NB: Open with an advanced pdf reader (e.g., Acrobat to have animations)

Conformal prediction - A tutorial

Aymeric DIEULEVEUT
Professor, École Polytechnique

July 8, 2025
Hi! PARIS Summer School

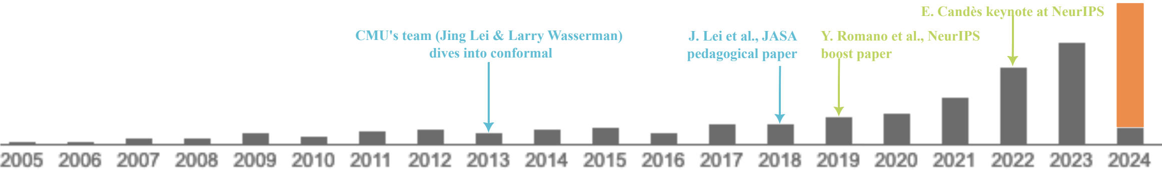
Why are we all here today?

(Slides available on my webpage)

Why are we all here today?

(Slides available on my webpage)

- Because Conformal Prediction has been a **popular** topic recently.



Vovk et al. (2005) algorithmic learning in a random world cite count.

Why are we all here today?

(Slides available on my webpage)

- Because Conformal Prediction has been a **popular** topic recently.

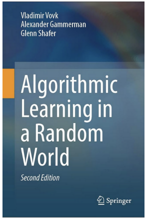
- Because we believe that conformal methods are **important** tools, whose strengths and limitations are sometimes misunderstood.

  Successfully applied to
  - Medical applications
  - Markets / demand forecasting
  - Computer Vision

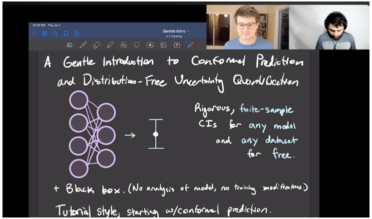
Why are we all here today?

(Slides available on my webpage)

- Because Conformal Prediction has been a **popular** topic recently.
- Because we believe that conformal methods are **important** tools, whose strengths and limitations are sometimes misunderstood.
- To be part of the **diffusion** effort that many colleagues are making.



Book reference: Vovk et al. (2005)
(new edition in 2022)



A gentle tutorial: Angelopoulos and Bates (2023)
+ Videos playlist

R. J. Tibshirani
introductive lecture's notes

Why are we all here today?

(Slides available on my webpage)

- Because Conformal Prediction has been a **popular** topic recently.
- Because we believe that conformal methods are **important** tools, whose strengths and limitations are sometimes misunderstood.
- To be part of the **diffusion** effort that many colleagues are making.

Margaux Zaffran
UC Berkeley
PhD at École Polytechnique,
Polytechnique Institute of Paris,
Inria, and EDF

→ Based on our tutorial with Margaux Zaffran at UAI and ICML, `slides here`
→ Builds upon earlier material accessible on `this webpage`

Goals and disclaimers

Goals

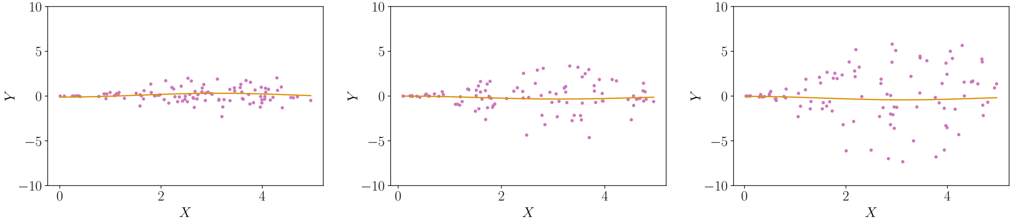
- Provide a detailed introduction to the basics
- Demystify the results: fair introduction with limits
- Give you insights on how to leverage those techniques in your own fields

On the importance of quantifying uncertainty

- **Obvious in most applications - weather, medical, markets**

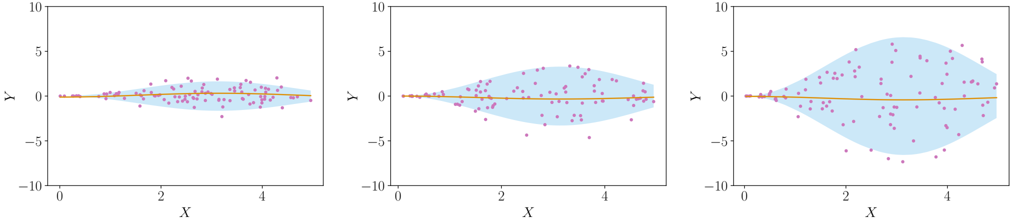
On the importance of quantifying uncertainty

- **Obvious in most applications - weather, medical, markets**
- **Mathematically**



↪ Same "best" predictor, yet 3 distinct underlying phenomena!

- **Obvious in most applications - weather, medical, markets**
- **Mathematically**



$\hookrightarrow$ Same "best" predictor, yet 3 distinct underlying phenomena!

$\Longrightarrow$ Quantifying uncertainty conveys this information.

Quantifying predictive uncertainty

- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables
- n training samples $(X_i, Y_i)_{i=1}^n$
- Goal: predict an unseen point Y_{n+1} at X_{n+1} with **confidence**
- How? Given a miscoverage level $\alpha \in [0, 1]$, build a predictive set $\mathcal{C}_\alpha$ such that:

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

and $\mathcal{C}_\alpha$ should be as small as possible, in order to be informative

For example: $\alpha = 0.1$ and obtain a 90% coverage interval

Quantifying predictive uncertainty

- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables
- n training samples $(X_i, Y_i)_{i=1}^n$
- Goal: predict an unseen point Y_{n+1} at X_{n+1} with **confidence**
- How? Given a miscoverage level $\alpha \in [0, 1]$, build a predictive set $\mathcal{C}_\alpha$ such that:

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

  and $\mathcal{C}_\alpha$ should be as small as possible, in order to be informative

  For example: $\alpha = 0.1$ and obtain a 90% coverage interval
- Construction of the predictive intervals should be
  - agnostic to the model
  - agnostic to the data distribution
- *Validity* should be ensured
  - in finite samples
  - for all data distribution and underlying model

Our afternoon!

Intro I: Split Conformal Prediction (SCP) - the simplest CP method

Intro II: Overview of some challenges in Conformal Prediction

Advanced I: Towards conditional coverage

Advanced II: Avoiding data splitting: full conformal and out-of-bags approaches

Advanced III: Beyond exchangeability

Applications & Methods I: Some case studies

Applications & Methods II: Some methodological advances

Concluding remarks

Intro I: Split Conformal Prediction (SCP) - the simplest CP method

- Standard regression case
- Conformalized Quantile Regression (CQR)
- SCP - Multi class Classification
- On the design choices of conformity scores and (empirical) conditional guarantees

Split Conformal Prediction (SCP) (Vovk et al., 2005)

without anim

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)

SCP: implementation details

without anim

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, get prediction values with $\hat{\mu}$

SCP: implementation details

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm $\mathcal{A}$ on the proper training set*
3. On the calibration set, get prediction values with $\hat{\mu}$
4. Obtain a set of $\#\mathrm{Cal} + 1$ conformity scores :

$$\mathcal{S} = \{S_i = |\hat{\mu}(X_i) - Y_i|, i \in \mathrm{Cal}\} \cup \{+\infty\}$$

(+ worst-case scenario)

SCP: implementation details

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm* $\mathcal{A}$ *on the proper training set*
3. On the calibration set, get prediction values with $\hat{\mu}$
4. Obtain a set of $\#\mathrm{Cal} + 1$ conformity scores :

$$\mathcal{S} = \{S_i = |\hat{\mu}(X_i) - Y_i|, i \in \mathrm{Cal}\} \cup \{+\infty\}$$

(+ worst-case scenario)

5. Compute the $1 - \alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$[1]

[1] Equivalently, let $\mathcal{S}$ be the set of $\#\mathrm{Cal}$ conformity scores (i.e. without adding $\{+\infty\}$). Compute the $(1-\alpha)(1/\#\mathrm{Cal} + 1)$ quantile of these scores $\mathcal{S}$.

SCP: implementation details

1. Randomly split the training data into a proper training set (size $\#\mathrm{Tr}$) and a calibration set (size $\#\mathrm{Cal}$)
2. Get $\hat{\mu}$ *by training the algorithm $\mathcal{A}$ on the proper training set*
3. On the calibration set, get prediction values with $\hat{\mu}$
4. Obtain a set of $\#\mathrm{Cal}+1$ conformity scores :

$$\mathcal{S} = \{S_i = |\hat{\mu}(X_i) - Y_i|, i \in \mathrm{Cal}\} \cup \{+\infty\}$$

(+ worst-case scenario)

5. Compute the $1-\alpha$ quantile of these scores, noted $q_{1-\alpha}(\mathcal{S})$[1]
6. For a new point X_{n+1}, return

$$\widehat{C}_\alpha(X_{n+1}) = [\hat{\mu}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}); \hat{\mu}(X_{n+1}) + q_{1-\alpha}(\mathcal{S})]$$

[1]Equivalently, let $\mathcal{S}$ be the set of $\#\mathrm{Cal}$ conformity scores (i.e. without adding $\{+\infty\}$). Compute the $(1-\alpha)(1/\#\mathrm{Cal}+1)$ quantile of these scores $\mathcal{S}$.

SCP: theoretical foundation

Exchangeability

$(X_i, Y_i)_{i=1}^n$ are exchangeable if, for any permutation σ of $[\![1, n]\!]$:

$$((X_1, Y_1), \dots, (X_n, Y_n)) \stackrel{d}{=} \left(\left(X_{\sigma(1)}, Y_{\sigma(1)}\right), \dots, \left(X_{\sigma(n)}, Y_{\sigma(n)}\right)\right) .$$

Exchangeability

$(X_i, Y_i)_{i=1}^n$ are exchangeable if, for any permutation σ of $[\![1, n]\!]$:

$$((X_1, Y_1), \ldots, (X_n, Y_n)) \stackrel{d}{=} \left(\left(X_{\sigma(1)}, Y_{\sigma(1)}\right), \ldots, \left(X_{\sigma(n)}, Y_{\sigma(n)}\right)\right).$$

Toy case: Z_1 and Z_2 are exchangeable if $(Z_1, Z_2) \stackrel{d}{=} (Z_2, Z_1)$.

Exchangeability

$(X_i, Y_i)_{i=1}^n$ are exchangeable if, for any permutation σ of $[\![1, n]\!]$:

$$((X_1, Y_1), \ldots, (X_n, Y_n)) \stackrel{d}{=} \left((X_{\sigma(1)}, Y_{\sigma(1)}), \ldots, (X_{\sigma(n)}, Y_{\sigma(n)})\right).$$

Toy case: Z_1 and Z_2 are exchangeable if $(Z_1, Z_2) \stackrel{d}{=} (Z_2, Z_1)$.

exchangeable sequences

- i.i.d. samples

Exchangeability

$(X_i, Y_i)_{i=1}^n$ are exchangeable if, for any permutation σ of $[\![1, n]\!]$:

$$((X_1, Y_1), \dots, (X_n, Y_n)) \overset{d}{=} \left((X_{\sigma(1)}, Y_{\sigma(1)}), \dots, (X_{\sigma(n)}, Y_{\sigma(n)})\right).$$

Toy case: Z_1 and Z_2 are exchangeable if $(Z_1, Z_2) \overset{d}{=} (Z_2, Z_1)$.

exchangeable sequences

- i.i.d. samples
- The components of $\mathcal{N}\left(\begin{pmatrix} m \\ \vdots \\ \vdots \\ m \end{pmatrix}, \begin{pmatrix} \sigma^2 & & & \\ & \ddots & \gamma^2 & \\ & \gamma^2 & \ddots & \\ & & & \sigma^2 \end{pmatrix}\right)$

SCP: theoretical guarantees

SCP enjoys finite sample guarantees proved in Vovk et al. (2005); Lei et al. (2018).

Marginal validity

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[a]. SCP applied on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

SCP enjoys finite sample guarantees proved in Vovk et al. (2005); Lei et al. (2018).

Marginal validity

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[a]. SCP applied on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

Additionally, if the scores $\{S_i\}_{i\in\mathrm{Cal}} \cup \{S_{n+1}\}$ are a.s. distinct:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\right\} \leq 1-\alpha+\frac{1}{\#\mathrm{Cal}+1}.$$

[a]Only the calibration and test data need to be exchangeable.

Quantile lemma

If $(U_1, \dots, U_n, U_{n+1})$ are exchangeable, then for any $\beta \in]0, 1[$:

$$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \dots, U_n, +\infty)\right) \geq \beta.$$

Additionally, if $U_1, \dots, U_n, U_{n+1}$ are almost surely distinct, then:

$$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \dots, U_n, +\infty)\right) \leq \beta + \frac{1}{n+1}.$$

Quantile lemma

If $(U_1, \ldots, U_n, U_{n+1})$ are exchangeable, then for any $\beta \in]0, 1[$:

$$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, +\infty)\right) \geq \beta.$$

Additionally, if $U_1, \ldots, U_n, U_{n+1}$ are almost surely distinct, then:

$$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, +\infty)\right) \leq \beta + \frac{1}{n+1}.$$

$(X_i, Y_i)_{i=1}^{n+1}$ exchangeable $\Rightarrow \{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ exchangeable.

Quantile lemma

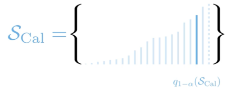
If $(U_1, \dots, U_n, U_{n+1})$ are exchangeable, then for any $\beta \in]0,1[$:

$$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \dots, U_n, +\infty)\right) \geq \beta.$$

Additionally, if $U_1, \dots, U_n, U_{n+1}$ are almost surely distinct, then:

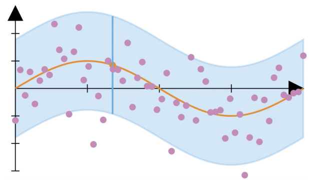
$$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \dots, U_n, +\infty)\right) \leq \beta + \frac{1}{n+1}.$$

$(X_i, Y_i)_{i=1}^{n+1}$ exchangeable $\Rightarrow \{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ exchangeable.



$\hookrightarrow$ quantile lemma to the scores gives the result.

$$\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\right\} = \left\{Y_{n+1} \in [\hat{\mu}(X_{n+1}) \pm q_{1-\alpha}(\mathcal{S})]\right\}$$
$$= \left\{|Y_{n+1} - \hat{\mu}(X_{n+1})| \leq q_{1-\alpha}(\mathcal{S})\right\}$$
$$\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\right\} = \left\{S_{n+1} \leq q_{1-\alpha}(\mathcal{S})\right\}.$$



Proof of the quantile lemma

Def: Empirical quantile $q_\beta(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

Proof of the quantile lemma

Def: Empirical quantile $q_\beta(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

First note that $U_{n+1} \le q_\beta(U_1, \ldots, U_n, +\infty) \iff U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})$.

Def: Empirical quantile $q_\beta(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

First note that $U_{n+1} \leq q_\beta(U_1, \ldots, U_n, +\infty) \iff U_{n+1} \leq q_\beta(U_1, \ldots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right) = \mathbb{P}\left(U_i \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right)$.

Def: Empirical quantile $q_\beta(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

First note that $U_{n+1} \le q_\beta(U_1, \ldots, U_n, +\infty) \Longleftrightarrow U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right) = \mathbb{P}\left(U_i \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right)$. Thus:

$$\mathbb{P}\left(U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(U_i \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right)$$

Proof of the quantile lemma

Def: Empirical quantile $q_\beta(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

First note that $U_{n+1} \le q_\beta(U_1, \ldots, U_n, +\infty) \iff U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right) = \mathbb{P}\left(U_i \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right)$. Thus:

$$\begin{aligned}
\mathbb{P}\left(U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(U_i \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right) \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}\left\{U_i \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right\}\right]
\end{aligned}$$

Proof of the quantile lemma

Def: Empirical quantile $q_{\beta}(U_1, \dots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \dots, U_k)$

First note that $U_{n+1} \leq q_{\beta}(U_1, \dots, U_n, +\infty) \iff U_{n+1} \leq q_{\beta}(U_1, \dots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \leq q_{\beta}(U_1, \dots, U_n, U_{n+1})\right) = \mathbb{P}\left(U_i \leq q_{\beta}(U_1, \dots, U_n, U_{n+1})\right)$. Thus:

$$
\begin{aligned}
\mathbb{P}\left(U_{n+1} \leq q_{\beta}(U_1, \dots, U_n, U_{n+1})\right) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(U_i \leq q_{\beta}(U_1, \dots, U_n, U_{n+1})\right) \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}\left\{U_i \leq q_{\beta}(U_1, \dots, U_n, U_{n+1})\right\}\right] \\
&\geq \frac{1}{n+1} \mathbb{E}\left[\lceil \beta(n+1) \rceil\right] \\
&= \frac{\lceil \beta(n+1) \rceil}{n+1} \geq \beta,
\end{aligned}
$$

proving the first statement.

Proof of the quantile lemma

Def: Empirical quantile $q_\beta(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

First note that $U_{n+1} \leq q_\beta(U_1, \ldots, U_n, +\infty) \iff U_{n+1} \leq q_\beta(U_1, \ldots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right) = \mathbb{P}\left(U_i \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right)$. Thus:

$$
\begin{aligned}
\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(U_i \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right) \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}\left\{U_i \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right\}\right] \\
&= \frac{1}{n+1} \mathbb{E}\left[\lceil \beta(n+1) \rceil\right] \quad \text{if all } (U_i) \text{ are distinct} \\
&= \frac{\lceil \beta(n+1) \rceil}{n+1} \leq \beta + \frac{1}{n+1},
\end{aligned}
$$

proving the second statement.

SCP enjoys finite sample guarantees proved in Vovk et al. (2005); Lei et al. (2018).

Marginal validity Vovk et al. (2005)

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[d]. SCP applied on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

Additionally, if the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ are a.s. distinct:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\right\} \leq 1 - \alpha + \frac{1}{\#\mathrm{Cal} + 1}.$$

[d] Only the calibration and test data need to be exchangeable.

SCP enjoys finite sample guarantees proved in Vovk et al. (2005); Lei et al. (2018).

Marginal validity Vovk et al. (2005)

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[d]. SCP applied on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

Additionally, if the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ are a.s. distinct:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\right\} \leq 1 - \alpha + \frac{1}{\#\mathrm{Cal} + 1}.$$

[d] Only the calibration and test data need to be exchangeable.

✓ Distribution free, model (regressor) free, finite sample average validity guarantee.

Standard mean-regression SCP – strength: validity – good vs bad estimator

SCP: theoretical guarantees

SCP enjoys finite sample guarantees proved in Vovk et al. (2005); Lei et al. (2018).

Marginal validity Vovk et al. (2005)

Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable[d]. SCP applied on $(X_i, Y_i)_{i=1}^{n}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

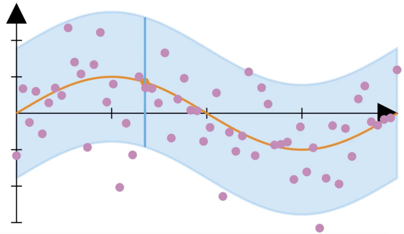
Additionally, if the scores $\{S_i\}_{i \in \mathrm{Cal}} \cup \{S_{n+1}\}$ are a.s. distinct:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\right\} \leq 1 - \alpha + \frac{1}{\#\mathrm{Cal} + 1}.$$

[d] Only the calibration and test data need to be exchangeable.

✗ Marginal coverage: $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \cancel{| X_{n+1} = x}\right\} \geq 1 - \alpha$

Standard mean-regression SCP – weakness: not adaptive

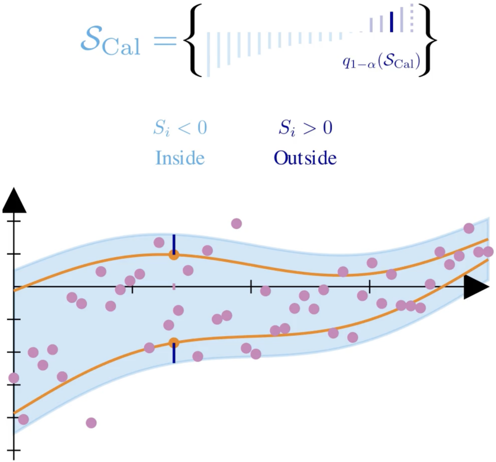


- Predict with $\hat{\mu}$
- Build $\widehat{C}_\alpha(x)$: $[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$

Intro I: Split Conformal Prediction (SCP) - the simplest CP method

- Standard regression case
- Conformalized Quantile Regression (CQR)
- SCP - Multi class Classification
- On the design choices of conformity scores and (empirical) conditional guarantees

Conformalized Quantile Regression (CQR) (Romano et al., 2019)



$$\widehat{C}_\alpha(x) = [\widehat{\mathrm{QR}}_{\mathrm{lower}}(x) - q_{1-\alpha}(\mathcal{S});$$
$$\widehat{\mathrm{QR}}_{\mathrm{upper}}(x) + q_{1-\alpha}(\mathcal{S})]$$

Thus

$$\left\{Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right)\right\} = \left\{S_{n+1} \leq q_{1-\alpha}\left(\mathcal{S}\right)\right\}.$$

$\rightsquigarrow$ Marginal validity is ensured, independently of the underlying quantile level or regressor quality. ✓

CQR: under vs over coverage

Intro I: Split Conformal Prediction (SCP) - the simplest CP method

- Standard regression case
- Conformalized Quantile Regression (CQR)
- SCP - Multi class Classification
- On the design choices of conformity scores and (empirical) conditional guarantees

SCP: standard classification

- $Y \in \{1, \ldots, C\}$ (C classes)
- $\hat{A}(X) = (\hat{p}_1(X), \ldots, \hat{p}_C(X))$ (estimated probabilities)

SCP: standard classification

- $Y \in \{1, \ldots, C\}$ (C classes)
- $\hat{A}(X) = (\hat{p}_1(X), \ldots, \hat{p}_C(X))$ (estimated probabilities)
- Score ?

SCP: standard classification

- $Y \in \{1, \ldots, C\}$ (C classes)
- $\hat{A}(X) = (\hat{p}_1(X), \ldots, \hat{p}_C(X))$ (estimated probabilities)
- Score

$$\mathrm{s}(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$$

SCP: standard classification

- $Y \in \{1, \dots, C\}$ (C classes)
- $\hat{A}(X) = (\hat{p}_1(X), \dots, \hat{p}_C(X))$ (estimated probabilities)
- Score

$$\mathrm{s}(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$$

- For a new point X_{n+1}, return

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \mathrm{s}(\hat{A}(X_{n+1}), y) \le q_{1-\alpha}(\mathcal{S})\}$$

i.e.,

$$\left\{Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})\right\} = \{S_{n+1} \le q_{1-\alpha}(\mathcal{S})\}.$$

SCP: standard classification in practice

Ex: $Y_i \in \{\text{“dog”}, \text{“tiger”}, \text{“cat”}\}$, with $\alpha = 0.1$

SCP: standard classification in practice

Ex: $Y_i \in \{$ "dog", "tiger", "cat" $\}$, with $\alpha = 0.1$

Pred on $\mathcal{C}$al	$i = 1$	2	3							$\mathcal{C}$al
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.40	0.35

SCP: standard classification in practice

Ex: $Y_i \in \{$ "dog" , "tiger" , "cat" $\}$, with $\alpha = 0.1$

Pred on Cal	$i=1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.40	0.35
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

Pred on Cal	$i = 1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.40	0.35
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"

SCP: standard classification in practice

Ex: $Y_i \in \{$ "dog" , "tiger" , "cat" $\}$, with $\alpha = 0.1$

Pred on Cal	$i = 1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.40	0.35
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"

- Scores on the calibration set $s(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

SCP: standard classification in practice

Ex: $Y_i \in \{$ "dog" , "tiger" , "cat" $\}$, with $\alpha = 0.1$

Pred on $\mathcal{C}al$	$i=1$	2	3							$\mathcal{C}al$
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.40	0.35
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
S_i	0.05	0.10	0.15	0.40	0.45	0.40	0.35	0.25	0.60	0.65

- Scores on the calibration set $s(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

Pred on Cal	$i=1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.40	0.35
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
S_i	0.05	0.10	0.15	0.40	0.45	0.40	0.35	0.25	0.60	0.65

- Scores on the calibration set $s(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

$\Rightarrow q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}}) = 0.65$

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

Pred on Cal	$i=1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.40	0.35
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
S_i	0.05	0.10	0.15	0.40	0.45	0.40	0.35	0.25	0.60	0.65

- Scores on the calibration set $s(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

$\Rightarrow q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}}) = 0.65$

Pred. on Test	$n+1$
$\hat{p}_{\text{dog}}(X_{n+1})$	0.03
$\hat{p}_{\text{tiger}}(X_{n+1})$	0.37
$\hat{p}_{\text{cat}}(X_{n+1})$	0.60

$\hat{A}(X_{n+1}) = (0.03, 0.37, 0.60)$

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

Pred on Cal	$i=1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.40	0.35
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
S_i	0.05	0.10	0.15	0.40	0.45	0.40	0.35	0.25	0.60	0.65

- Scores on the calibration set $s(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

$\Rightarrow q_{1-\alpha}(\mathcal{S}_{\text{Cal}}) = 0.65$

Pred. on Test	$n+1$	
$\hat{p}_{\text{dog}}(X_{n+1})$	0.03	$s(\hat{A}(X_{n+1}), \text{"dog"}) = 0.97$
$\hat{p}_{\text{tiger}}(X_{n+1})$	0.37	$s(\hat{A}(X_{n+1}), \text{"tiger"}) = 0.63$
$\hat{p}_{\text{cat}}(X_{n+1})$	0.60	$s(\hat{A}(X_{n+1}), \text{"cat"}) = 0.40$

$\hat{A}(X_{n+1}) = (0.03, 0.37, 0.60)$

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

Pred on Cal	$i=1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.40	0.35
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
S_i	0.05	0.10	0.15	0.40	0.45	0.40	0.35	0.25	0.60	0.65

- Scores on the calibration set $s(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

$\Rightarrow q_{1-\alpha}(\mathcal{S}_{\text{Cal}}) = \boxed{0.65}$

Pred. on Test	$n+1$		
$\hat{p}_{\text{dog}}(X_{n+1})$	0.03	$s(\hat{A}(X_{n+1}), \text{"dog"}) = 0.97$	$> q_{1-\alpha}(\mathcal{S}_{\text{Cal}})$
$\hat{p}_{\text{tiger}}(X_{n+1})$	0.37	$s(\hat{A}(X_{n+1}), \text{"tiger"}) = 0.63$	$\leq q_{1-\alpha}(\mathcal{S})$
$\hat{p}_{\text{cat}}(X_{n+1})$	0.60	$s(\hat{A}(X_{n+1}), \text{"cat"}) = 0.40$	$\leq q_{1-\alpha}(\mathcal{S}_{\text{Cal}})$

$\hat{A}(X_{n+1}) = (0.03, 0.37, 0.60)$

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

Pred on Cal	$i=1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.25	0.20
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.35	0.45
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.40	0.35
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
S_i	0.05	0.10	0.15	0.40	0.45	0.40	0.35	0.25	0.60	0.65

- Scores on the calibration set $\mathrm{s}(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

$\Rightarrow q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}}) = 0.65$

Pred. on Test	$n+1$			
$\hat{p}_{\text{dog}}(X_{n+1})$	0.03	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"dog"}) = 0.97$	$> q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"dog" $\notin \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{tiger}}(X_{n+1})$	0.37	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"tiger"}) = 0.63$	$\leq q_{1-\alpha}(\mathcal{S})$	"tiger" $\in \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{cat}}(X_{n+1})$	0.60	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"cat"}) = 0.40$	$\leq q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"cat" $\in \widehat{C}_\alpha(X_{n+1})$

$\hat{A}(X_{n+1}) = (0.03, 0.37, 0.60)$

$\widehat{C}_\alpha(X_{n+1}) = \{$"tiger", "cat"$\}$

SCP: standard classification in practice

Ex: $Y_i \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

Pred on Cal	$i=1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.05	0.05
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.15	0.05
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.80	0.90
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
S_i	0.05	0.10	0.15	0.40	0.45	0.40	0.35	0.25		-

- Scores on the calibration set $\mathrm{s}(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

$\Rightarrow q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}}) = 0.65$

Pred. on Test	$n+1$			
$\hat{p}_{\text{dog}}(X_{n+1})$	0.03	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"dog"}) = 0.97$	$> q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"dog" $\notin \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{tiger}}(X_{n+1})$	0.37	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"tiger"}) = 0.63$	$\leq q_{1-\alpha}(\mathcal{S})$	"tiger" $\in \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{cat}}(X_{n+1})$	0.60	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"cat"}) = 0.40$	$\leq q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"cat" $\in \widehat{C}_\alpha(X_{n+1})$

$\hat{A}(X_{n+1}) = (0.03, 0.37, 0.60)$

$\widehat{C}_\alpha(X_{n+1}) = \{$"tiger", "cat"$\}$

SCP: standard classification in practice

Ex: $Y_i \in \{$ "dog", "tiger", "cat" $\}$, with $\alpha = 0.1$

Pred on Cal	$i=1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.05	0.05
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.15	0.05
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.80	0.90
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
S_i	0.05	0.10	0.15	0.40	0.45	0.40	0.35	0.25	0.2	0.1

- Scores on the calibration set $\mathrm{s}(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

$\Rightarrow$ $q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}}) = 0.65$

Pred. on Test	$n+1$			
$\hat{p}_{\text{dog}}(X_{n+1})$	0.03	$\mathrm{s}(\hat{A}(X_{n+1}),$ "dog"$) = 0.97$	$> q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"dog" $\notin \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{tiger}}(X_{n+1})$	0.37	$\mathrm{s}(\hat{A}(X_{n+1}),$ "tiger"$) = 0.63$	$\leq q_{1-\alpha}(\mathcal{S})$	"tiger" $\in \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{cat}}(X_{n+1})$	0.60	$\mathrm{s}(\hat{A}(X_{n+1}),$ "cat"$) = 0.40$	$\leq q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"cat" $\in \widehat{C}_\alpha(X_{n+1})$

$\hat{A}(X_{n+1}) = (0.03, 0.37, 0.60)$ $\widehat{C}_\alpha(X_{n+1}) = \{$ "tiger", "cat" $\}$

SCP: standard classification in practice

Ex: $Y_i \in \{$ "dog", "tiger", "cat" $\}$, with $\alpha = 0.1$

Pred on Cal	$i=1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.05	0.05
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.15	0.05
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.80	0.90
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
S_i	0.05	0.10	0.15	0.40	0.45	0.40	0.35	0.25	0.2	0.1

- Scores on the calibration set $\mathrm{s}(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

$\Rightarrow q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}}) = 0.45$

Pred. on Test	$n+1$			
$\hat{p}_{\text{dog}}(X_{n+1})$	0.03	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"dog"}) = 0.97$	$> q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"dog" $\notin \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{tiger}}(X_{n+1})$	0.37	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"tiger"}) = 0.63$	$\leq q_{1-\alpha}(\mathcal{S})$	"tiger" $\in \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{cat}}(X_{n+1})$	0.60	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"cat"}) = 0.40$	$\leq q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"cat" $\in \widehat{C}_\alpha(X_{n+1})$

$\hat{A}(X_{n+1}) = (0.03, 0.37, 0.60)$ $\widehat{C}_\alpha(X_{n+1}) = \{$ "tiger", "cat" $\}$

SCP: standard classification in practice

Ex: $Y_i \in \{$ "dog" , "tiger" , "cat" $\}$, with $\alpha = 0.1$

Pred on Cal	$i=1$	2	3							Cal
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.05	0.05
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.15	0.05
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.80	0.90
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
S_i	0.05	0.10	0.15	0.40	0.45	0.40	0.35	0.25	0.2	0.1

- Scores on the calibration set $\mathrm{s}(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

$\Rightarrow q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}}) = 0.45$

Pred. on Test	$n+1$			
$\hat{p}_{\text{dog}}(X_{n+1})$	0.03	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"dog"}) = 0.97$	$> q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"dog" $\notin \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{tiger}}(X_{n+1})$	0.37	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"tiger"}) = 0.63$	$> q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"tiger" $\notin \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{cat}}(X_{n+1})$	0.60	$\mathrm{s}(\hat{A}(X_{n+1}), \text{"cat"}) = 0.40$	$\leq q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"cat" $\in \widehat{C}_\alpha(X_{n+1})$

$\hat{A}(X_{n+1}) = (0.03, 0.37, 0.60)$

$\widehat{C}_\alpha(X_{n+1}) = \{$ "tiger" , "cat" $\}$

SCP: standard classification in practice

Ex: $Y_i \in \{$ "dog" , "tiger" , "cat" $\}$, with $\alpha = 0.1$

Pred on $\mathcal{C}$al	$i=1$	2	3							$\mathcal{C}$al
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.15	0.15	0.20	0.15	0.15	0.05	0.05
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.60	0.55	0.60	0.65	0.10	0.15	0.05
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.25	0.30	0.20	0.20	0.75	0.80	0.90
Y_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
S_i	0.05	0.10	0.15	0.40	0.45	0.40	0.35	0.25	0.2	0.1

- Scores on the calibration set $s(\hat{A}(X), Y) := 1 - (\hat{A}(X))_Y$

$\Rightarrow \; q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}}) = 0.45$

Pred. on Test	$n+1$			
$\hat{p}_{\text{dog}}(X_{n+1})$	0.03	$s(\hat{A}(X_{n+1}), \text{"dog"}) = 0.97$	$> q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"dog" $\notin \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{tiger}}(X_{n+1})$	0.37	$s(\hat{A}(X_{n+1}), \text{"tiger"}) = 0.63$	$> q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"tiger" $\notin \widehat{C}_\alpha(X_{n+1})$
$\hat{p}_{\text{cat}}(X_{n+1})$	0.60	$s(\hat{A}(X_{n+1}), \text{"cat"}) = 0.40$	$\leq q_{1-\alpha}(\mathcal{S}_{\mathrm{Cal}})$	"cat" $\in \widehat{C}_\alpha(X_{n+1})$

$\hat{A}(X_{n+1}) = (0.03, 0.37, 0.60)$ $\qquad$ $\widehat{C}_\alpha(X_{n+1}) = \{$ "cat" $\}$

efficiency yet non-adaptivity of the simplest classification scores

✓ Outputs the most efficient set possible (i.e. achieving the smallest average set size, Sadinle et al., 2018),

✗ Does not allow to discriminate between "easy" and "hard" test point. In practice, it leads to predictive sets that under-cover (resp. over-cover) on "hard" (resp. "easy") subgroups. This is due to the fact that the same threshold $q_{1-\alpha}(\mathcal{S})$ is applied to any test point.

SCP: classification with Adaptive Prediction Sets[8]

1. Sort in decreasing order $\hat{p}_{\sigma_x(1)}(x) \geq \ldots \geq \hat{p}_{\sigma_x(C)}(x)$

[8] Romano et al. (2020b), *Classification with Valid and Adaptive Coverage*, NeurIPS

SCP: classification with Adaptive Prediction Sets[8]

1. Sort in decreasing order $\hat{p}_{\sigma_x(1)}(x) \geq \ldots \geq \hat{p}_{\sigma_x(C)}(x)$
2. $s(x, y; \hat{p}) := \sum_{k=1}^{\sigma_x^{-1}(y)} \hat{p}_{\sigma_x(k)}(x)$ (sum of the estimated probabilities associated to classes at least as large as that of the true class Y)

[8] Romano et al. (2020b), *Classification with Valid and Adaptive Coverage*, NeurIPS

SCP: classification with Adaptive Prediction Sets[8]

1. Sort in decreasing order $\hat{p}_{\sigma_x(1)}(x) \geq \ldots \geq \hat{p}_{\sigma_x(C)}(x)$
2. $s(x, y; \hat{p}) := \sum_{k=1}^{\sigma_x^{-1}(y)} \hat{p}_{\sigma_x(k)}(x)$ (sum of the estimated probabilities associated to classes at least as large as that of the true class Y)
3. Return the set of classes $\left\{\sigma_{X_{n+1}}(1), \ldots, \sigma_{X_{n+1}}(r^\star)\right\}$, where

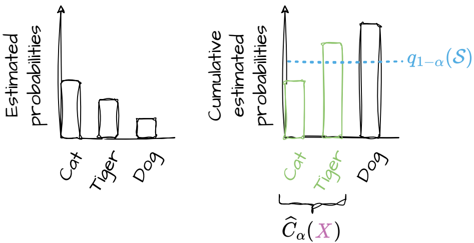
$$r^\star = \underset{1 \leq r \leq C}{\arg\max} \left\{ \sum_{k=1}^{r} \hat{p}_{\sigma_{X_{n+1}}(k)}(X_{n+1}) < q_{1-\alpha}(\mathcal{S}) \right\} + 1$$

[8] Romano et al. (2020b), *Classification with Valid and Adaptive Coverage*, NeurIPS

SCP: classification with Adaptive Prediction Sets[8]

1. Sort in decreasing order $\hat{p}_{\sigma_x(1)}(x) \geq \ldots \geq \hat{p}_{\sigma_x(C)}(x)$
2. $\boxed{s}(x, y; \hat{p}) := \sum_{k=1}^{\sigma_x^{-1}(y)} \hat{p}_{\sigma_x(k)}(x)$ (sum of the estimated probabilities associated to classes at least as large as that of the true class Y)
3. Return the set of classes $\left\{\sigma_{X_{n+1}}(1), \ldots, \sigma_{X_{n+1}}(r^\star)\right\}$, where

$$r^\star = \underset{1 \leq r \leq C}{\arg\max} \left\{ \sum_{k=1}^{r} \hat{p}_{\sigma_{X_{n+1}}(k)}(X_{n+1}) < q_{1-\alpha}(\mathcal{S}) \right\} + 1$$



[8] Romano et al. (2020b), *Classification with Valid and Adaptive Coverage*, NeurIPS
Figure highly inspired by Angelopoulos and Bates (2023).

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\mathrm{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\mathrm{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\mathrm{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

- $q_{1-\alpha}(\mathcal{S}) = 0.95$

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

- $q_{1-\alpha}(\mathcal{S}) = 0.95$

  $\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.05, 0.45, 0.5)$

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

- $q_{1-\alpha}(\mathcal{S}) = 0.95$

  $\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.05, 0.45, 0.5)$, $r^\star = 2$

$\widehat{C}_\alpha(X_{n+1}) = \{$"tiger", "cat"$\}$

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{\text{"dog"}, \text{"tiger"}, \text{"cat"}\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

- $q_{1-\alpha}(\mathcal{S}) = 0.95$
  - $\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.05, 0.45, 0.5)$, $r^\star = 2$

    $\widehat{C}_\alpha(X_{n+1}) = \{\text{"tiger"}, \text{"cat"}\}$
  - $\hookrightarrow$ Ex 2: $\hat{A}(X_{n+1}) = (0.03, 0.95, 0.02)$

SCP: classification with Adaptive Prediction Sets in practice

Ex: $Y \in \{$"dog", "tiger", "cat"$\}$, with $\alpha = 0.1$

- Scores on the calibration set

Cal_i	"dog"	"dog"	"dog"	"tiger"	"tiger"	"tiger"	"tiger"	"cat"	"cat"	"cat"
$\hat{p}_{\text{dog}}(X_i)$	0.95	0.90	0.85	0.05	0.05	0.05	0.10	0.25	0.10	0.15
$\hat{p}_{\text{tiger}}(X_i)$	0.02	0.05	0.10	0.85	0.80	0.75	0.75	0.40	0.30	0.30
$\hat{p}_{\text{cat}}(X_i)$	0.03	0.05	0.05	0.10	0.15	0.20	0.15	0.35	0.60	0.55
S_i	0.95	0.90	0.85	0.85	0.80	0.75	0.75	0.75	0.60	0.55

- $q_{1-\alpha}(\mathcal{S}) = 0.95$

  $\hookrightarrow$ Ex 1: $\hat{A}(X_{n+1}) = (0.05, 0.45, 0.5)$, $r^\star = 2$

  $\widehat{C}_\alpha(X_{n+1}) = \{$"tiger", "cat"$\}$

  $\hookrightarrow$ Ex 2: $\hat{A}(X_{n+1}) = (0.03, 0.95, 0.02)$, $r^\star = 1$

  $\widehat{C}_\alpha(X_{n+1}) = \{$"tiger"$\}$

Intro I: Split Conformal Prediction (SCP) - the simplest CP method

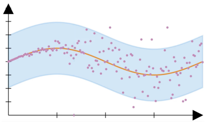
- Standard regression case
- Conformalized Quantile Regression (CQR)
- SCP - Multi class Classification
- On the design choices of conformity scores and (empirical) conditional guarantees

SCP: what choices for the regression scores?

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \boxed{\mathrm{s}} \left(X_{n+1}, y; \hat{A}\right) \leq q_{1-\alpha}(\mathcal{S})\}$$

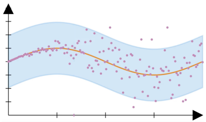
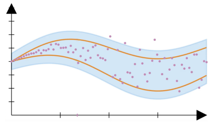
SCP: what choices for the regression scores?

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \boxed{\mathrm{s}} \left(X_{n+1}, y; \hat{A}\right) \leq q_{1-\alpha}(\mathcal{S})\}$$

	Standard SCP Vovk et al. (2005)		
$\boxed{\mathrm{s}}(\hat{A}(X), Y)$	$\lvert\hat{\mu}(X) - Y\rvert$		
$\widehat{C}_\alpha(x)$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$		
Visu.			
✓	black-box around a “usable” prediction		
✗	not adaptive		

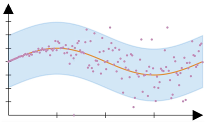
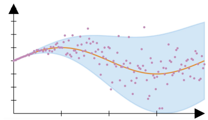
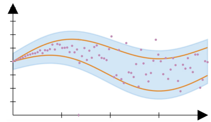
SCP: what choices for the regression scores?

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \mathrm{s}\left(X_{n+1}, y; \hat{A}\right) \leq q_{1-\alpha}(\mathcal{S})\}$$

	Standard SCP Vovk et al. (2005)		**CQR** Romano et al. (2019)
$\mathrm{s}(\hat{A}(X), Y)$	$\lvert\hat{\mu}(X) - Y\rvert$		$\max(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X) - Y, \; Y - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X))$
$\widehat{C}_\alpha(x)$	$[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$		$[\widehat{\mathrm{QR}}_{\mathrm{lower}}(x) - q_{1-\alpha}(\mathcal{S}); \; \widehat{\mathrm{QR}}_{\mathrm{upper}}(x) + q_{1-\alpha}(\mathcal{S})]$
Visu.			
✓	black-box around a "usable" prediction		adaptive
✗	not adaptive		no black-box around a "usable" prediction

SCP: what choices for the regression scores?

$$\widehat{C}_\alpha(X_{n+1}) = \{y \text{ such that } \boxed{\mathrm{s}}\left(X_{n+1}, y; \hat{A}\right) \leq q_{1-\alpha}\left(\mathcal{S}\right)\}$$

	Standard SCP Vovk et al. (2005)	**Locally weighted SCP** Lei et al. (2018)	**CQR** Romano et al. (2019)
$\boxed{\mathrm{s}}(\hat{A}(X), Y)$	$\lvert\hat{\mu}(X) - Y\rvert$	$\frac{\lvert\hat{\mu}(X)-Y\rvert}{\hat{\rho}(X)}$	$\max(\widehat{\mathrm{QR}}_{\mathrm{lower}}(X) - Y, Y - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X))$
$\widehat{C}_\alpha(x)$	$[\hat{\mu}(x) \pm q_{1-\alpha}\left(\mathcal{S}\right)]$	$[\hat{\mu}(x) \pm q_{1-\alpha}\left(\mathcal{S}\right)\hat{\rho}(x)]$	$[\widehat{\mathrm{QR}}_{\mathrm{lower}}(x) - q_{1-\alpha}\left(\mathcal{S}\right); \widehat{\mathrm{QR}}_{\mathrm{upper}}(x) + q_{1-\alpha}\left(\mathcal{S}\right)]$
Visu.			
✓	black-box around a "usable" prediction	black-box around a "usable" prediction	adaptive
✗	not adaptive	limited adaptiveness	not black-box around a "usable" prediction

Split Conformal Prediction: summary

- **Simple** procedure which quantifies the uncertainty of **any** predictive model $\hat{A}$ by returning predictive regions
- **Finite-sample** guarantees
- **Distribution-free** as long as the data are exchangeable (and so are the scores)

Split Conformal Prediction: summary

- **Simple** procedure which quantifies the uncertainty of **any** predictive model $\hat{A}$ by returning predictive regions
- **Finite-sample** guarantees
- **Distribution-free** as long as the data are exchangeable (and so are the scores)
- **Marginal** theoretical guarantee over the joint (X, Y) distribution, and not conditional, i.e., no guarantee that for any x:

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

- **Simple** procedure which quantifies the uncertainty of **any** predictive model $\hat{A}$ by returning predictive regions
- **Finite-sample** guarantees
- **Distribution-free** as long as the data are exchangeable (and so are the scores)
- **Marginal** theoretical guarantee over the joint (X, Y) distribution, and not conditional, i.e., no guarantee that for any x:

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

$\hookrightarrow$ marginal also over the whole calibration set and the test point!

Lab

Figure 1: Lab - CP - classification - with solution

Link

Figure 2: Lab - CP - classification - with parts to fill

Link

Intro I: Split Conformal Prediction (SCP) - the simplest CP method

Intro II: Overview of some challenges in Conformal Prediction

Advanced I: Towards conditional coverage

Advanced II: Avoiding data splitting: full conformal and out-of-bags approaches

Advanced III: Beyond exchangeability

Applications & Methods I: Some case studies

Applications & Methods II: Some methodological advances

Concluding remarks

Intro II: Overview of some challenges in Conformal Prediction

Conditional coverage

Data-splitting: Computational cost vs statistical power

Beyond exchangeability

Conditional coverage

Ultimately, we would love to get:

$$\mathbb{P}\left\{ Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) | X_{n+1} = x, \mathrm{Cal} \right\} \geq 1 - \alpha.$$

However:

- X_{n+1}-conditional validity is not possible (informatively) in full generality (hardness results)

Conditional coverage

Ultimately, we would love to get:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) \left| X_{n+1} = x, \mathrm{Cal} \right.\right\} \geq 1 - \alpha.$$

However:

- X_{n+1}-conditional validity is not possible (informatively) in full generality (hardness results)
- It is possible to aim for group-coditional validity, or asymptotic results, for example with universal quantile learner.

Conditional coverage

Ultimately, we would love to get:

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) \middle| X_{n+1} = x, \mathrm{Cal}\right\} \geq 1 - \alpha.$$

However:

- X_{n+1}-conditional validity is not possible (informatively) in full generality (hardness results)
- It is possible to aim for group-coditional validity, or asymptotic results, for example with universal quantile learner.
- It is possible to achieve PAC-type results that hold with high probability w.r.t. the Cal set.

Intro II: Overview of some challenges in Conformal Prediction

Conditional coverage

Data-splitting: Computational cost vs statistical power

Beyond exchangeability

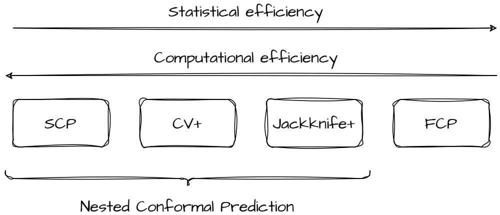
SCP suffers from data splitting:

- lower statistical efficiency (lower model accuracy and higher predictive set size)
- higher statistical variability

Avoiding data splitting: full conformal and out-of-bags approaches

SCP suffers from data splitting:

- lower statistical efficiency (lower model accuracy and higher predictive set size)
- higher statistical variability



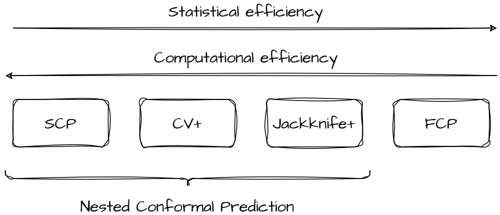
Can we avoid splitting the data set?

- Full Conformal Prediction
  - avoids data splitting
  - at the cost of many more model fits
- Jackknife+: (Barber et al., 2021b)
  - Based on leave-one-out (LOO) residuals

Avoiding data splitting: full conformal and out-of-bags approaches

SCP suffers from data splitting:

- lower statistical efficiency (lower model accuracy and higher predictive set size)
- higher statistical variability



Can we avoid splitting the data set?

- Full Conformal Prediction
  - avoids data splitting
  - at the cost of many more model fits
- Jackknife+: (Barber et al., 2021b)
  - Based on leave-one-out (LOO) residuals

Non exhaustive references.

Intro II: Overview of some challenges in Conformal Prediction

Conditional coverage

Data-splitting: Computational cost vs statistical power

Beyond exchangeability

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant
$\looparrowright$ (Tibshirani et al., 2019) (based on modeling and estimating the shifts, then reweighting)

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant
$\looparrowright$ (Tibshirani et al., 2019) (based on modeling and estimating the shifts, then reweighting)

✗ Label shift, i.e. $\mathcal{L}_Y$ changes but $\mathcal{L}_{X|Y}$ stays constant
$\looparrowright$ Similar approach by (Podkopaev and Ramdas, 2021).

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant
$\looparrowright$ (Tibshirani et al., 2019) (based on modeling and estimating the shifts, then reweighting)

✗ Label shift, i.e. $\mathcal{L}_Y$ changes but $\mathcal{L}_{X|Y}$ stays constant
$\looparrowright$ Similar approach by (Podkopaev and Ramdas, 2021).

✗ Arbitrary distribution shift
$\looparrowright$ see Cauchois et al. (2020) (distributionally robust optimization) Chernozhukov et al. (2018) (strongly mixing case), Barber et al. (2022) (quantifying non ex changeability)

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

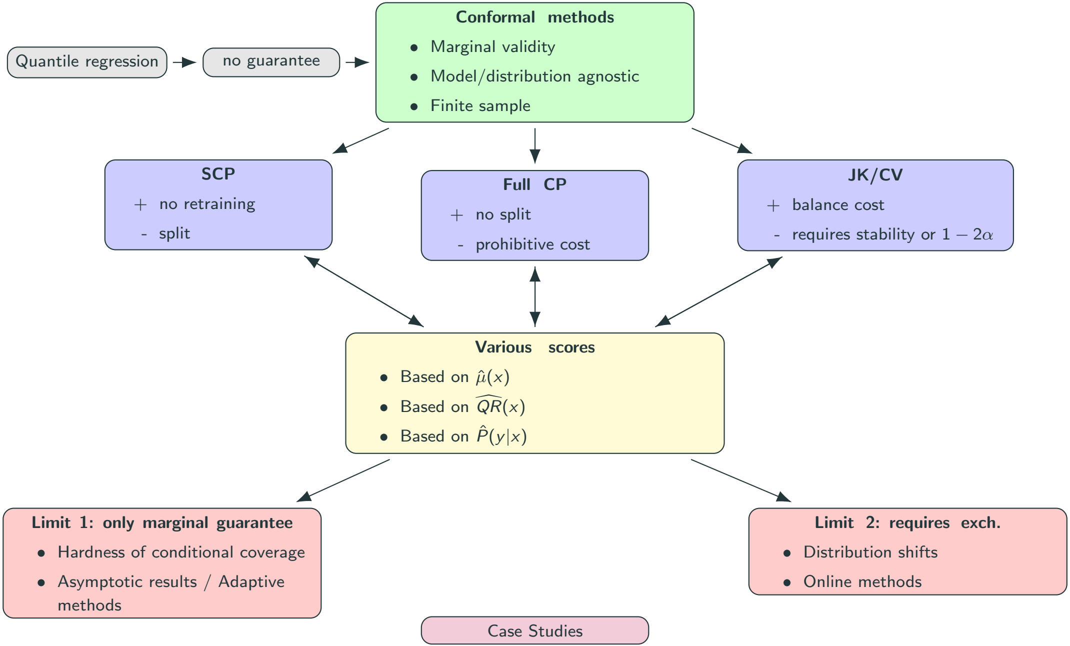
✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant
↬ (Tibshirani et al., 2019) (based on modeling and estimating the shifts, then reweighting)

✗ Label shift, i.e. $\mathcal{L}_Y$ changes but $\mathcal{L}_{X|Y}$ stays constant
↬ Similar approach by (Podkopaev and Ramdas, 2021).

✗ Arbitrary distribution shift
↬ see Cauchois et al. (2020) (distributionally robust optimization) Chernozhukov et al. (2018) (strongly mixing case), Barber et al. (2022) (quantifying non ex changeability)

✗ Possibly many shifts, not only one

Summary: →A more complete tutorial on conformal methods



Advanced I: Towards conditional coverage

On distribution-free X-conditional validity

Impact of the calibration set on the coverage

C_α = estimated predictive set based on n data points.

Distribution-free X-conditional validity

$\widehat{C}_\alpha$ achieves distribution-free X-conditional validity if:

- for any distribution $\mathcal{D}$,
- for any associated exchangeable joint distribution $\mathcal{D}^{\mathrm{exch}(n+1)}$,

we have that:

$$\mathbb{P}_{\mathcal{D}^{\mathrm{exch}(n+1)}}\left(Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) \middle| X_{n+1}\right) \overset{a.s.}{\geq} 1-\alpha.$$

Informative conditional coverage as such is impossible

Impossibility results Vovk (2012); Lei and Wasserman (2014)

If $\widehat{C}_\alpha$ is distribution-free X-conditionally valid, then, for any $\mathcal{D}$, for $\mathcal{D}_X$–almost all $\mathcal{D}_X$–non-atoms $x \in \mathcal{X}$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(\mathrm{mes}\left(\widehat{C}_\alpha\left(x\right)\right) = \infty\right) \geq 1 - \alpha.$$

Analogous statement is also available for the classification framework.

Informative conditional coverage as such is impossible

Impossibility results Vovk (2012); Lei and Wasserman (2014)

If $\widehat{C}_\alpha$ is distribution-free X-conditionally valid, then, for any $\mathcal{D}$, for $\mathcal{D}_X$–almost all $\mathcal{D}_X$–non-atoms $x \in \mathcal{X}$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(\mathrm{mes}\left(\widehat{C}_\alpha\left(x\right)\right) = \infty\right) \geq 1 - \alpha.$$

$\hookrightarrow$ distribution-free X-conditional hardness result applies beyond CP

Analogous statement is also available for the classification framework.

Informative conditional coverage as such is impossible

Impossibility results Vovk (2012); Lei and Wasserman (2014)

If $\widehat{C}_\alpha$ is distribution-free X-conditionally valid, then, **for any** $\mathcal{D}$, for $\mathcal{D}_X$–almost all $\mathcal{D}_X$–non-atoms $x \in \mathcal{X}$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(\mathrm{mes}\left(\widehat{C}_\alpha(x)\right) = \infty\right) \geq 1 - \alpha.$$

$\hookrightarrow$ distribution-free X-conditional hardness result applies beyond CP

$\hookrightarrow$ X-conditional estimators are overly large even on easy cases

Analogous statement is also available for the classification framework.

Informative conditional coverage as such is impossible

Impossibility results Vovk (2012); Lei and Wasserman (2014)

If $\widehat{C}_\alpha$ is distribution-free X-conditionally valid, then, for any $\mathcal{D}$, for $\mathcal{D}_X$–almost all $\mathcal{D}_X$–non-atoms $x \in \mathcal{X}$, it holds:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n)}}\left(\mathrm{mes}\left(\widehat{C}_\alpha\left(x\right)\right) = \infty\right) \geq 1 - \alpha.$$

$\hookrightarrow$ distribution-free X-conditional hardness result applies beyond CP

$\hookrightarrow$ X-conditional estimators are overly large even on easy cases

$\hookrightarrow$ the lower bound is tight

Naive estimator

$\mathcal{C}_\alpha(\cdot;\xi) \equiv \mathcal{Y}\mathbb{1}\left\{\xi \leq 1-\alpha\right\} + \emptyset\mathbb{1}\left\{\xi > \alpha\right\}$, where $\xi \sim \mathcal{U}\left([0,1]\right)$.

Analogous statement is also available for the classification framework.

distribution-free $(1-\alpha, \delta)$–X-conditional validity

Let $\delta > 0$ be a tolerance level.
An estimator $\widehat{C}_\alpha$ achieves distribution-free $(1-\alpha, \delta)$–X-conditional validity if for any distribution $\mathcal{D}$, for any $\mathscr{X} \subseteq \mathcal{X}$ such that $\mathbb{P}_{\mathcal{D}_X}(X \in \mathscr{X}) \geq \delta$, and for any associated exchangeable joint distribution $\mathcal{D}^{\text{exch}(n+1)}$, we have:

$$\mathbb{P}_{\mathcal{D}^{\text{exch}(n+1)}}\left(Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) \middle| X_{n+1} \in \mathscr{X}\right) \geq 1 - \alpha.$$

distribution-free $(1-\alpha,\delta)$–X-conditional validity

Let $\delta > 0$ be a tolerance level.
An estimator $\widehat{C}_\alpha$ achieves distribution-free $(1-\alpha,\delta)$–X-conditional validity if for any distribution $\mathcal{D}$, for any $\mathcal{X} \subseteq \mathcal{X}$ such that $\mathbb{P}_{\mathcal{D}_X}(X \in \mathcal{X}) \geq \delta$, and for any associated exchangeable joint distribution $\mathcal{D}^{\text{exch}(n+1)}$, we have:

$$\mathbb{P}_{\mathcal{D}^{\text{exch}(n+1)}}\left(Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) | X_{n+1} \in \mathcal{X}\right) \geq 1-\alpha.$$

Informal theorem (lower bound on $(1-\alpha,\delta)$–X-cond. valid efficiency)

An estimator achieving $(1-\alpha,\delta)$–X-conditional validity can not be more efficient than an estimator achieving **distribution-free marginal validity at the level** $1-\alpha\delta$.

$\hookrightarrow$ In practice, consider small $\delta \rightarrow$ unefficient predictive sets.

Getting closer to X-conditional coverage

Getting closer to X-conditional coverage

- Approximate conditional coverage

  $\hookrightarrow$ Romano et al. (2020a); Guan (2022); Jung et al. (2023); Gibbs et al. (2023)

  Target $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) | X_{n+1} \in \mathcal{R}(x)) \geq 1 - \alpha$

Non exhaustive references.

Getting closer to X-conditional coverage

- Approximate conditional coverage
  $\hookrightarrow$ Romano et al. (2020a); Guan (2022); Jung et al. (2023); Gibbs et al. (2023)
  Target $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) \,|X_{n+1} \in \mathcal{R}(x)) \geq 1 - \alpha$
- Asymptotic (with the sample size) conditional coverage
  $\hookrightarrow$ Romano et al. (2019); Kivaranovic et al. (2020); Chernozhukov et al. (2021); Sesia and Romano (2021); Izbicki et al. (2022)

Non exhaustive references.

- Approximate conditional coverage
  $\hookrightarrow$ Romano et al. (2020a); Guan (2022); Jung et al. (2023); Gibbs et al. (2023)
  Target $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1}) | X_{n+1} \in \mathcal{R}(x)) \geq 1 - \alpha$
- Asymptotic (with the sample size) conditional coverage
  $\hookrightarrow$ **Romano et al. (2019)**; Kivaranovic et al. (2020); Chernozhukov et al. (2021); Sesia and Romano (2021); Izbicki et al. (2022)

Non exhaustive references.

Advanced I: Towards conditional coverage

On distribution-free X-conditional validity

Impact of the calibration set on the coverage

Probably Approximately Correct bounds on calibration-conditional coverage (Vovk, 2012; Bian and Barber, 2023)

calibration conditional validity of SCP

SCP outputs $\widehat{C}_\alpha$ such that for any distribution $\mathcal{D}$ and any $0 < \delta \leq 0.5$:

$$\mathbb{P}_{\mathcal{D}^{\otimes(n+1)}}\left(\mathbb{P}_{\mathcal{D}}\left(Y_{n+1} \notin \widehat{C}_{n,\alpha}\left(X_{n+1}\right) \,|\, (X_i, Y_i)_{i=1}^n\right) \leq \alpha + \sqrt{\frac{\log(1/\delta)}{2\#\mathrm{Cal}}}\right) \geq 1-\delta.$$

$\hookrightarrow$ controls the deviation of miscoverage with respect to the nominal level of a predictive set built on a given calibration set.

Advanced II: Avoiding data splitting: full conformal and out-of-bags approaches

Full Conformal Prediction

Jackknife+

Splitting the data might not be desired

SCP suffers from data splitting:

- lower statistical efficiency (lower model accuracy and higher predictive set size)
- higher statistical variability

Splitting the data might not be desired

SCP suffers from data splitting:

- lower statistical efficiency (lower model accuracy and higher predictive set size)
- higher statistical variability

Can we avoid splitting the data set?

The naive idea does not enjoy valid coverage (even empirically)

- A naive idea:
  - Get $\hat{A}$ by training the algorithm $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

The naive idea does not enjoy valid coverage (even empirically)

- A naive idea:
  - Get $\hat{A}$ by training the algorithm $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
  - compute the empirical quantile $q_{1-\alpha}(\mathcal{S})$ of the set of scores

$$\mathcal{S} = \left\{ \mathrm{s}\left(X_i, Y_i; \hat{A}\right)\right\}_{i=1}^{n} \cup \{\infty\}.$$

The naive idea does not enjoy valid coverage (even empirically)

- A naive idea:
  - Get $\hat{A}$ by training the algorithm $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
  - compute the empirical quantile $q_{1-\alpha}(\mathcal{S})$ of the set of scores

$$\mathcal{S} = \left\{ \mathsf{s}\left(X_i, Y_i; \hat{A}\right) \right\}_{i=1}^{n} \cup \{\infty\}.$$

  - output the set $\left\{y \text{ such that } \mathsf{s}\left(X_{n+1}, y; \hat{A}\right) \leq q_{1-\alpha}(\mathcal{S})\right\}$.

The naive idea does not enjoy valid coverage (even empirically)

- A naive idea:
  - Get $\hat{A}$ by training the algorithm $\mathcal{A}$ on $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$.
  - compute the empirical quantile $q_{1-\alpha}(\mathcal{S})$ of the set of scores

$$\mathcal{S} = \left\{ \mathrm{s}\left(X_i, Y_i; \hat{A}\right) \right\}_{i=1}^{n} \cup \{\infty\}.$$

  - output the set $\left\{ y \text{ such that } \mathrm{s}\left(X_{n+1}, y; \hat{A}\right) \leq q_{1-\alpha}(\mathcal{S}) \right\}$.

✗ $\hat{A}$ obtained w. the training set $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ but not X_{n+1}.

"Naive Idea" sets with an interpolating algorithm

Assume $\mathcal{A}$ interpolates:

- $\hat{A} = \mathcal{A}\left((x_1, y_1), \dots, (x_n, y_n)\right)$
- $\hat{A}(x_k) - y_k = 0$ for any $k \in [\![1, n]\!]$

$\Rightarrow$ Naive method above *(with MAE score functions)* outputs $\{\hat{A}(X_{n+1})\}$ (a single point) for any new test point!

Full CP (Vovk et al., 2005) does not discard training points!

- Full Conformal Prediction
  - avoids data splitting

Full CP (Vovk et al., 2005) does not discard training points!

- Full Conformal Prediction
  - avoids data splitting
  - at the cost of many more model fits

Full CP (Vovk et al., 2005) does not discard training points!

- Full Conformal Prediction
  - avoids data splitting
  - at the cost of many more model fits
- **Idea**: the most probable labels Y_{n+1} live in $\mathcal{Y}$, and have a low enough conformity score. By looping over all possible $y \in \mathcal{Y}$, the ones leading to the smallest conformity scores will be found.

Full CP: recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$

Full CP: recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$
2. Obtain a set of training scores

$$\mathcal{S}_y^{(\text{train})} = \left\{ \mathsf{s} \left(X_i, Y_i; \hat{A}_y \right) \right\}_{i=1}^{n} \cup \left\{ \mathsf{s} \left(X_{n+1}, y; \hat{A}_y \right) \right\}$$

and compute their $1 - \alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)$

Full CP: recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$
2. Obtain a set of training scores

$$\mathcal{S}_y^{(\text{train})} = \left\{ \mathsf{s}\left(X_i, Y_i; \hat{A}_y\right) \right\}_{i=1}^{n} \cup \{ \mathsf{s}\left(X_{n+1}, y; \hat{A}_y\right) \}$$

and compute their $1 - \alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)$

Output the set $\left\{ y \text{ such that } \mathsf{s}\left(X_{n+1}, y; \hat{A}_y\right) \leq q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right) \right\}$.

Full CP: recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$
2. Obtain a set of training scores

$$\mathcal{S}_y^{\text{(train)}} = \left\{ \boxed{\mathrm{s}} \left(X_i, Y_i; \hat{A}_y \right) \right\}_{i=1}^{n} \cup \left\{ \boxed{\mathrm{s}} \left(X_{n+1}, y; \hat{A}_y \right) \right\}$$

   and compute their $1 - \alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{\text{(train)}}\right)$

Output the set $\left\{ y \text{ such that } \boxed{\mathrm{s}} \left(X_{n+1}, y; \hat{A}_y \right) \leq q_{1-\alpha}\left(\mathcal{S}_y^{\text{(train)}}\right) \right\}$.

✓ Test point treated in the same way than train points

Full CP: recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$
2. Obtain a set of training scores

$$\mathcal{S}_y^{(\text{train})} = \left\{ \mathrm{s}\left(X_i, Y_i; \hat{A}_y\right)\right\}_{i=1}^{n} \cup \{\mathrm{s}\left(X_{n+1}, y; \hat{A}_y\right)\}$$

and compute their $1-\alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)$

Output the set $\left\{ y \text{ such that } \mathrm{s}\left(X_{n+1}, y; \hat{A}_y\right) \leq q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)\right\}$.

✓ Test point treated in the same way than train points

✓ Any score works

Full CP: recovering exchangeability

For any candidate (X_{n+1}, y):

1. Get $\hat{A}_y$ by training $\mathcal{A}$ on $\{(X_1, Y_1), \dots, (X_n, Y_n)\} \cup \{(X_{n+1}, y)\}$
2. Obtain a set of training scores

$$\mathcal{S}_y^{(\text{train})} = \left\{ \mathrm{s}\left(X_i, Y_i; \hat{A}_y\right)\right\}_{i=1}^{n} \cup \left\{ \mathrm{s}\left(X_{n+1}, y; \hat{A}_y\right)\right\}$$

and compute their $1-\alpha$ empirical quantile $q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)$

Output the set $\left\{ y \text{ such that } \mathrm{s}\left(X_{n+1}, y; \hat{A}_y\right) \leq q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right)\right\}$.

- ✓ Test point treated in the same way than train points
- ✓ Any score works
- ✗ Computationally costly

Symmetrical algorithm

A deterministic algorithm $\mathcal{A} : (U_1, \ldots, U_n) \mapsto \hat{A}$ is symmetric if for any permutation σ of $[\![1, n]\!]$: $\mathcal{A}(U_1, \ldots, U_n) \overset{\text{a.s.}}{=} \mathcal{A}\left(U_{\sigma(1)}, \ldots, U_{\sigma(n)}\right)$.

Symmetrical algorithm

A deterministic algorithm $\mathcal{A} : (U_1, \ldots, U_n) \mapsto \hat{A}$ is symmetric if for any permutation σ of $[\![1, n]\!]$: $\mathcal{A}\left(U_1, \ldots, U_n\right) \overset{\text{a.s.}}{=} \mathcal{A}\left(U_{\sigma(1)}, \ldots, U_{\sigma(n)}\right)$.

Exchangeable scores

If the algorithm $\mathcal{A} : (U_1, \ldots, U_n) \mapsto \hat{A}$ is symmetric, and $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable, then $S_1, \ldots, S_{n+1}$ are exchangeable, with

$$S_i := \mathrm{s}\left(X_i, Y_i; \hat{A}_{Y_{n+1}}\right).$$

Symmetrical algorithm

A deterministic algorithm $\mathcal{A} : (U_1, \ldots, U_n) \mapsto \hat{A}$ is symmetric if for any permutation σ of $[\![1, n]\!]$: $\mathcal{A}\left(U_1, \ldots, U_n\right) \overset{\text{a.s.}}{=} \mathcal{A}\left(U_{\sigma(1)}, \ldots, U_{\sigma(n)}\right)$.

Exchangeable scores

If the algorithm $\mathcal{A} : (U_1, \ldots, U_n) \mapsto \hat{A}$ is symmetric, and $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable, then $S_1, \ldots, S_{n+1}$ are exchangeable, with

$$S_i := \mathrm{s}\left(X_i, Y_i; \hat{A}_{Y_{n+1}}\right).$$

Moreover

$$Y_{n+1} \in \widehat{C_\alpha^{\text{Full}}}(X_{n+1}) := \left\{ y \text{ such that } \mathrm{s}\left(X_{n+1}, y; \hat{A}_y\right) \leq q_{1-\alpha}\left(\mathcal{S}_y^{(\text{train})}\right) \right\}$$

$$\Leftrightarrow \mathrm{s}\left(X_{n+1}, Y_{n+1}; \hat{A}_{Y_{n+1}}\right) \leq q_{1-\alpha}\left(\mathcal{S}_{Y_{n+1}}^{(\text{train})}\right)$$

$$\Leftrightarrow S_{n+1} \leq q_{1-\alpha}(S_1, \ldots, S_n, S_{n+1}) \; !$$

Full CP: theoretical guarantees

Full CP enjoys finite sample guarantees proved in Vovk et al. (2005).

Marginal validity of Full CP Vovk et al. (2005)

Suppose that

(i) $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable,

(ii) the algorithm $\mathcal{A}$ is symmetric.

Full CP applied on $(X_i, Y_i)_{i=1}^{n} \cup \{X_{n+1}\}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

Additionally, if the scores are a.s. distinct:

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

Full CP: theoretical guarantees

Full CP enjoys finite sample guarantees proved in Vovk et al. (2005).

Marginal validity of Full CP Vovk et al. (2005)

Suppose that

(i) $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable,

(ii) the algorithm $\mathcal{A}$ is symmetric.

Full CP applied on $(X_i, Y_i)_{i=1}^{n} \cup \{X_{n+1}\}$ outputs $\widehat{C}_\alpha(\cdot)$ such that:

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

Additionally, if the scores are a.s. distinct:

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

✗ Marginal coverage: $\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_\alpha\left(X_{n+1}\right) \cancel{| X_{n+1} = x}\right\} \geq 1 - \alpha$

Interpolation regime

FCP sets with an interpolating algorithm

Assume $\mathcal{A}$ interpolates:

- $\hat{A} = \mathcal{A}\left((x_1, y_1), \ldots, (x_{n+1}, y_{n+1})\right)$
- $\hat{A}(x_k) - y_k = 0$ for any $k \in [\![1, n+1]\!]$

FCP sets with an interpolating algorithm

Assume $\mathcal{A}$ interpolates:

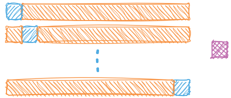
- $\hat{A} = \mathcal{A}\left((x_1, y_1), \ldots, (x_{n+1}, y_{n+1})\right)$
- $\hat{A}(x_k) - y_k = 0$ for any $k \in [\![1, n+1]\!]$

$\Rightarrow$ Full Conformal Prediction *(with standard score functions)* outputs $\mathcal{Y}$ (the whole label space) for any new test point!

Advanced II: Avoiding data splitting: full conformal and out-of-bags approaches

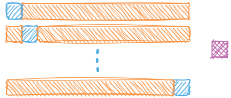
Full Conformal Prediction

Jackknife+



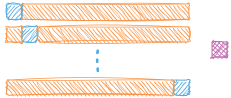
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$

Jackknife: the naive idea does not enjoy valid coverage



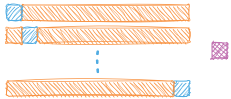
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean regression)

Jackknife: the naive idea does not enjoy valid coverage

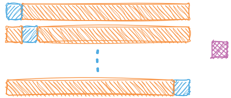


- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean regression)
- Get $\hat{A}$ by training $\mathcal{A}$ on $\mathcal{D}_n$

Jackknife: the naive idea does not enjoy valid coverage



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean regression)
- Get $\hat{A}$ by training $\mathcal{A}$ on $\mathcal{D}_n$
- Build the predictive interval: $\left[\hat{A}(X_{n+1}) \pm q_{1-\alpha}(\mathcal{S}) \right]$



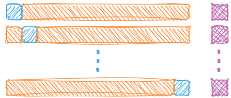
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean regression)
- Get $\hat{A}$ by training $\mathcal{A}$ on $\mathcal{D}_n$
- Build the predictive interval: $\left[\hat{A}(X_{n+1}) \pm q_{1-\alpha}(\mathcal{S}) \right]$

Warning

No guarantee on the prediction of $\hat{A}$ with scores based on $(\hat{A}_{-i})_i$, without assuming a form of **stability** on $\mathcal{A}$.

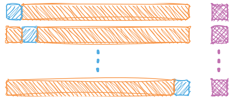
Jackknife+ (Barber et al., 2021b)

Jackknife+ (Barber et al., 2021b)



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$

Jackknife+ (Barber et al., 2021b)

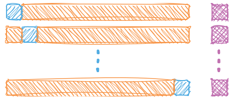


- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO predictions / predictive intervals

  $\mathcal{S}_{\text{up/down}} = \left\{ \hat{A}_{-i}(X_{n+1}) \pm |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{\pm\infty\}$

  (in standard mean regression)

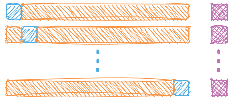
Jackknife+ (Barber et al., 2021b)



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO predictions / predictive intervals
  $\mathcal{S}_{\text{up/down}} = \left\{ \hat{A}_{-i}(X_{n+1}) \pm |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{\pm\infty\}$
  (in standard mean regression)
- Build the predictive interval: $[q_{\alpha,\inf}(\mathcal{S}_{\text{down}}); q_{1-\alpha}(\mathcal{S}_{\text{up}})]$

Recall $q_{\beta,\inf}(X_1, \ldots, X_k) := \lfloor \beta \times k \rfloor$ smallest value of $(X_1, \ldots, X_k)$

Jackknife+ (Barber et al., 2021b)



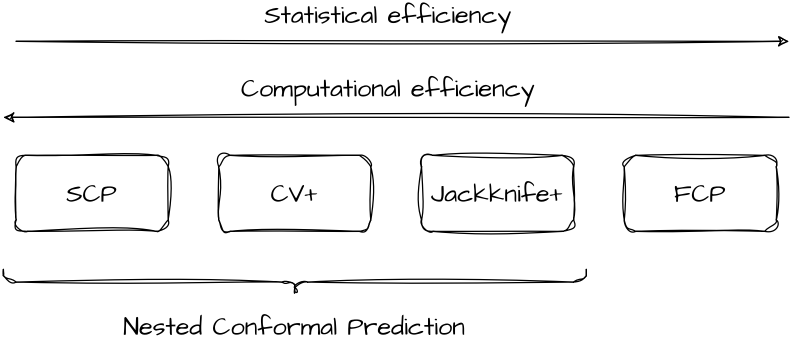
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO predictions / predictive intervals
  $\mathcal{S}_{\text{up/down}} = \left\{ \hat{A}_{-i}(X_{n+1}) \pm |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{\pm\infty\}$
  (in standard mean regression)
- Build the predictive interval: $[q_{\alpha,\inf}(\mathcal{S}_{\text{down}}); q_{1-\alpha}(\mathcal{S}_{\text{up}})]$

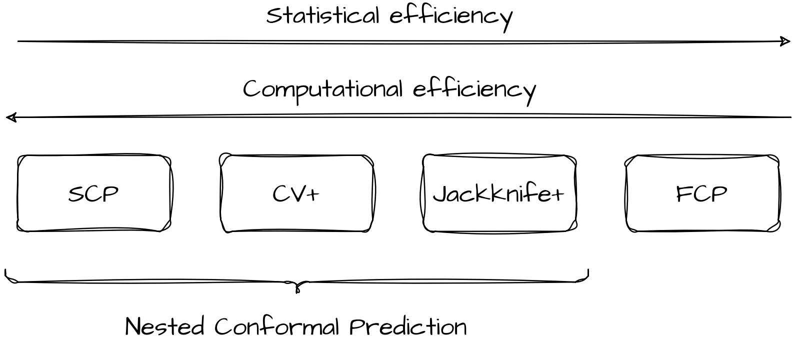
Marginal validity of Jackknife+ Barber et al. (2021b)

If $\mathcal{D}_n \cup (X_{n+1}, Y_{n+1})$ are exchangeable and $\mathcal{A}$ is symmetric: $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - 2\alpha$.

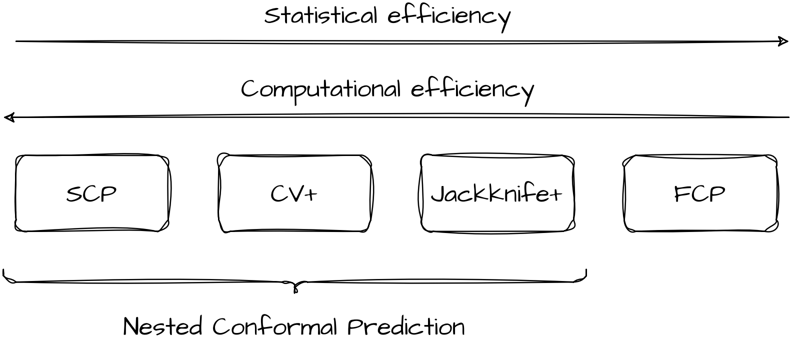
Recall $q_{\beta,\inf}(X_1, \ldots, X_k) := \lfloor \beta \times k \rfloor$ smallest value of $(X_1, \ldots, X_k)$

General overview





- Generalized framework encapsulating out-of-sample methods: Nested CP (Gupta et al., 2022) $\rightarrow$ extends *JK+/CV+ for any score.*



- Generalized framework encapsulating out-of-sample methods: Nested CP (Gupta et al., 2022) $\rightarrow$ extends *JK+/CV+ for any score.*
- Accelerating FCP: Nouretdinov et al. (2001); Lei (2019); Ndiaye and Takeuchi (2019); Cherubin et al. (2021); Ndiaye and Takeuchi (2022); Ndiaye (2022)

Non exhaustive references.

Intro I: Split Conformal Prediction (SCP) - the simplest CP method

Intro II: Overview of some challenges in Conformal Prediction

Advanced I: Towards conditional coverage

Advanced II: Avoiding data splitting: full conformal and out-of-bags approaches

Advanced III: Beyond exchangeability

Applications & Methods I: Some case studies

Applications & Methods II: Some methodological advances

Concluding remarks

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant

✗ Label shift, i.e. $\mathcal{L}_Y$ changes but $\mathcal{L}_{X|Y}$ stays constant

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity

✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant

✗ Label shift, i.e. $\mathcal{L}_Y$ changes but $\mathcal{L}_{X|Y}$ stays constant

✗ Arbitrary distribution shift

Exchangeability does not hold in many practical applications

- CP requires exchangeable data points to ensure validity
- ✗ Covariate shift, i.e. $\mathcal{L}_X$ changes but $\mathcal{L}_{Y|X}$ stays constant
- ✗ Label shift, i.e. $\mathcal{L}_Y$ changes but $\mathcal{L}_{X|Y}$ stays constant
- ✗ Arbitrary distribution shift
- ✗ Possibly many shifts, not only one

Covariate shift (Tibshirani et al., 2019)

- Setting:
  - $(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$
  - $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$

Covariate shift (Tibshirani et al., 2019)

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$
  - $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$
- Idea: give more importance to calibration points that are closer in distribution to the test point

Similar approach for Label shift (Podkopaev and Ramdas, 2021).

Covariate shift (Tibshirani et al., 2019)

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$
  - $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$
- Idea: give more importance to calibration points that are closer in distribution to the test point
- In practice:
  1. estimate the likelihood ratio $w(X_i) = \frac{\mathrm{d}\tilde{P}_X(X_i)}{\mathrm{d}P_X(X_i)}$

Similar approach for Label shift (Podkopaev and Ramdas, 2021).

Covariate shift (Tibshirani et al., 2019)

- Setting:
  - $(X_1, Y_1), \dots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$
  - $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$
- Idea: give more importance to calibration points that are closer in distribution to the test point
- In practice:
  1. estimate the likelihood ratio $w(X_i) = \frac{\mathrm{d}\tilde{P}_X(X_i)}{\mathrm{d}P_X(X_i)}$
  2. normalize the weights, i.e. $\omega_i = \omega(X_i) = \frac{w(X_i)}{\sum_{j=1}^{n+1} w(X_j)}$

Similar approach for Label shift (Podkopaev and Ramdas, 2021).

Covariate shift (Tibshirani et al., 2019)

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_X \times P_{Y|X}$
  - $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$
- Idea: give more importance to calibration points that are closer in distribution to the test point
- In practice:
  1. estimate the likelihood ratio $w(X_i) = \frac{\mathrm{d}\tilde{P}_X(X_i)}{\mathrm{d}P_X(X_i)}$
  2. normalize the weights, i.e. $\omega_i = \omega(X_i) = \frac{w(X_i)}{\sum_{j=1}^{n+1} w(X_j)}$
  3. outputs $\widehat{C}_\alpha(X_{n+1}) =$
  $$\left\{ y : \mathrm{s}\left(X_{n+1}, y; \hat{A}\right) \leq Q_{1-\alpha}\left(\sum_{i \in \mathrm{Cal}} \omega_i \delta_{S_i} + \omega_{n+1}\delta_\infty\right)\right\}$$

Similar approach for Label shift (Podkopaev and Ramdas, 2021).

- Arbitrary distribution shift: Cauchois et al. (2020) leverages ideas from the distributionally robust optimization literature
- Two major general theoretical results beyond exchangeability:

Generalizations

- Arbitrary distribution shift: Cauchois et al. (2020) leverages ideas from the distributionally robust optimization literature
- Two major general theoretical results beyond exchangeability:
  - Chernozhukov et al. (2018)
    $\hookrightarrow$ If the learnt model is accurate and the data noise is strongly mixing, then CP is valid asymptotically ✓

- Arbitrary distribution shift: Cauchois et al. (2020) leverages ideas from the distributionally robust optimization literature
- Two major general theoretical results beyond exchangeability:
  - Chernozhukov et al. (2018)
    $\hookrightarrow$ If the learnt model is accurate and the data noise is strongly mixing, then CP is valid asymptotically ✓
  - Barber et al. (2022)
    $\hookrightarrow$ Quantifies the coverage loss depending on the strength of exchangeability violation

    $$\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - \alpha - \text{\small average violation of exchangeability by each calibration point}$$

    $\hookrightarrow$ proposed algorithm: reweighting again!
    e.g., in a temporal setting, give higher weights to more recent points.

Online setting

- Data: T_0 random variables $(X_1, Y_1), \dots, (X_{T_0}, Y_{T_0})$ in $\mathbb{R}^d \times \mathbb{R}$
- Aim: predict the response values as well as predictive intervals for T_1 subsequent observations $X_{T_0+1}, \dots, X_{T_0+T_1}$ sequentially: at any prediction step $t \in [\![T_0 + 1, T_0 + T_1]\!]$, $Y_{t-T_0}, \dots, Y_{t-1}$ have been revealed
- Build the smallest interval $\widehat{C}_\alpha^t$ such that:

$$\mathbb{P}\left\{ Y_t \in \widehat{C}_\alpha^t(X_t) \right\} \geq 1 - \alpha, \quad \text{for } t \in [\![T_0 + 1, T_0 + T_1]\!],$$

often relaxed in:

$$\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{ Y_t \in \widehat{C}_\alpha^t(X_t) \right\} \approx 1 - \alpha.$$

Online setting

- Data: T_0 random variables $(X_1, Y_1), \ldots, (X_{T_0}, Y_{T_0})$ in $\mathbb{R}^d \times \mathbb{R}$
- Aim: predict the response values as well as predictive intervals for T_1 subsequent observations $X_{T_0+1}, \ldots, X_{T_0+T_1}$ sequentially: at any prediction step $t \in [\![T_0 + 1, T_0 + T_1]\!]$, $Y_{t-T_0}, \ldots, Y_{t-1}$ have been revealed
- Build the smallest interval $\widehat{C}_\alpha^t$ such that:

$$\mathbb{P}\left\{ Y_t \in \widehat{C}_\alpha^t(X_t) \right\} \geq 1 - \alpha, \quad \text{for } t \in [\![T_0 + 1, T_0 + T_1]\!],$$

  often relaxed in:

$$\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{ Y_t \in \widehat{C}_\alpha^t(X_t) \right\} \approx 1 - \alpha.$$

$\rightsquigarrow$ More during the case study!

Intro I: Split Conformal Prediction (SCP) - the simplest CP method

Intro II: Overview of some challenges in Conformal Prediction

Advanced I: Towards conditional coverage

Advanced II: Avoiding data splitting: full conformal and out-of-bags approaches

Advanced III: Beyond exchangeability

Applications & Methods I: Some case studies

Applications & Methods II: Some methodological advances

Concluding remarks

Applications & Methods I: Some case studies

Healthcare

Electricity

Image to Image regression with DF-UQ – Angelopoulos et al. (2022b)

- Medical application
- Image based task
- Pixel by pixel analysis $\rightsquigarrow$ applications to segmentation for self-driving cars

Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging

Anastasios N. Angelopoulos[*1] **Amit Kohli**[*1] **Stephen Bates**[1] **Michael I. Jordan**[1] **Jitendra Malik**[1] **Thayer Alshaabi**[2] **Srigokul Upadhyayula**[2,3] **Yaniv Romano**[4]

Image to Image regression with DF-UQ – Angelopoulos et al. (2022b)

- Medical application
- Image based task
- Pixel by pixel analysis $\rightsquigarrow$ applications to segmentation for self-driving cars

1. **Task**: *Image to Image regression* – for each pixel of an image, predict a real valued output from the entire image.
2. **UQ Goal:** provide a predictive interval for each pixel, such that the output is in the interval at least 90% of the time.

Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging

Anastasios N. Angelopoulos[*1] Amit Kohli[*1] Stephen Bates[1] Michael I. Jordan[1] Jitendra Malik[1]
Thayer Alshaabi[2] Srigokul Upadhyayula[2 3] Yaniv Romano[4]

- Medical application
- Image based task
- Pixel by pixel analysis $\rightsquigarrow$ applications to segmentation for self-driving cars

1. **Task**: *Image to Image regression* – for each pixel of an image, predict a real valued output from the entire image.
2. **UQ Goal:** provide a predictive interval for each pixel, such that the output is in the interval at least 90% of the time.

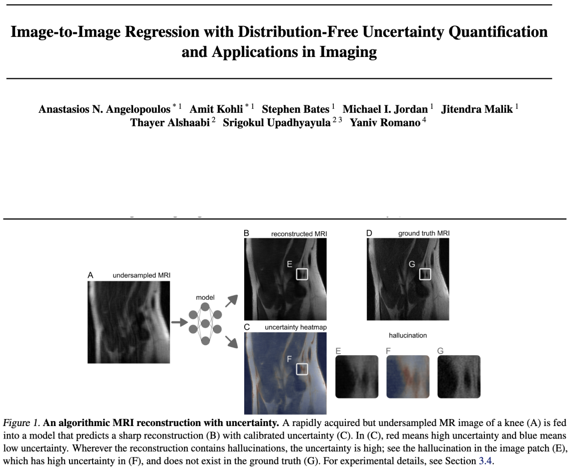


Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging

Anastasios N. Angelopoulos [* 1] Amit Kohli [* 1] Stephen Bates [1] Michael I. Jordan [1] Jitendra Malik [1] Thayer Alshaabi [2] Srigokul Upadhyayula [2 3] Yaniv Romano [4]

Figure 1. **An algorithmic MRI reconstruction with uncertainty.** A rapidly acquired but undersampled MR image of a knee (A) is fed into a model that predicts a sharp reconstruction (B) with calibrated uncertainty (C). In (C), red means high uncertainty and blue means low uncertainty. Wherever the reconstruction contains hallucinations, the uncertainty is high; see the hallucination in the image patch (E), which has high uncertainty in (F), and does not exist in the ground truth (G). For experimental details, see Section 3.4.

Figure 3: Image from Angelopoulos et al. (2022b)

Image to Image regression with DF-UQ – Angelopoulos et al. (2022b)

Method:

1. Split conformal prediction method - isolate calibration set
2. On the proper training set, learn:
   - Mean regressor - $\hat{\mu} : \mathbb{R}^{NM} \to [0; 1]$

Method:

1. Split conformal prediction method - isolate calibration set
2. On the proper training set, learn:
   - Mean regressor - $\hat{\mu} : \mathbb{R}^{NM} \to [0; 1]$
   - Heuristic notion of uncertainty: $\tilde{u}, \tilde{\ell} : \mathbb{R}^{NM} \to [0; 1]$, such that

$$[\hat{\mu}(X) - \tilde{\ell}(X); \hat{\mu}(X) + \tilde{u}(X)]$$

   $\to$ 3 regressors are used

   4 techniques are experimented for these regressors, including QR.

Method:

1. Split conformal prediction method - isolate calibration set
2. On the proper training set, learn:
   - Mean regressor - $\hat{\mu} : \mathbb{R}^{NM} \to [0; 1]$
   - Heuristic notion of uncertainty: $\tilde{u}, \tilde{\ell} : \mathbb{R}^{NM} \to [0; 1]$, such that

   $$[\hat{\mu}(X) - \tilde{\ell}(X); \hat{\mu}(X) + \tilde{u}(X)]$$

   $\to$ 3 regressors are used

   4 techniques are experimented for these regressors, including QR.
3. Calibration step: leverage the calibration set.
   - In spirit, almost equivalent to CQR but with a multiplicative form.
   - Precisely, relies on RCPS (Bates et al., 2021a)

Image to Image regression with DF-UQ – Angelopoulos et al. (2022b)

Method:

1. Split conformal prediction method - isolate calibration set
2. On the proper training set, learn:
   - Mean regressor - $\hat{\mu} : \mathbb{R}^{NM} \to [0;1]$
   - Heuristic notion of uncertainty: $\tilde{u}, \tilde{\ell} : \mathbb{R}^{NM} \to [0;1]$, such that

$$[\hat{\mu}(X) - \tilde{\ell}(X); \hat{\mu}(X) + \tilde{u}(X)]$$

   $\to$ 3 regressors are used

   4 techniques are experimented for these regressors, including QR.
3. Calibration step: leverage the calibration set.
   - In spirit, almost equivalent to CQR but with a multiplicative form.
   - Precisely, relies on RCPS (Bates et al., 2021a)

Guarantee:

$$\mathbb{P}\left[\mathbb{E}\left[\text{Average miscoverage on all pixels of a test image} | \text{Cal}\right] \geq \alpha\right] \leq \delta$$

$\to$ Marginal validity on the test, with high probability w.r.t. the calibration set.

Image to Image regression with DF-UQ – Angelopoulos et al. (2022b)

Abstract

Image-to-image regression is an important learning task, used frequently in biological imaging. Current algorithms, however, do not generally offer statistical guarantees that protect against a model's mistakes and hallucinations. To address this, we develop uncertainty quantification techniques with rigorous statistical guarantees for image-to-image regression problems. In particular, we show how to derive uncertainty intervals around each pixel that are guaranteed to contain the true value with a user-specified confidence probability. Our methods work in conjunction

2. Methods

We now formally describe the method for constructing uncertainty intervals. Each pixel in the image will get its own uncertainty interval, as in (1), that is statistically guaranteed to contain the true value with high probability.

How do you understand that?

Image to Image regression with DF-UQ – Angelopoulos et al. (2022b)

Abstract

Image-to-image regression is an important learning task, used frequently in biological imaging. Current algorithms, however, do not generally offer statistical guarantees that protect against a model's mistakes and hallucinations. To address this, we develop uncertainty quantification techniques with rigorous statistical guarantees for image-to-image regression problems. In particular, we show how to derive uncertainty intervals around each pixel that are guaranteed to contain the true value with a user-specified confidence probability. Our methods work in conjunction

- Not a conditional coverage claim!
- The statement is on-average on the test point - easy or hard.

Size-stratified risk. Next, we seek prediction sets that do not systematically make mistakes in difficult parts of the image. Our risk control requirement in Definition 2.1 may be satisfied even if the prediction sets systematically fail to contain the most difficult pixels. For example, if $\alpha = 0.1$ and 90% of pixels are covered by fixed-width intervals of size 0.01, then the requirement is satisfied—however, the sets no longer serve as useful notions of uncertainty. To

2. Methods

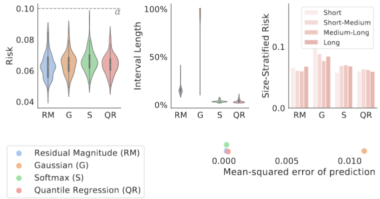
We now formally describe the method for constructing uncertainty intervals. Each pixel in the image will get its own uncertainty interval, as in (1), that is statistically guaranteed to contain the true value with high probability.

How do you understand that?

- Hard problem (impossibility results!)
- Introduce metrics to see *if* and *on which underlying regressors* such problem happens.

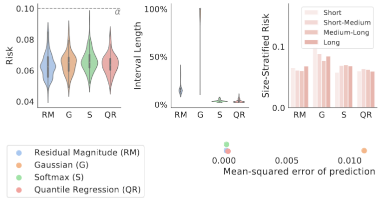
Example of such metrics (see also Feldman et al., 2021) **:**

- Link between the size of the PI and the coverage level $\longrightarrow$



Example of such metrics (see also Feldman et al., 2021) **:**

- Link between the size of the PI and the coverage level $\longrightarrow$
- Localization of the errors $\downarrow$



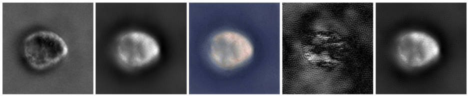


Figure 3. **Examples of quantitative phase reconstructions of leukocytes with uncertainty** shown in the following order: input (we only show one of the two illuminations), prediction, uncertainty visualization (produced with quantile regression), absolute difference between prediction and ground truth (renormalized for visualization), ground truth.

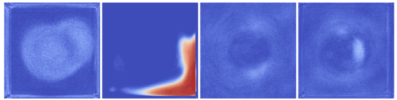


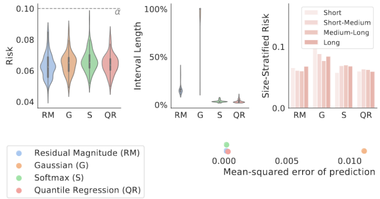
Figure 8. **Spatial variations in miscoverage** in the BSCCM dataset are shown for each of the four methods as a heatmap. Blue represents 0% miscoverage and red represents 100%. The methods are, in order, residual magnutude, gaussian, softmax, and quantile regression.

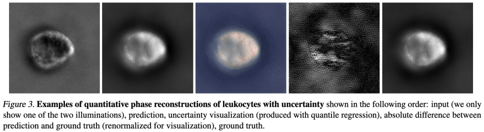
Figure 4: All images from Angelopoulos et al. (2022b)

Image to Image regression with DF-UQ – Angelopoulos et al. (2022b)

Example of such metrics (see also Feldman et al., 2021) **:**

- Link between the size of the PI and the coverage level $\longrightarrow$
- Localization of the errors $\downarrow$





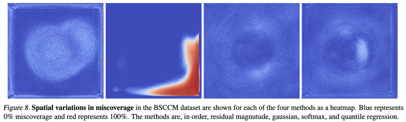


Figure 4: All images from Angelopoulos et al. (2022b)

Take aways:

- Elegant application of SCP with CQR type score
- Test marginal and calibration + train conditional validity guarantees with HP
- Main problem is Test conditionality $\rightarrow$ look at metrics to evaluate which methods performs best!

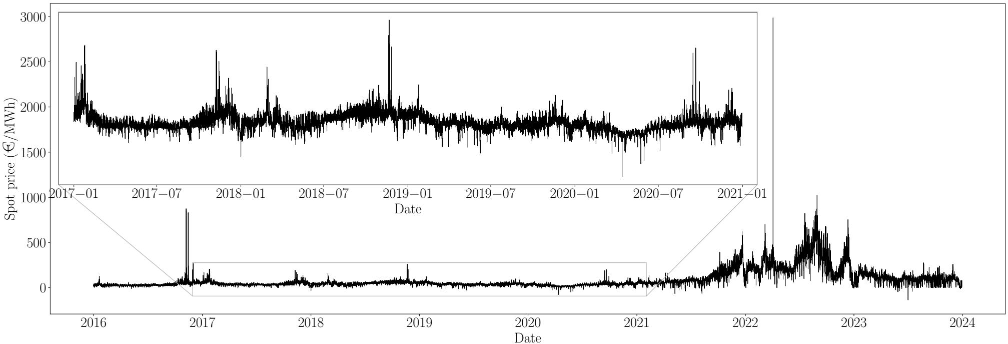
Applications & Methods I: Some case studies

Healthcare

Electricity

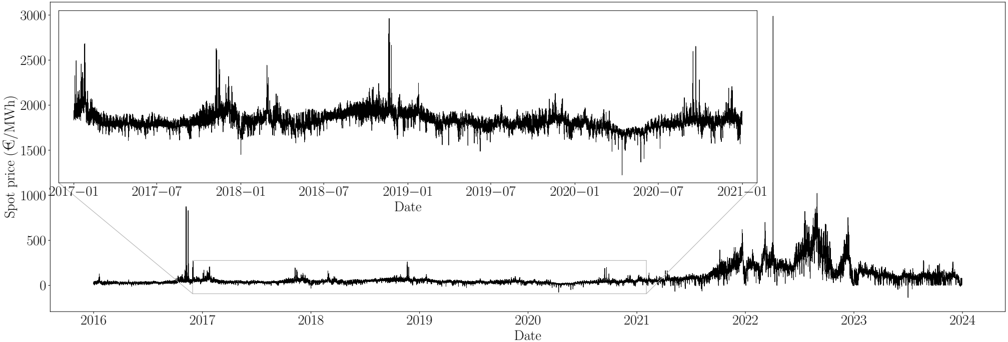
Hourly day-ahead market prices (between producers and suppliers)

Hourly day-ahead market prices (between producers and suppliers)



Forecasting French spot electricity prices

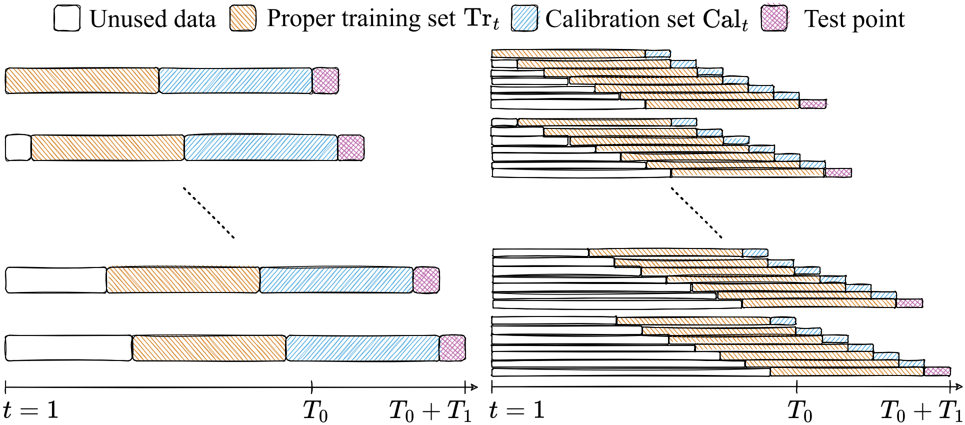
Hourly day-ahead market prices (between producers and suppliers)



To which extent are they forecastable?

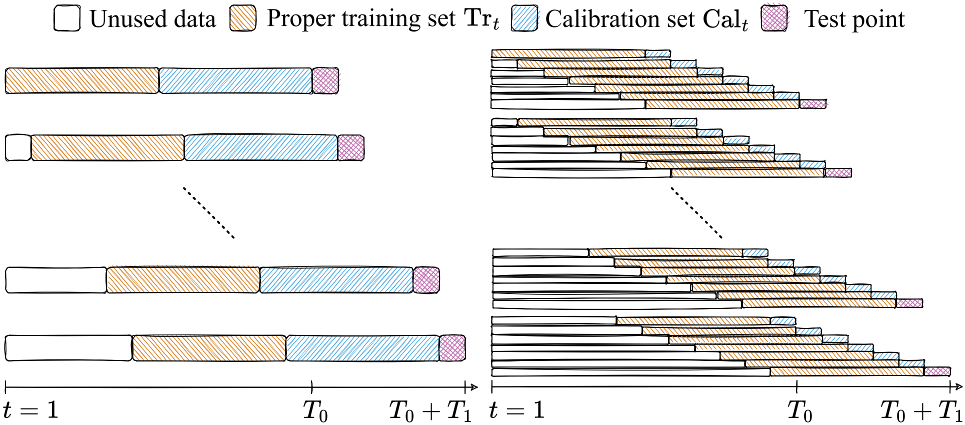
$\hookrightarrow$ forecasts errors **no lower than 10%** of the realized price!

Temporal splitting strategies: Online Sequential Split Conformal Prediction (OSSCP, Zaffran et al., 2022; Dutot et al., 2024)



(a) OSSCP (b) OSSCP-horizon

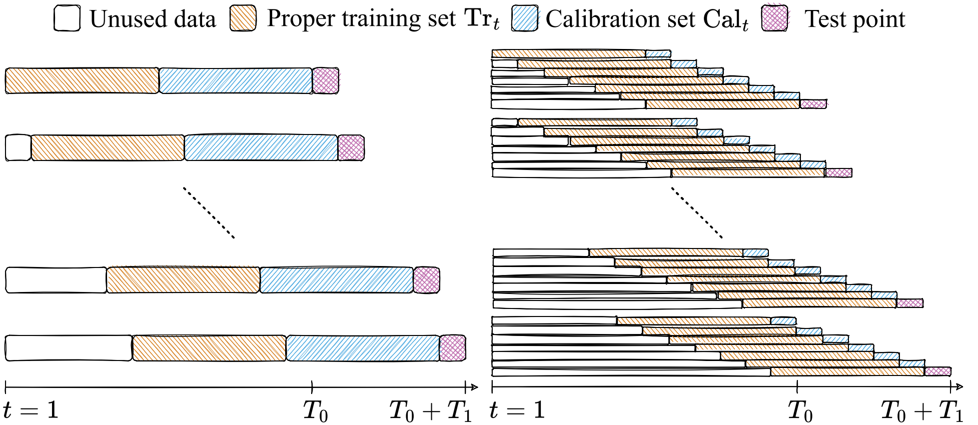
Temporal splitting strategies: Online Sequential Split Conformal Prediction (OSSCP, Zaffran et al., 2022; Dutot et al., 2024)



(a) OSSCP (b) OSSCP-horizon

$\hookrightarrow$ OSSCP improves robustness in temporal settings;

Temporal splitting strategies: Online Sequential Split Conformal Prediction (OSSCP, Zaffran et al., 2022; Dutot et al., 2024)



(a) OSSCP (b) OSSCP-horizon

$\hookrightarrow$ `OSSCP` improves robustness in temporal settings;

$\hookrightarrow$ `OSSCP-horizon` drastically improves robustness in non-stationary temporal settings.

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1} \left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t} \left(X^{(t)} \right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1} \left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t} \left(X^{(t)} \right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1}\left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t}\left(X^{(t)} \right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1} \left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t} \left(X^{(t)} \right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Guarantee: Asymptotic validity result for any sequence of observations.

$$\frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1} \left\{ Y^{(t)} \in \widehat{C}_{\alpha_t} \left(X^{(t)} \right) \right\} \underset{T_1 \to +\infty}{\longrightarrow} 1 - \alpha$$

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1}\left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t}\left(X^{(t)}\right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Guarantee: Asymptotic validity result for any sequence of observations.

$$\left| \frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1}\left\{ Y^{(t)} \in \widehat{C}_{\alpha_t}\left(X^{(t)}\right) \right\} - (1-\alpha) \right| \leq \frac{2}{\gamma T_1}$$

Zoom on Adaptive Conformal Inference (ACI, Gibbs and Candès, 2021)

Adaptive Conformal Inference (ACI) was initially proposed to handle distribution shift.

It relies on updating online an *effective miscoverage rate* α_t, with the scheme

$$\alpha_{t+1} := \alpha_t + \gamma \left(\alpha - \mathbb{1} \left\{ Y^{(t)} \notin \widehat{C}_{\alpha_t} \left(X^{(t)} \right) \right\} \right),$$

and $\alpha_1 = \alpha$, $\gamma \geq 0$.

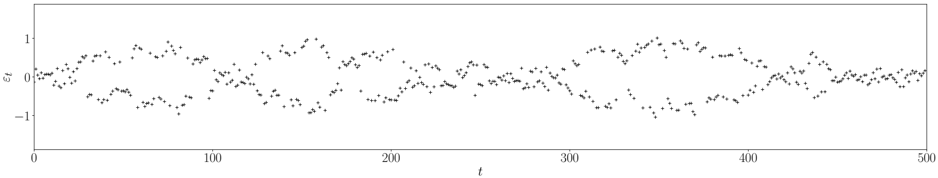
Intuition: if we did make an error, the interval was too small so we want to increase its length by taking a higher quantile (a smaller α_t). Reversely if we included the point.

Guarantee: Asymptotic validity result for any sequence of observations.

$$\left| \frac{1}{T_1} \sum_{t=T_0+1}^{T_0+T_1} \mathbb{1} \left\{ Y^{(t)} \in \widehat{C}_{\alpha_t} \left(X^{(t)} \right) \right\} - (1 - \alpha) \right| \leq \frac{2}{\gamma T_1}$$

$\Rightarrow$ favors large γ.

Visualisation of ACI procedure



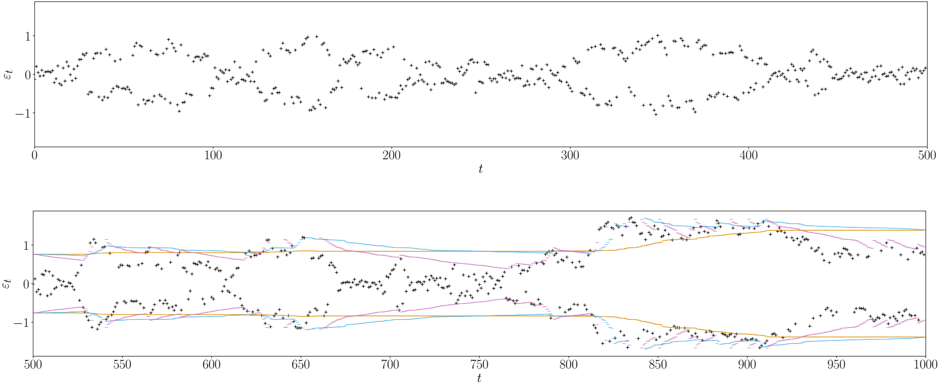
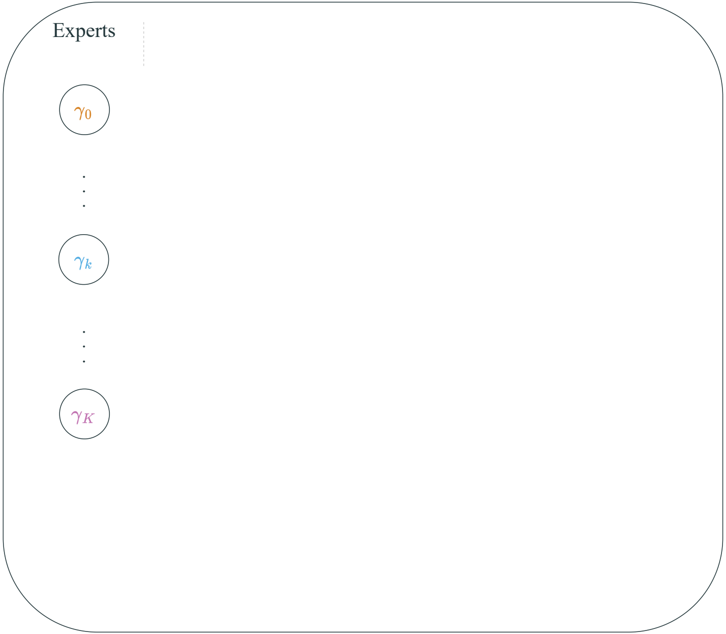
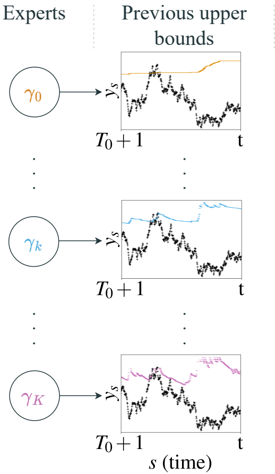


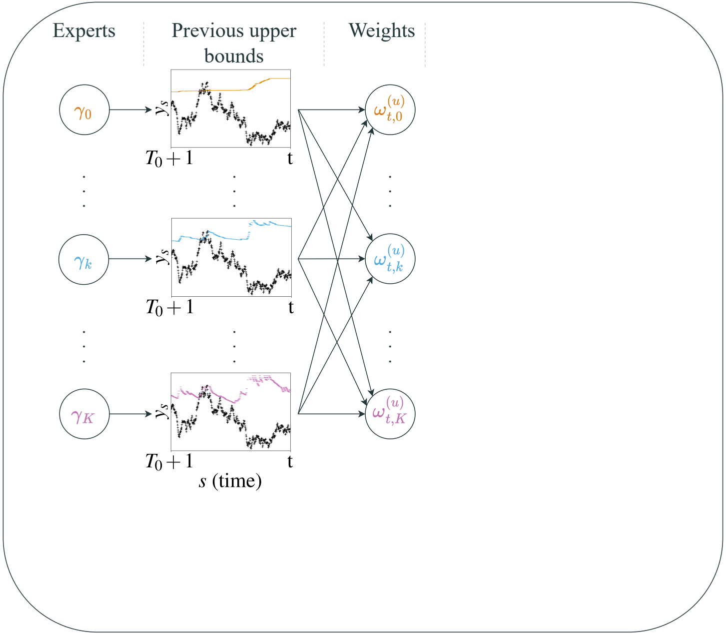
Figure 5: Visualisation of ACI with different values of γ ($\gamma = 0$, $\gamma = 0.01$, $\gamma = 0.05$)

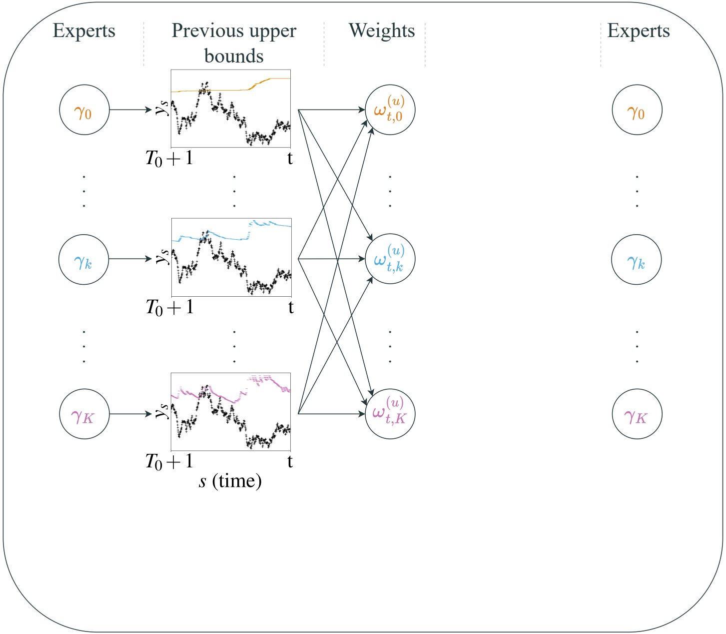
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



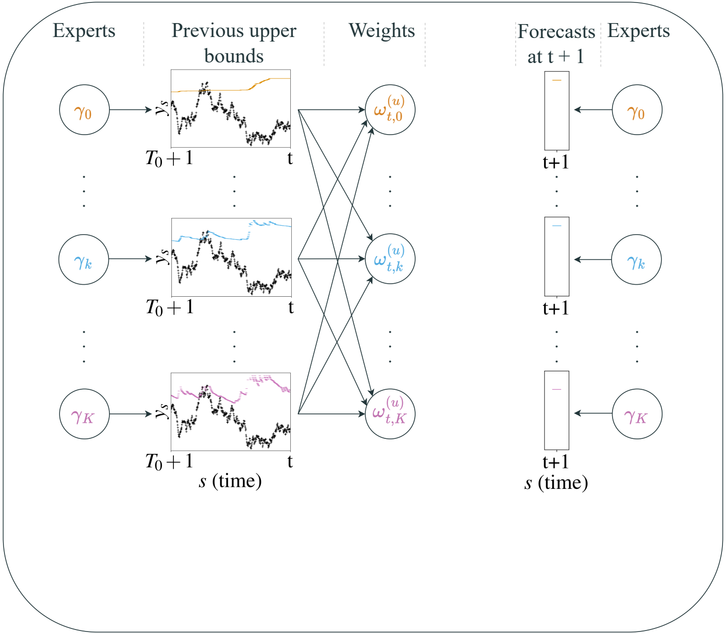

Experts
Previous upper bounds
γ_0
γ_k
γ_K
y_s
T_0+1
t
s (time)

AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)

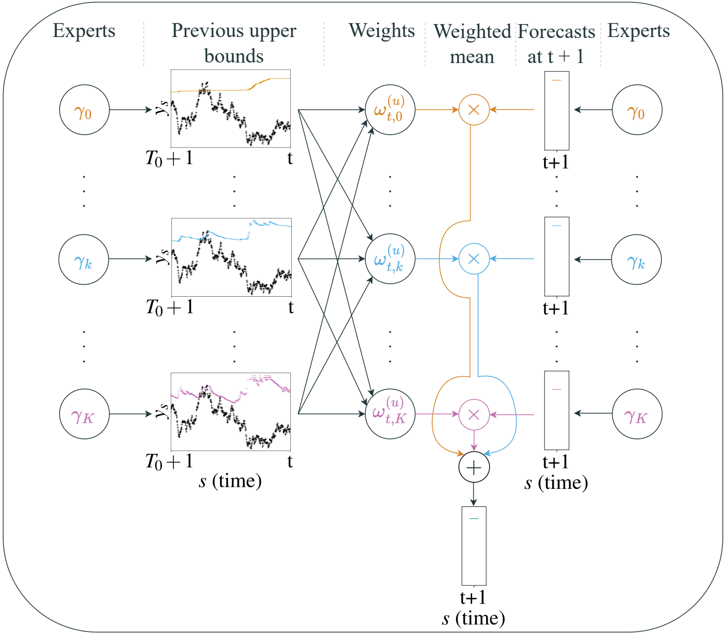



Experts
Previous upper bounds
Weights
Experts
γ_0
γ_k
γ_K
y_s
$T_0 + 1$
t
s (time)
$\omega_{t,0}^{(u)}$
$\omega_{t,k}^{(u)}$
$\omega_{t,K}^{(u)}$

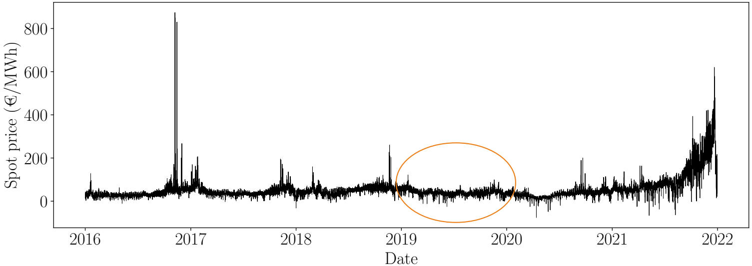
AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)



AgACI: adaptive wrapper around ACI, upper bound (Zaffran et al., 2022)

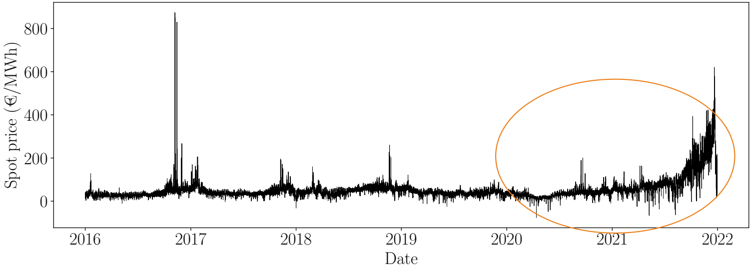


- 2019: `AgACI` provides validity with a reasonable efficiency;



Experimental take-away messages (Zaffran et al., 2022; Dutot et al., 2024)

- 2019: `AgACI` provides validity with a reasonable efficiency;
- 2020 and 2021: `AgACI` fails to ensure validity, and the various forecasting models considered[9] behave differently.



[9]Quantile Random Forests, Quantile Generalized Additive Models, Quantile Gradient Boosting, etc.

Improving adaptiveness for high non-stationarity (Dutot et al., 2024)

Online aggregation of various `AgACI`, each of them being trained with different underlying forecasting models, for each bound independently.

Improving adaptiveness for high non-stationarity (Dutot et al., 2024)

Online aggregation of various `AgACI`, each of them being trained with different underlying forecasting models, for each bound independently.

- ✓ Retrieves validity even in the most hazardous period of 2020 and 2021.

Improving adaptiveness for high non-stationarity (Dutot et al., 2024)

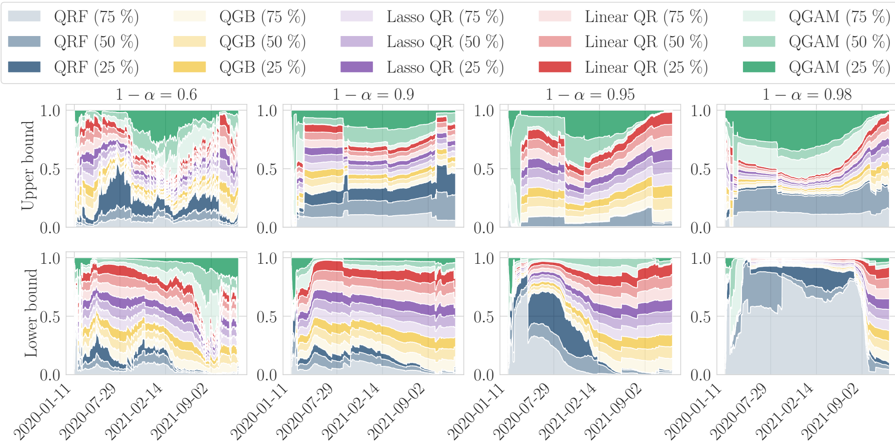
Online aggregation of various `AgACI`, each of them being trained with different underlying forecasting models, for each bound independently.

- ✓ Retrieves validity even in the most hazardous period of 2020 and 2021.
- ✓ Analyzing its weights provides interpretability.

Improving adaptiveness for high non-stationarity (Dutot et al., 2024)

Online aggregation of various `AgACI`, each of them being trained with different underlying forecasting models, for each bound independently.

- ✓ Retrieves validity even in the most hazardous period of 2020 and 2021.
- ✓ Analyzing its weights provides interpretability.



Aggregating the two bounds independently (as in `AgACI` and beyond):

Aggregating the two bounds independently (as in `AgACI` and beyond):

✓ Allows more flexible and adaptive behavior in practice, catching the varying nature of the predictive distribution tails

Highlights and perspectives

Aggregating the two bounds independently (as in `AgACI` and beyond):

- ✓ Allows more flexible and adaptive behavior in practice, catching the varying nature of the predictive distribution tails
- ✗ Prevents from obtaining theoretical guarantees (by opposition to Gibbs and Candès, 2022)

Highlights and perspectives

Aggregating the two bounds independently (as in `AgACI` and beyond):

✓ Allows more flexible and adaptive behavior in practice, catching the varying nature of the predictive distribution tails

✗ Prevents from obtaining theoretical guarantees (by opposition to Gibbs and Candès, 2022)

$\hookrightarrow$ Weaken the objective and consider a more practical theoretical aim?

Intro I: Split Conformal Prediction (SCP) - the simplest CP method

Intro II: Overview of some challenges in Conformal Prediction

Advanced I: Towards conditional coverage

Advanced II: Avoiding data splitting: full conformal and out-of-bags approaches

Advanced III: Beyond exchangeability

Applications & Methods I: Some case studies

Applications & Methods II: Some methodological advances

Concluding remarks

Applications & Methods II: Some methodological advances

- Conformal prediction and UQ with missing values
- Valid Selection among Conformal Sets

Conformal prediction and UQ with missing values

Margaux Zaffran

Yaniv Romano

Julie Josse

Conformal Prediction with Missing Values, ICML 2023, MZ, AD, JJ, YR.

Predictive Uncertainty Quantification with Missing Covariates, 2024, MZ, JJ, YR, AD.

Missing values are ubiquitous and challenging

Data: $\left(X^{(k)}, \qquad Y^{(k)}\right)_{k=1}^{n}$

Y	X_1	X_2	X_3
22	5	6	3
19	6	8	NA
19	5	3	6
7	NA	9	NA
13	4	9	0
20	NA	NA	1
9	8	NA	4

Missing values are ubiquitous and challenging

Data: $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$

Y	X_1	X_2	X_3
22	5	6	3
19	6	8	NA
19	5	3	6
7	NA	9	NA
13	4	9	0
20	NA	NA	1
9	8	NA	4

Mask $M =$

$(M_1$	M_2	$M_3)$
0	0	0
0	0	1
0	0	0
1	0	1
0	0	0
1	1	0
0	1	0

Missing values are ubiquitous and challenging

Data: $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$

Y	X_1	X_2	X_3
22	5	6	3
19	6	8	NA
19	5	3	6
7	NA	9	NA
13	4	9	0
20	NA	NA	1
9	8	NA	4

Mask $M =$

$(M_1$	M_2	$M_3)$
0	0	0
0	0	1
0	0	0
1	0	1
0	0	0
1	1	0
0	1	0

$\hookrightarrow$ 2^d potential masks.

Missing values are ubiquitous and challenging

Data: $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$

Y	X_1	X_2	X_3
22	5	6	3
19	6	8	NA
19	5	3	6
7	NA	9	NA
13	4	9	0
20	NA	NA	1
9	8	NA	4

Mask $M =$

$(M_1$	M_2	$M_3)$
0	0	0
0	0	1
0	0	0
1	0	1
0	0	0
1	1	0
0	1	0

$\hookrightarrow$ 2^d potential masks.

$\hookrightarrow$ M can depend on X or Y (depending on the missing mechanism (Rubin 1976)

Missing values are ubiquitous and challenging

Data: $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$

Y	X_1	X_2	X_3		Mask $M = (M_1$	M_2	$M_3)$
22	5	6	3		0	0	0
19	6	8	NA		0	0	1
19	5	3	6		0	0	0
7	NA	9	NA		1	0	1
13	4	9	0		0	0	0
20	NA	NA	1		1	1	0
9	8	NA	4		0	1	0

$\hookrightarrow$ 2^d potential masks.

$\hookrightarrow$ M can depend on X or Y (depending on the missing mechanism (Rubin 1976)

- **Missing Completely At Random** (MCAR): $M \perp\!\!\!\perp X$
- **Missing At Random** (MAR): missingness depends on the observed variables
- **Missing Non At Random** (MNAR)

Missing values are ubiquitous and challenging

Data: $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$

Y	X_1	X_2	X_3
22	5	6	3
19	6	8	NA
19	5	3	6
7	NA	9	NA
13	4	9	0
20	NA	NA	1
9	8	NA	4

Mask $M =$

$(M_1$	M_2	$M_3)$
0	0	0
0	0	1
0	0	0
1	0	1
0	0	0
1	1	0
0	1	0

$\hookrightarrow$ 2^d potential masks.

$\hookrightarrow$ M can depend on X or Y (depending on the missing mechanism (Rubin 1976)

- Missing Completely At Random (MCAR): $M \perp\!\!\!\perp X$
- Missing At Random (MAR): missingness depends on the observed variables
- Missing Non At Random (MNAR)
- $Y \perp\!\!\!\perp M \mid X$?

Conceptually: a structured distribution shift situation

1. For each pattern m, we have a different distribution for $(X_{\mathrm{obs(M)}}, Y)|M = m$.

2. Those distributions are connected.

3. Reasonable model of the link between pattern and the uncertainty.

Goals of predictive uncertainty quantification with missing values

Goal: predict $Y^{(n+1)}$ with **confidence** $1-\alpha$, i.e. build the smallest $\mathcal{C}_\alpha$ such that:

Goal: predict $Y^{(n+1)}$ with **confidence** $1-\alpha$, i.e. build the smallest $\mathcal{C}_\alpha$ such that:

1. Marginal Validity (MV)

$$\mathbb{P}\left\{Y^{(n+1)} \in \mathcal{C}_\alpha\left(X^{(n+1)}, M^{(n+1)}\right)\right\} \geq 1-\alpha. \tag{MV}$$

For example: $\alpha = 0.1$ and obtain a 90% coverage interval.

Goals of predictive uncertainty quantification with missing values

Goal: predict $Y^{(n+1)}$ with **confidence** $1-\alpha$, i.e. build the smallest $\mathcal{C}_\alpha$ such that:

1. Marginal Validity (MV)

$$\mathbb{P}\left\{Y^{(n+1)} \in \mathcal{C}_\alpha\left(X^{(n+1)}, M^{(n+1)}\right)\right\} \geq 1-\alpha. \qquad \text{(MV)}$$

2. Mask-Conditional-Validity (MCV)

$$\mathbb{P}\left\{Y^{(n+1)} \in \mathcal{C}_\alpha\left(X^{(n+1)}, M^{(n+1)}\right) | M^{(n+1)}\right\} \overset{a.s.}{\geq} 1-\alpha. \qquad \text{(MCV)}$$

Goals of predictive uncertainty quantification with missing values

Goal: predict $Y^{(n+1)}$ with **confidence** $1-\alpha$, i.e. build the smallest $\mathcal{C}_\alpha$ such that:

1. Marginal Validity (MV)

$$\mathbb{P}\left\{Y^{(n+1)} \in \mathcal{C}_\alpha\left(X^{(n+1)}, M^{(n+1)}\right)\right\} \geq 1-\alpha. \tag{MV}$$

2. Mask-Conditional-Validity (MCV)

$$\mathbb{P}\left\{Y^{(n+1)} \in \mathcal{C}_\alpha\left(X^{(n+1)}, M^{(n+1)}\right) | M^{(n+1)}\right\} \overset{a.s.}{\geq} 1-\alpha. \tag{MCV}$$

→ Let us start by considering marginal validity!

Achieving marginal validity through impute then conformalize

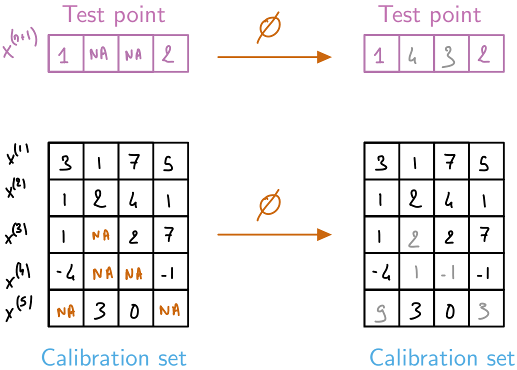
Test point

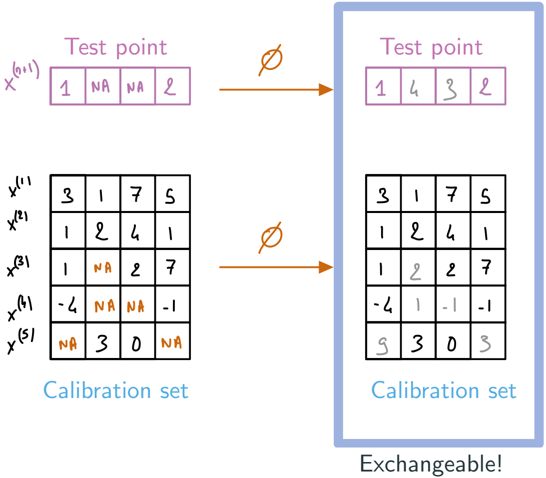
$X^{(n+1)}$	1	NA	NA	2

$X^{(1)}$	3	1	7	5
$X^{(2)}$	1	2	4	1
$X^{(3)}$	1	NA	2	7
$X^{(4)}$	-4	NA	NA	-1
$X^{(5)}$	NA	3	0	NA

Calibration set

Achieving marginal validity through impute then conformalize

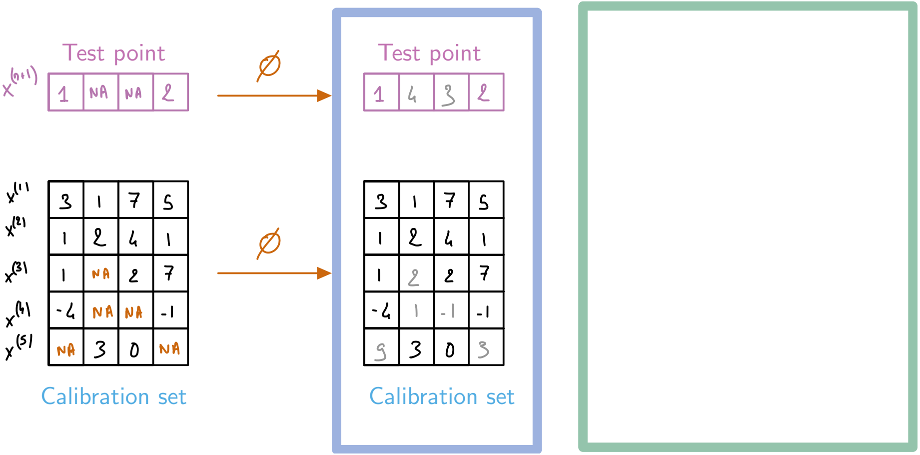




Lemma: Exchangeability after imputation (Zaffran, D., Josse and Romano, 2023)

Assume $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ are i.i.d. or exchangeable. Then, for any missing mechanism, for almost all imputation function ϕ: $\left(\phi(X^{(k)}_{\text{obs}(M^{(k)})}, M^{(k)}), Y^{(k)}\right)_{k=1}^{n}$ are **exchangeable**.

Achieving marginal validity through impute then conformalize



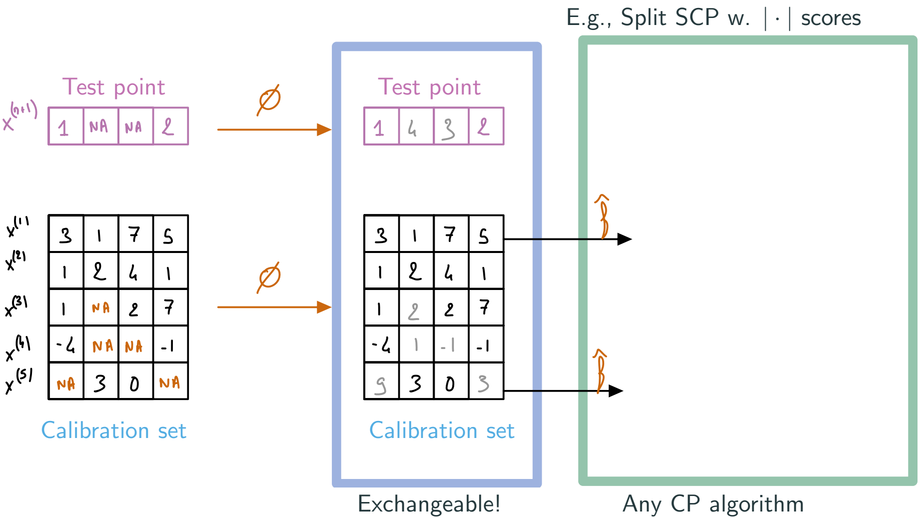
Exchangeable!

Any CP algorithm

Lemma: Exchangeability after imputation (Zaffran, D., Josse and Romano, 2023)

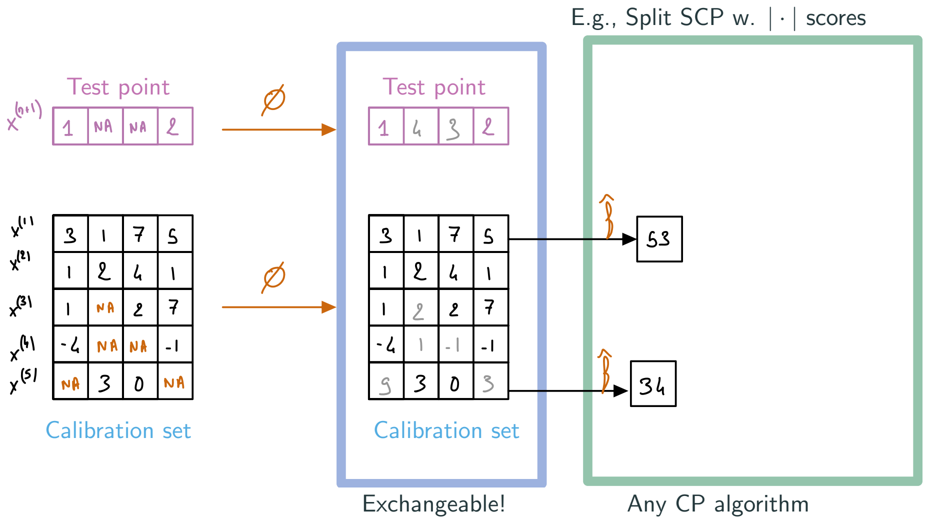
Assume $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ are i.i.d. or exchangeable. Then, for any missing mechanism, for almost all imputation function ϕ: $\left(\phi(X^{(k)}_{\text{obs}(M^{(k)})}, M^{(k)}), Y^{(k)}\right)_{k=1}^{n}$ are **exchangeable**.

Achieving marginal validity through impute then conformalize



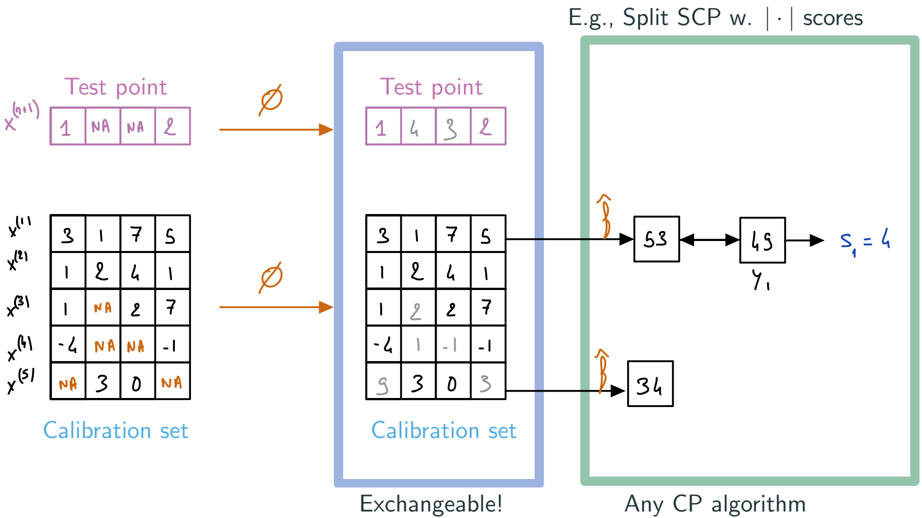
Lemma: Exchangeability after imputation (Zaffran, D., Josse and Romano, 2023)

Assume $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ are i.i.d. or exchangeable. Then, for any missing mechanism, for almost all imputation function ϕ: $\left(\phi(X^{(k)}_{\text{obs}(M^{(k)})}, M^{(k)}), Y^{(k)}\right)_{k=1}^{n}$ are **exchangeable**.



Lemma: Exchangeability after imputation (Zaffran, D., Josse and Romano, 2023)

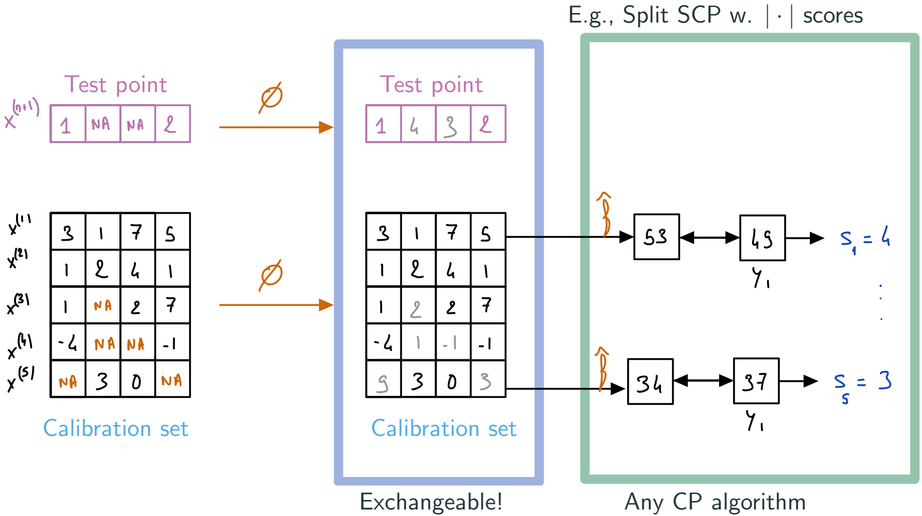
Assume $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ are i.i.d. or exchangeable. Then, for any missing mechanism, for almost all imputation function ϕ: $\left(\phi(X^{(k)}_{\text{obs}(M^{(k)})}, M^{(k)}), Y^{(k)}\right)_{k=1}^{n}$ are **exchangeable**.



Lemma: Exchangeability after imputation (Zaffran, D., Josse and Romano, 2023)

Assume $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ are i.i.d. or exchangeable. Then, for any missing mechanism, for almost all imputation function ϕ: $\left(\phi(X^{(k)}_{\mathrm{obs}(M^{(k)})}, M^{(k)}), Y^{(k)}\right)_{k=1}^{n}$ are **exchangeable**.

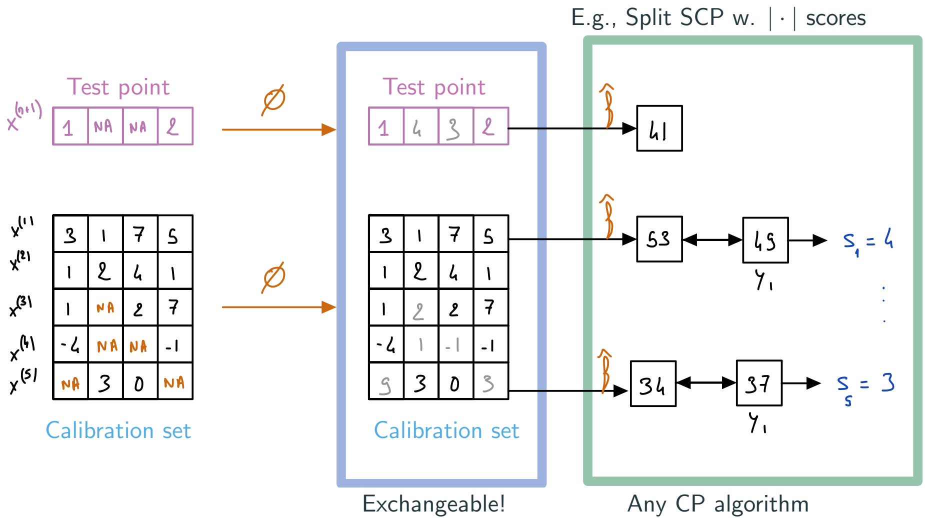
Achieving marginal validity through impute then conformalize



Lemma: Exchangeability after imputation (Zaffran, D., Josse and Romano, 2023)

Assume $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ are i.i.d. or exchangeable. Then, for any missing mechanism, for almost all imputation function ϕ: $\left(\phi(X^{(k)}_{\text{obs}(M^{(k)})}, M^{(k)}), Y^{(k)}\right)_{k=1}^{n}$ are **exchangeable**.

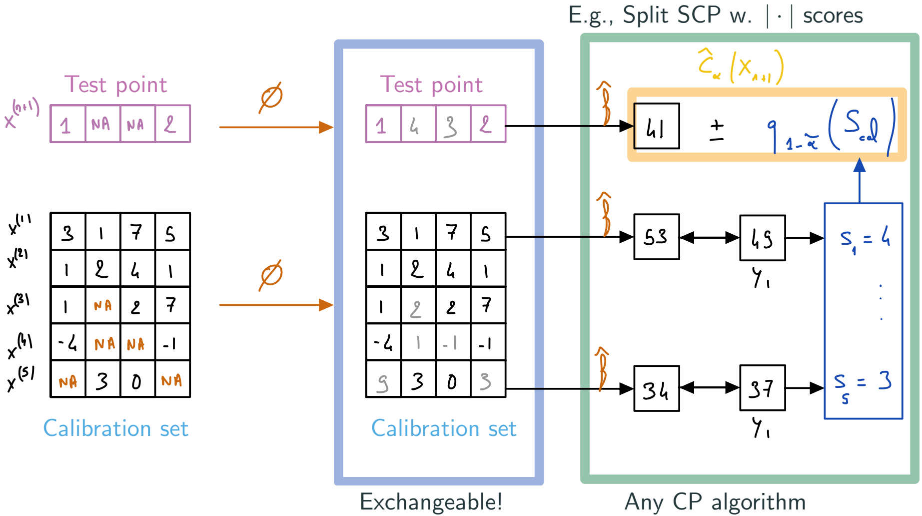
Achieving marginal validity through impute then conformalize



Lemma: Exchangeability after imputation (Zaffran, D., Josse and Romano, 2023)

Assume $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ are i.i.d. or exchangeable. Then, for any missing mechanism, for almost all imputation function ϕ: $\left(\phi(X^{(k)}_{\mathrm{obs}(M^{(k)})}, M^{(k)}), Y^{(k)}\right)_{k=1}^{n}$ are **exchangeable**.

Achieving marginal validity through impute then conformalize



Lemma: Exchangeability after imputation (Zaffran, D., Josse and Romano, 2023)

Assume $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ are i.i.d. or exchangeable. Then, for any missing mechanism, for almost all imputation function ϕ: $\left(\phi(X^{(k)}_{\text{obs}(M^{(k)})}, M^{(k)}), Y^{(k)}\right)_{k=1}^{n}$ are **exchangeable**.

CP is marginally valid (MV) after imputation

Lemma: Exchangeability after imputation (Zaffran, D., Josse and Romano, 2023)

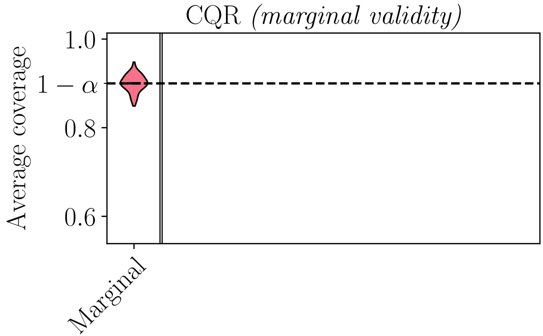
Assume $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n}$ are i.i.d. or exchangeable. Then, for any missing mechanism, for almost all imputation function ϕ: $\left(\phi(X^{(k)}_{\text{obs}(M^{(k)})}, M^{(k)}), Y^{(k)}\right)_{k=1}^{n}$ are **exchangeable**.

$\Rightarrow$ CQR, and Conformal Prediction, applied on an imputed data set still enjoys marginal guarantees

$$\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}, M^{(n+1)}\right)\right\} \geq 1 - \alpha.$$

CQR is marginally valid on imputed data sets

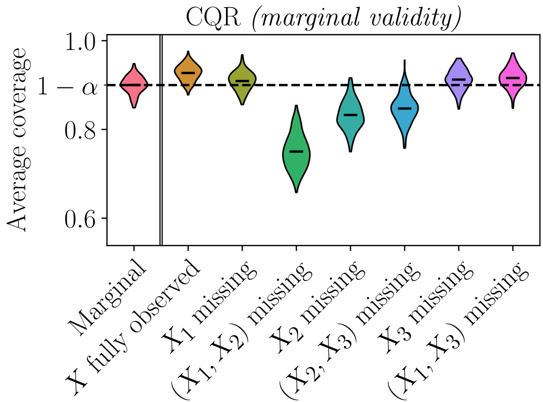
$$Y = \beta^T X + \varepsilon,\ \beta = (1, 2, -1)^T,\ X \text{ and } \varepsilon \text{ Gaussian.}$$



✓ Marginal (i.e. average) coverage (MV) is indeed recovered!

CQR is marginally valid on imputed data sets

$Y = \beta^T X + \varepsilon$, $\beta = (1, 2, -1)^T$, X and ε Gaussian.

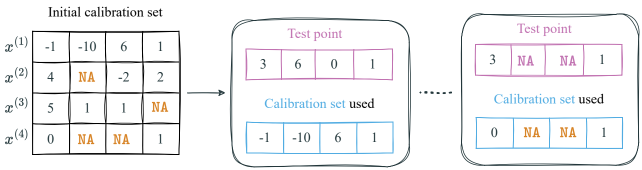


✓ Marginal (i.e. average) coverage (MV) is indeed recovered!

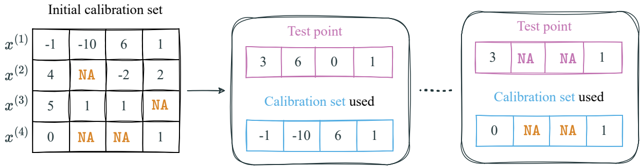
✗ Mask-conditional-validity (MCV) is not attained

$\hookrightarrow$ Missing values induce heteroskedasticity
(supported by theory under (non-)parametric assumptions)

1. Splitting the calibration set by mask is infeasible (lack of data)!



1. Splitting the calibration set by mask is infeasible (lack of data)!



2. Fully distribution-free MCV is necessarily uninformative

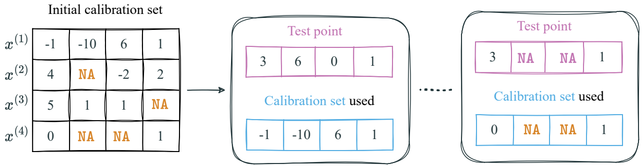
General MCV hardness result (Zaffran, Josse, Romano and D., 2024)[10]

If any $\widehat{C}_\alpha$ is distribution-free MCV then for any distribution P, for any mask m such that $P_M(m) > 0$, it holds:

$$\mathbb{P}_{P^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, m\right)\right) = \infty\right) \underset{\text{if} \quad P_M(m) \ll 1/\sqrt{n}}{\simeq} 1 - \alpha.$$

Challenges and limits

1. Splitting the calibration set by mask is infeasible (lack of data)!



2. Fully distribution-free MCV is necessarily uninformative

General MCV hardness result **(Zaffran, Josse, Romano and D., 2024)**[10]

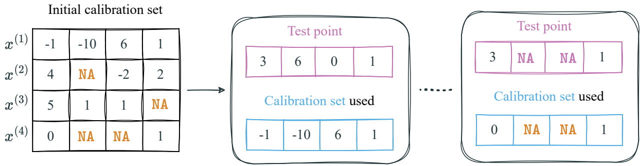
If any $\widehat{C}_\alpha$ is distribution-free MCV then for any distribution P, for any mask m such that $P_M(m) > 0$, it holds:

$$\mathbb{P}_{P^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, m\right)\right) = \infty\right) \underset{\text{if} \quad P_M(m) \ll 1/\sqrt{n}}{\simeq} 1 - \alpha.$$

1. Similar result if we assume $M \perp\!\!\!\perp X$.

Challenges and limits

1. Splitting the calibration set by mask is infeasible (lack of data)!



2. Fully distribution-free MCV is necessarily uninformative

General MCV hardness result (Zaffran, Josse, Romano and D., 2024)[10]

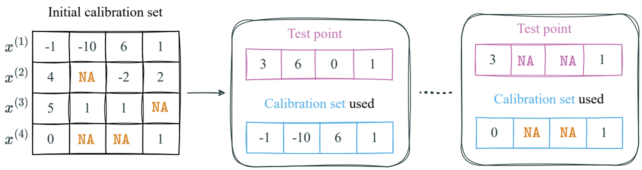
If any $\widehat{C}_\alpha$ is distribution-free MCV then for any distribution P, for any mask m such that $P_M(m) > 0$, it holds:

$$\mathbb{P}_{P^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, m\right)\right) = \infty\right) \underset{\text{if} \quad P_M(m) \ll 1/\sqrt{n}}{\simeq} 1 - \alpha.$$

1. Similar result if we assume $M \perp\!\!\!\perp X$.
2. Similar result if we assume $Y \perp\!\!\!\perp M \mid X$

Challenges and limits

1. Splitting the calibration set by mask is infeasible (lack of data)!



2. Fully distribution-free MCV is necessarily uninformative

General MCV hardness result (Zaffran, Josse, Romano and D., 2024)[10]

If any $\widehat{C}_\alpha$ is distribution-free MCV then for any distribution P, for any mask m such that $P_M(m) > 0$, it holds:

$$\mathbb{P}_{P^{\otimes(n+1)}}\left(\text{mes}\left(\widehat{C}_\alpha\left(X_{n+1}, m\right)\right) = \infty\right) \underset{\text{if} \quad P_M(m) \ll 1/\sqrt{n}}{\simeq} 1 - \alpha.$$

1. Similar result if we assume $M \perp\!\!\!\perp X$.
2. Similar result if we assume $Y \perp\!\!\!\perp M \,|X$

$\Rightarrow$ need to restrict both the link between M and X, as well as between M and Y.

Idea: for each test point, modify the calibration points to mimic the test mask

$\mapsto$ Solution 1: **Missing data augmentation**.

Test point

$x^{(n+1)}$	1	NA	NA	2

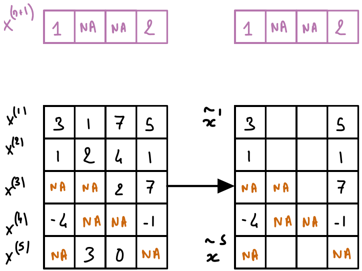
$x^{(1)}$	3	1	7	5
$x^{(2)}$	1	2	4	1
$x^{(3)}$	NA	NA	2	7
$x^{(4)}$	-4	NA	NA	-1
$x^{(5)}$	NA	3	0	NA

Calibration set

Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 1: **Missing data augmentation**.

Test point

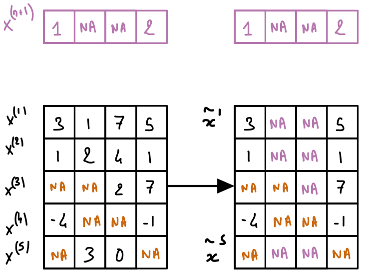


Calibration set

Idea: for each test point, modify the calibration points to mimic the test mask

$\mapsto$ Solution 1: **Missing data augmentation**.

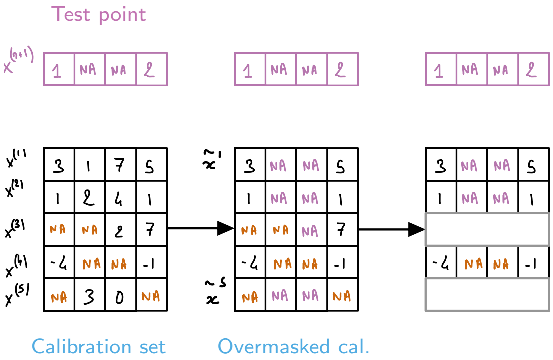
Test point



Calibration set Overmasked cal.

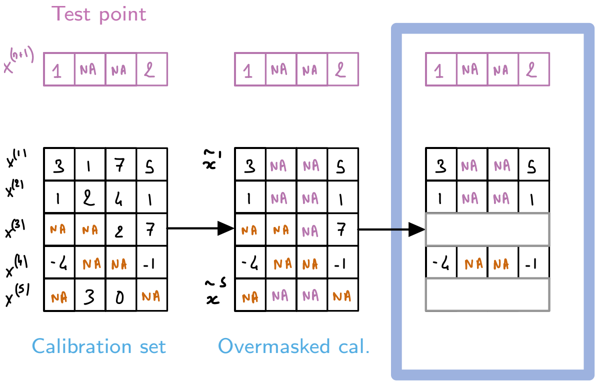
Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 1: **Missing data augmentation**.



Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 1: **Missing data augmentation**.



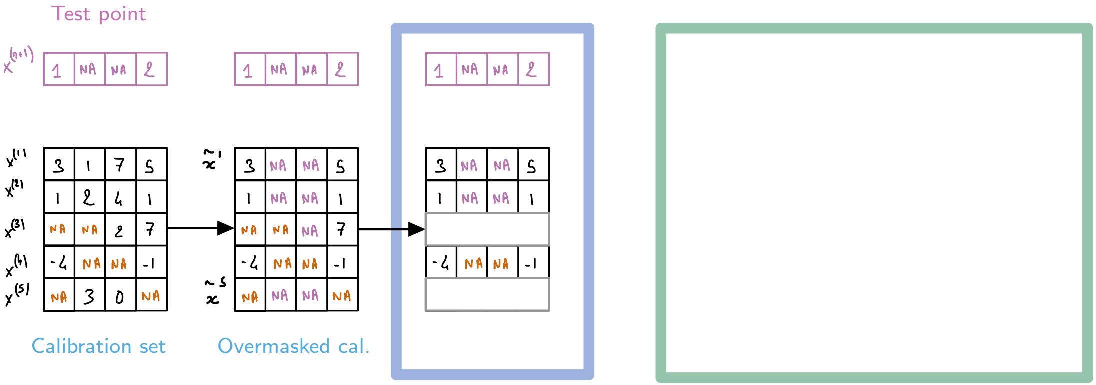
Exchangeable!

Marginal distribution

$= \text{dist of } (X^{(n+1)}, M^{(n+1)}, Y^{n+1}) | M^{n+1}$

Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 1: **Missing data augmentation**.

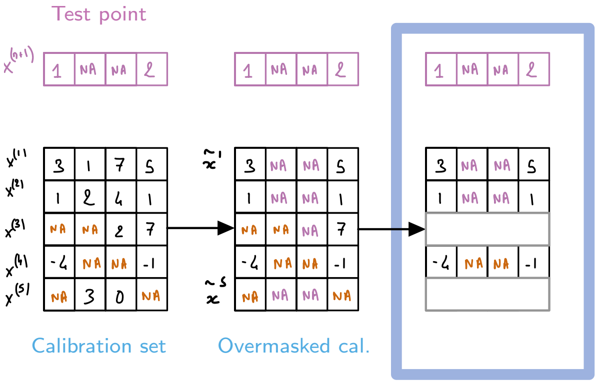


Exchangeable!
Marginal distribution
$= \text{dist of } (X^{(n+1)}, M^{(n+1)}, Y^{n+1}) | M^{n+1}$

→ Any CP algorithm ensures MCV

Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 1: **Missing data augmentation**.



$\mathbb{P}\left\{Y^{(n+1)} \notin \mathcal{C}_\alpha\left(X^{(n+1)}, m\right) | M^{(n+1)}\right\} \overset{a.s.}{\leq} \alpha.$

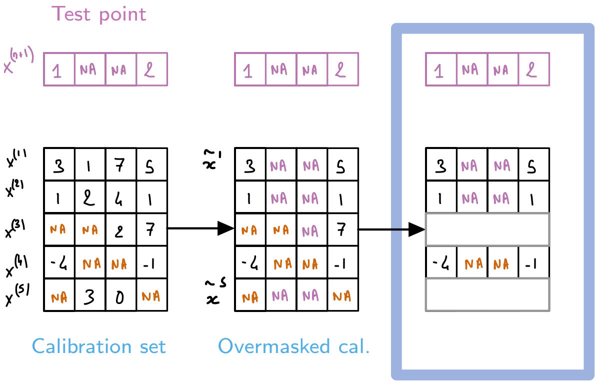
Exchangeable!
Marginal distribution
$= \text{dist of } (X^{(n+1)}, M^{(n+1)}, Y^{n+1}) | M^{n+1}$

→ Any CP algorithm ensures MCV

CP-MDA-Nested* (Missing Data Augmentation)

Idea: for each test point, modify the calibration points to mimic the test mask

$\mapsto$ Solution 1: **Missing data augmentation**.



$$\mathbb{P}\left\{Y^{(n+1)} \notin \mathcal{C}_\alpha\left(X^{(n+1)}, m\right) \middle| M^{(n+1)}\right\} \overset{a.s.}{\leq} \alpha.$$

Exchangeable!
Marginal distribution
$= \text{dist of } (X^{(n+1)}, M^{(n+1)}, Y^{n+1}) | M^{n+1}$

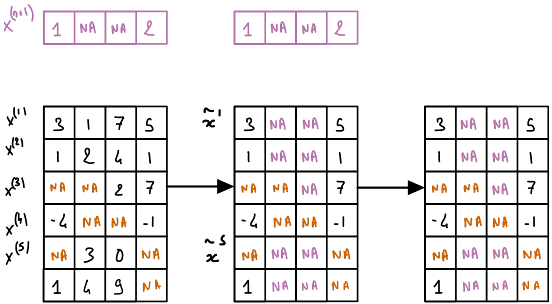
$\rightarrow$ Any CP algorithm ensures MCV

Problem: still far too few points for patterns with few missing data.

Idea: for each test point, modify the calibration points to mimic the test mask

$\mapsto$ Solution 2: **Missing data augmentation + keeping more points**.

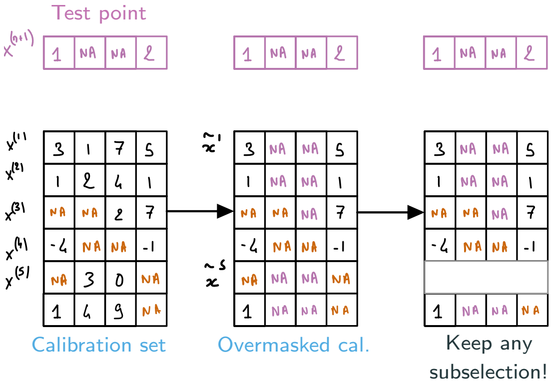
Test point



Calibration set Overmasked cal. Keep any subselection!

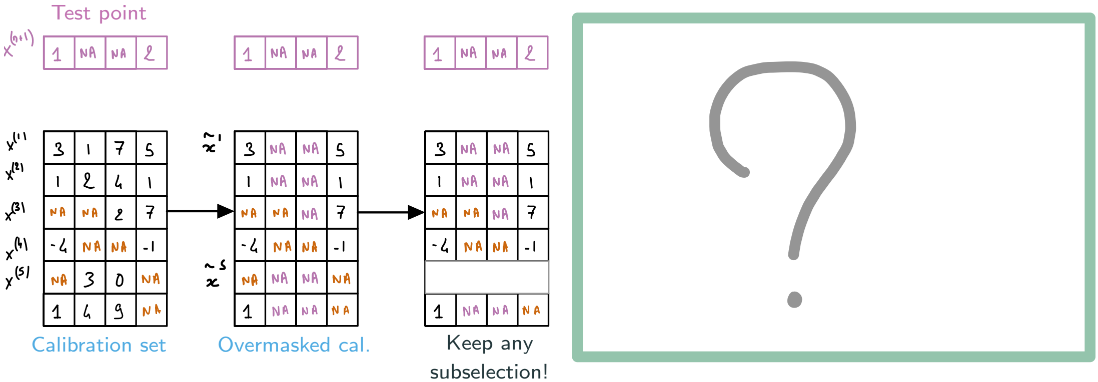
Idea: for each test point, modify the calibration points to mimic the test mask

$\mapsto$ Solution 2: **Missing data augmentation + keeping more points**.



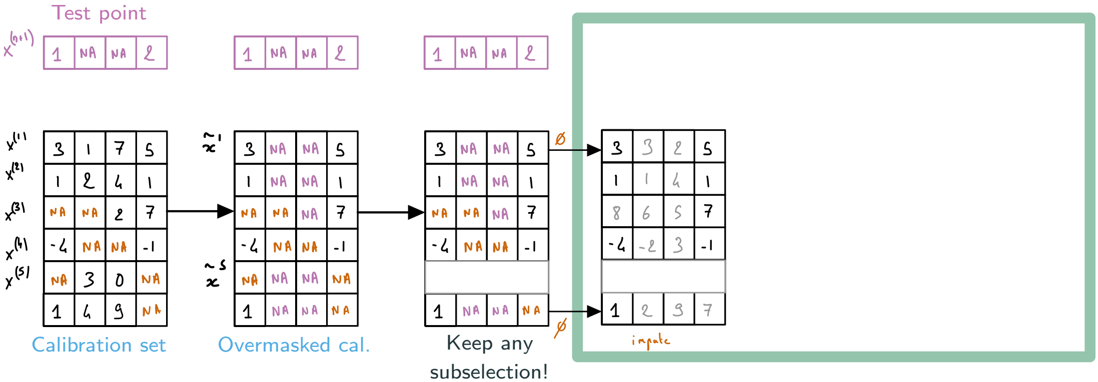
Idea: for each test point, modify the calibration points to mimic the test mask

$\mapsto$ Solution 2: **Missing data augmentation + keeping more points**.



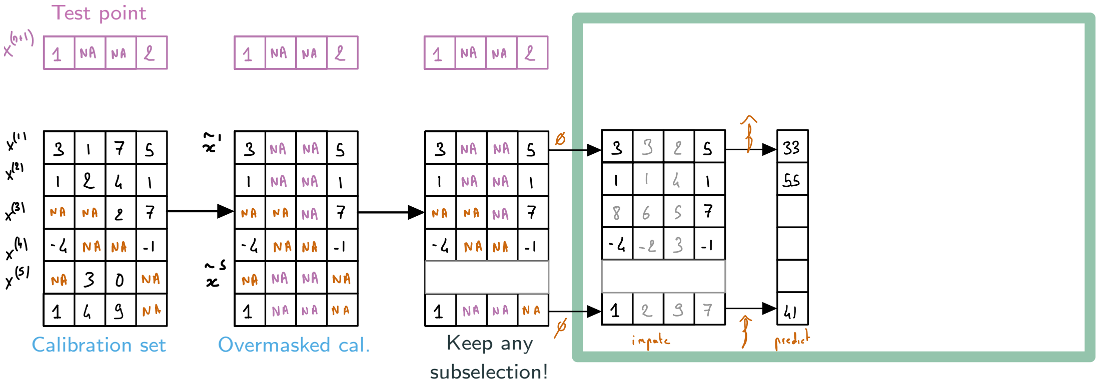
Idea: for each test point, modify the calibration points to mimic the test mask

$\mapsto$ Solution 2: **Missing data augmentation + keeping more points**.



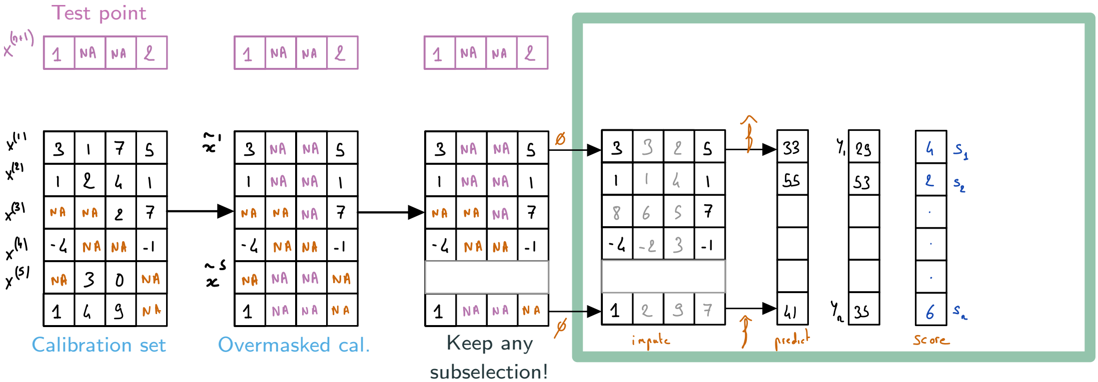
Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 2: **Missing data augmentation + keeping more points**.



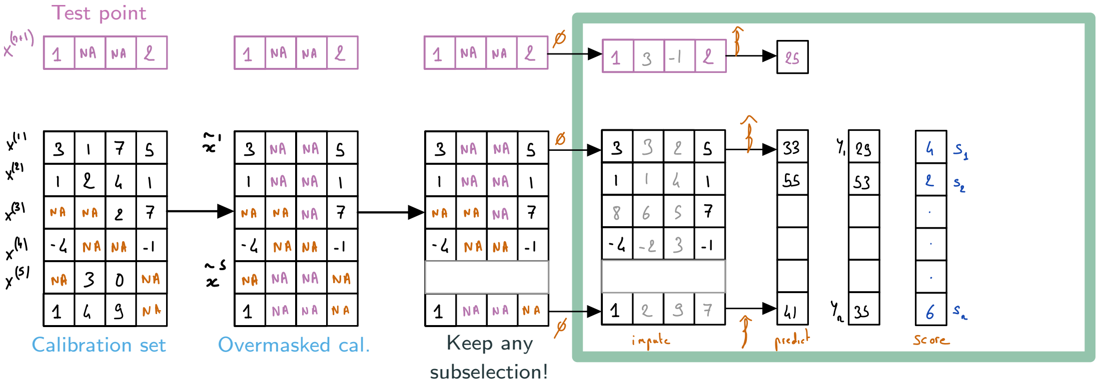
Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 2: **Missing data augmentation + keeping more points**.



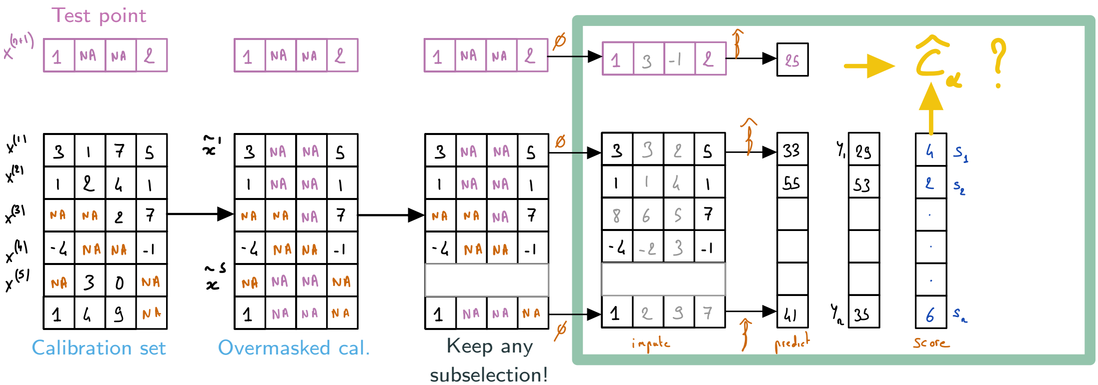
Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 2: **Missing data augmentation + keeping more points**.



Idea: for each test point, modify the calibration points to mimic the test mask

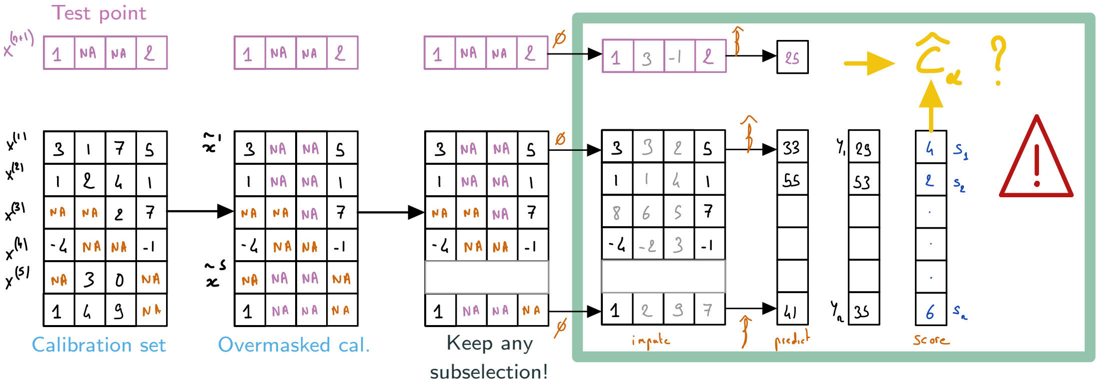
↦ Solution 2: **Missing data augmentation + keeping more points**.



CP-MDA-Nested* (Missing Data Augmentation)

Idea: for each test point, modify the calibration points to mimic the test mask

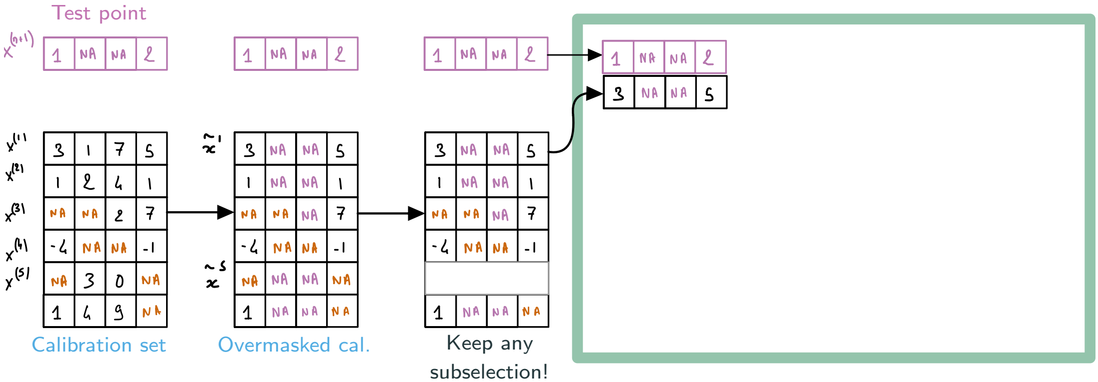
↦ Solution 2: **Missing data augmentation + keeping more points**.



Problem: scores cannot simply be aggregated and used, no guarantee on that!

Idea: for each test point, modify the calibration points to mimic the test mask

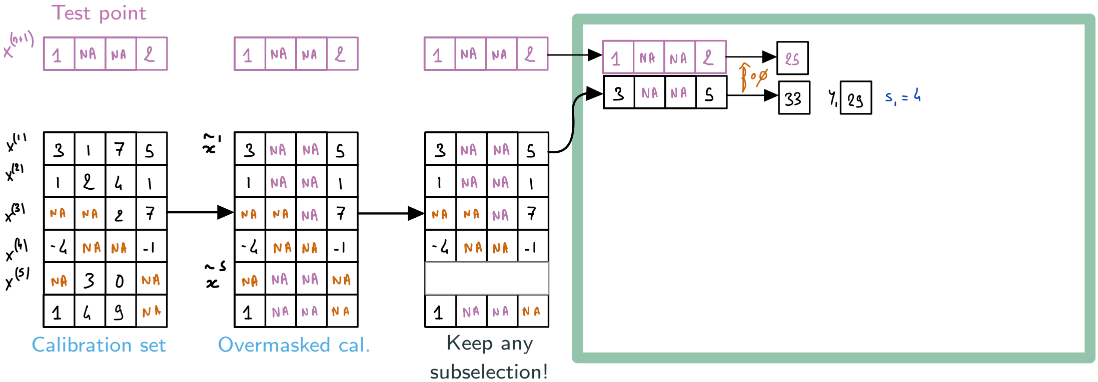
$\mapsto$ Solution 2: **Missing data augmentation + keeping more points**.



Combine *identical* missing patterns

Idea: for each test point, modify the calibration points to mimic the test mask

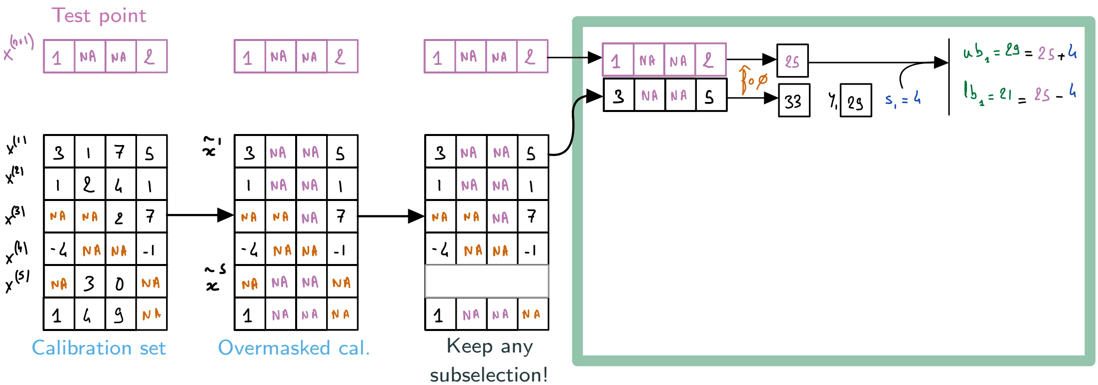
↦ Solution 2: **Missing data augmentation + keeping more points**.



Combine *identical* missing patterns

Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 2: **Missing data augmentation + keeping more points**.

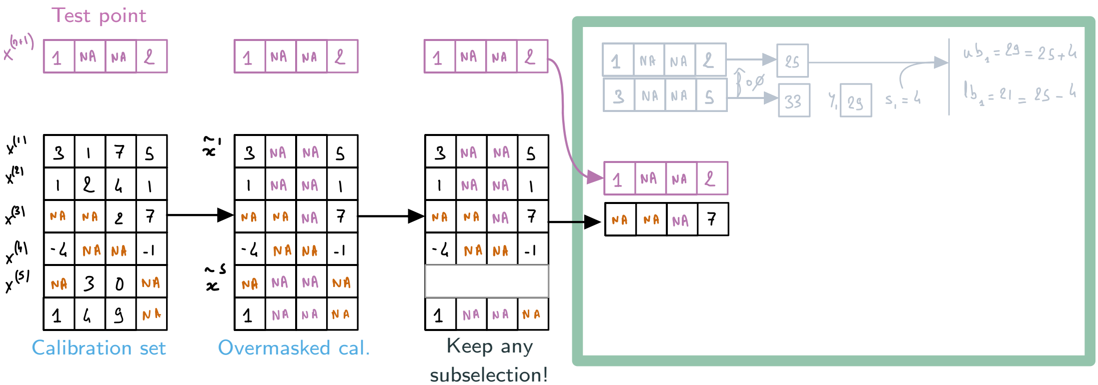


Combine *identical* missing patterns

CP-MDA-Nested* (Missing Data Augmentation)

Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 2: **Missing data augmentation + keeping more points**.

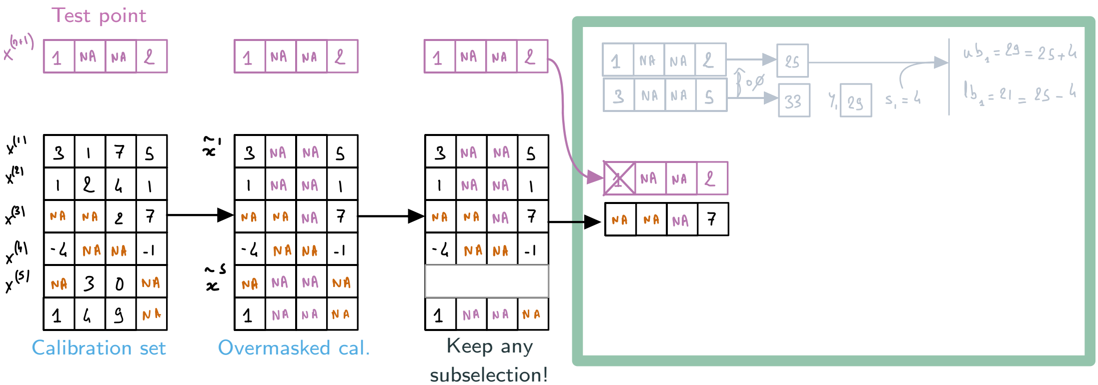


Combine *identical* missing patterns

CP-MDA-Nested* (Missing Data Augmentation)

Idea: for each test point, modify the calibration points to mimic the test mask

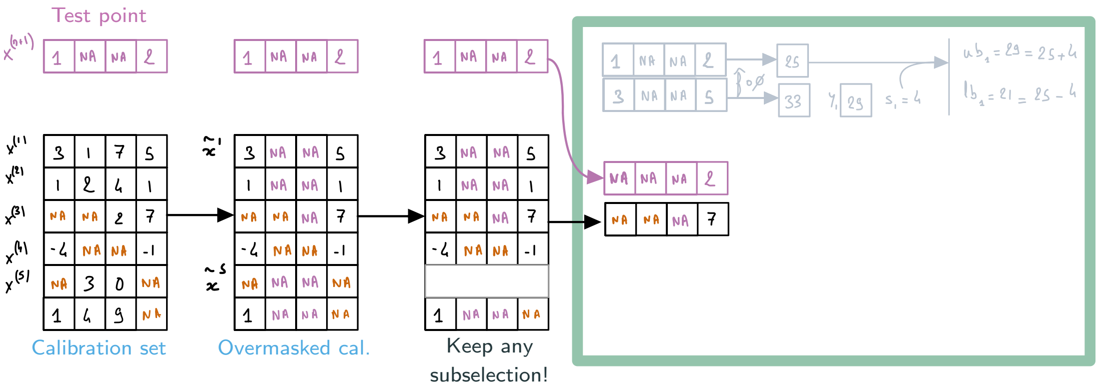
$\mapsto$ Solution 2: **Missing data augmentation + keeping more points**.



Combine *identical* missing patterns

Idea: for each test point, modify the calibration points to mimic the test mask

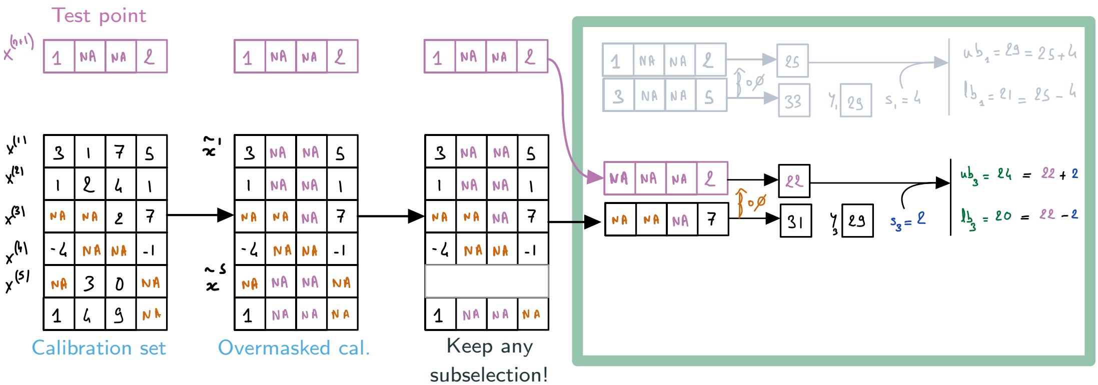
↦ Solution 2: **Missing data augmentation + keeping more points**.



Combine *identical* missing patterns

Idea: for each test point, modify the calibration points to mimic the test mask

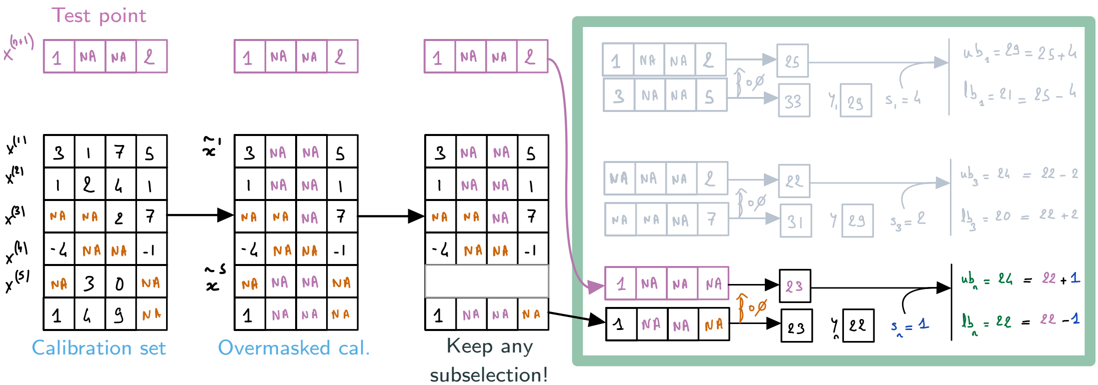
↦ Solution 2: **Missing data augmentation + keeping more points**.



Combine *identical* missing patterns

Idea: for each test point, modify the calibration points to mimic the test mask

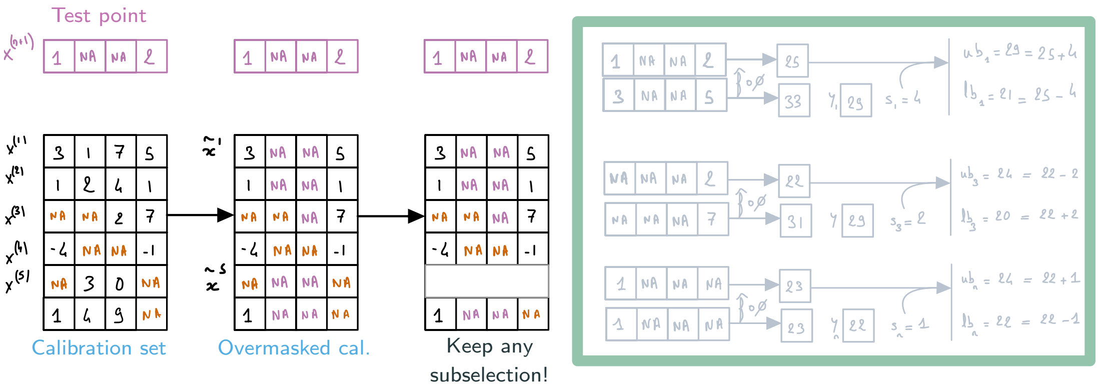
↦ Solution 2: **Missing data augmentation + keeping more points**.



Combine *identical* missing patterns

Idea: for each test point, modify the calibration points to mimic the test mask

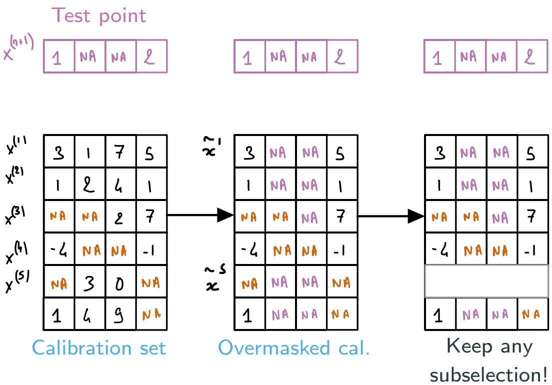
$\mapsto$ Solution 2: **Missing data augmentation + keeping more points**.

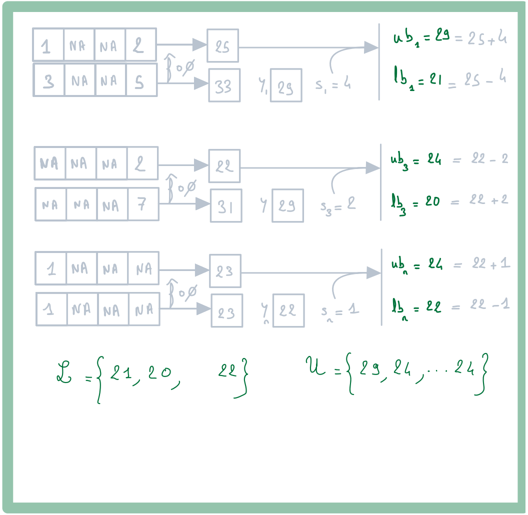


Combine *identical* missing patterns

Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 2: **Missing data augmentation + keeping more points**.



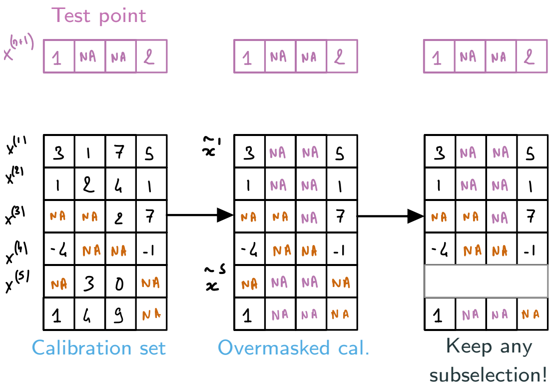


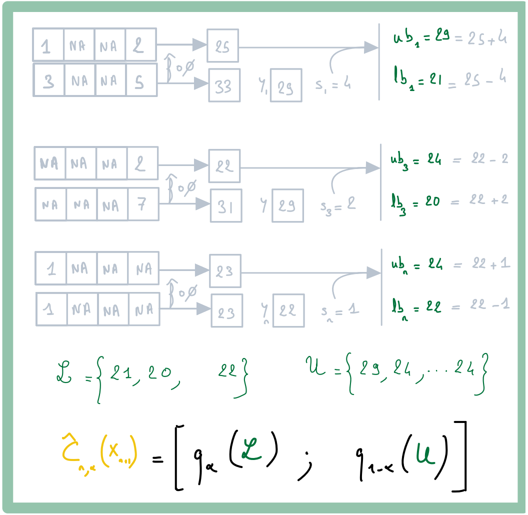
Combine *identical* missing patterns

CP-MDA-Nested* (Missing Data Augmentation)

Idea: for each test point, modify the calibration points to mimic the test mask

↦ Solution 2: **Missing data augmentation + keeping more points**.





Combine *identical* missing patterns

Mask-conditional-validity of $\texttt{CP-MDA-Nested}^\star$
(Zaffran, Josse, Romano and D., 2024)

Under the assumptions that:

- $M \perp\!\!\!\perp X, Y$,
- $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ are i.i.d.,
- the subsampling scheme is independent of $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$,

then, for almost all imputation function, $\texttt{CP-MDA-Nested}^\star$ reaches (MCV) at the level $1-2\alpha$, that is:

$$\mathbb{P}\left\{Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}, m\right) | M^{(n+1)}\right\} \overset{a.s.}{\geq} 1-2\alpha.$$

Mask-conditional-validity of CP-MDA-Nested*
(Zaffran, Josse, Romano and D., 2024)

Under the assumptions that:

- $M \perp\!\!\!\perp X, Y$,
- $\left(X^{(k)}, M^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$ are i.i.d.,
- the subsampling scheme is independent of $\left(X^{(k)}, Y^{(k)}\right)_{k=1}^{n+1}$,

then, for almost all imputation function, CP-MDA-Nested* reaches (MCV) at the level $1 - 2\alpha$, that is:

$$\mathbb{P}\left\{ Y^{(n+1)} \in \widehat{C}_\alpha\left(X^{(n+1)}, m\right) | M^{(n+1)} \right\} \overset{a.s.}{\geq} 1 - 2\alpha.$$

Proof relates CP-MDA-Nested* to JK+ type algorithms.

Non-absolute value scores and classification are also encompassed.

Link to Jackknife+ (Barber et al., 2021b)

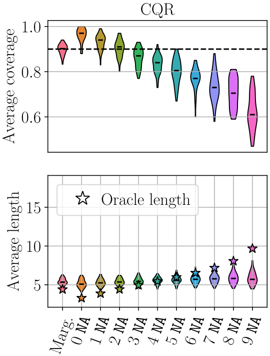
Link to JK+

1. When training on $\mathcal{D}_n^{-i}$, we sample a mask $M^{(i)}$, and the trained predictor is

$$\mathcal{A}(\mathcal{D}_n^{-i}) = \hat{f} \circ \Phi(X \odot M^{(i)})$$

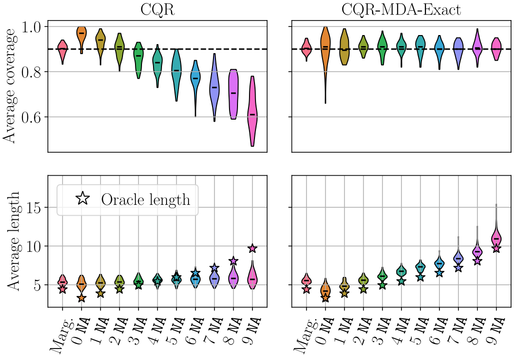
2. The randomness of the training is coupled with the randomness of the Masks in the calibration dataset.
3. We directly recognize an instance of JK+.

Experiments on $M \perp\!\!\!\perp (X, Y)$ Gaussian linear data in dimension 10



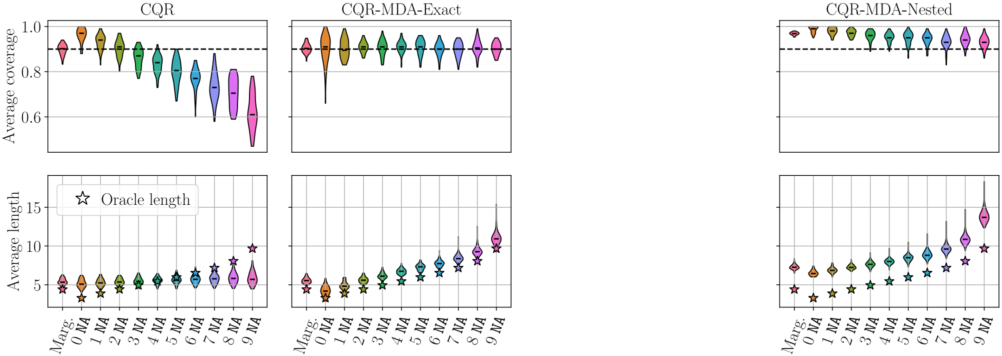
20% of missing values

Experiments on $M \perp\!\!\!\perp (X, Y)$ Gaussian linear data in dimension 10



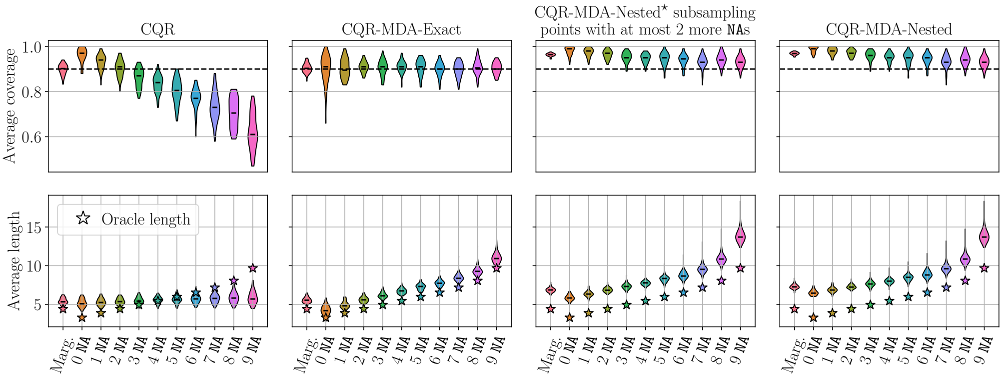
20% of missing values

Experiments on $M \perp\!\!\!\perp (X, Y)$ Gaussian linear data in dimension 10



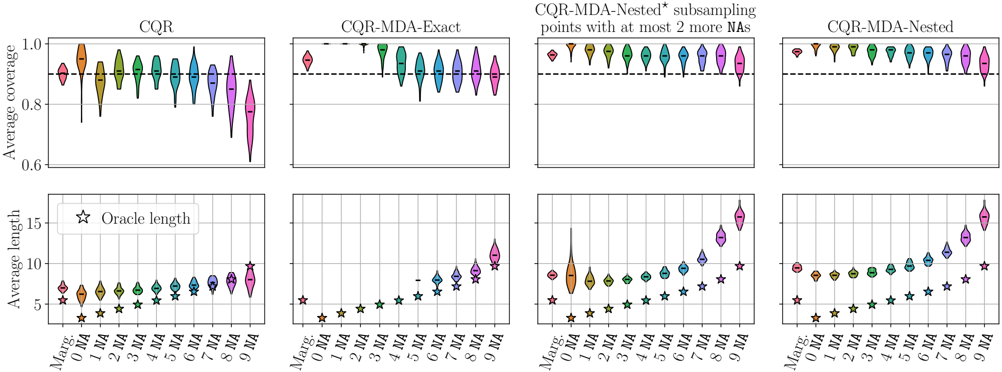
20% of missing values

Experiments on $M \perp\!\!\!\perp (X, Y)$ Gaussian linear data in dimension 10

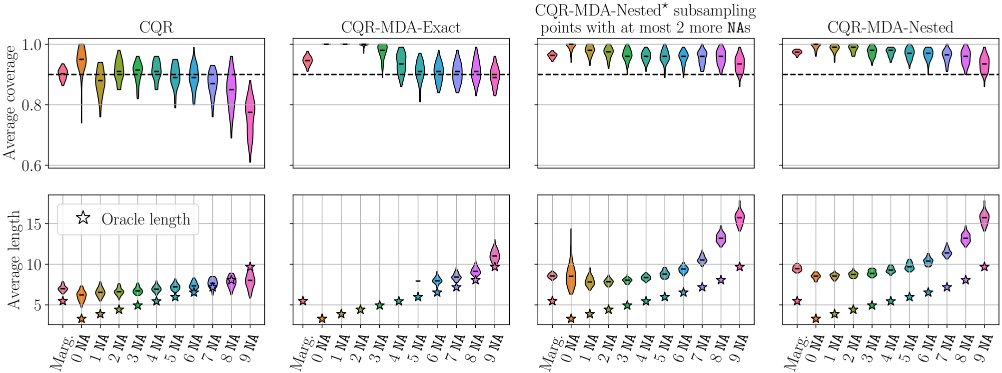


20% of missing values

Experiments on $M \perp\!\!\!\perp (X, Y)$ Gaussian linear data in dimension 10



40% of missing values



40% of missing values

Experiments beyond independence

→ Under various MAR and MNAR mechanisms, `CP-MDA-Nested`* maintains empirical MCV;

Take-home-messages : Conformal prediction and UQ with missing values

- CP marginal guarantees hold on the imputed data set.
- Missingness introduces additional heteroskedasticity and is a fun case to study shifts.
- CQR (and more generally CP) fails to attain coverage conditional on the missing pattern, i.e. MCV.
- MCV is impossible to ensure in an informative way without restricting both the dependence between M and X, and between M and Y.
- `CP-MDA-Nested`* (Missing Data Augmentation) is the first method to output predictive intervals with missing values.
- `CP-MDA-Nested`* attains conditional coverage with respect to the missing pattern (in MCAR and $Y \perp\!\!\!\perp M \,|X$ setting).
- `CP-MDA-Nested`* is empirically robust to non-MCAR scenarii.

Applications & Methods II: Some methodological advances

Conformal prediction and UQ with missing values

Valid Selection among Conformal Sets

Valid Selection among Conformal Sets



Mahmoud Hegazy
École Polytechnique

Liviu Aolaritei
UC Berkeley

Mickael I Jordan
UC Berkeley
INRIA Paris

Valid Selection among Conformal Sets, 2025, MH, LA, MJ, AD

Summary:

- **Question:** When multiple valid sets exist, to which extent can we select the smallest (most informative) one ?
  $\rightarrow$ Selection after observation of length can break the coverage guarantee.

Summary:

- **Question:** When multiple valid sets exist, to which extent can we select the smallest (most informative) one ?
  $\rightarrow$ Selection after observation of length can break the coverage guarantee.

Remark: line of work by Liang et al. (2024); Yang and Kuchibhotla (2024), that try to select the smallest set marginally on X, as well as Gasparin and Ramdas (2024a) in an online framework.

$\rightarrow$ We aim for pointwise selection.

$\rightarrow$ we show how to combine approaches

Summary:

- **Question:** When multiple valid sets exist, to which extent can we select the smallest (most informative) one ?
  $\rightarrow$ Selection after observation of length can break the coverage guarantee.

Remark: line of work by Liang et al. (2024); Yang and Kuchibhotla (2024), that try to select the smallest set marginally on X, as well as Gasparin and Ramdas (2024a) in an online framework.

$\rightarrow$ We aim for pointwise selection.

$\rightarrow$ we show how to combine approaches

Our Approach:

1. Use algorithmic stability in a randomized selection framework and adjust the coverage guarantee
   or
2. Define a meta-score incorporating the selection. [opt.]

Selection among Conformal Sets: Selecting the Smallest Prediction Set

- Given K prediction sets $\{C_i^\alpha(X)\}_{i=1}^K$, the goal is to select the smallest one.

Selection among Conformal Sets: Selecting the Smallest Prediction Set

- Given K prediction sets $\{C_i^\alpha(X)\}_{i=1}^K$, the goal is to select the smallest one.
- **Issue:** Naively selecting the smallest set can invalidate the coverage guarantee.

Selection among Conformal Sets: Selecting the Smallest Prediction Set

- Given K prediction sets $\{C_i^\alpha(X)\}_{i=1}^K$, the goal is to select the smallest one.
- **Issue:** Naively selecting the smallest set can invalidate the coverage guarantee.
- Let $i^*(X) = \mathrm{argmin}_i \lambda\left(C_i^\alpha(X)\right)$.

$$\mathbb{P}\left\{Y \in C_{i^*(X)}^\alpha(X)\right\} \ngeq 1 - \alpha$$

Selection among Conformal Sets: Selecting the Smallest Prediction Set

- Given K prediction sets $\{C_i^\alpha(X)\}_{i=1}^K$, the goal is to select the smallest one.
- **Issue:** Naively selecting the smallest set can invalidate the coverage guarantee.
- Let $i^*(X) = \mathrm{argmin}_i \lambda\left(C_i^\alpha(X)\right)$.

$$\mathbb{P}\left\{Y \in C^\alpha_{i^*(X)}(X)\right\} \ngeq 1 - \alpha$$

Obviously similar to *multiple tests*, or *inference after selection*, etc.

Definition: (Indistinguishability)

A r.v. S is η, τ indistinguishable from S_0, denoted $S \approx_{\eta,\tau} S_0$ if for all measurable $\mathcal{O}$,

$$\mathbb{P}\{S \in \mathcal{O}\} \le e^{\eta}\, \mathbb{P}\{S_0 \in \mathcal{O}\} + \tau.$$

Definition: (Indistinguishability)

A r.v. S is η, τ indistinguishable from S_0, denoted $S \approx_{\eta,\tau} S_0$ if for all measurable $\mathcal{O}$,

$$\mathbb{P}\{S \in \mathcal{O}\} \leq e^{\eta}\, \mathbb{P}\{S_0 \in \mathcal{O}\} + \tau.$$

- Intuitively: S and S_0 are "close" in distribution.

Definition: (Indistinguishability)

A r.v. S is η, τ indistinguishable from S_0, denoted $S \approx_{\eta,\tau} S_0$ if for all measurable $\mathcal{O}$,

$$\mathbb{P}\{S \in \mathcal{O}\} \le e^{\eta}\, \mathbb{P}\{S_0 \in \mathcal{O}\} + \tau.$$

- Intuitively: S and S_0 are "close" in distribution.
- Smaller η: greater similarity; τ: small additive difference.

Definition: (Indistinguishability)

A r.v. S is η, τ indistinguishable from S_0, denoted $S \approx_{\eta,\tau} S_0$ if for all measurable $\mathcal{O}$,

$$\mathbb{P}\{S \in \mathcal{O}\} \leq e^{\eta}\, \mathbb{P}\{S_0 \in \mathcal{O}\} + \tau.$$

- Intuitively: S and S_0 are "close" in distribution.
- Smaller η: greater similarity; τ: small additive difference.
- Conditional notation: $S \approx_{\eta,\tau}^{|\xi} S_0$ to hold conditionally to ξ.

We are ready to define the stability of a randomized algorithm,

$$\hat{S} : \Xi \times \mathcal{E} \to \mathcal{S}$$

(Zrnic and Jordan, 2023; Bassily et al., 2016; Bassily and Freund, 2016)[11].

- $\xi \in \Xi$: input data (e.g., sizes of conformal sets).
- $\mathcal{E}$: algorithm's internal randomness.

Algorithmic Stability (specific $\nu = 0$ case)

$\hat{S} : \Xi \times \mathcal{E} \to \mathcal{S}$ is $(\eta, \tau, \nu = 0)$-stable if $\exists S_0$

$$\hat{S}(\xi, \epsilon) \approx^{|\xi}_{\eta,\tau} S_0 \quad a.s.$$

$\to$ Quantifies how close the algorithm's output is from a fixed distribution for most (here all) inputs.

[11] Many similar definitions in many subfields of ML

From Stability to Valid Conformal Coverage

From Stability to Valid Conformal Coverage

Theorem: (Valid Stable Selection)

Assume each $C_i^\alpha(X)$ satisfies the coverage guarantee:

$$\forall i, \quad \mathbb{P}\{Y \notin C_i^\alpha(X)\} \leq \alpha.$$

If $\hat{S} : \Xi \times \mathcal{E} \to \mathcal{S}$ is (η, τ)-stable, then

$$\mathbb{P}\{Y \in \mathrm{C}^\alpha_{\hat{S}(\xi,\epsilon)}(X)\} \geq 1 - \alpha e^\eta - \tau.$$

- Standard $1 - \alpha$ coverage: adjust individual sets to $1 - (\alpha - \tau)e^{-\eta}$.

From Stability to Valid Conformal Coverage

Theorem: (Valid Stable Selection)

Assume each $C_i^\alpha(X)$ satisfies the coverage guarantee:

$$\forall i, \quad \mathbb{P}\{Y \notin C_i^\alpha(X)\} \leq \alpha.$$

If $\hat{S} : \Xi \times \mathcal{E} \to \mathcal{S}$ is (η, τ)-stable, then

$$\mathbb{P}\{Y \in \mathrm{C}^\alpha_{\hat{S}(\xi,\epsilon)}(X)\} \geq 1 - \alpha e^\eta - \tau.$$

- Standard $1-\alpha$ coverage: adjust individual sets to $1-(\alpha-\tau)e^{-\eta}$.
- Stable selection: favours smaller sets, controls coverage.

From Stability to Valid Conformal Coverage

Theorem: (Valid Stable Selection)

Assume each $C_i^\alpha(X)$ satisfies the coverage guarantee:

$$\forall i, \quad \mathbb{P}\{Y \notin C_i^\alpha(X)\} \leq \alpha.$$

If $\hat{S} : \Xi \times \mathcal{E} \to \mathcal{S}$ is (η, τ)-stable, then

$$\mathbb{P}\{Y \in \mathrm{C}^{\alpha}_{\hat{S}(\xi,\epsilon)}(X)\} \geq 1 - \alpha e^{\eta} - \tau.$$

- Standard $1 - \alpha$ coverage: adjust individual sets to $1 - (\alpha - \tau)e^{-\eta}$.
- Stable selection: favours smaller sets, controls coverage.
- Requires to increase set size

From Stability to Valid Conformal Coverage

Theorem: (Valid Stable Selection)

Assume each $C_i^\alpha(X)$ satisfies the coverage guarantee:

$$\forall i, \quad \mathbb{P}\{Y \notin C_i^\alpha(X)\} \leq \alpha.$$

If $\hat{S} : \Xi \times \mathcal{E} \to \mathcal{S}$ is (η, τ)-stable, then

$$\mathbb{P}\{Y \in \mathrm{C}^\alpha_{\hat{S}(\xi,\epsilon)}(X)\} \geq 1 - \alpha e^\eta - \tau.$$

- Standard $1 - \alpha$ coverage: adjust individual sets to $1 - (\alpha - \tau)e^{-\eta}$.
- Stable selection: favours smaller sets, controls coverage.
- Requires to increase set size
- Can be shown to be tight in the worst case.

From Stability to Valid Conformal Coverage

Theorem: (Valid Stable Selection)

Assume each $C_i^\alpha(X)$ satisfies the coverage guarantee:

$$\forall i, \quad \mathbb{P}\{Y \notin C_i^\alpha(X)\} \leq \alpha.$$

If $\hat{S} : \Xi \times \mathcal{E} \to \mathcal{S}$ is (η, τ)-stable, then

$$\mathbb{P}\{Y \in \mathrm{C}^\alpha_{\hat{S}(\xi,\epsilon)}(X)\} \geq 1 - \alpha e^\eta - \tau.$$

- Standard $1-\alpha$ coverage: adjust individual sets to $1-(\alpha-\tau)e^{-\eta}$.
- Stable selection: favours smaller sets, controls coverage.
- Requires to increase set size
- Can be shown to be tight in the worst case.

$\rightarrow$ Proposed Stable rule?

Algorithmic Stability: Stable Minimum Selection

Let
$$\xi(X) = \left(\lambda\left(C_1^\alpha(X)\right), \ldots, \lambda\left(C_K^\alpha(X)\right)\right).$$

$\rightarrow$ Minimum Stable Expectation (MinSE) mechanism achieves **stability**.

Idea: pick a prior $b \in \Delta^{K-1}$ (prior knowledge on set choice). Then, minimizes expected size while ensuring (η, τ)-stability. This turns out to be a linear program

Let

$$\xi(X) = \left(\lambda\left(C_1^\alpha(X)\right), \ldots, \lambda\left(C_K^\alpha(X)\right)\right).$$

→ Minimum Stable Expectation (MinSE) mechanism achieves **stability**.

Idea: pick a prior $b \in \Delta^{K-1}$ (prior knowledge on set choice). Then, minimizes expected size while ensuring (η, τ)-stability. This turns out to be a linear program

MinSE Mechanism

Let

$$p^\star(b, \xi) = \mathrm{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^\alpha(X))$$

$$\text{s.t.} p \in \Delta^{K-1}, s \in \mathrm{R}_+^K, \quad p_i \leq e^\eta b_i + s_i, \quad \sum_{i \in [K]} s_i \leq \tau$$

Selection rule such that

- $\mathbb{P}\left\{\hat{S}(\xi, \epsilon) = i | \xi\right\} = p^\star(b, \xi)_i$
- is η, τ-stable.

Limit cases of MinSE Mechanism (1)

1. $b = b_0 = (1/K, \dots 1/K)$ uniform prior and $e^\eta = K, \tau = 0$:

$$
\begin{aligned}
p^\star(b_0, \xi) =& \operatorname{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^\alpha(X)) \\
& \text{s.t.} p \in \Delta^{K-1}, \quad p_i \leq e^\eta b_i = 1 \\
=& e_{\operatorname{argmin}_i \lambda(C_i^\alpha(X))}
\end{aligned}
$$

1. $b = b_0 = (1/K, \dots 1/K)$ uniform prior and $\boxed{e^\eta = K, \tau = 0}$:

$$\begin{aligned} p^\star(b_0, \xi) =& \operatorname{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^\alpha(X)) \\ & \text{s.t.} p \in \Delta^{K-1}, \quad p_i \le e^\eta b_i = 1 \\ =& e_{\operatorname{argmin}_i \lambda(C_i^\alpha(X))} \end{aligned}$$

Thus $\hat{S}(\xi, \epsilon) = \operatorname{argmin}_i \lambda(C_i^\alpha(X))$ we select the minimal length set.

Limit cases of MinSE Mechanism (1)

1. $b = b_0 = (1/K, \dots 1/K)$ uniform prior and $e^\eta = K, \tau = 0$:

$$\begin{aligned} p^\star(b_0, \xi) =& \mathrm{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^\alpha(X)) \\ & \text{s.t.} p \in \Delta^{K-1}, \quad p_i \leq e^\eta b_i = 1 \\ =& e_{\mathrm{argmin}_i \lambda(C_i^\alpha(X))} \end{aligned}$$

Thus $\hat{S}(\xi, \epsilon) = \mathrm{argmin}_i \lambda(C_i^\alpha(X))$ we select the minimal length set.

→ To ensure $1 - \alpha$ coverage of the corresponding set, we need $1 - \alpha/K$ individual coverage.

Limit cases of MinSE Mechanism (1)

1. $b = b_0 = (1/K, \dots 1/K)$ uniform prior and $e^\eta = K, \tau = 0$:

$$
\begin{aligned}
p^\star(b_0, \xi) =& \mathrm{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^\alpha(X)) \\
& \mathrm{s.t.} p \in \Delta^{K-1}, \quad p_i \le e^\eta b_i = 1 \\
=& e_{\mathrm{argmin}_i \lambda(C_i^\alpha(X))}
\end{aligned}
$$

Thus $\hat{S}(\xi, \epsilon) = \mathrm{argmin}_i \lambda(C_i^\alpha(X))$ we select the minimal length set.

$\rightarrow$ To ensure $1-\alpha$ coverage of the corresponding set, we need $1 - \alpha/K$ individual coverage.

$\rightarrow$ Akin to Bonferoni correction

Limit cases of MinSE Mechanism (1)

1. $b = b_0 = (1/K, \dots 1/K)$ uniform prior and $e^\eta = K, \tau = 0$:

$$
\begin{aligned}
p^\star(b_0, \xi) =& \operatorname{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^\alpha(X)) \\
& \text{s.t.} p \in \Delta^{K-1}, \quad p_i \leq e^\eta b_i = 1 \\
=& e_{\operatorname{argmin}_i \lambda(C_i^\alpha(X))}
\end{aligned}
$$

Thus $\hat{S}(\xi, \epsilon) = \operatorname{argmin}_i \lambda(C_i^\alpha(X))$ we select the minimal length set.

→ To ensure $1-\alpha$ coverage of the corresponding set, we need $1-\alpha/K$ individual coverage.

→ Akin to Bonferoni correction

→ Beneficial if at any X, one set with coverage $1-\alpha/K$ is much smaller than the average length of sets with coverage $1-\alpha$.

Limit cases of MinSE Mechanism (2)

2 b any prior and $e^{\eta} = 1, \tau = 0$:

$$
\begin{aligned}
p^\star(b_0, \xi) =& \operatorname{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^\alpha(X)) \\
& \text{s.t.} p \in \Delta^{K-1}, \quad p_i \leq b_i \\
=& b
\end{aligned}
$$

Limit cases of MinSE Mechanism (2)

2 b any prior and $e^{\eta} = 1, \tau = 0$:

$$
\begin{aligned}
p^{\star}(b_0, \xi) =& \operatorname{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^{\alpha}(X)) \\
& \text{s.t.} p \in \Delta^{K-1}, \quad p_i \leq b_i \\
=& b
\end{aligned}
$$

$\Rightarrow \hat{S}(\xi, \epsilon)$ picks without data dependence, just sampling from a prior b.

Limit cases of MinSE Mechanism (2)

2 b any prior and $e^{\eta} = 1, \tau = 0$:

$$\begin{aligned} p^{\star}(b_0, \xi) = & \operatorname{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^{\alpha}(X)) \\ & \text{s.t.} p \in \Delta^{K-1}, \quad p_i \leq b_i \\ = & b \end{aligned}$$

$\Rightarrow \hat{S}(\xi, \epsilon)$ picks without data dependence, just sampling from a prior b.

$\rightarrow$ To ensure $1 - \alpha$ coverage of the corresponding set, we need $1 - \alpha$ individual coverage.

3 $b = b_0 = (1/K, \dots 1/K)$ uniform prior and $e^\eta = 2, \tau = 0$:

$$
\begin{aligned}
p^\star(b_0, \xi) = &\operatorname{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^\alpha(X)) \\
&\text{s.t.} p \in \Delta^{K-1}, \quad p_i \leq 2/K
\end{aligned}
$$

Limit cases of MinSE Mechanism (3)

3 $b = b_0 = (1/K, \dots 1/K)$ uniform prior and $\boxed{e^{\eta} = 2, \tau = 0}$:

$$p^{\star}(b_0, \xi) = \mathrm{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^{\alpha}(X))$$
$$\text{s.t.}\, p \in \Delta^{K-1}, \quad p_i \leq 2/K$$

$\Rightarrow \hat{S}(\xi, \epsilon)$ only picks among half the sets with smallest lengths.

Limit cases of MinSE Mechanism (3)

3 $b = b_0 = (1/K, \ldots 1/K)$ uniform prior and $e^\eta = 2, \tau = 0$:

$$p^\star(b_0, \xi) = \mathrm{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^\alpha(X))$$
$$\text{s.t.} p \in \Delta^{K-1}, \quad p_i \leq 2/K$$

$\Rightarrow \hat{S}(\xi, \epsilon)$ only picks among half the sets with smallest lengths.

$\rightarrow$ To ensure $1-\alpha$ coverage of the corresponding set, we need $1-\alpha/2$ individual coverage.

Limit cases of MinSE Mechanism (3)

3 $b = b_0 = (1/K, \ldots 1/K)$ uniform prior and $e^\eta = 2, \tau = 0$:

$$p^\star(b_0, \xi) = \mathrm{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^\alpha(X))$$
$$\text{s.t.} p \in \Delta^{K-1}, \quad p_i \leq 2/K$$

$\Rightarrow \hat{S}(\xi, \epsilon)$ only picks among half the sets with smallest lengths.

→ To ensure $1 - \alpha$ coverage of the corresponding set, we need $1 - \alpha/2$ individual coverage.

→ Beneficial if at any X, sets with coverage $1 - \alpha/2$ are much smaller than the average length of sets with coverage $1 - \alpha$.

Limit cases of MinSE Mechanism (3)

3 $b = b_0 = (1/K, \dots 1/K)$ uniform prior and $e^{\eta} = 2, \tau = 0$:

$$p^{\star}(b_0, \xi) = \mathrm{argmin}_p \sum_{i=1}^{K} p_i \lambda(C_i^{\alpha}(X))$$
$$\mathrm{s.t.} p \in \Delta^{K-1}, \quad p_i \leq 2/K$$

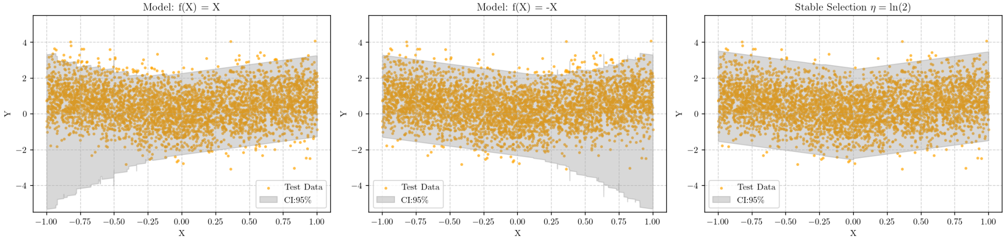
$\Rightarrow \hat{S}(\xi, \epsilon)$ only picks among half the sets with smallest lengths.

$\rightarrow$ To ensure $1 - \alpha$ coverage of the corresponding set, we need $1 - \alpha/2$ individual coverage.

$\rightarrow$ Beneficial if at any X, sets with coverage $1 - \alpha/2$ are much smaller than the average length of sets with coverage $1 - \alpha$.

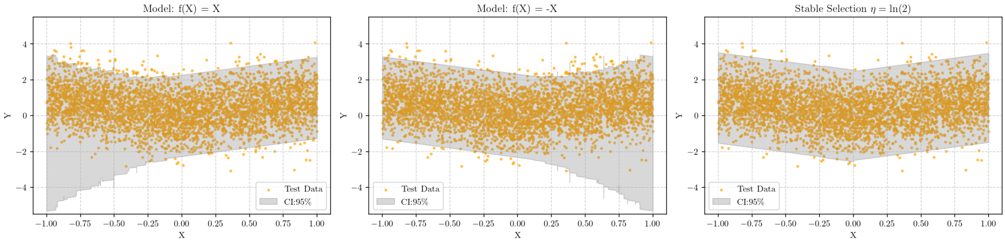
$\looparrowright$ Overall, as expected, useful when conformal sets perform "heterogeneously".

Toy example: Split Input



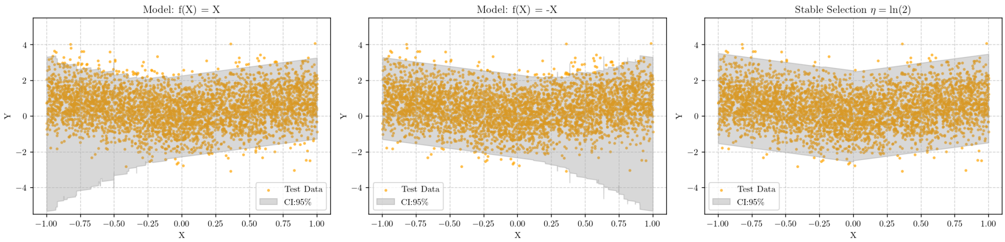
- **Setup:** $Y = |X| + \mathcal{N}(0, 0.25)$, $X \sim \mathrm{Uniform}([-1, 1])$.

Toy example: Split Input



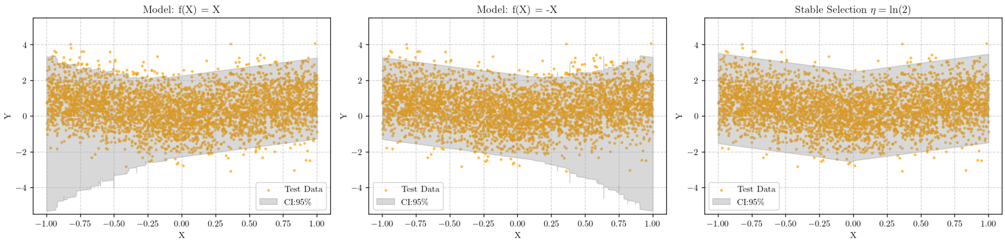
- **Setup:** $Y = |X| + \mathcal{N}(0, 0.25)$, $X \sim \text{Uniform}([-1, 1])$.
- **Predictors:**

Toy example: Split Input



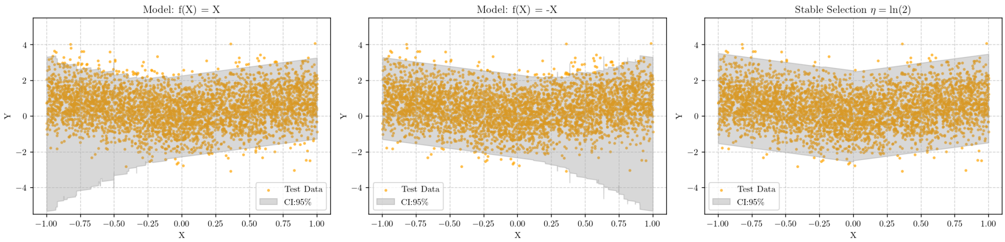
- **Setup:** $Y = |X| + \mathcal{N}(0, 0.25)$, $X \sim \mathrm{Uniform}([-1, 1])$.
- **Predictors:**
  - $f_1(X) = X$: Good for $X \geq 0$.

Toy example: Split Input



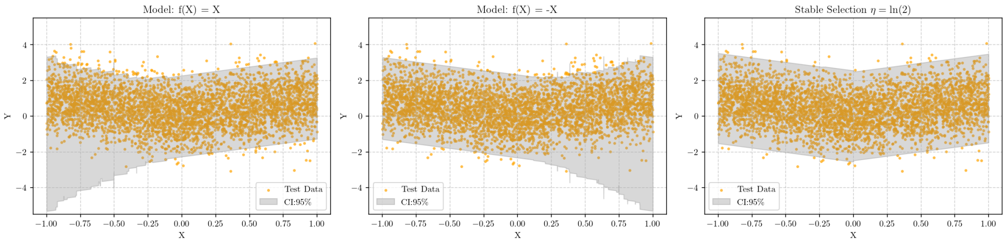
- **Setup:** $Y = |X| + \mathcal{N}(0, 0.25)$, $X \sim \mathrm{Uniform}([-1, 1])$.
- **Predictors:**
  - $f_1(X) = X$: Good for $X \geq 0$.
  - $f_2(X) = -X$: Good for $X \leq 0$.

Toy example: Split Input



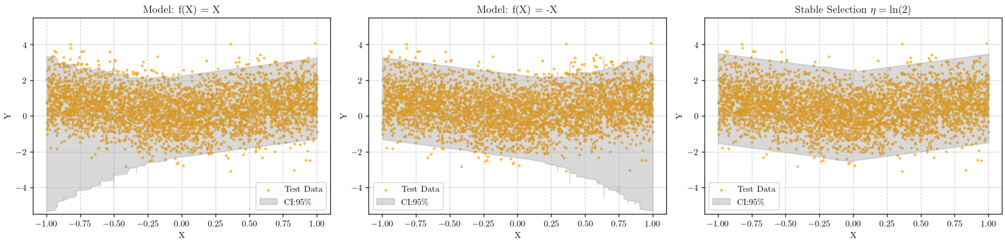
- **Setup:** $Y = |X| + \mathcal{N}(0, 0.25)$, $X \sim \mathrm{Uniform}([-1, 1])$.
- **Predictors:**
  - $f_1(X) = X$: Good for $X \geq 0$.
  - $f_2(X) = -X$: Good for $X \leq 0$.
- Sets obtained SCP with score $s(X, Y) = \frac{|Y - \hat{f}(X)|}{\hat{\rho}(X)}$

Toy example: Split Input



- **Setup:** $Y = |X| + \mathcal{N}(0, 0.25)$, $X \sim \mathrm{Uniform}([-1, 1])$.
- **Predictors:**
  - $f_1(X) = X$: Good for $X \geq 0$.
  - $f_2(X) = -X$: Good for $X \leq 0$.
- Sets obtained SCP with score $s(X, Y) = \frac{|Y - \hat{f}(X)|}{\hat{\rho}(X)}$
- Complementary strengths.

Toy example: Split Input



- **Setup:** $Y = |X| + \mathcal{N}(0, 0.25)$, $X \sim \text{Uniform}([-1, 1])$.
- **Predictors:**
  - $f_1(X) = X$: Good for $X \geq 0$.
  - $f_2(X) = -X$: Good for $X \leq 0$.
- Sets obtained SCP with score $s(X, Y) = \frac{|Y - \hat{f}(X)|}{\hat{\rho}(X)}$
- Complementary strengths.
- **Stable Selection:** allows pointwise selection which is better than picking the best on average as in Liang et al. (2024); Yang and Kuchibhotla (2024).

Extensions and improvements: 1 - AdaMinSE

1. Tradeoff between η and τ still not clear.
2. **Adaptive** MinSE (AdaMinSE): optimize over η and τ, to achieve a desired target miscoverage level α, given that original sets miscoverage is α'.
3. This is also a linear program:

AdaMinSE Mechanism

$$
\begin{aligned}
d^*(b,\xi) = \operatorname{argmin}_d \; & \sum_{i=1}^{K} d_i \lambda(C_i^{\alpha}(X)) \\
\text{s.t.} \quad & d \in \Delta^{K-1}, s \in \mathrm{R}_+^K, \tau, \eta \geq 0 \\
& d_i \leq e^{\eta} b_i + s_i, \quad \sum_{i \in [K]} s_i \leq \tau, \quad e^{\eta}\alpha' + \tau \leq \alpha
\end{aligned}
$$

Selection rule such that

- $\mathbb{P}\left\{\hat{S}(\xi,\epsilon) = i | \xi\right\} = d^*(b,\xi)_i$
- is η, τ-stable.

- **Data-Dependent Prior:** a uniform prior b_0 can be used in MinSE and AdaMinSE but can be suboptimal.

- **Data-Dependent Prior:** a uniform prior b_0 can be used in MinSE and AdaMinSE but can be suboptimal.
- Construct the prior using cross-validation on the training data, $\mathcal{D}_{\text{train}}$ or

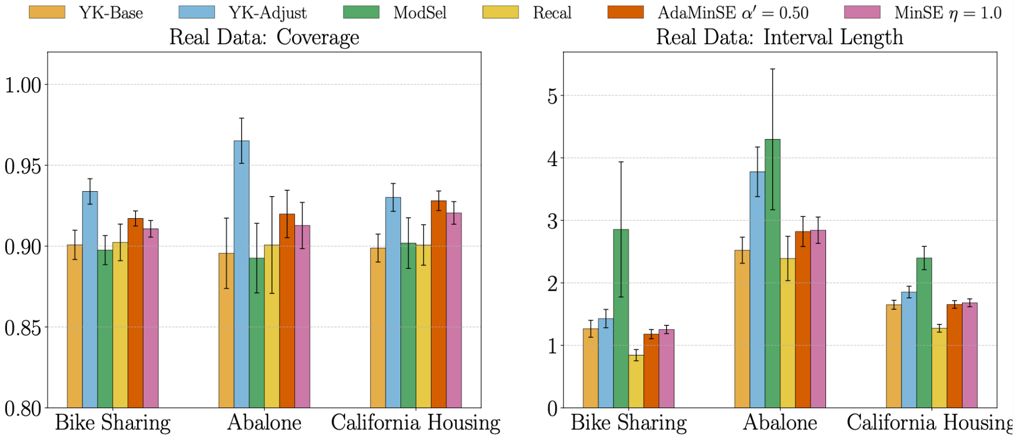
Extensions and improvements: 2- Improving the Prior

- **Data-Dependent Prior:** a uniform prior b_0 can be used in MinSE and AdaMinSE but can be suboptimal.
- Construct the prior using cross-validation on the training data, $\mathcal{D}_{\text{train}}$

or

- Construct the prior in an online fashion, incorporating techniques like COMA Gasparin and Ramdas (2024a).

Experiments: UCI Datasets Setup



Baselines: YK (Yang and Kuchibhotla, 2024, EFCP), LZB (Liang et al., 2024, ModSel-CP).

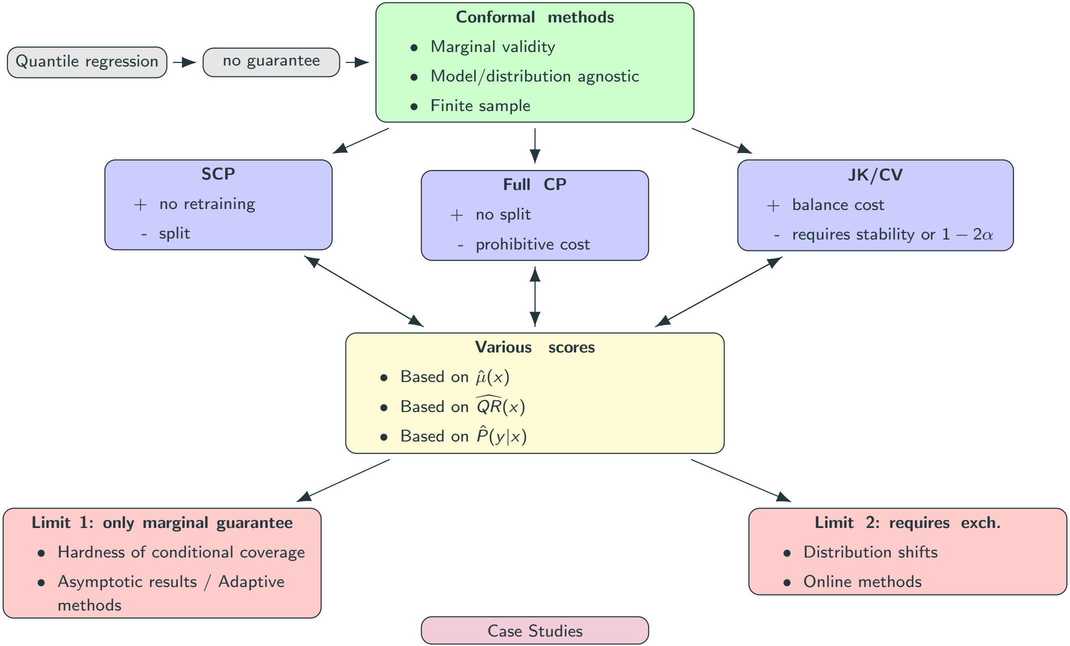
Heterogeneous training sets

Metrics: Coverage ($\geq 1 - \alpha$) & Normalized Interval Length (smaller is better)

Take-home-messages : Valid Selection among Conformal Sets

1. Coverage after selection requires care !
2. We leverage both stability based - and recalibration based methods can bring improvements.
3. Those techniques enable to favor pointwise smallest sets - they come at a cost.
4. Stability based method easily incorporate prior (e.g. in online) information.
5. Overall, expected to be favorable in heterogeneous training setups.

Summary: Uncertainty quantification through conformal methods



Some (other, non-exhaustives) current open directions

- Outlier detection (Vovk et al., 2003; Bates et al., 2023)
- Selective inference, false discovery rate guarantees (Marandon et al., 2024; Gazin et al., 2024)
- Beyond the indicator loss (Angelopoulos et al., 2022a; Bates et al., 2021b; Angelopoulos et al., 2023; Lekeufack et al., 2024)
- Aggregating predictive sets (Gasparin and Ramdas, 2024c,b; Gasparin et al., 2024)

Acknowledgments

For discussion and feedback, thanks to

- Julie Josse
- Claire Boyer
- Étienne Roquain

Questions?

Angelopoulos, A. N. and Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4).

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. (2022a). Learn then test: Calibrating predictive algorithms to achieve risk control.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. (2023). Conformal risk control.

Angelopoulos, A. N., Kohli, A. P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., and Romano, Y. (2022b). Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pages 717–730. PMLR.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021a). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2).

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2021b). Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1).

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. (2022). Conformal prediction beyond exchangeability. To appear in *Annals of Statistics (2023)*.

Bassily, R. and Freund, Y. (2016). Typical stability. *arXiv preprint arXiv:1604.03336*.

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. (2016). Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. (2021a). Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. (2021b). Distribution-free, risk-controlling prediction sets. *J. ACM*, 68(6).

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149 – 178.

Bian, M. and Barber, R. F. (2023). Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044 – 2066.

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2020). Robust Validation: Confident Predictions Even When Distributions Shift. arXiv: 2008.04267.

Chernozhukov, V., Wüthrich, K., and Yinchu, Z. (2018). Exact and Robust Conformal Inference Methods for Predictive Machine Learning with Dependent Data. In *Conference On Learning Theory*. PMLR.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48).

Cherubin, G., Chatzikokolakis, K., and Jaggi, M. (2021). Exact optimization of conformal predictors via incremental and decremental learning. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR.

Dutot, G., Zaffran, M., Féron, O., and Goude, Y. (2024). Adaptive probabilistic forecasting of french electricity spot prices.

Feldman, S., Bates, S., and Romano, Y. (2021). Improving Conditional Coverage via Orthogonal Quantile Regression. *arXiv:2106.00394 [cs]*. arXiv: 2106.00394.

Gasparin, M. and Ramdas, A. (2024a). Conformal online model aggregation. *arXiv preprint arXiv:2403.15527*.

Gasparin, M. and Ramdas, A. (2024b). Conformal online model aggregation.

Gasparin, M. and Ramdas, A. (2024c). Merging uncertainty sets via majority vote.

Gasparin, M., Wang, R., and Ramdas, A. (2024). Combining exchangeable p-values.

Gazin, U., Blanchard, G., and Roquain, E. (2024). Transductive conformal inference with adaptive scores.

Gibbs, I. and Candès, E. (2021). Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Gibbs, I. and Candès, E. (2022). Conformal inference for online prediction with arbitrary distribution shifts. arXiv: 2208.08401.

Gibbs, I., Cherian, J. J., and Candès, E. J. (2023). Conformal prediction with conditional guarantees. arXiv: 2305.12616.

Guan, L. (2022). Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1).

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127.

Izbicki, R., Shimizu, G., and Stern, R. B. (2022). CD-split and HPD-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87).

Jung, C., Noarov, G., Ramalingam, R., and Roth, A. (2023). Batch multivalid conformal prediction. In *International Conference on Learning Representations*.

Kivaranovic, D., Johnson, K. D., and Leeb, H. (2020). Adaptive, Distribution-Free Prediction Intervals for Deep Networks. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Lei, J. (2019). Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106(4).

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1).

Lekeufack, J., Angelopoulos, A. N., Bajcsy, A., Jordan, M. I., and Malik, J. (2024). Conformal decision theory: Safe autonomous decisions from imperfect predictions.

Liang, R., Zhu, W., and Barber, R. F. (2024). Conformal prediction after efficiency-oriented model selection. *arXiv preprint arXiv:2408.07066*.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2024). Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157 – 183.

Ndiaye, E. (2022). Stable conformal prediction sets. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR.

Ndiaye, E. and Takeuchi, I. (2019). Computing full conformal prediction set with approximate homotopy. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Ndiaye, E. and Takeuchi, I. (2022). Root-finding approaches for computing conformal prediction set. *Machine Learning*, 112(1).

Nouretdinov, I., Melluish, T., and Vovk, V. (2001). Ridge regression confidence machine. In *Proceedings of the 18th International Conference on Machine Learning*.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive Confidence Machines for Regression. In *Machine Learning: ECML*. Springer.

Podkopaev, A. and Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. PMLR.

Romano, Y., Barber, R. F., Sabatti, C., and Candès, E. (2020a). With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2(2).

Romano, Y., Patterson, E., and Candès, E. (2019). Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Romano, Y., Sesia, M., and Candes, E. (2020b). Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Sadinle, M., Lei, J., and Wasserman, L. (2018). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

Sesia, M. and Romano, Y. (2021). Conformal prediction using conditional histograms. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Tibshirani, R. J., Barber, R. F., Candes, E., and Ramdas, A. (2019). Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Vovk, V. (2012). Conditional Validity of Inductive Conformal Predictors. In *Asian Conference on Machine Learning*. PMLR.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer US.

Vovk, V., Nouretdinov, I., and Gammerman, A. (2003). Testing exchangeability on-line. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 768–775. AAAI Press.

Yang, Y. and Kuchibhotla, A. K. (2024). Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, pages 1–13.

Zaffran, M., Féron, O., Goude, Y., Josse, J., and Dieuleveut, A. (2022). Adaptive conformal predictions for time series. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR.

Zrnic, T. and Jordan, M. I. (2023). Post-selection inference via algorithmic stability. *The Annals of Statistics*, 51(4):1666–1691.

Proof of the quantile lemma

Def: Empirical quantile $q_{\beta}(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

Proof of the quantile lemma

Def: Empirical quantile $q_\beta(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

First note that $U_{n+1} \le q_\beta(U_1, \ldots, U_n, +\infty) \iff U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})$.

Def: Empirical quantile $q_\beta(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

First note that $U_{n+1} \leq q_\beta(U_1, \ldots, U_n, +\infty) \iff U_{n+1} \leq q_\beta(U_1, \ldots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right) = \mathbb{P}\left(U_i \leq q_\beta(U_1, \ldots, U_n, U_{n+1})\right)$.

Def: Empirical quantile $q_\beta(U_1, \dots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \dots, U_k)$

First note that $U_{n+1} \leq q_\beta(U_1, \dots, U_n, +\infty) \iff U_{n+1} \leq q_\beta(U_1, \dots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right) = \mathbb{P}\left(U_i \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right)$. Thus:

$$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(U_i \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right)$$

Proof of the quantile lemma

Def: Empirical quantile $q_\beta(U_1, \dots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \dots, U_k)$

First note that $U_{n+1} \leq q_\beta(U_1, \dots, U_n, +\infty) \iff U_{n+1} \leq q_\beta(U_1, \dots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right) = \mathbb{P}\left(U_i \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right)$. Thus:

$$\begin{aligned}
\mathbb{P}\left(U_{n+1} \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(U_i \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right) \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}\left\{U_i \leq q_\beta(U_1, \dots, U_n, U_{n+1})\right\}\right]
\end{aligned}$$

Proof of the quantile lemma

Def: Empirical quantile $q_\beta(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

First note that $U_{n+1} \le q_\beta(U_1, \ldots, U_n, +\infty) \iff U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right) = \mathbb{P}\left(U_i \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right)$. Thus:

$$
\begin{aligned}
\mathbb{P}\left(U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(U_i \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right) \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}\left\{U_i \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right\}\right] \\
&\ge \frac{1}{n+1} \mathbb{E}\left[\lceil \beta(n+1) \rceil\right] \\
&= \frac{\lceil \beta(n+1) \rceil}{n+1} \ge \beta,
\end{aligned}
$$

proving the first statement.

Proof of the quantile lemma

Def: Empirical quantile $q_\beta(U_1, \ldots, U_k) := \lceil \beta \times k \rceil$ smallest value of $(U_1, \ldots, U_k)$

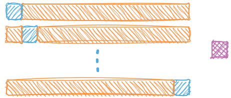
First note that $U_{n+1} \le q_\beta(U_1, \ldots, U_n, +\infty) \iff U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})$.

By exchangeability, for any $i \in [\![1, n+1]\!]$:
$\mathbb{P}\left(U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right) = \mathbb{P}\left(U_i \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right)$. Thus:

$$
\begin{aligned}
\mathbb{P}\left(U_{n+1} \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\left(U_i \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right) \\
&= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}\left\{U_i \le q_\beta(U_1, \ldots, U_n, U_{n+1})\right\}\right] \\
&= \frac{1}{n+1} \mathbb{E}\left[\lceil \beta(n+1) \rceil\right] \quad \text{if all } (U_i) \text{ are distinct} \\
&= \frac{\lceil \beta(n+1) \rceil}{n+1} \le \beta + \frac{1}{n+1},
\end{aligned}
$$

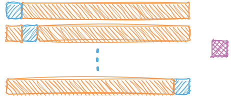
proving the second statement.

Jackknife: the naive idea does not enjoy valid coverage



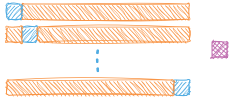
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$

Jackknife: the naive idea does not enjoy valid coverage



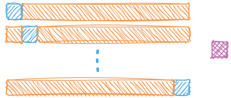
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean regression)

Jackknife: the naive idea does not enjoy valid coverage



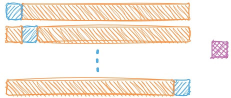
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean regression)
- Get $\hat{A}$ by training $\mathcal{A}$ on $\mathcal{D}_n$

Jackknife: the naive idea does not enjoy valid coverage



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean regression)
- Get $\hat{A}$ by training $\mathcal{A}$ on $\mathcal{D}_n$
- Build the predictive interval: $\left[\hat{A}(X_{n+1}) \pm q_{1-\alpha}(\mathcal{S}) \right]$

Jackknife: the naive idea does not enjoy valid coverage



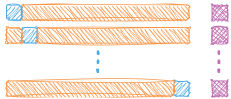
- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO scores $\mathcal{S} = \left\{ |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{+\infty\}$ (in standard mean regression)
- Get $\hat{A}$ by training $\mathcal{A}$ on $\mathcal{D}_n$
- Build the predictive interval: $\left[\hat{A}(X_{n+1}) \pm q_{1-\alpha}(\mathcal{S}) \right]$

Warning

No guarantee on the prediction of $\hat{A}$ with scores based on $(\hat{A}_{-i})_i$, without assuming a form of **stability** on $\mathcal{A}$.

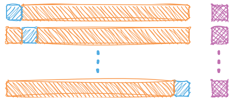
Jackknife+ (Barber et al., 2021b)

Jackknife+ (Barber et al., 2021b)



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$

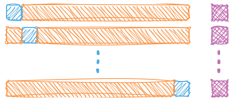
Jackknife+ (Barber et al., 2021b)



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO predictions / predictive intervals
  $$\mathcal{S}_{\mathrm{up/down}} = \left\{ \hat{A}_{-i}(X_{n+1}) \pm |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{\pm\infty\}$$

(in standard mean regression)

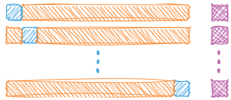
Jackknife+ (Barber et al., 2021b)



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO predictions / predictive intervals
  $\mathcal{S}_{\text{up/down}} = \left\{ \hat{A}_{-i}(X_{n+1}) \pm |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{\pm\infty\}$
  (in standard mean regression)
- Build the predictive interval: $[q_{\alpha,\inf}(\mathcal{S}_{\text{down}}); q_{1-\alpha}(\mathcal{S}_{\text{up}})]$

Recall $q_{\beta,\inf}(X_1, \dots, X_k) := \lfloor \beta \times k \rfloor$ smallest value of $(X_1, \dots, X_k)$

Jackknife+ (Barber et al., 2021b)



- Based on leave-one-out (LOO) residuals
- $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ training data
- Get $\hat{A}_{-i}$ by training $\mathcal{A}$ on $\mathcal{D}_n \setminus (X_i, Y_i)$
- LOO predictions / predictive intervals
  $\mathcal{S}_{\text{up/down}} = \left\{ \hat{A}_{-i}(X_{n+1}) \pm |\hat{A}_{-i}(X_i) - Y_i| \right\}_i \cup \{\pm\infty\}$
  (in standard mean regression)
- Build the predictive interval: $[q_{\alpha,\inf}(\mathcal{S}_{\text{down}}); q_{1-\alpha}(\mathcal{S}_{\text{up}})]$

Marginal validity of Jackknife+ Barber et al. (2021b)

If $\mathcal{D}_n \cup (X_{n+1}, Y_{n+1})$ are exchangeable and $\mathcal{A}$ is symmetric: $\mathbb{P}(Y_{n+1} \in \widehat{C}_\alpha(X_{n+1})) \geq 1 - 2\alpha$.

Recall $q_{\beta,\inf}(X_1, \ldots, X_k) := \lfloor \beta \times k \rfloor$ smallest value of $(X_1, \ldots, X_k)$

Recalibration: Approach

- **Split conformal** with calibration data $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{m}$ + test (X, Y), all $(m{+}1)$ points exchangeable.
- K base predictors $f_1, \ldots, f_K$ with non-conformity scores $(s_k)_{k \in [K]}$.
- **Rank–parameterised sets:**

$$C_k(X, R) = \big\{y : s_k(X, y, f_k) \le s_{k,(R)}\big\}.$$

- Choosing $R_\alpha = \lceil (1-\alpha)(m+1) \rceil$ gives $(C_k^\alpha)_{k \in K}$

$$\mathbb{P}\big\{Y \in C_k^\alpha(X, R_\alpha)\big\} \ge 1 - \alpha$$

- **After-selection challenge:** A stochastic rule $\hat{S}$ (depending on X) picks a predictor; vanilla coverage can break—needs new calibration.

Recalibration via Effective Ranks

- For each calibration point, set the (meta-score)

$$\hat{R}_i \;=\; R_{\hat{S}(X_i,\varepsilon_i),\,i}.$$

  i.e., the rank of the i-th point's score calculated using the selected predictor $S(X_i, \varepsilon_i)$
- Let $\hat{R}_{(1)} \le \cdots \le \hat{R}_{(m)}$ be the order statistics and $\tau_\alpha = \lceil (1-\alpha)(m+1) \rceil$.

Recalibration.

If $\hat{S} \;\perp\; \mathcal{D}_{\mathrm{cal}}$ then

$$\mathbb{P}\Big\{ Y \in C_{\hat{S}(X,\varepsilon)}\big(X, \hat{R}_{(\tau_\alpha)}\big) \Big\} \;\ge\; 1-\alpha.$$

Gives *exact*, finite-sample, distribution-free coverage for the *selected* predictor, without conservative inflation.

Implementing an Independent Selection Rule

- Independence is essential: meta-scores must stay exchangeable.
- Use an **auxiliary dataset** $\mathcal{D}_{\mathrm{aux}}$ (disjoint from $\mathcal{D}_{\mathrm{cal}}$).

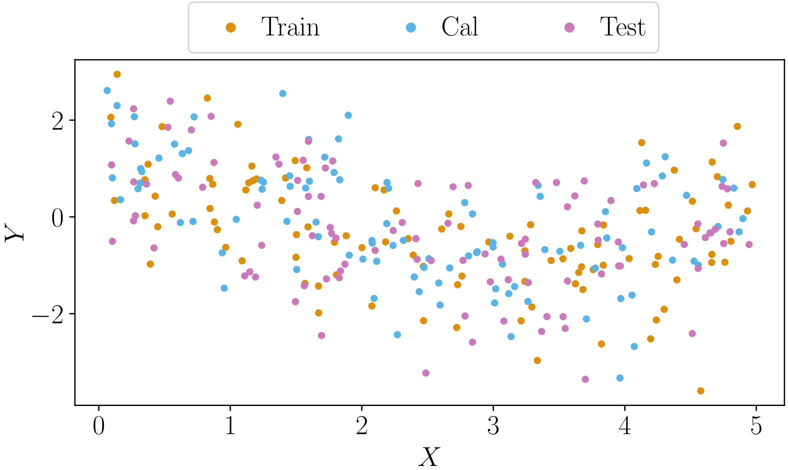
Extraslides no animations

Extraslides no animations

SCP

CQR

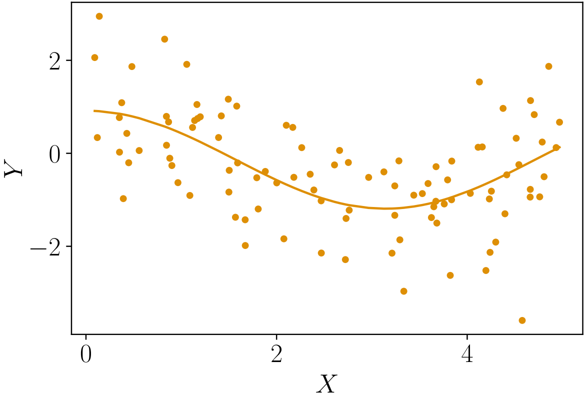
Split Conformal Prediction (SCP)[1,2,3]: toy example



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

[2]Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML

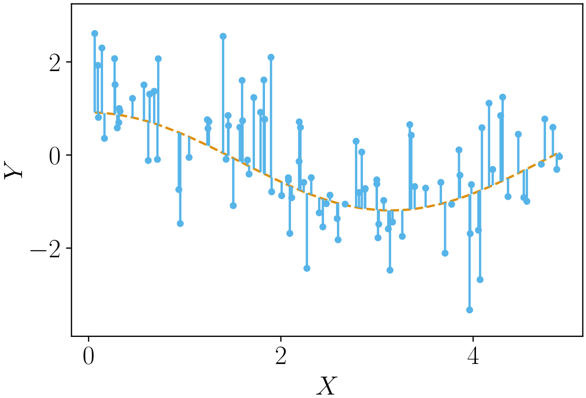
[3]Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B



- Learn (or get) $\hat{\mu}$

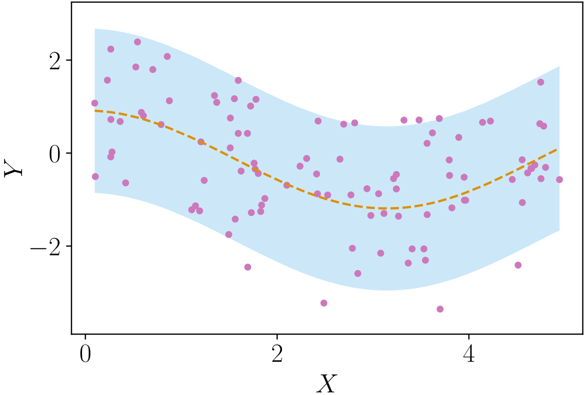
[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML
[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B

Split Conformal Prediction (SCP)[1,2,3]: toy example — calibration step



- Predict with $\hat{\mu}$
- Get the $|\text{residuals}|$, a.k.a. conformity scores
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S} = \{|\text{residuals}|\}_{\text{Cal}} \cup \{+\infty\}$, noted $q_{1-\alpha}(\mathcal{S})$

[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML
[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B



- Predict with $\hat{\mu}$
- Build $\widehat{C}_\alpha(x)$: $[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$

Back to SCP

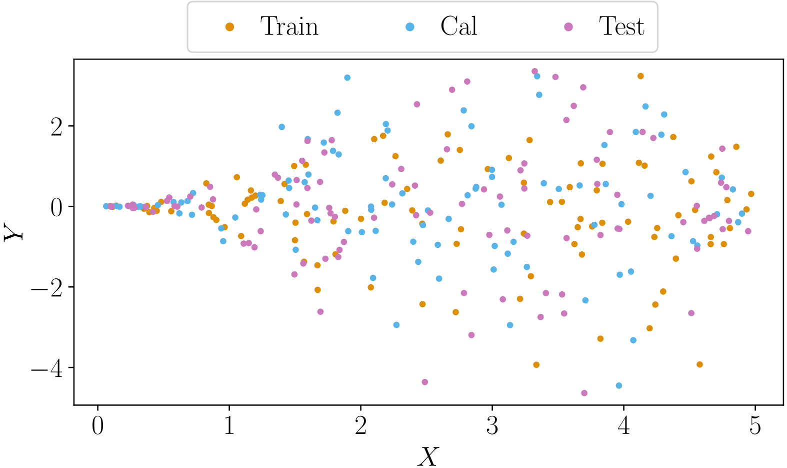
[1] Vovk et al. (2005), *Algorithmic Learning in a Random World*
[2] Papadopoulos et al. (2002), *Inductive Confidence Machines for Regression*, ECML
[3] Lei et al. (2018), *Distribution-Free Predictive Inference for Regression*, JRSS B

Extraslides no animations

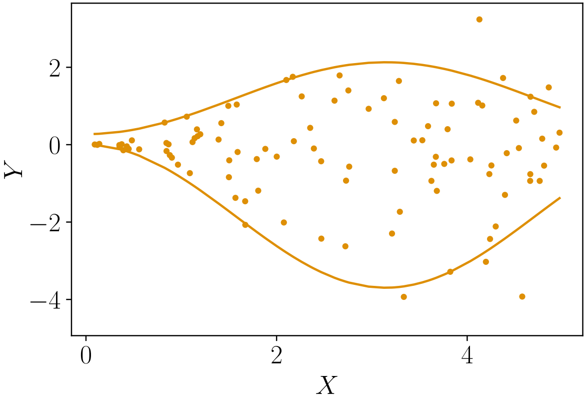
SCP

CQR

Conformalized Quantile Regression (CQR)[5]

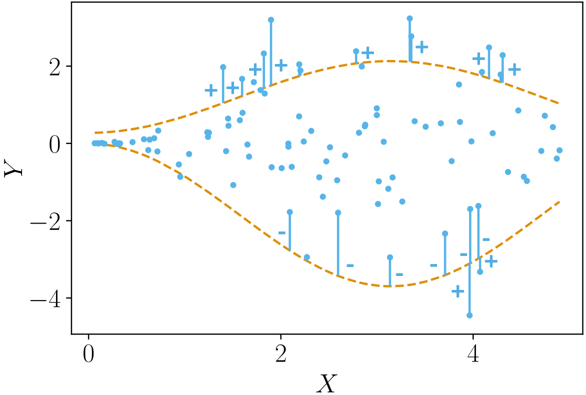


[5] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS



- Learn (or get) $\widehat{\mathrm{QR}}_{\mathrm{lower}}$ and $\widehat{\mathrm{QR}}_{\mathrm{upper}}$

[5] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS



- Predict with $\widehat{\mathrm{QR}}_{\mathrm{lower}}$ and $\widehat{\mathrm{QR}}_{\mathrm{upper}}$
- Get the scores $\mathcal{S} = \{S_i\}_{\mathrm{Cal}} \cup \{+\infty\}$
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S}$, noted $q_{1-\alpha}(\mathcal{S})$

$$\hookrightarrow \quad S_i := \max\left\{\widehat{\mathrm{QR}}_{\mathrm{lower}}(X_i) - Y_i, Y_i - \widehat{\mathrm{QR}}_{\mathrm{upper}}(X_i)\right\}$$

Back to Generalization SCP

[5] Romano et al. (2019), *Conformalized Quantile Regression*, NeurIPS

Label shift (Podkopaev and Ramdas, 2021)

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \stackrel{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification

Label shift (Podkopaev and Ramdas, 2021)

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification
- Idea: give more importance to calibration points that are closer in distribution to the test point

Label shift (Podkopaev and Ramdas, 2021)

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification
- Idea: give more importance to calibration points that are closer in distribution to the test point
- Trouble: the actual test labels are unknown

Label shift (Podkopaev and Ramdas, 2021)

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification
- Idea: give more importance to calibration points that are closer in distribution to the test point
- Trouble: the actual test labels are unknown
- In practice:
  1. estimate the likelihood ratio $w(Y_i) = \frac{\mathrm{d}\tilde{P}_Y(Y_i)}{\mathrm{d}P_Y(Y_i)}$ using algorithms from the existing label shift literature

Label shift (Podkopaev and Ramdas, 2021)

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification
- Idea: give more importance to calibration points that are closer in distribution to the test point
- Trouble: the actual test labels are unknown
- In practice:
  1. estimate the likelihood ratio $w(Y_i) = \frac{\mathrm{d}\tilde{P}_Y(Y_i)}{\mathrm{d}P_Y(Y_i)}$ using algorithms from the existing label shift literature
  2. normalize the weights, i.e. $\omega_i^y = \omega^y(X_i) = \frac{w(Y_i)}{\sum_{j=1}^n w(Y_j) + w(y)}$

Label shift (Podkopaev and Ramdas, 2021)

- Setting:
  - $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P_{X|Y} \times P_Y$
  - $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$
  - Classification
- Idea: give more importance to calibration points that are closer in distribution to the test point
- Trouble: the actual test labels are unknown
- In practice:
  1. estimate the likelihood ratio $w(Y_i) = \frac{\mathrm{d}\tilde{P}_Y(Y_i)}{\mathrm{d}P_Y(Y_i)}$ using algorithms from the existing label shift literature
  2. normalize the weights, i.e. $\omega_i^y = \omega^y(X_i) = \frac{w(Y_i)}{\sum_{j=1}^n w(Y_j) + w(y)}$
  3. outputs $\widehat{C}_\alpha(X_{n+1}) =$

$$\left\{ y : s\left(X_{n+1}, y; \hat{A}\right) \leq Q_{1-\alpha}\left(\sum_{i \in \mathrm{Cal}} \omega_i^y \delta_{S_i} + \omega_{n+1}^y \delta_\infty \right) \right\}$$