Prediction of future observations using belief functions: A likelihood-based approach

Orakanya Kanjanatarakul, Thierry Denoeux, Songsak Sriboonchitta

Prediction of future observations using belief functions: a likelihood-based approach

Orakanya Kanjanatarakul[1,2], Thierry Denœux[2]
and Songsak Sriboonchitta[3]

[1] Faculty of Management Sciences,
Chiang Mai Rajabhat University, Thailand
[2] Sorbonne Universités
Université de Technologie de Compiègne, CNRS,
UMR 7253 Heudiasyc, France
[3] Faculty of Economics, Chiang Mai University, Thailand

December 17, 2015

Abstract

We study a new approach to statistical prediction in the Dempster-Shafer framework. Given a parametric model, the random variable to be predicted is expressed as a function of the parameter and a pivotal random variable. A consonant belief function in the parameter space is constructed from the likelihood function, and combined with the pivotal distribution to yield a predictive belief function that quantifies the uncertainty about the future data. The method boils down to Bayesian prediction when a probabilistic prior is available. The asymptotic consistency of the method is established in the iid case, under some assumptions. The predictive belief function can be approximated to any desired accuracy using Monte Carlo simulation and nonlinear optimization. As an illustration, the method is applied to multiple linear regression.

Keywords: Dempster-Shafer Theory, Evidence Theory, Statistical inference, Likelihood, Uncertainty, Forecasting, Linear regression.

1 Introduction

The Dempster-Shafer theory of belief functions [11, 12, 58] is now a well established formal framework for reasoning with uncertainty. It has been successfully applied to many problems, including classification [21], function approximation [54, 64], clustering [20, 47], image segmentation [40], scene perception [39], multiple-attribute decision making [10], machine diagnosis and prognosis [56, 57], etc. To further extend the application of Dempster-Shafer theory to new problems, we need well-founded and computationally tractable methods to model different kinds of evidence in the belief function framework. The purpose of this paper, which builds on previous work by the authors [18,19,35,36], is to present such methods for statistical inference and prediction.

Although statistical inference provided the first motivation for introducing belief functions in the 1960's [11–13], applications in this area have remained limited. The reason might be that the approach initially introduced by Dempster [15], and further elaborated in recent years [41, 45, 46] under the name of the "weak belief" model, is computationally demanding and it cannot be applied easily to the complex statistical models encountered in many areas, such as machine learning or econometrics. For this reason, frequentist and Bayesian methods have remained by far the most popular. Yet, these approaches are not without defect. It is well known that frequentist methods provide pre-experimental measures of the accuracy of statistical evidence, which are not conditioned on specific data [8]. For instance, a 95% confidence interval contains the parameter of interest for 95% of the samples, but the 95% value is just an average, and the interval may certainly (or certainly not) contain the parameter for some specific samples [8, page 5]. For this reason, a confidence level or a p-value are not appropriate measures of the strength of statistical evidence (see more discussion on this point in [8]). Bayesian methods do implement some form of post-experimental reasoning. However, they require the statistician to provide a prior probability distribution, which is problematic when no prior knowledge, or only weak information, is available. These shortcomings of traditional methods of inference have motivated the development of alternative approaches up to these days. The theory of belief functions, which focuses on the concept of evidence [58], seems particularly well-suited as a model of statistical evidence. However, statistical methods based on belief functions will not gain widespread acceptance unless they are conceptually simple and easily applicable to a wide range of problems and models.

In this paper, we advocate another approach to statistical inference us-

ing belief functions, based on the concept of likelihood. This approach was initially introduced by Shafer in [58, Chapter 11] and was later studied by some authors [1, 65]. It was recently derived axiomatically from three principles: the likelihood principle, compatibility with Bayesian inference and the principle of maximum uncertainty [18, 19]. This approach is in line with likelihood-based inference as advocated by Fisher in his later work [28] and, later, by Birnbaum [9], Barnard [5], and Edwards [27], among others. It retains the idea that "all we need to know about the result of a random experiment is contained in the likelihood function", but reinterprets it as defining a consonant belief function. Combining this belief function by Dempster's rule with a Bayesian prior yields the Bayesian posterior distribution, which ensures compatibility with Bayesian inference. An important advantage of the belief function approach, however, is that it allows the statistician to use either a weaker form of prior information[1], as a general belief function, or even no prior information at all (which corresponds to providing a vacuous belief function as prior information).

In recent work [36], we have extended the likelihood-based approach to prediction problems. Prediction can be defined as the task of making statements about data that have not yet been observed. Assume, for instance, that we have drawn y balls out of n draws with replacement from an urn contain an unknown proportion θ of black balls, and a proportion $1 - \theta$ of white balls. Let z be a binary variable defined by $z = 1$ if the next ball to be drawn is black, and $z = 0$ otherwise. Guessing the value of z is a prediction problem. The general model for such problems involves a pair (y, z) of random quantities whose joint distribution depends on some parameter θ, where y is observed but z is not yet observed. In [36], we proposed a solution to this problem, using the likelihood-based approach outlined above, and we applied it to a very specific model in the field of marketing econometrics. The same approach was used in [67] to calibrate a certain kind of binary classifiers. In this paper, we further explore this method by proving that, under some mild assumptions, the predictive belief function converges, in some sense, to the true probability distribution of the not-yet observed data. We also address describe several simulation and approximation techniques to estimate the predictive belief function or an outer approximation thereof. Finally, we illustrate the practical application of the method using multiple linear regression. In particular, we show that the *ex ante* forecasting

[1] A similar goal is pursued by robust Bayes [7] and imprecise probability approaches (see, e.g., [42,48]), which attempt to represent weak prior information by sets of probability measures. A comparison with these alternative approaches is beyond the scope of this paper.

problem has a natural and simple solution using our approach.

The rest of this paper is organized as follows. Some background on belief functions will first be given in Section 2. The estimation and prediction methods will then be presented, respectively, in Sections 3 and 4. The application to linear regression will then be studied in Section 5. Finally, Section 6 will conclude the paper.

2 Background on belief functions

Most applications of Dempster-Shafer theory use belief functions defined on finite sets [58]. However, in statistical models, the parameter and sample spaces are often infinite. To make the paper self-contained, we will recall some basic definitions and results on belief functions defined on arbitrary spaces (finite or not).

2.1 Belief function induced by a source

Let $(\Omega, \mathcal{B})$ be a measurable space. A *belief function* on $\mathcal{B}$ is a mapping $Bel : \mathcal{B} \to [0,1]$ verifying the following three conditions:

1. $Bel(\emptyset) = 0$;
2. $Bel(\Omega) = 1$;
3. For any $k \geq 2$ and any collection $B_1, \ldots, B_k$ of elements of $\mathcal{B}$,

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

Similarly, a *plausibility function* can be defined as a function $Pl : \mathcal{B} \to [0,1]$ such that:

1. $Pl(\emptyset) = 0$;
2. $Pl(\Omega) = 1$;
3. For any $k \geq 2$ and any collection $B_1, \ldots, B_k$ of elements of $\mathcal{B}$,

$$Pl\left(\bigcap_{i=1}^{k} B_i\right) \leq \sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} Pl\left(\bigcup_{i \in I} B_i\right). \tag{2}$$

It is clear that, Bel is a belief function if and only if Pl defined by $Pl(B) = 1 - Bel(\overline{B})$ for all $B \in \mathcal{B}$ is a plausibility function. The function $pl : \Omega \to [0, 1]$ such that $pl(x) = Pl(\{x\})$ for any $x \in \Omega$ is called the *contour function* of Bel.

In Dempster-Shafer theory, a belief function Bel is used as a representation of a belief state about some question of interest, induced by a given piece of evidence. The number $Bel(B)$ is interpreted as the probability that the evidence supports (implies) B, while $Pl(B)$ is the probability that the evidence does not support the complement of B (is not contradictory with B).

A convenient way to create a belief function is through a multivalued mapping from a probability space to $\mathcal{B}$ [12]. More precisely, let $(S, \mathcal{A}, P)$ be a probability space and let $\Gamma : S \to 2^{\Omega}$ be a multi-valued mapping. We can define two inverses of Γ:

1. The lower inverse

$$\Gamma_*(B) = B_* = \{s \in S | \Gamma(s) \neq \emptyset, \Gamma(s) \subseteq B\}; \tag{3}$$

2. The upper inverse

$$\Gamma^*(B) = B^* = \{s \in S | \Gamma(s) \cap B \neq \emptyset\}, \tag{4}$$

for all $B \in \mathcal{B}$. We say that Γ is *strongly measurable* with respect to $\mathcal{A}$ and $\mathcal{B}$ iff, for all $B \in \mathcal{B}$, $B^* \in \mathcal{A}$. This implies that, for all $B \in \mathcal{B}$, $B_* \in \mathcal{A}$.

We then have the following important theorem [52].

Theorem 1 *Let $(S, \mathcal{A}, \mathbb{P})$ be a probability space, $(\Omega, \mathcal{B})$ a measurable space and Γ a strongly measurable mapping w.r.t. $\mathcal{A}$ and $\mathcal{B}$ such that $\mathbb{P}(\Omega^*) \neq 0$. Let the lower and upper probability measures be defined as follows: for all $B \in \mathcal{B}$,*

$$\mathbb{P}_*(B) = K \cdot \mathbb{P}(B_*), \tag{5a}$$

$$\mathbb{P}^*(B) = K \cdot \mathbb{P}(B^*) = 1 - \mathbb{P}_*(\overline{B}), \tag{5b}$$

where $K = [\mathbb{P}(\Omega^*)]^{-1}$. Then, $\mathbb{P}_*$ is a belief function and $\mathbb{P}^*$ is the dual plausibility function.

Under the conditions of Theorem 1, the four-tuple $(S, \mathcal{A}, \mathbb{P}, \Gamma)$ is called a *source* for the belief function $\mathbb{P}_*$. The set $\Gamma(s)$ are called the *focal sets* of Bel.

Given a source $(S, \mathcal{A}, \mathbb{P}, \Gamma)$, we can also define a third notion of inverse for Γ as

$$\widetilde{\Gamma}(B) = \widetilde{B} = \{s \in S | \Gamma(s) \supseteq B\}, \tag{6}$$

for all $B \in \mathcal{B}$. If $\widetilde{B} \in \mathcal{A}$ for all $B \in \mathcal{B}$, then we can define another function Q from $\mathcal{B}$ to $[0, 1]$, called the *commonality function*, as $Q(B) = K \cdot \mathbb{P}(\widetilde{B})$.

2.2 Practical models

In Section 3 below, we will encounter three important examples of sources defining belief functions of practical interest in $\Omega = \mathbb{R}^d$: random vectors, consonant random closed sets and random intervals.

Random vectors

Let $\boldsymbol{X}$ be a random vector from $(S, \mathcal{A}, \mathbb{P})$ to $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. It is clear that the mapping Γ from S to the power set of $\mathbb{R}^d$, defined by $\Gamma(s) = \{X(s)\}$, is strongly measurable. The induced belief function is the probability distribution $\mathbb{P}_{\boldsymbol{X}}$ of $\boldsymbol{X}$.

Consonant random closed sets

Let us assume that $\Omega = \mathbb{R}^d$. 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

$$^s\pi = \{x \in \mathbb{R}^d | \pi(x) \geq s\} \tag{7}$$

is closed. Furthermore, assume that $\pi(x) = 1$ for some x. Let S be the interval $[0, 1]$, $\mathcal{A}$ be the Borel σ-field on $[0, 1]$, λ the uniform probability measure on S, and Γ the mapping defined by $\Gamma(s) = {}^s\pi$. Then Γ is strongly measurable and it defines a random closed set [53]. We can observe that its focal sets are nested: it is said to be *consonant*. The corresponding plausibility and belief functions verify the following equalities, for any $B \subset \mathbb{R}^d$:

$$Pl(B) = \sup_{x \in B} \pi(x), \tag{8}$$

and

$$Bel(B) = 1 - Pl(\overline{B}) = \inf_{x \notin B} (1 - \pi(x)). \tag{9}$$

The corresponding contour function pl is equal to π.

Random closed intervals

Let (U, V) be a bi-dimensional random vector from a probability space $(S, \mathcal{A}, \mathbb{P})$ to $\mathbb{R}^2$ such that

$$\mathbb{P}(\{s \in S | U(s) \leq V(s)\}) = 1. \tag{10}$$

The mapping

$$\Gamma : s \to \Gamma(s) = [U(s), V(s)] \tag{11}$$

is strongly measurable. It defines a random closed interval [14], which is a source for a belief function in $\Omega = \mathbb{R}$.

The *lower and upper cumulative distribution functions* (cdf) of the random interval $[U, V]$ are the function F_* and F^* defined, respectively, as follows,

$$F_*(x) = Bel\left((-\infty, x]\right) \tag{12a}$$
$$F^*(x) = Pl\left((-\infty, x]\right), \tag{12b}$$

for all $x \in \mathbb{R}$, where Bel and Pl are the belief and plausibility functions associated to $[U, V]$. The following equalities hold,

$$F_*(x) = \mathbb{P}([U, V] \subseteq (-\infty, x]) = \mathbb{P}(V \leq x) = F_V(x), \tag{13}$$

where F_V is the cdf of V, and

$$F^*(x) = \mathbb{P}([U, V] \cap (-\infty, x] \neq \emptyset) = \mathbb{P}(U \leq x) = F_U(x). \tag{14}$$

2.3 Dempster's rule

Assume that we have n sources $(S_i, \mathcal{A}_i, \mathbb{P}_i, \Gamma_i)$ for $i = 1, \ldots, n$, where each Γ_i is a multi-valued mapping from S_i to 2^Ω. Then, the combined source $(S, \mathcal{A}, \mathbb{P}, \Gamma)$ can be defined as follows [12]:

$$S = S_1 \times S_2 \ldots \times S_n, \tag{15a}$$
$$\mathcal{A} = \mathcal{A}_1 \otimes \mathcal{A}_2 \ldots \otimes \mathcal{A}_n, \tag{15b}$$
$$\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2 \ldots \otimes \mathbb{P}_n, \tag{15c}$$
$$\Gamma_\cap(s) = \Gamma_1(s_1) \cap \Gamma_2(s_2) \cap \ldots \cap \Gamma_n(s_n), \tag{15d}$$

where $\mathcal{A}$ is the tensor product σ-algebra on the product space S, and $\mathbb{P}$ is the product measure. The belief function Bel induced by the source $(S, \mathcal{A}, \mathbb{P}, \Gamma_\cap)$ can then be written as $Bel_1 \oplus \ldots \oplus Bel_n$, where Bel_i is the belief function

induced by source i. For each $B \in \mathcal{B}$, $Bel(B)$ is the conditional probability that $\Gamma_\cap(s) \subseteq B$, given that $\Gamma_\cap(s) \neq \emptyset$,

$$Bel(B) = \frac{\mathbb{P}\left(\{s \in S | \Gamma_\cap(s) \neq \emptyset, \Gamma_\cap(s) \subseteq B\}\right)}{\mathbb{P}(\{s \in S | \Gamma_\cap(s) \neq \emptyset\})}, \tag{16}$$

which is well defined iff the denominator is non-null (i.e., if the n belief functions are not totally conflicting). The consideration of the product probability measure in (15c) corresponds to an assumption of independence between the items of evidence.

When $\Omega = \mathbb{R}^d$, the combined belief values $Bel(B)$ usually cannot be computed analytically, even when the individual belief functions Bel_i have one of the simple forms outlined in Section 2.2. However, they can easily be approximated by Monte Carlo simulation. The method, described in Algorithm 1, is to draw $(s_1, \ldots, s_n)$ from the product probability measure $\mathbb{P}$ and to compute the intersection of the sets $\Gamma(s_k)$ for $k = 1, \ldots, n$. If this intersection is non-empty, we keep it as a focal set of the combined belief function $Bel = Bel_1 \oplus \ldots \oplus Bel_n$. This process is repeated until we get N focal sets $B_1, \ldots, B_N$. These focal sets with probability masses $1/N$ constitute a belief function $\widehat{Bel}$ that approximates Bel. In particular, degrees of belief $Bel(B)$ and degrees of plausibility $Pl(B)$ can be approximated by $\widehat{Bel}(B)$ and $\widehat{Pl}(B)$ defined as follows,

$$\widehat{Bel}(B) = \frac{1}{N} \#\{i \in \{1, \ldots, N\} | B_i \subseteq B\}, \tag{17a}$$

$$\widehat{Pl}(B) = \frac{1}{N} \#\{i \in \{1, \ldots, N\} | B_i \cap B \neq \emptyset\}. \tag{17b}$$

If the degree of conflict $\mathbb{P}(\{s \in S | \Gamma_\cap(s) = \emptyset\})$ between the n belief functions is high, then Algorithm 1 will be slow, because the condition $\bigcap_{k=1}^n \Gamma_k(s_k) \neq \emptyset$ will often not be met, and a large number of draws will be needed to get N focal sets. Moral and Wilson have proposed Markov chain [50] and importance sampling [51] algorithms to approximate the combination of conflicting belief functions on a finite space more efficiently. We are not aware, however, of any extension of these methods to the case where Ω is infinite. In any case, conflict will not be an issue in paper, as the belief functions to be combined as part of the prediction method described in Section 4 have no conflict.

2.4 Least-commitment principle

In many cases, a belief function is underdetermined by some constraints. For instance, an expert might provide only the contour function pl. Usually,

Algorithm 1 Monte Carlo algorithm for Dempster's rule.

Require: Desired number of focal sets N
  $i \leftarrow 0$
  while $i < N$ **do**
    for $k = 1$ **to** n **do**
      Draw s_k from $\mathbb{P}_k$
    end for
    if $\bigcap_{k=1}^{n} \Gamma_k(s_k) \neq \emptyset$ **then**
      $i \leftarrow i + 1$
      $B_i \leftarrow \bigcap_{k=1}^{n} \Gamma_k(s_k)$
    end if
  end while

infinitely many belief functions are compatible with such constraints. The Least-Commitment Principle (LCP) [25, 61], or principle of maximal uncertainty [38], then prescribes to select the least informative (committed) one, if it exists. To make this principle operational, we need ways to compare the information contents of belief functions. Several partial informational orderings have been defined (see, e.g., [17,24,37,68]). For instance, Bel_1 can be considered to be less committed than Bel_2 if it assigns smaller degrees of belief to every statement, i.e., if $Bel_1 \leq Bel_2$. Alternatively, Bel_1 is said to be *Q-less committed* than Bel_2 if $Q_1 \geq Q_2$, where Q_1 and Q_2 are the commonality functions associated to Bel_1 and Bel_2, respectively. As shown in [24], these two notions are not equivalent.

2.5 Lower and upper expectations

Let Bel be a belief function on $(\Omega, \mathcal{B})$ induced by a source $(S, \mathcal{A}, \mathbb{P}, \Gamma)$, and let $\mathcal{P}(Bel)$ denote the set of probability measures P on $(\Omega, \mathcal{B})$ such that $Bel(B) \leq P(B) \leq Pl(B)$, for all $B \in \mathcal{B}$. For any measurable function X from Ω to $\mathbb{R}$, its lower and upper expectations with respect to Bel are defined as follows,

$$\mathbb{E}_*(X) = \inf_{P \in \mathcal{P}(Bel)} \mathbb{E}_P(X) \quad \text{and} \quad \mathbb{E}^*(X) = \sup_{P \in \mathcal{P}(Bel)} \mathbb{E}_P(X), \tag{18}$$

where $\mathbb{E}_P(\cdot)$ denotes the expectation with respect to P. It can be shown [65] that

$$\mathbb{E}_*(X) = \int_S \underline{X}(s)\mathbb{P}(ds) \quad \text{and} \quad \mathbb{E}^*(X) = \int_S \overline{X}(s)\mathbb{P}(ds), \tag{19}$$

where

$$\underline{X}(s) = \inf_{\omega \in \Gamma(s)} X(\omega) \quad \text{and} \quad \overline{X}(s) = \sup_{\omega \in \Gamma(s)} X(\omega). \tag{20}$$

In the special case where $\Omega = \mathbb{R}$ and Bel is induced by a random interval $[U, V]$, the lower and upper expectations of a non-decreasing function $X : \Omega \to \mathbb{R}$ are thus simply its expectation with respect to U and V,

$$\mathbb{E}_*(X) = \mathbb{E}_U(X) \quad \text{and} \quad \mathbb{E}^*(X) = \mathbb{E}_V(X). \tag{21}$$

As shown in [31], the lower and upper expectations are the Choquet integrals with respect to Bel and Pl, respectively. A Savage-like axiomatic justification of decision-making based on maximization of Choquet-expected utilities has been provided by Gilboa [30].

3 Estimation using belief functions

The definition of a belief function from the likelihood function will first be recalled in Section 3.1. Some connections with classical methods of inference will then be outlined in Section 3.2, and the consistency of the method will be studied in Section 3.3.

3.1 Likelihood-based belief function

Let $\boldsymbol{y} \in \mathbb{Y}$ denote the observed data, assumed to be a realization of a random vector $\boldsymbol{Y}$ with probability mass or density function $f_{\boldsymbol{\theta}}(\boldsymbol{y})$, where $\boldsymbol{\theta} \in \Theta$ is an unknown parameter. The likelihood function is a mapping $L_{\boldsymbol{y}}$ from Θ to $[0, +\infty)$ defined by

$$L_{\boldsymbol{y}}(\boldsymbol{\theta}) = c f_{\boldsymbol{\theta}}(\boldsymbol{y}), \tag{22}$$

where $c > 0$ is an arbitrary multiplicative constant. Several authors have defended the view that the likelihood function contains all the information about the parameter provided by the experiment, a thesis called the *Likelihood Principle* (LP) [5, 8, 9, 27]. In particular, this principle was shown by Birnbaum in [9] to follow from two principles generally accepted by most (but not all) statisticians: the conditionality principle (see also [8, page 25]) and the sufficiency principle. As likelihood is defined up to a multiplicative constant, it can conveniently be rescaled to the interval $[0, 1]$, by the transformation

$$R_{\boldsymbol{y}}(\boldsymbol{\theta}) = \frac{L_{\boldsymbol{y}}(\boldsymbol{\theta})}{L_{\boldsymbol{y}}(\widehat{\boldsymbol{\theta}})}, \tag{23}$$

where $\widehat{\boldsymbol{\theta}}$ is a maximizer of $L_{\boldsymbol{y}}(\boldsymbol{\theta})$, i.e., a maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$, and it is assumed that $L_{\boldsymbol{y}}(\widehat{\boldsymbol{\theta}}) < +\infty$. Plots of function $R_{\boldsymbol{y}}(\boldsymbol{\theta})$, called the relative likelihood, provide a complete graphical description of the result of a random experiment [63]. Likelihood regions, defined as the set of parameter values $\boldsymbol{\theta}$ whose relative likelihood exceeds some threshold,

$$^{s}R_{\boldsymbol{y}} = \{\boldsymbol{\theta} \in \Theta | R_{\boldsymbol{y}}(\boldsymbol{\theta}) \geq s\}, \tag{24}$$

for $s \in [0, 1]$, are useful summaries of function $R_{\boldsymbol{y}}$.

In [58], Shafer proposed to interpret $R_{\boldsymbol{y}}(\boldsymbol{\theta})$ as the plausibility of $\boldsymbol{\theta}$ after observing $\boldsymbol{y}$. The relative likelihood is then considered as the contour function $pl_{\boldsymbol{y}}(\boldsymbol{\theta})$ of a belief function $Bel_{\boldsymbol{y}}^{\Theta}$ on Θ:

$$pl_{\boldsymbol{y}}(\boldsymbol{\theta}) = R_{\boldsymbol{y}}(\boldsymbol{\theta}), \tag{25}$$

for all $\boldsymbol{\theta} \in \Theta$. If one further assumes $Bel_{\boldsymbol{y}}^{\Theta}$ to be consonant, then the plausibility of any hypothesis $H \subseteq \Theta$ is given by (8) as

$$Pl_{\boldsymbol{y}}^{\Theta}(H) = \sup_{\boldsymbol{\theta} \in H} pl_{\boldsymbol{y}}(\boldsymbol{\theta}). \tag{26}$$

As explained in Section 2.2, $Bel_{\boldsymbol{y}}^{\Theta}$ is then induced by the source $(S, \mathcal{A}, \lambda, \Gamma_{\boldsymbol{y}})$, where $S = [0, 1]$, $\mathcal{A}$ is the Borel sigma-field on S, λ is the uniform probability measure on $[0, 1]$ and $\Gamma_{\boldsymbol{y}}(s) = {}^{s}R_{\boldsymbol{y}}$ for all $s \in S$.

The so-called likelihood-based approach to belief function-based inference was introduced by Shafer on intuitive grounds. It was recently shown in [18] to be the only belief function $Bel_{\boldsymbol{y}}^{\Theta}$ on Θ verifying the following three requirements:

1. Likelihood principle: $Bel_{\boldsymbol{y}}^{\Theta}$ should only depend on the likelihood function $L_{\boldsymbol{y}}(\boldsymbol{\theta})$.

2. Compatibility with Bayesian inference: if a Bayesian prior $g(\theta)$ is available, combining it with $Bel_{\boldsymbol{y}}^{\Theta}$ using Dempster's rule (see Section 2.3) should yield the Bayesian posterior.

3. Least Commitment Principle (see Section 2.4): $Bel_{\boldsymbol{y}}^{\Theta}$ should be the least committed belief function (according to the Q-ordering), among all those satisfying the previous two requirements.

These principles are discussed at length in [18] and in the subsequent discussion [19, 22, 49]. They can be considered to provide a firm theoretical basis for the likelihood-based belief function approach.

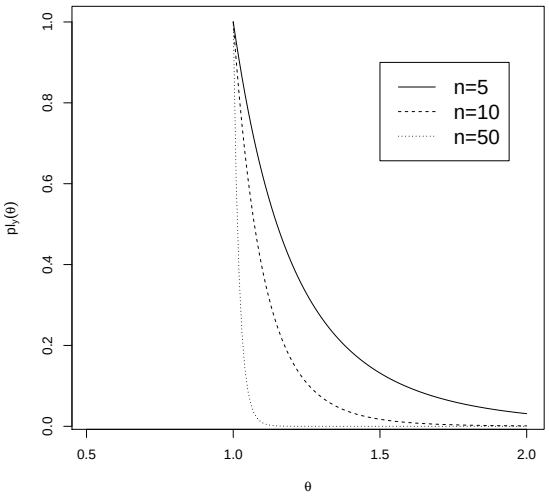


Figure 1: Contour function on θ for an iid sample from $\mathcal{U}([0,\theta])$, with $y_{(n)} = 1$ and $n \in \{5, 10, 50\}$.

Example 1 *Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a realization from an iid random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ from the uniform distribution $\mathcal{U}([0,\theta])$, where $\theta \in \Theta = [0, +\infty)$ is the unknown parameter. The likelihood function is*

$$L_{\boldsymbol{y}}(\theta) = \theta^{-n} \mathbb{1}_{[y_{(n)}, +\infty)}(\theta), \tag{27}$$

where $y_{(n)} = \max_{1 \leq i \leq n} y_i$, and the contour function is

$$pl_{\boldsymbol{y}}(\theta) = \left(\frac{y_{(n)}}{\theta}\right)^n \mathbb{1}_{[y_{(n)}, +\infty)}(\theta). \tag{28}$$

It is plotted in Figure 1 for $y_{(n)} = 1$ and $n \in \{5, 10, 50\}$. We note that, the contour function being unimodal and upper-semicontinous, the focal sets $\Gamma_{\boldsymbol{y}}(s)$ are close intervals $[\widehat{\theta}_{\boldsymbol{y}*}(s), \widehat{\theta}^*_{\boldsymbol{y}}(s)]$, with $\widehat{\theta}_{\boldsymbol{y}*}(s) = y_{(n)}$ and $\widehat{\theta}^*_{\boldsymbol{y}}(s) = y_{(n)} s^{-1/n}$ for all $s \in [0, 1]$. Consequently, the belief function $Bel^{\Theta}_{\boldsymbol{y}}$ is induced by the random closed interval $[y_{(n)}, y_{(n)} S^{-1/n}]$, with $S \sim \mathcal{U}([0, 1])$. □

3.2 Connections with classical statistical concepts

Likelihood-based inference The approach to statistical inference outlined in the previous section is very close to the "likelihoodist" approach

advocated by Birnbaum [9], Barnard [5], and Edwards [27], among others. The main difference resides in the interpretation of the likelihood function as defining a belief function. This interpretation allows us to quantify the uncertainty in statements of the form $\boldsymbol{\theta} \in H$, where H may contain multiple values. This is in contrast with the classical likelihood approach, in which only the likelihood of single hypotheses is defined. The belief function interpretation provides an easy and natural way to combine statistical information with expert opinions (see, e.g., [6]). It will also allow us to provide an original solution to the prediction problem, as will be shown in Section 4.

Frequentists tests and confidence regions We can also notice that $Pl_{\boldsymbol{y}}^{\Theta}(H)$ given by (26) is identical to the likelihood ratio statistic for H. From Wilk's theorem [66], we know that, under regularity conditions, the large sample distribution of $-2\ln Pl_{\boldsymbol{y}}(H)$, when H holds, is chi-squared, with degrees of freedom equal to the number r of restrictions imposed by H. Consequently, rejecting hypothesis H if its plausibility is smaller than $\exp(-\chi^2_{r;1-\alpha}/2)$, where $\chi^2_{r;1-\alpha}$ is the $1-\alpha$-quantile of the chi-square distribution with r degrees of freedom, is a testing procedure with significance level approximately equal to α. Another consequence is that the likelihood (or plausibility) regions (24) are approximate confidence regions [34]. Recently, Martin [43] proposed to define the plausibility of any hypothesis H not as (26), but as

$$Pl_{\boldsymbol{y}}^{\Theta}(H) = \sup_{\boldsymbol{\theta} \in H} F_{\boldsymbol{\theta}}(R_{\boldsymbol{y}}(\boldsymbol{\theta})), \tag{29}$$

where $F_{\boldsymbol{\theta}}$ is the cdf of $R_{\boldsymbol{Y}}(\boldsymbol{\theta})$ when $\boldsymbol{Y} \sim f_{\boldsymbol{\theta}}$. In this way, rejecting H when $Pl_{\boldsymbol{y}}^{\Theta}(H) \leq \alpha$ is an exact testing procedure with size α, and exact confidence regions can be constructed. However, this estimation method is no longer compatible with Bayesian inference, i.e., combining $Pl_{\boldsymbol{y}}^{\Theta}$ defined by (29) with a prior using Dempster's rule does not yield the Bayesian posterior. Imposing this condition, as done in this paper, thus rules out (29) as a valid definition for the plausibility of a hypothesis.

Profile likelihood Assume that $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\nu})$, where $\boldsymbol{\xi}$ is a (vector) parameter of interest and $\boldsymbol{\nu}$ is a nuisance parameter. Then, the marginal contour function for $\boldsymbol{\xi}$ is

$$pl_{\boldsymbol{y}}(\boldsymbol{\xi}) = \sup_{\boldsymbol{\nu}} pl_{\boldsymbol{y}}(\boldsymbol{\xi}, \boldsymbol{\nu}), \tag{30}$$

which is the profile relative likelihood function. The profiling method for eliminating nuisance parameter thus has a natural justification in our ap-

proach. When the quantities $pl_{\boldsymbol{y}}(\boldsymbol{\xi})$ cannot be derived analytically, they have to be computed numerically using an iterative optimization algorithm.

Rejection sampling The likelihood-based method described here does not require any prior knowledge of $\boldsymbol{\theta}$. However, by construction, this approach boils down to Bayesian inference if a prior probability $g(\boldsymbol{\theta})$ is provided and combined with $Bel_{\boldsymbol{y}}^{\Theta}$ by Dempster's rule. As it will usually not be possible to compute the analytical expression of the resulting posterior distribution, it can be approximated by Monte Carlo simulation, using Algorithm 1. The algorithm generates a sample $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ from the posterior distribution $g(\boldsymbol{\theta}|\boldsymbol{y})$. The particular form of this algorithm when the likelihood-based belief function is combined with a Bayesian prior is described as Algorithm 2. We can see that this is just the rejection sampling algorithm with the prior $g(\theta)$ as proposal distribution. The rejection sampling algorithm can thus be seen, in this case, as a Monte Carlo approximation to Dempster's rule of combination.

Algorithm 2 Monte Carlo algorithm for combining the likelihood-based belief function with a Bayesian prior by Dempster's rule.

Require: Desired number of focal sets N
  $i \leftarrow 0$
  while $i < N$ **do**
    Draw s in $[0, 1]$ from the uniform probability measure λ on $[0, 1]$
    Draw $\boldsymbol{\theta}$ from the prior probability distribution $g(\boldsymbol{\theta})$
    if $pl_{\boldsymbol{y}}(\boldsymbol{\theta}) \geq s$ **then**
      $i \leftarrow i + 1$
      $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}$
    end if
  end while

3.3 Consistency

In this section, we assume that the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a realization of an iid sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ from $Y \sim f_{\boldsymbol{\theta}}(y)$. In such a situation, it is generally required from any statistical procedure that it precisely identifies the true value $\boldsymbol{\theta}_0$ of parameter $\boldsymbol{\theta}$ in the limit, when n tends to infinity. Since, in our case, the result of the estimation is given in the form of a belief function on the parameter space Θ, this consistency property has to be given a precise definition. In Bayesian statistics, a posterior distribution μ_n is said to be consistent at θ_0 if, for every neighborhood N of θ_0, $\mu_n(N) \to 1$ almost surely under the law determined by θ_0. As we shall see, a similar property holds, under mild conditions, for the likelihood-based belief function defined in Section 3.1.

In the following, to emphasize the dependency on the sample size n, we will index by n all the quantities depending on $\boldsymbol{y}$. For instance, the likelihood and plausibility contour functions will be denoted, respectively, as $L_n(\boldsymbol{\theta})$ and $pl_n(\boldsymbol{\theta})$. The following theorem states that the plausibility of any value of $\boldsymbol{\theta}$ different from the true value $\boldsymbol{\theta}_0$ tends to 0 as the sample size tends to infinity. The simple proof given here follows closely that of Fraser [29, page 298]. We reproduce it for completeness.

Theorem 2 *If $\mathbb{E}_{\boldsymbol{\theta}_0}[\log f_{\boldsymbol{\theta}}(Y)]$ exists, is finite for all $\boldsymbol{\theta}$, and has a unique maximum at $\boldsymbol{\theta}_0$, then, for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, $pl_n(\boldsymbol{\theta}) \to 0$ almost surely under the law determined by $\boldsymbol{\theta}_0$.*

Proof. As $\mathbb{E}_{\boldsymbol{\theta}_0}[\log f_{\boldsymbol{\theta}}(Y)]$ has a unique maximum at $\boldsymbol{\theta}_0$, $\mathbb{E}_{\boldsymbol{\theta}_0}[\log f_{\boldsymbol{\theta}}(Y)] < \mathbb{E}_{\boldsymbol{\theta}_0}[\log f_{\boldsymbol{\theta}_0}(Y)]$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ or, equivalently,

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\log f_{\boldsymbol{\theta}}(Y) - \log f_{\boldsymbol{\theta}_0}(Y)] = \varepsilon < 0. \tag{31}$$

Hence, by the strong law of large numbers,

$$\mathbb{P}_{\boldsymbol{\theta}_0}\left[\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} [\log f_{\boldsymbol{\theta}}(Y_i) - \log f_{\boldsymbol{\theta}_0}(Y_i)] = \varepsilon\right] = 1, \tag{32}$$

Now,

$$\log \frac{L_n(\boldsymbol{\theta})}{L_n(\boldsymbol{\theta}_0)} = \sum_{i=1}^{n} [\log f_{\boldsymbol{\theta}}(Y_i) - \log f_{\boldsymbol{\theta}_0}(Y_i)], \tag{33}$$

so (32) can be written as

$$\mathbb{P}_{\boldsymbol{\theta}_0}\left[\lim_{n\to\infty} \frac{1}{n} \log \frac{L_n(\boldsymbol{\theta})}{L_n(\boldsymbol{\theta}_0)} = \varepsilon\right] = 1, \tag{34}$$

which implies that

$$\mathbb{P}_{\boldsymbol{\theta}_0}\left[\lim_{n\to\infty} \log \frac{L_n(\boldsymbol{\theta})}{L_n(\boldsymbol{\theta}_0)} = -\infty\right] = 1, \tag{35}$$

or, equivalently,

$$\mathbb{P}_{\boldsymbol{\theta}_0}\left[\lim_{n\to\infty} \frac{L_n(\boldsymbol{\theta})}{L_n(\boldsymbol{\theta}_0)} = 0\right] = 1. \tag{36}$$

Finally, $L_n(\widehat{\boldsymbol{\theta}}) \geq L_n(\boldsymbol{\theta}_0)$, hence

$$pl_n(\boldsymbol{\theta}) = \frac{L_n(\boldsymbol{\theta})}{L_n(\widehat{\boldsymbol{\theta}})} \leq \frac{L_n(\boldsymbol{\theta})}{L_n(\boldsymbol{\theta}_0)}, \tag{37}$$

from which we can deduce that $pl_n(\boldsymbol{\theta}) \to 0$ almost surely. □

From the consistency of the MLE, it might be expected that $pl_n(\boldsymbol{\theta}_0) \to 1$ almost surely. However, this is not the case in general. A recalled in Section 3.2, $-2\log pl_n(\boldsymbol{\theta}_0)$ converges in distribution to a chi square distribution with 1 degree of freedom, hence $pl_n(\boldsymbol{\theta}_0)$ does not converge to 1. However, it can be shown that, under mild conditions, the belief and plausibility functions become more and more concentrated around $\boldsymbol{\theta}_0$ when the sample size tends to infinity. This is a consequence of the following theorem, which follows directly a result proved by Fraser [29, page 301].

In the following, we assume that $\Theta = \mathbb{R}^d$ and we denote by $B_\rho(\boldsymbol{\theta})$ the ball of radius ρ about $\boldsymbol{\theta}$,

$$B_\rho(\boldsymbol{\theta}) = \{\boldsymbol{\theta}' \in \Theta : \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \rho\}. \tag{38}$$

We also denote by $B_\rho(\infty)$ a ball about ∞, defined as

$$B_\rho(\infty) = \{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta} - \mathbf{0}\| > 1/\rho\}. \tag{39}$$

Theorem 3 *If the following assumptions hold,*

1. *$f_{\boldsymbol{\theta}}(\boldsymbol{y})$ is a continuous function of $\boldsymbol{\theta}$ in $\mathbb{R}^d \cup \{\infty\}$;*
2. *For each $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, the distribution $f_{\boldsymbol{\theta}}(\boldsymbol{y})$ is different from $f_{\boldsymbol{\theta}_0}(\boldsymbol{y})$;*
3. *For each $\boldsymbol{\theta}'$ in $\mathbb{R}^d \cup \{\infty\}$, there is a neighborhood $B_\rho(\boldsymbol{\theta}')$ such that*

$$\sup_{\boldsymbol{\theta} \in B_\rho(\boldsymbol{\theta}')} \log f_{\boldsymbol{\theta}}(y) \leq M_{\boldsymbol{\theta}'}(y), \tag{40}$$

where $M_{\boldsymbol{\theta}'}(y)$ is a bounding function with finite mean value $\mathbb{E}_{\boldsymbol{\theta}_0}[M_{\boldsymbol{\theta}'}(Y)]$, and the expectation $\mathbb{E}_{\boldsymbol{\theta}_0}[\log f_{\boldsymbol{\theta}'}(Y)]$ is finite;

then, for any $\delta > 0$,

$$\mathbb{P}_{\boldsymbol{\theta}_0}\left[\lim_{n\to\infty} Pl_n^{\Theta}(\overline{B_\delta(\boldsymbol{\theta}_0)}) = 0\right] = 1. \tag{41}$$

Proof. Under the assumptions of the theorem, Fraser [29, page 301] shows that

$$\mathbb{P}_{\boldsymbol{\theta}_0}\left[\lim_{n\to\infty} \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\geq\delta} \log\frac{L_n(\boldsymbol{\theta})}{L_n(\boldsymbol{\theta}_0)} = -\infty\right] = 1. \tag{42}$$

The theorem follows directly from (37) and (42). □

As an immediate corollary, the belief and the plausibility of any neighborhood of $\boldsymbol{\theta}_0$ tends to 1 almost surely.

Corollary 1 *Under the assumptions of Theorem 3, for any neighborhood* N *of* $\boldsymbol{\theta}_0$, $Bel_n^{\Theta}(N) \to 1$ *and* $Pl_n^{\Theta}(N) \to 1$ *almost surely under the law determined by* $\boldsymbol{\theta}_0$.

Proof. For any $\delta > 0$, the following equality holds:

$$Bel_n^{\Theta}(B_\delta(\boldsymbol{\theta}_0)) = 1 - Pl_n^{\Theta}(\overline{B_\delta(\boldsymbol{\theta}_0)}). \tag{43}$$

Consequently, $Bel_n^{\Theta}(B_\delta(\boldsymbol{\theta}_0)) \to 1$ almost surely. For any neighborhood N of $\boldsymbol{\theta}_0$, there is some $\delta > 0$ such that $B_\delta(\boldsymbol{\theta}_0) \subseteq N$, so $Bel_n^{\Theta}(N) \geq Bel_n^{\Theta}(B_\delta(\boldsymbol{\theta}_0))$ and $Bel_n^{\Theta}(N) \to 1$ almost surely. As $Pl_n^{\Theta}(N) \geq Bel_n^{\Theta}(N)$, it also holds that $Pl_n^{\Theta}(N) \to 1$ almost surely. □

4 Prediction

The prediction method introduced in [36] will first be recalled in Section 4.1, and its consistency in the case of iid data will be established in Section 4.2. Practical calculation of the predictive belief function using Monte Carlo will then be addressed in Section 4.3. Finally, the method will be extended to the prediction of multidimensional quantities in Section 4.4.

4.1 Basic method

As we have seen in Section 3, the estimation problem is to make statements about some parameter $\boldsymbol{\theta}$ after observing some data $\boldsymbol{y}$ with distribution $f_{\boldsymbol{\theta}}(\boldsymbol{y})$. The *prediction* problem considered in this section is, in some sense, the inverse of the previous one: given some knowledge about $\boldsymbol{\theta}$ obtained by

observing $\boldsymbol{y}$ (represented here by a belief function), we wish to make statements about some future data $\boldsymbol{Z} \in \mathbb{Z}$, whose conditional distribution $f_{\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{z})$ given $\boldsymbol{y}$ depends on $\boldsymbol{\theta}$. In some cases, $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a vector composed of the n first observations of an i.i.d. sample, and $\boldsymbol{Z} = (Y_{n+1}, \ldots, Y_{n+m})$ is a vector containing m observations to be drawn independently from the same distribution. However, the model used here is more general. For instance, $\boldsymbol{y} = (y_0, y_1, \ldots, y_T)$ might be a time series and $\boldsymbol{Z} = (Z_{T+1}, \ldots, Z_{T+h})$ might represent h future values to be predicted. Vectors $\boldsymbol{y}$ and $\boldsymbol{Z}$ may also depend on some covariates, as in the regression model considered in Section 5.

To describe our prediction method, let us consider the case where the unobserved data Z is one-dimensional[2]. The multi-dimensional case will be addressed in Section 4.4. The main idea is to write Z, for fixed $\boldsymbol{y}$, as function of $\boldsymbol{\theta}$ and some pivotal variable W, whose distribution does not depend on $\boldsymbol{\theta}$,

$$Z = \varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, W). \tag{44}$$

Hereafter, such an equation will be called a φ-equation. In practice, function $\varphi_{\boldsymbol{y}}$ can be constructed canonically as follows. Let us first assume that Z is a continuous r.v. Let $F_{\boldsymbol{\theta},\boldsymbol{y}}(z)$ be its conditional cdf given $\boldsymbol{y}$. We know that $W = F_{\boldsymbol{\theta},\boldsymbol{y}}(Z)$ has a standard uniform distribution, and one can write Z as a function of $\boldsymbol{\theta}$ and W as

$$Z = F^{-1}_{\boldsymbol{\theta},\boldsymbol{y}}(W), \tag{45}$$

with $W \sim \mathcal{U}([0,1])$, which has the same form as (44). When W is discrete, (45) is still valid if $F^{-1}_{\boldsymbol{\theta},\boldsymbol{y}}$ now denotes the generalized inverse of $F_{\boldsymbol{\theta},\boldsymbol{y}}$,

$$F^{-1}_{\boldsymbol{\theta},\boldsymbol{y}}(W) = \inf\{z | F_{\boldsymbol{\theta},\boldsymbol{y}}(z) \geq W\}. \tag{46}$$

Example 2 *Assume that Z has a continuous uniform distribution on $[0, \theta]$ and is independent from $\boldsymbol{Y}$. Then $F_\theta(z) = z/\theta$ for all $0 \leq z \leq \theta$ and we can write $Z = \theta W$ with $W \sim \mathcal{U}([0,1])$.* □

Example 3 *Let Z be a normal r.v. with mean μ and standard deviation σ. Let $\boldsymbol{\theta} = (\mu, \sigma)$. Then*

$$F_\theta(Z) = \Phi\left(\frac{Z - \mu}{\sigma}\right) = W \sim \mathcal{U}([0,1]),$$

from which we get

$$Z = \mu + \sigma\Phi^{-1}(W).$$

□

[2]We use normal fonts for scalars, and bold fonts for vectors.

For fixed $\boldsymbol{y}$, Equation (44) describes a relation between z, $\boldsymbol{\theta}$ and an auxiliary variable W with standard uniform distribution. Dempster [12, 13, 15] used such an equation to construct a belief function on $\boldsymbol{\theta}$ after observing $Z = z$. Here, we will use it to construct a belief function on Z, given the belief function on $\boldsymbol{\theta}$ induced by the likelihood function $L_{\boldsymbol{y}}(\boldsymbol{\theta})$. For that purpose, we can notice that Equation (44) defines a multi-valued mapping

$$\Gamma'_{\boldsymbol{y}} : w \to \Gamma'_{\boldsymbol{y}}(w) = \{(z, \boldsymbol{\theta}) \in \mathbb{Z} \times \Theta | z = \varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, w)\}, \tag{47}$$

where $\mathbb{Z}$ is the sample space of z. The source $([0,1], \mathcal{B}([0,1]), \lambda, \Gamma'_{\boldsymbol{y}})$, where $\mathcal{B}([0,1])$ is the Borel sigma-field on $[0,1]$, defines a joint belief function $Bel_{\boldsymbol{y}}^{\mathbb{Z}\times\Theta}$ on $\mathbb{Z} \times \Theta$.

We now have two belief functions, $Bel_{\boldsymbol{y}}^{\Theta}$ and $Bel_{\boldsymbol{y}}^{\mathbb{Z}\times\Theta}$, induced by multi-valued mapping $s \to \Gamma_{\boldsymbol{y}}(s)$ and $w \to \Gamma'_{\boldsymbol{y}}(w)$. Given $\boldsymbol{y}$, the random variables S and W are independent: for instance, if we know that $S = s$, i.e., $\boldsymbol{\theta} \in \Gamma_{\boldsymbol{y}}(s)$, this information influences our beliefs about Z (because Z depends on $\boldsymbol{\theta}$), but it does not influence our beliefs about W, which continues to have a standard uniform distribution. Because of this independence property, the two belief functions $Bel_{\boldsymbol{y}}^{\Theta}$ and $Bel_{\boldsymbol{y}}^{\mathbb{Z}\times\Theta}$ can be combined by Dempster's rule. After marginalizing on $\mathbb{Z}$, we then get a predictive belief function $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ on $\mathbb{Z}$. The focal sets of the combined belief function are obtained by taking the intersections

$$(\mathbb{Z} \times \Gamma_{\boldsymbol{y}}(s)) \cap \Gamma'_{\boldsymbol{y}}(w) = \{(z, \boldsymbol{\theta}) \in \mathbb{Z} \times \Theta | z = \varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, w), \boldsymbol{\theta} \in \Gamma_{\boldsymbol{y}}(s)\}. \tag{48}$$

Projecting these sets on $\mathbb{Z}$, we get

$$\{z \in \mathbb{Z} | z = \varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, w), \boldsymbol{\theta} \in \Gamma_{\boldsymbol{y}}(s)\} = \varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), w). \tag{49}$$

Let $\Gamma''_{\boldsymbol{y}}$ denote the multi-valued mapping

$$(s, w) \to \varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), w). \tag{50}$$

The predictive belief function $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ is thus induced by the source

$$([0,1]^2, \mathcal{B}([0,1]^2), \lambda_2, \Gamma''_{\boldsymbol{y}}), \tag{51}$$

where λ_2 is the uniform probability measure in $[0,1]^2$. We then have

$$Bel_{\boldsymbol{y}}^{\mathbb{Z}}(A) = \lambda_2\left(\{(s,w) \in [0,1]^2 | \varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), w) \subseteq A\}\right), \tag{52a}$$

$$Pl_{\boldsymbol{y}}^{\mathbb{Z}}(A) = \lambda_2\left(\{(s,w) \in [0,1]^2 | \varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), w) \cap A \neq \emptyset\}\right), \tag{52b}$$

for any subset A of $\mathbb{Z}$ for which the above expressions are well-defined.

Example 4 *Continuing Examples 1 and 2, assume that $Y_1, \ldots, Y_n, Z$ is iid from $\mathcal{U}([0,\theta])$. We have seen that the belief function $Bel_{\boldsymbol{y}}^{\Theta}$ after observing $\boldsymbol{Y} = \boldsymbol{y}$ is induced by the random interval $[y_{(n)}, y_{(n)}S^{-1/n}]$. As function $\varphi(\theta, W) = \theta W$ is continuous in θ, each focal set of $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ is an interval*

$$\varphi(\Gamma_{\boldsymbol{y}}(s), w) = [y_{(n)}w, y_{(n)}s^{-1/n}w], \tag{53}$$

so that $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ is induced by the random interval

$$[\widehat{Z}_{\boldsymbol{y}*}, \widehat{Z}_{\boldsymbol{y}}^{*}] = [y_{(n)}W, y_{(n)}S^{-1/n}W]. \tag{54}$$

From (13)-(14), the upper and lower cdfs of $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ are, respectively, the cdfs of $\widehat{Z}_{\boldsymbol{y}*}$ and $\widehat{Z}_{\boldsymbol{y}}^{*}$. As $\widehat{Z}_{\boldsymbol{y}*} \sim \mathcal{U}([0, y_{(n)}])$, we have

$$F^*(z) = Pl_{\boldsymbol{y}}^{\mathbb{Z}}((-\infty, z]) = \mathbb{P}(\widehat{Z}_{\boldsymbol{y}*} \leq z) = \begin{cases} 0 & \text{if } z \leq 0, \\ z/y_{(n)} & \text{if } 0 < z \leq y_{(n)}, \\ 1 & \text{if } z > y_{(n)}, \end{cases} \tag{55}$$

and

$$\begin{aligned} F_*(z) &= Bel_{\boldsymbol{y}}^{\mathbb{Z}}((-\infty, z]) = \mathbb{P}(\widehat{Z}_{\boldsymbol{y}*} \leq z) && \text{(56a)} \\ &= \int_0^1 \mathbb{P}(y_{(n)}S^{-1/n}W \leq z | W = w)dw && \text{(56b)} \\ &= \int_0^1 \mathbb{P}(S \geq (wy_{(n)}/\boldsymbol{z})^n)dw && \text{(56c)} \\ &= \begin{cases} 0 & \text{if } z \leq 0, \\ \frac{nz}{(n+1)y_{(n)}} & \text{if } 0 < z \leq y_{(n)}, \\ 1 - \frac{1}{n+1}\left(\frac{z}{y_{(n)}}\right)^n & \text{if } z > y_{(n)}, \end{cases} && \text{(56d)} \end{aligned}$$

These functions are plotted in Figure 2 for $y_{(n)} = 1$ and $n = 5$.

The lower expectation of Z is $\mathbb{E}_*(Z) = y_{(n)}\mathbb{E}(W) = y_{(n)}/2$, and its upper expectation is $\mathbb{E}^*(Z) = y_{(n)}\mathbb{E}(S^{-1/n})\mathbb{E}(W)$. It is easy to show that $\mathbb{E}(S^{-1/n}) = n/(n-1)$. Hence,

$$\mathbb{E}^*(Z) = \frac{ny_{(n)}}{2(n-1)}. \tag{57}$$

It is interesting to study to the behavior of the predictive random interval (54) when the sample size n tends to infinity. From the consistency of the MLE, $Y_{(n)}$ converges in probability to θ_0, so

$$\widehat{Z}_{\boldsymbol{Y}*} = Y_{(n)}W \xrightarrow{d} \theta_0 W = Z, \tag{58}$$

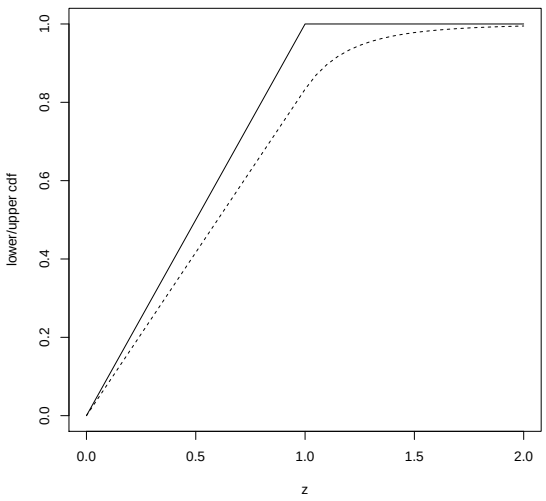


Figure 2: Lower and upper predictive cdf for $y_{(n)} = 1$ and $n = 5$ (Example 4).

where $\xrightarrow{d}$ *denotes convergence in distribution. We have seen that* $\mathbb{E}(S^{-1/n}) = n/(n-1)$*, and*

$$Var(S^{-1/n}) = \frac{n}{(n-2)(n-1)^2}. \tag{59}$$

Consequently, $\mathbb{E}(S^{-1/n}) \to 1$ *and* $Var(S^{-1/n}) \to 0$*, so* $S^{-1/n} \xrightarrow{P} 1$*. Hence,*

$$\widehat{Z}^*_{\boldsymbol{Y}} = Y_{(n)} S^{-1/n} W \xrightarrow{d} \theta_0 W = Z. \tag{60}$$

In this example, the predictive random interval is thus consistent, in the sense that its bounds converge in probability to the true distribution of Z*. In the next section, we will see that this property generally holds under mild conditions.* □

4.2 Consistency

In this section, we will assume that the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a realization of an iid sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Furthermore, we will assume that the likelihood function $L_n(\boldsymbol{\theta})$ is unimodal and upper-semicontinuous, so that its level sets $\Gamma_n(s)$ are closed and connected, and that function $\varphi(\boldsymbol{\theta}, w)$

is continuous. Under these conditions, the random set $\varphi(\Gamma_n(S), W)$ is a closed random interval $[\widehat{Z}_{*n}, \widehat{Z}_n^*]$. We then have the following theorem:

Theorem 4 *Assume that the conditions of Theorem 3 hold, and that the predictive belief function $Bel_n^{\mathbb{Z}}$ is induced by a random closed interval $[\widehat{Z}_{*n}, \widehat{Z}_n^*]$. Then $\widehat{Z}_{*n}$ and $\widehat{Z}_n^*$ both converge in distribution to Z when n tends to infinity.*

Proof. Let $S_1, S_2, \ldots$ be an iid sequence of random variables with a standard uniform distribution, and let $\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2, \ldots$ be a sequence of random variables such that $\widetilde{\boldsymbol{\theta}}_n \in \Gamma_n(S_n)$. According to Theorem 3, for any $\delta > 0$, we have,

$$\lim_{n\to\infty} Bel_n(B_\delta(\theta_0)) = \lim_{n\to\infty} \mathbb{P}(\Gamma(S_n) \subseteq B_\delta(\theta_0)) = 1, \tag{61}$$

almost surely under the law determined by $\boldsymbol{\theta}_0$. Since $\widetilde{\boldsymbol{\theta}}_n \in \Gamma_n(S_n)$, its follows that

$$\lim_{n\to\infty} \mathbb{P}(\|\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq \delta) = 1, \tag{62}$$

i.e., $\widetilde{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}_0$. Now,

$$\widehat{Z}_{*n} = \min \varphi(\Gamma_n(S_n), W) = \varphi(\widehat{\boldsymbol{\theta}}_{*n}, W), \tag{63}$$

for some $\widehat{\boldsymbol{\theta}}_{*n} \in \Gamma_n(S_n)$. As shown above, $\widehat{\boldsymbol{\theta}}_{*n}$ converges in probability to θ_0. By the continuity of φ, it results that $\widehat{Z}_{*n}$ converges in distribution to $\varphi(\widehat{\boldsymbol{\theta}}_{*n}, W) = Z$. The same line of reasoning leads to a similar conclusion for $\widehat{Z}_n^*$. □

4.3 Practical calculation

For most models, the calculations cannot be done analytically as in Example 4, and one often has to approximate the quantities defined in (52) by Monte Carlo simulation. Some computational issues are discussed in this section.

Hereafter, we will assume that each focal set $\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), w)$ to a closed interval $[z_*(s,w), z^*(s,w)]$, in which case the predictive belief function $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ is equivalent to a random closed interval. If this is not the case, due to, e.g., the multi-modality of the likelihood function $L_{\boldsymbol{y}}(\boldsymbol{\theta})$, then we can always represent each focal set $\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), w)$ by its interval hull $[\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), w)]$ (i.e., the smallest enclosing interval). This strategy yields a conservative approximation, in the sense that the approximating belief-plausibility intervals always contain the true ones.

Basically, the general Monte Carlo approach is to draw N pairs (s_i, w_i) independently from a uniform distribution, and to compute (or approximate)

the focal sets $\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s_i), w_i)$. The predictive belief and plausibility of any subset $A \subseteq \mathbb{Z}$ are then estimated by

$$\widehat{Bel}^{\mathbb{Z}}_{\boldsymbol{y}}(A) = \frac{1}{N}\#\{i \in \{1,\ldots,N\} | \varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s_i), w_i) \subseteq A\}, \tag{64a}$$

$$\widehat{Pl}^{\mathbb{Z}}_{\boldsymbol{y}}(A) = \frac{1}{N}\#\{i \in \{1,\ldots,N\} | \varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s_i), w_i) \cap A \neq \emptyset\}. \tag{64b}$$

When each focal set $\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s_i), w_i)$ is a closed interval $[z_*(s_i, w_i), z^*(s_i, w_i)]$, the lower and upper expectations of Z can be estimated, respectively, by the sample means of $z_*(s_i, w_i)$ and $z^*(s_i, w_i)$.

The main questions to be considered are (1) how to generate the pairs (s_i, w_i), and (2) how to compute the focal sets $\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s_i), w_i)$.

Generation of the pairs (s_i, w_i): To generate the pairs (s_i, w_i), we may use a uniform random number generator. However, better results can be obtained using quasi-random low-discrepancy sequences such as Halton sequences [33] [32, page 625]. A sequence of Halton draws is generated as follows. Let r be a prime number greater than two. An integer g can be expressed in terms of the base r as

$$g = \sum_{i=0}^{I} b_i r^i, \tag{65}$$

where $0 \leq b_i \leq r-1$ and $r^I \leq g \leq r^{I+1}$. The Halton sequence is then the series

$$H(g) = \sum_{i=0}^{I} b_i r^{-i-1}, \tag{66}$$

for $g = 1, 2, \ldots, N$. To generate a two-dimensional series, we select two prime numbers r and r'.

Computation of the focal sets $\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s_i), w_i)$ The basic method is to search for the minimum and the maximum of $\varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, w_i)$ under the constraint $pl_{\boldsymbol{y}}(\boldsymbol{\theta}) \geq s_i$, which can be achieved using an iterative constrained nonlinear optimization algorithm. In some cases, however, these optimization problems can be simplified.

First, consider the case where the parameter θ is a scalar. If the contour function is upper-semicontinuous and multimodal, the constraint $pl_{\boldsymbol{y}}(\theta) \geq s_i$ can be expressed as $\widehat{\theta}_*(s_i) \leq \theta \leq \widehat{\theta}^*(s_i)$, where $\widehat{\theta}_*(s_i)$ and $\widehat{\theta}^*(s_i)$ are the solutions of the equation $pl_{\boldsymbol{y}}(\theta) = s_i$. These solutions can be found by any

root-finding algorithm. If function $\varphi_{\boldsymbol{y}}(\theta, w_i)$ is monotone in θ, then the minimum and maximum of $\varphi_{\boldsymbol{y}}(\theta, w_i)$ can be found directly.

In the case where $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a realization of an iid sample, then the contour function will often be, for large n, approximately Gaussian [18],

$$pl_{\boldsymbol{y}}(\theta) \approx \exp\left[-\frac{1}{2} I_{\boldsymbol{y}}(\widehat{\theta})(\theta - \widehat{\theta})^2\right], \tag{67}$$

where $I_{\boldsymbol{y}}(\widehat{\theta})$ is the *observed information* defined as

$$I_{\boldsymbol{y}}(\widehat{\theta}) = -\left.\frac{\partial^2 \log pl_{\boldsymbol{y}}(\theta)}{\partial \theta^2}\right|_{\theta=\widehat{\theta}} = -\left.\frac{\partial^2 \log L_{\boldsymbol{y}}(\theta)}{\partial \theta^2}\right|_{\theta=\widehat{\theta}}. \tag{68}$$

The equation $pl_{\boldsymbol{y}}(\theta) = s_i$ then has the following two approximate solutions

$$\widehat{\theta}_*(s_i) \approx \widehat{\theta} - \sqrt{-\frac{2 \log s_i}{I_{\boldsymbol{y}}(\widehat{\theta})}} \tag{69a}$$

$$\widehat{\theta}^*(s_i) \approx \widehat{\theta} + \sqrt{-\frac{2 \log s_i}{I_{\boldsymbol{y}}(\widehat{\theta})}}. \tag{69b}$$

Before studying the multidimensional case, let us consider the following example.

Example 5 *Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a realization from an iid sample from the exponential distribution with rate parameter θ, with pdf*

$$f_\theta(y) = \theta \exp(-\theta y) \mathbb{1}_{[0,+\infty)}(y) \tag{70}$$

and cdf

$$F_\theta(y) = 1 - \exp(-\theta y). \tag{71}$$

Let Z be an independent rv with the same distribution as Y. By solving the equation $F_\theta(Z) = W$, we get the φ-equation

$$Z = -\frac{\log(1 - W)}{\theta}, \tag{72}$$

where W is a rv with a standard uniform distribution. The contour function is

$$pl_{\boldsymbol{y}}(\theta) = \left(\frac{\theta}{\widehat{\theta}}\right) \exp\left[(\widehat{\theta} - \theta) \sum_{i=1}^{n} y_i\right], \tag{73}$$

with $\widehat{\theta} = 1/\overline{y}$ is the inverse of the mean of the y_i. The focal sets $\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s_i), w_i)$ are the intervals $[z_*(s_i, w_i), z^*(s_i, w_i)]$ with

$$z_*(s_i, w_i) = -\frac{\log(1-w_i)}{\widehat{\theta}^*(s_i)} \quad \textit{and} \quad z^*(s_i, w_i) = -\frac{\log(1-w_i)}{\widehat{\theta}_*(s_i)}, \tag{74}$$

where $\theta^*(s_i)$ and $\theta_*(s_i)$ are the solutions of the equation $pl_{\boldsymbol{y}}(\theta) = s_i$. These solutions have no analytical expression, but they can be approximated by

$$\widehat{\theta}^*(s_i) \approx \widehat{\theta}\left(1 - \sqrt{-2\frac{\log s_i}{n}}\right) \tag{75a}$$

$$\widehat{\theta}_*(s_i) \approx \widehat{\theta}\left(1 + \sqrt{-2\frac{\log s_i}{n}}\right). \tag{75b}$$

Figure 3 shows lower and upper cdfs computed with $N = 10,000$ focal sets, for $n = 30$ observations drawn from the exponential distribution with $\theta = 1$. The MLE of θ was $\widehat{\theta} = 1.010396$. The solid and dotted lines correspond, respectively, to the exact bounds and to the normal approximations. We can see that the approximation is already very good for moderate n. Figure 4 shows the convergence of $\widehat{Bel}^{\mathbb{Z}}_{\boldsymbol{y}}([0,3])$ to $Bel^{\mathbb{Z}}_{\boldsymbol{y}}([0,3])$ for Halton draws and random draws from a uniform distribution. The convergence of the Halton estimator is clearly faster, which confirms previous findings [32, page 628]. □

Let us now consider the case where $\Theta = \mathbb{R}^p$ with $p > 1$. When p is large, the minimization and maximization of $\varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, w_i)$ under the constraint $pl_{\boldsymbol{y}}(\boldsymbol{\theta}) \geq s_i$, for each pair (s_i, w_i), may be time-consuming. However, an outer approximation of the predictive belief function can be computed efficiently as follows. Let $pl_{\boldsymbol{y}}(\theta_j)$ be the marginal contour function for component j of $\boldsymbol{\theta}$,

$$pl_{\boldsymbol{y}}(\theta_j) = \sup_{\boldsymbol{\theta}_{-j}} pl_{\boldsymbol{y}}(\boldsymbol{\theta}), \tag{76}$$

where $\boldsymbol{\theta}_{-j}$ is the subvector of $\boldsymbol{\theta}$ with component j removed. Assuming $pl_{\boldsymbol{y}}(\theta_j)$ to be unimodal and upper-semicontinuous, let $\widehat{\theta}_{j*}(s_i)$ and $\widehat{\theta}^*_j(s_i)$ be the two roots of the equation $pl_{\boldsymbol{y}}(\theta_j) = s_i$. Then, the Cartesian product of the intervals $[\widehat{\theta}_{j*}(s_i), \widehat{\theta}^*_j(s_i)]$ contains $\Gamma_{\boldsymbol{y}}(s_i)$,

$$\prod_{j=1}^{p} [\widehat{\theta}_{j*}(s_i), \widehat{\theta}^*_j(s_i)] \supseteq \Gamma_{\boldsymbol{y}}(s_i). \tag{77}$$

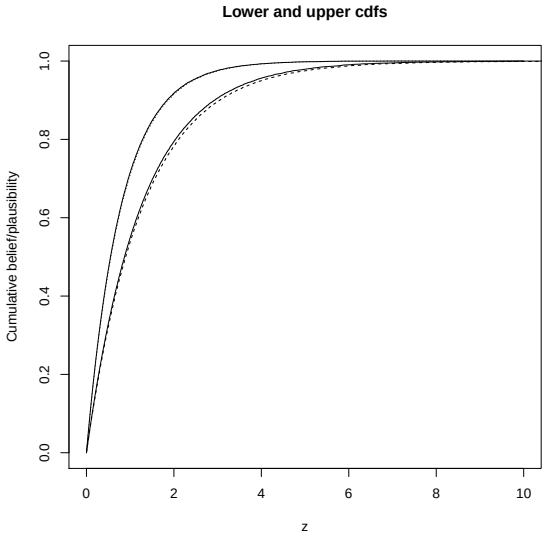


Figure 3: Lower and upper cdf for the exponential data (Example 5): exact bounds (plain curves) and normal approximations (dotted curves)

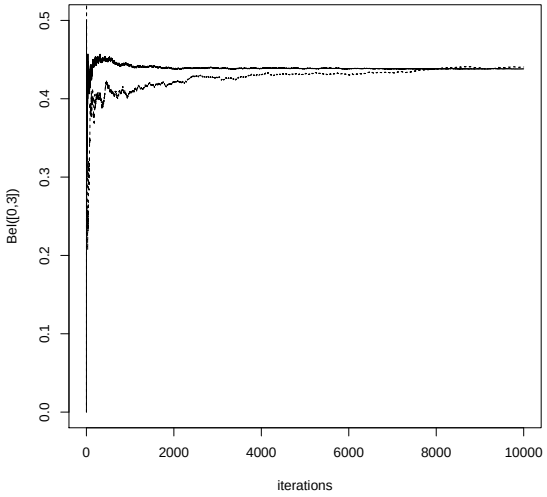


Figure 4: Example of convergence of $\widehat{Bel}_{\boldsymbol{y}}^{\mathbb{Z}}([0,3])$ to $Bel_{\boldsymbol{y}}^{\mathbb{Z}}([0,3])$ for Halton draws (upper solid line) and random draws (lower interrupted line) from a uniform distribution.

Let $\widetilde{z}_*(s_i, w_i)$ and $\widetilde{z}^*(s_i, w_i)$ be the minimum and the maximum of $\varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, w_i)$ under the constraints $\widehat{\theta}_{j*}(s_i) \leq \theta_j \leq \widehat{\theta}^*_j(s_i)$ for $j = 1, \ldots, p$. From (77), we have

$$\widetilde{z}_*(s_i, w_i) \leq z_*(s_i, w_i) \leq z^*(s_i, w_i) \leq \widetilde{z}^*(s_i, w_i), \tag{78}$$

where $z_*(s_i, w_i)$ and $z^*(s_i, w_i)$ are the minimum and the maximum of $\varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, w_i)$ under the constraint $pl_{\boldsymbol{y}}(\boldsymbol{\theta}) = s_i$. The approximating intervals

$$[\widetilde{z}_*(s_i, w_i), \widetilde{z}^*(s_i, w_i)]$$

thus contain the true focal intervals $[z_*(s_i, w_i), z^*(s_i, w_i)]$, resulting in an outer approximation of the true predictive belief function.

Example 6 *Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a realization of an iid sample from the normal distribution with mean μ and standard deviation σ. The vector of parameters is thus $\boldsymbol{\theta} = (\mu, \sigma)$. Let Z be an unobserved random quantity with the same distribution. As mentioned in Example 3, we can write Z as*

$$Z = \varphi(\boldsymbol{\theta}, W) = \mu + \sigma\Phi^{-1}(W), \tag{79}$$

where Φ denotes the cdf of the standard normal distribution and W has a standard uniform distribution. The contour function on Θ is

$$pl_{\boldsymbol{y}}(\mu, \sigma) = \left(\frac{s^2}{\sigma^2}\right)^{n/2} \exp\left(\frac{n}{2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right), \tag{80}$$

where s^2 is the sample variance. The focal sets $\varphi(\Gamma_{\boldsymbol{y}}(s_i), w_i)$ are closed intervals $[z_*(s_i, w_i), z^*(s_i, w_i)]$, where $z_*(s_i, w_i)$ and $z^*(s_i, w_i)$ are the minimum and the maximum of $\varphi(\boldsymbol{\theta}, w_i)$ under the nonlinear constraint $pl_{\boldsymbol{y}}(\mu, \sigma) = s_i$. Now, the marginal contour functions are

$$pl_{\boldsymbol{y}}(\mu) = pl_{\boldsymbol{y}}(\mu, \widehat{\sigma}^2(\mu)) = \left(\frac{s^2}{\widehat{\sigma}^2(\mu)}\right)^{n/2}, \tag{81}$$

where

$$\widehat{\sigma}^2(\mu) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2, \tag{82}$$

and

$$pl_{\boldsymbol{y}}(\sigma) = pl_{\boldsymbol{y}}(\overline{y}, \sigma^2) = \left(\frac{s^2}{\sigma^2}\right)^{n/2} \exp\left[\frac{n}{2}\left(1 - \frac{s^2}{\sigma^2}\right)\right]. \tag{83}$$

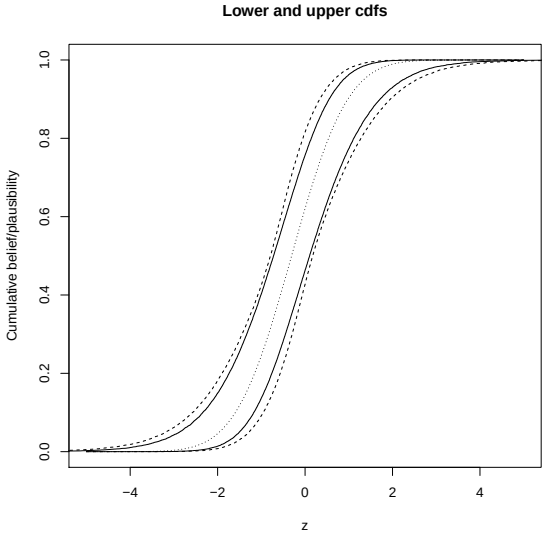


Figure 5: Lower and upper cdf for the normal data (Example 6): exact bounds (solid lines), approximations using the marginal contour functions (interrupted lines) and plug-in cdf $\Phi((z-\overline{y})/s^2)$ (central dotted line).

Let $\widehat{\mu}_*(s_i)$ *and* $\widehat{\mu}^*(s_i)$ *be the roots of the equation* $pl_{\boldsymbol{y}}(\mu) = s_i$. *Similarly, let* $\widehat{\sigma}_*(s_i)$ *and* $\widehat{\sigma}^*(s_i)$ *be the roots of the equation* $pl_{\boldsymbol{y}}(\sigma) = s_i$. *We have*

$$\widetilde{z}_*(s_i, w_i) = \begin{cases} \widehat{\mu}_*(s_i) + \widehat{\sigma}_*(s_i)\Phi^{-1}(w_i) & \text{if } w_i \geq 0.5 \\ \widehat{\mu}_*(s_i) + \widehat{\sigma}^*(s_i)\Phi^{-1}(w_i) & \text{if } w_i < 0.5, \end{cases} \tag{84a}$$

$$\widetilde{z}^*(s_i, w_i) = \begin{cases} \widehat{\mu}^*(s_i) + \widehat{\sigma}^*(s_i)\Phi^{-1}(w_i) & \text{if } w_i \geq 0.5 \\ \widehat{\mu}^*(s_i) + \widehat{\sigma}_*(s_i)\Phi^{-1}(w_i) & \text{if } w_i < 0.5. \end{cases} \tag{84b}$$

Figure 5 shows lower and upper cdfs computed using the exact focal intervals $[z_*(s_i, w_i), z^*(s_i, w_i)]$ *(solid lines) and using the approximations* $[\widetilde{z}_*(s_i, w_i), \widetilde{z}^*(s_i, w_i)]$ *(interrupted lines), for a normal sample of size* $n = 10$ *with* $\overline{y} = 0.3083027$ *and* $s^2 = 1.006766$. *We also show the plug-in cdf* $\Phi((z-\overline{y})/s^2)$ *(central dotted line). We can see that, in this case, the outer approximation is quite accurate. We will see later that this is not always true (see Example 9). Further research is needed to determine the conditions under which this approximation method produces acceptable results.* □

4.4 Prediction of a multidimensional variable

Until now, we have assumed, for simplicity, the predicted variable Z to be one-dimensional. However, the method can be extended in a straightforward way to the more general case where the data $\boldsymbol{Z}$ to be predicted is a vector. Assume, for instance, that $\boldsymbol{Z}$ is a two-dimensional vector (Z_1, Z_2). We can express the marginal distribution of Z_1 and the conditional distribution of Z_2 given z_1 as follows,

$$Z_1 = F^{-1}_{\boldsymbol{y},\boldsymbol{\theta}}(W_1) \tag{85a}$$

$$Z_2 = F^{-1}_{\boldsymbol{y},z_1,\boldsymbol{\theta}}(W_2), \tag{85b}$$

where $F_{\boldsymbol{y},\theta}$ is the conditional cdf of Z_1 given $\boldsymbol{y}$, $F_{\boldsymbol{y},z_1,\boldsymbol{\theta}}$ is the conditional cdf of Z_2 given $\boldsymbol{y}$ and z_1, and $\boldsymbol{W} = (W_1, W_2)$ has a uniform distribution in $[0,1]^2$. This line of reasoning shows that any d-dimensional vector z can be written as

$$\boldsymbol{Z} = \varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, \boldsymbol{W}), \tag{86}$$

where $\boldsymbol{W}$ has a uniform distribution on $[0,1]^d$. The predictive belief function $Bel^{\mathbb{Z}}_{\boldsymbol{y}}$ is then induced by the source

$$([0,1]^{d+1}, \mathcal{B}([0,1]^{d+1}), \lambda_{d+1}, \Gamma''_{\boldsymbol{y}}), \tag{87}$$

where λ_{d+1} is the uniform probability measure in $[0,1]^{d+1}$ and $\Gamma''_{\boldsymbol{y}}$ is the multi-valued mapping

$$(s, \boldsymbol{w}) \to \varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), \boldsymbol{w}). \tag{88}$$

In general, the focal sets $\Gamma''_{\boldsymbol{y}}(s, \boldsymbol{w})$ are subsets of $\mathbb{R}^d$ with arbitrary shape. They can be approximated by boxes

$$B(s, \boldsymbol{w}) = \prod_{k=1}^{d} [z_{*k}(s, \boldsymbol{w}), z^*_k(s, \boldsymbol{w})], \tag{89}$$

where $z_{*k}(s, \boldsymbol{w})$ and $z^*_k(s, \boldsymbol{w})$ are, respectively, the minimum and maximum of the k-th component of $\boldsymbol{z}$ under the constraint $pl_{\boldsymbol{y}}(\boldsymbol{\theta}) \geq s$.

Example 7 *Consider an AR(1) process*

$$X_t = c + \phi X_{t-1} + \varepsilon_t, \tag{90}$$

with $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. *Let* $\boldsymbol{y} = (x_1, \ldots, x_T)$ *the observed sequence between times 1 and* T. *Given that*

$$X_1 \sim \mathcal{N}\left(\frac{c}{1-\phi}, \frac{\sigma^2}{1-\phi^2}\right)$$

and

$$X_t | x_{t-1} \sim \mathcal{N}\left(c + \phi x_{t-1}, \sigma^2\right)$$

the likelihood function can easily computed as

$$L_{\boldsymbol{y}}(\boldsymbol{\theta}) = f(x_1) \prod_{t=2}^{T} f(x_t | x_{t-1}), \tag{91}$$

where $\boldsymbol{\theta} = (c, \phi, \sigma)$. *The contour function is then* $pl_{\boldsymbol{y}}(\boldsymbol{\theta}) = L_{\boldsymbol{y}}(\boldsymbol{\theta}) / L_{\boldsymbol{y}}(\widehat{\boldsymbol{\theta}})$, *where the MLE* $\widehat{\boldsymbol{\theta}}$ *has to be computed numerically. Assume now that we wish to predict the next two terms of the sequence, and let* $\boldsymbol{Z} = (X_{T+1}, X_{T+2})$. *We can write*

$$Z_1 = c + \phi x_t + \sigma \Phi^{-1}(W_1) \tag{92a}$$

$$Z_2 = c + \phi z_1 + \sigma \Phi^{-1}(W_2) \tag{92b}$$

$$= c + \phi c + \phi^2 x_t + \phi \sigma \Phi^{-1}(W_1) + \sigma \Phi^{-1}(W_2). \tag{92c}$$

Vector $\boldsymbol{Z}$ *can thus be written as* $\boldsymbol{Z} = \varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, W_1, W_2)$, *with* (W_1, W_2) *having a uniform distribution on* $[0,1]^2$. *The focal sets* $\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), \boldsymbol{w})$ *are regions of* $\mathbb{R}^2$ *defined as*

$$\begin{aligned}\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s), \boldsymbol{w}) = \{(z_1, z_2) : \exists (c, \phi, \sigma), pl_{\boldsymbol{y}}(c, \phi, \sigma) \geq s, \\ z_1 = c + \phi x_t + \sigma \Phi^{-1}(w_1), z_2 = c + \phi z_1 + \sigma \Phi^{-1}(z_2)\}.\end{aligned} \tag{93}$$

They can be approximated by boxes

$$B(s, \boldsymbol{w}) = [z_{*1}(s, \boldsymbol{w}), z_1^*(s, \boldsymbol{w})] \times [z_{*2}(s, \boldsymbol{w}), z_2^*(s, \boldsymbol{w})], \tag{94}$$

with

$$z_{*1}(s, \boldsymbol{w}) = \min_{c,\phi,\sigma} [c + \phi x_t + \sigma \Phi^{-1}(w_1)] \tag{95a}$$

$$z_1^*(s, \boldsymbol{w}) = \max_{c,\phi,\sigma} [c + \phi x_t + \sigma \Phi^{-1}(w_1)] \tag{95b}$$

$$z_{*2}(s, \boldsymbol{w}) = \min_{c,\phi,\sigma} [c + \phi c + \phi^2 x_t + \phi \sigma \Phi^{-1}(w_1) + \sigma \Phi^{-1}(w_2)] \tag{95c}$$

$$z_2^*(s, \boldsymbol{w}) = \max_{c,\phi,\sigma} [c + \phi c + \phi^2 x_t + \phi \sigma \Phi^{-1}(w_1) + \sigma \Phi^{-1}(w_2))], \tag{95d}$$

under the constraint $pl_{\boldsymbol{y}}(c, \phi, \sigma) \geq s$.

4.5 Comparison with previous work

In this section, we briefly discuss the relationships between our approach and previous work on statistical prediction in the belief function framework.

Dempster's approach [12, 13] is entirely based on φ-equations such as (44) (see also [2, page 251]. With our notation, this approach is to write $(\boldsymbol{Y}, Z)$ as function of $\boldsymbol{\theta}$ and a pivotal random variable W,

$$(\boldsymbol{Y}, Z) = \varphi(\boldsymbol{\theta}, W), \tag{96}$$

which induces a belief function of the product space $\mathbb{Y} \times \mathbb{Z} \times \Theta$. Conditioning on the observed data $\boldsymbol{y}$ and marginalizing on $\mathbb{Z}$ then yields a predictive belief function on $\mathbb{Z}$. This approach is appealing because of its conceptual simplicity. However, it has proved difficult to put at work in practice, except for very simple models. Our method combines Shafer's idea of building a consonant belief function from the likelihood function, with Dempster's φ-equation used only in the prediction step, resulting in a well-founded and yet computationally tractable method.

One could imagine using a different method to construct a belief function in the estimation step. Shafer [59] mentions three methods, including Smets' Generalized Bayes Theorem [60, 61]. However, this method is only applicable in the very specific situation where the parameter space Θ is finite, and we have independent datasets for each single parameter value $\theta_k \in \Theta$ [23]. These conditions are usually non satisfied in statistical inference problems. As argued in Section 3, the likelihood-based approach to representing statistical evidence is both well-founded theoretically and easy to use for a wide range of statistical problems.

In [16], Denoeux proposed a different method to build a predictive belief function for a discrete variable Z, based on multinomial confidence regions. This approach was extended to the case of a continuous variable in [3], using confidence bands on the cdf. Predictive belief functions constructed using this approach have the property that they are dominated by the true probability distribution of Z with some confidence level $1-\alpha$, i.e., for a proportion of at least $1-\alpha$ of the observed samples. In [4], a different approach based on the inverse pignistic transformation [26] was proposed. In this approach, we consider a set $\mathcal{P}$ of probability distributions that contains the true distribution of Z with some specified probability. We then demand that the pignistic probability distribution [62] of the predictive belief function be contained in $\mathcal{P}$, and we construct the most committed consonant belief function that is less committed than any belief function having this property. These two approaches have a frequentist flavor and can be implemented

using multinomial confidence regions and confidence bands in the case of iid data. In contrast with the method described here, they are, however, fundamentally incompatible with Bayesian reasoning. Our new approach is also more widely applicable, as it does not rely on the iid assumption.

In a recent paper[3] [44], Martin and Lingham propose a different approach in the Inferential Model framework [45, 46]. This approach starts with two separate φ-equations for $\boldsymbol{Y}$ and Z,

$$\boldsymbol{Y} = \varphi(\boldsymbol{\theta}, W) \quad \text{and} \quad Z = \varphi'(\boldsymbol{\theta}, W'). \tag{97}$$

Solving for $\boldsymbol{\theta}$ in the first equation and plugging in to the second one yields a new φ equation of the form

$$Z = \varphi'(\boldsymbol{\theta}(\boldsymbol{Y}, W), W'), \tag{98}$$

which they rewrite as $Z = \varphi''(\boldsymbol{Y}, W'')$. They then define a predictive random set for W'', which, for fixed $\boldsymbol{y}$, allows them to define the plausibility of any assertion $A \subset \mathbb{Y}$ of interest. The authors indicate that "the choice of a predictive random set ought to depend on the assertion A of interest". The approach thus departs from the classical Dempster-Shafer theory, in which belief functions quantify degrees of belief. In contrast, our method sticks strictly to this subjective interpretation. As both approaches have been developed independently and almost simultaneously, more work is needed to carry out a deep analysis of the relative merits of the two approaches, both from the theoretical and practical viewpoints.

5 Application to regression

To provide a more detailed illustration of the way the above estimation and prediction methods can be put at work, we will discuss their use for regression analysis. The estimation and prediction problems will first addressed, respectively, in Sections 5.1 and 5.2. Finally, the prediction problem with uncertain inputs will be studied in Section 5.3.

5.1 Estimation

We consider the following standard linear regression model,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{99}$$

[3] We thank an anonymous reviewer for bringing this work to our attention.

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the vector of n observations of the dependent variable, X is the fixed design matrix of size $n \times (p+1)$, such that the first column contains 1's and column j $(1 \leq j \leq p)$ contains the observations of the j-th covariate, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ is the vector of errors, assumed to be normally distributed with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I_n$, where I_n is the identity matrix of size n. The vector of coefficients is $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma)'$. The likelihood function for this model is

$$L_{\boldsymbol{y}}(\boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta})\right]. \tag{100}$$

Assuming X to have full column rank, the MLE of $\boldsymbol{\beta}$ is the ordinary least squares estimate

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y} \tag{101}$$

and the MLE of σ is the standard deviation of residuals:

$$\widehat{\sigma} = \sqrt{(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})/n}. \tag{102}$$

The contour function (25) can thus be readily calculated as

$$pl_{\boldsymbol{y}}(\boldsymbol{\theta}) = L_{\boldsymbol{y}}(\boldsymbol{\theta})/L_{\boldsymbol{y}}(\widehat{\boldsymbol{\theta}}), \tag{103}$$

with $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}', \widehat{\sigma})'$.

Let us now consider assertions (hypotheses) H of the form $R\boldsymbol{\beta} = \boldsymbol{q}$, where R is a $r \times (p+1)$ constant matrix and $\boldsymbol{q}$ is a constant vector of length r, for some $r \leq p+1$. (Equations of the form $R\boldsymbol{\beta} = \boldsymbol{q}$ are sometimes called "linear restrictions"). This general formulation includes as special cases usual assumptions of the forms $\{\beta_j = 0\}$, $\{\beta_j = 0, \forall j \in \{1, \ldots, p\}\}$, or $\{\beta_j = \beta_k\}$. The plausibility of H is

$$Pl^{\Theta}_{\boldsymbol{y}}(H) = \sup_{R\boldsymbol{\beta}=\boldsymbol{q}} pl_{\boldsymbol{y}}(\boldsymbol{\theta}). \tag{104}$$

The solution of this linearly constrained optimization problem is given by the restricted least-squares estimate $\widehat{\boldsymbol{\theta}}_* = (\widehat{\boldsymbol{\beta}}'_*, \widehat{\sigma}_*)'$, which has the following expression [32, page 122],

$$\widehat{\boldsymbol{\beta}}_* = \widehat{\boldsymbol{\beta}} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\widehat{\boldsymbol{\beta}} - \boldsymbol{q}), \tag{105}$$

and

$$\widehat{\sigma}_* = \sqrt{(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_*)'(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}_*)/n}. \tag{106}$$

We then have

$$Pl_{\boldsymbol{y}}^{\Theta}(H) = \frac{L_{\boldsymbol{y}}(\widehat{\boldsymbol{\theta}}_*)}{L_{\boldsymbol{y}}(\widehat{\boldsymbol{\theta}})}. \tag{107}$$

Equations (105)-(107) allow us, in particular, to compute the marginal contour functions $pl_{\boldsymbol{y}}(\beta_j)$. The marginal contour functions $pl_{\boldsymbol{y}}(\sigma^2)$ is

$$pl_{\boldsymbol{y}}(\sigma^2) = \left(\frac{\widehat{\sigma}^2}{\sigma^2}\right)^{n/2} \exp\left[\frac{n}{2}\left(1 - \frac{\widehat{\sigma}^2}{\sigma^2}\right)\right]. \tag{108}$$

We note that assertions of the form $c(\boldsymbol{\beta}) = 0$, where c is a nonlinear function, could be handled as well, the solving the corresponding nonlinear optimization problem numerically.

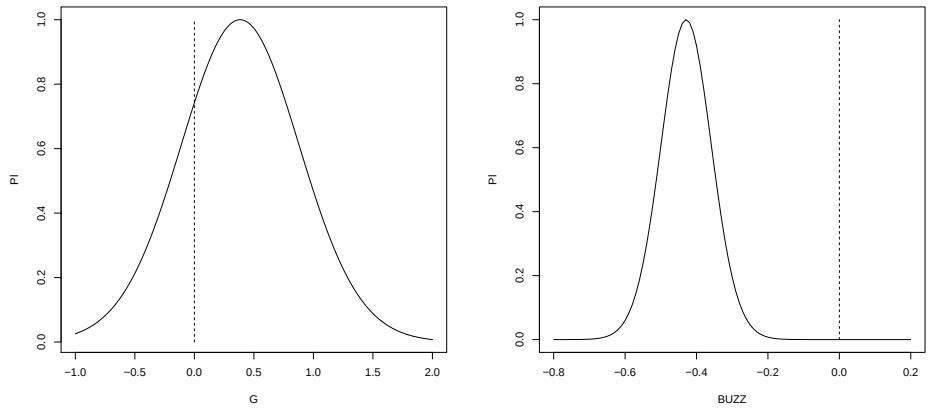
Example 8 *As an example, we considered the task of predicting the box office success of movies. We used the same dataset*[4] *as in [32, page 93], containing data about 62 movies released in 2009. We considered the logarithm of Box Office receipts as dependent variable, and 11 covariates: 3 dummy variables (G, PG, PG13) to encode the MPAA (Motion Picture Association of America) rating, logarithm of budget (LOGBUDGET), star power (STARPOWR), a dummy variable to indicate if the movie is a sequel (SEQUEL), four dummy variables to describe the genre (ACTION, COMEDY, ANIMATED, HORROR), and one variable to represent internet buzz (BUZZ). This last variable was constructed by aggregating four measures using principal component analysis, as described in [32, pages 93–94].*

Table 1 shows the MLEs of the coefficients, together with the usual statistics (standard errors, t and p-values) and the plausibilities $Pl(\beta_j = 0)$ computed using (105)-(26). We can see that, from a practical point of view, the p-values and the plausibilities provide similar information. Both measures identify variables BUZZ, ACTION and, to a lesser extent, ANIMATED as having a coefficient significantly different from zero. However, they have completely different interpretations: the p-value is the probability, under the hypothesis $\beta_j = 0$ and assuming new data to be repeatedly drawn, of observing an absolute value $|T|$ of the t statistics as least as large as the observed value $|t|$. It is thus based on the assumption of repeated sampling, and it takes into account, in the computation of the probability, values of statistics $|t|$ larger than the one that has actually been observed. In contrast, the assertion $Pl(\beta_j = 0) = \alpha$ means that there is a vector $\boldsymbol{\beta}$ of coefficients, with

[4]This dataset can be downloaded at `http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm`.

Table 1: Regression coefficients (movie example).

	Estimate	Std. Error	t-value	p-value	$Pl(\beta_j = 0)$
(Intercept)	15.400	0.643	23.960	$<$ 2e-16	1.0e-34
G	0.384	0.553	0.695	0.49	0.74
PG	0.534	0.300	1.780	0.081	0.15
PG13	0.215	0.219	0.983	0.33	0.55
LOGBUDGET	0.261	0.185	1.408	0.17	0.30
STARPOWR	4.32e-3	0.0128	0.337	0.74	0.93
SEQUEL	0.275	0.273	1.007	0.32	0.54
ACTION	-0.869	0.293	-2.964	4.7e-3	6.6e-3
COMEDY	-0.0162	0.256	-0.063	0.95	0.99
ANIMATED	-0.833	0.430	-1.937	0.058	0.11
HORROR	0.375	0.371	1.009	0.32	0.54
BUZZ	0.429	0.0784	5.473	1.4e-06	4.8e-07



(a) (b)

Figure 6: Marginal contour functions for the coefficients of variables G (a) and BUZZ (b).

$\beta_j = 0$, whose likelihood, given the data, is α times the maximum likelihood. This interpretation only involves the observed data (and not any other data that might have been observed). More complete information can be gained by plotting the marginal contour functions $pl_{\boldsymbol{y}}(\beta_j)$, such as displayed in Figure 6 for parameters G and BUZZ.

Finally, it is common practice in regression analysis to test the hypothesis that all coefficients (except the intercept) are equal to zero. If this hypothesis is rejected, the regression is said to be significant. In our approach, the plausibility of this hypothesis can be computed using (105)-(107), with the restriction matrix R of size $p \times (p+1)$ such that $R_{ij} = 1$ if $j = i+1$ and $R_{ij} = 1$ otherwise, and taking $\boldsymbol{q}$ as a vector of p components, each of which is equal to zero. In the movie example, we get $Pl\boldsymbol{y}^{\Theta}(H) = 10^{-12}$, which clearly indicates that this hypothesis is extremely unlikely, given the data. □

5.2 Prediction

Prediction with the linear regression can be easily handled using the method described in Section 4. Let Z be a not-yet observed value of the dependent variable for a vector $\boldsymbol{x}_0$ of covariates:

$$Z = \boldsymbol{x}_0'\boldsymbol{\beta} + \epsilon_0, \tag{109}$$

with $\epsilon_0 \sim \mathcal{N}(0, \sigma^2)$. We can write, equivalently,

$$Z = \boldsymbol{x}_0'\boldsymbol{\beta} + \sigma\Phi^{-1}(W), \tag{110a}$$

$$= \varphi_{\boldsymbol{y}}(\boldsymbol{\theta}, W), \tag{110b}$$

where W has a standard uniform distribution. The predictive belief function on Z can then be approximated using the methods described in Section 4.3. The exact method necessitates, for each pair (s_i, w_i), to compute the minimum and the maximum of the linear function $\boldsymbol{x}_0'\boldsymbol{\beta} + \sigma\Phi^{-1}(w_i)$ under the nonlinear constraint $pl_{\boldsymbol{y}} \geq s_i$. The outer approximation method is to compute the s_i-level sets $[\widehat{\beta}_{j*}(s_i), \widehat{\beta}_j^*(s_i)]$ and $[\widehat{\sigma}_*(s_i), \sigma^*(s_i)]$ of the marginal contour functions $pl_{\boldsymbol{y}}(\beta_j)$ and $pl_{\boldsymbol{y}}(\sigma)$. Each focal set $\varphi_{\boldsymbol{y}}(\Gamma_{\boldsymbol{y}}(s_i), w_i)$ is then approximated by the interval $[\widetilde{z}_*(s_i, w_i), \widetilde{z}^*(s_i, w_i)]$, with

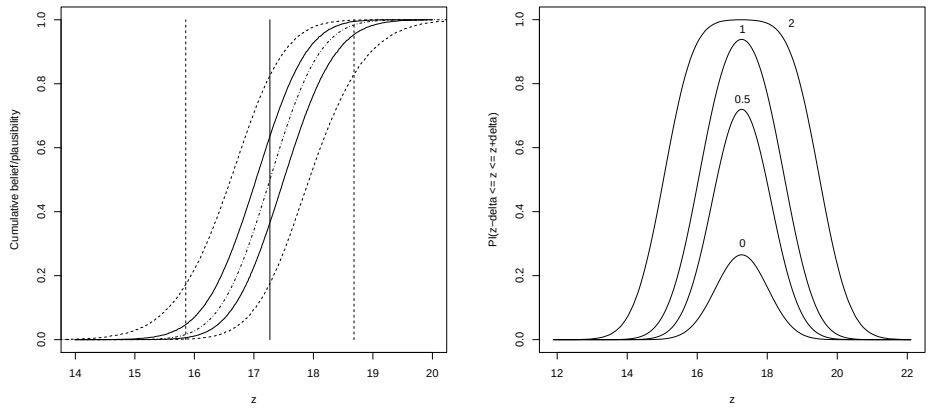
$$\widetilde{z}_*(s_i, w_i) = \sum_{j:x_{0j}>0} \widehat{\beta}_{j*}(s_i)x_{0j} + \sum_{j:x_{0j}<0} \widehat{\beta}_j^*(s_i)x_{0j} + (\widehat{\sigma}_*(s_i)\mathbb{1}_{w_i \geq 0.5} + \widehat{\sigma}^*(s_i)\mathbb{1}_{w_i<0.5})\Phi^{-1}(w_i) \tag{111a}$$

$$\widetilde{z}^*(s_i, w_i) = \sum_{j:x_{0j}>0} \widehat{\beta}^*_j(s_i)x_{0j} + \sum_{j:x_{0j}<0} \widehat{\beta}_{j*}(s_i)x_{0j} + (\widehat{\sigma}^*(s_i)\mathbb{1}_{w_i\geq 0.5} + \widehat{\sigma}_*(s_i)\mathbb{1}_{w_i<0.5})\Phi^{-1}(w_i). \quad (111b)$$

Example 9 *Continuing Example 8, let us consider a reduced model with only the variables ACTION and BUZZ as inputs (as only these two factors have a significant effect on the dependent variable). Assume that, for a particular movie, we have the following input vector* $\boldsymbol{x}_0 = (1, 12.81)'$*, meaning that it is an action film with buzz variable equal to 2.81. Figure 7(a) displays the lower and upper cdfs of the predictive belief function, approximated using* $N = 5000$ *randomly generated focal sets, using the exact method (solid lines) and using the outer approximation method (interrupted lines). We can see that, in this case, the outer approximation method is very conservative, which may be due to a high correlation between the coefficient estimates. The figure also shows the bounds of the 95% prediction interval, as well as the cdf of the plug-in distribution of* Z*, which is the normal distribution with mean* $\boldsymbol{x}_0'\widehat{\beta}$ *and standard deviation* $\widehat{\sigma}$*. Figure 7(b) is a "pl-plot", which shows the plausibilities* $Pl^{\mathbb{Z}}_{\boldsymbol{y}}([z-\delta, z+\delta])$ *(computed using the exact method) as a function of* z*, for different values of* δ*. The plug-in estimate of the expectation of* Z *is* 17.27*. The lower-upper expectation interval is* $[17.02, 17.51]$*. Its estimation using the outer approximation method is* $[16.57, 17.97]$*.* □

5.3 Prediction with uncertain inputs

From a practical point of view, a significant advantage of the predictive belief function formalism is the ease with which, being built upon the very general Dempster-Shafer framework, it can accommodate various sources of uncertain information. Consider, for example, the *ex ante* forecasting situation, in which some explanatory variables are unknown at the time of the forecast and have to be estimated or predicted. The classical way to handle this problem is to assume that $\boldsymbol{x}_0$ has been estimated with some variance, which has to be taken into account in the calculation of the forecast variance. However, as noted by Green [32, page 87], this problem is considered by many authors as "simply intractable" and, even with simplifying assumptions, "analytical results for the correct forecast variance remain to be derived except for simple special cases". In contrast, this problem can be handled very naturally in our approach by modeling partial knowledge of $\boldsymbol{x}_0$ by a belief function $Bel^{\mathbb{X}}$ in the sample space $\mathbb{X}$ of $\boldsymbol{x}_0$. Recall, from Section 4.1, that the predictive belief function $Bel^{\mathbb{Z}}_{\boldsymbol{y}}$ is obtained by combining the likelihood belief function $Bel^{\Theta}_{\boldsymbol{y}}$ with the joint belief function $Bel^{\mathbb{Z}\times\Theta}_{\boldsymbol{y}}$ induced



(a) (b)

Figure 7: (a): Lower and upper cdf for the prediction problem: exact values (solid lines) and outer approximations (interrupted lines); the central curve is the cdf of the plug-in distribution of Z; the vertical dotted lines correspond to the frequentist 95% prediction interval. (b): pl-plot: plausibility $Pl_{\boldsymbol{y}}^{\mathbb{Z}}([z-\delta, z+\delta])$ as a function of z, for $\delta \in \{0, 0.5, 1, 2\}$.

by (45). As the belief function $Bel^{\mathbb{X}}$ is just another piece of evidence, it can be combined with the other two by Dempster's rule. The combined belief function is then, as before, marginalized on $\mathbb{Z}$ to get the predictive belief function. To describe the corresponding algorithm, let us emphasize the dependence of Z on $\boldsymbol{x}_0$ by rewriting (110b) as

$$Z = \varphi_y(\boldsymbol{x}_0, \boldsymbol{\theta}, W). \tag{112}$$

Assume that the belief function $Bel^{\mathbb{X}}$ is induced by a source $(\Omega, \mathcal{A}, \mathbb{P}^{\Omega}, \Lambda)$, where Λ is a multi-valued mapping from Ω to $2^{\mathbb{X}}$. The predictive belief function $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ is then induced by the multi-valued mapping

$$(\omega, s, w) \to \varphi_y(\Lambda(\omega), \Gamma_{\boldsymbol{y}}(s), w). \tag{113}$$

If $\varphi_{\boldsymbol{y}}$ is continuous and if both $\Lambda(\omega)$ and $\Gamma_{\boldsymbol{y}}(s)$ are connected for all ω and s, then each set $\varphi_y(\Lambda(\omega), \Gamma_{\boldsymbol{y}}(s), w)$ is an interval, and $Bel_{\boldsymbol{y}}^{\mathbb{Z}}$ is equivalent to a random interval. It can be approximated by Monte Carlo simulation using Algorithm 3.

Example 10 *To illustrate the application of this algorithm, we considered the prediction of movie box office receipt with the same input vector* $\boldsymbol{x}_0$ *as*

Algorithm 3 Monte Carlo algorithm for approximating a predictive belief function with uncertain input vector $\boldsymbol{x}_0$ (random interval case).

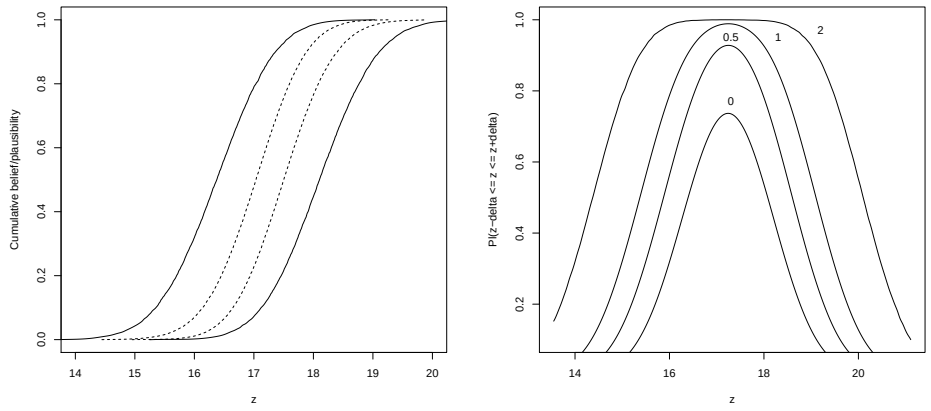
Require: Desired number of focal sets N
  for $i = 1$ **to** N **do**
    Draw (s_i, w_i) uniformly in $[0,1]^2$
    Draw ω from $\mathbb{P}^\Omega$
    Search for $z_{*i} = \min_{\boldsymbol{\theta}} \varphi_{\boldsymbol{y}}(\boldsymbol{x}_0, \boldsymbol{\theta}, w_i)$ such that $pl_{\boldsymbol{y}}(\boldsymbol{\theta}) \geq s_i$ and $\boldsymbol{x}_0 \in \Lambda(\omega)$.
    Search for $z_i^* = \max_{\boldsymbol{\theta}} \varphi_{\boldsymbol{y}}(\boldsymbol{x}_0, \boldsymbol{\theta}, w_i)$ such that $pl_{\boldsymbol{y}}(\boldsymbol{\theta}) \geq s_i$ and $\boldsymbol{x}_0 \in \Lambda(\omega)$.
    $B_i \leftarrow [z_{*i}, z_i^*]$
  end for

in Example 9, but assuming the buzz variable to be partially unknown. To model partial knowledge of BUZZ, we used a consonant random interval with the following triangular contour function,

$$pl(x) = \begin{cases} \dfrac{\widetilde{x} - x}{\widetilde{x} - x_*} & \text{if } x_* \leq x < \widetilde{x} \\ \dfrac{x - \widetilde{x}}{x^* - \widetilde{x}} & \text{if } \widetilde{x} \leq x < x^* \\ 0 & \text{otherwise,} \end{cases} \tag{114}$$

with $x_* = 0$, $x^* = 5$ *and* $\widetilde{x} = 2.81$. *The resulting lower and upper cdfs are shown as interrupted lines in Figure 8(a), and the corresponding pl-plot is shown in Figure 8(b). The estimated lower-upper expectation interval is* $[16.36, 18.11]$. *We can see that, as expected, the predictive belief function becomes more uncertain, as a result of the uncertainty on one of the covariates.* □

Another situation in which the regression analysis has to be combined with other sources of information is the case where, in addition to the statistical prediction computed from the linear regression model, some expert opinions about the future data Z are available. In our example, we can figure out that specialists of the movie industry (such as film critics) can provide a prediction of a movie's box office success, taking into account various pieces of evidence that can never be fully captured by a regression equation. Such non-statistical information can be represented in the belief function framework and combined with the predictive belief function using Dempster's rule. In contrast, it is not at all clear how prediction intervals



(a) (b)

Figure 8: (a): Lower and upper cdf for the prediction problem with uncertain inputs (solid lines) and with certain inputs (interrupted lines); (b): pl-plot of the predictive belief function with uncertain inputs.

and expert options, being of totally different natures, could be combined in a principled way.

6 Conclusions

In many areas, such as business and economics, forecasts are typically used for decision-making and strategic planning. When aggregating predictions from numerical models with other information, decision-makers need to assess the uncertainty of the forecasts. Describing this uncertainty in a faithful and accurate way is thus a very important issue. The approach advocated in this paper is to model estimation uncertainty using a belief function constructed from the likelihood, and to combine it with random uncertainty arising from the data-generating process, resulting in a predictive belief function. The practical use of this method has been illustrated using a simple but widely used model: standard linear regression.

Predictive belief functions constructed in this way have been argued to be better founded than frequentist prediction intervals, and to be more widely applicable than Bayesian posterior predictive distributions, which always require prior distributions. However, the latter are recovered when a prior distribution on the model parameters is specified. The proposed method also

has practical advantages over the frequentist approach, which often has to resort to asymptotic approximations. For instance, in linear regression with serial correlation, the variance of prediction errors cannot be determined exactly, making it difficult to compute prediction intervals [55, page 215]. In contrast, the predicted belief function can easily be approximated to any desired accuracy using Monte Carlo simulation, even for small sample sizes.

Acknowledgement

This research was supported by the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" by the National Agency for Research (reference ANR-11-IDEX-0004-02). It was also supported by the Center of Excellence in Econometrics at Chiang Mai University.

References

References

[1] M. Aickin. Connecting Dempster-Shafer belief functions with likelihood-based inference. *Synthese*, 123:347–364, 2000.

[2] R. G. Almond. *Graphical belief models*. Chapman and Hall, London, 1995.

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

[4] A. Aregui and T. Denœux. Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities. *International Journal of Approximate Reasoning*, 49(3):575–594, 2008.

[5] G. A. Barnard, G. M. Jenkins, and C. B. Winsten. Likelihood inference and time series. *Journal of the Royal Statistical Society*, 125(3):321–372, 1962.

[6] N. Ben Abdallah, N. Mouhous Voyneau, and T. Denœux. Combining statistical and expert evidence using belief functions: Application

to centennial sea level estimation taking into account climate change. *International Journal of Approximate Reasoning*, 55:341–354, 2014.

[7] J. O. Berger, E. Moreno, L. R. Pericchi, M. Bayarri, J. M. Bernardo, J. A. Cano, J. De la Horra, J. Martín, D. Ríos-Insúa, B. Betrò, A. Dasgupta, P. Gustafson, L. Wasserman, J. B. Kadane, C. Srinivasan, M. Lavine, A. O'Hagan, W. Polasek, C. P. Robert, C. Goutis, F. Ruggeri, G. Salinetti, and S. Sivaganesan. An overview of robust bayesian analysis. *Test*, 3(1):5–124, 1994.

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

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

[10] M. A. Boujelben and Y. De Smet. A choice model with imprecise ordinal evaluations. *International Journal of Approximate Reasoning*, 55(2):689–710, 2014.

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

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

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

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

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

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

[17] T. Denœux. Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *Artificial Intelligence*, 172:234–264, 2008.

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

[19] T. Denœux. Rejoinder on "likelihood-based belief function: Justification and some extensions to low-quality data". *International Journal of Approximate Reasoning*, 55(7):1614–1617, 2014.

[20] T. Denœux and M.-H. Masson. EVCLUS: Evidential clustering of proximity data. *IEEE Trans. on Systems, Man and Cybernetics B*, 34(1):95–109, 2004.

[21] T. Denœux and P. Smets. Classification using belief functions: the relationship between the case-based and model-based approaches. *IEEE Transactions on Systems, Man and Cybernetics B*, 36(6):1395–1406, 2006.

[22] D. Dubois. On various ways of tackling incomplete information in statistics. *International Journal of Approximate Reasoning*, 55(7):1570 – 1574, 2014.

[23] D. Dubois and T. Denœux. Statistical inference with belief functions and possibility measures: a discussion of basic assumptions. In C. Borgelt, G. G. Rodríguez, W. Trutschnig, M. A. Lubiano, M. A. Gil, P. Grzegorzewski, and O. Hryniewicz, editors, *Combining soft computing and statistical methods in data analysis (Proceedings of SMPS 2010)*, Advances in Intelligent and Soft Computing, pages 217–225, Oviedo, Spain, 2010. Springer.

[24] D. Dubois and H. Prade. A set-theoretic view of belief functions: logical operations and approximations by fuzzy sets. *International Journal of General Systems*, 12(3):193–226, 1986.

[25] D. Dubois and H. Prade. The principle of minimum specificity as a basis for evidential reasoning. In B. Bouchon and R. Yager, editors, *Uncertainty in Knowledge-Based Systems*, volume 286 of *Lecture Notes in Computer Science*, pages 75–84. Springer Berlin Heidelberg, 1987.

[26] D. Dubois, H. Prade, and P. Smets. A definition of subjective possibility. *International Journal of Approximate Reasoning*, 48(2):352–364, 2008.

[27] A. W. F. Edwards. *Likelihood (expanded edition)*. The John Hopkins University Press, Baltimore, USA, 1992.

[28] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A*, 222:309–368, 1922.

[29] D. A. S. Fraser. *The structure of inference*. Wiley, New-York, 1968.

[30] I. Gilboa. Expected utility with purely subjective non-additive probabilities. *Journal of Mathematical Economics*, 16(2):65–88, 1987.

[31] I. Gilboa and D. Schmeidler. Additive representations of non-additive measures and the choquet integral. *Annals of Operations Research*, 51:43–65, 1994.

[32] W. H. Greene. *Econometric analysis*. Prentice Hall, Upper Saddle River, NJ, USA, 7th edition, 2012.

[33] J. H. Halton. Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, 7(12):701–702, 2015.

[34] D. J. Hudson. Interval estimation from the likelihood function. *J. R. Statistical Society B*, 33(2):256–262, 1973.

[35] O. Kanjanatarakul, P. Lertpongpiroon, S. Singkharat, and S. Sriboonchitta. Econometric forecasting using linear regression and belief functions. In F. Cuzzolin, editor, *Belief functions: theory and applications 3*, Advances in Intelligent and Soft Computing, Oxford, UK, 2014. Springer.

[36] O. Kanjanatarakul, S. Sriboonchitta, and T. Denœux. Forecasting using belief functions: an application to marketing econometrics. *International Journal of Approximate Reasoning*, 55(5):1113–1128, 2014.

[37] F. Klawonn and P. Smets. The dynamic of belief in the transferable belief model and specialization-generalization matrices. In D. D. et al., editor, *Proc. of the 8th conference on Uncertainty in Artificial Intelligence*, pages 130–137. Morgan Kaufmann, San Mateo, CA, 1992.

[38] G. J. Klir and M. J. Wierman. *Uncertainty-Based Information. Elements of Generalized Information Theory*. Springer-Verlag, New-York, 1999.

[39] M. Kurdej, J. Moras, V. Cherfaoui, and P. Bonnifait. Controlling remanence in evidential grids using geodata for dynamic scene perception. *International Journal of Approximate Reasoning*, 55(1, Part 3):355 – 375, 2014.

[40] B. Lelandais, I. Gardin, L. Mouchard, P. Vera, and S. Ruan. Dealing with uncertainty and imprecision in image segmentation using belief function theory. *International Journal of Approximate Reasoning*, 55(1, Part 3):376–387, 2014.

[41] Z.-G. Liu, Q. Pan, and J. Dezert. Classification of uncertain and imprecise data based on evidence theory. *Neurocomputing*, 133:459 – 470, 2014.

[42] F. Mangili and A. Benavoli. New prior near-ignorance models on the simplex. *International Journal of Approximate Reasoning*, 56, Part B:278–306, 2015.

[43] R. Martin. Plausibility functions and exact frequentist inference. *Journal of the American Statistical Association (to appear)*, 2015.

[44] R. Martin and R. Lingham. Prior-free probabilistic prediction of future observations. *Technometrics (to appear)*, 2015.

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

[46] R. Martin, J. Zhang, and C. Liu. Dempster-Shafer theory and statistical inference with weak beliefs. *Statistical Science*, 25:72–87, 2010.

[47] M.-H. Masson and T. Denœux. ECM: an evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4):1384–1397, 2008.

[48] S. Moral. Imprecise probabilities for representing ignorance about a parameter. *International Journal of Approximate Reasoning*, 53(3):347–362, 2012.

[49] S. Moral. Comments on "likelihood-based belief function: Justification and some extensions to low-quality data" by thierry denœux. *International Journal of Approximate Reasoning*, 55(7):1591 – 1593, 2014.

[50] S. Moral and N. Wilson. Markov-chain Monte-Carlo algorithms for the calculation of Dempster-Shafer belief. In *Proc. of the Twelfth National Conference on Artificial intelligence (AAAI-94)*, volume 1, pages 269–274, 1994.

[51] S. Moral and N. Wilson. Importance sampling Monte-Carlo algorithms for the calculation of Dempster-Shafer belief. In *Proc. of IPMU'96*, volume III, pages 1337–1344, 1996.

[52] H. T. Nguyen. On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 65:531–542, 1978.

[53] H. T. Nguyen. *An Introduction to Random Sets*. Chapman and Hall/CRC Press, Boca Raton, Florida, 2006.

[54] S. Petit-Renaud and T. Denœux. Nonparametric regression analysis of uncertain and imprecise data using belief functions. *International Journal of Approximate Reasoning*, 35(1):1–28, 2004.

[55] R. S. Pindyck and D. L. Rubinfeld. *Econometric models and economic forecasts*. McGraw-Hill, Boston, Massachusetts, USA, 4th edition, 1998.

[56] E. Ramasso, M. Rombaut, and N. Zerhouni. Joint prediction of continuous and discrete states in time-series based on belief functions. *Cybernetics, IEEE Transactions on*, 43(1):37–50, 2013.

[57] L. Serir, E. Ramasso, P. Nectoux, and N. Zerhouni. E2GKpro: An evidential evolving multi-modeling approach for system behavior prediction with applications. *Mechanical Systems and Signal Processing*, 37(1–2):213 –228, 2013.

[58] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.

[59] G. Shafer. Belief functions and parametric models (with discussion). *J. Roy. Statist. Soc. Ser. B*, 44:322–352, 1982.

[60] P. Smets. *Un modèle mathématico-statistique simulant le processus du diagnostic médical*. PhD thesis, Université Libre de Bruxelles, Brussels, Belgium, 1978. (in French).

[61] P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[62] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.

[63] D. A. Sprott. *Statistical Inference in Science*. Springer-Verlag, Berlin, 2000.

[64] Z. G. Su, Y. F. Wang, and P. H. Wang. Parametric regression analysis of imprecise and uncertain data in the fuzzy belief function framework. *International Journal of Approximate Reasoning*, 54(8):1217–1242, 2013.

[65] L. A. Wasserman. Belief functions and statistical evidence. *The Canadian Journal of Statistics*, 18(3):183–196, 1990.

[66] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 1938.

[67] P. Xu, F. Davoine, H. Zha, and T. Denœux. Evidential calibration of binary SVM classifiers. *International Journal of Approximate Reasoning (this issue)*, 2015.

[68] R. R. Yager. The entailment principle for Dempster-Shafer granules. *Int. J. of Intelligent Systems*, 1:247–262, 1986.