Conformal prediction: A unified review of theory and new challenges

MATTEO FONTANA[1,2,3,c], GIANLUCA ZENI[1,a] and SIMONE VANTINI[1,b]

[1]*MOX-Department of Mathematics, Politecnico di Milano, Italy.* [a]*gianluca.zeni@mail.polimi.it,*
[b]*simone.vantini@polimi.it*
[2]*Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Italy*
[3]*now at European Commission, Joint Research Centre (JRC), Ispra, Italy.* [c]*matteo.fontana@ec.europa.eu*

In this work we provide a review of basic ideas and novel developments about Conformal Prediction — an innovative distribution-free, non-parametric forecasting method, based on minimal assumptions — that is able to yield in a very straightforward way prediction sets that are valid in a statistical sense also in the finite sample case. The discussion provided in the paper covers the theoretical underpinnings of Conformal Prediction, and then proceeds to list the more advanced developments and adaptations of the original idea.

Keywords: Conformal prediction; nonparametric statistics; prediction intervals; review

1. Introduction

In the the latest 25 years, a new method of prediction with confidence, called Conformal Prediction (CP), was introduced and developed. CP allows to produce prediction sets with guarantees on the error rate, exclusively under the simple i.i.d. assumption. Reliable estimation of prediction regions is a significant challenge in both machine learning and statistics, and the promising results generated by CP have generated further extensions of the original conformal framework. The increasing amount of real-world problems where robust predictions are needed is testified by the multitude of scientific works where CP is used.

In a nutshell, conformal prediction uses past experience in order to determine precise levels of confidence in new predictions. Using Gammerman, Vovk and Vapnik (1998)'s words in the very first work on the topic, it is "a practical measure of the evidence found in support of that prediction". In order to do this, it estimates how "unusual" a potential example looks with respect to the previous ones. Prediction regions are generated plainly by including the examples that have quite ordinary values, or better those ones that are not very unlikely. Conformal algorithms are proven to be always valid: the actual confidence level is the nominal one, without requiring any specific assumption on the distribution of the data except for the i.i.d. assumption. There are many conformal predictors for any particular prediction problem, whether it is a classification one or a regression one. Indeed, we can construct a conformal predictor from any method for scoring the similarity (conformity, as it is called) of a new example with respect to the old ones. For this reason, it can be used with any statistical and machine learning algorithm. Efficient performances let to understand the growing interest on the topic over the last few years.

The cornerstone of the related literature is "Algorithmic learning in a random world", published in 2005, where Vladimir Vovk and coauthors explains thoroughly all the theoretical fundamentals of CP There is only another, more recent, work that gives an overview on the topic, namely the book Conformal prediction for reliable machine learning, edited by Balasubramanian, Ho and Vovk (2014). The mentioned book addresses primarily applied researchers, showing them the practical results that

can be achieved and allowing them to embrace the possibilities CP is able to give. Therefore, its focus is almost totally on adaptations of conformal methods and the connected real-world applications.

In the latest years, an extensive research effort has been pursued with the aim of extending the framework, and several novel findings have been made. We have no knowledge of in-depth publications that aim to capture these developments and give a picture of recent theoretical breakthroughs. Moreover, there is a great deal of inconsistencies in the extensive literature that has been developed (e.g. regarding notation.), so the need for such an up-to-date review is then evident, and the aim of this work is to address this need of comprehensiveness and homogeneity.

As in recent papers, our discussion is focused on CP in the batch mode. Nonetheless, properties and results concerning the online setting, which tended to be the focus of the earliest works on the subject, are not omitted.

The paper is divided into two parts: Part 2 gives an introduction to CP for non-specialists, explaining the main algorithms, describing their scope and also their limitations, while Part 3 discusses more advanced methods and developments. Section 2.1 introduces comprehensively the original version of conformal algorithm, and let the reader familiarize with the topic and the notation. In Section 2.2, a simple generalization is introduced: each example is provided with a vector of covariates, like in classification or regression problems, which are tackled in the two related subsections. Section 1 of the Supplementary Material (Fontana, Zeni and Vantini, 2023) shows a comparison between CP and alternative ways of producing confidence predictions, namely the Bayesian framework and the statistical learning theory, and how CP is able to overcome their weak points. Moreover, we refer to an important result concerning the optimality of conformal predictors among valid predictors (in Section 1.1 of the Supplementary Material (Fontana, Zeni and Vantini, 2023)). Section 2.3 deals with the online framework, where examples arrive one by one and so predictions are based on an accumulating data set.

In Part 3, we focus on three important methodological themes. The first one is the concept of statistical validity: Section 3.1 is entirely devoted to this subject and introduces a class of conformal methods, namely Mondrian conformal predictors, suitable to gain object conditional validity in partial sense. Secondly, computational problems, and a different approach to conformal prediction — the inductive inference — to overcome the transductive nature of the basic algorithm (Section 3.2). Even in the inductive formulation, the application of conformal prediction in the case of regression is still complicated, but there are ways to face this problem (Section 2 of the Supplementary Material (Fontana, Zeni and Vantini, 2023)). Lastly, what is historically called as "the randomness assumption": conformal prediction is valid if examples are sampled independently from a fixed but unknown probability distribution. It actually works also under the slightly weaker assumption that examples are probabilistically exchangeable, and under other online compression models, as the widely used Gaussian linear model (Section 3 of the Supplementary Material (Fontana, Zeni and Vantini, 2023))).

The last section (Section 3.3) addresses interesting directions of further development and research. We describe extensions of the framework that improve the interpretability and applicability of conformal inference. CP has been applied to a variety of applied tasks and problems. For this reason it is not possible to refer to all of them here: the interested reader can find an exhaustive list in Balasubramanian, Ho and Vovk (2014).

2. Foundations of conformal prediction

2.1. Conformal predictors

We will now show how the basic version of CP works. In the basic setting, successive values $z_1, z_2, z_3, \cdots \in \mathbf{Z}$, called *examples*, are observed. $\mathbf{Z}$ is a measurable space, called the examples space.

We also assume that $\mathbf{Z}$ contains more than one element, and that each singleton is measurable. Before the $(n+1)$th value z_{n+1} is announced, the training set[1] consists of $(z_1, \ldots, z_n)$ and our goal is to predict the new example.

To be precise, we are concerned with a prediction algorithm that outputs a set of elements of $\mathbf{Z}$, implicitly meant to contain z_{n+1}. Formally, a *prediction set* is a measurable function γ that maps a sequence $(z_1, \ldots, z_n) \in \mathbf{Z}^n$ to a set $\gamma(z_1, \ldots, z_n) \subseteq \mathbf{Z}$, where the measurability condition reads as follow: the set $\{(z_1, \ldots, z_{n+1}) : z_{n+1} \in \gamma(z_1, \ldots, z_n)\}$ is measurable in $\mathbf{Z}^{n+1}$. A trade-off between reliability and informativeness has to be faced by the algorithm while giving as output the prediction sets. Indeed giving as a prediction set the whole examples space $\mathbf{Z}$ is not appealing nor useful: it is absolutely reliable but not informative.

Rather than a single set predictor, we are going to deal with nested families of set predictors depending on a parameter $\alpha \in [0,1]$, the *significance level* or target miscoverage level, reflecting the required reliability of the prediction. The smaller α is, the bigger the reliability in our guess. So, the quantity $1 - \alpha$ is usually called the confidence level. As a consequence, we define a *confidence predictor* to be a nested family of set predictors (γ^α), such that, given α_1, α_2 and $0 \le \alpha_1 \le \alpha_2 \le 1$,

$$\gamma^{\alpha_1}(z_1, \ldots, z_n) \supseteq \gamma^{\alpha_2}(z_1, \ldots, z_n). \tag{2.1}$$

Confidence predictors from old examples alone, without knowing anything else about them, may seem relatively uninteresting. But the simplicity of the setting makes it advantageous to explain and understand the rationale of the conformal algorithm, and, as we will see, it is then straightforward to take into account also features related to the examples.

In the greatest part of the literature concerning conformal prediction, and especially in the works of Vovk and coauthors, the symbol ε stands for the significance level. Nonetheless, we prefer to adopt the symbol α, as in Lei, Robins and Wasserman (2013), to be faithful to the statistical tradition and its classical notation. For the same reason, we want to predict the $(n+1)$th example, relying on the previous experience given by $(z_1, \ldots, z_n)$, still like Lei et al. and conversely to Vovk, Gammerman and Shafer. The latter is interested in the nth value given the previous $(n-1)$ ones.

2.1.1. The randomness assumption

We will make two main kinds of assumptions about the way examples are generated. The standard assumption is the usual i.i.d. one commonly employed in the statistical setting: the examples we observe are sampled independently from some unknown probability distribution P on $\mathbf{Z}$. Equivalently, the infinite sequence $z_1, z_2, \ldots$ is drawn from the power probability distribution P^∞ in $\mathbf{Z}^\infty$.

Under the exchangeability assumption, instead, the sequence $(z_1, \ldots, z_n)$ is generated from a probability distribution that is exchangeable: for any permutation π of the set $\{1, \ldots, n\}$, the joint probability distribution of the permuted sequence $(z_{\pi(1)}, \ldots, z_{\pi(n)})$ is the same as the distribution of the original sequence. In an identical way, the $n!$ different orderings are equally likely. It is possible to extend the definition of exchangeability to the case of an infinite sequence of variables: $z_1, z_2, \ldots$ are exchangeable if $z_1, \ldots, z_N$ are exchangeable for every N.

Exchangeability implies that variables have the same distribution. On the other hand, exchangeable variables need not to be independent. It is immediately evident how the exchangeability assumption is weaker than the randomness one. As we will see in Section 2.3, in the online setting the difference between the two assumptions almost disappears. For further discussions about exchangeability, including various definitions, a game-theoretic approach and a law of large numbers, refer to Section 3 of Shafer and Vovk (2008).

---

[1] From a mathematical point of view it is a sequence, not a set.

The randomness assumption is a standard assumption in machine learning. Conformal prediction, however, usually requires only the sequence $(z_1, \ldots, z_n)$ to be exchangeable. In addition, other models which do not require exchangeability can also use conformal prediction (Section 3 of the Supplementary Material ()).

2.1.2. Bags and nonconformity measures

First, the concept of a nonconformity (or strangeness) measure has to be introduced. In few words, it estimates how unusual an example looks with respect to the previous ones. The order in which old examples $(z_1, \ldots, z_n)$ appear should not make any difference. To underline this point, we will use the term *bag* (in short, *B*) and the notation $\langle z_1, \ldots, z_n \rangle$. A bag is defined exactly as a multiset. Therefore, $\langle z_1, \ldots, z_n \rangle$ is the bag we get from $(z_1, \ldots, z_n)$ when we ignore which value comes first, which second, and so on. To increase clarity, we specifiy that the main difference between a bag/multiset and a set is that an element in a bag/multiset can be repeated.

As mentioned, a *nonconformity measure* (NCM) $A \colon \mathbf{Z}^n \times \mathbf{Z} \to \mathbb{R}$ is a way of scoring how different an example $z \in \mathbb{R}$ is from a bag $B \in \mathbf{Z}^n \times \mathbf{Z}$. There is not just one nonconformity measure. For instance, once the sequence of old examples $(z_1, \ldots, z_n)$ is at hand, a natural choice is to take the average as the simple predictor of the new example, and then compute the nonconformity score as the absolute value of the difference from the average. In more general terms, the distance from the central tendency of the bag might be considered. As pointed out in Vovk, Gammerman and Shafer (2005), whether a particular function A is an appropriate way of measuring nonconformity will always be open to discussion, as it greatly depends on contextual factors.

We have previously remarked that α represents our significance level. Now, for a given nonconformity measure A, we set $R = A(B, z)$ to stand for the nonconformity score — where R is related in a certain way to the word "residual". On the contrary, most of the literature uses ε and $\alpha = A(B, z)$, respectively. We still prefer Lei et al.'s notation.

Instead of a nonconformity measure, a conformity one might be chosen. The line of reasoning does not change at all: we could compute the scores and resume to the first framework just by changing the sign, or computing the inverse. However, conformity measures are not a common choice.

2.1.3. Conformal prediction

The idea behind conformal methods is extremely simple. Consider $n + 1$ exchangeable observations of a scalar random variable, let's say $u_1, \ldots, u_{n+1}$. In the absence of ties, the rank of the $n + 1$th observation u_{n+1} among $u_1, \ldots, u_{n+1}$ is uniformly distributed over the set $\{1, \ldots, n+1\}$, due to exchangeability.

Back to the nonconformity framework, under the assumption that the z_i are exchangeable, we define, for a given $z \in \mathbf{Z}$:

$$p_z := \frac{|\{i = 1, \ldots, n+1 : R_i \geq R_{n+1}\}|}{n+1} \tag{2.2}$$

where

$$
\begin{aligned}
R_i &= A\left(\langle z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, z \rangle, z_i\right) \\
&= A\left(\langle z_1, \ldots, z_n, z \rangle \setminus \langle z_i \rangle, z_i\right) \qquad \forall\, i = 1, \ldots, n
\end{aligned}
\tag{2.3}
$$

and

$$R_{n+1} := A\left(\langle z_1, \ldots, z_n \rangle, z\right). \tag{2.4}$$

It is straightforward that p_z stands for the fraction of examples that are as or more different from the all the others than z actually is. This fraction, which lies between $\frac{1}{n+1}$ and 1, is defined as the *p-value*

for z. If p_z is small, then z is very nonconforming with respect to the past experience, represented by $(z_1, \ldots, z_n)$. On the contrary, if large, then z is very conforming and likely to appear as the next observation. Hence, it is reasonable to include it in the prediction set.

As a result, we define the prediction set $\gamma^\alpha(z_1, \ldots, z_n)$ by including all the zs that conform with the previous examples. In a formula, $\gamma^\alpha(z_1, \ldots, z_n) := \{z \in \mathbf{Z} : p_z > \alpha\}$. To summarize, the algorithm tells us to form a prediction region consisting of all the zs that are not among the fraction α most out of place with respect to the bag of old examples. Shafer and Vovk (2008) give also a clear interpretation of $\gamma^\alpha(z_1, \ldots, z_n)$ as an application of the Neyman-Pearson theory for hypothesis testing and confidence intervals.

2.1.4. Validity and efficiency

The two main indicators of how good confidence predictors behave are validity and efficiency, respectively an index of reliability and informativeness. A set predictor γ is *exactly valid* at a significance level $\alpha \in [0, 1]$, if the probability of making an error — namely the event $z_{n+1} \notin \gamma^\alpha$ — is α, under any probability distribution on $\mathbf{Z}^{n+1}$. If the probability does not exceed α, under the same conditions, a set predictor is defined as conservatively valid. If the properties hold at each of the significance level α, the confidence predictor $(\gamma^\alpha : \alpha \in [0, 1])$ is respectively valid and conservatively valid. The following result, concerning conformal prediction, holds (Vovk, Gammerman and Shafer, 2005):

Proposition 2.1. *Under the exchangeability assumption, the probability of error, $z_{n+1} \notin \gamma^\alpha(z_1, \ldots, z_n)$, will not exceed α, for any α and any conformal predictor γ.*

In an intuitive way, due to exchangeability, the distribution of $(z_1, \ldots, z_{n+1})$ and so the distribution of the nonconformity scores $(R_1, \ldots, R_{n+1})$ are invariant under permutations; in particular, all permutations are equiprobable. This simple concept is the bulk of the proof and the key of conformal methods.

From a practical point of view, the conservativeness of the validity is often not ideal, especially when n is large, and so we get long-run frequency of errors very close to α. From a theoretical prospective, Lei et al. (2018) indeed prove, under minimal assumptions on the residuals, that conformal prediction intervals are accurate, meaning that they do not substantially over-cover.

A conformal predictor is always conservatively valid. Is it possible to achieve exact validity by introducting a randomized component into the algorithm. The *smoothed conformal predictor* is defined in the same way as before, except that the p-values (2.2) are replaced by the smoothed p-values:

$$p_z := \frac{|\{i : R_i > R_{n+1}\}| + \tau |\{i : R_i = R_{n+1}\}|}{n+1}, \tag{2.5}$$

where the *tie-breaking* random variable τ is uniformly distributed on $[0, 1]$ (τ can be the same for all zs). For a smoothed conformal predictor, as wished, the probability of a prediction error is exactly α (Vovk, Gammerman and Shafer (2005), Proposition 2.4).

Alongside validity, prediction algorithms should be efficient too, that is to say, the uncertainty related to predictions should be as small as possible. Validity is the priority: without it, the meaning of predictive regions is lost, and it becomes easy to achieve the best possible performance. Without constraints, in fact, the trivial $\gamma^\alpha(z_1, \ldots, z_{n-1}) := \varnothing$ is the most efficient one. Efficiency may appear as a vague notion, but in any case it can be meaningful only if we impose some restrictions on the predictors that we consider.

Among the main problems solved by Machine Learning and Statistics we can find two types of problems: classification, when predictions deal with a small finite set (often binary), and regression, when instead the real line is considered. In classification problems, two criteria for efficiency have been

used most often in literature. One criterion takes account of whether the prediction is a single class (the most efficient case), a multiple set of classes, or the entire set of possible classes (the most inefficient case) at a given significance level α. Alternatively, the confidence and *credibility* of the prediction — which do not depend on the choice of a significance level α — are considered. The former is the greatest $1 - \alpha$ for which γ^{α} is a single label, while the latter, helpful to avoid overconfidence when the object x is unusual, is the largest α for which the prediction set is empty. Vovk et al. (2016) show several other criteria, giving a detailed depiction of the framework. In regression problems instead, the prediction set is often an interval of values, and a natural measure of efficiency of such a prediction is simply the length of the interval. The smaller it is, the better its performance.

We will be looking for the most efficient confidence predictors in the class of valid, or in an equivalent term well-calibrated, confidence predictors; different notions of validity (including conditional validity, examined in Section 3.1) and different formalisations of the notion of efficiency will lead to different solutions to the problem.

2.2. Objects and labels

In this section, we introduce a generalization of the basic CP setting. A sequence of successive examples $z_1, z_2, z_3, \ldots$ is still observed, but each example consists of an *object* x_i and its *label* y_i, i.e $z_i = (x_i, y_i)$. The objects are elements of a measurable space $\mathbf{X}$ called the object space, and the labels of a measurable space $\mathbf{Y}$ called the label space (both in the classification and the regression contexts). As before, we take for granted that $|\mathbf{Y}| > 1$. In a more compact way, let z_i stand for (x_i, y_i), and $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ be the example space.

At the $(n + 1)$th trial, the object x_{n+1} is given, and we are interested in predicting its label y_{n+1}. The general scheme of reasoning is unchanged. Under the randomness assumption, examples, i.e. (x_i, y_i) couples, are assumed to be i.i.d. First, we need to choose a nonconformity measure in order to compute nonconformity scores. Then, p-values are computed, too. Last, the prediction set Γ^{α} turns out to be defined as follow:

$$\Gamma^{\alpha}(z_1, \ldots, z_n, x_{n+1}) := \{y : (x_{n+1}, y) \in \Gamma^{\alpha}(z_1, \ldots, z_n)\}$$
$$= \{y \in \mathbf{Y} : p^{(x_{n+1}, y)} > \alpha\} \qquad (2.6)$$

In most cases, the way to proceed, when defining how much a new example conforms with the bag B of old examples, is relying on a simple predictor f. The only condition to hold is that f must be invariant to permutations in its arguments — equivalently, the output does not depend on the order in which they are presented. The method f defines a prediction rule. It is natural then to measure the nonconformity of z by looking at the deviation of the predicted label $\hat{y}_i = f_{\langle z_1, \ldots, z_n \rangle}(x_i)$ from the true one. For instance, in regression problems, we can just take the absolute value of the difference between $\hat{y}_i$ and y_i. That's exactly what we have suggested in the previous (unstructured) case (Section 2.1), when we proposed to take the mean or the median as the simple predictor for the next observation. Nevertheless, there are instances in which other NCMs may be more powerful and/or useful: a notable example is Multi-Level Conformal Clustering, and the use of the k-NN distance (Nouretdinov et al., 2020) or the use of density estimators in Lei, Rinaldo and Wasserman (2015).

Following these steps any simple predictor, combined with a suitable measure of deviation of $\hat{y}_i$ from y_i, leads to a nonconformity measure and, therefore, to a conformal predictor. The algorithm will always produce valid nested prediction regions. But the prediction regions will be efficient (i.e. small) only if $A(B, z)$ measures well how different z is from the examples in B. And consequently only if the underlying algorithm is appropriate. Conformal prediction ends up to be a powerful meta-algorithm, created on top of any point predictor — very powerful but yet extremely simple in its rationale.

A useful remark in Shafer and Vovk (2008) points out that the prediction regions produced by the conformal algorithm do not change when the nonconformity measure A is transformed monotonically. For instance, if A is positive, choosing A or its square A^2 will make no difference. While comparing the scores to compute p_z, indeed, the interest is on the relative values and their reciprocal position — whether one is bigger than another or not, but not on the single absolute values. As a result, the choice of the deviation measure is relatively unimportant. In this framework, the important step in determining the nonconformity measure in this setting is choosing the point predictor f.

2.2.1. Classification

In the broader literature, CP has been proposed and implemented with different nonconformity measures for classification — i.e, when $|\mathbf{Y}| < \infty$. As an illustration, given the sequence of old examples $(x_1, y_1), \ldots, (x_n, y_n)$ representing past experience, nonconformity scores R_i can be computed as follow:

$$R_i := \frac{\min_{j=1,\ldots,n\,:\,j\neq i\,\&\,y_j=y_i}\ \Delta(x_i, x_j)}{\min_{j=1,\ldots,n\,:\,j\neq i}\ \Delta(x_i, x_j)} \tag{2.7}$$

where Δ is a metric on $\mathbf{X}$, usually the Euclidean distance in an Euclidean setting. The rationale behind the scores (2.7) — in the spirit of the 1-nearest neighbor algorithm — is that an example is considered nonconforming to the sequence if it is close to examples labeled in a different way and far from the ones with the same label. In a different way, we could use a nonconformity measure that takes account of the average values for the different labels, and the score R_i is simply the distance to the average of its label.

As an alternative, nonconformity scores can be extracted from the support vector machines trained on $(z_1, \ldots, z_n)$. We consider in particular the case of binary classification, as the first works actually did to face this problem (Gammerman, Vovk and Vapnik, 1998, Saunders, Gammerman and Vovk, 1999), but there are also ways to adapt it to solve multi-label classification problems (Balasubramanian, Ho and Vovk, 2014). A plain approach is defining nonconformity scores as the values of the Lagrange multipliers, that stand somehow for the margins of the probability estimating model. If an example's true class is not clearly separable from other classes, then its score R_i is higher and, as desired, we tend to classify it as strange.

Another example of nonconformity measure for classification problems is Devetyarov and Nouretdinov (2010), who rely on random forests. For instance, a random forest is constructed from the data sequence, and the conformity score of an example z_i is just equal to the percentage of correct predictions for its features x_i given by decision trees.

2.2.2. Regression

In regression problems, a very natural nonconformity measure is:

$$R_i := \Delta(y_i, f(x_i)) \tag{2.8}$$

where Δ is a measure of difference between two labels (usually a metric) and f is a prediction rule (for predicting the label given the object) trained on the set $(z_1, \ldots, z_n)$.

It is evident how there is a fundamental problem in implementing conformal prediction for regression tasks: to form the prediction set (2.6), examining each potential label y is needed. Nonetheless, there is often a feasible way to compute (2.6) which does not require to examine infinitely many cases; in particular, this happens when the underlying simple predictor is ridge regression or nearest neighbors regression. We are going to provide a sketch of how it works, to give an idea of the way used to circumvent the unfeasible brute-force, testing-all approach. Besides, a slightly different approach to conformal

prediction has been developed and carried on to overcome this difficulty (Section 3.2, Section 2 of the Supplementary Material (Fontana, Zeni and Vantini, 2023)).

In the case where $\Delta(y_i, \hat{y}_i) = |y_i - \hat{y}_i|$ and f is the ridge regression procedure, the conformal predictor is called the ridge regression confidence machine (RRCM). The initial attempts to apply conformal prediction in the case of regression involve exactly ridge regression (Melluish, Vovk and Gammerman (1999), and soon after, in a more in-depth version, Nouretdinov, Melluish and Vovk (2001)). Suppose that objects are vectors consisting of d attributes in a Euclidean space, say $\mathbf{X} \subseteq \mathbb{R}^d$, and let λ^2 be the non-negative constant called the ridge parameter — least squares is the special case corresponding to $\lambda = 0$ The explicit representation, in matrix form, of this nonconformity measure is:

$$R_i := |y_i - x_i'(X'X + \lambda I)^{-1} X'Y|, \tag{2.9}$$

where X is the $n + 1 \times d$ object matrix whose rows are $x_1', x_2', \ldots, x_{n+1}'$, Y is the label vector $(y_1, \ldots, y_{n+1})'$, I is the unit $d \times d$ matrix. Hence, the vector of nonconformity scores $(R_1, \ldots, R_{n+1})'$ can be written as $|Y - HY| = |(I - H)Y|$, where H is the hat matrix.

Let $y = y_{n+1}$ be a possible label for x_{n+1}, and $(z_1, \ldots, z_n, (x_{n+1}, y))$ the augmented data set. Now, $Y := (y_1, \ldots, y_n, y)'$. Note that $Y = (y_1, \ldots, y_n, 0)' + (0, \ldots, 0, y)'$ and so the vector of nonconformity scores can be represented as $|A + By|$, where: $A = (I - H)(y_1, \ldots, y_n, 0)'$ and $B = (I - H)(0, \ldots, 0, y)'$. Therefore, each $R_i = R_i(y)$ has a linear dependence on y. As a consequence, since the p-value $p_z(y)$ simply counts how many scores R_i are greater than R_{n+1}, it can only change at points where $R_i(y) - R_{n+1}(y)$ changes sign for some $i = 1, \ldots, n$. This means that we can calculate the set of points y on the real line whose corresponding p-value $p_z(y)$ exceeds α rather than trying all possible y, leading to a feasible prediction. Precise computations can be found in Vovk, Gammerman and Shafer (2005), chap 2.

Before going on in the discussion, a clarification is required. The reader has surely noticed that the proposed non-conformity measure is computed on the i-th example too, contrary to what has been stated before. In fact, the statement of the conformal algorithm, we define the nonconformity score for the ith example by: $R_i := A(\langle z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, z \rangle, z_i)$ (2.3), apparently specifying that we do not want to include z_i in the bag to which it is compared. But then, in the RRCM, we use the nonconformity scores (2.9), as if: $R_i := A(\langle z_1, \ldots, z_n, z \rangle, z_i)$.

The point is whether to include the new example in the bag with which we are comparing it or not is a delicate question that is for instance tackled in Shafer and Vovk (2008). First of all, it is noteworthy to assert that both of them are valid choices. They both indeed satisfy the symmetry property required for valid non-conformity measures. In some cases (e.g., when there is a monotonic relation between $A(\{z_1, \ldots, z_{i-1}, z_i, z_{i+1}, z_n, z\}, z_i)$ and $A(\{z_1, \ldots, z_{i-1}, z_{i+1}, z_n, z\}, z_i)$) they can even lead to the same prediction sets. For example, if R_i is the absolute value of the difference between z_i and the mean value of the bag B, including or not z_i in the bag is absolute invariant, as in both cases Simple computations show that the two scores are the same, except for a scale factor $\frac{n}{n+1}$.

From an efficiency viewpoint, one may correctly wonder if z_i should be included in the training set or not in order to maximize the efficiency. To our knowledge there are no works dealing in detail on this issue. Even though the answer could possibly depend on the specific non-conformity measure, underlying model, and data distribution, we believe that this issue could and should be an object of further future investigation.

We have introduced conformal prediction with the formula (2.3), as the reference book of Vovk, Gammerman and Shafer (2005) and the first works did. Moreover, in this form conformal prediction generalizes to online compression models (Section 3 of the Supplementary Material (Fontana, Zeni and Vantini, 2023)). In general, however, the inclusion of the ith example simplifies the implementation or

---

[2] λ can be selected in a non-adaptive fashion, or an in adaptive one, exploiting techniques such as cross-validation.

at least the explanation of the conformal algorithm. From now on, we rely on this approach when using conformal prediction, and define instead the methods relying on (2.3) as jackknife procedures.

Conformal predictors can be implemented in a feasible and at the same time particularly simple way for nonconformity measures based on the nearest neighbors algorithm, too. Recently, an efficient method to compute in an exact way conformal prediction with the Lasso, i.e. considering the quadratic loss function and the l_1 norm penalty, has been provided by Lei (2019). A straight extension to the elastic net — which considers both a l_1 and l_2 penalty, is also given.

Recently, novel methods to perform regression, that follow a different paradigm to compute NCMs have been proposed: we provide a discussion about them in Section 3.1

2.3. The online framework

The first works on Conformal algorithms tended to focus on the online framework, where examples arrive one by one and so predictions are based on an accumulating data set. The predictions these algorithms make are *hedged*: they incorporate a valid indication of their own accuracy and reliability. Vovk, Gammerman and Shafer (2005) claim that most existing algorithms for hedged prediction first learn from a training data set and then predict without ever learning again. The few algorithms that do learn and predict simultaneously, instead, do not provide confidence information.

Moreover, the property of validity of conformal predictors can be stated in an especially strong form in the online framework. Classically, a method for finding $(1 - \alpha)$ prediction regions is considered valid if it has a $(1 - \alpha)$ probability of containing the label predicted, because by the law of large numbers it would then be correct $(1 - \alpha)\%$ of the times when repeatedly applied to independent data sets. The online story may seem more complicated, because we repeatedly apply a method not to independent data sets, but to an accumulating data set. After using $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and x_{n+1} to predict y_{n+1}, we use $(x_1, y_1), \ldots, (x_{n+1}, y_{n+1})$ and x_{n+2} to predict y_{n+2}, and so on. For a $(1 - \alpha)$ online method to be valid, $(1 - \alpha)\%$ of these predictions must be correct. Under minimal assumptions, conformal prediction is valid in this new and powerful sense.

The intermediate step behind this result is that successive errors are probabilistically independent. Let us start with some definitions, following Vovk (2002) and Vovk, Gammerman and Shafer (2005): being $\omega = (z_1, z_2, \ldots) = (x_1, y_1, x_2, y_2, \ldots$ an infinite data sequence and Γ^α a generic randomised conformal predictor, we define the number of errors made by Γ^α on the infinite sequence ω during the first n trials to be

$$Err_n(\Gamma^\alpha, \omega) := \#\{i = 1, \ldots, n : y_i \notin \Gamma^\alpha(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i)\} \tag{2.10}$$

where $\#B$ is the size of set B. The individual prediction result is defined instead as

$$err_n(\Gamma^\alpha, \omega) := Err_n - Err_{n-1} = \begin{cases} 1 & y_n \notin \Gamma^\alpha(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n) \\ 0 & otherwise \end{cases} \tag{2.11}$$

and the whole infinite sequence of prediction results is then defined as

$$err(\Gamma^\alpha, \omega) := (err_1(\Gamma^\alpha, \omega), err_2(\Gamma^\alpha, \omega), \ldots) \tag{2.12}$$

It is shown for the first time in Vovk (2002) that, being and Γ^α a generic randomised conformal predictor of level $1 - \alpha$, the following holds:

Proposition 2.2. *For any randomised conformal predictor* Γ *of any confidence level* $1 - \alpha$, *and any exchangeable probability distribution P in* $\mathbf{Z}^\infty$, *the image of* $P \times U^\infty$ *under the mapping*

$$(\omega \in \mathbf{Z}^\infty, \tau \in [0,1]^\infty) \rightarrow err(\Gamma^\alpha, \omega, \tau)$$

is the probability distribution B_α^∞ *of independent Bernoulli random trial of parameter* α.

In a more intuitive sense, Proposition 2.2 states, under the exchangeability assumption and in the online mode, that the errors made at different steps by randomised conformal predictors are independent. The experiment involved in predicting z_{101} from $z_1, \ldots, z_{100}$ is not entirely independent of the experiment involved in predicting, say, z_{105} from $z_1, \ldots, z_{104}$. The 101 random numbers involved in the first experiment are all also involved in the second one, but this overlap does not actually matter.

Summing up, we already know that the probability of error is below the significance level α. In addition to that, events for successive n are probabilistically independent notwithstanding the overlap. Hence, $(1 - \alpha)\%$ of consecutive predictions must be correct. In other words, the random variables $\mathbb{1}_{z_{n+1} \notin \gamma^\alpha(z_1, \ldots, z_n)}$ are independent Bernoulli variables with parameter α. Vovk, Nouretdinov and Gammerman (2009) focuses on the prediction of consecutive responses, especially when the number of observations does not exceed the number of parameters.

It should be noted that the assumption of exchangeability rather than randomness makes Proposition 2.2 stronger: it is very easy to give examples of exchangeable distributions on $\mathbf{Z}^N$ that are not of the form P^N — where it is worth recalling that P is the unknown distribution of examples. Nonetheless, in the infinite-horizon case (which is the standard setting for the online mode of prediction) the difference between the exchangeability and randomness assumptions essentially disappears: according to a well-known theorem by de Finetti, each exchangeable probability distribution on $\mathbf{Z}^\infty$ is a mixture of power probability distributions P^∞, provided $\mathbf{Z}$ is a Borel space (Hewitt and Savage, 1955). In particular, using the assumption of randomness rather than exchangeability in the case of the infinite sequence hardly weakens it: the two forms are equivalent when $\mathbf{Z}$ is a Borel space.

3. Recent advances in conformal prediction

3.1. Different notions of validity

An appealing property of conformal predictors is their automatic validity under the exchangeability assumption:

$$\mathbb{P}(Y_{n+1} \in \Gamma^\alpha(Z_1, \ldots, Z_n, X_{n+1})) \geq 1 - \alpha \qquad \text{for all } P, \tag{3.1}$$

where $\mathbb{P} = P^{n+1}$ is the joint measure of $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$. A major focus of this section will be on conditional versions of the notion of validity.

The idea of conditional inference in statistics is about the wish to make conclusions that are as much conditional on the available information as possible. Although finite sample coverage defined in (3.1) is a desirable property, this might not be enough to guarantee good prediction bands, even in very simple cases. We refer to (3.1) as *marginal* coverage, which is different from (in fact, weaker than) the *conditional* coverage as usually sought in prediction problems. As a result, a good estimator must satisfy something more than marginal coverage. A natural criterion would be conditional coverage.

However, distribution-free *conditional coverage*, that is:

$$\mathbb{P}(Y_{n+1} \in \Gamma^\alpha(x) \mid X_{n+1} = x) \geq 1 - \alpha \qquad \text{for all } P \text{ and a.a } x, \tag{3.2}$$

with $\Gamma^\alpha(x) \equiv \Gamma^\alpha(Z_1, \ldots, Z_n, x)$ is impossible to achieve with a finite sample for rich object spaces, such as $\mathbf{X} = \mathbb{R}$ (Lei and Wasserman (2014), Lemma 1). Indeed, the requirement of precise object conditional validity cannot be satisfied in a nontrivial way, unless we know the true probability distribution generating the data (or we are willing to use a subjective or postulated probability distribution, as in Bayesian theory), or unless the test object is an atom of the data-generating distribution. If we impose that requirement, the prediction interval is expected to have infinite length (Vovk (2013a) and for general background related to distribution-free inference Bahadur and Savage (1956), Donoho (1988)).

The negative result — that conditional coverage cannot be achieved by finite-length prediction intervals without regularity and consistency assumptions on the model and the estimator f — does not prevent set predictors to be (object) conditionally valid in a partial and asymptotic sense, and simultaneously asymptotically efficient.

Therefore, as an alternative solution, Lei and Wasserman (2014) develop a new notion, called *local validity*, that naturally interpolates between marginal and conditional validity, and is achievable in the finite sample case. Formally, given a partition $\mathcal{A} = \{A_j : j \geq 1\}$ of supp(P_X), a prediction band Γ^α is locally valid with respect to $\mathcal{A}$ if:

$$\mathbb{P}(Y_{n+1} \in \Gamma^\alpha(X_{n+1}) \mid X_{n+1} \in A_j) \geq 1 - \alpha \qquad \text{for all } j \text{ and all } P. \tag{3.3}$$

Then, their work is focused on defining a method that shows both finite sample (marginal and local) coverage and asymptotic conditional coverage (i.e., when the sample size goes to ∞, the prediction band give arbitrarily accurate conditional coverage). At the same time, they prove it to be asymptotic efficient. The finite sample marginal and local validity is distribution free: no assumptions on P are required. Then, under mild regularity conditions, local validity implies asymptotically conditionally validity.

The way Lei and Wasserman (2014) built the prediction bands to achieve local validity can be seen as a particular case of a bigger class of predictors, which now we introduce and explain, the so called Mondrian conformal predictors. The discussion about validity is definetely not closed: recently Barber et al. (2021) reflect again on the idea of a proper intermediate definition, while Romano, Patterson and Candès (2019) proposes the so called Conformalised Quantile Regression (CQR), a novel method that, by fusing a Conformal way of thinking with quantile regression is able to yield valid distribution-free bands, yet remarkably capable to adapt to any heteroschedasticity of the data and with increased efficiency with respect to standard proposals, as thoroughly shown by the simulation studies in the article.

As in split conformal prediction, the procedure is based on splitting the data into a proper training set, indexed by $\mathcal{I}_1$ and a calibration set indexed by $\mathcal{I}_2$. Given a quantile regression method A, two conditional quantiles $\hat{q}_{\alpha_{hi}}$ and $\hat{q}_{\alpha_{lo}}$ are fitted using the proper training set:

$$\{\hat{q}_{\alpha_{hi}}, \hat{q}_{\alpha_{lo}}\} \longleftarrow A(\{(X_i, Y_i : i \in \mathcal{I}_\infty\}) \tag{3.4}$$

The non-conformity scores are then calculated as:

$$E_i := max \{\hat{q}_{\alpha_{lo}}(X_i) - Y_i, Y_i - \hat{q}_{\alpha_{hi}}(X_i) - Y_i\} \tag{3.5}$$

Finally, given a new input data X_{n+1} the prediction interval for Y_{n+1} is.

$$C(X_{n+1}) = [\hat{q}_{\alpha_{lo}}(X_{n+1}) - Q_{1-\alpha}(E, \mathcal{I}_2), \hat{q}_{\alpha_{hi}}(X_{n+1}) + Q_{1-\alpha}(E, \mathcal{I}_2)] \tag{3.6}$$

where $Q_{1-\alpha}(E, \mathcal{I}_2) := (1-\alpha)(1 + 1/|\mathcal{I}_2|)$-th empirical quantile of $E_i, i \in \mathcal{I}_2$ Using the same line of reasoning, an extension of this context to the calculation of conditionally valid confidence (with respect to prediction) intervals for parameters has been very recently proposed in Medarametla and Candès (2021)

3.1.1. *Mondrian conformal predictors*

We start from an example. In handwritten digit recognition problems, some digits (such as "5") are more difficult to recognize correctly than other digits (such as "0"), and it is natural to expect that at the confidence level 95% the error rate will be significantly greater than 5% for the difficult digits; our usual, unconditional, notion of validity only ensures that the average error rate over all digits will be close to 5%.

We might not be satisfied by the way the conformal predictors work. If our set predictor is valid at the significance level 5% but makes an error with probability 10% for men and 0% for women, both men and women can be unhappy with calling 5% the probability of error. It is clear that whenever the size of the training set is sufficient for making conditional claims, we should aim for this. The requirement of object conditional validity is a little bit more than what we can ask a predictor to be, but it can be considered as a special case: for somehow important events E we do not want the conditional probability of error given E to be very different from the given significance level α.

We are going to deal with a natural division of examples into several *categories*: e.g., different categories can correspond to different labels, or kinds of objects, or just be determined by the ordinal number of the example. As pointed out in the examples above, conformal predictors — as we have seen so far — do not guarantee validity within categories: the fraction of errors can be much larger than the nominal significance level for some categories, if this is compensated by a smaller fraction of errors for other categories. A stronger kind of validity, validity within categories, which is especially relevant in the situation of asymmetric classification, is the main property of *Mondrian conformal predictors* (MCPs), first introduced in Vovk et al. (2003). The exchangeable framework is the assumption under which MCPs are proved to be valid; in Section 3 of the Supplementary Material (Fontana, Zeni and Vantini, 2023), again, we will have a more general setting, relaxing the hypothesis.

When the term categories comes into play, we are referring to a given division of the example space $\mathbf{Z}$: a measurable function κ maps each z to its category k, belonging to the (usually finite) measurable space $\mathbf{K}$ of all categories. In many instances, it is a kind of classification of z_i. The category $\kappa_i = \kappa(z_i)$ might depend on the other examples in the data sequence $(z_1, \ldots, z_n)$, but disregarding their order. Such a function κ is called a Mondrian taxonomy, as a tribute to the Dutch painter Piet Mondrian. Indeed, the taxonomy that κ defines in the space $\mathbf{Z}$ recalls the grid-based paintings and the style for which the artist is renowned.

To underline the dependence of $\kappa(z_i)$ on the bag of the entire dataset, Balasubramanian, Ho and Vovk (2014) introduce the n-taxonomy $K : \mathbf{Z}^n \Rightarrow \mathbf{K}^n$, which maps a vector of examples to the vector of corresponding categories. Using this notation, it is required that the n-taxonomy K is equivariant with respect to permutations, that is:

$$(\kappa_1, \ldots, \kappa_n) = K(z_1, \ldots, z_n) \Rightarrow (\kappa_{\pi(1)}, \ldots, \kappa_{\pi(n)}) = K(z_{\pi(1)}, \ldots, z_{\pi(n)}).$$

We prefer however to let the dependence implicit and remain stuck to the simpler notation of Vovk, Gammerman and Shafer (2005).

Given a Mondrian taxonomy κ, to use conformal prediction we have to modify slightly some of the definitions seen in the previous chapter. To be precise, a Mondrian nonconformity measure might take into account also the categories $\kappa, \ldots, \kappa_n$, while the p-values (2.2) should be computed as:

$$p_z := \frac{|\{i = 1, \ldots, n+1 : \kappa_i = \kappa_{n+1} \,\&\, R_i \geq R_{n+1}\}|}{|\{i = 1, \ldots, n+1 : \kappa_i = \kappa_{n+1}\}|}, \tag{3.7}$$

where $\kappa_{n+1} = \kappa(z)$. As a remark, we would like to point out and stress what we are exactly doing in the formula just defined. Although one can choose any conformity measure, in order to have local

validity the ranking must be based on a local subset of the sample. Hence, the algorithm selects only the examples among the past experience that have the same category of the new one, and makes its decision based on them.

At this point, the reader is able to write by himself the smoothed version of the MCP, which satisfies the required level of reliability in an exact way.

Proposition 3.1. *If examples $z_1, \ldots, z_{n+1}$ are generated from an exchangeable probability distribution on $\mathbf{Z}^{n+1}$, any smoothed MCP based on a Mondrian taxonomy κ is category-wise exact with respect to κ.*

In the previous proposition, category-wise exact means that $\forall n$, the probability of p_{n+1} conditional to $k(1, z_1), p_1, \ldots, k(n, z_n), p_n$ is a uniform on $[0, 1]$ where $z_1, \ldots, z_n, z_{n+1}$ are generated from an exchangeable distribution Z^∞ and $p_1, \ldots, p_n$ are the p-values generated by f (Vovk, Gammerman and Shafer, 2005, Section 4.5).

Moreover, we might want to have different significance levels α_k for different categories k. In some contexts, certain kinds of errors are more costly than others. For example, it may be more costly to classify a high-risk credit applicant as low risk (one kind of error) than it is to classify a low-risk applicant as high risk (a different kind of error). In an analogous way, we could be required to distinguish between useful messages and spam in the problem of mail filtering: classifying a useful message as spam is a more serious error than vice versa. We do not have misclassification costs to take into account, but setting in a proper way the significance levels allow us to specify the relative importance of different kinds of prediction errors. And MCPs still do the job (Vovk, Gammerman and Shafer, 2005).

Last, a brief discussion of an important question: how to select a good taxonomy? While choosing the partitions that determine a Mondrian taxonomy κ, it comes out indeed a dilemma that is often called the "problem of the reference class". We want the categories into which we divide the examples to be large, in order to have a reasonable sample size for estimating the probabilities. But we also want them to be small and homogeneous, to make the inferences as specific as possible. Balasubramanian, Ho and Vovk (2014) points out a possible strategy for conditional conformal predictors in the problem of classification in the online setting. The idea is to adapt the method as the process goes on. At first, the conformal predictor should not be conditional at all. Then, as the number of examples grows, it should be label conditional. As the number of examples grows further, we could split the objects into clusters (using a label independent taxonomy) and make the prediction sets conditional on them as well.

3.2. Inductive prediction

While being very straightforward from a mathematical point of view, Transductive (or "Full") Conformal predictors are very computationally intensive and, over time, an extensive literature has developed to address this issue. In particular, *inductive conformal predictors* (ICPs), also referred as *split conformal predictors* have been proposed

ICPs were first proposed by Papadopoulos et al. (2002) for regression and by Papadopoulos, Vovk and Gammerman (2002) for classification, and in the online setting by Vovk (2002). Before the appearance of inductive conformal predictors, several other possibilities had been studied, but not with great success. To speed computations up in a multi-class pattern recognition problem which uses support vector machines in its implementation, Saunders, Gammerman and Vovk (2000) used a hashing function to split the training set into smaller subsets, of roughly equal size, which are then used to construct a number of support vector machines. In a different way, just to mention but a few, Ho and Wechsler (2004) exploit the adiabatic version of incremental support vector machine, and lately Vovk (2013b)
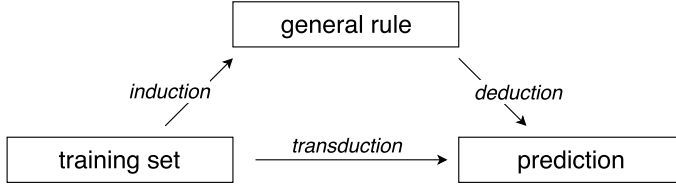
Figure 1. Inductive and transductive approach to prediction.

introduces Bonferroni predictors, a simple modification based on the idea of the Bonferroni adjustment of p-values.

We now spend some words to recall the concepts of transduction and induction (figure 1), as introduced in Vapnik (1998). In inductive prediction we first move from the training data to some general rule: a prediction or decision rule, a model, or a theory (inductive step). When a new object comes out, we derive a prediction based on the general rule (deductive step). On the contrary, in transductive prediction, we take a shortcut, going directly from the old examples to the prediction for the new object. The practical distinction between them is whether we extract the general rule or not. A side-effect of using a transductive method is computational inefficiency; computations need to be started from scratch every time.

Combining the inductive approach with conformal prediction, the data sequence $(z_1, \ldots, z_n)$ is *split* into two parts, the proper training set $(z_1, \ldots, z_m)$ of size $m < n$ and the *calibration set* $(z_{m+1}, \ldots, z_n)$. We use the proper training set to feed the underlying algorithm, and, using the derived rule, we compute the non-conformity scores for each example in the calibration set. For every potential label y of the new unlabelled object x_{n+1}, its score R_{n+1} is calculated and is compared to the ones of the calibration set. In more mathematical terms, we define the nonconformity scores as:

$$R_j := A_{m_k+1} \left(\langle (x_1, y_1), \ldots, (x_{m_k}, y_{m_k}) \rangle, (x_j, y_j) \right), \, for \, j = m_k + 1, \ldots, n-1$$
$$R_{n+1} := A_{m_k+1} \left(\langle (x_1, y_1), \ldots, (x_{m_k}, y_{m_k}) \rangle, (x_{n+1}, y) \right), \tag{3.8}$$

Therefore the p-value is:

$$p_z := \frac{|\{ i = m+1, \ldots, n+1 : R_i \geq R_{n+1} \}|}{n - m + 1}. \tag{3.9}$$

Inductive conformal predictors can be smoothed in exactly the same way as conformal predictors. As in the transductive approach, under the exchangeability assumption, p_z is a valid p-value. All is working as before. For a discussion of conditional validity and various ways to achieve it using inductive conformal predictors, see Vovk (2013a).

A greater computational efficiency of inductive conformal predictors is now evident. The computational overhead of ICPs is light: they are almost as efficient as the underlying algorithm. The decision rule is computed from the proper training set only once, and it is applied to the calibration set also only once. Several studies related to this fact are reported in the literature. For instance, a computational complexity analysis can be found in the work of Papadopoulos (2008), where conformal prediction on top of neural networks for classification has been closely examined.

With such a dramatically reduced computation cost, it is possible to combine easily conformal algorithms with computationally heavy estimators. While validity is taken for granted in conformal framework, efficiency is related to the underlying algorithm. Taking advantage of the bargain ICPs represent, we can compensate the savings in computational terms and, in metaphor, invest a lot of resources in the choice of f.

Moreover, this computational effectiveness can be exploited further and fix conformal prediction as a tool in Big Data frameworks, where the increasing size of datasets represents a challenge for machine learning and statistics. The inductive approach makes the task feasible, but can we ask for anything more? Actually, the (trivially parallelizable) serial code might be run on multiple CPUs. Capuccini et al. (2015) propose and analyze a parallel implementation of the conformal algorithm, where multiple processors are employed simultaneously in the Apache Spark framework.

Achieving computational efficiency does not come for free. A drawback of inductive conformal predictors is their potential prediction inefficiency. In actual fact, we waste the calibration set when developing the prediction rule f, and we do not use the proper training set when computing the p-values. An interesting attempt to cure this disadvantage is made in Vovk (2015). *Cross-conformal prediction*, a hybrid of the methods of inductive conformal prediction and cross-validation, consists, in a nutshell, in dividing the data sequence into K folds, constructing a separate ICP using the kth fold as the calibration set and the rest of the training set as the proper training set. Then the different p-values, which are the outcome of the procedure, are merged in a proper way.

Of course, it is also possible to use a uneven split, using a larger portion of data for model fitting and a smaller set for the inference step. This will produce sharper prediction intervals, but the method will have higher variance; this trade-off is unavoidable for data splitting methods. Common choices found in the applied literature for the dimension of the calibration set, providing a good balance between underlying model performance and calibration accuracy, lie between 25% and 33% of the dataset. The problem related to how many examples the calibration set should contain is faced meticulously in Linusson et al. (2014b). To maximize the efficiency of inductive conformal classifiers, they suggest to keep it small relative to the amount of available data (approximately $15\% - 30\%$ of the total). At the same time, at least a few hundred examples should be used for calibration (to make it granular enough), unless this leaves too few examples in the proper training set. Techniques that try to handle the problems associated with small calibration sets are suggested and evaluated in both Johansson et al. (2015) and Carlsson et al. (2015), using interpolation of calibration instances and a different notion of (approximate) p-value, respectively.

Splitting improves dramatically on the speed of conformal inference, but it introduces additional noise into the procedure. One way to reduce this extra randomness is to combine inferences from N several splits, each of them — using a Bonferroni-type argument — built at level $1 - \alpha/N$. Multiple splitting on one hand decreases the variability as expected, but on the other hand this may produce, as a side effect, the width of Γ_N^α to grow with N. As described in Shafer and Vovk (2008), under rather general conditions, the Bonferroni effect is dominant and hence intervals get larger and larger with N. For this reason, they suggest using a single split.

Linusson et al. (2014b) even raise doubts about the commonly accepted claim that transductive conformal predictors are by default more efficient than inductive ones. It is known indeed that an unstable nonconformity function — one that is heavily influenced by an outlier example, e.g., an erroneously labeled new example (x_{n+1}, y) — can cause (transductive) conformal confidence predictors to become inefficient. They compare the efficiency of transductive and inductive conformal classifiers using decision tree, random forest and support vector machine models as the underlying algorithm, to find out that the full approach is not always the most efficient. Their position is actually the same of Papadopoulos (2008), where the loss of accuracy introduced by induction is claimed to be small, and usually negligible. And not only for large data sets, which clearly contain enough training examples so that the removal of the calibration examples does not make any difference to the training of the algorithm. Some further explorations on the topic can be found in Linusson et al. (2014a) but, to our knowledge, the question remains still open.

From another perspective, lying between the computational complexities of the full and split conformal methods is *jackknife prediction*. This method wish to make a better use of the training data than

the split approach does and to cure as much as possible the connected loss of informational efficiency, when constructing the absolute residuals, due to the partition of old examples into two parts, without resorting at the same time to the extensive computations of the full conformal prediction. With this intention, it uses leave-one-out residuals to define prediction intervals. That is to say, for each example z_i it trains a model f_{-i} on the rest of the data sequence $(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$ and computes the nonconformity score R_i with respect to f_{-i}.

The advantage of the jackknife method over the split conformal method is that it can often produce regions of smaller size. However, in regression problems it is not guaranteed to have valid coverage in finite samples. As Lei et al. (2018) observe, the jackknife method has the finite sample in-sample coverage property:

$$\mathbb{P}(Y_i \in \Gamma^\alpha_{jack}(X_i)) \geq 1 - \alpha, \tag{3.10}$$

where $\Gamma^\alpha_{jack}(X_i)$ is the Jackknife prediction set. When dealing with out-of-sample coverage (actually, true predictive inference), the properties of the Jackknife are much more fragile. In fact, even asymptotically, its coverage properties do not hold without requiring nontrivial conditions on the base estimator f. It is actually due to the approximation required to avoid the unfeasible enumeration approach, that we are going to tackle in a while, precisely in the next section. The predictive accuracy of the jackknife under assumptions of algorithm stability is explored by Steinberger and Leeb (2016) for the linear regression setting, and in a more general setting by Steinberger and Leeb (2018). Hence, while the full and split conformal intervals are valid under essentially no assumptions, the same is not true for the jackknife ones.

The key to speed up the learning process is to employ a fast and accurate learning method as the underlying algorithm. This is exactly what Wang, Wang and Shi (2018) do, proposing a novel, fast and efficient conformal regressor, with combines the local-weighted (see Section 3.3) jackknife prediction, and the regularized extreme learning machine. Extreme learning machine (ELM) addresses the task of training feed-forward neural networks fast without losing learning ability and predicting performance. The underlying learning process and the outstanding learning ability of ELM make the conformal regressor very fast and informationally efficient.

Recently, a slight but crucial and remarkable modification to the algorithm gives life to the jackknife+ methods, able to restore rigorous coverage guarantees (Barber et al., 2019a). We expect many possible extension of these methods to stem from this seminal piece of research.

3.3. Other interesting developments

Full conformal and split conformal methods, combined with basically any fitting procedure in regression, provide finite sample distribution-free predictive inference. We are now going to introduce generalizations and further explorations of the possibilities of CP along different directions.

In the pure online setting, we get an immediate feedback (the true label) for every example that we predict. While this scenario is convenient for theoretical studies, in practice, however, rarely one immediately gets the true label for every object. On the contrary weak teachers are allowed to provide the true label with a delay or sometimes not to provide it at all. In this case, we have to accept a weaker (actually, an asymptotic) notion of validity, but conformal confidence predictors adapt and keep at it (Nouretdinov and Vovk, 2006, Ryabko, Vovk and Gammerman, 2003).

Moreover, we may want something more than just providing p-values associated with the various labels to which a new observation could belong. We might be interested in the problem of probability forecasting: we observe n pairs of objects and labels, and after observing the $(n + 1)$th object x_{n+1}, the goal is to give a probability distribution p_{n+1} for its label. It represents clearly a more challenging

task (Vovk, Gammerman and Shafer (2005), chap 5), therefore a suitable method is necessary to handle carefully the reliability-resolution trade-off. A class of algorithms called *Venn predictors* (Vovk, Shafer and Nouretdinov, 2004) satisfies the criterion for validity when the label space is finite, while only among recent developments there are adaptations in the context of regression, i.e. with continuous labels — namely Nouretdinov et al. (2018) and in a different way, following the work of Shen, Liu and Xie (2018), Vovk et al. (2019). For many underlying algorithms, Venn predictors (like conformal methods in general) are computationally inefficient. Therefore Lambrou et al. (2012), and as an extension Lambrou, Nouretdinov and Papadopoulos (2015), combine Venn predictors and the inductive approach, while Vovk et al. (2018) introduce cross-conformal predictive systems.

Online compression models is not the only framework where CP does not require examples to be exchangeable. Dunn and Wasserman (2018) extend the conformal method to construct valid distribution-free prediction sets when there are random effects, and Barber et al. (2019b) to handle weighted exchangeable data, as in the setting of covariate shift (Chen et al., 2016b, Shimodaira, 2000). Dashevskiy and Luo (2011) robustify the conformal inference method by extending its validity to settings with dependent data. They indeed propose an interesting blocking procedure for times series data, whose theoretical performance guarantees are provided in Chernozhukov, Wuthrich and Zhu (2018).

Now, we describe more in detail a couple of other recent advances, while advances on High-Dimensional Regression in a conformal setting are described in Section 4 of the Supplementary Material (Fontana, Zeni and Vantini, 2023).

3.3.1. Normalized nonconformity scores

In conformal algorithms seen so far, the width of $\Gamma^\alpha(x)$ is roughly immune to x (figure 2, left). This property is desirable if the spread of the residual $Y - \hat{Y}$, where $\hat{Y} = f(X)$, does not vary substantially as X varies. However, in some scenarios this will not be true, and we wish conformal bands to adapt correspondingly. Actually, it is possible to have individual bounds for the new example which take into account the difficulty of predicting a certain y_{n+1}. The rationale for this, from a conformal prediction standpoint, is that if two examples have the same nonconformity scores using (2.8), but one is expected
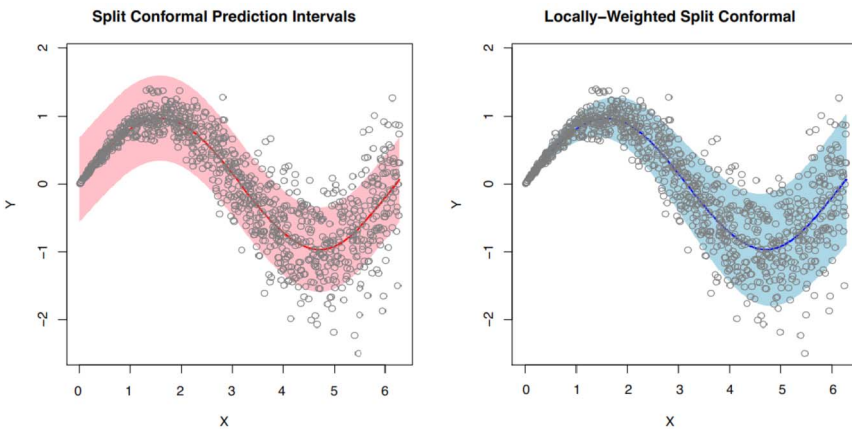

Figure 2. Conformal predictors do not contemplate heteroskedasticity in the data distribution. In such a case, one would expect the length of the output interval to be an increasing function of the corresponding variance of the output value, which can give more information of the target label. To tackle this problem, local-weighted conformal inference has been introduced. *Source: Lei et al. (2018)*.

to be more accurate than the other, then the former is actually stranger (more nonconforming) than the latter. We are interested in resulting prediction intervals that are smaller for objects that are deemed easy to predict and larger for harder objects.

To reach the goal, normalized nonconformity functions come into play (figure 2, right), that is:

$$R_i = \frac{|\hat{y}_i - y_i|}{\sigma_i}, \tag{3.11}$$

where the absolute error concerning the ith example is scaled using the expected accuracy σ_i of the underlying model; see, e.g., Papadopoulos and Haralambous (2011), and Papadopoulos, Vovk and Gammerman (2011). Choosing Equation (3.11), the confidence predictor (Equation 2.1 in the Supplementary Material (Fontana, Zeni and Vantini, 2023)) becomes:

$$\Gamma^\alpha(x_{n+1}) = [f(x_{n+1}) - R_s\,\sigma_{n+1}, f(x_{n+1}) + R_s\,\sigma_{n+1}]. \tag{3.12}$$

As a consequence, the resulting predictive regions tend to be tighter than those produced by the simple conformal methods. Anyway, using locally-weighted residuals, as in (3.11), the validity and accuracy properties of the conformal methods, both finite sample and asymptotic, again carry over.

As said, σ_i is an estimate of the difficulty of predicting the label y_i. There is a wide choice of estimates of the accuracy available in the literature. A common practice is to train another model to predict errors, as in Papadopoulos and Haralambous (2010). More in details, once f has been trained and the residual errors computed, a different model g is fit using the object $x_1, \ldots, x_n$ and the residuals. Then, σ_i could be set equal to $g(x_i) + \beta$, where β is a sensitivity parameter that regulates the impact of normalization.

Other approaches use, in a more direct way, properties of the underlying model f; for instance, it is the case of Papadopoulos, Gammerman and Vovk (2008). In the paper, they consider conformal prediction with k-NNR method, which computes the weighted average of the k nearest examples, and as a measure of expected accuracy they simply use the distance of the examined example from its nearest neighbours. Namely,

$$d_i^k = \sum_{j=1}^{k} \text{distance}\,(x_i, x_j). \tag{3.13}$$

The nearer an example is to its neighbours, the more accurate this prediction is indeed expected to be.

3.3.2. Functional prediction bands

Functional Data Analysis (FDA) is a branch of statistics that analyses data that exist over a continuous domain, broadly speaking functions. Functional data are intrinsically infinite dimensional. This is a rich source of information, which brings many opportunities for research and data analysis — a powerful modeling tool. Meanwhile the high or infinite dimensional structure of the data, however, poses challenges both for theory and computations. Therefore, FDA has been the focus of much research efforts in the statistics and machine learning community in the last decade.

There are few publications in the conformal prediction literature that deal with functional data. We are going to give just some details about a simple scenario that could be reasonably typical. In the following, the work of Lei, Rinaldo and Wasserman (2015) guide us. The sequence $z_1(\cdot), \ldots, z_n(\cdot)$ consists now of $L^2[0,1]$ functions. The definition of validity for a confidence predictor γ^α is:

$$\mathbb{P}\,(z_{n+1}(t) \in \gamma^\alpha(z_1, \ldots, z_n)(t)\ \forall t) \geq 1 - \alpha \qquad \text{for all } P. \tag{3.14}$$

Then, as always, to apply conformal prediction, a nonconformity measure is needed. A fair choice might be:

$$R_i = \int (z_i(t) - \bar{z}(t))^2 dt, \tag{3.15}$$

where $\bar{z}(t)$ is the average of the augmented data set. Due to the dimension of the problem, an inductive approach is more desirable. Therefore, once the nonconformity scores R_i are computed for the example functions of the calibration set, the conformal prediction set is given by all the functions z whose score is smaller than the suitable quantile R_s.

Then, one more step is mandatory. Given a conformal prediction set γ^α, the inherent prediction bands are defined in terms of lower and upper bounds:

$$l(t) = \inf_{z \in \gamma^\alpha} z(t) \qquad \text{and} \qquad u(t) = \sup_{z \in \gamma^\alpha} z(t). \tag{3.16}$$

Consequently, thanks to provable conformal properties,

$$\mathbb{P}\left(l(t) \le z_{n+1}(t) \le u(t), \ \forall t \right) \ge 1 - \alpha. \tag{3.17}$$

However, γ^α could contain very disparate elements, hence no close form for $l(t)$ and $u(t)$ is available in general and these bounds may be hard to compute.

To sum up, the key features to be able to handle functional data efficiently are the nonconformity measure and a proper way to make use of the prediction set in order to extract useful information. The question is still an open challenge, but the topic stands out as a natural way for conformal prediction to grow up and face bigger problems.

An intermediate work in this sense is Lei, Rinaldo and Wasserman (2015), which studies prediction and visualization of functional data paying specific attention to finite sample guarantees. As far as we know, it is the only analysis up to now that applies conformal prediction to the functional setting. In particular, their focal point is exploratory analysis, exploiting conformal techniques to compute clustering trees and simultaneous prediction bands — that is, for a given level of confidence $1 - \alpha$, the bands that covers a random curve drawn from the underlying process (as in (3.14)).

However, satisfying (this formulation of) validity could be really a tough task in the functional setting. Since their focus is on the main structural features of the curve, they lower the bar and set the concept in a revised form, that is:

$$\mathbb{P}(\Pi(z_{n+1})(t) \in \gamma^\alpha(z_1, \ldots, z_n)(t) \ \forall t) \ge 1 - \alpha \qquad \text{for all } P, \tag{3.18}$$

where Π is a mapping into a finite dimensional function space $\Omega_p \subseteq L^2[0,1]$.

The prediction bands they propose are constructed, as shown previously in (3.18), adopting a finite dimensional projection approach. Once a basis of functions $\{\phi_1, \ldots, \phi_p\}$ is chosen — let it be a fixed one, like the Fourier basis, or a data-driven basis, such as functional principal components — the vector of projection coefficients ξ_i is computed for each of the m examples in the proper training set. Then, the scores R_i measure how different the projection coefficients are with respect to the ones of the training set, that is, for the ith calibration example, $R_i = A(\xi_1, \ldots, \xi_m; \xi_i)$. Let:

$$\gamma_\xi = \{\xi \in \mathbb{R}^p : R_\xi \le R_s\} \tag{3.19}$$

and

$$\gamma^\alpha(t) = \left\{ \sum_{i=1}^p \varsigma_i \phi_i(t) : (\varsigma_1, \ldots, \varsigma_p) \in \gamma_\xi \right\}. \tag{3.20}$$

As a consequence, γ^α is valid, i.e. (3.18) holds.

Exploiting the finite dimensional projection, the conformity measure handles vectors, so all the experience seen in these two chapters gives a hand. A density estimator indeed is usually selected to assess conformity. Nevertheless, picking A out is critical in the sense that a not suitable one may give a lot of trouble in computing $\gamma^\alpha(t)$. It is the case, for instance, of kernel density estimators. In their work, the first p elements of the eigenbasis — i.e. the eigenfunctions of the autocovariance operator — constitute the basis, while A is a Gaussian mixture density estimator. In this set up, approximations are available, and lead to the results they obtain.

Though their work can be deemed as remarkable, the way used to proceed simplifies a lot the scenario. It is a step forward in order to extend conformal prediction to functional data, but not a complete solution. Very promising results in extending and overcoming the approach in Lei, Rinaldo and Wasserman (2015), can be found in Diquigiovanni, Fontana and Vantini (2021b), where the authors propose a family of novel conformal predictors for functional data that yield closed-form simultaneous and finite-sample valid prediction bands, as well as proposing an approach to modulate the amplitude of the bands in relationship to the problem at hand. The same authors moreover propose an extension of such approach to a multivariate functional data setting and in a regression framework in Diquigiovanni, Fontana and Vantini (2021a). Moving from Chernozhukov, Wuthrich and Zhu (2018) it is also proposed, in Diquigiovanni, Fontana and Vantini (2021c), an extension of the proposed method to a functional time-series setting, with an application to the modelling of the gas market.

Acknowledgements

Funding

Supplementary Material

Supplementary Material to "Conformal prediction: A unified review of theory and new challenges" (DOI: 10.3150/21-BEJ1447SUPP; .pdf). Additional Sections and Miscellaneous Topics.

References

Bahadur, R.R. and Savage, L.J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Stat.* **27** 1115–1122. MR0084241 https://doi.org/10.1214/aoms/1177728077

Balasubramanian, V., Ho, S.-S. and Vovk, V. (2014). *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. London: Newnes.

Barber, R.F., Candes, E.J., Ramdas, A. and Tibshirani, R.J. (2019a). Predictive inference with the jackknife+. arXiv preprint. Available at arXiv:1904.06019.

Barber, R.F., Candes, E.J., Ramdas, A. and Tibshirani, R.J. (2019b). Conformal prediction under covariate shift. arXiv preprint. Available at arXiv:1904.06019.

Barber, R.F., Candès, E.J., Ramdas, A. and Tibshirani, R.J. (2021). The limits of distribution-free conditional predictive inference. *Inf. Inference* **10** 455–482. MR4270755 https://doi.org/10.1093/imaiai/iaaa017

Burnaev, E. and Vovk, V. (2014). Efficiency of conformalized ridge regression. In *Conference on Learning Theory* 605–622.

Capuccini, M., Carlsson, L., Norinder, U. and Spjuth, O. (2015). Conformal prediction in Spark: Large-scale machine learning with confidence. In *2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)* 61–67. IEEE.

Carlsson, L., Ahlberg, E., Boström, H., Johansson, U. and Linusson, H. (2015). Modifications to p-values of conformal predictors. In *International Symposium on Statistical Learning and Data Sciences* 251–259. Springer.

Chen, W., Chun, K.-J. and Barber, R.F. (2018). Discretized conformal prediction for efficient distribution-free inference. *Stat* **7** e173, 11. MR3769053 https://doi.org/10.1002/sta4.173

Chen, W., Wang, Z., Ha, W. and Barber, R.F. (2016a). Trimmed conformal prediction for high-dimensional models. arXiv preprint. Available at arXiv:1611.09933.

Chen, X., Monfort, M., Liu, A. and Ziebart, B.D. (2016b). Robust covariate shift regression. In *Artificial Intelligence and Statistics* 1270–1279.

Chernozhukov, V., Wuthrich, K. and Zhu, Y. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. arXiv preprint. Available at arXiv:1802.06300.

Dashevskiy, M. and Luo, Z. (2011). Time series prediction with performance guarantee. *IET Commun.* **5** 1044–1051.

Devetyarov, D. and Nouretdinov, I. (2010). Prediction with confidence based on a random forest classifier. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* 37–44. Springer.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–67. With a discussion and a rejoinder by the authors. MR0829555 https://doi.org/10.1214/aos/1176349830

Diquigiovanni, J., Fontana, M. and Vantini, S. (2021a). Conformal prediction bands for multivariate functional data. *J. Multivariate Anal.* in press.

Diquigiovanni, J., Fontana, M. and Vantini, S. (2021b). The importance of being a band: finite-sample exact distribution-free prediction sets for functional data. arXiv preprint. Available at arXiv:2102.06746.

Diquigiovanni, J., Fontana, M. and Vantini, S. (2021c). Distribution-free prediction bands for multivariate functional time series: an application to the Italian gas market. arXiv preprint. Available at arXiv:2107.00527.

Donoho, D.L. (1988). One-sided inference about functionals of a density. *Ann. Statist.* **16** 1390–1420. MR0964930 https://doi.org/10.1214/aos/1176351045

Dunn, R. and Wasserman, L. (2018). Distribution-free prediction sets with random effects. arXiv preprint. Available at arXiv:1809.07441.

Fontana, M., Zeni, G. and Vantini, S. (2023). Supplement to "Conformal prediction: A unified review of theory and new challenges." https://doi.org/10.3150/21-BEJ1447SUPP

Gammerman, A., Vovk, V. and Vapnik, V. (1998). Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* 148–155. Morgan Kaufmann Publishers Inc.

Hebiri, M. (2010). Sparse conformal predictors. *Stat. Comput.* **20** 253–266. MR2610776 https://doi.org/10.1007/s11222-009-9167-2

Hewitt, E. and Savage, L.J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.* **80** 470–501. MR0076206 https://doi.org/10.2307/1992999

Ho, S.-S. and Wechsler, H. (2004). Learning from data streams via online transduction. Ma et al, 45–52.

Johansson, U., Boström, H., Löfström, T. and Linusson, H. (2014). Regression conformal prediction with random forests. *Mach. Learn.* **97** 155–176. MR3252831 https://doi.org/10.1007/s10994-014-5453-0

Johansson, U., Ahlberg, E., Boström, H., Carlsson, L., Linusson, H. and Sönströd, C. (2015). Handling small calibration sets in Mondrian inductive conformal regressors. In *International Symposium on Statistical Learning and Data Sciences* 271–280. Springer.

Lambrou, A., Nouretdinov, I. and Papadopoulos, H. (2015). Inductive Venn Prediction. *Ann. Math. Artif. Intell.* **74** 181–201. MR3353902 https://doi.org/10.1007/s10472-014-9420-z

Lambrou, A., Papadopoulos, H., Nouretdinov, I. and Gammerman, A. (2012). Reliable probability estimates based on support vector machines for large multiclass datasets. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* 182–191. Springer.

Lei, J. (2019). Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika* **106** 749–764. MR4031197 https://doi.org/10.1093/biomet/asz046

Lei, J., Rinaldo, A. and Wasserman, L. (2015). A conformal prediction approach to explore functional data. *Ann. Math. Artif. Intell.* **74** 29–43. MR3353895 https://doi.org/10.1007/s10472-013-9366-6

Lei, J., Robins, J. and Wasserman, L. (2013). Distribution-free prediction sets. *J. Amer. Statist. Assoc.* **108** 278–287. MR3174619 https://doi.org/10.1080/01621459.2012.751873

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 71–96. MR3153934 https://doi.org/10.1111/rssb.12021

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J. and Wasserman, L. (2018). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* **113** 1094–1111. MR3862342 https://doi.org/10.1080/01621459.2017.1307116

Linusson, H., Norinder, U., Boström, H., Johansson, U. and Löfström, T. (2014a). On the calibration of aggregated conformal predictors. *Proc. Mach. Learn. Res.* **60** 261–270.

Linusson, H., Johansson, U., Boström, H. and Löfström, T. (2014b). Efficiency comparison of unstable transductive and inductive conformal classifiers. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* 261–270. Springer.

Medarametla, D. and Candès, E. (2021). Distribution-free conditional median inference. *Electron. J. Stat.* **15** 4625–4658. MR4316665 https://doi.org/10.1214/21-ejs1910

Melluish, T., Vovk, V. and Gammerman, A. (1999). Transduction for regression estimation with confidence. In *Neural Information Processing Systems, NIPS'99*.

Melluish, T., Saunders, C., Nouretdinov, I. and Vovk, V. (2001). Comparing the Bayes and typicalness frameworks. In *European Conference on Machine Learning* 360–371. Springer.

Nouretdinov, I., Melluish, T. and Vovk, V. (2001). Ridge regression confidence machine. In *ICML* 385–392.

Nouretdinov, I. and Vovk, V. (2006). Criterion of calibration for transductive confidence machine with limited feedback. *Theoret. Comput. Sci.* **364** 3–9. MR2268298 https://doi.org/10.1016/j.tcs.2006.07.037

Nouretdinov, I., Vovk, V., Vyugin, M. and Gammerman, A. (2001). Pattern recognition and density estimation under the general i.i.d. assumption. In *Computational Learning Theory (Amsterdam, 2001). Lecture Notes in Computer Science* **2111** 337–353. Berlin: Springer. MR2042045 https://doi.org/10.1007/3-540-44581-1_22

Nouretdinov, I., Volkhonskiy, D., Lim, P., Toccaceli, P. and Gammerman, A. (2018). Inductive Venn-Abers predictive distribution. *Proc. Mach. Learn. Res.* **91** 1–22.

Nouretdinov, I., Gammerman, J., Fontana, M. and Rehal, D. (2020). Multi-level conformal clustering: A distribution-free technique for clustering and anomaly detection. *Neurocomputing* **397** 279–291.

Papadopoulos, H. (2008). Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence*. InTech.

Papadopoulos, H., Gammerman, A. and Vovk, V. (2008). Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)* 64–69.

Papadopoulos, H. and Haralambous, H. (2010). Neural networks regression inductive conformal predictor and its application to total electron content prediction. In *International Conference on Artificial Neural Networks* 32–41. Springer.

Papadopoulos, H. and Haralambous, H. (2011). Reliable prediction intervals with regression neural networks. *Neural Netw.* **24** 842–851.

Papadopoulos, H., Vovk, V. and Gammerman, A. (2002). Qualified prediction for large data sets in the case of pattern recognition. In *ICMLA* 159–163. MR2805257

Papadopoulos, H., Vovk, V. and Gammerman, A. (2011). Regression conformal prediction with nearest neighbours. *J. Artificial Intelligence Res.* **40** 815–840. MR2805257 https://doi.org/10.1613/jair.3198

Papadopoulos, H., Proedrou, K., Vovk, V. and Gammerman, A. (2002). Inductive confidence machines for regression. In *Machine Learning: ECML 2002. Lecture Notes in Computer Science* **2430** 345–356. Berlin: Springer. MR2050303 https://doi.org/10.1007/3-540-36755-1_29

Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2168993

Riabko, D. (2005). On the flexibility of theoretical models for pattern recognition, Ph.D. thesis, Citeseer.

Romano, Y., Patterson, E. and Candès, E.J. (2019). Conformalized quantile regression. arXiv preprint. Available at arXiv:1905.03222.

Ryabko, D., Vovk, V. and Gammerman, A. (2003). Online region prediction with real teachers. Submitted for publication. Criterion of Calibration for Transductive Confidence Machine, 267.

Saunders, C., Gammerman, A. and Vovk, V. (1999). Transduction with confidence and credibility. In *Proceedings of the International Joint Conference on Artificial Intelligence* **2** 722–726.

Saunders, C., Gammerman, A. and Vovk, V. (2000). Computationally efficient transductive machines. In *International Conference on Algorithmic Learning Theory* 325–337. Springer.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.* **9** 371–421. MR2417240

Shen, J., Liu, R.Y. and Xie, M. (2018). Prediction with confidence—a general framework for predictive inference. *J. Statist. Plann. Inference* **195** 126–140. MR3760843 https://doi.org/10.1016/j.jspi.2017.09.012

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference* **90** 227–244. MR1795598 https://doi.org/10.1016/S0378-3758(00)00115-4

Steinberger, L. and Leeb, H. (2016). Leave-one-out prediction intervals in linear regression models with many variables. arXiv preprint. Available at arXiv:1602.05801.

Steinberger, L. and Leeb, H. (2018). Conditional predictive inference for high-dimensional stable algorithms. arXiv preprint. Available at arXiv:1809.01412.

Valiant, L.G. (1984). A theory of the learnable. *Commun. ACM* **27** 1134–1142.

Vapnik, V.N. (1998). *Statistical Learning Theory*. *Adaptive and Learning Systems for Signal Processing, Communications, and Control*. New York: Wiley. MR1641250

Vovk, V. (2002). Online confidence machines are well-calibrated. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on* 187–196. IEEE.

Vovk, V. (2013a). Conditional validity of inductive conformal predictors. *Mach. Learn.* **92** 349–376. MR3080332 https://doi.org/10.1007/s10994-013-5355-6

Vovk, V. (2013b). Transductive conformal predictors. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* 348–360. Springer.

Vovk, V. (2015). Cross-conformal predictors. *Ann. Math. Artif. Intell.* **74** 9–28. MR3353894 https://doi.org/10.1007/s10472-013-9368-4

Vovk, V., Gammerman, A. and Shafer, G. (2005). *Algorithmic Learning in a Random World*. New York: Springer. MR2161220

Vovk, V., Nouretdinov, I. and Gammerman, A. (2009). On-line predictive linear regression. *Ann. Statist.* **37** 1566–1590. MR2509084 https://doi.org/10.1214/08-AOS622

Vovk, V., Shafer, G. and Nouretdinov, I. (2004). Self-calibrating probability forecasting. In *Advances in Neural Information Processing Systems* 1133–1140.

Vovk, V., Lindsay, D., Nouretdinov, I. and Gammerman, A. (2003). Mondrian confidence machine. Technical report.

Vovk, V., Fedorova, V., Nouretdinov, I. and Gammerman, A. (2016). Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications* 23–39. Springer International Publishing.

Vovk, V., Nouretdinov, I., Manokhin, V. and Gammerman, A. (2018). Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications* 37–51.

Vovk, V., Shen, J., Manokhin, V. and Xie, M. (2019). Nonparametric predictive distributions based on conformal prediction. *Mach. Learn.* **108** 445–474. MR3917055 https://doi.org/10.1007/s10994-018-5755-8

Wang, D., Wang, P. and Shi, J. (2018). A fast and efficient conformal regressor with regularized extreme learning machine. *Neurocomputing* **304** 1–11.