

Valid inferential models for prediction in supervised learning problems*

Leonardo Cella[†] and Ryan Martin[†]

December 21, 2021

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 data-dependent (imprecise) probability distribution for the to-be-predicted observation given the observed data. It is essential that the probabilistic predictor be reliable or valid, and here we offer a notion of validity and explore its behavioral and statistical implications. In particular, we show that valid probabilistic predictors avoid sure loss and lead to prediction procedures with desirable frequentist error rate control properties. We also provide a general inferential model construction that yields a provably valid probabilistic predictor, and we illustrate this construction in regression and classification applications.

Keywords and phrases: 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 applications where the data $Z = (X, Y)$ consists of explanatory variables $X \in \mathbb{X} \subseteq \mathbb{R}^d$, for some $d \geq 1$, and a response variable $Y \in \mathbb{Y}$. The two most common examples of such applications are *regression* and *classification*, where \mathbb{Y} is an open and finite subset of \mathbb{R} , respectively. We consider both cases in what follows.

More specifically, we observe a collection $Z^n = \{Z_i = (X_i, Y_i) : i = 1, \dots, n\}$ of n pairs from an exchangeable process \mathbf{P} , a value $x_{n+1} \in \mathbb{X}$ for the next explanatory variables X_{n+1} , and then the goal is to predict the corresponding future response $Y_{n+1} \in \mathbb{Y}$. By “prediction” here we mean quantifying uncertainty about Y_{n+1} in a data-dependent way, i.e., depending on the observed data Z^n and the given value x_{n+1} of X_{n+1} . One perspective on prediction uncertainty quantification is the construction of a suitable family of

*This is an extended version of the 2021 International Symposium on Imprecise Probability Theory and Applications (ISIPTA) proceedings paper, Cella and Martin (2021b).

[†]Department of Statistics, North Carolina State University; lolivei@ncsu.edu, rgmarti3@ncsu.edu

prediction sets representing collections of sufficiently plausible values for Y_{n+1} ; see, e.g., Vovk et al. (2005), Campi et al. (2009), Kuleshov et al. (2018), and Equation (21) below. While prediction sets are practically useful, 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 or hypotheses “ $Y_{n+1} \in A$,” for $A \subseteq \mathbb{Y}$. An alternative approach is to develop what we refer to here as a *probabilistic predictor*, i.e., a probability-like structure (precise or imprecise probability) defined 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 Equation (17) below. The most common approach to probabilistic prediction is Bayesian, where a prior distribution for \mathbf{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).

Before moving forward, it is important to distinguish between uncertainty quantification with prediction sets and with probabilistic predictors. One does not need a full (precise or imprecise) probability distribution to construct prediction sets and, moreover, sets derived from a probabilistic predictor are not guaranteed to satisfy the frequentist coverage probability property that warrants calling them genuine “prediction sets.” Therefore, the only possible motivation for going through the trouble of constructing probabilistic predictor, Bayesian or otherwise, is that there are important prediction-related tasks that prediction sets cannot satisfactorily handle. So it must be that the belief assignments provided by a (precise or imprecise) probability are a high priority. Strangely, however, the reliability of probabilistic predictors is only ever assessed in terms of (asymptotic) coverage probability properties of their corresponding prediction sets. Our unique perspective is that, since belief assignments are a priority in applications, there ought to be a way to directly assess the reliability of a probabilistic predictor’s belief assignments.

For prediction problems without explanatory variables, where only the (response) variables Y_1, \dots, Y_n are observed, Cella and Martin (2021c) 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; more precise statements are given in Definitions 1–2 below. It turns out these notions of validity have some important consequences, imposing certain constraints on the mathematical structure of the probabilistic predictor. Indeed, we show that in order for a probabilistic predictor to achieve validity in the sense of Definition 2, and to achieve the desirable behavioral and statistical properties described next, it must take the form of an imprecise probability.

After describing the basic problem setup and introducing these notions of validity, we explore their behavioral and statistical consequences. First, we show that even the weaker validity property in Definition 1 implies that the probabilistic predictor avoids sure loss, hence is not internally irrational from a behavioral point of view. We go on to show that prediction-related “tests” derived by a valid probabilistic predictor control frequentist Type I error. Moreover, under the stronger notion of validity in Definition 2, prediction sets derived from the probabilistic predictor achieve the nominal frequentist coverage probability. The take-away message is that a valid probabilistic predictor provides the “best of both worlds,” in the sense that it allows the data analyst to simultaneously achieve both desirable behavioral and statistical properties.

Given the desirable properties of a valid probabilistic predictor, the natural question is *how to construct one?* 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 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. Section 6 gives some concluding remarks.

2 Prediction validity

2.1 Definitions

Recall that there is an exchangeable process Z_1, Z_2, \dots with distribution \mathbf{P} , where each Z_i is a pair $(X_i, Y_i) \in \mathbb{Z} = \mathbb{X} \times \mathbb{Y}$. The distribution \mathbf{P} is completely unknown, beyond that it is exchangeable and supported on \mathbb{Z}^∞ . 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, a common strategy is to construct so-called *prediction sets* that achieve the nominal frequentist coverage probability. That is, a collection of functions $C_{n,\alpha}$, from $\mathbb{Z}^n \times \mathbb{X}$ to subsets of \mathbb{Y} , indexed by $\alpha \in [0, 1]$ and $n \geq 1$, defines a family of $100(1 - \alpha)\%$ prediction sets for Y_{n+1} if

$$\mathbf{P}\{C_{n,\alpha}(Z^n, X_{n+1}) \ni Y_{n+1}\} \geq 1 - \alpha, \quad \text{for all } \alpha, n, \text{ and } \mathbf{P}, \quad (1)$$

where “for all \mathbf{P} ” means “for all exchangeable distributions \mathbf{P} supported on \mathbb{Z}^∞ .” Note that the probability on the left-hand side above is with respect to the joint distribution of Z^n and $Z_{n+1} = (X_{n+1}, Y_{n+1})$ under \mathbf{P} . This uniformity in \mathbf{P} is needed because the true \mathbf{P} is unknown so we cannot tailor $C_{n,\alpha}$ to the specific \mathbf{P} in order to achieve (1).

However, the continued interest in the construction of a probability distribution to quantify uncertainty about Y_{n+1} implies that there are prediction-related problems that are both practically relevant and not fully/satisfactorily resolved through the use of prediction sets. In particular, quantifying uncertainty about claims of the form “ $Y_{n+1} \in A$,”

for relevant $A \subseteq \mathbb{Y}$, in a reliable way is desirable. To formalize this, we follow Cella and Martin (2021c) and define a *probabilistic predictor* as a map $(z^n, x) \mapsto (\underline{\Pi}_x^n, \overline{\Pi}_x^n)$, where $(\underline{\Pi}_x^n, \overline{\Pi}_x^n)$ is a pair of lower and upper predictive probabilities for the corresponding Y_{n+1} ; for notational simplicity, the probabilistic predictor's dependence on the observed 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 large enough to estimate standard errors. As a rule in what follows, if n is smaller than the necessary sample size, then we will silently take the probabilistic predictor to be vacuous, i.e., assign lower and upper probabilities 0 and 1, respectively, to every assertion.

What kind of mathematical form does the function $A \mapsto (\underline{\Pi}_x^n(A), \overline{\Pi}_x^n(A))$ take? Let \mathcal{A} denote the σ -algebra of subsets of \mathbb{Y} that are measurable with respect to the (common) marginal of the Y_i 's under \mathbf{P} . We will assume that \mathcal{A} is rich enough to contain the singletons, e.g., like the Borel σ -algebra. Then the lower and upper probabilities are capacities defined on \mathcal{A} , i.e., monotone set functions, taking value 1 at \mathbb{Y} and value 0 at \emptyset . However, being “lower” and “upper” suggests a link between the two. We formalize by requiring that, for each z^n and new value x of the feature X_{n+1} , the upper probability $\overline{\Pi}_x^n$ for Y_{n+1} is sub-additive. Then the lower probability $\underline{\Pi}_x^n$ is defined as the dual or conjugate to the upper probability,

$$\underline{\Pi}_x^n(A) = 1 - \overline{\Pi}_x^n(A^c), \quad A \in \mathcal{A}, \quad (2)$$

and from sub-additivity it follows that

$$\underline{\Pi}_x^n(A) \leq \overline{\Pi}_x^n(A), \quad A \in \mathcal{A},$$

hence the name “lower” and “upper” probabilities. Ordinary or precise probabilities are additive—and hence sub-additive—so they satisfy these conditions with $\underline{\Pi}_x^n \equiv \overline{\Pi}_x^n$. Moreover, all of the standard imprecise probability models—belief functions, possibility measures, lower/upper previsions—satisfy these conditions, so our assumptions corresponding to no loss of generality. Since we will be interested in the statistical properties of the probabilistic predictor as functions of the data, we will assume that $(Z^n, X_{n+1}) \mapsto (\underline{\Pi}_{X_{n+1}}^n(A), \overline{\Pi}_{X_{n+1}}^n(A))$ is measurable for each $n \geq 1$ and for each $A \in \mathcal{A}$.

The interpretation of the probabilistic predictor's output is subjective and goes as follows. For given data z^n and a new value x of the feature X_{n+1} , the lower and upper probabilities represent

$$\begin{aligned} \underline{\Pi}_x^n(A) &= \text{maximum buying price for the gamble } \$1(Y_{n+1} \in A) \\ \overline{\Pi}_x^n(A) &= \text{minimum selling price for the gamble } \$1(Y_{n+1} \in A), \end{aligned}$$

where $1(B)$ denotes the indicator of the event B . Therefore, based on data z^n and new feature x , if the investigator's $\underline{\Pi}_x^n(A)$ is large, then she would be inclined to buy the gamble $\$1(Y_{n+1} \in A)$, whereas, if her $\overline{\Pi}_x^n(A)$ is small, then she would be inclined to sell the gamble $\$1(Y_{n+1} \in A)$; otherwise, she might choose to neither buy nor sell the gamble. For this reason, $\underline{\Pi}_x^n(A)$ measures the subjective degree of belief and $\overline{\Pi}_x^n(A)$

the plausibility of the event “ $Y_{n+1} \in A$.” Below we introduce an element of objectivity through a requirement that its predictions be reliable in a statistical sense.

So far, we have imposed effectively no mathematical constraints on the probabilistic predictor, plus its interpretation is subjective, so virtually no construction can be ruled out at this point. However, the probabilistic predictor’s 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. Based on the interpretations of the lower and upper probabilities described above, events of the general form

$$\{(z^n, x_{n+1}, y_{n+1}) : \underline{\Pi}_{x_{n+1}}^n(A) \text{ is large and } y_{n+1} \notin A\}$$

and

$$\{(z^n, x_{n+1}, y_{n+1}) : \overline{\Pi}_{x_{n+1}}^n(A) \text{ is small and } y_{n+1} \in A\},$$

should they occur, put the investigator at risk of making erroneous predictions and incurring losses, monetary or otherwise. To protect the investigator from this risk, we impose the following condition on probabilistic predictors, ensuring that the aforementioned undesirable, risk-creating events are controllably rare.

Definition 1. The probabilistic predictor $(Z^n, x) \mapsto (\underline{\Pi}_x^n, \overline{\Pi}_x^n)$ is *valid* if one and, hence, both of the following equivalent conditions hold:

$$\mathbb{P}\{\underline{\Pi}_{X_{n+1}}^n(A) \geq 1 - \alpha, Y_{n+1} \notin A\} \leq \alpha, \quad \text{for all } (\alpha, n, A, \mathbb{P}) \quad (3)$$

$$\mathbb{P}\{\overline{\Pi}_{X_{n+1}}^n(A) \leq \alpha, Y_{n+1} \in A\} \leq \alpha, \quad \text{for all } (\alpha, n, A, \mathbb{P}). \quad (4)$$

Here “for all $(\alpha, n, A, \mathbb{P})$ ” is short for “for all $\alpha \in [0, 1]$, all $n \geq 1$, all $A \in \mathcal{A}$, and all distributions \mathbb{P} for the exchangeable process Z_1, Z_2, \dots .” The two conditions are equivalent by the duality in (2) and the “for all A ” clause.

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, or large lower probability to assertions about Y_{n+1} that happen to be false. Practically, this ensures that the data analyst is not making systematically misleading predictions.

That the calibration property imposed in (4) is required to hold for all $A \subseteq \mathbb{Y}$ might seem overly strong, but it turns out that there is an even stronger property that is needed and can readily be attained. This stronger property requires validity to hold uniformly in A , not just point-wise in A . In particular, the “for all A ” clause on the outside of the probability statement in (4) will now be moved to the inside, allowing the choice of A in the bound to be data-dependent. This generalization will be important for technical reasons—see Proposition 3 below—but here we give some intuition in the gambling scenario described above. Consider that the agent’s opponents have access to the data (z^n, x) at the time of prediction. This allows them to use the data to make strategic choices about which assertions A to negotiate with the agent. Of course, if the agent’s opponents can make these more sophisticated data-dependent plays and he is only able control errors for predictions specified in advance, then that could make his error rates unacceptably large. The following uniform-in-assertions validity guarantees protection against these strategic choices.

Definition 2. The probabilistic predictor $(Z^n, x) \mapsto (\underline{\Pi}_x^n, \overline{\Pi}_x^n)$ is *uniformly valid* if

$$\mathbb{P}\{\underline{\Pi}_{X_{n+1}}^n(A) \geq 1 - \alpha \text{ and } Y_{n+1} \notin A \text{ for some } A\} \leq \alpha, \quad \text{for all } (\alpha, n, \mathbb{P}) \quad (5)$$

$$\mathbb{P}\{\overline{\Pi}_{X_{n+1}}^n(A) \leq \alpha \text{ and } Y_{n+1} \in A \text{ for some } A\} \leq \alpha, \quad \text{for all } (\alpha, n, \mathbb{P}). \quad (6)$$

As the name and the discussion above suggest, uniform validity in the sense of Definition 2 is stronger than validity in the sense of Definition 1. Indeed, the “for some A ” inside the probability statement in (6) is effectively a union of A -dependent events like those in (4) over all A . So if the union over A of these A -dependent events has probability bounded by α , then so would any individual event in that union. In general, this union is uncountable so one might anticipate issues with measurability, but it turns out that there are no such issues; see Section 2.3 below.

2.2 Practical implications

The validity conditions above have a number of interesting implications. Below are two results of a very different nature. First, despite validity’s focus on frequentist-style operating characteristics, it turns out that it has some important behavioral consequences, à la de Finetti, Walley, and others. One example of this is Proposition 1, a generalization of the result presented in Cella and Martin (2021c).

To state the result precisely, define

$$\underline{\gamma}_n(A) = \inf_{z^n, x} \underline{\Pi}_x^n(A) \quad \text{and} \quad \overline{\gamma}_n(A) = \sup_{z^n, x} \overline{\Pi}_x^n(A),$$

the lower/upper probabilistic predictor evaluated at A , minimized/maximized over all of its data inputs. Then an especially poor specification of prediction probabilities is a situation where, for some $A \subseteq \mathbb{Y}$,

$$\underline{\gamma}_n(A) > \mathbb{P}(Y_{n+1} \in A) \quad \text{or} \quad \overline{\gamma}_n(A) < \mathbb{P}(Y_{n+1} \in A). \quad (7)$$

Ideally, the probabilistic predictor would mimic the true conditional probability at least in the sense that its average over data inputs would not be far from $\mathbb{P}(Y_{n+1} \in A)$. So a situation like in (7), where the probabilistic predictor is uniformly bounded away from $\mathbb{P}(Y_{n+1} \in A)$ in one way or the other, is a clear sign of trouble. More precisely, the undesirable outcome in (7) 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 (2021). Fortunately, as we show below, the validity property in Definition 1 implies no sure loss.

Proposition 1. *Suppose that the probabilistic predictor, $(z^n, x) \mapsto (\underline{\Pi}_x^n, \overline{\Pi}_x^n)$, suffers from sure loss in the sense that (7) holds for some $A \subseteq \mathbb{Y}$. Then validity in the sense of Definition 1 fails.*

Proof. We present the argument here for the case where $\overline{\gamma}_n(A) < \mathbb{P}(Y_{n+1} \in A)$; the argument for the $\underline{\gamma}_n(A)$ bound is very similar. For the assertion A in (7), define

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

so that (4) is equivalent to

$$\xi_n(A, \alpha) \leq \alpha \quad \text{for all } (A, \alpha, n, \mathbf{P}). \quad (8)$$

Using iterated expectation, by conditioning on (Z^n, X_{n+1}) , it is easy to see that

$$\xi_n(A, \alpha) = \mathbb{E}[1\{\bar{\Pi}_{X_{n+1}}^n(A) \leq \alpha\} \mathbf{P}(Y_{n+1} \in A \mid Z^n, X_{n+1})]. \quad (9)$$

From this alternative representation of $\xi_n(A, \alpha)$, it is also clear that

$$\xi_n(A, \alpha) \geq 1\{\bar{\gamma}_n(A) \leq \alpha\} \mathbf{P}(Y_{n+1} \in A). \quad (10)$$

According to (7), there exists an $A \subseteq \mathbb{Y}$ and a threshold $\alpha \in [0, 1]$ such that

$$\bar{\gamma}_n(A) < \alpha < \mathbf{P}(Y_{n+1} \in A).$$

Then from (10), with this choice of (A, α) ,

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

Then (8) and, hence, (4) fails, so the claim follows. \square

It is interesting to see that a frequentist calibration property can have meaningful behaviorist consequences. This gives mathematical support to the following intuition: a method that is externally reliable across applications ought not be internally irrational in any one application.

The second implication concerns the more classical frequentist-style prediction properties, and we will consider two different questions. The first concerns testing certain “hypotheses” about the next observation Y_{n+1} . For example, an investor may want to sell a certain asset when the price exceeds some fixed level, say y^* . So she would like to quantify uncertainty about an assertion or hypothesis of the form $Y_{n+1} \in A$ for $A = [0, y^*]$ and, in particular, decide if the new price being below the y^* threshold is too plausible to warrant taking quick action to sell. The following proposition establishes that the test

$$\text{reject “} Y_{n+1} \in A \text{” if and only if } \bar{\Pi}_x^n(A) \leq \alpha, \quad (11)$$

derived from a valid probabilistic predictor controls frequentist Type I error at level α .

Proposition 2. *If the probabilistic predictor $(z^n, x) \mapsto (\underline{\Pi}_x^n, \bar{\Pi}_x^n)$ is valid in the sense of Definition 1, then the test described in (11) controls frequentist Type I error at level α in the sense that*

$$\mathbf{P}\{\text{test based on } (Z^n, X_{n+1}) \text{ rejects and } Y_{n+1} \in A\} \leq \alpha.$$

Proof. This is an immediate consequence of the definition of validity. In particular, compare the Type I error probability above to the left-hand-side of (4). \square

Perhaps a more common—and arguably more challenging—prediction-related task is the construction of a prediction set, i.e., a set of sufficiently plausible values for Y_{n+1}

given the observed data. A natural way to construct a candidate prediction set from a probabilistic predictor is as follows:

$$C_{n,\alpha}(z^n, x) = \bigcap \{A : \underline{\Pi}_x^n(A) \geq 1 - \alpha\}. \quad (12)$$

That is, $C_{n,\alpha}(z^n, x)$ is the smallest assertion A about Y_{n+1} to which the probabilistic predictor assigns lower probability at least $1 - \alpha$. This is consistent with strategies used in the Bayesian literature for constructing posterior credible regions for inference and prediction. The following proposition shows that uniform validity implies that this is a genuine $100(1 - \alpha)\%$ prediction set in the sense that its frequentist coverage probability is at least the advertise/nominal level $1 - \alpha$.

Proposition 3. *If the probabilistic predictor $(z^n, x) \mapsto (\underline{\Pi}_x^n, \overline{\Pi}_x^n)$ is uniformly valid in the sense of Definition 2, then the derived set (12) is a genuine $100(1 - \alpha)\%$ prediction set in the sense that (1) holds or, equivalently, that*

$$\mathbf{P}\{C_{n,\alpha}(Z^n, X_{n+1}) \not\ni Y_{n+1}\} \leq \alpha, \quad \text{for all } (\alpha, n, \mathbf{P}).$$

Proof. The event where $C_{n,\alpha}(Z^n, X_{n+1})$ misses Y_{n+1} can be written as

$$\begin{aligned} C_{n,\alpha}(Z^n, X_{n+1}) \not\ni Y_{n+1} &\iff Y_{n+1} \notin \bigcap \{A : \underline{\Pi}_{X_{n+1}}^n(A) \geq 1 - \alpha\} \\ &\iff Y_{n+1} \in \bigcup \{A^c : \underline{\Pi}_{X_{n+1}}^n(A) \geq 1 - \alpha\} \\ &\iff Y_{n+1} \in A^c \text{ and } \underline{\Pi}_{X_{n+1}}^n(A) \geq 1 - \alpha \text{ for some } A. \end{aligned}$$

By uniform validity, in particular, Equation (5), the right-most event has \mathbf{P} -probability no more than α , proving the claim. \square

As expected, a uniformly valid probabilistic predictor leads to reliable frequentist prediction sets. This is in addition to the behaviorist no-sure-loss property that is implied by the weaker validity condition. After some additional high-level implications of (uniform) validity in Section 2.3, a general construction of a (uniformly) valid probabilistic predictor is presented in Section 3.

2.3 High-level implications

Before proceeding to our proposed construction of a (uniformly) valid probabilistic predictor, we believe that it is helpful to discuss how some other developments in the literature compare to what was presented above.

If \mathbf{P} was assumed known, then it follows immediately from (9) that the probabilistic predictor equal to the true conditional probability, i.e.,

$$\underline{\Pi}_x^n(A) = \overline{\Pi}_x^n(A) = \mathbf{P}(Y_{n+1} \in A \mid Z^n, X_{n+1} = x), \quad A \in \mathcal{A},$$

satisfies the validity-related properties (3) and (4), *but only at the known \mathbf{P}* . Similarly, one can also easily show that the set $C_{n,\alpha}$ defined in (12) has coverage probability of at least $1 - \alpha$, *but only at the known \mathbf{P}* . However, the validity property, as stated in Definition 1, requires those inequalities to hold for all \mathbf{P} ; this uniformity is important because \mathbf{P} is virtually always unknown applications, so its knowledge cannot be assumed

when constructing a probabilistic predictor. While a formal statement and proof of the following claim presently escapes us, it seems intuitively clear that there are no probabilistic predictors with the mathematical form of a precise/additive probability that are valid in the sense of Definition 1. If this claim is true, then it provides a version of the *false confidence theorem* (Balch et al. 2019; Martin 2019) in the context of prediction: that is, the only way to achieve valid probabilistic prediction is via a proper imprecise probability $(\underline{\Pi}_x^n, \overline{\Pi}_x^n)$.

It is not just precise/additive probabilities that have trouble achieving the validity property. More generally, one could start with a suitable credal set \mathbb{P} , a collection of candidate distributions \mathbf{P} , and define the probabilistic predictor’s upper probability as

$$\overline{\Pi}_x^n(A) = \sup_{\mathbf{P} \in \mathbb{P}} \mathbf{P}(Y_{n+1} \in A \mid Z^n, X_{n+1} = x), \quad A \in \mathcal{A}, \quad (13)$$

and lower probability with sup replaced by inf. Then the same argument presented above for the known/assumed- \mathbf{P} case implies that the validity-related properties (3) and (4) hold, *but only for* $\mathbf{P} \in \mathbb{P}$. If reliable choice of set \mathbb{P} were available in a given problem, then of course it should be used, and one way it could be used, while maintaining validity, is via the lower and upper envelop construction described above. However, if no such information is available, then taking \mathbb{P} to be all distributions or blindly assuming that the chosen proper subset \mathbb{P} contains the true \mathbf{P} would not be satisfactory options. An interesting open question is how genuine prior information in the form of a proper subset \mathbb{P} of probabilities could be incorporated into the proposed construction in Section 3.

A closer look at the uniform validity property can provide some helpful insights. Indeed, a necessary and sufficient condition for uniform validity is that

$$\mathbf{P}\{\pi_{X_{n+1}}^n(Y_{n+1}) \leq \alpha\} \leq \alpha, \quad \text{for all } (\alpha, n, \mathbf{P}), \quad (14)$$

where

$$\pi_x^n(y) = \overline{\Pi}_x^n(\{y\}), \quad x \in \mathbb{X}, \quad y \in \mathbb{Y}. \quad (15)$$

The proof—which follows by showing that both probabilities on the left-hand sides of (5) and (6) are equal to the left-hand side of (14)—is an immediate consequence of the probabilistic predictor’s monotonicity property. This explains why there are no measurability issues when working with the (potentially) uncountable unions related to the “for some A ” in (5) and (6). At least in the case of an absolutely continuous additive/precise probabilistic predictor, even that based on the true \mathbf{P} , it is clear that (14) cannot hold. So uniformly valid probabilistic prediction requires imprecise-probabilistic considerations.

The condition (14) is familiar, at least when connections are drawn to other contexts. In particular, (14) closely resembles the properties satisfied by p-values from hypothesis testing in classical statistics. It is also effectively the same as the so-called *fundamental frequentist principle*, or *FFP*, in Walley (2002). This connection to classical frequentist statistics strongly suggests that, to achieve uniform validity, the probabilistic predictor have the mathematical form of a consonant belief/plausibility function or a necessity/possibility measure. Our proposed probabilistic predictor construction in Section 3 indeed implies this special form.

Finally, we mention here connections with certain notions of “frequency-calibration” in the imprecise probability literature. In particular, using our terminology and notation, Denceux (2006) defines a probabilistic predictor to have a “100(1 − α)% confidence

property,” for a fixed $\alpha \in [0, 1]$, if

$$\mathbb{P}\{\bar{\Pi}_{X_{n+1}}^n(A) \geq \mathbb{P}(Y_{n+1} \in A \mid Z^n, X_{n+1}) \text{ for all } A\} \geq 1 - \alpha.$$

This and other variations are discussed more recently in Denceux and Li (2018). Obviously, since the event on the left-hand side does not explicitly depend on α , it must be that the probabilistic predictor depends implicitly on the specified α value, and various approaches to incorporate this α -dependence so that the above property can be achieved are given in the aforementioned references. The key observation is that calibration requires some relation between the probabilistic predictor for Y_{n+1} and the true conditional distribution of Y_{n+1} . In particular, the prediction upper probability ought to dominate the true conditional probability in a strict sense as in (13) or a less-strict sense as in the above display. This same kind of dominance appears in our definition of validity, but through the alternative formulation in (9), specifically,

$$\mathbb{E}[1\{\bar{\Pi}_{X_{n+1}}^n(A) \leq \alpha\} \mathbb{P}(Y_{n+1} \in A \mid Z^n, X_{n+1})] \leq \alpha.$$

That is, our notion of validity implies that, when restricted to data sets (Z^n, X_{n+1}) for which $\bar{\Pi}_{X_{n+1}}^n(A)$ is small, the true conditional probability $\mathbb{P}(Y_{n+1} \in A \mid Z^n, X_{n+1})$ cannot be any bigger on average.

3 Inferential models

An inferential model (IM) is a data-dependent probabilistic quantification of uncertainty about unknowns, like the probabilistic predictor described above. The difference is, as the name suggests, that IMs are focused on the *statistical inference* problem, where the unknowns are fixed quantities. IMs have connections to various other approaches to statistical inference, some that quantify uncertainties with ordinary probabilities, e.g., Bayesian inference (Martin and Liu 2015a), fiducial inference (Fisher 1935; Taraldsen and Lindqvist 2013), and generalized fiducial inference (Hannig et al. 2016), as well as with imprecise probabilities, e.g., Dempster–Shafer theory (Dempster 1967, 1968, 2008, 2014; Shafer 1976) and other belief function frameworks (Denceux 2014; Denceux and Li 2018). While there are some technical differences resulting from the unknown being fixed in the inference case and random in the prediction case, the common goal of providing valid uncertainty quantification is more or less the same. Therefore, we expect that the key ideas behind the construction of a valid IM for inference ought to be applicable to the prediction problem as well, modulo a few adjustments. Below we describe the key ideas behind the IM construction and how they can be employed to construct a probabilistic predictor that is valid in the sense described in Section 2.

The general 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. For example, suppose we have, say, n independent and identically distributed (iid) observations Z_1, \dots, Z_n , collected into the vector Z^n , from a statistical model with unknown parameter θ . Then an association would effectively be a description of how to generate data Z^n from that model, i.e.,

$$Z^n = a(\theta, U^n),$$

where U^n would typically be a vector of iid latent/auxiliary variables with a known distribution, e.g., $\text{Unif}(0,1)$. While such an association can always be written down, there are a few obstacles one might face when trying to complete the IM construction:

- When the dimension of U^n is greater than that of θ , as is often the case, an efficiency-motivated dimension-reduction step is recommended by Martin and Liu (2015a), but this can be challenging.
- The association itself requires (more than) a fully specified statistical model for data, which may not be available in the application at hand.

However, Martin (2015, 2018) showed that the A-step’s requirements can be relaxed. All that is needed is an association that relates a suitable function of both the observable data and the unknowns to an unobservable auxiliary variable. Besides the examples presented in Martin (2015, 2018), this idea has been applied to meta-analysis and survival analysis in Cahoon and Martin (2020, 2021). An extension of that initial generalization, which can avoid even specification of a statistical model was recently developed in Cella and Martin (2021a), with a focus on machine learning applications.

Once a generalized association has been set, the remain steps of the (generalized) IM construction proceed exactly as described in, say, Martin and Liu (2013). Roughly, 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 generalize IM construction in more detail for the prediction problem at hand.

For prediction, the unknown is Y_{n+1} , not a model parameter in the IM and generalized IM formulation described above. So the kind of association needed is one that identifies a function of (Z^n, Z_{n+1}) that has a known distribution. Once found, the three-step (generalized) IM construction proceeds as follows.

A-step. Suppose there exists a function $\phi_n : \mathbb{Z}^n \times \mathbb{Z} \rightarrow \mathbb{R}$ such that the distribution, say, \mathbf{Q}_n , of the random variable $\phi_n(Z^n, Z_{n+1})$ is known, i.e., does not depend on the unknown P . Then associate the observable data Z^n and the yet-to-be-observed Z_{n+1} with the unobservable auxiliary variable U as follows:

$$\phi_n(Z^n, Z_{n+1}) = U, \quad U \sim \mathbf{Q}_n. \quad (16)$$

For our case where $Z_{n+1} = (X_{n+1}, Y_{n+1})$ and interest is in Y_{n+1} for a given $X_{n+1} = x$, 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{U} (see below) on the space \mathbb{U} of the auxiliary variable U , designed to reliably contain realizations of $U \sim \mathbf{Q}_n$ in the sense of (20) below. The distribution of the random set \mathcal{U} will be denoted by \mathbf{R}_n .

C-step. Combine the results of the A- and P-steps to get the data-dependent random set

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

Then the distribution of this new random set, derived from the distribution of \mathcal{U} , determines the probabilistic predictor for Y_{n+1} , i.e.,

$$\begin{aligned} \underline{\Pi}_x^n(A) &= \mathbf{R}_n\{\mathbb{Y}_x^n(\mathcal{U}) \subseteq A\} \\ \overline{\Pi}_x^n(A) &= \mathbf{R}_n\{\mathbb{Y}_x^n(\mathcal{U}) \cap A \neq \emptyset\}. \end{aligned} \tag{17}$$

Remark 1. If $\mathbb{Y}_x^n(\mathcal{U})$ is empty with positive \mathbf{R}_n -probability, then some adjustment to the probabilistic predictor in (17) 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 \mathbf{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 \mathbf{P} are being made, so different techniques are required. The remainder of this section investigates the properties of the abstract probabilistic predictor construction above.

The random set \mathcal{U} 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

$$\pi_x^n(y) = \mathbf{R}_n\{\mathbb{Y}_x^n(\mathcal{U}) \ni y\}, \quad y \in \mathbb{Y}. \tag{18}$$

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

$$\overline{\Pi}_x^n(A) = \sup_{y \in A} \pi_x^n(y), \quad A \subseteq \mathbb{Y}. \tag{19}$$

This alternative expression is important for at least two reasons. First, 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 (uniformly) valid. This requires stating the required conditions on \mathcal{U} 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

$$f(u) = \mathbf{R}_n(\mathcal{U} \ni u), \quad u \in \mathbb{U}.$$

Then the required link between \mathbf{Q}_n and \mathbf{R}_n is that

$$U \sim \mathbf{Q}_n \implies f(U) \sim \text{Unif}(\mathcal{J}_{n+1}), \quad (20)$$

where $\mathcal{J}_{n+1} = \{1, \dots, n, n+1\}$, so that this uniform distribution is discrete. With this link between the auxiliary variable's distribution \mathbf{Q}_n and the random set's distribution \mathbf{R}_n , we are ready to state and prove the main result.

Theorem 1. *If the random set \mathcal{U} satisfies (20), and if $\mathbb{Y}_{X_{n+1}}^n(\mathcal{U})$ is non-empty with \mathbf{R}_n -probability 1 for \mathbf{P} -almost all (Z^n, X_{n+1}) , then the probabilistic predictor defined in (17), or equivalently (19), is uniformly valid in the sense of Definition 2.*

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

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

The \mathbf{R}_n -probability of the left- and right-hand side events are $\pi_{X_{n+1}}^n(Y_{n+1})$ and $f(U)$, respectively, so these two random variables—the first as a function of $(Z^n, Z_{n+1}) \sim \mathbf{P}$ and the second as a function of $U \sim \mathbf{Q}_n$ —have the same distribution. Equation (20) states that $f(U)$ is uniform and, therefore, so is $\pi_{X_{n+1}}^n(Y_{n+1})$, which proves the claim. \square

The non-emptiness condition is not necessary for validity, but some adjustment is needed to the definition in (17), as mentioned in Remark 1, to address this. We will discuss this below in the specific application to classification in Section 5.

The following is an immediate consequence of the uniform validity conclusion above and the general results in Propositions 1–3 in the previous section.

Corollary 1. *Under the conditions of Theorem 1, the probabilistic predictor defined in (17) or, equivalently, in (19)*

- (a) *avoids sure loss in the sense of (7), and*
- (b) *admits a prediction set $C_{n,\alpha}$ as in (12) that achieves the nominal frequentist coverage probability in the sense of (1).*

Moreover, there is an equivalent form of that prediction set in terms of the plausibility contour, namely,

$$C_\alpha^n(x) = \{y \in \mathbb{Y} : \pi_x^n(y) > \alpha\} \quad (21)$$

that is computationally more convenient and, of course, also achieves the nominal frequentist coverage probability.

Consequently, the proposed probabilistic predictor construction achieves the desired subjective/behaviorist and objective/frequentist properties simultaneously. Two specific and practically relevant applications of this construction in the context of regression and classification will be presented in Section 4 and 5, respectively.

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 calibrated prediction sets as in (1). That is, the conditional coverage probability of the prediction set is

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

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 \mathbf{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 $Z^{n+1} \rightarrow T^{n+1}$, defined by

$$T_i = \Psi(Z_{-i}^{n+1}, Z_i), \quad i \in \mathcal{I}_{n+1}, \quad (22)$$

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

$$T_i = |Y_i - \hat{\mu}_{-i}^{n+1}(X_i)|, \quad i \in \mathcal{I}_{n+1}. \quad (23)$$

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, \dots 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. In regression, where the Y_i 's are continuous and Ψ is non-constant on sets of Y^{n+1} with positive \mathbf{P} -probability, like the one in (23), so that there are no ties, a well-known consequence of exchangeability of T_1, \dots, T_{n+1} is that their ranks are marginally distributed according to $\text{Unif}(\mathcal{I}_{n+1})$, the discrete uniform law 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 (16) as follows:

$$r(T_{n+1}) = U, \quad U \sim \text{Unif}(\mathcal{I}_{n+1}), \quad (24)$$

where $r(\cdot)$ is the ascending ranking operator. The choice of T_{n+1} instead of any of the other T_i 's in (24) 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

$$\mathcal{U} = \{1, 2, \dots, U'\}, \quad U' \sim \text{Unif}(\mathcal{I}_{n+1}). \quad (25)$$

It is straightforward to show that this random set satisfies the critical calibration property (20). Moreover, this choice also makes intuitive sense, as \mathcal{U} always includes the value 1.

This is desirable given the ascending ranking operator in (24) 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{U} is combined with the u -indexed collection of sets

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

that arise from the association (24). 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, It is easy to see that $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$'s corresponding contour function for Y_{n+1} is given by

$$\begin{aligned} \pi_{x_{n+1}}^n(y_{n+1}) &= \mathbf{R}_n\{\mathbb{Y}_{x_{n+1}}^n(\mathcal{U}) \ni y_{n+1}\} \\ &= \Pr\{\text{Unif}(\mathcal{J}_{n+1}) \geq r(T_{n+1}(z^{n+1}))\} \\ &= \frac{1}{n+1} \sum_{i=1}^{n+1} 1\{T_i(z^{n+1}) \geq T_{n+1}(z^{n+1})\}. \end{aligned} \quad (26)$$

As $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$ 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 (19), which can easily be approximated by

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

Uniform validity of the probabilistic predictor derived in this Section is a direct consequence of Theorem 1. Consequently, this probabilistic predictor satisfies (4), 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.

For illustration, consider the following example. Let X_1, \dots, X_n be iid $\text{Unif}(0, 1)$, with $n = 200$, and let Y_1, \dots, Y_n be independent, where $Y_i = \mu(X_i) + 0.1\varepsilon_i$, where $\mu(x) = \sin^3(2\pi x^3)$, and $\varepsilon_1, \dots, \varepsilon_n$ are iid from a Student-t distribution with $\text{df} = 5$. 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 (21) and x_{n+1} taking values along the observed x^n .

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). 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 (26), which is precisely conformal prediction's p-value or transducer. The theory in Vovk et al. (2005) takes this conformal transducer, which is uniformly distributed as stated in Theorem 1, and constructs a prediction set as in (21) with the prediction coverage probability property as in (1). Apparently it was recognized only recently (Cella and Martin 2021c) 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

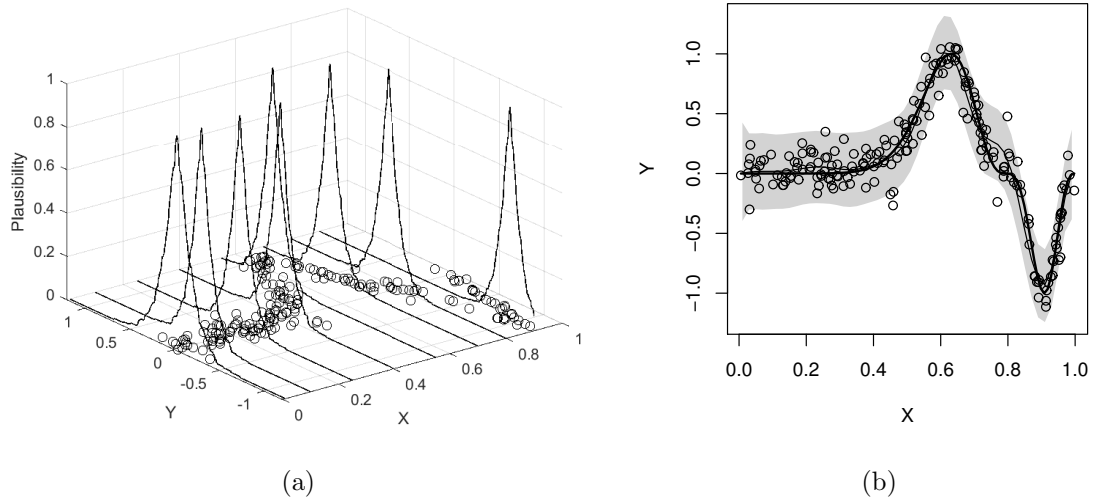


Figure 1: Panel (a): Data and the plausibility contours at selected values of x . Panel (b): Data, the true mean curve (heavy line), the fitted B-spline regression curve (thin line), and the 95% pointwise prediction band.

plausibility function via (19). We refer to this general probabilistic predictor construction as “conformal + consonance,” and all it requires is that the conformal transducer π_x^n be a plausibility contour function, i.e.,

$$\sup_y \pi_x^n(y) = 1, \quad \text{for all } (Z^n, x). \quad (27)$$

This is easy to verify in cases like regression where Y is a continuous random variable. Indeed, for the Ψ function in (23), 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 (23) 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:

$$\Psi(Z_{-i}^{n+1}, Z_i) = \frac{\min_{j \in \mathcal{J}_{n+1} \setminus \{i\}: Y_j = Y_i} d(X_j, X_i)}{\min_{j \in \mathcal{J}_{n+1} \setminus \{i\}: Y_j \neq Y_i} d(X_j, X_i)}, \quad (28)$$

where d is the Euclidean distance. In words, $\Psi(Z_{-i}^{n+1}, Z_i)$ is large if X_i is close to an element in X_{-i}^{n+1} with a label different from Y_i and far from any element in X_{-i}^{n+1} with label equal to Y_i . If both the numerator and the denominator in (28) are 0, Shafer and Vovk (2007) recommend taking the ratio also to be 0. Other non-conformity measures for classification problems can be found in Vovk et al. (2005).

Two factors were fundamental to the specification of the association (24) in Section 4, namely the identification of Ψ , so that Z^{n+1} can be mapped to T^{n+1} preserving exchangeability, and the continuity of the T_i 's. In classification, however, the Y_i 's are not continuous, so there could be ties in the T_i 's. Consequently, their ranks would be no longer uniform distributed on \mathcal{J}_{n+1} . Luckily, when ties are possible, $r(T_{n+1})$ is stochastically no larger than the discrete uniform distribution it would take if there were no ties. This leads to an “association” of the form

$$r(T_{n+1}) = U, \quad U \leq_{\text{st}} \text{Unif}(\mathcal{J}_{n+1}).$$

But for situations like this where the association involves a stochastic inequality, the general arguments in Martin and Liu (2015c, Sec. 5) imply that the inequality can be ignored and the association (24)—with stochastic equality—can still be used.

Having identified the appropriate association, the IM construction proceeds analogously to that in the previous section: the A-step is completed by writing (24), the random set (25) 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 $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$, a data-dependent random subset of \mathbb{Y} . However, due to the discreteness of \mathbb{Y} , it is possible that $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$ is empty with positive \mathbf{R}_n -probability. As discussed in Section 3, in these cases, some adjustment to the probabilistic predictor in (17) is necessary to avoid the counter-intuitive “conflict” cases where realizations of the random set $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$ 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 $\mathbb{Y}_{x_{n+1}}^n(\mathcal{S})$. 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

$$y_{n+1} \mapsto \mathbf{R}_n\{\mathbb{Y}_{x_{n+1}}^n(\mathcal{U}) \ni y_{n+1} \mid \mathbb{Y}_{x_{n+1}}^n(\mathcal{U}) \neq \emptyset\}.$$

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 fail to be valid. This inflation does, however, suggest a potential loss of efficiency, e.g., larger prediction sets in (21).

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). Roughly, those \mathcal{U} such that $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U}) = \emptyset$ correspond to “conflict cases,” and Dempster’s conditioning rule simply removes these conflict cases and renormalizes the \mathcal{U} -probabilities. As an alternative, Ermini Leaf and Liu (2012) suggested to stretch those conflict \mathcal{U} cases just enough so that $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$ is

non-empty. Their formulation was in the context of inference under non-trivial parameter constraints, but here we apply this to classification.

Start by defining the set

$$\mathbb{U}_{x_{n+1}}^n = \bigcup_{y_{n+1} \in \mathbb{Y}} \{r(T_{n+1}(z^n, z_{n+1}))\} \subseteq \mathcal{J}_{n+1}. \quad (29)$$

There are only finitely many y_{n+1} values, and the set $\mathbb{U}_{x_{n+1}}^n$ defined above is just the collection of ranks that are possible for the given Z^n and x_{n+1} . Note that $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$ is empty if and only if \mathcal{U} has empty intersection with $\mathbb{U}_{x_{n+1}}^n$. Therefore, the conflict cases mentioned above can be alternatively defined as realizations of \mathcal{U} that have empty intersection with $\mathbb{U}_{x_{n+1}}^n$. This conflicting situation can be avoided if, instead of throwing out the conflict \mathcal{U} , we stretch it to a suitable \mathcal{U}_e , with $e \geq 0$ a stretching parameter that controls how far \mathcal{U} is stretched toward $\mathbb{U}_{x_{n+1}}^n$. In particular, we take

$$\mathcal{U}_e = \{1, 2, \dots, U' + e\}, \quad U' \sim \text{Unif}(\mathcal{J}_{n+1}).$$

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

$$\hat{e} = \min\{e : \mathcal{U}_e \cap \mathbb{U}_{x_{n+1}}^n \neq \emptyset\} = \begin{cases} \min \mathbb{U}_{x_{n+1}}^n - U' & \text{if } U' < \min \mathbb{U}_{x_{n+1}}^n \\ 0 & \text{otherwise.} \end{cases}$$

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

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

In summary, in the stretching IM, the IM's original random set output $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U})$ is replaced with $\mathbb{Y}_{x_{n+1}}^n(\mathcal{U}_{\hat{e}})$, and its guaranteed non-emptiness makes the probabilistic predictor derived from it valid. It is also more efficient than conditioning since it avoids globally inflating the plausibility contour via renormalization, as the following example highlights.

For illustration, consider the data in Table 1, taken from Agresti (2003, p. 304), describing the primary food choices and lengths of $n = 39$ male alligators 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

$$\mathbb{Y}_{x_{n+1}}^n(\mathcal{U}) = \begin{cases} \{I\} & \text{with probability 0.1} \\ \{I, F\} & \text{with probability 0.2} \\ \{I, F, O\} & \text{with probability 0.3} \\ \emptyset & \text{with probability 0.4.} \end{cases} \quad (30)$$

The corresponding plausibility contour, as given in (18), 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 (30) $\neq \emptyset$ is easy to evaluate, and is represented by the dashed lines in Figure 2(a). To calculate the plausibility contour under

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

Table 1: Primary food choice (I, invertebrates; F, fish; O, other) and lengths (in meters) for $n = 39$ male alligators (Agresti 2003, p. 304).

the stretching approach, we obtain, after some calculations, $\mathbb{U}_{x_{n+1}}^n = \{17, 21, 29\}$. As $\min \mathbb{U}_{x_{n+1}}^n = 17$,

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

where $U' \sim \text{Unif}(1, 2, \dots, 40)$. Therefore,

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

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, stretching leads to a more efficient probabilistic predictor.

To further see this gain in efficiency we consider the *Glass Identification* data set from the USA Forensic Science Service, available in the UCI Machine Learning Repository (Dua and Graff 2017).¹ It has 10 attributes associated with 214 glasses. The type of glass,

¹<https://archive.ics.uci.edu/ml/datasets/glass+identification>

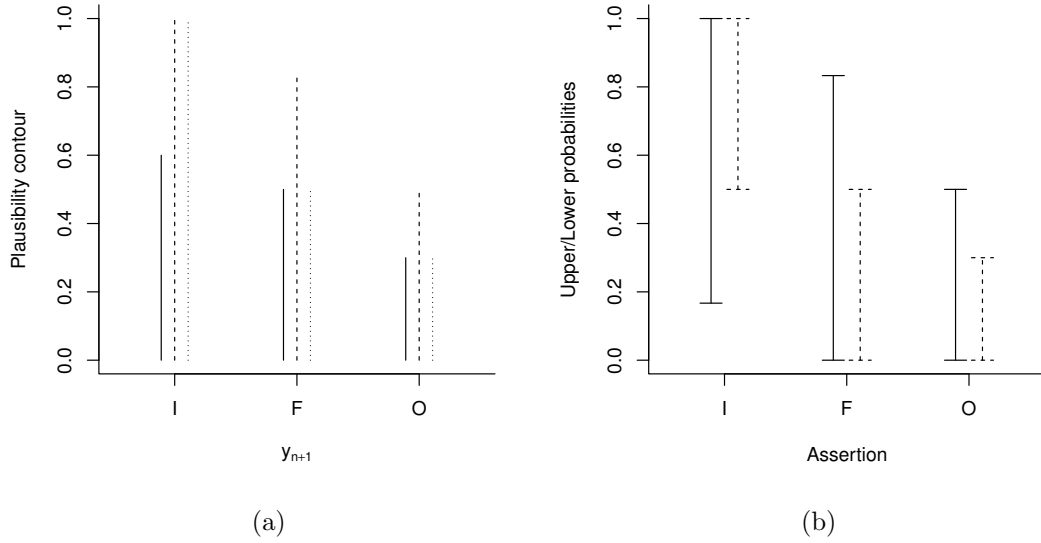


Figure 2: Panel (a): Plausibility contours in Equation (18), derived from an IM construction with no adjustment (solid lines), conditioning adjustment (dashed lines) and stretching adjustment (dotted lines). Panel (b): Upper and lower probabilities for the singleton assertions $\{I\}$, $\{F\}$ and $\{O\}$ derived from an IM construction with the conditioning adjustment (solid lines) and the stretching adjustment (dashed lines). These predictions are based on a new alligator of length $x_{n+1} = 2$ meters.

Strategy	Coverage	Size
Conditioning	0.96	3.07
Stretching	0.96	2.73

Table 2: Coverage probabilities and average size of 95% prediction sets in (21) derived from an IM construction with conditioning and stretching adjustment.

a categorical variable—with six categories, including “containers” and “headlamps”—is the response variable. The nine remaining variables, which describe the oxide content, i.e., Na, Fe, K, etc., are the explanatory variables. Classification of types of glass is relevant in criminology applications, where glass fragments left at the scene of the crime may be important evidence if correctly identified. To evaluate the performance of the IM in classifying glass fragments, we randomly split the data in half and train both the conditioning and stretching strategies in the first half, with Ψ function as in (28). Table 2 shows the empirical coverage probabilities and the average sizes (cardinality) of 95% prediction sets for the responses in the second half of the data. As expected, both strategies lead to valid predictions, but stretching is slightly more efficient.

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_x^n satisfies (27). 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, (27) 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 following two adjusted conformal transducers:

$$\dot{\pi}_x^n(y) = \frac{\pi_x^n(y)}{\max_y \pi_x^n(y)},$$

and

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

where $\hat{y} = \arg \max_y \pi_x^n(y)$ and $y \in \mathbb{Y}$. In words, $\dot{\pi}_x^n(y)$ takes the conformal transducers for the different $y \in \mathbb{Y}$ and divide them by their maximum, and $\ddot{\pi}_x^n(y)$ maintains all the conformal transducer values except for its maximum, which is assigned the value 1. That both adjusted transducers reach the value 1 makes the probabilistic predictors derived by them, through (19), valid in the sense of Definition 1. It is also easy to see that these probabilistic predictors obtained 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 an 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.

Exchangeability was crucial to our IM construction, that is, without exchangeability, we cannot establish the distribution of the auxiliary variables. While exchangeability is a relatively weak assumption compared to iid from a parametric family, there are, of course, situations where exchangeability is inappropriate, such as time series or spatial applications. Work to develop conformal prediction methods in not-exactly-exchangeable

settings is an active area of current research (e.g., Mao et al. 2020), and it would be interesting to see what the IM perspective has to offer here.

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

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

and, of course, for all $(\alpha, n, A, \mathbf{P})$ 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.

Acknowledgments

This work is partially supported by the U.S. National Science Foundation, grants DMS-1811802 and SES-2051225. The authors thank the reviewers of our ISIPTA’21 conference proceedings submission for their feedback, and the *IJAR* guest editors—Andrés Cano, Jasper De Bock, and Enrique Miranda—for the invitation to submit an extended version of our conference proceedings paper to the special journal issue.

References

- Agresti, A. (2003). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley.
- Balch, M. S., Martin, R., and Ferson, S. (2019). Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2227):1–20.
- Cahoon, J. and Martin, R. (2020). Generalized inferential models for meta-analyses based on few studies. *Statistics and Applications*, 18(2):299–316.
- Cahoon, J. and Martin, R. (2021). Generalized inferential models for censored data. *International Journal of Approximate Reasoning*, 137:51–66.
- Campi, M., Calafiore, G., and Garatti, S. (2009). Interval predictor models: Identification and reliability. *Automatica*, 45(2):382–392.

- Cella, L. and Martin, R. (2019). Incorporating expert opinion in an inferential model while maintaining validity. In De Bock, J., de Campos, C. P., de Cooman, G., Quaeghebeur, E., and Wheeler, G., 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. PMLR.
- Cella, L. and Martin, R. (2021a). Approximately valid and model-free possibilistic inference. In Denœux, T., Lefèvre, E., Liu, Z., and Pichon, F., editors, *Belief Functions: Theory and Applications*, pages 127–136, Cham. Springer International Publishing.
- Cella, L. and Martin, R. (2021b). Valid inferential models for prediction in supervised learning problems. In Cano, A., De Bock, J., Miranda, E., and Moral, S., editors, *Proceedings of the Twelfth International Symposium on Imprecise Probability: Theories and Applications*, volume 147 of *Proceedings of Machine Learning Research*, pages 72–82. PMLR.
- Cella, L. and Martin, R. (2021c). Validity, consonant plausibility measures, and conformal prediction. *International Journal of Approximate Reasoning*, to appear.
- Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2019). Distributional conformal prediction. [arXiv:1909.07889](https://arxiv.org/abs/1909.07889).
- Coolen, F. P. A. (2006). On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language, and Information*, 15(1/2):21–47.
- Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339.
- Dempster, A. P. (1968). A generalization of Bayesian inference. (With discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 30:205–247.
- Dempster, A. P. (2008). The Dempster–Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365–377.
- Dempster, A. P. (2014). Statistical inference from a Dempster–Shafer perspective. In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L., editors, *Past, Present, and Future of Statistical Science*, chapter 24. Chapman & Hall/CRC Press.
- Denœux, T. (2006). Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3):228–252.
- Denœux, T. (2014). Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning*, 55(7):1535–1547.
- Denœux, T. and Li, S. (2018). Frequency-calibrated belief functions: review and new insights. *International Journal of Approximate Reasoning*, 92:232–254.
- Dua, D. and Graff, C. (2017). UCI machine learning repository. <http://archive.ics.uci.edu/ml>. University of California, Irvine, School of Information and Computer Sciences.

- Dubois, D. and Prade, H. (1988). *Possibility Theory*. Plenum Press, New York.
- Ermini Leaf, D. and Liu, C. (2012). Inference about constrained parameters using the elastic belief method. *International Journal of Approximate Reasoning*, 53(5):709–727.
- Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, 6(4):391–398.
- Gong, R. and Meng, X.-L. (2021). Judicious judgment meets unsettling updating: Dilation, sure loss, and Simpson’s paradox. *Statistical Science*, 36(2):169–190.
- Hannig, J., Iyer, H., Lai, R. C. S., and Lee, T. C. M. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515):1346–1361.
- Kuleshov, V., Fenner, N., and Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. [arXiv:1807.00263](https://arxiv.org/abs/1807.00263).
- Lawless, J. F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92(3):529–542.
- Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96.
- Mao, H., Martin, R., and Reich, B. (2020). Valid model-free spatial prediction. [arXiv:2006.15640](https://arxiv.org/abs/2006.15640).
- Martin, R. (2015). Plausibility functions and exact frequentist inference. *Journal of the American Statistical Association*, 110(512):1552–1561.
- Martin, R. (2018). On an inferential model construction using generalized associations. *Journal of Statistical Planning and Inference*, 195:105–115.
- Martin, R. (2019). False confidence, non-additive beliefs, and valid statistical inference. *International Journal of Approximate Reasoning*, 113:39–73.
- Martin, R. (2021). An imprecise-probabilistic characterization of frequentist statistical inference. *Researchers.One*, <https://researchers.one/articles/21.01.00002>.
- Martin, R. and Lingham, R. T. (2016). Prior-free probabilistic prediction of future observations. *Technometrics*, 58:225–235.
- Martin, R. and Liu, C. (2013). Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association*, 108:301–313.
- Martin, R. and Liu, C. (2015a). Conditional inferential models: combining information for prior-free probabilistic inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:195–217.
- Martin, R. and Liu, C. (2015b). *Inferential Models: Reasoning with Uncertainty*. Monographs in Statistics and Applied Probability Series. Chapman & Hall/CRC Press.

- Martin, R. and Liu, C. (2015c). Marginal inferential models: Prior-free probabilistic inference on interest parameters. *Journal of the American Statistical Association*, 110(512):1621–1631.
- Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J.
- Shafer, G. and Vovk, V. (2007). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421.
- Taraldsen, G. and Lindqvist, B. H. (2013). Fiducial theory and optimal inference. *Annals of Statistics*, 41(1):323–341.
- Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg.
- Vovk, V., Shen, J., Manokhin, V., and Xie, M. (2018). Nonparametric predictive distributions based on conformal prediction. *Machine Learning*, 108:445–474.
- Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.
- Walley, P. (2002). Reconciling frequentist properties with the likelihood principle. *Journal of Statistical Planning and Inference*, 105:35–65.
- Wang, C. M., Hannig, J., and Iyer, H. K. (2012). Fiducial prediction intervals. *Journal of Statistical Planning and Inference*, 142(7):1980–1990.