Inferential models: A framework for prior-free posterior probabilistic inference

Ryan Martin

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
rgmartin@math.uic.edu

Chuanhai Liu
Department of Statistics
Purdue University
chuanhai@stat.purdue.edu

March 26, 2013

Abstract

Posterior probabilistic statistical inference without priors is an important but so far elusive goal. Fisher's fiducial inference, Dempster–Shafer theory of belief functions, and Bayesian inference with default priors are attempts to achieve this goal but, to date, none has given a completely satisfactory picture. This paper presents a new framework for probabilistic inference, based on *inferential models* (IMs), which not only provides data-dependent probabilistic measures of uncertainty about the unknown parameter, but does so with an automatic long-run frequency calibration property. The key to this new approach is the identification of an unobservable auxiliary variable associated with observable data and unknown parameter, and the prediction of this auxiliary variable with a random set before conditioning on data. Here we present a three-step IM construction, and prove a frequency-calibration property of the IM's belief function under mild conditions. A corresponding optimality theory is developed, which helps to resolve the non-uniqueness issue. Several examples are presented to illustrate this new approach.

Keywords and phrases: Belief function; plausibility function; predictive random set; score function; validity.

1 Introduction

In a statistical inference problem, one attempts to convert *experience*, in the form of observed data, to *knowledge* about the unknown parameter of interest. The fact that observed data is surely limited implies that there will be some uncertainty in this conversion, and probability is a natural tool to describe this uncertainty. But a statistical inference

problem is different from the classical probability setting because everything—observed data and unknown parameter—is fixed, and so it is unclear where these probabilistic assessments of uncertainty should come from, and how they should be interpreted. For example, the classical frequentist approach assigns probabilistic assessments of uncertainty (e.g., confidence levels) by considering repeated sampling from the super-population of possible data sets. These uncertainty measures do not depend on the observed data, so their meaningfulness in a given problem is questionable. The Bayesian approach, on the other hand, is able to produce meaningful data-dependent probabilistic measures of uncertainty, but the cost is that a prior probability distribution for the unknown parameter is required. Early efforts to get probabilistic inference without prior specification include Fisher's fiducial inference (Zabell 1992) and its variants (Hannig 2009, 2012; Hannig and Lee 2009), confidence distributions (Xie and Singh 2012; Xie et al. 2011), Fraser's structural inference (Fraser 1968), and the Dempster–Shafer theory (Dempster 2008; Shafer 1976). These methods generate probabilities for inference, but these probabilities may not be easy to interpret, e.g., they may not be properly calibrated across users or experiments. So recent efforts have focused on incorporating a frequentist element. In particular, objective Bayes analysis with default/reference priors (Berger 2006; Berger et al. 2009; Bernardo 1979; Ghosh 2011) attempts to construct priors for which certain posterior inferences, such as credible intervals, closely match that of a frequentist (Fraser 2011; Fraser et al. 2010). Calibrated Bayes (Dawid 1985; Little 2011; Rubin 1984) has similar motivations. But difficulties remain in choosing good reference priors for high-dimensional problems so, despite these efforts, a fully satisfactory framework of objective Bayes inference has yet to emerge.

The goal of this paper is to develop a new framework for statistical inference, called *inferential models* (IMs). The seeds for this idea were first planted in Martin et al. (2010) and Zhang and Liu (2011); here we formalize and extend these ideas towards a cohesive framework for statistical inference. The jumping off point is a simple association of the observable data X and unknown parameter $\theta \in \Theta$ with an unobservable auxiliary variable U. For example, consider the simple signal plus noise model, $X = \theta + U$, where $U \sim \mathsf{N}(0, 1)$. If $X = x$ is observed, then we know that $x = \theta + u^\star$, where $u^\star$ is some *unobserved* realization of U. From this it is clear that knowing $u^\star$ is equivalent to knowing θ. So the IM approach attempts to accurately predict the value $u^\star$ before conditioning on $X = x$. The benefit of focusing on $u^\star$ rather than θ is that more information is available about $u^\star$: indeed, all that is known about θ is that it sits in Θ, while $u^\star$ is known to be a realization of a draw U from an *a priori* distribution, in this case $\mathsf{N}(0, 1)$, that is fully specified by the postulated sampling model. However, this *a priori* distribution alone is insufficient for accurate prediction of $u^\star$. Therefore, we adopt a so-called *predictive random set* for predicting $u^\star$, which amounts to a sort of "smearing" of this distribution for U. When combined with the association between observed data, parameters, and auxiliary variables, these random sets produce prior-free, data-dependent probabilistic assessments of uncertainty about θ.

To summarize, an IM starts with an association between data, parameters, and auxiliary variables and a predictive random set, and produces prior-free, post-data probabilistic measures of uncertainty about the unknown parameter. The following associate-predict-combine steps provide a simple yet formal IM construction. The details of each of these three steps will be fleshed out in Section 2.

A-step. Associate the unknown parameter θ to each possible (x, u) pair to obtain a collection of sets $\Theta_x(u)$ of candidate parameter values.

P-step. Predict $u^\star$ with a valid predictive random set $\mathcal{S}$.

C-step. Combine $X = x$, $\Theta_x(u)$, and $\mathcal{S}$ to obtain a random set $\Theta_x(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_x(u)$. Then, for any assertion $A \subseteq \Theta$, compute the probability that the random set $\Theta_x(\mathcal{S})$ is a subset of A as a measure of the available evidence in x supporting A.

The A-step is meant to emphasize the use of unobservable but predictable auxiliary variables in the statistical modeling step. These auxiliary variables make it possible to introduce posterior probability-like quantities without a prior distribution for θ. The P-step is new and unique to the inferential model framework. The key is that $\Theta_x(\mathcal{S})$ contains the true θ if and only if $\mathcal{S}$ contains $u^\star$. Then the validity condition in the P-step ensures that $\mathcal{S}$ will hit its target with large probability which, in turn, guarantees that probabilistic output from the C-step has a desirable frequency-calibration property. This, together with its dependence on the observed data x, makes the IM's probabilistic output meaningful both within and across experiments.

The remainder of the paper is organized as follows. Section 2 provides the details of the IM analysis, specifically the three-step construction outlined above, as well as a description of calculation and interpretation of the IM output: a posterior belief function. These ideas are illustrated with a simple Poisson mean example. After arguing, in Section 2, that the IM output provides a meaningful summary of one's uncertainty about θ after seeing $X = x$, we prove a frequency calibration property of the posterior belief functions in Section 3 which establishes the meaningfulness of the posterior belief function across different users and experiments. As a consequence of this frequency-calibration property, we show in Section 3.4 that the IM output can easily be used to design new frequentist decision procedures having the desired control on error probabilities, etc. Some basic but fundamental results on IM optimality are presented in Section 4. Section 5 gives IM-based solutions to two non-trivial examples, both involving some sort of marginalization. Nonetheless, these examples are relatively simple and they illustrate the advantages of the IM approach. Concluding remarks are given in Section 6, and R codes for the examples are available on the first author's website: `www.math.uic.edu/~rgmartin`.

2 Inferential models

2.1 Auxiliary variable associations

If X denotes the observable sample data, then the sampling model is a probability distribution $\mathsf{P}_{X|\theta}$ on the sample space $\mathbb{X}$, indexed by a parameter $\theta \in \Theta$. Here X may consist of a collection of n (possibly vector-valued) data points, in which case both $\mathsf{P}_{X|\theta}$ and $\mathbb{X}$ would depend on n. The sampling model for X is induced by an auxiliary variable U, for given θ. Let $\mathbb{U}$ be an (arbitrary) auxiliary space, equipped with a probability measure P_U. In applications, $\mathbb{U}$ can often be a unit hyper-cube and P_U Lebesgue measure. The sampling model $\mathsf{P}_{X|\theta}$ shall be determined by the following "algorithm:"

$$\text{sample } U \sim \mathsf{P}_U \text{ and set } X = a(U, \theta), \tag{2.1}$$

for an appropriate mapping $a : \mathbb{U} \times \Theta \to \mathbb{X}$. The key is the association of the observable X, the unknown θ, and the auxiliary variable U through the relation $X = a(U, \theta)$. This particular formulation of the sampling model is not really a restriction. In fact, the two-step construction of the observable X in (2.1) is often consistent with scientific understanding of the underlying process under investigation; linear models form an interesting class of examples. As another example, suppose $X = (X_1, \ldots, X_n)$ consists of an independent sample from a continuous distribution. If the corresponding distribution function F_θ is invertible, then $a(\theta, U)$ may be written as

$$a(\theta, U) = \left(F_\theta^{-1}(U_1), \ldots, F_\theta^{-1}(U_n) \right), \tag{2.2}$$

where $U = (U_1, \ldots, U_n)$ is a set of independent $\mathsf{Unif}(0, 1)$ random variables.

The notation $X = a(\theta, U)$ chosen to represent the association between (X, θ, U) is just for simplicity. In fact, this association need not be described by a formal equation. As the Poisson example below shows, all we need is a recipe, like that in (2.1), describing how to produce a sample X, for a given θ, based on a realization $U \sim \mathsf{P}_U$.

Gaussian Example. Consider the problem of inference on the mean θ based on a single sample $X \sim \mathsf{N}(\theta, 1)$. In this case, the association linking X, θ, and an auxiliary variable U may be written as $U = \Phi(X - \theta)$ or, equivalently, $X = \theta + \Phi^{-1}(U)$, where $U \sim \mathsf{Unif}(0, 1)$, and Φ is the standard Gaussian distribution function.

Poisson Example. Consider the problem of inference on the mean θ of a Poisson population based on a single observation X. For this discrete problem, the association for X, given θ, may be written as

$$F_\theta(X - 1) \leq 1 - U < F_\theta(X), \quad U \sim \mathsf{Unif}(0, 1), \tag{2.3}$$

where F_θ denotes the $\mathsf{Pois}(\theta)$ distribution function. This representation is familiar for simulating $X \sim \mathsf{Pois}(\theta)$, i.e., one can first sample $U \sim \mathsf{Unif}(0, 1)$ and then choose X so that the inequalities in (2.3) are satisfied. But here we also interpret (2.3) as a means to link data, parameter, and auxiliary variable.

It should not be surprising that, in general, there are many associations for a given sampling model. In fact, for a given sampling model $\mathsf{P}_{X|\theta}$, there are as many associations as there are triplets $(\mathbb{U}, \mathsf{P}_U, a)$ such that $\mathsf{P}_{X|\theta}$ equals the push-forward measure $\mathsf{P}_U a_\theta^{-1}$, with $a_\theta(\cdot) = a(\theta, \cdot)$. For example, if $X \sim \mathsf{N}(\theta, 1)$, then each of the following defines an association: $X = \theta + U$ with $U \sim \mathsf{N}(0, 1)$, $X = \theta + \Phi^{-1}(U)$ with $U \sim \mathsf{Unif}(0, 1)$, and

$$X = \begin{cases} \theta + U & \text{if } \theta \geq 0, \\ \theta - U & \text{if } \theta < 0, \end{cases} \quad \text{with } U \sim \mathsf{N}(0, 1).$$

Presently, there appears to be no strong reason to choose one of these associations over the other. However, the optimality theory presented in Section 4 helps to resolve this non-uniqueness issue, that is, the optimal IM depends only on the sampling model, and not on the chosen association. From a practical point of view, we prefer, for continuous data problems, associations which are continuous in both θ and U, which rules out the latter of the three associations above. Also, we tend to prefer the representation with a uniform U, any other choice being viewed as just a reparametrization of this one. It will become evident that this view is without loss of generality for simple problems with a one-dimensional auxiliary variable. The case when U is moderate- to high-dimensional is more challenging and we defer its discussion to Section 6.

4

2.2 Three-step IM construction

2.2.1 Association step

The association (2.1) plays two distinct roles. Before the experiment, the association characterizes the predictive probabilities of the observable X. But once $X = x$ is observed, the role of the association changes. The key idea is that the observed x and the unknown θ must satisfy

$$x = a(u^\star, \theta) \tag{2.4}$$

for some unobserved realization $u^\star$ of U. Although $u^\star$ is unobserved, there is information available about the nature of this quantity; in particular, we know exactly the distribution P_U from which it came.

Of course, the value of $u^\star$ can never be known, *but if it were*, the inference problem would be simple: given $X = x$, just solve the equation $x = a(u^\star, \theta)$ for θ. More generally, one could construct the set of solutions $\Theta_x(u^\star)$, where

$$\Theta_x(u) = \{\theta : x = a(u, \theta)\}, \quad x \in \mathbb{X}, \quad u \in \mathbb{U}. \tag{2.5}$$

For continuous-data problems, $\Theta_x(u)$ is typically a singleton for each u; for other problems, it could be a set. In either case, given $X = x$, $\Theta_x(u^\star)$ represents the best possible inference in the sense that *the true θ is guaranteed to be in $\Theta_x(u^\star)$*.

Gaussian Example (cont). The Gaussian mean problem is continuous, so the association $x = \theta + \Phi^{-1}(u)$ identifies a single θ for each fixed (x, u) pair. Therefore, $\Theta_x(u) = \{x - \Phi^{-1}(u)\}$. In this case, clearly, if $u^\star$ were somehow observed, then the true θ could be determined with complete certainty.

Poisson Example (cont). Integration-by-parts reveals that the $\mathsf{Pois}(\theta)$ distribution function F_θ satisfies $F_\theta(x) = 1 - G_{x+1}(\theta)$, where G_a is a $\mathsf{Gamma}(a, 1)$ distribution function. Therefore, from (2.3), we get the u-interval $G_{x+1}(\theta) < u \leq G_x(\theta)$. Inverting this u-interval produces the following θ-interval:

$$\Theta_x(u) = \left(G_x^{-1}(u), G_{x+1}^{-1}(u) \right]. \tag{2.6}$$

If $u^\star$ was available, then $\Theta_x(u^\star)$ would provide the best possible inference in the sense that the true value of θ is guaranteed to sit inside this interval. But even in this ideal case there is no information available to identify the exact location of θ in $\Theta_x(u^\star)$.

2.2.2 Prediction step

The above discussion highlights the importance of the auxiliary variable for inference. It is, therefore, only natural that the inference problem should focus on accurately predicting the unobserved $u^\star$. To predict $u^\star$ with a certain desired accuracy, we employ a so-called a *predictive random set*. First we give the simplest description of a predictive random set and provide a useful example. More general descriptions will be given later.

Let $u \mapsto S(u)$ be a mapping from $\mathbb{U}$ to a collection of P_U-measurable subsets of $\mathbb{U}$; one decent example of such a mapping S is given in equation (2.7) below. Then the predictive random set $\mathcal{S}$ is obtained by applying the set-valued mapping S to a draw $U \sim \mathsf{P}_U$, i.e., $\mathcal{S} = S(U)$ with $U \sim \mathsf{P}_U$. The intuition is that if a draw $U \sim \mathsf{P}_U$ is a good prediction for the unobserved $u^\star$, then the random set $\mathcal{S} = S(U)$ should be even better in the sense that there is high probability that $\mathcal{S} \ni u^\star$.

Gaussian Example (cont). In this example we may predict the unobserved $u^\star$ with a predictive random set $\mathcal{S}$ defined by the set-valued mapping

$$S(u) = \big\{u' \in (0,1) : |u' - 0.5| < |u - 0.5|\big\}, \quad u \in (0,1). \tag{2.7}$$

As this predictive random set is designed to predict an unobserved uniform variate, we may also employ (2.7) in other problems, including the Poisson example.

There are, of course, other choices of $S(u)$, e.g., $[0, u)$, $(u, 1]$, $(0.5u, 0.5 + 0.5u)$ and more. Although some other choice of $\mathcal{S} = S(U)$ might perform slightly better depending on the assertion of interest, (2.7) seems to be a good default choice, provided that the association satisfies certain monotonicity conditions. See Sections 3 and 4 for more on the choice predictive random sets.

For the remainder of this paper, we shall mostly omit the set-valued mapping S from the notation and speak directly about the predictive random set $\mathcal{S}$. That is, the predictive random set $\mathcal{S}$ will be just a random subset of $\mathbb{U}$ with distribution $\mathsf{P}_{\mathcal{S}}$. In the above description, $\mathsf{P}_{\mathcal{S}}$ is just the push-forward measure $\mathsf{P}_U S^{-1}$.

2.2.3 Combination step

For the time being, let us assume that the predictive random set $\mathcal{S}$ is satisfactory for predicting the unobserved $u^\star$; this is actually easy to arrange, but we defer discussion until Section 3. To transfer the available information about $u^\star$ to the θ-space, our last step is to combine the information in the association, the observed $X = x$, and the predictive random set $\mathcal{S}$. The intuition is that, if $u^\star \in \mathcal{S}$, then the true θ must be in the set $\Theta_x(u)$, from (2.5), for at least one $u \in \mathcal{S}$. So, logically, it makes sense to consider, for inference about θ, the expanded set

$$\Theta_x(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_x(u). \tag{2.8}$$

The set $\Theta_x(\mathcal{S})$ contains those values of θ which are consistent with the observed data and sampling model for at least one candidate $u^\star$ value $u \in \mathcal{S}$. Since $\theta \in \Theta_x(\mathcal{S})$ if and only if the unobserved $u^\star \in \mathcal{S}$, if we are willing to accept that the predictive random set $\mathcal{S}$ is satisfactory for predicting $u^\star$, then $\Theta_x(\mathcal{S})$ will do equally well at capturing θ.

Now consider an assertion A about the parameter of interest θ. Mathematically, an assertion is just a subset of Θ, but it acts much like a hypothesis in the context of classical statistics. To summarize the evidence in x that supports the assertion A, we calculate the probability that $\Theta_x(\mathcal{S})$ is a subset of A, i.e.,

$$\mathsf{bel}_x(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \subseteq A \mid \Theta_x(\mathcal{S}) \neq \varnothing\}. \tag{2.9}$$

We refer to $\mathsf{bel}_x(A)$ as the *belief function* at A. Naturally, bel_x also depends on the choice of association and predictive random set, but for now we suppress this dependence in the notation. There are some similarities between our belief function and that of Dempster–Shafer theory (Shafer 1976). For example, bel_x is subadditive in the sense that if A is a non-trivial subset of Θ, then $\mathsf{bel}_x(A) + \mathsf{bel}_x(A^c) \leq 1$ with equality if and only if $\Theta_x(\mathcal{S})$ is a singleton with $\mathsf{P}_{\mathcal{S}}$-probability 1. However, our use of the predictive random set (and our emphasis on validity in Section 3) separates our approach from that of Dempster–Shafer; see Martin et al. (2010).

Here we make two technical remarks about the belief function in (2.9). First, in the problems considered in this paper, the case $\Theta_x(\mathcal{S}) = \varnothing$ is a $\mathsf{P}_{\mathcal{S}}$-null event, so the belief function can be simplified as $\mathsf{bel}_x(A) = \mathsf{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \subseteq A\}$, no conditioning. This simplification may not hold in problems where the observation $X = x$ can induce a constraint on the auxiliary variable u. For example, consider the Gaussian example from above, but suppose that the mean is known to satisfy $\theta \geq 0$. In this case, it is easy to check that $\Theta_x(\mathcal{S}) = \varnothing$ iff $\Phi^{-1}(\inf \mathcal{S}) > x$, an event which generally has positive $\mathsf{P}_{\mathcal{S}}$-probability. So, in general, we can ignore conditioning provided that

$$\Theta_x(u) \neq \varnothing \quad \text{for all } x \text{ and all } u. \tag{2.10}$$

The IM framework can be modified in cases where (2.10) fails, but we will not discuss this here; see Ermini Leaf and Liu (2012). Second, measurability of $\mathsf{bel}_x(A)$, as a function of x for given A, which is important in what follows, is not immediately clear from the definition and should be assessed case-by-case. However, in our experience and in all examples herein, $\mathsf{bel}_x(A)$ is a nice measurable function of x.

Unlike with an ordinary additive probability measure, to reach conclusions about A based on bel_x one must know *both* $\mathsf{bel}_x(A)$ and $\mathsf{bel}_x(A^c)$; for example, in the extreme case of "total ignorance" about A, one has $\mathsf{bel}_x(A) = \mathsf{bel}_x(A^c) = 0$. It is often more convenient to work with a different but related function

$$\mathsf{pl}_x(A) = 1 - \mathsf{bel}_x(A^c) = \mathsf{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \not\subseteq A^c \mid \Theta_x(\mathcal{S}) \neq \varnothing\}, \tag{2.11}$$

called the *plausibility function* at A; when $A = \{\theta\}$ is a singleton, we write $\mathsf{pl}_x(\theta)$ instead of $\mathsf{pl}_x(\{\theta\})$. From the subadditivity of the belief function, it follows that $\mathsf{bel}_x(A) \leq \mathsf{pl}_x(A)$ for all A. In what follows, to summarize the evidence in x supporting A, we shall report the pair $\mathsf{bel}_x(A)$ and $\mathsf{pl}_x(A)$, also known as lower and upper probabilities.

Gaussian Example (cont). With the predictive random set $\mathcal{S}$ in (2.7), the random set $\Theta_x(\mathcal{S})$ is given by

$$\begin{aligned}
\Theta_x(\mathcal{S}) &= \bigcup_{u \in \mathcal{S}}\{x - \Phi^{-1}(u)\} \\
&= \left(x - \Phi^{-1}(0.5 + |U - 0.5|),\, x - \Phi^{-1}(0.5 - |U - 0.5|)\right) \\
&= \left(\underline{\Theta}_x(U),\, \overline{\Theta}_x(U)\right), \quad \text{say,}
\end{aligned}$$

where $U \sim \mathsf{Unif}(0, 1)$. For a singleton assertion $A = \{\theta\}$, it is easy to see that the belief function is zero. But the plausibility function is

$$\begin{aligned}
\mathsf{pl}_x(\theta) &= 1 - \mathsf{P}_U\{\underline{\Theta}_x(U) > \theta\} - \mathsf{P}_U\{\overline{\Theta}_x(U) < \theta\} \\
&= 1 - |2\Phi(x - \theta) - 1|. \tag{2.12}
\end{aligned}$$

A plot of $\mathsf{pl}_x(\theta)$, with $x = 5$, as a function of θ, is shown in Figure 1(a). The symmetry around the observed x is apparent, and all those θ values in a neighborhood of $x = 5$ are relatively plausible. See Section 3.4 for more statistical applications of this graph.

Poisson Example (cont). With the same predictive random set as in the previous example, the random set $\Theta_x(\mathcal{S})$ is given by

$$\begin{aligned}
\Theta_x(\mathcal{S}) &= \bigcup_{u \in \mathcal{S}}\left(G_x^{-1}(u), G_{x+1}^{-1}(u)\right] \\
&= \left(G_x^{-1}(0.5 - |U - 0.5|), G_{x+1}^{-1}(0.5 + |U - 0.5|)\right) \\
&= \left(\underline{\Theta}_x(U),\, \overline{\Theta}_x(U)\right), \quad \text{say,}
\end{aligned}$$

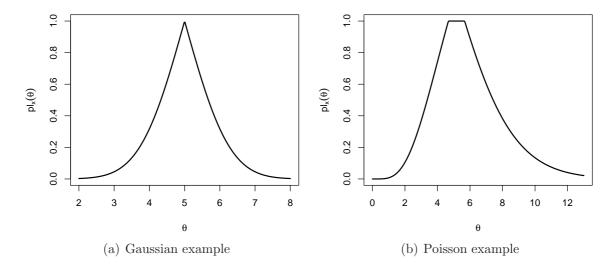(a) Gaussian example   (b) Poisson example

Figure 1: Plot of the plausibility functions $\mathsf{pl}_x(\theta)$, as functions of θ, in (a) the Gaussian example and (b) the Poisson example. In both cases, $X = 5$ is observed.

where U is a random draw from $\mathsf{Unif}(0,1)$. For a singleton assertion $A = \{\theta\}$, again the belief function is zero, but the plausibility function is

$$\mathsf{pl}_x(\theta) = 1 - \mathsf{P}_U\{\underline{\Theta}_x(U) > \theta\} - \mathsf{P}_U\{\overline{\Theta}_x(U) < \theta\}$$
$$= 1 - \max\{1 - 2G_x(\theta), 0\} - \max\{2G_{x+1}(\theta) - 1, 0\}. \tag{2.13}$$

A graph of $\mathsf{pl}_x(\theta)$, with $x = 5$, as a function of θ is shown in Figure 1(b). The plateau indicates that no θ values in a neighborhood of 5 can be ruled out. Like in the Gaussian example, θ values in an interval around 5 are all relatively plausible.

 Dempster (2008) gives a different analysis of this same Poisson problem. His plausibility function for the singleton assertion $A = \{\theta\}$ is $r_x(\theta) = e^{-\theta}\theta^x/x!$, which is the Poisson mass function treated as a function of θ. This function has a similar shape to that in Figure 1(b), but the scale is much smaller. For example, $r_5(5) = 0.175$, suggesting that the assertion $\{\theta = 5\}$ is relatively implausible, even though $X = 5$ was observed. Compare this to $\mathsf{pl}_5(5) = 1$. We would argue that, if $X = 5$ is observed, then no plausibility function threshold should be able to rule out $\{\theta = 5\}$; in that case, $\mathsf{pl}_5(5) = 1$ makes more sense. Furthermore, as Dempster's analysis is similar to ours but with an invalid predictive random set, namely, $\mathcal{S} = \{U\}$, with $U \sim \mathsf{Unif}(0,1)$, the corresponding plausibility function is not properly calibrated for all assertions.

2.3 Interpretation of the belief function

It is clear that the belief function depends on the observed x and so must be meaningful within the problem at hand. But while it is data-dependent, $\mathsf{bel}_x(A)$ is not a posterior probability for A in the familiar Bayesian sense. In fact, under our assumption that θ is fixed and non-random, there can be no non-trivial posterior distribution on Θ. The way around this limitation is to drop the requirement that posterior inference be based on a bona fide probability measure (e.g., Heath and Sudderth 1978; Walley 1996; Wasserman

1990). Therefore, we recommend interpreting $\mathsf{bel}_x(A)$ and $\mathsf{bel}_x(A^c)$ as degrees of belief, rather than ordinary probabilities, even though they manifest from P_U-probability calculations. More precisely, $\mathsf{bel}_x(A)$ and $\mathsf{bel}_x(A^c)$ represent the knowledge gained about the respective claims $\theta \in A$ and $\theta \notin A$ based on both the observed x and prediction of the auxiliary variable.

2.4   Summary

The familiar sampling model appears in the A-step, but it is the corresponding association which is of primary importance. This association, in turn, determines the auxiliary variable which is to be the focus of the IM framework. We propose to predict the unobserved value of this auxiliary variable in the P-step with a predictive random set $\mathcal{S}$, which is chosen to have certain desirable properties (see Definition 1 below). This use of a predictive random set is likely the aspect of the IM framework which is most difficult to swallow, but the intuition should be clear: one cannot hope to accurately predict a fixed value $u^\star$ by an ordinary continuous random variable. With the association, predictive random set, and observed $X = x$ in hand, one proceeds to the C-step where a random set $\Theta_x(\mathcal{S})$ on the parameter space is obtained. As this random set corresponds to a set of "reasonable" θ values, given x, it is natural to summarize the support of an assertion A by the probability that $\Theta_x(\mathcal{S})$ is a subset of A. This probability is exactly the belief function that characterizes the output of the IM and an argument is presented that justifies the meaningfulness of $\mathsf{bel}_x(A)$ and $\mathsf{pl}_x(A)$ as summaries of the evidence in favor of A.

Finally, we mention that the predictive random set $\mathcal{S}$ can depend on the assertion A of interest. That is, one might consider using one predictive random set, say $\mathcal{S}_A$, to evaluate $\mathsf{bel}_x(A)$, and another predictive random set, say $\mathcal{S}_{A^c}$, to evaluate $\mathsf{pl}_x(A) = 1 - \mathsf{bel}_x(A^c)$. In Section 4 we show that this is actually a desirable strategy, in the sense that the optimal predictive random set depends on the assertion in question. In what follows, this dependence of the predictive random set on the assertion will be kept implicit.

3   Theoretical validity of IMs

3.1   Intuition

In Section 2 we argued that $\mathsf{bel}_x(A; \mathcal{S})$ and $\mathsf{pl}_x(A; \mathcal{S})$ together provide a meaningful summary of evidence in favor of A for the given $X = x$; our notation now explicitly indicates the dependence of these function on the predictive random set $\mathcal{S}$. In this section we show that $\mathsf{bel}_X(A; \mathcal{S})$ and $\mathsf{pl}_X(A; \mathcal{S})$ are also meaningful as functions of the random variable $X \sim \mathsf{P}_{X|\theta}$ for a fixed assertion A. For example, we show that $\mathsf{bel}_X(A)$ is frequency-calibrated in the following sense: if $\theta \notin A$, then $\mathsf{P}_{X|\theta}\{\mathsf{bel}_X(A; \mathcal{S}) \geq 1 - \alpha\} \leq \alpha$ for each $\alpha \in [0, 1]$. In other words, the amount of evidence in favor of a false A can be large with only small probability. This property means that the belief function is appropriately scaled for objective scientific inference. A similar property also holds for $\mathsf{pl}_X(A)$. We refer to this frequency-calibration property as *validity* (Definition 2). Bernardo (1979), Rubin (1984) and Dawid (1985) give similar investigations of frequency-calibration and of objective priors for Bayesian inference.

3.2 Predictive random sets

We start with a few definitions, similar to those found in Martin et al. (2010) and Zhang and Liu (2011). Set $Q_\mathcal{S}(u) = \mathsf{P}_\mathcal{S}\{\mathcal{S} \not\ni u\}$, for $u \in \mathbb{U}$, which is the probability that the predictive random set $\mathcal{S}$ misses the specified target u. Ideally, $\mathcal{S}$ will be such that the random variable $Q_\mathcal{S}(U)$, a function of $U \sim \mathsf{P}_U$, will be probabilistically small. This suggests a connection between $\mathsf{P}_\mathcal{S}$ and P_U which will be made precise in Theorem 1.

Definition 1. A predictive random set $\mathcal{S}$ is *valid* for predicting the unobserved auxiliary variable if $Q_\mathcal{S}(U)$, as a function of $U \sim \mathsf{P}_U$, is stochastically no larger than $\mathsf{Unif}(0, 1)$, i.e., for each $\alpha \in (0, 1)$, $\mathsf{P}_U\{Q_\mathcal{S}(U) \geq 1 - \alpha\} \leq \alpha$. If "$\leq \alpha$" can be replaced by "$= \alpha$," then $\mathcal{S}$ is *efficient*.

In words, validity of $\mathcal{S}$ implies that the probability that it misses a target u is large for only a small P_U-proportion of possible u values. The predictive random set $\mathcal{S}$ defined by the mapping (2.7) is both valid and efficient. Indeed, it is easy to check that, in this case, $Q_\mathcal{S}(u) = |2u - 1|$. Therefore, if $U \sim \mathsf{Unif}(0, 1)$ then $Q_\mathcal{S}(U) \sim \mathsf{Unif}(0, 1)$ too. Corollary 1 below gives a simple and general recipe for constructing a valid and efficient $\mathcal{S}$.

There is an important and apparently fundamental concept related to validity of predictive random sets, namely, *nesting*. We say that a collection of sets $\mathbb{S} \subseteq 2^\mathbb{U}$ is nested if, for any pair of sets S and S' in $\mathbb{S}$, we have $S \subseteq S'$ or $S' \subseteq S$. We shall also implicitly assume, without loss of generality, that $\mathsf{P}_U(S) > 0$ for some $S \in \mathbb{S}$; the user can easily arrange this. The following theorem shows that if the predictive random set $\mathcal{S}$ is nested, i.e., if $\mathcal{S}$ is supported on a nested collection of sets $\mathbb{S}$, then it is valid.

Theorem 1. *Let $\mathbb{S} \subseteq 2^\mathbb{U}$ be a nested collection of P_U-measurable subsets S of $\mathbb{U}$. Define a predictive random set $\mathcal{S}$ with distribution $\mathsf{P}_\mathcal{S}$, supported on $\mathbb{S}$, such that*

$$\mathsf{P}_\mathcal{S}\{\mathcal{S} \subseteq K\} = \sup_{S \in \mathbb{S}: S \subseteq K} \overline{\mathsf{P}}_U(S), \quad K \subseteq \mathbb{U},$$

where $\overline{\mathsf{P}}_U(\cdot) = \mathsf{P}_U(\cdot) / \sup_{S \in \mathbb{S}} \mathsf{P}_U(S)$. Then $\mathcal{S}$ is valid in the sense of Definition 1.

Proof. The idea of the proof is that $Q_\mathcal{S}(u) = \mathsf{P}_\mathcal{S}\{\mathcal{S} \not\ni u\}$ is large iff u sits outside a set that contains most realizations of $\mathcal{S}$. To make this formal, take any $\alpha \in (0, 1)$ and let $S_\alpha = \bigcap\{S \in \mathbb{S} : \mathsf{P}_U(S) \geq 1 - \alpha\}$ be the smallest set in $\mathbb{S}$ with P_U-probability no less than $1 - \alpha$; here, the intersection over an empty collection of sets is taken to be $\mathbb{U}$. Since $\mathbb{S}$ is nested, $S_\alpha \in \mathbb{S}$, $\mathsf{P}_U(S_\alpha) \geq 1 - \alpha$, and

$$\mathsf{P}_\mathcal{S}\{\mathcal{S} \subseteq S_\alpha\} = \sup_{S \in \mathbb{S}: S \subseteq S_\alpha} \overline{\mathsf{P}}_U(S) = \overline{\mathsf{P}}_U(S_\alpha) \geq \mathsf{P}_U(S_\alpha) \geq 1 - \alpha.$$

Therefore, since $Q_\mathcal{S}(u) > 1 - \alpha$ iff $u \notin S_\alpha$, we get $\mathsf{P}_U\{Q_\mathcal{S}(U) > 1 - \alpha\} = \mathsf{P}_U(S_\alpha^c) = 1 - \mathsf{P}_U(S_\alpha) \leq \alpha$. Finally, validity follows since α was arbitrary. $\qquad\square$

It is clear that, if P_U is absolutely continuous and the nested support $\mathbb{S}$ is sufficiently rich, then the predictive random set defined above is also efficient. Specifically, if $\mathbb{U} \in \mathbb{S}$ and, for S_α defined in the proof above, $\mathsf{P}_U(S_\alpha) = 1 - \alpha$ for every $\alpha \in (0, 1)$. This vague argument for efficiency is made more precise in the next important special case.

Corollary 1. *Suppose the P_U is non-atomic, and let h be a real-valued function on $\mathbb{U}$. Then the predictive random set $\mathcal{S} = \{u \in \mathbb{U} : h(u) < h(U)\}$, with $U \sim \mathsf{P}_U$, is valid. If h is continuous and constant only on P_U-null sets, then it is also efficient.*

Proof. Validity is a consequence of Theorem 1 and the fact that this $\mathcal{S}$ is nested. To prove the efficiency claim, let H be the distribution function of $h(U)$ when $U \sim \mathsf{P}_U$. Then, for $u \in \mathbb{U}$, $Q_{\mathcal{S}}(u) = \mathsf{P}_U\{h(U) \leq h(u)\} = H(h(u))$. If h satisfies the stated conditions, then $h(U)$ is a continuous random variable. Therefore, if $U \sim \mathsf{P}_U$, then $Q_{\mathcal{S}}(U) = H(h(U)) \sim \mathsf{Unif}(0,1)$, so efficiency follows. $\qquad\square$

The above results demonstrate that nesting is a sufficient condition for predictive random set validity. But nesting is not a necessary condition (Martin et al. 2010). The real issue, however, is the performance of the corresponding IM. We show in Section 4 that for any non-nested predictive random set $\mathcal{S}$, there is a nested predictive random set $\mathcal{S}'$ such that the IM based on $\mathcal{S}'$ is "at least as good" as that based on $\mathcal{S}$.

3.3   IM validity

Validity of the underlying predictive random set $\mathcal{S}$ is essentially all that is needed to prove the meaningfulness of the corresponding IM/belief function. Here meaningfulness refers to a calibration property of the belief function.

Definition 2. Suppose $X \sim \mathsf{P}_{X|\theta}$ and let A be an assertion of interest. Then the IM with belief function bel_x is *valid for A* if, for each $\alpha \in (0,1)$,

$$\sup_{\theta \notin A} \mathsf{P}_{X|\theta}\big\{\mathsf{bel}_X(A;\mathcal{S}) \geq 1-\alpha\big\} \leq \alpha. \qquad (3.1)$$

The IM is *valid* if it is valid for all A.

By (2.11), the validity property can also be stated in terms of the plausibility function. That is, the IM is valid if, for all assertions A and for any $\alpha \in (0,1)$,

$$\sup_{\theta \in A} \mathsf{P}_{X|\theta}\big\{\mathsf{pl}_X(A;\mathcal{S}) \leq \alpha\big\} \leq \alpha. \qquad (3.2)$$

Theorem 2. *Suppose the predictive random set $\mathcal{S}$ is valid, and $\Theta_x(\mathcal{S}) \neq \varnothing$ with $\mathsf{P}_{\mathcal{S}}$-probability 1 for all x. Then the IM is valid.*

Proof. For any A, take (x, u, θ) such that $\theta \notin A$ and $x = a(\theta, u)$. Since $A \subseteq \{\theta\}^c$, $\mathsf{bel}_x(A;\mathcal{S}) \leq \mathsf{bel}_x(\{\theta\}^c;\mathcal{S}) = \mathsf{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \not\ni \theta\} = \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \not\ni u\}$ by monotonicity. Validity of $\mathcal{S}$ implies that the right-hand side, as a function of $U \sim \mathsf{P}_U$, is stochastically smaller than $\mathsf{Unif}(0,1)$. This, in turn, implies the same of $\mathsf{bel}_X(A;\mathcal{S})$ as a function of $X \sim \mathsf{P}_{X|\theta}$. Therefore, $\mathsf{P}_{X|\theta}\{\mathsf{bel}_X(A;\mathcal{S}) \geq 1-\alpha\} \leq \mathsf{P}\{\mathsf{Unif}(0,1) \geq 1-\alpha\} = \alpha$. Taking a supremum over $\theta \notin A$ on the left-hand side completes the proof. $\qquad\square$

A key feature of the validity theorem above is that it holds under minimal conditions on the predictive random set. Validity of the IM does not depend on the particular form of predictive random set, only that it is valid. Recall that the condition "$\Theta_x(\mathcal{S}) \neq \varnothing$ with $\mathsf{P}_{\mathcal{S}}$-probability 1" holds whenever (2.10) holds. See, also, Ermini Leaf and Liu (2012).

The following corollary states that the validity theorem remains true even after a suitable—possibly θ-dependent—change of auxiliary variable. In other words, the validity property is independent of the choice of auxiliary variable parametrization. This reparametrization comes in handy in examples, including those in Section 5.

Corollary 2. *Consider a one-to-one transformation $v = \varphi_\theta(u)$ such that the push-forward measure $\mathsf{P}_V = \mathsf{P}_U \varphi_\theta^{-1}$ on $\mathbb{V} = \varphi_\theta(\mathbb{U})$ does not depend on θ. Suppose $\mathcal{S}$ is valid for predicting $v^\star = \varphi_\theta(u^\star)$, and $\Theta_x(\mathcal{S}) \neq \varnothing$ with $\mathsf{P}_\mathcal{S}$-probability 1 for all x. Then the corresponding belief function satisfies* (3.1) *and the transformed IM is valid.*

3.4   Application: IM-based frequentist procedures

In addition to providing problem-specific measures of certainty about various assertions of interest, the belief/plausibility functions can easily be used to create frequentist procedures. First consider testing $H_0 : \theta \in A$ versus $H_1 : \theta \in A^c$. Then an IM-based counterpart to a frequentist testing rule is of the following form:

$$\text{Reject } H_0 \text{ if } \mathsf{pl}_x(A) \leq \alpha, \text{ for a specified } \alpha \in (0,1). \tag{3.3}$$

According to (3.2) and Theorem 2, if the predictive random set $\mathcal{S}$ is valid, then the probability of a Type I error for such a rejection rule is $\sup_{\theta \in A} \mathsf{P}_{X|\theta}\{\mathsf{pl}_X(A) \leq \alpha\} \leq \alpha$. Therefore, the test (3.3) controls the probability of a Type I error at level α.

Next consider the class of singleton assertions $\{\theta\}$, with $\theta \in \Theta$. As a counterpart to a frequentist confidence region, define the $100(1-\alpha)\%$ *plausibility region*

$$\Pi_x(\alpha) = \{\theta : \mathsf{pl}_x(\theta) > \alpha\}. \tag{3.4}$$

Now the coverage probability of the plausibility region (3.4) is

$$\mathsf{P}_{X|\theta}\{\Pi_X(\alpha) \ni \theta\} = \mathsf{P}_{X|\theta}\{\mathsf{pl}_X(\theta) > \alpha\} = 1 - \mathsf{P}_{X|\theta}\{\mathsf{pl}_X(\theta) \leq \alpha\} \geq 1 - \alpha,$$

where the last inequality follows from Theorem 2. Therefore, this plausibility region has at least the nominal coverage probability.

Gaussian Example (cont). Suppose $X = 5$. Then, using the predictive random set $\mathcal{S}$ in (2.7), the plausibility function is $\mathsf{pl}_5(\theta) = 1 - |2\Phi(5 - \theta) - 1|$. The 90% plausibility interval for θ, determined by the inequality $\mathsf{pl}_5(\theta) > 0.10$, is $5 \pm \Phi^{-1}(0.05)$, the same as the classical 90% z-interval for θ given in standard textbooks.

Poisson Example (cont). For the predictive random set determined by $\mathcal{S}$ in (2.7), the plausibility function $\mathsf{pl}_x(\theta)$ is displayed in (2.13). For observed $X = 5$, a 90% plausibility interval for θ, characterized by the inequality $\mathsf{pl}_5(\theta) > 0.10$, is $(1.97, 10.51)$. This interval is not the best possible; in fact, the one presented in Section 4.3.2 is better. But these plausibility intervals have exact coverage properties, which means that they may be too conservative at certain θ values for practical use. This is the case for all exact intervals in discrete data problems (e.g., Brown et al. 2003; Cai 2005).

4 Theoretical optimality of IMs

4.1 Intuition

Martin et al. (2010) showed that Fisher's fiducial inference and Dempster–Shafer theory are special cases of the IM framework corresponding to a singleton predictive random set. But it is easy to show that, for some assertions A, the fiducial probability is not valid in the sense of Definition 2. To correct for this bias, we propose to replace the singleton with some larger $\mathcal{S}$. But taking $\mathcal{S}$ to be too large will lead to inefficient inference. So the goal is to take $\mathcal{S}$ just large enough that validity is achieved.

4.2 Preliminaries

Throughout the subsequent discussion, we shall assume (2.10), i.e., $\Theta_x(u) \neq \varnothing$ for all x and u. This allows us to ignore conditioning in the definition of belief functions.

For the predictive random set $\mathcal{S}_0 = \{U\}$, with $U \sim \mathsf{Unif}(0,1)$, the belief function at A is $\mathsf{bel}_x(A; \mathcal{S}_0) = \mathsf{P}_U\{\Theta_x(U) \subseteq A\}$, where $\Theta_x(u) = \{\theta : x = a(\theta, u)\}$ as in (2.5). This is exactly the fiducial probability for A given $X = x$. For a general predictive random set $\mathcal{S}$, we have $\mathsf{bel}_x(A; \mathcal{S}) = \mathsf{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \subseteq A\}$, where $\Theta_x(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_x(u)$ is defined in (2.8). In light of the discussion in Section 4.1, we shall compare the two belief functions $\mathsf{bel}_x(A; \mathcal{S})$ and $\mathsf{bel}_x(A; \mathcal{S}_0)$. Towards this, we have the following result which says that the fiducial probability is an upper bound for the belief function.

Proposition 1. *If* (2.10) *holds and the predictive random set $\mathcal{S}$ is valid in the sense of Definition 1, then $\mathsf{bel}_x(A; \mathcal{S}) \leq \mathsf{bel}_x(A; \mathcal{S}_0)$ for each fixed x.*

Proof. Let $\mathbb{U}_x(A) = \{u : \Theta_x(u) \subseteq A\}$; note that $\mathcal{S} \subseteq \mathbb{U}_x(A)$ iff $\Theta_x(\mathcal{S}) \subseteq A$. Also, put $b = \mathsf{bel}_x(A; \mathcal{S})$ and $b_0 = \mathsf{bel}_x(A; \mathcal{S}_0) \equiv \mathsf{P}_U\{\mathbb{U}_x(A)\}$. If $u \notin \mathbb{U}_x(A)$, then

$$Q_{\mathcal{S}}(u) = \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \not\ni u\} \geq \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq \mathbb{U}_x(A)\} = \mathsf{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \subseteq A\} = b.$$

Therefore, $\mathsf{P}_U\{Q_{\mathcal{S}}(U) \geq b\} \geq \mathsf{P}_U\{\mathbb{U}_x(A)^c\} = 1 - b_0$. Also, validity of $\mathcal{S}$ implies $\mathsf{P}_U\{Q_{\mathcal{S}}(U) \geq b\} \leq 1 - b$. Consequently, $1 - b_0 \leq 1 - b$, i.e., $\mathsf{bel}_x(A; \mathcal{S}) \leq \mathsf{bel}_x(A; \mathcal{S}_0)$. $\square$

For given assertion A and predictive random set $\mathcal{S}$, consider the ratio

$$R_A(x; \mathcal{S}) = \mathsf{bel}_x(A; \mathcal{S})/\mathsf{bel}_x(A; \mathcal{S}_0), \quad x \in \mathbb{X}. \tag{4.1}$$

We call this the relative efficiency of the IM based on $\mathcal{S}$ compared to fiducial. Proposition 1 guarantees that this ratio is bounded by unity, provided that the denominator $\mathsf{bel}_x(A; \mathcal{S}_0)$ is non-zero. Our main goal is to choose $\mathcal{S}$ to make this ratio large in some sense. Towards this goal, we have the following "complete-class theorem" which says that nested predictive random sets—which, by Theorems 1 and 2, produce valid IMs—are the only kind of predictive random sets under consideration.

Theorem 3. *Fix $A \subseteq \Theta$ and assume* (2.10). *Given any predictive random set $\mathcal{S}$, there exists a nested predictive random set $\mathcal{S}'$ such that $R_A(x; \mathcal{S}') \geq R_A(x; \mathcal{S})$ for each x.*

13

Proof. Given $\mathcal{S}$, construct a collection $\mathbb{S} = \{S_x : x \in \mathbb{X}\}$ as follows:

$$S_x = \bigcap_{x':\mathsf{bel}_{x'}(A;\mathcal{S}) \geq \mathsf{bel}_x(A;\mathcal{S})} \mathbb{U}_{x'}(A),$$

where $\mathbb{U}_x(A)$ is defined in the proof of Proposition 1. This collection $\mathbb{S}$, which will serve as the support for the new $\mathcal{S}'$, is clearly nested. Indeed, if $\mathsf{bel}_{x_1}(A;\mathcal{S}) \leq \mathsf{bel}_{x_2}(A;\mathcal{S})$, then $S_{x_2} \subseteq S_{x_1}$. The distribution $\mathsf{P}_{\mathcal{S}'}$ of $\mathcal{S}'$ is defined as

$$\mathsf{P}_{\mathcal{S}'}\{\mathcal{S}' \subseteq K\} = \sup_{x:S_x \subseteq K} \bar{b}(x), \quad K \subseteq \mathbb{U},$$

where $\bar{b}(t) = \mathsf{bel}_t(A;\mathcal{S})/\sup_x \mathsf{bel}_x(A;\mathcal{S})$ is the normalized belief function. In particular, for $K = S_x$, we have $\mathsf{P}_{\mathcal{S}'}\{\mathcal{S}' \subseteq S_x\} = \bar{b}(x) \geq \mathsf{bel}_x(A;\mathcal{S})$. Then we have

$$\begin{aligned}
\mathsf{bel}_x(A;\mathcal{S}') &= \mathsf{P}_{\mathcal{S}'}\{\Theta_x(\mathcal{S}') \subseteq A\} \\
&= \mathsf{P}_{\mathcal{S}'}\{\mathcal{S}' \subseteq \mathbb{U}_x(A)\} \\
&\geq \mathsf{P}_{\mathcal{S}'}\{\mathcal{S}' \subseteq S_x\} \\
&\geq \mathsf{bel}_x(A;\mathcal{S}),
\end{aligned}$$

where the second equality is due to the fact that $\Theta_x(\mathcal{S}') \subseteq A$ iff $\mathcal{S}' \subseteq \mathbb{U}_x(A)$, and the first inequality is by monotonicity of $\mathsf{P}_{\mathcal{S}'}\{\mathcal{S}' \subseteq \cdot\}$ and the fact that $S_x \subseteq \mathbb{U}_x(A)$ for each x. Therefore, $R_A(x;\mathcal{S}')$ can be no less than $R_A(x;\mathcal{S})$ for each x, proving the claim. $\square$

We omit the details, but if the collection $\mathbb{U}_x(A)$ in the proof of Proposition 1 is itself nested, then, in general, one can construct an "optimal" $\mathcal{S}$ such that $R_A(X;\mathcal{S}) \equiv 1$. This is done explicitly for the special case in Section 4.3.1 below.

4.3 Optimality in special cases

Throughout this section, we will focus on scalar X and θ. However, this is just for simplicity, and not a limitation of the method; see Section 5. Indeed, special dimension-reduction techniques, akin to Fisher's theory of sufficient statistics, are available to reduce the dimension of observed X to that of θ within the IM framework; see Martin and Liu (2012, 2013). Also, there is no conceptual difference between scalar and vector θ problems so, since the ideas are new, we prefer to keep the presentation as simple as possible.

4.3.1 One-sided assertions

Here we consider a one-sided assertion, e.g., $A = \{\theta \in \Theta : \theta < \theta_0\}$, where θ_0 is fixed. This "left-sided" assertion is the kind we shall focus on, but other one-sided assertions can be handled similarly. In this context, we can consider a very strong definition of optimality.

Definition 3. Fix a left-sided assertion A. For two nested predictive random sets $\mathcal{S}$ and $\mathcal{S}'$, the IM based on $\mathcal{S}$ is said to be more efficient than that based on $\mathcal{S}'$ if, as functions of $X \sim \mathsf{P}_{X|\theta}$ for any $\theta \in A$, $R_A(X;\mathcal{S})$ is stochastically larger than $R_A(X;\mathcal{S}')$. The IM based on $\mathcal{S}^\star$ is optimal, or most efficient, if $R_A(X;\mathcal{S}^\star)$ is stochastically largest, in the sense above, among all nested predictive random sets.

14

That optimality here is described via a stochastic ordering property is natural in light of the notion of validity used throughout. This definition particularly strong because it concerns the full distribution of $R_A(X, \mathcal{S})$ as a function of $X \sim \mathsf{P}_{X|\theta}$, not just a functional thereof. Next we establish a strong optimality result for one-sided assertions; when the assertion is not one-sided, it may not be possible to establish such a strong result.

Theorem 4. *Let $A = \{\theta \in \Theta : \theta < \theta_0\}$ be a left-sided assertion. Suppose that $\Theta_x(u)$, defined in (2.5), is such that, for each x, the right endpoint $\sup \Theta_x(u)$ is a non-decreasing (resp. non-increasing) function of u. Then, for the given A, the optimal predictive random set is $\mathcal{S}^\star = [0, U]$ (resp. $\mathcal{S}^\star = [U, 1]$), where $U \sim \mathsf{Unif}(0, 1)$.*

Proof. First observe that both forms of $\mathcal{S}^\star$ are nested. We shall focus on the non-decreasing case only; the other case is similar. Since $\sup \Theta_x(u)$ is non-decreasing in u, it follows that $\sup \Theta_x([0, U]) = \sup \Theta_x(U)$. Therefore,

$$\mathsf{bel}_x(A; \mathcal{S}^\star) = \mathsf{P}_U\{\sup \Theta_x([0, U]) < \theta_0\} = \mathsf{P}_U\{\sup \Theta_x(U) < \theta_0\} = \mathsf{bel}_x(A; \mathcal{S}_0).$$

This holds for all x, so $R_A(\cdot; \mathcal{S}^\star) \equiv 1$, its upper bound. Consequently, $R_A(X; \mathcal{S}^\star)$ is stochastically larger than $R_A(X; \mathcal{S})$ for any other $\mathcal{S}$, so optimality of $\mathcal{S}^\star$ obtains. □

Gaussian Example (cont). We showed previously that $\Theta_x(u) = \{x - \Phi^{-1}(u)\}$. If we treat this as a degenerate interval, then we see that the right endpoint $x - \Phi^{-1}(u)$ is a strictly decreasing function of u. Therefore, by Theorem 4, the optimal predictive random set for a left-sided assertion is $\mathcal{S}^\star = [U, 1]$, $U \sim \mathsf{Unif}(0, 1)$.

As an application, consider the testing problem $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$. If we take $A = (-\infty, \theta_0)$, then the IM-based rule (3.3) rejects H_0 iff $1 - \mathsf{bel}_x(A; \mathcal{S}^\star) \leq \alpha$. With the optimal $\mathcal{S}^\star = [U, 1]$ as above, we get $\mathsf{bel}_x(A; \mathcal{S}^\star) = \Phi(\theta_0 - x)$. So the IM-based testing rule rejects H_0 iff $\Phi(\theta_0 - x) \geq 1 - \alpha$ or, equivalently, iff $x \leq \theta_0 - \Phi^{-1}(1 - \alpha)$. The reader will recognize this as the uniformly most powerful size-α test based on the classical Neyman–Pearson theory.

Poisson Example (cont). In this case, $\Theta_x(u) = (G_x^{-1}(u), G_{x+1}^{-1}(u)]$; see (2.6). The right endpoint $G_{x+1}^{-1}(u)$ is strictly increasing in u. So Theorem 4 states that, for left-sided assertions, the optimal predictive random set is $\mathcal{S}^\star = [0, U]$, $U \sim \mathsf{Unif}(0, 1)$. The same connection with the Neyman–Pearson uniformly most powerful test in the Gaussian example holds here as well, but we omit the details.

4.3.2 Two-sided assertions

Consider the case where $A = \{\theta_0\}^c$ is the two-sided assertion of interest, with θ_0 a fixed interior point of $\Theta \subseteq \mathbb{R}$. This is an important case, which we have already considered in Section 2, just in a different form. These problems are apparently more difficult than their one-sided counterparts, just like in the classical hypothesis testing context. Here we present some basic results and intuitions on local IM optimality for two-sided assertions.

Assume $\mathsf{P}_{X|\theta}$ is continuous. Then the fiducial probability $\mathsf{bel}_X(\{\theta_0\}^c; \mathcal{S}_0)$ for the two-sided assertion is unity, and so the relative efficiency (4.1) is simply $\mathsf{bel}_x(\{\theta_0\}^c; \mathcal{S})$. Here we focus on predictive random sets $\mathcal{S}$ with the property that $\mathsf{bel}_X(\{\theta_0\}^c; \mathcal{S}) \sim \mathsf{Unif}(0, 1)$ under $\mathsf{P}_{X|\theta_0}$; see Corollary 1. Based on the intuition developed in Section 2, $\mathsf{bel}_X(\{\theta_0\}^c; \mathcal{S})$

should be smallest (probabilistically) under $\mathsf{P}_{X|\theta}$ for $\theta = \theta_0$. We shall, therefore, impose the following condition on the predictive random set $\mathcal{S}$:

$$\mathsf{P}_{X|\theta}\{\mathsf{bel}_X(\{\theta_0\}^c; \mathcal{S}) \leq \alpha\} < \alpha, \quad \forall\, \theta \neq \theta_0, \quad \forall\, \alpha \in (0,1). \tag{4.2}$$

Roughly speaking, condition (4.2) states that the belief function at $\{\theta_0\}^c$ is stochastically larger under $\mathsf{P}_{X|\theta}$ than under $\mathsf{P}_{X|\theta_0}$. There is also a loose connection between (4.2) and the classical unbiasedness condition imposed to construct optimal tests when the alternative hypothesis is two-sided (Lehmann and Romano 2005, Ch. 4). Our goal in what follows is to find a "best" predictive random set that satisfies (4.2).

To make things formal, suppose that both $\mathbb{X}$ and Θ are one-dimensional, that $\mathsf{P}_{X|\theta}$ is continuous with distribution function $F_\theta(x)$ and density function $f_\theta(x)$, and that the usual regularity conditions hold; in particular, we assume that the order of expectation with respect to $\mathsf{P}_{X|\theta}$ and differentiation with respect to θ can be interchanged. Note that we have fixed a parametrization, and the analysis that follows depends on this selection. Let $T_\theta(x) = (\partial/\partial\theta) \log f_\theta(x)$ be the score function, an important quantity in what follows. Also, let $V_\theta(x) = T_\theta(x)^2 + (\partial/\partial\theta)T_\theta(x)$. Then, under the usual regularity conditions, we have $\mathsf{E}_{X|\theta}\{T_\theta(X)\} = 0$ and $\mathsf{E}_{X|\theta}\{V_\theta(X)\} = 0$ for all θ.

In Appendix A we argue that a good predictive random set $\mathcal{S}$ must have a support with certain symmetry or balance properties with respect to the sampling distribution of $T_{\theta_0}(X)$. In particular, let $B = \{B_t : t \in \mathbb{T}\}$ be a generic collection of nested measurable subsets of $\mathbb{T} = T_{\theta_0}(\mathbb{X})$. The collection B shall be called *score-balanced* if

$$\mathsf{E}_{X|\theta_0}\{T_{\theta_0}(X)I_{B_t}(T_{\theta_0}(X))\} = 0, \quad \forall\, t \in \mathbb{T}. \tag{4.3}$$

For a score-balanced collection $B = \{B_t\}$ satisfying (4.3) we can define a corresponding *score-balanced predictive random set* $\mathcal{S} = \mathcal{S}_B$ as follows. Define the class $\mathbb{S} = \{S_t : t \in \mathbb{T}\}$ of subsets of $\mathbb{U} = [0,1]$ given by

$$S_t = F_{\theta_0}\big(\{x : T_{\theta_0}(x) \in B_t\}\big).$$

For simplicity, and without loss of generality, assume $\mathbb{S}$ contains $\varnothing$ and $\mathbb{U}$. Now take a predictive random set $\mathcal{S}_B$, supported on $\mathbb{S}$, such that its measure $\mathsf{P}_{\mathcal{S}_B}$ satisfies

$$\mathsf{P}_{\mathcal{S}_B}\{\mathcal{S}_B \subseteq K\} = \sup_{t:S_t \subseteq K} \mathsf{P}_U(S_t), \quad K \subseteq [0,1],$$

where P_U is the $\mathsf{Unif}(0,1)$ measure. (The set S_t is P_U-measurable for all t by the assumed measurability of B_t, T_{θ_0}, and F_{θ_0}.) The corresponding score-balanced belief function is

$$\begin{aligned}
\mathsf{bel}_x(\{\theta_0\}^c; \mathcal{S}_B) &= \mathsf{P}_{\mathcal{S}_B}\{\mathcal{S}_B \not\supseteq F_{\theta_0}(x)\} \\
&= \mathsf{P}_{X|\theta_0}\{B_{T_{\theta_0}(X)} \not\supseteq T_{\theta_0}(x)\} \\
&= \mathsf{P}_{X|\theta_0}\{T_{\theta_0}(X) \in B_{T_{\theta_0}(x)}\},
\end{aligned}$$

where the last equality follows from the assumed nesting of $\{B_t\}$. Proposition 3 in Appendix A shows that predictive random sets which are good in the sense that (4.2) holds (at least locally) must be score-balanced.

But there are many such $\mathcal{S}_B$ to choose from, so we now consider finding a "best" one. A reasonable definition of optimal score-balanced predictive random set is one that makes the difference between the right- and left-hand sides of (4.2) as large as possible for each θ in a neighborhood of θ_0. Then, for two-sided assertions, we have

16

Definition 4. Let $B^\star = \{B_t^\star : t \in \mathbb{T}\}$ be such that, for each t,

$$\int_{T_{\theta_0}(x) \in B_t^\star} V_{\theta_0}(x) f_{\theta_0}(x) \, dx \qquad (4.4)$$

is minimized subject to the score-balance constraint (4.3). Then $\mathcal{S}^\star = \mathcal{S}_{B^\star}$ is the optimal score-balanced predictive random set.

Here we give a general construction of an an optimal score-balanced predictive random sets. Proving that the predictive random sets satisfy the conditions of Definition 4 will require assumptions about the model. Start with the following class of intervals:

$$B_t^\star = \big(\xi_-(t), \xi_+(t)\big), \quad t \in T_{\theta_0}(\mathbb{X}), \qquad (4.5)$$

where the functions ξ_-, ξ_+ (which depend implicitly on θ_0) are such that (4.3) holds. In addition, we shall assume these functions are continuous and satisfy

- $\xi_-(t)$ is non-positive, $\xi_-(t) = t$ for $t \in (-\infty, 0)$ and is decreasing for $t \in [0, \infty)$;
- $\xi_+(t)$ is non-negative, $\xi_+(t) = t$ for $t \in [0, \infty)$ and is increasing for $t \in (-\infty, 0)$.

The functions ξ_-, ξ_+ describe a sort of symmetry/balance in the distribution of $T_{\theta_0}(X)$: they satisfy $\xi_+(\xi_-(t)) = t$ and $\xi_-(\xi_+(-t)) = -t$ for all $t \geq 0$. In some cases, for given t, expressions for $\xi_-(t)$ and $\xi_+(t)$ can be found analytically, but typically numerical solutions are required. Set $\mathcal{S}^\star = \mathcal{S}_{B^\star}$. We claim that, under certain conditions on $V_{\theta_0}(x)$, $\mathcal{S}^\star$ is optimal in the sense of Definition 4.

Before we get to the optimality considerations, we first verify the assumption that $\mathsf{bel}_X(\{\theta_0\}^c; \mathcal{S}^\star) \sim \mathsf{Unif}(0,1)$ under $\mathsf{P}_{X|\theta_0}$. From the definition of $B_t^\star$, it is clear that

$$
\begin{aligned}
T \in B_t^\star &\iff \xi_-(t) < T < \xi_+(t) \\
&\iff \xi_-(t) < \xi_-(T) < \xi_+(T) < \xi_+(t) \\
&\iff \xi_+(T) - \xi_-(T) < \xi_+(t) - \xi_-(t).
\end{aligned}
$$

Consequently, if $D_{\theta_0}(X) = \xi_+(T_{\theta_0}(X)) - \xi_-(T_{\theta_0}(X))$, then

$$\mathsf{bel}_x(\{\theta_0\}^c; \mathcal{S}^\star) = \mathsf{P}_{X|\theta_0}\{T_{\theta_0}(X) \in B_{T_{\theta_0}}^\star(x)\} = \mathsf{P}_{X|\theta_0}\{D_{\theta_0}(X) < D_{\theta_0}(x)\}.$$

Therefore, since $D_{\theta_0}(X)$ is a continuous random variable, an argument like that in Corollary 1 shows that $\mathsf{bel}_X(\{\theta_0\}^c; \mathcal{S}^\star) \sim \mathsf{Unif}(0,1)$ under $\mathsf{P}_{X|\theta_0}$.

We are now ready for optimality of $\mathcal{S}^\star$. Write $V(t)$ for $V_{\theta_0}(x)$, when treated as a function of $t = T_{\theta_0}(x)$. The condition to be imposed is:

$$V(t) \text{ is uniquely minimized at } t = 0, \text{ and } V(0) < 0. \qquad (4.6)$$

This condition holds, e.g., for all exponential families with θ the natural parameter.

Proposition 2. *Under condition (4.6), the score balanced predictive random set $\mathcal{S}^\star = \mathcal{S}_{B^\star}$, with $B^\star$ described above, is optimal in the sense of Definition 4.*

Proof. The proof is simple but tedious so here we just sketch the main idea. Under (4.6), the intervals $B_t^\star$ which are "balanced" around $T_{\theta_0}(x) = 0$, make most efficient use of the space where $V_{\theta_0}(x)$ is smallest in the following sense. They are exactly the right size to make $\mathcal{S}_{B^\star}$ efficient, so any other efficient score-balanced predictive random set $\mathcal{S}_B$ must be determined by sets $B = \{B_t\}$ other than intervals concentrated around $T_{\theta_0}(x) = 0$. Since such intervals are where $V_{\theta_0}(x)$ is smallest, the integral in (4.4) corresponding to B_t must be larger than that corresponding to $B_t^\star$. Therefore, $\mathcal{S}^\star$ satisfies the conditions of Definition 4 and, hence, is optimal. □

Unfortunately, (4.6) is not always satisfied. For example, it can fail for exponential families not in natural form. But we claim that (4.6) is not absolutely essential. Assume $V(t)$ is convex and $V(0) < 0$. This relaxed assumption holds, e.g., for all exponential families. To keep things simple, suppose that $V(t)$ is minimized at $\hat{t} > 0$. Although the argument to be given is general, Figure 2(a) illustrates the phenomenon for the exponential distribution with mean $\theta_0 = 1$. The heavy line there represents $V(t)$, and the thin lines represent $th(t)$ (black) and $V(t)h(t)$ (gray), where $h(t)$ is the density of T. The horizontal lines represent the intervals $B_t^\star$ in (4.5) for select t. By convexity of $V(t)$, there exists t_0 such that $\hat{t} \in (0, t_0)$ and $V(t) < V(0)$ for each $t \in (0, t_0)$; this is $(0, 0.5)$ in the figure. For $t \in (0, t_0)$, the intervals $B_t^\star$ do not contain $(0, t_0)$; these intervals are shown in black. In such cases, the integral (4.4) can be reduced by breaking $B_t^\star$ into two parts: one part takes more of $(0, t_0)$, where $V(t)$ is smallest, and the other part is chosen to satisfy the score-balance condition (4.3). But when $t \geq t_0$, no improvement can be made by changing $B_t^\star$; these cases are shown in gray. So, in this sense, the intervals $B_t^\star$ in (4.5) are not too bad even if (4.6) fails.

On the other hand, violations of (4.6) are due to the choice of the parametrization. Indeed, under mild assumptions, there exists a transformation $\eta = \eta(\theta)$ such that the corresponding $V(t)$ function for η satisfies (4.6). Then the predictive random set $\mathcal{S}^\star$ in Proposition 2 is the optimal for this transformed problem.

Gaussian Example (cont). This is a natural exponential family distribution, so Proposition 2 holds, and $\mathcal{S}^\star$ is the optimal score-balanced predictive random set. Here the score function is $T_\theta(x) = x - \theta$. Under $X \sim \mathsf{N}(\theta, 1)$, the distribution of $T_\theta(X)$ is symmetric about 0. Therefore, $B_t^\star = (-|t|, |t|)$, and the corresponding predictive random set is supported on subsets S_t given by

$$S_t = F_{\theta_0}\big(\{x : |x - \theta_0| \leq |t|\}\big) = \big(\Phi(-|t|), \Phi(|t|)\big),$$

with belief function $\mathsf{bel}_x(\{\theta_0\}^c; \mathcal{S}^\star) = 2\Phi(|x - \theta_0|) - 1$. This is exactly one minus the plausibility function in (2.12) based on the default predictive random set (2.7). Therefore, we conclude that the (2.7) is, in fact, the optimal score-balanced predictive random set in the Gaussian problem. This is consistent with our intuition, given that the results based on this default choice in the Gaussian example match up with good classical results.

Exponential Example. Suppose X is an exponential random variable with mean θ, as discussed above. Unlike the Gaussian, this distribution is asymmetric, so, for the optimal score-balanced IM, a numerical method is needed to identify the set $B_{T_{\theta_0}(x)}$ for each observed x. Plots of the corresponding plausibility functions $\mathsf{pl}_x(\theta; \mathcal{S}) = 1 - \mathsf{bel}_x(\{\theta\}^c; \mathcal{S})$ for two different predictive random sets based on $X = 5$ are shown in Figure 2(b). The

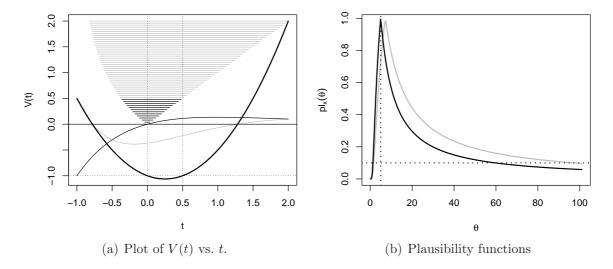(a) Plot of $V(t)$ vs. t.   (b) Plausibility functions

Figure 2: Specifics of Panel (a) are discussed in the text. Panel (b) shows $\mathsf{pl}_x(\theta; \mathcal{S})$, as a function of the exponential scale parameter θ, for two predictive random sets $\mathcal{S}$: optimal score-balanced (black) and default (gray). Vertical line marks the observed $X = 5$.

black line is based on the optimal score-balanced predictive random set, and the gray line is based on the default predictive random set in (2.7). 90% plausibility intervals, determined by the horizontal line at $\alpha = 0.1$, are much shorter for the score-balanced IM compared to the default in this case. For comparison, one might consider a crude nominal 90% confidence interval for θ, namely, $(Xe^{-1.65}, Xe^{1.65})$, based on a variance-stabilizing transformation and normal approximation. These intervals tend to be shorter than both plausibility intervals, but their coverage probability (≈ 0.82) is too small.

Poisson Example (cont). Although the theory above holds only for continuous models, the score-balanced predictive random set performs well in discrete problems too. For the sake of space, we refer the reader to Martin et al. (2012) for the details.

5   Two more examples

5.1   A standardized mean problem

Suppose that $X_1, \ldots, X_n$ are independent $\mathsf{N}(\mu, \sigma^2)$ observations. The goal is to make inference on $\psi = \mu/\sigma$, the standardized mean, or signal-to-noise ratio. Following Dempster (1963), we start with a reduction of the full data to the sufficient statistics for $\theta = (\mu, \sigma^2)$, namely $(\overline{X}, S^2)$, the sample mean and variance. Formal IM-based justification for this reduction is available, though we shall not discuss this here.

For the A-step, we take the association to be

$$\overline{X} = \mu + n^{-1/2}\sigma U_1 \quad \text{and} \quad S = \sigma U_2, \tag{5.1}$$

where $U = (U_1, U_2) \sim \mathsf{P}_U = \mathsf{N}(0, 1) \times \{\mathsf{ChiSq}(n-1)/(n-1)\}^{1/2}$. After replacing σ in the left-most identity in (5.1) with S/U_2, a bit of algebra reveals that

$$n^{1/2}\overline{X}/S = (n^{1/2}\psi + U_1)/U_2 \quad \text{and} \quad S = \sigma U_2.$$

19

For $\theta = (\psi, \sigma)$, make a change of auxiliary variable $v = \varphi_\theta(u)$, given by

$$v_1 = F_\psi\left(\frac{n^{1/2}\psi + u_1}{u_2}\right) \quad \text{and} \quad v_2 = \frac{\exp(u_2)}{1 + \exp(u_2)},$$

where F_ψ is the distribution function for $\mathsf{t}_{n-1}(n^{1/2}\psi)$, a non-central Student-t distribution with $n - 1$ degrees of freedom and non-centrality parameter $n^{1/2}\psi$. Note that the full generality of the parameter-dependent change-of-variables in Corollary 2 is needed here. Then the transformed association is

$$n^{1/2}\overline{X}/S = F_\psi^{-1}(V_1) \quad \text{and} \quad S = \sigma\log\{V_2/(1 - V_2)\},$$

and the measure P_V on the space of $V = (V_1, V_2)$ has a $\mathsf{Unif}(0, 1)$ marginal on the V_1-space; the distribution on V_1-slices of the V_2 space can be worked out, but it is not needed in what follows. For the P-step, we predict $v^\star = \varphi_\theta(u^\star)$ with a rectangle predictive random set $\mathcal{S}$ defined by the following set-valued mapping, similar to (2.7):

$$v = (v_1, v_2) \mapsto \{v_1' : |v_1' - 0.5| < |v_1 - 0.5|\} \times [0, 1]. \tag{5.2}$$

Optimality considerations along the lines in Section 4.3.2 could be pursued here, but we choose to keep things simple since analysis of the non-central Student-t distribution is non-trivial. An important direction of future research is to develop numerical methods for evaluating optimal IMs. Using a predictive random set that spans the entire v_2-space for each v has the effect of "integrating out" the nuisance parameter σ. For the predictive random set $\mathcal{S}$ in (5.2), if $z = n^{1/2}\overline{x}/s$, then the C-step gives the following set $\Theta_x(\mathcal{S}) = \Psi_x(\mathcal{S}) \times \Sigma_x(\mathcal{S})$ of candidate (ψ, σ) pairs:

$$\{\psi : |F_\psi(z) - 0.5| < |V_1 - 0.5|\} \times \{\sigma : \sigma > 0\}, \quad V \sim \mathsf{P}_V. \tag{5.3}$$

For assertions $A = \{(\psi, \sigma) : \sigma > 0\}$ the plausibility function is given by

$$\mathsf{pl}_x(A) = \mathsf{P}_\mathcal{S}\{\Theta_x(\mathcal{S}) \not\subseteq A^c\} = \mathsf{P}_\mathcal{S}\{\Psi_x(\mathcal{S}) \ni \psi\} = 1 - |2F_\psi(z) - 1|.$$

In this case, the $100(1 - \alpha)\%$ plausibility interval $\Pi_x(\alpha)$ for ψ is obtained by inverting the inequality $1 - |2F_\psi(z) - 1| > \alpha$, i.e., $\Pi_x(\alpha) = \{\psi : \alpha/2 < F_\psi(z) < 1 - \alpha/2\}$.

This is exactly the usual frequentist confidence interval based on the sampling distribution of the standardized sample mean; it also agrees with the fiducial intervals obtained by Dempster (1963) and Dawid and Stone (1982). The standard frequentist approach relies on an informal "plug-in style" marginalization, whereas the IM approach above shows exactly how σ is ignored via cylinder assertions. More sophisticated IM marginalization techniques are available, but we do not discuss these here.

5.2 A many-exponential-rates problem

For our last example, we consider a high-dimensional problem. Suppose that $X = (X_1, \ldots, X_n)$ consists of independent observations $X_i \sim \mathsf{Exp}(\theta_i)$, $i = 1, \ldots, n$, with unknown rates $\theta_1, \ldots, \theta_n$. The goal is to give a probabilistic measure of the support in $X = x$ for the assertion $A = \{\theta_1 = \cdots = \theta_n\}$ that the rates are equal. A version of this

problem was also discussed in Martin et al. (2010), but here we simplify the presentation, emphasize the three-step IM construction, and produce much better results.

Start, in the A-step, with the association $X_i = U_i/\theta_i$, $i = 1, \ldots, n$, where P_U is the product measure $\mathsf{Exp}(1)^{\times n}$. Make a change of auxiliary variables $v = \varphi(u)$:

$$v_0 = \sum_{i=1}^{n} u_i \quad \text{and} \quad v_i = u_i/v_0, \quad i = 1, \ldots, n.$$

The new vector $v = (v_0, v_1, \ldots, v_n)$ takes values in $\mathbb{V} = (0, \infty) \times \mathbb{P}_{n-1}$, where $\mathbb{P}_{n-1}$ is the $(n-1)$-dimensional probability simplex in $\mathbb{R}^n$, and $\mathsf{P}_V = \mathsf{P}_U \varphi^{-1}$ is the product measure $\mathsf{Gamma}(n, 1) \times \mathsf{Dir}_n(1_n)$. Then the modified association is

$$X_i = V_0 V_i/\theta_i, \quad i = 1, \ldots, n, \quad \text{where} \quad V = (V_0, V_1, \ldots, V_n) \sim \mathsf{P}_V. \qquad (5.4)$$

For the P-step, we shall consider the following predictive random set $\mathcal{S}$ characterized by $V \sim \mathsf{P}_V$ and the set-valued mapping $v \mapsto \{v' : h(v') < h(v)\}$. In this case, we take

$$h(v) = -\sum_{i=1}^{n-1} \big[a_i \log t_i(v) + b_i \log\{1 - t_i(v)\} \big],$$

with $t_i(v) = \sum_{j=1}^{i} v_i$, $a_i = 1/(n - i - 0.3)$, and $b_i = 1/(i - 0.3)$. A few remarks on this choice of $\mathcal{S}$ are in order. First, it follows from Corollary 1 that $\mathcal{S}$ is efficient. Second, the random vector $(t_1(V), \ldots, t_{n-1}(V))$, for $V \sim \mathsf{P}_V$, has the distribution of a vector of $n - 1$ sorted $\mathsf{Unif}(0, 1)$ random variables, and Zhang (2010, Sec. 3.4.2) shows that $\mathcal{S}$ provides an easy-to-compute alternative to the well-performing hierarchical predictive random set for predicting sorted uniforms used in Martin et al. (2010). Finally, that the first component v_0 of v is essentially ignored in $\mathcal{S}$ is partly for convenience, and partly because v_0 is related to the overall scale of the problem which is irrelevant to the assertion A of interest.

For the C-step, combining the observed data, the association model (5.4), and the predictive random set $\mathcal{S}$ above, we get the following random set for θ:

$$\Theta_x(\mathcal{S}) = \{\theta : h(v(x, \theta)) < h(V)\}, \quad V \sim \mathsf{P}_V,$$

where $v(x, \theta) = (\theta_1 x_1, \ldots, \theta_n x_n)/\sum_{j=1}^{n} \theta_j x_j$. Since the assertion $A = \{\theta_1 = \cdots = \theta_n\}$ is a one-dimensional subset of Θ, the belief function is zero. It is also important to note that when θ is a constant vector, $v(x, \theta)$ is independent of that constant, i.e., $v(x, \theta) = v(x, 1_n)$, which greatly simplifies computation of the plausibility function at A. Indeed,

$$\mathsf{pl}_x(A) = \mathsf{P}_V\{h(V) > h(v(x, 1))\},$$

which can easily be evaluated using Monte Carlo. As described in Section 3.4, the level α IM-based tests rejects the assertion A if and only if $\mathsf{pl}_x(A) \leq \alpha$.

For illustration, we compare our results with those of Martin et al. (2010). They consider the basic likelihood ratio test, which is based on the test statistic $\big\{ \big(\prod_{i=1}^{n} x_i\big)^{1/n}/\overline{x} \big\}^n$. They also consider a different sort of IM solution, based on thresholding the plausibility function, but with a default type of predictive random set that uses a Kullback–Leibler neighborhood for predicting the component $(V_1, \ldots, V_n)$ of V. We compare the power of these three tests in several different cases. In each setup, $n = n_1 + n_2 = 100$ observations are available, but the first n_1 exponential rates equal 1 while the last n_2 equal θ. Figure 3

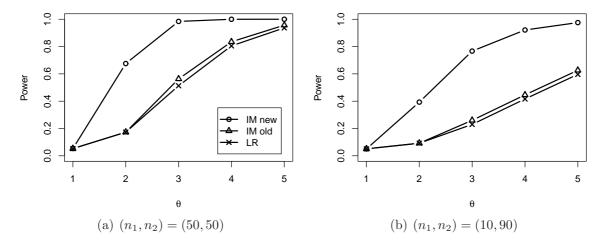(a) $(n_1, n_2) = (50, 50)$ (b) $(n_1, n_2) = (10, 90)$

Figure 3: Estimated powers of the likelihood ratio and two IM-based tests for the simulation described in Section 5.2. Here θ is the ratio of the rate of the last n_2 observations to that of the first n_1.

shows the power functions over a range of θ values for two configurations of (n_1, n_2). Here we see that, in both cases, the likelihood ratio and old IM tests have similar power, possibly because of the common connection to the Kullback–Leibler divergence. On the other hand, the new IM-based test presented above has strikingly larger power than the other two. This substantial improvement in power is likely due to the close relationship between our choice of $\mathcal{S}$ and the assertion of interest. So while the comparison between the new IM results and those of the other "default" methods is not entirely fair, it is interesting to see that an assertion-specific choice of predictive random set can lead to drastically improved performance.

6   Discussion

The conversion of experience to knowledge is fundamental to the advancement of science, and statistical inference plays a crucial role. For ages, there has been disagreement about which statistical paradigm to choose. Both the frequentist and Bayesian paradigms have their own set of advantages and disadvantages, so it would be worthwhile to identify something new which combines the respective advantages but loses, or at least weakens, the disadvantages. Here we have described a three-step procedure to construct IMs for prior-free, post-data probabilistic inference, and proved that IMs yield frequency-calibrated probabilities under very general conditions. The point is that the values of the corresponding belief/plausibility function are meaningful both within and across experiments, accomplishing both the frequentist and Bayesian goals simultaneously.

The proposed IM approach is surely new, but since new is not always better, it is natural to ask what is the benefit of using IMs. Our response is that, although it will take time for users to familiarize themselves with the thought process, the IM framework is logical, intuitive, and able to produce meaningful and frequency-calibrated probabilistic

measures of uncertainty about θ without a prior distribution. The latter property is something that no other inferential framework is able to achieve.

Admittedly, the final IM depends on the user's choice of association and predictive random set, but we do not believe that this is particularly damning. Section 4 laid the foundation for a theory of optimal predictive random sets, and further efforts to develop "default" predictive random sets are ongoing, particularly for multi-parameter problems. But a case can be made to prefer the ambiguity of the choice of predictive random set over that of a frequentist's choice of statistic or Bayesian's choice of prior. The point is that neither a frequentist sampling distribution nor a Bayesian prior distribution adequately describes the source of uncertainty about θ. As we argued above, this uncertainty is fully characterized by the fact that, whatever the association, the value of $u^\star$ is missing. Therefore, it seems only natural to prefer the IM framework that features a direct attack on the source of uncertainty over another that attacks the problem indirectly. Moreover, as was demonstrated in Section 5.2, choosing the predictive random set that depends on the problem and/or assertion of interest can lead to drastically improved results.

We note that differences between IM outputs from different predictive random sets are slight for assertions involving one-dimensional quantities. However, for high-dimensional auxiliary variables, the choice of predictive random set deserves special attention. In such cases, our approach is to construct predictive random sets for functions of auxiliary variables that are most relevant to the assertions of interest. This leads to a practically useful auxiliary variable dimension reduction. It is interesting that this approach has some close connections to Fisher's theory of sufficient statistics (Martin and Liu 2012). For nuisance parameter problems, like those in Section 5, there is a different form of dimension reduction required (Martin and Liu 2013).

Of course, compared to the well-developed Bayesian and frequentist methods, IMs have many open problems. Both theoretical work and applications have shown that the IM framework is promising. Given the attractive properties of IMs developed here and in the references above, we expect to see more exciting advancements in IMs or new inferential frameworks (e.g., Martin 2012) that are probabilistic and have desirable frequency properties.

Acknowledgments

A   Details from Section 4.3.2

If we assume that $\mathsf{bel}_X(\{\theta_0\}^c; \mathcal{S}) \sim \mathsf{Unif}(0,1)$ under $\mathsf{P}_{X|\theta_0}$, then there exists a collection of measurable subsets $\mathbb{X}(\alpha) \subseteq \mathbb{X}$, depending implicitly on θ_0 and $\mathcal{S}$, such that, for each α, $\mathsf{P}_{X|\theta_0}\{\mathbb{X}(\alpha)\} = \alpha$, and $\mathsf{bel}_x(\{\theta_0\}^c; \mathcal{S}) \leq \alpha$ iff $x \in \mathbb{X}(\alpha)$. It follows that, for any θ,

$$\mathsf{P}_{X|\theta}\{\mathsf{bel}_X(\{\theta_0\}^c; \mathcal{S}) \leq \alpha\} = \psi_\alpha(\theta) := \int_{\mathbb{X}(\alpha)} f_\theta(x)\, dx.$$

By definition, $\psi_\alpha(\theta_0) = \alpha$. Now, (4.2) is equivalent to $\psi_\alpha(\theta) < \psi_\alpha(\theta_0)$ for all α, or, to put it another way, $\psi_\alpha(\theta)$ is maximized at $\theta = \theta_0$ for all α. Under the stated regularity conditions, this maximization is equivalent to the claim that, for all $\alpha \in (0, 1)$, the first and second derivatives of $\psi_\alpha(\theta)$ at $\theta = \theta_0$ satisfy

$$\psi_\alpha'(\theta_0) = \int_{\mathbb{X}(\alpha)} T_{\theta_0}(x) f_{\theta_0}(x) \, dx = 0, \tag{A.1}$$

$$\psi_\alpha''(\theta_0) = \int_{\mathbb{X}(\alpha)} V_{\theta_0}(x) f_{\theta_0}(x) \, dx < 0. \tag{A.2}$$

Since $T_{\theta_0}(X)$ has mean zero under $\mathsf{P}_{X|\theta_0}$, we can see that (A.1) requires $\mathbb{X}(\alpha)$ to be somehow symmetric, or balanced, with respect to the distribution of $T_{\theta_0}(X)$. We, therefore, refer to (A.1) as the *score-balance* condition. This condition, expressed in terms of $\mathbb{X}(\alpha)$ in (A.1), can be traced back to a corresponding condition on the predictive random set.

Let us now assume that $\mathcal{S}_B$ is such that $\mathsf{bel}_X(\{\theta_0\}^c; \mathcal{S}_B) \sim \mathsf{Unif}(0, 1)$ under $\mathsf{P}_{X|\theta_0}$; in the main text we construct a particular score-balanced predictive random set and show that that this assumption holds. Then, as we argued above, for any $\alpha \in (0, 1)$, there exists $t(\alpha) \in \mathbb{T}$ such that $\mathsf{bel}_x(\{\theta_0\}^c; \mathcal{S}_B) \le \alpha$ iff $T_{\theta_0}(x) \in B_{t(\alpha)}$. In this case, for any θ,

$$\mathsf{P}_{X|\theta}\{\mathsf{bel}_X(\{\theta_0\}^c; \mathcal{S}_B) \le \alpha\} = \int_{T_{\theta_0}(x) \in B_{t(\alpha)}} f_\theta(x) \, dx,$$

and the right-hand side is $\psi_\alpha(\theta)$ as defined previously. From the definition of B, differentiating under the integral sign reveals that (A.1) holds. We can now prove

Proposition 3. *Focus on predictive random sets $\mathcal{S}$ such that $\mathsf{bel}_X(\{\theta_0\}^c; \mathcal{S}) \sim \mathsf{Unif}(0, 1)$ under $\mathsf{P}_{X|\theta_0}$. Then condition (4.2) holds for all θ in a neighborhood of θ_0 iff the predictive random set $\mathcal{S} = \mathcal{S}_B$ is score-balanced and*

$$\int_{T_{\theta_0}(x) \in B_t} V_{\theta_0}(x) f_{\theta_0}(x) \, dx < 0, \quad \forall \, t \in \mathbb{T}. \tag{A.3}$$

Proof. Take θ close enough to θ_0 such that the remainder terms in a second-order Taylor approximation of $\psi_\alpha(\theta)$ about $\theta = \theta_0$ can be ignored. That is, for any α,

$$\psi_\alpha(\theta) - \psi_\alpha(\theta_0) = \int_{T_{\theta_0}(x) \in B_{t(\alpha)}} T_{\theta_0}(x) f_{\theta_0}(x) \, dx \cdot (\theta - \theta_0)$$
$$+ \frac{1}{2} \int_{T_{\theta_0}(x) \in B_{t(\alpha)}} V_{\theta_0}(x) f_{\theta_0}(x) \, dx \cdot (\theta - \theta_0)^2.$$

The first terms vanishes and the second term is negative by (A.3). Therefore $\psi_\alpha(\theta) < \psi_\alpha(\theta_0)$ for all α and, hence, (4.2) holds for all θ in a neighborhood of θ_0. $\square$

References

Berger, J. (2006), "The case for objective Bayesian analysis," *Bayesian Anal.*, 1, 385–402.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009), "The formal definition of reference priors," *Ann. Statist.*, 37, 905–938.

Bernardo, J.-M. (1979), "Reference posterior distributions for Bayesian inference," *J. Roy. Statist. Soc. Ser. B*, 41, 113–147.

Brown, L. D., Cai, T. T., and DasGupta, A. (2003), "Interval estimation in exponential families," *Statist. Sinica*, 13, 19–49.

Cai, T. T. (2005), "One-sided confidence intervals in discrete distributions," *J. Statist. Plann. Inference*, 131, 63–88.

Dawid, A. P. (1985), "Calibration-based empirical probability," *Ann. Statist.*, 13, 1251–1285, with discussion.

Dawid, A. P. and Stone, M. (1982), "The functional-model basis of fiducial inference," *Ann. Statist.*, 10, 1054–1074, with discussion.

Dempster, A. P. (1963), "Further examples of inconsistencies in the fiducial argument," *Ann. Math. Statist.*, 34, 884–891.

— (2008), "Dempster–Shafer calculus for statisticians," *Internat. J. of Approx. Reason.*, 48, 265–277.

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

Fraser, D. A. S. (1968), *The Structure of Inference*, New York: John Wiley & Sons Inc.

— (2011), "Is Bayes posterior just quick and dirty confidence?" *Statist. Sci.*, 26, 299–316.

Fraser, D. A. S., Reid, N., Marras, E., and Yi, G. Y. (2010), "Default priors for Bayesian and frequentist inference," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72, 631–654.

Ghosh, M. (2011), "Objective priors: an introduction for frequentists," *Statist. Sci.*, 26, 187–202.

Hannig, J. (2009), "On generalized fiducial inference," *Statist. Sinica*, 19, 491–544.

— (2012), "Generalized fiducial inference via discretization," *Statist. Sinica*, to appear.

Hannig, J. and Lee, T. C. M. (2009), "Generalized fiducial inference for wavelet regression," *Biometrika*, 96, 847–860.

Heath, D. and Sudderth, W. (1978), "On finitely additive priors, coherence, and extended admissibility," *Ann. Statist.*, 6, 333–345.

Lehmann, E. L. and Romano, J. P. (2005), *Testing statistical hypotheses*, Springer Texts in Statistics, New York: Springer, 3rd ed.

Little, R. (2011), "Calibrated Bayes, for statistics in general, and missing data in particular," *Statist. Sci.*, 26, 162–174.

Martin, R. (2012), "Plausibility functions and exact frequentist inference," Unpublished manuscript, arXiv:1203.6665.

Martin, R., Ermini Leaf, D., and Liu, C. (2012), "Optimal inferential models for a Poisson mean," Unpublished manuscript, arXiv:1207.0105.

Martin, R. and Liu, C. (2012), "Conditional inferential models: combining information for prior-free probabilistic inference," Unpublished manuscript, arXiv:1211.1530.

— (2013), "Marginal inferential models: optimal prior-free probabilistic inference on interest parameters," Unpublished manuscript.

Martin, R., Zhang, J., and Liu, C. (2010), "Dempster–Shafer theory and statistical inference with weak beliefs," *Statist. Sci.*, 25, 72–87.

Rubin, D. B. (1984), "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *Ann. Statist.*, 12, 1151–1172.

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

Walley, P. (1996), "Inferences from multinomial data: learning about a bag of marbles," *J. Roy. Statist. Soc. Ser. B*, 58, 3–57, with discussion and a reply by the author.

Wasserman, L. A. (1990), "Prior envelopes based on belief functions," *Ann. Statist.*, 18, 454–464.

Xie, M. and Singh, K. (2012), "Confidence distribution, the frequentist distribution of a parameter – a review," *Int. Statist. Rev.*, to appear.

Xie, M., Singh, K., and Strawderman, W. E. (2011), "Confidence distributions and a unifying framework for meta-analysis," *J. Amer. Statist. Assoc.*, 106, 320–333.

Zabell, S. L. (1992), "R. A. Fisher and the fiducial argument," *Statist. Sci.*, 7, 369–387.

Zhang, J. (2010), "Statistical inference with weak beliefs," Ph.D. thesis, Purdue University, West Lafayette, IN.

Zhang, J. and Liu, C. (2011), "Dempster–Shafer inference with weak beliefs," *Statist. Sinica*, 21, 475–494.

B   Corrections—added post-publication

B.1   Correction of Theorem 1

In the main text, for validity of the predictive random set $\mathcal{S}$, the support $\mathbb{S}$ was assumed only to be nested, i.e., for any $S, S' \in \mathbb{S}$, either $S \subseteq S'$ or $S' \subseteq S$. However, some additional technical conditions are required for the proof to go through.

Fix a topology on the auxiliary variable space $\mathbb{U}$, and let the σ-algebra defined there contain all the open sets. In addition to being nested, we shall assume that $\mathbb{S}$ contains both $\varnothing$ and $\mathbb{U}$, and that all of its contents are closed subsets of $\mathbb{U}$. These additional requirements result in no real loss of generality. Indeed, those predictive random sets in Corollary 1 of the main text already satisfy these. These extra conditions also make the statement and proof of the theorem more transparent.

Theorem 1′. *Let $\mathbb{S}$ be a nested collection of closed P_U-measurable subsets of $\mathbb{U}$ that contains $\varnothing$ and $\mathbb{U}$. Define a predictive random set $\mathcal{S}$, with distribution $\mathsf{P}_{\mathcal{S}}$, supported on $\mathbb{S}$, such that*

$$\mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq K\} = \sup_{S \in \mathbb{S}: S \subseteq K} \mathsf{P}_U(S), \quad K \subseteq \mathbb{U}.$$

Then $\mathcal{S}$ is valid in the sense of Definition 1 in the main text.

Proof. Set $Q(u) = \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \not\supseteq u\}$. For any $\alpha \in (0,1)$, let S_α be the smallest $S \in \mathbb{S}$ such that $\mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq S\} \equiv \mathsf{P}_U(S) \geq 1 - \alpha$. In particular, $S_\alpha = \bigcap\{S \in \mathbb{S} : \mathsf{P}_U(S) \geq 1 - \alpha\}$. Since each S is closed, so is S_α; it is also measurable by our assumptions about the richness of the σ-algebra on $\mathbb{U}$. The key observation is that $Q(u) > 1 - \alpha$ iff $u \in S_\alpha^c$. Therefore, by continuity of P_U from above, we get

$$\mathsf{P}_U\{Q(U) > 1 - \alpha\} = \mathsf{P}_U(S_\alpha^c) = 1 - \mathsf{P}_U(S_\alpha) = 1 - \lim \mathsf{P}_U(S),$$

where the limit is over all S decreasing to S_α. By construction, each such S satisfies $\mathsf{P}_U(S) \geq 1 - \alpha$. So, finally, we get $\mathsf{P}_U\{Q(U) > 1 - \alpha\} \leq \alpha$ and, since α is arbitrary, the claimed validity is proved. $\qquad\square$

B.2   Correction/extension of Theorem 3

Theorem 3 in the main text says that nested predictive random sets are more efficient than those which are not nested. However, the nested predictive random set constructed in that theorem is not necessarily valid. Since validity is a key to the IM analysis, it would be desirable if the new nested predictive random set $\mathcal{S}'$ was also valid. We accomplish this in Theorem 3′ below. First, we need the following lemma.

Lemma 0. *On a space $\mathbb{U}$ equipped with probability P_U, let $\mathcal{S}$ be a valid predictive random set for $U \sim \mathsf{P}_U$. Choose a collection of P_U-measurable subsets $\{\mathbb{U}_x : x \in \mathbb{X}\}$ of $\mathbb{U}$, and set $\eta(x) = \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq \mathbb{U}_x\}$. Then*

$$\inf_{x \in \mathbb{X}_0} \eta(x) \leq \mathsf{P}_U\Big\{\bigcap_{x \in \mathbb{X}_0} \mathbb{U}_x\Big\}$$

for any subset $\mathbb{X}_0$ of $\mathbb{X}$ such that $\bigcap_{x \in \mathbb{X}_0} \mathbb{U}_x$ is P_U-measurable.

Proof. First, note that if $u \in \mathbb{U}_x^c$, then $Q(u) \equiv \mathsf{P}_{\mathcal{S}}\{\mathcal{S} \not\ni u\} \geq \eta(x)$. Therefore, if $u \in \bigcup_{x \in \mathbb{X}_0} \mathbb{U}_x^c$, then $Q(u) \geq \inf_{x \in \mathbb{X}_0} \eta(x)$. This argument implies

$$\mathsf{P}_U\Big\{Q(U) \geq \inf_{x \in \mathbb{X}_0} \eta(x)\Big\} \geq \mathsf{P}_U\Big\{\bigcup_{x \in \mathbb{X}_0} \mathbb{U}_x^c\Big\} = 1 - \mathsf{P}_U\Big\{\bigcap_{x \in \mathbb{X}_0} \mathbb{U}_x\Big\}.$$

Since $\mathcal{S}$ is valid, we have

$$\mathsf{P}_U\Big\{Q(U) \geq \inf_{x \in \mathbb{X}_0} \eta(x)\Big\} \leq 1 - \inf_{x \in \mathbb{X}_0} \eta(x);$$

Combining this with the inequality in the previous display, we get

$$1 - \inf_{x \in \mathbb{X}_0} \eta(x) \geq 1 - \mathsf{P}_U\Big\{\bigcap_{x \in \mathbb{X}_0} \mathbb{U}_x\Big\},$$

which implies $\inf_{x \in \mathbb{X}_0} \eta(x) \leq \mathsf{P}_U\{\bigcap_{x \in \mathbb{X}_0} \mathbb{U}_x\}$. $\qquad\square$

A measurability question was overlooked in the main text. In particular, the sets in (B.1) below are not automatically measurable. To confirm this, we shall add one more modification; note that this is not needed if the sampling model $\mathsf{P}_{X|\theta}$ is discrete. To start, for the given topology on $\mathbb{U}$, keep the same assumptions about the corresponding σ-algebra as above. Now, recall the a-events $\mathbb{U}_x(A) = \{u \in \mathbb{U} : \Theta_x(u) \subseteq A\}$ defined in the proof of Proposition 1 in the main text. Here we shall replace $\mathbb{U}_x(A)$ with its closure. This does not affect any properties of the resulting belief function when P_U is non-atomic. In all the examples we have considered, P_U can be taken as continuous; this is a particularly convenient choice, in light of Corollary 1 in the main text.

Theorem 3′. *Suppose that either $\mathbb{X}$ is a discrete space, or that the assumptions in the previous paragraph hold. Fix $A \subseteq \Theta$ and assume condition (2.10) in the main text. Given any valid predictive random set $\mathcal{S}$, there exists a nested and valid predictive random set $\mathcal{S}'$ such that $\mathsf{bel}_x(A; \mathcal{S}') \geq \mathsf{bel}_x(A; \mathcal{S})$ for each $x \in \mathbb{X}$.*

Proof. For the given A and $\mathcal{S}$, set $b(x) \equiv \mathsf{bel}_x(A; \mathcal{S})$. Define a collection $\mathbb{S}' = \{S_x' : x \in \mathbb{X}\}$ of subsets of $\mathbb{U}$ as follows:

$$S_x' = \bigcap_{y \in \mathbb{X}:b(y) \geq b(x)} \mathbb{U}_y(A), \tag{B.1}$$

where $\mathbb{U}_x(A)$ is the new closed a-event. If necessary, add $\varnothing$ and $\mathbb{U}$ to $\mathbb{S}'$ to satisfy the requirement in Theorem 1′. This collection $\mathbb{S}'$ will serve as the support for the new predictive random set $\mathcal{S}'$. First, we can see that $\mathbb{S}'$ is nested: if $b(y) \geq b(x)$, then $S_y \supseteq S_x$. Second, since the new a-events are closed, each S_x' in (B.1) is closed and, hence, P_U-measurable. Third, define the measure $\mathsf{P}_{\mathcal{S}'}$ for $\mathcal{S}'$ to satisfy

$$\mathsf{P}_{\mathcal{S}'}\{\mathcal{S}' \subseteq K\} = \sup_{x:S_x' \subseteq K} \mathsf{P}_U(S_x').$$

According to Theorem 1′, the new $\mathcal{S}'$ is valid. Moreover, by Lemma 0 and the definition of S_x', we have

$$\mathsf{P}_U(S_x') \geq \inf_{y \in \mathbb{X}:b(y) \geq b(x)} b(y) = b(x) \equiv \mathsf{bel}_x(A; \mathcal{S}). \tag{B.2}$$

28

Finally, we have a comparison of the belief functions corresponding to $\mathcal{S}$ and $\mathcal{S}'$:

$$\mathsf{bel}_x(A; \mathcal{S}') = \mathsf{P}_{\mathcal{S}'}\{\mathcal{S}' \subseteq \mathbb{U}_x(A)\} \geq \mathsf{P}_{\mathcal{S}'}\{\mathcal{S}' \subseteq S'_x\} = \mathsf{P}_U(S'_x) \geq \mathsf{bel}_x(A; \mathcal{S});$$

the first inequality follows from monotonicity of $\mathsf{P}_{\mathcal{S}'}\{\mathcal{S}' \subseteq \cdot\}$ and the fact that $S'_x \subseteq \mathbb{U}_x(A)$ for each x, and the second inequality follows from (B.2). $\qquad\square$