehat{\mu} \to \mu$, then[21]

$$\mathcal{C}(X_{n+1}) \approx \mathcal{C}^*(X_{n+1}) \text{ as } n \to \infty$$

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

# Model-based theory for conformal: examples

❷ In the regression setting, suppose we would like an equal-tailed, conditional coverage guarantee.

Then the optimal set is

$$\mathcal{C}^*(X_{n+1}) = \big[\underbrace{\gamma_{\alpha/2}(X_{n+1}), \gamma_{1-\alpha/2}(X_{n+1})}_{\text{true quantiles of } Y \mid X}\big]$$

[22] Romano et al 2019, *Conformalized quantile regression*; Sesia & Candès 2020, *A comparison of some conformal quantile regression methods*

# Model-based theory for conformal: examples

❷ In the regression setting, suppose we would like an equal-tailed, conditional coverage guarantee.

Then the optimal set is

true quantiles of $Y \mid X$

$$\mathcal{C}^*(X_{n+1}) = \left[\gamma_{\alpha/2}(X_{n+1}), \gamma_{1-\alpha/2}(X_{n+1})\right]$$

Can rewrite as

$$\mathcal{C}^*(X_{n+1}) = \{y : s^*(X_{n+1}, y) \leq 0\}$$

where $s^*(x, y) = \max\left\{y - \gamma_{1-\alpha/2}(x), \gamma_{\alpha/2}(x) - y\right\}$

[22]Romano et al 2019, *Conformalized quantile regression*; Sesia & Candès 2020, *A comparison of some conformal quantile regression methods*

# Model-based theory for conformal: examples

❷ In the regression setting, suppose we would like an equal-tailed, conditional coverage guarantee.

Then the optimal set is

true quantiles of $Y \mid X$

$$\mathcal{C}^*(X_{n+1}) = \left[\gamma_{\alpha/2}(X_{n+1}), \gamma_{1-\alpha/2}(X_{n+1})\right]$$

Can rewrite as

$$\mathcal{C}^*(X_{n+1}) = \{y : s^*(X_{n+1}, y) \leq 0\}$$

where $s^*(x, y) = \max\left\{y - \gamma_{1-\alpha/2}(x), \gamma_{\alpha/2}(x) - y\right\}$

If $\widehat{\gamma}_{\text{lo}} \to \gamma_{\alpha/2}$ and $\widehat{\gamma}_{\text{hi}} \to \gamma_{1-\alpha/2}$,[22]

$$\mathcal{C}(X_{n+1}) \approx \mathcal{C}^*(X_{n+1}) \text{ as } n \to \infty$$

[22] Romano et al 2019, *Conformalized quantile regression*; Sesia & Candès 2020, *A comparison of some conformal quantile regression methods*

# Model-based theory for conformal: examples

❸ In the categorical setting, with conditional PMF $p(y \mid x)$ — smallest possible prediction set with *marginal* coverage is

$$\mathcal{C}^*(X_{n+1}) = \{y : p(y \mid x) \geq t\}$$
$$= \{y : s^*(x,y) \leq -t\} \text{ where } s^*(x,y) = -p(y \mid x)$$

$\Longrightarrow$ if we use the high-probability score, & $\widehat{p} \to p$, then[23]

$$\mathcal{C}(X_{n+1}) \approx \mathcal{C}^*(X_{n+1}) \text{ as } n \to \infty$$

[23] Sadinle et al 2019, *Least ambiguous set-valued classifiers with bounded error levels*

# Model-based theory for conformal

**Theorem (informal): asymptotic results for split CP**[24]

Assume $Z_1, Z_2, \dots$ are i.i.d., & the data split satisfies $n_0, n_1 \to \infty$.
If $s_n \to s^*$, then

$$|\mathcal{C}(X_{n+1}) \triangle \mathcal{C}^*(X_{n+1})| \to 0.$$

[24] Duchi et al 2024, *Predictive inference in multi-environment scenarios*; Angelopoulos, B., Bates 2024, *Theoretical Foundations of Conformal Prediction*

# Summary

## Summary: part 1

- Conformal allows us to start with any algorithm,
  & calibrate it to achieve (marginal) predictive coverage
- Tradeoff between statistical & computational efficiency:
  Split CP, full CP, and CV-based versions
- Conformal + model-based methods $\rightsquigarrow$ "best of both worlds"

## Preview: part 2

In part 2, we will ask if conformal can be extended to handle:

- The streaming-data setting
- Distribution shift & distribution drift
- Conditional coverage rather than only marginal coverage
- & other extensions