Conformal calibrators

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 February 19, 2019.

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 and cross-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. It is interesting that the method may also work without the IID assumption, standard in conformal prediction.

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 [3]. 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. 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, both asymptotically and in practical problems (cf. Proposition 1 below). Whereas the emphasis in [10] is on asymptotic optimality only, one of the purposes of this paper is to propose practically useful solutions.

This is a very preliminary version of the paper; we plan to submit a more mature and self-contained version to COPA 2019. For now we will freely use the terminology and notation of [13].

2 Predictive systems and randomized predictive systems

Let us fix (until Section 6) a nonempty measurable space $\mathbf{X}$, which 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 $Q : \mathbf{Z}^{n+1} \to [0, 1]$ is called a *predictive system* (PS) if:

1. For 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 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* 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 $Q : \mathbf{Z}^{n+1} \times [0, 1] \to [0, 1]$ is called a *randomized predictive system* (RPS) if:

1. For 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 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 [10, 12–14]: in this paper we follow the terminology of [7, Chapter 12] rather than [8].

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:
$$\forall \alpha \in [0, 1] : \mathbb{P}\left(Q(Z_1, \dots, Z_n, Z, \tau) \le \alpha\right) = \alpha. \tag{2}$$
(This was included as Requirement R2 in the definition of an RPS in [10,12–14].)

Algorithm 1 Split-Conformal Calibrator

Require: A training sequence $(x_i, y_i) \in \mathbf{Z}$, $i = 1, \dots, n$, and $m < n$.
Require: A test object $x \in \mathbf{X}$ and random number $\tau \in [0, 1]$.
  for $i \in \{1, \dots, n-m\}$ **do**
    $p_i := A(z_1, \dots, z_m, z_{m+i})$
  end for
  sort $p_1, \dots, p_{n-m}$ in the increasing order obtaining $p_{(1)} < \dots < p_{(k)}$
  for $j \in \{1, \dots, k\}$ **do**
    $n_j := \left|\left\{i = 1, \dots, n-m \mid p_i = p_{(j)}\right\}\right|$
    $m_j := \sup\{y \mid A(z_1, \dots, z_m, (x, y)) < p_{(j)}\}$
    $M_j := \inf\{y \mid A(z_1, \dots, z_m, (x, y)) > p_{(j)}\}$
  end for
  return the predictive distribution C^A given by (4) for the label y of x.

3 Split-conformal calibrators

When considered as a split conformity measure each predictive system is balanced and isotonic (at least if we ignore its values 0 and 1), which makes it possible to apply Proposition 3.1 in [13].

If A is a predictive system, the *split-conformalized predictive system* (SCPS) corresponding to A is defined as follows (following the definition of a split-conformal transducer in [13]). 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* $z_{m+1}, \dots, z_n$; we are given a test object x. The output of C^A is defined as

$$C^A(z_1, \dots, z_n, (x, y), \tau) := \frac{1}{n-m+1} \left|\{i = m+1, \dots, n \mid \alpha_i < \alpha^y\}\right| + \frac{\tau}{n-m+1} \left|\{i = m+1, \dots, n \mid \alpha_i = \alpha^y\}\right| + \frac{\tau}{n-m+1}, \quad (3)$$

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

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

For simplicity, let us assume that A never takes values 0 and 1. By [13, Proposition 3.1], every split-conformalized predictive system is an RPS. The functional mapping predictive systems to the corresponding split-conformalized predictive systems are *split-conformal calibrators*.

The SCPS C^A can be implemented by directly coding the definition (3) 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 affect, e.g., CRPS); we can set the probability interval $\operatorname{conv}(\{C^A(z_1, \dots, z_n, (x, y), \tau) \mid \tau \in [0, 1]\})$ at those points y to the union of the probability intervals at the adjacent points without a substantial change to the

predictive system. Some of the p_i, $i = 1, \ldots, n-m$, may coincide, so we can only say that $k \in \{1, \ldots, n-m\}$ (notice that the sequence $p_{(j)}$, $j = 1, \ldots, k$, is strictly increasing). The predictive distribution that it outputs is

$$C^A(z_1, \ldots, z_n, (x, y), \tau) = \begin{cases} \frac{\tau}{n-m+1} & \text{if } y < m_1 \\ \frac{n_1 + \cdots + n_{j-1} + \tau n_j + \tau}{n-m+1} & \text{if } y \in (m_j, M_j),\ j \in \{1, \ldots, k\} \\ \frac{n_1 + \cdots + n_j + \tau}{n-m+1} & \text{if } y \in (M_j, m_{j+1}),\ j \in \{1, \ldots, k-1\} \\ \frac{n_1 + \cdots + n_k + \tau}{n-m+1} = \frac{n-m+\tau}{n-m+1} & \text{if } y > M_k. \end{cases} \tag{4}$$

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

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

Notice that 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 the sigmoid function to (1) to make sure it takes values in $[0, 1]$.

Split-conformal predictive systems are automatically calibrated in probability, in the sense of satisfying (2), under the IID assumption. If F is the distribution function produced for a test object X^*, $F := C^A_{Z_1,\ldots,Z_n,X^*,\tau}$, 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 $Z_i^* = (X_i^*, Y_i^*)$, $i = 1, \ldots, k$, $F_i(Y_i^*)$ will not be independent, even though distributed uniformly on $[0, 1]$, where $F_i := C^A_{Z_1,\ldots,Z_n,X_i^*,\tau_i}$ 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^*, $i = 1, \ldots, k$, sequentially and adding Z_i^* to the calibration sequence as soon as it is processed. (This assumes that Z_i, Z_i^*, and τ_i are all independent.) This remark might be useful for debugging implementations of split-conformal calibrators.

4 Cross-conformal calibrators

We can easily combine several split-conformal calibrators into a cross-conformal calibrator, exactly in the same way as in [13, Section 4]. The resulting RPS will lose automatic calibration in probability (2) but will use the available data more efficiently.

5 Conformalizing ideal predictive systems

In this section we will explore the efficiency of conformal calibrators 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. (This is the question asked in a slightly different context independently by Evgeny Burnaev and Larry Wasserman.) If, for any A, C^A is almost as good as A, we can say that our conformal calibrator is fully adaptive.

In this section we only consider the IID case. Let P be the true probability measure on $\mathbf{Z}$ generating the observations $z_1, z_2, \ldots$. A *conditional distribution function* for P is a right-continuous function $A : \mathbf{Z} \to [0,1]$ satisfying, for each $y \in \mathbb{R}$,

$$A(X, y) = \mathbb{P}(Y \le y \mid X) \quad \text{a.s.} \tag{5}$$

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., [2, Theorem 10.2.2]).

Consider a sequence $\xi_1, \xi_2, \ldots$ of independent and uniformly distributed random variables $\xi_i \sim U$. Let $\mathbb{G}_n$ be the empirical distribution function of $\xi_1, \ldots, \xi_n$; 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 as base predictive system (roughly, this corresponds to an infinitely long training sequence proper, $m = \infty$). The corresponding *ideal conformalized predictive system* (ICPS) is defined as

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

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 CCPS.

The following two propositions say that C^A will be close to A and that the distance between them will be of order $n^{-1/2}$.

Proposition 1. *Suppose the conditional distribution function $A_x := A(x, \cdot)$ (for the true probability measure) is continuous and strictly increasing for almost all $x \in \mathbf{X}$. Then the ICPS C^A satisfies*

$$\left(C^A_{Z_1, \ldots, Z_n, X, \tau} \circ A_X^{-1} \right)_{n=1}^{\infty} \overset{d}{=} (\mathbb{G}_n + \eta_n)_{n=1}^{\infty},$$

where $\overset{d}{=}$ 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,

$$\begin{aligned} C^A_{Z_1, \ldots, Z_n, X, \tau} \left(A_X^{-1}(t) \right) = {} & \frac{1}{n+1} \left| \{ i \in \{1, \ldots, n\} \mid A_{X_i}(Y_i) < t \} \right| \\ & + \frac{\tau}{n+1} \left| \{ i \in \{1, \ldots, n\} \mid A_{X_i}(Y_i) = t \} \right| + \frac{\tau}{n+1} = \frac{k}{n+1} + \frac{\tau}{n+1}, \end{aligned}$$

where the second equality holds almost surely and

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

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

$$\sup_{\tau \in [0,1], k \in \{0,\dots,n\}} \left| \frac{k}{n+1} + \frac{\tau}{n+1} - \frac{k}{n} \right| = \frac{1}{n+1}. \qquad \square$$

Corollary 2. *Suppose the conditional distribution function A_x is continuous and strictly increasing for almost all $x \in \mathbf{X}$. Then the ICPS C^A corresponding to A approaches A in the sense of*

$$\sqrt{n} \left(C^A_{Z_1,\dots,Z_n,X,\tau} \circ A^{-1}_{Z_1,\dots,Z_n,X} - I \right) \Rightarrow \mathbb{U} \quad \text{as } n \to \infty, \tag{6}$$

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.

Proof. This follows from Proposition 1 by the standard result $n^{1/2}(\mathbb{G}_n - I) \Rightarrow \mathbb{U}$ about the weak convergence of empirical processes (see, e.g., [1, Theorem 16.4]) and the invariance of weak convergence under small perturbations (e.g., [1, Theorem 4.1]). $\square$

According to (6), the speed of convergence of C^A to A is indeed $O(n^{-1/2})$. This speed of convergence is the same as for the Dempster–Hill procedure [14, Section 5.1]. In the case of Gaussian y_i (and with x_i absent), this is stated in [14, Theorem 4], but it is true without any parametric assumptions. Notice that the Dempster–Hill procedure is a special case of our procedure corresponding to x_i absent and any continuous and strictly increasing

$$A_{z_1,\dots,z_m,x} = A_{y_1,\dots,y_m} = A$$

(the first equality saying that x_i are absent and the second being our restriction on A). Since conformity measures A and $\phi(A)$ lead to the same conformal transducer provided ϕ is strictly increasing, we can just set $A(y) := y$, $y \in \mathbb{R}$.

6 Experimental results

The main question that we plan to explore in this section is whether our conformalization procedure improves the performance of standard predictive systems for artificial and benchmark data sets. (Alternatively, it might happen that standard predictive systems are calibrated or almost calibrated automatically, and the extra calibration step does not help.) In this version of the paper we only consider one standard predictive system and one toy artificial data set.

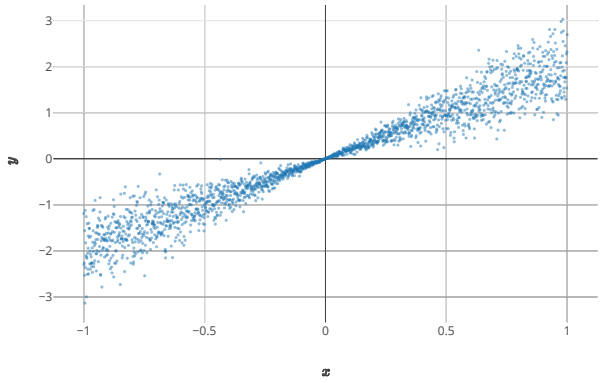

Figure 1: The toy training set.

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

$$F(y \mid x) = \frac{\sum_{i=1}^{n} \sigma\left(\frac{y-y_i}{h}\right) G\left(\frac{x-x_i}{g}\right)}{\sum_{i=1}^{n} G\left(\frac{x-x_i}{g}\right)}, \tag{7}$$

where we will take σ to be the sigmoid distribution function

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

and G the Gaussian kernel

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

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 $|x_i|/2$, and the objects x_i are drawn from the uniform distribution on $[-1, 1]$; x_i and ϵ_i, $i = 1, 2, \ldots$, are all independent. A training set of size 2000 is shown in Figure 1.

The loss of predictions is measured by CRPS (continuous ranked probability score), as described in [13, Section 7]. The left panel of Figure 2 shows the loss, averaged over 1000 test observations, of the Nadaraya–Watson predictive system (7) 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.

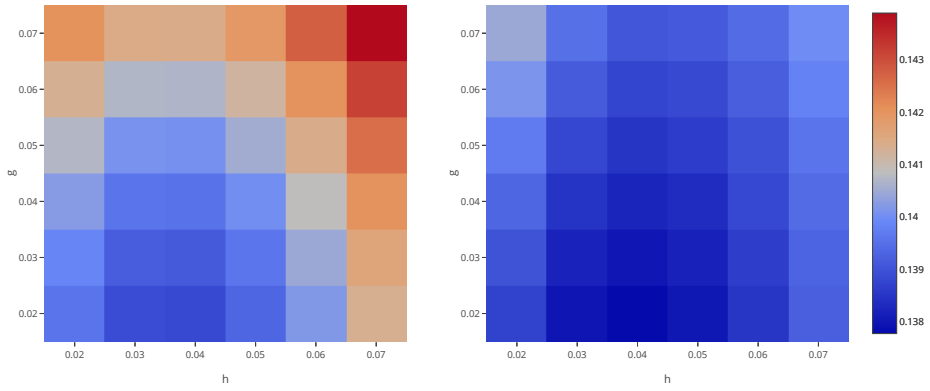

Figure 2: Performance of the Nadaraya–Watson predictive system (left) and its conformalized version (right) for a range of g and h.

7 Calibration without the IID assumption

A standard assumption 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]). Therefore, it is interesting that Proposition 1 continues to hold in the absence of this assumption. Indeed, the proof only depends on the probability integral transforms $A_{X_i}(Y_i)$ being distributed uniformly on $[0,1]$ and independent, which does not require the IID assumption. This is a well-known fact going back to Lévy [4, Section 39], who only assumed that the distribution functions A_x are continuous for all $x \in \mathbf{X}$. (Modern papers usually refer to Rosenblatt [5], who disentangled Lévy's argument from his concern with the foundations of probability, but Rosenblatt referred to Lévy [4] in his paper.)

To see an example where the conformalization procedure works very well in the absence of the IID assumption, suppose the base PS outputs the predictive distribution $\phi(A_x)$ for each test object x, where A is the true conditional distribution function (defined by (5)) and $\phi : [0,1] \to [0,1]$ is a very non-linear increasing function, such as $\phi(u) := u^2$. (So that the base PS has perfect resolution but is badly miscalibrated.) By Corollary 2, the conformalized version (not depending on ϕ) of the base PS will quickly converge to A, whereas the base PS will always remain poor.

On the negative side, in the absence of the IID assumption conformal calibrators have no validity guarantees.

8 Conclusion

There are many directions of further research, including:

- applying conformal calibrators to a wider range of artificial data and to benchmark datasets;
- analyzing the predictive performance of conformal calibrators conditional on the test object x; optimizing conditional performance might require using Mondrian (namely, object-conditional) conformal calibrators and their modifications;
- analyzing the predictive performance of conformal calibrators 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. This work has been supported by AstraZeneca (grant number R10911, "Machine Learning for Chemical Synthesis") and Centrica.

References

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

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

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

[4] Paul Lévy. *Théorie de l'addition des variables aléatoires*. Gauthier-Villars, Paris, 1937. Second edition: 1954.

[5] Murray Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23:470–472, 1952.

[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, September 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, October 2017.

[13] Vladimir Vovk, Ilia Nouretdinov, Valery Manokhin, and Alex Gammerman. Cross-conformal predictive distributions, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 22, 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, April 2017.