Quantifying predictive uncertainty using belief functions: different approaches and practical construction

Thierry Denœux

Abstract We consider the problem of quantifying prediction uncertainty using the formalism of belief functions. Three requirements for predictive belief functions are reviewed, each one of them inducing a distinct interpretation: compatibility with Bayesian inference, approximation of the true distribution, and frequency calibration. Construction procedures allowing us to build belief functions meeting each of these three requirements are described and illustrated using simple examples.

1 Introduction

Statistical prediction is the task of making statements about a not-yet-observed realization y of a random variable Y, based on past observations x. An important issue in statistical prediction is the quantification of uncertainty. Typically, prediction uncertainty has two components:

1. *Estimation uncertainty*, arising from the partial ignorance of the probability distribution of Y, and
2. *Random uncertainty*, due to the variability of Y.

If the distribution of Y is completely known, there is no estimation uncertainty. If Y is a constant, there is no random uncertainty: this is the case in parameter estimation problems. In all practical problems of interest, both sources of uncertainty coexist, and should be accounted for in the prediction method.

In this paper, we assume the past data X and the future data Y to be independent, and we consider sampling models $X \sim P_X(\cdot;\theta)$ and $Y \sim P_Y(\cdot;\theta)$, where θ is a parameter known only to belong to some set Θ. The sample spaces of X and Y will be

Thierry Denœux
Sorbonne Universités, Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, Compiègne, France, e-mail: thierry.denoeux@utc.fr

denoted by $\mathscr{X}$ and $\mathscr{Y}$, respectively. To keep the exposition simple, we will assume Y to be a real random variable, with $\mathscr{Y} \subseteq \mathbb{R}$.

The statistical prediction problem is treated differently in the Bayesian and frequentist frameworks. Here, we briefly outline the main approaches within each of these two frameworks.

Bayesian approach

In the Bayesian framework, X, Y and θ are considered as random variables. A Bayesian posterior predictive distribution $F_B(y|x)$ can then be computed from the conditional distribution $F(y|x;\theta) = F(y|\theta)$ by integrating out θ,

$$F_B(y|x) = \int F(y|\theta)p(\theta|x)d\theta, \tag{1}$$

where $p(\theta|x)$ is the posterior density of θ. The main limitation of this approach is the necessity to specify a prior distribution $p(\theta)$ on θ. In many cases, prior knowledge on θ is either nonexistent, or too vague to be reliably described by a single probability distribution.

Frequentist approach

In the frequentist framework, the prediction problem can be addressed in several ways. The so-called *plug-in* approach is to replace θ in the model by a point estimate $\widehat{\theta}$ and to estimate the distribution of Y by $P_Y(\cdot;\widehat{\theta})$. This approach amounts to neglecting estimation uncertainty. Consequently, it will typically underestimate the prediction uncertainty, unless the sample size is very large. Another approach is to consider *prediction intervals* $[L_1(X), L_2(X)]$ such that the coverage probability

$$CP(\theta) = P_{X,Y}(L_1(X) \leq Y \leq L_2(X);\theta) \tag{2}$$

has some specified value, perhaps approximately. The coverage probability can take any value only if Y is continuous; consequently, we often make this assumption when using this approach. Confidence intervals do account for estimation and prediction uncertainty, but they do not provide any information about the relative plausibility of values inside or outside that set. To address the issue, we may consider one-sided confidence intervals $(-\infty, L_\alpha(X)]$ indexed by $\alpha \in (0,1)$, such that

$$CP(\theta) = P_{X,Y}(Y \leq L_\alpha(X);\theta) \tag{3}$$

is equal to α, at least approximately. Then, we may treat α-prediction limits $L_\alpha(x)$ as the α-quantiles of some *predictive distribution function* $\widetilde{F}_p(y|x)$ [2, 16]. Such a predictive distribution is not a frequentist probability distribution; rather, it can

be seen as a compact way of describing one or two-sided (perhaps, approximate) prediction intervals on Y at any level.

In all the approaches summarized above, uncertainty about Y is represented either as a set (in the case of prediction intervals), or as a probability distribution (such as a frequentist or Bayesian predictive distribution). In this paper, we consider approaches to the prediction problem where uncertainty about Y is represented by a *belief function*. In Dempster-Shafer theory, belief functions are expressions of *degrees of support* for statements about the unknown quantity under consideration, based on evidence. Any subset $A \subseteq \mathscr{Y}$ can be canonically represented by a belief function, and any probability measure is also a particular belief function: consequently, the Dempster-Shafer formalism is more general and flexible than the set-membership or probabilistic representations. The problem addressed in this paper is to exploit this flexibility to represent the prediction uncertainty on Y based on the evidence of observed data x.

The interpretation of a *predictive belief function* will typically depend on the requirements imposed on the construction procedure. There is, however, no general agreement as to which properties should be imposed. The purpose of this paper is to review some desired properties, and to describe practical construction procedures allowing us to build predictive belief functions that verify these properties. As we shall see, three main properties have been proposed in previous work, resulting in three main types of predictive belief functions.

The rest of this paper is organized as follows. Some general definitions and results related to belief functions are first recalled in Section 2. The requirements are then presented in Section 3, and construction procedures for the three types of predictive belief functions considered in this paper are described in Section 4. Section 5 contains conclusions.

2 Background on belief functions

In this section, we provide a brief reminder of the main concepts and results from the theory of belief functions that will be used in this paper. The definitions of belief and plausibility functions will first be recalled in Section 2.1. The connection with random sets will be explained in Section 2.2, and Dempster's rule will be introduced in Section 2.4.

2.1 Belief and plausibility functions

Let Ω be a set, and $\mathscr{B}$ an algebra of subsets of Ω. A belief function on $(\Omega, \mathscr{B})$ is a mapping $Bel : \mathscr{B} \to [0,1]$ such that $Bel(\emptyset) = 0$, $Bel(\Omega) = 1$, and for any $k \geq 2$ and any collection $B_1, \ldots, B_k$ of elements of $\mathscr{B}$,

$$Bel\left(\bigcup_{i=1}^{k} B_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} B_i\right). \tag{4}$$

Given a belief function Bel, the dual *plausibility function* $Pl : \mathscr{B} \to [0,1]$ is defined by $Pl(B) = 1 - Bel(\overline{B})$, for any $B \in \mathscr{B}$. In the Dempster-Shafer theory of belief functions [22], $Bel(B)$ is interpreted as the degree of support in the proposition $Y \in B$ based on some evidence, while $Pl(B)$ is a degree of consistency between that proposition and the evidence.

If the inequalities in (4) are replaced by equalities, then Bel is a finitely additive probability measure, and $Pl = Bel$. If the evidence tells us that $Y \in A$ for some $A \in \mathscr{B}$, and nothing more, then it can be represented by a function Bel_A that gives full degree of support to any $B \in \mathscr{B}$ such that $B \subseteq A$, and zero degree of support to any other subset. It can easily be verified that Bel_A is a belief function. If $A = \Omega$, the belief function is said to be *vacuous*: it represent complete ignorance on Y.

Given two belief functions Bel_1 and Bel_2, we say that Bel_1 is *less committed* than Bel_2 if $Bel_1 \leq Bel_2$; equivalently, $Pl_1 \geq Pl_2$. The meaning of this notion is that Bel_1 represents a weaker state of knowledge than that represented by Bel_2.

2.2 Connection with random sets

A belief function is typically induced by a *source*, defined as a four-tuple $(\mathscr{S}, \mathscr{A}, P, \Gamma)$, where $\mathscr{S}$ is a set, $\mathscr{A}$ an algebra of subsets of $\mathscr{S}$, P a finitely additive probability measure on $(\mathscr{S}, \mathscr{A})$, and Γ a mapping from $\mathscr{S}$ to 2^{Ω}. The mapping Γ is strongly measurable with respect to $\mathscr{A}$ and $\mathscr{B}$ if, for any $B \in \mathscr{B}$, we have

$$\{s \in \mathscr{S} | \Gamma(s) \neq \emptyset, \Gamma(s) \subseteq A\} \in \mathscr{A}.$$

We can then show [19], that the function Bel defined by

$$Bel(B) = \frac{P(\{s \in \mathscr{S} | \Gamma(s) \neq \emptyset, \Gamma(s) \subseteq B\})}{P(\{s \in \mathscr{S} | \Gamma(s) \neq \emptyset\})}, \tag{5}$$

for all $A \subseteq \mathscr{B}$ is a belief function. The dual plausibility function is

$$Pl(B) = \frac{P(\{s \in \mathscr{S} | \Gamma(s) \cap B \neq \emptyset\})}{P(\{s \in \mathscr{S} | \Gamma(s) \neq \emptyset\})}. \tag{6}$$

The mapping Γ is called a *random set*. We should not, however, get abused by the term "random": most of the time, the probability measure P defined on $(\mathscr{S}, \mathscr{A})$ is subjective, and there is no notion of randomness involved.

2.3 Consonant random closed sets

Let us assume that $\Omega = \mathbb{R}^d$ and $\mathscr{B} = 2^\Omega$. Let π be an upper semi-continuous map from $\mathbb{R}^d$ to $[0,1]$, i.e., for any $s \in [0,1]$, the set ${}^s\pi \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d | \pi(x) \geq s\}$ is closed. Furthermore, assume that $\pi(x) = 1$ for some x. Let $S = [0,1]$, $\mathscr{A}$ be the Borel σ-field on $[0,1]$, μ the uniform measure, and Γ the mapping defined by $\Gamma(s) = {}^s\pi$. Then Γ is a random closed set [20]. We can observe that its focal sets are nested: it is said to be *consonant*. The plausibility function is then a possibility measure [25], and π is the corresponding possibility distribution. Function Pl can be computed as $Pl(B) = \sup_{x \in B} \pi(x)$, for any $B \subseteq \mathbb{R}^d$. In particular , $Pl\{x\} = \pi(x)$ for all $x \in \Omega$.

2.4 Dempster's rule

Assume that we have two sources $(\mathscr{S}_i, \mathscr{A}_i, P_i, \Gamma_i)$ for $i = 1,2$, where each Γ_i is a multi-valued mapping from $\mathscr{S}_i$ to 2^Ω, and each source induces a belief function Bel_i on $\mathscr{Y}$. Then, the orthogonal sum of Bel_1 and Bel_2, denoted as $Bel_1 \oplus Bel_2$ is induced by the source $(\mathscr{S}_1 \times \mathscr{S}_2, \mathscr{A}_1 \otimes \mathscr{A}_2, P_1 \otimes P_2, \Gamma_\cap)$, where $\mathscr{A}_1 \otimes \mathscr{A}_2$ is the tensor product algebra on the product space $\mathscr{S}_1 \times \mathscr{S}_2$, $P_1 \otimes P_2$ is the product measure, and $\Gamma_\cap(s_1, s_2) = \Gamma_1(s_1) \cap \Gamma_2(s_2)$. This operation is called Dempster's rule of combination [7]. It is the fundamental operation to combine belief functions induced by independent pieces of evidence in Dempster-Shafer theory.

3 Predictive belief functions

In this paper, we are concerned with the construction of predictive belief functions (PBF), i.e., belief functions that quantify the uncertain on future data Y, given the evidence of past data x. This problem can be illustrated by the following examples, which will be used throughout this paper.

Example 1 *We have observed the times between successive failures of an air-conditioning (AC) system, as shown in Table 1 [21]. We assume the time ξ between failures to have an exponential distribution $\mathscr{E}(\theta)$, with cdf*

$$F(\xi;\theta) = \left[1 - \exp(-\theta x)\right] I(\xi \geq 0),$$

where θ is the rate parameter. Here, the past data $x = (\xi_1, \ldots, \xi_n)$ is a realization of an iid sample $X = (\Xi_1, \ldots, \Xi_n)$, with $\Xi_i \sim \mathscr{E}(\theta)$, and Y is a random variable independent from X, also distributed as $\mathscr{E}(\theta)$. Based on these data and this model, what can we say about the time to the next failure of the system? □

Example 2 *The data shown in Figure 1(a) are annual maximum sea-levels recorded at Port Pirie, a location just north of Adelaide, South Australia, over the period*

Table 1 Times between successive failures of an air-conditioning system, from [21].

23	261	87	7	120	14	62	47	225	71
246	21	42	20	5	12	120	11	3	14
71	11	14	11	16	90	1	16	52	95

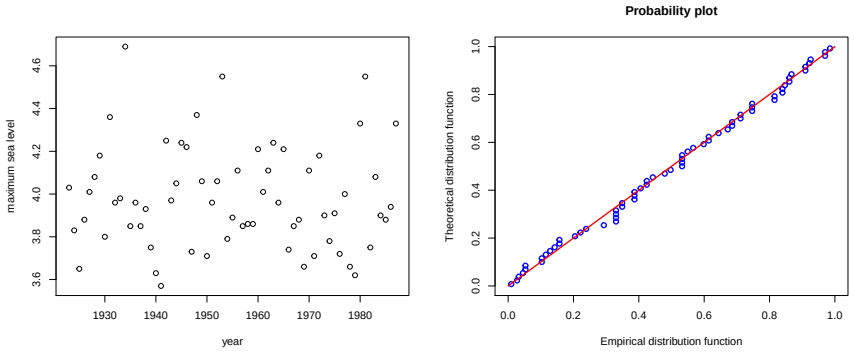
1923-1987 [5]. The probability plot in Figure 1(b) shows a good fit with the Gumbel distribution, with cdf

$$F_X(\xi;\theta) = \exp\left(-\exp\left(-\frac{\xi-\mu}{\sigma}\right)\right), \tag{7}$$

where μ is the mode of the distribution, σ a scale parameter, and $\theta = (\mu,\sigma)$. Suppose that, based on these data, we want to predict the maximum sea level Y in the next $m = 10$ years. Assuming that the distribution of sea level will remain unchanged in the near future (i.e., neglecting, for instance, the effect of sea level rise due to climate change), the cdf of Y is

$$F_Y(y;\theta) = F_X(y;\theta)^m = \exp\left(-m\exp\left(-\frac{y-\mu}{\sigma}\right)\right). \tag{8}$$

The parameter θ is unknown, but the observed data provides information about it. How to represent this information, so as to quantify the uncertainty on Y? What can be, for instance, a sound definition for the degree of belief in the proposition $Y \geq 5$? □



(a) (b)

Fig. 1 Annual maximum sea-levels recorded at Port Pirie over the period 1923-1987 (a), and probability plot for the Gumbel fit to the data (c).

In general, the evidence on Y may consist in (1) the observed data x and (2) prior knowledge on θ, which can be assumed to be represented by a belief function Bel_θ^0. A predictive belief function on Y can thus be denoted as $Bel_Y(\cdot;x,Bel_\theta^0)$. If Bel_θ^0 is vacuous, we simple write $Bel_Y(\cdot;x)$. The following three requirements have been proposed for Bel_Y.

R0: Likelihood principle

As we assume X and Y to be independent, the observation of X provides information on Y only through the parameter θ. The likelihood principle [4, 13] states that all relevant information about θ, after observing $X = x$, is contained in the likelihood function $L(\theta;x) = p(x;\theta)$. Formally, this principle means that two observations X and X' generated by two different random experiments, with probability distributions $p(x;\theta)$ and $p(x';\theta)$, provide the same information about θ as long as $p(x;\theta)$ and $p(x';\theta)$ are proportional, i.e., there is some constant $c = c(x,x')$ not depending on θ, such that $p(x;\theta) = c \cdot p(x';\theta)$ for all $\theta \in \Theta$. Consequently, we should also have

$$\big(\forall\theta \in \Theta, p(x;\theta) = c \cdot p(x';\theta)\big) \Rightarrow Bel_Y(\cdot;x) = Bel_Y(\cdot;x'). \tag{9}$$

The likelihood principle was shown by Birnbaum in [4] to follow from two principles generally accepted by most (but not all) statisticians: the conditionality principle (see also [3, page 25]) and the sufficiency principle.

R1: Compatibility with Bayes

For some statisticians, Bayesian reasoning is a perfectly valid approach to statistical inference provided a prior probability distribution is available, but is questionable in the absence of such prior information. Many authors have attempted to generalize Bayesian inference to some "prior-free" method of inference. This was, in particular, Dempster's motivation in his early papers on belief functions [6, 8]. If we adopt this point of view, then a predictive belief function should coincide with the Bayesian posterior predictive distribution if a probabilistic prior is available. Formally, if $Bel_\theta^0 = P_\theta^0$ is a probability measure, then the following equality should hold,

$$Bel_Y(A;x,P_\theta^0) = P_B(A|x) \tag{10}$$

for all measurable event $A \subseteq \mathscr{Y}$, where $P_B(\cdot|x)$ is the Bayesian posterior predictive probability measure corresponding to (1). This requirement ensures that the Bayesian and belief function approaches yield the same predictions when they are provided with exactly the same information.

A PBF verifying requirements (9) and (10) will be called a Type-I PBF. It can be seen as a representation of the evidence about Y from the observation of X, and possibly additional information on θ; it becomes the Bayesian predictive posterior distribution when combined with a probabilistic prior.

R2: Approximation of the true future data distribution

We may also consider that, if we knew the true value of parameter θ, then we would equate the predictive belief function with the true distribution $P_Y(\cdot;\theta)$ of Y. If we do not know θ, but we have only observed a sample x of X, then the predictive belief function should most of the time (i.e., for most of the observed samples) be less committed than $P_Y(\cdot;\theta)$ [1, 10]. Formally, we may thus fix some $\alpha \in (0,1)$, and require that, for any $\theta \in \Theta$,

$$P_X\left(Bel_Y(\cdot;X) \leq P_Y(\cdot;\theta);\theta\right) \geq 1-\alpha. \tag{11}$$

If $X = (\Xi_1,\ldots,\Xi_n)$ is a sequence of observations, a weaker requirement is to demand that (11) holds in the limit, as $n \to \infty$. A PBF verifying (11), at least asymptotyically, will be called a type-II PBF. For most of the samples, a type-II PBF is a lower approximation of the true probability distribution of Y. It can thus be compared to the plug-in distribution $P_Y(\cdot;\widehat{\theta})$, which is also an approximation of $P_Y(\cdot;\theta)$. However, the PBF will generally be non-additive, as a consequence of accounting not only for random uncertainty, but also for estimation uncertainty.

R3: Calibration

Another line of reasoning, advocated by Martin and Liu [18], is to consider that plausibility values be *calibrated*, in the sense that the plausibility of the true value Y should be small with only a small probability [18, Chapter 9]. More precisely, for any $\theta \in \Theta$ and any $\alpha \in (0,1)$, we my impose the following condition,

$$P_{X,Y}(pl_Y(Y;X) \leq \alpha;\theta) \leq \alpha, \tag{12}$$

or, equivalently,

$$P_{X,Y}(pl_Y(Y;X) > \alpha;\theta) \geq 1-\alpha, \tag{13}$$

where $pl_Y(Y;X) = Pl_Y(\{Y\};X)$ is the *contour function* evaluated at Y. Eqs. (12) and (13) may hold only asymptotically, as the sample size tends to infinity. It follows from (13) that the sets $\{y \in \mathscr{Y} | pl_Y(y;X) > \alpha\}$ are prediction sets at level $1-\alpha$ (maybe, approximately). A PBF verifying (13) will be called a type-III PBF. It can be seen as encoding prediction sets at all levels; as such, it is somewhat similar to a frequentist predictive distribution; however, it is not required to be additive. Requirement (13) is very different from the previous two. In particular, a type-III PBF has no connection with the Bayesian predictive distribution, and it does not approximate the true distribution of Y. Rather, (12) establishes a correspondence between plausibilities and frequencies. A type III-PBF can be seen as a generalized prediction interval.

In the following section, we introduce a simple scheme that will allow us to construct PBF of each of the three kinds above, for any parametric model. We will also mention some alternative methods.

4 Construction of predictive belief functions

In [14, 15], the authors introduced a general method to construct PBFs, by writing the future data Y in the form

$$Y = \varphi(\theta, V), \tag{14}$$

where V is a pivotal variable with known distribution [6, 15, 18]. Equation (14) is called a φ-equation. It can be obtained by inverting the cdf of Y. More precisely, let us first assume that Y is continuous; we can then observe that $V = F_Y(Y;\theta)$ has a standard uniform distribution. Denoting by $F_Y^{-1}(\cdot;\theta)$ the inverse of the cdf $F_Y(\cdot;\theta)$, we get

$$Y = F_Y^{-1}(V;\theta), \tag{15}$$

with $V \sim \mathscr{U}([0,1])$, which has the same form as (14). When Y is discrete, (15) is still valid if F_Y^{-1} now denotes the generalized inverse of F_Y,

$$F_Y^{-1}(V;\theta) = \inf\{y | F_Y(y;\theta) \geq V\}. \tag{16}$$

Example 3 *In the Air Conditioning example, it is assumed that $Y \sim \mathscr{E}(\theta)$, i.e., $F_Y(y;\theta) = 1 - \exp(-\theta y)$. From the equality $F_Y(Y;\theta) = V$, we get*

$$Y = -\frac{\log(1-V)}{\theta}, \tag{17}$$

with $V \sim \mathscr{U}([0,1])$. □

Example 4 *Let Y be the maximum sea level in the next m years, with cdf given by (8). From the equality $F_Y(Y;\theta) = V$, we get $Y = \mu - \sigma \log\log(V^{-1/m})$, with $V \sim \mathscr{U}([0,1])$.* □

The plug-in prediction is obtained by plugging the MLE $\widehat{\theta}$ in (14),

$$\widehat{Y} = \varphi(\widehat{\theta}, V). \tag{18}$$

Now, the Bayesian posterior predictive distribution can be obtained by replacing the constant θ in (14) by a random variable θ_B with the posterior cdf $F_\theta(\cdot;x)$. We then get a random variable Y_B with cdf $F_B(y|x)$ given by (1). We can write

$$Y_B = \varphi(F_\theta^{-1}(U|x), V). \tag{19}$$

The three methods described in the sequel somehow generalize the above methods. They are based on (14), and on belief functions Bel_θ and Bel_V on θ and V induced, respectively, by random sets $\Gamma(U;x)$ and $\Lambda(W)$, where U and W are random variables. The predictive belief function on Y is then induced by the random set

$$\Pi(U,W;x) = \varphi(\Gamma(U;x), \Lambda(W)). \tag{20}$$

Assuming that $\Pi(u,w;x) \neq \emptyset$ for any u, v and x, we thus have

$$Bel_Y(A;x) = P_{U,W}\left\{\Pi(U,W;x) \subseteq A\right\}$$

and

$$Pl_Y(A;x) = P_{U,W}\left\{\Pi(U,W;x) \cap A \neq \emptyset\right\}$$

for all subset $A \subseteq \mathscr{Y}$ for which these expressions are well-defined.

The three methods described below differ in the choice of the random sets $\Gamma(U;x)$ and $\Lambda(W)$. As will we see, each of the three types of PBF described in Section 3 can be obtained by suitably choosing these two random sets.

4.1 Type-I predictive belief functions

As shown in [11], Requirements R0 and R1 jointly imply that the contour function $pl(\theta,x)$ associated to $Bel_\theta(\cdot;x)$ should be proportional to the likelihood function $L(\cdot;x)$. The least committed belief function (in some sense, see [11]) that meets this constraint is the consonant belief function defined by the following contour function,

$$pl(\theta;x) = \frac{L(\theta;x)}{L(\widehat{\theta};x)}, \tag{21}$$

where $\widehat{\theta}$ is a maximizer of $L(\theta;x)$, i.e., a maximum likelihood estimate (MLE) of θ, and it is assumed that $L(\widehat{\theta};x) < +\infty$. As it is consonant, the plausibility of any hypothesis $H \subseteq \Theta$ is the supremum of the plausibilities of each individual values of θ inside H,

$$Pl_\theta(H;x) = \sup_{\theta \in H} pl(\theta;x). \tag{22}$$

The corresponding random set is defined by

$$\Gamma_\ell(U) = \{\theta \in \Theta | pl(\theta;x) \geq U\} \tag{23}$$

with $U \sim \mathscr{U}([0,1])$, i.e., it is the set of values of θ whose relative likelihood is larger than a uniformly distributed random variable U. This *likelihood-based belief function* was first introduced by Shafer [22], and it has been studied by Wasserman [23], among others.

The prediction method proposed in [14, 15] consists in choosing Bel_θ defined by (21)-(22) as the belief function on θ, and P_V, the uniform probability distribution of V, as the belief function on V. The resulting PBF $Bel_{Y,\ell}(\cdot;x)$ is induced by the random set

$$\Pi_\ell(U,V;x) = \varphi(\Gamma_\ell(U;x),V), \tag{24}$$

where (U,V) has a uniform distribution in $[0,1]^2$.

By construction, combining $Bel_\theta(\cdot;x)$ with a Bayesian prior P_θ^0 by Dempster's rule yields the Bayesian posterior $P_B(\cdot|x)$. The random set (24) then becomes

$$\Pi_B(U,V;x) = \varphi(F_B^{-1}(U|x),V), \tag{25}$$

with (U,V) uniformly distribution in $[0,1]^2$. This random set is actually a random point, i.e., a random variable, and this rv is identical to (19): its distribution is the Bayesian posterior predictive distribution. Consequently, the PBF $Bel_{Y,\ell}$ constructed by this method meets requirements R0 and R1.

Example 5 *The contour function for the AC data of Example 1, assuming an exponential distribution, is shown in Figure 2(a). As it is unimodal and continuous, the sets $\Gamma_\ell(u;x)$ are closed intervals $[\theta^-(u),\theta^+(u)]$, whose bounds can be approximated numerically as the roots of the equation $pl(\theta;x) = u$. From (17), the random set $\Pi_\ell(U,V;x)$ is then the random closed interval*

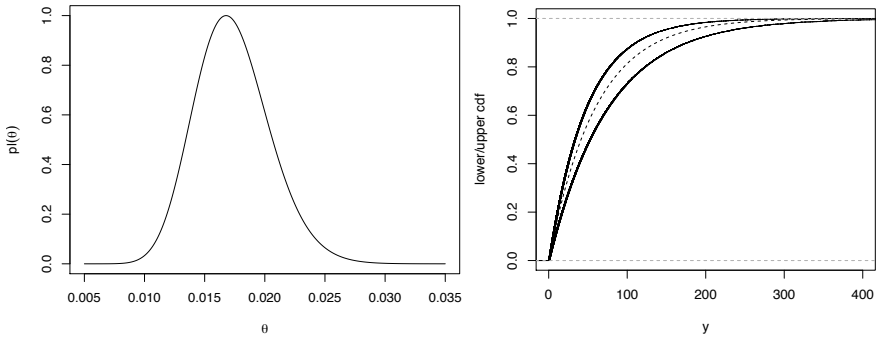
$$\Pi_\ell(U,V;x) = [Y^-(U,V;x), Y^+(U,V;x)],$$

with

$$Y^-(U,V;x) = -\frac{\log(1-V)}{\theta^+(U)}$$

and

$$Y^+(U,V;x) = -\frac{\log(1-V)}{\theta^-(U)}.$$



(a) (b)

Fig. 2 AC data. (a): Contour function; (b): Lower and upper cdf (solid lines) and plug-in cdf (dotted line)

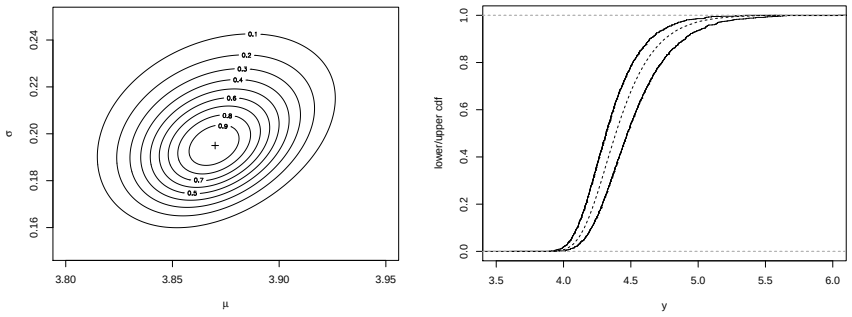
As shown by Dempster [9], the following equalities hold, for any $y \geq 0$,

$$Bel_Y((-\infty,y]) = P_{U,V}(Y^+(U,V;x) \leq y)$$
$$Pl_Y((-\infty,y]) = P_{U,V}(Y^-(U,V;x) \leq y),$$

i.e., they are the cdfs of, respectively, the upper and lower bounds of Π_ℓ. Functions $Bel_Y((-\infty,y])$ and $Pl_Y((-\infty,y])$ are called the lower and upper cdfs of the random set $\Pi_\ell(U,V;x)$. As explained in [15], they can be approximated by Monte Carlo simulation: let (u_i,v_i), $i = 1,\ldots,N$ be a pseudo-random sequence generated

independently from the uniform distribution in $[0,1]^2$. *Let* $y_i^- = y^-(u_i, v_i; x)$ *and* $y_i^+ = y^+(u_i, v_i; x)$ *be the corresponding realizations of the bounds of* Π_ℓ. *Then, the lower and upper cdfs can be approximated by the empirical cdfs of the* y_i^+ *and the* y_i^-, *respectively. These functions are plotted in Figure 2(b), together with the plug-in cdf* $F_Y(y; \widehat{\theta})$, *with* $\widehat{\theta} = 1/\bar{x}$. *We can observe that the plug-in cdf is always included in the band defined by the lower and upper cdf, which is a consequence of the inequalities* $\theta^-(u) \leq \widehat{\theta} \leq \theta^-(u)$ *for any* $u \in (0,1]$. *We note that* $\widehat{\theta} = \theta^-(1) = \theta^+(1)$. □

Example 6 *Let us now consider the Sea Level data of Example 2. The contour function (21) for these data is plotted in Figure 3(a). As the level sets* $\Gamma_\ell(u; x)$ *of this function are closed and connected, the sets* $\Pi_\ell(U, V; x)$ *still are closed intervals in this case [15]. To find the bounds* $Y^-(u, v; x)$ *and* $Y^+(u, v; x)$ *for any pair* (u, v), *we now need to search for the minimum an the maximum of* $\varphi(\theta, v)$, *under the constraint* $pl(\theta; x) \geq u$. *This task can be performed by a nonlinear constrained optimization algorithm. The lower and upper cdfs computed using this method are shown in Figure 3(b).* □



(a) (b)

Fig. 3 Port Pirie sea-level data: (a): Contour plot of the relative likelihood function; (b): Lower and upper cdfs of the type-I PBF; the central broken line corresponds to the plug-in prediction.

4.2 Type-II predictive belief functions

The φ-equation (14) also allows us to construct a type-II PBF, such as defined in [10]. Let $C(X)$ be a confidence set for θ at level $1-\alpha$, i.e.,

$$P_X(C(X) \ni \theta; \theta) = 1 - \alpha. \tag{26}$$

Consider the following random set,

$$\Pi_{Y,c}(V;x) = \varphi(C(x),V), \tag{27}$$

which is a special case of the general expression (20), with $\Gamma(U;x) = C(x)$ for all $U \in [0,1]$, $W = V$ and $\Lambda(V) = V$. The following theorem states that the belief function induced by the random set (27) is a type-II PBF.

Theorem 1 *Let $Y = \varphi(\theta,V)$ be a random variable, and $C(X)$ a confidence region for θ at level $1-\alpha$. Then, the belief function $Bel_{U,c}(\cdot;x)$ induced by the random set $\Pi_{Y,c}(V;x) = \varphi(C(x),V)$ verifies*

$$P_X(Bel_{Y,c}(\cdot;X) \leq P_Y(\cdot;\theta);\theta) \geq 1-\alpha, \tag{28}$$

i.e., it is a type-II PBF.

Proof. If $\theta \in C(x)$, then $\varphi(\theta,V) \in \varphi(C(x),V)$ for any V. Consequently, the following implication holds for any measurable subset $A \subseteq \mathscr{Y}$, and any $x \in \mathscr{X}$,

$$\varphi(C(x),V) \subseteq A \Rightarrow \varphi(\theta,V) \in A.$$

Hence,

$$P_V(\varphi(C(x),V) \subseteq A) \leq P_V(\varphi(\theta,V) \subseteq A),$$

or, equivalently,

$$Bel_{Y,c}(A;x) \leq P_Y(A;\theta). \tag{29}$$

As (29) holds whenever $\theta \in C(x)$, and $P_X(C(X) \ni \theta;\theta) = 1-\alpha$, it follows that (29) holds for any measurable event A with probability at least $1-\alpha$, i.e.,

$$P_X\left(Bel_{Y,c}(\cdot;X) \leq P_Y(\cdot;\theta);\theta\right) \geq 1-\alpha.$$

□

If $C(X)$ is an approximate confidence region, then obviously (28) will hold only approximately. In the case where $X = (X_1,\ldots,X_n)$ is iid, the likelihood function will often provide us with a means to obtain a confidence region on θ. From Wilks' theorem [24], we know that, under regularity conditions, $-2\log pl(\theta;X)$ has approximately, for large n, a chi square distribution with p degrees of freedom, where p is the dimension of θ. Consequently, the sets

$$\Gamma_\ell(c;X) = \{\theta \in \Theta | pl(\theta;X) \geq c,$$

with $c = \exp(-0.5\chi^2_{p;1-\alpha})$, are approximate confidence regions at level $1-\alpha$. The corresponding predictive random set is

$$\Pi_{Y,c}(V;x) = \varphi(\Gamma_\ell(c;x),V). \tag{30}$$

We can see that this expression is similar to (24), except that, in (30), the relative likelihood function is cut at a fixed level c. A similar idea was explored in Ref. [26].

Table 2 gives values of c for different values of p and $\alpha = 0.05$. We can see that c decreases quickly with p, which means that the likelihood-based confidence regions and, consequently, the corresponding PBFs will become increasing imprecise as p increases. In particular, the likelihood-based type-II PBFs will typically be less committed than the type-I PBFs.

Table 2 Likelihood levels c defining approximate 95% confidence regions.

p	1	2	5	10	15
c	0.15	0.5	3.9e-03	1.1e-04	3.7e-06

Example 7 *For the AC data, the likelihood-based confidence level at level* $1-\alpha = 0.95$ *is*

$$[\theta^-(c), \theta^+(c)] = [0.01147, 0.02352],$$

with $c = 0.15$. *It is very close to the exact confidence level at the same level,*

$$\left[\frac{\widehat{\theta}\chi^2_{\alpha/2,2n}}{2n}, \frac{\widehat{\theta}\chi^2_{1-\alpha/2,2n}}{2n}\right] = [0.01132, 0.02329].$$

The corresponding Type-II PBF is induced by the random interval

$$\Pi_{Y,c}(V;x) = \left[-\frac{\log(1-V)}{\theta^+(c)}, -\frac{\log(1-V)}{\theta^-(c)}\right].$$

The lower and upper bounds of this interval have exponential distributions with rates $\theta^+(c)$ *and* $\theta^-(c)$, *respectively. Figure 4 shows the corresponding lower and upper cdfs, together with those of the Type-I PBF computed in Example 5. We can see that the Type-II PBF at the 95% confidence level is less committed than the Type-I PBF.*

Example 8 *Figure 5 shows the lower and upper cdfs of the type-II PBF constructed from the likelihood-based confidence region with* $\alpha = 0.05$. *The estimate of the true coverage probability, obtained using the parametric bootstrap method with* $B = 5000$ *bootstrap samples, was 0.94998, which is remarkably close to the nominal level. The simulation method to compute these functions is similar to that explained in Example 6, except that we now have* $u_i = c = 0.05$ *for* $i = 1, \ldots, n$. *The lower and upper cdfs form a confidence band on the true cdf of* Y. *Again, we observe that this band is larger than the one corresponding to the type-I PBF.*

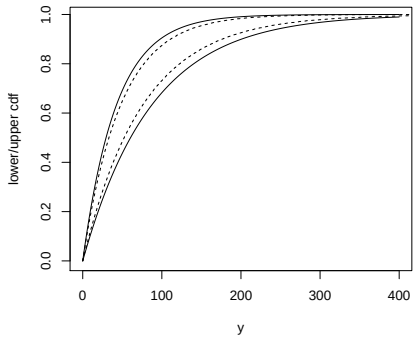


Fig. 4 Lower and upper cdfs of the type-II PBF for the AC data (solid lines). The type-I lower and upper cdf are shown as broken lines.

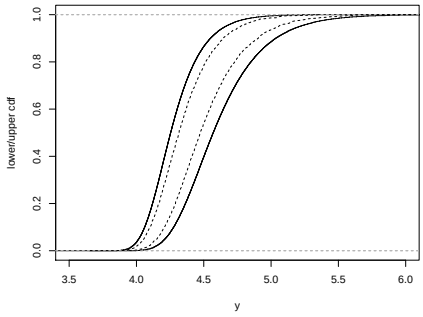


Fig. 5 Lower and upper cdfs of the type-II PBF for the sea-level example (solid lines). The type-I lower and upper cdf are shown as broken lines.

4.3 Type-III predictive belief functions

The calibration condition (12) was introduced by Martin and Liu [17, 18], in the context of their theory of Inferential Models (IMs). An equivalent formulation is to require that the random variable $pl_Y(Y;X)$ be stochastically not less than a random variable having a standard uniform distribution. In [18, Chapter 9], Martin and Liu propose a quite complex method for constructing PBFs verifying this requirements, based on IMs. It turns out that such Type-III PBFs (as we call them in this paper) can be generated by a simple construction procedure based on the φ-equation (14) and suitable belief functions on θ and V. Because the notion of Type-III PBFs is intimately related to prediction sets, and prediction sets at a given level can only be defined for continuous random variables, we will assume Y to be continuous in this section.

Let us first assume that θ is known. In that case, predicting $Y = \varphi(\theta, V)$ boils down to predicting $V = F(Y;\theta)$. Consider the random interval

$$\Lambda(W) = \left[\frac{W}{2}, 1 - \frac{W}{2}\right],$$

with $W \sim \mathscr{U}([0,1])$. It is easy to check that the induced contour function is $pl(v) = 1 - |2v - 1|$ (it is a triangular possibility distribution with support $[0,1]$ and mode 0.5), and $pl(V) \sim \mathscr{U}([0,1])$. Consider the predictive random set

$$\Pi_Y(W) = \varphi(\theta, \Lambda(W)) \tag{31}$$

and the associated contour function

$$pl(y) = 1 - |1 - 2F(y;\theta)|. \tag{32}$$

It is clear that $pl(Y) = pl(V) \sim \mathscr{U}([0,1])$, and the consonant belief function with contour function (32) verifies the calibration property (12). We can observe that the transformation (32) from the probability distribution of Y to this possibility distribution is an instance of the family of probability-possibility transformations studied in [12]. The mode of the possibility distribution is the median $y_{0.5} = \varphi(\theta, 0.5)$, and each α-cut $\Pi_Y(\alpha) = [y_{\alpha/2}, y_{1-\alpha/2}]$ with $\alpha \in (0,1)$ is a prediction interval for Y, at level $1-\alpha$.

Until now, we have assume θ to be known. When θ is unknown, we could think of replacing it by its MLE $\widehat{\theta}$, and proceed as above by applying the same probability-possibility distribution to the plug-in predictive distribution $F_Y(u;\widehat{\theta})$. As already mentioned, this approach would amount to neglecting the estimation uncertainty, and the α-cuts of the resulting possibility distribution could have a coverage probability significantly smaller than $1-\alpha$. A better approach, following [16], is to consider the exact or approximate pivotal quantity $\widetilde{V} = F(Y;\widehat{\theta}(X))$. We assume that $\widehat{\theta}$ is a consistent estimator of θ as the information about θ increases, and $\widetilde{V}$ is asymptotically distributed as $\mathscr{U}([0,1])$ [16]. However, for finite sample size, the distribution of $\widetilde{V}$ will generally not be uniform. Let G be the cdf of $\widetilde{V}$, assuming that it is pivotal, and let $\widetilde{\Lambda}(W)$ be the random interval

$$\widetilde{\Lambda}(W) = \left[G^{-1}(W/2), G^{-1}(1 - W/2)\right]$$

with $W \sim \mathscr{U}([0,1])$ and corresponding contour function

$$pl(\widetilde{v}) = 1 - \left|1 - 2G(\widetilde{v})\right|.$$

The random set

$$\widetilde{\Pi}_Y(W;x) = \varphi(\widehat{\theta}(x), \widetilde{\Lambda}(W))$$

induces the contour function

$$pl(y;x) = 1 - \left|1 - 2G\{F[y;\widehat{\theta}(x)]\}\right|. \tag{33}$$

As $G(F(Y;\widehat{\theta}(X))) \sim \mathscr{U}([0,1])$, we have $pl(Y;X) \sim \mathscr{U}([0,1])$, and the focal sets $\widetilde{\Pi}_Y(\alpha;X)$ are exact prediction intervals at level $1-\alpha$. Consequently, the consonant belief function with contour function (33) is a type-III PBF. We can remark that it is obtained by applying the probability-possibility transformation (32) to the predictive confidence distribution $\widetilde{F}(y;x) = G\{F[y;\widehat{\theta}(x)]\}$.

When an analytical expression of the cdf G is not available, or $\widetilde{V}$ is only asymptotically pivotal, an approximate distribution $\widetilde{G}$ can be determined by a parametric bootstrap approach [16]. Specifically, let $x_1^*,\ldots,x_B^*$ be B and $y_1^*,\ldots,y_B^*$ be B bootstrap replicates of x and y, respectively. We can compute the corresponding values $\widetilde{v}_b^* = F(y_i^*;\widehat{\theta}(x_b^*))$, $b=1,\ldots,B$, and the distribution of $\widetilde{V}$ can be approximated by the empirical cdf

$$\widetilde{G}(v) = \frac{1}{B}\sum_{b=1}^{B} I(\widetilde{v}_b^* \leq v).$$

Example 9 *Consider again the AC example. For the exponential distribution, it has been shown [16] that the quantity*

$$\widetilde{V} = F(Y,\widehat{\theta}(X)) = 1-\exp(-Y\widehat{\theta}(X))$$

is pivotal, and has the following cdf,

$$G(\widetilde{v}) = 1-\left\{1-\frac{1}{n}\log(1-\widetilde{v})\right\}^{-n}.$$

The predictive cdf is then

$$\widetilde{F}(y;x) = G\{F(y,\widehat{\theta}(x))\} = 1-\left(1+\frac{y\widehat{\theta}(x)}{n}\right)^{-n}$$

and the contour function of the type-III PBF is

$$pl(y;x) = 1-\left|2\left(1+\frac{y\widehat{\theta}(x)}{n}\right)^{-n}-1\right|. \tag{34}$$

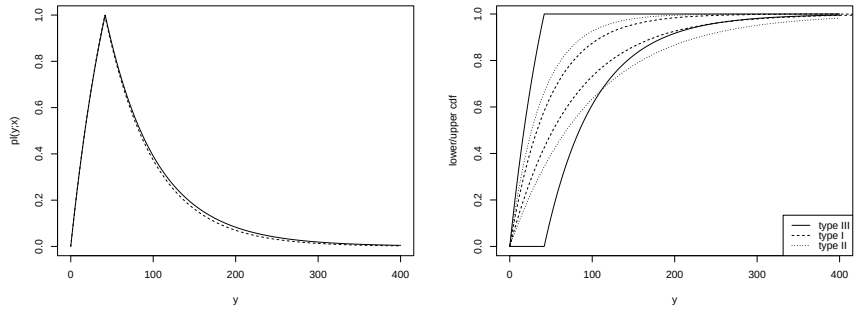
Figure 6(a) shows the contour function (34) for the AC data (solid line), together with the contour function induced by the plug-in distribution (interrupted line). The two curves are quite close in this case: for $n=30$, the distribution of $\widetilde{V}$ is already very close to the standard uniform distribution. Figure 6(b) shows the lower and upper cdfs of the PBF, together with the Type-I and Type-II ($1-\alpha=0.95$) lower and upper cdfs for the same data. As the Type-III PBF is consonant, the lower and upper cdfs can computed from the contour function as

$$Pl_Y((-\infty,y]) = \sup_{y'\leq y} pl(y') = \begin{cases} pl(y;x) & \text{if } y \leq \widetilde{F}^{-1}(0.5;x) \\ 1 & \text{otherwise,} \end{cases}$$

and

$$Bel_Y((-\infty, y]) = 1 - \sup_{y'>y} pl(y') = \left[1 - pl(y;x)\right] I\left(y > \widetilde{F}^{-1}(0.5;x)\right).$$

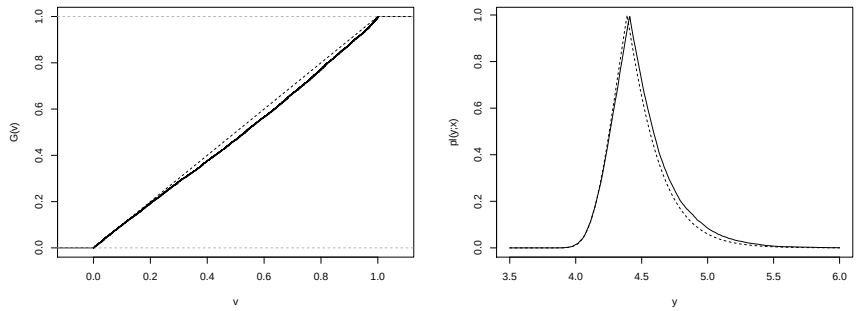
□



(a) (b)

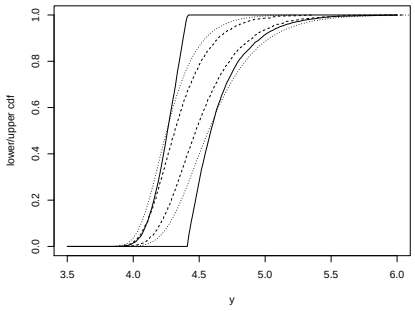
Fig. 6 AC example. (a): Contour function of the type-III PBF (solid line), and contour function induced by the plug-in distribution (interrupted line); (b): Lower and upper cdfs of the type-III PBF (solid lines). The type-I and type-II lower and upper cdf are shown, respectively, as interrupted and dotted lines.

Example 10 *Let us now consider again the sea-level data. Here, the exact distribution of the quantity* $\widetilde{V} = F_Y(Y;\widehat{\theta}(X))$ *is intractable, but it can be estimated by the parametric bootstrap technique. Figure 7(a) shows the bootstrap estimate of the distribution of* $\widetilde{V}$*, with* $B = 10000$*. There is clearly a small, but discernible departure from the uniform distribution. Figure 7(b) shows the contour function of the type-III PBF, together with that induced by the plug-in predictive distribution (corresponding to the approximation* $G(\widetilde{v}) = \widetilde{v}$*). Again, the two curves are close, but clearly discernible. With* $n = 65$*, the prediction intervals computed from the plug-in distribution have true coverage probabilities quite close to the stated ones. Finally, the lower and upper cdf of the type-III PBF for the Port-Pirie data are shown in Figure 7(c), together with the corresponding functions for the type-I and type-II PBFs. Comparing Figures 6(b) and 7(c), we can see that, in both cases, the type-I PBF is less committed than the type-III PBF. It is not clear, however, whether this result holds in general.* □



(a)

(b)



(c)

Fig. 7 Sea-level example. (a): Bootstrap estimate of the cdf G of $\widetilde{V} = F(Y, \widehat{\theta}(X))$; (b): Contour function of the type-III PBF (solid line), and contour function induced by the plug-in distribution (interrupted line); (c): Lower and upper cdfs of the type-III PBF (solid lines). The type-I and type-II lower and upper cdf are shown, respectively, as interrupted and dotted lines.

5 Conclusions

Being related to random sets, belief functions have greater expressivity than probability measures. In particular, the additional degrees of freedom of the belief function framework make it possible to distinguish between lack of information and randomness. In this paper, we have considered different ways of exploiting this high expressivity to quantify prediction uncertainty. Based on three distinct requirements, three different kinds of predictive belief functions have been distinguished, and construction procedures for each of them have been proposed. Type-I belief functions have a Bayesian flavor, and boil down to Bayesian posterior predictive belief functions when a prior probability distribution on the parameter is provided. In contrast, belief functions of the other types are frequentist in spirit. Type-II belief functions correspond to a family of probability measures, which contain the

true distribution of the random variable of interest with some probability, in a repeated sampling setting. Type-III belief functions are "frequency-calibrated", in so far as the true value of the variable of interest rarely receives a small plausibility. It should be noticed by "frequentist" predictive belief functions (of types II and III) are not compatible with Bayesian inference, i.e., they do not allow us to recover the Bayesian posterior predictive distribution when combined with a Bayesian prior. It thus seems that the Bayesian and frequentist views cannot be easily reconciled, and different inference procedures have to coexist, just as frequentist and Bayesian procedures in mainstream statistics. Beyond philosophical arguments, the practicality of these construction procedures, as well as their interpretability and acceptability by decision-makers remain to be investigated.

References

[1] Aregui, A., Denœux, T.: Constructing predictive belief functions from continuous sample data using confidence bands. In: G. De Cooman, J. Vejnarová, M. Zaffalon (eds.) Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA '07), pp. 11–20. Prague, Czech Republic (2007)

[2] Barndorff-Nielsen, O.E., Cox, D.R.: Prediction and asymptotics. Bernoulli **2**(4), 319–340 (1996)

[3] Berger, J.O., Wolpert, R.L.: The likelihood principle: a review, generalizations, and statistical implications, *Lecture Notes–Monograph Series*, vol. 6, 2nd edn. Institute of Mathematical Statistics, Hayward, CA (1988)

[4] Birnbaum, A.: On the foundations of statistical inference. Journal of the American Statistical Association **57**(298), 269–306 (1962)

[5] Coles, S.G.: An Introduction to Statistical Modelling of Extreme Values. Springer, London (2001)

[6] Dempster, A.P.: New methods for reasoning towards posterior distributions based on sample data. Annals of Mathematical Statistics **37**, 355–374 (1966)

[7] Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics **38**, 325–339 (1967)

[8] Dempster, A.P.: A generalization of Bayesian inference (with discussion). J. R. Statistical Society B **30**, 205–247 (1968)

[9] Dempster, A.P.: Upper and lower probabilities generated by a random closed interval. Annals of Mathematical Statistics **39**(3), 957–966 (1968)

[10] Denœux, T.: Constructing belief functions from sample data using multinomial confidence regions. International Journal of Approximate Reasoning **42**(3), 228–252 (2006)

[11] Denœux, T.: Likelihood-based belief function: justification and some extensions to low-quality data. International Journal of Approximate Reasoning **55**(7), 1535–1547 (2014)

[12] Dubois, D., Foulloy, L., Mauris, G., Prade, H.: Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities. Reliable Computing **10**(4), 273–297 (2004)
[13] Edwards, A.W.F.: Likelihood (expanded edition). The John Hopkins University Press, Baltimore, USA (1992)
[14] Kanjanatarakul, O., Sriboonchitta, S., Denœux, T.: Forecasting using belief functions: an application to marketing econometrics. International Journal of Approximate Reasoning **55**(5), 1113–1128 (2014)
[15] Kanjanatarakul, O., Sriboonchitta, S., Denœux, T.: Statistical estimation and prediction using belief functions: principles and application to some econometric models. International Journal of Approximate Reasoning **72**, 71–94 (2016)
[16] Lawless, J.F., Fredette, M.: Frequentist prediction intervals and predictive distribution. Biometrika **92**(3), 529–542 (2005)
[17] Martin, R., Lingham, R.T.: Prior-free probabilistic prediction of future observations. Technometrics **58**(2), 225–235 (2016)
[18] Martin, R., Liu, C.: Inferential Models: Reasoning with Uncertainty. CRC Press, Boca Raton (2016)
[19] Nguyen, H.T.: On random sets and belief functions. Journal of Mathematical Analysis and Applications **65**, 531–542 (1978)
[20] Nguyen, H.T.: An Introduction to Random Sets. Chapman and Hall/CRC Press, Boca Raton, Florida (2006)
[21] Olkin, I., Gleser, L., Derman, C.: Probability Models and Applications. Macmillan (1994)
[22] Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton, N.J. (1976)
[23] Wasserman, L.A.: Belief functions and statistical evidence. The Canadian Journal of Statistics **18**(3), 183–196 (1990)
[24] Wilks, S.S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Statist. **9**(1), 60–62 (1938)
[25] Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems **1**, 3–28 (1978)
[26] Zhu, K., Thianpaen, N., Kreinovich, V.: How to make plausibility-based forecasting more accurate. In: V. Kreinovich, S. Sriboonchitta, V.N. Huynh (eds.) Robustness in Econometrics, pp. 99–110. Springer Berlin, Cham, Switzerland (2017)