Likelihood-based belief function: justification and some extensions to low-quality data

Thierry Denoeux

Likelihood-based belief function: justification and some extensions to low-quality data

Thierry Denœux

Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc

Abstract

Given a parametric statistical model, evidential methods of statistical inference aim at constructing a belief function on the parameter space from observations. The two main approaches are Dempster's method, which regards the observed variable as a function of the parameter and an auxiliary variable with known probability distribution, and the likelihood-based approach, which considers the relative likelihood as the contour function of a consonant belief function. In this paper, we revisit the latter approach and prove that it can be derived from three basic principles: the likelihood principle, compatibility with Bayes' rule and the minimal commitment principle. We then show how this method can be extended to handle low-quality data. Two cases are considered: observations that are only partially relevant to the population of interest, and data acquired through an imperfect observation process.

Keywords: Statistical Inference, Dempster-Shafer Theory, Evidence Theory, Likelihood Principle, Uncertain data, Partially relevant data.

1. Introduction

Belief functions were initially introduced by Dempster as part of a new approach to statistical inference, in which lower and upper posterior probabilities can be constructed without having to postulate the existence of prior probabilities [10, 11, 12]. Dempster's initial ideas were later formalized by Shafer [37] and transformed into a general system for reasoning under uncertainty, now usually referred to as Dempster-Shafer theory[1]. Whereas this

[1]The fundamental notions of Dempster-Shafer theory will be assumed to be known thoughout this paper. The reader is referred to Shafer's monograph [37] for a thorough

work has had a strong impact in some research fields such as Pattern Recognition, Information Fusion and Artificial Intelligence, statistical applications have until now remained quite limited. Recently, however, there has been revived interest in this topic, perhaps under the influence of Dempster's more recent work along his initial ideas [14]. New variants of Dempster's method of inference leading to belief functions having some well-defined long-run frequency properties have been proposed, such as the Weak Belief approach [32], the Elastic Belief method [30] and the Inferential Models [31].

Let $X \in \mathbb{X}$ denote the observable data, $\theta \in \Theta$ the parameter of interest and $f(x;\theta)$ the probability mass or density function describing the data-generating mechanism. The key idea underlying Dempster's method of inference and its variants is to represent such a sampling model by an equation

$$X = a(\theta, U), \tag{1}$$

where U is an unobserved auxiliary variable with known probability distribution μ independent of θ. This representation is similar to that of Fraser [27]. It is quite natural in the context of data generation. For instance, to generate a continuous random variable X with cumulative distribution function (cdf) F_θ, one might draw U from a continuous uniform distribution in $[0,1]$ and set $X = F_\theta^{-1}(U)$. Equation (1) defines a multi-valued mapping

$$\Gamma : U \to \Gamma(U) = \{(X,\theta) \in \mathbb{X} \times \Theta | X = a(\theta, U)\}. \tag{2}$$

This mapping, together with measure μ on $\mathbb{U}$, defines a random set, i.e., a belief function on $\mathbb{X} \times \Theta$ [33]. Conditioning this belief function on θ yields the sampling distribution $f(\cdot;\theta)$ on $\mathbb{X}$, while conditioning it on $X = x$ gives a belief function[2] $Bel_\Theta(\cdot;x)$ on Θ.

While this inference method is conceptually elegant, it often leads to cumbersome or even intractable calculations except for simple models, which imposes the use of Monte-Carlo simulations. A more fundamental criticism that may be raised against this approach is the fact that representation (1) and, consequently, $Bel_\Theta(\cdot;x)$ are not unique for a given statistical model $\{f(\cdot;\theta), \theta \in \Theta\}$. As the auxiliary variable U is not observable, it is not clear how one could argue for one model or another, except from practical considerations.

exposition and to some papers cited in reference such as, e.g, [42, 18] for short introductions.

[2]Throughout this paper, the domain of set functions will be indicated as subscript.

An alternative approach to statistical inference was proposed by Shafer in [37]. In this approach, the evidence about Θ, after ofter observing $X = x$, is postulated to be represented by a consonant "likelihood-based" belief function, whose contour function equals the normalized likelihood function:

$$pl(\theta; x) = \frac{L(\theta; x)}{\sup_{\theta' \in \Theta} L(\theta'; x)}, \tag{3}$$

where $L(\cdot; x) : \theta \to f(x; \theta)$ is the likelihood function. The corresponding plausibility function is thus defined as:

$$Pl_\Theta(A; x) = \sup_{\theta \in A} pl(\theta; x) = \frac{\sup_{\theta \in A} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)}, \tag{4}$$

for all $A \subseteq \Theta$.

In [37], Shafer did not present any formal justification for (3) and (4). Moreover, in later writings, he rejected the likelihood-based approach approach on the ground that it is not compatible with Dempster's rule of combination in the case of independent observations [38]. More precisely, assume that $\mathbf{X}$ is an independent sample $(X_1, \ldots, X_n)$. We could combine the n observations at the "aleatory level" by computing $Pl_\Theta(\cdot; \mathbf{x})$ using (4), or we could combine them at the "epistemic level" by first computing the consonant plausibility functions $Pl_\Theta(\cdot; x_i)$ induced by each of the independent observations and using Dempster's rule. Obviously, these two procedures yield different results in general, as consonance is not preserved by Dempster's rule.

In spite of this apparent deficiency, Wasserman [47] provided an axiomatic justification of (3) and (4) in the case where Θ is finite and showed this approach to yield interesting results for realistic inference problems. Other authors questioned the use of Dempster's rule for combining independent observations in statistical inference problems [44][1].

The objective of this paper is two-fold. First, we provide new arguments in favor of the evidential likelihood-based approach, by showing that it can be derived from three principles: the likelihood principle [22], compatibility with Bayesian inference when a prior probability distribution is available and the least-commitment principle [40]. The second goal of this paper is to extend likelihood-based belief functions to low-quality data. Specifically, two cases will considered. In the first case, data are only partially relevant to the problem at hand, a situation previously addressed using, e.g., the concept of weighted likelihood [28, 46]. The second case concerns uncertain data, i.e., data that have been generated by a random mechanism and a non-random, imperfect observation process [17, 18].

The rest of this paper is organized as follows. In Section 2, we provide a new justification for the construction of a consonant belief function based on the likelihood and we demonstrate its application to some simple statistical inference problems. Extensions to partially relevant and uncertain data are then introduced in Sections 3 and 4, respectively. Finally, Section 5 concludes the paper.

2. Likelihood-based belief function

A key argument in favor of the likelihood-based approach to the construction of belief functions for statistical inference is the Likelihood Principle, which will be discussed in Subsection 2.1. Once this principle is accepted, Equations (3) and (4) follow directly from two additional requirements: compatibility with Bayes' rule and the least commitment principle, as will be shown in Subsection 2.2. The practical application of this approach will be demonstrated using some simple inference problems in Subsection 2.5.

2.1. Likelihood principle

As is well know, most approaches to statistical inference fall in two main categories: Bayesian approaches, assuming the existence of a prior probability distribution on Θ and frequentist methods relying on confidence intervals and tests of hypotheses with well-defined long-run frequency properties. Yet, a third tradition can be traced back from Fisher's later work [26] to Barnard [3], Birnbaum [6] and Edward [22], among others. This third approach to statistical inference centers on direct inspection of the likelihood function $L(\theta; x)$ alone, without relying on the concept of repeated sampling (underlying long-run frequency considerations) and without assuming the existence of a prior probability distribution. For proponents of this approach as Birnbaum [6], "reports of experimental results in scientific journals should in principle be descriptions of likelihood functions, when adequate mathematical models can be assumed, rather than reports of significance levels or interval estimates".

The likelihood principle underlies the likelihood-based approach to statistical inference [6]. Let E denote a statistical model representing an experimental situation. Typically, E is composed of the parameter space Θ, the sample space $\mathbb{X}$ and a probability mass or density function $f(x; \theta)$ for each $\theta \in \Theta$. Following Birnbaum [6], let us denote by $Ev(E, x)$ the *evidential meaning* of the specified instance (E, x) of statistical evidence. The likelihood Principle (L) can be stated as follows:

The likelihood principle (L). If E and E' are any two experiments with the same parameter space Θ, represented by probability mass or density functions $f(x;\theta)$ and $g(y;\theta)$, and if x and y are any two respective outcomes which determine likelihood functions satisfying $f(x;\theta) = cg(x;\theta)$ for some positive constant $c = c(x,y)$ and all $\theta \in \Theta$, then $Ev(E,x) = Ev(E',y)$.

As noted by Birnbaum [6], (L) is an immediate consequence of Bayes' principle, which implies that the evidential meaning of (E,x) is contained in the posteriori probability distribution distribution $p(\theta|x) \propto f(x,\theta)p(\theta)$, where $p(\theta)$ is the prior probability distribution. However, it was also accepted as self-evident by statisticians who did not adhere to the Bayesian school, including Fisher [26] and Barnard [3]. From a non Bayesian perspective, it was placed on firm ground by Birnbaum [6], who showed that (L) can be derived from the principles of sufficiency (S) and conditionality (C), which can be stated as follows.

The principle of sufficiency (S). Let E be an experiment, with sample space $\{x\}$, and let $t(x)$ is any sufficient statistic (i.e., any statistic such that the conditional distribution of x given t does not depend on θ). Let E' be an experiment, derived from E, having the same parameter space, such that when any outcome x of E is observed the corresponding outcome $t = t(x)$ of E' is observed. Then for each x, $Ev(,x) = Ev(E',t)$, where $t = t(x)$.

The principle of conditionality (C). If E is mathematically equivalent to a mixture of component experiments E_h, with possible outcomes (E_h, x_h), then $Ev(E,(E_h,x_h)) = Ev(E_h,x_h)$.

In short, (C) means that component experiments that might have been carried out, but in fact were not, are irrelevant once we know that E_h has been carried out. For instance, assume that two measuring instruments provide measurements x_1 and x_2 of an unknown quantity θ. An instrument is picked at random (experiment E). Assume we know that the first instrument ($h = 1$) is selected and we observe x_1. Then, the fact that the second instrument could have been selected is relevant and the over-all structure of the original experiment E can be ignored.

Birnbaum [6] showed that (S) and (C) are jointly equivalent to (L). Unless we reject (S) or (C), which very few statisticians would be willing to do, we are thus compelled to accept (L), i.e., to accept the idea that all the information which the data provide about the parameter is contained in the

likelihood function.

Fisher, who introduced the likelihood function [22, 2], repeatedly stressed that "probability and likelihood are quantities of an entirely different nature" [24] as, in particular, likelihoods are not additive. Yet, Fisher held the view that "in an important class of cases the likelihood may be held to measure the degree of our rational belief in a conclusion" [25]. This conception of a measure of belief that does not obey the rules of probability theory seems to have been welcomed with skepticism in some statistical circles where it was sometimes claimed that the concept of likelihood does not have a clear meaning [22]. The idea of a non additive measure of belief is, of course, more easily understood now than it was in the 1950's. However, the concept of likelihood needs to be linked to that of Dempster-Shafer belief function, as was done on intuitive grounds by Shafer [37] (see also [39]). An attempt to achieve this goal will be presented in the next section.

2.2. From likelihoods to beliefs

In this section, as in the rest of this paper, we accept the Dempster-Shafer theory of belief functions as a suitable framework for representing evidence about any unknown quantity. It follows that statistical evidence x about a parameter θ should be representable by a belief function $Bel_\Theta(\cdot; x)$ defined on the parameter space Θ. According the likelihood principle (L), whose justification has been recalled in the previous section, $Bel_\Theta(\cdot; x)$ should be defined only from the likelihood function. It now remains to be decided which additional requirement should be imposed on $Bel_\Theta(\cdot; x)$ for it to qualify as a sound representation of statistical evidence.

As noticed by Wasserman [47], the most natural requirement is compatibility with Bayesian inference when a Bayesian prior is available. More precisely, let $\pi(\theta)$ be a prior probability mass or density function on Θ, representing what was known about θ before observing the result of the random experiment. Combining it with $Bel_\Theta(\cdot; x)$ using Dempster's rule yields a Bayesian belief function with probability mass of density function given by

$$p(\theta|x) \propto pl(\theta; x)\pi(\theta), \tag{5}$$

where, as before, $pl(\theta; x) = Pl_\Theta(\{\theta\}; x)$ denotes the contour function associated to $Bel_\Theta(\cdot; x)$. Since the posterior probability function on θ verifies

$$p(\theta|x) \propto f(x; \theta)\pi(\theta), \tag{6}$$

it is clear that compatibility with Bayes's rule imposes that

$$pl(\theta; x) = cf(x; \theta) = cL(\theta; x) \tag{7}$$

for some constant $c > 0$ depending only on the likelihood function $L(\cdot; x)$.

Let $\mathcal{B}_x$ denote the set of belief function on θ verifying (7). Assuming that $Bel_\Theta(\cdot; x) \in \mathcal{B}_x$ and in the absence of additional requirements, the Least Commitment Principle [40] leads to selecting the *least informative* element in $\mathcal{B}_x$ (assuming such an element exists), according to some informational ordering between belief functions. As noted in [20], several such meaningful orderings can be defined. In particular, Bel_1 is said to be q-least 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. Intuitively, this ordering can be justified as follows: the commonality function associated to $Bel_1 \oplus Bel_2$, where $\oplus$ denotes Dempster's rule, is proportional to $Q_1 Q_2$. The closer is $Q_1(A)$ to 1, for any $A \subseteq \Theta$, the less influence has the combination with Bel_1 on Bel_2 and, hence, the less informative is Bel_1. Now, the following proposition holds.

Proposition 1. *The q-least committed element in $\mathcal{B}_x$ is the consonant belief function whose contour function is equal to the relative likelihood function:*

$$pl(\theta; x) = \frac{L(\theta; x)}{\sup_{\theta' \in \Theta} L(\theta'; x)}. \tag{8}$$

Proof. Let $Bel_\Theta(\cdot; x)$ be the consonant belief function whose contour function is defined by (8), and let $Q_\Theta(\cdot; x)$ be the corresponding commonality function. Similarly, let $Bel'_\Theta(\cdot; x)$ and $Q'_\Theta(\cdot; x)$ denote any corresponding belief and commonality functions verifying (7) for some constant c. For any $\theta \in \Theta$, we have $Q_\Theta(\{\theta\}; x) = pl(\theta; x)$, $Q'_\Theta(\{\theta\}; x) = pl'(\theta; x)$ and

$$Q_\Theta(\{\theta\}; x) \geq Q'_\Theta(\{\theta\}; x),$$

as $c \leq [\sup_{\theta' \in \Theta} L(\theta'; x)]^{-1}$. Now, let us consider any nonempty $A \subseteq \Theta$ such that $Q'_\Theta(A; x)$ exists. For any $\theta \in A$,

$$Q'_\Theta(A; x) \leq Q'_\Theta(\{\theta\}; x).$$

Hence,

$$Q'_\Theta(A; x) \leq \inf_{\theta \in A} Q'_\Theta(\{\theta\}; x).$$

Now, it follows from the consonance of $Q_\Theta(\cdot; x)$ that

$$Q_\Theta(A; x) = \inf_{\theta \in A} Q_\Theta(\{\theta\}; x) \geq \inf_{\theta \in A} Q'_\Theta(\{\theta\}; x) \geq Q'_\Theta(A; x),$$

which completes the proof. □

The focal sets of $Bel_\Theta(\cdot; x)$ are the levels sets of $pl(\theta; x)$ defined as follows:

$$\Gamma_x(\omega) = \{\theta \in \Theta | pl(\theta; x) \geq \omega\}, \tag{9}$$

for $\omega \in [0, 1]$. These sets may be called plausibility regions and can be interpreted as sets of parameter values whose plausibility is greater than some threshold ω.

$Bel_\Theta(\cdot; x)$ can be regarded as being induced by a random set [33] defined by a probability measure P_Ω on $\Omega = [0, 1]$ and the multi-valued mapping Γ_x, in the sense that

$$Bel_\Theta(A; x) = P_\Omega(\{\omega \in [0, 1] | \Gamma_x(\omega) \subseteq A\}) \tag{10}$$

and

$$Pl_\Theta(A; x) = P_\Omega(\{\omega \in [0, 1] | \Gamma_x(\omega) \cap A \neq \emptyset\}) \tag{11}$$

for all measurable subsets A of Θ. When $pl(\theta; x)$ is continuous, P_Ω can be taken as the Lebesgue measure on Ω.

2.3. Gaussian approximation

The construction of a belief function on Θ from the normalized likelihoods as justified above is in line with Barnard's precept to "have regard to the whole course of the likelihood function" [3]. However, plotting and manipulating the contour function $pl(\theta; x)$ when Θ has more than two dimensions may be difficult. In particular, plausibility regions may be difficult to describe analytically.

This problem can sometimes be tackled by computing a Taylor expansion of the logarithm of the contour function about the maximum likelihood estimate (or maximum plausibility estimate – MPE) $\hat{\theta}$ up to the second order [43]:

$$\log pl(\theta; x) = \log pl(\hat{\theta}; x) + (\theta - \hat{\theta})^T \left.\frac{\partial \log pl(\theta; x)}{\partial \theta}\right|_{\theta=\hat{\theta}} + \frac{1}{2}(\theta - \hat{\theta})^T \left.\frac{\partial^2 \log pl(\theta; x)}{\partial \theta \partial \theta^T}\right|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \cdots \tag{12}$$

The first term on the right-hand side of the above equation is zero by definition, and the second term is zero in the usual case where $\hat{\theta}$ is a stationary point of $pl(\theta; x)$. Neglecting the remaining terms of the Taylor expansion, we get the following approximation

$$pl(\theta; x) \approx \exp\left[-\frac{1}{2}(\theta - \hat{\theta})^T I(\hat{\theta}; x)(\theta - \hat{\theta})\right], \tag{13}$$

where $I(\hat{\theta}; x)$ is the *observed information matrix* defined as

$$I(\hat{\theta}; x) = -\left.\frac{\partial^2 \log pl(\theta; x)}{\partial\theta\partial\theta^T}\right|_{\theta=\hat{\theta}} = -\left.\frac{\partial^2 \log L(\theta; x)}{\partial\theta\partial\theta^T}\right|_{\theta=\hat{\theta}}. \tag{14}$$

Equation (13) defines a Gaussian approximation to the contour function. As noted in [43], this approximation is usually well verified when $X = (X_1, \ldots, X_n)$ is an independent and identically distributed (iid) random vector and n is large.

We may observe that function $pl(\theta; x)$ tends to be all the more "peaked" around $\hat{\theta}$ and, consequently, all the more informative, that the diagonal elements of $I(\hat{\theta}; x)$ are large. It thus makes sense to regard $I(\hat{\theta}; x)$ as reflecting proportional to the amount of information contained in the observations.

2.4. Discussion

Consistency with statistical practice. In support of the method outlined above for constructing belief functions from data, we can remark that viewing the relative likelihood function as the contour function of a consonant belief function or, equivalently, as a possibility distribution [49, 21] is, to a large extent, consistent with statistical practice. For instance, likelihood intervals [29, 43] are focal intervals of the relative likelihood viewed as a possibility distribution. In the case where $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$ is a vector parameter, the marginal contour function on Θ_1 is

$$pl(\theta_1; x) = \sup_{\theta_2 \in \Theta_2} pl(\theta_1, \theta_2; x), \tag{15}$$

which is the relative profile likelihood function when θ_2 is considered as a nuisance parameter. As another example, the usual likelihood ratio statistics $\Lambda(\mathbf{x})$ for a composite hypothesis $H_0 \subset \Theta$ can be seen as the plausibility of H_0, as

$$\Lambda(x) = \frac{\sup_{\theta \in H_0} L(\theta; x)}{\sup_{\theta' \in \Theta} L(\theta'; x)} = \sup_{\theta \in H_0} pl(\theta; x) = Pl_\Theta(H_0; x). \tag{16}$$

Incompatibility with Dempster's rule. As already mentioned above, Shafer [38] questioned the likelihood-based construction of belief functions because it is not compatible with Dempster's rule. One might as well question Dempster's rule for combining independent statistical evidence, as was done by Aickin [1] and Walley [44], among others. Let E and E' be two independent random experiments with the same parameter space Θ, producing outcomes x and y according to frequency distributions $f(x, \theta)$ and $g(y, \theta)$. Let $Bel_\Theta(\cdot; x)$ and $Bel_\Theta(\cdot; y)$ denote the belief functions on Θ obtained after

observing x and y, respectively. It is clear that $Bel_\Theta(\cdot; x) \oplus Bel_\Theta(\cdot; y)$ and $Bel_\Theta(\cdot; xy)$ are different, although they have the same contour function. However, $Bel_\Theta(\cdot; xy)$ can be obtained from $Bel_\Theta(\cdot; x)$ and $Bel_\Theta(\cdot; y)$ using the product rule of Possibility theory [21], which amounts to multiplying the contour functions (or possibility distributions) and renormalizing:

$$pl(\theta; xy) = \frac{pl(\theta; x) pl(\theta; y)}{\sup_{\theta' \in \Theta} pl(\theta'; x) pl(\theta'; y)}. \tag{17}$$

The apparent inadequacy of Dempster's rule in this case remains to be convincingly explained. It might be that different kinds of evidence require different combination mechanisms, as suggested by Dubois et al. in [19].

Inconsistency with the imprecise probability view. Linking the relative likelihood function with more general uncertainty representation formalisms has been attempted by other authors, within different frameworks. For instance, Walley and Moral [45] addressed the problem of defining lower and upper posterior probabilities from likelihoods in the case of a finite parameter space Θ. They showed that the plausibility function defined by (4) does not qualify as a valid upper probability measure because it may be strongly inconsistent, in the sense that it may incur sure loss. However, this objection does not apply in our case, since we are not placing ourselves in the imprecise probability framework. In particular, we do not consider the plausibility function (4) as the upper envelope of a set of probability measures.

2.5. Examples

We will conclude this section with a few examples showing the application of the method outlined above to some simple statistical problems. To avoid technicalities and emphasize basic principles, we will restrict ourselves to the very simple binomial model.

Assume that we observe a random variable X having a binomial distribution:

$$f(x; \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}. \tag{18}$$

The likelihood-based belief function induced by x has the following contour function:

$$pl(\theta; x) = \frac{\theta^x (1-\theta)^{n-x}}{\hat{\theta}^x (1-\hat{\theta})^{n-x}} = \left(\frac{\theta}{\hat{\theta}}\right)^x \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{n-x}, \tag{19}$$

for all $\theta \in \Theta = [0, 1]$, where $\hat{\theta} = x/n$ is the MPE. The observed information is

$$I(\hat{\theta}; x) = \frac{n}{\hat{\theta}(1-\hat{\theta})}, \tag{20}$$

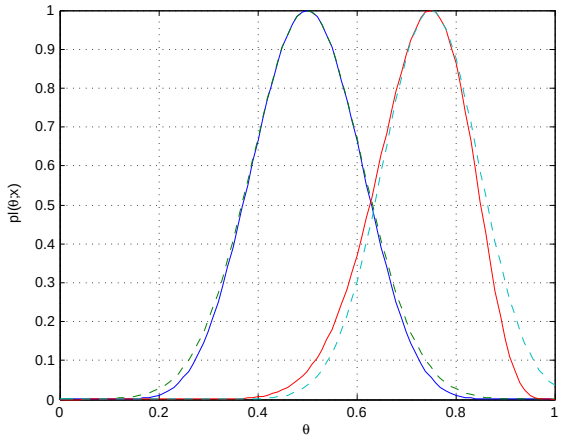

Figure 1: Contour functions (normalized likelihoods) for the binomial distribution (solid lines) and their Gaussian approximations (interrupted lines) with $n = 20$ and $x = 10$ (left) and $x = 15$ (right).

from which we get the following Gaussian approximation of pl_x:

$$pl(\theta; x) \approx \exp\left(-\frac{\hat{\theta}(1-\hat{\theta})(\theta-\hat{\theta})^2}{2n}\right). \tag{21}$$

The contour function $pl(\theta; x)$ and its Gaussian approximation are plotted in Figure 1 for $n = 20$ and $x = 10, 15$. We can see that the Gaussian approximation is already good for small n, especially when the likelihood function is symmetric.

As $pl(\theta; x)$ is unimodal and continuous, each level set $\Gamma_x(\omega)$ for $\omega \in [0, 1]$ is a closed interval and $Bel_\Theta(\cdot; x)$ is a closed random interval [13].

Example 1. *As a first example of a very simple problem, assume that we wish to assess the extent to which the data supports a hypothesis $H \subset [0, 1]$. This may be achieved by computing the plausibility of that hypothesis:*

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

For instance, assume that $H = [\theta_0, \theta_1]$. We have

$$Pl_\Theta([\theta_0, \theta_1]; x) = \begin{cases} pl(\theta_1; x) & \text{if } \theta_1 < \hat{\theta} \\ 1 & \text{if } \theta_0 \leq \hat{\theta} \leq \theta_1, \\ pl(\theta_0; x) & \text{if } \hat{\theta} < \theta_0. \end{cases} \tag{23}$$

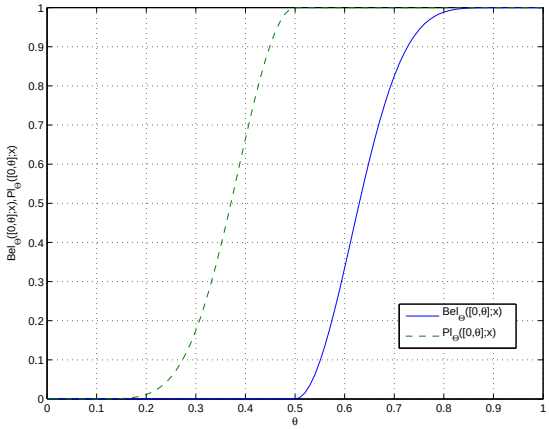

Figure 2: Upper and lower cumulative distributions $Pl_\Theta([0,\theta])$ and $Bel_\Theta([0,\theta]) = 1 - Pl_\Theta((\theta,1])$ for $n = 20$ and $x = 10$.

Table 1: Data of Example 2.

Treatment	S	F	Total
Ramipril	834	170	1004
Placebo	760	222	982
Total	1594	392	1986

The upper and lower cumulative distributions $Pl_\Theta([0,\theta])$ *and* $Bel_\Theta([0,\theta]) = 1 - Pl_\Theta((\theta,1])$ *for* $n = 20$ *and* $x = 10$ *are plotted in Figure 2.*

Example 2. *Let us now assume that we wish to compare two proportions. For instance, Table 1 shows data, reported in [43], from a clinical trial to investigate the efficacy of ramipril in enhancing survival after an acute myocardial infection. There were 1986 subjects, of which 1004 randomly chosen subjects were given ramipril, and the remaining 982 were given a placebo. Let* θ_1 *and* θ_2 *denote the survival probability in the ramipril and control group, respectively. We wish to compute the plausibility that the two probabilities are equal, i.e.,* $Pl(H)$ *with*

$$H = \{(\theta_1, \theta_2) \in \Theta_1 \times \Theta_2 | \theta_1 = \theta_2\}. \tag{24}$$

Let x *and* y *denote the number of survivals in each group; let* Γ_x *and* Γ_y *be the multi-valued mappings corresponding to* $pl(\theta_1; x)$ *and* $pl(\theta_2; y)$ *(Figure*

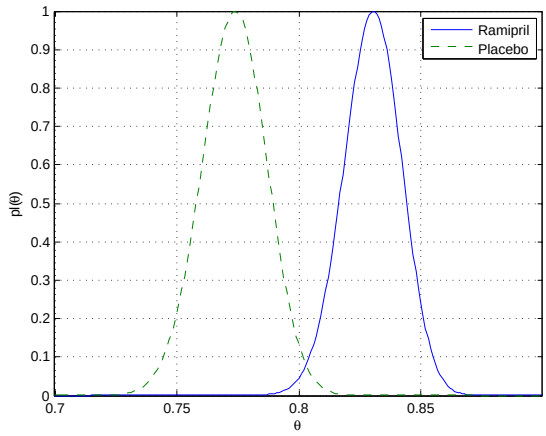

Figure 3: Contour functions (normalized likelihoods) for the data of Example 2.

3). We have

$$\begin{aligned} Pl(H) &= P\left(\{(\omega_1,\omega_2)\in[0,1]^2|\,(\Gamma_x(\omega_1)\times\Gamma_y(\omega_2))\cap H\neq\emptyset\right) && \text{(25a)}\\ &= P\left(\{(\omega_1,\omega_2)\in[0,1]^2|\Gamma_x(\omega_1)\cap\Gamma_y(\omega_2)\neq\emptyset\right) && \text{(25b)}\\ &= 1-P\left(\{(\omega_1,\omega_2)\in[0,1]^2|\Gamma_x(\omega_1)\cap\Gamma_y(\omega_2)=\emptyset\right). && \text{(25c)}\end{aligned}$$

We can see that $Pl(H)$ is equal to one minus the degree of conflict [37] between Bel_x and Bel_y. This quantity can easily be approximated by Monte Carlo simulation. Here, we find $Pl(H)\approx 0.0227$.

3. Partially relevant data

In the above framework, we have assumed, as usually done in statistics, that the data generation process is well defined and that its outcomes are perfectly observed. There are practical situations, however, where such assumptions are not realistic. Sometimes, we are interested in some parameter θ of a certain population, but we collect data from one or several populations that are only known to "resemble" the population of interest. For instance, we may have hospital admissions for different geographical regions that are close together. In such a situation, some of the data are only "partially relevant" to the problem at hand [46]. This situation will be studied in this section.

Assume that we are interested in a parameter $\theta\in\Theta$ related to a certain population and we observe a random variable X with probability density or mass function $f(x;\theta')$, where $\theta'\in\Theta$ is a parameter believed to be "close"

to θ, in some not necessarily well defined sense. Having observed $X = x$, our belief about θ' is represented by the contour function

$$pl'(\theta'; x) = \frac{L(\theta'; x)}{\sup_{\theta'} L(\theta'; x)}. \tag{26}$$

What does x tell us about θ? Arguably, the contour function $pl(\theta; x)$ representing the information on Θ provided by x should be less committed than $pl'(\theta'; x)$, i.e., we should have $pl(\cdot; x) \geq pl'(\cdot; x)$. In the following, we propose two solutions to this problem, described in the following two subsections.

3.1. Weighted likelihood-based belief function

Assume that X is distributed according to $f(x; \theta)$ and we observe n times the same value x in n independent outcomes. The induced contour function would be:

$$pl(\theta; x, \cdots, x) = \frac{L(\theta; x)^n}{\sup_\theta L(\theta; x)^n} = pl(\theta; x)^n. \tag{27}$$

Thus, for a compound observation $x, \ldots, x$ containing n times more information than x, the contour function is raised to power n. Conversely, if x contains n times *less* information than it would do if we had $\theta' = \theta$, it makes sense to define $pl(\theta; x)$ as:

$$pl(\theta; x) = pl'(\theta; x)^w = \frac{L(\theta; x)^w}{\sup_\theta L(\theta; x)^w}, \tag{28}$$

with $w = 1/n$. Equation (28) can be extended to any value $w \in [0, 1]$ and w can be interpreted as a relevance coefficient: $w = 1$ means that x is completely relevant to estimate θ (as it is the case when $\theta' = \theta$), whereas $w = 0$ means that x is totally irrelevant, in which case the belief function $Bel_\Theta(\cdot; x)$ induced by $pl(\theta; x)$ is vacuous. We can observe that the numerator of the fraction on the right-hand side of (28) is the relevance weighted likelihood [28, 46]. Hu and Zidek showed that it can be derived from the maximum entropy principle in the discrete case, and Wang [46] extended this derivation in the continuous case.

Examples of weighted contour functions for the binomial case are plotted in Figure 4.

3.2. Delation of the contour function

An alternative and arguably more principled approach to inference from partially relevant data can now be described as follows. Assume that the statement "θ' is close to θ" can be formalized as $d(\theta, \theta') \leq \delta$, where d is a

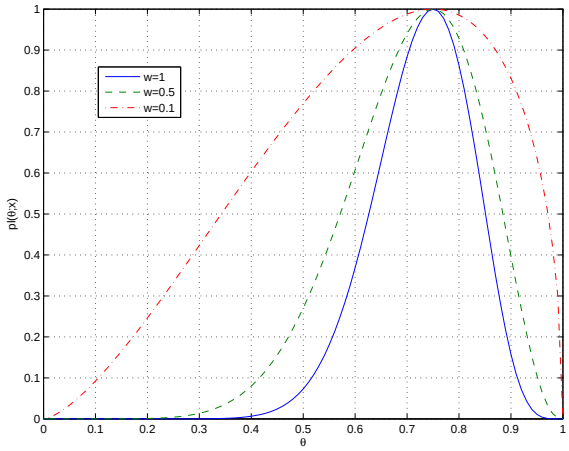

Figure 4: Weighted contour functions for the binomial distribution with $n = 20$ and $x = 15$.

dissimilarity measure defined on Θ and δ is a known constant. This piece of information can be modeled by a logical belief function with focal set $S_\delta = \{(\theta, \theta') | d(\theta, \theta') \leq \delta\} \subset \Theta^2$. Combining it with $pl'(\theta'; x)$ using Dempster's rule yields a consonant belief function on Θ^2, with contour function

$$pl(\theta, \theta'; x) = pl'(\theta'; x) \mathbb{1}_{S_\delta}(\theta, \theta'). \tag{29}$$

Marginalizing out θ' yields the following contour function for θ:

$$pl(\theta; x) = \sup_{\theta'} pl(\theta, \theta'; x) = \sup_{\theta' \in B_\delta(\theta)} pl'(\theta'; x), \tag{30}$$

where

$$B_\delta(\theta) = \{\theta' \in \Theta | d(\theta, \theta') \leq \delta\}. \tag{31}$$

It is obvious that $pl(\theta; x)$ defined by (30) verifies $pl(\cdot; x) \geq pl'(\cdot; x)$. Its focal sets are

$$\begin{aligned} \Gamma_x(\omega) &= \{\theta \in \Theta | pl(\theta; x) \geq \omega\} && \text{(32a)} \\ &= \{\theta \in \Theta | \exists \theta' \in \theta, d(\theta, \theta') \leq \delta \text{ and } pl'(\theta; x) \geq \omega\} && \text{(32b)} \\ &= \bigcup_{\theta' \in \Gamma'_x(\omega)} B_\delta(\theta'), && \text{(32c)} \end{aligned}$$

where $\Gamma'_x(\omega)$ is the ω-level cut of $pl'(\cdot; x)$. Each level set $\Gamma_x(\omega)$ of $pl(\cdot; x)$ is thus obtained from the corresponding level set $\Gamma'_x(\omega)$ of $pl'(\cdot; x)$ by a *delation* operation, as defined in mathematical morphology [36, 7].

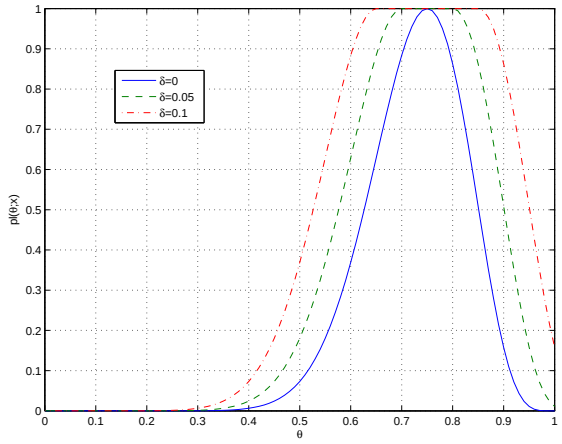

Figure 5: Delated contour functions for the binomial distribution with $n = 20$ and $x = 15$.

Figure 5 shows "delated" contour functions for the binomial distribution with $n = 20$, $x = 15$, $d(\theta, \theta') = |\theta - \theta'|$ and $\delta \in \{0, 0.05, 0.1\}$.

4. Uncertain data

Another "non standard" situation that may arise in practice is that were the data are observed with some uncertainty. For instance, assume that θ is the proportion of patients with some disease in a population. Let X be a Bernoulli variable indicated if a patient randomly selected from the population has the disease. Sometimes, the value of X cannot be determined with certainty. A physician may examine the patient and give, say, a degree plausibility $pl(1)$ that the patient has the disease, and a degree of plausibility $pl(0)$ that he/she does not have the disease. How can we extend the above inference framework to such uncertain data? This issue has been addressed in [16, 17, 18], with emphasis on point estimation using en Expectation-Maximization (EM) algorithm. In Subsection 4.1, we will first reexamine some previously introduced notions from the viewpoint adopted in this paper. A comparison with Bayesian analysis of the same problem will then be undertaken in Subsection 4.2.

4.1. Contour function on Θ induced by uncertain data

We consider in this section the situation where the data x have been generated by a random process but have been imperfectly observed, i.e., *after* the random experiment has taken place and a value x has been generated,

we gather evidence on x. Such evidence can be described by a belief function $Bel_{\mathbb{X}}$ on the sample space $\mathbb{X}$. To simplify exposition, and because the emphasis in this paper is on principles and not on technical aspects, we will assume $\mathbb{X}$ to be finite, so that $Bel_{\mathbb{X}}$ can be described by a mass function $m_{\mathbb{X}}$ on $\mathbb{X}$. All the notions introduced here can be extended to the continuous case, as was done in [17, 18].

It must be stressed that, given x, the uncertain data $m_{\mathbb{X}}$ is *not* assumed to be randomly generated, i.e., no repeatable mechanism for producing mass functions $m_{\mathbb{X}}$ with given frequencies is postulated in our model. This is in sharp contrast with other approaches based on random sets [34] or fuzzy random variables [23], in which a crisp or fuzzy set is assumed to be generated at random. In practice, $m_{\mathbb{X}}$ will usually be obtained from human experts or using indirect methods. A real-world application in the area of technical diagnosis where this formalism has been used has been described in [8].

Definition of the belief function on Θ induced by $m_{\mathbb{X}}$. As our approach is based on likelihoods, let us first extend the likelihood function to uncertain data $m_{\mathbb{X}}$. The likelihood of a hypothesis given data is usually defined as a quantity proportional to the probability of observing the data, given the hypothesis [22]. Here, the data are uncertain, i.e., we do not know exactly what has been observed. Let $(\Omega, 2^{\Omega}, P_{\Omega}, \Gamma)$ denote the finite random set inducing $m_{\mathbb{X}}$, where Ω is seen as a finite set of possible interpretations of the evidence about x. If interpretation ω holds, then the evidence tells us that $x \in \Gamma(\omega)$. The conditional probability of observing this event is

$$P_{\mathbb{X}}(\Gamma(\omega); \theta) = \sum_{x \in \Gamma(\omega)} f(x; \theta). \tag{33}$$

Averaging over ω yields the mean probability:

$$P(m_{\mathbb{X}}; \theta) = \sum_{\omega \in \Omega} P_{\Omega}(\{\omega\}) P_{\mathbb{X}}(\Gamma(\omega); \theta) = \sum_{A \subseteq \mathbb{X}} m_{\mathbb{X}}(A) P_{\mathbb{X}}(A; \theta), \tag{34}$$

which can be seen as the "probability of mass function $m_{\mathbb{X}}$", defined as the mean probability of its focal sets. The likelihood function given the uncertain observation $m_{\mathbb{X}}$ can then be defined as $L(\theta; m_{\mathbb{X}}) = P(m_{\mathbb{X}}; \theta)$ for all $\theta \in \Theta$. It is easy to show that $L(\theta; m_{\mathbb{X}})$ only depends on the contour

function $pl(x)$ associated to $m_{\mathbb{X}}$. To see this, we may write:

$$\begin{aligned} L(\theta; m_{\mathbb{X}}) &= \sum_{A \subseteq \mathbb{X}} m_{\mathbb{X}}(A) \left(\sum_{x \in A} f(x; \theta) \right), && \text{(35a)} \\ &= \sum_{x \in \mathbb{X}} f(x; \theta) \left(\sum_{\{A \subseteq \mathbb{X} | A \ni x\}} m_{\mathbb{X}}(A) \right), && \text{(35b)} \\ &= \sum_{x \in \mathbb{X}} f(x; \theta) pl(x) && \text{(35c)} \\ &= \mathbb{E}_\theta \left[pl(X) \right]. && \text{(35d)} \end{aligned}$$

When $m_{\mathbb{X}}$ is consonant, $pl_{\mathbb{X}}$ can be interpreted as a possibility distribution or, equivalently, as the membership function of the fuzzy set F of values that may be taken by x. $L(\theta; m_{\mathbb{X}})$ is then the probability of that fuzzy set, according to Zadeh's definition for the probability of a fuzzy event [48, 17].

As a natural extension of (3), we propose to represent the information on θ provided by the uncertain data by the consonant plausibility function with the following contour function:

$$pl(\theta; m_{\mathbb{X}}) = \frac{L(\theta; m_{\mathbb{X}})}{\sup_{\theta \in \Theta} L(\theta; m_{\mathbb{X}})}. \tag{36}$$

It is obvious that (36) is a proper generalization of (8), which is recovered when $m_{\mathbb{X}}$ is a logical mass function with focal set $\{x\}$.

An iterative procedure for finding a value $\widehat{\theta}$ of θ that maximizes $pl(\theta; m_{\mathbb{X}})$ has been introduced in [9] and generalized in [16, 18]. This procedure, called the Evidential Expectation Maximization ($\mathrm{E}^2\mathrm{M}$) algorithm, is an extension of the EM algorithm [15].

Independence assumptions. Let us assume that the observable data are a random vector $\mathbf{X} = (X_1, \ldots, X_n)$, where each X_i is a random variable taking values in $\mathbb{X}_i$. Similarly, its realization can be written as $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{X} = \mathbb{X}_1 \times \ldots \times \mathbb{X}_n$. Two different independence assumptions can then be made:

1. Under the *stochastic independence* of the random variables $X_1, \ldots, X_n$, the probability mass function $f(\mathbf{x}; \theta)$ can be decomposed as:

$$f(\mathbf{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta), \tag{37}$$

for all $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{X}$;

2. Under the *cognitive independence* of $x_1, \ldots, x_n$ with respect to m (see [37, page 149]), we can write:

$$pl(\mathbf{x}) = \prod_{i=1}^{n} pl_i(x_i), \tag{38}$$

for all $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{X}$, where pl_i is the contour function corresponding to the mass function $m_{\mathbb{X}_i}$ obtained by marginalizing $m_{\mathbb{X}}$ on $\mathbb{X}_i$.

We can remark here that the two assumptions above are totally unrelated as they are of different natures: stochastic independence of the random variables X_i is an objective property of the random data generating process, whereas cognitive independence pertains to our state of knowledge about the unknown realization $\mathbf{x}$ of $\mathbf{X}$.

If both assumptions hold, the likelihood criterion (35c) can be written as a product of n terms:

$$L(\theta; m_{\mathbb{X}}) = \prod_{i=1}^{n} \mathbb{E}_\theta \left[pl_i(X_i) \right] \tag{39}$$

and $pl(\theta; m_{\mathbb{X}})$ can be obtained as the normalized product of the contour functions $pl(\theta; m_{\mathbb{X}_i})$:

$$pl(\theta; m_{\mathbb{X}}) = \frac{\prod_{i=1}^{n} pl(\theta; m_{\mathbb{X}_i})}{\sup_{\theta \in \Theta} \prod_{i=1}^{n} pl(\theta; m_{\mathbb{X}_i})}, \tag{40}$$

which generalizes (17).

Example 3. *Assume that $\mathbf{X} = (X_1, \ldots, X_n)$ is an iid random vector and each X_i has a Bernoulli distribution with parameter θ. Let $pl_i(1)$ and $pl_i(0)$ denote, respectively, the plausibilities that $x_i = 1$ and $x_i = 0$. We have*

$$\mathbb{E}_\theta \left[pl_i(X_i) \right] = \theta pl_i(1) + (1 - \theta) pl_i(0) \tag{41}$$

and, assuming that (38) holds:

$$pl(\theta; m_{\mathbb{X}}) = \frac{\prod_{i=1}^{n} \left[\theta pl_i(1) + (1 - \theta) pl_i(0) \right]}{\sup_{\theta \in \Theta} \prod_{i=1}^{n} \left[\theta pl_i(1) + (1 - \theta) pl_i(0) \right]}. \tag{42}$$

Figure 6 shows the contour functions for $n = 20$, $pl_i(1) = 1$, $pl_i(0) = r$ for $i = 1, \ldots, 10$, and $pl_i(1) = r$, $pl_i(0) = 1$ for $i = 11, \ldots, 20$. Clearly, there is no data uncertainty when $r = 0$, in which case we recover the usual contour function with $\hat{\theta} = 0.5$. The uncertainty increases as $r \to 1$. In the limit case where $r = 1$, the belief function $Bel_\Theta(\cdot; m_{\mathbb{X}})$ becomes vacuous.

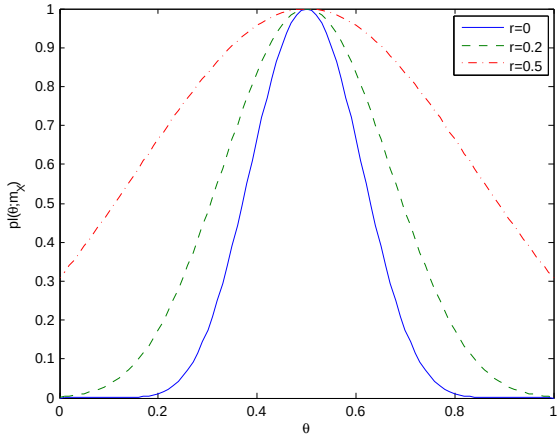

Figure 6: Contour functions for the Bernoulli distribution with uncertain data.

4.2. Comparison with Bayesian analysis

It is interesting to study how the problem considered in this section can be treated in the Bayesian framework. The posterior probability distribution given $m_{\mathbb{X}}$ can be written as

$$f(\theta|m_{\mathbb{X}}) = \sum_{x\in\mathbb{X}} f(\theta|x, m_{\mathbb{X}})p(x|m_{\mathbb{X}}). \tag{43}$$

We have

$$f(\theta|x, m_{\mathbb{X}}) = f(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{p(x)}, \tag{44}$$

where π is the prior probability distribution on θ and $p(x) = \int f(x|\theta)\pi(\theta)d\theta$. Now,

$$p(x|m_{\mathbb{X}}) = \sum_{A\subseteq\mathbb{X}} p(x|A)m_{\mathbb{X}}(A). \tag{45}$$

From the indifference principle:

$$p(x|A) = \frac{\mathbb{1}_A(x)}{|A|}, \tag{46}$$

where $|A|$ denotes the cardinality of A. Hence, we can see that

$$p(x|m_{\mathbb{X}}) = \sum_{\{A\subseteq\mathbb{X}|x\in A\}} \frac{m_{\mathbb{X}}(A)}{|A|} = betp_{\mathbb{X}}(x), \tag{47}$$

where $betp_{\mathbb{X}}$ is the pignistic probability distribution [42, 41] associated to $m_{\mathbb{X}}$.

Consequently, the following equalities hold:

$$f(\theta|m_{\mathbb{X}}) = \sum_{x\in\mathbb{X}} f(\theta|x) betp_{\mathbb{X}}(x) = \pi(\theta) \sum_{x\in\mathbb{X}} f(x|\theta) \frac{betp_{\mathbb{X}}(x)}{p(x)}. \tag{48}$$

By comparing Equations (48) and (35c), and it is clear that $f(\theta|m_{\mathbb{X}})$ and $pl(\theta|m_{\mathbb{X}})$ are not proportional, in general, even when the prior distribution π is uniform. Hence, the belief function and Bayesian frameworks will lead to different inferences about θ, in general. An argument in favor of the belief function approach can be found by considering the impact of a refinement of the sample space.

As an example, let us consider the simple Bernoulli model already studied in Example 3. Assume that $X = 1$ means that a patient taken at random from a given population has some disease D, and $X = 0$ means that the patient does not have the disease. After examining a patient, a physician gives a mass function $m_{\mathbb{X}}$ quantifying his/her beliefs about the state of that particular patient. Now, after some time, we discover that disease D exists in two equiprobable forms D_a and D_b. Should this piece of information change our beliefs on θ? We can argue that it should not, as evidence collected about the patient does not point specifically to D_a or D_b anyway. Let us see this problem can be formally treated in the two models.

Let $\mathbb{X} = \{0, 1\}$ be the initial sample space, and let $\mathbb{X}' = \{0, 1a, 1b\}$ be the refined sample space, with obvious notations. Let us introduce the following notations: $m_{\mathbb{X}}(\{0\}) = u$, $m_{\mathbb{X}}(\{1\}) = v$ and $m_{\mathbb{X}}(\mathbb{X}) = w$. The same piece of evidence can expressed in the refined space $\mathbb{X}'$ as $m_{\mathbb{X}'}(\{0\}) = u$, $m_{\mathbb{X}'}(\{1a, 1b\}) = v$ and $m_{\mathbb{X}'}(\mathbb{X}') = w$.

Let us first consider the belief function analysis. We have

$$pl(\theta; m_{\mathbb{X}}) \propto \theta pl_{\mathbb{X}}(1) + (1-\theta) pl_{\mathbb{X}}(0) = \theta(v+w) + (1-\theta)(u+w) \tag{49}$$

and

$$\begin{aligned} pl(\theta; m_{\mathbb{X}'}) &\propto \frac{\theta}{2} pl_{\mathbb{X}'}(1a) + \frac{\theta}{2} pl_{\mathbb{X}'}(1b) + (1-\theta) pl_{\mathbb{X}'}(0) \\ &= \frac{\theta}{2}(v+w) + \frac{\theta}{2}(v+w) + (1-\theta)(u+w) \\ &= \theta(v+w) + (1-\theta)(u+w) \end{aligned} \tag{50}$$

Consequently, $pl(\theta; m_{\mathbb{X}}) = pl(\theta; m_{\mathbb{X}'})$, as expected: refining the sample space after the observation has been made does not change the inference about θ.

Let us now conduct the Bayesian analysis of the same problem. We have

$$f(\theta|m_{\mathbb{X}}) = f(\theta|1)betp_{\mathbb{X}}(1) + f(\theta|0)betp_{\mathbb{X}}(0). \tag{51}$$

Assuming a uniform prior on Θ, we have

$$f(\theta|1) = \frac{f(1|\theta)\pi(\theta)}{\int f(1|\theta)\pi(\theta)d\theta} = 2\theta \tag{52}$$

and

$$f(\theta|0) = \frac{f(0|\theta)\pi(\theta)}{\int f(0|\theta)\pi(\theta)d\theta} = 2(1-\theta), \tag{53}$$

hence:

$$f(\theta|m_{\mathbb{X}}) = 2\theta(v + \frac{w}{2}) + 2(1-\theta)(u + \frac{w}{2}). \tag{54}$$

Considering the refined sample space, we have

$$f(\theta|m_{\mathbb{X}'}) = f(\theta|1a)betp_{\mathbb{X}'}(1a) + f(\theta|1b)betp_{\mathbb{X}'}(1b) + f(\theta|0)betp_{\mathbb{X}'}(0). \tag{55}$$

We can easily check that $f(\theta|1a) = f(\theta|1b) = 2\theta$, hence

$$f(\theta|m_{\mathbb{X}'}) = 2\theta(v + \frac{2w}{3}) + 2(1-\theta)(u + \frac{w}{3}). \tag{56}$$

Comparing (54) and (56), we can see that refining the sample space a posteriori has changed the posterior probability distribution of θ, a result that can seriously be questioned.

5. Conclusions

In the classical view of likelihood-based inference, the likelihood function is defined up to a multiplicative constant, and the likelihood of a hypothesis is meaningless: only the likelihood ratios have meaning. Furthermore, the likelihood of a compound hypothesis, defined as the disjunction of several simple hypotheses, is generally considered to be undefined, because a compound hypothesis does not specify numerically the probability of the observations. As noted by Sprott [43, page 13], "The fact that a likelihood of a disjunction of exclusive alternatives cannot be determined from the likelihoods of the individual values gives rise to the principal difficulty in using likelihoods for inference".

The method for transforming the likelihood function into a consonant belief function, introduced by Shafer [37] and revisited in this paper, resolves this difficulty. In this paper, we have provided some new arguments in

support of this approach, by showing that it can be derived from three basic principles: the likelihood principle, compatibility with Bayesian inference, and the least commitment principle. We have also shown that this method can be easily generalized to handle data that are only partially relevant to the population of interest, or that have been acquired through an imperfect observation process. Although the method has been demonstrated here using the simplest binomial model, it has been successfully applied to more complex models such as Gaussian mixture models [9, 18], independent factor analysis [8] and hidden Markov models [35].

One of the main advantages of expressing statistical evidence in the belief function framework is the possibility to combine it with expert opinions expressed in the same language. An example of such combination for quantifying the uncertainty of sea level rise due to climate change has been presented in [4, 5], and further work along this line is under way.

References

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

[2] J. Aldrich. R. A. Fisher and the making of the maximum likelihood 1912-1922. *Statistical Science*, 12:162–176, 1997.

[3] 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.

[4] N. Ben Abdallah, N. Mouhous Voyneau, and T. Denœux. Combining statistical and expert evidence within the D-S framework: Application to hydrological return level estimation. In T. Denoeux and M.-H. Masson, editors, *Belief functions: theory and applications (Proc. of the 2nd Int. Conf. on Belief Functions)*, number AISC 164 in Advances in Intelligent and Soft Computing, pages 393–400, Compiègne, France, 2012. Springer.

[5] 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 (accepted)*, 2013.

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

[7] I. Bloch. Defining belief functions using mathematical morphology – application to image fusion under imprecision. *International Journal of Approximate Reasoning*, 48(2):437–465, 2008.

[8] Z. L. Cherfi, L. Oukhellou, E. Côme, T. Denœux, and P. Aknin. Partially supervised independent factor analysis using soft labels elicited from multiple experts: Application to railway track circuit diagnosis. *Soft Computing*, 16(5):741–754, 2012.

[9] E. Côme, L. Oukhellou, T. Denœux, and P. Aknin. Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition*, 42(3):334–348, 2009.

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

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

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

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

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

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38, 1977.

[16] T. Denœux. Maximum likelihood from evidential data: an extension of the EM algorithm. In C. Borgelt et al., editor, *Combining soft computing and statistical methods in data analysis (Proceedings of SMPS 2010)*, Advances in Intelligent and Soft Computing, pages 181–188, Oviedeo, Spain, 2010. Springer.

[17] T. Denœux. Maximum likelihood estimation from fuzzy data using the fuzzy EM algorithm. *Fuzzy Sets and Systems*, 18(1):72–91, 2011.

[18] T. Denœux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):119–130, 2013.

[19] D. Dubois, S. Moral, and H. Prade. *Belief change*, volume 3 of *Handbook of defeasible reasoning and uncertainty management systems*, chapter Belief change rules in ordinal and numerical uncertainty theories, pages 311–392. Kluwer Academic Publishers, Boston, 1998.

[20] 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.

[21] D. Dubois and H. Prade. *Possibility Theory: An approach to computerized processing of uncertainty*. Plenum Press, New-York, 1988.

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

[23] M. B. Ferraro, R. Coppi, G. González Rodríguez, and A. Colubi. A linear regression model for imprecise response. *International Journal of Approximate Reasoning*, 51(7):759–770, 2010.

[24] R. A. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

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

[26] R. A. Fisher. *Statistical methods and scientific inference*. Oliver and Boyd, Edinburgh, 1956.

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

[28] F. Hu and J. V. Zidek. The relevance weighted likelihood with applications. in: Empirical bayes and likelihood inference. In S. E. Ahmed and N Reid, editors, *Empirical Bayes and Likelihood Inference*, pages 211–235. Springer Verlag, New York, 1997.

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

[30] Duncan Ermini Leaf and Chuanhai Liu. Inference about constrained parameters using the elastic belief method. *International Journal of Approximate Reasoning*, 53(5):709 – 727, 2012.

[31] R. Martin and C. Liu. Inferential models: A framework for prior-free posterior probabilistic inference. *Journal of the American Statistical Association (to appear)*, 2012.

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

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

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

[35] E. Ramasso and T. Denœux. Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions. *IEEE Transactions on Fuzzy Systems (submitted)*, 2013.

[36] J. Serra. *Image Analysis and Mathematical Morphology*, volume 1. Academic Press, London, 1982.

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

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

[39] Ph. Smets. Possibilistic inference from statistical data. In *Second World Conference on Mathematics at the service of Man*, pages 611–613, Universidad Politecnica de Las Palmas, 1982.

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

[41] Ph. Smets. Decision making in the TBM: the necessity of the pignistic transformation. *International Journal of Approximate Reasoning*, 38:133–147, 2005.

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

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

[44] P. Walley. Belief function representations of statistical evidence. *The Annals of Statistics*, 15(4):1439–1465, 1987.

[45] P. Walley and S. Moral. Upper probabilities based on the likelihood function. *Journal of the Royal Statistical Society B*, 161:831–847, 1999.

[46] S. X. Wang. *Maximum Weighted Likelihood Estimation.* PhD thesis, University of Bristish Columbia, Department of Statistics, 2001.

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

[48] L. A. Zadeh. Probability measures of fuzzy events. *J. Math. Analysis and Appl.*, 10:421–427, 1968.

[49] L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.