UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Conformal prediction and beyond

Author:
Gerard CASTRO CASTILLO

Supervisor:
Dr. Jordi VITRIÀ

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

July 11, 2024

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Conformal prediction and beyond

by Gerard CASTRO CASTILLO

The role of uncertainty quantification (UQ) has become indispensable with the advent of artificial intelligence and its application to the decision-making. This thesis leverages conformal prediction (CP) as its cornerstone, a pivotal methodology in the field of distribution-free and model-agnostic UQ, which stems from the notion of "*conformalizing*" predictions to data using the residuals to understand the errors distribution.

In particular, in this work some strategies within the CP approach are theoretically justified, and its guarantees and limitations presented. Even though the CP paradigm was classically applied only under "*data exchangeability*" conditions, this work also reviews some of the most recent and non-trivial efforts to enable CP when this hypothesis is not fulfilled.

Lastly, to practically demonstrate CP ability to provide prediction intervals with statistically valid coverage, different strategies are successfully applied both to a tabular data regression problem and to a time series forecasting problem.

Acknowledgements

I would like to express my deepest gratitude to my thesis supervisor, Jordi Vitrià, for his invaluable guidance and mentorship throughout the development of this thesis. His early direction and thoughtful corrections have been instrumental in shaping this work, providing me with continuous insights that deeply enhanced my understanding and execution of the project.

Special thanks are extended to my family, and in particular to my mother, Mercè. Their unwavering support, patience, and boundless love have been the cornerstone of my strength and perseverance throughout this academic journey. Their belief in my capabilities has been a constant source of encouragement.

I am also immensely grateful to my friends, who have supported me unconditionally. Their encouragement, through both kind words and endless supplies of cheap coffees, along with their patience in listening to my frequent explanations, have made this journey not only bearable but also enjoyable.

This thesis would not have been possible without the contributions and support of each one of you. Thank you.

Contents

Chapter 1

Introduction

In the modern era of data-driven decision-making, the need for uncertainty quantification has increased across various disciplines: finance, autonomous driving, medicine... and any other high-stake application in which there is a huge error cost. Understanding and quantifying uncertainty is not merely about acknowledging the limits of predictive models, but about enhancing the reliability of decisions made based on these.

At its core, uncertainty quantification enables us to assess the confidence in the predictions made by models and to comprehend the potential variability in these predictions.

Formally, let us suppose we have $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables from which n samples $(X_i, Y_i)_{i=1}^n$ were obtained; and assume unknown both the X and Y marginal underlying probability distributions, as well as its joint distribution[1]. Given a new sample X_{n+1} and a miscoverage level $\alpha \in [0, 1]$, we would like to estimate a predictive interval $\mathcal{C}_\alpha$ such that the probability of Y_{n+1} falling into $\mathcal{C}_\alpha$ is at least $1 - \alpha$, *i.e.*

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

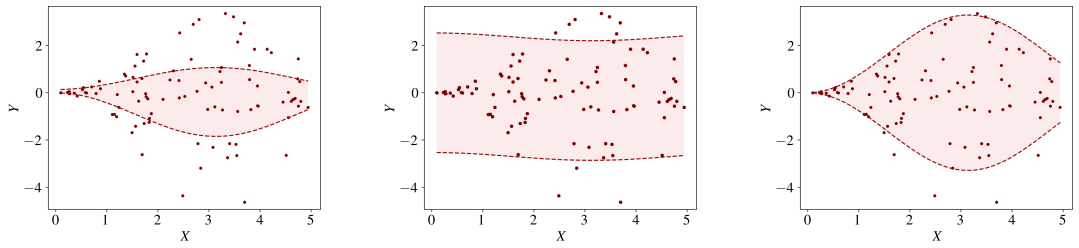
Note that, while retaining statistical coverage, $\mathcal{C}_\alpha$ should be as small as possible to be informative. Furthermore, the intervals should be as much adaptive as possible, seeking for conditional coverage as upper limit (contrarily to non-adaptive marginal coverage). These are defined as:

- **Marginal** coverage: $\mathbb{P}\left\{Y_{n+1} \in \hat{\mathcal{C}}_\alpha (X_{n+1})\right\}$ the errors may differ across regions of the input space (*i.e.* non-adaptive)
- **Conditional** coverage: $\mathbb{P}\left\{Y_{n+1} \in \hat{\mathcal{C}}_\alpha (X_{n+1}) \mid X_{n+1}\right\}$ errors are evenly distributed (*i.e.* fully adaptive)

Obviously, conditional coverage is stronger than marginal coverage as it allows to better represent the underlying phenomena, see Figure 1.1. However, as it will be discussed in 2.5.4, without any distributional assumption conditional coverage cannot be guaranteed; see Lei and Wasserman, 2014; Vovk, 2012 and Foygel Barber et al., 2020.

Now, in the general setup, these predictions intervals $\mathcal{C}_\alpha$ should be valid in finite samples and agnostic not only to the data distribution but also to the model used to predict $\hat{Y}_{n+1}$.

[1] Throughout this work, the distribution-free case will be considered: *i.e.* no prior knowledge of the variables' underlying probability distributions is assumed and, thus, all Bayesian approaches will be disregarded.



(A) No coverage (B) Marginal coverage (C) Conditional coverage

FIGURE 1.1: Different types of coverage. Plots based in Zaffran, 2023.

Within this data-agnostic setup, in section 1.1 quantile regression is reviewed as a historical baseline; then, in section 1.2 conformal prediction will be presented as means of producing $\mathcal{C}_\alpha$ with statistical valid coverage.

1.1 Quantile regression

Introduced by Koenker and Bassett, 1978, quantile regression focuses on estimating either boundaries of the distribution (quantiles) for a regression problem. Unlike classical linear regression, which predicts dependent variable's median given independent variables; quantile regression predicts a quantile, providing a more comprehensive analysis of possible outcome distributions.

Essentially, given the quantile level $\beta \in [0,\ 1]$, one can find a estimator for β by adapting the loss function using the so-called *pinball loss*:

$$\ell_\beta\left(Y, Y'\right) = \beta\left|Y - Y'\right| \mathbb{1}_{\{|Y-Y'|\geq 0\}} + (1-\beta)\left|Y - Y'\right| \mathbb{1}_{\{|Y-Y'|\leq 0\}}$$

Thence, during the fitting process of a estimator μ_β for the quantile β - where $Y \sim \mu\left(X\right)$ -, the following loss function can be defined:

$$L_{\mu_\beta}(X, Y) := \mathbb{E}\left[\ell_\beta\left(Y, \hat{\mu_\beta}(X)\right)\right],$$

as the loss function.

As discussed by Zaffran, 2023, if $\hat{\mu} \in \operatorname{argmin}_\mu L_{\mu_\beta}$, then $\hat{\mu} \equiv Q_{X|Y}(\beta) := \inf\{x \in \mathbb{R}, \mathbb{P}(X \leq x|Y) \geq \beta\}$ is the β quantile function.

Thus, the naive approach to obtain a predictive interval with α miscoverage level ($1-\alpha$ confidence) could be to fit the dependent variable's distribution, in terms of the independents X, with this adapted loss function and, then, prescribe:

$$\mathcal{C}_\alpha(X_{n+1}) = \left[\hat{\mu}_{\frac{\alpha}{2}}\left(X_{n+1}\right),\ \hat{\mu}_{1-\frac{\alpha}{2}}\left(X_{n+1}\right)\right],$$

for a new sample X_{n+1}.

However, since $\mu_{\frac{\alpha}{2}}$ & $\mu_{1-\frac{\alpha}{2}}$ have been trained just with $\{(X_i, Y_i)_{i=1}^n\}$, it could happen $\mathcal{C}_\alpha$ is under/over-confident out of training. Namely, for a finite sample, there is no theoretical guarantee $\mathcal{C}_\alpha$ has statistical valid coverage:

$$\mathbb{P}\left(Y_{n+1} \in \left[\hat{\mu}_{\frac{\alpha}{2}}\left(X_{n+1}\right),\ \hat{\mu}_{1-\frac{\alpha}{2}}\left(X_{n+1}\right)\right]\right) \neq 1-\alpha$$

In conclusion, though providing an approximate measure of the target's variability, quantile regression (QR) does not inherently constitute a framework for providing statistically valid predictive sets.

1.2 Conformal prediction

Initially developed in the early 2000s by Vladimir Vovk, Alexander Gammerman, and Glenn Shafer; conformal prediction (CP) was born as a framework for producing statistically valid, data-agnostic & distribution-free predictive sets.

Heuristically, CP is based on the idea of using the samples data, not only to train the estimator μ, but also to "conformalize" the model with the data so that the predictive sets attain the expected coverage (unlike in QR 1.1).

In this sense, CP is based on the idea of using this "past" (sample) data to determine how the model errors are distributed and, thus, its conformity to the "reality". This is measured in terms of the so-called conformity score.

Notice, nevertheless, CP requires at least one assumption to ensure its validity. Since the model conformity measure is evaluated on the same samples from which the training set is created, the data needs to be **exchangeable** (*i.e.* any permutation of samples should not affect the joint distribution). A thorough explanation and theoretical justification of CP, along the different methodologies to implement CP (with options for trading-off computational or statistical efficiency), is presented in chapter 2.

Chapter 2

Conformal prediction

Leveraging same notation as in chapter 1, and given $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables with unknown marginal and joint probability distributions, we want to obtain a $\mathcal{C}_\alpha$ predictive interval. Let us address the case in which $Y = \mu(X) + \epsilon$, where μ is the model function to be determined and $\epsilon_i \sim P_{Y|X}$ the noise. From now on, we will only assume the data to be **exchangeable**.

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

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

where $\mathcal{L}$ designates the joint distribution.

Example 1. Independent identically distributed (*i.i.d*) samples, or the components of a multidimensional normal distribution, are exchangeable data.

Example 2. Time series data, as well as data obtained from a random variable under any kind of distribution shift, are examples of non-exchangeable data.

Furthermore, as mentioned in section 1.2, conformal prediction is ultimately based in using the so-called *conformality scores* to construct the predictive interval. These scores, $s_{\hat{\mu}}$ or directly s, allow to transform a heuristic notion of uncertainty from a model $\hat{\mu}$ into a rigorous measure of it.

Formally, any function $s(X, Y) \in \mathbb{R}$ can be chosen as score if it returns larger values the worse the agreement between X and Y is. Note the choice of conformity score will determine the way confidence intervals are built. In particular, the simplest choice is the absolute residual score (namely, the residual) as score for the regression problems: $s_i := s_{\hat{\mu}}(X_i, Y_i) = |Y_i - \hat{\mu}(X_i)|$.

In this case, then, the predictive intervals is built as

$$\hat{\mathcal{C}_\alpha} = [\hat{\mu}(X) - q(\mathcal{S}),\ \hat{\mu}(X) + q(\mathcal{S})],$$

where $q(\mathcal{S})$ is the $1 - \alpha$ empirical quantile of the conformity scores $\mathcal{S} = \{s_i\}_i$.

Note. The *conformity score* election can play a pivotal role in the uncertainty quantification process. In particular, useless or no informative intervals can be obtained in function of the chosen score, as explained in Angelopoulos and Bates, 2021.

Example 3. There are other well-known scores, for instance those implemented by MAPIE developers, 2024:

- Gamma score $s_{\hat{\mu}}(X, Y) := \frac{|Y - \hat{\mu}(X)|}{\hat{\mu}(X)}$, such that:

$$\hat{\mathcal{C}_\alpha} = [\hat{\mu}(X)\,(1 - q(s))\,,\ \hat{\mu}(X)\,(1 + q(s))]$$

- Residual normalized score $s_{\hat{\mu}}(X,Y) := \frac{|Y-\hat{\mu}(X)|}{\hat{\sigma}(X)}$, where $\hat{\sigma}(X)$ is another model which predicts residuals from X (trained on $(X, |Y-\hat{\mu}(X)|)$), and it is such that:

$$\hat{\mathcal{C}}_\alpha = [\hat{\mu}(X) - q(s)\hat{\sigma}(X),\ \hat{\mu}(X) + q(s)\hat{\sigma}(X)]$$

Within the same conformity score setup, however, several approaches for CP can be applied in function of the user needs regarding computational and statistical efficiency. In this sense, from sections 2.1 to 2.3, different CP flavours will be presented; while, in section 2.4, a way of "*conformalizing*" the quantile regression method is reviewed. Finally, section 2.5 is devoted the theoretical guarantees and limits of CP and its flavours.

2.1 Split Conformal Prediction

Split Conformal Prediction (SCP) is the most widely-used flavour of conformal prediction and it is based in splitting the data in a training Tr set and a calibration Cal set. Thus, if we used the absolute residual as conformity score, SCP would prescribe as it follows:

1. Split the data set into a **training set** of size #Tr and a **calibration set** of size #Cal
2. Obtain $\hat{\mu}$ by training the algorithm in the Tr set
3. Obtain a set $\mathcal{S}$ of conformity scores by using the Cal set: $\mathcal{S}_{\text{Cal}} := \{s_i\}_{i\in\text{Cal}} = \{|Y_i - \hat{\mu}(X_i)|, i \in \text{Cal}\}$
4. Compute $q_{1-\alpha^{\text{SCP}}}(\mathcal{S}_{\text{Cal}})$, namely the $(1-\alpha)\left(\frac{1}{\#\text{Cal}}+1\right)$ quantile of $\mathcal{S}_{\text{Cal}}$. From now on, we will indistinctly write $q_{1-\alpha^{\text{SCP}}}(\mathcal{S}_{\text{Cal}}) \equiv q_{1-\alpha}(\mathcal{S}_{\text{Cal}})$
5. For a new sample X_{n+1}, return the predictive interval

$$\hat{\mathcal{C}}_\alpha = [\hat{\mu}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}_{\text{Cal}}),\ \hat{\mu}(X_{n+1}) + q_{1-\alpha}(\mathcal{S}_{\text{Cal}})] \tag{2.1}$$

Note. To attain at least $1-\alpha$ coverage taking into account the finite number of samples in the Cal set, in step 4 the quantile of conformity scores $q_{1-\alpha^{\text{SCP}}}(\mathcal{S}) \equiv q_{1-\alpha}(\mathcal{S})$ must be computed as a $(1-\alpha)\left(\frac{1}{\#\text{Cal}}+1\right)$-quantile (instead of a $(1-\alpha)$-quantile), the so-called $1-\alpha$ **empirical quantile**.

Let us, in Algorithm 1, state the algorithm independently of the conformity score s choice.

Notice SCP requires that one must have enough observations to split its original dataset into train and calibration, but at least it attains the expected coverage of $\geq 1-\alpha$. This will be theoretically backed later at section 2.5.1, particularly through Theorem 1.

2.2 Full Conformal Prediction

Even though SCP attains expected coverage and just needs to fit a model μ once (thus, it is not computationally demanding), the dataset needs to be large enough for $\mathcal{C}_\alpha$ to be informative.

Full Conformal Prediction (FCP) is born as a workaround to this problem. Unlike SCP, FCP leverages the whole dataset as training and its core idea is as stated at

Algorithm 1 SCP algorithm

Input: Regression algorithm $\mathcal{A}$, miscoverage level α, data samples $\{(X_t, Y_t)\}_{t=1}^{T}$.
Output: Prediction interval $\mathcal{C}_\alpha(X)$ for any $X \in \mathbb{R}^d$.

1: Randomly split $\{1, \ldots, T\}$ into two disjoint sets Tr and Cal.
2: Fit a mean regression function: $\hat{\mu}(\cdot) \leftarrow \mathcal{A}\left(\{(X_t, Y_t), t \in \text{Tr}\}\right)$.
3: **for** $j \in$ Cal **do**
4: Set s_j the *conformity scores*.
Note if we chose *e.g.* the absolute residual as conformity score, then $s_j := s_{\hat{\mu}}(X_j, Y_j) = |Y_j - \hat{\mu}(X_j)|$.
5: **end for**
6: Set $\mathcal{S}_{\text{Cal}} = \{s_j, j \in \text{Cal}\}$.
7: Compute $q_{1-\alpha}\left(\mathcal{S}_{\text{Cal}}\right)$, the $(1-\alpha)\left(1 + 1/|\text{Cal}|\right)$ quantile of $\mathcal{S}_{\text{Cal}}$ (*i.e.* the $1-\alpha$ empirical quantile).
8: Return $\mathcal{C}_\alpha(X) = \{Y \in \mathbb{R} \mid s_{\hat{\mu}}(X, Y) \leq q_{1-\alpha}(\mathcal{S}_{\text{Cal}})\}$, for any $X \in \mathbb{R}^d$.
Note the explicit form of $\mathcal{C}_\alpha(x)$ depends on the conformality score; *e.g.* if we chose the absolute residual, then $\mathcal{C}_\alpha(X) = [\hat{\mu}(X) \pm q_{1-\alpha}\left(\mathcal{S}_{\text{Cal}}\right)]$.

Angelopoulos and Bates, 2021.

Let us assume the dataset $\{(X_t, Y_t)\}_{t=1}^{T}$ is available and, for a new sample X_{T+1}, the user wants to provide the interval $\mathcal{C}_\alpha(X_{T+1})$. Then, since the true label Y_{T+1} lies somewhere in $\mathcal{Y} := \text{Im}(\mu) \subset \mathbb{R}$, looping over all possible $Y \in \mathcal{Y}$ will eventually hit in the (X_{T+1}, Y_{T+1}) data point which is exchangeable with the first T points; specifically, the most probable labels will have a low enough conformity score.

FCP, as explained in Algorithm 2, consists in: "*discretizing*" the target space $\mathcal{Y}$ into N candidates Y_j, traversing this loop fitting the estimator to the data & (X_{T+1}, Y_j) as new data point, and finally returning those candidates Y_j such that they "conform enough" to the data. The latter condition will be translated to asserting whether the conformity score of Y_j candidate is "low enough" (lower than the $1-\alpha$ empirical quantile of all conformity scores $\mathcal{S}$).

Note FCP solves the problem of effectively reducing the dataset through a split, at expenses of a huge computational efforts. In particular, FCP requires to re-fit the model N times for every new X_{T+1} feature sample.
Of course, the existence of N is due to the fact we need to discretize $\mathcal{Y}$ to compute $\mathcal{C}_\alpha(X_{T+1})$. In this sense, the larger N the more accurate FCP will be, but then the more time it will take to infer the predictive set.
Notice that in the non-discrete case (*i.e.* when $N \to +\infty$) we would be returning the continuous set

$$\mathcal{C}_\alpha(X_{T+1}) = \{Y \in \mathbb{R} \mid s_{\hat{\mu}_j}(X_{T+1}, Y) \leq q^j_{1-\alpha}(\mathcal{S}_j)\} .$$

2.3 Other flavours

On the one hand, in 2.1 it is shown split conformal prediction requires only one model fitting step, but sacrifices statistical efficiency. On the other hand, in 2.2 is reviewed how full conformal prediction requires a very large number of model fitting steps, but has high statistical efficiency. These are not the only two achievable points

Algorithm 2 FCP algorithm

Input: Regression algorithm $\mathcal{A}$, miscoverage level α, data samples $\{(X_t, Y_t)\}_{t=1}^{T}$ and new X_{T+1} feature sample.
Output: Prediction interval $\mathcal{C}_\alpha(X_{T+1})$ for any given $X_{T+1} \in \mathbb{R}^d$.

1: Discretize the target space $\mathcal{Y}$ reducing into N candidates Y_j.
2: Initialize $\hat{\mathcal{Y}}_{\text{low}} = \{\}$ the array for candidates with "low enough" conformity score
3: **for** $j \in \{1, \ldots, N\}$, Y_j candidate **do**
4: Fit a mean regression function μ_j using $\text{Tr}_j = \{(X_t, Y_t)\}_{t=1}^{T} \cup \{(X_{T+1}, Y_j)\}$ as training data:

$$\hat{\mu}_j(\cdot) \leftarrow \mathcal{A}\left(\{(X_t, Y_t), t \in \text{Tr}_j\}\right) .$$

5: Set $\mathcal{S}_j = \{s_{\hat{\mu}_j}(X_i, Y_i)\}_{i=1}^{T} \cup \{s_{\hat{\mu}_j}(X_{T+1}, Y_j)\}$ the conformity scores obtained in the same Tr_j training data.
6: Set $q^j_{1-\alpha}(\mathcal{S}_j)$ the $(1-\alpha)\left(1+\frac{1}{T+1}\right)$ quantile of $\mathcal{S}_j$ (*i.e.* the $1-\alpha$ empirical quantile)
7: **if** $s_{\hat{\mu}_j}(X_{T+1}, Y_j) \leq q^j_{1-\alpha}(\mathcal{S}_j)$ **then**
8: Add Y_j to $\hat{\mathcal{Y}}_{\text{low}}$
9: **end if**
10: **end for**
11: Return $\mathcal{C}_\alpha(X_{T+1}) = \hat{\mathcal{Y}}_{\text{low}}$.

on the spectrum: but there are techniques that precisely fall in between, trading off statistical efficiency and computational efficiency differently.

As represented in Figure 2.1, this is the case of cross-conformal prediction CV+ (Vovk, 2015) and Jackknife+ (Barber et al., 2021), methods which both use a small number of model fits, but still leveraging all data for both model fitting and calibration.

On the one hand, the so-called Jackknife+ method is based on leave-one-out (LOO) and, for a new X_{n+1}, it heuristically prescribes:

1. For each $i \in \mathcal{D} := \{1, \ldots, T\}$ sample of the training data:

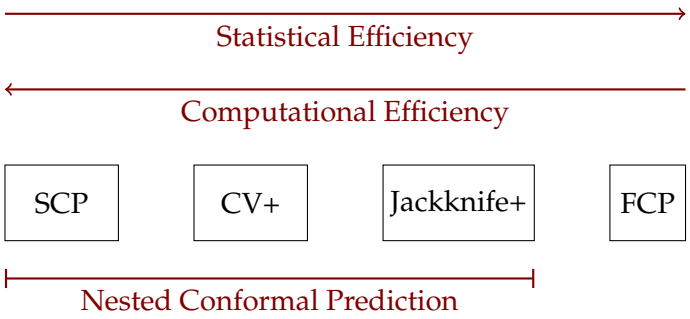


FIGURE 2.1: Representation of the trade-off between statistical and computational efficiencies for the different approaches. Based on Zaffran, 2023.

- Fit a mean regression function μ_{-i} training $\mathcal{A}$ in $\mathcal{D} \setminus (X_i, Y_i)$:

$$\hat{\mu}_{-i}(\cdot) \leftarrow \mathcal{A}\left(\{(X_t, Y_t), t \in \mathcal{D} \setminus (X_i, Y_i)\}\right) .$$

- Get the conformity scores

$$\mathcal{S}^i_{\text{up/down}} = \hat{\mu}_{-i}(X_{n+1}) \pm s_{\hat{\mu}_{-i}}(X_i, Y_i)$$

2. Set the conformity scores' sets: $\mathcal{S}_{\text{up}} = \{\mathcal{S}^i_{\text{up}}\}_{i\in\mathcal{D}}$ and $\mathcal{S}_{\text{down}} = \{\mathcal{S}^i_{\text{down}}\}_{i\in\mathcal{D}}$
3. Defining $q_{\beta,\text{inf}}(Z_1, \ldots, Z_n)$ as the $\lfloor \beta \times n \rfloor$ smallest value of $(Z_1, \ldots, Z_n)$, and $q_{1-\alpha}$ the $1-\alpha$ empirical quantile; the following predictive interval is returned:

$$\hat{\mathcal{C}}_\alpha(X_{n+1}) = \left[q_{\alpha,\text{inf}}(\mathcal{S}_{\text{down}}),\ q_{1-\alpha}(\mathcal{S}_{\text{up}})\right]$$

Note, however, J+aB may be more computationally demanding even than FCP if the dataset is large such $T > N$ (more samples T than N points needed to *discretize* the $\mathcal{Y}$ space in FCP).

On the other hand, the CV+ method is based on cross-validation residuals and precisely extends the previous idea into a "batch" of samples. Thence, the following differences must be taken into account:

- Instead of leaving one out of $\mathcal{D}$, CV+ splits the data into K folds $F_1, \ldots, F_K$.
- Then, for each $k \in \{1, \ldots, K\}$ fold:
  - A mean regression function μ_{-F_K} is fit training $\mathcal{A}$ in $\mathcal{D} \setminus F_k$, instead of $\mathcal{D} \setminus (X_i, Y_i)$.
  - The conformity scores are no longer a value (for each $i \in \{1, \ldots, N\}$) but a subset (obtained with samples within each fold k):

$$\mathcal{S}^k_{\text{up/down}} = \{\hat{\mu}_{-k}(X_{n+1}) \pm s_{\hat{\mu}_{-k}}(X_i, Y_i)\}_{i\in F_k}$$

Notice this method enhances the computational efficiency at expenses of the statistical's, by training $\mathcal{A}$ less times (and with less data). For a precise and complete description of the algorithms, we refer the reader to the former works (Vovk, 2015 and Barber et al., 2021).

2.4 Conformalized Quantile Regression

While the former flavours of conformal prediction have theoretical guarantees for its statistical coverage, just marginal coverage is sought thus providing non-adaptive predictive intervals:

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

In this section, thence, we present the Conformalized Quantile Regression (CQR) as means of obtaining more adaptive intervals $\mathcal{C}_\alpha$. As discussed in 2.5.4, notice that while approximate and asymptotic conditional coverage can be sought, conformal prediction (even within CQR) does not guarantee it without further assumptions than data exchangeability. Nevertheless, at practice, CQR will allow us to obtain more informative intervals.

CQR proposes to fit not one μ but two models μ_{down} and μ_{up} and adjusting the conformality score s_μ to:

$$s_{\mathcal{A}}(X_i, Y_i) := \max\left(\hat{\mu}_{\text{down}}(X_i) - Y_i,\ Y_i - \hat{\mu}_{\text{up}}(X_i)\right) \tag{2.2}$$

The model μ_{down} and μ_{up} are no longer trying to capture the mean value of (X, Y), but rather the low and high quantiles of their distribution (*e.g.* μ_{down} & μ_{up} could be estimators of the $\alpha/2$ and $1-\alpha/2$ quantiles). In this sense, CQR achieve adaptive predictive intervals by adding/subtracting the conformity score to the inferred values of 2 different estimators.

Thus, CQR is not a different methodology but another "choice" of conformity score definition and it is compatible with the SCP or FCP approaches (or others such as Jackknife+, CV+...).

Let us see, as an example in Algorithm 3, what CQR prescribes if we use the SCP approach:

Algorithm 3 CQR algorithm (using split prediction)

Input: Regression algorithm $\mathcal{A}$, miscoverage level α, data samples $\{(X_t, Y_t)\}_{t=1}^{T}$.
Output: Prediction interval $\mathcal{C}_\alpha(X)$ for any $X \in \mathbb{R}^d$.

1: Randomly split $\{1, \ldots, T\}$ into two disjoint sets Tr and Cal.
2: Fit 2 regression functions, one for the lower quantile $\hat{\mu}_{\text{down}}$ and one for the upper $\hat{\mu}_{\text{up}}$:

$$\hat{\mu}_{\text{down}}(\cdot), \hat{\mu}_{\text{up}}(\cdot) \leftarrow \mathcal{A}\left(\{(X_t, Y_t), t \in \text{Tr}\}\right) .$$

3: **for** $j \in$ Cal **do**
4: Set s_j the *conformity score* using 2.2:

$$s_j := s_{\mathcal{A}}(X_j, Y_j) = \max\left(\hat{\mu}_{\text{down}}(X_i) - Y_i,\ Y_i - \hat{\mu}_{\text{up}}(X_i)\right)$$

5: **end for**
6: Set $\mathcal{S}_{\text{Cal}} = \{s_j, j \in \text{Cal}\}$.
7: Compute $q_{1-\alpha}(\mathcal{S}_{\text{Cal}})$, the $(1-\alpha)(1+1/|\text{Cal}|)$ quantile of $\mathcal{S}_{\text{Cal}}$ (*i.e.* the $1-\alpha$ empirical quantile).
8: Return $\mathcal{C}_\alpha(X_{n+1}) = \left[\hat{\mu}_{\text{down}}(X_{n+1}) - q_{1-\alpha}(\mathcal{S}_{\text{Cal}}),\ \hat{\mu}_{\text{up}}(X_{n+1}) + q_{1-\alpha}(\mathcal{S}_{\text{Cal}})\right]$ for any new $X_{n+1} \in \mathbb{R}^d$.

2.5 Theoretical guarantees and limits

Even though the heuristic notion justifying the validity of the different paradigms of conformal prediction may be clear, we have not discussed its theoretical guarantees and limits yet. Yet, this is precisely the objective in this section: while in 2.5.1 and 2.5.2 we will proof $1-\alpha$ coverage can be sought in SCP and FCP respectively, in 2.5.3 we will discuss which are the guarantees for Jackknife+ and CV+, and we will conclude by reviewing the known limits of conformal prediction in 2.5.4.

2.5.1 Split Conformal Prediction

Henceforth, in Theorem 1 and through Lemma 1, the theoretical guarantee of SCP (algorithm 1) is proven. We reproduce the proof found at Zaffran, 2023, which is general for any conformity score s choice; however, note that:

- The proof for the lower bound initially appeared at Papadopoulos et al., 2002 for *i.i.d.* data. Then, Vovk, Gammerman, and Shafer, 2005 proved that the theorem also holds if the observations satisfy the weaker condition of exchangeability.
- The upper bound case was initially proved with Theorem 2.2 of Lei, G'Sell, et al., 2018 (along the lower bound case, see A.1) but specifically for the absolute residual score case. The proof required the residuals to have a continuous joint distribution, but as mentioned in Angelopoulos and Bates, 2021 this condition is not important because the user can always add a vanishing amount of random noise to the score.
- Both lower and upper bound cases were later stated for the CQR case in Theorem 1 of Romano, Patterson, and Candès, 2019 (and proved in its supplementary material).

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

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

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

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

Proof. First note that $U_{n+1} \leq q_\beta\left(U_1, \ldots, U_n, +\infty\right) \iff U_{n+1} \leq q_\beta\left(U_1, \ldots, U_n, U_{n+1}\right)$. Then, by definition of q_β :

$$U_{n+1} \leq q_\beta\left(U_1, \ldots, U_n, U_{n+1}\right) \iff \operatorname{rank}\left(U_{n+1}\right) \leq \lceil \beta(n+1) \rceil$$

By exchangeability, $\operatorname{rank}\left(U_{n+1}\right) \sim \mathcal{U}\{1, \ldots, n+1\}$. Thus:

$$\mathbb{P}\left(\operatorname{rank}\left(U_{n+1}\right) \leq \lceil \beta(n+1) \rceil\right) \geq \frac{\lceil \beta(n+1) \rceil}{n+1} \geq \beta.$$

If $U_1, \ldots, U_n, U_{n+1}$ are almost surely distinct (without ties):

$$\begin{aligned}\mathbb{P}\left(\operatorname{rank}\left(U_{n+1}\right) \leq \lceil \beta(n+1) \rceil\right) &= \frac{\lceil \beta(n+1) \rceil}{n+1} \\ &\leq \frac{1+\beta(n+1)}{n+1} = \beta + \frac{1}{n+1}.\end{aligned}$$

□

Theorem 1 (SCP guarantees). *Suppose $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable. SCP applied on $(X_i, Y_i)_{i=1}^{n}$ yields $\hat{\mathcal{C}}_\alpha(\cdot)$ such that:*

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

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

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

Proof. When $(X_i, Y_i)_{i=1}^{n+1}$ are exchangeable, the scores $\{S_i\}_{i \in \text{Cal}} \cup \{S_{n+1}\}$ are also exchangeable. Applying Lemma 1 to the scores concludes the proof. □

2.5.2 Full Conformal Prediction

Even though it leverages all data samples, FCP also attains the expected $1-\alpha$ statistical coverage and the idea in which the proof relies is really similar to the SCP proof (in particular, it strongly relies on the exchangeability of the s_{n+1} conformity score *w.r.t.* $s_1, \ldots, s_n$).

However, an additional hypothesis is needed and, although related to exchangeability, consists in the algorithm $\mathcal{A}$ being *symmetrical*.

Definition 2 (Symmetrical algorithm)**.** A deterministic algorithm $\mathcal{A} : (U_1, \ldots, U_n) \to \hat{A}$ is symmetric if, for any permutation σ of $[1, n]$:

$$\mathcal{A}(U_1, \ldots, U_n) \overset{\text{a.s.}}{=} \mathcal{A}\left(U_{\sigma(1)}, \ldots, U_{\sigma(n)}\right) .$$

With this new restraint, the following theorem can now be announced:

Theorem 2 (FCP guarantees)**.** *Suppose* $(X_i, Y_i)_{i=1}^{n+1}$ *are exchangeable and* $\mathcal{A}$ *is symmetric. FCP applied on* $(X_i, Y_i)_{i=1}^{n} \cup \{X_{n+1}\}$ *yields* $\hat{\mathcal{C}}_\alpha(\cdot)$ *such that:*

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

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

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

For sake of length and in view of the similarity of FCP proof *w.r.t.* to SCP's, Theorem 2 proof will be skipped. Nevertheless, for a detailed version, the reader can refer to Vovk, Gammerman, and Shafer, 2005 or Lei, G'Sell, et al., 2018 (A.1 appendix proof of Theorem 2.1).

2.5.3 Other flavours

Whereas SCP sacrificed statistical efficiency at expenses of computational efforts by splitting the available dataset in two, FCP proposed the contrary leveraging all the data but at the price of much more model fits; and, so far, we have seen these both two opposite methodologies guarantee $1-\alpha$ statistical coverage.

In this section, we briefly discuss the state of 2 approaches in-between the spectrum: the Jackknife+ & the CV+.

In this sense, in general, it is seen the Jackknife+ method does not guarantee $1-\alpha$ statistical coverage. Nevertheless, as proved in Barber et al., 2021, if $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable and $\mathcal{A}$ symmetric, then:

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

Yet, Barber et al., 2021 proposes the so-called "*jackknife-minmax*" method to attain $1-\alpha$ coverage within the jackknife setting; in practice, it is seen the yielded prediction intervals are too conservative.

We can refer to the same former reference to find out the CV+ method, proposed in Vovk, 2015, is another case in which $1-\alpha$ coverage is no attained without further assumptions.

Actually, within the same hypothesis of $\{(X_i, Y_i)\}_{i=1}^{n+1}$ being exchangeable and $\mathcal{A}$ symmetric, one can prove that, for CV+:

$$\mathbb{P}\left(Y_{n+1} \in \hat{\mathcal{C}}_\alpha(X_{n+1})\right) \geq 1 - 2\alpha - \min\left(\frac{2\,(1-1/K)}{n/K+1}, \frac{1-K/n}{K+1}\right) \geq 1 - 2\alpha - \sqrt{2/n}$$

It is beyond the scope of this work offer complete proof for these results; however, to obtain an exhaustive overview, the user is referred to Barber et al., 2021.

2.5.4 Impossibility results

So far, it has been shown SCP (and FCP) just requires data exchangeability (as well as $\mathcal{A}$ to be symmetrical) in order to achieve guaranteed $1-\alpha$ statistical coverage.

Of course, even though useful and informative predictive intervals can be obtained (precisely, this is the purpose of CP), the type of coverage we can expect is marginal rather than conditional.

This is precisely the impossibility result Lei and Wasserman, 2014; Vovk, 2012 and Foygel Barber et al., 2020 have reached to: without distribution assumption, in finite sample, a perfectly conditionally valid $\hat{\mathcal{C}}_\alpha$ is such that $\mathbb{P}\left\{\text{mes}\left(\hat{\mathcal{C}}_\alpha(x)\right) = \infty\right\} = 1$ for any non-atomic x. Namely, the measure of $\hat{\mathcal{C}}_\alpha(x)$ will be, with all probability, infinite.

However, for certain problems and in function of the sample size, approximate or even asymptotic conditional coverage can be reached. Below, several of these results are cited so that the user can refer to them:

- **Approximate conditional coverage**: $\mathbb{P}\left(Y_{n+1} \in \hat{\mathcal{C}}_\alpha \mid X_{n+1} \in \mathcal{R}(x)\right) \geq 1 - \alpha$ for certain $\mathcal{R}(x)$ clusters or aggregations of the features samples.
  - Romano, Sesia, and Candès, 2020: in classification problems, approximate conditional coverage can be sought for specialized versions of Jackknife+ and CV+ techniques and categorical & unordered response labels.
  - Guan, 2022: additional local coverage guarantees (under suitable assumptions) by offering a single-test-sample adaptive construction that emphasizes a local region around this test sample.
  - Jung et al., 2022: CP algorithms are proposed in order to attain multivalid coverage on exchangeable data in the batch setting (a little bit stronger than conditional coverage on group membership).
  - Gibbs, Cherian, and Candès, 2023: reformulates conditional coverage as coverage over a class of covariate shifts and, when the target class of shifts is finite dimensional, finite sample coverage over all possible shifts is achieved.
- **Asymptotic (with the sample size) conditional coverage**:
  - Kivaranovic, Johnson, and Leeb, 2020: a neural network is proposed so that it outputs three values instead of a single point estimate and optimizes a loss function motivated by the standard quantile regression loss, achieving stronger coverage.

- Izbicki, Shimizu, and R. Stern, 2020 and Izbicki, Shimizu, and R. B. Stern, 2022: both introduce some conformal methods based on conditional density estimators to obtain asymptotic conditional coverage; most based on the idea of creating prediction bands (in a data-driven way) locally on a partition of the features space.
- Chernozhukov, Wüthrich, and Zhu, 2021: a method is proposed to construct conditionally valid prediction intervals based on models for conditional distributions (such as quantile and distribution regression).
- Sesia and Romano, 2021: a conformal method is proposed to compute prediction intervals for non-parametric regression that can automatically adapt to skewed data (with probably marginal coverage in finite samples, while asymptotically achieving conditional coverage if the model is consistent).

Chapter 3

Beyond exchangeability

As discussed in chapters 1 & 2, the (data) exchangeability assumption is crucial as it ensures the error rates of the predictions are controlled across all possible partitions of the data. However, real-world data often challenge this assumption, presenting scenarios where exchangeability is not preserved.

In particular, we can distinguish two principal cases where exchangeability fails: distribution shifts and auto-correlation.

On one hand, distribution shifts occur when the process generating the test data differs from the process that generated the training data. These shifts can significantly impact the performance of predictive models, not only by rendering the previously learned patterns obsolete or less effective; but also in the CP case, by affecting the distribution of error rates.

For instance, some common types of shifts are:

- **Covariate shift**: changes in the input features' distribution.
- **Label shift**: changes in the target variable's distribution.

On the other hand, when there is auto-correlation between samples, training and calibration samples might be similar in such a way the model error rates is not evenly distributed. For instance, in the case of time series and their temporal nature of data.

In this chapter, we delve into the application of CP in such contexts. In particular, in sections 3.1 and 3.2 we review the heuristic ideas on how to make CP to work under shifts in data distribution (covariate & label shifts, respectively). Then, in section 3.3, we present how can CP be adapted to work with time series data.

3.1 Covariate shift

Covariate shift describes the phenomenon by which the input features' distribution has suffered changes $P_X \to \tilde{P}_X$.

Thus, within this case, we can formally state the following setting:

- $\{(X_i, Y_i)\}_{i=1}^{n} \overset{\text{exch.}}{\sim} P_X \times P_{Y|X}$
- $(X_{n+1}, Y_{n+1}) \sim \tilde{P}_X \times P_{Y|X}$

Then, the heuristic idea in order to keep the error rates of the predictions somewhat controlled is to give more "relevance" to those calibration points that are closer in distribution to the test point.

In practice, this can be translated in the following algorithm:

1. Estimate how "close" a sample X_i ($\sim P_X$) is *w.r.t.* to the test point ($\sim \tilde{P}_X$) using the likelihood ratio: $w(X_i) := \frac{d\tilde{P}_X(X_i)}{dP_X(X_i)}$.
2. Normalize the weights: $\omega_i := \frac{w(X_i)}{\sum_{j=1}^{n+1} w(X_j)}$.
3. Build the predictive interval $\mathcal{C}_\alpha$ using the weighted calibration samples:

$$\hat{\mathcal{C}}_\alpha(X_{n+1}) = \{Y : s_{\hat{\mu}}(X_{n+1}, Y) \leq q_{1-\alpha}(\{\omega_i S_i\}_{i \in \text{Cal}})\} \tag{3.1}$$

This approach not only works in practice, but also theoretically attains **at least** $1 - \alpha$ coverage if the samples are *i.i.d.* drawn, as announced by Theorem 3:

Theorem 3 (CP guarantees under covariate shift)**.** *Suppose* $\{(X_i, Y_i)\}_{i=1}^n$ *are drawn i.i.d. from* $P_X \times P_{Y|X}$*, and* (X_{n+1}, Y_{n+1}) *is drawn independently from* $\tilde{P}_X \times P_{Y|X}$*. Then,* $\hat{\mathcal{C}}_\alpha$ *from 3.1 is such that:*

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

Further details of the implementation and theoretical guarantees can be found at Tibshirani et al., 2019.

3.2 Label shift

Label shift refers to the change in the target variable's distribution $P_Y \to \tilde{P}_Y$.

Formally, thus, this case can be stated as it follows:

- $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{exch.}}{\sim} P_{X|Y} \times P_Y$
- $(X_{n+1}, Y_{n+1}) \sim P_{X|Y} \times \tilde{P}_Y$

The essential idea here is analogous to section 3.1: to give more "relevance" to those calibration points that are closer to the test point; with the added difficulty, however, that for a new sample X_{n+1}, its label Y_{n+1} is unknown.

In practice, this can be circumvented letting the weights ω_i as function of Y:

1. Estimate how "close" a label Y_i ($\sim P_Y$) is *w.r.t.* to the hypothetical point ($\sim \tilde{P}_Y$) using the likelihood ratio: $w(Y_i) := \frac{d\tilde{P}_Y(Y_i)}{dP_Y(Y_i)}$.
2. Normalize the weights: $\omega_i^Y := \frac{w(Y_i)}{\sum_{j=1}^n w(Y_j) + w(Y)}$.
3. Build the predictive interval $\mathcal{C}_\alpha$ traversing all the variable output's space and using the weighted calibration samples:

$$\hat{\mathcal{C}}_\alpha(X_{n+1}) = \{Y : s_{\hat{\mu}}(X_{n+1}, Y) \leq q_{1-\alpha}\left(\{\omega_i^Y S_i\}_{i \in \text{Cal}}\right)\} \tag{3.2}$$

For a more exhaustive overview, the user can refer to Podkopaev and Ramdas, 2021. Apart from a detailed implementation of the procedure, the author adapts the covariate-shift results of Tibshirani et al., 2019 (based on the "*weighted exchangeability*" concept) into this label-shift case through Theorem 4:

Theorem 4 (CP guarantees under label shift)**.** *Suppose* $\{(X_i, Y_i)\}_{i=1}^n$ *are drawn i.i.d. from* $P_{X|Y} \times P_Y$*,* (X_{n+1}, Y_{n+1}) *is drawn independently from* $P_{X|Y} \times \tilde{P}_Y$ *and the true likelihood ratios* ω_i^y *are known for all* Y*. Then,* $\hat{\mathcal{C}}_\alpha$ *from 3.2 is such that:*

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

3.3 Conformal prediction for time-series

Time series are, in general, a great example of data with strong auto-correlation amongst samples. Examples abound *e.g.* in financial markets (stock prices are influenced by their historical values) or in meteorology (with spatial auto-correlation as well, since weather conditions are influenced by their own regimes and variability modes).

In this case we have a setup such that $Y_t = \mu(X_t) + \epsilon_t$, where ϵ_t are identically distributed according to a common cumulative distribution function F. Assuming the first T sample points $\{(X_t, Y_t)_{t=1}^{T}\}$ are training data, we want to construct a sequence of $s \geq 1$ prediction intervals of α miscoverage level, $\{\mathcal{C}^{\alpha}_{T,T+i}\}_{i=1}^{s}$, for the unknown labels $\{Y^{\alpha}_{T+i}\}_{i=1}^{s}$ (being s a fixed batch size corresponding to how many steps we want to look ahead).

Once new samples $\{(X_{T+i}, Y_{T+i})\}_{i=1}^{s}$ become available, we would like to also leverage them using the most recent $T+s$ points for the predictive intervals of Y_j for $j = T+s+1$ onward.

We will focus on the so-called "*EnbPI*" methodology, proposed by Xu and Xie, 2021, which allows the use of CP when there is samples auto-correlation, particularly specializing in the case of time series.

This flavour resembles the "*Jackknife+ after bootstrap*" technique (Kim, Xu, and Barber, 2020), in the sense it applies CP to ensemble methods; but unlike the former work, EnbPI does not assume exchangeability and it does leverage new (sequentially) revealed observations.

Yet, as in section 2.3, the i-th "leave-one-out" (LOO) estimator[1] of μ will be denoted by $\hat{\mu}_{-i}$.

Besides, as discussed below, EnbPI has several other benefits: it requires no data-splitting, avoids model overfitting and does not refit models during test time.

3.3.1 EnbPI implementation

Below, it is listed the essential idea which EnbPI algorithm[2] uses to return the T_1 future predictive intervals (indices $T+1, \ldots, T+T_1$; as there are T training samples):

- Obtain B bootstrapped models (henceforth denoted by μ^b) by:
  - Sampling, with replacement, an index set $S_b := (i_1, \ldots, i_T)$ from indices $(1, \ldots, T)$.
  - Fit the bootstrapped model, with S_b:

$$\hat{\mu}^b(\cdot) \leftarrow \mathcal{A}\left(\{(X_i, Y_i), i \in S_b\}\right) .$$

- For each i of the T training samples, aggregate the bootstrapped models (with an aggregation function denoted ϕ) obtaining: $\hat{\mu}^{\phi}_{-i}$. Then, compute the conformity scores using the absolute residual: $\epsilon^{\phi}_i := \left|Y_i - \hat{\mu}^{\phi}_{-i}(X_i)\right|$.

[1] Note its training data, then, will include the rest of $T-1$ points and just exclude the i-th (X_i, Y_i) point).

[2] We no longer use S for the conformity scores, but for the S index sets. In EnbPI, the idea of conformity score is represented through ϵ and w.

- For each t of the future T_1 timestamps (test data), return in a s-sized batch the predictive interval:

$$\hat{C}^{\alpha}_{T,t}(X_t) = \left[\hat{\mu}^{\phi}_{-t}(X_t) - w^{\phi}_t,\ \hat{\mu}^{\phi}_{-t}(X_t) + w^{\phi}_t\right],$$

  where:

  - $\hat{\mu}^{\phi}_{-t}(X_t)$ is the $1-\alpha$ quantile of $\{\hat{\mu}^{\phi}_{-i}(X_t)\}^{T}_{i=1}$.
  - w^{ϕ}_t is the $1-\alpha$ quantile of $\{\epsilon^{\phi}_i\}^{T}_{i=1}$.

- Lastly, note this interval's retrieval is made sequentially using "batch". This means that, for each s returned intervals, the conformity score w^{ϕ}_t will be re-computed leveraging the most recent observations as well (steps 15-19 in Algorithm 4).

More formally, the algorithm is adapted from Xu and Xie, 2021 and stated in Algorithm 4.

Note the authors later proposed, in Xu and Xie, 2023, a slight modification in the calculation of $\hat{\mu}^{\phi}_{-t}(X_t)$ (leveraging ϕ) and w^{ϕ}_t (introducing a new quantity $\hat{\beta}$ to affect the quantile level).

3.3.2 Theoretical guarantees

An exhaustive review of the theoretical proofs of EnbPI coverage is much beyond of the scope of this work.

Nevertheless, the reader can refer to Xu and Xie, 2023 where the main results are provided.

In particular, conditional coverage is attained not in an absolute sense but in an asymptotic sense; all this up to two hypotheses:

- Errors are short-term *i.i.d* (independent and identically distributed) according to a common CDF Lipschitz continuous. See Assumption 1 of Xu and Xie, 2023.

- Estimation quality (Assumption 2): there exists a real sequence $\{\delta_T\}_{T>0}$ that converges to zero such that:

$$\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\mu}_{-t}(X_t) - \mu(X_t)\right)^2 \geq \delta_T^2 \text{ and}$$
$$\left|\hat{\mu}_{-t}(X_{T+1}) - \mu(X_{T+1})\right| \leq \delta_T.$$

As discussed, these are mild assumptions. However, since are talking about asymptotic coverage, notice the coverage level depends on the size of the training set and on $(\delta_T)_{T>0}$.

Furthermore, though weaker, the authors also present a theorem showing marginal asymptotic coverage is sought too.

The proof (see Appendix A of Xu and Xie, 2023) removes the assumptions on data exchangeability by replacing them with general and verifiable assumptions on the error process and estimation quality. Without loss of generality, just guarantees are shown for $t = T+1$ (one-step-ahead prediction); but, in Remark 1, it is explained how it can be extended to $t = T+2, \ldots, T+T_1$.

Algorithm 4 EnbPI algorithm

Input: Regression algorithm $\mathcal{A}$, miscoverage level α, aggregation function ϕ, number of bootstrap models B, batch size s, training data $\{(X_i, Y_i)\}_{i=1}^{T}$ and test data $\{(X_t, Y_t)\}_{t=T+1}^{T+T_1}$ with Y_t revealed only after the batch of s prediction intervals with t in the batch are constructed.

Output: Ensemble prediction interval $\{\hat{\mathcal{C}}_{T,t}^{\alpha}(X_t)\}_{t=T+1}^{T+T_1}$.

for $b = 1, \ldots, B$ **do**
  Sample with replacement an index set $S_b = (i_1, \ldots, i_T)$ from indices $(1, \ldots, T)$
  Compute $\hat{\mu}^b(\cdot) \leftarrow \mathcal{A}(\{(X_i, Y_i), i \in S_b\})$.
end for
Initialise $\epsilon = \{\}$
for $i = 1, \ldots, T$ **do**
  $\hat{\mu}_{-i}^{\phi}(X_i) = \phi\left(\{\hat{\mu}^b(X_i) \mid i \notin S_b\}\right)$
  Compute $\hat{\epsilon}_i^{\phi} = \left|Y_i - \hat{\mu}_{-i}^{\phi}(X_i)\right|$
  $\epsilon = \epsilon \cup \left\{\hat{\epsilon}_i^{\phi}\right\}$
end for
for $t = T+1, \ldots, T+T_1$ **do**
  Let $\hat{\mu}_{-t}^{\phi}(X_t) = (1-\alpha)$ quantile of $\left\{\hat{\mu}_{-i}^{\phi}(X_t)\right\}_{i=1}^{T}$
  Let $w_t^{\phi} = (1-\alpha)$ quantile of ϵ
  Return $\hat{\mathcal{C}}_{T,t}^{\alpha}(X_t) = \left[\hat{\mu}_{-t}^{\phi}(X_t) \pm w_t^{\phi}\right]$
  if $t - T = 0 \bmod s$ **then**
    for $j = t-s, \ldots, t-1$ **do**
      Compute $\hat{\epsilon}_j^{\phi} = \left|Y_j - \hat{\mu}_{-j}^{\phi}(X_t)\right|$
      $\epsilon = \left(\epsilon - \left\{\hat{\epsilon}_1^{\phi}\right\}\right) \cup \left\{\hat{\epsilon}_i^{\phi}\right\}$ and reset index of ϵ
    end for
  end if
end for

Chapter 4

Results

After a survey of the main theoretical CP findings & modern flavours (Chapter 2), and the recent workarounds to enable this framework in the non-exchangeable case (Chapter 3); this chapter will be devoted to the assessment of two different practical use-cases.

First, in section 4.1 we will briefly review the code implementation in Python and, in section 4.2, we will discuss how the user can assess several important attributes of the obtained predictive intervals.

Lastly, in section 4.3 we will present the results of applying CP to a tabular data regression problem; while the non-exchangeable case will be covered in section 4.4, in which CP will be applied to a energy demand forecasting problem.

4.1 Implementation

The code developed for this work can be found at the corresponding author's repository. Essentially, the scripts leverage the Python library `mapie`, both for the computation of predictive intervals and the assessment of results.

In particular, regarding the uncertainty quantification procedure, we can distinguish the following Python classes:

- `mapie.regression.MapieRegressor`: it deduces valid confidence intervals by evaluating out-of-fold conformity scores on hold-out validation sets. Depending on the different parameters election, one strategy or another can be implemented. In particular, to apply
  - SCP: `method="base"` & `cv="split"` must be selected.
  - CV+: `method="plus"` & `cv=K` must be selected, with `K` the number of cross-validation folds (*e.g.* 10).
  - Jackknife's J+aB: `method="plus"` & `cv=Subsample(n_resamplings=n)` must be selected, initializing the `mapie.subsample.Subsample` class with a certain `n` number of bootstrapped resamples (*e.g.* 50).

  All the former is implemented in the `cp.exchangeable` author's module.

- `mapie.regression.MapieQuantileRegressor`: enables CQR strategy, as proposed by Romano, Patterson, and Candès, 2019, with the only valid selections: `method="quantile"` & `cv="split"`. The author implements it in the same `cp.exchangeable` module.

- `mapie.regression.MapieTimeSeriesRegressor`: enables the conformal prediction framework for single-output time series data by predicting intervals calibrated with out-of-fold residuals. The `method="enbpi"` implements the

EnbPI strategy, as proposed by Xu and Xie, 2021, which allows you to continuously update conformal scores using the `partial_fit` class method. The author leverages this class in the `cp.ts` module.

4.2 Assessment

When it comes to assessing the benefits of each strategy certain attributes must be taken into account. For instance:

- **Coverage level**: *i.e.* the fraction of true labels which lie within the prediction intervals, with `mapie.metrics.regression_coverage_score_v2` as metric.
- **Interval width**: the intervals' mean width, which can be turned into score as prescribed by `mapie.metrics.regression_mean_width_score`.
- **"*Informativeness*"**: a trade-off exists between the interval's width (the smaller, the more informative they are) and the statistical coverage. A clear way of assessing this, combining both the w mean width score and the c coverage score, is through the CWC score.

  This metric, implemented by `mapie.metrics.coverage_width_based`, was proposed by Khosravi, Nahavandi, and Creighton, 2010 and is computed as:

  $$\text{CWC} = (1 - w) * \exp\left(-\eta(c - (1 - \alpha))^2\right),$$

  where η is a balancing term devoted to reward narrow intervals and penalize those that do not achieve a specified coverage probability.
- **Adaptability**: the ability of achieving (approximate) conditional coverage, measured by the score `mapie.metrics.regression_ssc_score`. The SSC (Size Stratified Coverage) score computes the maximum violation of the coverage. In particular, the intervals are grouped by width and the coverage is computed for each group. The lower coverage is the maximum coverage violation. An adaptive method is one where this maximum violation is as close as possible to the global coverage.

  However, it is very important to check that the intervals widths are well spread before drawing conclusions, because this metric is only usable if the predicted intervals have non-constant width.
- **Computational efficiency**: the amount of computational resources needed to implement each strategy, which could be measured in terms of CPU time (both training and inference).

In this work, we use the author's repository `cp.validate` & `cp.visualize` modules to respectively compute and represent all the former metrics.

4.3 Exchangeable data

In this work, we use (through the author's `cp.data` module) the same dataset as the `mapie`'s CQR tutorial to present the exchangeable data use-case: the `sklearn` built-in California Housing dataset.

Chosen in view of being simple and reproducible, in particular no feature engineering is needed; it is composed of 20,640 samples of the following 8 different features:

- The median income in block group
- The median house age in block group
- The average number of rooms per household
- The average number of bedrooms per household
- The block group population
- The average number of household members
- The location (latitude & longitude) of the block group
- The label variable: the median house price for a given block group.

The marginal distributions of the dataset are shown in Figure A.1 at Appendix A, where the complete set of visualizations and plots for the results' analysis can also be found.

Due to the dataset complexity and its potential non-linear relationships, a gradient boosting model is chosen as base estimator. In particular, the `LGBM` regressor is implemented through the library's `lightgbm.LGBMRegressor` Python class.

Furthermore, to automate the hyper-parameters fine-tuning task, a randomized grid-search is implemented using a 5-fold cross-validation. For this particular problem, the found best settings are:

- `learning_rate`: 0.34318
- `max_depth`: 18
- `n_estimators`: 75
- `num_leaves`: 29

Then, 4 strategies are implemented with `mapie` according to section 4.1 configuration; these are: the Split Conformal Prediction (SCP), the Cross-Validation + (CV+), the Jackknife+ after Bootstrapping (J+aB) and the Conformalized Quantile Regression (CQR).

While all the 4 strategies are able to provide informative prediction intervals, they present differences in their attributes as shown in Table 4.1. All the code used to generate these results and visualizations can be found at the author's regression.ipynb notebook.

In particular, some remarks can be drawn:

- CQR & SCP are those with most statistical efficiency in terms of coverage (the attained "*Coverage*" is closer to the expected 0.80).
- On the one hand, CQR & SCP are both based in a 1-fold split of the dataset. Consequently, while their training and prediction times are significantly lower than CV+ & J+aB strategies; the predictive power of the base estimator (trained just with 70% of the data) is also slightly lower.
- On the other hand, and spite of its large training and inference times, & J+aB offers the best coverage-width ratio (and thus, informative intervals) according to the CWC score.
- Unlike J+aB or SCP, CV+ & CQR offers some interval adaptability. In this sense, according to the SSC score, CQR is better not only achieving global coverage but also approximate conditional coverage.

Strategy	Coverage	RMSE	Training time	Inference time
SCP	0.806 ± 0.008	0.472 ± 0.007	1.602 ± 0.174	0.068 ± 0.054
CV+	0.853 ± 0.004	0.467 ± 0.009	9.329 ± 2.804	7.986 ± 0.302
J+aB	0.734 ± 0.007	0.467 ± 0.009	51.210 ± 5.609	9.698 ± 0.424
CQR	0.805 ± 0.010	0.494 ± 0.013	2.601 ± 0.087	0.095 ± 0.044

(A) Coverage, RMSE, training & inference times.

Strategy	Coverage	Width	CWC	SSC
SCP	0.806 ± 0.008	0.971 ± 0.015	0.798 ± 0.004	—
CV+	0.853 ± 0.004	1.042 ± 0.005	0.784 ± 0.002	0.650 ± 0.012
J+aB	0.734 ± 0.007	0.710 ± 0.003	0.853 ± 0.001	—
CQR	0.805 ± 0.010	1.013 ± 0.013	0.790 ± 0.004	0.745 ± 0.043

(B) Coverage, width, coverage width-based criterion (CWC) score & size-stratified coverage (SSC) score.

TABLE 4.1: Different strategies' metrics after a 5-fold cross-validation for $\alpha = 0.2$ regression problem.

In conclusion, CQR seems the best strategy in order to achieve the best marginal and conditional coverage; thus, resulting a specially good choice in those applications needing for a conservative and statistical efficient tool. The training and inference times constitute good reasons for its election too, but it should be taken into account a large enough dataset is needed for the method to be informative (since the dataset will be split).

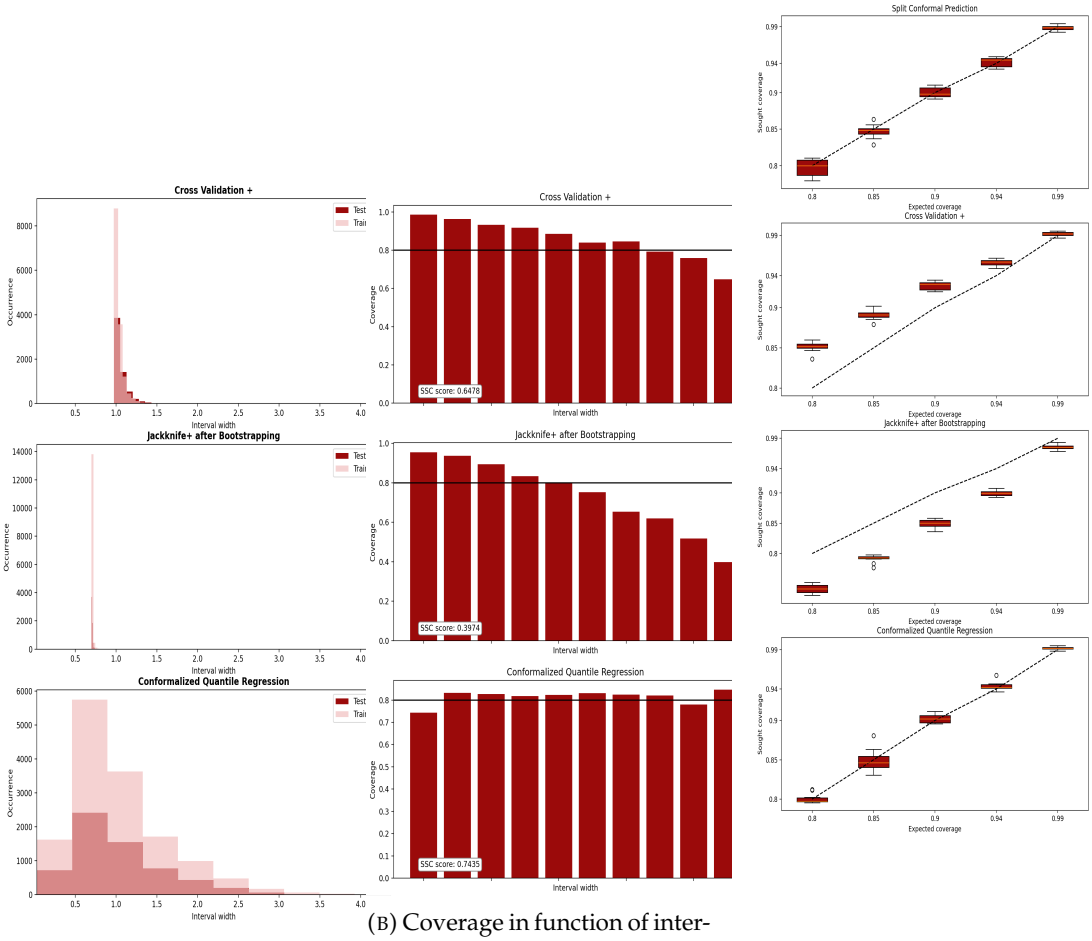
On the contrary, in case predictive power is to be maximized, while minimizing intervals width, and at expenses of some potential coverage loss and almost no adaptability, then J+aB should be chosen. Thus, this strategy seems suitable for those applications in which more reckless guesses can be afforded and the highest predictive power with the minimal width is desired.

Finally, to provide more in-depth detail about the coverage capabilities of the former strategies, some plots regarding a specific experiment are displayed in Figure 4.1.

In particular, at sub-figures 4.1a & 4.1b, the ability of CQR & CV+ to adapt the interval width to the situation is displayed opposed to J+aB's. Besides, more adaptability is achieved by CQR, because sub-figure 4.1a shows the intervals' width histograms and how CQR features much more variability than CV+ (also wider intervals' bins are occupied more frequently).

Then, sub-figure 4.1b presents the attained coverage in function of the interval widths each strategy yielded, showing how CQR effectively attains more coverage per different width.

And lastly, sub-figure 4.1c shows the ability to attain global coverage but in function of different α values (through a 5-fold cross-validation for 5 α values evenly spaced from 0.20 to 0.01). The same conclusions for the $\alpha = 0.2$ analysis apply: CQR & SCP, and independently of α, are the best when it comes to attaining global coverage.



(A) Intervals' width histograms (B) Coverage in function of intervals' width (C) Coverage in function of α

FIGURE 4.1: Visualizations related to width & coverage distributions for the test data and 4 different strategies, from top to bottom: SCP, CV+, J+aB & CQR. The first 2 plots just display the last 3 strategies, since SCP displays no adaptability at all (whereas J+aB slight to none adaptability).

4.4 Time series data

For the non-exchangeable data use-case, the same dataset as the `mapie`'s time series tutorial was chosen. This is the Victoria electricity demand dataset, used in the book "*Forecasting: Principles and Practice*" (Hyndman and Athanasopoulos, 2014), and contains a total of 1340 hourly samples. It deals with an electricity demand forecasting problem, which not only features daily and weekly seasonality, but it is also impacted by temperature. Thus, apart from the demand lagged up to 7 days (and other time features), temperature will be used as exogenous variable.

The dataset can be visualized in Figure B.1 at Appendix B, where the complete set of visualizations and plots for the results' analysis can also be found.

In this case, a different ensemble model than gradient boosting (section 4.3) is chosen as base estimator; in particular, a random forest regressor is implemented

through `sklearn` library's `sklearn.ensemble.RandomForestRegressor` Python class.

Similarly to section 4.3, a randomized grid-search is implemented using a 5-fold cross-validation to automate the hyper-parameters fine-tuning task. For this particular problem, the found best settings are:

- `max_depth`: 23
- `n_estimators`: 99

Then, 2 strategies based in `mapie`'s `EnbPI` implementation (see the configuration at section 4.1) are implemented, these are: `EnbPI` without partial fit (EnbPI_nP), *i.e.* the test residuals are not used to further adjust the model (steps 15-19 of Algorithm 4); and `EnbPI` with partial fit (EnbPI).

In both cases, the prediction batches were implemented using the `mapie`'s class for bootstrapping blocks of data, `mapie.subsample.BlockBootstrap`, with:

- `n_resamplings`= 100
- `length`= 48 (i.e. batch size of $s = 48$ h samples).

All the code used to generate these results and visualizations can be found at the author's timeseries.ipynb notebook.

4.4.1 Original dataset

For the original Victoria electricity demand dataset, both EnbPI_nP and EnbPI are able to provide informative prediction intervals, presenting some minor differences in their attributes as shown in Table 4.2.

Strategy	Coverage	RMSE	Total time (train + infer)
EnbPI_nP	0.780 ± 0.069	0.165 ± 0.067	6.157 ± 0.334
EnbPI	0.789 ± 0.058	0.165 ± 0.067	528.343 ± 0.359

(A) Coverage, RMSE, total time (training & inference with residuals adjustment if applies).

Strategy	Coverage	Width	CWC	SSC
EnbPI_nP	0.780 ± 0.069	0.293 ± 0.013	0.935 ± 0.018	—
EnbPI	0.789 ± 0.058	0.300 ± 0.007	0.934 ± 0.016	0.518 ± 0.209

(B) Coverage, width, coverage width-based criterion (CWC) score & size-stratified coverage (SSC) score

TABLE 4.2: Different strategies' metrics after the Figure 4.2's 5-fold cross-validation for the time series problem with $\alpha = 0.2$.

Note these results stem from a 5-fold cross-validation, with the batches (not shuffled samples, to break temporal auto-correlation) shown in Figure 4.2.

In particular, it is easy to conclude the benefits of adjusting the intervals with the test residuals (steps 15-19 of Algorithm 4) since EnbPI is better than EnbPI_nP. Despite a slight 0.001 loss width-coverage ratio score, the global coverage improved. Not only global coverage was improved, but also adaptive intervals were enabled achieving a 0.518 ± 0.209 SSC score (with respect to the non-adaptive intervals of EnbPI_nP).

The latter can be easily checked in sub-figures 4.3a & 4.3b, where not only EnbPI features several sized intervals, but it also attains uniform coverage in every of them

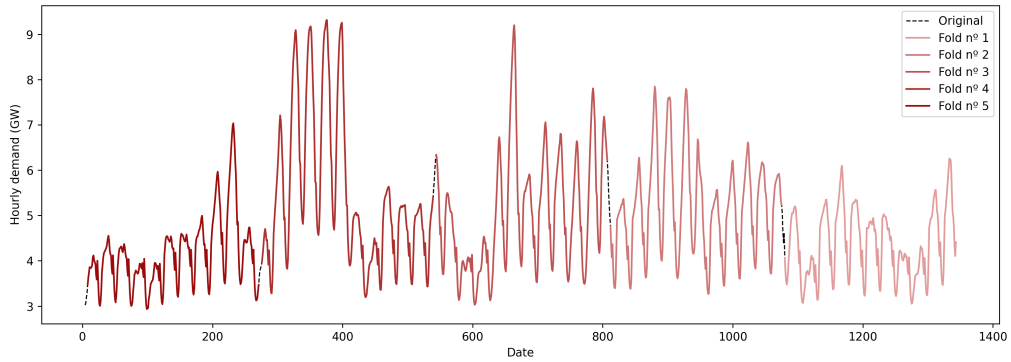
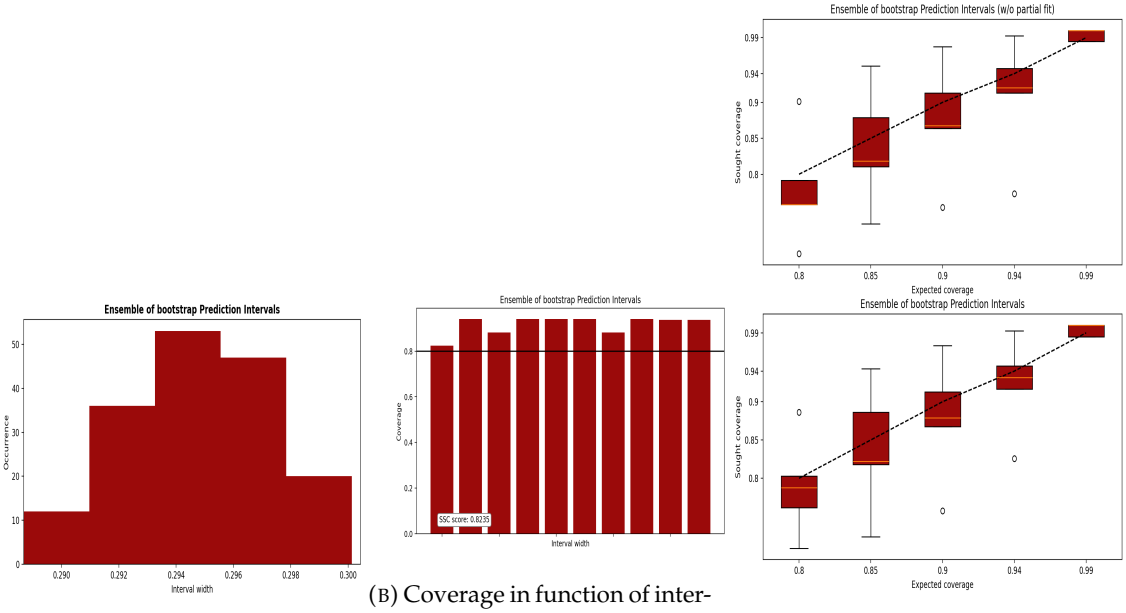


FIGURE 4.2: 5-fold splits from the original dataset.



(A) Intervals' width histograms (B) Coverage in function of intervals' width (C) Coverage in function of α

FIGURE 4.3: Intervals width & coverage distributions for the test data and the EnbPI strategy (without partial fit, top; and with it, bottom).

(and almost the expected one). Also, the ability to attain global coverage for different α values is shown in sub-figure 4.3c.

Of course, despite its complexity and multiple seasonality, this dataset is very consistent presenting almost neither trend nor strong distribution shifts as shown in Figure 4.2. Due to this reason, EnbPI_nP & EnbPI performances are really similar.

To avoid outshining EnbPI adaptive ability, a special strong-shift case is presented in subsection 4.4.2.

4.4.2 Change point in the test data

Henceforth in this subsection, a special case of the time series problem is considered. In particular, a change point will be added in the test split (by artificially subtracting 2 GW) to mock off the situation in which there is a sudden strong distribution shift in the middle of the test data. Also, in this case and unless otherwise is specified, $\alpha = 0.05$ is chosen as miscoverage level.

The dataset with this change point can be visualized in Figure C.1 at Appendix C (where the rest of visualizations can also be found).

For this particular scenario, EnbPI_nP and EnbPI do present significance differences in their performance, as shown in Table 4.3.

Strategy	Coverage	RMSE	Total time (train + infer)
EnbPI_nP	0.439 ± 0.075	1.431 ± 0.024	6.047 ± 0.307
EnbPI	0.696 ± 0.042	1.431 ± 0.024	529.902 ± 1.319

(A) Coverage, RMSE, total time (training & inference with residuals adjustment if applies).

Strategy	Coverage	Width	SSC
EnbPI_nP	0.439 ± 0.075	0.569 ± 0.043	—
EnbPI	0.696 ± 0.042	1.300 ± 0.034	0.069 ± 0.120

(B) Coverage, width & size-stratified coverage (SSC) score.

TABLE 4.3: Different strategies' metrics after the Figure 4.4's 5-fold cross-validation for the time series problem (with a change point in test) and $\alpha = 0.05$.

Note these metrics are obtained from the batched 5-fold cross-validation splits shown in Figure 4.4 (just the test splits are shown, the train split is the rest of the dataset for each case).

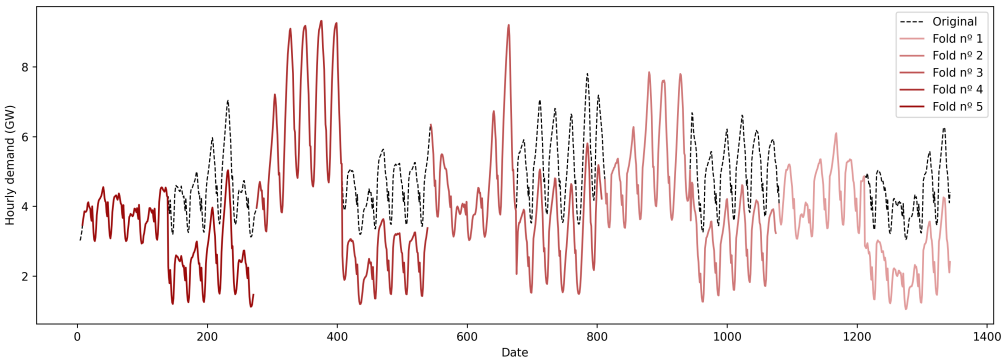


FIGURE 4.4: 5-fold splits from the dataset with change point in the test split(s).

These results clearly justify the need of adjusting the intervals with the test residuals (EnbPI) despite of significantly increasing the computational times.

Specifically, EnbPI's ability of increasing the intervals width after the strong-shift in the middle of the test split, allows the methodology to recover and yields a 58.54% coverage increase. This justified need of suddenly increasing intervals width (making them less informative) is the reason why the CWC score is not shown in Table 4.3, since it could lead to wrong interpretations (the metric would penalize this behavior).

This behavior by which EnbPI, and after the change point, tries to compensate the lack of coverage in their future predictions (by increasing intervals width) is easily observed at Figure 4.5.

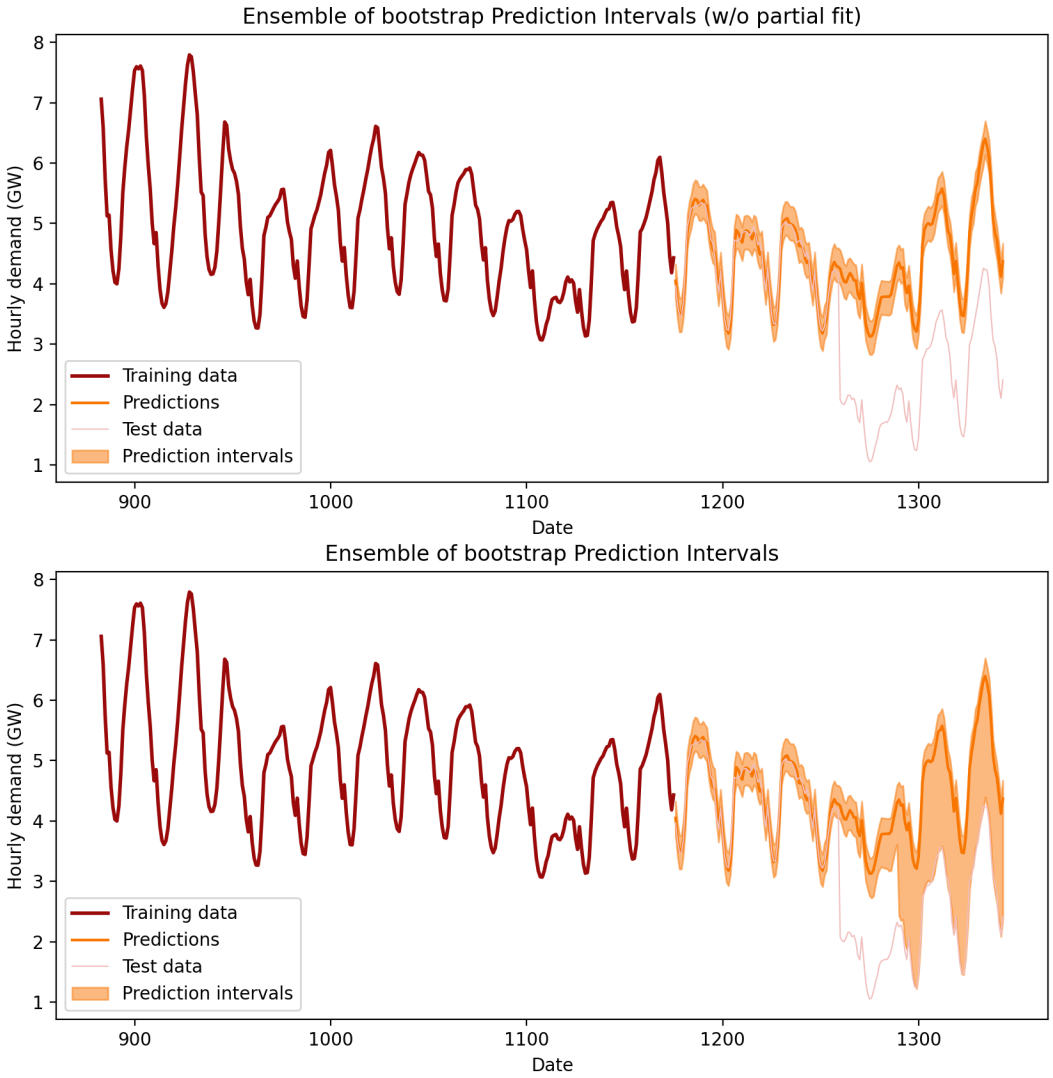


FIGURE 4.5: EnbPI increasing interval width through its partial fit feature, for a particular experiment with $\alpha = 0.05$.

However, note that this recover speed is directly related to the tolerated miscoverage level. Namely, the lower the miscoverage α level, the quicker this change will be featured. In particular, in Figure 4.6, it can be seen how EnbPI has no time to recover for $\alpha = 0.20, 0.15$ (80%, 85% confidence levels; top & middle sub-figures), while it effectively does for $\alpha = 0.10$ (but, of course, later than Figure 4.5's $\alpha = 0.05$).

These different speeds can also be noted at sub-figure 4.7c, in which the attained global coverage for the EnbPI strategy varies non-linearly with α.

Finally, while in sub-figures 4.7a & 4.7b the adaptive feature of EnbPI intervals[1] is shown; in Figure 4.8 the EnbPI's progressive coverage recovery is featured in function of time with a rolling window (while indeed EnbPI_nP does not recover at all).

[1]Note the change point is also perceived with these 2 visualizations: the widest intervals correspond to those with less conditional coverage, since those were issued after the change point's recovery.

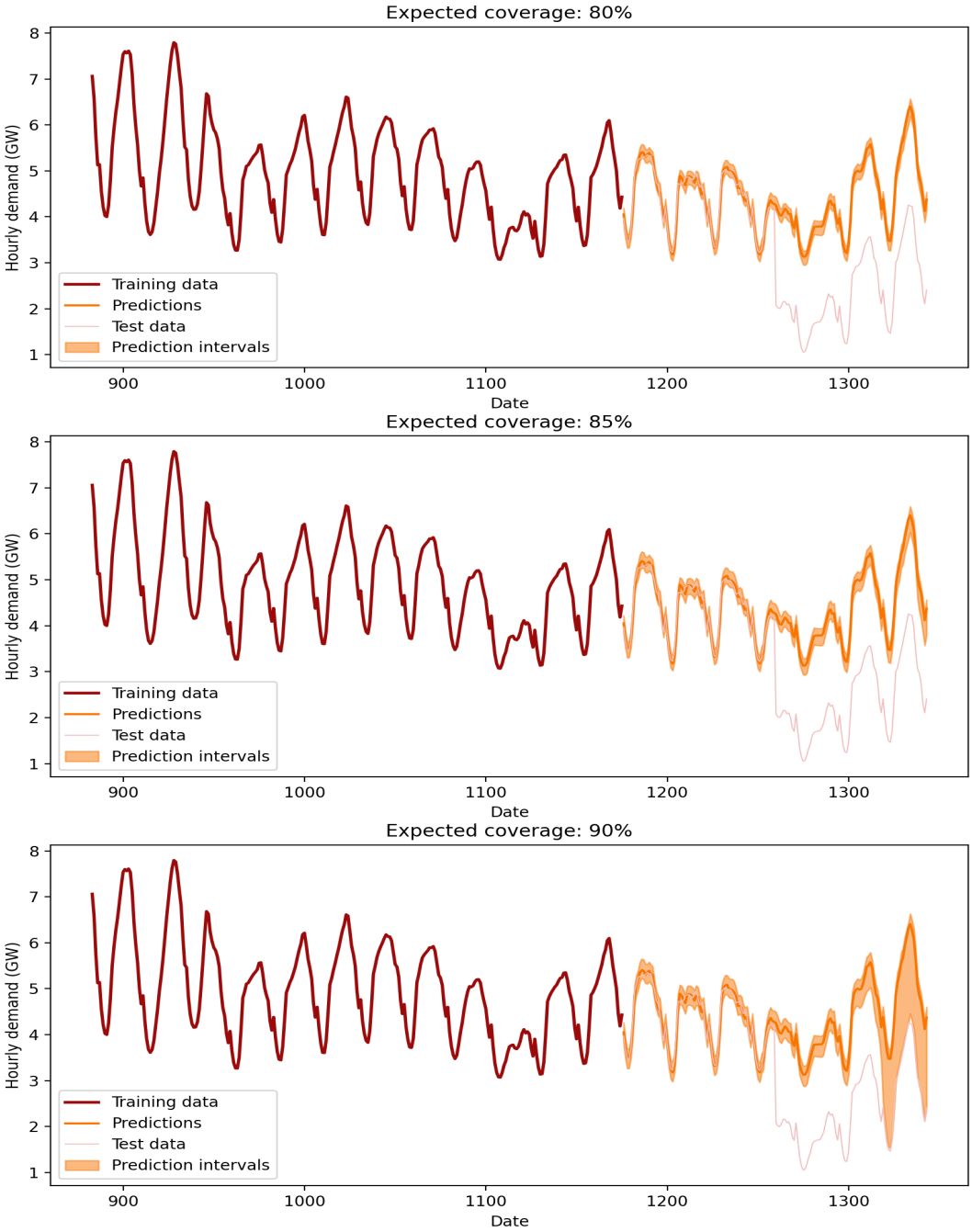
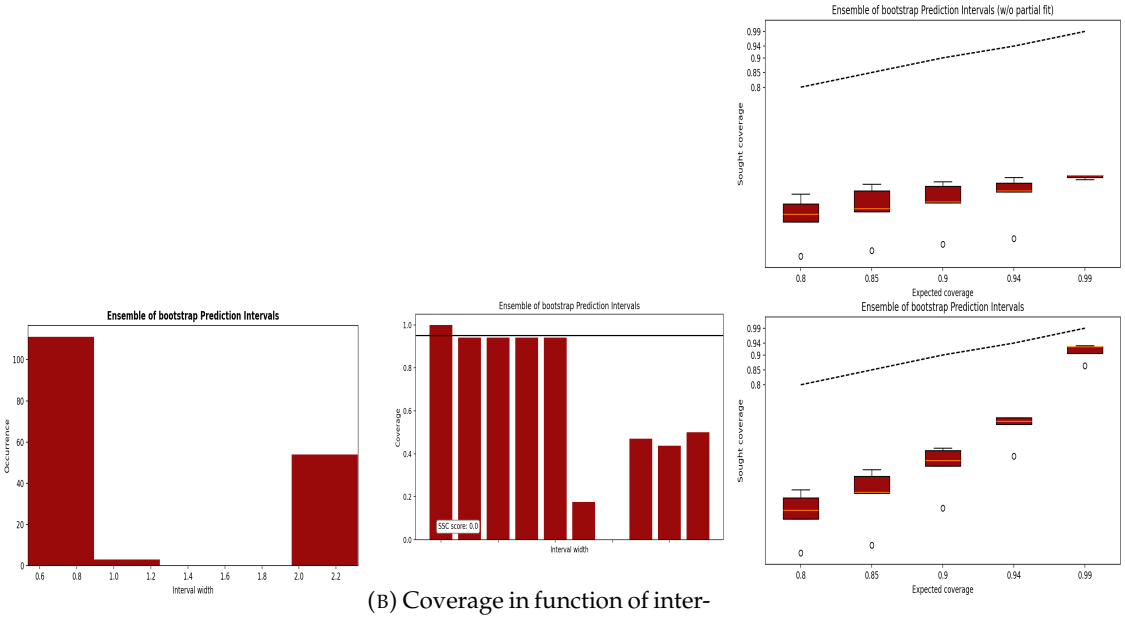


FIGURE 4.6: EnbPI with partial fit recovering intervals' width at different pace, for different α values (from top to bottom: $\alpha = 0.20, 0.15, 0.10$).



(A) Intervals' width histograms (B) Coverage in function of intervals' width (C) Coverage in function of α

FIGURE 4.7: Width & coverage distributions for the change point's (test) data ($\alpha = 0.05$) for EnbPI. At subfigure 4.7c, EnbPI_nP & EnbPI are displayed (top & bottom, respectively).

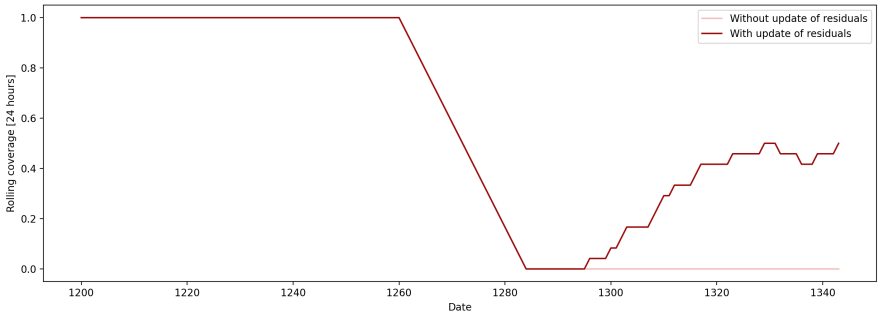


FIGURE 4.8: Coverage in function of time (grouped within 24h rolling windows) for the EnbPI strategies (and change point's data).

Chapter 5

Conclusions

Throughout this work, several methods have been theoretically justified and successfully applied to quantify the prediction uncertainty both for regression and time series problem. These strategies are distribution-free & model-agnostic and stem from the notion of "*conformalizing*" predictions to data and using the residuals to understand the errors distribution. That is why they are grouped within the so-called "conformal prediction" (CP) methodologies.

Even though CP paradigm was classically applied only under "*data exchangeability*" conditions, this work has reviewed some of the most recent & non-trivial efforts to enable CP when this hypothesis is not fulfilled.

In particular, while SCP, CV+, J+aB & CQR were studied for the exchangeable case; regarding the time series case, "*EnbPI*" (Xu and Xie, 2021) was presented as the strategy to effectively obtain prediction intervals with statistically valid coverage.

While all the former strategies were successfully applied to practical case, generally providing valid intervals, below the more fine-grained conclusions are listed:

- **Exchangeable case** (regression problem):
  - The best strategies, **decreasingly ordered** by:
    - *Statistical efficiency* are: CQR, SCP, CV+, J+aB. This is fulfilled independently of α.
    - *Computational efficiency* are: SCP, CQR, CV+, J+aB.
    - *Predictive power* are: CV+ & J+aB, SCP, CQR.
    - "*Informativeness*" (coverage-width ratio) are: J+aB, SCP, CQR, CV+.
    - *Adaptability* are: CQR, CV+, J+aB (slight to none). Contrarily, SCP intervals are not adaptive at all.
  - **CQR** seems the best **strategy** to achieve the best **marginal** & **conditional coverage**, when dataset is large enough.

    Thus, it may result suitable when a conservative and statistical efficient tool is needed.
  - **J+aB** seems the best **strategy** to achieve the best **informative intervals** (maximizing predictive power, while minimizing intervals width), at expenses of no-adaptability & losing some coverage.

    Thus, it may result suitable when more reckless guesses can be afforded and low training & inference times are not a requirement.
- **Non-exchangeable case** (time series problem):
  - **EnbPI** is a **suitable option** to provide valid intervals for **time series problems**.

  - In general, and particularly when there might be strong shifts in data, EnbPI's intervals adjustment using test residuals (its "partial fit" feature) is of crucial importance.
  - This "**partial fit**" option will not only allow the intervals' coverage **recover from change points**, but also will allow all the issued **intervals** to be **adaptive**.

5.1 Further research

There a huge number of relevant other inquiries and research lines which could extend this work, but were out of the scope of this thesis. Below, some of them are reviewed:

- Leverage cross-validation folds in the CQR strategy, instead of a simple train-test split of the dataset, to improve the predictive power and reduce the need of a large dataset.
- Implement other contemporary CP methodologies for time series problems, such as *Adaptive Conformal Inference*, ACI (Gibbs and Candes, 2021) and the more recent *Hopfield Conformal Prediction Trees*, HopCPT (Auer et al., 2023); in order to compare performance differences and their suitability.
- Extend all these methods to the multi-dimensional output variables' case, to broaden their applicability to multi-output regression & time series problems.
- Apply the former methodologies to classification problems (discrete target variables) and discuss whether similar conclusions to their continuous counterpart can be drawn.

Appendix A

Regression problem

Unless it is specified otherwise, the miscoverage level was set to $\alpha = 0.2$ (*i.e.* 80% of expected coverage) for the visualizations.

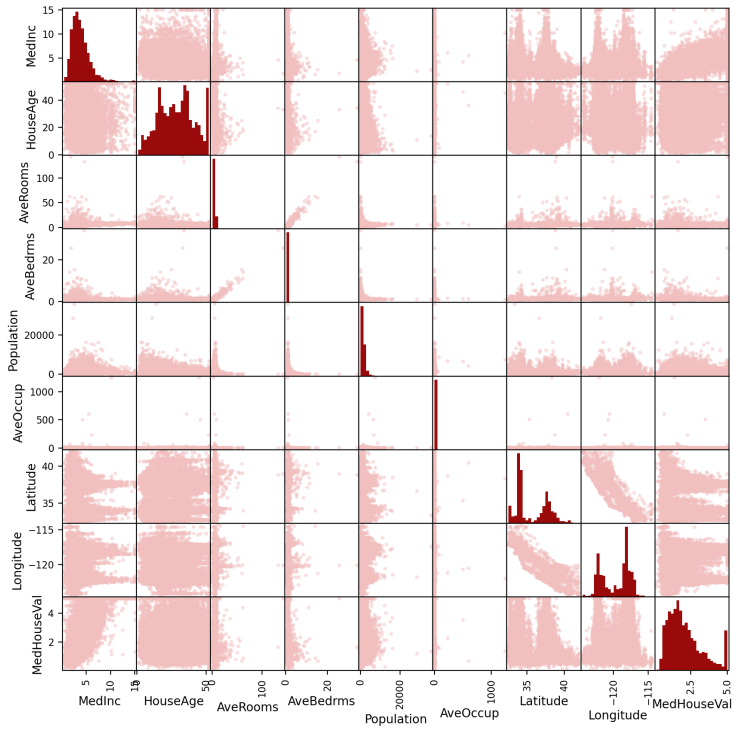
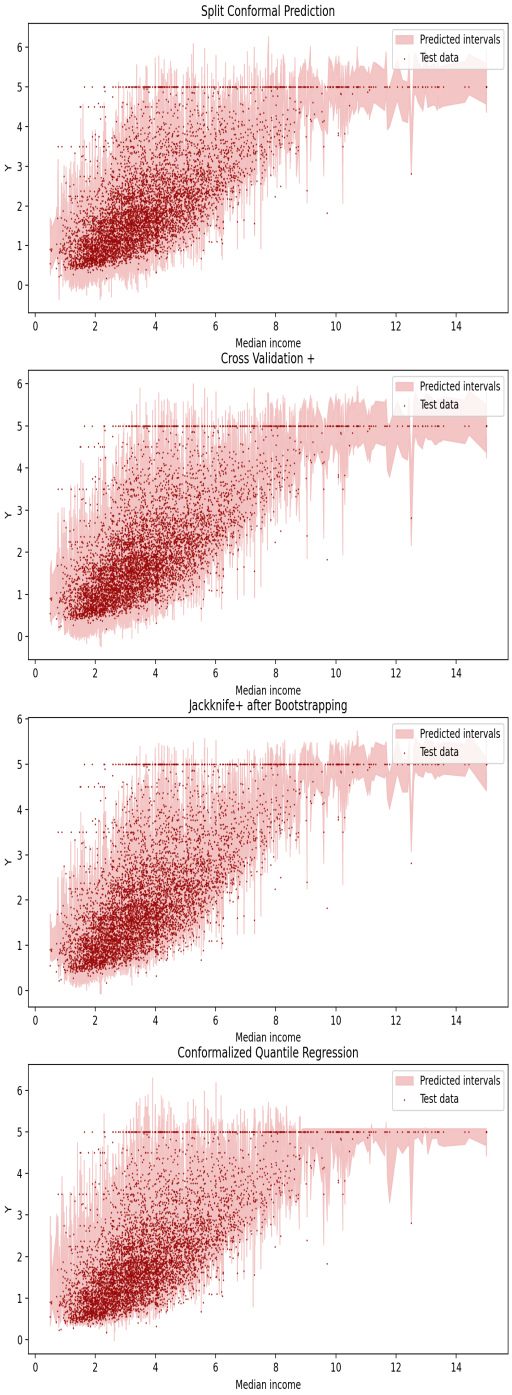
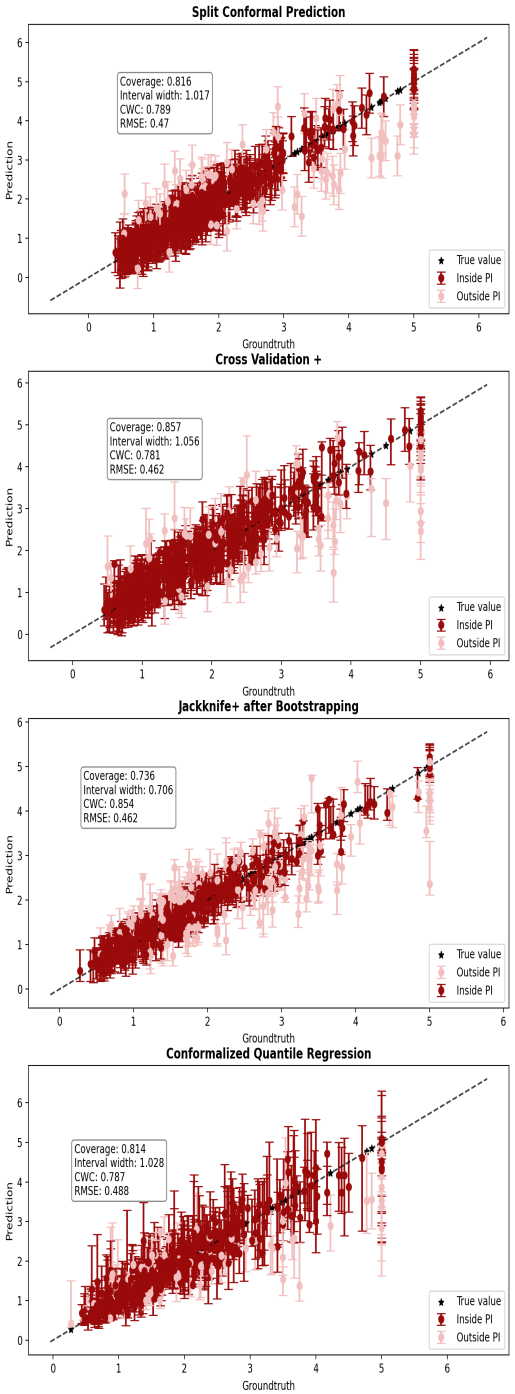


FIGURE A.1: Marginal distributions for each of the possible combinations of the regression problem's features.

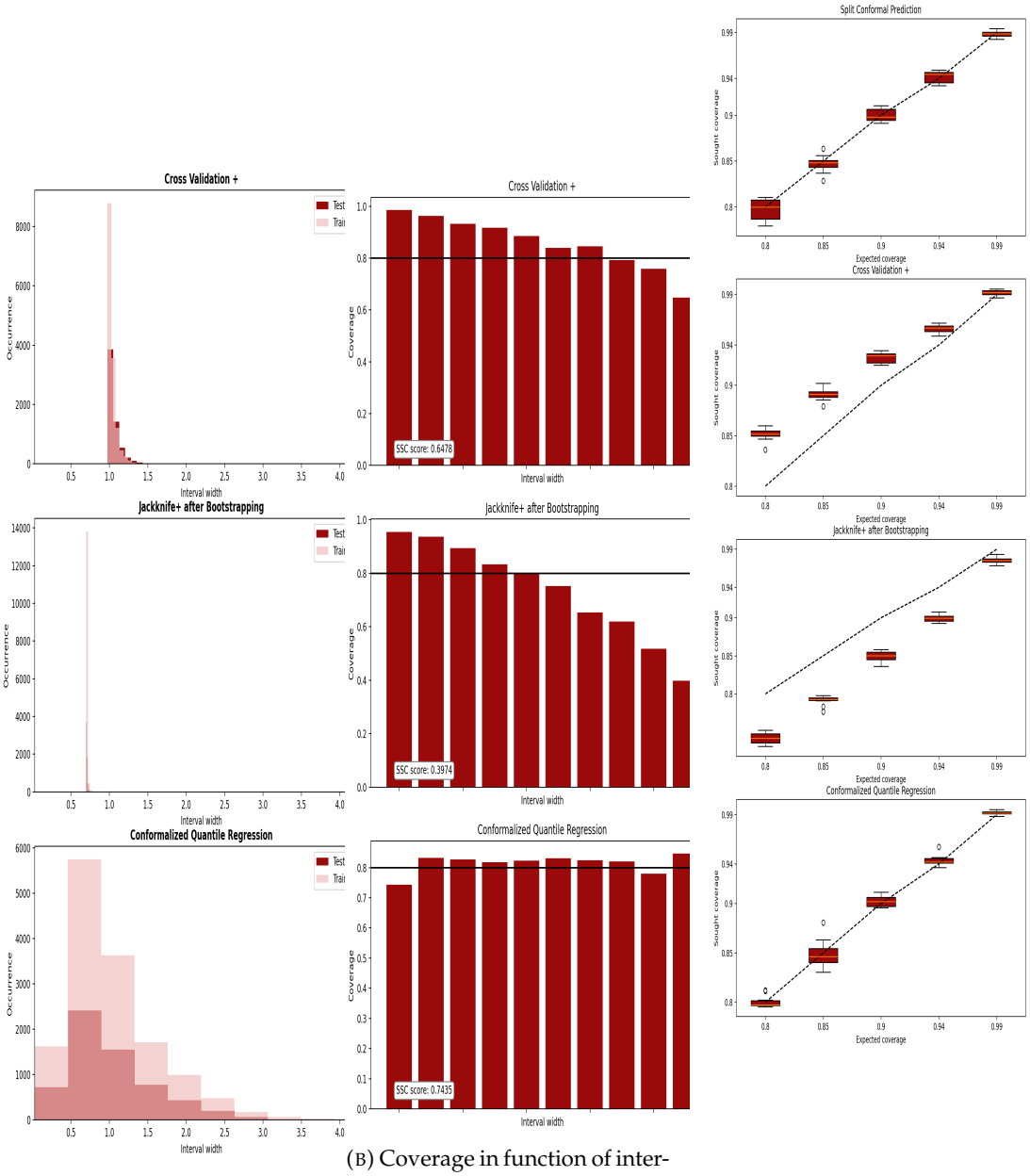


(A) Prediction intervals



(B) Goodness of the prediction intervals (prediction vs. ground truth). Just 7.5% was used for visualization purposes.

FIGURE A.2: Visualizations related to the prediction intervals for the test data and 4 different strategies, from top to bottom: SCP, CV+, J+aB and CQR.



(A) Intervals' width histograms (B) Coverage in function of intervals' width (C) Coverage in function of α

FIGURE A.3: Visualizations related to width & coverage distributions for the test data and 4 different strategies; from top to bottom: SCP, CV+, J+aB & CQR. The first 2 plots just display the last 3 strategies, since SCP displays no adaptability at all (whereas J+aB slight to none adaptability).

	Coverage	Int. width	RMSE	CWC	SSC score
SCP	0.816	1.017	0.47	0.789	0.0
CV+	0.857	1.056	0.462	0.781	0.648
J+aB	0.736	0.706	0.462	0.854	0.397
CQR	0.814	1.028	0.488	0.787	0.744

FIGURE A.4: Test data metrics for the 4 different strategies and 1 particular experiment (no 5-folds CV) for $\alpha = 0.20$. From top to bottom: SCP, CV+, J+aB and CQR.

Appendix B

Time series original problem

Unless it is specified otherwise, the miscoverage level was set to $\alpha = 0.2$ (*i.e.* 80% of expected coverage) for the visualizations.

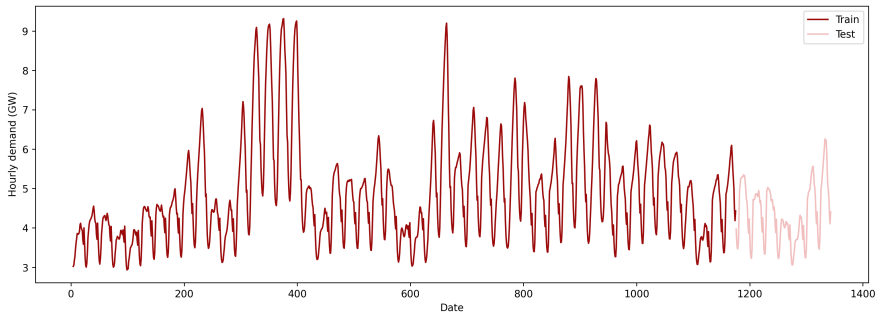


FIGURE B.1: Time series problem's dataset.

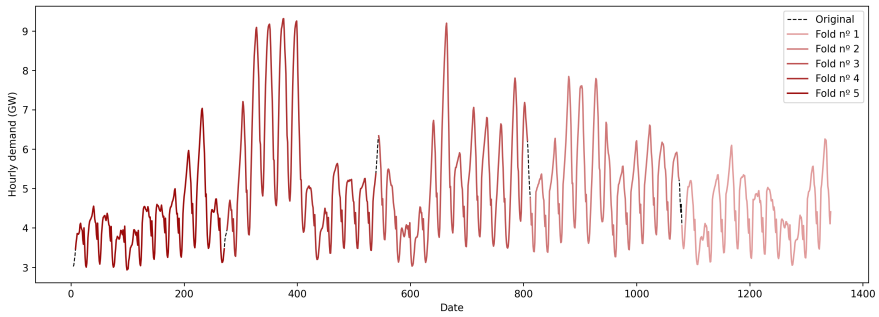
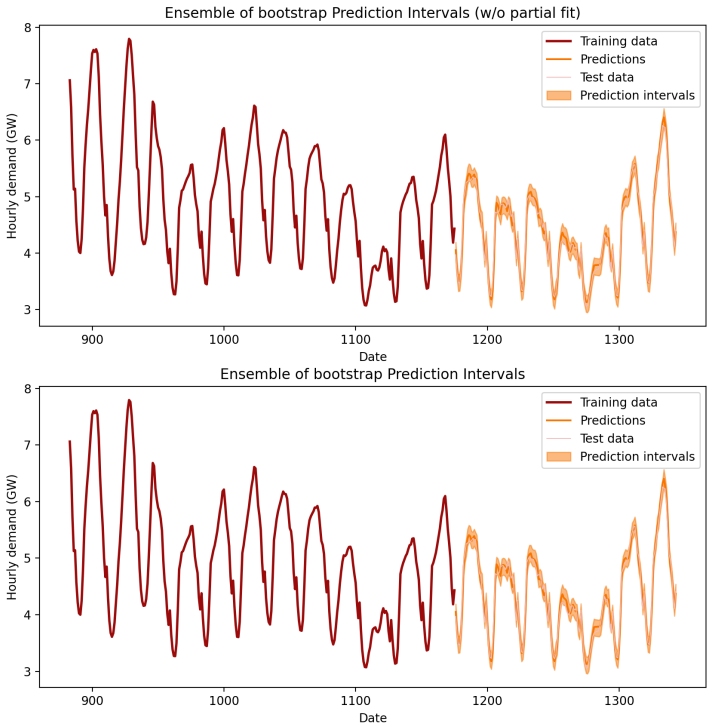
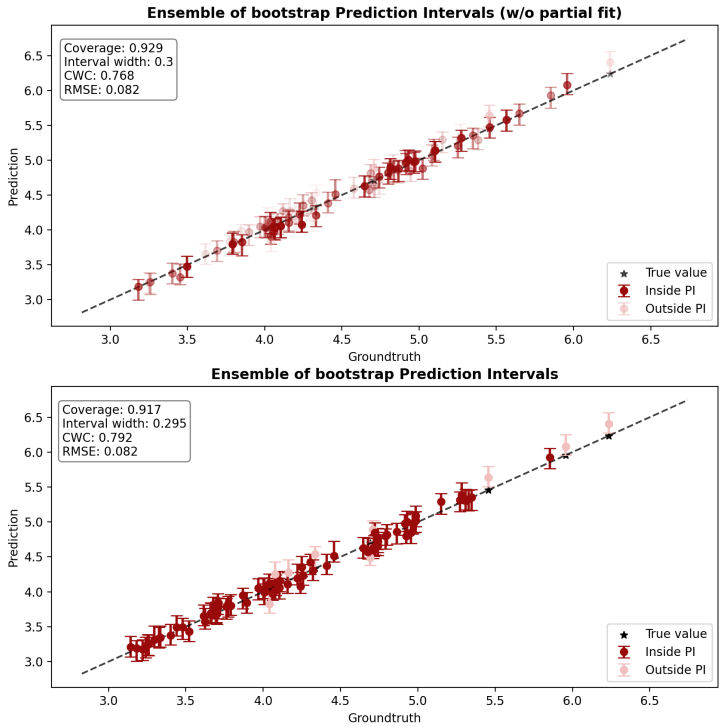


FIGURE B.2: 5-fold splits from the original dataset and for the assessments based on cross-validations.

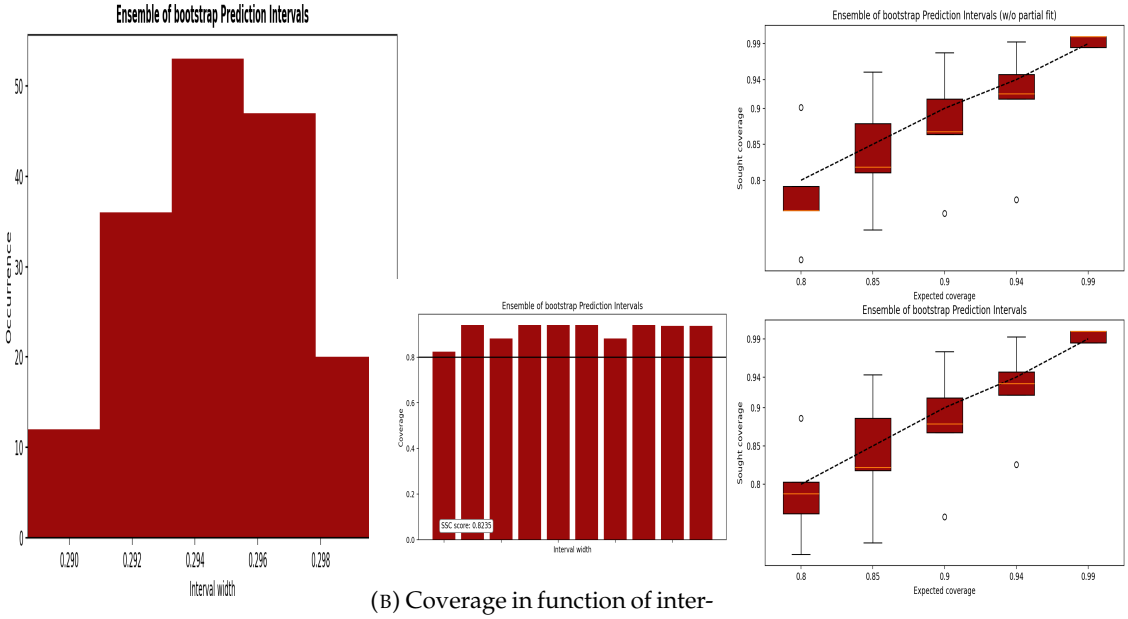


(A) Prediction intervals



(B) Goodness of the prediction intervals (prediction vs. ground truth). Just a 50% of the data is shown due to visualization reasons.

FIGURE B.3: Visualizations related to the prediction intervals for the test data and the EnbPI strategy (without partial fit, top; and with it, bottom).



(A) Intervals' width histograms (B) Coverage in function of intervals' width (C) Coverage in function of α

FIGURE B.4: Visualizations related to width & coverage distributions for the test data and the EnbPI strategy (without partial fit, top; and with it, bottom).

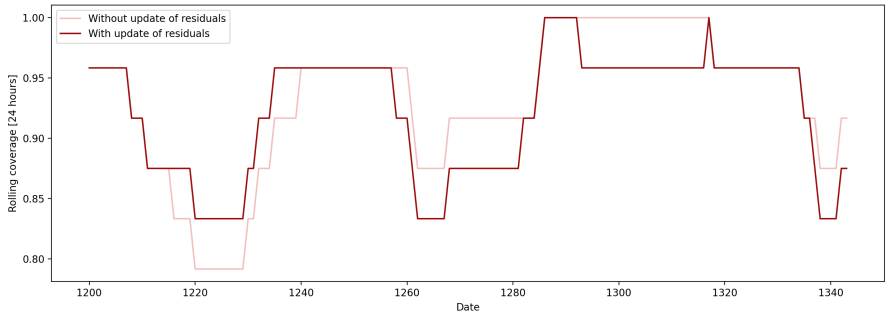


FIGURE B.5: Test data coverage, for the EnbPI strategies, in function of time (grouped within 24h rolling windows).

	Coverage	Int. width	RMSE	CWC	SSC score
EnbPI_nP	0.929	0.3	0.082	0.768	0.0
EnbPI	0.917	0.295	0.082	0.792	0.824

FIGURE B.6: Test data metrics for the EnbPI strategy and 1 particular experiment (no 5-folds CV) for $\alpha = 0.20$. From top to bottom: EnbPI without partial fit (EnbPI_nP), EnbPI with it (EnbPI).

Appendix C

Time series problem with change point in test

Unless it is specified otherwise, the miscoverage level was set to $\alpha = 0.05$ (*i.e.* 95% of expected coverage) for the visualizations.

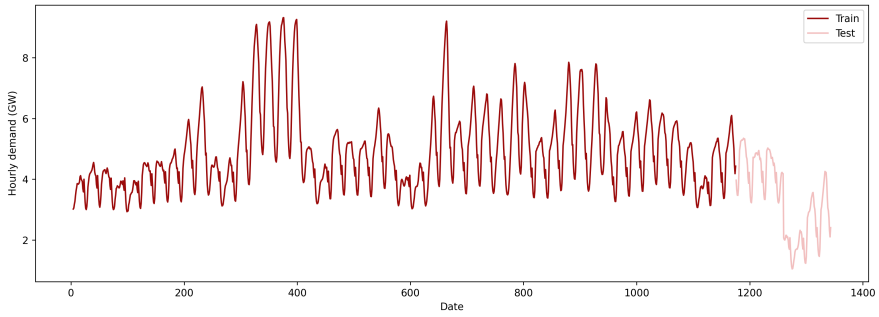


FIGURE C.1: Time series problem's dataset when a change point is added to the test split.

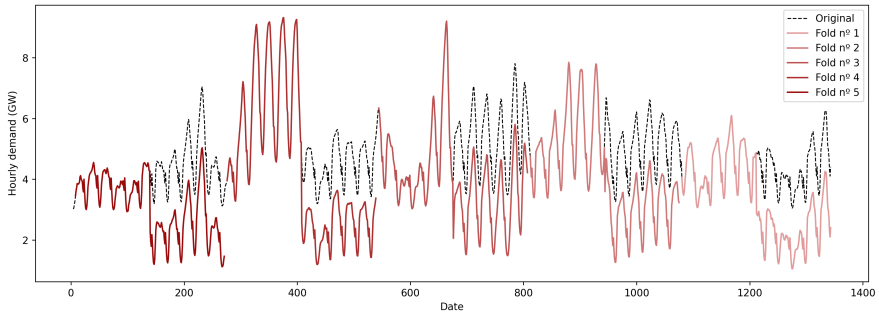
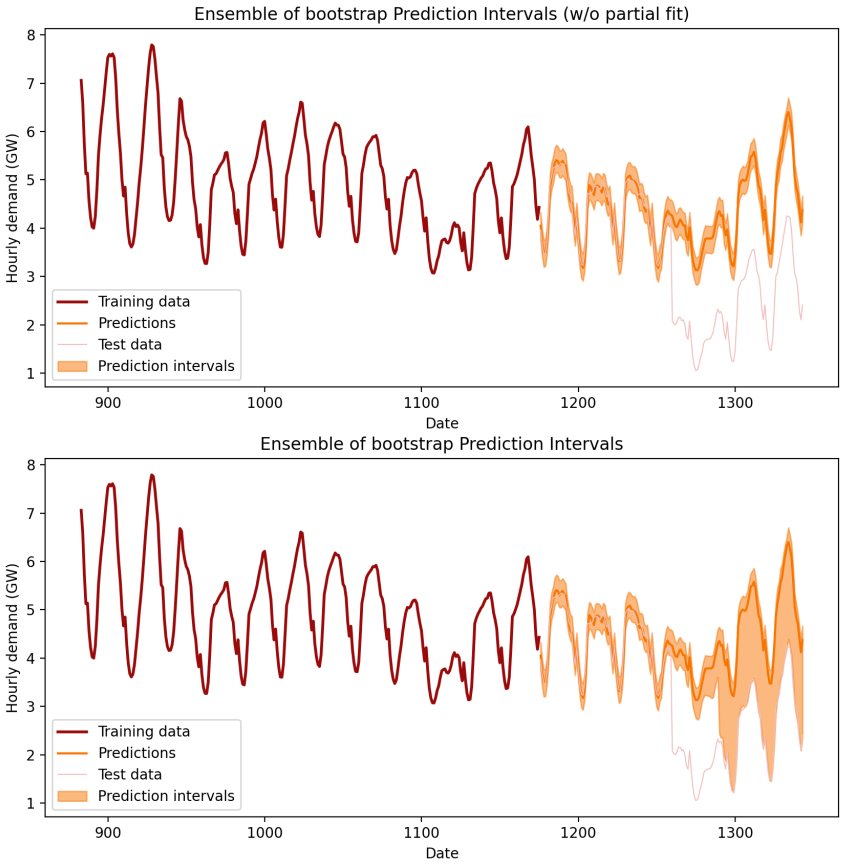
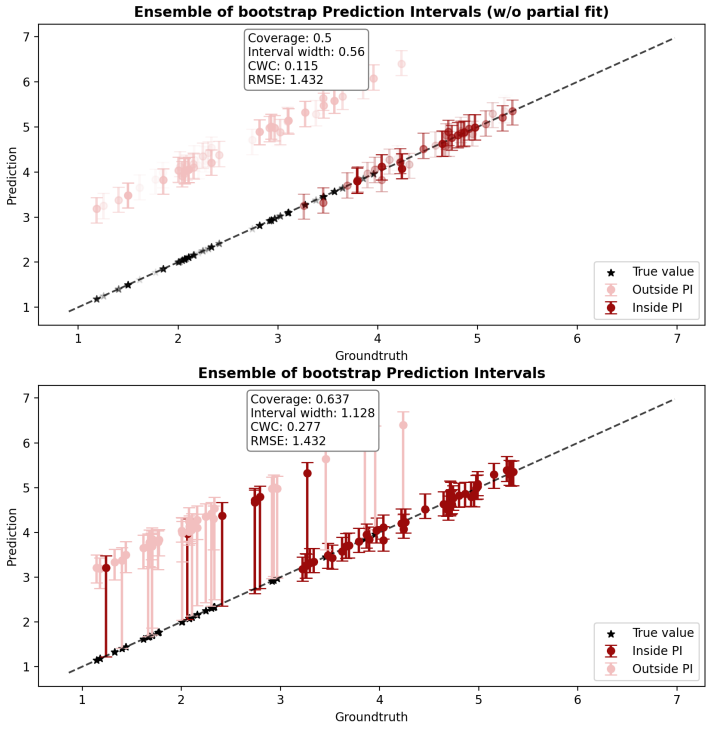


FIGURE C.2: 5-fold splits from the original dataset and with a change point in the test split.

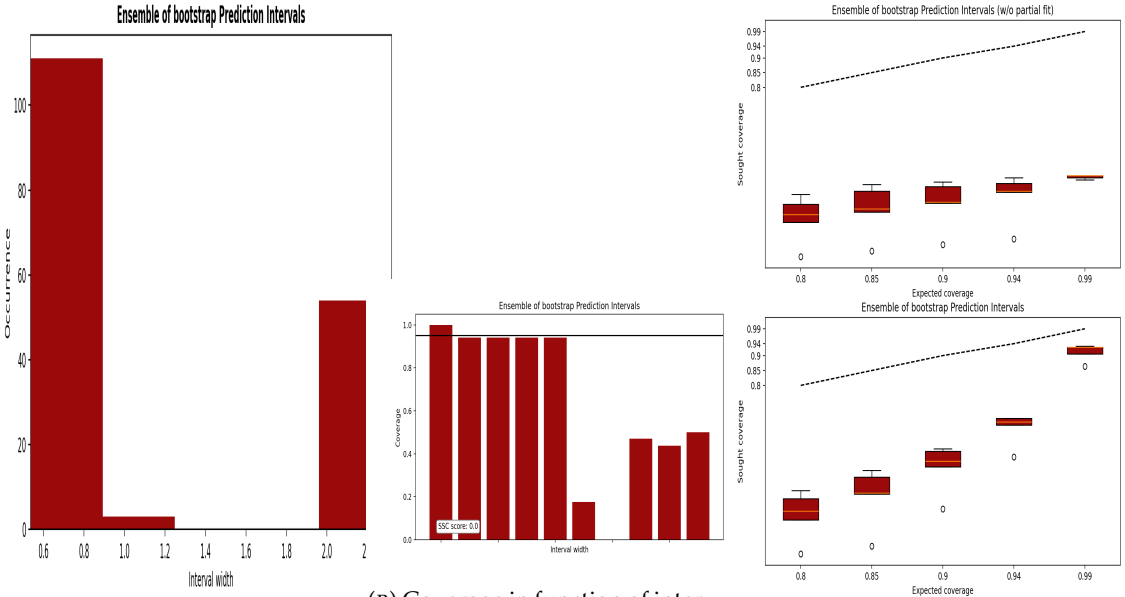


(A) Prediction intervals



(B) Goodness of the prediction intervals (prediction vs. ground truth). Just a 50% of the data is shown due to visualization reasons.

FIGURE C.3: Visualizations related to the prediction intervals for the test data (with a change point) & the EnbPI strategy (without partial fit, top; and with it, bottom).



(A) Intervals' width histograms (B) Coverage in function of intervals' width (C) Coverage in function of α

FIGURE C.4: Visualizations related to width & coverage distributions for the test data (with a change point) & the EnbPI strategy (without partial fit, top; and with it, bottom).

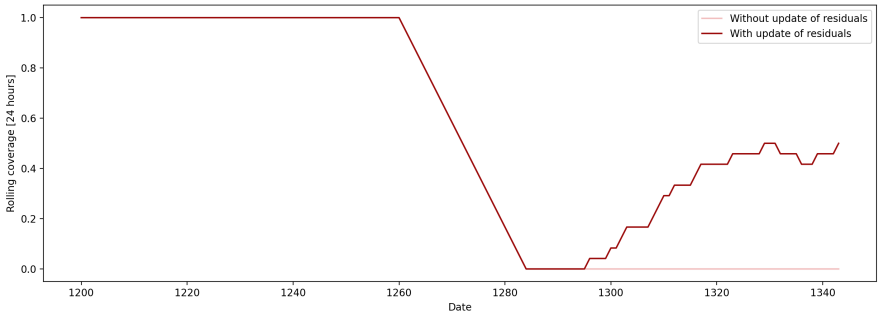


FIGURE C.5: Test data coverage (change point's dataset), for the EnbPI strategies, in function of time (grouped within 24h rolling windows).

	Coverage	Int. width	RMSE	CWC	SSC score
EnbPI_nP	0.5	0.56	1.432	0.115	0.0
EnbPI	0.637	1.128	1.432	0.277	0.0

FIGURE C.6: Test data metrics for the EnbPI strategy and 1 particular experiment (no 5-folds CV) for $\alpha = 0.05$ and the chage point dataset. From top to bottom: EnbPI without partial fit (EnbPI_nP), EnbPI with it (EnbPI).

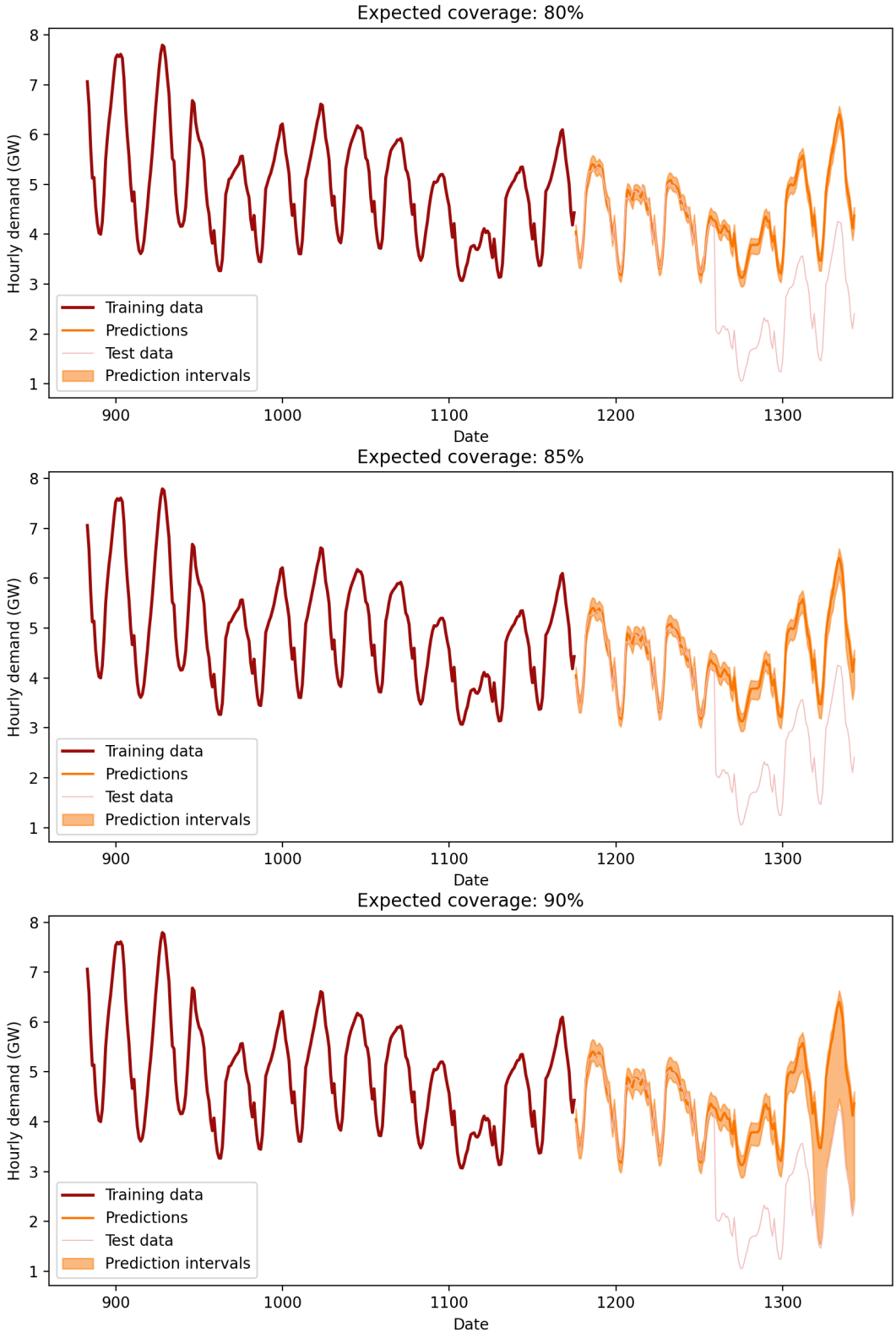


FIGURE C.7: Prediction intervals (change point's dataset) in function of time and for different values of α.

Bibliography

Angelopoulos, Anastasios N. and Stephen Bates (2021). "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification". In: *CoRR* abs/2107.07511. arXiv: 2107.07511. URL: https://arxiv.org/abs/2107.07511.

Auer, Andreas et al. (2023). "Conformal Prediction for Time Series with Modern Hopfield Networks". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., pp. 56027–56074. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/aef75887979ae1287b5deb54a1e3cbda-Paper-Conference.pdf.

Barber, Rina Foygel et al. (2021). "Predictive inference with the jackknife+". In: *The Annals of Statistics* 49.1, pp. 486–507. DOI: 10.1214/20-AOS1965. URL: https://doi.org/10.1214/20-AOS1965.

Chernozhukov, Victor, Kaspar Wüthrich, and Yinchu Zhu (2021). "Distributional conformal prediction". In: *Proceedings of the National Academy of Sciences* 118.48, e2107794118. DOI: 10.1073/pnas.2107794118. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2107794118. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2107794118.

Foygel Barber, Rina et al. (Aug. 2020). "The limits of distribution-free conditional predictive inference". In: *Information and Inference: A Journal of the IMA* 10.2, pp. 455–482. ISSN: 2049-8772. DOI: 10.1093/imaiai/iaaa017. eprint: https://academic.oup.com/imaiai/article-pdf/10/2/455/38549621/iaaa017.pdf. URL: https://doi.org/10.1093/imaiai/iaaa017.

Gibbs, Isaac and Emmanuel Candes (2021). "Adaptive Conformal Inference Under Distribution Shift". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 1660–1672. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf.

Gibbs, Isaac, John J. Cherian, and Emmanuel J. Candès (2023). *Conformal Prediction With Conditional Guarantees*. arXiv: 2305.12616 [stat.ME].

Guan, Leying (July 2022). "Localized conformal prediction: a generalized inference framework for conformal prediction". In: *Biometrika* 110.1, pp. 33–50. ISSN: 1464-3510. DOI: 10.1093/biomet/asac040. eprint: https://academic.oup.com/biomet/article-pdf/110/1/33/49160126/asac040.pdf. URL: https://doi.org/10.1093/biomet/asac040.

Hyndman, R.J. and G. Athanasopoulos (2014). *Forecasting: principles and practice*. OTexts. ISBN: 9780987507105. URL: https://books.google.es/books?id=gDuRBAAAQBAJ.

Izbicki, Rafael, Gilson Shimizu, and Rafael Stern (Aug. 2020). "Flexible distribution-free conditional predictive bands using density estimators". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 3068–3077. URL: https://proceedings.mlr.press/v108/izbicki20a.html.

Izbicki, Rafael, Gilson Shimizu, and Rafael B. Stern (2022). "CD-split and HPD-split: Efficient Conformal Regions in High Dimensions". In: *Journal of Machine Learning Research* 23.87, pp. 1–32. URL: http://jmlr.org/papers/v23/20-797.html.

Jung, Christopher et al. (2022). *Batch Multivalid Conformal Prediction*. arXiv: 2209.15145 [cs.LG].

Khosravi, Abbas, Saeid Nahavandi, and Doug Creighton (2010). "Construction of Optimal Prediction Intervals for Load Forecasting Problems". In: *IEEE Transactions on Power Systems* 25.3, pp. 1496–1503. DOI: 10.1109/TPWRS.2010.2042309.

Kim, Byol, Chen Xu, and Rina Foygel Barber (2020). "Predictive inference is free with the jackknife+-after-bootstrap". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. , Vancouver, BC, Canada, Curran Associates Inc. ISBN: 9781713829546.

Kivaranovic, Danijel, Kory D. Johnson, and Hannes Leeb (Aug. 2020). "Adaptive, Distribution-Free Prediction Intervals for Deep Networks". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 4346–4356. URL: https://proceedings.mlr.press/v108/kivaranovic20a.html.

Koenker, Roger and Gilbert Bassett (1978). "Regression Quantiles". In: *Econometrica* 46.1, pp. 33–50. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1913643 (visited on 04/27/2024).

Lei, Jing, Max G'Sell, et al. (July 2018). "Distribution-Free Predictive Inference for Regression". In: *Journal of the American Statistical Association* 113.523, pp. 1094–1111. DOI: 10.1080/01621459.2017.130. URL: https://ideas.repec.org/a/taf/jnlasa/v113y2018i523p1094-1111.html.

Lei, Jing and Larry Wasserman (Jan. 2014). "Distribution-free prediction bands for non-parametric regression". In: *Journal of the Royal Statistical Society Series B* 76.1, pp. 71–96. URL: https://ideas.repec.org/a/bla/jorssb/v76y2014i1p71-96.html.

MAPIE developers (2024). *Theoretical Description for Conformity Scores - MAPIE Documentation*. [Online; accessed 2024/03/25]. URL: https://mapie.readthedocs.io/en/stable/theoretical_description_conformity_scores.html.

Papadopoulos, Harris et al. (2002). "Inductive Confidence Machines for Regression". In: *Machine Learning: ECML 2002*. Ed. by Tapio Elomaa, Heikki Mannila, and Hannu Toivonen. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 345–356. ISBN: 978-3-540-36755-0.

Podkopaev, Aleksandr and Aaditya Ramdas (July 2021). "Distribution-free uncertainty quantification for classification under label shift". In: *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*. Ed. by Cassio de Campos and Marloes H. Maathuis. Vol. 161. Proceedings of Machine Learning Research. PMLR, pp. 844–853. URL: https://proceedings.mlr.press/v161/podkopaev21a.html.

Romano, Yaniv, Evan Patterson, and Emmanuel J. Candès (2019). *Conformalized Quantile Regression*. arXiv: 1905.03222 [stat.ME].

Romano, Yaniv, Matteo Sesia, and Emmanuel J. Candès (2020). *Classification with Valid and Adaptive Coverage*. arXiv: 2006.02544 [stat.ME].

Sesia, Matteo and Yaniv Romano (2021). "Conformal Prediction using Conditional Histograms". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., pp. 6304–6315.

Tibshirani, Ryan J et al. (2019). "Conformal Prediction Under Covariate Shift". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32.

Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/8fb21ee7a2207526da55a679f0332de2-Paper.pdf.

Vovk, Vladimir (Nov. 2012). "Conditional Validity of Inductive Conformal Predictors". In: *Proceedings of the Asian Conference on Machine Learning*. Ed. by Steven C. H. Hoi and Wray Buntine. Vol. 25. Proceedings of Machine Learning Research. Singapore Management University, Singapore: PMLR, pp. 475–490. URL: https://proceedings.mlr.press/v25/vovk12.html.

— (June 2015). "Cross-conformal predictors". In: *Annals of Mathematics and Artificial Intelligence* 74.1, pp. 9–28. ISSN: 1573-7470. DOI: 10.1007/s10472-013-9368-4. URL: https://doi.org/10.1007/s10472-013-9368-4.

Vovk, Vladimir, Alex Gammerman, and Glenn Shafer (Jan. 2005). "Algorithmic Learning in a Random World". In: DOI: 10.1007/b106715.

Xu, Chen and Yao Xie (July 2021). "Conformal prediction interval for dynamic time-series". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 11559–11569. URL: https://proceedings.mlr.press/v139/xu21h.html.

— (2023). *Conformal prediction for time series*. arXiv: 2010.09107 [stat.ME].

Zaffran, Margaux (2023). *Introduction to Conformal Prediction*. [Online; accessed 25-March-2024]. URL: https://conformalpredictionintro.github.io/assets/files/cptuto.pdf.