Statistical inference from ill-known data using belief functions

Thierry Denœux
UMR CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529 - F-60205 Compiègne cedex - France
Thierry.Denoeux@hds.utc.fr

January 2, 2013

1 Introduction

Whereas current research in statistics and econometrics mainly focuses on the development of more complex models and inference procedures, data quality is recognized by applied statisticians as a key factor influencing the validity of the conclusions drawn from a statistical analysis. As noted by Cox [5], "issues of data quality and relevance, while underemphasized in the theoretical statistical and econometric literature, are certainly of great concern in much statistical work". Arguing for better consideration of empirical practice in econometric theory, Heckman [21] also remarked that "Data quality, data collection and economic interpretation of statistical evidence are perceived as topics off limits to econometricians, but central to the field of empirical economics".

One of the reasons why data quality, in spite of its importance, has received relatively little attention in the statistical literature, may be that its evaluation often requires subjective judgements that do not easily fit with the standard likelihood-based or Bayesian frameworks. While the latter approach allows for the introduction of personalistic prior information, it does so in a very specific and questionable manner (by treating all unknown quantities as random variables), which raises a number of theoretical and practical issues [35, 15].

In the past thirty years, alternatives to the Bayesian framework for reasoning from weak information have emerged, including Possibility Theory [38], Imprecise Probabilities [35] and the theory of Belief Functions [7, 24]. In particular, the latter approach, also referred to as Dempster-Shafer or Evidence theory, was introduced by Dempster [6, 8] with the objective to reconcile Bayesian and fiducial inference. Shafer [24] later formalized this approach as a general method for representing and combining evidence, not necessarily statistical. Smets [28, 32] emphasized the singularity of the theory of belief functions as opposed to related but distinct frameworks such as imprecise probabilities [35] and random sets [23].

The main feature of theory of belief function is that is subsumes both the logical and probabilistic approaches to uncertainty: a belief function may be seen as a non-additive probability measure [24] and as a generalized set [17]. Also, basic mechanisms for reasoning with belief functions extend both probabilistic operations (such as marginalization and conditioning) and set-theoretic operations (such as intersection and union). In particular, the belief function approach coincides with the Bayesian approach when all variables are described by probability distributions, while allowing for considerably more flexibility when the available knowledge does not allow for the specification of a reasonable probability distributions without introducing unsupported assumptions.

In this paper, the theory of belief function is advocated as a suitable framework for statistical analysis of low quality, i.e., imprecise and/or partially reliable data. The main concepts of the theory will first be recalled in Section 2 and its application to the representation of statistical evidence will be discussed in Section 3. The use of belief functions for representing data uncertainty and corresponding inferential procedures will be introduced in Section 4. Finally, Section 5 will conclude the paper with a summary of the main results and the presentation of some research challenges.

2 Belief functions

This section recalls the necessary background notions related to Dempster-Shafer theory. Belief functions on finite domains and Dempster's rule of combination are first presented in Subsections 2.1 and 2.2, respectively. Some notions regarding the definition and manipulation of belief functions on continuous domains are then recalled in Subsection 2.3.

2.1 Belief functions on finite domains

Let $\boldsymbol{\theta}$ be a variable taking values in a finite domain Θ, called the *frame of discernment*. Uncertain evidence about $\boldsymbol{\theta}$ may be represented by a *mass function* m on Θ, defined as a function from the powerset of Θ, denoted as 2^Θ, to the interval $[0,1]$, such that $m(\emptyset)=0$ and

$$\sum_{A\subseteq\Theta} m(A) = 1. \tag{1}$$

Any subset A of Θ such that $m(A)>0$ is called a *focal set* of m. A categorical mass function has only one focal set (it is thus equivalent to a set), while a Bayesian mass function has only focal sets of cardinality one and is thus equivalent to a probability distribution. The mass function m such that $m(\Theta)=1$ is said to be vacuous.

Each number $m(A)$ is interpreted as a *degree of belief* attached to the proposition $\boldsymbol{\theta}\in A$ and to *no more specific proposition*, based on some evidence. As argued by Shafer [26], the meaning of such degrees of belief can be better understood by assuming that we have compared our evidence to a canonical chance set-up. The set-up proposed by Shafer consists of an encoded message and a set of codes $\Omega=\{\omega_1,\ldots,\omega_n\}$, exactly one of which is selected at random. We know the list of codes as well as the chance p_i of each code ω_i being selected. Decoding the encoded message using code ω_i produces a message of the form "$\boldsymbol{\theta}\in A_i$" for some $A_i\subseteq\Theta$. Then

$$m(A) = \sum_{\{1\le i\le n: A_i=A\}} p_i \tag{2}$$

is the chance that the original message was "$\boldsymbol{\theta}\in A$". Stated differently, it is the probability of knowing that $\boldsymbol{\theta}\in A$. In particular, $m(\Theta)$ is, in this setting, the probability that the original message was vacuous, i.e., the probability of knowing nothing.

The above setting thus consists of a set Ω, a probability measure P on Ω and a multi-valued mapping $\Gamma:\Omega\rightarrow 2^\Theta\setminus\{\emptyset\}$ such that $A_i=\Gamma(\omega_i)$ for each $\omega_i\in\Omega$. This is the framework initially considered by Dempster in [7]. The triple (Ω,P,Γ) formally defines a finite *random set* [23]: mass functions are thus exactly equivalent to random sets from a mathematical point of view. However, the meaning of mass functions differs from the usual interpretation of a random set as the outcome of a random experiment: here, $m(A)$ is *not* the chance that A was selected, but it can be viewed as the chance of the evidence meaning that $\boldsymbol{\theta}$ is in A [26].

To each normalized mass function m, we may associate belief and plausibility functions from 2^Θ to $[0,1]$ defined as follows:

$$Bel(A) = P\left(\{\omega\in\Omega|\Gamma(\omega)\subseteq A\}\right) = \sum_{B\subseteq A} m(B) \tag{3a}$$

$$Pl(A) = P\left(\{\omega\in\Omega|\Gamma(\omega)\cap A\neq\emptyset\}\right) = \sum_{B\cap A\neq\emptyset} m(B), \tag{3b}$$

for all $A \subseteq \Theta$. These two functions are linked by the relation $Pl(A) = 1 - Bel(\overline{A})$, for all $A \subseteq \Theta$. Each quantity $Bel(A)$ may be interpreted as the degree to which the evidence *supports* A, while $Pl(A)$ can be interpreted as the degree to which the evidence *is not contradictory* with A. The following inequalities always hold: $Bel(A) \leq Pl(A)$, for all $A \subseteq \Theta$. The function $pl : \Theta \rightarrow [0, 1]$ such that $pl(\theta) = Pl(\{\theta\})$ is called the *contour function* associated to m.

If m is Bayesian, then function Bel is identical to Pl and it is a probability measure, and pl is the corresponding probability mass function. Another special case of interest is that where m is *consonant*, i.e., its focal elements are nested. The plausibility function is then a *possibility measure* [38, 18] with possibility distribution pl, i.e., the plausibility function can be recovered from the contour function as follows: [24]:

$$Pl(A) = \max_{\theta \in A} pl(\theta). \tag{4}$$

for all $A \subseteq \Theta$.

Given two mass functions m_1 and m_2, m_1 is said to be *less specific* than m_2 if it can be obtained from m_2 by transferring belief masses $m_2(A)$ to supersets $B \supseteq A$ [37, 17]. In this case, m_1 can be considered as less informative, or less committed[1] than m_2. The *Least Commitment Principle* (LCP) [30] states that, given some constraints on an unknown mass function, the least committed should be selected. This principle provides a justification of consonant mass functions: given a function $\pi : \Theta \rightarrow [0, 1]$ such that $\max \pi = 1$, the least specific mass function m with contour function pl such that $pl = \pi$ is consonant; its plausibility function, given by (4), will be denoted as pl^*.

2.2 Dempster's rule

A key idea in Dempster-Shafer theory is that beliefs are elaborated by aggregating different items of evidence. The basic mechanism for evidence combination is Dempster's rule of combination, which can be naturally derived using the random code metaphor as follows.

Let m_1 and m_2 be two mass functions induced by triples $(\Omega_1, P_1, \Gamma_1)$ and $(\Omega_2, P_2, \Gamma_2)$ interpreted under the random code framework as before. Let us further assume that the codes are selected independently. For any two codes $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$, the probability that they both are selected is then $P_1(\{\omega_1\})P_2(\{\omega_2\})$, in which case we can conclude that $\boldsymbol{\theta} \in \Gamma_1(\omega_1) \cap \Gamma_2(\omega_2)$. If $\Gamma_1(\omega_1) \cap \Gamma_2(\omega_2) = \emptyset$, we know that the pair of codes (ω_1, ω_2) could not have been selected: consequently, the joint probability distribution on $\Omega_1 \times \Omega_2$ must be conditioned, eliminating such pairs [26]. This line of reasoning yields the following combination rule, referred to as Dempster's rule [24]:

$$(m_1 \oplus m_2)(A) = \frac{1}{1 - \kappa} \sum_{B \cap C = A} m_1(B) m_2(C) \tag{5}$$

for all $A \subseteq \Theta$, $A \neq \emptyset$ and $(m_1 \oplus m_2)(\emptyset) = 0$, where

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{6}$$

is the *degree of conflict* between m_1 and m_2. If $\kappa = 1$, there is a logical contradiction between the two pieces of evidence and they cannot be combined. Dempster's rule is commutative, associative, and it admits as neutral element the *vacuous* mass function defined as $m(\Omega) = 1$.

Dempster's rule can be easily expressed in terms of contour functions: if pl_1 and pl_2 are the contour functions of two mass functions m_1 and m_2, then the contour function of $m_1 \oplus m_2$ is, using the same symbol $\oplus$ as used for mass functions and contour functions

$$(pl_1 \oplus pl_2)(\theta) = \frac{pl_1(\theta) pl_2(\theta)}{1 - \kappa} \tag{7}$$

[1] Alternative comparative orderings between belief functions have been proposed, see, e.g., [17].

for all $\theta \in \Theta$, where κ is the degree of conflict. If m_1 or m_2 is Bayesian, then so is m_1 and m_2 and the degree of conflict is then

$$\kappa = 1 - \sum_{\theta \in \Theta} pl_1(\theta) pl_2(\theta). \tag{8}$$

2.3 Random real intervals

The definition of belief functions and random sets in infinite spaces implies greater mathematical sophistication than it does in finite spaces [25, 23]. Here, we will restrict our discussion to random closed intervals on the real line (see, e.g., [9, 31, 11]), which constitute a simple yet sufficiently general framework for expressing beliefs on a real variable.

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $(U, V) : \Omega \to \mathbb{R}^2$ a two-dimensional real random vector such that $P(\{\omega \in \Omega | U(\omega) \leq V(\omega)\}) = 1$. Let Γ be the multi-valued mapping that maps each $\omega \in \Omega$ to the closed interval $[U(\omega), V(\omega)]$. This setting defines a random interval, as well as belief and plausibility functions on $\mathbb{R}$ defined, respectively, by

$$\begin{aligned} Bel(A) &= P(\{\omega \in \Omega | [U(\omega), V(\omega)] \subseteq A\}) & (9) \\ Pl(A) &= P(\{\omega \in \Omega | [U(\omega), V(\omega)] \cap A \neq \emptyset\}) & (10) \end{aligned}$$

for all elements A of the Borel sigma-algebra $\mathcal{B}(\mathbb{R})$ on the real line [9]. The intervals $[U(\omega), V(\omega)]$ are referred to as the focal intervals of $[U, V]$. We note that, when U and V are continuous, the notion of mass function should be replaced by that of mass density function defined by $m([u, v]) = p(u, v)$, where $p(u, v)$ denotes the joint probability density function (pdf) of (U, V). To simplify the terminology, we will continue to use the term "mass function" in this case.

If $U = V$, then we have a random point, which is equivalent to a real random variable. Another special case of interest is that of consonant random closed intervals defined as follows. Let $\pi : \mathbb{R} \to [0, 1]$ be an upper semi-continuous function and let $\Omega = [0, 1]$. For each $\omega \in \Omega$, let

$$\Gamma(\omega) = \{x \in \mathbb{R} | \pi(x) \leq \omega\},$$

which is a closed interval $[U(\omega), V(\omega)]$. Finally, let P denote the Lebesgue measure on Ω. Then, $[U, V]$ is a random interval and π is its contour function, i.e., $pl(x) = Pl(\{x\}) = \pi(x)$ for all $x \in \mathbb{R}$. Such a random interval is said to be consonant because its focal intervals $\Gamma(\omega)$ are nested.

Dempster's rule can be defined for random intervals as follows. Let us assume that we have two random intervals $(\Omega_i, \mathcal{A}_i, P_i, \Gamma_i)$ with $i = 1, 2$ and $[U_i(\omega), V_i(\omega)] = \Gamma_i(\omega)$. Let Γ_{12} be the mapping from $\Omega_1 \times \Omega_2$ to the set of closed real intervals defined by

$$\Gamma_{12}(\omega_1, \omega_2) = \Gamma_1(\omega_1) \cap \Gamma_2(\omega_2), \quad \forall (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$$

and let P_{12} be the product measure $P_1 \times P_2$ conditioned on the set $\{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 | \Gamma_{12}(\omega_1, \omega_2) \neq \emptyset\}$. Then, $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2, P_{12}, \Gamma_{12})$ define a random interval $[U_{12}, V_{12}] = [U_1, V_1] \oplus [U_2, V_2]$. Its contour function is

$$(pl_1 \oplus pl_2)(x) = \frac{pl_1(x) pl_2(x)}{1 - \kappa}$$

for all $x \in \mathbb{R}$, where κ is the degree of conflict between the two random intervals defined as:

$$\kappa = P\left(\{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 | \Gamma_{12}(\omega_1, \omega_2) \neq \emptyset\}\right).$$

In general, the combination of two random intervals by Dempster's rule is not easy to compute analytically. However, a special case in which the computations are very simple is that were a random point with pdf p_1 is combined with a random interval with contour function pl_2. The results is a random point with pdf

$$(p_1 \oplus pl_2)(x) = \frac{p_1(x) pl_2(x)}{1 - \kappa}, \tag{11}$$

where the degree of conflict κ is

$$\kappa = 1 - \int_{-\infty}^{+\infty} p_1(x) pl_2(x) dx. \tag{12}$$

3 Modeling statistical evidence

Let us now turn our attention to the representation of statistical evidence. Assume that we have observed a realization $\mathbf{x}$ of a random vector $\mathbf{X}$ with pdf $p(\mathbf{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$ is an unknown parameter. What does this item of evidence tell us about $\boldsymbol{\theta}$? Shafer's solution [24] derived from the Likelihood and Least Commitment principles will first be recalled in Subsection 3.1. Arguments for and against this solution will then be discussed in Subsection 3.2 and an illustrative example will be presented in Subsection 3.3.

3.1 Least committed solution based on likelihoods

In the standard statistical framework, information about $\boldsymbol{\theta}$ is typically assumed to be represented by the likelihood function defined by $L(\theta; \mathbf{x}) = p(\mathbf{x}; \theta)$ for all $\theta \in \Theta$. More precisely, the likelihood principle [2] [3] [19, chapter 3] states that "Within the framework of a statistical model, *all* the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of these hypotheses on the data". In statistical parlance, the likelihood ratio is often referred to as the "relative plausibility", which suggests translating the likelihood ratio in the belief function framework as follows:

$$\frac{pl(\theta_1; \mathbf{x})}{pl(\theta_2; \mathbf{x})} = \frac{L(\theta_1; \mathbf{x})}{L(\theta_2; \mathbf{x})},$$

for all $(\theta_1, \theta_2) \in \Theta^2$ or, equivalently,

$$pl(\theta; \mathbf{x}) = cL(\theta; \mathbf{x})$$

for all $\theta \in \Theta$ and some positive constant c. The LCP then leads us to giving the highest possible value to constant c, i.e., defining pl as the relative likelihood :

$$pl(\theta; \mathbf{x}) = \frac{L(\theta; \mathbf{x})}{\sup_{\theta \in \Theta} L(\theta; \mathbf{x})} \tag{13}$$

and representing evidence about $\boldsymbol{\theta}$ by the least committed plausibility function induced by pl, i.e.,

$$Pl(A; \mathbf{x}) = \sup_{\theta \in A} pl(\theta; \mathbf{x}) = \frac{\sup_{\theta \in A} L(\theta; \mathbf{x})}{\sup_{\theta \in \Theta} L(\theta; \mathbf{x})}, \tag{14}$$

for all $A \subseteq \Theta$. The corresponding belief function is called a likelihood-based belief function by Wasserman [36].

3.2 Discussion

Equation (14) was first proposed by Shafer in [24, chapter 11] who, however, did not justify it by the LCP, but by the more questionable requirement that the belief function on Θ be consonant. In the special case where $\Theta = \{\theta_1, \theta_2\}$ has only two points, Wasserman [36] showed that the plausibility function (14) corresponds to the unique belief function $Bel(\cdot; \mathbf{x})$ verifying the following requirements:

1. If $L(\theta_1; \mathbf{x}) = L(\theta_2; \mathbf{x})$, then $Bel(\cdot; \mathbf{x})$ should be vacuous;

2. $Bel(\{\theta\}; \mathbf{x})$ should be nondecreasing in $L(\theta; \mathbf{x})$;

3. If $Bel = Bel(\cdot;\mathbf{x}) \oplus P_0$ and P_0 is a probability measure, then Bel should be equal to the Bayesian posterior.

This argument can be extended to the case where Θ is a complete, separable metric space [36].

One of the main criticisms against the use of the likelihood-based plausibility function (14) for represented statistical evidence is its incompatibility with Dempster's rule in the case of independent observations [27]. More precisely, assume that $\mathbf{X}$ is an independent sample $(X_1, \ldots, X_n)$ and each observation X_i has a marginal pdf $p(x_i;\boldsymbol{\theta})$ depending on $\boldsymbol{\theta}$. We could combine the n observations at the "aleatory level" by computing $Pl(\cdot;\mathbf{x})$ using (14), or we could combine them at the "epistemic level" by first computing the consonant plausibility functions $Pl(\cdot;x_i)$ induced by each of the independent observations and applying Dempster's rule. Obviously, these two procedures yield different results in general, as consonance is not preserved by Dempster's rule.

Shafer [27] seems to have regarded the above argument as strong enough to reject (14) as a reasonable method to represent statistical evidence. However, Aickin [1] proposed to keep (14) but questioned Dempster's rule as a mechanism for combining statistical evidence. Additional arguments against the use of Dempster's rule for combining evidence from independent observations can be found in [34].

Based on the above discussion, we propose to adopt (13) and (14) as models of statistical evidence. Further arguments in favor of this approach are summarized below:

1. This method of inference is considerably simpler than other methods such as Dempster's initial proposal [8] and other methods discussed in [27], while being more widely applicable than Smets' Generalized Bayesian Theorem [29, 16].

2. Combining $Pl(\cdot;\mathbf{x})$ given by (14) with a Bayesian prior P_0 on Θ using Dempster's rule yields a Bayesian plausibility function $Pl(\cdot;\mathbf{x}) \oplus P_0$ which is identical to the posterior probability obtained using Bayes' rule: consequently, the proposed method of inference boils down to Bayesian inference when a Bayesian prior is available.

3. Finally, viewing the relative likelihood function as a possibility distribution seems to be consistent with statistical practice, although this point of view has not been adopted explicitly in the statistical literature. For instance, likelihood intervals [22, 33] are focal intervals of the relative likelihood viewed as a possibility distribution. In the case where $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$ and θ_2 is considered as a nuisance parameter, the relative profile likelihood function can be written

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

which is the marginal possibility distribution on Θ_1. Eventually, we can remark that the usual likelihood ratio statistics $\Lambda(\mathbf{x})$ for a composite hypothesis $H_0 \subset \Theta$ is nothing but the plausibility of H_0, as

$$\Lambda(\mathbf{x}) = \frac{\sup_{\theta \in H_0} L(\theta;\mathbf{x})}{\sup_{\theta \in \theta} L(\theta;\mathbf{x})} = \sup_{\theta \in H_0} pl(\theta;\mathbf{x}) = Pl(H_0;\mathbf{x}).$$

3.3 Illustrative example

As a concrete example, let us consider the following problem using real dataset. Average public teacher pay and spending on public schools per pupil in 1985 for 49 states and the District of Columbia were reported by the Albuquerque Tribune[2]. The data are plotted in Figure 1 for each of the three areas: Northeast and North Central, South and West. We can see that public teacher pay is approximately linearly related to spending on public schools. Is there any statistical evidence of different relations holding in the three regions?

[2] The dataset can be downloaded from the Data and Story Library at `http://lib.stat.cmu.edu/DASL`. The data for Alaska is an outlier and was not considered in this analysis

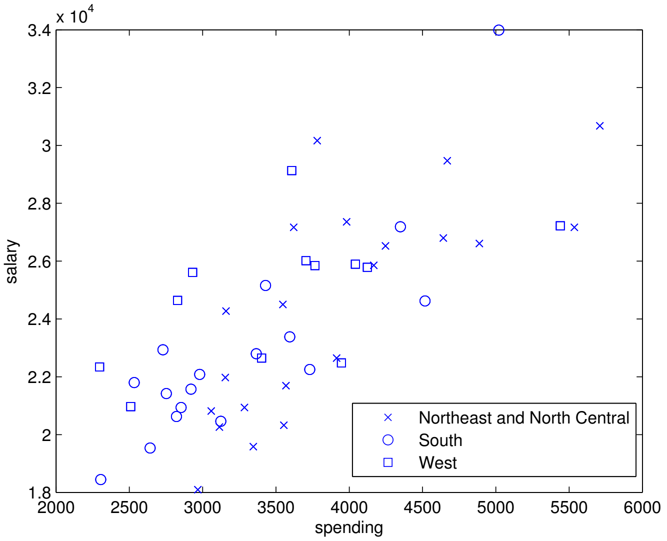


Figure 1: Average public school teacher annual salary (\$) as a function of spending on public schools per pupil (\$) for 49 states and the District of Columbia.

Let y_{ki} and x_{ki} denote, respectively, the teacher pay and spending on public schools in state i of region k ($k = 1, 2, 3$). We assume that $\mathbf{y}_k = \{y_{ki}\}_{i=1}^{n_k}$ is a realization of a Gaussian random vector $\mathbf{Y}_k \sim \mathcal{N}(\mathbf{X}_k\mathbf{b}_k, \sigma_k^2 I_n)$, where $\mathbf{X}_k$ is the fixed design matrix with line i equal to $(1, x_{ki})$, I_n is the identity matrix of size n, and $\boldsymbol{\theta}_k = (\mathbf{b}_k, \sigma_k)'$ is the parameter vector.

Figure 2 shows the contour functions $pl(\mathbf{b}_k; \mathbf{y}_k)$. We recall that this function is obtained as the relative profile likelihood function considering variance as a nuisance parameter, i.e.,

$$pl(\mathbf{b}_k; \mathbf{y}_k) = \sup_{\sigma_k > 0} pl(\mathbf{b}_i, \sigma_k; \mathbf{y}_k) = \frac{\sup_{\sigma_k>0} L(\mathbf{b}_k, \sigma_k; \mathbf{y}_k)}{\sup_{\mathbf{b}_k \in \mathbb{R}^2, \sigma_i > 0} L(\mathbf{b}_k, \sigma_k; \mathbf{y}_k)},$$

with

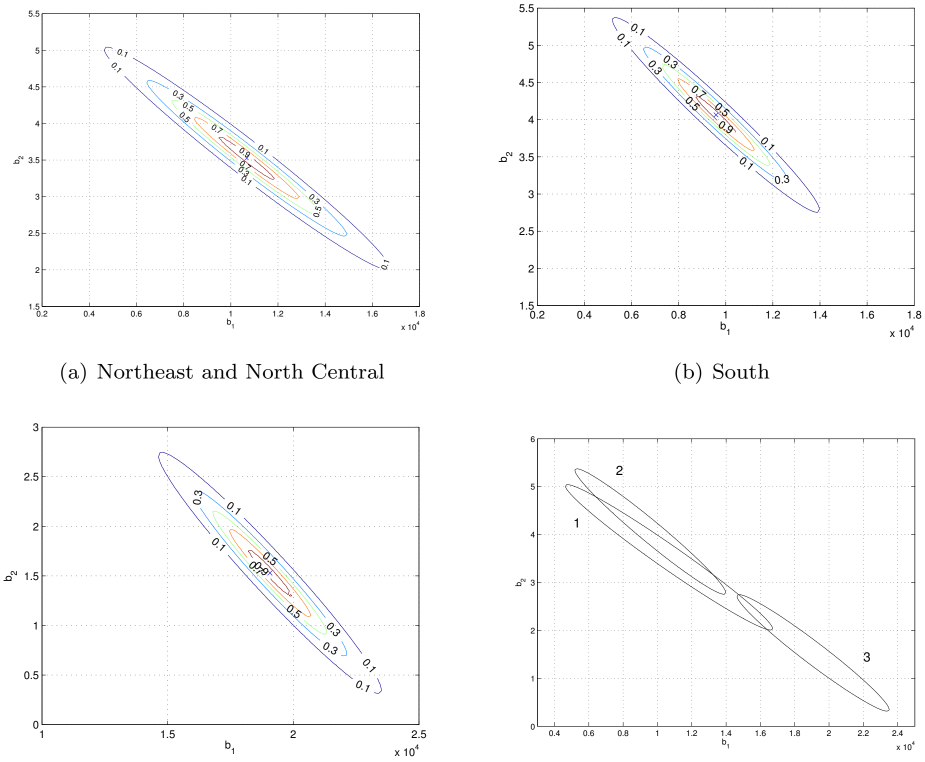
$$L(\mathbf{b}_k, \sigma_k; \mathbf{y}_k) = \phi(\mathbf{y}_k; \mathbf{X}_k'\mathbf{b}_k, \sigma_k^2 I_n) = \prod_{i=1}^{n} \phi(y_{ki}; (1, x_{ki})'\mathbf{b}_k, \sigma_k^2),$$

We can see from Figure 2(d) that the contour at level 0.1 for region 3 does not intersect the corresponding contour for region 2, which suggests that $\mathbf{b}_3$ is different from $\mathbf{b}_2$ with a high plausibility. To carry the analysis further, we can compute the plausibilities $Pl(\mathbf{b}_i = \mathbf{b}_j)$ for each pair of regions, as well as the plausibility $Pl(\mathbf{b}_1 = \mathbf{b}_2 = \mathbf{b}_3)$ that the three parameters are equal. It is easy to see [14] that these plausibilities are equal to one minus the degree of conflict between the belief functions related to each parameter. These degrees of conflict are not easy to compute analytically, but they can be estimated by Monte-Carlo simulation. This is achieved by picking a focal set at random independently for each of the belief function, and estimating the probability for the focal sets to be disjoint. We obtain the following values:

$$Pl(\mathbf{b}_1 = \mathbf{b}_2) = 0.70, \quad Pl(\mathbf{b}_1 = \mathbf{b}_3) = 0.12, \quad Pl(\mathbf{b}_2 = \mathbf{b}_3) = 0.02$$

$$Pl(\mathbf{b}_1 = \mathbf{b}_2 = \mathbf{b}_3) = 0.01.$$

which confirms that the hypotheses $\mathbf{b}_2 = \mathbf{b}_3$ and $\mathbf{b}_1 = \mathbf{b}_2 = \mathbf{b}_3$ can be discarded as having very small plausibility.



(a) Northeast and North Central

(b) South

(c) West

(d) 0.1-level contours

Figure 2: Contour functions $pl(\mathbf{b}_k; \mathbf{y}_k)$ for each of the three regions (a-c) and 0.1-level contours (d). Please note that the x and y axes have different ranges in the four plots.

4 Inference from uncertain data

We consider in this section the situation where the data $\mathbf{x}$ have been generated by a random process but have been imperfectly observed [12, 13, ?]. Our partial knowledge of $\mathbf{x}$ will then be described by a mass function m on the data space $\Omega_X \subseteq \mathbb{R}^d$. Our objective will be to find a suitable representation of the information about the parameter provided by such data, in the belief function framework. Our approach will be to generalize the likelihood function and, as before, to consider the relative likelihood as the contour function of a consonant plausibility measure.

Before we describe our approach, it must be emphasized that, in this model, the pdf or probability mass function $p(\mathbf{x}, \boldsymbol{\theta})$ and the Dempster-Shafer mass function m represent two different pieces of knowledge:

- $p(\mathbf{x}, \boldsymbol{\theta})$ represents *generic* knowledge about the data generating process or, equivalently, about the underlying population; it corresponds to *random uncertainty*;

- m represents *specific* knowledge about a given realization $\mathbf{x}$ of $\mathbf{X}$; this knowledge is only partial because the observation process is imperfect; function m captures *epistemic uncertainty*, i.e., uncertainty due to lack of knowledge.

The uncertain data m is thus not assumed to be produced by a random experiment, which is in sharp contrast with other approaches based on random sets [23] or fuzzy random variables [20].

Our approach will first be described in Subsection 4.1. The impact of stochastic and cognitive independence assumptions will then be examined in Subsection 4.2.

4.1 Representation of uncertain statistical evidence

Let us assume that the mass function m is induced by a random set $(\Omega, \mathcal{A}, P, \Gamma)$. We will further assume that one of the following two conditions holds:

- $\mathbf{X}$ is discrete, or

- $\mathbf{X}$ is continuous an $\Gamma(\omega)$ is not reduced to a point (which would correspond an infinite precision).

Under these assumptions, the probability of observing the result $\Gamma(\omega)$ given that the interpretation $\omega \in \Omega$ holds is

$$P(\Gamma(\omega); \theta) = \int_{\Gamma(\omega)} p(\mathbf{x}; \theta) d\mathbf{x},$$

assuming that the integral in the right-hand side is well defined. The probability of the uncertain observation m may then defined as the average of $P(\Gamma(\omega); \theta)$ over $\omega \in \Omega$, which can be written as

$$P(m; \theta) = \sum_{\omega \in \Omega} p(\omega) P(\Gamma(\omega); \theta)$$

if Ω is finite and

$$P(m; \theta) = \int_{\Omega} p(\omega) P(\Gamma(\omega); \theta) d\omega$$

otherwise, assuming this integral to be defined. The likelihood function given the uncertain observation m can then be defined as $L(\theta; m) = P(m; \theta)$ for all $\theta \in \Theta$. It is easy to show that

$L(\theta; m)$ only depends on the contour function. To see this, we may write:

$$\begin{aligned} L(\theta; m) &= \int_{\Omega} p(\omega) \left(\int_{\Gamma(\omega)} p(\mathbf{x}; \theta) d\mathbf{x} \right) d\omega, & (15) \\ &= \int_{\Omega_X} p(\mathbf{x}; \theta) \left(\int_{\{\omega | \Gamma(\omega) \ni \mathbf{x}\}} p(\omega) d\omega \right) d\mathbf{x}, & (16) \\ &= \int_{\Omega_X} p(\mathbf{x}; \theta) pl(\mathbf{x}) d\mathbf{x} & (17) \\ &= \mathbb{E}_{\theta} \left[pl(\mathbf{X}) \right]. & (18) \end{aligned}$$

As a natural extension of (13), we propose to represent the information on $\boldsymbol{\theta}$ provided by the uncertain data by the consonant plausibility function with the following contour function:

$$pl(\theta; m) = \frac{L(\theta; m)}{\sup_{\theta \in \Theta} L(\theta; m)}. \tag{19}$$

An iterative procedure for finding a value $\widehat{\theta}$ of θ that maximizes $pl(\theta; m)$ has been introduced in [4] and generalized in [12, **?**]. This procedure, called the Evidential Expectation Maximization (E^{2}M) algorithm, is an extension of the EM algorithm [10].

4.2 Independence assumptions

Let us assume that the random vector $\mathbf{X}$ can be written as $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, where each $\mathbf{X}_i$ is a p-dimensional random vector taking values in $\Omega_{\mathbf{X}_i}$. Similarly, its realization can be written as $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \Omega_{\mathbf{X}}$. Two different independence assumptions can then be made:

1. Under the *stochastic independence* of the random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$, the pdf or probability mass function $p(\mathbf{x}; \boldsymbol{\theta})$ can be decomposed as:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i; \boldsymbol{\theta}), \quad \forall \mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \Omega_{\mathbf{X}} \tag{20}$$

2. Under the *cognitive independence* of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with respect to m (see [24, page 149]), we can write:

$$pl(\mathbf{x}) = \prod_{i=1}^{n} pl_i(\mathbf{x}_i), \quad \forall \mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \Omega_{\mathbf{X}}, \tag{21}$$

   where pl_i is the contour function corresponding to the mass function m_i obtained by marginalizing m on $\Omega_{\mathbf{X}_i}$.

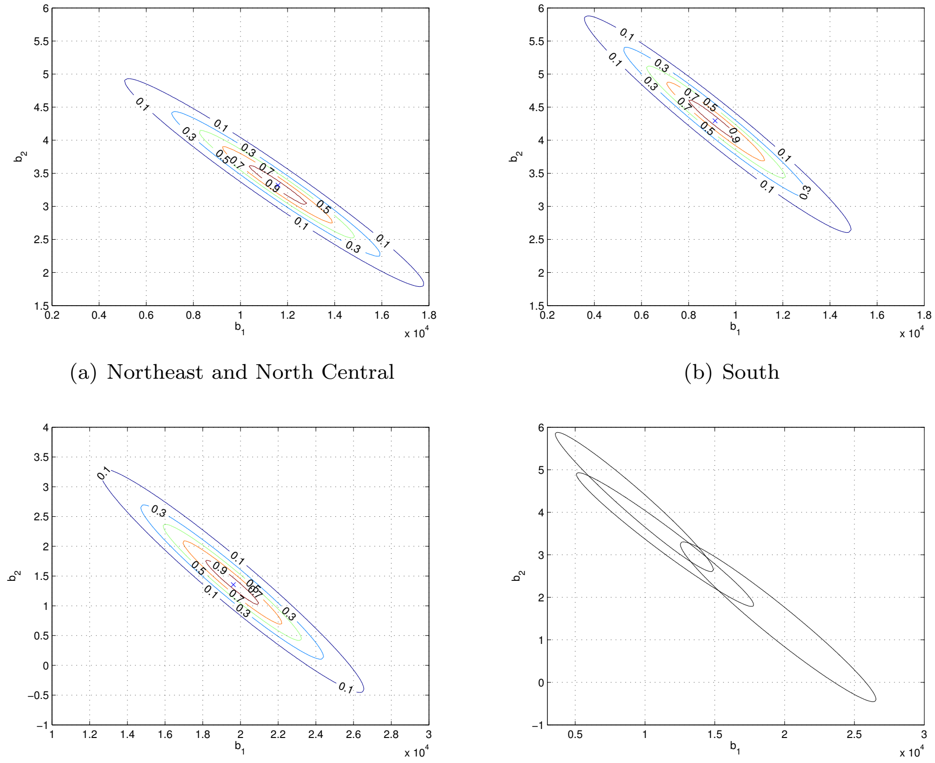
We can remark here that the two assumptions above are totally unrelated as they are of different natures: stochastic independence of the random variables $\mathbf{X}_i$ is an objective property of the random data generating process, whereas cognitive independence pertains to our state of knowledge about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

If both assumptions hold, the likelihood criterion (18) can be written as a product of n terms:

$$L(\boldsymbol{\theta}; m) = \prod_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}} \left[pl_i(\mathbf{X}_i) \right] \tag{22}$$

and $pl(\theta; m)$ can be written as:

$$pl(\theta; m) = \frac{\prod_{i=1}^{n} pl(\theta; m_i)}{\sup_{\theta \in \Theta} \prod_{i=1}^{n} pl(\theta; m_i)}. \tag{23}$$



(a) Northeast and North Central

(b) South

(c) West

(d) 0.1-level contours

Figure 3: Contour functions $pl(\mathbf{b}_k; \mathbf{y}_k)$ for each of the three regions (a-c) and 0.1-level contours (d), with simulated data uncertainty. Please note that the x and y axes have different ranges in the four plots.

Example 1 Let us come back to the analysis of Subsection 3.3, this time assuming that the observations of the dependent variable are uncertain. This is reasonable if we assume that teacher pay data for each state are not known exactly but are estimated by surveys carried out with samples of different sizes and under different conditions. As we do not know in which conditions the data were collected, we simulated data uncertainty by assuming the contour functions $pl_{ki}(y_{ki})$ to be normalized Gaussians centered at each data point and with standard deviation s_{ki} selected at random from a uniform distribution in $[0, 5000]$.

The results are shown in Figure 3. We can see that the consideration of data uncertainty actually leads to less committed plausibility functions in the parameter space. The plausibility values for the same hypotheses as considered in Subsection 3.3 are now:

$$Pl(\mathbf{b}_1 = \mathbf{b}_2) = 0.61, \quad Pl(\mathbf{b}_1 = \mathbf{b}_3) = 0.39, \quad Pl(\mathbf{b}_2 = \mathbf{b}_3) = 0.13,$$

$$Pl(\mathbf{b}_1 = \mathbf{b}_2 = \mathbf{b}_3) = 0.08,$$

which shows that the hypotheses $\mathbf{b}_2 = \mathbf{b}_3$ and $\mathbf{b}_1 = \mathbf{b}_2 = \mathbf{b}_3$ can no longer be rejected based on the uncertain statistical evidence.

5 Conclusion

The Dempster-Shafer theory of belief functions places emphasis on the representation of evidence for evaluating degrees of belief. This generality and flexibility of this framework makes it suitable for representing and combining expert judgments and statistical evidence.

In this paper, we have focused on the representation of statistical evidence, seeing the relative likelihood function as the contour function of a consonant belief function in the parameter space, as originally proposed by Shafer. Likelihood-based and Bayesian inference schemes can both be seen as special cases of this approach.

We have shown that this method can be extended in a simple way to the representation of uncertain statistical evidence or ill-known data, where lack of knowledge comes from imperfectness of the observation process. Maximum plausibility estimation can still be performed in this case using a computationally simple iterative procedure that extends the EM algorithm.

A interesting perspective of this approach concerns situations in which statistical evidence needs to be combined with expert judgements. Such problems typically arise in climate change studies, in which statistical data cannot be considered as a unique source of information but have to be pooled with expert opinions summarizing findings from physical modeling. Results concerning the application of the belief approach to such problems will be reported in future publications.

References

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

[2] 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.

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

[4] 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.

[5] D. R. Cox. Some remarks on statistical aspects of econometrics. In J. Panaretos, editor, *Stochastics Musings*. Lawrence Erlbaum, Mahwah, NJ, 2003.

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

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

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

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

[10] 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.

[11] T. Denœux. Extending stochastic ordering to belief functions on the real line. *Information Sciences*, 179:1362–1376, 2009.

[12] 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.

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

[14] 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.

[15] D. Dubois. Possibility theory and statistical reasoning. *Computational Statistics and Data Analysis*, 51(1):47–69, 2006.

[16] D. Dubois and T. Denœux. Statistical inference with belief functions and possibility measures: a discussion of basic assumptions. 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 217–225, Oviedo, Spain, 2010. Springer.

[17] 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.

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

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

[20] 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.

[21] J. J. Heckman. Econometrics and empirical economics. *Journal of Econometrics*, 100:3–5, 2001.

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

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

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

[25] G. Shafer. Allocations of probability. *Annals of Probability*, 7(5):827–839, 1979.

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

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

[28] Ph. Smets. Resolving misunderstandings about belief functions. *International Journal of Approximate Reasoning*, 6:321–344, 1990.

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

[30] Ph. Smets. The Transferable Belief Model for quantified belief representation. In D. M. Gabbay and Ph. Smets, editors, *Handbook of Defeasible reasoning and uncertainty management systems*, volume 1, pages 267–301. Kluwer Academic Publishers, Dordrecht, 1998.

[31] Ph. Smets. Belief functions on real numbers. *International Journal of Approximate Reasoning*, 40(3):181–223, 2005.

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

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

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

[35] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

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

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

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