Conformal Inference Beyond Exchangeability

Alessadro Ciancetta, Ramón Talvi,
Alessandro Tenderini, Akash Yadav

Barcelona School of Economics

March 9, 2023

Abstract

The report builds upon the paper "Conformal Prediction Beyond Exchangeability" by R.F. Barber et al. and aims to explore the impact of lifting the usual assumptions of exchangeability and symmetry on both split and full conformal inference methods. Additionally, throughout the report we conduct a constant comparison of split and full conformal inference methods, taking into account the fact that there is a trade off between computational expensiveness/feasibility and statistical efficiency (split-full trade off). After explaining the split and full conformal inference methods through theory and application, we unravel what happens when only the exchangeability assumption is lifted but symmetry still holds. Subsequently, we analyze the setting where both assumptions are violated and its theoretical repercussions on the coverage bounds. We then discuss a hybrid method -Jackknife+- that addresses the split-full trade off. Finally, we show empirical applications on simulated data points and a real data set to convey the main findings of the paper.

Contents

1 Introduction[1]

Machine learning algorithms usually have remarkable forecasting potential, often at the cost of sacrificing interpretability and making inference more abstruse. One of the main challenges, specially in deep learning architectures, is to quantify the uncertainty of a given prediction. Conformal inference provides a framework for constructing prediction intervals or sets no matter the underlying distribution of the data. In other words, conformal inference is a distribution-free method and enables us to construct valid prediction sets or intervals (for classification and regression problems respectively).

1.1 Assumptions

Standard conformal inference relies on two main assumptions: exchangeability and symmetry. Doing without a mathematical formalization, data points are exchangeable if the joint probability distribution is invariant to any permutation of them. Symmetry, on the other hand, is a property of the algorithm: a symmetric algorithm implies that any reordering of the data points produces the same output.

Data point exchangeability is not equivalent but related to the independence and identically distributed assumption: i.i.d is a slightly stronger assumption than exchangeability. Hence, when the data distribution drifts or changes for different data points or the data points exhibit some dependence we are in a scenario where exchangeability is violated. In the case of symmetry, the assumption is violated if, for instance, recent observations get assigned a different weight than those form the distant past (time). Alternatively, data points collected from a nearby spatial domain can be assigned a different weight than more distant ones (location).

1.2 General Setting

Suppose we have as input an image $x_i \in R^d$ for $i \in \{1, \dots, N\}$ and k possible labels such that y $\in \{1, \dots, k\}$. The standard approach[2] implies dividing out data points in three sets: a training set where we fit the algorithm and obtain $\hat{u} : X \rightarrow \mathbb{R}$, a callibration or hold-out set of pairs $(x_i, y_i) \dots (x_n, y_n)$ for i $\in \{1, \dots, n\}$ and a test point (x_{n+1}, y_{n+1}). Note that, consistently with the paper which the report builds upon, we have n points belonging to the calibration set.

In our specific example we are in a classification setting, therefore, the main goal is to predict a set $\tau(x_{n+1}) \subseteq y$ (subset of possible labels) that contains the true class y_{n+1} with a certain degree of confidence. Analogously, in a hypothetical regression setting our goal is to construct prediction intervals that contain the true label y_{n+1} with a given confidence level.

1.3 Baseline Concepts

In this subsection we are going to cover the main concepts that will be referenced throughout the report. One key concept in conformal inference is the idea of *coverage*. We will define it as the probability that our prediction belongs to the prediction set or is contained in the prediction interval. The coverage guarantee -later formalized as a theorem- implies that the latter probability is, at least, (1 - α):

$$P\left(y_{n+1} \in \tau(x_{n+1})\right) \geq 1 - \alpha \tag{1}$$

Note that α is the significance level, usually 5% or 10%, that is a measure of the degree of confidence.

Another structural concept in conformal inference is *non conformity score*. The score function $S(x_i, y_i) \in \mathbb{R} \quad \forall \quad i = 1, \dots, n$ is a measure of the "weirdness" or how unusual a given observation is. Examples of score functions are (absolute) residuals $R_i = \left|y_i - \hat{\mu}(x_{n+1})\right|$ in a regression setting and $E_i = \sum_{j=1}^{k} \hat{\pi}(x_{n+1})_{(j)}$ in a classification setting, where E_i is the amount of probability mass you need to include in your model for it to contain the true label.

[1] The theory of this section is based upon two main papers referenced in the bibliography: "A gentle introduction to conformal prediction and distribution-free uncertainty quantification" and "Conformal Inference beyond exchangeability."

[2] For the sake of clarity, we explain the general setting assuming the split conformal inference method is used.

1.4 Objectives

Correct Coverage: We want to make sure that the conformal inference has attained the correct coverage. In a subsequent section we will explore a diagnostic method to evaluate if our procedure has indeed the correct coverage.

Adaptivity: If we choose all the possible labels as our predicted set with probability $(1 - \alpha)$ and the empty set with probability α, we would be right 90% of the time but our prediction is completely useless. Hence, we would like our set size $|\tau(X)|$ to be small. More precisely, we would like to have adaptive sets: set size should be small for easy examples and bigger for harder inputs in order to adequately capture the model's uncertainty.

2 Exchangeability and Symmetric Algorithm

All the analysis in this section presumes that both exchangeability and symmetry assumptions hold. For both split and full conformal inference, we will explain in detail the main idea of each method and propose an algorithm for classification and for regression. Moreover, for split conformal inference, as it is a more intuitive method to grasp, we will present two illustrative examples (one for regression and another for classification). Afterwards, we will present the coverage theorem -which holds for both methods- and finally, mention the basic guidelines to evaluate conformal inference.

2.1 Split conformal inference

2.1.1 Main idea[3]

Split conformal inference implies splitting our data set into training and calibration in order to predict a test point. The general setting resembles the one already proposed in the preceding section. We run our algorithm on the training data and obtain prediction points $\hat{u} : X \rightarrow \mathbb{R}$, and then use the calibration set to compute the non conformity score for each observation. Assuming we are in a regression setting, we choose the score function to be the residuals: $R_i = |y_i - \hat{\mu}(x_{n+1})|$.

Subsequently, we compute the prediction intervals $\hat{C}_n$ for the target vector (x_{n+1}):

$$\hat{C}_n(x_{n+1}) = \hat{\mu}(x_{n+1}) \pm Q_{1-\alpha}\left(\sum_{i=1}^{n} \frac{1}{n+1}\delta_{R_i} + \frac{1}{n+1}\delta_{+\infty}\right) \tag{2}$$

where $\hat{\mu}(x_{n+1})$ is the point estimate of the test point provided by the trained model, $Q_{(1-\alpha)}$ is the $(1 - \alpha)$ quantile of the sorted residuals and δ_a is the point mass at a. In essence, to obtain the interval we sort the residuals in increasing order and select the $\lceil(1 - \alpha)(n + 1)\rceil$-th smallest of $R_1, \dots, R_n$. Finally, we use the preceding to compute the prediction interval for the target vector.

To get comfortable with this notation let us analyze what these delta point masses are and what is their functionality in this notation. Since we want to select the $(1 - \alpha)$ quantile of our residuals, these delta point masses take value 1 if our quantile demarcation encompasses that residual. For normalization, we have $\delta_{+\infty}$ term to account for the odd case where $R_1, \dots.., R_n$ i.e. δ_{R_i} takes value 1 for all $i = 1, \dots., n$ and we have not yet reached our $1 - \alpha$ quantile. In that case, we see our quantile function $Q_{1-\alpha} \rightarrow \infty$. This becomes more intuitive by using the properties of cdf F namely, $F(x) \rightarrow 1$ as $x \rightarrow +\infty$ and the quantile function is the inverse of the cdf function.

In section 2.3 we will present the coverage theorem to show that in this context the coverage is guaranteed to be at least $(1 - \alpha)$.

2.1.2 Algorithm

The following algorithm portrays how we would implement split conformal inference in a regression setting.

Let $Z = (X, Y)$ be the historical data.

[3] Based on section 3.1 of paper "Conformal Prediction beyond exchangeability".

1. Divide Z into two disjoint sets:
   - Z_t training set where cardinality $|Z_t| = m$
   - Z_c callibration set where cardinality $|Z_c| = n$
2. Fit model $\hat{\mu}(x)$ using Z_t
3. Define non conformity score $S(x, y)$ function
4. Apply $S(x, y)$ to each element of Z_c and obtain callibration score $s_1 \dots s_n$
5. Sort callibration scores $S(x_i, y_i)$ for $i = 1, \dots, n$
6. Compute $Q_{1-\alpha}\left(\sum_{i=1}^{n} \frac{1}{n+1}\delta_{S_i} + \frac{1}{n+1}\delta_{+\inf}\right)$
7. Compute prediction interval:
$$\hat{C}_n(x_{n+1}) = \hat{\mu}(x_{n+1}) \pm Q_{1-\alpha}\left(\sum_{i=1}^{n} \frac{1}{n+1}\delta_{R_i} + \frac{1}{n+1}\delta_{+\inf}\right)$$

The following algorithm shows how we would implement split conformal inference in a classification task. We start the at step 5 given the first four steps of the algorithm are equivalent than those just stated for a continuous target variable.

6. Take $(1-\alpha)$ quantile of these scores and compute:
$$\hat{q} = \left\lceil (1-\alpha)\frac{n+1}{n} \right\rceil$$
where $\frac{n+1}{n}$ is a finite sample correction.
7. Determine prediction set:
$$\tau(x_{n+1}) = \{y : S(x, y) \leq \hat{q}\}$$

2.1.3 Application

Example 1: Adaptive Prediction Sets for Classification[4]

A naive approach for establishing the prediction set would be to include the top scoring classes until the total probability mass exceeds $(1-\alpha)$. Unless we have a perfect model for the conditional probability distribution $P(Y|X = x)$, this would most likely lead to over fitting and prediction intervals being too narrow.

Consider the case where we fit a model on the training set and use the calibration set obtain the softmax output: that is, the probability that observation x_i belongs to class k. Define as conformal score $E_i = \sum_{j=1}^{k} \hat{\pi}(x_i)_{(j)}$, where $\hat{\pi}(x_i)$ is the sorted softmax output and k is the rank of the class. The conformal score E_i is the total mass of softmax output until you reach the true class. In other words, the amount of estimated probability mass you need to include in your set in order to include the true label. The set $\{E_i\}_{i=1}^{n}$ is a calibration of the model's uncertainty.

The following step involves computing the $(1-\alpha)$ quantile of the conformal scores. More formally:
$$\hat{q}_{1-\alpha} = \lceil (1-\alpha)\frac{(n+1)}{n} \rceil$$
where $\frac{n+1}{n}$ is a finite sample correction.

The finally step involves forming the prediction sets. We select {the k most likely classes where $\sum_{j=1}^{k} \hat{\pi}(x_{n+1})_{(Y_{n+1})} \leq \hat{q}$}. In simple terms, we include in our prediction set the most likely classes up until the cumulative softmax score is above the threshold $\hat{q}$. If the total softmax mass exceeds $\hat{q}$, then with a confidence level of $(1-\alpha)$ our prediction set contains the true label.

[4] Example is adapted from section 2.2 of the paper "A Gentle Introduction to conformal inference and ".

Example 2: Conformalized Quantile Regression[5]

Suppose we are in the usual setting we have been working on. We have a model for the 5% quantile $\hat{t}_{\frac{\alpha}{2}}(x)$ and a model for the 95% quantile $\hat{t}_{1-\frac{\alpha}{2}}(x)$. Using the approximate quantiles as prediction intervals is an heuristic measure of uncertainty that requires conformal calibration to turn it into a more rigorous notion of uncertainty. The main idea is to inflate the bands by a constant to get the right coverage.

After estimating both models, for each calibration observation we compute the non conformity scores E_i. This measures the vertical distance from observation i to the closest prediction band, considering observations that fall outside the bands as positive and the ones that fall within as negative.

We take approximately the 90% quantile $\hat{q}$ of the non conformity scores to construct the prediction set:

$$\tau(X_{n+1}) = [\hat{t}_{\frac{\alpha}{2}}(x_{n+1}) - \hat{q}, \hat{t}_{1-\frac{\alpha}{2}}(x_{n+1}) + \hat{q}]$$

If $\hat{q} \geq 0$, then quantile bands get pushed apart, meaning the original bands had a coverage slightly below 90%. If $\hat{q} \leq 0$, then the initial bands were too conservative so the intervals are tighten.

2.2 Full conformal inference

Full conformal inference is an alternative approach to quantify uncertainty and obtain prediction intervals or sets. Recall the split-full trade off: while split conformal inference is a method that is much less computationally expensive than the full one, the full conformal method achieves a higher degree of statistical efficiency. We will explain the main idea of full conformal inference, propose an algorithm for its implementation and present two examples -one for regression and another for classification.

2.2.1 Main idea

Unlike split conformal inference, the full conformal method does not require to split the data points into different sets. In fact, we will train our model on training and test data together. Assume we observe $Z = (X_1, Y_1), \dots (X_n, Y_n), (X_{n+1}, Y_{n+1})$ and note that in this context n denotes the number of data points in the training set and $n+1$ refers to the test point. Also, recall that we are not only assuming exchangeability but also that the algorithm is symmetric. This implies that any permutation π of $[n+1]$ data points will yield exactly the same output:

$$\mathcal{A}((X_{\pi(1)}, Y_{\pi(1)}), \dots, (X_{\pi(n+1)}, Y_{\pi(n+1)})) = \mathcal{A}((X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}))$$

Supposing we have k labels, we will fit the model k times, one for each label. In more formal terms: for each label $Y_j \in \mathbb{R}$ where $j = \{1, \dots, k\}$, we fit the model using the $n+1$ data points:

$$\hat{\mu}^y = \mathcal{A}((X_1, Y_1), \dots (X_n, Y_n), (X_{n+1}, Y_{n+1})) \tag{3}$$

We then proceed to compute the non conformity score -in this specific case, the residuals-, for each data point in the training set and for the test set across all possible labels:

$$R_i^y = \begin{cases} |y_i - \hat{\mu}^y(x_i)| & i = 1, \dots, n \\ |y_i - \hat{\mu}^y(x_{n+1})| & i = n+1 \end{cases}$$

Once we have computed the residuals for each label of y for all data points, we will determine the prediction set. In essence, we will include the label in the prediction set if R_{n+1}^y is smaller or equal than the $(1-\alpha)$ quantile of the sorted residuals in increasing order. Formally, the prediction set $\hat{C}_n$ for the test point (X_{n+1}) is given by:

$$\hat{C}_n(X_{n+1}) = \left\{ y \in \mathbb{R} \,:\, R_{n+1}^y \leq Q_{1-\alpha}\left(\sum_{i=1}^{n+1} \frac{1}{n+1} \delta_{R_i^y}\right)\right\} \tag{4}$$

[5] Example is taken from section 2.1 of the paper "A gentle introduction to conformal inference and distribution-free uncertainty quantification".

2.2.2 Algorithm

Let Z be the full training and test set. For each $y_j \in \mathbb{R}$ where $j = \{1, \dots, k\}$:

1. Fit model $\hat{\mu}^y$ using Z
2. Define non conformity score $S(x, y)$ function
3. Apply $S(x, y)$ to each element of Z but compute by separate the residuals in training and test set.
4. Sort non conformity scores $S(x_i, y_i)$ for $i = 1, \dots, n+1$
5. Compute $(1-\alpha)$ quantile: $Q_{1-\alpha}(\sum_{i=1}^{n+1} \frac{1}{n+1} \delta_{R_i^y})$
6. Compute prediction set: $\hat{C}_n(X_{n+1}) \leq \{y \in \mathbb{R} : Q_{1-\alpha}(\sum_{i=1}^{n+1} \frac{1}{n+1} \delta_{R_i^y})\}$

2.3 Coverage Theorem

We will prove the coverage theorem for the case where our assumptions are valid. Note that similar coverage guarantees have been proven for more robust methods where we take away these assumptions one by one. We will cover these in next sections without giving a formal rigorous proof and focusing only on the main idea.

Theorem 1 (Full Conformal Coverage Guarantee [Vovk et al., 2005]). *If the data points $(X_1, Y_1), \dots\dots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are i.i.d. (or more generally exchangeable), and the algorithm $\mathcal{A}$ treats the input data points symmetrically, then the full conformal prediction set satisfies*

$$P\left\{Y_{n+1} \in \hat{C}_n(X_{n+1})\right\} \geq 1 - \alpha \tag{5}$$

By considering a trivial algorithm $\mathcal{A}$ that returns the same fixed 'pre-fitted' $\hat{\mu}$ regardless of the input data, this theorem also implies that same coverage result holds for the split conformal method. This is evident in our simulation example where split conformal method gives rise to larger confidence intervals but gives approximately same coverage.

Proof. Let us denote the i-th residual, at the hypothesized value $y = Y_{n+1}$ by $R_i = R_i^{Y_{n+1}}$. Since our data points are i.i.d., and the fitted model is constructed via algorithm $\mathcal{A}$ that treats $n+1$ data points symmetrically, the residuals $R_i = |Y_i - \hat{\mu}(X_1)|$ are exchangeable.

Define the set of "strange" points as

$$S(R) = \left\{ i \in [n+1] : R_i > Q_{1-\alpha}\left(\sum_{j=1}^{n+1} \frac{1}{n+1} \cdot \delta_{R_j}\right)\right\}$$

Index i corresponds to a "strange" point if its residual is one of the largest elements of the list of residuals. We observe, $|S(R)| \leq \alpha(n+1)$. Coverage fails if and only if our test point $n+1$ is "strange", i.e., $n+1 \in S(R)$ i.e. if and only if $R_{n+1} > Q_{1-\alpha}\left(\sum_{i=1}^{n+1} \frac{1}{n+1} \cdot \delta_{R_i}\right)$. Therefore, the probability that coverage fails is given by

$$P\left\{Y_{n+1} \notin \hat{C}_n(X_{n+1})\right\} = P\{n+1 \in S(R)\} = \frac{1}{n+1} \sum_{i=1}^{n+1} P\{i \in S(R)\}$$

This is due to the exchangeability of $R_1, R_2, \dots\dots, R_{n+1}$.

Probability in last step can equivalently be written as expectation of indicator random variable

$$P\left\{Y_{n+1} \notin \hat{C}_n(X_{n+1})\right\} = \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} 1\{i \in S(R)\}\right] = \frac{1}{n+1} \mathbb{E}[|S(R)|] \leq \frac{1}{n+1} \cdot \alpha(n+1) = \alpha$$

$1 - P\left\{Y_{n+1} \notin \hat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$ which proves our theorem.

2.4 Evaluating Conformal Inference

In section 1.4 we covered the two of the main objectives pursued when implementing conformal inference. On one hand, attaining the correct coverage and the other, predicted sets that are adaptive i.e the size of the predicted set reflects the degree of difficulty of classifying the observation.

We want to run a posterior diagnostics to ensure the objectives have been fulfilled. The evaluation assessments of conformal inference can be classified into two main categories[6]: coverage correctness and adaptiveness evaluation. Correct coverage implies empirically checking if the attained coverage indeed satisfies Theorem 1. One of the main challenges involves accounting for the finite sample variability -in the example in section 2.1.3 we made finite sample correction when computing the quantile- inherent to real data sets. Regarding adaptiveness evaluation, note that having the smallest possible prediction set is not desirable per se: we seek to find a prediction set that faithfully reflects the model's uncertainty. This implies the prediction set's cardinality should reflect how challenging an input is in terms of prediction label.

2.4.1 Coverage Correctness

The first most intuitive diagnostic aims at verifying that, indeed, the conformal procedure has the correct coverage. The idea is to compare the empirical coverage to the theoretical one. Assuming that n is sufficiently large[7], Vladimir Vovk[8] showed the coverage distribution has an analytical form:

$$P(Y_{test} \in C(X_{test}) | \{(X_i, Y_i)\}_{i=1}^{n}) \sim Beta(n+1-l, l)$$

where $l = \lfloor (n+1)\alpha \rfloor$. Notice that the conditional expectation above is the coverage with an hypothetical infinite validation set and holding the calibration set fixed.

We now want to construct the empirical coverage that should look like the theoretical one. For a real data set containing $n + n_{val}$ data points, we randomly split the $n + n_{val}$ data points R times into calibration and validation set running conformal inference each time. Mathematically, the empirical coverage is given by:

$$C_j = \frac{1}{n_{val}} \sum_{i=1}^{n_{val}} 1\{y_{i,j}^{\text{val}} \in C_j(x_{i,j}^{\text{val}})\}, \quad j = 1 \ldots, R$$

where $(x_{i,j}^{val}, y_{i,j}^{val})$ is the i-th validation example in trial j.

If the histogram of C_j is skewed or biased there it signals that there is a sub-optimal performance in certain regions of the space. The histogram should be nearly symmetric and centered at $(1-\alpha)$. If properly implemented, conformal prediction is guaranteed to satisfy:

$$(1-\alpha) \leq P(y_{test} \in C(x_{test})) \leq (1-\alpha) + \frac{1}{n+1}$$

.

2.4.2 Adaptivity

The first slightly informal method to evaluate adaptivity is to plot the histograms of the set size. Two signaling alerts can arise: on one hand, the average test size might be too big indicating a potential problem with the conformal procedure (score or model), and on the other, the spread of the set sizes may be too narrow which could suggest that set size is not accurately reflecting the difficulty of the observation. Even if both signaling alerts are not present, we still need a more rigorous method to evaluate adaptiveness.

Evaluating adaptivity is usually formalized by asking for conditional coverage. As stated in theorem 1, under certain conditions marginal coverage -that is, the overall coverage among all groups- is guaranteed to be at least $(1-\alpha)$. Conditional coverage is a much stronger assumption: it enforces the coverage to be at least $(1-\alpha)$ for every possible observation. The property of conditional coverage can be defined as:

[6] Based on section 3 of the paper "A gentle introduction to conformal prediction and distribution-free uncertainty quantification".

[7] For the nuances of the size of the calibration set and its impact on coverage see section 3.2 of "A gentle introduction to conformal prediction and distribution-free uncertainty quantification".

[8] V. Vovk; "Conditional validity of inductive conformal predictors" in Proceedings of the Asian Conference on Machine Learning, vol. 25, 2012, pp. 475-490.

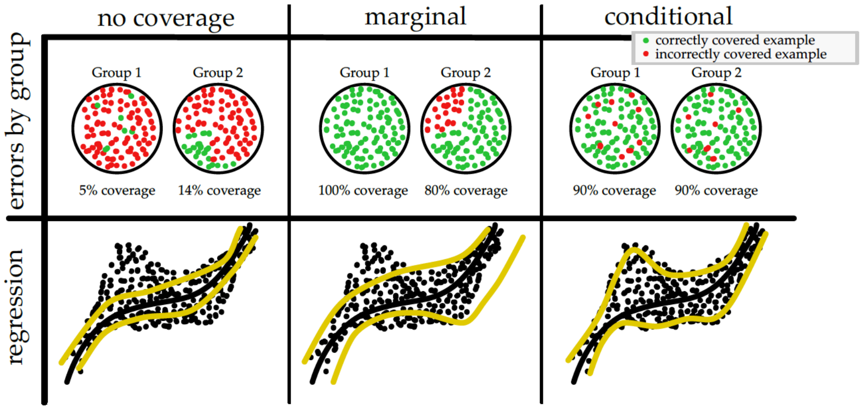

Figure 1: Difference between no coverage (left), marginal coverage (center) and conditional coverage (right). For classification, the red points correspond to misclassified observations, and green ones for correctly classified. Extension to regression is analogous.

$$P(Y_{test} \in C(X_{test})|X_{test}) \geq 1 - \alpha$$

Figure 1 presents a practical example to illustrate the conceptual difference between marginal and conditional coverage[9].

Let's focus on the second column and analyze marginal coverage. A coverage of 100% for group A and 80% for group B would yield a marginal or overall coverage of 90%. Nonetheless, a coverage of 90% can be achieved but misclassifications (or points outside the prediction interval for regression) can be concentrated in a specific region of the space. The third column portrays the idea of conditional coverage: the coverage is at least 90% for each of the groups. In essence, we want to achieve a balanced coverage among all possible prediction labels or regions in the space.

Conditional coverage is, in the general case, almost impossible to achieve: we hope to approximate the conformal procedure approximates to it as much as possible.

3 Non Exchangeability and Symmetric Algorithm

In this and the following section, we will follow the method proposed by Barber et al. (2023) to relax the assumption of exchangeability. This assumption was at the core of Theorem 1. The intuition of why exchangeability is so crucial in guaranteeing the $1 - \alpha$ coverage of the confidence set in Eq. 4 is readily seen in the case of split conformal inference. The basic idea of split conformal inference is to assess the uncertainty around the prediction of a trained model using a set of fresh data forming the calibration set. In essence, what this method does is to compute some transformation of the models error on this calibration set and to consider the quantiles on the calibration set as a valid quantile of the transformed error also on the test data. However, it is easy to see that when the distribution of the data changes, then the distribution of the residuals on the calibration and the test sets are no longer guaranteed to be comparable. Therefore, it is not sensible to try to evaluate the uncertainty of the model on the test set based on the thresholds obtained in the calibration set.

3.1 Robust Inference through weighted quantiles

Non-exchangeable data are very common in practice. For instance, spatial and time-series data are non-exchangeable, as dependence is a distinguishing feature of this kind of data. Also, in many cases data can present change-points where the distribution of the data suddenly change, or we can observe a slow variation in the distribution of the data, even if the observations are independent. In cases where exchangeability is

[9] The example was extracted from section 3.1 of the paper "A gentle introduction to conformal prediction and distribution-free uncertainty quantification".

violated, the distribution of an algorithm's residuals changes for different subset of the data, so that adjustments are needed to guarantee the theoretical bounds of conformal inference. A key idea proposed by Barber et al. (2023) is that of weighting differently the scores depending on their distance from the distribution of the target point. For example, in time series data we would prefer to weight more observations nearer in time to the target for building a confidence interval, and similarly we would prefer to weight more observations from contiguous units in case of spatial data. To formalize this idea, let $w_1, \dots, w_n \in [0,1]$ denote the weights associated to each observation. The weights are here considered as fixed: for instance, in a time series setting we could use an exponentially decreasing weighting $w_i = \rho^{n+1-i}, \rho \in (0,1)$, where $n+1$ is the target point. For the ease of notation, we can denote the observations with let's denote the observations with $Z_i = (X_i, Y_i)$ and consider the standardized residuals

$$\tilde{w}_i = \begin{cases} \dfrac{w_i}{w_1 + \cdots + w_n + 1} & \text{for } i = 1, \dots n \\ \\ \dfrac{1}{w_1 + \cdots + w_n + 1} & \text{for } i = n+1 \end{cases}$$

We can use these standardized weights to define robust confidence intervals in the non-exchangeable setting.

Split Conformal Inference In this case, consider the following confidence interval:

$$\hat{C}_n(X_{n+1}) = \hat{\mu}(X_{n+1}) \pm Q_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{w}_i \delta_{R_i} + \tilde{w}_{n+1}\delta_{+\infty}\right), \tag{6}$$

where $R_i = |Y_i - \hat{\mu}(X_i)$ and $\hat{\mu}$ is the trained model.

Full Conformal Inference In this case, consider this alternative definition of the confidence interval:

$$\hat{C}_n(X_{n+1}) = \left\{ y : R^y_{n+1} \leq Q_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{w}_i \delta_{R^y_i}\right)\right\} \tag{7}$$

where $R_i = |Y_i - \hat{\mu}(X_i)^y$ and $\hat{\mu}$ is the model trained on $Z_1, \dots, Z_n, (X_{n+1}, y)$.

This sets are indeed valid $1-\alpha$ confidence sets (as it will be shown in Theorem 2):

$$P\left(Y_{n+1} \in \hat{C}(X_{n+1})\right) \geq 1 - \alpha - \sum_{i=1}^{n} \tilde{w}_i d_{TV}(R(Z), R(Z^i)) \tag{8}$$

4 Non Exchangeability and Non Symmetry

In this section, we introduce the application of non-symmetric algorithms for conformal prediction. This extension increases the robustness of the model toward non-exchangeable data by assigning greater weights to observations that are more representative of the new data point that we try to predict. Specifically, the non-symmetric algorithm that we are going to implement in the empirical part is the weighted least square (WLS) method.

In general, to define a non symmetric algorithm $\mathcal{A}$, we need to specify a fixed sequence of weights that we call tags t_i. We then assign each t_i to the i^{th} observation, resulting in a new set of data point (y_i, x_i, t_i). With the inclusion of these tags, the algorithm is no more required to treat data points symmetrically as the presence of $t_1, .., t_n$ will result in different fits for different permutations.

4.1 Example

As the non-symmetric nature of an algorithm comes from the fact that model's parameters vary when fitted on permuted data points, then WLS is a clear example of this class of algorithms. Such an algorithm is an extension of least square regression and is implemented by assigning a weight to each observation, where the weight should reflect the impact of the corresponding observation when estimating $\hat{\beta}$. In practice, let's define a weight sequence t of length n and consider a diagonal matrix T of size n where the diagonal entries T_{ii} equals t_i. Then, $\hat{\beta}$ is obtained by minimizing the following expression

$$(y - X\beta)'T(y - X\beta)$$

which results in

$$\hat{\beta} = (X'TX)^{-1}X'Ty$$

If we select weights that decrease as we go backward in the data sequence, such as $t_i = p^{n+1-i}$ with $p \in (0, 1)$, then we downweight the covariance between distanced points. This means that the prediction of a new point will be more heavily influenced by the information from nearby points.

4.2 Split Conformal Inference

The model $\hat{\mu}$ is fitted exclusively on training data; it does not depend on the holdout set i.e $\hat{\mu}$ is trivially a symmetric function of the holdout set. Consequently, prediction interval (6) for the case of non exchangeability and symetric algorithm still holds.

4.3 Full Conformal Inference

When implementing full conformal inference in the setting where both assumptions are violated, we introduce a new idea called the swap step[10]. The swap step requires us to run our algorithm $\mathcal{A}$ after swapping data point $(X_k.Y_k)$ and $(X_{n+1}.Y_{n+1})$ but keeping the tags. More concretely, we run $\mathcal{A}$ on data points $(X_k.Y_k, t_{n+1})$ and $(X_{n+1}.Y_{n+1}, t_k)$ instead of $(X_k.Y_k, t_k)$ and $(X_{n+1}.Y_{n+1}, t_{n+1})$. As explicitly stated in the referenced paper, the swap step is only necessary for the theoretical guarantees to hold.

In this setting, the model $\hat{\mu}^{y,k}$ is fitted by applying $\mathcal{A}$ on the whole data -including test point- but with the k-th and $(n+1)$ data point swapped:

$$\hat{\mu}^{y,k} = \mathcal{A}((X_{\pi_k(i)}, Y^y_{\pi_k(i)}, t_i) \, : \, i \in [n+1])) \tag{9}$$

where π_k is a permutation of $[n+1]$ data points after swapping observation k and $n+1$.

In order to compute the residuals, we first define:

$$Y_i^y = \begin{cases} Y_i & i = 1, \dots, n \\ y & i = n+1 \end{cases}$$

Concomitantly, we define the residuals similarly than in the setting of full conformal baseline case where both assumptions hold, but we include a super index k due to the additional swapping step.

$$R_i^{y,k} = \begin{cases} \left| y_i - \mu^{\hat{y,k}}(x_i) \right| & i = 1, \dots, n \\ \\ \left| y_i - \mu^{\hat{y,k}}(x_{n+1}) \right| & i = n+1 \end{cases}$$

Finally, we are in conditions to compute the prediction interval:

$$\hat{C}_n(X_{n+1}) = \left\{ y \, : \, R_{n+1}^{y,k} \leq Q_{1-\alpha} \left(\sum_{i=1}^{n+1} \tilde{w}_i \delta_{R_i^{y,k}} \right) \right\} \tag{10}$$

The major difference with 4 is the inclusion of the swapping super index k in the residuals.

[10] Section 3.2 in paper "Conformal prediction beyond exchangeability".

4.4 Coverage Bounds

In this subsection, we will briefly touch upon the coverage bounds achieved by our algorithms in this non-exchangeable non-symmetric setting. We will state the theorems pertaining to the lower and upper bounds on coverage. Once again, if we evaluate these bounds for full conformal, these apply to split-conformal since it is a special case.

Here, we will discuss some of the challenges that one can encounter while following the setup of Theorem 1 to find a bound in this general setting. In alter sections, we will state theorems on lower and upper bounds of coverage without proving them. This however, has been described in detail in R.F. Barber et al. (2023).

Challenges Our earlier proof (Theorem 1) was quintessentially based on the equality

$$P\left\{R_{n+1} > Q_{1-\alpha}\left(\sum_{j=1}^{n+1}\frac{1}{n+1}\cdot\delta_{R_j}\right)\right\} = P\left\{R_i > Q_{1-\alpha}\left(\sum_{j=1}^{n+1}\frac{1}{n+1}\cdot\delta_{R_j}\right)\right\}$$

because of the exchangeability of the residuals R_i, a facet of the exchangeability assumption on the data and symmetric algorithm. Does this hold for the non-exchangeable full conformal case?

Even when the residuals are exchangeable, the weighted quantile $Q_{1-\alpha}(\sum_{j=1}^{n+1}\tilde{w}_j\cdot\delta_{R_j})$ is no longer a symmetric function of $R_1, ..., R_{n+1}$ if the weights are not all equal. Therefore, the equality above wouldn't hold in general.

Also, if we use non-symmetric algorithms that take tagged data points as input, situation becomes even more complicated, since our residuals depend on the fitted model $\hat{\mu}$ that treat data points non-symmetrically they may no longer be exchangeable.

4.4.1 Lower Bound

Theorem 2 (Nonexchangeable full conformal prediction). *Let $\mathcal{A}$ be an algorithm mapping a sequence of triplets (X_i, Y_i, t_i) to a fitted function $\hat{\mu}$.* Then the nonexchangeable full conformal method defined in 10 satisfies

$$P\left(Y_{n+1}\in\hat{C}(X_{n+1})\right) \geq 1-\alpha-\sum_{i=1}^{n}\tilde{w}_i d_{TV}(R(Z), R(Z^i)) \tag{11}$$

To prove this bound, we primarily rely on the swap step. Then defining "strange" points as we did in Theorem 1, we argue that noncoverage of Y_{n+1} implies strangeness of point k. This theorem is proved in Barber et al. (2023).

4.4.2 Upper Bound

Theorem 3 (Coverage upper bound in the non-exchangeable case). *For any algorithm $\mathcal{A}$, if $R_1^{Y_{n+1},K}, \ldots, R_n^{Y_{n+1},K}, R_{n+1}^{Y_{n+1},K}$ are almost surely distinct, then the nonexchangeable full conformal method in Eq. 10 satisfies*

$$P\left(Y_{n+1}\in\hat{C}(X_{n+1})\right) < 1-\alpha+\tilde{w}_{n+1}+\sum_{i=1}^{n}\tilde{w}_i d_{TV}(R(Z), R(Z^i))$$

Defining the coverage event in a similar way as Theorem 2, we can get the upper bound on our coverage for the non-exchangeable case.

Note: This applies in general to both symmetric and non-symmetric algorithm as by simply ignoring the tag, we recover the symmetric case.

5 Jackknife+

5.1 Jackknife review

Let us go back to our two assumptions setting and start with a stricter assumption of having i.i.d. training data $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, ..., n$ and a new test point (X_{n+1}, Y_{n+1}) drawn independently from the same distribution. Our goal is to fit a regression model to the training data, a function $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}$ where $\hat{\mu}(x)$

predicts Y_{n+1} given a new feature vector $X_{n+1} = x$, and then use the framework of conformal inference to provide a prediction interval for the test point that should contain true test response value.

As before, we can formalise this statement by constructing a prediction interval $\hat{C}_{n,\alpha}$ as a function of the n training data points and target coverage level $1 - \alpha$, such that

$$P\left\{Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1})\right\} \geq 1 - \alpha \tag{12}$$

where probability is taken over the joint training data and test point.

Overfitting problem: If we naively use residuals on the training data and consider the $1 - \alpha$ quantile of the residuals for all n points, we run into overfitting problem where residuals on our training data points are typically smaller than residual on our previously unseen test point.

Leave One Out to the rescue: Computing margin of error with a leave-one-out construction we can get rid of this problem (LOO-CV deja vu!):

- For each $i = 1, ..., n$ fit the regression function $\hat{\mu}_{-i}$ removing i^{th} point and compute the corresponding LOO residual, R_i given by $|Y_i - \hat{\mu}_{-i}(X_i)|$.

- Fit the regression function $\hat{\mu}$ to the full training data and get the output prediction interval as

$$\hat{\mu}(X_{n+1}) \pm \big((1-\alpha) \text{ quantile of } |Y_1 - \hat{\mu}_{-1}(X_1)|, ..., |Y_n - \hat{\mu}_{-n}(X_n)|\big)\,.$$

$$\hat{C}^{jk}_{n,\alpha}(X_{n+1}) = \left[\hat{q}^{-}_{n,\alpha}\{\hat{\mu}(X_{n+1}) - R_i^{LOO}\}, \hat{q}^{+}_{n,\alpha}\{\hat{\mu}(X_{n+1}) + R_i^{LOO}\}\right] \tag{13}$$

where $\hat{q}^{+}_{n,\alpha} = \lceil (1-\alpha)(n+1) \rceil$ smallest value of its argument. However, for high dimensional setting where $\hat{\mu}$ is unstable, jackknife method looses predictive coverage which has been the benchmark of our previous algorithms. To rectify this, we will look at a new method jackknife+ that provides such coverage guarantees, first keeping the exchangeability condition intact and then in a more general setting (R. F. Barber et al, 2021).

5.2 Jackknife+

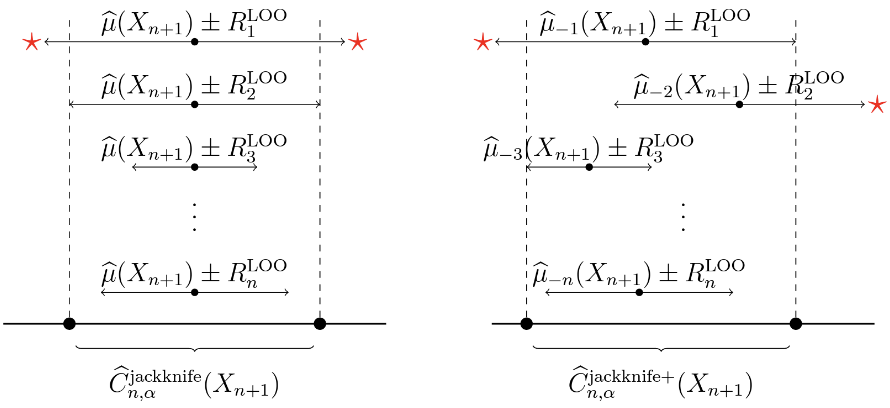
$\widehat{\mu}(X_{n+1}) \pm R_1^{\text{LOO}}$
$\widehat{\mu}(X_{n+1}) \pm R_2^{\text{LOO}}$
$\widehat{\mu}(X_{n+1}) \pm R_3^{\text{LOO}}$
$\widehat{\mu}(X_{n+1}) \pm R_n^{\text{LOO}}$
$\widehat{C}^{\text{jackknife}}_{n,\alpha}(X_{n+1})$
$\widehat{\mu}_{-1}(X_{n+1}) \pm R_1^{\text{LOO}}$
$\widehat{\mu}_{-2}(X_{n+1}) \pm R_2^{\text{LOO}}$
$\widehat{\mu}_{-3}(X_{n+1}) \pm R_3^{\text{LOO}}$
$\widehat{\mu}_{-n}(X_{n+1}) \pm R_n^{\text{LOO}}$
$\widehat{C}^{\text{jackknife+}}_{n,\alpha}(X_{n+1})$

Figure 2: Jackknife vs Jackknife+. Prediction intervals are so chosen that only a sufficiently small proportion of extremes of these arrows exceed the boundary (marked by star). The difference is where we center our interval, for jackknife we center on the predicted value $\hat{\mu}$ whereas for jackknife+ we use the LOO predictions $\hat{\mu}_{-i}$

Defining the LOO predictions as we did in the previous section, we use these predictions, given by $\hat{\mu}_{-i}$, for the test point to center our interval.

Jackknife+ interval prediction in our original notation is given by

$$\hat{C}^{jk+}_{n,\alpha} = \left[Q_\alpha \left(\sum_{i=1}^{n} \frac{1}{n+1} \cdot \delta_{\hat{\mu}_{-i}(X_{n+1}) - R_i^{LOO}} + \frac{1}{n+1} \cdot \delta_{-\infty} \right), Q_{1-\alpha} \left(\sum_{i=1}^{n} \frac{1}{n+1} \cdot \delta_{\hat{\mu}_{-i}(X_{n+1}) - R_i^{LOO}} + \frac{1}{n+1} \cdot \delta_{+\infty} \right) \right] \tag{14}$$

While the jackknife interval $\hat{C}^{jk}_{n,\alpha}(X_{n+1})$ is defined as a symmetric interval around predicted value fitted on the full training data $\hat{\mu}(X_{n+1})$, the jacknnife+ interval can be interpreted as an interval around the median prediction, Median $\left(\hat{\mu}_{-1}(X_{n+1}), ..., \hat{\mu}_{-n}(X_{n+1})\right)$ which is guaranteed to lie inside $\hat{C}^{jk+}_{n,\alpha}(X_{n+1})$ for $\alpha \leq \frac{1}{2}$.

Theorem 5.1 The jackknife+ prediction interval satisfies

$$P\left\{ Y_{n+1} \in \hat{C}^{jk+}_{n,\alpha}(X_{n+1}) \right\} \geq 1 - 2\alpha \tag{15}$$

Proof of this theorem with exchangeability condition intact is presented in the reference mentioned above. There, a notion of "strange" data points with unusually large residuals is defined to argue that the probability of $Y_{n+1} \notin$ interval prediction is $\leq$ probability of point indexed $n+1$ being strange which is bounded by 2α. Then, since probability over the whole outcome space is 1, we get this result.

5.3 Non-exchangeable Jackknife+ with a symmetric algorithm

We use the framework laid out in section 3 to present non-exchangeable jackknife+ method for the setting where algorithm $\mathcal{A}$ is symmetric.

Begin by choosing weights fixed ahead of time, this gives rise to normalized weights. Then, prediction interval is given (analogously) by

$$\left[Q_\alpha \left(\sum_{i=1}^{n} \tilde{w_i} \cdot \delta_{\hat{\mu}_{-i}(X_{n+1}) - R_i^{LOO}} + \tilde{w_{n+1}} \cdot \delta_{-\infty} \right), Q_{1-\alpha} \left(\sum_{i=1}^{n} \tilde{w_i} \cdot \delta_{\hat{\mu}_{-i}(X_{n+1}) - R_i^{LOO}} + \tilde{w_{n+1}} \cdot \delta_{+\infty} \right) \right] \tag{16}$$

Unweighted version of Jackknife+ is recovered by choosing weights $w_1 = ... = w_n = 1$ here.

5.4 Non-exchangeable Jackknife+ with nonsymmetric algorithm

Once again the framework laid out in section 4 works almost exactly, however we have another important index to worry about. This is the index i corresponding to removed ith point. Since we are going to tag every point this point is assigned a tag but we explicitly form a case where identity permutation is used in both cases, $k = i$ or $k = n+1$. Therefore, we have

$$\hat{\mu}^k_{-i} = \begin{cases} \mathcal{A}((X_j, Y_j, t_j) : j \in [n] \setminus \{i, k\}, (X_k, Y_k, t_{n+1})), & \text{if } k \in [n] \text{ and } k \neq i, \\ \mathcal{A}((X_j, Y_j, t_j) : j \in [n] \setminus \{i\}), & \text{if } k = n+1 \text{ or } k = i. \end{cases}$$

We define the corresponding LOO residuals as

$$R_i^{k,LOO} = |Y_i - \hat{\mu}^k_{-i}(X_i)| \tag{17}$$

To run the algorithm, we first draw a random index K then compute the non-exchangeable jackknife+ prediction interval as

$$\left[Q_\alpha \left(\sum_{i=1}^{n} \tilde{w_i} \cdot \delta_{\hat{\mu}^K_{-i}(X_{n+1}) - R_i^{K,LOO}} + \tilde{w_{n+1}} \cdot \delta_{-\infty} \right), Q_{1-\alpha} \left(\sum_{i=1}^{n} \tilde{w_i} \cdot \delta_{\hat{\mu}^K_{-i}(X_{n+1}) - R_i^{K,LOO}} + \tilde{w_{n+1}} \cdot \delta_{+\infty} \right) \right] \tag{18}$$

By simply ignoring the tag, we recover the symmetric case mentioned above.

6 Example using simulated data

We now present a simple example that clarifies the advantages of the method proposed by Barber et al. (2023). Consider 800 data points $((X_1, Y_1), \dots, (X_{800}, Y_{800}))$ generated from the following univariate linear model with distribution shift:

$$X_i \sim \mathcal{N}(0,4), \quad u_i \sim \mathcal{N}(0,1) \qquad Y_i = X_i\beta_i + u_i$$
$$\beta_1, \dots, \beta_{800} \quad \text{defined over a regular grid from -2 to 2.}$$

Notice that these data are independent but come from different distributions. Indeed, even if any X_i is drawn independently from the same distribution, then the mapping to Y_i occurs through varying parameters β_i. This means that the order in which the data are considered actually matters, as two data points extracted consequently come from similar distributions while two points generated with the initial and final β's do not. Therefore, the data are not exchangeable. From the previous sections, and in particular from Eq. 8, we know that non-exchangeability may compromise the coverage guarantees of the conformal prediction, and that a suitable weighting of the observations is needed to achieve the right coverage. The weighting should be designed so that the points having very distant distributions are almost not accounted in computing the confidence bounds. For this example we compare three procedures for assessing the uncertainty around a prediction based on least-squares estimation:

1. The standard unweighted method, showing how non-exchangeability can affect the coverage properties of the conformal bounds if not explicitly considered.

2. The robust method, with exponentially decreasing weights $w_i = 0.99^{n+1-i}$. This will be helpful in adapting the conformal bounds to take into account the drift in the distribution of the data.

3. The robust method using weighted least squares (WLS) as the algorithm. The coverage properties of this estimator are guaranteed by the discussion about asymmetric algorithms in Section 4. We consider as weights for the the WLS algorithm $t_i = 0.9^{n+1-i}$. In principle, the WLS estimator should give better point estimates, by weighting more subsequently-drawn observations that have similar distributions. These preciser estimates potentially make the distribution of the residuals of the fitted model more similar, thus increasing the coverage capacity of the conformal intervals.

Each row in Figure 3 reports the estimated conformal intervals computed using the procedures in the points above. Each method is repeated using full, split and jackknife+ procedures. The algorithm is estimated with a burn-in of the first 300 observations (reported in gray). The aim is to estimate a confidence interval covering 95% of the target points (reported in blue). Each target point is predicted using the algorithm; then a confidence bound is put around its prediction; and finally the point is added to the training set for predicting the next target value, in an expanding-training approach.

The plots in the first row show that the standard unweighted method is unable to achieve the desired coverage. Indeed, in this case the least-squares prediction gives very poor performances, as it get stuck to the distribution of the burn-in points without adapting to the drifting distribution of the target points. In the second row, we can see that the robust method is indeed able of improving the coverage, but the confidence set is still not valid, with a coverage of about 80%, against the desired 95%. Moreover, the way the better coverage is achieved is simply by widening the confidence intervals around a substantially wrong prediction. This provides a very interesting insight about conformal prediction: even if theoretically we can find a set of weights that makes the interval reach the desired coverage, nothing assures that the estimated intervals are useful at all. Indeed, if the algorithm that we use gives bad point predictions, the coverage will be reached with extremely large confidence intervals that may be of little practical relevance. The latter point is further confirmed by the plots in the third row, where we consider a different, asymmetric algorithm. The WLS algorithm better retrieves the underlying drift in the data distribution, giving better point-predictions for the target points. This results in better confidence bounds, that are narrower and achieve higher coverage.

In all the three rows, the different variants of conformal inference achieve similar results. As expected, jackknife+ achieves almost identical results than full conformal inference. Split conformal has good performances, but slightly larger confidence intervals, as expected in theory. However, it is worth noticing that in more realistic scenarios than this simple 1-dimensional example, the three algorithms may diverge more (and the computational cost of full conformal inference may become unsustainable).

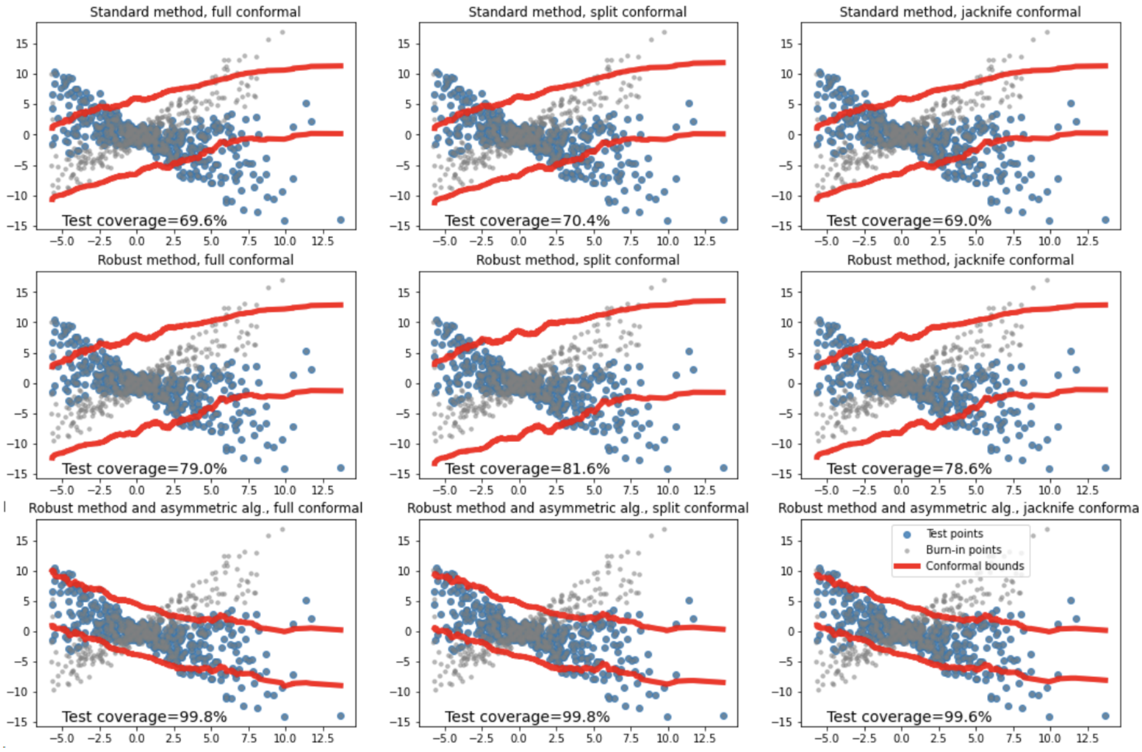

Figure 3: Example of standard conformal inference VS. nonexchangeability-robust conformal inference.

7 Electricity Transformer Data Set

Our empirical study focuses on the ETT dataset, which contains data from an Electricity Transformer in China [11]. This dataset includes 6 types of load power measures and oil temperature, which is our target variable. The data was collected over two years, with one data point recorded every minute. In total, the dataset consists of 70,080 data points but we only consider the last 2000 observations due to computational complexity.

We use this data to compare the coverage performance of the jackknife and full conformal models. We implement three versions of each model: exchangeable data with least square, non-exchangeable data with least square, and non-exchangeable data with weighted least square.

Figure 4 shows the coverage of each model with a rolling average with a window of size 200. We can see that the exchangeable version of both models performs pretty bad, which is what we expected since we are dealing with non-exchangeable data. Indeed, the coverage is not adaptive to changes in distribution and often falls to 0.8. Dealing with the non-exchangeable models, the performances clearly improve and become adaptive as we considered weighted quantiles with exponentially decaying weights. Finally, the implementation of weighted quantiles together with the weighted least square is clearly the best model as it is mostly stable around 0.9 coverage, showing robustness against distribution dirfts.

[11] The data retrieved from https://github.com/zhouhaoyi/ETDataset

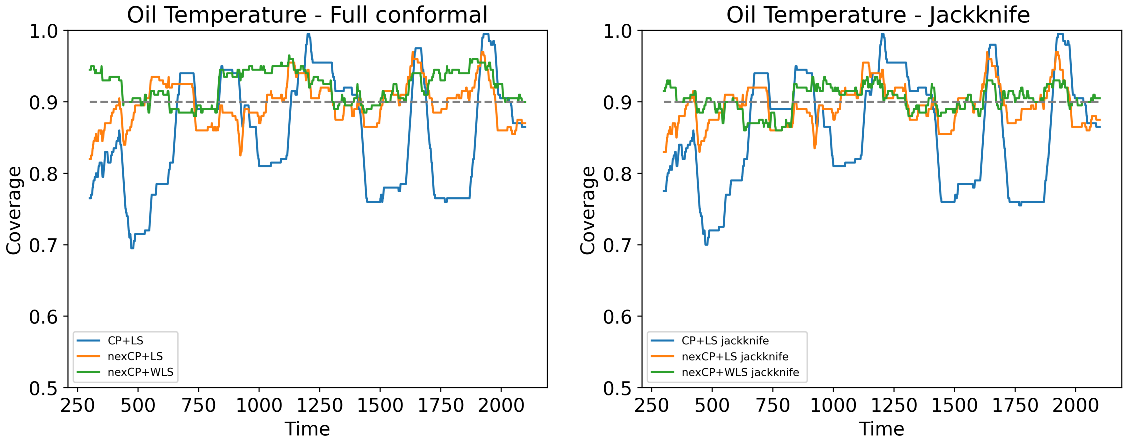

Figure 4: Coverage for Chinese transformer data. The lines are smoothed by taking a rolling average with a window of size 200

Moreover, Table 1 shows the average coverage for each model. The results are consistent with our expectations as the average coverage increases for both methods when implementing weighted quantiles and it further increases in the case of the asymmetric algorithm.

Model	Mean Coverage
CP+LS	0.8500
nexCP+LS	0.8940
nexCP+WLS	0.9210
CP+LS+jackknife	0.8515
nexCP+LS+jackknife	0.8915
nexCP+WLS+jackknife	0.9035

Table 1: Average coverage

Finally, in Figure 5 we compare the optimal models for both full conformal and jackknife approaches. As we can see, the full conformal coverage is almost always above the jackknife. This could be due to an overfitting effect of the full conformal approach which results in wider prediction intervals. Therefore, since our objective was to find a 90% coverage set, the jackknife approach would be the preferable choice as it prevents the overfitting effect of the full conformal and is also less computationally expensive.

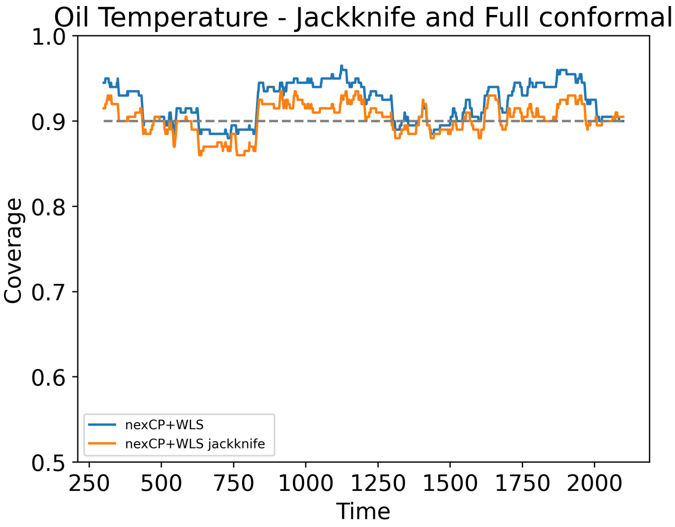

Figure 5: Comparison between full conformal and jackknife for the coverage on Chinese transformer data. The lines are smoothed by taking a rolling average with a window of size 200

References

R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani; "Conformal prediction beyond exchangeability"; arXiv:2202.13415v4, 2023.

Lecture slides from co-author of preceding paper Professor R. F. Barber, Statistician, University of Chicago.

Anastasios N. Angelopoulos and Stephen Bates; "A gentle introduction to conformal prediction and distribution-free uncertainty quantification"; arXiv preprint arXiv:2107.07511, 2021.

R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani; “Predictive inference with the jackknife+”; The Annals of Statistics, vol. 49, no. 1, pp. 486–507, 2021.

V. Vovk; "Conditional validity of inductive conformal predictors," in Proceedings of the Asian Conference on Machine Learning, vol. 25, 2012, pp. 475-490.

V. Vovk, A. Gammerman, G. Shafer; "Algorithmic Learning in a Random World", Springer Second Edition, 2022.