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

Thierry Denœux

Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc

Abstract

This note is a rejoinder to comments by Dubois and Moral about my paper "Likelihood-based belief function: justification and some extensions to low-quality data" published in this issue. The main comments concern (1) the axiomatic justification for defining a consonant belief function in the parameter space from the likelihood function and (2) the Bayesian treatment of statistical inference from uncertain observations, when uncertainty is quantified by belief functions. Both issues are discussed in this note, in response to the discussants' comments.

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

I thank the discussants, Dr. Dubois and Prof. Moral, for there insightful and thought-provoking comments about my paper "Likelihood-based belief function: justification and some extensions to low-quality data" published in this issue [6][7][11]. Both discussants questioned some of the axioms I proposed in order to justify the method for defining a consonant belief function from the likelihood function introduced by Shafer [13]. Additionally, Moral [11] argued for a Bayesian analysis of statistical inference from uncertain observations, different from the approach presented in [6]. These two issues are further examined below, in response to the discussants' comments.

1. Axiomatic justification of likelihood-based belief functions

In [13], Shafer introduced a method for modeling statistical evidence using a consonant belief function. In [6], I attempted to justify this method axiomatically by reasoning as follows. 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. Having observed a realization x of X, we wish to quantify the uncertainty on Θ using a belief function $Bel_\Theta(\cdot; x)$. Which properties should $Bel_\Theta(\cdot; x)$ have? In [6], I proposed the following three requirements:

1. Likelihood principle: $Bel_\Theta(\cdot; x)$ should only depend on the likelihood function, defined by $L(\theta; x) = \alpha f(x; \theta)$ for all $\theta \in \Theta$, where α is any positive multiplicative constant.
2. Compatibility with Bayesian inference: if a Bayesian prior $\pi(\theta)$ is available, combining it with $Bel_\Theta(\cdot; x)$ using Dempster's rule should yield the Bayesian posterior.
3. Least Commitment Principle: $Bel_\Theta(\cdot; x)$ should be the least committed belief function, among all those satisfying the previous two requirements.

The first two requirements jointly imply that the contour function $pl(\theta; x)$ associated to $Bel_\Theta(\cdot; x)$ should be proportional to the likelihood function:

$$pl(\theta; x) \propto L(\theta; x). \tag{1}$$

Assuming the commonality ordering [8] to be relevant for comparing the informational content of belief functions, the third requirement, together with the other two, then yields a unique solution, which is the consonant belief function whose contour function is the relative likelihood function:

$$pl(\theta; x) = \frac{L(\theta; x)}{\sup_{\theta' \in \Theta} L(\theta'; x)}. \tag{2}$$

Dubois [7] and Moral [11] both acknowledge that this formal result provides a justification for viewing the normalized likelihood function as defining a consonant belief function, a view that has been shared by several authors (see, e.g., [17], [21], [1], among others). However, they also claim that the axioms that support this solution (in particular, requirements 2 and 3 above) are questionable and that, consequently, they are not sufficiently compelling to rule out alternative methods of statistical inference in the belief function framework.

I can only agree with the discussants on this point. Indeed, I never claimed that the approach advocated in [6] is the only acceptable one for statistical inference in the belief function framework. As a matter of fact, formal requirements justifying a solution to some problem are rarely unquestionable or self-evident (see, e.g., the long-lasting debates on Savage's or Cox's axioms [15], [19], [16]). The axiomatic approach only lays bare some basic principles, from which the solution can be derived. These principles should be as "natural" or "reasonable" as possible. However, accepting or rejecting them is a matter of personal judgement.

Dubois [7] is, of course, right to point out that the informational ordering of belief functions based on their commonalities is not the only "reasonable" one. Other examples are the ordering based on plausibilities, the specialization ordering, as well as orderings based on the canonical decomposition of belief functions, as introduced in [5]. None of these orderings seems to qualify as the unique way of comparing the information content of belief functions. Similarly, there is no unique equivalent of Shannon's entropy for measuring the degree of uncertainty of belief functions (see, e.g., [12], [10]). Consequently, any application of the Least commitment Principle [9, 18] can be criticized as being based on a more or less arbitrary choice of an ordering or an uncertainty measure. Some arguments in favor of one informational ordering or another might be found in the future. It seems unlikely, though, that such arguments will be strong enough to discard other solutions. I am more inclined to accept the coexistence of several notions of information content for belief functions, just as there exist several notions of independence (see, e.g., [2], [3], [4]).

In his comments, Moral [11] focuses his criticism on the second requirement, namely, that combining the belief function induced by observations with a Bayesian prior *using Dempster's rule* yield the Bayesian posterior. Indeed, the belief function induced by two independent samples is not equal to the orthogonal sum of the belief functions induced by each of the samples. As noted in [6], this remark suggests that Dempster's rule is not the only reasonable mechanism for pooling statistical evidence, which leaves open the possibility of defining other ways of combining a Bayesian prior with a belief function induced by the data. Starting from this observation, Moral [11] proposes two alternatives to the consonant belief function induced by (2). However, while these two alternative solutions make sense in the imprecise probability framework, when interpreting upper probabilities as lower selling prices [20], they appear to be weakly justified in the belief function framework. Actually, one of the proposed solution (denoted by $\overline{P}_4$) is not a plausibility function, and it is not known how consistency of the other solution (denoted by pl_3) with Bayesian inference could be achieved.

To conclude this section, we might observe that, while the discussants' criticisms are partly well-founded, they do not point to any better justified alternative to the method advocated in [6] for representing statistical evidence in the belief function framework.

2. Bayesian analysis of inference from uncertain data

In [6], I considered the situation where the data x are imperfectly observed, and uncertainty about x is described by a mass function $m_{\mathbb{X}}$ induced by a finite random set $(\Omega, 2^{\Omega}, P_{\Omega}, \Gamma)$ (to simplify the discussion, the sample space $\mathbb{X}$ is assumed to be finite). I then proposed an extension of the likelihood function to such uncertain data, as:

$$L(\theta; m_{\mathbb{X}}) = \sum_{x \in \mathbb{X}} f(x; \theta) pl(x), \tag{3}$$

where $pl(x)$ is the contour function associated to $m_{\mathbb{X}}$. This corresponding contour function in the parameter space is then

$$pl(\theta; m_{\mathbb{X}}) = \frac{L(\theta; m_{\mathbb{X}})}{\sup_{\theta \in \Theta} L(\theta; m_{\mathbb{X}})}, \tag{4}$$

which is a proper generalization of (2). In [6], I also proposed a Bayesian analysis of this problem, and arrived at the conclusion that the posterior probability function $f(\theta | m_{\mathbb{X}})$ given $m_{\mathbb{X}}$ is not proportional to $pl(\theta; m_{\mathbb{X}})$, which means that the Dempster-Shafer and Bayesian analyses of this problem yield different solutions.

In his comments [11], Moral proposes an alternative Bayesian analysis, which he claims to be “the most natural” and which yields a solution equivalent to (3). To clarify the discussion, we need to examine in detail the assumptions underlying the two solutions.

Moral models a piece of evidence as a propositional variable E depending on x and ω, such that

$$P(E | x, \omega) = \begin{cases} 1 & \text{if } x \in \Gamma(\omega), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Assuming (a) the independence of $\mathbb{X}$ and Ω and (b) the conditional independence of E and θ given x and ω, he then finds $P(E|\theta)$ to be identical to $L(\theta; m_{\mathbb{X}})$ in (3). The posterior probability mass or density function $f(\theta|E)$ (with our notation) is then proportional to $P(E|\theta)\pi(\theta)$.

The meaning of these assumptions can be better understood using a simple concrete example. Consider an urn containing red and green balls. A ball is drawn from the urn, and an observer reports that the ball is green. However, there is 20 % chance that the observer is color-blind. What can we say about the proportion θ of green balls?

Here, $x = 1$ if the drawn ball is green, and $x = 0$ otherwise; $\Omega = \{c, nc\}$, where c stands for "color-blind" and nc for "not color-blind"; and the probability distribution on Ω is defined by $p(c) = 0.2$ and $p(nc) = 0.8$. If the observer is not color-blind, we know that the ball that has been drawn from the urn is green; otherwise, we only know that it is either red or green. Such evidence can then be modeled by a multi-valued mapping Γ from Ω to $2^{\mathbb{X}}$ such that $\Gamma(nc) = \{1\}$ and $\Gamma(c) = \{0, 1\}$.

In this example, it can be assumed that E corresponds to the proposition "The observer reports that the ball is green". The two independence assumptions make sense: the fact that the observer is color-blind does not depend on the color of the ball being drawn from the urn, and the probability that the observer will report the ball to be green does not depend on the proportion of green balls in the urn, if we already know the color of the ball as well as the state of the observer (color-blind or not). Assumption (5), though, is more problematic. In this example, it implies that $P(E|0, c) = 1$ and $P(E|1, c) = 1$, which means that, if the observer is color-blind, he will report for sure the ball to be green, whatever its real color. This assumption is quite strong and it does not seem to be implied by the initial story. We may simply not know what a color-blind observer would report if the ball was green or red.

The analysis presented in [6] is based on the interpretation of a piece of evidence as a message with uncertain meaning [14]. Here, the report that the ball is green can be seen as a message with two possible meanings: if the observer is not color-blind, it means that the ball is really green; otherwise, it means that the ball is either red or green. Let us detail the analysis in [6] by explicitly introducing the meaning M of the message, taking values in $2^{\mathbb{X}}$. We note that the introduction of this variable taking values in the power set of $\mathbb{X}$ is not artificial, as suggested in [11]; it is inherent in the interpretation of belief functions as representing pieces of evidence. The relation between ω and M given E is determined by the multi-valued mapping Γ:

$$p(M = A|E, \omega) = \begin{cases} 1 & \text{if } \Gamma(\omega) = A, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Now,

$$f(\theta|E) = \sum_{\omega \in \Omega, A \subseteq \mathbb{X}} f(\theta|\omega, M = A, E) p(\omega, M = A|E). \tag{7}$$

We have

$$p(\omega, M = A|E) = p(M = A|\omega, E)p(\omega|E) \tag{8a}$$

$$= \begin{cases} m_{\mathbb{X}}(A) & \text{if } \Gamma(\omega) = A, \\ 0 & \text{otherwise,} \end{cases} \tag{8b}$$

and $f(\theta|\omega, M = A, E) = f(\theta|M = A)$ (once the meaning of the message is known, the state of the observer becomes irrelevant). Consequently,

$$f(\theta|E) = \sum_{A \subseteq \mathbb{X}} f(\theta|M = A) m_{\mathbb{X}}(A). \tag{9}$$

Now, knowing that the meaning of the message is A is the same as knowing that $X \in A$, and nothing more. Hence,

$$f(\theta|M = A) = f(\theta|X \in A) = \frac{P_{\mathbb{X}}(A|\theta)\pi(\theta)}{P_{\mathbb{X}}(A)}, \tag{10}$$

and

$$f(\theta|E) = \pi(\theta) \sum_{A \subseteq \mathbb{X}} \frac{P_{\mathbb{X}}(A|\theta)}{P_{\mathbb{X}}(A)} m_{\mathbb{X}}(A), \tag{11}$$

which is identical to Equation (42) in [6]. In the example considered here, we have, assuming a uniform prior:

$$f(\theta|E) = \frac{P_{\mathbb{X}}(\{1\}|\theta)}{P_{\mathbb{X}}(\{1\})} m_{\mathbb{X}}(\{1\}) + \frac{P_{\mathbb{X}}(\{0,1\}|\theta)}{P_{\mathbb{X}}(\{0,1\})} m_{\mathbb{X}}(\{0,1\}). \tag{12}$$

As $P_{\mathbb{X}}(\{1\}|\theta) = \theta$ and $P_{\mathbb{X}}(\{1\}) = \int \theta\pi(\theta)d\theta = 1/2$, we have

$$f(\theta|E) = 1.6\theta + 0.2. \tag{13}$$

In contrast, the generalized likelihood function (3) is

$$L(\theta|E) = L(\theta|1)pl(1) + L(\theta|0)pl(0) \tag{14a}$$

$$= \theta + 0.2(1 - \theta) \tag{14b}$$

$$= 0.8\theta + 0.2. \tag{14c}$$

The Bayesian analysis (with the above assumptions) and the D-S analysis thus yield different results even in the case of a uniform prior, as claimed in [6].

References

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

[2] B. Ben Yaghlane, Ph. Smets, and K. Mellouli. Belief function independence: I. the marginal case. *International Journal of Approximate Reasoning*, 29:47–70, 2002.

[3] B. Ben Yaghlane, Ph. Smets, and K. Mellouli. Belief function independence: II. the conditional case. *International Journal of Approximate Reasoning*, 31:31–75, 2002.

[4] I. Couso and S. Moral. Independence concepts in evidence theory. *International Journal of Approximate Reasoning*, 51(7):748–758, 2010.

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

[6] T. Denœux. Likelihood-based belief function: justification and some extensions to low-quality data. *International Journal of Approximate Reasoning (in press)*, 2014. http://dx.doi.org/10.1016/j.ijar.2013.06.007.

[7] D. Dubois. On various ways of tackling incomplete information in statistics. *International Journal of Approximate Reasoning*, this issue, 2014.

[8] 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.

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

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

[11] S. Moral. Comments on "Likelihood-based belief function: Justification and some extension to low-quality data". *International Journal of Approximate Reasoning*, this issue, 2014.

[12] N. R. Pal, J. C. Bezdek, and R. Hemasinha. Uncertainty measures for evidential reasoning I: A review. *International Journal of Approximate Reasoning*, 7:165–183, 1992.

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

[14] G. Shafer. Constructive probability. *Synthese*, 48(1):1–60, 1981.

[15] G. Shafer. Savage revisited (with discussion). *Statistical Science*, 1:463–501, 1986.

[16] G. Shafer. Comments on "Constructing a logic of plausible inference: a guide to Cox's Theorem", by Kevin S. Van Horn . *International Journal of Approximate Reasoning*, 35(1):97–105, 2004.

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

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

[19] K. S. Van Horn. Constructing a logic of plausible inference: a guide to Cox's theorem. *International Journal of Approximate Reasoning*, 34(1):3–24, 2003.

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

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