ConForME: Multi-horizon conformal time series forecasting

PhD Student:
Aloysio Galvão Lopes

Advisors:
Eric Goubault
Laurent Pautet
Sylvie Putot

Outline

- Work accepted recently at the *13th Symposium on Conformal and Probabilistic Prediction with Applications (COPA 2024)*
- Recap on conformal prediction
- ConForME

Conformal Prediction

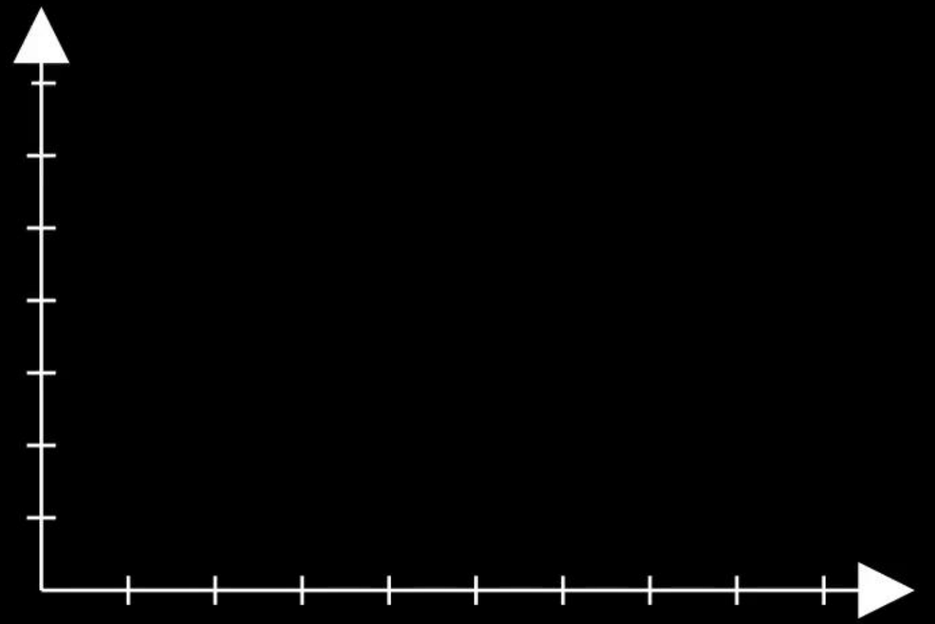
Given a dataset of 1-D points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to compute valid prediction regions for a new point $\hat{y}_{n+1} = f(x_{n+1})$, for a given desired coverage probability $1 - \alpha$.

Conformal Prediction

Given a dataset of 1-D points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to compute valid prediction regions for a new point $\hat{y}_{n+1} = f(x_{n+1})$, for a given desired coverage probability $1 - \alpha$.

$$f(x_{n+1}) = \hat{y}_{n+1}$$

$$(x_{n+1}, y_{n+1})$$

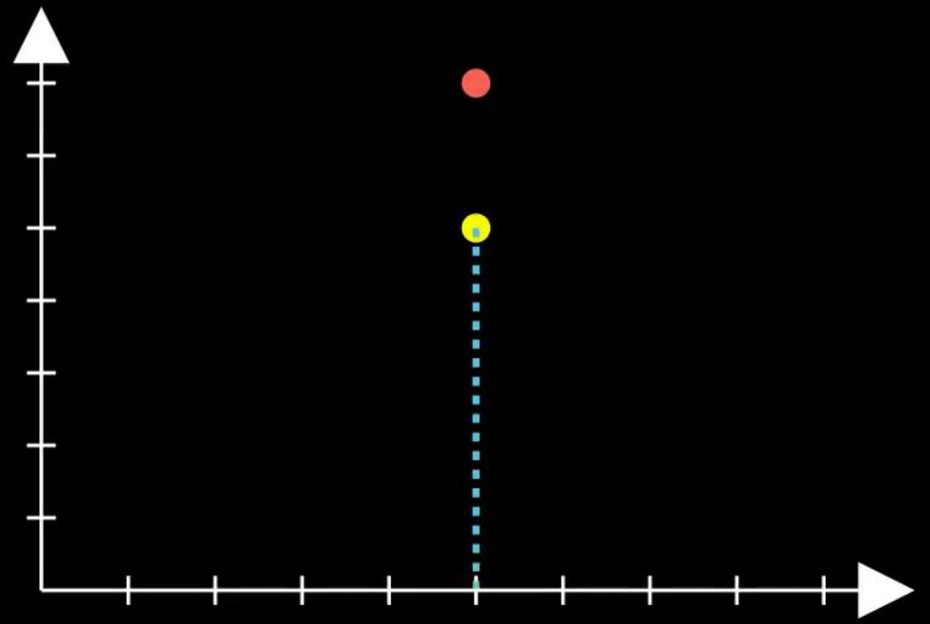


Conformal Prediction

Given a dataset of 1-D points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to compute valid prediction regions for a new point $\hat{y}_{n+1} = f(x_{n+1})$, for a given desired coverage probability $1 - \alpha$.

$$f(x_{n+1}) = \hat{y}_{n+1}$$

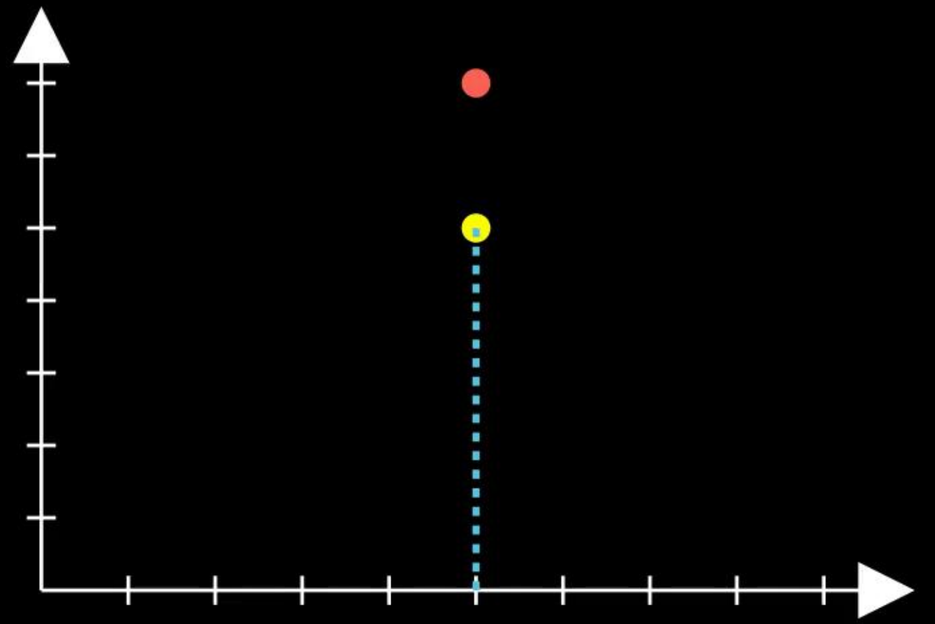
$$(x_{n+1}, y_{n+1})$$



Conformal Prediction

Given a dataset of 1-D points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to compute valid prediction regions for a new point $\hat{y}_{n+1} = f(x_{n+1})$, for a given desired coverage probability $1 - \alpha$.

$$f(\ 5.0\ ) = \ 5.0$$
$$(\ 5.0\ ,\ 7.0\ )$$

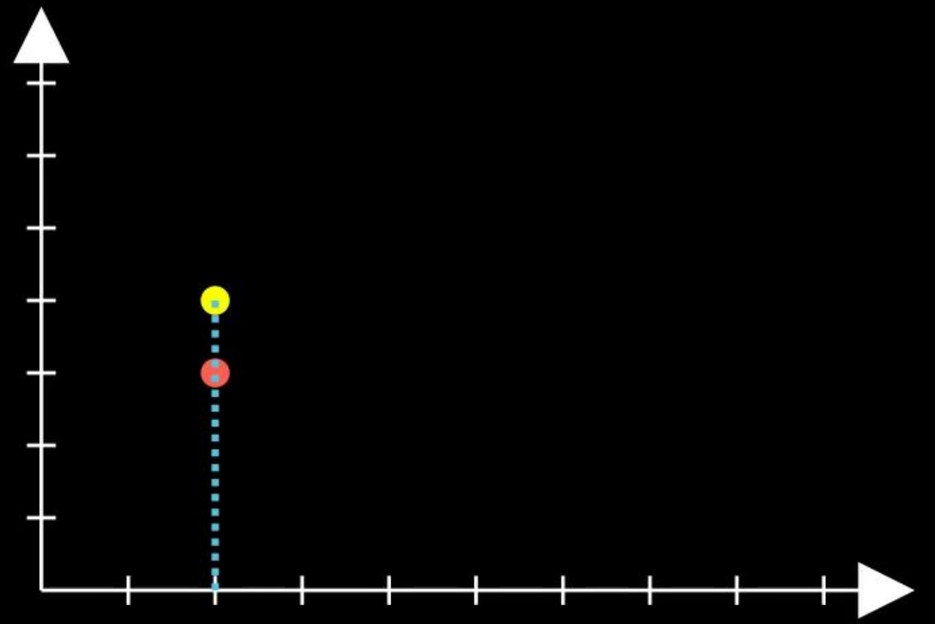


Conformal Prediction

Given a dataset of 1-D points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to compute valid prediction regions for a new point $\hat{y}_{n+1} = f(x_{n+1})$, for a given desired coverage probability $1 - \alpha$.

$$f(\ 2.0\ ) = \ 4.0$$

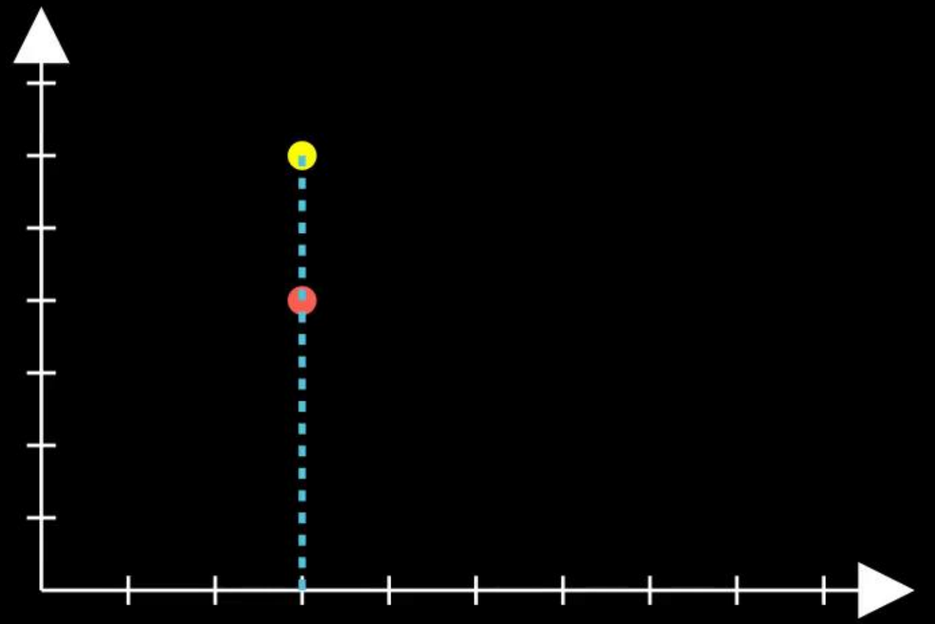
$$(\ 2.0\ ,\ 3.0\ )$$



Conformal Prediction

Given a dataset of 1-D points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to compute valid prediction regions for a new point $\hat{y}_{n+1} = f(x_{n+1})$, for a given desired coverage probability $1 - \alpha$.

$$f(\ 3.0\ ) = \ 6.0$$
$$(\ 3.0\ ,\ 4.0\ )$$

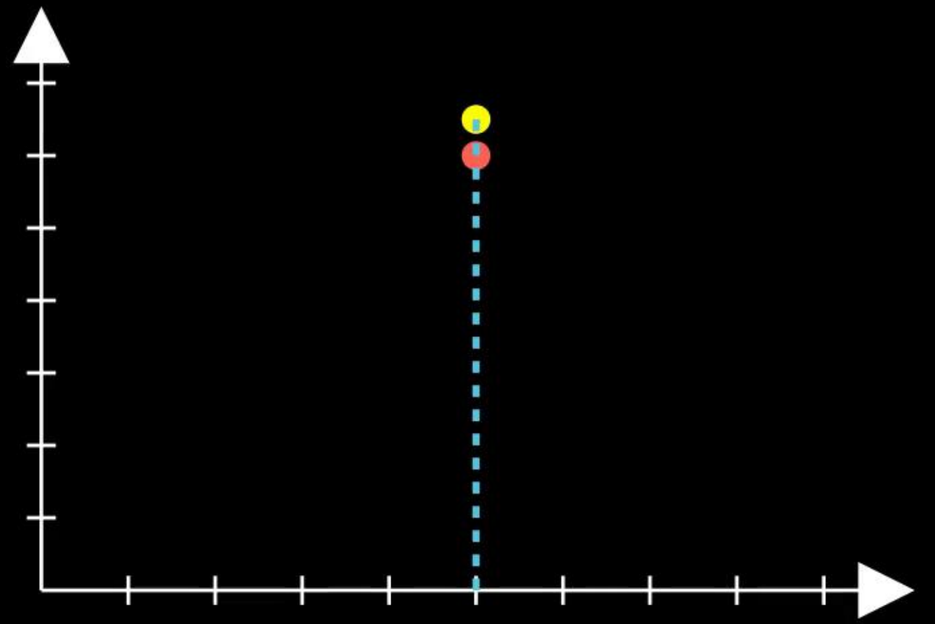


Conformal Prediction

Given a dataset of 1-D points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to compute valid prediction regions for a new point $\hat{y}_{n+1} = f(x_{n+1})$, for a given desired coverage probability $1 - \alpha$.

$$f(\ 5.0\ ) = \ 6.5$$

$$(\ 5.0\ ,\ 6.0\ )$$

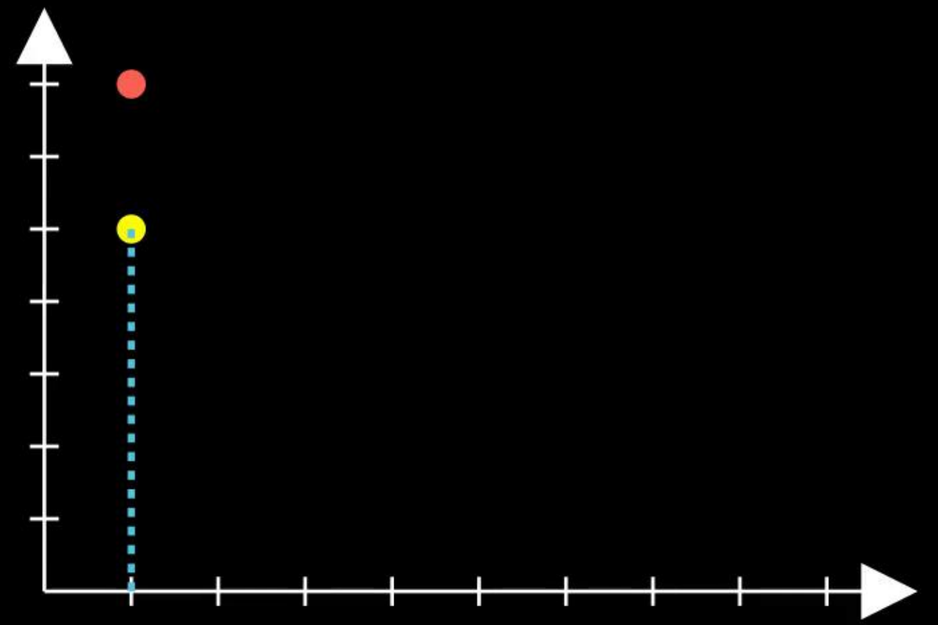


Conformal Prediction

Given a dataset of 1-D points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to compute valid prediction regions for a new point $\hat{y}_{n+1} = f(x_{n+1})$, for a given desired coverage probability $1 - \alpha$.

$$f(\ 1.0\ ) = \ 5.0$$

$$(\ 1.0\ ,\ 7.0\ )$$

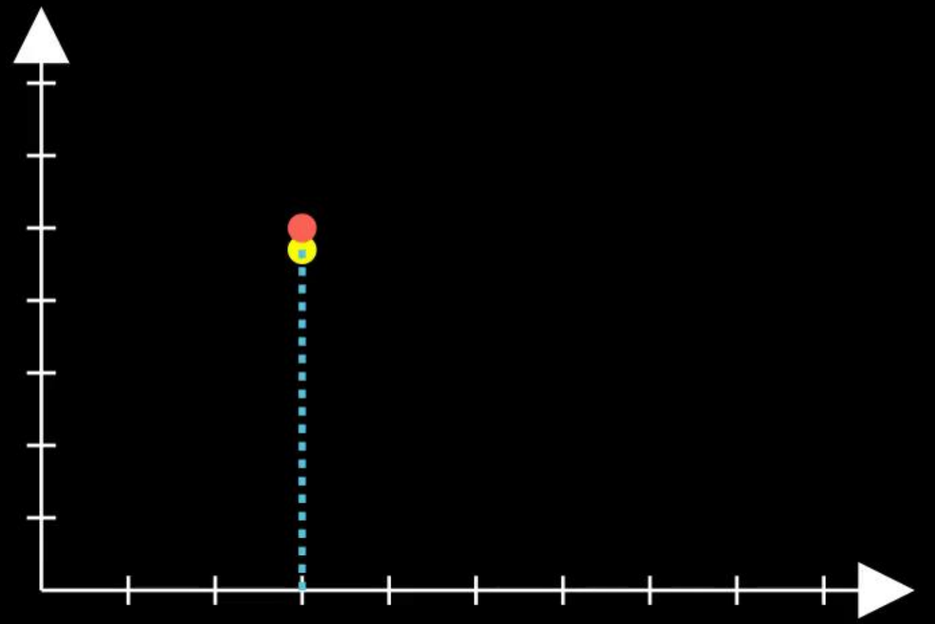


Conformal Prediction

Given a dataset of 1-D points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to compute valid prediction regions for a new point $\hat{y}_{n+1} = f(x_{n+1})$, for a given desired coverage probability $1 - \alpha$.

$$f(\ 3.0\ ) = \ 4.7$$

$$(\ 3.0\ ,\ 5.0\ )$$



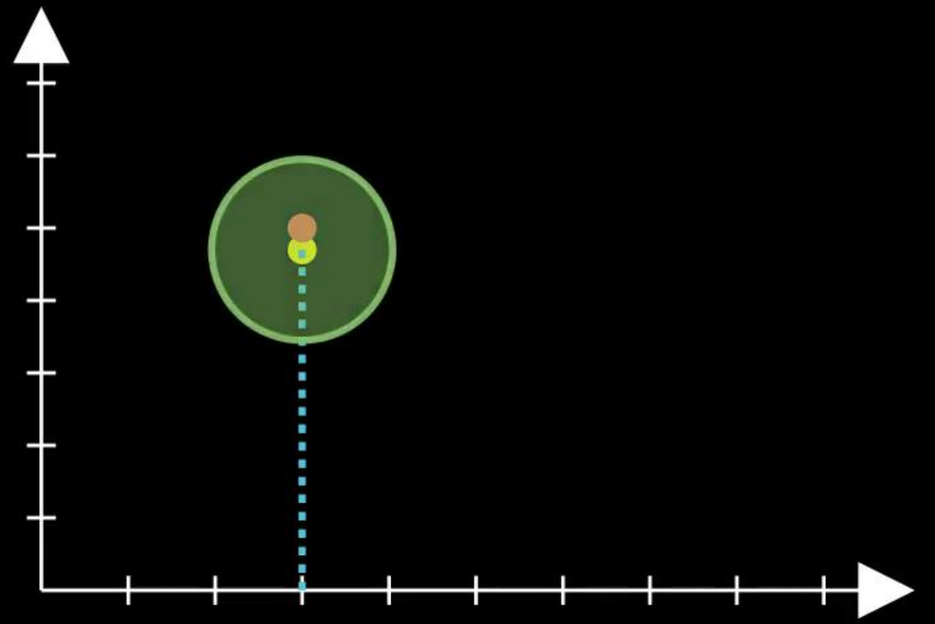
Conformal Prediction

Given a dataset of 1-D points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we want to compute valid prediction regions for a new point $\hat{y}_{n+1} = f(x_{n+1})$, for a given desired coverage probability $1 - \alpha$.

$$f(\ 3.0\ ) = \ 4.7$$

$$(\ 3.0\ ,\ 5.0\ )$$

$$\hat{\mathbf{y}} = C_f^{\alpha}(x)$$



Conformal Prediction

Split the dataset into a training set and a calibration set $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{cal}}$

Conformal Prediction

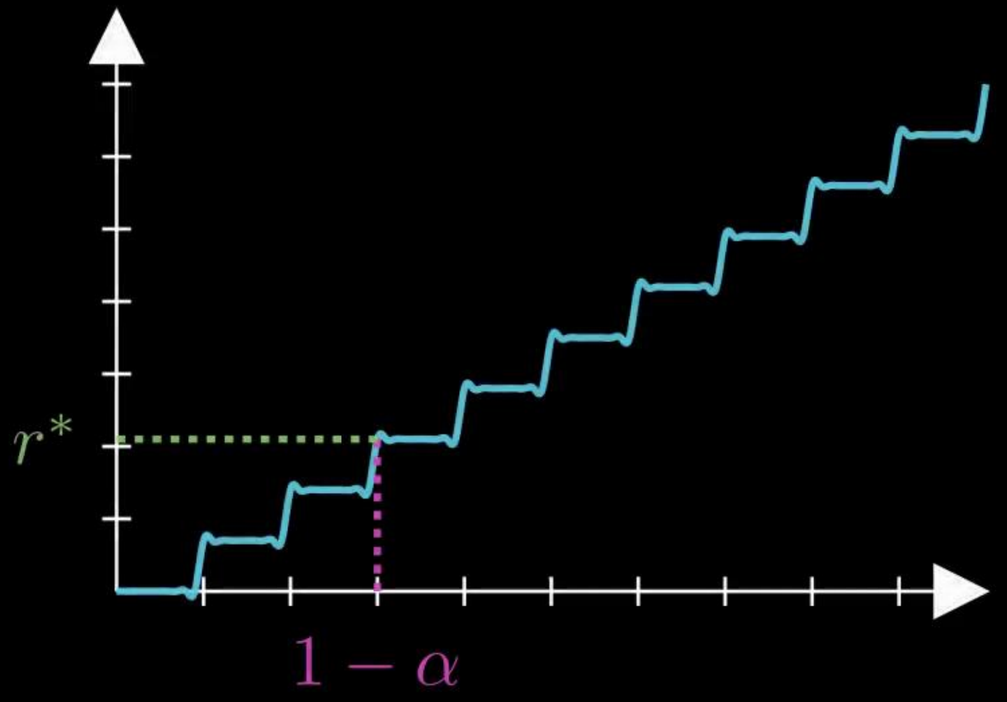
Split the dataset into a training set and a calibration set $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{cal}}$

Compute the residuals $r_i = |y_i - f(x_i)|$ for $i \in \mathcal{D}_{\text{cal}}$ and build their empirical quantiles

Conformal Prediction

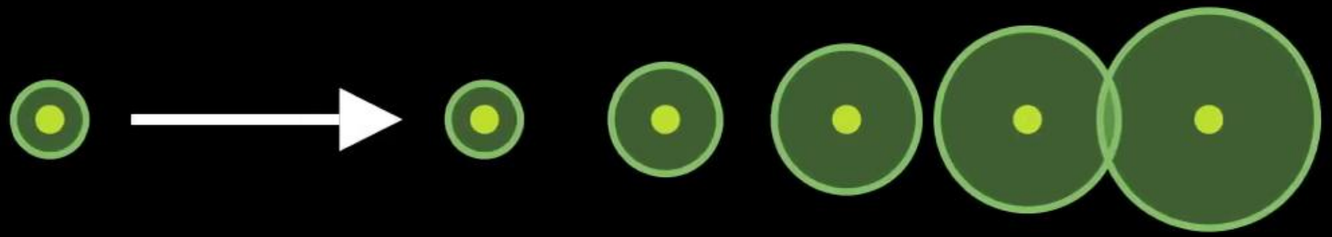
Split the dataset into a training set and a calibration set $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{cal}}$

Compute the residuals $r_i = |y_i - f(x_i)|$ for $i \in \mathcal{D}_{\text{cal}}$ and build their empirical quantiles

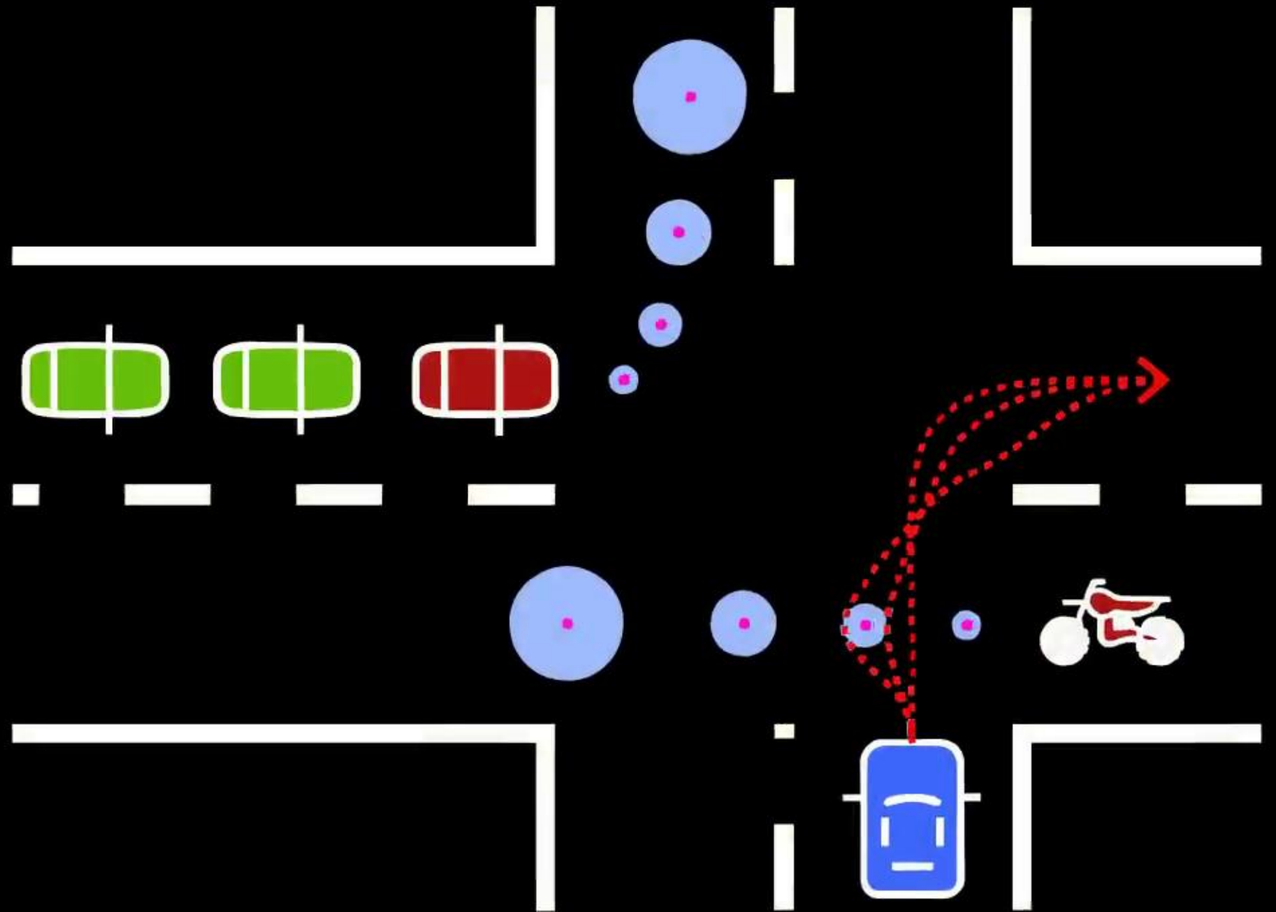


$$C_f^\alpha(x) = [f(x) - r^*, f(x) + r^*]$$

$$\mathbb{P}(y \in C_f^\alpha(x)) = 1 - \alpha$$

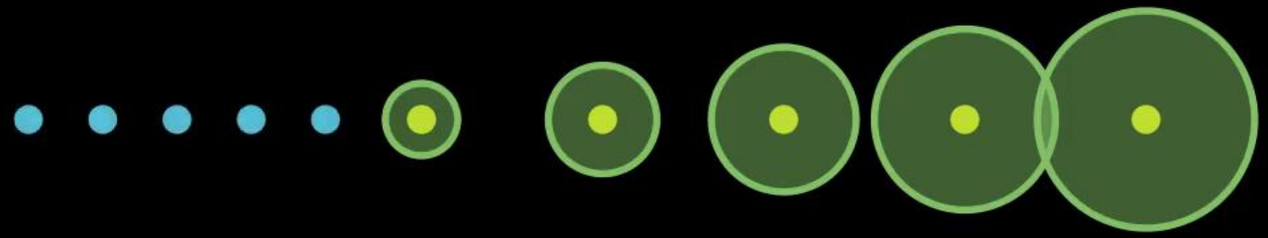


Given trajectory predictions in magenta, compute *valid* and *efficient* (i.e. tight) prediction regions in blue, for a given desired coverage probability $1 - \alpha$ (probability true trajectory is completely inside the blue regions).



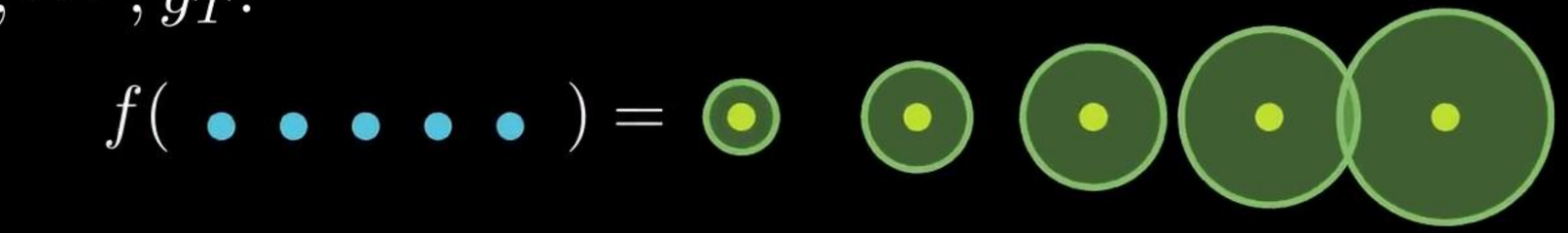
Problem Settings

Extending conformal prediction to multi-horizon forecasting. We consider we are given observations $y_1, \cdots, y_{T-H}$ and a predictor f which produces the predictions $\hat{y}_{T-H+1}, \cdots, \hat{y}_T$.



Problem Settings

Extending conformal prediction to multi-horizon forecasting. We consider we are given observations $y_1, \cdots, y_{T-H}$ and a predictor f which produces the predictions $\hat{y}_{T-H+1}, \cdots, \hat{y}_T$.



Problem Settings

Extending conformal prediction to multi-horizon forecasting. We consider we are given observations $y_1, \cdots, y_{T-H}$ and a predictor f which produces the predictions $\hat{y}_{T-H+1}, \cdots, \hat{y}_T$.

$$f(y_1, \cdots, y_{T-H}) = (\hat{y})_{i=T-H+1}^{T}$$

Problem Settings

Extending conformal prediction to multi-horizon forecasting. We consider we are given observations $y_1, \cdots, y_{T-H}$ and a predictor f which produces the predictions $\hat{y}_{T-H+1}, \cdots, \hat{y}_T$.

$$f(y_1, \ldots, y_{T-H}) = (\hat{y})_{i=T-H+1}^{T}$$

Compute valid prediction intervals $\hat{\mathbf{y}}_{\mathbf{T-H+1}}, \ldots, \hat{\mathbf{y}}_{\mathbf{T}}$ where validity is defined below:

$$\mathbb{P}\left(\bigcap_{i=T-H+1}^{T} (y_i \in \hat{\mathbf{y}}_{\mathbf{i}})\right) > 1 - \alpha$$

Problem Settings

Extending conformal prediction to multi-horizon forecasting. We consider we are given observations $y_1, \cdots, y_{T-H}$ and a predictor f which produces the predictions $\hat{y}_{T-H+1}, \cdots, \hat{y}_T$.

$$f(y_1, \ldots, y_{T-H}) = (\hat{y})_{i=T-H+1}^{T}$$

Compute valid prediction intervals $\hat{\mathbf{y}}_{\mathbf{T-H+1}}, \ldots, \hat{\mathbf{y}}_{\mathbf{T}}$ where validity is defined below:

$$\mathbb{P}\left(\bigcap_{i=T-H+1}^{T} (y_i \in \hat{\mathbf{y}}_{\mathbf{i}})\right) > 1 - \alpha$$

Use the mean interval size as performance metric.

Other Works

We have at our disposal a dataset $\mathcal{D}$ of sequences of length T which is i.i.d. with the observed data. This dataset is then split into a training set $\mathcal{D}_{train}$ and a calibration set $\mathcal{D}_{cal}$.

- A branch of the literature focuses on a dataset of past points and the predicted interval is just around a single prediction ($H = 1$). In this case, the guarantees are only asymptotic.
- The branch of the literature that shares the same setting is based on the work of *Stankeviciute et al. (2021)* (CF-RNN).

CF-RNN

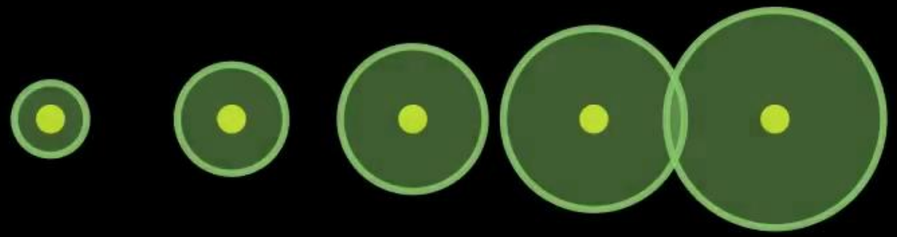
Each individual prediction interval is computed as follows:

$$\hat{\mathbf{y}}_i = C_{f_i}^{\alpha/H}(y_1, \ldots, y_{T-H})$$

CF-RNN

Each individual prediction interval is computed as follows:

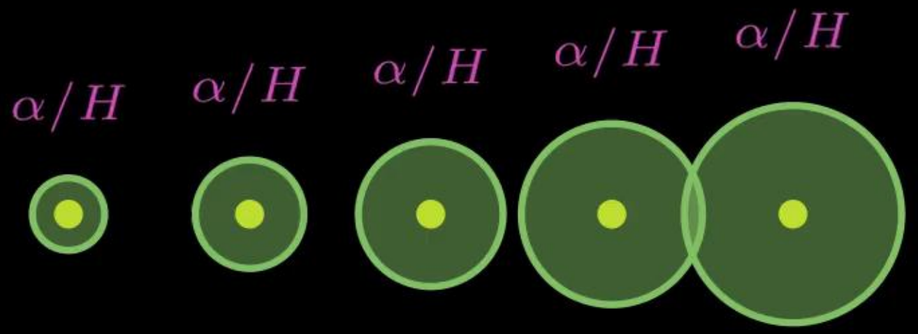
$$\hat{\mathbf{y}}_i = C_{f_i}^{\alpha/H}(y_1, \ldots, y_{T-H})$$



CF-RNN

Each individual prediction interval is computed as follows:

$$\hat{\mathbf{y}}_i = C_{f_i}^{\alpha/H}(y_1, \ldots, y_{T-H})$$



CF-RNN

Each individual prediction interval is computed as follows:

$$\hat{\mathbf{y}}_i = C_{f_i}^{\alpha/H}(y_1, \ldots, y_{T-H})$$

Which means that, by Boole's inequality, the probability of at least one error is at most α.

$$\mathbb{P}\left\{ \bigcup_{i=T-H+1}^{T} y_i \notin C_{f_i}^{\alpha/H}(y_1, \ldots, y_{T-H}) \right\} \leq \sum_{i=T-H+1}^{T} \frac{\alpha}{H} = \alpha$$

Though effective against other methods, CF-RNN introduces a significant approximation error, especially when there's a lot of dependence in time as the events are not disjoint.

CF-RNN

Each individual prediction interval is computed as follows:

$$\hat{\mathbf{y}}_i = C_{f_i}^{\alpha/H}(y_1, \ldots, y_{T-H})$$

Which means that, by Boole's inequality, the probability of at least one error is at most α.

$$\mathbb{P}\{A \cup B\} \leq \mathbb{P}(A) + \mathbb{P}(B)$$

Though effective against other methods, CF-RNN introduces a significant approximation error, especially when there's a lot of dependence in time as the events are not disjoint.

CF-RNN

Each individual prediction interval is computed as follows:

$$\hat{\mathbf{y}}_i = C_{f_i}^{\alpha/H}(y_1, \ldots, y_{T-H})$$

Which means that, by Boole's inequality, the probability of at least one error is at most α.

$$\mathbb{P}\left\{ \bigcup_{i=T-H+1}^{T} y_i \notin C_{f_i}^{\alpha/H}(y_1, \ldots, y_{T-H}) \right\} \leq \sum_{i=T-H+1}^{T} \frac{\alpha}{H} = \alpha$$

Though effective against other methods, CF-RNN introduces a significant approximation error, especially when there's a lot of dependence in time as the events are not disjoint.

ConForME

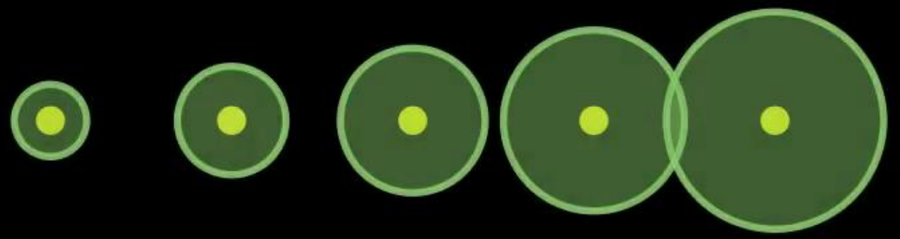
We group the prediction intervals into blocks of size b_j:

$$(\hat{\mathbf{y}}_{\mathbf{i}})_{i=T-H+1}^{T} = \Big(\underbrace{\hat{\mathbf{y}}_{\mathbf{T-H+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T-H+b_1}}}_{B_1} \underbrace{\cdots\cdots}_{B_j} \underbrace{\hat{\mathbf{y}}_{\mathbf{T-b_k+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T}}}_{B_k}\Big)$$

ConForME

We group the prediction intervals into blocks of size b_j:

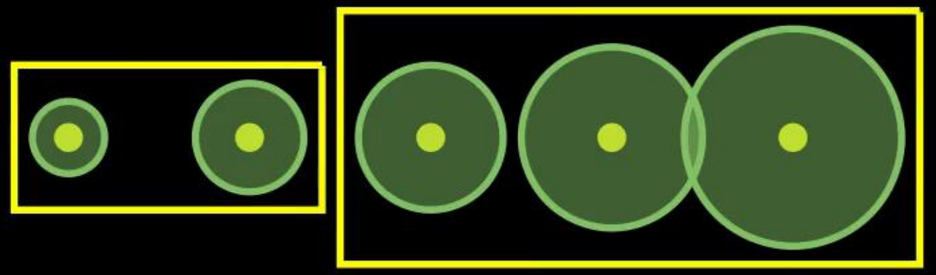
$$(\hat{\mathbf{y}}_{\mathbf{i}})_{i=T-H+1}^{T} = \big(\underbrace{\hat{\mathbf{y}}_{\mathbf{T-H+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T-H+b_1}}}_{B_1} \underbrace{\cdots\cdots}_{B_j} \underbrace{\hat{\mathbf{y}}_{\mathbf{T-b_k+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T}}}_{B_k}\big)$$



ConForME

We group the prediction intervals into blocks of size b_j:

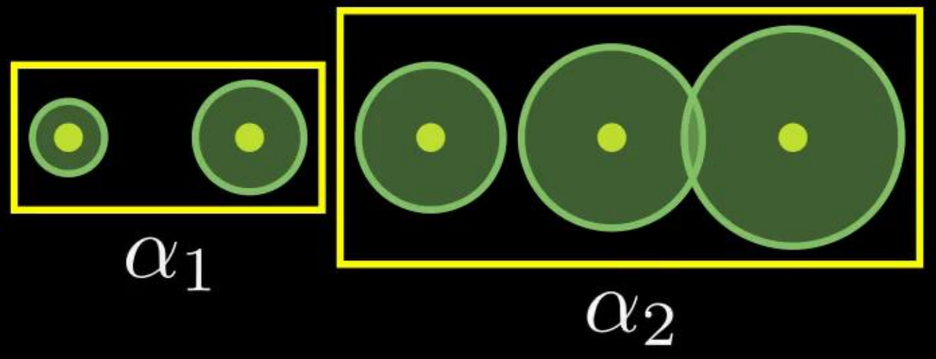
$$(\hat{\mathbf{y}}_\mathbf{i})_{i=T-H+1}^{T} = (\underbrace{\hat{\mathbf{y}}_{\mathbf{T-H+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T-H+b_1}}}_{B_1} \underbrace{\cdots\cdots}_{B_j} \underbrace{\hat{\mathbf{y}}_{\mathbf{T-b_k+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T}}}_{B_k})$$



ConForME

We group the prediction intervals into blocks of size b_j:

$$(\hat{\mathbf{y}}_\mathbf{i})_{i=T-H+1}^{T} = (\underbrace{\hat{\mathbf{y}}_{\mathbf{T-H+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T-H+b_1}}}_{B_1} \underbrace{\cdots\cdots}_{B_j} \underbrace{\hat{\mathbf{y}}_{\mathbf{T-b_k+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T}}}_{B_k})$$

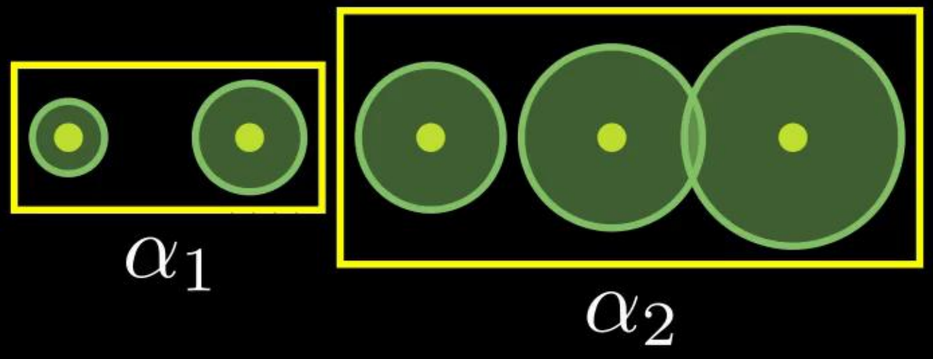


$$\alpha_1 + \alpha_2 = \alpha$$

ConForME

We group the prediction intervals into blocks of size b_j:

$$(\hat{\mathbf{y}}_\mathbf{i})_{i=T-H+1}^{T} = (\underbrace{\hat{\mathbf{y}}_{\mathbf{T-H+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T-H+b_1}}}_{B_1} \underbrace{\cdots\cdots}_{B_j} \underbrace{\hat{\mathbf{y}}_{\mathbf{T-b_k+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T}}}_{B_k})$$



$$\alpha_j > 0, \ \sum_{j=1}^{k} \alpha_j = \alpha$$

ConForME

We group the prediction intervals into blocks of size b_j:

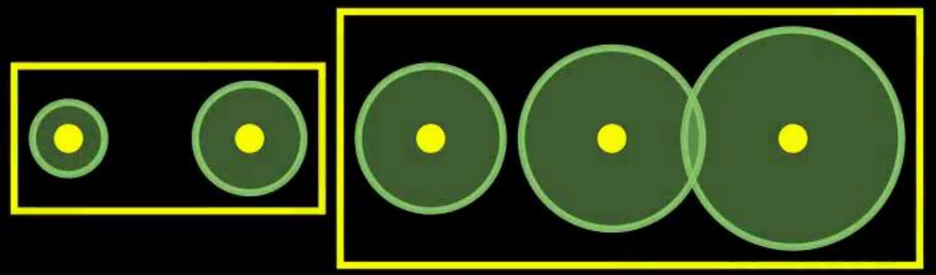
$$(\hat{\mathbf{y}}_{\mathbf{i}})_{i=T-H+1}^{T} = (\underbrace{\hat{\mathbf{y}}_{\mathbf{T-H+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T-H+b_1}}}_{B_1} \underbrace{\cdots\cdots}_{B_j} \underbrace{\hat{\mathbf{y}}_{\mathbf{T-b_k+1}} \cdots \hat{\mathbf{y}}_{\mathbf{T}}}_{B_k})$$

Consider separately validity in each block to enforce the validity as a whole:

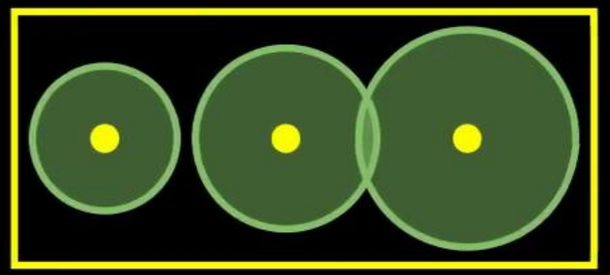
$$\mathbb{P}\left(\bigcup_{l=1}^{b_j} \left(y_{(l)^j} \notin \hat{\mathbf{y}}_{(\mathbf{l})^{\mathbf{j}}}\right)\right) \leq \alpha_j$$

$$\alpha_j > 0, \ \sum_{j=1}^{k} \alpha_j = \alpha$$

ConForME



ConForME



ConForME

Given $\mathcal{D}_{cal}$ with $(y_i)_{i=1}^{T} \in \mathcal{D}_{cal}$

ConFormE

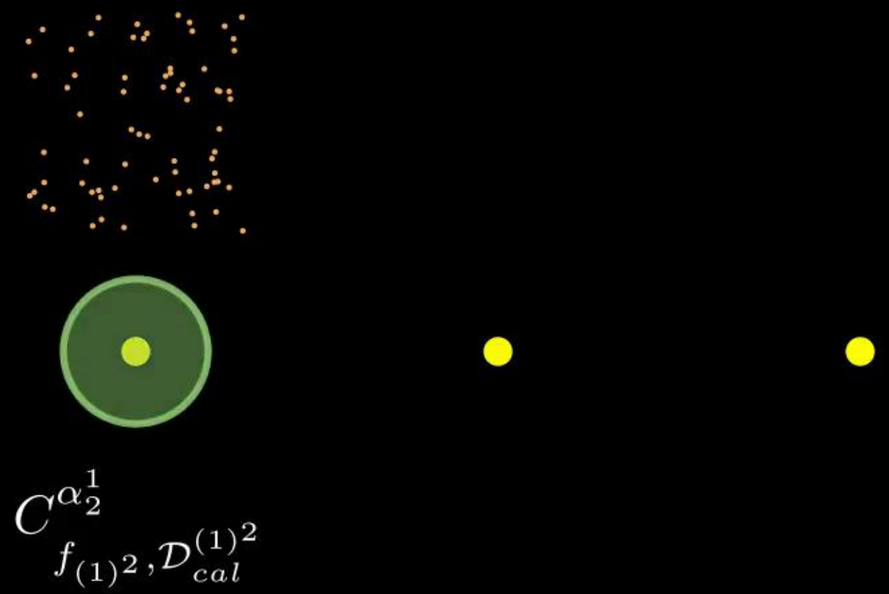
Given $\mathcal{D}_{cal}$ with  $\in \mathcal{D}_{cal}$

ConFormE

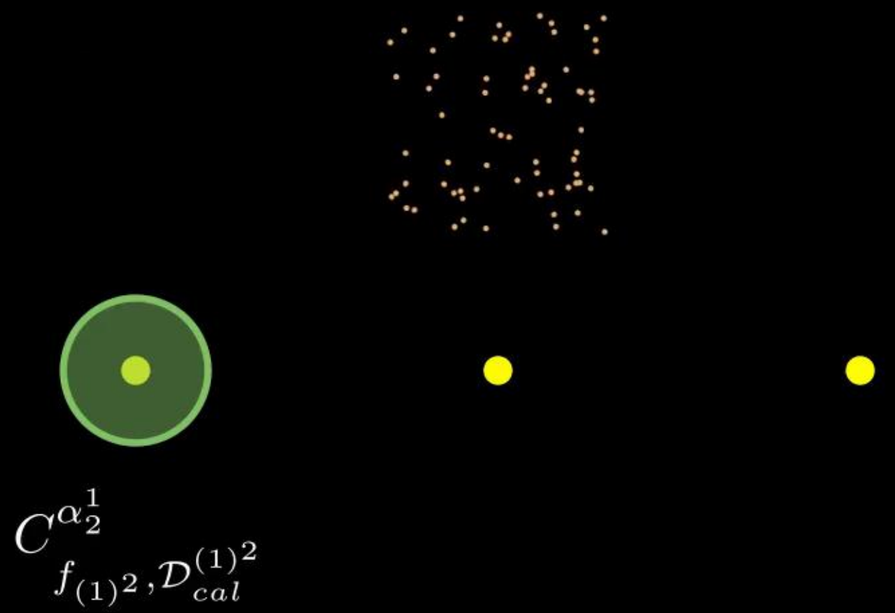
ConFormE

Given $\mathcal{D}_{cal}$ with  $\in$

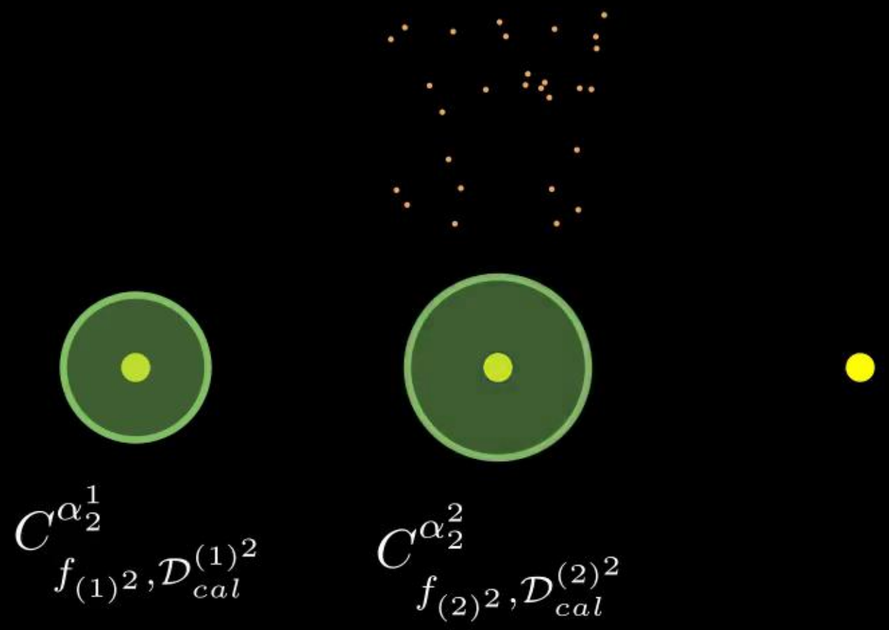
ConForME



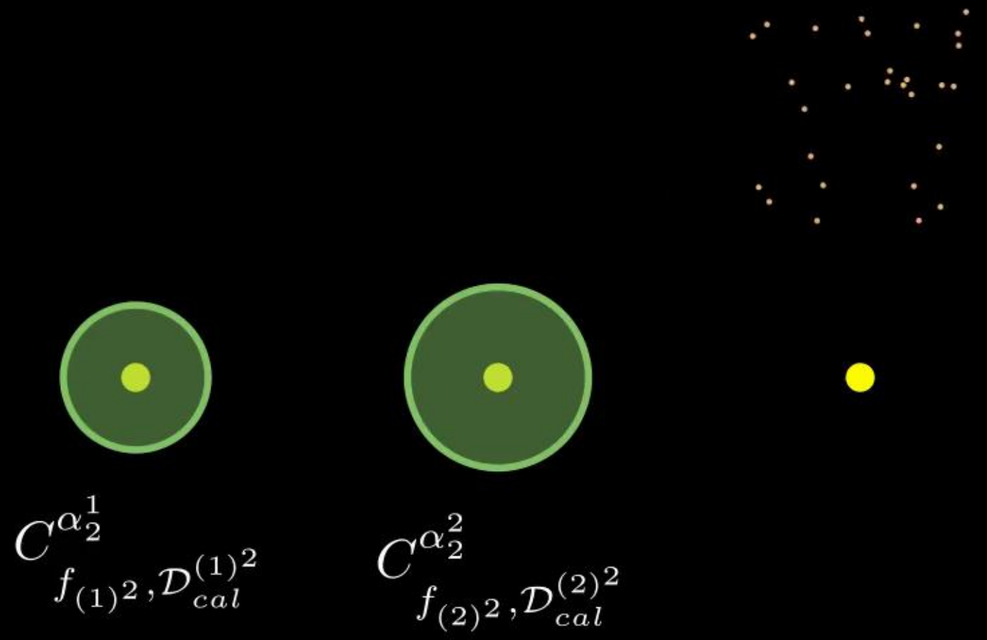
ConForME



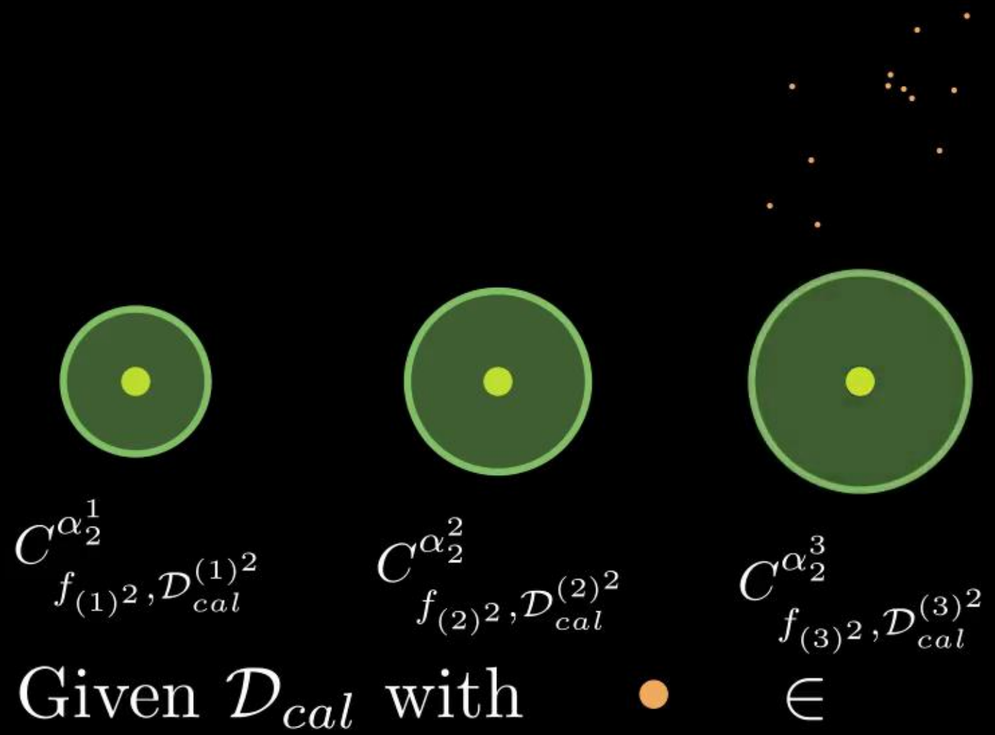
ConForME


$C^{\alpha_2^1}_{f_{(1)^2}, \mathcal{D}^{(1)^2}_{cal}}$
$C^{\alpha_2^2}_{f_{(2)^2}, \mathcal{D}^{(2)^2}_{cal}}$
Given $\mathcal{D}_{cal}$ with
$\in$

ConForME


$C^{\alpha_2^1}_{f_{(1)^2}, \mathcal{D}^{(1)^2}_{cal}}$
$C^{\alpha_2^2}_{f_{(2)^2}, \mathcal{D}^{(2)^2}_{cal}}$
Given $\mathcal{D}_{cal}$ with
$\in$

ConForME


$C^{\alpha_2^1}_{f_{(1)^2}, \mathcal{D}^{(1)^2}_{cal}}$
$C^{\alpha_2^2}_{f_{(2)^2}, \mathcal{D}^{(2)^2}_{cal}}$
$C^{\alpha_2^3}_{f_{(3)^2}, \mathcal{D}^{(3)^2}_{cal}}$
Given $\mathcal{D}_{cal}$ with $\in$

ConForME

We compute the intervals as:

$$\hat{\mathbf{y}}_{(l)^j} = C^{\alpha_j^l}_{f_{(l)^j}, \mathcal{D}_{cal}^{(l)^j}}(y_1, \ldots, y_{T-H})$$

where $\mathcal{D}_{cal}^{(l)^j} = \{(y_i) \in \mathcal{D}_{cal} \mid y_{(m)^j} \in \hat{\mathbf{y}}_{(\mathbf{m})^\mathbf{j}} \forall m \in ((1)^j, \ldots, (l)^j - 1)\}$

With:

$$\sum_{l=1}^{b_j} \alpha_j^l = \alpha_j$$

ConForME

How to choose the block sizes b_j and the α_j^l's? We propose three methods:

Evenly distributed blocks:

$$\alpha_j^l = \frac{\alpha}{H}$$

Pairwise evenly distributed:

$$\alpha_j^l = \beta \frac{\alpha}{\lceil H/2 \rceil} \text{ if } l \text{ is odd, } \alpha_j^l = (1-\beta)\frac{\alpha}{\lceil H/2 \rceil} \text{ otherwise}$$

Optimized:

$$loss = mean\_size\left(conforme\left(\left(\alpha_j^l\right)\right), k, \left(b_j\right)\right) + \lambda \cdot tan\left(\sum_{j=1}^{k}\sum_{l=1}^{b_j}\alpha_j^l - \alpha\right)$$

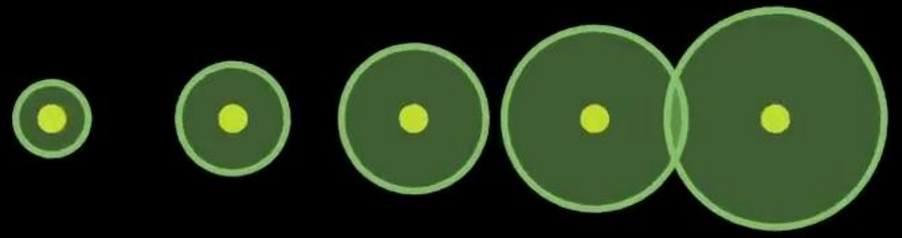
ConForME

How to choose the block sizes b_j and the α_j^l's? We propose three methods:

Evenly distributed blocks:

$$\alpha_j^l = \frac{\alpha}{H}$$

Pairwise evenly distributed:



Optimized:

$$loss = mean\_size\left(conforme\left(\left(\alpha_j^l\right)\right), k, \left(b_j\right)\right) + \lambda \cdot tan\left(\sum_{j=1}^{k}\sum_{l=1}^{b_j}\alpha_j^l - \alpha\right)$$

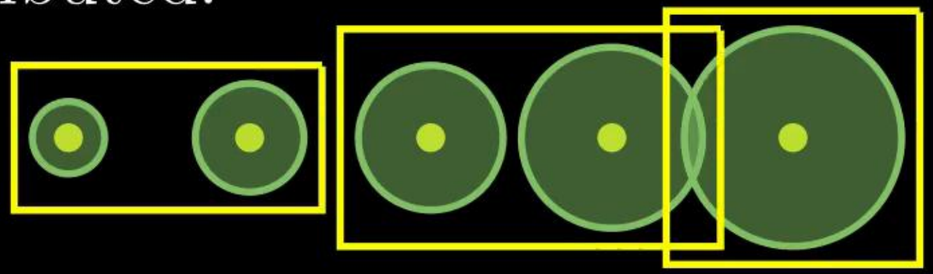
ConForME

How to choose the block sizes b_j and the α_j^l's? We propose three methods:

Evenly distributed blocks:

$$\alpha_j^l = \frac{\alpha}{H}$$

Pairwise evenly distributed:



Optimized:

$$loss = mean\_size\left(conforme\left(\left(\alpha_j^l\right)\right), k, \left(b_j\right)\right) + \lambda \cdot tan\left(\sum_{j=1}^{k}\sum_{l=1}^{b_j}\alpha_j^l - \alpha\right)$$

ConForME

How to choose the block sizes b_j and the α_j^l's? We propose three methods:

Evenly distributed blocks:

$$\alpha_j^l = \frac{\alpha}{H}$$

Pairwise evenly distributed:

$$\alpha_j^l = \beta \frac{\alpha}{\lceil H/2 \rceil} \text{ if } l \text{ is odd, } \alpha_j^l = (1-\beta)\frac{\alpha}{\lceil H/2 \rceil} \text{ otherwise}$$

Optimized:

$$loss = mean\_size\left(conforme\left(\left(\alpha_j^l\right)\right), k, \left(b_j\right)\right) + \lambda \cdot tan\left(\sum_{j=1}^{k}\sum_{l=1}^{b_j}\alpha_j^l - \alpha\right)$$

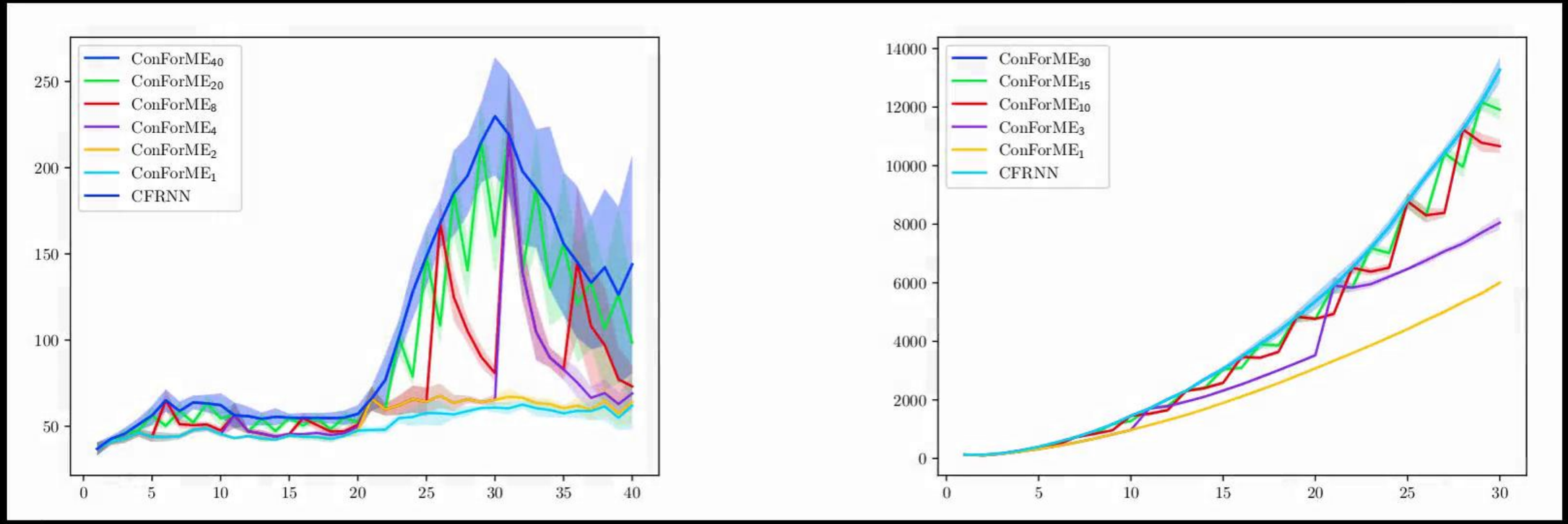
Results

Comparison with CF-RNN on the following datasets:

Dataset	Description	Size
Synthetic	Programatically generated data	2500
EEG	Electroencephalograms from visual stimuli	38400
Argoverse	Car trajectories	218693
COVID-19	Covid cases in different regions	380

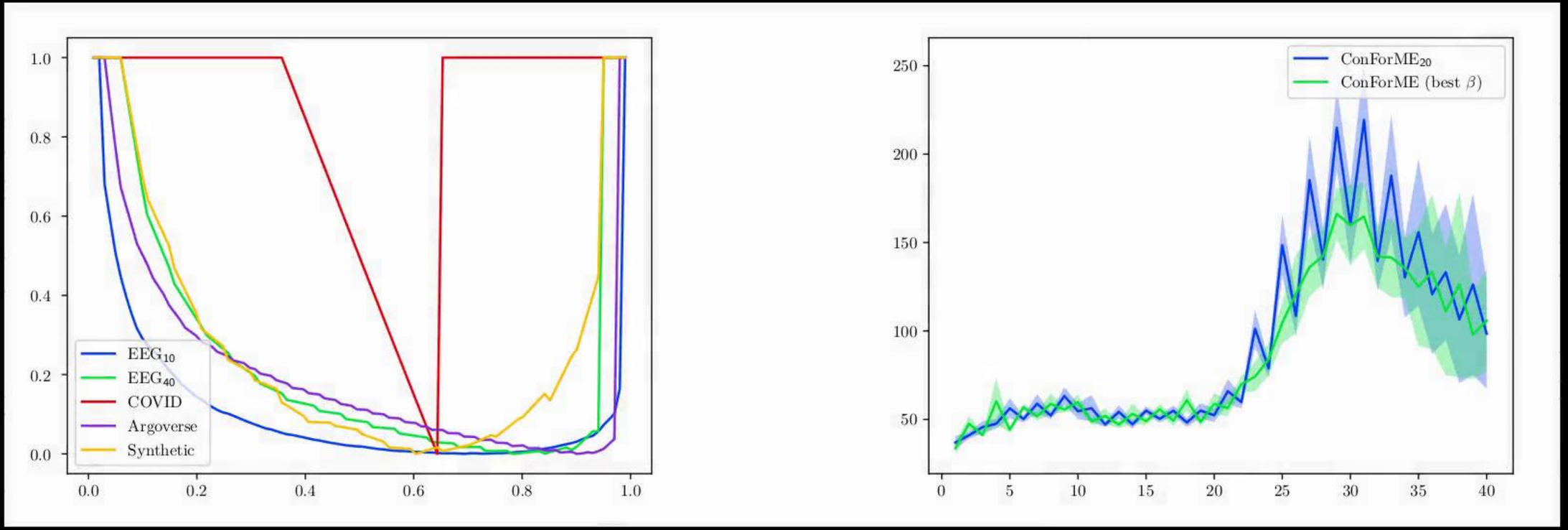
Results

Up to 52% smaller interval sizes on the EEG data, at least 35% smaller intervals on real world data, 9.7% smaller intervals on synthetic data. The image below shows the interval sizes for EEG 40 (left) and Argoverse (right) datasets.



Results

For the pairwise evenly distributed method, we study the effect of β on the left. On the right, for the EEG_{40} dataset, we compare the mean interval sizes per horizon for the optimal β with ConFormE_{20}.



Future Work

- Integrate my method with planning in real-time.
- Better understand the hyperparameters choice: use better algorithm for optimal hyperparameter choice.
- Prove that pairwise evenly distributed can be always computed efficiently.