A Course on Transductive Conformal Inference

Created for the Linear Models in High Dimensions course of the Sorbonne University Master of Statistics.

Author :
Guillhem Artis

13/01/2024

Contents

1 Introduction

The purpose of this course is to provide insights to M2 students on some aspects of transductive conformal inference. Mostly based on Roquain et al. article [GBR23], it aims at explaining the control of the false coverage rate enabled by this theory and give an illustrative study with actual data. [GBR23] article made quite a significant step within statistical learning literature. Its mild (not to say "weak") assumptions on the observations allow the results, derived here, to be used in transfer learning context (thanks to adaptive scores, as we will see after). This article is organized as follows 1. we motivate our study on real problematic, 2. we provide a mathematical model to deploy the theory in 3. whose results will finally be used (to implement method) and illustrated in 4. and 5.

2 Motivation

Assume we want to predict the price of electricity in Germany given its daily consumption, averaged on the entire German poplulation. In addition, assume we have been provided historical data on the average daily electricity consumption and prices in France :

COUNTRY	FR_CONSUMPTION	TARGET
FR	-0.427458	0.028313
FR	-1.003452	-0.112516
FR	1.978665	-0.180840
FR	-0.617038	-0.071733
FR	-0.765120	0.932105

Table 1: France Consumption and Price Data

DE_CONSUMPTION
-0.849198
-0.811337
-0.331101
-1.062255
1.629315

Table 2: Germany Consumption Data

(The interested reader may find this data on the website [Par23]. We have extracted only few columns, so as to transform the more general prediction problem into a transfer learning problem.)

The task we want to perform is to predict and give simultaneous confidence intervals around the predictions. This is a regression learning problem with tabular data. The strategy will be, at first, to come up with a transfer learning algorithm, enabling us to make use of French Data, to estimate the prices in Germany., and subsquently, use conformal prediction theory to compute confidence intervals around our predictions. Warning

: such confidence intervals must be controlled and guaranteed. Here are the steps we follow to tackle the problem :

1. we provide a mathematical desription of this task, to
2. introduce the theoritical results on,
3. the predicting intervals.

3 Method and mathematical modelling

As being said, the mathematical framework embracing the difficulty of the problem are statistics uncertainty analysis and transfer learning (in a regression context).

3.1 Mathematical settings

At first, the reader may forget the transfer learning setting and assume we possess n independent observations $\mathcal{D}_{cal} = ((X_1, Y_1), ..., (X_n, Y_n))$ identically distributed as $(X, Y) \sim \mathcal{P}$ where $\mathcal{P}$ is the unknown joint distribution supported in $\mathbb{R}^d \times \mathbb{R}$ (for the sake of simplicity one may keep in mind : $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$. Regression learning seeks to estimate a measurable function for predicting numerical outcomes : $f : \mathbb{R}^d \to \mathbb{R}$ minimizing the expected risk $\mathbb{E}[l(f(X, Y)]$, for a given loss function $l : \mathbb{R}^2 \to \mathbb{R}$. Setting aside predicting intervals construction, the we want to perform is slightly different. We are not required to estimate f on the all domain of X, but to predict the value of m new data points (i.e. estimate the regression funtion on a finite discrete subset of $\mathbb{R}^d$).

The right formulation for our problem is provided by Roquain et al. article [GBR23] : given m new independent data points $\mathcal{D}_{test} = ((X_{n+1}, Y_{n+1}), ..., (X_{n+m}, Y_{n+m}))$ generated from $\mathcal{P}$, build m prediction intervals for $Y_{n+1}, ..., Y_{n+m}$ given $X_{n+1}, ..., X_{n+m}$. More formally, the goal is to construct $\mathcal{I} = (\mathcal{I}_i)_{1 \leq i \leq m}$, a familily of m random (since they a priori depend on the random variables) intervals of $\mathbb{R}$ such that the amount of coverage errors $(\mathbb{1}_{[Y_{n+1} \notin \mathcal{I}_i]})_{1 \leq i \leq m}$ is controlled.

Remark 3.1. In our "real world" problem the test variables $(X_{n+1}, ..., X_{n+m})$ are unlabeled, i.e. $(Y_{n+1}, ..., Y_{n+m})$ are not provided (2).

3.2 Tools and Assumptions

Since it provides sharp uncertainty quantification, the conformal inference framework is conducive for our study. A key tool for this theory to build prediction intervals is *(non)-conformity score* :

Definition 3.2. Let $i \in [\![1, n+m]\!]$. We call i-th *(non)-conformity score* denoted S_i the real-valued random interval :

$$S_i := \left| Y_i - \hat{\mu}\left(X_i; (\mathcal{D}_{train}, \mathcal{D}^X_{cal+test})\right)\right|,$$

where the value $\hat{\mu}\left(x; (\mathcal{D}_{train}, \mathcal{D}^X_{cal+test})\right)$ is the output of a machine learning prediction of Y_i at point $X_i = x$

Remark 3.3. The non conformity score S_i corresponds to a residual between Y_i and the prediction at point X_i. An important observation is that our machine learning regressor $\hat{\mu}$ also depends on the calibration and test sets. This point is crucial since we want our results to be applicable to the class of transfer learning problems. As a consequence the scores are not i.i.d., but providing an hypothesis on the regressor's dependency towards its arguments, we assume them to be exchangable. *(exch.)*

Definition 3.4. When *(exch.)* 3.3 is assumed, the $(n+m)$-uplet of (non)-conformity scores are such that, for any $J \subset [\![1+n+m]\!]$, and for any permutation σ on J,

$$\mathcal{L}((S_{\sigma(i)})_{i\in J}) = \mathcal{L}((S_i)_{i\in J}).$$

The assumption on the regressor's depencies, ensuring scores exchangeability, is called the permutation invariant hypothesis *(PermInv)* and is defined as follow :

Definition 3.5. We say that the predictor $\hat{\mu}$ is permutation invariant when :

$$\forall x \in \mathbb{R}^d \quad \hat{\mu}(x; (\mathcal{D}_{train}, \mathcal{D}^X_{cal+test}) \text{ is invariant by permutation on } \mathcal{D}^X_{cal+test}.$$

To avoid undesirable cases we assume that the $(n+m)$-uplet of scores has no ties almost surely *(NoTies.)*. The combination of the above assumptions *(exch.)* 3.2 and *(NoTies.)* is called *mild density assumption* , in the sense that *(PermInv)* 3.2 holds true in general (otherwise adding an independent gaussian noise suffices to ensure it). Conversely, when i.i.d. *(i.i.d.)* assumption is made upon the scores, this very assumption combined with *(NoTies)* will be refered as *strong density assumption* .

Finally, the concept of p-values will be highly recquired in the next parts. Indeed, the statistical inference will rely upon the so-called split conformal p-values :

$$p_i = \frac{1}{n+1}\left(1 + \sum_{j=1}^{n} \mathbb{1}_{[S_j \geq S_i]}\right), \quad i \in [\![1, m]\!] \tag{1}$$

since they enable the construction of the empirical cumulative distribution function :

$$\begin{array}{rccc} \hat{F}_m & : & [0,1] & \to & [0,1] \\ & & t & \mapsto & \frac{1}{m}\sum_{i=1}^{n} \mathbb{1}_{[p_i \leq t]} \end{array} \tag{2}$$

Remark 3.6. $(n+1)p_i$ is equal to the rank of S_{n+i} in the set of values $\{S_1, ..., S_n, S_{n+i}\}$. The smaller the i-th p-value is, the higher the residual test score S_i is, compared to the calibration scores $(S_j)_{1\leq j\leq n}$. They are called "split"-conformal p-values since they computation involves a comparaison to a fraction of the scores provided by the calibration set.

3.3 Strategy and Preliminary Results

To come up with uncertainty bounds on our electric energy prices in Germany, we have to construct predicting intervals $\mathcal{I} = (\mathcal{I}_i)_{1\leq m}$ and subsequently controlling their false coverage proportion :

$$FDP(\mathcal{I}) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{[Y_{n+i} \notin \mathcal{I}_i]}. \tag{3}$$

The strategy we propose here is the one from Roquain et al. [GBR23]. Having at hand the tools introduced previously, we construct predicting intervals using the split-conformal p-values 1. then we will upper bound the false covering proportion using a concentration inequality on the empirical cumulative distribution function 2. In order to provide such a concentration-type inequality we would have beforehand studied the joint distribution for th m plit-conformal p-values. That is because the distribution function $\hat{F}_m$ is set by those p-values.

Under the assumptions above, some propositions can be derived. First under the *stong density assumptions*, the p-values are independent and identically distributed.

Proposition 3.7. *Under the strong density assumption (3.2), conditionnally on $\mathcal{D}_{cal}$, the p-values are i.i.d and of common distribution :*

$$p_1|\mathcal{D}_{cal} \sim P^U$$

where,

$$P^U\left(\left\{\frac{l}{n+1}\right\}\right) = U_{(l)} - U_{(l-1)}, \quad l \in [\![1, n-1]\!]$$

with $0 = U_{(0)} \leq U_{(1)} \leq ... \leq U_{(n)} \leq U_{(n+1)} = 1$ the increasing values of $U = (U_1, ..., U_n) = (1 - F(S_1), ..., 1 - F(S_n))$, F being the common cumulative distribution of the calibration scores (which we recall are random valued $\mathcal{D}_{cal}$-measurable). In addition the pseudo vector U has independent coordinates uniformly distributed on $[0, 1]$.

Proof. Straightforwardly, since U_i is a deterministic function of S_i and the calibration scores are i.i.d coordinates of U. The proof can be separated in two steps.

- Step 1 : Show that $U_1 \sim Unif[0, 1]$.

  Thanks to above observation, it suffices to show that $\mathbb{P}(U_1 \leq t) = t$ for each $t \in [0, 1]$. So for a fixed t, compute :

  $$\begin{aligned}\mathbb{P}(U_1 \leq t) &= \mathbb{P}(1 - F(S_1) \leq t) \\ &= 1 - \mathbb{P}(F(S_1) < 1 - t)\end{aligned}$$

  (NoTies) assumption implies score distribution to be atomless. Hence F is continous, and one may use the pseudo-inverse denoted F^{-1},

  $$\begin{aligned}\mathbb{P}(U_1 \leq t) &= 1 - \mathbb{P}(F(S_1) \leq 1 - t) \\ &= 1 - \mathbb{P}(S_1 \leq F^{-1}(1 - t)) \\ &= 1 - F(F^{-1}(1 - t)) \\ &= 1 - (1 - t) = t.\end{aligned}$$

- Step 2 : Prove that $p_1|\mathcal{D}_{cal} \sim P^U$.

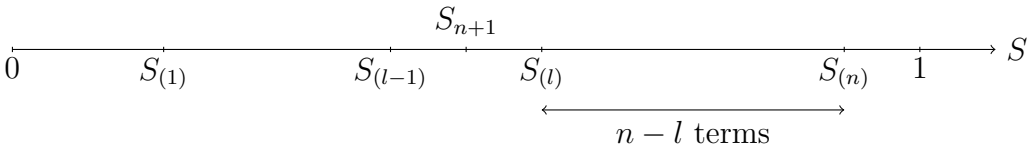
  For $x \in [0, 1]$, by its definition p_1 takes value in a discrete set of $[0, 1]$.

  $$\{p_1 \leq x\} \iff p_1 \leq \frac{\lfloor x(n+1) \rfloor}{n+1}$$

It follows that,

$$\begin{aligned}
\{p_1 \leq x\} &\iff (n+1)p_1 \leq \lfloor x(n+1) \rfloor \\
&\iff \sum_{j=1}^{n} \mathbb{1}_{[S_j \geq S_{n+1}]} + 1 \leq \lfloor x(n+1) \rfloor \\
&\iff \#\{j \in [\![1, n]\!] \mid S_j \leq S_{n+1}\} \leq \lfloor x(n+1) \rfloor - 1 \\
&\iff S_{n+1} > S_{(l-1)},
\end{aligned}$$

with l defined below :



Since $n-l$ must be, at most equal to $\lfloor x(n+1) \rfloor - 1$, then $l-1 = n-(\lfloor x(n+1) \rfloor - 1)$
Hence

$$\begin{aligned}
\mathbb{P}(p_1 \leq x \mid \mathcal{D}_{cal}) &= \mathbb{P}(S_{n+1} > S_{(n+1-\lfloor x(n+1) \rfloor)} \mid \mathcal{D}_{cal}) \\
&= 1 - \mathbb{P}(S_{n+1} \leq S_{(n+1-\lfloor x(n+1) \rfloor)} \mid \mathcal{D}_{cal}) \\
&= 1 - F(S_{(n+1-\lfloor x(n+1) \rfloor)}) \quad \text{with } S_{(n+1-\lfloor x(n+1) \rfloor)} \text{ deterministic} \\
&= 1 - F\left(U_{n+1-(n+1-\lfloor x(n+1) \rfloor)}\right)
\end{aligned}$$

By noticing that $F(U_j) = 1 - F_{(n+1-j)}$,

$$\begin{aligned}
\mathbb{P}(p_1 \leq x \mid \mathcal{D}_{cal}) &= 1 - F\left(U_{(\lfloor x(n+1) \rfloor)}\right) \\
&= P^U\left(\frac{\lfloor x(n+1) \rfloor}{n+1}\right) \\
&= P^U(x).
\end{aligned}$$

□

Additionnaly, by integrating over $U \sim Unif[0,1]$ the family of split-conformal p-values haas distribution $P_{n,m}$ supported on $[0,1]^m$. $P_{n,m}$ is then defined as follow : $P_{n,m} = \mathcal{D}(q_i, i \in [\![1, m]\!])$, where $(q_1, ..., q_m|U) \sim P^U$ (i.i.d) and $U = (U_1, ..., U_n) \sim Unif[0,1]$ (i.i.d).

Remark 3.8. The attentive reader may notice that p-values family only depends on the the $n+m$ scores ranks. This remark lead to the last result of this part :

Proposition 3.9. *Under the (mild density assumption) (3.2), the family of m split conformal p-values has joint distribution $P_{n,m}$ and is independent of the score distributions.*

Remark 3.10. The last remark unveils the mechanism of result generalization from *stron density assumption* toward *mild density assumption*, but it is not considered as a proof. $P_{n,m}$ being independent of scores distribution, we say that $P_{n,m}$ is "universal".

4 Main theoritical results

Thanks to preliminary part, we can now address on the main results of Roquain et al. article [GBR23]. They will enable us to construct predicting intervals and control false coverage proportion, ensuring guaranteed confidence level for future predictions. First, let's study the joint distribution of the p-values.

4.1 On the joint distribution of p-values

The article, this coursed was based on, has contributed to the staticians community by providing the following result :

Theorem 4.1. *Under the mild density assumption, the joint distribution $P_{n,m}$ of the m split-conformal p-values, is the distribution of the colors of m successive draws in a standard Pólya urn model with $n+1$ colors labeled $\left\{\frac{l}{n+1}, l \in [\![1, n+1]\!]\right\}$.*

In the standard Pólya urn model with $n+1$ colors, we consider an urn containing at the begining $n+1$ balls of different colors (one different color for each ball). At each draw, a ball is selected uniformly at random and its color is noted. The Pólya urn scheme is characterized by the rule that after each draw, the selected ball is returned to the urn along with an additional ball of the same color. This process is repeated m times successively, which influences the likelihood of drawing each color in future draws. The probability of drawing a ball of a particular color becomes higher when this very ball has already been drawn. Such scheme leads in to a richly interconnected stochastic process.

To grasp the significancy of this result, the formulas given in appendix A [GBR23] reveal some particular behaviors of this p-value family. First, they tend to take the same value. To be more precise the more equal p-values there are, the higher they joint distribution is (positive depedence behaviour). And second, when n is bounded by above, the p-values family displays a non-concentration behavior. They would tend to take different each value of the color set : $\left\{\frac{l}{n+1}, l \in [\![1, n+1]\!]\right\}$ equally likely. Such result seems to harden our initial problem. Indeed, since taking uniform values on the all possible universe (probability) is characteristic to a chaotic behaviour. Those behaviours are difficult to understand, let alone controlling their evolution. Nevertheless, the next result will demonstrate, the family concentrate around its mean, as long as n and m tend to infinity.

4.2 On the false coverage proportion

We now provide a DKW-type envelope for the empirical distribution function (2) of conformal p-values. Let us introduce the discretized identity function

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

and the following bound:

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

where $\tau_{n,m} := \frac{nm}{n+m} \in \left[\frac{\min(n,m)}{2}, \min(n,m)\right]$ is an "effective sample size".

Theorem 4.2 (DKW-type Envelope)**.** *Let us consider the process $\hat{F}_m$ defined by (2). Under the mild density assumption, we have for all $\lambda > 0$, $n, m \geq 1$,*

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

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

$$\begin{aligned}
&\lambda^{DKW_{\delta,n,m}} = \Psi^{(r)}(1);\\
&\Psi(x) = 1 \wedge \left(\frac{log(1 + \sqrt{2\pi}\frac{2\tau_{n,m}x}{(n+m)^{0.5}}}{2\tau_{n,m}}\right)^{0.5}
\end{aligned}$$

where $\Psi(r)$ denotes the function Ψ iterated r times $\Psi^{(r)}(x) = \underbrace{\Psi(\Psi(\ldots\Psi}_{r\ times}(x)\ldots))$.

Proof. The proof of the DKW-type enveloppe is rather technic. We explains its starting point. Since the empirical cumulative distribution function $\hat{F}_m(t)$ involves the p-values, one may want to introduce the cumulative distributive function F^U of P^U (in the sense that if $A \mid P^U$ then the c.d.f. of A is F^U). Recall U is the vector composed of the n i.i.d. uniformly supported on $[0,1]$ random variables $U_i = 1 - F(S_i)$. For a fixed $x \in [0,1]$

$$\begin{aligned}
F^U(x) = \mathbb{P}(A \sim P^U \leq x) &= \sum_{k=1}^{\lfloor x(n+1)\rfloor} \mathbb{P}(A = \frac{k}{n+1})\\
&= \sum_{k=1}^{\lfloor x(n+1)\rfloor} P^U(\frac{k}{n+1}) = \sum_{k=1}^{\lfloor x(n+1)\rfloor} U_{(k)} - U_{(k-1)}\\
&= U_{(\lfloor x(n+1)\rfloor)}
\end{aligned}$$

Write :

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

Since,

$$\sup_{t\in[0,1]}(\hat{F}_m(t) - F^U(t) + F^U(t) - I_n(t)) \leq \sup_{t\in[0,1]}(\hat{F}_m(t) - F^U(t)) + \sup_{t\in[0,1]}(F^U(t) - I_n(t)),$$

by denoting $Z = \sup_{t\in[0,1]}(F^U(t) - I_n(t))$, it follows :

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

since $F^U(0) = 0$, $\sup_{t\in[0,1]}(\hat{F}_m(t) - I_n(t)) \geq 0$ and equal to 0 with probability 0. Henceforth,

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

Since, $\hat{F}_m(t)$ is the empirical pendant of $F^U(t)$, by using a Dvoretzky, Kiefer and Wolfowitz concentration inequality, see theorem 1. in [Mas90] :

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

From then on, it turns out to upper bound the RHS expectation. Recall that if A is a posive random variable : $\mathbb{E}[A]=\int_0^\infty\mathbb{P}(A\geq t)dt$. Since $e^{-2m(\lambda-Z)_+^2}$ is supported in $[0,1]$, one has :

$$\begin{aligned}\mathbb{E}\left[e^{-2m(\lambda-Z)_+^2}\right]&=\int_0^1\mathbb{P}\left(e^{-2m(\lambda-Z)_+^2}\geq v\right)dv\\&\leq\int_0^{e^{-2m\lambda^2}}dv+\int_{e^{-2m\lambda^2}}^1\mathbb{P}\left(e^{-2m(\lambda-Z)_+^2}\geq v\right)dv\end{aligned}$$

By inverting in the probability term, we get :

$$\mathbb{E}\left[e^{-2m(\lambda-Z)_+^2}\right]\leq e^{-2m\lambda^2}+\int_{e^{-2m\lambda^2}}^1\mathbb{P}\left(Z>\lambda-\sqrt{-log(v)/2m}\right)dv.$$

So it boils down controlling Z survival function,

$$\mathbb{P}(Z>x)=\mathbb{P}\left(\sup_{t\in[0,1]}(U_{(\lfloor(n+1)t\rfloor)}-(n+1)t/(n+1))>x\right)$$

After some computations, we come up with :

$$\mathbb{P}(Z>x)\leq\mathbb{P}\left(\exists k\in[\![1,n]\!]:\hat{H}_n(x+k/(n+1))-\lceil x+k/(n+1)\rceil\leq-x\right)$$

where $\hat{H}_n$ denotes the empirical cdf of $(U_1,...,U_n)$. Still admiting this DKW result (here adapted to the left tail), one has $\mathbb{P}(Z>x)\leq e^{-2nx^2}$. Finally injecting this result in the upper bound above we obtain :

$$\mathbb{E}\left[e^{-2m(\lambda-Z)_+^2}\right]\leq e^{-2m\lambda^2}+\int_{e^{-2m\lambda^2}}^1 e^{-2n(\lambda-\sqrt{-log(v)/(2m)})^2}dv$$

Up to this stage, all previous results have been used (to obtained the first bound of the theorem at least). That is why we stop the proof here and let the reader see the end of the proof in appendix C of [GBR23].

□

Remark 4.3. As hard to embrace as the DKW-type bound is, it will enable us to control the FCP of our predicting intervals. This result the might be seen as the conceptual climax of this course.

Let's obtain the confidence intervals. To begin, we determine one predicting interval with confidence $1-\alpha$ for $Y_{n+i} : \mathcal{C}_i$. The strategy is to invert in α $\{p_i > \alpha$ with respect to Y_{n+i} :

$$\begin{aligned}\{p_i \leq \alpha\} &= \{n+1+\sum_{1\leq j\leq n} \mathbb{1}_{S_j \geq S_{n+i}} \leq (n+1)\alpha\} \\ &= \{\sum_{1\leq j\leq n} \mathbb{1}_{S_j \geq S_{n+i}} \geq (n+1)(\alpha-1)\}\end{aligned}$$

By substracting $n+1$ on both sides :

$$\begin{aligned}\{p_i \leq \alpha\} &= \{\sum_{1\leq j\leq n} \mathbb{1}_{S_j < S_{n+i}} \geq (n+1)(1-\alpha)\} \\ &= \{\sum_{1\leq j\leq n} \mathbb{1}_{S_j < S_{n+i}} \geq \lceil (n+1)(1-\alpha)\rceil\} \\ &= \{S_{(\lceil (n+1)(1-\alpha)\rceil)} < S_{n+i}\}\end{aligned}$$

And by using definition of S_{n+i}, we get :

$$C_i(\alpha) = [\hat{\mu}(X_{n+i}; (\mathcal{D}_{train}, \mathcal{D}^X_{cal+test})) \pm S_{(\lceil (n+1)(1-\alpha)\rceil)}$$

Under *(mild density assumption)*, proposition 3.7 ensures that :

$$\mathbb{P}(Y_{n+i} \in \mathcal{C}_i(\alpha)) \geq 1-\alpha.$$

The conformal predicting region is defined by $\mathcal{C}(\alpha) = (\mathcal{C}_i)_{1\leq i\leq m}$. We have controlled each interval individually, we want to control them simultaneously.

$$\mathbb{P}(FCP(\mathcal{C}(\alpha)) \leq \beta) \geq 1-\delta,$$

is the quantity we aim to keep under track. Three parameters : α the confidence level of the prediction intervals (individually), β the maximum false coverage proportion we accept and δ, which controls false coverage proportion fluctuation around its mean. Basically they have 2 degrees of freedom. We aim at finding a family of random variables $(F\bar{C}P_{\alpha,\delta})_{\alpha\in(0,1)}$, such that :

$$\mathbb{P}(\forall \alpha \in (0,1), \quad FCP(\mathcal{C}(\alpha) \leq F\bar{C}P_{\alpha,\delta}) \geq 1-\delta. \tag{4}$$

The initial price prediction problem recquires to bound by above the FCP (i.e set a maximum value of $F\bar{C}P_{\alpha,\delta}$) while guaranteeing it (i.e. set a minimum value of δ). The level, in order to get a theoritical guaranteed will have to be adjusted. This final result provides just that :

Corollary 4.4. *Let $n, m \geq 1$, and keep the same notations. Under (mild density assumption) : for any $\bar{\alpha} \in [0,1], \delta \in (0,1), \mathcal{C}(\alpha = t_{\bar{\alpha},\delta})$ satisfies 4 provided that $t_{\bar{\alpha},\delta}$ is chosen such that :*

$$\mathbb{P}_{(p_1,\ldots,p_m)\sim P_{n,m}}(p_{(\lfloor \bar{\alpha} m\rfloor+1)} \leq t_{\bar{\alpha},\delta}) \geq \delta$$

Remark 4.5. Indeed, this result enables to guarantee (in probability δ) a maximum false covering proportion $\bar{\alpha}$ by adjusting the confidence intervals $(t_{\bar{\alpha},\delta})$.

5 Preliminary application - on simulated data

The breakthrough made by [GBR23] is to have developed theoretical bounds under the *mild density assumption* on the false covering proportion. Such an assumption only requires the machine learning estimate $\hat{mu}$ to be *PermInv*. Thus machine learning procedures can be adaptativ in the sense that they may depend on the calibration and test set. Such a capacity allows to tackel wider classes of supervised learning problems such as transfer learning. Even though, we introduce briefly the transfert learning setting, the reader may find additionnal details in [Cou+17]. Assume we have at hand a training set $\mathcal{D}_{train}$ distributed as $\mathcal{P}_S$ in a source domain, and a calibration+test set in a target domain distributed as $\mathcal{P}_t$. All the covariables are supposed to be i.i.d.. Besides, we make the "covariate shift assumption" : the differences between the domains are characterized by a change in the feature distributions $P(X)$, while the conditional distributions $P(Y|X)$ remain unchanged.

To apply our theory (and as a warmup for the real task), we simulate data such as part 3.5 in [GBR23], and reproduce the same exercices. Let (W_i, Y_i) i.i.d. with $Y_i|W_i \sim N(\mu(W_i), \sigma^2)$ for some function μ and parameter $\sigma > 0$. We programm a transfer learning algorithm based on optimal transport, see the github of [Cou+17], by performing a RBF kernel ridge regression with : $|D_{\text{train}}| = 1000$, $n = m = 50$, $\mu(x) = \cos(x)$, $W_i \sim U(0, 5)$, $f_1(x) = x$, $f_2(x) = 0.6x + \frac{x^2}{25}$ and $\sigma = 0.1$. Then we compute marginal predicting intervals with a confidence-level of 95% (having made use of some functions from [Boy]). Note that the confidence levels are not simultaneously guaranted here.

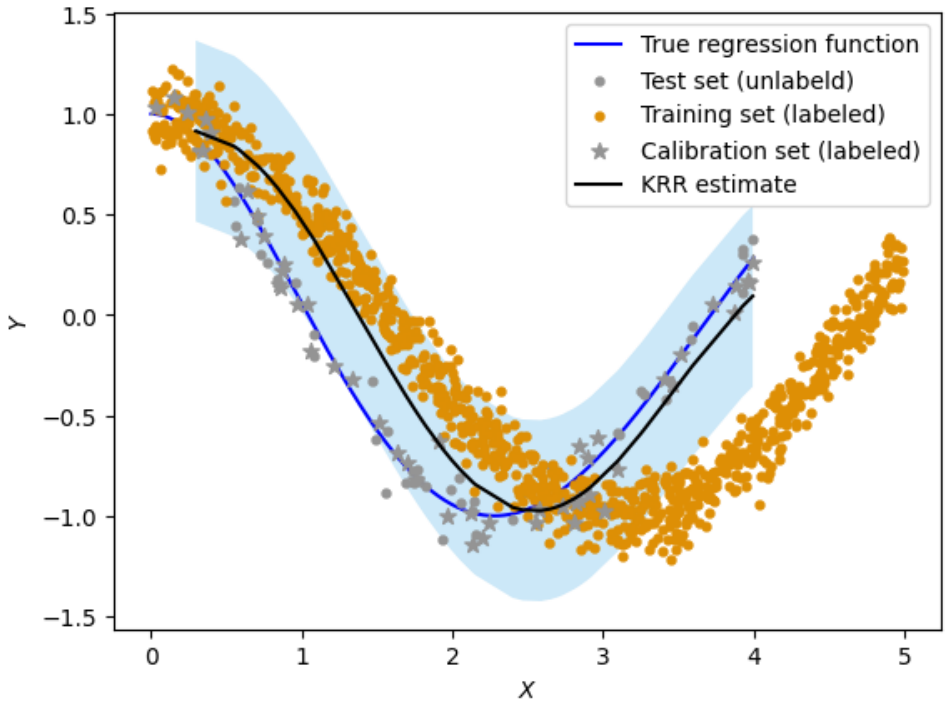


Figure 1: Predicting 95% confidence intervals from a RBF kernel ridge regressor with simulated dataset. At first glance, few test points are outide the intervals. Because the KRR estimate is rather far (in a least square sense) from the true regression function. The learning part seems more error providind than the conformal prediction.

(The reader may find all the programms in "Simulations notebook" [Gui]).

6 Application - What is the electric energy price in Germany ?

This problem was introduced in part (2). The datas were cleaned up so as to get a problem suitable for our analysis recquirements (domain shift hypothesis close enough to reality,...). Formally we apply the same learning algorithm as with simulations and the same procedures to construct the predicting intervals. The reader may find the codes in "Actualdataset notebook" [Gui].

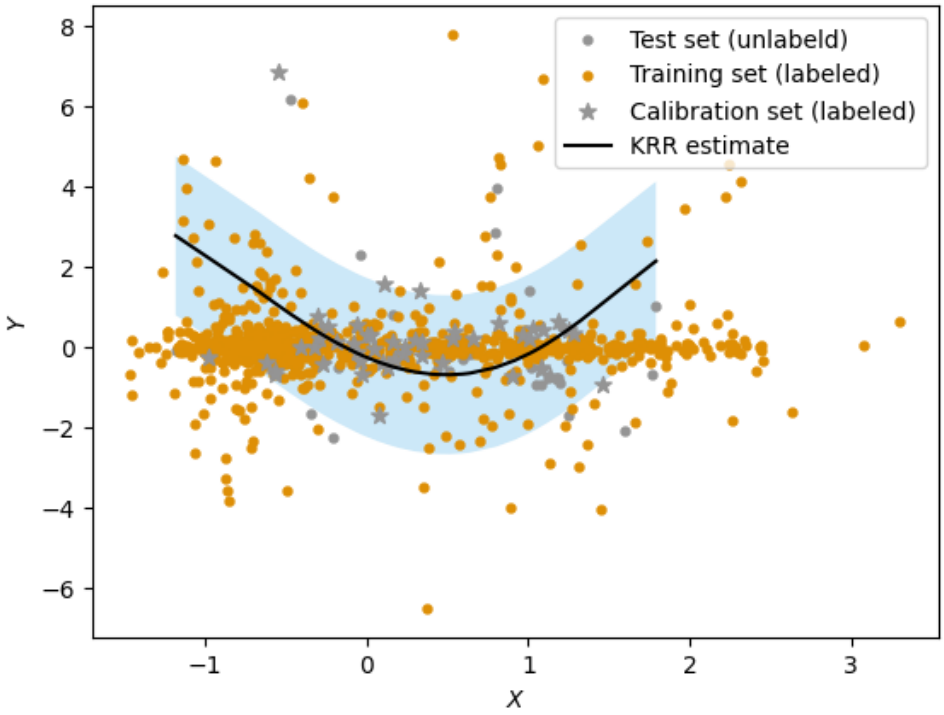


Figure 2: Predicting 95% confidence intervals from a RBF kernel ridge regressor with actual data.

The frequency of test points covered by the predicting region, we get : 0.8, which is rather good. Nevertheless, the machine learning algorithm seems not to perform well. The real issue stems from an oversimplification of the dataset. We did not keep enough features to transport price knowledge from France to Germany.

References

[Boy] Claire Boyer. *tutorial-conformal-prediction*. https://github.com/claireBoyer/tutorial-conformal-prediction.

[Cou+17] Nicolas Courty et al. *Joint Distribution Optimal Transportation for Domain Adaptation*. 2017. arXiv: 1705.08848 [stat.ML].

[GBR23] Ulysse Gazin, Gilles Blanchard, and Etienne Roquain. *Transductive conformal inference with adaptive scores*. 2023. arXiv: 2310.18108 [stat.ME].

[Gui] Guillhem. *conformal inference*. https://github.com/GuillhemArtis/Conformal_Inference.

[Mas90] Massart. *The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality*. 1990.

[Par23] ENS Paris. *Data Challenge QRT*. 2023. URL: https://challengedata.ens.fr/challenges/97.