Transductive conformal inference with adaptive scores

Myriam Frikha

December 2023

Contents

1 Introduction

This article is a brief analysis of *Transductive conformal inference with adaptive scores* written by Ulysse Gazin, Gilles Blanchard and Etienne Roquain. [1]. Conformal inference is a general framework used to make reliable and consistent predictions with a quantifiable level of certainty. At its core, it's about creating prediction intervals or sets that are likely to contain the true outcome with a specified probability.

We consider a framework where predictions have to be made on a set of m different points, which leads to m conformal p-values. In the first section, we focus on the joint distribution of conformal p-values which follows a Polya urn model as well as conformal prediction intervals. Then, in the second section, we establish a concentration inequality for their empirical distribution function which serves as an introduction to novelty detection. Finally, a simulation on a real dataset is done in order to back up some of the theoretical results.

2 Motivations with data set

We consider the airfoil dataset from the UCI Machine Learning Repository (Dua and Graff, 2019) [2]. It has $N = 1503$ observations of a response Y representing the scaled sound pressure level of Nasa airfoils and a covariate X with 5 different dimensions : log frequency, angle of attack, chord length, free-stream velocity and suction side log displacement thickness. Our aim is to create conformal prediction intervals on this data set. Before tackling this complex task, we'll first introduce the concept of transductive conformal prediction in the next section.

3 Prediction intervals

3.1 Setting and method

We consider i.i.d. regression data $Z_1, ..., Z_p \sim P$ where each $Z_i = (X_i, Y_i)$ is a random variable in $\mathbb{R}^d \times \mathbb{R}$ composed of a d-dimensional vector of features X_i and a response variable Y_i. In addition, we distinguish three main samples for this framework :

- Calibration sample $\mathcal{D}_{cal} = \{(X_i, Y_i), i \in [\![n]\!]\}$ where Z_i is observed and used to calibrate the sizes of the prediction intervals.
- Training sample $\mathcal{D}_{train}$, where Z_i is observed and specifically used to build predictors.
- Test sample $\mathcal{D}_{test} = \{(X_{n+i}, Y_{n+i}), i \in [\![n]\!]\}$ where only the X_i's are observed.

Transductive conformal aims at building m prediction intervals for $Y_{n+1}, .., Y_{n+m}$ given $X_{n+1}, ..., X_{n+m}$ from $\mathcal{D}_{test}$. More precisely, we want to build a family of m random intervals of $\mathbb{R}$, $\mathcal{I} = (\mathcal{I}_i)_{i\in[\![m]\!]}$, such that the amount of coverage errors $(\mathbb{1}\{Y_{n+i} \notin \mathcal{I}_i\})_{i\in[\![m]\!]}$ is under control. In addition, we consider a configuration that is specific to *transfer learning* : The distribution of $\mathcal{D}_{train}$ might differ from $\mathcal{D}_{cal}$ and $\mathcal{D}_{test}$ where the data points are i.i.d within each sample. A good prediction is still guaranteed in this case.

The construction of $\mathcal{I}$ is done trough *non conformity scores* $(S_i)_{1\le i\le n+m}$:

$$S_i := |Y_i - \hat{\mu}(X_i, (\mathcal{D}_{train}, \mathcal{D}^X_{cal+test})|, \ i \in [\![n+m]\!], \tag{1}$$

where $\hat{\mu}$ is a predictor of Y_i that uses $\mathcal{D}_{train}$ as well as the features $(X_1, ..., X_{n+m})$ of $\mathcal{D}^X_{cal+test}$. Usually, classical scores only focus on $\mathcal{D}_{train}$. The particularity of these ones is that they can take into account the unlabeled data $\mathcal{D}^X_{cal+test}$ which makes them suitable for *transfer learning.* This led to the appellation *adaptive scores.*

3.2 Theoretical results

3.2.1 p-values

One of the key quantities of this topic is split conformal p-values. It's defined as follows :

$$p_i = (n+1)^{-1}\left(1 + \sum_{j=1}^{n} \mathbb{1}(S_j \ge S_{n+i})\right), \ i \in [\![m]\!] = \{1, ..., m\}. \tag{2}$$

Classical results usually only concern the marginal distribution of the p-values. In this section, we try to do an analysis on the joint distribution.

Let $(S_i)_{i\in[\![n+m]\!]}$ be real random variables corresponding to non-conformity scores, for which $(S_j)_{j\in[\![n]\!]}$ are the "reference" scores and $(S_{n+i})_{i\in[\![m]\!]}$ are the "test" scores. We start with two initial assumptions regarding non-conformity scores :

$$\text{The variables } S_i, \ i \in [\![n+m]\!] \text{ are i.i.d} \tag{H1}$$

$$\text{The score vector } (S_i)_{i\in[\![n+m]\!]} \text{ has no ties a.s.} \tag{H2}$$

Let P^U be a discrete distribution on the set $\left\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\right\}$, for any vector $U = (U_1, ..., U_n) \in [0,1]^n$, defined as

$$P^U(\{\ell/(n+1)\}) = U_\ell - U_{\ell-1}, \quad \ell \in n+1.$$

where $0 = U_{(0)} \le U_{(1)} \le \cdots \le U_{(n)} \le U_{(n+1)} = 1$ are the increasingly ordered values of $U = (U_1, \ldots, U_n)$. The probability distribution represents the likelihood of a uniformly distributed variable over the interval [0, 1] landing in any of the sections $\left\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\right\}$.

In addition, P^U has for c.d.f.

$$F^U(x) = U(\lfloor (n+1)x \rfloor), \quad x \in [0,1].$$

Proposition 1. *Assume (H1) and (H2) and consider the p-values $(p_i, i \in m)$ given by (2). Then conditionally on $D_{cal} = (S_1, \ldots, S_n)$ the p-values are i.i.d. of common distribution given by*

$$p_1 | D_{cal} \sim P^U$$

where $U = (U_1, \ldots, U_n) = (1 - F(S_1), \ldots, 1 - F(S_n))$ are pseudo-scores and F is the common c.d.f. of the scores of D_{cal}, that is $F(s) = P(S_1 \leq s)$, $s \in \mathbb{R}$. In addition, the pseudo-score vector U is i.i.d. Unif$[0, 1]$ distributed.

Proof. We first draw these observations concerning F and p_i:

- Assumption (H2) implies that F is continuous and $1 - F(S_i)$ has Unif$[0, 1]$ distribution. Therefore, $(U_1, \ldots, U_{n+m}) = (1 - F(S_1), \ldots, 1 - F(S_{n+m}))$ are i.i.d. $\sim$ Unif$[0, 1]$.
- Since $p_i = g(S_{n+i}, D_{\text{cal}})$, it follows that conditionally on D_{cal}, the variables $p_1, \ldots, p_m$ are independent and identically distributed.
- Since F is continuous, it holds that $F^{-1}(F(S_i)) = S_i$ almost surely, where F^{-1} is the generalized inverse of F.

Therefore, $\mathbb{1}_{\{S_j \geq S_{n+i}\}} = \mathbb{1}_{\{U_j \leq U_{n+i}\}}$ almost surely and p_1 is distributed as

$$(n+1)^{-1} \left(1 + \sum_{j=1}^{n} \mathbb{1}_{\{U_j \leq U_{n+1}\}} \right) = (n+1)^{-1} \left(1 + \sum_{j=1}^{n} \mathbb{1}_{\{U_{(j)} \leq U_{n+1}\}} \right),$$

Hence, we have for all $x \in [0, 1]$,

$$\mathbb{P}(p_1 \leq x | D_{\text{cal}}) = \mathbb{P} \left(1 + \sum_{j=1}^{n} \mathbb{1}_{\{U_{(j)} \leq U_{n+1}\}} \leq x(n+1) | D_{\text{cal}} \right)$$

We recognize that the sum $1 + \sum_{j=1}^{n} \mathbb{1}_{\{U_{(j)} \leq U_{n+1}\}}$ effectively counts the number of $(U_{(j)})$ that are less than or equal to (U_{n+1}). This count is less than or equal to $x(n+1)$ precisely when U_{n+1} is less than the $\lfloor x(n+1) \rfloor$-th order statistic $U_{(\lfloor x(n+1) \rfloor)}$:

$$\mathbb{P} \left(1 + \sum_{j=1}^{n} \mathbb{1}_{\{U_{(j)} \leq U_{n+1}\}} \leq x(n+1) | D_{\text{cal}} \right) = \mathbb{P} \left(U_{n+1} < U_{(\lfloor x(n+1) \rfloor)} | D_{\text{cal}} \right)$$

Given the uniform distribution of U, we finally have :

$$\mathbb{P} \left(U_{n+1} < U_{(\lfloor x(n+1) \rfloor)} | D_{\text{cal}} \right) = U_{(\lfloor x(n+1) \rfloor)} = F^U(x)$$

We recongnize the c.d.f of P^U. □

This proposition is very effective because it allows to control the failure probability $\mathbb{P}(p_1 \leq \alpha | \mathcal{D}_{cal})$ around its expectation. However, it has limitations since it can't handle *adaptive score training* : The score functions only depend on $\mathcal{D}_{cal}$

In order to solve this problem, we consider a weaker assumption instead of (H1) :

$$\text{The score vector } (S_i)_{i \in [\![n+m]\!]} \text{ is exchangeable.} \tag{H3}$$

Proposition 2. *Assume (H3),(H2), then the family of p-values $(pi, i \in [\![m]\!])$ given by (2) has joint distribution $P_{n,m}$ which is independent of the specific score distribution, defined by :*

$$P_{n,m} = \mathcal{D}(q_i, i \in [\![m]\!]), where$$
$$\begin{cases} (q_1, ..., q_m | U) \overset{i.i.d.}{\sim} P^U \\ and\ U = (U_1, ..., U_n) \overset{i.i.d.}{\sim} Unif([0, 1]) \end{cases}$$

Proof. Start by considering the assumption (H2) which, by its definition, occurs with probability 1. This ensures that the scores are distinct, leading to well-defined ranks R_i for the ordered scores. Consequently, the vector $(p_1, \ldots, p_m)$ is determined solely by the rank vector $(R_1, \ldots, R_{n+m})$. Specifically, the ranking condition can be expressed as:

$$R_i \leq R_j \iff S_i \leq S_j$$

and the conformal p-values from (1) can be rewritten as:

$$p_i = (n+1)^{-1} \left(1 + \sum_{j=1}^{n} \mathbb{1}_{\{R_j \geq R_{n+i}\}} \right), \quad \text{for } i \in [\![m]\!].$$

The next key aspect is the assumption (H3), which implies that the vector $(R_1, \ldots, R_{n+m})$ is uniformly distributed over the permutations of $[\![n+m]\!]$. This uniform distribution over permutations, combined with (H2), results in the same rank distribution for any score distribution that satisfies these assumptions. Therefore, the joint distribution of the p-values is the same for any such score distribution. This joint distribution can be obtained thank to *proposition 1*, by an integration over U.

□

3.2.2 Transductive error rates

Given the scores (1) we define the specific conformal procedure $\mathcal{C}(\alpha) = (\mathcal{C}_i(\alpha))_{i \in m}$ with respect to $\{p_i > \alpha\} = \{Y_{n+i} \in \mathcal{C}_i(\alpha)\}$ almost surely :

$$\{p_i \leq \alpha\} = \{\sum_{j=1}^{n} \mathbb{1}\,(S_j < S_{n+i}) \geq (n+1)(1-\alpha)\} = \{S\lceil(n+1)(1-\alpha)\rceil) < S_{n+i}\}$$

Therefore, we obtain an explicit form :

$$\mathcal{C}_i(\alpha) := [\hat{\mu}(X_{n+i}; (D_{\text{train}}, D^X_{\text{cal+test}})) \pm S(\lceil(n+1)(1-\alpha)\rceil)] \tag{3}$$

where $S(1) \leq \ldots \leq S(n) \leq S(n+1) := +\infty$ denote the order statistics of the calibration scores $(S_1, \ldots, S_n)$. In order to take into account the prediction multiplicity, we consider the *false coverage proportion* FCP of procedure $\mathcal{I} = (\mathcal{I}_i)_{i \in m}$ given by

$$\text{FCP}(\mathcal{I}) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{Y_{n+i} \notin \mathcal{I}_i\}. \tag{4}$$

We can clearly see that

$$\text{FCR}(\mathcal{C}(\alpha)) := \mathbb{E}[\text{FCP}(\mathcal{C}(\alpha))] \leq \alpha. \tag{5}$$

However, controlling the FCP directly is not a good idea since it fluctuates around its mean. Thus, $\{\text{FCP}(C(\alpha)) \leq \alpha\}$ is not always assured. Instead, we aim at controlling it in probability :

$$\mathbb{P}[\text{FCP}(\mathcal{C}(\alpha)) \leq \overline{\alpha}] \geq 1 - \delta. \tag{6}$$

We can tackle this control with several methods. One of those is to find a suitable family of random variables $\overline{\text{FCP}}_{\alpha,\delta}$ such that :

$$\mathbb{P}[\forall \alpha \in (0,1), \text{FCP}(\mathcal{C}(\alpha)) \leq \overline{\text{FCP}}_{\alpha,\delta}] \geq 1 - \delta. \tag{7}$$

Let $\hat{F}_m$ be the empirical distribution function of the p-value family :

$$\hat{F}_m(t) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{p_i \leq t\}, \; t \in [0,1]. \tag{8}$$

From (3), (4), (8) we observe that $\text{FCP}(C(t)) = \hat{F}_m(t)$, and thus for all $t \in [0,1]$:

$$\begin{aligned}\{\ \mathrm{FCP}(\mathcal{C}(t)) \leq \overline{\alpha}\} &= \{\hat{F}_m(t) \leq \bar{\alpha}\} \\ &= \{m\hat{F}_m(t) \leq \lfloor \bar{\alpha} m \rfloor\} \\ &= \{p_{(\lfloor \bar{\alpha} m \rfloor + 1)} > t\}\end{aligned}$$

where $p(1) \leq \cdots \leq p(m)$ denote the ordered conformal p-values. We deduce the following result.

Corollary 1. *Let $n, m \geq 1$. Consider the setting above, the conformal procedure $C(\alpha)$ given by (3) and $P_{n,m}$ given by Proposition 2. Then the following holds:*

(i) for any $\overline{\alpha} \in [0,1]$, $\delta \in (0,1)$, $C(\alpha = t_{\overline{\alpha}\delta})$ satisfies 6 provided that $t_{\alpha\delta}$ is chosen s.t.

$$\mathbb{P}_{p \sim P_{nm}}(p_{(\lfloor \overline{\alpha} m \rfloor + 1)} \leq t_{\overline{\alpha},\delta}) \leq \delta.$$

(ii) for any $\delta \in (0,1)$, $(\overline{FCP}_{\alpha\delta})_{\alpha \in (0,1)}$, satisfies 7 provided that

$$\mathbb{P}_{p \sim P_{nm}}(\exists \alpha \in (0,1) : \hat{F}_m(\alpha) > \overline{FCP}_{\alpha,\delta}) \leq \delta.$$

Thanks to this corollary and *Theorem 1* (discussed further in section 2), the following family satisfies 7:

$$\overline{\mathrm{FCP}}^{DKW}_{\alpha,\delta} = (\alpha + \lambda^{DKW}_{\delta,n,m})\mathbb{1}_{\{\alpha \geq \frac{1}{n+1}\}} \tag{9}$$

Where $\lambda^{DKW}_{\delta,n,m}$ is the parameter described in *Theorem 1* as well.

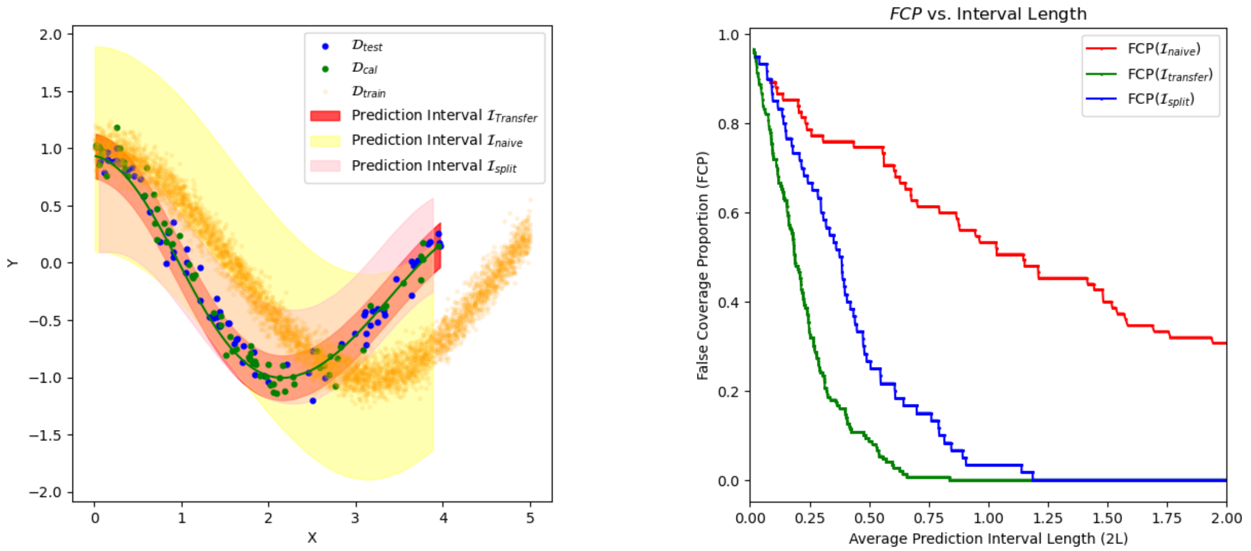
3.3 Simulations

As a first illustration, we consider the same framework described in [1] : We consider a regression model (W_i, Y_i) i.i.d. with $Y_i|W_i \sim \mathcal{N}(\mu(W_i), \sigma^2)$ where μ and σ are unknown. In order to assure a domain shift between $\mathcal{D}_{train}$ and the other samples, we assume that we observe $X_i = f_1(W_i)$ in $\mathcal{D}_{train}$ and $X_i = f_2(W_i)$ in $\mathcal{D}_{cal} \cup \mathcal{D}_{test}$ where f_1, f_2 are two functions. For the simulation, we set $|\mathcal{D}_{train}| = 5000$, $n = m = 75$, $\mu(x) = cos(x)$, $W_i \sim \mathcal{U}(0,5)$, $f_1(x) = x$, $f_2(x) = 0.6x + x^2/25$ and $\sigma = 0.1$. The implementation code is based on Boyer and Zaffran (2023) [5].

We consider three conformal procedures $\mathcal{I} = \mathcal{C}(\alpha)$ where the construction of the scores is not the same.

- $\mathcal{I}_{naive}$ uses a predictor $\hat{\mu}(., \mathcal{D}_{train})$ ans thus completely ignores domain shift.
- $\mathcal{I}_{split}$ splits $\mathcal{D}_{cal}$ into two samples of the same size and applies usual conformal approach.
- $\mathcal{I}_{transfer}$ which is backed up by previous theoretical results and uses $\hat{\mu}(., \mathcal{D}_{train}, \mathcal{D}^X_{cal+test})$.

For the first two procedures, an RBF kernel ridge regression is applied. For the last procedure, a transfer method is applied based on Courty et al. (2017) [6]. The prediction uses a *j.dot* transfer predictor from a specific library[7]. Let's note that this predictor satisfies H2. In addition, H3 is satisfied in this framework.



(a) Conformal prediction intervals with domain shift

(b) Plot of FCP($\mathcal{I}$) in function of interval length $2L$

Figure 1: Comparaison of three conformal procedures

Figure 1(a) shows that all methods provide a correct marginal coverage. However, $\mathcal{I}_{transfer}$ is much more precise which proves the relevance of transfer learning. Figure 1(b) shows that each bound is uniformly valid in L. $\mathcal{I}_{transfer}$ seems once again to be the best choice because it reduces effectively the FCP. A similar plot can be done for 9 (see notebook for more details).

4 Novelty detection

4.1 Setting and method

We observe a sample of nominal data points in $\mathbb{R}^d$; $\mathcal{D}_{null} = (X_1, ..., X_n)$ drawn i.i.d. from a distribution P_0 and a test sample $\mathcal{D}_{test} = (X_{n+1}, ..., X_{n+m})$ of independent points. The aim of novelty detection is to decide if each of X_{n+i} is distributed as the training sample or not, in which case it is considered as a *Novelty*.

One approach is to consider the proportion of mistakes within the items identified as novelties. This idea was originally presented by *Bates et al.* [4] introducing the *False discovery proportion* :

$$FDP(R) = \frac{|R \cap \mathcal{H}_0|}{|R| \vee 1}$$

where $\mathcal{H}_0 = \{i \in [m] : X_i \sim P_0\}$ the set of non-novelty in the test sample and $R \subset [\![m]\!]$ the subset of the X_i's declared as novelties. One particularity to note is that this quantity still holds in large scale settings because the number of errors $|R \cap \mathcal{H}_0|$ is rescaled by $|R|$.

The construction of the *adaptive scores* differs from the first section. We introduce the new strategy as follows :

- First, we choose $n \in (1, n_0)$. Then, we split $\mathcal{D}_{null}$ into $\mathcal{D}_{train}$ and $\mathcal{D}_{cal} = (X_i, i \in [\![n]\!])$.
- We compute novelty scores $S_i = g(X_i), i \in [\![n+m]\!]$ where $g : \mathbb{R}^d \to \mathbb{R}$.
- We compute the conformal p-values defined in (2).

In this framework, H2 and H3 are satisfied and the theoretical results of section 3 still hold.

4.2 Theoretical results

In this section, we discuss a DKW-type envelope for $\hat{F}_m$ 8. We start by introducing the discretized identity function

$$I_n(t) = \lfloor t(n+1) \rfloor / (n+1) = \mathbb{E}[\hat{F}_m(t)], t \in [0,1], \tag{10}$$

as well as the B^{DKW} bound

$$B^{DKW}(\lambda, n, m) := \mathbb{1}_{\lambda<1} \left[1 + \frac{2\sqrt{2\pi}\lambda\tau_{n,m}}{(n+m)^{1/2}} \right] e^{-2\lambda^2\tau_{n,m}}, \tag{11}$$

where $\tau_{n,m} = nm/(n+m) \in [(n \wedge m)/2, n \wedge m]$ is an "effective sample size".

Theorem 1. *Let us consider the process $\hat{F}_m$ 8 and the discrete identity function $I_n(t)$ 10. Let's assume H2 and H3. Then we have for all $\lambda > 0$, $n, m \geq 1$,*

$$\mathcal{P}\left(\sup_{t\in[0,1]} (\hat{F}_m(t) - I_n(t)) > \lambda \right) \leq B^{DKW}(\lambda, n, m). \tag{12}$$

In addition, $B^{DKW}(\lambda, n, m) \leq \delta$ for

$$\lambda^{DKW}_{\delta,n,m} = \psi^{(r)}(1); \tag{13}$$

$$\psi(x) = 1 \wedge \left(\frac{log(\frac{1}{\delta}) + log(1 + \sqrt{2\pi}\frac{2\tau_{n,m}x}{(n+m)^{1/2}})}{2\tau_{n,m}} \right)^{1/2}$$

where $\psi^{(r)}$ denotes the function ψ iterated r times (for an arbitrary integer $r \geq 1$).

Proof. The proof relies essentially on *Proposition 2*. First, observe that the LHS of (12) is 0 if $\lambda \geq 1$ so that we can assume $\lambda < 1$. We prove (12) with a more complex bound :

$$B^{DKWfull}(\lambda, n, m) := \frac{n}{n+m}e^{-2m\lambda} + \frac{m}{m+n}e^{-2n\lambda^2} + C_{\lambda,n,m}\frac{2\sqrt{2\pi}\lambda nm}{(n+m)^{3/2}}e^{-\frac{2nm}{n+m}\lambda^2},$$

where $C_{\lambda,n,m} = \mathbb{P}(\mathcal{N}(\lambda\mu, \sigma^2 \in [0,\lambda]) < 1$, for $\sigma^2 = (4(n+m))^{-1}$ and $\mu = n(n+m)^{-1}$.

Let $U = (U_1, ..., U_n)$ i.i.d. $\sim U(0,1)$, and denote $F^U(x) = U_{(\lfloor (n+1)x \rfloor)}, x \in [0,1]$. Conditionally on U, draw $(q_i(U), i \in [m])$ i.i.d. of common c.d.f F^U and let

$$\hat{G}_m(t) = m^{-1} \sum_{i=1}^{m} \mathbb{1}\left\{q_i(U) \leq t\right\}, \ t \in [0,1]$$

the empirical c.d.f. of $(q_i(U), i \in m)$. According to *Proposition 2*, we know that $\hat{F_m}$ has the same distribution as $\hat{G_m}$ unconditionally on U. Thus, for any fixed $n, m \geq 1$ and $\lambda > 0$,

$$\mathbb{P}\left(\sup_{t\in[0,1]} \hat{F}_m(t) - I_n(t) > \lambda \right) = \mathbb{E}\left[\mathbb{P}\left(\sup_{t\in[0,1]} \hat{G}_m(t) - I_n(t) > \lambda \mid U \right) \right].$$

Denote $Z = \sup_{t\in[0,1]}(FU(t) - I_n(t))$, which lies in $[0,1]$.

$$\mathbb{P}\left(\sup_{t\in[0,1]} \hat{F}_m(t) - I_n(t) > \lambda \right) \leq \mathbb{E}\left[\mathbb{P}\left(\sup_{t\in[0,1]} (\hat{F}_m(t) - F^U(t)) + Z > \lambda \mid U \right) \right] \quad \text{(by triangle inequality)}$$

$$\leq \mathbb{E}\left[\mathbb{P}\left(\sup_{t\in[0,1]} (\hat{F}_m(t) - F^U(t)) \geq (\lambda - Z)_+ \mid U \right) \right] \tag{P1}$$

Let's remark that Z conditional to U is a constant. In addition, since $(q_i(U), i \in m)$ are i.i.d $\sim F^U$ and their empirical cumulative distribution function (e.c.d.f.) conditionally to U is $\hat{F_m}$, we can apply the *DKW inequality (Massart 1990)* [3] :

$$\mathbb{P}\left(\sup_{t\in[0,1]} (\hat{F}_m(t) - F^U(t)) \geq (\lambda - Z)_+ \mid U\right) \leq e^{-2m(\lambda-Z)_+^2}.$$

Thus, we have

$$\mathbb{E}\left[\mathbb{P}\left(\sup_{t\in[0,1]} (\hat{F}_m(t) - F^U(t)) \geq (\lambda - Z)_+ \mid U\right)\right] \leq \mathbb{E}\left[e^{-2m(\lambda-Z)_+^2}\right].$$

We can show that the last bound can be rewritten, using Fubini's Theorem, as

$$\frac{n}{n+m}e^{-2m\lambda} + \frac{m}{m+n}e^{-2n\lambda^2} + C_{\lambda,n,m}\frac{2\sqrt{2\pi}\lambda nm}{(n+m)^{3/2}}e^{-\frac{2nm}{n+m}\lambda^2}$$

see [1], appendix C.4 for detailed calculus. Which finally leads to, according to (P1),

$$\mathbb{P}\left(\sup_{t\in[0,1]} (\hat{F}_m(t) - I_n(t)) \geq \lambda\right) \leq B^{DKWfull}(\lambda, n, m)$$

Since $n \vee m \geq nm/(n+m)$ and $C_{\lambda,n,m} \leq 1$, $B^{DKWfull}(\lambda, n, m) \leq B^{DKW}(\lambda, n, m)$ and we deduce :

$$\mathbb{P}\left(\sup_{t\in[0,1]} (\hat{F}_m(t) - I_n(t)) \geq \lambda\right) \leq B^{DKW}(\lambda, n, m)$$

Which concludes the proof of (12).
Let's now prove (13). If $\Psi(1) = 1$ then $\Psi^{(r)}(1) = 1$ for all r and the announced claim holds since $\mathrm{BDKW}(1, n, m) = 0$ by definition. We therefore assume $\Psi(1) < 1$ from now on. Since Ψ is non-decreasing, by an immediate recursion we have $\Psi^{(r+1)}(1) \leq \Psi^{(r)}(1) < 1$ for all integers r.

On the other hand, note that for any $x \in (0, 1)$ satisfying $\Psi(x) \leq x < 1$, it holds

$$\Psi(x) := \left(\frac{\log(1/\delta) + \log(1 + \sqrt{2\pi/2\tau_{nm}}x)}{\sqrt{2\tau_{nm}}(n+m)^{1/2}}\right)^{1/2}$$

and thus

$$\mathrm{BDKW}(\Psi(x), n, m) = \left[1 + \frac{2\sqrt{2\pi}\psi(x)\tau_n, m}{(n+m)^1/2}\right]\left[1 + \frac{2\sqrt{2\pi}x\tau_n, m}{(n+m)^{1/2}}\right]^{-1}\delta \leq \delta.$$

Since we established that $x = \Psi^{(r)}(1)$ satisfies $\Psi(x) \leq x$ for any integer r, the claim follows. □

Let us consider any thresholding novelty procedure

$$R(t) := \{i \in m : p_i \leq t\} \quad t \in (0, 1). \tag{14}$$

Then the following result holds true.

Corollary 2. *In the above novelty detection setting and under H2, the family of thresholding novelty procedures 14 is such that with probability at least* $1 - \delta$ *we have for all* $t \in (0, 1)$

$$FDP(R(t)) \leq \frac{\hat{m}_0}{In(t) + \hat{m}_0\lambda_{DKW}^{\delta n\hat{m}_0}\max\{1, |R(t)|\}} = FDP_{DKW}^{t\delta} \tag{15}$$

where $\lambda_{DKW}^{\delta n\hat{m}_0}$ *is given by (10) and* $\hat{m}_0$ *is any random variable such that*

$$\hat{m}_0 \geq \max\left\{r : \inf_t\left(\sum_{i=1}^m 1\{p_i > t\} + r\lambda_{DKW}^{\delta nr}(1 - In(t)) \geq r\right)\right\} \tag{16}$$

where r *is in the range* m *and the maximum is equal to* m *if the set is empty.*

The proof is admitted.

5 Application to the dataset

Now that we established the efficiency of transfer learning in the construction of conformal prediction intervals, we try, in this section, to adapt the strategies seen in previous sections on the Nasa data set. We reproduce the same framework as before, where $\mathcal{I}_{transfer}$ uses $\hat{\mu}(., \mathcal{D}_{train}, \mathcal{D}^X_{cal+test})$ with j-dot.

First, we apply transfer learning without creating a shift in the sample $\mathcal{D}_{train}$. This serves as an introductive simulation (see notebook).

Then, we create a shift in $\mathcal{D}_{train}$ thanks to the function : $w(x) = exp(x^T\beta)$ where $\beta = (-1, 0, 0, 0, 1)$. This allows to reweight the samples based on their feature values and the coefficients in β. We are thus able to apply transfer learning and create conformal prediction intervals. Figure 2 shows an attempt at this.

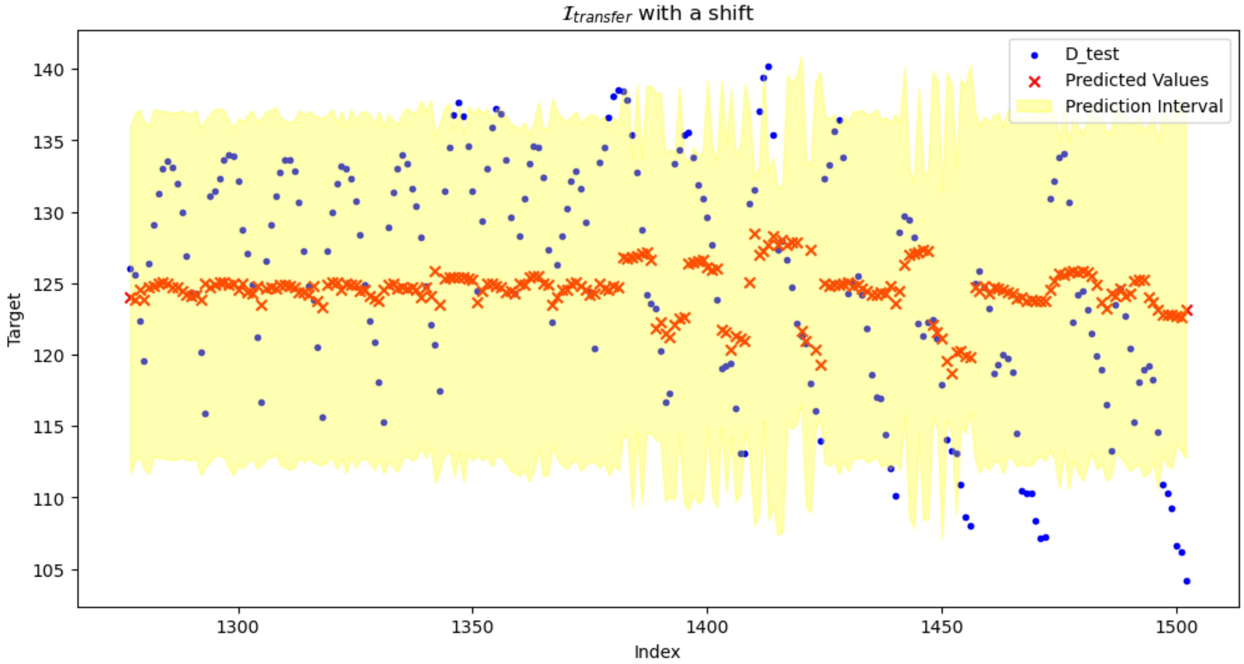


Figure 2: Conformal prediction intervals with domain shift

To further visualize the impact of the domain shift, we suggest comparing both $\mathcal{D}_{train}$ and $\mathcal{D}_{train\ shifted}$ for some features X_i:

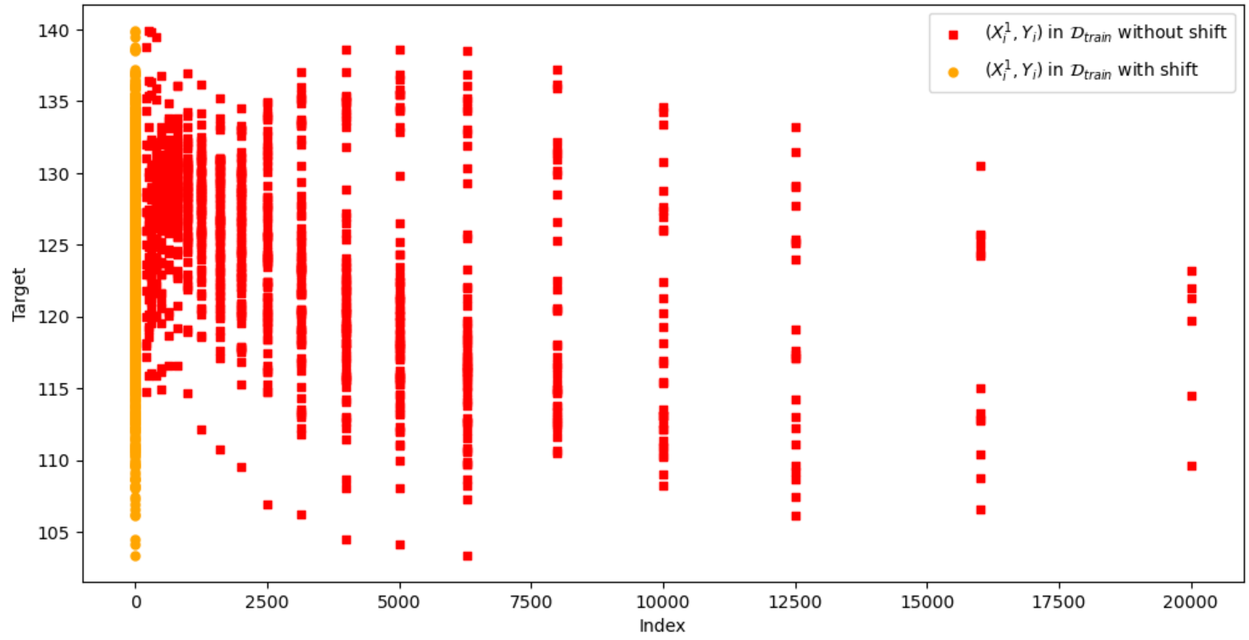


Figure 3: Comparison of $\mathcal{D}_{train}$ and $\mathcal{D}_{train\ shifted}$ for the 1st covariate

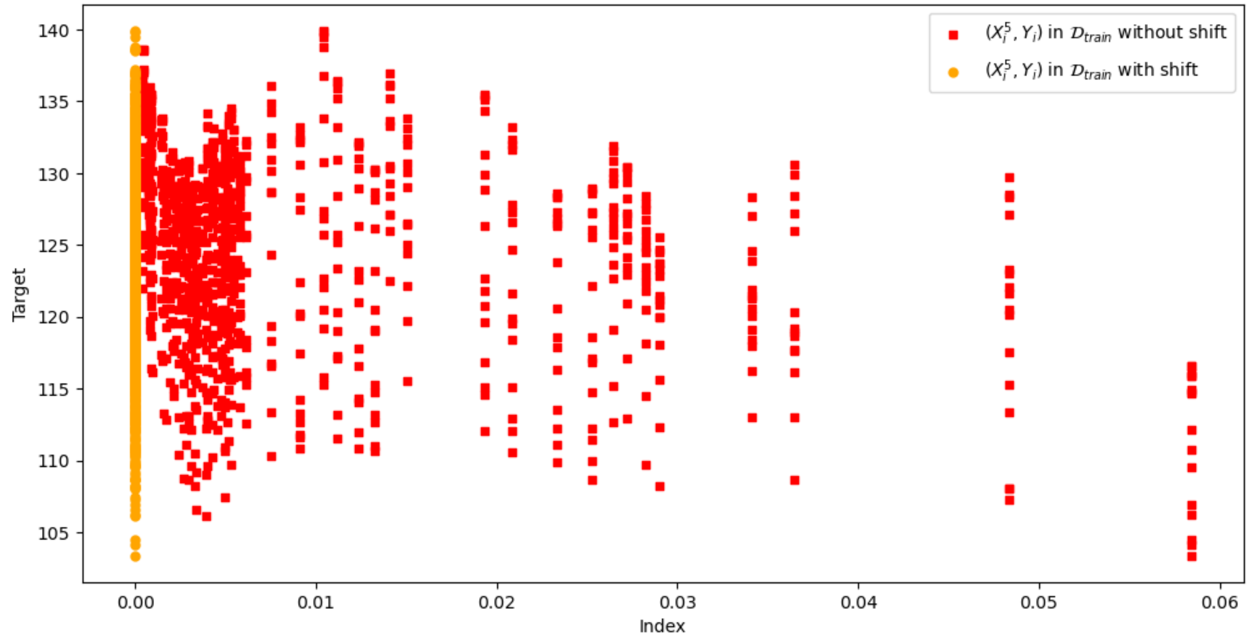


Figure 4: Comparison of $\mathcal{D}_{train}$ and $\mathcal{D}_{train\ shifted}$ for the 5th covariate

The transfer predictor seems to perform well when the shift is applied. However, one could criticize the length of the conformal prediction interval in figure 2. A change in the shift technique might leed to some improvements. Let's also note that the dimension of the features $d = 5$ makes the problem slightly more complex than the simulation done in section 3, where the features were one-dimensional.

References

[1] Roquain E., Blanchard G., Gazin U., (2023). *Transductive conformal inference with adaptive scores.*

[2] Dheeru Dua and Casey Graff. (2019) *UCI machine learning repository.* https://archive.ics.uci.edu/dataset/291/airfoil+self+noise

[3] Massart, P. (1990). *The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality.* Ann. Probab., 18(3):1269–1283.

[4] Bates, S., Cand'es, E., Lei, L., Romano, Y., and Sesia, M. (2023) *Testing for outliers with conformal p-values. Ann. Statist., 51(1):149–178.*

[5] Boyer, C. and Zaffran, M. (2023). *Tutorial on conformal prediction.* https://claireboyer.github.io/tutorial-conformal-prediction/

[6] Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). *Joint distribution optimal transportation for domain adaptation. In Advances in neural information processing systems 30 (NIPS 2017), volume 30.*

[7] Open Source Python implementation of JDOT. https://github.com/rflamary/JDOT