Belief function and multivalued mapping robustness in statistical estimation

Alessio Benavoli

"Dalle Molle" Institute for Artificial Intelligence (IDSIA),
Galleria 2, Via Cantonale, CH-6928 Manno-Lugano (Switzerland)

Abstract

We consider the case in which the available knowledge does not allow to specify a precise probabilistic model for the prior and/or likelihood in statistical estimation. We assume that this imprecision can be represented by belief functions models. Thus, we exploit the mathematical structure of belief functions and their equivalent representation in terms of closed convex sets of probabilities to derive robust posterior inferences using Walley's theory of imprecise probabilities. Then, we apply these robust models to practical inference problems and we show the connections of the proposed inference method with interval estimation and statistical inference with missing data.

Key words: Belief function, multivalued mapping, closed convex set of probabilities, robust inference

1. Introduction

A multivalued function (also called multifunction or correspondence) is a set-valued function, i.e., it assigns to each point in one set a set of points in a possibly different set. Multivalued functions are interesting because they arise in various practical applications.

- In economics, the budget set is defined as:

$$\Gamma(p,m) = \{x \in \mathbb{R}^n_+ : \ p^T x \leq m\},$$

  i.e, the set of commodity vectors x that can be bought with income $m \in \mathbb{R}_+$ at the vector of prices $p \in \mathbb{R}^n_+$. Here, Γ is a multivalued function that maps points of $\mathbb{R}^n_+ \times \mathbb{R}_+$ to subsets of $\mathbb{R}^n_+$.

- In statistics, a missingness process is defined as:

$$\Gamma(o) = \{o, ?\},$$

  i.e., the observation $o \in \mathscr{O}$ is mapped into itself or is missing (?). For instance, in a sequence of dice rolls the process that turned the sequence $\{2,4,1,5,2,6\}$ into the sequence $\{2,4,1,5,?,6\}$ is a missing process (the outcome of the fifth roll is missing). It can be regarded as a multivalued function that maps the fifth roll to the set of all possible dice roll outcomes (here denoted by ?).

- In optimal control theory, it may happen that for a given state x the optimal control value u is not unique, i.e., there exists a set $\Gamma(x)$ of equivalent optimal controls. The map from the state space to the set of equivalent optimal controls is thus a multivalued function.

Let Γ denote generically a multivalued function from the set $\mathscr{Z}$ to subsets of the set $\mathscr{X}$. Assume that P_Z is a measure that assigns probabilities to the members of a class $\mathscr{F}$ of subsets of $\mathscr{Z}$. If $P_Z(A)$ is the probability of $A \in \mathscr{F}$ and if $B \subseteq \Gamma(A)$, then what is the probability of B? While a single valued function (under general conditions) carries P_Z to a unique probability measure P_X over subsets of $\mathscr{X}$, a multivalued function leads to a set of probability measures on $\mathscr{X}$.

A way to characterize such set of probability measures is to determine lower and upper bounds for the probabilities induced from the multivalued mapping; this has been the approach first proposed by Dempster in [1]. Later, Shafer [2, 3] has called the lower and upper probabilities induced from a multivalued mapping belief (*Bel*) and, respectively, plausibility (*Pl*) functions.

The aim of this paper is to show how belief function models can be used in robust statistics. Hereafter, we use the term robust as in Bayesian robustness analysis, i.e., the robustness of the posterior inferences to the choice of the involved probabilistic models, namely the prior and the likelihood. In case of total or partial lack of information about the probabilistic models, an issue in Bayesian analysis is how to select the prior and the likelihood. Consider for instance the choice of the prior, there are two main avenues that can be followed (for both avenues we will distinguish the cases of total or partial lack of information).

The first assumes that the lack of prior information can be managed satisfactorily by considering a single prior probability. For instance in case of total prior ignorance a common choice is to consider so-called "noninformative priors", i.e.,

Laplace's prior, Jeffreys' prior, or the reference prior of Bernardo (see [4, Sec. 5.6.2] for a review). This view has been questioned on diverse grounds. Non-informative priors are typically improper and may lead to an improper posterior. Moreover, even if the posterior is proper, it can be inconsistent with the likelihood model (i.e., incoherent in the subjective interpretation of probability [5, Ch. 7]). Furthermore, the need of selecting a single probability limits the expressiveness of the probabilistic model. For instance, the most important criticism of non-informative priors is that they are not expressive enough to represent ignorance [5, Ch. 5]. For the case of partial lack of prior information, a common choice is to consider so-called "fat-tail" priors (e.g., the t-Student or the Cauchy distribution) or distributions selected according to some external criterion, e.g., maximum entropy. This approach can also be questioned, since it usually leads to a unimodal distribution while the available prior information may also be compatible with a multimodal distribution. The result is that the inferences may be not robust, in the sense that a point estimate based on a unimodal distribution can be in a region of lower probability if the true distribution is multimodal.

An alternative is to use a set of prior distributions, $\mathscr{M}$, rather than a single distribution, to model prior ignorance about statistical parameters. Each prior distribution in $\mathscr{M}$ is updated by Bayes' rule, producing a set of posterior distributions. In fact there are two distinct approaches of this kind, which have been compared by Walley [5]. The first approach, known as *Bayesian robustness* [6, 7], considers a set of priors which is built around a candidate (ideal) distribution which is compatible with, but does not match completely, the available prior information. The resulting set of priors is in general a *neighbourhood model*, i.e., the set of all distributions that are close (w.r.t. some criterion) to this ideal distribution. Examples of neighbourhood models are: ε-contamination models [8, 9]; restricted ε-contamination models [10]; intervals of measures [9, 11]; the density ratio class [5, 11], etc. Note that this approach is not suitable in case of total lack of prior information, because in this case there is no ideal prior distribution, since no single prior distribution can adequately model the lack of prior information. Therefore, in this case, also a neighbourhood model can be inadequate.
In case of total lack of prior information, Walley [5] has proposed the use of the so-called "near-ignorance" priors. This approach revises Bayesian robustness by directly emphasizing the upper and lower expectations that are generated by $\mathscr{M}$. In choosing a set $\mathscr{M}$ to model total prior ignorance, the main aim is to generate lower and upper expectations with the property that $\underline{E}(g) = \inf g$ and $\overline{E}(g) = \sup g$ for a specific class of gambles g of interest in the statistical analysis. This means

that the only information about $E(g)$ is that it belongs to $[\inf g, \sup g]$, which is equivalent to state a condition of complete prior ignorance about the value of g (this is the reason why we said that a single, however noninformative, prior cannot model prior ignorance). However, such condition of prior ignorance can only be imposed on a subset of the possible functions g (for this reason the model is called near-ignorance prior) otherwise it produces vacuous posterior inferences [5, Ch. 5]. Based on this idea, Walley [5, 12] has developed near-ignorance prior models for various statistical models: for inferences with categorical data (i.e., the so-called Imprecise Dirichlet Model); for inferences with real data [12, 13]. Starting from this work, in [14] we have derived near-ignorance prior models for all the members of the regular exponential families, which include the most used probabilistc models in statistical analysis. An issue with near-ignorance prior models is that, in some cases they may produce too uninformative inferences, for example when the observations are not precise [15]. In [16], to overcome this issue in the case of a bounded parameter space, Moral has proposed some alternative models to the Imprecise Dirichlet Model that do not satisfy near-ignorance, but that produce more meaningful inferences in those cases where the ones produced by the Imprecise Dirichlet Model seem to be too weak. Observe that, for inferences with sets of probabilities, an estimate is called robust when either it does not depend on the choice of a particular probability in the set (i.e., we return the set of all point estimates computed by considering any probability in the set) or it is calculated based on a worst-case scenario (i.e., a minimax estimate computed with respect to the most adverse probability in the set).

In this paper, we consider the case in which some partial information about the probabilistic models is available and we assume that this information can be modelled by belief functions. In this respect, the statistical models developed in this paper are close to the neighbourhood models discussed previously and, in some cases, they coincide with these models. For instance, it will be shown that the ε-contamination models are indeed belief functions [17, 18].

The use of belief functions for statistical inference has been investigated by several authors, see for instance [17–23]. Most of these approaches consider a belief function model for the likelihood and then use frequentist approaches to derive inferences. In other case, both likelihood and priors are modelled by belief functions and, then, the Dempster-Shafer calculus is used to compute inferences. An issue of these approaches is that when the belief functions and probabilities are given a betting interpretation [5, 24], then these models can incur a sure loss and,

thus, be inconsistent under a betting interpretation. This issue can be avoided if we perform the analysis in the Bayesian framework as shown by Walley [5].
In the context of Bayesian analysis, the use of belief functions in robust statistics was first proposed by Wasserman [17, 18] with the aim of building new robust prior models. In this paper, we extend this analysis by considering the case in which also the likelihood model can be modelled by a belief function. To obtain this goal, we will exploit the tools of the theory of Imprecise Probability developed by Walley. In particular, we will employ the multivalued mapping mechanism to build robust belief function models for the likelihood and the prior. Then, we will exploit the interpretation of belief functions as closed convex sets of probabilities:

$$\mathscr{P}_X = \{P : \ Bel(A) \leq P(A) \leq Pl(A), \ \ \forall A \in \mathscr{F}\},$$

and apply Walley's theory of imprecise probabilities to these sets to derive inferences. In particular, we will exploit two tools of Walley's theory: (i) marginal extension, (ii) regular extension. Given an unconditional closed[1] convex set of probabilities $\mathscr{P}_X$ and a conditional one $\mathscr{P}_{Y|X}$ (i.e., a collection of conditional closed convex sets of probabilities of Y for each given value of the conditioning variable X in $\mathscr{X}$), marginal extension builds a joint set $\mathscr{P}_{X,Y}$ which is obtained by applying the law of total probability to all pairs of probabilities in the closed convex sets $\mathscr{P}_X$ and $\mathscr{P}_{Y|X}$. Conversely, given a closed convex set of joint probabilities $\mathscr{P}_{X,Y}$ and an observation $Y = \tilde{y}$, we compute the conditional closed convex set $\mathscr{P}_{X|\tilde{y}}$ by applying Bayes' rule to all elements of the set $\mathscr{P}_{X,Y}$, which assign positive probability to the observation $\tilde{y}$. This approach is thus a straightforward generalisation of Bayesian inference to closed convex sets of probabilities.

It should be pointed out that marginal extension does not preserve the ∞-monotonicity of the set of probabilities to be combined (we will seen an example later in the paper). In other words, if the prior $\mathscr{P}_X$ and the likelihood model $\mathscr{P}_{Y|X}$ are closed convex sets of probabilities defined by a multivalued mechanism (they are belief functions), the resulting posterior set $\mathscr{P}_{X|\tilde{y}}$, that we obtain by applying first marginal extension and then generalised Bayes' rule, may be not a belief function. This means that the lower probability induced by $\mathscr{P}_{X|\tilde{y}}$, i.e.,

$$\underline{P}(A|\tilde{y}) = \inf_{P(\cdot|\tilde{y}) \in \mathscr{P}_{X|\tilde{y}}} P(A|\tilde{y}), \ \ \forall A \in \mathscr{F},$$

may be not a belief function.

[1] In the weak* topology; see [5, Sec. 3.6] for more details.

We do not see this as a weak point of the proposed inference method. In fact, many of the models used in robust statistics and based on sets of probabilities are not belief functions. For instance the lower probability defined by the following set of Normal densities with bounded mean:

$$\mathscr{P}_X = \{N(x;m,1) : \ m \in [a,b]\},$$

is not a belief function (this is shown later in the paper). This is one of the most used robust models in engineering applications. Many other counterexamples can be provided. The advantage of using Walley's theory of imprecise probability is that it can be applied to general closed convex sets of probabilities and, thus, we are not obliged to limit ourself to belief function models.

Although many useful sets of probabilities are not belief functions, the multivalued mapping mechanism is a very useful tool to build robust models. Furthermore, belief functions are advantageous from a computational point of view as it will be explained later. For these reasons, it is worth to investigate the application of belief function models to statistical inference problems; at least for all the cases in which the expressiveness of belief function is enough to model the statistical problem we are considering. The paper is organized as follows. Section 2 revises the interpretation of belief functions in terms of closed convex sets of probabilities and presents some examples. Section 3 includes the main results of the paper for the application of belief functions to statistical inference. Section 4 presents new models for robust inference based on belief functions and shows the connections of the proposed inference method based on belief functions with interval estimation and statistical inference with missing data. Finally Section 5 ends the paper.

2. Belief function

In this section we revise some properties of belief functions [17]. Let $\mathscr{X}$ be a Polish space (e.g., Euclidean space) with Borel σ-algebra $\mathscr{B}(\mathscr{X})$ and let $\mathscr{Z}$ be a convex, compact, metrizable subset of a locally convex topological vector space with Borel σ-algebra $\mathscr{B}(\mathscr{Z})$ [17]. Let P_Z be a probability measure on $(\mathscr{Z},\mathscr{B}(\mathscr{Z}))$ and let Γ be a multivalued mapping from $\mathscr{Z}$ to $2^{\mathscr{X}}$ (i.e., the power set of $\mathscr{X}$) such that, by defining $A^* = \{z \in \mathscr{Z} : \Gamma(z) \cap A \neq \emptyset\}$ for a given $A \in \mathscr{B}(\mathscr{X})$, it satisfies that $A^* \in \mathscr{B}(\mathscr{Z})$ for each $A \in \mathscr{B}(\mathscr{X})$.[2]

[2]This property of Γ is called strong measurability, we point the reader to [25] for more details.

For each $A \subseteq \mathscr{X}$, define the belief and plausibility function as [1, 17]:

$$\begin{array}{rcccl} \underline{P}(A) & = & Bel(A) & = & P_Z(\{z \in \mathscr{Z} : \Gamma(z) \subset A\}), \\ \overline{P}(A) & = & Pl(A) & = & P_Z(\{z \in \mathscr{Z} : \Gamma(z) \cap A \neq \emptyset\}). \end{array} \tag{1}$$

The fourtuple $(\mathscr{Z}, \mathscr{B}(\mathscr{Z}), P_Z, \Gamma)$ is called a source for Bel. Bel and Pl are related by $Bel(A) = 1 - Pl(A^c)$, where A^c is the complement of A. An intuitive explanation [17] of Bel and Pl is as follows. Draw z randomly according to P_Z. Then $Bel(A)$ is the probability that the random set $\Gamma(z)$ is contained in A and $Pl(A)$ is the probability that the random set $\Gamma(z)$ hits A [26]. Here, a simple example [5, Sec. 5.13.3] that explains the construction of belief functions through multivalued mappings.

Example 1. *Suppose that our information on $\mathscr{X}$ is a report from an unreliable witness that the event $B \subset \mathscr{X}$ has occurred. We might consider two possible explanations: either the witness really observed B, or he observed nothing at all. These hypotheses are represented by z_1 and z_2, with multivalued mapping $\Gamma(z_1) = B$ and $\Gamma(z_2) = \mathscr{X}$. If we assess the probability $P_Z(z_1) = q$ and $P_Z(z_2) = 1 - q$ with $q \in (0,1)$, this corresponds to the belief function $Bel(A) = q$ if $A \supseteq B$ and $A \neq \mathscr{X}$; $Bel(A) = 1$ if $A = \mathscr{X}$ and zero otherwise.*

The multivalued mapping mechanism can be used to define belief functions also in the case the sets $\mathscr{Z}$ and $\mathscr{X}$ are infinite.

Example 2. *Consider the case $\mathscr{Z} = \mathbb{R}_+$, $\mathscr{B}(\mathscr{Z})$ is the Borel σ-algebra on $\mathbb{R}_+$ and $P_Z(dz) = p_z(z)dz$, where p_z is the chi-square density function (w.r.t. the Lebesgue measure on $\mathbb{R}_+$) with one degree of freedom:*

$$p_Z(z) = \frac{z^{-1/2} e^{-z/2}}{2^{1/2} G\left(\frac{1}{2}\right)}, \;\; z > 0,$$

where $G(z)$ is the Gamma function. Furthermore, assume that $\mathscr{X} = \mathbb{R}$, $\mathscr{B}(\mathscr{X})$ is the Borel σ-algebra on $\mathbb{R}$ and consider the multivalued mapping $\Gamma(z) = \pm x = \pm\sqrt{z}$, i.e, $z = x^2$. We aim to compute the lower and upper probability of the following intervals $A = (-\infty, x]$ for each $x \in \mathscr{X}$. By definition of cumulative distribution function (CDF), the lower and upper probability of $A = (-\infty, x]$ correspond to the lower and upper CDF. By exploiting (1) it follows that:

$$\underline{P}(A) = \underline{F}(x) = \begin{cases} 0, & x \leq 0, \\ \int_0^{x^2} p_Z(z)dz, & x > 0, \end{cases} \qquad \overline{P}(A) = \overline{F}(x) = \begin{cases} \int_{x^2}^{\infty} p_Z(z)dz, & x \leq 0, \\ 1, & x > 0. \end{cases} \tag{2}$$

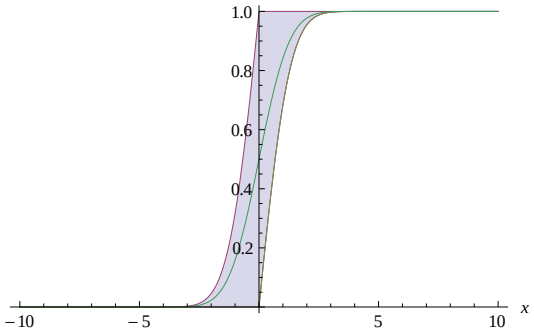

Figure 1: Lower and upper distribution function. The central line is the Normal distribution.

Figure 1 shows the lower and upper CDF together with the CDF of the standard Normal distribution. It is well known that if X is standard Normal distributed then $Z = X^2$ is chi-square distributed with one degree of freedom. Since the inverse of the relation $Z = X^2$ is a multivalued function, the converse does not hold. For instance, if $X > 0$ is chi distributed with one degree of freedom, i.e., it has density

$$p(x) = \frac{2^{\frac{1}{2}} e^{-\frac{x^2}{2}}}{G(\frac{1}{2})},$$

then X^2 is again chi-square distributed with one degree of freedom. Thus, both Normal and the chi distribution with are mapped into a chi-square distribution with one degree of freedom from the relation $Z = X^2$. Summarizing, if we know that Z is chi-square distributed and that $X = \pm\sqrt{Z}$, we can only say that the CDF of X is bounded by the lower and upper CDF in Figure 1. The CDF of the chi distribution coincides with the lower CDF. Note that, the area between the lower and upper CDF in Figure 1 may include CDFs whose transformation to $Z = X^2$ is not a chi-square distribution. Thus, if we keep only the set of CDFs bounded by the upper and lower distribution functions, we may lose information, in the sense that there can be intermediate CDFs which do not correspond with this multivalued function. This issue can be avoided by working directly with the set of probabilities induced by the multivalued mapping from the probability measure P_z. ■

Example 3. *Assume $\mathscr{X} = \mathscr{Z} = \mathbb{R}$, we discuss another model (see [17] for the derivations) generated by a multivalued map. Assume that*

$$p_Z(x) = (1-\varepsilon)\pi'_Z(x) + \varepsilon\delta_{\{z_0\}}(x),$$

where $\delta_{\{z_0\}}$ is a Dirac's delta on z_0 and π' is a probability density function (PDF) such that $\pi'_Z(z_0) = 0$ and $\pi'_Z = \pi_Z$ if $x \neq z_0$. Consider then the multivalued map $\Gamma(x) = x$ if $x \neq z_0$ and $\Gamma(x) = \mathbb{R}$ if $x = z_0$. From (1) it follows that:

$$\underline{P}(A) = (1-\varepsilon)\int_A \pi'_Z(x)dx, \;\; \overline{P}(A) = (1-\varepsilon)\int_A \pi'_Z(x)dx + \varepsilon.$$

This is the so called ε-contamination model [8] or linear-vacuous model [5, Sec. 2.9.2]. When $\varepsilon = 1$, it reduces to a vacuous model $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$ for all $A \neq \mathscr{X}$ and $\underline{P}(\mathscr{X}) = \overline{P}(\mathscr{X}) = 1$.

This lack of knowledge expressed via a belief function can equivalently[3] be represented through a set of probability measures, i.e., the set of all probabilities on X that are compatible with the bounds *Bel* and *Pl* [1]:

$$\mathscr{P}_X = \{P_X : \; Bel(A) \leq P_X(A) \leq Pl(A) \text{ for any } A \subseteq \mathscr{X}\}. \tag{3}$$

Example 4. *Consider the Example 1 with $\mathscr{X} = \{x_1, x_2, x_3\}$ and $B = \{x_1, x_2\}$. The set of probability measures induced by the belief function is the following closed convex set:*

$$\mathscr{P}_X = \left\{ p : \; p = \sum_{i=1}^{4} \alpha_i p_i, \; \alpha_i > 0, \; \sum_{i=1}^{4} \alpha_i = 1 \right\}, \tag{4}$$

where p, p_i denote probability mass functions in $\mathscr{P}_X$ and $p_1(x_1) = 1$, $p_1(x_2) = 0$, $p_1(x_3) = 0$ and $p_2(x_1) = 0$, $p_2(x_2) = 1$, $p_2(x_3) = 0$ and $p_3(x_1) = q$, $p_3(x_2) = 0$, $p_3(x_3) = 1-q$ and $p_4(x_1) = 0$, $p_4(x_2) = q$, $p_4(x_3) = 1-q$. In other words, $\mathscr{P}_x$ is the convex hull of the set of extreme probabilities:

$$Ext(\mathscr{P}_X) = \{p_1, \ldots, p_4\}.$$

Here p_1 considers the case in which all the mass is assigned to x_1, which belongs to both B and $\mathscr{X}$ (similar p_2 for x_2). p_3 and p_4 consider the case in which the mass $1 - Bel(B)$ is assigned to x_3, while the mass $Bel(B)$ is assigned respectively to x_1 or x_2. It can easily be verified that the belief function (and the plausibility function) in Example 1 satisfy:

$$Bel(A) = \underline{P}(A) = \min_{p \in \mathscr{P}_X} \sum_{x_i \in A} p(x_i), \;\; Pl(A) = \overline{P}(A) = \max_{p \in \mathscr{P}} \sum_{x_i \in A} p(x_i).$$

[3] Conditions for the equivalence between the class of the distributions induced by a random set and the lower (or upper) probability it induces are discussed [25]. In this paper we focus the attention to cases for which this equivalence holds. We point the reader to [25] for more details.

For this reason, Bel (Pl) is also called lower (upper) probability, since it is the lower (upper) envelope of a set of probability measures. Observe that the minimum/maximum are always attained by the extremes $p_1, \ldots, p_4$; this follows from the fundamental theorem of linear programming. ■

Example 5. *Consider Example 3, in this case the set of extreme probabilities is*

$$Ext(\mathscr{P}_X) = \left\{ p = (1-\varepsilon)\pi'_Z + \varepsilon \delta_{\{x_d\}} : \; x_d \in \mathbb{R} \right\}, \tag{5}$$

that is the extreme probabilities are convex combinations of the density π'_Z with all the possible Dirac's delta in $\mathbb{R}$. Given any real valued function g, the lower and upper expectations of g are obtained by these extreme probabilities. ■

Thus, associated to each belief function, there is a closed convex set of probability measures of which a belief function is a lower bound but, on the other hand, the lower bound $\underline{P}$ of a closed convex set of probability measures is not necessarily a belief function, see for instance [5]. Assuming that $\underline{P}(\emptyset) = 0$, $\underline{P}(\mathscr{X}) = 1$ and $\underline{P}(X) \geq 0$ for all $X \in \mathscr{X}$ then, to be a belief function, the lower bound $\underline{P}$ of a closed convex set of probability measures has to satisfy the property of ∞-monotonicity [2, Theorem 2.1], i.e., for every positive integer $n \geq 2$ and every collection $\mathscr{X}_1, \ldots, \mathscr{X}_n$ of elements of $\mathscr{B}(\mathscr{X})$,

$$\begin{array}{rcl} \underline{P}(\mathscr{X}_1 \cup \mathscr{X}_2 \cup \cdots \cup \mathscr{X}_n) & \geq & \sum_i \underline{P}(\mathscr{X}_i) - \sum_{i<j} \underline{P}(\mathscr{X}_i \cap \mathscr{X}_j) + \ldots \\ & + & (-1)^{n+1} \underline{P}(\mathscr{X}_1 \cap \cdots \cap \mathscr{X}_n). \end{array} \tag{6}$$

There are many closed convex sets of probabilities that are used in practice in statistical inference that are not belief functions, see next Examples 6 and 7. By restricting closed convex sets of probabilities to be belief functions one looses in generality but can gain in tractability.[4] In fact, because of the ∞-monotonicity property, belief functions satisfy several nice properties. We discuss one of the most useful properties in the next section. Besides tractability, belief functions are also an useful source of closed convex set of probabilities. This is mainly because of the multivalued mapping mechanism that can be used to define belief functions.

[4] This is not always the case as shown by the simple counterexamples 6 and 7. Inferences from both these models are relatively easy to derive.

Example 6. *Consider a Normal distribution $N(x;m,1)$ with $x \in \mathbb{R}$, mean m and variance 1. Assume that we do not know m but we know that it belongs to $[-1,1]$. Our uncertainty can be modelled by the following set of probability density functions:*

$$\mathscr{P}_X = \{N(x;m,1) : \ m \in [-1,1]\},$$

i.e., the set of all Normal densities with mean varying in $[-1,1]$. Consider then the events $\mathscr{X}_1 = [-1,-0.9] \cup [0.9,1]$ and $\mathscr{X}_2 = [-0.6,-0.5] \cup [0.9,1]$, one has that

$$\underline{P}(\mathscr{X}_1) = \inf_{m \in [-1,1]} \int_{\mathscr{X}_1} N(x;m,1)dx = 0.0456,$$

and the minimum is obtained for $m = -1$ or $m = 1$. Similarly, one has that:

$$\begin{aligned} \underline{P}(\mathscr{X}_2) &= 0.0420, \quad \textit{for } m = -1, \\ \underline{P}(\mathscr{X}_1 \cap \mathscr{X}_2) &= 0.0060, \quad \textit{for } m = -1, \\ \underline{P}(\mathscr{X}_1 \cup \mathscr{X}_2) &= 0.0578, \quad \textit{for } m = 1. \end{aligned}$$

Thus, it results that

$$0.0578 = \underline{P}(\mathscr{X}_1 \cup \mathscr{X}_2) < \underline{P}(\mathscr{X}_1) + \underline{P}(\mathscr{X}_2) - \underline{P}(\mathscr{X}_1 \cap \mathscr{X}_2) = 0.0818,$$

which violates (6) for $n = 2$. We thus conclude that the lower probability defined by the set of densities $\mathscr{P}_X$ is not a belief function (it is not ∞-monotone). ■

Example 7. *This example has been adapted from [5, Sec. 5.13.4]. Consider a multivariate Normal distribution $N(x;m,\Sigma)$ with $x = [x_1,x_2]^T \in \mathbb{R}^2$, mean $m = [0,0]^T$ and covariance matrix*

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

where the correlation coefficient ρ is completely unknown. Thus, we only know that it satisfies $\rho \in (-1,1)$.[5] *Our uncertainty can be modelled by the following set of probability density functions:*

$$\mathscr{P}_X = \{N(x;m,\Sigma) : \ \rho \in (-1,1)\}.$$

[5] We have considered an open interval $\rho \in (-1,1)$ to avoid degenerate situations in which the covariance matrix Σ has zero determinant, i.e., the cases $\rho = \pm 1$.

Consider then the events $\mathscr{X}_1 = \{x:\ x_1 \leq 0\}$ *and* $\mathscr{X}_2 = \{x:\ x_2 \leq 0\}$*, one has that*

$$\begin{aligned} \underline{P}(\mathscr{X}_1) = \overline{P}(\mathscr{X}_1) = 0.5, &\quad \textit{for any } \rho, \\ \underline{P}(\mathscr{X}_2) = \overline{P}(\mathscr{X}_2) = 0.5, &\quad \textit{for any } \rho, \\ \underline{P}(\mathscr{X}_1 \cap \mathscr{X}_2) = 0, &\quad \textit{for } \rho \to -1, \\ \overline{P}(\mathscr{X}_1 \cap \mathscr{X}_2) = 0.5, &\quad \textit{for } \rho \to 1. \end{aligned}$$

By exploiting the inequality [5, Property 2.7.4(h)]:

$$\underline{P}(\mathscr{X}_1 \cup \mathscr{X}_2) \leq \overline{P}(\mathscr{X}_1) + \overline{P}(\mathscr{X}_2) - \overline{P}(\mathscr{X}_1 \cap \mathscr{X}_2),$$

it follows that $\underline{P}(\mathscr{X}_1 \cup \mathscr{X}_2) \leq 0.5$*. Then, since*

$$\underline{P}(\mathscr{X}_1) + \underline{P}(\mathscr{X}_2) - \underline{P}(\mathscr{X}_1 \cap \mathscr{X}_2) = 1,$$

one gets

$$\underline{P}(\mathscr{X}_1 \cup \mathscr{X}_2) < \underline{P}(\mathscr{X}_1) + \underline{P}(\mathscr{X}_2) - \underline{P}(\mathscr{X}_1 \cap \mathscr{X}_2),$$

which violates (6). We thus conclude that the lower probability defined by the set of densities $\mathscr{P}_X$ *is not a belief function.* ■

2.1. Upper and lower expectation

The previous section has discussed several belief functions generated through multivalued mappings. We have also seen that a belief function can equivalently be interpreted as a lower probability model defined on the subsets of $\mathscr{X}$ and, thus, as a lower expectation model defined on the indicator functions over the subsets of $\mathscr{X}$, i.e., $\underline{E}(I_{\{A\}}) = \underline{P}(A)$. Assume that we know the functional $\underline{P}(A) = \underline{E}(I_{\{A\}})$ for any subset A of $\mathscr{X}$ how can we extend this lower probability model to compute $\underline{E}(g)$ for any bounded real-valued function of interest g.[6] The lower and upper expectations can be obtained as follows:

$$\underline{E}_X(g) = \inf_{P_X \in \mathscr{P}_X} \int g(x) P_X(dx), \ \ \overline{E}_X(g) = \sup_{P_X \in \mathscr{P}_X} \int g(x) P_X(dx). \tag{7}$$

Thus, the interpretation of belief functions as closed convex sets of probability measures allows to compute lower and upper expectations for any bounded

[6] For unbounded functions, we can define the lower (equivalently upper) expectation as the limit of a bounded restriction of g as the restriction vanishes. For instance, for $g = X$ with $X \in \mathbb{R}$, we can consider the limit for $a \to \infty$ of the lower expectation of $gI_{[-a,a]}$. We point the reader to [27] for a more rigorous definition and for issues concerning the choice of the restriction.

real valued function. Since belief functions are multivalued mapping, it has been proved in [17] that (7) is equal to:

$$\underline{E}_X(g) = \int g_*(z) P_Z(dx), \;\; \overline{E}_X(g) = \int g^*(z) P_Z(dz), \tag{8}$$

where $g_*(z) = \inf_{x \in \Gamma(z)} g(x)$ and $g^*(z) = \sup_{x \in \Gamma(z)} g(x)$. This fact has important implications for computation because it reduces the problem of calculating extrema over the set of probability measures $\mathscr{P}_X$ to that of finding extrema of g over subsets of $\mathscr{X}$ followed by a single integral over Z.

Example 8. *Consider for instance the ε-contamination model discussed in the previous section, then*

$$\begin{aligned} \underline{E}_X(g) = \int g_*(z) P_Z(dz) &= \int dz \left[(1-\varepsilon)\pi'_Z(z) + \varepsilon \delta_{\{z_0\}}(z) \right] \inf_{x \in \Gamma(z)} g(x), \\ &= \int_{\mathscr{Z} - \{z_0\}} (1-\varepsilon)\pi'_Z(z) g(z) dz + \varepsilon \inf_{x \in \mathbb{R}} g(x) \\ &= \int (1-\varepsilon)\pi_Z(z) g(z) dz + \varepsilon \inf_{x \in \mathbb{R}} g(x). \end{aligned} \tag{9}$$

■

2.2. Conditional models

In the previous sections, we have discussed unconditional models generated from a probability space through a multivalued mapping. We can easily extend the previous results to conditional models. Let $P_Z(\cdot|z_o)$ be a conditional probability measure on $(\mathscr{Z}, \mathscr{B}(\mathscr{Z}))$ for each value $z_o \in \mathscr{Z}_o$ and let $\{\Gamma(\cdot|z_o) : \; z_o \in \mathscr{Z}_o\}$ be a set of multivalued maps parametrized by z_o and taking points in $\mathscr{Z}$ to nonempty, closed subsets of $\mathscr{X}$. For each $A \subseteq \mathscr{X}$, define the conditional belief and plausibility function as:

$$\begin{array}{rcccl} \underline{P}(A|z_o) & = & Bel(A|z_o) & = & P_Z(\{z_i \in \mathscr{Z} : \Gamma(z_i|z_o) \subset A\}|z_o), \\ \overline{P}(A|z_o) & = & Pl(A|z_o) & = & P_Z(\{z_i \in \mathscr{Z} : \Gamma(z_i|z_o) \cap A \neq \emptyset\}|z_o), \end{array} \tag{10}$$

for each value $z_o \in \mathscr{Z}_o$. Hence, equation (8) becomes:

$$\underline{E}(g|z_o) = \int g_*(z) P_Z(dx|z_o), \;\; \overline{E}(g|z_o) = \int g^*(z) P_Z(dz|z_o), \tag{11}$$

where $g_*(z) = \inf_{x \in \Gamma(z|z_o)} g(x)$ and $g^*(z) = \sup_{x \in \Gamma(z|z_o)} g(x)$.

Example 9. *Consider again Example 8 in case* $\pi_Z(z|z_0) = \mathscr{N}(z; z_o, 1)$ *and* $\Gamma(z|z_o) = [z_0 - a, z_0 + a]$ *with* $a > 0$*. Then, one gets*

$$\underline{E}_X(g|z_o) = (1-\varepsilon)\int g(z)\mathscr{N}(z; z_o, 1)dz + \varepsilon \inf_{x\in[z_o-a,z_o+a]} g(x).$$

■

3. Statistical inference

Consider the problem of statistical inference about a variable X from measurements $\tilde{y}^n = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$ of the variables $Y_1, \ldots, Y_n$. Assume that we have some prior information over X which is expressed through a belief function or, equivalently, through the closed convex set of probability measures associated to the belief function. A belief function model can also be assumed for the observation process. How can we compute the lower/upper posterior expectation of a bounded real-valued function g of X given the observations $\tilde{y}_1, \ldots, \tilde{y}_n$?

Before stating the solution of the above inference for belief functions, it is useful to show how standard Bayesian inference can be formulated in terms of expectations.

Theorem 1. *Assume that our information on the initial state and observation model is represented by the expectation* E_X *and, respectively, conditional expectations* $E_{Y_k}[\cdot|X]$ *for any* $k = 1, \ldots, n$*. Furthermore, assume that, the variables* $Y_1, \ldots, Y_n$ *are conditionally independent given* X*, which implies that*

$$E_{Y^n}[h|x] \;\; = \;\; h_0(x)\prod_{k=1}^{n} E_{Y_k}[h_k|x], \tag{12}$$

for any given $x \in \mathscr{X}$ *and for any bounded real-valued function* $h : \mathscr{X} \times \mathscr{Y}^n \to \mathbb{R}$ *such that* $h = h_0 \prod_{k=1}^{n} h_k$ *with* $h_0 : \mathscr{X} \to \mathbb{R}$ *and* $h_k : \mathscr{Y}_k \to \mathbb{R}$*. Then, assuming that* $E_{X,Y^n}[\prod_{k=1}^{n} I_{\{\tilde{y}_k\}}] > 0$*, where* $I_{\{\tilde{y}_k\}}$ *denotes the indicator function of the observation* $\tilde{y}_k$ *and given the sequence of measurements* $\tilde{y}^n$*, the posterior expectation* $E_X[g|\tilde{y}^n]$ *for any bounded real-valued function* $g : \mathscr{X} \to \mathbb{R}$ *is equal to the unique value* $\nu \in \mathbb{R}$ *that solves the following optimization problem:*

$$\sup \nu \;\; s.t. \;\; E_{X,Y^n}\left[(g-\nu)\prod_{k=1}^{n} I_{\{\tilde{y}_k\}}\right] \geq 0, \tag{13}$$

where the above joint expectation is given by:

$$E_X\left[(g-\nu)\prod_{k=1}^{n}E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right]. \tag{14}$$

■

PROOF. By exploiting the law of total expectation (also called law of iterated expectations), the joint in (13) can be rewritten as:

$$E_{X,Y^n}\left[(g-\nu)\prod_{k=1}^{n}I_{\{\tilde{y}_k\}}\right]=E_X\left[E_{Y^n}\left[(g-\nu)\prod_{k=1}^{n}I_{\{\tilde{y}_k\}}\middle|X\right]\right].$$

Since $(g-\nu)$ is a function of X only, from (12) one has that:

$$E_X\left[E_{Y^n}\left[(g-\nu)\prod_{k=1}^{n}I_{\{\tilde{y}_k\}}\middle|X\right]\right]=E_X\left[(g-\nu)\prod_{k=1}^{n}E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right].$$

By linearity of the expectation, since ν is a constant, one has that

$$\begin{aligned}E_X\left[(g-\nu)\prod_{k=1}^{n}E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right] &= E_X\left[g\prod_{k=1}^{n}E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right]\\ &- \nu E_X\left[\prod_{k=1}^{n}E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right]\end{aligned}. \tag{15}$$

Equation (13) states that the posterior expectation $E_X[g|\tilde{y}^n]$ is the supremum value of ν such that (15) is non-negative. Since by assumption:

$$E_{X,Y^n}[\prod_{k=1}^{n}I_{\{\tilde{y}_k\}}]=E_X\left[\prod_{k=1}^{n}E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right]>0,$$

we can divide (15) by $E_{X,Y^n}[\prod_{k=1}^{n}I_{\{\tilde{y}_k\}}]>0$ and thus from (13) obtain

$$\sup\nu\ \ s.t.\ \ \frac{E_X\left[g\prod_{k=1}^{n}E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right]}{E_X\left[\prod_{k=1}^{n}E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right]}-\nu\geq 0. \tag{16}$$

The solution of (16) is thus

$$\nu = E_X[g|\tilde{y}^n] = \frac{E_X\left[g\prod_{k=1}^{n} E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right]}{E_X\left[\prod_{k=1}^{n} E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right]}, \tag{17}$$

this in fact is the supremum value of ν such that the inequality in (16) is satisfied. ■

Assume for instance that $\mathscr{X}, \mathscr{Y}_i$ are finite sets and, thus, the expectations are completely defined by the probability mass functions $p(x)$, $p(y_i|x)$, then the solution of (16) is:

$$\begin{aligned}\nu &= E_X[g|\tilde{y}^n] = \frac{E_X\left[g\prod_{k=1}^{n} E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right]}{E_X\left[\prod_{k=1}^{n} E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right]} = \frac{\sum_{x\in\mathscr{X}} g(x)p(x)\prod_{k=1}^{n} p(\tilde{y}_k|x)}{\sum_{x\in\mathscr{X}} p(x)\prod_{k=1}^{n} p(\tilde{y}_k|x)} \\ &= \sum_{x\in\mathscr{X}} g(x)p(x|\tilde{y}^n) = E_X[g|\tilde{y}^n],\end{aligned}$$

where $p(x|\tilde{y}^n)$ denotes the posterior probability of X given $\tilde{y}^n$. This shows that (13) is just the formulation of Bayes' rule in terms of expectations. Observe that, in case $\mathscr{Y}_k \subseteq \mathbb{R}^m$ and $E_{Y_k}[\cdot|X]$ is the expectation w.r.t. a probability measure that is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^m$, i.e., $E_{Y_k}[h|x] = \int h(y_k)p(y_k|x)dy_k$ where $p(y_k|x)$ is a probability density function, then $E_{Y_k}[I_{\{\tilde{y}_k\}}|x] = 0$ since any singleton set has zero measure. Thus, Theorem 1 cannot be applied because the condition $E_{X,Y^n}[\prod_{k=1}^{n} I_{\{\tilde{y}_k\}}] > 0$ is not met. A way to overcome such issue in Bayesian estimation is to replace the observation $\tilde{y}_k$ with nested neighbourhoods $B(\tilde{y}_k, \gamma)$ (a ball of radius γ, which should not depend on x). Then, if the density $p(y_k|x)$ is continuous, bounded and positive in these neighbourhoods, the posterior is obtained by taking the limit of the fraction in (17) for $\gamma \to 0$. This gives Bayes' rule for density functions, for more details see [5, Sec. 6.10]. In this case, the posterior can equivalently be obtained from Theorem 1 by replacing the indicator $I_{\{\tilde{y}_k\}}$ by a Dirac's delta on $\tilde{y}_k$ and by exploiting $E_{Y_k}[\delta_{\{\tilde{y}_k\}}|x] = p(\tilde{y}_k|x)$, i.e.,

$$E_X[g|\tilde{y}^n] = \frac{E_X\left[g\prod_{k=1}^{n} E_{Y_k}\left[\delta_{\{\tilde{y}_k\}}\middle|X\right]\right]}{E_X\left[\prod_{k=1}^{n} E_{Y_k}\left[\delta_{\{\tilde{y}_k\}}\middle|X\right]\right]} = \frac{\int_{\mathscr{X}} g(x)p(x)\prod_{k=1}^{n} p(\tilde{y}_k|x)dx}{\int_{\mathscr{X}} p(x)\prod_{k=1}^{n} p(\tilde{y}_k|x)dx}.$$

In the following we extend Theorem 1 to the case where the prior and/or the likelihood are modelled by lower expectation models (for instance induced by a belief function). To obtain such extension, we must generalise the law of iterated expectations and conditional independence to lower expectation models. There is not a unique way to perform such generalisation, we will use marginal extension [5, Th. 6.7.2] and, respectively, strong independence because they have a Bayesian sensitivity analysis interpretation: (i) to apply marginal extension is equivalent to apply the law of iterated expectations to all the expectations obtained from the closed convex sets of probabilities $\mathscr{P}_X$ and $\mathscr{P}_{Y|X}$; (ii) to say that Y_1 and Y_2 are strongly independent given X is equivalent to say that they are stochastically independent for all extreme points of the joint closed convex set of probabilities $\mathscr{P}_{X,Y_1,Y_2}$.

Definition 1. *Let $\underline{E}_X$ and $\underline{E}_Y[\cdot|X]$ be respectively an unconditional and conditional lower expectation model defined by $\mathscr{P}_X$ and, respectively, $\mathscr{P}_{Y|X}$. The marginal extension of $\underline{E}_X$ and $\underline{E}_Y[\cdot|X]$ is the joint lower expectation:*

$$\underline{E}_{X,Y}[h] = \underline{E}_X[\underline{E}_Y[h|X]], \tag{18}$$

for any bounded real-valued function h on $\mathscr{X} \times \mathscr{Y}$ [5, Sec. 6.7].

Definition 2. *Let $\underline{E}_{Y_i}[\cdot|X]$ be conditional lower expectation models for $i = 1,\ldots,m$ defined by the closed convex sets $\mathscr{P}_{Y_i|X}$. We call strong extension of the $\mathscr{P}_{Y_i|X}$ the joint conditional closed convex set of probabilities $\mathscr{P}_{Y^m}[\cdot|x] = \mathscr{P}_{Y_1,\ldots,Y_m}[\cdot|x]$ whose extremes are obtained by element-wise combining all the extremes of the $\mathscr{P}_{Y_i|x}$ for all $x \in \mathscr{X}$ [28].*

Consider three variables, X, Y_1, Y_2, and the joint closed convex set of probabilities $\mathscr{P}_{X,Y_1,Y_2}$ obtained from $\mathscr{P}_{Y_1|X}$, $\mathscr{P}_{Y_2|X}$ and $\mathscr{P}_X$ by applying strong extension and marginal extension. Then, Y_1 and Y_2 are strongly independent conditional on X, meaning that Y_1 and Y_2 are stochastically independent given $X = x$ for all the extreme points of $\mathscr{P}_{X,Y_1,Y_2}$ and for all $x \in \mathscr{X}$ [28].

Now we are ready to extend Theorem 1 to lower expectations.

Theorem 2. *Assume that our information on the initial state and observation model are represented by the lower expectation $\underline{E}_X$ and, respectively, conditional lower expectation $\underline{E}_{Y_k}[\cdot|X]$ for any $k = 1,\ldots,n$. Furthermore, assume that, the joint $\underline{E}_{Y^n}[\cdot|x]$ is obtained by strong extension from the $\underline{E}_{Y_k}[\cdot|X]$, which implies that*

$$\underline{E}_{Y^n}[h|x] \;=\; h_0^+(x)\prod_{k=1}^{n}\underline{E}_{Y_k}[h_k|x] + h_0^-(x)\prod_{k=1}^{n}\overline{E}_{Y_k}[h_k|x], \tag{19}$$

for any given $x \in \mathscr{X}$ and for any bounded real-valued function $h : \mathscr{X} \times \mathscr{Y}^n \to \mathbb{R}$ such that $h = h_0 \prod_{k=1}^n h_k$ with $h_0 : \mathscr{X} \to \mathbb{R}$, $h_k : \mathscr{Y}_k \to \mathbb{R}$ and where $h_0^+(x) = \max(h_0(x), 0)$ and $h_0^-(x) = \min(h_0(x), 0)$. Then, assuming that $\overline{E}_{X,Y^n}[\prod_{k=1}^n I_{\{\tilde{y}_k\}}] > 0$ and given the sequence of measurements $\tilde{y}^n$, the posterior lower expectation $\underline{E}_X[g|\tilde{y}^n]$ for any bounded real-valued function $g : \mathscr{X} \to \mathbb{R}$ is defined as the value $\nu \in \mathbb{R}$ that solves the following optimization problem:

$$\sup \nu \;\; s.t. \;\; \underline{E}_{X,Y^n}\left[(g-\nu)\prod_{k=1}^n I_{\{\tilde{y}_k\}}\right] \geq 0, \tag{20}$$

where the above joint lower expectation is given by:

$$\underline{E}_X\left[(g-\nu)^+ \prod_{k=1}^n \underline{E}_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right] + (g-\nu)^- \prod_{k=1}^n \overline{E}_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right]. \tag{21}$$

■

PROOF. By exploiting marginal extension, the joint in (20) can be rewritten as:

$$\underline{E}_{X,Y^n}\left[(g-\nu)\prod_{k=1}^n I_{\{\tilde{y}_k\}}\right] = \underline{E}_X\left[\underline{E}_{Y^n}\left[(g-\nu)\prod_{k=1}^n I_{\{\tilde{y}_k\}}\middle|X\right]\right].$$

Since $g-\nu$ is a function of X only, from the definition of lower expectation in (7) it follows that:

$$\begin{aligned}
&\underline{E}_X\left[\underline{E}_{Y^n}\left[(g-\nu)\prod_{k=1}^n I_{\{\tilde{y}_k\}}\middle|X\right]\right] \\
&= \underline{E}_X\left[(g-\nu)^+\underline{E}_{Y^n}\left[\prod_{k=1}^n I_{\{\tilde{y}_k\}}\middle|X\right] + (g-\nu)^-\overline{E}_{Y^n}\left[\prod_{k=1}^n I_{\{\tilde{y}_k\}}\middle|X\right]\right],
\end{aligned}$$

where $(g(x)-\nu)^+ = \max(g(x)-\nu, 0)$ and $(g(x)-\nu)^+ = \min(g(x)-\nu, 0)$. Note in fact that, since $g-\nu$ is a function of X only, one has that:

$$\begin{aligned}
&\underline{E}_{Y^n}\left[(g-\nu)\prod_{k=1}^n I_{\{\tilde{y}_k\}}\middle|X\right] = \inf_{p\in\mathscr{P}_{Y^n|x}} \int (g(x)-\nu)\prod_{k=1}^n I_{\{\tilde{y}_k\}}(y_k)dP(y^n|x) \\
&= (g(x)-\nu)^+ \inf_{p\in\mathscr{P}_{Y^n|X}} \int \prod_{k=1}^n I_{\{\tilde{y}_k\}}(y_k)dP(y^n|x) \\
&+(g(x)-\nu)^- \sup_{p\in\mathscr{P}_{Y^n|X}} \int \prod_{k=1}^n I_{\{\tilde{y}_k\}}(y_k)dP(y^n|x).
\end{aligned}$$

By exploiting (19), one has that:

$$\underline{E}_{Y^n}\left[\prod_{k=1}^{n} I_{\{\tilde{y}_k\}}\middle|X\right] = \prod_{k=1}^{n} \underline{E}_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right],$$

and that

$$\overline{E}_{Y^n}\left[\prod_{k=1}^{n} I_{\{\tilde{y}_k\}}\middle|X\right] = \prod_{k=1}^{n} \overline{E}_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right].$$

The lower posterior expectation $\underline{E}_X[g|\tilde{y}^n]$ is obtained through regular extension [5, Appendix J], i.e., by solving:

$$\sup \nu \;\; s.t. \;\; \underline{E}_X\left[(g-\nu)^+ \prod_{k=1}^{n} \underline{E}_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right] + (g-\nu)^- \prod_{k=1}^{n} \overline{E}_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|X\right]\right] \geq 0. \tag{22}$$

■

Observe that, the lower posterior expectation of g computed from (20) can equivalently be obtained by (i) applying Bayes' rule to all members of the closed convex set of joint probabilities associated to $\underline{E}_{X,Y^n}$ which assign positive mass to the observations; (ii) computing the expectation of g for each posterior obtained at the previous step; (iii) compute the lower envelope of all the posterior expectations computed at the previous step. Note that (20) is a version of Walley's generalised Bayes' rule which is called Regular Extension [5, Appendix J].

In Section 2.1, we have seen that the mathematical structure of belief functions allows to simplify the computations of lower and upper expectations. Hereafter, we use this fact to specialize the results of Theorem 2 to the following cases:

- $\underline{E}_X$ and $\underline{E}_{Y_k}[\cdot|X]$ are lower expectations induced by a multivalued mapping (i.e., lower expectations w.r.t. a belief function);
- $\underline{E}_X$ is a lower expectation induced by a multivalued mapping while $\underline{E}_{Y_k}[\cdot|X] = \overline{E}_{Y_k}[\cdot|X] = E_{Y_k}[\cdot|X]$, i.e., the likelihood model can be described by a single probabilistic model.

Corollary 1. *Consider Theorem 2 in case $\underline{E}_X$ and $\underline{E}_{Y_k}[\cdot|X]$ are lower expectations induced by a multivalued mapping then the joint in (21) is*

$$\begin{aligned}&\int_{\mathscr{Z}} P_Z(dz) \inf_{x\in\Gamma_Z(z)} \left[(g(x)-\nu)^+ \prod_{i=1}^{k} \int_{\mathscr{U}_k} P_{U_k|x}(du_k|x) \inf_{y_k\in\Gamma_{Y_k}(u_k|x)} I_{\{\tilde{y}_k\}}(y_k)\right.\\ &\left.+(g(x)-\nu)^- \prod_{i=1}^{k} \int_{\mathscr{U}_k} P_{U_k|x}(du_k|x) \sup_{y_k\in\Gamma_{Y_k}(u_k|x)} I_{\{\tilde{y}_k\}}(y_k)\right],\end{aligned} \tag{23}$$

where Γ_Z maps the probability measure P_Z defined in $(\mathscr{Z}, \mathscr{B}(\mathscr{Z}))$ to $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$ and $\Gamma_{Y_k}(\cdot|x)$ maps $P_{U_k}(\cdot|x)$ defined in $(\mathscr{U}_k, \mathscr{B}(\mathscr{U}_k))$ to $(\mathscr{Y}_k, \mathscr{B}(\mathscr{Y}_k))$. ■

PROOF. This can be derived from (21) by using the expression for the lower expectation in (8).

Corollary 2. *Consider Theorem 2 in case the observation process satisfies $\underline{E}_{Y_k}[\cdot|X] = \overline{E}_{Y_k}[\cdot|X] = E_{Y_k}[\cdot|X]$, i.e., the likelihood model can be described by a single probability. Then, in case $\mathscr{Y}_k$ is discrete, since $E_{Y_k}\left[I_{\{\tilde{y}_k\}}\middle|x\right] = p(\tilde{y}_k|x)$, where $p(\tilde{y}_k|x)$ is the conditional probability mass function of $\tilde{y}_k$ given x, the joint in (21) becomes:*

$$\underline{E}_X\left[(g-v)\prod_{k=1}^{n} p(\tilde{y}_k|x)\right], \tag{24}$$

and in case $\underline{E}_X$ is the lower expectation induced by a multivalued mapping it becomes

$$\int_{\mathscr{Z}} P_Z(dz) \inf_{x\in\Gamma_Z(z)} (g(x)-v)\prod_{k=1}^{n} p(\tilde{y}_k|x). \tag{25}$$

Finally, in case $\mathscr{Y}_k \subseteq \mathbb{R}^m$, the expressions (24) and (25) still hold but now $p(\tilde{y}_k|x)$ must be interpreted as the probability density function of $\tilde{y}_k$ given x. ■

PROOF. This can easily be derived from Corollary 1.

4. Applications

In this section, by exploiting the results of Section 3, we will show how belief functions can be employed to draw robust inferences in practical statistical problems.

4.1. Interval estimation

Consider the following statistical model:

$$\begin{cases} x \in \mathscr{X}, \\ y_i = x + v_i, \quad v_i \in \mathscr{V}, \end{cases} \tag{26}$$

for $i = 1,\ldots,n$, where the only prior information on the value x of X is that $x \in \mathscr{X} \subset \mathbb{R}$ and also the noise term satisfies $v_i \in \mathscr{V} \subset \mathbb{R}$. The goal is to estimate x given n observations $\tilde{y}_i$. We assume that the sets $\mathscr{X}$ and $\mathscr{V}$ are closed symmetric

intervals including the origin and with known width $\rho_0 > 0$ and, respectively, $\rho_v > 0$. We can model this information by the following (belief function) lower expectation models:

$$\begin{array}{rcl} \underline{E}_X[g] & = & \inf\limits_{|x|\leq\rho_0} g(x), \\ \underline{E}_{Y_i}[h|x] & = & \inf\limits_{|y_i-x|\leq\rho_v} h(x,y_i), \end{array} \tag{27}$$

for any given x and bounded real-valued functions g and h. Observe in fact that, the information that x belongs to $\mathscr{X}$ can be modelled by the set of all probability measures with support in $\mathscr{X}$. Hence, it follows that $\underline{E}_X[g] = \inf_{|x|\leq\rho_0} g(x)$, since the infimum is obtained by a Dirac's delta centred at $x_\ell = \arg\inf_{|x|\leq\rho_0} g(x)$. This is called vacuous restricted model. A similar consideration holds for the conditional model of Y_i given x.

We can now exploit Corollary 1 to compute the posterior lower expectation of a bounded real-valued function g of X given n observations $\tilde{y}^n$. Let us start to write down $\underline{E}_{X,Y^n}$. From (23), one has that

$$\begin{array}{rcl} \underline{E}_{X,Y^n}\left[(g-\nu)\prod\limits_{i=1}^n I_{\{\tilde{y}_i\}}\right] & = & \inf\limits_{\{x:|x|\leq\rho_0\}}\Big[(g(x)-\nu)^+\prod\limits_{i=1}^n \inf\limits_{\{y_i:\ |y_i-x_i|\leq\rho_v\}} I_{\{\tilde{y}_i\}}(y_i) \\ & + & (g(x)-\nu)^-\prod\limits_{i=1}^n \sup\limits_{\{y_i:\ |y_i-x_i|\leq\rho_v\}} I_{\{\tilde{y}_i\}}(y_i)\Big]. \end{array} \tag{28}$$

Since

$$\inf_{\{y_i:\ |y_i-x|\leq\rho_v\}} I_{\{\tilde{y}_i\}}(y_i) = 0,$$

and

$$\sup_{\{y_i:\ |y_i-x|\leq\rho_v\}} I_{\{\tilde{y}_i\}}(y_i) = I_{\{x:\ |\tilde{y}_i-x|\leq\rho_v\}}, \tag{29}$$

equation (28) reduces to

$$\underline{E}_{X,Y^n}\left[(g-\nu)\prod_{i=1}^n I_{\{\tilde{y}_i\}}\right] = \inf_{\{x:|x|\leq\rho_0\}}\Big[(g(x)-\nu)^-\prod_{i=1}^n I_{\{x:\ |\tilde{y}_i-x|\leq\rho_v\}}(x)\Big]. \tag{30}$$

Observe that the product of the indicators is only different from zero for the values of x which satisfy $x \in \{\bigcap_{i=1}^n \mathscr{X}_i(\tilde{y}_i)\}$, $\mathscr{X}_i(\tilde{y}_i) = \{x:\ |\tilde{y}_i - x| \leq \rho_v\}$. Thus, (30) can be rewritten as

$$\underline{E}_{X,Y^n}\left[(g-\nu)\prod_{i=1}^n I_{\{\tilde{y}_i\}}\right] = \inf_{\{\bigcap_{i=1}^n \mathscr{X}_i(\tilde{y}_i)\}\cap\{x:|x|\leq\rho_0\}}(g(x)-\nu)^-. \tag{31}$$

Since $(g(x)-\nu)^-$ is always non-positive, the supremum value of ν such that (31) is non-negative is

$$\nu = \inf_{\{\bigcap_{i=1}^n \mathscr{X}_i(\tilde{y}_i)\}\cap\{x:|x|\leq\rho_0\}} g(x), \tag{32}$$

which, by (20), is the lower posterior expectation of g given the observations $\tilde{y}^n$. Observe, that the above intersections are non empty because of the assumption in Theorem 2:

$$\overline{E}_{X,Y^n}\left[\prod_{i=1}^n I_{\{\tilde{y}_i\}}\right] > 0.$$

Consider for instance the case $g = X$, i.e., we aim to compute the lower posterior mean of X, i.e., $\underline{E}_X[X|\tilde{y}^n]$. From (32), one has that

$$\nu = \underline{E}_X[X|\tilde{y}^n] = \max(-\rho_0, \tilde{y}_1 - \rho_v, \ldots, \tilde{y}_n - \rho_v),$$

and $\overline{E}_X[X|\tilde{y}^n] = -\underline{E}_X[-X|\tilde{y}^n] = \min(\rho_0, \tilde{y}_1 + \rho_v, \ldots, \tilde{y}_n + \rho_v)$. It can be noticed that the lower/upper posterior mean can be simply computed by applying interval estimation.

Actually, in interval estimation one aims to compute the posterior support of X and not the lower and upper posterior mean. We can easily show that the interval $[\underline{E}_X[X|\tilde{y}^n], \overline{E}_X[X|\tilde{y}^n]]$ coincides with the posterior support.

The posterior support can in fact be obtained by setting $g(x) = I_{[r,s]}(x) - 1$ and looking for the smallest interval $[r,s]$ that has lower posterior probability of g equal to zero. For Bayesian models, this is in fact equivalent to determine the 100% posterior Bayesian credible interval:

$$\begin{aligned} &\min s - r, \quad s.t. \\ &0 = E_X[I_{[r,s]}(x) - 1|\tilde{y}^n] = \int_{\mathscr{X}} (I_{[r,s]}(x) - 1) dP(x|\tilde{y}^n) \\ &= \int_r^s dP(x|\tilde{y}^n) - 1. \end{aligned} \tag{33}$$

For a set of probabilities, we just determine the smallest interval that has lower probability equal to 1 so that the constraint in (33) holds for any probability in the set. Thus for $g(x) = I_{[r,s]}(x) - 1$, (32) becomes

$$\underline{E}_{X_1}[g|\tilde{y}^n] = \min_{\{\bigcap_{i=1}^n \mathscr{X}_i\}\cap\{x:|x|\leq\rho_0\}} I_{[r,s]}(x) - 1.$$

It is clear that the smallest interval $[r,s]$ such that $\underline{E}_{X_1}[g|\tilde{y}^n] = 0$ is again

$$\max(-\rho_0, \tilde{y}_1 - \rho_v, \ldots, \tilde{y}_n - \rho_v) \leq x \leq \min(\rho_0, \tilde{y}_1 + \rho_v, \ldots, \tilde{y}_n + \rho_v).$$

Thus, interval estimation can be seen as the application of the the statistical inference procedure discussed in Section 3 to the simplest belief function models, i.e., restricted vacuous models. Here, we have discussed the solution of interval estimation in the univariate (scalar) case. In the multivariate case, instead of the constraints (26) we would have something like $||x||_p \leq \gamma_0$ for some p-norm (same for the measurement noise). The above derivations can straightforwardly be generalised to this case. This Section has shown that interval estimation can be formulated in the realm of probability. This is an interesting result because it shows that interval estimation is not a deterministic approach, but it can be interpreted as a probabilistic approach based on belief functions (more in general closed convex sets of probabilities). The advantage of seeing interval estimation under this view is that we can generalize it in case some additional information is available (besides the support) that can be modelled by belief functions. Then, we can easily include this information in our model and use Corollary 1 to derive less conservative posterior inferences.

4.2. Fat tailed prior model

Assume that our information about $X \in \mathbb{R}$ can be described by a Normal density $p(x) = \mathcal{N}(x;0,1)$ with zero mean and unit variance in the interval $[-1.65, 1.65]$ (this the 90% credible interval for X based on $p(x)$) but we do not know how to assign the remaining mass. We model this lack of information by means of the following multivalued map Γ.

$$\Gamma(x) = \begin{cases} x, & \textit{if } x \in (-1.65, 1.65), \\ [1.65, \infty), & \text{if } x \in [1.65, \infty), \\ (-\infty, -1.65], & \text{if } x \in (-\infty, -1.65]. \end{cases}$$

Observe that, here we are considering a multivalued map from $\mathscr{Z} = \mathbb{R}$ to $\mathscr{X} = \mathbb{R}$, with $p_z(z) = \mathcal{N}(z;0,1)$ and Γ defined above. Since $\mathscr{Z} = \mathscr{X}$ with a bit of abuse of notation, we have defined Γ as a map from $\mathscr{X}$ to itself. In fact, our aim is to use the multivalued mapping mechanism to model our uncertainty about the probability of X on the tails $[1.65, \infty)$ and $(-\infty, -1.65]$. This uncertainty is modelled by Γ that maps all the points x in the right-tail (left-tail) of the Normal distribution to the interval $[1.65, \infty)$ (respectively $(-\infty, -1.65]$).

The closed-convex set of priors defined by the above multivalued map mechanism can equivalently be characterized by the lower expectation model:

$$\underline{E}_X[g] = \int dx \mathcal{N}(x;0,1) \inf_{w \in \Gamma(x)} g(w), \tag{34}$$

for any bounded real-valued function g of X. Consider for instance the following two cases:

1. for $g = X$, one obtains $\underline{E}_X[X] = -\infty$ and $\overline{E}_X[X] = \infty$;
2. for $g = I_{(-\infty,x']}$, one obtains

$$\underline{E}_X[I_{(-\infty,x']}] = \begin{cases} 0, & \text{if } x' \in (-\infty, -1.65), \\ \int_{-\infty}^{x'} \mathscr{N}(x;0,1)dx, & \text{if } x' \geq -1.65; \end{cases}$$

$$\overline{E}_X[I_{(-\infty,x']}] = \begin{cases} \int_{-\infty}^{-1.65} \mathscr{N}(x;0,1)dx, & \text{if } x' \in (-\infty, -1.65), \\ \int_{-\infty}^{x'} \mathscr{N}(x;0,1)dx, & \text{if } x' \in [-1.65, 1.65], \\ 1 & \text{if } x' \geq 1.65. \end{cases}$$

which are respectively the lower and upper prior CDF of X.

The set of extreme priors which, for any g, attain the lower expectation $\underline{E}_X$ is:

$$Ext(\mathscr{P}_X) = \Big\{ p \;=\; I_{(-1.65,1.65)}\mathscr{N}(x;0,1) + r_l\delta_{x_l} + r_u\delta_{x_u} : \\ x_l \in [1.65,\infty),\ x_u \in (-\infty,-1.65] \Big\},$$

with $r_l = r_u = \int_{-\infty}^{-1.65} \mathscr{N}(x;0,1)dx$.

Assume the likelihood distribution is the Normal $p(y|x) = \mathscr{N}(y;x,1)$ and that a sequence of $n = 4$ observations, with sample mean $\hat{y}_n$, is available for inference. This means that $p(\hat{y}_n|x) = \mathscr{N}(\hat{y}_n;x,1/4)$. We can then exploit Corollary 2 to compute the lower (upper) posterior expectation of any bounded real-valued function g of X given the n observations, i.e.,:

$$\sup \nu \;\; s.t. \;\; \int dx \mathscr{N}(x;0,1) \inf_{w \in \Gamma(x)} (g(w) - \nu)\mathscr{N}(\hat{y}_n;w,1/n) \geq 0. \tag{35}$$

The solution of (35) gives the lower posterior expectation of $\underline{E}_X[g|\hat{y}_n]$; the upper can be computed as $\overline{E}_X[g|\hat{y}_n] = -\underline{E}_X[-g|\hat{y}_n]$. Figure 2 compares the lower and upper posterior mean obtained by solving (35) with the posterior mean computed by applying Bayesian inference to the likelihood $\mathscr{N}(\hat{y}_n;x,1/4)$ and the Normal prior $\mathscr{N}(x;0,1)$ (in this case the posterior mean is $E[X|\hat{y}_n] = \frac{4}{5}\hat{y}_n$) and, respectively, the posterior mean computed by applying Bayesian inference to the likelihood $\mathscr{N}(\hat{y}_n;x,1/4)$ and a t-Student prior with zero mean and 1 degree of freedom (the posterior mean has been computed numerically in this case). Note that, when the sample mean $\hat{y}_n$ belongs to $[0,1.65]$, the difference between upper and

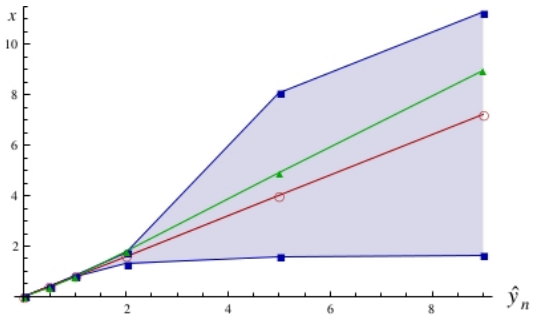

Figure 2: Lower and upper posterior mean (blue-square), posterior mean based on the Normal prior (red-circle) and posterior mean based on the t-Student prior (green-triangle).

lower mean is approximately zero. Conversely, when $\hat{y}_n$ gets larger, this difference grows highlighting the conflict between prior and likelihood model. Thus, when $\hat{y}_n \in [0, 1.65]$, the tail of prior affects only minimally the posterior inference and, thus, we can equivalently choose a Normal prior or a t-student prior or our belief function model, since they produce the same inferences, i.e., there is no issue of robustness. Conversely, when $\hat{y}_n > 1.65$, the choice of the tail behaviour is critical and our inference strongly depends on this choice. The advantage of using our belief model in this case is that it has the worst tail behaviour and, thus, it includes all the posterior inferences computed via a Bayesian analysis w.r.t. any choice of the tails of the prior. Thus, inferences based on the belief function model are maximally robust to the tail behaviour.

4.3. Hierarchical Belief function priors

The set of priors considered in the previous section can sometimes be too conservative because of the presence of the Dirac's deltas. In some cases, we may want to restrict the closed-convex set of priors to include only absolutely continuous probability measures, i.e., a closed-convex set of probability density functions. We have already seen some examples of this kind of models in Section 2, for instance the set of Normal densities with mean belonging to an interval. The lower probability induced by this set of densities is not a belief function, however we can see it as a hierarchical model generated by a belief function.

Consider the Normal prior $N(x; m, \sigma^2)$ with $x \in \mathbb{R}$ and assume that we do not know the mean m but we know that

$$m \in [a, b] \subset \mathbb{R}, \tag{36}$$

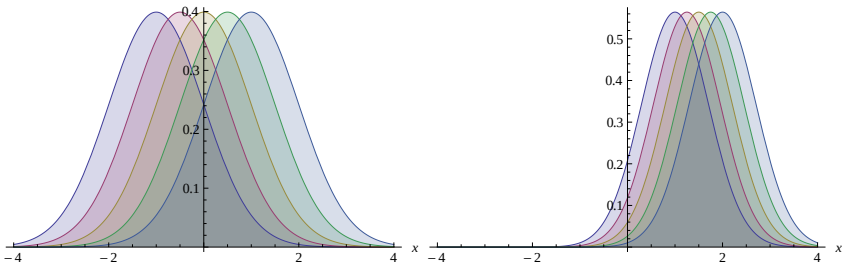

Figure 3: Set of priors (left) and posterior (right) for $m \in \{-1, -0.5, 0, 0.5, 1\}$.

with $a < b$. We can model the information about the value m of the mean M by a restricted vacuous Belief function on $[a,b]$:

$$\underline{E}_m[f] = \inf_{m\in[a,b]} f(m), \tag{37}$$

for any bounded real-valued function f of M. By combining the Normal prior with the Belief function (37) using marginal extension we can define the following lower expectation on X:

$$\underline{E}_X[g] = \underline{E}_M\left[\underline{E}_X\left[g|M\right]\right] = \inf_{m\in[a,b]} \int g(x)N(x;m,\sigma^2)dx, \tag{38}$$

for any bounded real-valued function g of X.[7] For instance in case $g = X$ we obtain

$$\underline{E}_X[X] = \inf_{m\in[a,b]} \int xN(x;m,\sigma^2)dx = \inf_{m\in[a,b]} m = a, \tag{39}$$

and $\overline{E}_X[X] = -\underline{E}_X[-X] = b$, which are respectively the lower and upper mean of X. Figure 3 shows the set of priors in (39) for $[a,b] = [-1,1]$, $\sigma = 1$ and the resulting set of posteriors computed w.r.t the likelihood $N(y;x,\sigma_y^2)$ with $y = 3$ and $\sigma_y^2 = 1$. We can include more information in the above model by considering instead of (37) the following linear-vacuous mixture:

$$\underline{E}_m[f] = (1-\varepsilon)\int_a^b f(m)TN(m;m_0,\sigma_0^2)dm + \varepsilon \inf_{m\in[a,b]} f(m), \tag{40}$$

where $TN(m;m_0,\sigma_0^2)$ denotes a truncated Normal in $[a,b]$ with $m_0 \in [a,b]$. In this case, we are assuming that some probabilistic information about the mean m is

[7] We have already seen that the lower probability induced by (38) is not a belief function. This shows that marginal extension does not preserve the monotonicity of (37).

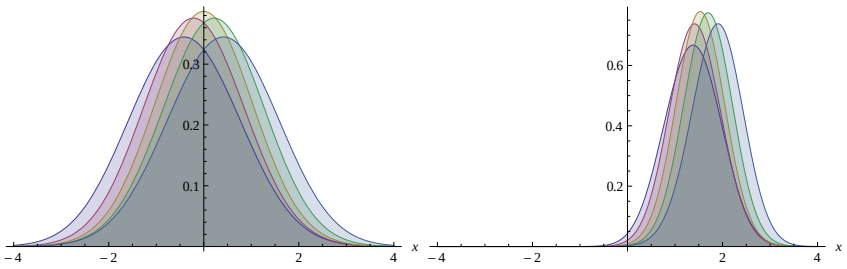

Figure 4: Set of priors (left) and posterior (right) for $m \in \{-1,-0.5,0,0.5,1\}$ and $\varepsilon = 0.4$.

available (the truncated Normal) but we are not completely certain ($\varepsilon > 0$) about it and, thus, we still allow the possibility that m can assume any value in $[a,b]$. By combining the Normal prior $N(x;m,\sigma^2)$ with the Belief function (40) using marginal extension we can define the following lower expectation on X:

$$\begin{aligned} \underline{E}_X[g] = \underline{E}_M\left[\underline{E}_X\left[g|M\right]\right] &= (1-\varepsilon)\int_a^b g(x)N(x;m,\sigma^2)TN(m;m_0,\sigma_0^2)dmdx \\ &+ \varepsilon \inf_{m\in[a,b]} \int g(x)N(x;m,\sigma^2)dx. \end{aligned} \tag{41}$$

Assuming that $[a,b]$ is larger than the 3σ credible interval of the Normal $N(m;m_0,\sigma_0^2)$, we can approximate $TN(m;m_0,\sigma_0^2)$ with $N(m;m_0,\sigma_0^2)$. Then by exploiting the following result:

$$\int\int g(x)N(x;m,\sigma^2)N(m;m_0,\sigma_0^2)dxdm = \int g(x)N(x;m_0,\sigma^2+\sigma_0^2),$$

and assuming that $g = X$, one obtains that

$$\begin{aligned} \underline{E}_X[X] &= (1-\varepsilon)\int xN(x;m_0,\sigma^2+\sigma_0^2)dx \\ &+ \varepsilon \inf_{m\in[a,b]} \int xN(x;m,\sigma^2)dx = (1-\varepsilon)m_0 + \varepsilon a. \end{aligned} \tag{42}$$

and similarly $\overline{E}_X[X] = (1-\varepsilon)m_0 + \varepsilon b$. Figure 4 shows the set of priors in (41) for $[a,b] = [-1,1]$, $\varepsilon = 0.4$, $m_0 = 1$, $\sigma = 1$ and $\sigma_0 = 1/3$ and the resulting set of posteriors computed w.r.t the likelihood $N(y;x,\sigma_y^2)$ with $y = 3$ and $\sigma_y^2 = 1$ by using Theorem 2. Observe that in this case the set of posteriors is more concentrated around the posterior $N(x,1.5,0.5)$ that we would obtain from the prior $N(x,0,1)$ and the likelihood $N(y;x,\sigma_y^2)$ using Bayes' rule. Threfore, the additional information about m reduces the posterior imprecision w.r.t the mean.

4.4. Incomplete observations

The term incomplete observations is used in literature to refer to set-valued observations. It describes a situation where we want to measure the value of a certain variable Y, but for some reason we can only determine it in an imperfect manner: we perform some kind of measurement whose outcome is W, but this does not allow us to completely determine the value of Y. Let us consider some examples.

1. Suppose we want to measure the voltage Y across a resistor, but the readout W of our digital voltage meter rounds this voltage to the next millivolt (mV). So if, say, we read that $W = 12mV$, we only know that the voltage Y belongs to the interval $(11mV, 12mV]$ [29].
2. Suppose that in the sequence of coin tosses H, H, T, H, T, T the fifth toss is missing, i.e., $H, H, T, H, ?, T$. We only know that the fifth observation Y belongs to the possibility space $\{H, T\}$.

The second example represents a particular kind of incomplete observation called missing process. This is an important case because the problem of missing data is ubiquitous in statistics.

It can be noticed that an incomplete observation mechanism can be described by a multivalued function

$$\Gamma(y) = \mathscr{W}',$$

where $\mathscr{W}'$ is a subset of the observation space $\mathscr{W}$ [29, 30]. Thus, it can be modelled by a belief function:

$$\underline{E}_W[h|x] = \int_{\mathscr{Y}} dP(y|x) \inf_{w \in \Gamma(y|x)} h(w),$$

for any bounded real-valued function h of W and value of the conditional variable x. It is often assumed that $\Gamma(y|x) = \Gamma(y)$, i.e., it does not depend on x but only on the value of y.

A formalisation of the incomplete observation mechanism was first derived in 1985 by Shafer [31]. Shafer showed that the right way to update probabilities with incomplete observations requires knowledge of the incompleteness mechanism, i.e., the mechanism that is responsible for turning a complete observation into an incomplete one. Shafers result states that neglecting the incompleteness mechanism can lead to unreliable conclusions. To overcome to this issue, first de Cooman and Zaffalon [29] and then Zaffalon and Miranda [30] proposed a conservative inference rule to deal with incomplete observations. The basic idea is

to consider the most conservative assumption about the incompleteness mechanism and, thus, to derive lower and upper bounds for the inferences based on this assumption. We point the reader to [29, 30] for a deeper discussion about conservative inferences with incomplete observations. In the next section, we show the main ideas with a simple example.

4.4.1. The coin example

Let X denote the probability of obtaining head in a coin toss. Let Y_i denote the outcome of the i-th coin toss ($Y_i = 1$ means head and $Y_i = 0$ tail). We assume that we cannot observe directly Y_i but we can read the record W_i of the coin toss reported by a witness. The problem is that some records of the witness' are missing.

We assume that:

- our prior information about X is described by the Beta distribution

$$p(x) = Beta(x; \alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1},$$

  for some $\alpha, \beta > 0$;

- the observations Y_i are i.i.d. with distribution:

$$p(Y_i = y_i|x) = x^{y_i}(1-x)^{1-y_i},$$

  with $y_i \in \{0, 1\}$.

Given for instance the following sequence of observations

$$\tilde{w}^n = \{1, 1, 0, 1, ?, 0\},$$

our goal is to draw inferences about X.

The likelihood model of the observation $W_i = w_i$ given x is:

$$p(W_i = w_i|x) = \sum_{y_i \in \{0,1\}} p(w_i|y_i)x^{y_i}(1-x)^{1-y_i} = p(w_i|1)x + p(w_i|0)(1-x). \quad (43)$$

Since the incomplete observation mechanism is a missing process, we can assume that $p(W_i = 0|1) = 0$ and $p(W_i = 1|0) = 0$. Then, since $\mathscr{W}_i = \{1, 0, ?\}$, from (43) it results that

$$\begin{array}{rcl} p(W_i = 1|x) & = & p(W_i = 1|1)x, \\ p(W_i = 0|x) & = & p(W_i = 0|0)(1-x), \\ p(W_i = ?|x) & = & p(W_i = ?|1)x + p(W_i = ?|0)(1-x). \end{array} \quad (44)$$

If we assume that the missing of an observation does not depend on the outcome of the coin toss (this assumption is called *missingness at random*), i.e., $p(W_i = ?|1) = (W_i = ?|0) = \rho$, then:

$$p(W_i = ?|x) = \rho.$$

In this case, it does not matter the values of $p(W_i = 1|1) > 0$, $p(W_i = 0|0) > 0$ and $\rho > 0$, by applying Bayes' rule to (43) and to the Beta prior, we obtain:

$$p(x|\tilde{w}^n) = Beta(x; n_1 + \alpha, n_0 + \beta), \tag{45}$$

where n_1 is the number of ones in $\tilde{w}^n$ and n_0 is the number of zeros in $\tilde{w}^n$. Hence, the posterior expectation of X is:

$$E[X|\tilde{w}^n] = \frac{n_1 + \alpha}{n_1 + n_0 + \alpha + \beta}. \tag{46}$$

That is, we can simply neglect the missing observations and use the remaining ones to derive inferences about X.

In many practical cases the missing at random assumption is not justified because we do not know if the missing mechanism depends or not on the outcome of the coin toss, i.e., we only know that:

$$p(W_i = ?|1) + p(W_i = 1|1) = 1, \;\; p(W_i = ?|0) + p(W_i = 0|0) = 1.$$

These two equalities define two closed and convex conditional sets of probabilities

$$\mathscr{P}_{W_i|Y_i=1} = \{p : \;\; p(W_i = ?|1) + p(W_i = 1|1) = 1\},$$

$$\mathscr{P}_{W_i|Y_i=0} = \{p : \;\; p(W_i = ?|0) + p(W_i = 0|0) = 1\}.$$

We can then compute lower and upper bounds for $E[X|\tilde{w}^n]$ by considering the extreme distributions in the above sets. These distributions can be obtained by considering the two extreme cases: (i) $p(W_i = ?|1) = 0$ and $p(W_i = ?|0) = 1$ and (ii) $p(W_i = ?|1) = 1$ and $p(W_i = ?|0) = 0$, which gives the lower and, respectively, upper posterior expectation of X:

$$\underline{E}_X[X|\tilde{w}^n] = \frac{n_1 + \alpha}{n + \alpha + \beta}, \;\; \overline{E}[X|\tilde{w}^n] = \frac{n_1 + n_? + \alpha}{n + \alpha + \beta}. \tag{47}$$

The lower posterior expectation is obtained by replacing all question marks with 0, while the upper posterior expectation by replacing all question marks with 1. In

other words, we consider the two extreme cases: (i) all the question marks were 0, (ii) all the question marks were 1. This is the least committal approach in case we do not know anything about the missing process.[8] This result is an application of the the Conservative Inference Rule presented in [30, Sec. 4].

The inferences (47) can equivalently be derived by exploiting the results in Corollary 1 in the special case in which $\underline{E}_X = \overline{E}_X = E_X$, where E_X is the expectation w.r.t. the prior Beta density.

Note in fact that the missing observation mechanism can be described by the following multivalued map:

$$\Gamma(Y_i|x) = \{Y_i, ?\},$$

for all $x \in (0,1)$. This means that the observation $Y_i \in \{0,1\}$ either is mapped into itself or it is missing and that this does not depend on the value of x. Hence, it results that:

$$\underline{E}_{W_i}[I_{\{\tilde{w}_i\}}|x] = \sum_{y_i \in \{0,1\}} x^{y_i}(1-x)^{1-y_i} \inf_{w_i \in \{y_i,?\}} I_{\{\tilde{w}_i\}}(w_i), \tag{48}$$

and

$$\overline{E}_{W_i}[I_{\{\tilde{w}_i\}}|x] = \sum_{y_i \in \{0,1\}} x^{y_i}(1-x)^{1-y_i} \sup_{w_i \in \{y_i,?\}} I_{\{\tilde{w}_i\}}(w_i), \tag{49}$$

These lower and upper expectations define bounds for $E_{W_i}[I_{\{\tilde{w}_i\}}|x]$. For instance, for $\tilde{w}_i = 0$, one has that

$$\underline{E}_{W_i}[I_{\{0\}}|0] = 0 \leq E_{W_i}[I_{\{0\}}|0] \leq 1 - x = \overline{E}_{W_i}[I_{\{0\}}|0].$$

Thus, the lower and upper posterior mean in (47) can equivalently be obtained by applying Corollary 1 to the case in which the prior expectation E_X is precise (i.e., it is the expectation w.r.t. the Beta density on X) and, the likelihood model is given by $\underline{E}_{W_i}[\cdot|X]$. Assuming that a sequence of n values $\tilde{w}_1, \ldots, \tilde{w}_n$ is observed, from (23) for $g(x) = x$ and (48)–(49) one gets

$$\int_0^1 Beta(x;\alpha,\beta) \left[\prod_{i=1}^{n} \sum_{y_i \in \{0,1\}} x^{y_i}(1-x)^{1-y_i} \inf_{w_i \in \{y_i,?\}} (x-\nu) I_{\{\tilde{w}_i\}}(w_i) \right] dx. \tag{50}$$

[8] Note that, we are also assuming that the missing probability may be not stationary, i.e., it can change from draw to draw.

For all the three possible values of $\tilde{w}_i$, i.e., $0, 1, ?$, we can parametrize the expectations included in the bounds (48)–(49) in the following way:

$$E_{W_i}[I_{\{1\}}|x] = \varepsilon_1 x, \quad E_{W_i}[I_{\{0\}}|x] = \varepsilon_2(1-x), \quad E_{W_i}[I_{\{?\}}|x] = \varepsilon_3 x + \varepsilon_4(1-x) \quad (51)$$

where $\varepsilon_i \in (0,1)$ and does not depend on x (because Γ does not depend on x).[9] In fact, from (48)–(49) it results that

$$\underline{E}_{W_i}[I_{\{0\}}|x] = 0, \;\; \overline{E}_{W_i}[I_{\{0\}}|x] = 1-x,$$

and, thus, $0 \leq E_{W_i}[I_{\{1\}}|x] \leq 1-x$, which can be rewritten as $E_{W_i}[I_{\{1\}}|x] = \varepsilon_2(1-x)$ for $\varepsilon_2 \in (0,1)$. Similar expressions can be derived for $E_{W_i}[I_{\{1\}}|x]$ and $E_{W_i}[I_{\{?\}}|x]$. Then, (50) can be rewritten as:

$$\begin{aligned} \int_0^1 Beta(x;\alpha,\beta) \quad & \left[\prod_{i=1}^n \inf_{\varepsilon_i \in (0,1)} (x-\nu) \Big(I_{\{\tilde{w}_i\}}(1)\varepsilon_1 x + I_{\{\tilde{w}_i\}}(0)\varepsilon_2(1-x) \right. \\ & \left. + I_{\{\tilde{w}_i\}}(?)(\varepsilon_3 x + \varepsilon_4(1-x)) \Big) \right] \end{aligned} \quad (52)$$

Observe that

$$I_{\{\tilde{w}_i\}}(1)\varepsilon_1 x + I_{\{\tilde{w}_i\}}(0)\varepsilon_2(1-x) + I_{\{\tilde{w}_i\}}(?)(\varepsilon_3 x + \varepsilon_4(1-x))$$

for $\tilde{w}_i = 0$, $\tilde{w}_i = 1$ and $\tilde{w}_i = ?$ is respectively equal to $\varepsilon_1 x$, $\varepsilon_2(1-x)$ and $\varepsilon_3 x + \varepsilon_4(1-x)$. We are looking for the supremum value of ν such that (52) is greater than or equal to zero. Such value of ν is obtained for $\varepsilon_3 = 0$, $\varepsilon_4 = 1$ and $\varepsilon_1, \varepsilon_2 > 0$ (it does not matter their value provided that it is positive). These are the values that give the posterior lower mean of X. In fact, by comparing (51) with (44), it follows that $\varepsilon_1 = p(W_i = 1|1)$, $\varepsilon_2 = p(W_i = 0|0)$, $\varepsilon_3 = p(W_i = ?|1)$ and $\varepsilon_4 = p(W_i = ?|1)$. Thus, the posterior lower and upper mean is again obtained by the two extreme cases: (i) $p(W_i = ?|1) = 0$ and $p(W_i = ?|0) = 1$ and (ii) $p(W_i = ?|1) = 1$ and $p(W_i = ?|0) = 0$. It does not matter the value of ε_1 and ε_2 provided that they are positive. This example shows that belief functions are fundamental to treat missing data when the missingness mechanism is unknown.

[9] Note that the fact that Γ does not depend on x, implies that W_i is also strongly independent of X given Y_i, see [30, Sec. 4].

5. Conclusions

We have shown that, although many useful models used in robust statistics cannot be represented by belief functions, the multivalued mapping mechanism that induces a belief function is a very useful tool to design robust model. Furthermore, belief functions give advantages from a computational point view. By using the multivalued mapping mechanism, we have derived several belief function models that we have used to derive robust statistical inferences by using Walley's theory of imprecise probabilities. We have also shown the connection of the proposed approach with interval estimation and statistical inference with missing data. As future work, we intend to apply this work to more practical estimation problems and to derive more closed convex sets of probability measures by using the multivalued mapping mechanism of belief functions.

Acknowledgements

This work has been partially supported by the Swiss NSF grant n. 200020-137680/1.

References

[1] A. P. Dempster, Upper and lower probabilities induced by a multiple-valued mapping, Ann. Math. Stat. 38 (1967) 325–339.

[2] G. Shafer, A mathematical theory of evidence, Princeton University Press, 1976.

[3] G. Shafer, Allocations of probability, Ann. Probability 7 (5) (1979) 827–839.

[4] J. Bernardo, A. Smith, Bayesian theory, John Wiley & Sons, 1994.

[5] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, New York, 1991.

[6] J. Berger, E. Moreno, L. Pericchi, M. Bayarri, Bernardo, et al., An overview of robust Bayesian analysis, Test 3 (1) (1994) 5–124.

[7] L. Wasserman, Invariance properties of density ratio priors, The Annals of Statistics 20 (4) (1992) 2177–2182.

[8] J. O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer Series in Statistics, New York, 1985.

[9] P. Huber, Robust estimation of a location parameter, The Annals of Mathematical Statistics (1964) 73–101.

[10] S. Sivaganesan, J. Berger, Ranges of posterior measures for priors with unimodal contaminations, The Annals of Statistics (1989) 868–889.

[11] L. DeRoberts, J. Hartigan, Bayesian inference using intervals of measures, The Annals of Statistics 9 (2) (1981) 235–244.

[12] L. Pericchi, P. Walley, Robust Bayesian credible intervals and prior ignorance, International Statistical Review (1991) 1–23.

[13] P. Walley, A bounded derivative model for prior ignorance about a real-valued parameter, Scandinavian Journal of Statistics 24 (4) (1997) 463–483.

[14] A. Benavoli, M. Zaffalon, A model of prior ignorance for inferences in the one-parameter exponential family, Journal of Statistical Planning and Inference 142 (7) (2012) 1960 – 1979.

[15] A. Piatti, M. Zaffalon, F. Trojani, M. Hutter, Limits of learning about a categorical latent variable under prior near-ignorance, Int. Journal of Approximate Reasoning 50 (4) (2009) 597–611.

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

[17] L. A. Wasserman, Prior envelopes based on belief functions, The Annals of Statistics 18 (1) (1990) pp. 454–464.

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

[19] A. Dempster, The dempstershafer calculus for statisticians, International Journal of Approximate Reasoning 48 (2) (2008) 365 – 377.

[20] D. E. Leaf, C. Liu, Inference about constrained parameters using the elastic belief method, International Journal of Approximate Reasoning 53 (5) (2012) 709 – 727.

[21] M. S. Balch, Mathematical foundations for a theory of confidence structures, International Journal of Approximate Reasoning 53 (7) (2012) 1003 – 1019.

[22] S. Moral, Calculating uncertainty intervals from conditional convex sets of probabilities, in: Proceedings of the Eighth international conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 1992, pp. 199–206.

[23] R. Hable, A minimum distance estimator in an imprecise probability model computational aspects and applications, International Journal of Approximate Reasoning 51 (9) (2010) 1114 – 1128.

[24] P. Walley, S. Moral, Upper probabilities based only on the likelihood function, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61 (4) (1999) 831–847.

[25] E. Miranda, I. Couso, P. Gil, Study of the probabilistic information of a random set, in: Proc. of the 3th Int. Symp. on Imprecise Probability: Theories and Applications, Madrid (Spain), Vol. 18, 2003, pp. 382–394.

[26] I. Molchanov, Theory of random sets, Springer Verlag, 2005.

[27] M. Troffaes, G. De Cooman, Extension of coherent lower previsions to unbounded random variables, Intelligent systems for information processing: from representation to applications (2002) 277–288.

[28] I. Couso, S. Moral, P. Walley, A survey of concepts of independence for imprecise probabilities, Risk Decision and Policy 5 (2) (2000) 165–181.

[29] G. De Cooman, M. Zaffalon, Updating beliefs with incomplete observations, Artificial Intelligence 159 (1) (2004) 75–125.

[30] M. Zaffalon, E. Miranda, Conservative inference rule for uncertain reasoning under incompleteness, Journal of Artificial Intelligence Research 34 (2) (2009) 757–821.

[31] G. Shafer, Conditional probability, International Statistical Review/Revue Internationale de Statistique (1985) 261–275.