Valid Inferential Models for Prediction in Supervised Learning Problems

Leonardo Cella LOLIVEI@NCSU.EDU
Ryan Martin RGMARTI3@NCSU.EDU
Department of Statistics, North Carolina State University, USA

Abstract

Prediction, where observed data is used to quantify uncertainty about a future observation, is a fundamental problem in statistics. Prediction sets with coverage probability guarantees are a common solution, but these do not provide probabilistic uncertainty quantification in the sense of assigning beliefs to relevant assertions about the future observable. Alternatively, we recommend the use of a *probabilistic predictor*, a fully-specified (imprecise) probability distribution for the to-be-predicted observation given the observed data. It is essential that the probabilistic predictor is reliable or valid in some sense, and here we offer a notion of validity and explore its implications. We also provide a general inferential model construction that yields a provably valid probabilistic predictor, with illustrations in regression and classification.

Keywords: classification; conformal prediction; plausibility contour; random sets; regression

1. Introduction

Data-driven prediction of future observations is a fundamental problem. Here our focus is on so-called *supervised learning* applications, where the data $Z = (X,Y)$ consists of a feature or explanatory variable $X \in \mathbb{R}^d$, for some $d \geq 1$, and a label or response variable $Y \in \mathbb{Y}$; the complementary case of *unsupervised learning* is when data consists of only labels Y. The two most common supervised learning examples are *regression* and *classification*, where $\mathbb{Y}$ is an open and finite subset of $\mathbb{R}$, respectively. We consider both regression and classification in what follows.

The focus of our investigation is prediction. That is, we observe a collection $Z^n = \{Z_i = (X_i,Y_i) : i = 1,\ldots,n\}$ of n feature–label pairs from an exchangeable process P, a value x_{n+1} for the next feature X_{n+1}, and the goal is to predict the corresponding Y_{n+1}. By "prediction" here we mean a Z^n-dependent quantification of uncertainty about the value of Y_{n+1}, given $X_{n+1} = x_{n+1}$. Most commonly, this quantification of uncertainty is carried out by producing a suitable family of prediction sets representing collections of sufficiently plausible values for Y_{n+1}; see (12) below. While prediction sets are practically useful tools, there are prediction-related tasks that they cannot perform, in particular, it cannot assign degrees of belief (or betting odds, etc.) to all relevant assertions $A \subseteq \mathbb{Y}$ about Y_{n+1}. An alternative approach is to develop what we refer to here as a *probabilistic predictor*, i.e., a fully-specified (imprecise) probability distribution on $\mathbb{Y}$, depending on Z^n and x_{n+1}, designed to quantify uncertainty about Y_{n+1} by directly assigning degrees of belief to relevant assertions; see (7) below. The most common approach to probabilistic predictor construction is Bayesian, where a prior distribution for P is specified and uncertainty is quantified by the posterior predictive distribution of Y_{n+1}, given Z^n and $X_{n+1} = x_{n+1}$. Other non-Bayesian approaches leading to predictive distributions include Lawless and Fredette (2005), Coolen (2006), Wang et al. (2012), and Vovk et al. (2018).

The advantage of probabilistic predictors, compared to prediction sets, is that they provide belief assignments to all sorts of assertions about Y_{n+1}. So there would be no motivation to go to the trouble of constructing a probabilistic predictor in a given application, compared to simply reporting prediction sets, unless having these belief assignments were a high priority. The typical property boasted by probabilistic predictors, however, is that prediction sets derived from them achieve the nominal coverage probability, at least approximately. That is, no claims are made about the validity or reliability of the probabilistic predictor's belief assignments. If belief assignments are a priority in applications, then we ought to have a way to directly assess the reliability of a probabilistic predictor's belief assignments.

For prediction in the context of unsupervised learning, Cella and Martin (2020) introduced a notion of validity for probabilistic predictors. Roughly, their validity condition requires that the event "the probabilistic predictor, depending on the observed data, assigns a relatively high degree of belief to A *and* $Y_{n+1} \notin A$" has relatively low probability; a more precise statement is given in Definition 1 below. It turns out this notion of validity has some important consequences, including some constraints on the mathematical structure of the probabilistic predictor. Indeed, our analysis indicates that in order for a probabilistic predictor to achieve validity in realistic applications, it must be an imprecise probability. The main goals of this paper are

- to develop an analogous validity property for prediction in supervised learning applications;
- to investigate the consequences of validity;

- to provide a construction of a probabilistic predictor that achieves thie validity property;
- and to illustrate it in regression and classification.

The probabilistic predictor we construct here is largely based on the general construction of a valid *inferential model* (IM) as described in Martin and Liu (2013, 2015b). The setup in the aforementioned references assumes a parametric family of distributions for Y or for Y given X—the prediction problem under such assumptions was addressed in Martin and Lingham (2016). Here, however, we aim to avoid such parametric assumptions and, for this, we use particular extension of the so-called *generalized IM* approach developed in Martin (2015, 2018). The basic idea is that a link—or association—between observable data, quantities of interest, and an unobservable auxiliary variable with known distribution can be made without fully specifying the data-generating process. Like the the conformal prediction approach of Vovk et al. (2005) and others, we establish this association using only the assumption of exchangeability, hence we can avoid any parametric model assumptions. There is also an interesting connection between conformal prediction and our proposed solution.

The remainder of the paper is organized as follows. In Section 2, the validity property for probabilistic predictors is defined and its consequences are investigated. After a brief background on the general IM theory, a generic construction from which the derived probabilistic predictor would be provably valid is given in Section 3. The specifics of this construction are presented in Section 4, in the context of regression. In Section 5, we show that the discreteness of Y in classification problems may cause the IM random set output, from which the probabilistic predictor is derived, to be empty with positive probability. Two possible adjustments are provided, with the one based on "stretching" the random set being most efficient. Finally, Section 6 gives some concluding remarks.

2. Prediction Validity

Recall that there is an exchangeable process $Z_1, Z_2, \ldots$, with distribution P, where each Z_i is a pair $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{Y}$. Given the observed data Z^n and a value x_{n+1} of X_{n+1}, the goal is to reliably predict the corresponding Y_{n+1}. As discussed in Section 1, one way to perform reliable prediction is via prediction sets that achieve the nominal coverage probability. However, the continued interest in the construction of probability distributions for predicting Y_{n+1} implies that prediction-related tasks other than the construction of prediction sets are practically relevant. In particular, quantifying uncertainty about claims of the form "$Y_{n+1} \in A$" in a reliable way ought to be desirable. To formalize this, we follow Cella and Martin (2020) and define a *probabilistic predictor* as a map $(z^n, x) \mapsto (\underline{\Pi}_x^n, \overline{\Pi}_x^n)$ that takes data z^n and a new feature $X_{n+1} = x$ to a pair of lower and upper predictive probabilities for the corresponding Y_{n+1}; for notational simplicity, the probabilistic predictor's dependence on the data z^n is encoded in the superscript "n" only. Then uncertainty quantification about Y_{n+1}, given z^n and $X_{n+1} = x$, is provided by the function $A \mapsto (\underline{\Pi}_x^n(A), \overline{\Pi}_x^n(A))$.

We are defining the probabilistic predictor for all n, but it could be that some minimum sample size is needed in order to properly define it. For example, if some standardization procedure is being employed, then it would be necessary to have $n \geq 2$. In what follows, if n is smaller than the necessary sample size, then we will take the probabilistic predictor to be vacuous, i.e., assign lower and upper probabilities 0 and 1, respectively, to every assertion.

In principle, it is easy to construct a probabilistic predictor. However, its practical utility requires that the uncertainty quantification derived from it be reliable in a certain sense. The particular sense we have in mind is *statistical*, but see below for some behavioral consequences. That is, we require that inferences drawn based on the probabilistic predictor not be systematically misleading. For example, in a particular application, with data z^n and candidate value $X_{n+1} = x$, suppose $\overline{\Pi}_x^n(A)$ happened to be small. Then the data analyst would be inclined to infer "$Y_{n+1} \notin A$." But if it happened that $(Z^n, X_{n+1}) \mapsto \overline{\Pi}_{X_{n+1}}^n(A)$ had a tendency to be small even in cases where $Y_{n+1} \in A$, then this probabilistic predictor—and, hence, the data analyst using it—would be making unreliable predictions. To protect the data analyst from this risk, we impose the following condition on probabilistic predictors, ensuring that those aforementioned undesirable events are controllably rare.

In what follows, discrete uniform distributions will become relevant. So, in certain places where one might see a threshold "α" in the case of continuous variables, we will instead need

$$k_n(\alpha) = (n+1)^{-1} \lfloor (n+1)\alpha \rfloor, \quad \alpha \in [0,1], \tag{1}$$

where $\lfloor a \rfloor$ is the greatest integer less than or equal to a. Of course, $k_n(\alpha) \leq \alpha$ and $k_n(\alpha) \approx \alpha$ when n is large, so this adjustment is of little practical consequence. Nevertheless, it is needed to make the theory precise.

Definition 1 *The probabilistic predictor* $(Z^n, x) \mapsto (\underline{\Pi}_x^n, \overline{\Pi}_x^n)$ *is* valid *if*

$$\mathsf{P}\{\overline{\Pi}_{X_{n+1}}^n(A) \leq k_n(\alpha), Y_{n+1} \in A\} \leq \alpha, \tag{2}$$

for all $\alpha \in [0,1]$, *all* $A \subseteq \mathbb{Y}$, *all* $n \geq 1$, *and all distributions* P *for the exchangeable process* $Z_1, Z_2, \ldots$.

Given the duality between the lower and upper probabilities, i.e., $\underline{\Pi}_x^n(A) = 1 - \overline{\Pi}_x^n(A^c)$, and the fact that (2) is required to hold for all A, there is an equivalent definition of validity in terms of the probabilistic predictor's lower probability:

$$\mathsf{P}\{\underline{\Pi}_{X_{n+1}}^n(A) > 1 - k_n(\alpha), Y_{n+1} \notin A\} \leq \alpha.$$

The version in terms of the upper probability is more intuitive for us, so that is the one we focus on.

The key point, again, is that validity ensures the probabilistic predictor will not tend to assign small upper probability to assertions about Y_{n+1} that happen to be true. Practically, this ensures that the data analyst is not making systematically misleading predictions.

The validity condition has implications beyond this notion of "reliability." In particular, it has some behavioral consequences in the sense of de Finetti, Walley, and others. One example of this is the following, a generalization of the result presented in Cella and Martin (2020).

To state the result precisely, define

$$\beta_n(A) = \sup_{z^n, x_{n+1}} \overline{\Pi}^n_{x_{n+1}}(A),$$

the upper probabilistic predictor evaluated at A, maximized over all of its data inputs. Then a particularly gross misspecification of prediction probabilities is a situation where

$$\beta_n(A) < \mathsf{P}(Y_{n+1} \in A), \quad \text{for some } A. \tag{3}$$

This leads to *sure loss* in the sense of, e.g., Condition (C7) in Walley (1991, Sec. 6.5.2) or Definition 3.3 in Gong and Meng (2020). Fortunately, as we show below, validity in the sense of Definition 1 implies no sure loss.

Proposition 2 *Suppose that the probabilistic predictor falls victim to sure loss in the sense that* (3) *holds, and that n is sufficiently large. Then validity in the sense of Definition 1 fails.*

Proof Define the function

$$\xi_n(A,\alpha) = \mathsf{P}\{\overline{\Pi}^n_{X_{n+1}}(A) \leq k_n(\alpha), Y_{n+1} \in A\},$$

so that (2) is equivalent to

$$\xi_n(A,\alpha) \leq \alpha \quad \forall\, (A,\alpha,n,\mathsf{P}). \tag{4}$$

Using iterated expectation, by conditioning on (Z^n, X_{n+1}), it is easy to see that $\xi_n(A,\alpha)$ equals

$$\mathsf{E}\big[1\{\overline{\Pi}^n_{X_{n+1}}(A) \leq k_n(\alpha)\}\,\mathsf{P}(Y_{n+1} \in A \mid Z^n, X_{n+1})\big],$$

where $1\{E\}$ is the indicator function of the event E. From this alternative representation of $\xi_n(A,\alpha)$ in the above display, it is also clear that

$$\xi_n(A,\alpha) \geq 1\{\beta_n(A) \leq k_n(\alpha)\}\mathsf{P}(Y_{n+1} \in A). \tag{5}$$

According to (3), for sufficiently large n, there exists an assertion A and a threshold α such that

$$\beta_n(A) < k_n(\alpha) \leq \alpha < \mathsf{P}(Y_{n+1} \in A).$$

[We need n large so that $k_n(\alpha) \approx \alpha$ also fits in the range defined by (3).] Then from (5), with this choice of (A,α),

$$\xi_n(A,\alpha) \geq \mathsf{P}(Y_{n+1} \in A) > \alpha.$$

Then (4) and, hence, (2) fails, so the claim follows. ■

This *validity implies no sure loss* property helps to provide a behavioral interpretation to validity, which connects it to the more familiar properties in the imprecise probability literature. In fact, a yet-to-be-settled conjecture (Cella and Martin, 2020) is that the only valid probabilistic predictor with output being a precise probability distribution on $\mathbb{Y}$ is the true conditional distribution, i.e.,

$$\underline{\Pi}^n_x(A) = \overline{\Pi}^n_x(A) = \mathsf{P}(Y_{n+1} \in A \mid Z^n, X_{n+1} = x).$$

If this conjecture is true, then this and the fact that the conditional distribution above is generally unavailable—because P is unknown—implies that valid prediction can only be achieved by an imprecise probability model. In the context of statistical inference, the *false confidence theorem* (Balch et al., 2019; Martin, 2019) implies that precise probabilities fail to be valid, and the above conjecture suggests a similar conclusion in the context of prediction. Moreover, it also seems that the kinds of imprecise probability models that can achieve validity in general are very special, namely, those whose upper probabilities satisfy a consonance property; see, e.g., Martin (2021).

3. Inferential Models

The IM approach aims to provide valid, in a sense similar to that described in Section 2 above, data-dependent uncertainty quantification about unknown quantities of interest. It has connections to various other approaches to statistical inference, some that quantify uncertainties with ordinary probabilities, e.g., Bayesian inference (Martin and Liu, 2015a, Remark 4), fiducial inference (Fisher, 1935; Taraldsen and Lindqvist, 2013), generalized fiducial inference (Hannig et al., 2016); and imprecise probabilities, e.g., Dempster–Shafer theory (Dempster, 2014, 1967, 1968, 2008; Shafer, 1976), other belief function frameworks (Denœux and Li, 2018; Denœux, 2014). The IM construction is composed of three steps. The A-step *associates* the observable data and unknown quantity of interest with an unobservable auxiliary variable whose distribution is fully known. In the early work on IMs, this association was usually a complete description of the data-generating process. However, Martin (2015, 2018) showed that this can be actually generalized. Suitable functions relating the three aforementioned components, namely observable data, unknown quantity of interest and auxiliary variable, is all that is needed in the A-step. The P-step introduces a random set that aims to *predict* or guess the unobserved value of the auxiliary variable. Easy to arrange properties of this user-specified random set ensure that the guessing of the auxiliary variable is done in a reliable way, which turns out to be fundamental for validity. Next, the C-step *combines* the results of the A- and P-steps, yielding a new, data-dependent random set on the space where the quantity of interest resides. Finally, this random set's distribution determines lower and upper probabilities that can be used to assign degrees of belief and plausibility to any relevant assertion about the unknown quantities of interest. Below we describe the general construction in more detail.

A-step *Suppose there exists a real-valued function ϕ_n such that the distribution of $\phi_n(Z^n, Z_{n+1})$ is known, i.e., does not depend on the unknown* P. *That distribution may depend on n, so denote it by Q_n. Then associate the observable data Z^n and the yet-to-be-observed Z_{n+1} with the unobservable auxiliary variable U as follows:*

$$\phi_n(Z^n, Z_{n+1}) = U, \quad U \sim \mathsf{Q}_n. \tag{6}$$

For our case where $Z_{n+1} = (X_{n+1}, Y_{n+1})$ and interest is in Y_{n+1}, the association defines a set-valued mapping

$$(Z^n, x, u) \mapsto \mathbb{Y}_x^n(u) = \{y \in \mathbb{Y} : \phi_n(Z^n, (x,y)) = u\}.$$

P-step *Define a nested random set $\mathcal{S}$ on the space $\mathbb{U}$ of the auxiliary variable U, designed to reliably contain realizations of $U \sim \mathsf{Q}_n$ in the precise sense described in (10) below. We will denote the distribution of $\mathcal{S}$ by $\mathsf{Q}_{n,\mathcal{S}}$.*

C-step *Combine the results of the A- and P-steps to get a new, data-dependent random set*

$$\mathbb{Y}_x^n(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \mathbb{Y}_x^n(u) = \{y \in \mathbb{Y} : \phi_n(Z^n, (x,y)) \in \mathcal{S}\}.$$

And the distribution of this new random set determines the probabilistic predictor for Y_{n+1}, i.e.,

$$\begin{aligned} \underline{\Pi}_x^n(A) &= \mathsf{Q}_{n,\mathcal{S}}\{\mathbb{Y}_x^n(\mathcal{S}) \subseteq A\} \\ \overline{\Pi}_x^n(A) &= \mathsf{Q}_{n,\mathcal{S}}\{\mathbb{Y}_x^n(\mathcal{S}) \cap A \neq \varnothing\}. \end{aligned} \tag{7}$$

Remark 3 *If $\mathbb{Y}_x^n(\mathcal{S})$ is empty with positive $\mathsf{Q}_{n,\mathcal{S}}$-probability, then some adjustment to the probabilistic predictor in (7) is needed. This will be relevant for the classification problem in Section 5.*

The above construction is abstract for the purpose of generality. The challenge is in identifying the function ϕ_n. Specific constructions will be given in Sections 4–5 below. Other examples were explored previously in Martin and Lingham (2016) where P was assumed to belong to a specified parametric family of distributions. The additional structure provided by the parametric family makes it possible to borrow much of the theory in Martin and Liu (2015b). Here, however, no parametric assumptions about P are being made, so different techniques are required. The remainder of this section investigates the general properties of the abstract prediction IM construction above.

The random set $\mathcal{S}$ is assumed to be nested in the sense that, for any two sets in its support, one is a subset of the other. As a consequence, the derived probabilistic predictor is a consonant plausibility function (Shafer, 1976, Ch. 10) or, equivalently, a possibility measure (Dubois and Prade, 1988), which means it is completely determined by its corresponding plausibility contour function. That is, define

$$\pi_x^n(y) = \mathsf{Q}_{n,\mathcal{S}}\{\mathbb{Y}_x^n(\mathcal{S}) \ni y\}, \quad y \in \mathbb{Y}. \tag{8}$$

Then the probabilistic predictor's upper probability can be equivalently written as

$$\overline{\Pi}_x^n(A) = \sup_{y \in A} \pi_x^n(y), \quad A \subseteq \mathbb{Y}. \tag{9}$$

This alternative expression is important for at least two reasons. First, since the plausibility contour is an ordinary function, which makes it relatively easy to visualize and compute with compared to a set-function. Second, the consonance property appears to be fundamental to achieving validity both in the inference and prediction contexts; see, for example, Martin (2021).

It remains to establish that the probabilistic predictor resulting from the above construction is, in fact, valid. This requires stating the required conditions on $\mathcal{S}$ more precisely. Since the cases in the following sections involve an auxiliary variable U that is discrete, we will focus on the discrete case. For the corresponding theory when U has a continuous distribution, see Martin and Liu (2015b). First, define the random set's containment function

$$f(u) = \mathsf{Q}_{n,\mathcal{S}}(\mathcal{S} \ni u), \quad u \in \mathbb{U}.$$

Then the required link between Q_n and $\mathsf{Q}_{n,\mathcal{S}}$ is that

$$\text{if } U \sim \mathsf{Q}_n, \text{ then } f(U) \sim \mathsf{Unif}(\mathcal{I}_{n+1}), \tag{10}$$

where $\mathcal{I}_{n+1} = \{1, \ldots, n, n+1\}$, so that the uniform distribution on the right is of the discrete variety.

Theorem 4 *If the random set $\mathcal{S}$ in the above construction satisfies (10), and if $\mathbb{Y}_x^n(\mathcal{S})$ is non-empty with $\mathsf{Q}_{n,\mathcal{S}}$-probability 1 for P-almost all (Z^n, x), then the plausibility contour (8) that characterizes the probabilistic predictor satisfies*

$$\mathsf{P}\{\pi_{X_{n+1}}^n(Y_{n+1}) \leq k_n(\alpha)\} \leq \alpha, \tag{11}$$

where the P-probability is with respect to the joint distribution of (Z^n, Z_{n+1}), with $Z_{n+1} = (X_{n+1}, Y_{n+1})$.

Proof First, for Z^n and $Z_{n+1} = (X_{n+1}, Y_{n+1})$, set $U = \phi_n(Z^n, Z_{n+1})$. Then it is easy to see that

$$\mathbb{Y}_{X_{n+1}}^n(\mathcal{S}) \ni Y_{n+1} \iff \mathcal{S} \ni U.$$

The $\mathsf{Q}_{n,\mathcal{S}}$-probability of the left- and right-hand side events is $\pi_{X_{n+1}}^n(Y_{n+1})$ and $f(U)$, respectively, so these two random variables—the first as a function of $(Z^n, Z_{n+1}) \sim \mathsf{P}$ and the second as a function of $U \sim \mathsf{Q}_n$—have the same distribution. Equation (10) states that $f(U)$ is uniform and, therefore, so is $\pi_{X_{n+1}}^n(Y_{n+1})$, which proves the claim. ■

Corollary 5 *Under the conditions of Theorem 4, the probabilistic predictor in (7) is valid in the sense of Definition 1.*

Proof Take any $A \subseteq \mathbb{Y}$ and note that, by (9),

$$\{\overline{\Pi}^n_{X_{n+1}}(A) \leq k_n(\alpha), Y_{n+1} \in A\} \implies \pi^n_{X_{n+1}}(Y_{n+1}) \leq k_n(\alpha).$$

Therefore, since the right-hand side has P-probability less than α from (11), the validity property follows. ■

Two brief remarks concerning Theorem 4 and its consequences. First, the non-emptiness condition is not necessary for validity, but some adjustment is needed to the definition in (7), as mentioned in Remark 3, to address this. We will discuss this in Section 5. Second, note that the property (11) is stronger than the notion of validity in Definition 1. In fact, (11) is analogous to the familiar property satisfied by p-values, a key element to most, if not all, methods for inference and prediction with frequentist error rate guarantees. Naturally, there is another important consequence concerning the coverage probability of prediction regions.

Given $\alpha \in [0,1]$, use the plausibility contour (8) that characterizes the IM output in (7) to define a $100(1-\alpha)\%$ prediction plausibility set

$$\mathcal{P}^n_\alpha(x) = \{y \in \mathbb{Y} : \pi^n_x(y) > k_n(\alpha)\} \tag{12}$$

Then the following prediction coverage probability result follows immediately from Theorem 4.

Corollary 6 *The prediction plausibility set defined in (12) is a genuine $100(1-\alpha)\%$ prediction set in the sense that it satisfies*

$$\mathsf{P}\{\mathcal{P}^n_\alpha(X_{n+1}) \ni Y_{n+1}\} \geq 1-\alpha.$$

It is important to point out that the kind of validity being considered here is *marginal*, which is easiest to understand in the context of Corollary 6. That is, the conditional coverage probability of the prediction set is

$$x_{n+1} \mapsto \mathsf{P}\{\mathcal{P}^n_\alpha(x_{n+1}) \ni Y_{n+1} \mid X_{n+1} = x_{n+1}\},$$

a function of x_{n+1}. Then the validity property implies that the expected value of this function, with respect to the marginal distribution of X_{n+1} under P, is at least $1-\alpha$. This marginal coverage guarantee, of course, says nothing about the conditional coverage at any particular x_{n+1} values. Conditional validity is both challenging and practically relevant, and we discuss this briefly in Section 6.

4. Probabilistic Prediction in Regression

Recall that the A-step requires the specification of a real-valued function ϕ_n, such that the distribution of $\phi_n(Z^n, Z_{n+1})$ is known. Towards this, given $Z^{n+1} = (Z^n, Z_{n+1})$ consisting of the observable (Z^n, X_{n+1}) and the yet-to-be-observed Y_{n+1}, consider first a transformation $Z^{n+1} \to T^{n+1}$, defined by

$$T_i = \Psi(Z^{n+1}_{-i}, Z_i), \quad i \in \mathcal{I}_{n+1}, \tag{13}$$

where $Z^{n+1}_{-i} = Z^{n+1} \setminus \{(Y_i, X_i)\}$, and Ψ is a suitable real-valued function that compares Y_i to a prediction derived from Z^{n+1}_{-i} at X_i, being small if they agree and large if they disagree. For example, to each Z^{n+1}_{-i}, one could fit a linear or non-linear regression model to get an estimated mean response $\hat{\mu}^{n+1}_{-i}(X_i)$ and take T_i as the corresponding absolute residual

$$T_i = \big|Y_i - \hat{\mu}^{n+1}_{-i}(X_i)\big|, \quad i \in \mathcal{I}_{n+1}. \tag{14}$$

The critical property of Ψ is that it be symmetric in the elements of its first vector argument. This symmetry guarantees that the assumed exchangeability in $Z_1, Z_2, \ldots$ is preserved when Z^{n+1} get mapped to T^{n+1}. As T_i depends on the entire data Z^{n+1}, we will write $T_i(Z^{n+1})$ where necessary to highlight that dependence. A well-known consequence of exchangeability of $T_1, \ldots, T_{n+1}$ is that their ranks are marginally distributed according to $\mathsf{Unif}(\mathcal{I}_{n+1})$, a discrete uniform distribution on $\mathcal{I}_{n+1}$,

Having identified a function of (Z^n, Z_{n+1}) whose distribution is known, we can complete the A-step of the IM construction by writing a version of (6) as follows:

$$r(T_{n+1}) = U, \quad U \sim \mathsf{Unif}(\mathcal{I}_{n+1}), \tag{15}$$

where $r(\cdot)$ is the ascending ranking operator. The choice of T_{n+1} instead of any of the other T_i's in (15) is simply because T_{n+1} is the one that holds the to-be-predicted value, Y_{n+1}, in special status. Note that, while it appears this expression only depends on T_{n+1}, it does implicitly depend on all the T_i's and, hence, all of Z^{n+1}, through the ranking procedure. In summary, to complete the A-step, the only task for the data analyst is the specification of Ψ.

For the P-step, the specification of a nested random set targeting the unobserved realization of the auxiliary variable U, introduced above, is needed. Consider

$$\mathcal{S} = \{1, 2, \ldots, \tilde{U}\}, \quad \tilde{U} \sim \mathsf{Unif}(\mathcal{I}_{n+1}). \tag{16}$$

It is straightforward to show that this random set satisfies the critical calibration property (10). Moreover, this choice also makes intuitive sense, as $\mathcal{S}$ always includes the value 1. This is desirable given the ascending ranking operator in (15) because it implies values of Y_{n+1} that make the residual T_{n+1} small will be assigned high plausibility.

Finally, in the C-step, $\mathcal{S}$ is combined with the u-indexed collection of sets

$$\mathbb{Y}^n_{x_{n+1}}(u) = \{y_{n+1} : r\big(T_{n+1}(z^{n+1})\big) = u\}$$

that arise from the association (15). Here and below, note that z^{n+1} consists of the observed z^n values with $z_{n+1} =$

(x_{n+1}, y_{n+1}) appended to it. The particular combination, as described in the previous section, leads to the following data-dependent random subset of $\mathbb{Y}$:

$$\mathbb{Y}^n_{x_{n+1}}(\mathcal{S}) = \{y_{n+1} : r(T_{n+1}(z^{n+1})) \le \tilde{U}\} \tag{17}$$

It is easy to see that $\mathbb{Y}^n_{x_{n+1}}(\mathcal{S})$'s corresponding contour function for Y_{n+1} is given by

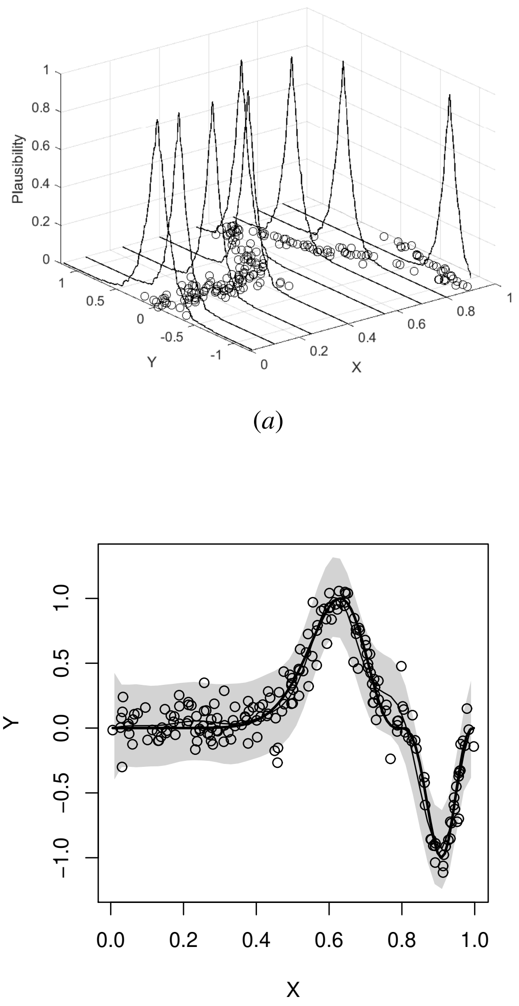
$$\begin{aligned}\pi^n_{x_{n+1}}(y_{n+1}) &= \mathsf{Q}_{n,\mathcal{S}}\{\mathbb{Y}^n_{x_{n+1}}(\mathcal{S}) \ni y_{n+1}\} \\ &= \mathsf{P}_{\tilde{U}}\{\tilde{U} \ge r(T_{n+1}(z^{n+1}))\} \\ &= \frac{1}{n+1}\sum_{i=1}^{n+1} 1\{T_i(z^{n+1}) \ge T_{n+1}(z^{n+1})\}. \end{aligned} \tag{18}$$

As $\mathbb{Y}^n_{x_{n+1}}(\mathcal{S})$ is both nested and non-empty, its contour function above is all that is needed to define a probabilistic predictor and, consequently, quantify uncertainty about any assertion $A \subset \mathbb{Y}$ of interest. For example, an upper probability about A would be given by (9), which can easily be approximated by

$$\overline{\Pi}^n_{x_{n+1}}(A) \approx \max_{y \text{ on a grid and in } A} \pi^n_{x_{n+1}}(y).$$

Validity of the probabilistic predictor derived in this Section is a direct consequence of Theorem 4. Consequently, this probabilistic predictor satisfies (2), so we are guaranteed that the assignment of small (large) upper (lower) probabilities that happen to be true (false) will be controllably rare, which prevents the data analyst from making systematically misleading predictions.

To illustrate the practicality and flexibility of this approach, consider the following example. Let $X_1, \ldots, X_n$ be iid $\mathsf{Unif}(0,1)$, with $n = 200$, and let $Y_1, \ldots, Y_n$ be independent, where $Y_i = \mu(X_i) + 0.1\varepsilon_i$, where $\mu(x) = \sin^3(2\pi x^3)$, and $\varepsilon_1, \ldots, \varepsilon_n$ are iid from a Student-t distribution with df $= 5$. Figure 1 displays the data, the true regression function $\mu(x)$ and the fitted regression curve $\hat{\mu}(x)$ based on a B-spline with 12 degrees of freedom. A 95% prediction band is also displayed, derived by (12) and x_{n+1} taking values along the observed x^{200}.



(*a*)

(*b*)

Figure 1: Panel (a):Data and the plausibility contours at selected values of x. Panel (b): Data, the true mean curve (heavy line), the fitted B-spline regression curve (thin line), and the 95% pointwise prediction band.

We end this section pointing out an important connection between the prediction IM developed here and the powerful *conformal prediction* presented in Vovk et al. (2005) and summarized in Shafer and Vovk (2007). The careful reader may have recognized the Ψ function in the A-step of our construction as the so-called *non-conformity measure*, an essential component in the conformal prediction framework. Moreover, the basic output from the IM construction presented below is the plausibility contour in (18), which is precisely conformal prediction's p-value or transducer. The theory in Vovk et al. (2005) takes this conformal transducer, which satisfies the property (11) in Theorem 4, and constructs a prediction set as in (12) with the prediction coverage probability property as in Corollary 6. Apparently it was recognized only recently (Cella and Martin, 2020) that the conformal prediction output could be converted into a valid probabilistic predictor in the sense of Definition 1, one that can make valid belief assignments, by treating the transducer as the contour of a consonant plausibility function via (9). We refer to this probabilistic predictor construction generically as "conformal + consonance," and all it requires is that the conformal transducer π^n_x satisfy the properties of a contour function, namely, that

$$\sup_y \pi^n_x(y) = 1, \quad \text{for all } (Z^n, x). \tag{19}$$

This is easy to verify in cases like regression where Y is a continuous random variable. Indeed, for the Ψ function in (14), the supremum is attained at $y = \hat{\mu}^{n+1}_{-(n+1)}(x)$. In other cases, like in classification where Y is discrete, the "conformal + consonance" construction is not so straightforward. We discuss these considerations next in Section 5.

5. Probabilistic Prediction in Classification

In Section 4, we found that the A-step boils down to the specification of a suitable real-valued, exchangeability-preserving function Ψ, which Vovk et al. (2005) refer as a non-conformity measure. In binary classification problems, a Ψ function like in (14) can also be used here by encoding the two possible values of Y_i by two different real numbers. However, when $\mathbb{Y}$ has more than two labels and they are not in an ordinal scale where the assignment of different numbers to them is justified, there is no natural way to measure the distance between labels. Consequently, we cannot measure how wrong a prediction is—it is simply right or wrong (Shafer and Vovk, 2007). To circumvent this, Vovk et al. (2005) suggest the following non-conformity measure based on the nearest-neighbor method for classification:

$$\Psi(Z^{n+1}_{-i}, Z_i) = \frac{\min_{j \in \mathcal{I}_{n+1} \setminus \{i\}: Y_j = Y_i} d(X_j, X_i)}{\min_{j \in \mathcal{I}_{n+1} \setminus \{i\}: Y_j \neq Y_i} d(X_j, X_i)}, \qquad (20)$$

where d is the Euclidean distance. In words, $\Psi(Z^{n+1}_{-i}, Z_i)$ is large if X_i is close to an element in X^{n+1}_{-i} with a label different from Y_i and far from any element in X^{n+1}_{-i} with label equal to Y_i. If both the numerator and the denominator in (20) are 0, Shafer and Vovk (2007) recommend taking the ratio also to be 0. Other nonconformity measures suitable for classification problems can be found in Vovk et al. (2005); Shafer and Vovk (2007).

Having identified Ψ, so that Z^{n+1} can be mapped to T^{n+1} preserving exchangeability, the IM construction proceeds analogously to that in the previous section: the A-step is completed by writing (15), the random set (16) is chosen in the P-step to target the unobserved realization of the auxiliary variable U, and, in the C-step, the ingredients in the A- and P-steps are combined to get $\mathbb{Y}^n_{x_{n+1}}(\mathcal{S})$ in (17), a data-dependent random subset of $\mathbb{Y}$. However, due to the discreteness of $\mathbb{Y}$, it is possible that $\mathbb{Y}^n_{x_{n+1}}(\mathcal{S})$ is empty with positive $\mathsf{Q}_{n,\mathcal{S}}$-probability. As discussed in Section 3, in these cases, some adjustment to the probabilistic predictor in (7) is necessary to avoid both violations to the validity condition and counter-intuitive cases where the realized prediction set happens to be empty. There is a sense in which empty prediction sets could be meaningful, but we defer this discussion to Section 6.

There are two available adjustments to account for the potentially empty realizations of the random set $\mathbb{Y}^n_{x_{n+1}}(\mathcal{S})$. The first, and probably most intuitive, is *conditioning* on the event that the random set is non-empty, which happens to be equivalent to Dempster's rule of combination (e.g., Shafer, 1976, Chap. 3). For example, the post-conditioning plausibility contour is given by

$$y_{n+1} \mapsto \mathsf{Q}_{n,\mathcal{S}}\{\mathbb{Y}^n_{x_{n+1}}(\mathcal{S}) \ni y_{n+1} \mid \mathbb{Y}^n_{x_{n+1}}(\mathcal{S}) \neq \varnothing\}.$$

It is easy to see that conditioning simply rescales the original plausibility contour, making it larger at each $y_{n+1} \in \mathbb{Y}$. Clearly, if the unadjusted probabilistic predictor is valid, then this conditioning adjustment—which only inflates its plausibility contour values—cannot affect its validity. This inflation does, however, suggest a potential loss of efficiency, e.g., larger prediction sets in (12).

The second adjustment strategy, designed to preserve validity without sacrificing efficiency, is based on a suitable *stretching* of the original random set; see, e.g., Ermini Leaf and Liu (2012) and Cella and Martin (2019). To see how this stretching strategy works, start by defining the set

$$\mathbb{U}^n_{x_{n+1}} = \bigcup_{y_{n+1} \in \mathbb{Y}} \{r(T_{n+1}(z^n, z_{n+1}))\} \subseteq \mathcal{I}_{n+1}. \qquad (21)$$

There are only finitely many y_{n+1} values, and the set $\mathbb{U}^n_{x_{n+1}}$ defined above is just the collection of ranks that are possible for the given Z^n and x_{n+1}. Note that $\mathbb{Y}^n_{x_{n+1}}(\mathcal{S})$ is empty if and only if $\mathcal{S}$ does not intersect with $\mathbb{U}^n_{x_{n+1}}$. This conflicting situation can be avoided if, instead of $\mathcal{S}$, we adopt a *stretched* random set $\mathcal{S}_e$, obtained through equipping $\mathcal{S}$ with a stretching parameter, $e \geq 0$, that controls how far $\mathcal{S}$ is stretched toward $\mathbb{U}^n_{x_{n+1}}$:

$$\mathcal{S}_e = \{1, 2, \ldots, \tilde{U} + e\}, \quad \tilde{U} \sim \mathsf{Unif}(\mathcal{I}_{n+1}).$$

Following Ermini Leaf and Liu (2012), the parameter e is chosen as the smallest value at which the intersection of $\mathcal{S}_e$ and $\mathbb{U}^n_{x_{n+1}}$ is non-empty, i.e.,

$$\begin{aligned} \hat{e} &= \min\{e : \mathcal{S}_e \cap \mathbb{U}^n_{x_{n+1}} \neq \varnothing\} \\ &= \begin{cases} \min \mathbb{U}^n_{x_{n+1}} - \tilde{U} & \text{if } \tilde{U} < \min \mathbb{U}^n_{x_{n+1}} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Consequently, $\mathcal{S}_{\hat{e}}$ would be

$$\mathcal{S}_{\hat{e}} = \begin{cases} \{1, 2, \ldots, \min \mathbb{U}^n_{x_{n+1}}\} & \text{if } \tilde{U} < \min \mathbb{U}^n_{x_{n+1}} \\ \{1, 2, \ldots, \tilde{U}\} & \text{otherwise.} \end{cases}$$

In summary, in the stretching IM, the IM's original random set output $\mathbb{Y}^n_{x_{n+1}}(\mathcal{S})$ is replaced with $\mathbb{Y}^n_{x_{n+1}}(\mathcal{S}_e)$, and the non-emptiness of $\mathbb{Y}^n_{x_{n+1}}(\mathcal{S}_e)$ makes the probabilistic predictor derived from it valid.

To better understand how the conditioning and stretching adjustments are done in practice, we consider the data in Table 1, taken from Agresti, p. 304, corresponding to the primary food choices and lengths of $n = 39$ male alligators

Length (m)	Choice	Length (m)	Choice
1.30	I	1.80	F
1.32	F	1.85	F
1.32	F	1.93	I
1.40	F	1.93	F
1.42	I	1.98	I
1.42	F	2.03	F
1.47	I	2.03	F
1.47	F	2.31	F
1.50	I	2.36	F
1.52	I	2.46	F
1.63	I	3.25	O
1.65	O	3.28	O
1.65	O	3.33	F
1.65	I	3.56	F
1.65	F	3.58	F
1.68	F	3.66	F
1.70	I	3.68	O
1.73	O	3.71	F
1.78	F	3.89	F
1.78	O		

Table 1: Primary food choice (I, invertebrates; F, fish; O, other) and lengths (in meters) for $n = 39$ male alligators caught in Lake George, Florida (Agresti, p. 304).

caught in Lake George, Florida. Assume the 40th caught alligator is two meters long, i.e., $X_{n+1} = 2$. The goal is to predict Y_{n+1}, its primary food choice.

Note that

$$\mathbb{Y}^n_{x_{n+1}}(\mathcal{S}) = \begin{cases} \{I\} & \text{with probability } 0.1 \\ \{I,F\} & \text{with probability } 0.2 \\ \{I,F,O\} & \text{with probability } 0.3 \\ \varnothing & \text{with probability } 0.4. \end{cases} \tag{22}$$

The corresponding plausibility contour, as given in (8), is represented by the solid lines in Figure 2(a). By thresholding it at any $\alpha > 0.6$ we obtain $100(1-\alpha)\%$ prediction sets that are empty, which is undesirable.

The plausibility contour, conditioned on (22) $\neq \varnothing$ is easy to evaluate, and is represented by the dashed lines in Figure 2(a). To calculate the plausibility contour under the stretching approach, we obtain, after some calculations, $\mathbb{U}^n_{x_{n+1}} = \{17, 21, 29\}$. As $\min \mathbb{U}^n_{x_{n+1}} = 17$,

$$\mathcal{S}_{\hat{e}} = \begin{cases} \{1,2,\ldots,17\} & \text{if } \tilde{U} < 17 \\ \{1,2,\ldots,\tilde{U}\} & \text{otherwise.} \end{cases}$$

where $\tilde{U} \sim \mathsf{Unif}(1,2,\ldots,40)$. Therefore,

$$\mathbb{Y}^n_{x_{n+1}}(\mathcal{S}_{\hat{e}}) = \begin{cases} \{I\} & \text{with probability } 0.5 \\ \{I,F\} & \text{with probability } 0.2 \\ \{I,F,O\} & \text{with probability } 0.3, \end{cases}$$

and the dotted lines in Figure 2(a) illustrate its corresponding plausibility contour. Note, first, that empty prediction sets are eliminated with both the conditioning and the stretching adjustments. Second, for any α, the $100(1-\alpha)\%$ prediction sets derived from the stretching adjustment are no larger than the corresponding ones derived from the conditioning adjustment, which indicates that the former is no less efficient than the latter. Another way to see this is through the difference between the upper and lower probabilities derived by the respective probabilistic predictors. Dempster (2008) referred to this gap as the "don't know" probability. Of course, between two valid probabilistic predictors, the one with less "don't know" is preferred because it is more efficient. Figure 2(b) shows the upper and lower probabilities for the singleton assertions $\{I\}$, $\{O\}$ and $\{F\}$, for both strategies. Clearly, the stretching strategy leads to a more efficient probabilistic predictor.

Recall from Section 4 that the probabilistic predictor derived from the "conformal + consonance" construction is valid according to Definition 1, given that the conformal transducer π^n_x satisfies (19). In regression problems, this condition follows naturally from the continuity of Y, and the derived probabilistic predictor is equivalent to the one that would be obtained from an IM construction (assuming both use the same Ψ function). In classification problems, however, (19) may not hold because Y is discrete. This implies the "conformal + consonance" cannot be applied directly without some adjustment. This is not surprising given that similar adjustments were needed in the IM construction discussed above too.

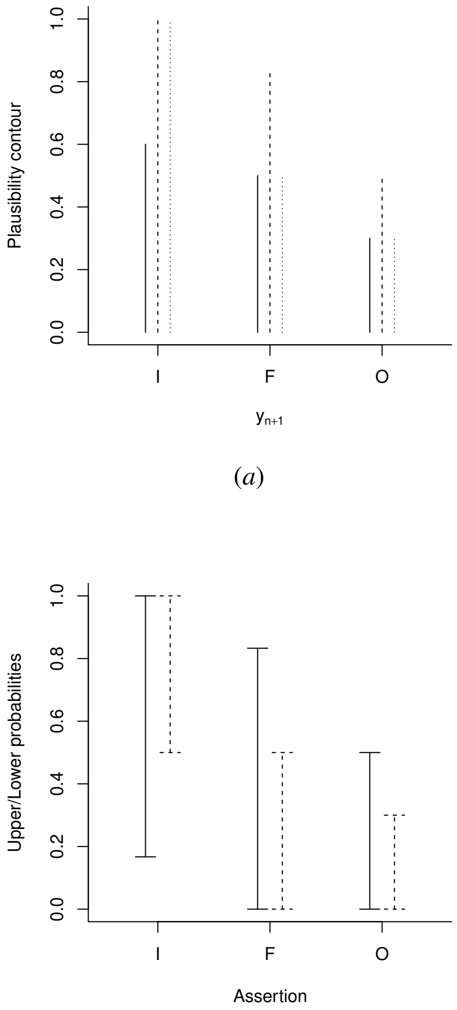
A natural adjustment is to force the conformal transducer to attain the value 1. Consider the two following rescaled conformal transducers:

$$\dot{\pi}^n_x(y) = \frac{\pi^n_x(y)}{\max_y \pi^n_x(y)},$$

and

$$\ddot{\pi}^n_x(y) = \begin{cases} 1 & \text{if } y = \hat{y}, \\ \pi^n_x(y) & \text{otherwise,} \end{cases}$$

where $\hat{y} = \arg\max_y \pi^n_x(y)$ and $y \in \mathbb{Y}$. In words, $\dot{\pi}^n_x(y)$ takes the conformal transducers for the different $y \in \mathbb{Y}$ and divide them by their maximum, and $\ddot{\pi}^n_x(y)$ maintains all the conformal transducers except for their maximum, which is assigned the value 1. The fact that both rescaled transducers reach the value 1 make the probabilistic predictors derived by them, through (9), valid in the sense of Definition 1. It is also easy to see that these probabilistic predictors obtained



(a)

(b)

Figure 2: Panel (a): Plausibility contours in Equation (8), derived from an IM construction with no adjustment (solid lines), conditioning adjustment (dashed lines) and stretching adjustment (dotted lines). Panel (b): Upper and lower probabilities for the singleton assertions {I}, {F} and {O} derived from an IM construction with the conditioning adjustment (solid lines) and the stretching adjustment (dashed lines).

from $\dot{\pi}_x^n(y)$ and $\ddot{\pi}_x^n(y)$ are equivalent to the ones derived from the IM construction with, respectively, the conditioning and the stretching adjustments. This shows that forcing consonance of the conformal transducer is not an ad hoc strategy; it is justified by the corresponding operations on random sets. Moreover, in light of this connection to the IM's random set adjustments, we find that the second adjustment to the conformal predictor, i.e., setting the maximum value equal to 1, is the more efficient adjustment.

6. Conclusion

Here we focused on the important problem of prediction in supervised learning applications with no model assumptions (except exchangeability). We presented a notion of prediction validity, one that goes beyond the usual coverage probability guarantees of prediction sets. This condition assures the reliability of the degrees of belief, obtained from a imprecise probability distribution, assigned to all relevant assertions about the yet-to-be-observed quantity of interest. We also showed that, by following a new variation on the (generalized) IM construction first presented in Martin (2015, 2018), this validity property can be easily achieved. We also noted the connection between this new IM construction and the conformal prediction strategy in, e.g., Vovk et al. (2005), and presented illustrations in both regression and classification settings.

In Section 3 we noted that the IM construction there leads naturally to a notion of *marginal* validity, which is different (and weaker) than the so-called *conditional* validity property. While this is usually framed in the context of prediction sets, the corresponding definition in the context of probabilistic predictors is

$$\mathsf{P}\{\overline{\Pi}_x^n(A) \le k_n(\alpha), Y_{n+1} \in A \mid X_{n+1} = x\} \le \alpha \quad \forall\, x,$$

and, of course, for all $(\alpha, n, A, \mathsf{P})$ as before. Given the impossibility results in, e.g., Lei and Wasserman (2014), it seems unlikely that conditional validity can be achieved by any non-trivial probabilistic predictor. Asymptotic validity is possible, and some promising ideas are given in Chernozhukov et al. (2019).

We mentioned in Section 5 that, surprisingly, empty random sets may have some practical value. This concerns the so-called *open-* versus *closed-world* view of the prediction problem. If the world is closed in the sense that all the possible labels are known, then it makes sense to remove the empty set cases and, hence, force consonance. However, if the world is open in the sense that other labels are possible, then the empty set realization is an indication that the new object being classified may be of previously-unknown type, which itself is valuable information. How this open-world view can be captured by the IM framework developed here remains an open question.

References

A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley.

M. S. Balch, R. Martin, and S. Ferson. Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2227):1–20, 2019.

L. Cella and R. Martin. Incorporating expert opinion in an inferential model while maintaining validity. In Jasper De Bock, Cassio P. de Campos, Gert de Cooman, Erik Quaeghebeur, and Gregory Wheeler, editors, *Proceedings of the Eleventh International Symposium on Imprecise Probabilities: Theories and Applications*, volume 103 of *Proceedings of Machine Learning Research*, pages 68–77, Thagaste, Ghent, Belgium, 03–06 Jul 2019. PMLR. URL http://proceedings.mlr.press/v103/cella19a.html.

L. Cella and R. Martin. Validity, consonant plausibility measures, and conformal prediction. https://researchers.one/articles/20.01.00010, 2020.

V. Chernozhukov, K. Wüthrich, and Y. Zhu. Distributional conformal prediction. arXiv:1909.07889, 2019.

F. P. A. Coolen. On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language, and Information*, 15(1/2):21–47, 2006.

A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statististics*, 38:325–339, 1967. ISSN 0003-4851.

A. P. Dempster. A generalization of Bayesian inference. (With discussion). *Journal of the Royal Statistical Society - Series B*, 30:205–247, 1968. ISSN 0035-9246.

A. P. Dempster. The Dempster–Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365–377, 2008.

A. P. Dempster. Statistical inference from a Dempster–Shafer perspective. In Xihong Lin, Christian Genest, David L. Banks, Geert Molenberghs, David W. Scott, and Jane-Ling Wang, editors, *Past, Present, and Future of Statistical Science*, chapter 24. Chapman & Hall/CRC Press, 2014.

T. Denœux. Likelihood-based belief function: justification and some extensions to low-quality data. *Internat. J. Approx. Reason.*, 55(7):1535–1547, 2014. ISSN 0888-613X.

T. Denœux and S. Li. Frequency-calibrated belief functions: review and new insights. *Internat. J. Approx. Reason.*, 92:232–254, 2018.

D. Dubois and H. Prade. *Possibility Theory*. Plenum Press, New York, 1988.

D. Ermini Leaf and C. Liu. Inference about constrained parameters using the elastic belief method. *Internat. J. Approx. Reason.*, 53(5):709–727, 2012.

R. A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6(4):391–398, 1935.

R. Gong and X.-L. Meng. Judicious judgment meets unsettling updating: Dilation, sure loss, and Simpson's paradox. *Statist. Sci.*, to appear, arXiv:1712.08946, 2020.

J. Hannig, H. Iyer, R. C. S. Lai, and T. C. M. Lee. Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515): 1346–1361, 2016.

J. F. Lawless and M. Fredette. Frequentist prediction intervals and predictive distributions. *Biometrika*, 92(3): 529–542, 2005. ISSN 00063444.

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

R. Martin. Plausibility functions and exact frequentist inference. *Journal of the American Statistical Association*, 110(512):1552–1561, 2015.

R. Martin. On an inferential model construction using generalized associations. *Journal of Statistical Planning and Inference*, 195:105–115, 2018.

R. Martin. False confidence, non-additive beliefs, and valid statistical inference. *International Journal of Approximate Reasoning*, 113:39–73, 2019.

R. Martin. An imprecise-probabilistic characterization of frequentist statistical inference. *Researchers.One*, https://researchers.one/articles/21.01.00002, 2021.

R. Martin and R. T. Lingham. Prior-free probabilistic prediction of future observations. *Technometrics*, 58:225–235, 2016.

R. Martin and C. Liu. Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108:301–313, 2013.

R. Martin and C. Liu. Conditional inferential models: combining information for prior-free probabilistic inference. *Journal of the Royal Statistical Society - Series B*, 77: 195–217, 2015a.

R. Martin and C. Liu. *Inferential Models: Reasoning with Uncertainty*. Monographs in Statistics and Applied Probability Series. Chapman & Hall/CRC Press, 2015b.

G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J, 1976.

G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2007.

G. Taraldsen and B. H. Lindqvist. Fiducial theory and optimal inference. *Annals of Statistics*, 41(1):323–341, 2013.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

V. Vovk, J. Shen, V. Manokhin, and M. Xie. Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 108:445–474, 2018.

P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1991.

C. M. Wang, J. Hannig, and H. K. Iyer. Fiducial prediction intervals. *J. Statist. Plann. Inference*, 142(7):1980–1990, 2012. ISSN 0378-3758.