Conformal calibration

Vladimir Vovk, Ivan Petej, Paolo Toccaceli, and Alex Gammerman

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #23

First posted February 15, 2019. Last revised August 2, 2020.

Project web site:
http://alrw.net

Abstract

Most existing examples of full conformal predictive systems, split conformal predictive systems, and cross-conformal predictive systems impose severe restrictions on the adaptation of predictive distributions to the test object at hand. In this paper we develop split conformal predictive systems that are fully adaptive. Our method consists in calibrating existing predictive systems; the input predictive system is not supposed to satisfy any properties of validity, whereas the output predictive system is guaranteed to be calibrated in probability.

Contents

1 Introduction

Conformal predictive distributions were inspired by the work on predictive distributions in parametric statistics (see, e.g., [7], Chapter 12, and [8]) and first suggested in [14]. As usual, we will refer to algorithms producing conformal predictive distributions as conformal predictive systems (CPS, used in both singular and plural senses).

Conformal predictive systems are built on top of traditional prediction algorithms to ensure a property of validity usually referred to as calibration in probability [5]. Several versions of the Least Squares Prediction Machine, CPS based on the method of Least Squares, are constructed in [14]. This construction is slightly extended to cover ridge regression and then further extended to nonlinear settings by applying the kernel trick in [12]. However, even after this extension the method is not fully adaptive, even for a universal kernel. As explained in [12, Section 7], the universality of the kernel shows in the ability of the predictive distribution function to take any shape; however, the CPS is still inflexible in that the shape does not depend, or depends weakly, on the test object.

For many base algorithms full CPS (like full conformal predictors in general) are computationally inefficient, and [13] define and study computationally efficient versions of CPS, namely split conformal predictive systems (SCPS) and cross-conformal predictive systems (CCPS). However, specific SCPS and CCPS proposed in [13] are based on the split conformity measure

$$A(z_1, \dots, z_m, (x, y)) := \frac{y - \hat{y}}{\hat{\sigma}}, \tag{1}$$

where $\hat{y}$ is a prediction for y computed from x as test object and $z_1, \dots, z_m$ as training sequence, and $\hat{\sigma} > 0$ is an estimate of the quality of $\hat{y}$ computed from the same data. The predictive distributions corresponding to (1) are slightly more adaptive: not only their location but also their scale depends on the test object x. (The conformity measures used in [14] and [12] correspond to (1) with $\hat{\sigma} := 1$ and so implicitly assume homoscedasticity.) Ideally, however, we would like to allow a stronger dependence on the test object. This paper follows [10, Section 10] in using a method that is fully flexible and, for a suitable base algorithm, adapts fully to the test object (cf. Proposition 2 below). Whereas the emphasis in [10] is on asymptotic optimality only, one of the purposes of this paper is to propose practically useful solutions.

We start by defining, in Section 2, algorithms outputting predictive distributions, which we call predictive systems (when randomization is not allowed) or randomized predictive systems (when it is allowed). In the next section we define split conformal predictive systems. Section 4 is devoted to validity. In particular, we explain that split conformal predictive systems are always valid, in the sense of being calibrated in probability, under the IID assumption. The IID assumption, standard in conformal prediction, is that the observations are generated in the IID fashion (sometimes this assumption is slightly weakened to assuming an online compression model, as in [11, Chapter 8]). In Section 5 we discuss conformalizing ideal predictive systems under the IID assumption, although in this context this assumption becomes less essential. Section 6 contains some experimental results. Section 7 states directions of further research.

This paper and [10] establish very different versions of the generic notion of efficiency. Whereas [10] studies an asymptotic version of efficiency, this paper concentrate on a rather narrow but small-sample version. It is a less conservative form of the medical principle “first, do no harm”: if a predictive system is already perfect, conformalizing it should not make it much worse.

2 Predictive systems and randomized predictive systems

Let us fix (until Section 6) a nonempty measurable space $\mathbf{X}$ that will serve as our *object space*, and let $\mathbf{Z} := \mathbf{X} \times \mathbb{R}$ stand for our *observation space*. Each observation $z = (x, y) \in \mathbf{Z}$ consists of an object $x \in \mathbf{X}$ and its label $y \in \mathbb{R}$.

Definition 1. A measurable function $Q : \cup_{n=1}^{\infty} \mathbf{Z}^{n+1} \to [0, 1]$ is called a *predictive system* (PS) if:

1. For each $n \in \{1, 2, \dots\}$, each training sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$, and each test object $x \in \mathbf{X}$, the function $Q(z_1, \dots, z_n, (x, y))$ is monotonically increasing in y (where “monotonically increasing” is understood in the wide sense allowing intervals of constancy).

2. For each $n \in \{1, 2, \dots\}$, each training sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$, and each test object $x \in \mathbf{X}$,

$$\lim_{y \to -\infty} Q(z_1, \dots, z_n, (x, y)) = 0$$

and

$$\lim_{y \to \infty} Q(z_1, \dots, z_n, (x, y)) = 1.$$

The output $y \in \mathbb{R} \mapsto Q(z_1, \dots, z_n, (x, y))$ of a PS on a given training sequence $z_1, \dots, z_n$ and test object x will be referred to as a *predictive distribution* (function) and will sometimes be denoted $Q_{z_1,\dots,z_n,x}$. It is a distribution function in the sense of probability theory except that we do not require that it be right-continuous.

We also need the notion of a randomized predictive system.

Definition 2. A measurable function $Q : \cup_{n=1}^{\infty} \mathbf{Z}^{n+1} \times [0, 1] \to [0, 1]$ is called a *randomized predictive system* (RPS) if:

1. For each $n \in \{1, 2, \dots\}$, each training sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$, and each test object $x \in \mathbf{X}$, the function $Q(z_1, \dots, z_n, (x, y), \tau)$ is monotonically increasing in y and monotonically increasing in τ.

2. For each $n \in \{1, 2, \dots\}$, each training sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$, and each test object $x \in \mathbf{X}$,

$$\lim_{y \to -\infty} Q(z_1, \dots, z_n, (x, y), 0) = 0$$

and

$$\lim_{y \to \infty} Q(z_1, \dots, z_n, (x, y), 1) = 1.$$

The output $y \in \mathbb{R} \mapsto Q(z_1, \dots, z_n, (x, y), \tau)$ of an RPS on a given training sequence $z_1, \dots, z_n$, test object x, and (random) number τ will be referred to as a *predictive distribution* (function) and will sometimes be denoted $Q_{z_1, \dots, z_n, x, \tau}$.

Notice that Definition 2 does not include any requirement of validity, unlike the corresponding definitions in [12–14] and [10]: in this paper we follow the terminology of [7, Chapter 12] rather than [8]. Validity will be discussed in Section 4.

3 Split conformal calibration

If A is a predictive system, the *split conformal predictive system* (SCPS) corresponding to A (or the *split-conformalized version* of A) is defined as follows. The training sequence $z_1, \dots, z_n$ is split into two parts: the *training sequence proper* $z_1, \dots, z_m$ and the *calibration sequence* $z_{m+1}, \dots, z_n$, where we assume $1 \le m < n$. Given a test object x, the output of C^A is defined as

$$\begin{aligned} C^A_{z_1, \dots, z_n, x, \tau}(y) := {} & \frac{1}{n - m + 1} \left| \{ i = m + 1, \dots, n \mid \alpha_i < \alpha^y \} \right| \\ & + \frac{\tau}{n - m + 1} \left| \{ i = m + 1, \dots, n \mid \alpha_i = \alpha^y \} \right| + \frac{\tau}{n - m + 1}, \end{aligned} \qquad (2)$$

where the *conformity scores* α_i, $i = m + 1, \dots, n$, and α^y, $y \in \mathbb{R}$, are defined by

$$\begin{aligned} \alpha_i &:= A(z_1, \dots, z_m, (x_i, y_i)), \qquad i = m + 1, \dots, n, \\ \alpha^y &:= A(z_1, \dots, z_m, (x, y)). \end{aligned}$$

This follows the definition of a split conformal transducer in [13].

For simplicity, let us assume that A never takes values 0 and 1. When considered as a split conformity measure, as defined in [13, Section 3], such a predictive system is balanced and isotonic, which makes it possible to apply Proposition 3.1 in [13] and conclude that the SCPS C^A is an RPS (and satisfies the property of validity introduced in Section 4 below). We refer to this method, namely transforming predictive systems to the corresponding split conformal predictive systems, as *split conformal calibration*.

The SCPS C^A can be implemented by directly coding the definition (2) using a grid of values of y (as we do for the experiments in Section 6). Algorithm 1 describes another implementation of C^A. It defines the predictive distribution apart from a finite number of points y (and so the values at those points do not

Algorithm 1 Split Conformal Calibration

```
Require: Training sequence $(x_i, y_i) \in \mathbf{Z}$, $i = 1, \ldots, n$, and positive integer $m < n$.
Require: Test object $x \in \mathbf{X}$ and random number $\tau \in [0, 1]$.
  for $i \in \{1, \ldots, n - m\}$ do
    $p_i := A(z_1, \ldots, z_m, z_{m+i})$
  end for
  sort $p_1, \ldots, p_{n-m}$ in the increasing order obtaining $p_{(1)} < \cdots < p_{(k)}$
  for $j \in \{1, \ldots, k\}$ do
    $n_j := \left|\{i = 1, \ldots, n - m \mid p_i = p_{(j)}\}\right|$
    $m_j := \sup\{y \mid A(z_1, \ldots, z_m, (x, y)) < p_{(j)}\}$
    $M_j := \inf\{y \mid A(z_1, \ldots, z_m, (x, y)) > p_{(j)}\}$
  end for
  return the predictive distribution $C^A$ given by (3) for the label $y$ of $x$.
```

affect, e.g., CRPS as defined by (10) in Section 6); we can set the probability interval $\operatorname{conv}(\{C^A_{z_1,\ldots,z_n,x,\tau}(y) \mid \tau \in [0, 1]\})$ at those points y to the union of the prediction intervals at the adjacent points without a substantial change to the predictive system. Some of the p_i, $i = 1, \ldots, n-m$, in Algorithm 1 may coincide, so we can only say that $k \in \{1, \ldots, n-m\}$ rather than $k = n-m$ (notice that the sequence $p_{(j)}$, $j = 1, \ldots, k$, is strictly increasing). The predictive distribution that it outputs is

$$C^A_{z_1,\ldots,z_n,x,\tau}(y) = \begin{cases} \frac{\tau}{n-m+1} & \text{if } y < m_1 \\ \frac{n_1+\cdots+n_{j-1}+\tau n_j+\tau}{n-m+1} & \text{if } y \in (m_j, M_j),\ j \in \{1, \ldots, k\} \\ \frac{n_1+\cdots+n_j+\tau}{n-m+1} & \text{if } y \in (M_j, m_{j+1}),\ j \in \{1, \ldots, k-1\} \\ \frac{n_1+\cdots+n_k+\tau}{n-m+1} = \frac{n-m+\tau}{n-m+1} & \text{if } y > M_k. \end{cases} \tag{3}$$

Algorithm 1 is a slight generalization of Algorithm 1 in [13]. The latter makes an assumption (the base distribution functions $A_{z_1,\ldots,z_n,x}$ being continuous and strictly increasing) implying that $m_j = M_j$ for all $j \in \{1, \ldots, k\}$; in our current general context we can only say that

$$m_1 \le M_1 \le m_2 \le M_2 \le \cdots \le m_k \le M_k.$$

Location-scale models

The split conformity measure (1), which is used in [13], is not covered directly by our definition since it does not have to take values in $[0, 1]$. But this can be easily arranged: e.g., we can apply a fixed strictly increasing distribution function $F : \mathbb{R} \to [0, 1]$ to (1) to make sure the split conformity measure takes values in $[0, 1]$. This makes the approach based on (1) a special case of this paper's approach corresponding to the location-scale families

$$F_{\mu,\sigma}(y) := F\left(\frac{y-\mu}{\sigma}\right). \tag{4}$$

Notice that the conformalized predictive distributions will not depend on the choice of F as long as F is strictly increasing; e.g., we can fix it to the standard Gaussian distribution function (and this will not mean that we are relying on the Gaussian assumption).

Specializing (4) by setting $\sigma := 1$, we obtain a class containing the predictive systems considered in [12, 14]. This class will be used in our experiments in Section 6.

4 Validity of conformal calibration

An RPS Q is *calibrated in probability* if, for any probability measure P on $\mathbf{Z}$, as function of random training observations $Z_1 \sim P, \dots, Z_n \sim P$, a random test observation $Z \sim P$, and a random number $\tau \sim U$ (U being the uniform probability measure on $[0,1]$), all assumed independent, the distribution of Q is uniform:

$$\forall \alpha \in [0,1] : \mathbb{P}\left(Q(Z_1,\dots,Z_n,Z,\tau) \le \alpha\right) = \alpha. \tag{5}$$

(This was included as Requirement R2 in the definition of an RPS in [12–14] and [10].)

Split conformal predictive systems are automatically calibrated in probability, in the sense of satisfying (5), under the IID assumption. If F is the distribution function produced for a test object X^*, $F := C^A_{Z_1,\dots,Z_n,X^*,\tau}$, then $F(Y^*)$ will be distributed uniformly on $[0,1]$, where Y^* is the true label of X^*. Notice, however, that for a test sequence $Z_i^* = (X_i^*, Y_i^*)$, $i = 1,\dots,l$, $F_i(Y_i^*)$ will not be independent, even though distributed uniformly on $[0,1]$, where $F_i := C^A_{Z_1,\dots,Z_n,X_i^*,\tau_i}$ is the distribution function produced for X_i^*. To make $F_i(Y_i^*)$ not only distributed uniformly on $[0,1]$ but also independent, we can use the *semi-online protocol*, predicting the labels Y_i^* of X_i^*, $i = 1,\dots,k$, sequentially and adding Z_i^* to the calibration sequence as soon as it is processed. This is asserted in the following proposition and might be useful for debugging implementations of split-conformalized predictive systems.

Proposition 1. *Suppose $Z_1,\dots,Z_n,Z_1^*,Z_2^*,\dots$ is an IID sequence of observations and $(\tau_1,\tau_2,\dots) \in [0,1]^\infty$ is independent and uniformly distributed. Then the random variables*

$$C^A_{Z_1,\dots,Z_n,Z_1^*,\dots,Z_{i-1}^*,X_i^*,\tau_i}(Y_i^*)$$

are independent and uniformly distributed on $[0,1]$*, where X_i^* and Y_i^* are the components of $Z_i^* = (X_i^*, Y_i^*)$.*

This proposition gives a stronger property of validity, online calibration in probability, for split conformal prediction.

Cross-conformal calibration

We can easily combine several split conformal predictive systems as defined in the previous section into a cross-conformal predictive system, exactly in the same way as in [13, Section 4]. The resulting RPS will lose automatic calibration in probability (5) but will use the available data more efficiently.

Full conformal calibration

Let us say that a predictive system A is *invariant* if, for any $n \in \{1, 2, \dots\}$ and any $z_1, \dots, z_n, z \in \mathbf{Z}$, the value $A(z_1, \dots, z_n, z)$ does not depend on the order of $z_1, \dots, z_n$. The *full conformal predictive system* (or simply *conformal predictive system*) corresponding to an invariant predictive system A is defined in [14, Section 2]. This definition, however, is applicable to a narrower class of predictive systems than that in the definition of the split conformal predictive systems. For example, it will be applicable if we assume, additionally, that, for any $n \in \{1, 2, \dots\}$, any $x_1 \in \mathbf{X}$, and any $z_2, \dots, z_{n+1}$, $A((x_1, y_1), z_2, \dots, z_{n+1})$ is monotonically decreasing in $y_1 \in \mathbb{R}$ [14, Section 2.2].

Full conformal predictive systems are automatically calibrated in probability [14, Section 2].

5 Efficiency of conformalizing ideal predictive systems

In this section we will explore the efficiency of conformal calibration in the situation where the base predictive system A is the ideal one. In this case we cannot improve A, and we are interested in how much worse C^A can become as compared with A. (Similar questions were asked by Wasserman and by [2].) If, for any A, C^A is almost as good as A, we can say that the calibration method is fully adaptive.

Let P be the true probability measure on $\mathbf{Z}$ generating the observations $z_1, z_2, \dots$ in the IID manner. A *conditional distribution function* for P is a function $A : \mathbf{Z} \to [0, 1]$ such that:

- for each $x \in \mathbf{X}$, as function of $y \in \mathbb{R}$, $A(x, y)$ is a distribution function (i.e., is increasing, is right-continuous, and satisfies $A(x, -y) \to 0$ and $A(x, y) \to 1$ as $y \to \infty$);
- for each $y \in \mathbb{R}$,
$$A(X, y) = \mathbb{P}(Y \le y \mid X) \quad \text{a.s.} \tag{6}$$
when $(X, Y) \sim P$.

The existence and a.s. uniqueness of a conditional distribution function follows from standard results about the existence of regular probability distributions (e.g., [4, Theorem 10.2.2]).

Consider a sequence $\xi_1, \xi_2, \dots$ of independent and uniformly distributed random variables $\xi_i \sim U$. Let $\mathbb{G}_n$ be the empirical distribution function of $\xi_1, \dots, \xi_n$; we are using the notation of [9], who refer to $\mathbb{G}_n$ as the *uniform empirical distribution function*. For large n and with high probability, $\mathbb{G}_n$ is close to the main diagonal of the unit square $[0,1]^2$.

Let us use the true conditional distribution function A, satisfying (6), as base predictive system (roughly, this corresponds to an infinitely long training sequence proper). The corresponding *ideal conformalized predictive system* (ICPS) is defined as

$$C^A_{z_1,\dots,z_n,x,\tau}(y) := \frac{1}{n+1} \left|\{i = 1, \dots, n \mid A(x_i, y_i) < A(x, y)\}\right| \\ + \frac{\tau}{n+1} \left|\{i = 1, \dots, n \mid A(x_i, y_i) = A(x, y)\}\right| + \frac{\tau}{n+1},$$

where x is the test object. Intuitively, the whole training sequence is used as the calibration sequence (we do not need a training sequence proper as A is already perfect). An ICPS is an idealization of both SCPS and cross-conformal predictive systems.

The following proposition says that C^A will be close to A and that the distance between them will be of order $n^{-1/2}$. We will state it in a semi-online protocol and further discuss the intuition behind it after its proof.

Proposition 2. *Suppose the conditional distribution function $A_x := A(x, \cdot)$ (for the true probability measure P) is continuous and strictly increasing for almost all $x \in \mathbf{X}$, and $Z_1, Z_2, \dots \sim P$ and $\tau_1, \tau_2, \dots \sim U$ are all independent. Then the ICPS C^A satisfies*

$$\left(C^A_{Z_1,\dots,Z_n,X_{n+1},\tau_{n+1}} \circ A^{-1}_{X_{n+1}}\right)_{n=1}^{\infty} \overset{d}{=} (\mathbb{G}_n + \eta_n)_{n=1}^{\infty}, \qquad (7)$$

where X_{n+1} is the first component of Z_{n+1}, $\overset{d}{=}$ means the equality of distributions, and η_n are random functions in the Skorokhod space $D[0,1]$ satisfying $\|\eta_n\|_\infty \le 1/(n+1)$ a.s.

Proof. For given $t \in [0,1]$ and n,

$$C^A_{Z_1,\dots,Z_n,X,\tau}\left(A^{-1}_X(t)\right) = \frac{1}{n+1} \left|\{i \in \{1,\dots,n\} \mid A_{X_i}(Y_i) < t\}\right| \\ + \frac{\tau}{n+1} \left|\{i \in \{1,\dots,n\} \mid A_{X_i}(Y_i) = t\}\right| + \frac{\tau}{n+1} = \frac{k}{n+1} + \frac{\tau}{n+1},$$

where the second equality holds almost surely and

$$k := \left|\{i \in \{1,\dots,n\} \mid A_{X_i}(Y_i) \le t\}\right|.$$

It remains to notice that the probability integral transforms $A_{X_i}(Y_i) \sim U$ are IID and that

$$\sup_{\tau,k} \left|\frac{k}{n+1} + \frac{\tau}{n+1} - \frac{k}{n}\right| = \sup_{\tau,k} \left|\frac{\tau - k/n}{n+1}\right| = \frac{1}{n+1},$$

where τ ranges over $[0,1]$ and k over $\{0,\dots,n\}$. □

As mentioned earlier, Proposition 2 can be interpreted as saying that conformal calibration is a fully adaptive system. Nothing like it holds for the earlier methods, such as Least Squares Prediction Machine [14] or its kernelized versions [12]. Equation (7) implies that

$$C^A_{Z_1,\dots,Z_n,X_{n+1},\tau_{n+1}} \approx A_{X_{n+1}}. \tag{8}$$

The difference between the two sides of (8) is of the order $n^{-1/2}$; this follows from the standard result $n^{1/2}(\mathbb{G}_n - I) \Rightarrow \mathbb{U}$, where $I : [0,1] \to [0,1]$ is the identity function $I(t) = t$, $t \in [0,1]$, and $\mathbb{U}$ is a Brownian bridge (see, e.g., [1, Theorem 16.4]) and the invariance of weak convergence under small perturbations such as η_n (e.g., [1, Theorem 4.1]).

6 Experimental results

In this section we explore whether our conformalization procedure improves the performance of standard predictive systems for artificial and real datasets and how it compares to earlier methods. Following the standard usage, we will often say "training set" and "test set" when the order of elements in a training sequence or test sequence is not important. We begin by considering a standard predictive system and a toy artificial dataset.

The predictive system that we consider is the Nadaraya–Watson predictive system (first introduced in the density form in [6])

$$F(y \mid x) = \frac{\sum_{i=1}^n H\left(\frac{y-y_i}{h}\right) G\left(\frac{x-x_i}{g}\right)}{\sum_{i=1}^n G\left(\frac{x-x_i}{g}\right)}, \tag{9}$$

where we will take H to be the sigmoid distribution function

$$H(u) := \frac{1}{1+e^{-u}}$$

and G the Gaussian kernel

$$G(u) := \frac{1}{\sqrt{2\pi}} e^{-u^2/2}.$$

There are two positive parameters in (9), g and h.

The labels y_i are generated as

$$y_i := 2x_i + \epsilon_i,$$

where the objects x_i are drawn from the uniform distribution on $[-1,1]$, ϵ_i is Gaussian noise with mean 0 and standard deviation $|x_i|/2$, and x_i and ϵ_i,

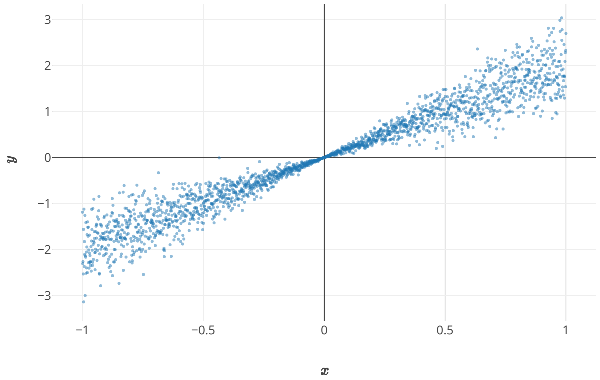

Figure 1: The toy training set.

$i = 1, 2, \dots,$ are all independent. A training set proper of size 2000 is shown in Figure 1.

The quality of predictions will be measured by a popular loss function known as CRPS (continuous ranked probability score). The CRPS loss suffered by a distribution function F (as prediction) on a real number y (as label) is

$$\mathrm{CRPS}(F, y) := \int_{-\infty}^{\infty} \left(F(u) - \mathbf{1}_{\{u \ge y\}}\right)^2 \, \mathrm{d}u, \tag{10}$$

where $\mathbf{1}$ stands for the indicator function. It takes its minimal value 0 when $F = \mathbf{1}_{[y,\infty)}$, and it is ∞ when F has a fat tail. The loss of a sequence of distribution functions F_i on a test sequence (x_i, y_i), $i = 1, \dots, l$, is measured by the average

$$\frac{1}{l} \sum_{i=1}^{l} \mathrm{CRPS}(F_i, y_i),$$

where F_i is the predictive distribution function output for the label of the ith test observation (found from the training set and the test object) and y_i is the true label of the ith test observation.

Our definition (2) gives a function typically ranging between $\frac{\tau}{n-m+1} \approx 0$ and $\frac{n-m+\tau}{n-m+1} \approx 1$, and for the purpose of computing CRPS we turn it into a function ranging between 0 and 1 by applying the appropriate (unique) linear transformation to its values.

The left panel of Figure 2 shows the loss, averaged over 1000 test observations, of the Nadaraya–Watson predictive system (9) for various values of parameters g and h. The right panel shows the loss of the Nadaraya–Watson predictive system calibrated using a separate calibration sequence of size 1000. We can see that calibration improves the performance of the base predictive system for a wide range of parameter values.

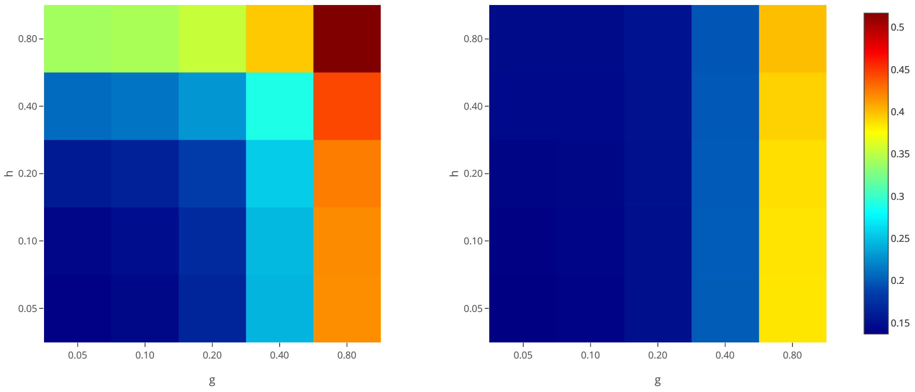

Figure 2: Performance (in the sense of CRPS) of the Nadaraya–Watson predictive system (left) and its split-conformalized version (right) for a range of g and h.

Next, we extend our analysis using different versions of the same toy dataset shown in Figure 1 and applying four different base predictive systems: random forest regression (RF), Gaussian process with a radial basis function kernel (GRBF), Gaussian process with a Matérn kernel (GM), and TensorFlow probability module (TF). We apply each of the four predictive systems to the artificially generated toy dataset across four different scenarios, designed to test whether our procedure improves the performance of standard predictive systems:

Normal (Norm): the labels y_i are generated as $y_i := 2x_i + \epsilon_i$, where ϵ_i is Gaussian noise with mean 0 and standard deviation 0.5 for the training and test sets, and the training and test objects are drawn from the uniform distribution on $[-1, 1]$.

Heteroscedasticity (Het): the labels y_i are generated as $y_i := 2x_i + \epsilon_i$, where ϵ_i is Gaussian noise with mean 0 and standard deviation $|x_i|/2$, and the training and test objects x_i are drawn from the uniform distribution on $[-1, 1]$; x_i and ϵ_i, $i = 1, 2, \dots$, are all independent. See Figure 1.

Heteroscedasticity and covariate shift (HetCov$_1$): the labels y_i are generated as $y_i := 2x_i + \epsilon_i$, where ϵ_i is Gaussian noise with mean 0 and standard deviation 0.5 for the training set and mean 0 and standard deviation 2 for the test set, the training objects are drawn from the uniform distribution on $[-1, 0]$, and the test objects are drawn from the uniform distribution on $[0, 1]$.

Heteroscedasticity and covariate shift (HetCov$_2$): the labels y_i are generated as $y_i := 2x_i + \epsilon_i$, where ϵ_i is Gaussian noise with mean 0 and standard deviation $|x_i|/2$, the training objects are drawn from the uniform distribution on $[-1,0]$, and the test objects are drawn from the uniform distribution on $[0,1]$.

For each of these four scenarios we generate a total of $n+l=500$ observations with $n=400$ for the training set and $l=100$ for the test set. For split conformal prediction we further divide the training set into a training set proper of size m and a calibration set of size $n-m$ with a ratio of $m/n \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. We train the RF predictive system by using 3-fold cross-validation on the full training set with a range of hyperparameters and using individual tree predictions to construct probabilistic predictions by allocating an equal weight to each individual tree. Alongside the Matérn and radial basis function kernels, each Gaussian process predictive system also contains constant and white noise kernels with default parameters with the hyperparameter for `lengthscale` for the former two set to the number of input features (which is the recommended practice). The RF, GRBF, and GM algorithms have been programmed using the Python `scikit-learn` library, and TensorFlow probability TF has been sourced from a recently released `TensorFlow 2.0` module [3]. For TF, we assume a 4-layer sequential network with the first two layers containing the number of densely connected neurons equal to the number of features and the third layer containing a densely connected neuron with two outputs to the probabilistic layer, one for mean and the second for variance. The source code for the experiments, programmed in Python 3.7, can be found on GitHub (`https://github.com/ip200/conformal-calibrators.git`).

In the rest of this section we discuss three groups of prediction algorithms:

- the four base predictive systems, as described in the previous paragraph;
- the SCPS corresponding to the conformity scores (1), where $\hat{\sigma} := 1$ and $\hat{y}$ is the mean of the predictive distribution output by one of the four base predictive systems (for all four base predictive systems $\hat{y}$ is defined unambiguously; e.g., it is the mean prediction of the component decision trees in the case of the RF predictive system); we will refer to them as *nSCPS*, where "n" is a reminder that these RPS produce predictive distribution whose shape is not (sufficiently) adaptive;
- the SCPS corresponding to the four base predictive systems, as described in Section 3; we will refer to them as *aSCPS* (where "a" stands for "adaptive").

The results in Table 1 show the comparison of median CRPS values with $m/n = 0.5$ for the three groups of prediction algorithms. It is interesting that, whereas calibration typically improves the performance of predictive algorithms, the more adaptive method is not obviously better. In addition, Figure 3 shows scatter plots of CRPS values across all splits for the nSCPS and aSCPS methods. In the three cases where heteroscedasticity is present the more adaptive method tends to work better for difficult observations, i.e., those with higher losses (represented by points towards the North-East in each of the four plots).

		base	nSCPS	aSCPS
Norm	RF	0.1209	**0.1066**	0.1193
	GM	0.0964	0.0947	**0.0932**
	GRBF	**0.1127**	0.1160	0.1158
	TF	0.1138	0.0942	**0.0931**
Het	RF	0.1323	**0.1289**	0.1370
	GM	0.0940	0.0914	**0.0891**
	GRBF	0.0917	**0.0813**	0.0818
	TF	0.1339	0.1046	**0.1035**
HetCov$_1$	RF	1.2750	**1.1417**	1.2074
	GM	0.9135	1.0355	**0.9055**
	GRBF	**0.9657**	1.0179	0.9816
	TF	1.1857	0.9425	**0.7746**
HetCov$_2$	RF	0.8738	**0.7325**	0.8001
	GM	0.2551	0.2561	**0.2503**
	GRBF	0.1288	**0.1155**	0.1225
	TF	0.1084	**0.0961**	0.1210

Table 1: Median CRPS values for the base predictive systems, nSCPS, and aSCPS for the artificial datasets with $m/n = 0.5$. In each row the best result is set in boldface

For our real-life prediction problem we apply our method to the prediction of total number of ferry passengers using a dataset from Stena Line. Each year the company operates a large number of sailing routes, and one of their goals is to predict the final number of passengers at departure some time ahead of sailing. The dataset contains transformed and standardised input features for the route: the number of days ahead of departure, the total number of bookings and the corresponding passengers booked to date, the month, week, and day of the week of the departure, whether the departure is occurring during a weekend or a special event, and the ferry identifier. The dataset covers a total of four years of sailing for a representative route (namely, Gothenburg, Sweden – Kiel, Germany) with three years' worth of data (randomly chosen) as the training set ($n = 94{,}691$) and one year as the test set ($l = 31{,}795$). We apply the methods of nSCPS and aSCPS to each of the four base predictive systems described above using a range of splits for the training set proper of size m and calibration set of size $n - m$ with a ratio of $m/n \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ to a randomly sampled set of 1000 training and 100 test observations, with the experiment repeated 10 times.

Table 2 shows the comparison of the CRPS values between the nSCPS and aSCPS methods and the base predictive systems (applied to the full training set of size n). It is interesting that the less adaptive method of calibration works better for this particular dataset (this was also our experience for many benchmark datasets). This is true for a range of fractions m/n used for the training set proper (the optimal value of m/n will depend on the size of the

	m/n				
	0.1	*0.25*	*0.5*	*0.75*	*0.9*
RF	14.285	14.560	16.098	18.397	20.667
GM	12.759	13.274	14.229	16.358	16.373
GRBF	13.776	14.486	20.711	17.091	23.251
TF	41.218	41.591	41.287	41.191	41.285
RF	13.130	12.636	13.817	**15.343**	17.478
GM	**12.165**	**12.473**	**13.409**	15.588	**15.323**
GRBF	13.107	13.543	20.440	16.955	22.002
TF	17.561	17.340	17.038	17.657	16.819
RF	15.756	16.267	17.364	19.537	22.005
GM	12.326	12.705	13.844	15.845	15.824
GRBF	13.439	13.620	20.478	16.971	21.870
TF	17.669	17.465	17.358	17.819	17.180

Table 2: CRPS values for the base predictive systems (first four rows), nSCPS (next four rows), and aSCPS (last four rows) for the Stena Line passenger dataset. In each column the best result is set in boldface

dataset).

Figure 4 shows the calibration curves for $m/n = 0.5$. Each calibration curve plots the percentage of values $F_i(y_i)$ that are less than or equal to p (on the vertical axis) against $p \in [0, 1]$ (on the horizontal axis), where F_i is the predictive distribution output for the label of the ith test observation and y_i is the true label of the ith test observation. The improvement in calibration is particularly noticeable for TF; this is the base predictive method that can be seen to benefit from calibration greatly in Table 2.

7 Conclusion

This paper proposes fully adaptive versions of split conformal predictive systems and discusses their validity and efficiency. The provable property of efficiency (established in Section 5) is that, if the underlying predictive system is already ideal, conformalizing it with our new method will not make it worse (or at least significantly worse). When the underlying predictive system is not ideal, as in Section 6, our proposed fully flexible method does not always outperform the older less flexible methods. Asymptotically, as the size of the training set tends to infinity, fully flexible methods achieve optimal performance [10], but for moderate sized datasets it appears that restricting flexibility can provide useful regularization. This is an interesting phenomenon that needs to be understood and explored further.

There are many other directions of further research, including:

- applying conformal calibration to a wider range of artificial and benchmark datasets;

- analyzing the predictive performance of conformal calibration conditional on the test object x; optimizing conditional performance might require using Mondrian (namely, object-conditional) conformal predictive systems and their modifications;
- analyzing the predictive performance of conformal calibration when applied to benchmark time series and in other non-IID situations.

Acknowledgments

Thanks to Claus Bendtsen and the rest of the AstraZeneca team and to Philip Dawid for useful discussions. We are grateful to the COPA 2020 reviewers for their comments. This work has been supported by Stena Line, AstraZeneca (grant number R10911, "Machine Learning for Chemical Synthesis"), and Centrica.

References

[1] Patrick Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1968.

[2] Evgeny Burnaev and Vladimir Vovk. Efficiency of conformalized ridge regression. *JMLR: Workshop and Conference Proceedings*, 35:605–622, 2014. COLT 2014.

[3] Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A. Saurous. TensorFlow distributions. Technical report, arXiv:1711.10604 [cs.LG], November 2017.

[4] Richard M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, revised edition, 2002.

[5] Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

[6] Murray Rosenblatt. Conditional probability density and regression estimators. In Paruchuri R. Krishnaiah, editor, *Multivariate Analysis II*, pages 25–31. Academic Press, New York, 1969.

[7] Tore Schweder and Nils L. Hjort. *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge University Press, Cambridge, 2016.

[8] Jieli Shen, Regina Liu, and Minge Xie. Prediction with confidence—a general framework for predictive inference. *Journal of Statistical Planning and Inference*, 195:126–140, 2018.

[9] Galen R. Shorack and Jon A. Wellner. *Empirical Processes with Applications to Statistics*. Wiley, New York, 1986.

[10] Vladimir Vovk. Universally consistent predictive distributions, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 18, August 2019 (first posted April 2017).

[11] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[12] Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alex Gammerman. Conformal predictive distributions with kernels, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 20, January 2019 (first posted October 2017).

[13] Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Valery Manokhin, and Alex Gammerman. Computationally efficient versions of conformal predictive distributions, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 22, November 2019 (first posted March 2018).

[14] Vladimir Vovk, Jieli Shen, Valery Manokhin, and Minge Xie. Nonparametric predictive distributions based on conformal prediction, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 17, March 2019 (first posted April 2017).

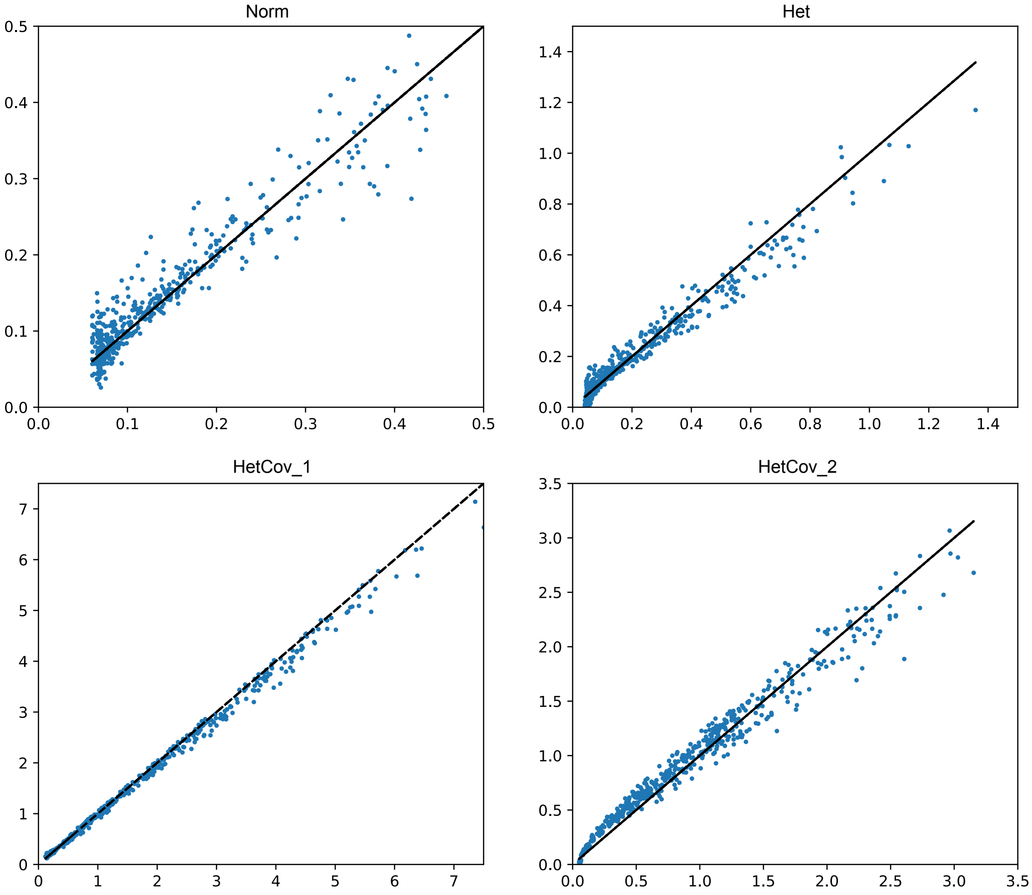

Figure 3: CRPS comparison between nSCPS and aSCPS methods applied to the RF base predictive system across the four different datasets for all splits, where the horizontal axis represents the nSCPS and the vertical axis the aSCPS

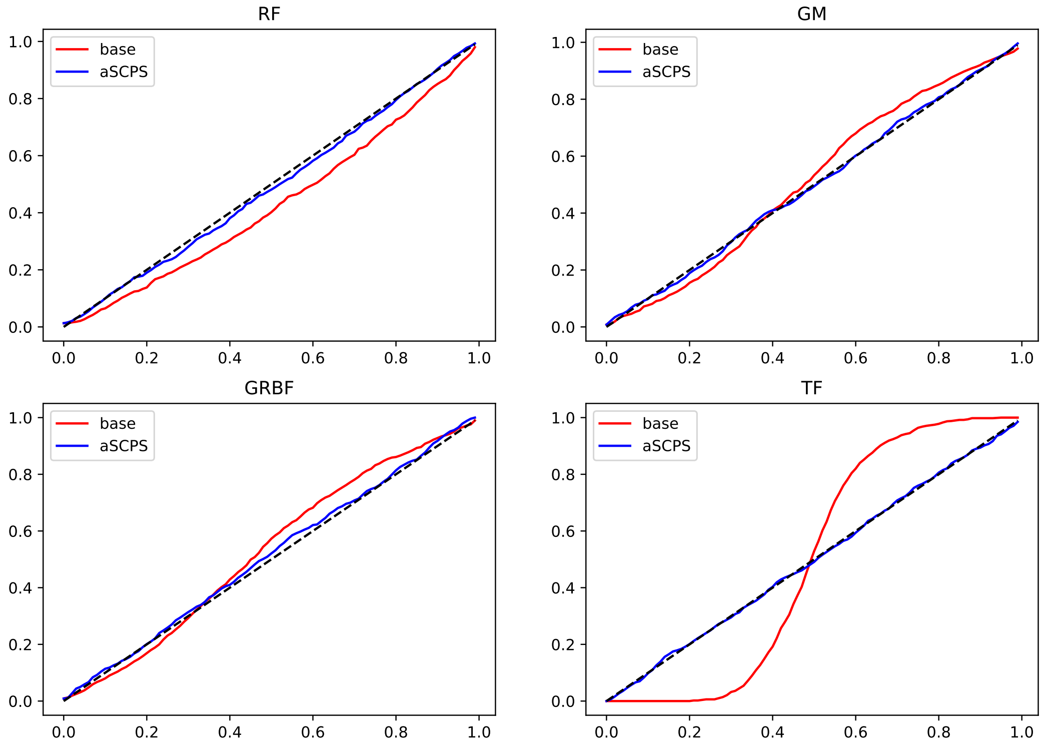

Figure 4: Calibration curves for the base predictive systems and aSCPS for the Stena Line passenger dataset