MOX-Report No. 22/2020

Conformal Prediction: a Unified Review of Theory and New Challenges

Zeni, G.; Fontana, M.; Vantini, F.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Conformal Prediction: a Unified Review of Theory and New Challenges

GIANLUCA ZENI[1,*] MATTEO FONTANA[1,2,**] and SIMONE VANTINI[1,†]

[1] *MOX-Department of Mathematics, Politecnico di Milano, Italy*
E-mail: [*] gianluca.zeni@polimi.it; [†] simone.vantini@polimi.it

[2] *Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Italy E-mail:* [**] matteo.fontana@polimi.it

Abstract

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 predictions sets that are valid in a statistical sense also in in the finite sample case. The in-depth 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

At the beginning of the third millennium, a new method of prediction with confidence, called Conformal Prediction (CP), was introduced and developed. It allows to produce prediction sets with the guaranteed error rate, exclusively under the simple i.i.d. assumption of the sample. Reliable estimation of prediction confidence is a significant challenge in both machine learning and statistics, and the promising results generated by CP have resulted in further extensions of the original conformal framework. The increasing amount of real-world problems where robust predictions are needed had yielded a plethora of new articles 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 et al. (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 problem or a regression

problem. 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 milestone in the related literature is the book entitled Algorithmic learning in a random world, written by Vovk et al. (2005). It explains thoroughly all the theoretical fundamentals, and it was published in 2005. There is only another work that gives an overview on the topic, a more recent one actually, namely the book Conformal prediction for reliable machine learning, by Balasubramanian et al. (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 to give a picture of recent theoretical breakthroughs. Moreover, there is a great deal of inconsistencies in the extensive literature that has been developed regarding notation. 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, where it was initially proposed, 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 2.3 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. Section 2.4 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 partially object conditional validity. 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 3.3). Lastly, the randomness assumption: confor-

mal 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 3.4).

The last section (Section 3.5) 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 here to refer to all of them: the interested reader can find an exhaustive selection in Balasubramanian et al. (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 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 $0 \leq \alpha_1 \leq \alpha_1 \leq 1$,

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

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.

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

In the greatest part of the literature concerning conformal prediction, from the beginning and the very first works of Vovk et al. (1998), the symbol ε stands for the significance level. Nonetheless, we prefer to adopt the symbol α, as in Lei et al. (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 $(z_1, \dots, z_n)$, still like Lei et al. and conversely to Vovk et al.. 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 randomness one (to be clear, the usual i.i.d. assumption 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 $z_1, z_2, \dots$ is drawn from the power probability distribution P^∞ in $\mathbf{Z}^\infty$.

Under the exchangeability assumption, instead, the sequence $(z_1, \dots, z_n)$ is generated from a probability distribution that is exchangeable: for any permutation π of the set $\{1, \dots, n\}$, the joint probability distribution of the permuted sequence $(z_{\pi(1)}, \dots, z_{\pi(n)})$ 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: $z_1, z_2, \dots$ are exchangeable if $z_1, \dots, z_N$ 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 much weaker than the randomness one. As we will see in Section 2.4, in the online setting the difference between the two assumptions almost disappears. For further discussion about exchangeability, including various definitions, a game-theoretic approach and a law of large numbers, refer to Section 3 of Shafer and Vovk (2008).

The randomness assumption is a standard assumption in machine learning. Conformal prediction, however, usually requires only the sequence $(z_1, \dots, z_n)$ to be exchangeable. In addition, other models which do not require exchangeability can also use conformal prediction (Section 3.4).

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 $(z_1, \dots, z_n)$ appear should not make any difference. To underline this point, we will use the term *bag* (in short, B) and the notation $\wr z_1, \dots, z_n \wr$. A bag is defined exactly as a multiset. Therefore, $\wr z_1, \dots, z_n \wr$ is the bag we get from $(z_1, \dots, z_n)$ when we ignore which value comes first, which second, and so on.

As mentioned, a *nonconformity measure* $A(B, z) \colon \mathbf{Z}^n \times \mathbf{Z} \to \mathbb{R}$ is a way of scoring how different an example z is from a bag B. There is not just one nonconformity measure. For instance, once the sequence of old examples $(z_1, \dots, z_n)$ 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 et al. (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 miscoverage 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 i.i.d. (or even exchangeable) observations of a scalar random variable, let's say $u_1, \ldots, u_n$. The rank of another i.i.d. 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 $z \in \mathbf{Z}$:

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

where

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

and

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

It is straightforward that p_z stands for the fraction of examples that are 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, $\gamma^\alpha(z_1, \ldots, z_n) \coloneqq \{z \in \mathbf{Z} : p_z > \alpha\}$. 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 et al., 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. Therefore, the coverage of conformal intervals is highly concentrated around $1 - \alpha$.

A conformal predictor is always conservatively valid. Is it possible to achieve exact validity, in some way? Adding a bit of randomization into the algorithm, actually, it is. 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 et al. (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 restrictions, indeed, the trivial $\gamma^\alpha(z_1, \ldots, z_{n-1}) := \emptyset$ 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 singleton (the ideal case), multiple (an inefficient prediction), or empty (a superefficient prediction) 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 formalizations 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 $\mathbf{Z} \coloneqq \mathbf{X} \times \mathbf{Y}$ 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:

$$\begin{aligned} \Gamma^\alpha(z_1, \ldots, z_n, x_{n+1}) &\coloneqq \{y : (x_{n+1}, y) \in \Gamma^\alpha(z_1, \ldots, z_n)\} \\ &\coloneqq \{y \in \mathbf{Y} : p^{(x_{n+1}, y)} > \alpha\} \end{aligned} \tag{2.6}$$

In most cases, the way to proceed, when defining how much a new example is conform 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 $\hat{y}_i = f_{\wr z_1, \ldots, z_n \wr}(x_i)$ 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.

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. The really crucial step in determining the nonconformity measure, again, 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 et al., 1998; Saunders et al., 1999), but there are also ways to adapt it to solve multi-label classification problems (Balasubramanian et al., 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:

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

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 3.3).

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 et al. (1999), and soon after, in a much better version, Nouretdinov et al. (2001a)). Suppose that objects are vectors consisting of d attributes in a Euclidean space, say $\mathbf{X} \subseteq \mathbb{R}^d$, and let λ 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:

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

where X is the $n \times d$ object matrix whose rows are $x_1', x_2', \ldots, x_n'$, Y is the label vector $(y_1, \ldots, y_n)'$, I is the unit $d \times d$ matrix. Hence, the vector of nonconformity scores $(R_1, \ldots, R_n)'$ can be written as $|Y - HY| = |(I - H)Y|$, where H is the hat matrix.
Let y be a possible label for x_{n+1}, and $(z_1, \ldots, z_n, (x_{n+1}, y))$ the augmented data set. Now, $Y \coloneqq (y_1, \ldots, y_n, y)'$. 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 et al. (2005), chap 2.

Before going on in the discussion, a clarification is required. The point is whether to include the new example in the bag with which we are comparing it or not — a delicate question, that Shafer and Vovk (2008) do not overlook in their precise work. In the statement of the conformal algorithm, we define the nonconformity score for the ith example by: $R_i \coloneqq A(\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, z\}, z_i)$ (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: $R_i \coloneqq A(\{z_1, \ldots, z_n, z\}, z_i)$. First of all, it is noteworthy to assert that both of them are valid. That's the most important thing. Moreover, the two ways of defining nonconformity scores are equivalent, to the extent that whatever we can get with one of them we can get from the other by changing the nonconformity measure. 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 equivalent. Simple computations show that the two scores are the same, except for a scale factor $\frac{n}{n+1}$. But we know that conformal prediction makes no difference for a monotone transformation of the scores. It does not indeed change the prediction regions. Analogous result holds, in regression problems, when the distance from the least square line or some other regression line is chosen.
There are cases where (2.3) might be more convenient, and cases where not. We have introduced conformal prediction with the formula (2.3), as the reference book of Vovk et al. (2005) and the first works did. Moreover, in this form conformal prediction generalizes to online compression models (Section 3.4). In general, however, the inclusion of the ith example simplifies the implementation or 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 (2017). A straight extension to the elastic net — which considers both a l_1 and l_2 penalty, is also given.

2.3. Novelty of Conformal Prediction

The problem of prediction sets is well studied in the context of linear regression, where they are usually constructed under linear and Gaussian assumptions. The Gaussian assumption can be relaxed by using, for example, quantile regression. These linear-model-based methods usually have reasonable finite sample performance. However, the coverage is valid only when the regression model is correctly specified. In contrast, non-parametric methods have the potential to work for any smooth distribution, but only asymptotic results are available and the finite sample behaviour remains unclear. To sum up, none of these methods yields prediction bands with distribution-free, finite sample validity. Furthermore, the output is a prediction set in the form of an interval, which may not be optimal to catch the structure of the data (figure 1). Conformal prediction instead is a general approach to construct valid and distribution-free prediction sets (and sequentially, in the online setting).

There are two other areas in statistics and machine learning that produce some kind of confidence information — a guarantee of the prediction error: the Bayesian framework and the theory of Probably Approximately Correct learning — PAC theory, in short (Valiant, 1984). Specifically, the Bayesian framework is able to complement individual

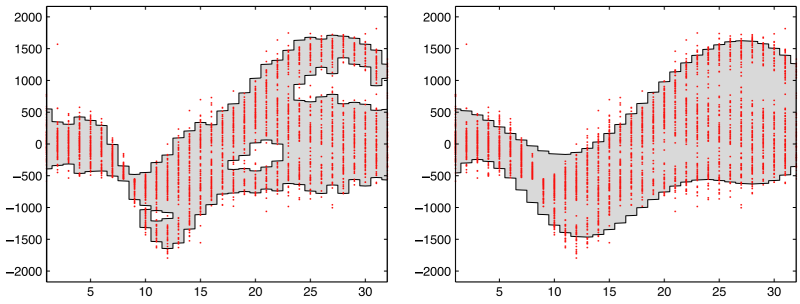

Figure 1. A comparison between conformal prediction bands, on the left, and quantile regression bands, on the right, for a selected confidence level $1 - \alpha = 0.9$. There are clear gaps in the data, indicating that the high density regions of the conditional density of Y given X are not connected. The quantile regression approach obscures these features. *Source: Lei and Wasserman (2014).*

predictions with probabilistic measures of their quality. These measures are, however, based on some a priori assumption about the underlying distribution. Burnaev and Vovk (2014) show that when the (artificial) data set satisfies the prior, the intervals produced are valid, and slightly tighter than the corresponding intervals produced by conformal algorithms. The problem is that for real-world data, the required knowledge is typically not available and as a result, one is forced to assume the existence of some arbitrarily chosen prior. In this case, since the assumed prior is most probably violated, the outputs of Bayesian methods may become quite misleading, due to the loss of validity (Melluish et al., 2001).

If we measure the efficiency of a prediction interval by its length, we can see that there is a certain dualism between Bayes and conformal prediction intervals: as the Bayesian assumption becomes less and less satisfied, the Bayes prediction intervals lose their validity while maintaining their efficiency, and, on the contrary, the conformal ones lose their efficiency while maintaining their validity. However, validity is more important than efficiency. Hence, if we believe the examples to be generated by a certain model, then we may want to use a nonconformity measure based on a method of prediction that is optimal for that model. This will be efficient if the proposed model is right, but valid in any case. Conformal prediction only assumes exchangeability. In the extreme case, paradoxically, even a function that returns a random nonconformity score (like $\text{rand}(0, 1)$) for all examples will be valid, but the prediction regions will be very wide. The dependence of the validity of prediction intervals on the Bayesian assumption is particularly serious in nonparametric statistics (Diaconis and Freedman, 1986).

On the other hand, PAC-learning can be applied to an algorithm in order to produce upper bounds on the probability of its error with respect to some confidence level. It only assumes that examples are generated independently by some unknown distribution, but for the resulting bounds to be interesting in practice, the data set must be particularly clean. As this is rarely the case, the bounds are typically very loose and therefore not particularly useful for real-world applications (Nouretdinov et al., 2001b). In addition,

PAC theory has two more drawbacks: the majority of relevant results either involve large explicit constants or do not specify the relevant constants at all; the obtained bounds are for the overall error and not for individual predictions. Nevertheless, there are less theoretical and more effective ways of estimating the confidence in predictions, like the *hold-out* estimates. They are attained by randomly dividing examples in two separate partitions, one that is used for obtaining the prediction model and the other for testing it. The observed rate of errors on the test set then allows to assess the confidence to have in the prediction rule when new examples are considered. Conformal methods turn out to be a different way of producing hedged predictions.

Aside from the elegance of conformal prediction methods, at least in comparison with the procedure that relies on a hold-out sample, other features constitute important advantages (Vovk et al., 2005). First, there is no rigid separation between learning and prediction, which is the feature of the traditional approaches that makes hedged prediction feasible. Moreover, the hedged predictions produced by conformal algorithms are more accurate, without involving variable transformations or specifying a model. In addition, the confidence with which the label of a new object is predicted is always tailored not only to the previously seen examples but also to that object. Hence, rather than just providing a bound on the prediction error for the entire distribution, it allows to get different bounds for different instances, something which may be very valuable in many practical applications. For instance, in the medical domain, it is clearly more important to be able to evaluate the confidence in predictions related to individual patients instead of groups of patients.

To sum up, in contrast to Bayesian techniques, CP produces well-calibrated outputs as they are only based on the general randomness assumption, and no assumptions *a priori* about the distribution generating the data is needed. Moreover, unlike the PAC theory, they produce confidence measures that are useful in practice and are associated with individual predictions.

2.3.1. Optimality

The current literature highlights that conformal predictors are essentially the *best* confidence predictors (in the sense we are going to specify), when not the only ones, in a very natural class that satisfy the strong non-asymptotic property of validity. A couple of definitions are required. A confidence predictor γ^α is *invariant* if $\gamma^\alpha(z_1, \ldots, z_n) = \gamma^\alpha(z_{\pi(1)}, \ldots, z_{\pi(n)})$, for any permutation π of the indices $1, \ldots, n$, i.e. it does not depend on the order in which $z_1, \ldots, z_n$ are listed. Under the exchangeability assumption, this is a very natural class of confidence predictors. Later, however, we will also study confidence predictors that are not invariant, such as Mondrian and inductive conformal predictors, respectively in Section 3.1.1 and 3.2. In second place, given a couple of confidence predictors γ_1 and γ_2, we say that γ_2 is *at least as good* as γ_1 if, for any significance level α, $\gamma_2^\alpha(z_1, \ldots, z_n) \subseteq \gamma_1^\alpha(z_1, \ldots, z_n)$ holds for almost all $(z_1, \ldots, z_n)$ generated by any exchangeable distribution on $\mathbf{Z}^n$.

It turns out that any valid invariant confidence predictor is a conformal predictor or can be improved to become a conformal predictor (Shafer and Vovk, 2008).

Proposition 2.2. *Assume* $\mathbf{Z}$ *is a Borel space. Let* γ_1 *be an invariant confidence predictor that is conservatively valid under exchangeability. Then there is a conformal predictor* γ_2 *that is at least as good as* γ_1.

2.4. The Online Framework

Conformal algorithms were originally introduced in 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 et al. (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. However in the online picture, 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. In the spirit of comparison, consider i.i.d. random variables $z_1, z_2, \ldots$, drawn from a gaussian distribution. In a classical framework, Fisher's well known prediction interval reads as:

$$\bar{z}_n \pm t_{n-1}^{1-\alpha/2} s_n \sqrt{\frac{n+1}{n}}, \tag{2.10}$$

where

$$s_n = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z}_n)^2. \tag{2.11}$$

The formula defined in (2.10) is assumed to be used several times, but in entirely separate problems. The online story may seem more complicated, because 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 $z_1, \ldots, z_{104}$. but this overlap does not actually matter. As shown for the first time in Shafer and Vovk (2008), the following holds:

Proposition 2.3. *Under the exchangeability assumption, in the online mode, predictors make errors at different steps independently.*

Going back to conformal predictors, we already know that the probability of error is below the miscoverage 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 et al. (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 2.3 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, 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 (2012) and for general background related to distribution-free inference Bahadur and Savage (1956), Donoho (1988)).

As a remark, it has been said that the distribution-free coverage offered by conformal intervals is marginal. The conditional coverage may be larger than $1-\alpha$ at some values $X_{n+1} = x$ and smaller than $1-\alpha$ at other values. This should not be considered as a disadvantage of conformal inference, because the statistical accuracy of conformal prediction bands crucially depends on the base estimator. In a sense, conformal inference broadens the scope and the value of any point estimator with nearly no costs: if the estimator is accurate (which usually requires an approximately correctly specified model, and a proper choice of tuning parameters), then the conformal prediction band is near-optimal; if the estimator is bad, then we still have valid marginal coverage. As a result, it makes sense to use a conformal prediction band as a diagnostic and comparison tool for regression function estimators.

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 $\text{supp}(P_X)$, 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. Still on validity, recently Barber et al. (2019) reflect again on the idea of a proper intermediate definition.

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 (3.4), 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 et al. (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 et al. (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:

$$p_z \coloneqq \frac{|\{\, i = 1, \ldots, n+1 : \kappa_i = \kappa_{n+1} \,\&\, R_i \geq R_{n+1}\}|}{|\{\, i : \kappa_i = \kappa_{n+1}\}|}, \tag{3.4}$$

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. Indeed,

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.*

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 miscoverage levels allow us to specify the relative importance of different kinds of prediction errors. And MCPs still do the job (Vovk et al., 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 et al. (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

A relevant problem of conformal predictors is their computational inefficiency. Over time, an extensive literature has developed to address this issue. In particular, *inductive conformal predictors* (ICPs) have been proposed.

ICPs were first proposed by Papadopoulos et al. (2002a) for regression and by Papadopoulos et al. (2002b) 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

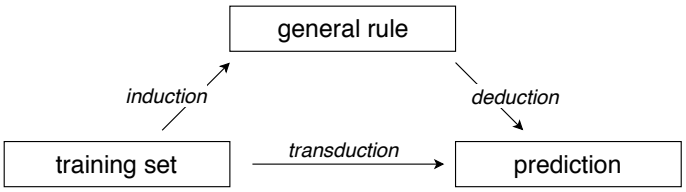

Figure 2. Inductive and transductive approach to prediction.

pattern recognition problem which uses support vector machines in its implementation, Saunders et al. (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 (2013) 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 2), 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 $(z_1, \dots, z_n)$ is *split* into two parts, the proper training set $(z_1, \dots, z_m)$ of size $m < n$ and the *calibration set* $(z_{m+1}, \dots, z_n)$. 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. Therefore the p-value is:

$$p_z := \frac{|\{\, i = m+1, \dots, n+1 : R_i \geq R_{n+1}\}|}{n - m + 1}. \tag{3.5}$$

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 (2012).

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. (2014). 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. (2014) 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.

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 $(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n)$ 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 shorter 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:

$$\mathbb{P}\left(Y_i \in \Gamma^{\alpha}_{jack}(X_i)\right) \geq 1 - \alpha, \tag{3.6}$$

but when dealing with out-of-sample coverage (actually, true predictive inference), its properties 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.

Although not theoretically valid, the jackknife procedures are shown to be empirically valid and informationally efficient. 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 et al. (2018) do, proposing a novel, fast and efficient conformal regressor, with combines

the local-weighted (see Section 3.5) 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 modification to the algorithm gives life to the jackknife+ methods, able to restore rigorous coverage guarantees (Barber et al., 2019b).

3.3. Regression and Approximations

While examining the CP algorithm, the reader may notice that for each possible value $y \in \mathbb{R}$ (that is, for each potential value y for the test data point Y_{n+1}), we must refit the model f. Depending on the setting, each run may be fairly expensive — but even disregarding cost, in general we cannot hope to run it infinitely many times, one for each $y \in \mathbb{R}$.

In some settings, this problem can be circumvented using specific regularities within the model fitting algorithm (as the RRCM, Section 2.2). In nearly any other setting, however, we must instead turn to approximations of the full conformal prediction method.

Efficient approximations are available for kernel density estimator, as in Lei et al. (2013), and kernel nonparametric regression (Lei and Wasserman, 2014). They exploit a result, known as the "sandwich lemma", which provides a simple characterization of the conformal prediction set in terms of the plug-in estimators of density level set. Indeed, the set predictor, whose analytical form may be intractable, is "sandwiched" by two kernel density level sets, with carefully tuned cut-off parameters, that can be computed quickly and maintain finite sample validity.

Except on these situations, two approaches are available. A straightforward way to approximate the algorithm is to fit it only for a finite set of y values — for instance, taking a fine grid over some interval $[a, b]$ that includes the empirical range of the observed response values. That's exactly how the `conformalInference` R package, developed in Lei et al. (2018), is implemented: in order to compute the conformal confidence predictor at a new covariate vector x_{n+1}, it scans a set of grid points in the space $\mathbf{Y}$. Chen et al. (2018) formalize this rounding procedure, proving that rounding can be done without losing the coverage guarantee of the method.

The second approach, commonly used in the inductive setting, relies instead on the quantiles of the fitted residual distribution. Let R_s be the sth smallest value among the nonconformity scores $R_1, \ldots, R_n$, where $s = \lceil (n+1)(1-\alpha) \rceil$. Actually, R_s forms a probabilistic bound for the residuals at significance level α; that is, with probability $1-\alpha$, the nonconformity score of x_{n+1} will be at most R_s. The conformal set predictor is then:

$$\Gamma^{\alpha}(x_{n+1}) = [f(x_{n+1}) - R_s, f(x_{n+1}) + R_s]. \tag{3.7}$$

It is self-evident how, as we improve the estimate of the underlying regression function $f(x)$, residuals get smaller, and the resulting prediction interval decreases in length.

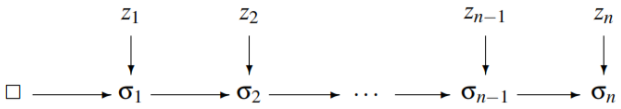

Figure 3. Updating summaries in online compression models.

3.4. Online Compression Models

The idea of conformal prediction can be generalized from learning under randomness, where examples are independent and identically distributed, to *online compression models*. These models include, besides the exchangeability model, the Gaussian model, the Markov model, and many others.

In an online compression model (OCM), it is assumed that data can be summarized in way that can be updated as new examples show up, and the only probabilities given are backward probabilities — probabilities that explain how the updated summary might have been obtained. It is usually impossible to restore all the statistical information from the OCM's summary (so they perform lossy compression), but it can be argued that the only information lost is noise, and the summary is a sufficient statistic, which store knowledge related to data, useful for predicting future examples, in an efficient way.

In general, an online compression model for an example space $\mathbf{Z}$ consists of a space S, whose elements we call *summaries*, and:

- a sequence $U_1, U_2, \ldots$ of updating functions, to bring up to date the summary statistics as new examples come up. At the $(n+1)$th trial, the function U_{n+1}, given a summary σ and a new example z, outputs the new summary $U_{n+1}(\sigma, z)$;
- a sequence of one-step kernels $R_1, R_2, \ldots$. For each summary σ, the kernel R_n defines a joint probability distribution $R_n(\sigma', z|\sigma)$, for an unknown summary σ' and unknown example z. It is required that the set of pairs (σ', z) such that $U_n(\sigma', z) = \sigma$ has probability one.

The intuition behind the concept of OCM is that they are a way of summarizing statistical information. At the beginning we do not have any information, which is represented by the empty summary denoted with □. When the first example z_1 arrives, we update our summary to $\sigma_1 := U_1(\square, z_1)$, and so on, as depicted in figure 3.

Moreover, we can also define the sequence of summarizing functions $\Sigma_1, \Sigma_2 \ldots$ and of full kernels $P_1, P_2, \ldots$. Σ_n maps a n-tuple of examples $(z_1, \ldots, z_n)$ to the summary σ_n, and it can be derived from the updating functions just by composition, while P_n is equivalent to looking back all the way and so it can be carried out by combining, backwards from σ_n, one-step look-backs. Actually, P_n is a Markov kernel, of the form $P_n(z_1, \ldots, z_n \mid \sigma_n)$. Such a kernel — and that's the relevant detail — gives probabilities for the different $z_1, \ldots, z_n$ that could have produced σ_n. Usually, online compression models are initially specified in terms of their summarizing functions Σ_n and their full kernels P_n, since these are in most of the cases easy to describe.

A more careful look at the exchangeability model is sufficient to identify the general structure of an online compression model. Indeed, we summarize examples simply by omitting information about their ordering; the ordered examples are summarized by a bag containing them. With the notation introduced above, $\Sigma_n(z_1, \ldots, z_n) = \wr z_1, \ldots, z_n \int$. The backward-looking probabilities are equally simple: given the bag, the different possible orderings all have equal probability, as if the ordering resulted from drawing the examples successively at random from the bag without replacement. The probability of the event $\{z_1 = a_1, \ldots, z_n = a_n\}$ is:

$$P_n(a_1, \ldots, a_n \mid \sigma_n) = \frac{n_1! \ldots n_k!}{n!} \qquad \text{if } \wr a_1, \ldots, a_n \int = \sigma_n, \tag{3.8}$$

and 0 otherwise, where the bag σ_n consists of k different elements, each with cardinality n_j. Other OCMs compress more or less drastically but have a similar structure.

As usual, to use conformal prediction, the starting point is a nonconformity measure, which in this case must be a function $A(\sigma; z)$ such that its value is small if z seems very similar to the examples that might be summarized by σ, and vice versa. In the base case, without labels (as in Section 2.1), we have to decide whether to include z in $\gamma^\alpha(z_1, \ldots, z_n)$ or not. Let $\tilde{\sigma}_n$ and $\tilde{z}_{n+1}$ stand for random variables. The p-value p_z is computed as:

$$p_z := R_{n+1}\left(A\left(\tilde{\sigma}_n, \tilde{z}_{n+1}\right) \geq A\left(\sigma_n, z\right) \mid \sigma_{n+1}\right). \tag{3.9}$$

Hence, as always, $\gamma^\alpha(z_1, \ldots, z_n) = \{z : p_z > \alpha\}$. In the structured case, as presented in Section 2.2, the algorithm is exactly the same of the base case, once setting $z = (x_{n+1}, y)$. Like under the randomness (or exchangeable) assumption, a law of large numbers for backward-looking probabilities holds too, and again we use it to justify confidence in conformal prediction regions. Nevertheless, in this general setting, there is no guarantee any more that conformal prediction regions are optimal.

3.4.1. Exchangeability-Within-Label

The first example of OCM we are going to introduce is still connected to the exchangeability assumption, but it is actually a relaxation of the hypothesis. Suppose only that the examples of each label are exchangeable with each other — so, the appearance of one label might change the probabilities for the next label. For instance, as in the work of Riabko (2005) aimed at relaxing the randomness assumption in online pattern recognition, consider the problem of hand-written character recognition in a text. The stream of characters is far from exchangeable (we strongly expect to meet "u" after "q"). However, the model here presented can be close to be correct: different instances of the character "a", for example, can be almost exchangeable.

As explained in the book of Vovk et al. (2005), chap 8, the exchangeability-within-label model is a Mondrian model, where the category of an example is the label itself. Mondrian models are really interesting when we are willing to assume exchangeability across the categories, because the conformal predictions they produce will always be calibrated within categories.

3.4.2. Online Gaussian Linear Model

The *online Gaussian linear model* overlaps the exchangeability model, in the sense that the assumptions for both of the models can hold at the same time, but the assumptions for one of them can hold without the assumptions for the other holding. It is closely related to the classical Gaussian linear model. The strong result we report in the following is that conformal prediction, under these general assumptions, leads to the same prediction regions that are used for the classical model.

Consider examples $z_1, \ldots, z_n$ of the form $z_i = (x_i, y_i)$, with the label space being the real line $\mathbf{Y} = \mathbb{R}$ and the object space being the p-dimensional Euclidean space, $\mathbf{X} = \mathbb{R}^p$. The OCM here introduced is defined by the sequence of summarizing functions:

$$\Sigma_n = \left(x_1, \ldots, x_n, \sum_{i=1}^{n} y_i x_i, \sum_{i=1}^{n} y_i^2 \right) = (X_n, X_n' Y_n, Y_n' Y_n), \tag{3.10}$$

and the full kernel $P_n(z_1, \ldots, z_n \mid \sigma_n)$ is the *uniform* probability distribution over the set of vectors $(y_1, \ldots, y_n)$ consistent with the summary σ_n. Let Σ_n be (X_n, C, r^2), in short. A vector $(y_1, \ldots, y_n)$ is consistent with σ_n if it belongs to $\Sigma_n^{-1}(\sigma_n) = \Sigma_n^{-1}(X_n, C, r^2)$, namely if $\sum_i y_i = C$ and $\sum_i y_i^2 = r^2$. This is the intersection of a hyperplane with a sphere, may it be a lower-dimensional sphere or, if they are tangent, a point, and the kernel $P_n(\cdot \mid \sigma_n)$ distributes all its probability uniformly over it.

It is interesting, as Vovk et al. (2005) makes clear, that the probability distribution of $z_1, \ldots, z_n$ under the linear regression statistical model $y_i = \beta \cdot x_i + \xi_i$, where β is the constant vector of regression coefficients and ξ_i are the errors, independent random variables with the same zero-mean normal distribution, does agree with the Gauss linear model. Still in the classical framework, it is useful to recall the following theoretical result: given an object x_{n+1}, once computed $\hat{y}_{n+1}$, that is the least squares prediction of its label y based on the examples summarized in σ_n, the interval containing y_{n+1} with probability $1 - \alpha$ reads as:

$$\left[\hat{y}_{n+1} \pm t_{n-p-1,\alpha/2}\, S_n \sqrt{1 + x_{n+1}'(X_n' X_n)^{-1} x_{n+1}} \right], \tag{3.11}$$

with S_n the (standard) unbiased estimate of the noise variance. For details, refer to any statistical book.

The online Gaussian linear model is tightly connected to the classical Gaussian linear model. We are going to give some results about the (close) relation between the classical and the online models (Shafer and Vovk, 2008). First, as just mentioned, but still worth repeating, if $z_1, \ldots, z_n$ fulfill the assumptions of the classical Gaussian linear model, then they satisfy the assumptions of the online model. That is, assuming errors ξ_i to be i.id., with mean zero, a common variance and a normal distribution, implies that, conditional on the summary σ_n, i.e. on $X_n' Y_n = C$ and $Y_n' Y_n = r^2$, the vector Y is distributed uniformly over the sphere defined by C and r^2. Second, the assumption of the online Gaussian linear model is sufficient to guarantee that

$$\frac{y_{n+1} - \hat{y}_{n+1}}{S_n \sqrt{1 + x_{n+1}'(X_n' X_n)^{-1} x_{n+1}}} \tag{3.12}$$

has the t-distribution with $n - p - 1$ degrees of freedom. Third, suppose $z_1, z_2, \ldots$ is an infinite sequence of random variables. Then $z_1, \ldots, z_n$ satisfy the assumptions of the online Gaussian linear model for every integer n if and only if the joint distribution of $z_1, \ldots, z_n$ is a mixture of distributions given by the classical Gaussian linear model, each model in the mixture possibly having a different β and a different variance for the errors.

Therefore, it can be proved that, when the nonconformal measure is $A(\sigma, z) = |y - \hat{y}|$, which is a natural choice, the related conformal prediction region $\Gamma^\alpha(z_1, \ldots, z_n, x_{n+1})$ is exactly the classical (3.11)! Moreover, it has to be kept in mind that in the online setting these intervals are valid, in the sense that they are right $(1 - \alpha)\%$ of the times even though used on accumulating data (Section 2.4).

3.5. 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 (Ryabko et al., 2003; Nouretdinov and Vovk, 2006).

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 et al. (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 et al., 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 et al. (2018), Vovk et al. (2017). For many underlying algorithms, Venn predictors (like conformal methods in general) are computationally inefficient. Therefore Lambrou et al. (2012), and as an extension Lambrou et al. (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. (2019a) to handle weighted exchangeable data, as in the setting of covariate shift (Shimodaira, 2000; Chen et al., 2016b). 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 et al. (2018).

Now, we describe more in details a couple of other recent advances.

3.5.1. Normalized Nonconformity Scores

In conformal algorithms seen so far, the width of $\Gamma^\alpha(x)$ is roughly immune to x (figure 4, 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 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 4, right), that is:

$$R_i = \frac{|\hat{y}_i - y_i|}{\sigma_i}, \tag{3.13}$$

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 et al. (2011). Choosing (3.13), the confidence predictor (3.7) becomes:

$$\Gamma^\alpha(x_{n+1}) = [f(x_{n+1}) - R_s\, \sigma_{n+1}, f(x_{n+1}) + R_s\, \sigma_{n+1}]. \tag{3.14}$$

As a consequence, the resulting predictive regions are in most cases much tighter than those produced by the simple conformal methods. Anyway, using locally-weighted residuals, as in (3.13), 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 et al. (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,

$$d_i^k = \sum_{j=1}^{k} \text{distance}\,(x_i, x_j). \tag{3.15}$$

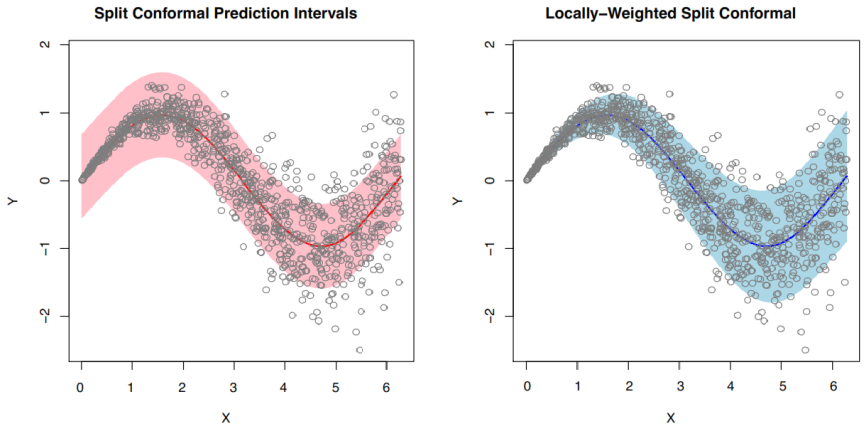

Figure 4. Conformal predictors do not contemplate heteroskedasticity in the data distribution. In such a case, one would expect the length of the output interval to be an increasing function of the corresponding variance of the output value, which can give more information of the target label. To tackle this problem, local-weighted conformal inference has been introduced. *Source: Lei et al. (2018).*

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

3.5.2. High-Dimensional Regression

Only a few works in literature deal with prediction sets in high-dimensional regression, where $x \in \mathbb{R}^p$ and $p \gg n$. Current high-dimensional inference methods make strong assumptions while little is known about their robustness against model misspecification. Common approaches in this setting include greedy methods like forward step-wise regression, and l_1-penalty based methods like the lasso. There is an enormous amount of work dedicated to studying various properties of these methods, but to our knowledge, not the same on set predictors.

In high-dimensional problems, estimators are inevitably more complicated and so the corresponding conformal prediction sets are much harder to characterize. On the other hand, conformal prediction is arguably most useful in such scenarios: model assumptions such as sparsity and low-intrinsic dimensionality are often not true, and the inferential tools developed under such hypotheses are often invalid under model misspecification.

Without any doubt, the most common way to proceed is based on combining the principle of conformal prediction with the l_1-penalized least squares estimator. Over time, an extensive literature has developed on the topic. Hebiri (2010) describes an approximation of the conformalized lasso estimator — a partial conformalization indeed. This approximation leads to a big speedup over the original conformal prediction method build on top of the lasso, but loses the key property of conformal inference, the model free coverage guarantee. Recently, Steinberger and Leeb (2016) analyze the jackknife

conformal method in the high-dimensional setting, but asymptotic validity is not for free and requires some assumptions on the base estimator (of the regression parameters). Meanwhile, Chen et al. (2016) propose a method which explores a smaller search space. Computational costs are so reduced by a constant fraction, but it still evaluates the prediction set on a grid of points. Lastly, as already mentioned, Lei (2017) develop an algorithm that efficiently and exactly computes the conformal prediction set for the lasso, in an analogous way, to a certain extent, to RRCM (Section 2.2.2).

More in general, Lei et al. (2018) think that the main way to approach high-dimensional problems lies in the simple, computationally efficient, and yet powerful method that split conformal inference represents. In their work, empirical properties of conformal methods under different simulated data settings are examined — from a simple (linear and classical) setup, to a heteroskedastic and heavy-tailed one, with correlated features. In particular, they compare performances between conformal prediction based on the ordinary linear regression estimator and classical parametric prediction intervals for linear models. Actually, in high-dimensional problems, the full conformal interval outperforms the parametric one in terms of both length and coverage across all settings, due to the poor accuracy of linear regression estimators when p is large. Even the use of ridge regression does not change things. Moreover, looking at the different implementations of conformal prediction, the split method exhibits a clear computational advantage compared to the full one, guaranteeing similar performance. With such a dramatically reduced computation cost, as already mentioned but even more precious here, adopting split conformal in combination with computationally heavy estimators that involve cross-validation or bootstrap is considered as the best approach.

In the same work, they cast light on an interesting topic, i.e. how conformal inference can help with model-free *variable selection*. The aim is to construct model-free, prediction-based inferential statements about the importance of each covariate in the prediction model for Y_{n+1} given X_{n+1}. To do so, they propose a notion of variable importance, called leave-one-covariate-out (or LOCO) inference. A random variable Δ_j, for each covariate j, $j = 1, \ldots, p$, is properly defined to measure the increase in prediction error due to not having access to that covariate in the data set. And consequently inferential statements about variable importance are carried out, based on these variables.

3.5.3. 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 et al. (2015) guide us. The se-

quence $z_1(\cdot), \ldots, z_n(\cdot)$ consists now of $L^2[0,1]$ functions. The definition of validity for a confidence predictor γ^α is:

$$\mathbb{P}\left(z_{n+1}(t) \in \gamma^\alpha(z_1, \ldots, z_n)(t)\ \forall t\right) \geq 1 - \alpha \qquad \text{for all } P. \tag{3.16}$$

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

$$R_i = \int (z_i(t) - \bar{z}(t))^2 dt, \tag{3.17}$$

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:

$$l(t) = \inf_{z \in \gamma^\alpha} z(t) \qquad \text{and} \qquad u(t) = \sup_{z \in \gamma^\alpha} z(t). \tag{3.18}$$

Consequently, thanks to provable conformal properties,

$$\mathbb{P}\left(l(t) \leq z_{n+1}(t) \leq u(t),\ \forall t\right) \geq 1 - \alpha. \tag{3.19}$$

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 et al. (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 3.16).

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:

$$\mathbb{P}\left(\Pi(z_{n+1})(t) \in \gamma^\alpha(z_1, \ldots, z_n)(t)\ \forall t\right) \geq 1 - \alpha \qquad \text{for all } P, \tag{3.20}$$

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 (3.20) let it known in advance, 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:

$$\gamma_\xi = \{\xi \in \mathbb{R}^p : R_\xi \leq R_s\} \tag{3.21}$$

and

$$\gamma^\alpha(t) = \left\{ \sum_{i=1}^{p} \varsigma_i \phi_i(t) : (\varsigma_1, \ldots, \varsigma_p) \in \gamma_\xi \right\}. \tag{3.22}$$

As a consequence, γ^α is valid, i.e. (3.20) holds.

Exploiting the finite dimensional projection, the nonconformity 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 (the inverse of) 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. So, the extension of CP to FDA is still considered an important open question.

Acknowledgments

The authors acknowledge financial support from: ACCORDO Quadro ASI-POLIMI "Attività di Ricerca e Innovazione" n. 2018-5-HH.0, collaboration agreement between the Italian Space Agency and Politecnico di Milano; the European Research Council, ERC grant agreement no 336155-project COBHAM "The role of consumer behaviour and heterogeneity in the integrated assessment of energy and climate policies"; the "Safari Njema Project - From informal mobility to mobility policies through big data analysis", funded by Polisocial Award 2018 - Politecnico di Milano.

References

Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122. MR0084241

Balasubramanian, V., Ho, S.-S., and Vovk, V. (2014). *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

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

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2019). The limits of distribution-free conditional predictive inference. *arXiv preprint arXiv:1903.04684*.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2019b). Predictive inference with the jackknife+. *arXiv preprint arXiv:1905.02928*.

Burnaev, E. and Vovk, V. (2014). Efficiency of conformalized ridge regression. In *Conference on Learning Theory*, pages 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)*, pages 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*, pages 251–259. Springer.

Chen, W., Chun, K.-J., and Barber, R. F. (2018). Discretized conformal prediction for efficient distribution-free inference. *Stat*, 7(1):e173. MR3769053

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

Chen, X., Monfort, M., Liu, A., and Ziebart, B. D. (2016b). Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pages 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 arXiv:1802.06300*.

Dashevskiy, M., and Luo, Z. (2011). Time series prediction with performance guarantee. *IET communications*, 5(8):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*, pages 37–44. Springer.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, pages 1–26. MR0829555

Donoho, D. L. (1988). One-sided inference about functionals of a density. *The Annals of Statistics*, 16(4):1390–1420. MR0964930

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

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

Hebiri, M. (2010). Sparse conformal predictors. *Statistics and Computing*, 20(2):253–266. MR2610776

Hewitt, E. and Savage, L.J. (1955). Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501. MR0076206

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

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*, pages 271–280. Springer.

Johansson, U., Boström, H., Löfström, T., and Linusson, H. (2014). Regression conformal prediction with random forests. *Machine Learning*, 97(1-2):155–176. MR3252831

Lambrou, A., Nouretdinov, I., and Papadopoulos, H. (2015). Inductive Venn prediction. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):181–201. MR3353902

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*, pages 182–191. Springer.

Lei, J. (2017). Fast exact conformalization of lasso using piecewise linear homotopy. *arXiv preprint arXiv:1708.00427*.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111. MR3862342

Lei, J., Rinaldo, A., and Wasserman, L. (2015). A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):29–43. MR3353895

Lei, J., Robins, J., and Wasserman, L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287. MR3174619

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96. MR3153934

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

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

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

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

Nouretdinov, I., Volkhonskiy, D., Lim, P., Toccaceli, P., and Gammerman, A. (2018). Inductive Venn-Abers predictive distribution. *Proceedings of Machine Learning Research*, 91:1–22.

Nouretdinov, I. and Vovk, V. (2006). Criterion of calibration for transductive confidence machine with limited feedback. *Theoretical computer science*, 364(1):3–9. MR2268298

Nouretdinov, I., Vovk, V., Vyugin, M., and Gammerman, A. (2001b). Pattern recognition and density estimation under the general iid assumption. In *International Conference on Computational Learning Theory*, pages 337–353. Springer.

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)*, pages 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*, pages 32–41. Springer.

Papadopoulos, H. and Haralambous, H. (2011). Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002a). Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer. MR2050303

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

Papadopoulos, H., Vovk, V., and Gammerman, A. (2011). Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer. MR1889966

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

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*, volume 2, pages 722–726.

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

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421. MR2417240

Shen, J., Liu, R. Y., and Xie, M.-g. (2018). Prediction with confidence - a general framework for predictive inference. *Journal of Statistical Planning and Inference*, 195:126–140. MR3760843

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244. MR1795598

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

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

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. (1998). *Statistical learning theory. 1998*, volume 3. Wiley, New York.

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

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. MR3080332

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

Vovk, V. (2015). Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):9–28. MR3353894

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

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

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

Vovk, V., Nouretdinov, I., Gammerman, A., et al. (2009). Online predictive linear regression. *The Annals of Statistics*, 37(3):1566–1590. MR2509084

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

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

Vovk, V., Shen, J., Manokhin, V., and Xie, M. (2017). Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, pages 1–30. MR3917055

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

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

21/2020 Benacchio, T.; Bonaventura, L.; Altenbernd, M.; Cantwell, C.D.; Düben, P.D.; Gillard, M.; Gir
Resilience and fault-tolerance in high-performance computing for numerical weather and climate prediction

20/2020 Almi, S.; Belz, S.; Micheletti, S.; Perotto, S.
A DIMENSION-REDUCTION MODEL FOR BRITTLE FRACTURES ON THIN SHELLS WITH MESH ADAPTIVITY

19/2020 Stella, S.; Vergara, C.; Maines, M.; Catanzariti, D.; Africa, P.; Demattè, C.; Centonze, M.; Nob
Integration of maps of activation times in computational cardiac electrophysiology

17/2020 Cerroni, D.; Formaggia, L.; Scotti, A.
A control problem approach to Coulomb's friction

18/2020 Fumagalli, A.; Scotti, A.; Formaggia, L.
Performances of the mixed virtual element method on complex grids for underground flow

16/2020 Paolucci, R.; Mazzieri, I.; Piunno, G.; Smerzini, C.; Vanini, M.; Ozcebe, A.G.
Earthquake ground motion modelling of induced seismicity in the Groningen gas field

15/2020 Fumagalli, I.; Fedele, M.; Vergara, C.; Dede', L.; Ippolito, S.; Nicolò, F.; Antona, C.; Scrofani,
An Image-based Computational Hemodynamics Study of the Systolic Anterior Motion of the Mitral Valve

14/2020 Calissano, A.; Feragen, A; Vantini, S.
Populations of Unlabeled Networks: Graph Space Geometry and Geodesic Principal Components

13/2020 Pozzi S.; Domanin M.; Forzenigo L.; Votta E.; Zunino P.; Redaelli A.; Vergara C.
A data-driven surrogate model for fluid-structure interaction in carotid arteries with plaque

11/2020 Antonietti, P.F.; Facciola', C.; Houston, P.; Mazzieri, I.; Pennes, G.; Verani, M.
High-order discontinuous Galerkin methods on polyhedral grids for geophysical applications: seismic wave propagation and fractured reservoir simulations