A Tutorial on Conformal Inference

Zhimei Ren
Department of Statistics and Data Science, University of Pennsylvania
zren@wharton.upenn.edu

Nordic Probabilistic AI School, June 2025

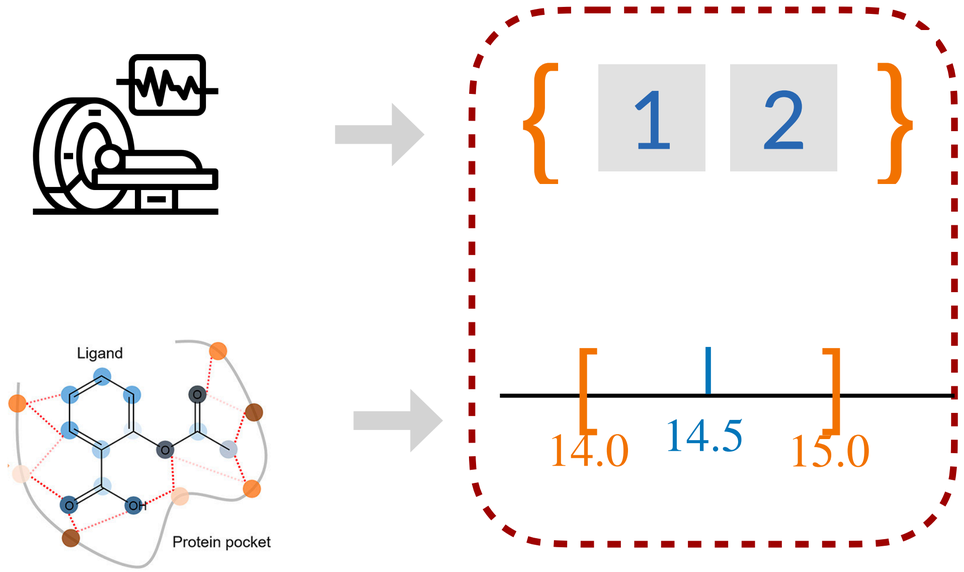
Reference books

- Vovk, Gammerman & Shafer, 2005 - *Algorithmic Learning in a Random World*
- Angelopoulos & Bates, 2023 - *Conformal Prediction: A Gentle Introduction*
- Angelopoulos, Barber & Bates, 2025 - *Theoretical Foundations of Conformal Prediction*

Part I: split conformal, full conformal, Jackknife+/CV+

Distribution-free predictive inference

- Predict the status of a disease
  - w/ medical history, examination results, …
- Predict efficacy of drug candidates
  - w/ structure of compound, type of target, …



How much do we trust the predictions?

Predictive inference aims at providing uncertainty quantification for predictions

Distribution-free predictive inference

- Quality of predictions depends on the assumptions on model / data-generation process / …
  - e.g. sparsity / smoothness / parametric / convex / …
- What if the assumptions do not hold?

Distribution-free predictive inference: providing uncertainty quantification with universally valid guarantees over all distributions

Prediction set: marginal coverage guarantees

Feature vector Response variable

- A sequence of data points $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1,2,\ldots,n$
- A new feature $X_{n+1} \rightsquigarrow$ wish to predict Y_{n+1}
- Point prediction $\hat{f}(X_{n+1})$ - margin of error?

A prediction interval $\mathcal{C}(X_{n+1}) \subseteq \mathcal{Y}$ satisfies the **marginal coverage guarantee** if

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

User-specified error level

Prediction set: marginal coverage guarantees

A prediction interval $\mathscr{C}(X_{n+1}) \subseteq \mathscr{Y}$ satisfies the **marginal coverage guarantee** if

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

- The prediction set $\mathscr{C}(\,\cdot\,)$ may depend on $(X_1, Y_1) \ldots, (X_n, Y_n)$
- The probability is over the randomness of $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$

Prediction set construction

- For now, focus on the regression task & assume $\mathcal{Y} = \mathbb{R}$
- Typically, fit a prediction model $\hat{f}$ with $(X_i, Y_i)_{i=1}^n$
- **Idea:** construct a interval via the range of $|Y_{n+1} - \hat{f}(X_{n+1})|$

Can we use the $\{|Y_i - \hat{f}(X_i)|\}_{i=1}^n$ to approximate the distribution of $|Y_{n+1} - \hat{f}(X_{n+1})|$?

Not really, can be prone to over-fitting.

The model is fitted on $(X_i, Y_i)_{i=1}^n$

Split conformal prediction[1,2]

- Split the data into two folds:[3] $\{1,\ldots,n/2\}$ and $\{n/2+1,\ldots,n\}$
- Use $(X_i, Y_i)_{1\leq i\leq n/2}$ to fit the prediction model $\hat{f}$
- Compute the residuals $S_i = |Y_i - \hat{f}(X_i)|$, for $i = n/2+1,\ldots,n$
- Let $\hat{q}$ be $\lceil (n/2+1)(1-\alpha) \rceil$ smallest element in $S_{n/2+1}, \ldots, S_n$
- Return the prediction interval

$$\mathscr{C}(X_{n+1}) = [\hat{f}(X_{n+1}) - \hat{q}, \hat{f}(X_{n+1}) + \hat{q}]$$

---

1. Papadopoulos et al 2002, *Inductive confidence machines for regression*.
2. Lei et al 2018, *Distribution-free predictive inference for regression*.
3. Assume without loss of generality that n is even.

Split conformal prediction: validity

Can be relaxed to exchangeable

Theorem.[4] If $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are i.i.d., then

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

Proof. $\mathbb{P}(Y_{n+1} \in \mathscr{C}(X_{n+1})) = \mathbb{P}(S_{n+1} \leq \hat{q})$

$= \mathbb{P}\big(S_{n+1} \leq \text{the } \lceil (1-\alpha)(n/2+1) \rceil \text{-th smallest of } S_{n/2+1}, \ldots, S_n\big)$

$\geq \mathbb{P}\big(S_{n+1} \leq \text{the } \lceil (1-\alpha)(n/2+1) \rceil \text{-th smallest of } S_{n/2+1}, \ldots, S_n, S_{n+1}\big)$

$\geq 1 - \alpha$ Because $S_{n/2+1}, \ldots, S_{n+1}$ are exchangeable conditional on the $(X_1, Y_1), \ldots, (X_{n/2}, Y_{n/2})$

4. Vovk, Gammerman, Shafer 2005, *Algorithmic learning in a random world.*

Split conformal prediction: tightness

Theorem.[5] If $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are i.i.d., and that there are no ties a.s., then

$$\mathbb{P}(Y_{n+1} \in \mathscr{C}(X_{n+1})) \leq 1 - \alpha + \frac{1}{1 + n/2}.$$

Proof. $\mathbb{P}(Y_{n+1} \in \mathscr{C}(X_{n+1})) = \mathbb{P}(S_{n+1} \leq \hat{q})$

$$= \mathbb{P}\Big(S_{n+1} \leq \text{the } \lceil (1-\alpha)(n/2+1) \rceil \text{-th smallest of } S_{n/2+1}, \ldots, S_n\Big)$$

Since no ties →

$$= \mathbb{P}\Big(S_{n+1} < \text{the } \lceil (1-\alpha)(n/2+1) \rceil \text{-th smallest of } S_{n/2+1}, \ldots, S_n\Big)$$

$$\leq \mathbb{P}\Big(S_{n+1} < \text{the } (\lceil (1-\alpha)(n/2+1) \rceil + 1) \text{-th smallest of } S_{n/2+1}, \ldots, S_n, S_{n+1}\Big)$$

$$= \frac{\lceil (1-\alpha)(n/2+1) \rceil}{n/2+1} \leq 1 - \alpha + \frac{1}{n/2+1}$$

5. Lei, et al 2018, *Distribution-free predictive inference for regression.*

Some discussion

- The core of SCP is the exchangeability among data calibration and test data points
- Holds for any prediction model / any (marginal) distribution
- Evaluation of prediction sets:
  - Validity ⇝ does the prediction set achieve the promised coverage guarantee?
  - Efficiency ⇝ how large is the set?

→
  - The form of $\mathscr{C}$
  - The "base" foundation model
  - The data-generating process
  - ...

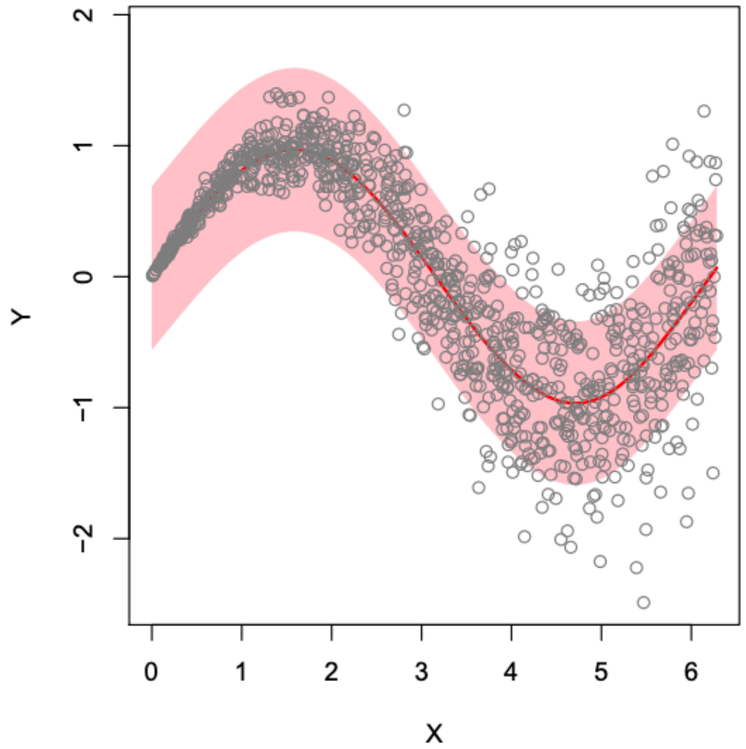
Conformity scores

Recall the prediction set we derived in the form of $\hat{f}(X_{n+1}) \pm \hat{q}$

Does not change w.r.t. X_{n+1}

Heterogeneous noise? Discrete y?

Anything unsatisfactory?



$$Y_i = \sin(X_i) + \frac{\pi |X_i|}{20} \epsilon_i$$

[Figure from Lei et al 2017]

Conformity scores

Another look at the prediction interval

$$\left[\hat{f}(X_{n+1}) - \hat{q}, \hat{f}(X_{n+1}) + \hat{q}\right] = \left\{y \in \mathbb{R} : |y - \hat{f}(X_{n+1})| \leq \hat{q}\right\}$$
$$= \left\{y \in \mathbb{R} : |y - \hat{f}(X_{n+1})| \leq Q^+_{1-\alpha}(S_{n/2+1}, \ldots, S_n)\right\}$$

↑

the $\lceil (1-\alpha)(n/2+1) \rceil$ smallest of $S_{n/2+1}, \ldots, S_n$

Comparing $|y - \hat{f}(X_{n+1})|$ with $|Y_1 - \hat{f}(X_1)|, \ldots, |Y_n - \hat{f}(X_n)|$

Conformity score function

- **Conformity score function:** $s(x, y)$
  - measures whether (x, y) "conforms" to trend observed in other data
  - A special case: the residual score $s(x, y) = |\hat{f}(x) - y|$

$$\begin{aligned}\left[\hat{f}(X_{n+1}) - \hat{q}, \hat{f}(X_{n+1}) + \hat{q}\right] &= \left\{y \in \mathbb{R} : |y - \hat{f}(X_{n+1})| \leq Q^{+}_{1-\alpha}(S_{n/2+1}, \ldots, S_n)\right\} \\ &= \left\{y \in \mathbb{R} : s(X_{n+1}, y) \leq Q^{+}_{1-\alpha}\left(S(X_{n/2+1}, Y_{n/2+1}), \ldots, S(X_n, Y_n)\right)\right\}\end{aligned}$$

Split conformal prediction: general form

- Split the data into two folds:[3] $\{1, \ldots, n/2\}$ and $\{n/2+1, \ldots, n\}$
- Use $(X_i, Y_i)_{1 \le i \le n/2}$ to fit the prediction model $\hat{f}$
- Compute the residuals $S_i = s(X_i, Y_i; \hat{f})$, for $i = n/2+1, \ldots, n$
- Let $\hat{q}$ be $\lceil (n/2+1)(1-\alpha) \rceil$ smallest element in $S_{n/2+1}, \ldots, S_n$
- Return the prediction interval

$$\mathscr{C}(X_{n+1}) = \left\{ y \in \mathscr{Y} : s(X_{n+1}, y; \hat{f}) \le \hat{q} \right\}$$

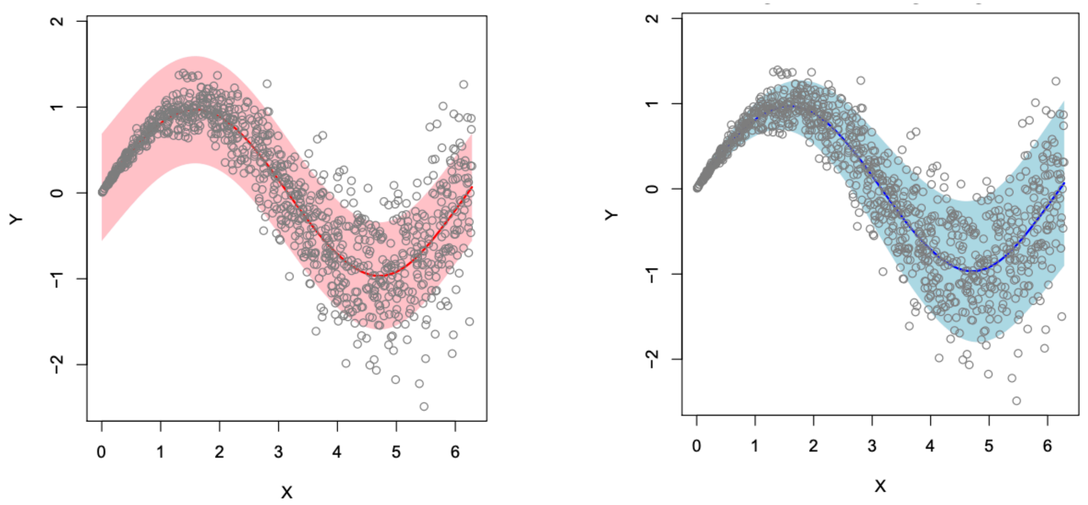
↑

Any challenge you could see in obtaining $\mathscr{C}(X_{n+1})$?

Conformity scores: scaled residual

The scaled residual score $s(x, y) = \dfrac{|y - \hat{f}(x)|}{\hat{\sigma}(x)}$

The resulting prediction set $\mathscr{C}(x) = \hat{f}(x) \pm \hat{\sigma}(x) \cdot \hat{q}$



Previous example cont'd.

✓ Improved Coverage per-x

✓ Improved length of PI

[Figure from Lei et al 2017]

Conformity scores: CQR

- The CQR (conformalized quantile regression) score[6]

$$s(x, y) = \max\left\{\hat{\tau}_{\alpha/2}(x) - y, y - \hat{\tau}_{1-\alpha/2}(x)\right\}$$

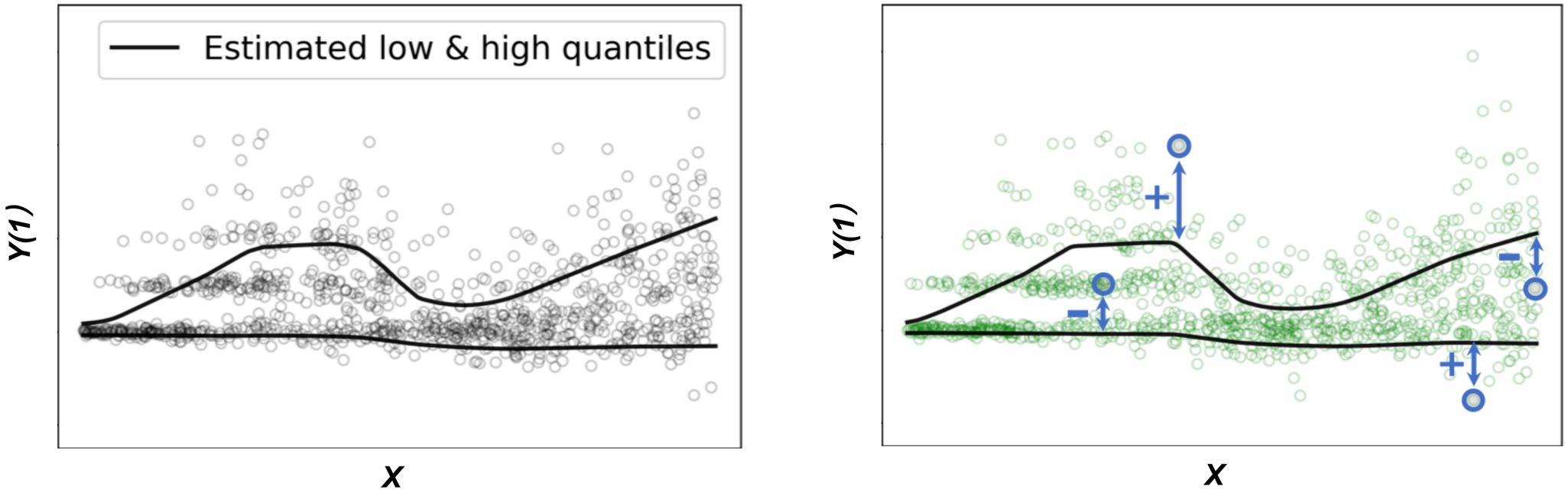
$\hat{\tau}_\beta(x)$: estimated β quantile of $Y \mid X = x$

- The resulting prediction set $\mathscr{C}(x) = \left[\hat{\tau}_\alpha(x) - \hat{q}, \hat{\tau}_{1-\alpha/2}(x) + \hat{q}\right]$

6. Romano, Patterson, Candès 2019, *Conformalized quantile regression*.

Conformity scores: CQR

Illustration of CQR

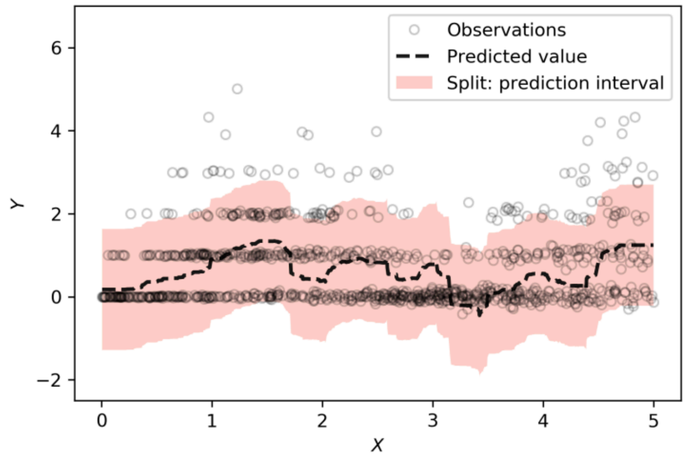


Training

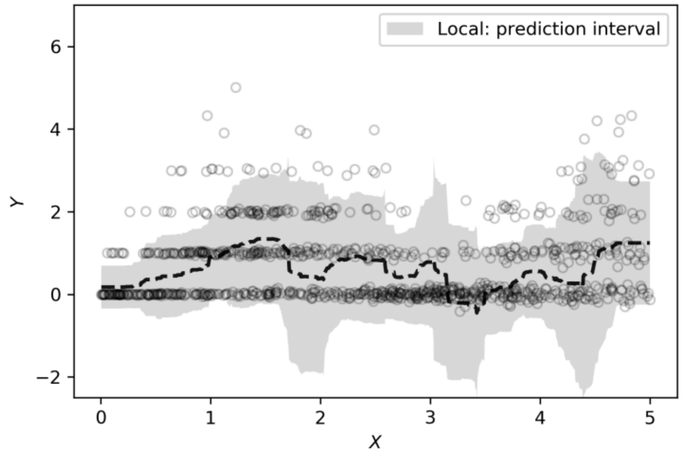
Calibration

[Figure credit: Lihua Lei]

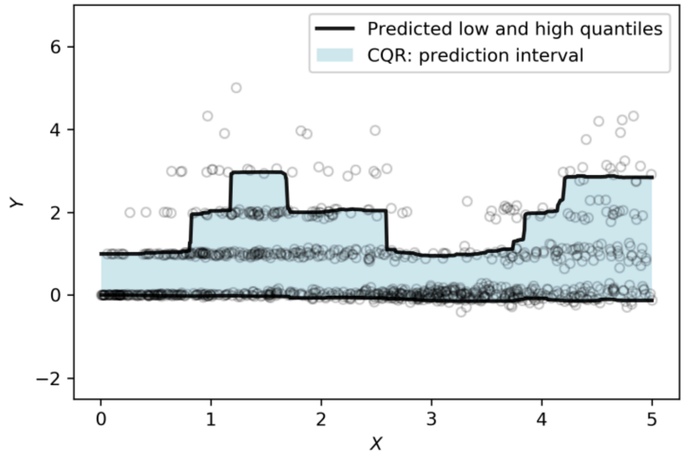
Conformity scores: CQR



Residual score



Rescaled residual score



CQR score

[Figure from Romano, Patterson, Candès 2019]

Conformity scores: high-probability

- The high-probability score[7]

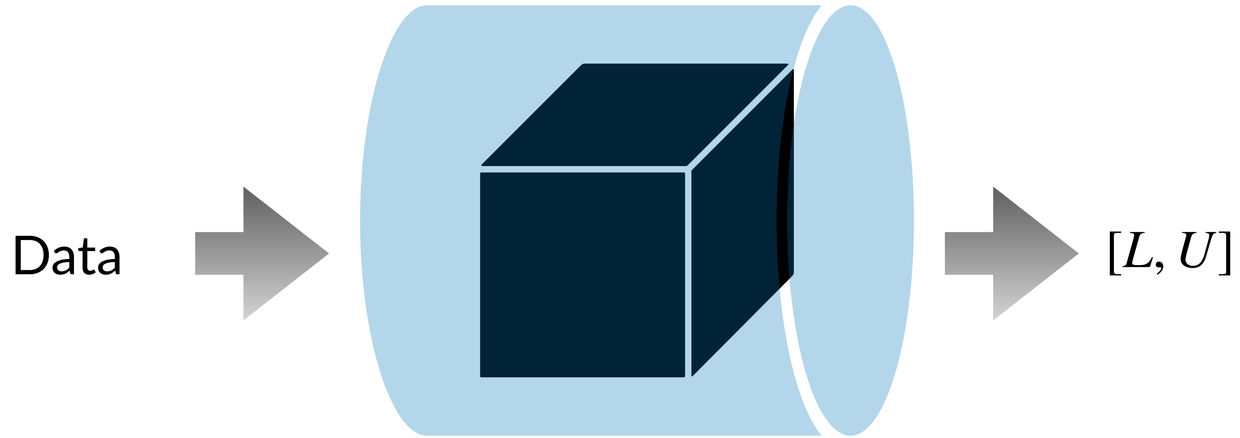
$$s(x, y) = -\hat{\pi}(y \mid x)$$

- The resulting prediction set $\mathscr{C}(X_{n+1}) = \left\{y \in \mathscr{Y} : \hat{\pi}(y \mid X_{n+1}) \geq -\hat{q}\right\}$

- If $\mathscr{Y} = \{1, \ldots, K\}$, $\hat{\pi}(k \mid x) \approx \mathbb{P}(Y = k \mid X = x)$
- If Y continuous, $\hat{\pi}(y \mid x) \approx$ conditional density

7. Sadinle, Lei, Wasserman 2019, *Least ambiguous set-valued classifiers with bounded error levels*.

Conformity scores: design principle and choice



Conformal inference: a wrapper of an arbitrary black-box algorithm

$\rightsquigarrow$ how to choose the prediction model / conformity scores / ...?

Conformity scores: design principle and choice

- “Optimal” conformity scores: depends on desiderata & modeling assumptions
- Typical goals: minimizing length / coverage in subgroups / …

Model-based design principle

- Assume a working model
- Derive the optimal form of prediction set and the corresponding conformity score
- “Conformalize” the oracle form: find the empirical version

Model-free validity + model-based efficiency

Split conformal prediction

- Training data: $n/2$ data points $\rightsquigarrow$ fitting model $\hat{f}$
- Calibration data: $n/2$ data points $\rightsquigarrow$ evaluate & calibrate $\hat{f}$
- New test point $X_{n+1} \rightsquigarrow$ prediction & UQ

Drawback: need to split the data into two halves

- Reduce the number of samples for model fitting
- Introduce external randomness (harms interpretability / reproducibility)

Full conformal prediction

- A sequence of data points $(X_i, Y_i)_{i=1,\ldots,n}$
- A new feature X_{n+1} $\rightsquigarrow$ wish to predict Y_{n+1}
- Assume $(X_i, Y_i)_{i=1}^{n+1}$ are i.i.d.

Use all the data for training & calibration:

- Imagine observing Y_{n+1}
- Fit a model $\hat{f}$ using all the data $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$
  - The model-fitting procedure cannot use the order of the data points
- Let $S_i = |Y_i - \hat{f}(X_i)|$, for $1 \leq i \leq n+1$

$$\mathbb{P}\left(S_{n+1} \leq \text{the } \lceil (1-\alpha)(n+1) \rceil \text{-th smallest of } S_1 \ldots, S_{n+1}\right) \geq 1-\alpha$$

Because $S_1, \ldots, S_{n+1}$ are exchangeable

Are we done? Still need to get a set $\mathscr{C}(X_{n+1})$!

Full conformal prediction

$$S_{n+1} \leq \text{the } \lceil (1-\alpha)(n+1) \rceil \text{-th smallest of } S_1 \ldots, S_{n+1}$$

$$\iff Y_{n+1} \in \left\{ y \in \mathcal{Y} : S^y_{n+1} \leq Q_{1-\alpha}(S^y_1, \ldots, S^y_{n+1}) \right\}$$

$$\uparrow$$

$$\mathscr{C}(X_{n+1})$$

$$\downarrow$$

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

Computation of $\mathscr{C}(X_{n+1})$ can be hard

- Need to enumerate all possible $y \in \mathcal{Y}$ satisfying the conditions
- The score function $s(\cdot, \cdot)$ also defends on $y \rightsquigarrow$ refit for every value $y \in \mathcal{Y}$
- Approximation by discretization or fast computation for special fitting algo., e.g., LASSO[8]

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

CV+ / Jackknife+

So far ...

Split conformal

- Computationally light
- Loses $n/2$ data points

Full conformal

- Computationally challenging
- Uses full data

Interpolate the two versions, approaching the "best of two worlds"?

CV+ / Jackknife+

Recap: cross validation (CV)

- Split the data into K folds: $S_1 \cup \cdots \cup S_K$
- For $i \in S_k$, fit a model $\hat{f}_{-k}$ on all the data except S_k
- Let $S_i = |\hat{f}_{-k}(X_i) - Y_i|$

Fitted on all data

Can we proceed as before and build $\mathscr{C}(X_{n+1}) = \hat{f}(X_{n+1}) \pm Q_{1-\alpha}(S_1, \ldots, S_n)$?

Not really, since $S_1, \ldots, S_n, S_{n+1}$ are not exchangeable

CV+ / Jackknife+

CV+[9]

- Split the data into K folds: $S_1 \cup \cdots \cup S_K$
- For $i \in S_k$, fit a model $\hat{f}_{-k}$ on all the data except S_k
- Let $S_i = |\hat{f}_{-k}(X_i) - Y_i|$
- Construct the prediction set

$$\mathscr{C}^{\text{CV+}}(X_{n+1}) = \left[Q^-_\alpha\left(\hat{f}_{-k(i)}(X_{n+1}) - S_i\right), Q^+_{1-\alpha}\left(\hat{f}_{-k(i)}(X_{n+1}) + S_i\right) \right]$$

Fold of unit i

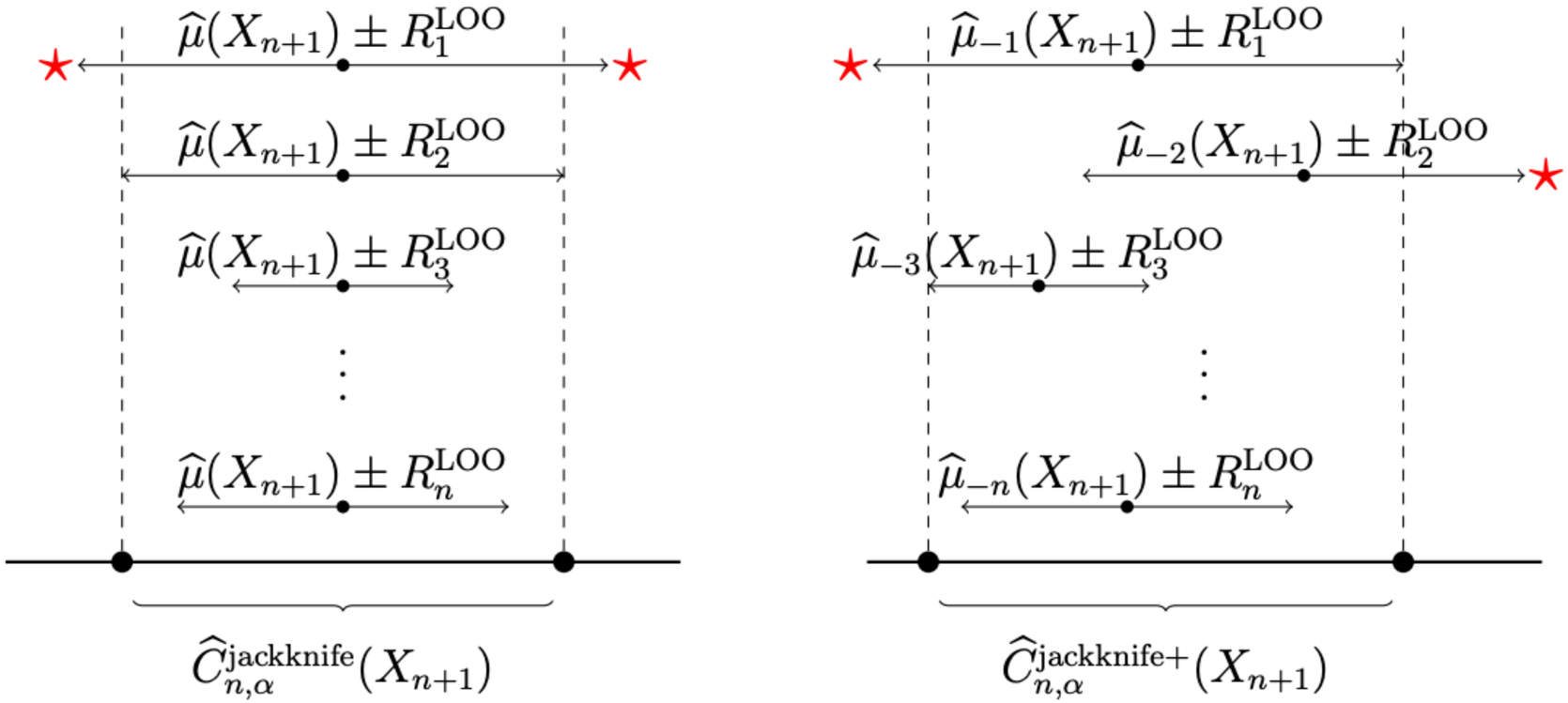
$Q^-_\beta(\ldots)$: the $\lfloor (1-\beta)(n+1) \rfloor$ smallest of $\ldots$

$Q^+_\beta(\ldots)$: the $\lceil (1-\beta)(n+1) \rceil$ smallest of $\ldots$

- Jackknife+: take $K = n$

9. Barber, Candès, Ramdas, Tibshirani 2019, *Predictive Inference with the Jackknife+*.

CV+ / Jackknife+



[Figure from Barber, Candès, Ramdas, Tibshirani 2019]

CV+ / Jackknife+

Theorem.[10] If $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are i.i.d., then

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

Theorem.[10] If $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are i.i.d., then

$$\mathbb{P}\left(Y_{n+1} \in \mathscr{C}^{\text{CV+}}(X_{n+1})\right) \geq 1 - 2\alpha - \min\left\{\frac{2(1-1/K)}{n/K+1}, \frac{1-K/n}{K+1}\right\}.$$

10. Barber, Candès, Ramdas, Tibshirani 2019, *Predictive Inference with the Jackknife+*.

Proof sketch for the validity of Jackknife+

Proof.

- Again, imagine we have access to Y_{n+1} for now
- Let $\tilde{f}_{-(i,j)}$ denote the model with $\{1,\ldots,n,n+1\}\backslash\{i,j\} \Rightarrow \hat{f}_{-i} = \tilde{f}_{-(i,n+1)}$
- Intuitively, $|\tilde{f}_{-(i,j)}(X_i) - Y_i|$ should be comparable with $|\tilde{f}_{-(i,j)}(X_j) - Y_j|$

$$R_{ij} = \begin{cases} +\infty, & \text{if } i = j, \\ |\tilde{f}_{-(i,j)}(X_i) - Y_i| & \text{if } i \neq j \end{cases} \qquad A_{ij} = \mathbf{1}\{R_{ij} > R_{ji}\}$$

⇝ A game between player i and player j

Tournament Lemma. Let $A \in \{0,1\}^{(n+1)\times(n+1)}$ satisfy $A_{ij} + A_{ji} \leq 1$ for all i, j. The for any $\alpha \in (0,1)$,

$$\sum_{i=1}^{n+1} \mathbf{1}\left\{A_{i,\cdot} \geq (n+1)(1-\alpha)\right\} \leq 2\alpha(n+1)$$

$\sum_j A_{ij}$

Proof sketch for the validity of Jackknife+

Proof (cont'd).

- By exchangeability, every point has equal probability of being a "winner"

$$\mathbb{P}\left(A_{n+1,\cdot} \geq (n+1)(1-\alpha)\right) \leq 2\alpha$$

$$\updownarrow$$

$$R_{n+1,i} > R_{i,n+1} \text{ for at least } \lceil (n+1)(1-\alpha) \rceil \text{ points}$$

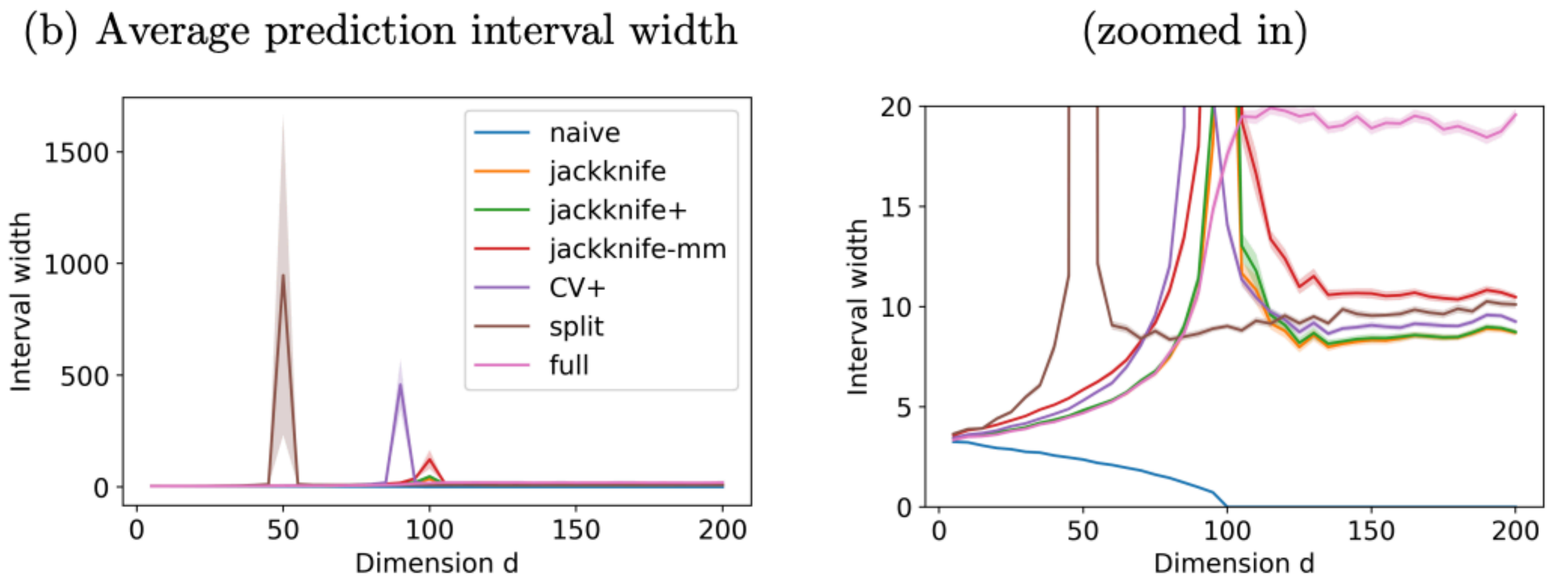
$$\updownarrow$$

$$Y_{n+1} > \hat{f}_{-i}(X_{n+1}) + S_i \text{ or } Y_{n+1} < \hat{f}_{-i}(X_{n+1}) - S_i \text{ for at least } \lceil (n+1)(1-\alpha) \rceil \text{ points}$$

Recall that $$\mathscr{C}^{\text{jackknife+}}(X_{n+1}) = \left[Q^-_\alpha\left(\hat{f}_{-i}(X_{n+1}) - S_i\right), Q^+_{1-\alpha}\left(\hat{f}_{-i}(X_{n+1}) + S_i\right)\right]$$

CV+ / Jackknife+

$n = 100$



[Figure from Barber, Candès, Ramdas, and Tibshirani, 2021.]

Summary

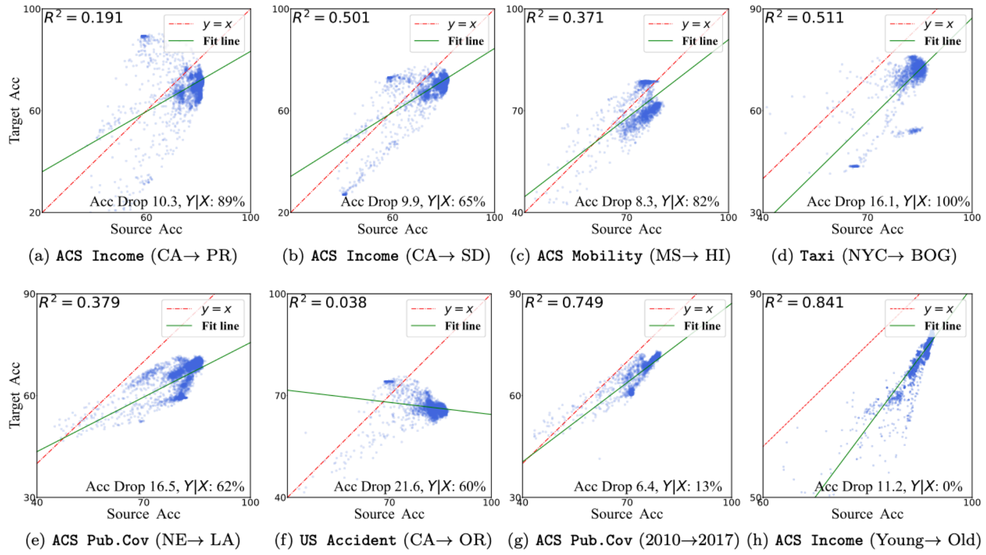
- Conformal prediction provides prediction sets w/ distribution-free guarantees
- Conformity scores are essential for the property of CP sets
- The comparison among Split CP / full CP / Jackknife+/ CV+ in terms of computation, efficiency, guarantees

Part II: distribution shift, robustness

Distribution shifts

- Key assumption in CP: training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and testing point (X_{n+1}, Y_{n+1}) jointly exchangeable
- What if this assumption is violated?

Training distribution $\neq$ Target distribution



(a) `ACS Income` (CA→ PR) (b) `ACS Income` (CA→ SD) (c) `ACS Mobility` (MS→ HI) (d) `Taxi` (NYC→ BOG)

(e) `ACS Pub.Cov` (NE→ LA) (f) `US Accident` (CA→ OR) (g) `ACS Pub.Cov` (2010→2017) (h) `ACS Income` (Young→ Old)

[Figure from Liu, Wang, Cui and Namkoong '23]

Uncertainty quantification under distribution shift

Training data	Test data
$(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$	$(X_{n+1}, Y_{n+1}) \sim Q$

Goal: use the training data to construct a prediction interval $\mathscr{C}(\cdot)$ such that

$$\mathbb{P}_{(X_{n+1},Y_{n+1})\sim Q}(Y_{n+1} \in \mathscr{C}(X_{n+1})) \geq 1 - \alpha$$

- Hopeless if Q is arbitrary
- Will consider structured distribution shift

Covariate shift

- The marginal distribution of X is different under P and Q
  - Selection based on X
  - different populations
- But $Y \mid X$ remains the same

$$\frac{dQ_X}{dP_X}(x) \propto w(x) \rightsquigarrow \text{known function}$$

  - $S_1, \ldots, S_n, S_{n+1}$ are no longer exchangeable
  - Under covariate shift $\rightsquigarrow$ weighted exchangeable

Covariate shift

Weighted split CP[11]

- Split the data into two folds: $\{1,\ldots,n/2\}$ and $\{n/2+1,\ldots,n\}$
- Use $(X_i, Y_i)_{1\le i\le n/2}$ to fit the prediction model $\hat{f}$
- Compute the conformity score S_i for $i = n/2+1,\ldots,n$
- Let $\hat{q}$ be $(1-\alpha)$ quantile of the following distribution:

δ_x: point mass at x

$$\sum_{i=n/2+1}^{n} \frac{w(X_i)}{\sum_{j=n/2+1}^{n} w(X_j) + w(X_{n+1})} \cdot \delta_{S_i} + \frac{w(X_n+1)}{\sum_{j=n/2+1}^{n} w(X_j) + w(X_{n+1})} \cdot \delta_{+\infty}$$

- Return the prediction interval

$$\mathscr{C}(X_{n+1}) = \left\{ y \in \mathscr{Y} : s(X_{n+1}, y) \le \hat{q} \right\}$$

11. Barber, Candès, Ramdas, Tibshirani 2019, *Conformal prediction under covariate shift*.

Weighted split CP

Theorem.[11] If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. samples from P and (X_{n+1}, Y_{n+1}) is an independent sample from Q, where (1) Q_X is absolutely continuous with respect to P_X with $dQ_X/dP_X(x) \propto w(x)$, and (2) $Q_{Y|X} = P_{Y|X}$, then the prediction set of weighted split CP satisfies

$$\mathbb{P}_{(X_{n+1},Y_{n+1})\sim Q}\left(Y_{n+1} \in \mathscr{C}(X_{n+1})\right) \geq 1 - \alpha$$

- Assumption can be relaxed to weighted exchangeability
- Full conformal version

11. Barber, Candès, Ramdas, Tibshirani 2019, *Conformal prediction under covariate shift*.

Covariate shift

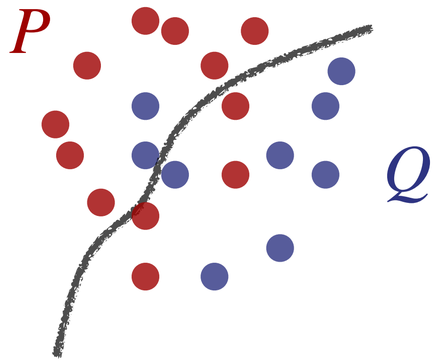
- Have assumed the weight function w to be known
- What if w unknown?

Training data	Test data
$(X_1, Y_1), \ldots, (X_n . Y_n) \sim P$	$(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m}) \sim Q$
$A_i = 0,\ i = 1,\ldots, n$	$A_i = 1,\ i = n+1,\ldots, n+m$

- Estimate w with $(X_1, A_1), \ldots, (X_{n+m}, A_{n+m})$

$$\frac{\mathbb{P}(A = 1 \mid X = x)}{\mathbb{P}(A = 0 \mid X = x)} = \frac{\mathbb{P}(A = 1)}{\mathbb{P}(A = 0)} \frac{Q_X}{P_X}(x) \propto w(x)$$

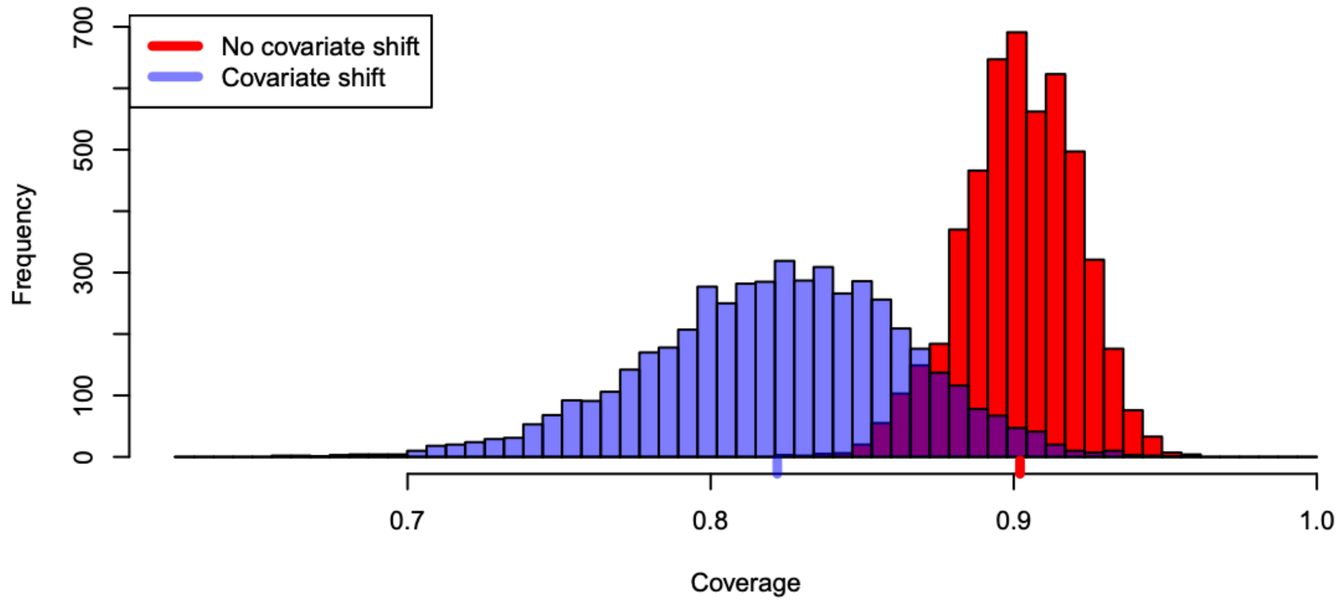
Can be estimated with classification algo.



- Coverage guarantee $\rightsquigarrow 1 - \alpha - \frac{1}{2}\mathbb{E}_{X \sim P_X}\left[\,|\, w(X) - \hat{w}(X) \,|\,\right]$. [12]

12. Lei and Candès 2021, *Conformal inference of counterfactuals and individual treatment effects*.

Weighted split CP



[Figure from Barber, Candès, Ramdas, Tibshirani, 2019]

Weighted split CP: application

Application to causal inference $\rightsquigarrow$ inference on individual treatment effect (ITE)[12]

- Covariate X, treatment $T \in \{0,1\}$, outcome Y
  - Potential outcomes $Y(1)$, $Y(0)$
  - Assumption (1): $Y = T \cdot Y(1) + (1 - T) \cdot Y(0)$
  - Assumption (2): $(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$
- Data $(X_i, T_i, Y_i) \overset{\text{i.i.d.}}{\sim} (X, T, Y)$
- Goal: prediction interval for the ITE $Y(1) - Y(0)$
- Fundamental Challenge: never simultaneously observe $Y(0)$ and $Y(1)$ on the same unit
  - For some unit with $T = 0 \rightsquigarrow$ need to predict $Y(1)$

12. Lei and Candès 2021, *Conformal inference of counterfactuals and individual treatment effects*.

Weighted split CP: application

Application to causal inference $\rightsquigarrow$ inference on individual treatment effect (ITE)[12]

- Goal: prediction set for $Y(1)$ of a control unit ($T = 0$)
- Only observe $Y(1)$ for units with $T = 1$
- Training distribution: $(X, Y(1)) \mid T = 1$
- Test distribution: $(X, Y(1)) \mid T = 0$

Distribution shift!

$$\frac{dP_{Y(1),X|T=0}}{dP_{Y(1),X|T=1}} = \frac{dP_{X|T=0}}{dP_{X|T=1}} \times \frac{dP_{Y(1)|X,T=0}}{dP_{Y(1)|X,T=1}} = \frac{dP_{X|T=0}}{dP_{X|T=1}} = \frac{P(T=1)}{P(T=0)} \times \frac{P(T=0 \mid X)}{P(T=1 \mid X)}$$

Assumption (2)

✓ Covariate shift!

12. Lei and Candès 2021, *Conformal inference of counterfactuals and individual treatment effects*.

Weighted split CP: application

Application to survival analysis

- Covariate X, survival time T, censoring time $C \rightsquigarrow$ only observe X, C, and $\tilde{T} = \min(T, C)$.[13]
  - Assumption: $T \perp\!\!\!\perp C \mid X$
- Data $(X_i, C, \tilde{T}_i) \overset{\text{i.i.d.}}{\sim} (X, C, \tilde{T})$
- Goal: prediction lower bound for the survival time T
- Challenge: only partially observe T

A lower bound on $\tilde{T}$ via CP?

Can be very conservative!

13. Type-I censoring.

Weighted split CP: application

Conformalized survival analysis[13]

- Focus on a subset w/ $C \geq c_0$ ($\tilde{T}$ closer to T)
- On the subset, $\min(\tilde{T}, c_0) = \min(T, c_0)$ $\rightsquigarrow$ aim at building lower bound for $\min(T, c_0)$
- Observe $X, C, \min(T, c_0)$ on the subset & aim at predicting $\min(T, c_0)$ $\rightsquigarrow$ apply CP?
- Caution! Subsetting causes distribution shift
- Training distribution: $(X, \min(T, c_0)) \mid C \geq c_0$
- Test distribution: $(X, \min(T, c_0))$

$$\frac{dP_{X,\min(T,c_0)}}{dP_{X,\min(T,c_0)|C\geq c_0}} = \frac{P(C \geq c_0)}{P(C \geq c_0 \mid X)}$$

Assumption

✓ Covariate shift!

14. Candès, Lei, R. 2023, Conformalized survival analysis

Partially identifiable distribution shift

- In the previous case, the target distribution is specified by w (known / estimable)
- What if target distribution Q unknown but restricted in some range?
- One way to measure the distance between P and Q is through the f-divergence

$$D_f(Q \,\|\, P) = \mathbb{E}_P[f(dQ/dP)]$$ [KL divergence, χ^2 divergence, TV distance ...]

- Fixing a robust parameter ρ, consider $Q : D_f(Q \,\|\, P) \leq \rho$.[15]

Goal: Use the training data to construct a prediction interval $\mathscr{C}(\cdot)$ such that

$$\inf_{Q: D_f(Q\|P) \leq \rho} \mathbb{P}_{(X_{n+1}, Y_{n+1}) \sim Q}\Big(Y_{n+1} \in \mathscr{C}(X_{n+1})\Big) \geq 1 - \alpha$$

15. Cauchois et al 2024, *Robust validation: confident predictions even when distributions shift*.

Distributional robust CP

Goal: Use the training data to construct a prediction interval $\mathscr{C}(\cdot)$ such that

$$\inf_{Q: D_f(Q\|P)\leq\rho} \mathbb{P}(Y_{n+1} \in \mathscr{C}(X_{n+1})) \geq 1-\alpha$$

- Follow the split CP framework $\rightsquigarrow S_{n/2}, \ldots, S_n$
- Key: finding the range of S_{n+1}
- Idea: finding the largest quantile $q(Q)$ among all the possible Q

Distributional robust CP

Goal: Use the training data to construct a prediction interval $\mathscr{C}(\cdot)$ such that

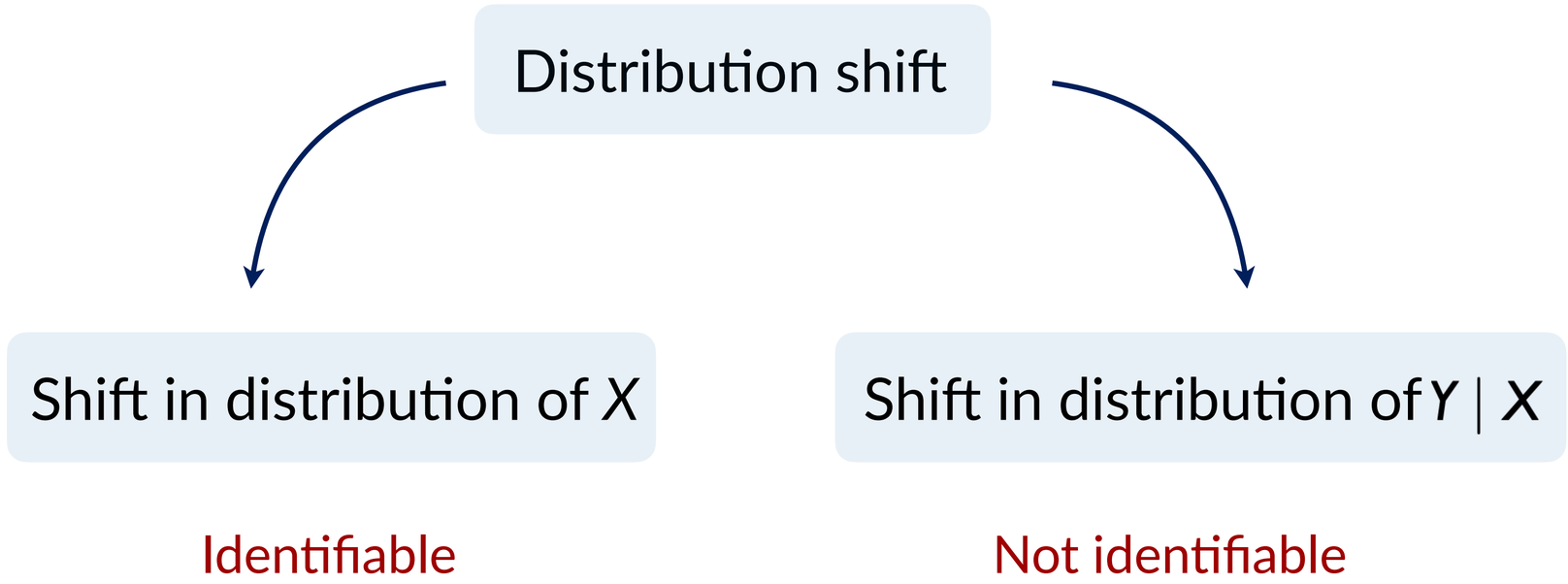
$$\inf_{Q:D_f(Q\|P)\leq\rho} \mathbb{P}(Y_{n+1} \in \mathscr{C}(X_{n+1})) \geq 1-\alpha$$

- Result: For $Q \in \{Q : D_f(Q\|P) \leq \rho\}$, the largest quantile of $s(X,Y)$ is an inflated quantile under P[15]
- Inflation level: $g_{f,\rho}^{-1}(1-\alpha) > 1-\alpha$
- $g_{f,\rho}^{-1}$ only depends on f and ρ

$$\mathscr{C}(X_{n+1}) = \left\{y \in \mathscr{Y} : s(X_{n+1}, y) \leq Q^{+}_{g_{f,\rho}^{-1}(1-\alpha)}(S_{n/2+1}, \ldots, S_n)\right\}$$

15. Cauchois et al 2024, *Robust validation: confident predictions even when distributions shift*.

A mixture approach



Guarding against the worst-case joint distribution shift can be too conservative

Fine-grained robust CP[16]

- Learn the X shift and guard against the worst-case $Y \mid X$ shift
- **Model:**
  - No constraints on the X-shift
  - $Y \mid X$-shift bounded in f-divergence: $D_f(Q_{Y|X} \,\|\, P_{Y|X}) \leq \rho$
- **Idea:** adjust for the worst-case $Y \mid X$ shift and weight the samples according to $w(x)$

$$\mathscr{C}(X_{n+1}) = \left\{ y : s(X_{n+1}, y) \leq \text{Quantile}\left(g_{f,\rho}^{-1}(1-\alpha), \sum_i \frac{w(X_i)}{\sum_{i'} w(X_{i'}) + w(X_{n+1})} \cdot \delta_{S_i} + \frac{w(X_{n+1})}{\sum_{i'} w(X_{i'}) + w(X_{n+1})} \cdot \delta_{+\infty} \right) \right\}$$

16. Ai and R. 2024, *Not all distributional shifts are equal: Fine-grained robust conformal inference.*

Fine-grained robust CP[16]

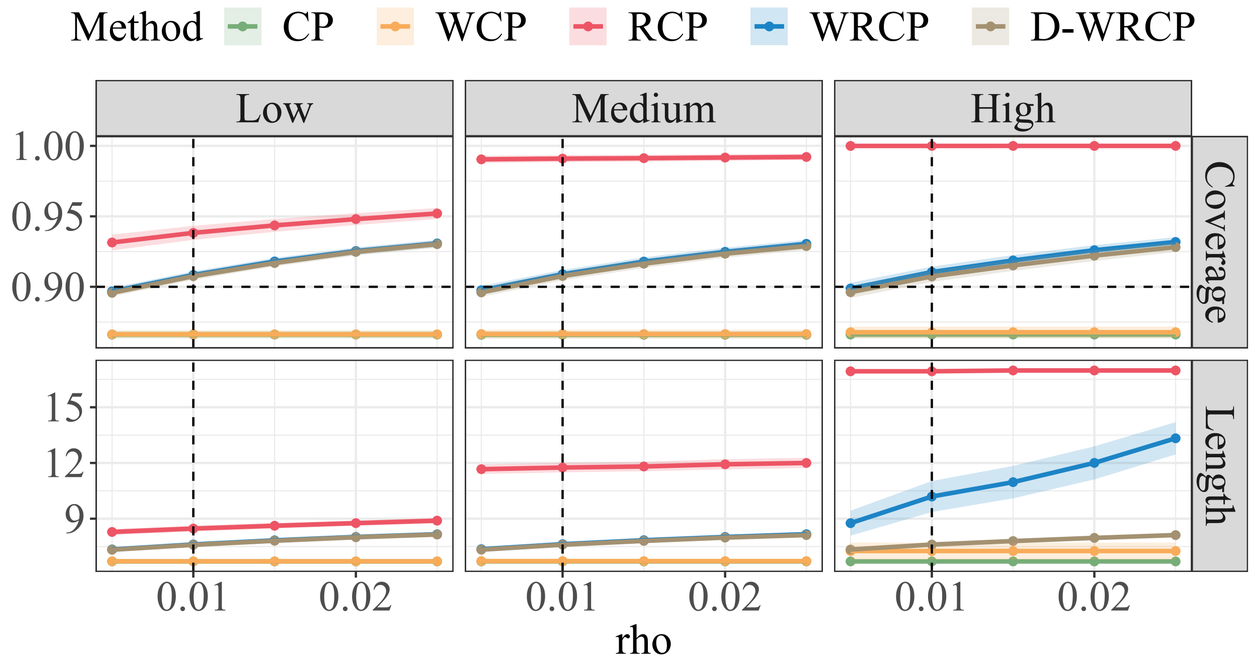
Theorem.[16] Assume the training data $(X_i, Y_i)_{i=1}^n$ are i.i.d. sampled from P and $(X_{n+1}, Y_{n+1}) \sim Q$ is independent of the training data. Assume that Q is absolutely continuous with respect to P and $w(x) \propto dQ_X/dP_X(x)$. For any $\alpha \in (0,1)$ and common choices of *f*, there is

$$\mathbb{P}_{(X_{n+1},Y_{n+1})\sim Q}\Big(Y_{n+1} \in \mathscr{C}(X_{n+1})\Big) \geq 1-\alpha$$

16. Ai and R. 2024, *Not all distributional shifts are equal: Fine-grained robust conformal inference.*

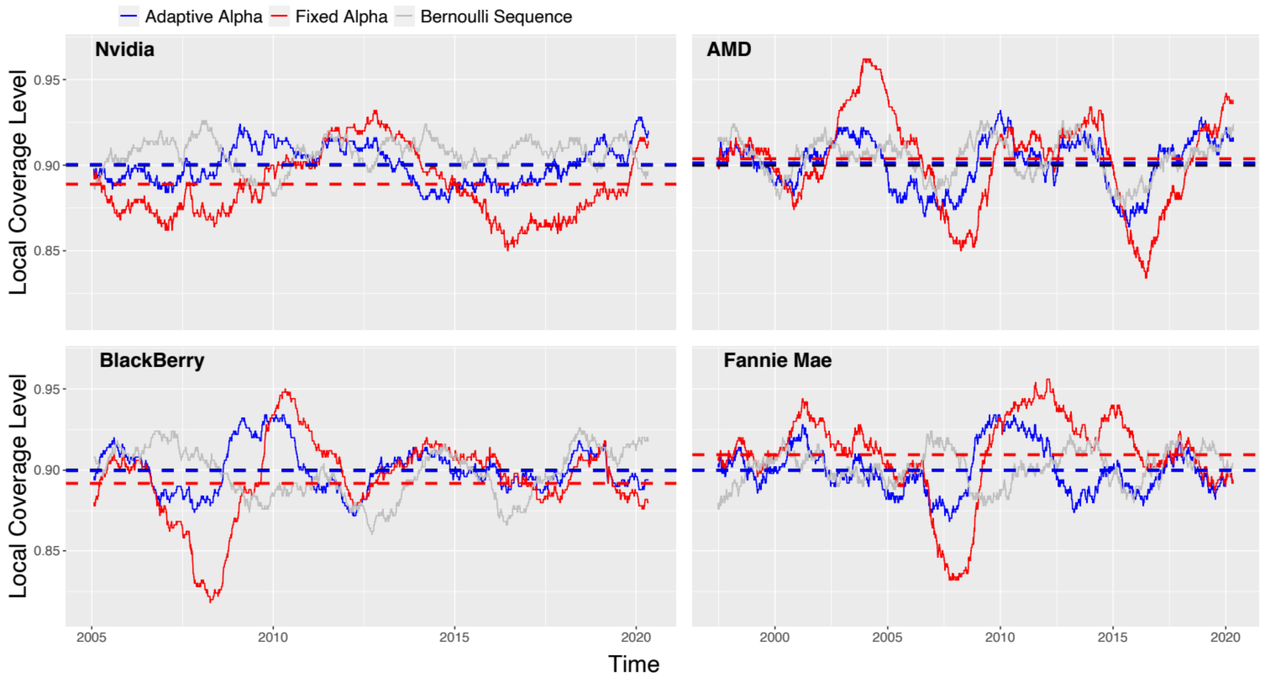
Empirically…

- **Dimension:** $X \in \mathbb{R}^{50}, Y \in \mathbb{R}$
- **Ground-truth:** $\rho^* = D_{\mathrm{KL}}(Q_{Y|X} \,\|\, P_{Y|X}) = 0.01$
- # training data = # test data = 2000
- **Target confidence level:** 90%



Non-exchangeable data

- So far … the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid from P (not necessarily $= Q$)
- Model-fitting & calibration treats the points in an equal way
- What if … asymmetry among training points / drift over time / …



[Figure from Gibbs and Candès 2021]

Non-exchangeable data

Nonexchangeable (split) CP (NexCP)[17]

- Split the data into training and calibration fold
- Fit a model $\hat{f}$ on the training fold obtain the conformity scores on the calibration fold
- Weight the scores $S_{n/2+1}, \ldots, S_n$ with weights $w_{n/2+1}, \ldots, w_n$
  - $w_i \in [0,1]$ & weights need to be fixed a-priori / independent of the data

$$\mathscr{C}^{\text{nexCP}}(X_{n+1}) = \left\{ y : s(X_{n+1}, y) \leq \text{Quantile}\left(1-\alpha, \sum_i \frac{w_i}{\sum_{i'} w_{i'} + 1} \cdot \delta_{S_i} + \frac{1}{\sum_{i'} w_i + 1} \cdot \delta_{+\infty}\right)\right\}$$

17. Barber, Candès, Ramdas, Tibshirani 2023, *Conformal Prediction Beyond Exchangeability.*

Non-exchangeable data

Theorem.[17] The nexCP prediction set satisfies

$$\mathbb{P}\left(Y_{n+1} \in \mathscr{C}^{\text{nexCP}}(X_{n+1})\right) \geq 1 - \alpha - \sum_i w_i \cdot \mathrm{d_{TV}}\left(S, S^{\text{swap }(i)}\right)$$

All conformity scores computed on the original data

All conformity scores computed on the data after swapping i and $n+1$

- If the data is indeed exchangeable $\rightsquigarrow$ achieves $1 - \alpha$ coverage
- Choose small weights to offset large exchangeability violation (requires prior knowledge)
- Can be combined w/ asymmetric algorithm

Non-exchangeable data

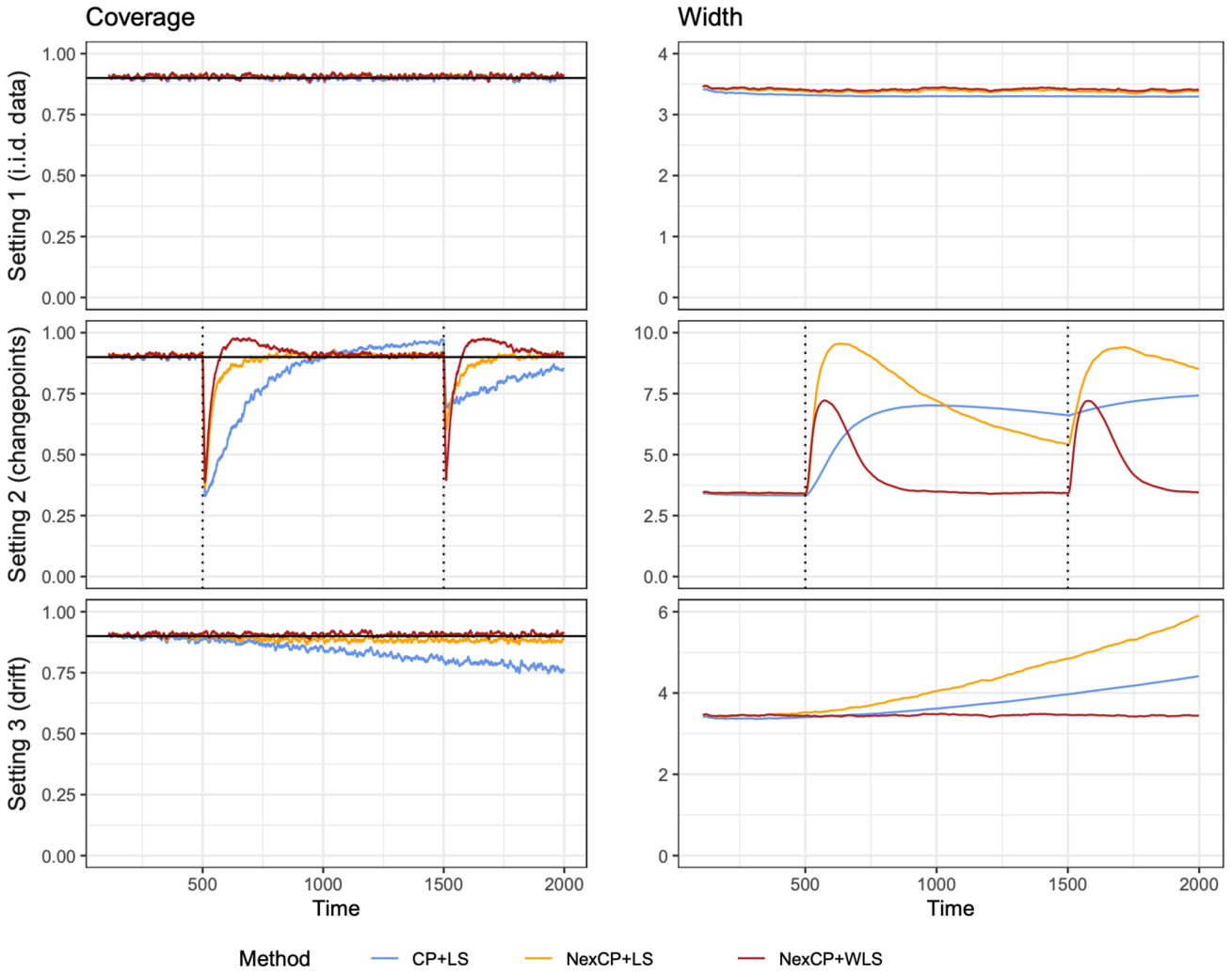


Figure from Barber et al. 2023

Summary

Conformal prediction under (structured) distribution shift

- Covariate shift: weight the data points
- Distributional shift bounded in f-divergence: finding the worst-case shift
- Mixture shift: weighting + worst-case bound
- General drift: weighting w/ fixed budget

Figure from Barber et al. 2023

Related works

- Label shift
  - Podkopaev and Ramdas, 2021, *Distribution-free uncertainty quantification for classification under label shift*
- Charactering distribution shift via Wasserstein distance
- Time series data
  - Xu and Xie 2023, *Conformal prediction for time series*
  - Xu and Xie 2023, *Sequential predictive conformal inference for time series*
- Online predictive inference
  - Gibbs and Candès 2021, *Adaptive conformal inference under distribution shift*
  - Gibbs and Candès 2022, *Conformal inference for online prediction with arbitrary distribution shifts*
- …

Part III: Beyond marginal coverage guarantees

Variants of coverage guarantees

- Recall that A prediction interval $\mathscr{C}(X_{n+1}) \subseteq \mathscr{Y}$ satisfies the marginal coverage guarantee if

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

- Randomness taken over all the training data & (X_{n+1}, Y_{n+1})

  - What conclusions can be drawn given a specific realization of training data?

  - What conclusion can be drawn given a specific realization of test point

Training-conditional coverage guarantee

- Consider the split CP framework & assume the prediction model is given
- Write the calibration data as $\mathscr{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

$$\mathbb{P}(Y_{n+1} \in \mathscr{C}(X_{n+1}) \mid \mathscr{D}_n) \longrightarrow \text{a random quantity}$$

"Future coverage given the current calibration data and model"

- If $\mathbb{P}(Y_{n+1} \in \mathscr{C}(X_{n+1}) \mid \mathscr{D}_n) \geq 1 - \alpha$ almost surely, then marginal coverage holds
- $\mathbb{P}(Y_{n+1} \in \mathscr{C}(X_{n+1})) \geq 1 - \alpha \not\Rightarrow \mathbb{P}(Y_{n+1} \in \mathscr{C}(X_{n+1}) \mid \mathscr{D}_n) \geq 1 - \alpha$

Training-conditional coverage guarantee

- To make things simpler, assume the conformity score is continuous and $(1-\alpha)(n+1)$ is an integer[18]

$$\mathbb{P}\Big(Y_{n+1} \in \mathscr{C}(X_{n+1}) \mid \mathscr{D}_n\Big) \sim \text{beta}\Big((1-\alpha)(n+1), \alpha(n+1)\Big)$$

$$\Rightarrow \mathbb{P}\left(\mathbb{P}\Big(Y_{n+1} \in \mathscr{C}(X_{n+1}) \mid \mathscr{D}_n\Big) \leq 1-\alpha-\Delta\right) = F_{(1-\alpha)(n+1),\alpha(n+1)}(1-\alpha-\Delta) \leq e^{-2n\Delta}$$

For any $\delta \in (0,1)$, take $\Delta = \log(1/\delta)/(2n)$; then with probability at least $1-\delta$,

$$\mathbb{P}(Y_{n+1} \in \mathscr{C}(X_{n+1}) \mid \mathscr{D}_n) > 1-\alpha-\frac{\log(1/\delta)}{2n}$$

18. Elder 2016, *Bayesian adaptive data analysis guarantees from subgaussianity*.

Training-conditional coverage guarantee

Related notion: Probably approximately correct (PAC)

$$\mathbb{P}\Big(\mathbb{P}(Y_{n+1} \notin \mathscr{C}(X_{n+1}) \mid \mathscr{D}_n\Big) \leq \alpha) \geq 1 - \delta$$

- Adjust the quantile in constructing $\mathscr{C}$ to achieve PAC coverage

18. Elder 2016, *Bayesian adaptive data analysis guarantees from subgaussianity*.

Test-conditional coverage guarantee

- Test-conditional coverage: for any distribution P on $\mathcal{X} \times \mathcal{Y}$,

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1}\right) \geq 1 - \alpha, \text{ holds almost surely.}$$

Hardness depends on the richness of $\mathcal{X}$

- If X is discrete, i.e., $\mathcal{X} = \{x_1, \ldots, x_k\}$
- Possible $\rightsquigarrow$ stratify on the groups

- Let $\hat{q}_k$ be $\lceil (n_k + 1)(1 - \alpha) \rceil$ smallest element in $\{S_i : X_i = k\}$
- Return the prediction interval

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

Test-conditional coverage guarantee

- Test-conditional coverage: for any distribution P on $\mathcal{X} \times \mathcal{Y}$,

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1}\right) \geq 1 - \alpha, \text{ holds almost surely.}$$

Hardness depends on the richness of $\mathcal{X}$

- If X is continuous
- Impossible $\rightsquigarrow$ in what sense?

$$\mathcal{C}(x) = \begin{cases} \mathcal{Y} & \text{w.p. } 1 - \alpha, \\ \varnothing & \text{w.p. } \alpha\,. \end{cases}$$

Always valid but non-informative

Impossibility of test-conditional coverage guarantee

Theorem.[19] Suppose $\mathscr{C}$ a prediction set constructed with calibration points and for any P over $\mathscr{X} \times \mathscr{Y}$,

$$\mathbb{P}(Y \in \mathscr{C}(X_{n+1}) \mid X_{n+1}) \geq 1 - \alpha, \text{ almost surely.}$$

Suppose $\mathscr{Y} = \mathbb{R}$. For any distribution P on $\mathscr{X} \times \mathscr{Y}$, where P_X is non-atomic, there is

$$\mathbb{P}(\text{Leb}(\mathscr{C}(x)) = \infty) \geq 1 - \alpha,$$

for any $x \in \mathscr{X}$.

19. Barber, Candès, Ramdas, Tibshirani, 2021, The limits of distribution-free conditional predictive inference.

Selection-conditional coverage guarantees

- Calibration data $\mathscr{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$
- Test samples $\mathscr{D}_{\text{test}} = \{(X_{n+j}, Y_{n+j})\}_{j=1}^{m}$ with unknown $\{Y_{n+j}\}_{j=1}^{m}$ $(m \geq 1)$
- A selection rule $\mathscr{S} : \mathscr{D}_{\text{cal}} \cup \mathscr{D}_{\text{test}} \to 2^{\{n+1,\ldots,n+m\}}$ that picks the focal units $\hat{S} \subseteq \{n+1,\ldots,n+m\}$

It is often useful to ensure

$$\mathbb{P}\left(Y_{n+1} \in \hat{C}(X_{n+1}) \mid n+1 \in \hat{S}\right) \geq 95\,\%$$

Solution: calibrating with an exchangeable subset

JOMI (JOint Mondrian conformal Inference)[20]

- Nonconformity score $V\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, e.g., $V(x, y) = |y - \hat{\mu}(x)|$
- Compute calibration scores $V_i = V(X_i, Y_i)$, $i = 1, \ldots, n$
- For each hypothesized value $y \in \mathcal{Y}$, find

$$\hat{C}^{\alpha}_{n+1} = \left\{ y\colon s(X_{n+1}, y) \leq \text{Quantile}(0.95;\ \{S_i\}_{i \in \mathcal{R}_{n+1}(y)} \cup \{s(X_{n+1}, y)\} \right\}$$

where $\mathcal{R}_{n+j}(y) \subseteq \{1, \ldots, n\}$ is a reference set determined by our method.

Idea for constructing $\mathcal{R}_{n+j}(y)$

Find calibration points which, when posited as a test point, lead to the same selection event

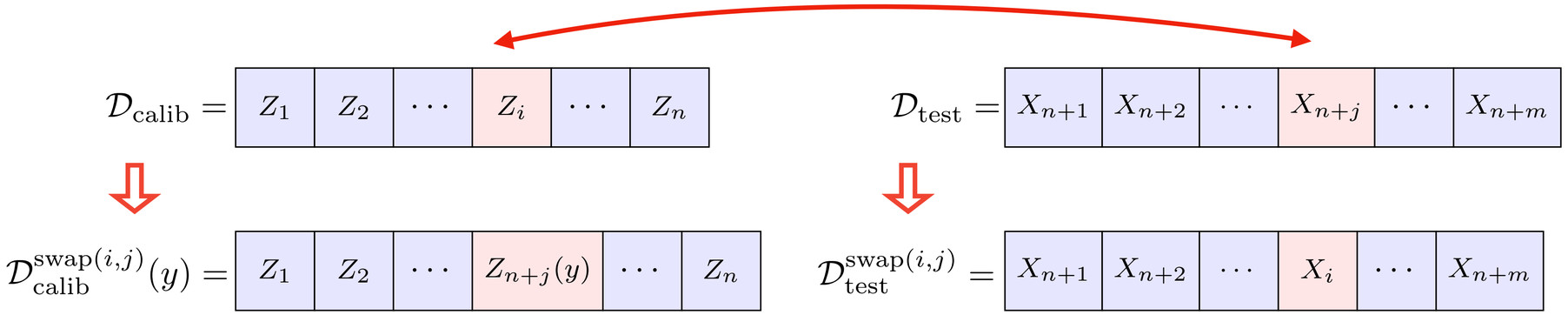
---

20. Jin and R, 2025. *Confidence on the focal: Conformal prediction with selection-conditional coverage.*

Finding the exchangeable subset

Full data $Z = (X, Y)$, and $Z_{n+j}(y) = (X_{n+j}, y)$ for any hypothesized $y \in \mathcal{Y}$.

Given a test point j, for each $i \in \{1, \ldots, n\}$, define "swapped" datasets:



Define reference set

$$\hat{\mathcal{R}}_{n+j}(y) = \left\{ i \in [n] : j \in \hat{\mathcal{S}}^{\text{swap}(i,j)}(y) \right\}$$

Selection set using "swapped" datasets

Can be computed for a number of selection algorithms

Other types of coverage guarantees

- Label-conditional coverage: suppose $\mathcal{Y} = \{1,2,\ldots,K\}$;

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

  - Vovk et al 2003, *Mondrian confidence machine.*
  - Löfström et al 2015, *Bias reduction through conditional conformal prediction.*
  - Ding et al 2024, *Class-conditional conformal prediction with many classes.*

- Relaxed test-conditional coverage:

  - Isaac, Cherian, Candès, 2025, *Conformal prediction with conditional guarantees.*
  - Hore and Barber, 2025, *Conformal prediction with local weights: randomization enables local guarantees.*

Summary

- The scope and aim of distribution-free predictive inference
- Introduction to split, full CP, CV+/Jackknife+
- Considerations in CP & choices of conformity scores
- CP under distributional shifts
- Relaxing the marginal coverage guarantees

Thank you!

zren@wharton.upenn.edu | https://zhimeir.github.io/