Learning with Imprecise Probabilities as Model Selection and Averaging

Serafín Moral

Dpto. Ciencias de la Computación e Inteligencia Artificial
ETSI Informática, Universidad de Granada
18071, Granada, Spain
smc@decsai.ugr.es

Abstract

This paper presents a general framework for learning with imprecise probabilities, consisting of a hierarchical approach with two sets of parameters. In the top set we have imprecise information, and conditioned on this set we have precise Bayesian information about the other set of parameters. Given a set of observations, the information about both sets of parameters is updated by conditioning, and a model selection method is applied to compute a reduced top set. This model selection method is based on decisions with imprecise probabilities. It will be shown that many existing approaches can be fitted in this general procedure, and a theoretical justification will be provided. Finally, the method will be applied to the problem of learning credal networks.

Keywords: imprecise probability, decision making, learning, model selection, probability estimation, likelihood, Bayesian networks, credal networks.

1. Introduction

This paper proposes a hierarchical approach to learning. It considers that learning is equivalent to selecting the parameters for a given model. These parameters should be understood in a wide sense. For example, if we are considering classification trees, the parameters can represent the tree structure too. So, this learning also implies model selection [1, 2].

The procedure used in this paper follows the basic scheme of Gärdenfors and Shalin [3]. The parameters are classified in two sets: Θ and B. Θ is the

top set and conditioned to each $\theta \in \Theta$ we have a precise Bayesian model for the parameters in B and the variables of interest, in order to follow a model averaging approach (by averaging with respect to the posterior probability in B given the observations) [1]. At the same time, a set of observations defines a likelihood in top set Θ. Different methods of using this likelihood have been proposed in the literature [4, 5, 3], but in general they do not provide a sound and well-founded integration model. The idea in this paper is essentially similar to the α-cut conditioning by Cattaneo [5], but it gives a justification based on decision making with imprecise probabilities. The final procedure proposed here is quite flexible and can accommodate very different available procedures, such as maximum likelihood, likelihood intervals [6], imprecise probability methods such as the imprecise Dirichlet model [7], and others. It will be also the basis for proposing new imprecise probabilities methods for learning credal networks [8]. There are also relations with Bayesian high-density regions [9, 10] that can be justified within our framework.

An important fact about the procedure is that the prior information on Θ is given by a coherent set of desirable gambles [11, 12, 13]. This is a quite general model to represent imprecise probabilistic information with a behavioural interpretation. In our case, the use of desirable gambles is essential in the case of an infinite Θ, since there will not be an equivalent representation as a set of probability measures or credal set. The model will be based on the discounting of uniform information on Θ. This uniform information will be a generalization of the uniform probability on a finite Θ but quite different from the usual uniform density on a continuous Θ. The proposed discounting is a generalization of the concept of discounting a belief function [14] or the ϵ-contaminated robust models [15], but here a behavioural interpretation based on sets of desirable gambles is provided following the ideas presented in [16].

This paper is organized as follows: Section 2 provides the basic framework and introduces the main concepts of coherent sets of desirable gambles; Section 3 introduces the discounted uniform model, which will be the prior information on Θ; Section 4 provides examples of the general procedure used to estimate multinomial probabilities, including cases such as maximum likelihood, maximum likelihood intervals, and the imprecise Dirichlet model, among others; Section 5 proposes a modification of the prior model for cases in which Θ is partitioned into sets of different dimensions to provide the basis for selecting among models of different complexity; Section 6 applies these ideas to propose methods for learning credal networks; and finally Section 7

is devoted to the conclusions and future work.

2. The Basic Framework for Learning

Uncertain variables will be denoted by $X, Y, Z, \ldots$. A set of uncertain variables will be denoted in bold-face $\mathbf{X} = \{X, Y, U, \ldots\}$. The set of possible values of X is denoted by Ω_X, and analogously for sets of variables.

To learn an imprecise probabilities model, assume that we have a parameter space $\Theta \times B$ and that for each $\theta \in \Theta$, there is a precise probabilistic model for β and a set of variables $\mathbf{X}$, i.e. we have precise probabilities $P(\beta|\theta)$ and $P(\mathbf{X}|\beta, \theta)$.

We are interested in learning a model for variables $\mathbf{X}$ given a dataset $\mathcal{O}$ of independent observations for variables $\mathbf{X}$. An element of $\mathcal{O}$ can be a full or partial observation of the variables in $\mathbf{X}$. Since conditioned to $\theta \in \Theta$ we have a precise Bayesian specification of B and $\mathbf{X}$, we can consider that learning with a fixed parameter $\theta \in \Theta$ is equivalent to conditioning the parameter β to the observations $\mathcal{O}$, i.e. computing $P(\beta|\theta, \mathcal{O})$, so that the model after the observations is[1]

$$P(\mathbf{X}|\mathcal{O}, \theta) = \int_B P(\beta|\theta, \mathcal{O}) P(\mathbf{X}|\beta, \theta) d\beta \tag{1}$$

For this expression, it has been assumed that $\mathcal{O}$ and $\mathbf{X}$ are conditionally independent given the parameters (θ, β).

In this way we follow a model averaging Bayesian approach for parameter β conditioned to $\theta \in \Theta$. Note that this does not imply that a single value or a region is selected from B. In fact, a pure Bayesian approach is followed, computing the posterior density on B given the observations and predicting future observations by averaging with respect to this posterior density (assuming that $\mathbf{X}$ is conditionally independent of $\mathcal{O}$ given β, θ).

But for parameter $\theta \in \Theta$ we will follow a model selection procedure [1, 2, 17]. Model selection is a decision problem that in our case will consist in selecting a parameter $\theta \in \Theta$: We will assume that we have a decision d_θ for each parameter $\theta \in \Theta$. To solve it, some prior information on Θ will be necessary. This prior information will be encoded using a coherent set of desirable gambles [11, 12, 13, 18]. A gamble on Θ is a bounded mapping

[1]When the parameter space B is finite, the integral is the summation on $\beta \in B$, but we will always keep the integral notation for the general case.

$f : \Theta \to \mathbb{R}$. A number $x \in \mathbb{R}$ will denote the gamble that is constantly equal to x for any $\theta \in \Theta$. If f is a gamble, the support of f, denoted by Support(f) is the set $\{\theta \in \Theta \ : \ f(\theta) \neq 0\}$.

Operations on gambles (sum, scalar multiplication) will mean a pointwise application of them. $f \leq g$ will mean that $f(\theta) \leq g(\theta), \forall \theta \in \Theta$, $f < g$ will denote $f \leq g$ and $f(\theta) < g(\theta)$ for at least one $\theta \in \Theta$. The set of all the gambles in Θ will be denoted by $\mathcal{L}$. A set $\mathcal{D} \subset \mathcal{L}$ is said to be a coherent set of desirable gambles if and only if it satisfies the following properties,

D1. $0 \notin \mathcal{D}$,

D2. if $f \in \mathcal{L}$ and $f > 0$ then $f \in \mathcal{D}$,

D3. if $f \in \mathcal{D}$ and $c \in \mathbb{R}$ with $c > 0$ then $cf \in \mathcal{D}$,

D4. if $f \in \mathcal{D}$ and $g \in \mathcal{D}$ then $f + g \in \mathcal{D}$.

Given an arbitrary set of gambles $\mathcal{A}$, then the natural extension of $\mathcal{A}$ will be the intersection of all the coherent sets of gambles containing $\mathcal{A}$, denoted by $\overline{\mathcal{A}}$. If there is no coherent set of gambles containing $\mathcal{A}$ then, the natural extension is $\mathcal{L}$. In other case, it is said that $\mathcal{A}$ avoids partial loss. If $\mathcal{K} \subseteq \mathcal{L}$, where posi($\mathcal{K}$) is the set of gambles

$$\text{posi}(\mathcal{K}) = \{\sum_{i=1}^{k} c_i f_i \ : \ c_i > 0, f_i \in \mathcal{K}, k \geq 1\},$$

then if $\mathcal{A}$ avoids partial loss, it can be proved that the natural extension of $\mathcal{A}$ is equal to posi($\mathcal{A} \cup \mathcal{L}^+$), where $\mathcal{L}^+$ is the set of gambles $f > 0$.

If we have a coherent set of desirable gambles $\mathcal{D}$, then the conditioning of $\mathcal{D}$ to a likelihood function $L : \Theta \to \mathbb{R}$, is the set of gambles $\mathcal{D}_L = \{f \ : \ f.L \in \mathcal{D}\}$, where $f.L$ stands for pointwise multiplication [12]. If A is a subset of Θ, then conditioning $\mathcal{D}$ to A, $\mathcal{D}_A$, means conditioning to the likelihood equal to the indicator function of A, I_A. This conditioning to a likelihood is the counterpart in terms of desirable gambles of the conditioning to a likelihood for lower previsions (see [19, Sec. 8.4]). For example, if you find that $aI_\theta - bI_{\theta'}$ desirable, and our observations induce a likelihood such that $L(\theta) = 1$ $L(\theta') = 0.5$, then our odds in favour of θ against θ' double, so we are ready to accept $aI_\theta - 2bI_{\theta'}$ as desirable (multiplying this gamble by the likelihood is unconditionally desirable).

Given a set of desirable gambles $\mathcal{D}$, we can associate with it a set of almost desirable gambles $\mathcal{D}^*$, which is given by:

$$\mathcal{D}^* = \{f \ : \ f + \alpha \in \mathcal{D}, \forall \alpha > 0\}. \tag{2}$$

A set of desirable gambles on Θ defines a convex set of finitely additive probability measures (a credal set) on Θ, namely

$$\mathcal{M}_{\mathcal{D}} = \{P \mid P(f) \geq 0, \forall f \in \mathcal{D}\}, \tag{3}$$

where $P(f)$ is the expectation of f with respect to P [19, Section 3.2].

Using the same expression, we can associate a credal set to $\mathcal{D}^*$, which is identical to the one associated with $\mathcal{D}$: $\mathcal{M}_{\mathcal{D}} = \mathcal{M}_{\mathcal{D}^*}$ [19]. There is a one-to-one correspondence between sets of almost desirable gambles and credal sets, in the sense that $\mathcal{D}^* = \{f \mid P(f) \geq 0, \forall P \in \mathcal{M}_{\mathcal{D}^*}\}$ [19].

We will assume that each decision d_θ is identified with a utility gamble denoted in the same way, in such a way that $d_\theta(\theta')$ denotes the utility of selecting $\theta \in \Theta$ when the true value of the parameter is θ'. Hereafter, a 0-1 utility will be considered: $d_\theta(\theta') = 1$ if $\theta = \theta'$ and 0, otherwise. Other utilities could be used, but this one does not assume a distance in Θ and will be enough to explain many of the model selection procedures used in practice.

In our full setting, each observation $\mathcal{O}$ defines a likelihood (the marginal likelihood with respect to $\beta \in B$) on $\mathcal{D}$ given by:

$$L(\theta) = \int_B P(\beta|\theta).P(\mathcal{O}|\beta, \theta)d\beta. \tag{4}$$

Finally, our full procedure to learning can be expressed with the following steps:

- Solve the decision problem in Θ: Since we have a set of observations, we must first condition $\mathcal{D}$ on these observations, and then compute the non-dominated decisions, i.e.
  1. Compute the conditional set of desirable gambles $\mathcal{D}_L$ for the likelihood associated to the observations, which is given by expression (4).
  2. Select the set of maximal (non-dominated) decisions on the parameter space Θ [20], i.e. the set

$$H_L = \{d_\theta \ : \ d_{\theta'} - d_\theta \notin \mathcal{D}_L, \forall \theta' \in \Theta\}. \tag{5}$$

- Finally, consider the set of models about $\mathbf{X}$ associated with these decisions,

$$\mathcal{M}_{\mathcal{O}} = \{P(\mathbf{X}|\mathcal{O}, \theta) \ : \ d_\theta \in H_L\}, \tag{6}$$

where $P(\mathbf{X}|\mathcal{O}, \theta)$ is given by expression (1).

The final result is a set of models, not a single model. Since the information about Θ is imprecise, the set of optimal decisions will be in most of the cases imprecise too.

3. The Discounted Uniform Prior

This section is focused on the prior information $\mathcal{D}$ on Θ. Several possibilities can be considered, but our approach is based on the uniform prior. This uniform prior, $\mathcal{D}_u$, is the natural extension of the set of gambles:

$$\mathcal{K}_u = \{I_\theta - \alpha I_{\theta'} \ : \ \theta, \theta' \in \Theta, \alpha < 1\} \tag{7}$$

where I_θ is the indicator function of θ. Note that I_θ is identical to d_θ, but here we have preferred to use a different notation, to distinguish it from decision d_θ.

Proposition 1. *If Θ is finite, $\mathcal{D}_u$ is equal to the set of gambles f such that $\sum_{\theta\in\Theta} f(\theta) > 0$.*

Proof. Any gamble f belonging to posi($\mathcal{K}_u$) can be expressed as

$$f = \sum_{i=1}^{k} c_i f_i, \tag{8}$$

where $f_i > 0$ or $f_i \in \mathcal{K}_u$ and $c_i > 0$. Since for any f_i, $\sum_{\theta\in\Theta} f_i(\theta) = r_i > 0$, it follows immediately that $\sum_{\theta\in\Theta} f(\theta) = \sum_{\theta\in\Theta} \sum_{i=1}^{k} c_i f_i(\theta) = \sum_{i=1}^{k} c_i \sum_{\theta\in\Theta} f_i(\theta) = \sum_{i=1}^{k} c_i r_i > 0$.

On the other hand, if f is such that $\sum_{\theta\in\Theta} f(\theta) > 0$, let us prove that it can be expressed as in equation (8) by induction in the number of elements of Support(f).

If Support(f) contains only one element $\{\theta\}$, then we have that $f = cI_\theta$, where the indicator $I_\theta > 0$, in which case it is in the natural extension of $\mathcal{K}_u$.

If the number of elements in Support(f) is greater than 1, then let us consider $\theta_0 \in$ Support(f) such that $|f(\theta_0)| = \min_{\theta \in \text{Support}(f)} |f(\theta)| = h$, and $r = \sum_{\theta \in \Theta} f(\theta) > 0$. Let us consider $\theta_1 \in$ Support(f) and $\theta_1 \neq \theta_0$. In this case we have

- If $f(\theta_0) < 0$, let us consider $f' = f - c(I_{\theta_1} - \alpha I_{\theta_0})$, where $c = r/2 - f(\theta_0) > 0$ and $\alpha = \frac{-f(\theta_0)}{r/2 - f(\theta_0)} < 1$.

  As $\sum_{\theta \in \Theta} c(I_{\theta_1}(\theta) - \alpha I_{\theta_0}(\theta)) = c(1-\alpha) = r/2$, we have that $\sum_{\theta \in \Theta} f'(\theta) = r - r/2 > 0$.

  Taking into account that f' is obtained from f by modifying its value in two points belonging to its support, then Support$(f') \subseteq$ Support(f), but $f'(\theta_0) = f(\theta_0) + c\alpha = 0$, so Support$(f')$ has less elements than Support(f), and we can apply the induction hypothesis to f'. As $f = f' + c(I_{\theta_1} - \alpha I_{\theta_0})$, and $I_{\theta_1} - \alpha I_{\theta_0} \in \mathcal{K}_u$, the result follows.

- If $f(\theta_0) > 0$, let us consider $f' = f - c(I_{\theta_0} - \alpha I_{\theta_1})$, were $c = 1/f(\theta_0) > 1$ and $\alpha = 1 - \frac{f(\theta_0) r}{2} < 1$. In this case too it follows immediately that Support$(f') =$ Support$(f) \setminus \{\theta_0\}$ and $\sum_{\theta \in \Theta} f'(\theta) = r - c(1-\alpha) = r/2 > 0$. So, the result follows by induction taking into account that $(I_{\theta_0} - \alpha I_{\theta_1}) \in \mathcal{K}_u$.

□

In this case (a finite Θ), set $\mathcal{M}_{\mathcal{D}_u}$ of finitely additive probability measures in Θ contains only one element: P_u, the uniform probability in Θ. If there are two values θ_1, θ_2 such that $P(\theta_1) > P(\theta_2)$, it is easy to select a gamble from $\mathcal{K}_u$, $f = I_{\theta_2} - \alpha I_{\theta_1}$ with $P(f) < 0$.

For the following result, Support$^+(f)$ will denote the set $\{\theta \in \Theta : \ f(\theta) > 0\}$ and Support$^-(f)$ will denote the set $\{\theta \in \Theta : \ f(\theta) < 0\}$.

Proposition 2. *If Θ is infinite, D_u is the set of gambles f for which* Support$^-(f)$ *is finite and there is $H \subseteq$* Support$^+(f)$ *with H finite and $\sum_{\theta \in (H \cup \text{Support}^-(f))} f(\theta) > 0$.*

Proof. The natural extension is posi$(\mathcal{K}_u \cup \mathcal{L}^+)$, as any element of $\mathcal{K}$ and $\mathcal{L}^+$ has finite negative support. Then, any finite positive linear combination of them will have a finite negative support too. Also, if $f \in$ posi$(\mathcal{K}_u \cup \mathcal{L}^+)$, then $f = \sum_{i=1}^k c_i f_i + \sum_{j=1}^l d_j g_l$, where $f_i \in \mathcal{K}_u$, $g_j \in \mathcal{L}^+$, $c_i > 0$, $d_j > 0$.

Let H be $\text{Support}^+(f) \cap (\bigcup_{i=1}^k \text{Support}(f_i))$. We have that H is finite, and $\sum_{\theta \in (H \cup \text{Support}^-(f))} f(\theta) \geq \sum_{\theta \in (H \cup \text{Support}^-(f))} \sum_{i=1}^k c_i f_i$. And the result follows taking into account that $\text{Support}(f_i) \subseteq (H \cup \text{Support}^-(f))$ and $\sum_{\theta \in \Theta} f_i = 1 - \alpha > 0$.

On the other hand, if f is a gamble for which $\text{Support}^-(f)$ is finite and there is $H \subseteq \text{Support}^+(f)$ with H finite and $\sum_{\theta \in (H \cup \text{Support}^-(f))} f(\theta) > 0$, we can consider the gamble $g = I_{(H \cup \text{Support}^-(f))} \cdot f$, where $I_{(H \cup \text{Support}^-(f))}$ is the indicator function of $(H \cup \text{Support}^-(f))$. This gamble has a finite support and so we can follow a reasoning completely analogous to the previous proposition to show that $g \in \mathcal{D}_u$, and since $f \geq g$, we have that $f \in \mathcal{D}_u$. □

The following result shows that moving from the finite to the infinite case changes the nature of $\mathcal{M}_{\mathcal{D}_u}$. In fact, now $\mathcal{M}_{\mathcal{D}_u}$ contains many probability measures.

Proposition 3. *If Θ is infinite, then $\mathcal{M}_{\mathcal{D}_u}$ is equal to the set of all the finitely additive probability measures in Θ such that $P(H) = 0$ for any $H \subseteq \Theta$ finite.*

Proof. If P satisfies $P(H) = 0$ for any finite set H, since every gamble f in $\mathcal{D}_u$ has a finite negative support, we have that $P(f) \geq 0$ and $P \in \mathcal{M}_{\mathcal{D}_u}$.

On the other hand, if P is such that $P(H) > 0$ for a finite set, then there is $\theta_0 \in \Theta$ with $P(\theta_0) > 0$. Since Θ is infinite, we cannot have $P(\theta) \geq P(\theta_0)$ for every $\theta \in \Theta$. Therefore, we can conclude that there is a $\theta_1 \in \Theta$ such that $P(\theta_1) < P(\theta_0)$.

Consider $\alpha = \frac{P(\theta_0) - P(\theta_1)}{2P(\theta_0)} < 1$, and the gamble $f = I_{\theta_1} - \alpha I_{\theta_0} \in \mathcal{D}_u$. We have that $P(f) = P(\theta_1) - \alpha P(\theta_0) = P(\theta_1) - (P(\theta_0) - P(\theta_1))/2 = (P(\theta_1) - P(\theta_0))/2 < 0$, and therefore $P \notin \mathcal{D}_u$. □

It is important to note that $\mathcal{M}_{\mathcal{D}_u}$ contains many finitely additive probability measures, in particular if Θ is the $[0, 1]$ interval, it contains all the finitely additive probability measures compatible with the uniform density, but also all the finitely additive probability measures compatible with any other continuous distribution in $[0, 1]$, as all of them assign probability 0 to finite sets. For example, it contains the probabilities associated with density $h(x) = 2x$. However, $\mathcal{D}_u$ contains more information than $\mathcal{M}_{\mathcal{D}_u}$, as sets of desirable gambles may include information conditional to sets of probability 0. In particular, if H is finite, the conditioning of $\mathcal{D}_u$ to H, D_{uH}, will contain all the gambles f such that $\sum_{\theta \in H} f(\theta) > 0$, and the associated credal set is given by the uniform probability in H.

Now, we define discounting for sets of desirable gambles following the approach in [16]. If $\mathcal{D}$ is a set of desirable gambles and $\epsilon \in [0,1]$, then the discounting of $\mathcal{D}$ by ϵ is the set of desirable gambles given by:

$$\mathcal{D}^\epsilon = \{f - \epsilon \inf(f) I_{\text{Support}(f)} \ : \ f \in \mathcal{D}\} \setminus \{0\}. \tag{9}$$

It is immediate to prove that if $\mathcal{D}$ is coherent, then $\mathcal{D}^\epsilon$ is coherent too. Also, we have that $\mathcal{D}^\epsilon \subseteq \mathcal{D}$, i.e. we are losing information in $\mathcal{D}$ when moving to $\mathcal{D}^\epsilon$. On the other hand, $\mathcal{D}^0 = \mathcal{D}$ and $\mathcal{D}^1$ is the vacuous set of gambles: it only contains the positive gambles.

The use of the support in the definition is important to keep conditional information in the discounting. If sets of almost desirable gambles are to be considered, the use of the support in the definition is not significant (in the sense of the following result).

Proposition 4. *If $\mathcal{D}$ is a coherent set of desirable gambles, then if $\mathcal{D}'^\epsilon = \{f - \epsilon \inf(f) \ : \ f \in \mathcal{D}\} \setminus \{0\}$, we have that $(\mathcal{D}^\epsilon)^* = (\mathcal{D}'^\epsilon)^*$.*

Proof. First, we are going to prove that $\mathcal{D}'^\epsilon \subseteq \mathcal{D}^\epsilon$. If $g \in \mathcal{D}'^\epsilon$ and $g \geq 0$, then $g \in \mathcal{D}^\epsilon$, as $\mathcal{D}^\epsilon$ is coherent. If $g \not\geq 0$, then $\inf(g) < 0$, and there is $f \in \mathcal{D}$ with $\inf(f) < 0$, such that $g = f - \epsilon \inf(f)$. As $f \in \mathcal{D}$, then $h = f - \epsilon \inf(f) I_{\text{Support}(f)} \in \mathcal{D}^\epsilon$. Since $\inf(f) < 0$, we have that $g \geq h$, and therefore $g \in \mathcal{D}^\epsilon$.

Having proved that $\mathcal{D}'^\epsilon \subseteq \mathcal{D}^\epsilon$, we also have that $(\mathcal{D}'^\epsilon)^* \subseteq (\mathcal{D}^\epsilon)^*$.

Consider now $f \in (\mathcal{D}^\epsilon)^*$. Again, if $f \geq 0$, then $f \in (\mathcal{D}'^\epsilon)^*$. In other case we can assume without loss of generality that $\inf(f) < -1$ (we could multiply by a positive constant, c, and make the proof with cf, and if $cf \in (\mathcal{D}'^\epsilon)^*$ we will also have $f \in (\mathcal{D}'^\epsilon)^*$). Let us consider the sequence $\{\alpha_n\}_{n \in \mathbb{N}}$ where $\alpha_n = 1/n$. We have that, for any n, $f + \alpha_n \in \mathcal{D}^\epsilon$, and there is $g_n \in \mathcal{D}$ such that

$$f + \alpha_n = g_n - \epsilon \inf(g_n) I_{\text{Support}(g_n)}. \tag{10}$$

Since $\inf(f) < -1$, we have $\inf(f + \alpha_n) < 0$, and therefore $\inf(g_n - \epsilon \inf(g_n) I_{\text{Support}(g_n)}) < 0$ obtaining

$$\inf(f) + \alpha_n = \inf(g_n)(1 - \epsilon),$$

and

$$\inf(g_n) = \frac{\inf(f) + \alpha_n}{1-\epsilon}.$$

So, $\{\inf(g_n)\}_{n\in\mathbb{N}}$ is a decreasing sequence converging to $\frac{\inf(f)}{1-\epsilon}$.

Let us denote by A_n, B_n the supports of $f+\alpha_n$ and g_n respectively, and by A_n^c, B_n^c their complementary sets. We have that $A_n^c = \{\theta \in \Theta \ : \ f(\theta) = \alpha_n\}$ and then $A_n^c \cap A_m^c = \emptyset$ if $n \neq m$. We also have that $B_n^c \subseteq A_n^c$. Therefore, $B_n^c \cap B_m^c = \emptyset$ if $n \neq m$.

Going back to equation (10) and subtracting $\epsilon \inf(g_n) I_{B_n^c}$ in both sides, we get:

$$f + \alpha_n - \epsilon \inf(g_n) I_{B_n^c} = g_n - \epsilon \inf(g_n) I_{\text{Support}(g_n)} - \epsilon \inf(g_n) I_{B_n^c} = g_n - \epsilon \inf(g_n).$$

As $g_n \in \mathcal{D}$, we get $f + \alpha_n - \epsilon \inf(g_n) I_{B_n^c} \in \mathcal{D}'^{\epsilon}$.

Now consider $m \in \mathbf{N}$ and the average of the gambles $f_n = f + \alpha_n - \epsilon \inf(g_n) I_{B_n^c}$, i.e. $t_m = (1/m) \sum_{n=1}^m f_n$. We have that $t_m \in \mathcal{D}'^{\epsilon}$ too, since this set is coherent.

As all sets B_n^c $(n = 1, \ldots, m)$ are disjoint and $\inf(g_n) < 0$, we have

$$t_m \leq f + (1/m) \sum_{n=1,\ldots,m} \alpha_n - (1/m)\epsilon \inf(g_n).$$

Taking into account that $\{\inf(g_n)\}_{n\in\mathbb{N}}$ is a decreasing sequence converging to $\frac{\inf(f)}{1-\epsilon}$, we also have

$$t_m \leq f + (1/m) \sum_{n=1}^{m} \alpha_n - (1/m)\epsilon \inf(f).$$

So, for any m, we have $f + (1/m) \sum_{n=1}^m \alpha_n - (1/m)\epsilon \inf(f) \in \mathcal{D}'^{\epsilon}$. As the sequences $\{(1/m) \sum_{n=1,\ldots,m} \alpha_n\}_{m\in\mathbb{N}}$ and $\{(1/m)\epsilon \inf(f)\}_{m\in\mathbb{N}}$, converge to 0, we have that, for any $\alpha > 0$, there is an m such that $f + (1/m) \sum_{n=1,\ldots,m} \alpha_n - (1/m)\epsilon \inf(f) \leq f + \alpha$, and then $f + \alpha \in \mathcal{D}'^{\epsilon}$, and $f \in (\mathcal{D}'^{\epsilon})^*$. □

This discounting is consistent with the one used in belief functions [14] or the ϵ-contaminated model [15], as proved by the following proposition.

Proposition 5. *If $\mathcal{D}$ is a coherent set of desirable gambles and $\epsilon \in [0, 1]$, then $\mathcal{M}_{\mathcal{D}^\epsilon} = (1-\epsilon)\mathcal{M}_{\mathcal{D}} + \epsilon\mathcal{M}_0 = \{(1-\epsilon)P + \epsilon Q \ : \ P \in M_{\mathcal{D}}, Q \in \mathcal{M}_0\}$, where $\mathcal{M}_0$ is the vacuous set of finitely additive probability measures in Θ, i.e. it contains all the finitely additive probability measures.*

Proof. If $\epsilon = 1$, the result is trivial as both sets are $\mathcal{M}_0$. Assume now $\epsilon < 1$.

Assume $P \in (1-\epsilon)\mathcal{M}_{\mathcal{D}} + \epsilon\mathcal{M}_0$, then $P = (1-\epsilon)P_1 + \epsilon P_2$ where $P_1 \in \mathcal{M}_{\mathcal{D}}$ and P_2 is an arbitrary finitely additive probability measure.

If $g \in \mathcal{D}^\epsilon$,then $g = f - \alpha(\inf(f))\text{Support}(f)$, where $f \in \mathcal{D}$.

As $f \in \mathcal{D}$ and $P_1 \in \mathcal{M}_{\mathcal{D}}$, we have that $P_1(f) \geq 0$.

In these conditions, $P(g) = P(f) - \epsilon(\inf(f))P(\text{Support}(f)) = (1-\epsilon)P_1(f) + \epsilon P_2(f) - \epsilon(\inf(f))P(\text{Support}(f))$.

As $P_2(f) \geq \inf(f)$, we have

$$P(g) \geq (1-\epsilon)P_1(f) + \epsilon\inf(f) - \epsilon(\inf(f))P(\text{Support}(f)) \geq (1-\epsilon)P_1(f) \geq 0.$$

As $P(g) \geq 0$ for any $g \in \mathcal{D}^\epsilon$, we have that $P \in \mathcal{M}_{\mathcal{D}^\epsilon}$.

For the other inclusion, given that a coherent set of desirable gambles and its associated set of almost desirable gambles define the same credal set and the result of Proposition 4, it is enough to prove that $\mathcal{M}_{(\mathcal{D}'^\epsilon)^*} \subseteq (1-\epsilon)\mathcal{M}_{\mathcal{D}} + \epsilon\mathcal{M}_0$.

Assume now $P \notin (1-\epsilon)\mathcal{M}_{\mathcal{D}} + \epsilon\mathcal{M}_0$. As this set is closed and convex, then as a consequence of the weak*-compactness theorem [19, Section 3.6.1] there is a gamble f such that $P(f) < 0$ and $P'(f) \geq 0$ for any $P' \in (1-\epsilon)\mathcal{M}_{\mathcal{D}} + \epsilon\mathcal{M}_0$.

This implies that

$$(1-\epsilon)P_1(f) + \epsilon P_2(f) \geq 0, \quad \forall P_1 \in \mathcal{M}_{\mathcal{D}}, P_2 \in \mathcal{M}_0.$$

Let $\lambda > 0$ be an arbitrary value. Since $\mathcal{M}_0$ contains all the probability measures, consider that P_2 is the probability that assigns probability 1.0 to the points $\theta \in \Theta$ in which $f(\theta) \leq \inf(f) + \lambda$. We get

$$(1-\epsilon)P_1(f) + \epsilon(\inf(f) + \lambda) \geq 0, \quad \forall P_1 \in \mathcal{M}_{\mathcal{D}}.$$

Since $\lambda > 0$ is arbitrary, we get

$$(1-\epsilon)P_1(f) + \epsilon\inf(f) \geq 0, \quad \forall P_1 \in \mathcal{M}_{\mathcal{D}}.$$

Given the duality between credal sets and sets of almost desirable gambles, we have that $g = (1-\epsilon)f + \epsilon\inf(f) \in \mathcal{D}^*$.

We also have that $\inf(g) = (1-\epsilon)\inf(f) + \epsilon\inf(f) = \inf(f)$.

For any $\alpha > 0$, let $\alpha' = \alpha/(1-\epsilon)$, then $g + \alpha' \in \mathcal{D}$, and therefore $g + \alpha' - \epsilon(\inf(g) + \alpha') \in \mathcal{D}'^\epsilon$, so $g - \epsilon\inf(g) + \alpha'(1-\epsilon) = g - \epsilon\inf(g) + \alpha \in \mathcal{D}'^\epsilon$, and therefore, $g - \epsilon\inf(g) \in (\mathcal{D}'^\epsilon)^*$. On the other hand, $P(g - \epsilon\inf(g)) = (1-\epsilon)P(f) + \epsilon\inf(f) - \epsilon\inf(g) = (1-\epsilon)P(f) < 0$. So, $P \notin \mathcal{M}_{(\mathcal{D}'^\epsilon)^*}$. □

Now we are going to analyse the gambles in $\mathcal{D}_u^\epsilon$, focusing on gambles with finite support.

Proposition 6. *If f is a gamble with finite support, then $f \in \mathcal{D}_u^\epsilon$ if and only if $(1-\epsilon)\sum_{\theta\in\text{Support}(f)} f(\theta) + \epsilon \inf(f)|\text{Support}(f)| > 0$.*

Proof. If $f > 0$, then $f \in \mathcal{D}_u^\epsilon$ and $(1-\epsilon)\sum_{\theta\in\text{Support}(f)} f(\theta) + \epsilon\inf(f) > 0$, so let us assume that there is a point θ with $f(\theta) < 0$.

$f \in \mathcal{D}_u^\epsilon$ then $f = g - \epsilon\inf(g) I_{\text{Support}(g)}$, where $g \in \mathcal{D}_u$. We have that $\text{Support}(g)$ must be finite. Otherwise $\text{Support}^-(g)$ should be infinite and this is not possible if $g \in \mathcal{D}_u$.

We also get $\inf(f) = \inf(g) - \epsilon\inf(g) = (1-\epsilon)\inf(g)$.

Now, $\sum_{\theta\in\text{Support}(f)} f(\theta) = \sum_{\theta\in\text{Support}(g)} g(\theta) - \epsilon|\text{Support}(g)|\inf(g)$ and $\sum_{\theta\in\text{Support}(g)} g(\theta) = \sum_{\theta\in\text{Support}(f)} f(\theta) + \epsilon|\text{Support}(g)|\inf(g) = \sum_{\theta\in\text{Support}(f)} f(\theta) + \epsilon|\text{Support}(g)|\inf(f)/(1-\epsilon) > 0$.

Therefore, $\sum_{\theta\in\text{Support}(f)} f(\theta) + \epsilon|\text{Support}(g)|\inf(f)/(1-\epsilon) > 0$. Since $\text{Support}(f) \subseteq \text{Support}(g)$ and $\inf(f) < 0$, we also have $\sum_{\theta\in\text{Support}(f)} f(\theta) + \epsilon|\text{Support}(f)|\inf(f)/(1-\epsilon) > 0$, and multiplying by $1-\epsilon$ we get $(1-\epsilon)\sum_{\theta\in\text{Support}(f)} f(\theta) + \epsilon\inf(f)|\text{Support}(f)| > 0$.

On the other hand, assume that $(1-\epsilon)\sum_{\theta\in\text{Support}(f)} f(\theta) + \epsilon\inf(f)|\text{Support}(f)| > 0$, and then consider

$$g = (1-\epsilon)f + \epsilon(\inf(f) - \alpha) I_{\text{Support}(f)},$$

where $\alpha > 0$ is a number small enough such that $(1-\epsilon)\sum_{\theta\in\text{Support}(f)} f(\theta) + \epsilon(\inf(f)-\alpha)|\text{Support}(f)| > 0$ and $f(\theta) + \epsilon(\inf(f)-\alpha) \neq 0$, for any $\theta \in \text{Support}(f)$ (this is always possible, as $\text{Support}(f)$ is finite).

We have that $\text{Support}(f) = \text{Support}(g)$ and $g \in \mathcal{D}_u$, so $g - \epsilon\inf(g) I_{\text{Support}(g)} \in \mathcal{D}_u^\epsilon$.

Since $\inf(g) = (1-\epsilon)\inf(f) + \epsilon(\inf(f)-\alpha) = \inf(f) - \epsilon\alpha$,

if $\theta \in \text{Support}(g)$, $g(\theta) - \epsilon\inf(g) = (1-\epsilon)f(\theta) + \epsilon(\inf(f)-\alpha) - \epsilon(\inf(f) - \epsilon\alpha) = (1-\epsilon)f(\theta) - \epsilon\alpha(1-\epsilon) \leq (1-\epsilon)f(\theta)$, so as $g - \epsilon\inf(g) I_{\text{Support}(g)} \in \mathcal{D}_u^\epsilon$ we have that $f \in \mathcal{D}_u^\epsilon$ too. □

To finish this section, let us consider the gambles f in $\mathcal{D}_u^\epsilon$ with support in two points θ_1, θ_2 and such that $f \not> 0$. Without loss of generality, we can assume that $f(\theta_1) > 0 > f(\theta_2)$. These assumptions are important, as they will determine which decisions are dominated in our setting. In these

conditions, taking into account that $\inf(f) = f(\theta_2)$ gamble f is desirable in $\mathcal{D}_u^\epsilon$ if and only if

$$(1-\epsilon)f(\theta_1) + (1+\epsilon)f(\theta_2) > 0.$$

Or, equivalently, if

$$\frac{-f(\theta_2)}{f(\theta_1)} < \frac{1-\epsilon}{1+\epsilon}.$$

If we have a set of observations $\mathcal{O}$, then we must transform $\mathcal{D}_u^\epsilon$ by conditioning using the associated likelihood L, thus producing $\mathcal{D}_{u,L}^\epsilon$. Then g with support in $\{\theta_1, \theta_2\}$ and $g(\theta_1) > 0 > g(\theta_2)$ is desirable in this set if and only if $gL \in \mathcal{D}_u^\epsilon$, i.e.

$$\frac{-g(\theta_2)L(\theta_2)}{g(\theta_1)L(\theta_1)} < \frac{1-\epsilon}{1+\epsilon}.$$

If $g(\theta_2) = -1$, $g(\theta_1) = 1$, this condition is

$$\frac{L(\theta_2)}{L(\theta_1)} < \frac{1-\epsilon}{1+\epsilon}. \tag{11}$$

4. Applications to Multinomial Probabilities Estimation

In this section we will assume that we have a single variable X with K possible values $\{x_1, x_2, \ldots, x_K\}$ for which we want to estimate $P(X = x_i) = \theta_i, (i = 1, \ldots, K)$. This is a simple setting, but most of the procedures shown here can be extended to other statistical problems such as estimating the parameters of a Gaussian variable or a regression model.

4.1. Maximum Likelihood Estimation

In this case, $\Theta = \{\theta = (\theta_1, \ldots, \theta_K) \mid \sum_{i=1}^K \theta_i = 1, \theta_i \geq 0\}$, and $B = \{\beta\}$ (a single value is equivalent to the non-existence of the set: there is no uncertainty and it does not have any effect in the observation process). If we consider the uniform set of desirable gambles in Θ without discounting, then d_θ is dominated if and only if there is a $d_{\theta'}$ such that

$$\frac{L(\theta)}{L(\theta')} < 1,$$

i.e. we obtain maximum likelihood estimation. If we observe a sample of size N for which n_i values correspond to the observation $[X = x_i]$, it is well known that the only non-dominated decision d_θ corresponds to the maximum likelihood estimation: $\hat{\theta}_i = n_i/N$, being the likelihood function:

$$L(\theta) = \theta_1^{n_1} \dots . \theta_K^{n_K}. \tag{12}$$

4.2. Maximum Likelihood Confidence Regions

In the above setting, if the uniform model is ϵ discounted, the non-dominated decisions θ are those vectors, such that

$$\frac{L(\theta)}{L(\hat{\theta})} \geq \frac{1-\epsilon}{1+\epsilon}, \tag{13}$$

where $\hat{\theta}$ is the maximum likelihood estimation.

This corresponds to the so-called likelihood-based confidence regions [6], consisting in selecting all the parameters θ for which $\frac{L(\theta)}{L(\hat{\theta})} > c$. This provides a framework for justifying the use of these intervals with an ϵ-discounted uniform model and $c = \frac{1-\epsilon}{1+\epsilon}$. It is important to notice, that in some cases, Expression (13) gives rise to regions that are not convex, as it will be shown with an example. In the case of a single parameter with a convex region, we obtain what is called a pure likelihood interval [6].

The result in this case is an imprecise model, since many decisions are not dominated, namely all the vectors $(\theta_1, \dots, \theta_K)$ such that

$$\theta_1^{n_1} \dots . \theta_K^{n_K} \geq \left(\frac{n_1}{N}\right)^{n_1} \dots . \left(\frac{n_K}{N}\right)^{n_K} \frac{1-\epsilon}{1+\epsilon}.$$

Maximum likelihood estimation can be too precise, especially with small samples: If $N = 1$ and we observe x_i, then we estimate the probability $\theta_i = P(x_i) = 1$. Maximum likelihood intervals can be used to solve this problem. For a sample of size 1, $P(x_i)$ will be in the interval $[\frac{1-\epsilon}{1+\epsilon}, 1]$ if the only observation has been x_i and in the interval $[0, 1 - \frac{1-\epsilon}{1+\epsilon}]$ for $P(X = x_j), j \neq i$.

4.3. Bayesian High-Density Regions

In the Bayesian setting there are also procedures for selecting a parameter region instead of doing a point estimation or a model averaging: the so-called high-density regions, or Bayesian interval estimation [9, 10]. To describe this in our framework we will consider a set Θ of parameters and $B = \{beta\}$

(equivalent to non-existence of the set). It will also be assumed that there is a prior density f on Θ. In these conditions, it is possible to compute a posterior density $f(\theta|\mathcal{O})$. Then a high-density region is computed by fixing a threshold γ and then selecting a region $H \subseteq \Theta$ (a measurable set) such that $P(H||\mathcal{O}) \geq \gamma$, being $H = \{\theta \,:\, f(\theta|\mathcal{O}) \geq d\}$ of minimum size (with respect to the Lebesgue measure). Usually, this method is presented as a procedure to summarize the posterior density. A justification of this methodology in terms of Bayesian decision theory is given in [10, Subset. 5.2.5], but the problem is specified in a different way: The set of decisions is the set of all measurable sets $H \subseteq \Theta$, instead of the set of parameters. Under suitable loss functions linking a true value of the parameter θ to a selected region H, the best decision is a high-density region.

If a prior uniform density is considered, this procedure is very similar to maximum likelihood confidence regions, since $f(\theta|\mathcal{O}) \propto L(\theta)$ in that case. However, in practice there is a difference: In likelihood regions a value c is fixed and then $d = cL(\hat{\theta})$. In high-density regions γ is selected and d is the maximum value such that $P(H||\mathcal{O}) \geq \gamma$. So, in both cases the confidence region will be of the form $H = \{\theta \,:\, f(\theta|\mathcal{O}) \geq d\}$, but the value d is computed in different ways from the defining parameter in both cases. Therefore, a 0.95 likelihood region will be in general different from a 0.95 high-density region: One is included in the other, depending of the values of d in both cases.

If the prior density f is not uniform, then a very similar procedure to the high-density regions can be obtained by considering that the prior information is given taking into account that

$$\mathcal{K}_f = \{f(\theta')I_\theta - \alpha f(\theta)I_{\theta'} \,:\, \theta, \theta' \in \Theta, \alpha < 1\}$$

In this situation, if observations give rise to likelihood L as in Equation (**??**), then it follows that the non-dominated decision corresponds to the maximum posterior point:

$$\hat{\theta}_f = \arg\max_{\theta \in \Theta} f(\theta)L(\theta).$$

If a previous discounting by ϵ is carried out for $\mathcal{K}_f$, then what we get is a high-density region:

$$H = \{\theta \in \Theta \mid f(\theta)L(\theta) \geq cf(\hat{\theta}_f)L(\hat{\theta}_f), \}$$

where $c = \frac{1-\epsilon}{1+\epsilon}$, as usual. Note that this is a high-density region, but the defining parameter is $\gamma = P(H|\mathcal{O})$, instead of ϵ or c.

Finally, let us remark that this approach makes sense when the posterior density is bounded.

4.4. Bayesian Point Estimation

In this case, $\Theta = \{s\}$ is a single hyperparameter (the equivalent sample size), and $B = \{\theta = (\theta_1, \ldots, \theta_K) \mid \sum_{i=1}^{K} \theta_i = 1, \theta_i \geq 0\}$ (the former Θ in maximum likelihood estimation).

The prior probability in B conditioned to s is associated to a symmetrical Dirichlet distribution:

$$P(\theta_1, \ldots, \theta_K|s) = \frac{(\Gamma(s/K))^K}{\Gamma(s)} \theta_1^{s/K} \ldots . \theta_1^{s/K},$$

where $\Gamma()$ is the Gamma function.

The probability of the observations is $P(X = x_i|\theta, s) = \theta_i$.

Now, there is no model selection, since Θ has only a single value, and we only have a Bayesian averaging of parameters in B, estimating the probabilities using Equation (1): $P(X = x_i|\mathcal{O}, s)$ by $\frac{n_i + s/K}{N+s}$.

Model averaging is appropriate in many situations, but there are cases in which likelihood intervals or general high-density regions can be more intuitive.

Example 1. *Consider that we are estimating the probability of a binary variable X with values in $\{0, 1\}$. Imagine that we have a set of 100 independent values of X ($X_1, \ldots, X_{100}$) which have been partially observed: We only know the difference between the number of 0s and the number of 1s, which in our particular case was 100, i.e. all the observations were of the same value. If $P(X = 0) = \theta$, then the likelihood function is:*

$$L(\theta) = P(X_1 = X_2 = \cdots = X_{100}|\theta) =$$
$$P(X_1 = X_2 = \cdots = X_{100} = 1|\theta) + P(X_1 = X_2 = \cdots = X_{100} = 0|\theta) =$$
$$\theta^{100} + (1-\theta)^{100}. \quad (14)$$

As this likelihood is symmetrical with respect to $\theta = 0.5$, if prior information is symmetrical too, the posterior will be also symmetrical, and the averaging is 0.5. However, given the observations we have a strong belief

that the probabilities will be extreme. In fact, with $\frac{1-\epsilon}{1+\epsilon} = 0.95$*, the likelihood confidence region is* $[0, 0.02951305] \cup [0.970487, 1.0]$*, showing that the parameter value should be close to 0 or 1.*

The averaging Bayesian approach can make sense if we want to predict the probability of a single future outcome of the variables, but often we are estimating a probability, which remains fixed for later use. This use can be very diverse and not known at this moment. For example, we could need to know whether two future independent observations of X *will be both 0. In that case, the use of Bayesian probabilities could lead to very bad decisions.*

4.5. The Imprecise Dirichlet Model (IDM)

The imprecise Dirichlet model (IDM) was introduced by Walley [7]. In our setting, it can be described by considering a fixed hyperparameter s (the equivalent sample size) and considering $\Theta = \{\alpha = (\alpha_1, \ldots, \alpha_k) \ : \ \sum_{i=1}^{K} \alpha_i = s, \alpha_i > 0\}$; $B = \{\theta = (\theta_1, \ldots, \theta_K) \mid \sum_{i=1}^{K} \theta_i = 1, \theta_i \geq 0\}$ as in the Bayesian procedure, and $P(X = x_i|\alpha, \theta) = \theta_i$. The prior information in Θ is the 1-discounted uniform, i.e. the vacuous set of desirable gambles. In this setting, there is no dominance of decisions in Θ and the model is equivalent to averaging with respect to all the parameters in this set. If we have a dataset of N independent observations for variable X, the result for $P(X = x_i|\alpha, \mathcal{O})$ is $\frac{n_i+\alpha_i}{N+s}$. As there is no dominated α, the result for $P(X = x_i|\mathcal{O})$ is the set of values corresponding to all the α values. Taking supremum and infimum in these values, we get $P(X = x_i|\mathcal{O}) \in [\frac{n_i}{n_i+s}, \frac{n_i+s}{n_i+s}]$.

4.6. The Empirical Bayes Approach

In the Bayesian approach we must fix the hyperparameter s. However, there is no unified method to select its value. The empirical Bayes approach [21] considers a set of possible values, for example, an interval $[s_1, s_2]$, and makes the estimation using the value $s \in [s_1, s_2]$ with maximum likelihood (marginal likelihood on $\beta \in B$ as in Equation (4)). In our setting, $\Theta = [s_1, s_2]$, $B = \{\theta = (\theta_1, \ldots, \theta_K) \mid \sum_{i=1}^{K} \theta_i = 1, \theta_i \geq 0\}$, with

$$P(\theta_1, \ldots, \theta_K|s) = \theta_1^{s/K} \ldots . \theta_1^{s/K},$$
$$P(X = x_i|\theta, s) = \theta_i$$

In Θ we consider the uniform model without discounting ($\epsilon = 0$). If we have a set of observations about X, then the likelihood in Θ is [22],

$$L(s) = \frac{\Gamma(s)}{\Gamma(N+s)} \prod_{i=1}^{K} \frac{\Gamma(n_i + s/k)}{\Gamma(s/K)}.$$

So, the only parameter in Θ that is not dominated is parameter $\hat{s}$, maximizing $L(s)$ and the final estimation $\frac{n_i+\hat{s}/K}{N+\hat{s}}$.

Of course, this model can be also applied by discounting the uniform distribution on $\Theta = [s_1, s_2]$. If $\epsilon = 1$, the vacuous set of desirable gambles in Θ is obtained, and the result is the imprecise sample size Dirichlet model (ISSDM) proposed in Masegosa, Moral [23]. In this model, the estimated interval for $P(X_i = x_i|\mathcal{O})$ is given by

$$[\min\{\frac{n_i + s_1/K}{N+s_2}, \frac{n_i + s_2/K}{N+s_1}\}, \max\{\frac{n_i + s_1/K}{N+s_1}, \frac{n_i + s_2/K}{N+s_2}\}].$$

But we could also consider an ϵ-discounted uniform, in which case the estimation of probabilities would be done by computing $\hat{s}$, the value of s with greatest likelihood, and then reducing Θ to $H_L = \{s \in \Theta \mid L(s) \geq \frac{1-\epsilon}{1+\epsilon}\}$. Finally, $P(X_i = x_i|\mathcal{O})$ is given by

$$[\min_{s \in H_L} \frac{n_i + s/K}{N+s}, \max_{s \in \Theta'} \frac{n_i + s/K}{N+s}]$$

.

We do not know a closed form for this interval, or even under which conditions H_L is convex, but in any case, an approximate computation based on considering a finite set of points $s \in [s_1, s_2]$ seems feasible.

4.7. The Imprecise Dirichlet Model with α-cut Conditioning

It is well known that IDM has a good behaviour when we have precise observations of variable X, but it has difficulties when we have indirect observations with some probability of error [24]. Let us assume that we have now two variables X, Y with values $\{0, 1\}$, where X is a binomial variable for which we want to learn about its probabilities, and we have a Θ and B as in the IDM. Y is a variable that is conditional independent on the parameters (θ, β), given the value of X, i.e. it only depends of the concrete value of X and represents indirect observations of the values of X, with $P(Y = 1|X = 1) = 1 - \lambda_1$, $P(Y = 0|X = 0) = 1 - \lambda_0$, where λ_0, λ_1 are small positive numbers. Imagine that we have $N = 100$ observations of variable Y

and that in all of them we have observed $Y = 1$. We should expect a low value for the probability $P(X = 0) = \theta_0$, but the IDM provides the vacuous interval for this probability [24]: $P(X = 0|\mathcal{O}) \in [0, 1]$.

In fact, given the probabilities $\theta_0, 1 - \theta_0$ for the two values of X, we have

$$L(\theta_0|\mathcal{O}) = P(\mathcal{O}|\theta_1, \theta_2) = (\theta_0\lambda_0 + (1 - \theta_0)(1 - \lambda_1))^{100}.$$

Given a couple of values $(\alpha_0, \alpha_1) \in \Theta$ with $\alpha_0 + \alpha_1 = s$, the posterior probability of θ_0 is proportional to:

$$P(\theta_0|\mathcal{O}) \propto \theta_0^{\alpha_0 - 1}(1 - \theta_0)^{\alpha_1 - 1}(\theta_0\lambda_0 + (1 - \theta_0)(1 - \lambda_1))^{100}$$

For a very small value of α_0, we have the product of two functions, $\theta_0^{\alpha_0 - 1}(1 - \theta_0)^{\alpha_1 - 1}$, which is concentrated in low values of θ_0 and $(\theta_0\lambda_0 + (1 - \theta_0)(1 - \lambda_1))^{100}$ which in turn is concentrated in values of θ_0 close to 1. However, the second term does not depend on α_0 and it is possible to make α_0 small enough to concentrate the product in low values of θ_0, with which the expected value of $P(\theta_0|\mathcal{O})$ can be made as close to 0 as desirable, so that the final interval for $P(\theta_0|\mathcal{O})$ is $[0, 1]$.

The induced likelihood in Θ is given by

$$L(\alpha_0) = \int_0^1 \theta_0^{\alpha_0 - 1}(1 - \theta_0)^{\alpha_1 - 1}(\theta_0\lambda_0 + (1 - \theta_0)(1 - \lambda_1))^{100} d\theta_0.$$

For low values of α_0, as the two functions we are multiplying are concentrated in different parts of $[0, 1]$, their product is small in all the intervals, and the likelihood is small. So, even if the infimum of the expected values of θ_0 is 0, this infimum is obtained for values α_0 with very low likelihood. But the problem is that, since in the IDM we have the vacuous information in Θ (1-discounted uniform model), the information provided by this likelihood is not being used.

To solve this problem, Cattaneo [5] proposes the α-cut conditioning consisting in selecting a threshold $c \in [0, 1]$ and discarding the values α_0' for which $L(\alpha_0') < c \max_{\alpha_0} L(\alpha_0)$. In this way, low values of α_0 should be discarded in our example and the interval for $P(X = 1|\mathcal{O})$ is not vacuous. This approach corresponds to our model with ϵ-discounted uniform and $c = \frac{1-\epsilon}{1+\epsilon}$.

In general, the ϵ-discounted uniform can be applied to the general IDM with direct observations. In that situation, we must compute $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_K)$ maximizing the marginal likelihood,

$$L(\alpha) = L(\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(s)}{\Gamma(N+s)} \prod_{i=1}^{K} \frac{\Gamma(n_i + \alpha_i)}{\Gamma(\alpha_i)}.$$

Then, Θ is reduced to the values $H_L = \{\alpha \ : \ L(\alpha) \geq L(\hat{\alpha})\frac{1-\epsilon}{1+\epsilon}\}$. Finally, the estimated probabilities are

$$\left[\min_{\alpha \in H_L} \frac{n_i + \alpha_i}{N+s}, \max_{\alpha \in H_L} \frac{n_i + \alpha_i}{N+s}\right].$$

We have not studied this computational problem, but it is intuitive to think that the maximum of likelihood should be obtained in a point α, where α_i/s is close (or equal) to the relative frequencies n_i/N, and that the points in H_L will be in a neighbourhood around this point. As the value of $\frac{n_i+\alpha_i}{N+s}$ is increasing in α_i, it should not be very difficult to compute the maximum and minimum, at least in an approximate way.

5. Partitioned Sets of Parameters

In some cases, Θ is not homogeneous and some alternative to the uniform discounted model can be more reasonable.

Example 2. *Imagine that we have two observed variables (X, Y) that represent the colours of two balls extracted from urns with 10 balls in two different colours, red (R) and white (W). Two situations are possible:*

- *X and Y are selected from different urns of unknown composition.*
- *X and Y are selected from the same urn (with replacement, so that the extractions are conditionally independent given the urn).*

In this setting, the set of parameters Θ can be decomposed in two parts:

- *$\Theta_1 = \{(D, r_{i1}, r_{j2}) \mid r_{i1}, r_{i2} \in \{0, 1, \ldots, 10\}\}$, representing the first situation, in which r_{i1}, r_{i2} are the number of red balls in urns 1 and 2 respectively.*
- *$\Theta_2 = \{(E, r_i) \mid r_i \in \{0, 1, \ldots, 10\}\}$, representing the case of one urn for the two extractions, being r_i the number of red balls.*

Θ_1 and Θ_2 do not have the same size (121 for Θ_1 and 11 for Θ_2). Furthermore, for each parameter $(E, r_i) \in \Theta_2$ there is a parameter $(D, r_{i1}, r_{j2}) \in \Theta_1$ with $L(E, r_i) = L(D, r_{i1}, r_{j2})$, considering $r_{i1} = r_{i2} = r_i$, as in this case the probabilities for the observations are the same. In this situation, (E, r_i) will not dominate (D, r_i, r_i) for any set of observations $\mathcal{O}$. So, at least from the point of view of the likelihood information, the set of parameters Θ_2 is included in Θ_1. However, if the maximum likelihood obtained in Θ_1 is similar to the maximum likelihood obtained in Θ_2, it could be reasonable to discard parameters in Θ_1 and select the case of the same urn, since it is simpler and explains the data reasonably well.

The situation described in the example above could be handled for $\Theta = \bigcup_{i=1}^{I} \Theta_i$, being Θ_i disjoint finite sets, by considering the following modification of the uniform model:

- Given Θ_i, a uniform model is considered inside Θ_i. This implies that all the gambles $aI_{\theta_{i1}} + bI_{\theta_{i2}}$ with $a + b > 0$ and $\theta_{i1}, \theta_{i2} \in \Theta_i$, are considered desirable.
- A gamble $aI_{\theta_i} + bI_{\theta_j}$ is desirable with $\theta_i \in \Theta_i, \theta_j \in \Theta_j$, when $a/|\Theta_i| + b/|\Theta_j| > 0$.

In this way, we consider that there is no preference for any Θ_i over Θ_j, i.e. no gamble $I_{\Theta_i} - I_{\Theta_j}$ should be desirable. In fact, we have that for any probability in the associated credal set $P(\Theta_i) = P(\Theta_j)$. In that situation, parameters in larger sets Θ_i get probabilities lower than parameters in the smaller set Θ_2. In the case of the example we have that $P(\theta_1)/121 = P(\theta_2)/11$, if $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$.

The final result is a hierarchical uniform model: The parts have globally the same associated probability and there is a uniform probability conditioned to each one of the parts. After the observations, in the example, we could have a parameter in Θ_2 that dominates all the parameters in Θ_1.

Consider now that $\Theta = \Theta_1 \cup \Theta_2$, and that each parameter in $\theta_i \in \Theta_i$ is a vector of M_i elementary parameters $(\theta_{i1}, \ldots, \theta_{iM_i})$ and each θ_{ij} is one element of a set of m values (the same for every component of every parameter and every set). The case in the example corresponds to this situation with $m = 11, M_1 = 2, M_2 = 1$. In these conditions, a gamble $I_{\theta_1} - I_{\theta_2}$ is desirable conditioned to likelihood $L(\theta)$ if and only if $L(\theta_1)I_{\theta_1} - I_{\theta_2}L(\theta_2)$ is desirable,

and this happens if $L(\theta_1)/|\Theta_1| > L(\theta_2)/|\Theta_2|$. Taking into account that $|\Theta_i| = m^{M_i}$, and taking logarithms, this condition is equivalent to:

$$\log(L(\theta_1)) - M_1 \log(m) > \log(L(\theta_2)) - M_2 \log(m). \tag{15}$$

Note the similarity with classical model selection procedures like AIC or BIC, having a part related to the logarithm of the likelihood (how the model explains the data) and other part that is a penalty for the complexity of the associated model and which is proportional to the number of parameters.

This approach cannot be directly generalized to the infinite case, as the expression $a/|\Theta_i| > b/|\Theta_j|$ makes only sense in the finite case. Furthermore, there is no single way of penalizing complexity in this framework, in the sense that there can be many different coherent methods for comparing gambles from two different subsets of parameters. But following the analysis in the finite case, it seems reasonable to assume a virtual value for $|\Theta_i|$, which is equal to r^{M_i} where M_i is the number of continuous parameters in Θ_i and r is a parameter. Finally, a gamble $I_{\theta_i} - I_{\theta_j}$ is desirable conditioned to likelihood $L(\theta)$, if and only if

$$\frac{L(\theta_i)}{r^{M_i}} > \frac{L(\theta_j)}{r^{M_j}}. \tag{16}$$

Note that this criterion is also valid for $i = j$, since in that case $r^{M_i} = r^{M_j}$. Different criteria are obtained depending on the selection of parameter r. More specifically, if $r = e$ (Euler's number), what we obtain is Akaike's information criterion [25]. Another possibility is $r = \sqrt{N}$, i.e. the number of virtual cases of a continuous parameter is the square root of the sample size. This can make sense, since the precision with which a parameter can be determined for a sample of size N is, in general, proportional to $\sqrt{N}$ (under very general regularity conditions, the error of the maximum likelihood estimation is asymptotically Gaussian, with standard deviation proportional to $1/\sqrt{N}$). Assuming that $r = \sqrt{N}$, what we get is equivalent to the Bayesian information criterion [26]. In any case, other selections for r are also possible.

If this partitioned uniform model is discounted, the result will be a set of models, in particular all θ_i computed as follows:

- Compute θ_i^* maximizing $L(\theta_i)/|\Theta_i|$, where $|\Theta_i|$ is the cardinality of Θ_i in the finite case, or the virtual cardinal r^{M_i} in the infinite case, being M_i the number of continuous parameters in Θ_i and r a parameter.

- Select all $\theta_j \in \Theta_j$, such that

$$L(\theta_j)/|\Theta_j| \geq \frac{1-\epsilon}{1+\epsilon} L(\theta_i^*)/|\Theta_i|. \tag{17}$$

5.1. Asymmetrical Discounting

The above procedure can be hard from a computational point of view, as in some situations there can be too many parameters satisfying Equation (17). A possible solution would be to favour the simpler models satisfying the equation. The basis for this can be obtained using the following set of desirable gambles based on asymmetrical discounting.

We will assume that $\Theta = \bigcup_{i=1}^{I} \Theta_i$. Then we will consider the following preorder relation in sets Θ_i: $\Theta_i \preceq \Theta_j$, if and only if for each $\theta_i \in \Theta_i$, there is $\theta_j \in \Theta_j$ such that for each possible set of observations $\mathcal{O}$ in the model defining likelihood function $L(\theta) = P(\mathcal{O}|\theta)$, we have that $L(\theta_i) = L(\theta_j)$. We say that $\Theta_i \prec \Theta_j$, when $\Theta_i \preceq \Theta_j$ and $\Theta_j \not\preceq \Theta_i$. This preorder depends on the full model, including the observation procedure. The intuitive idea is that $\Theta_i \preceq \Theta_j$ when Θ_i is simpler than Θ_j, since for any parameter in Θ_i there is another parameter in Θ_j giving always rise to the same likelihood (i.e., defining the same probability for the observations). In Example 2 we have that $\Theta_2 \preceq \Theta_1$, as for each parameter $(E, r_i) \in \Theta_2$ there is a parameter $(D, r_{i1}, r_{j2}) \in \Theta_1$ with $L(E, r_i) = L(D, r_{i1}, r_{j2})$.

So, the asymmetrical discounting considers the set of desirable gambles, which is the natural extension of gambles

$$\begin{aligned}\{a\frac{I_{\theta_i}}{|\Theta_i|} - b\frac{I_{\theta_j}}{|\Theta_j|} + b\epsilon I_{\{\theta_i,\theta_j\}} \ : \ \theta_i \in \Theta_i, \theta_j \in \Theta_j, \Theta_i \not\prec \Theta_j, a, b > 0\} \cup \\ \{a\frac{I_{\theta_i}}{|\Theta_i|} - b\frac{I_{\theta_j}}{|\Theta_j|} \ : \ \theta_i \in \Theta_i, \theta_j \in \Theta_j, \Theta_i \prec \Theta_j, a + b \geq 0\}.\end{aligned} \tag{18}$$

In this case, the gambles expressing the dominance of a simpler model by a larger model are not discounted. As a result, if we use, for example, BIC (with the corresponding virtual cardinality of Θ_i) then if a model $\theta_i \in \Theta_i$ has the same or greater BIC than another model $\theta_j \in \Theta_j$ and $\Theta_i \prec \Theta_j$, then θ_j will be dominated: Never consider models that are more complex than necessary to obtain the same fitting value. In this way, the number of non-dominated parameters is lower, as the set of desirable gambles is larger.

6. Applications to Learning Credal Networks

A Bayesian network [27] for a set of variables $\mathbf{X} = (X_1, \ldots, X_m)$ is a pair (G, Π) where G is a directed acyclic graph with a node for each variable X_i and Π is a list of conditional probability distributions $P(X_1|Pa_1), \ldots, P(X_m|Pa_m)$, one for each variable X_i, given its parents in G, Pa_i.

- The graph G encodes a set of independent relationships: Given its parents, each variable X_i is independent of its non-descendant variables.
- The Bayesian network encodes the joint probability distribution:

$$P(X_1, \ldots, X_m) = \prod_{i=1}^{m} P(X_i|Pa_i) \tag{19}$$

Let us call K_i the number of values of variable X_i. The number of possible values or configurations of parent variables Pa_i is equal to $R_i = \prod_{X_j \in Pa_i} K_j$. The conditional probability distribution $P(X_i|Pa_i)$ consists in a probability distribution of X_i for each configuration of the parents $Pa_i = pa_j$. The value $P(X_i = x_k|Pa_i = pa_j), j = 1, \ldots, R_i$, will be denoted by θ_{ijk}, $(i = 1, \ldots, m,\ j = 1, \ldots, R_i,\ k = 1, \ldots, K_i)$. The vector of all these values $(\theta_{ijk})_{ijk}$ will be denoted by θ_G, and it is the set of parameters associated to a graph G that are required to specify the conditional probability distributions. We add the subscript G to make explicit the fact that this vector depends on G. We will also denote as θ_{ij} the vector $(\theta_{ij1}, \ldots, \theta_{ijK_i})$, i.e. the probabilities associated with the distribution of probability of X_i given the jth configuration of its parents $Pa_i = pa_j$. Also, we will also denote by $\mathcal{G}$ the set of all possible directed acyclic graphs for variables $\mathbf{X}$ and by Θ_G the set of all possible parameter values θ_G for graph G.

Learning is the process of selecting a model (G, θ_G) given a set of observations $\mathcal{O}$ of the variables. We will consider that we have a set of full observations of the variables $\mathbf{X}$ and that n_{ijk} is the number of observations of $X_i = x_k$ and the jth configuration of the parents Pa_i and $n_{ij} = \sum_{k=1}^{K_i} n_{ijk}$.

In our setting, learning Bayesian networks can be considered in different ways:

- One possibility is considering Θ the set of all directed acyclic graphs G for variables $\mathbf{X} = (X_1, \ldots, X_m)$, $\mathcal{G}$, and B being the set of parameters

$\{\theta_G \in \Theta_G \mid G \in \mathcal{G}\}$. Given G, there is a probability density concentrated in θ_G: It is assumed that each θ_{ij} is an independent Dirichlet distribution $D(\alpha_1, \ldots, \alpha_{K_i})$ with parameters $\alpha_k = s/(R_i K_i)$, where s is a global hyperparameter (the global sample size).

With these conditions, in the uniform model without discounting we must select a graph G maximizing $L(G) = P(\mathcal{O}|G)$. This value is the well-known BDEu score given by [22]:

$$BDEu(G) = P(\mathcal{O}|G) = \prod_{i=1}^{m} \prod_{j=1}^{R_i} \frac{\Gamma(s/R_i)}{\Gamma(n_{ij} + s/R_i)} \prod_{k=1}^{K_i} \frac{\Gamma(n_{ijk} + s/(R_i K_i))}{\Gamma(s/(R_i K_i))}. \tag{20}$$

As the parameters of the graph, θ_G, are in set B, the approach implies an averaging in these parameters. Taking into account that the parameters of each conditional probability in $P(X_i|Pa_i)$ are a Dirichlet distribution with $\alpha_k = s/(R_i K_i)$, then the estimated probabilities for $X_i = x_k$ given the jth configuration of its parents are

$$\hat{\theta}_{ijk} = \frac{n_{ijk} + s/(R_i K_i)}{n_{ij} + s/(R_i)}. \tag{21}$$

- Another possibility is to consider $\Theta = \{(G, \theta_G) \ : \ G \in \mathcal{G}, \theta_G \in \Theta_G\}$, which is partitioned as $\Theta = \bigcup_{G \in \mathcal{G}} \Theta'_G$, where $\Theta'_G = \{(G, \theta_G) \mid \theta_G \in \Theta_G\} = \{G\} \times \Theta_G$ and then the uniform partitioned model with different versions of the virtual $|\Theta'_G|$. When $|\Theta'_G|$ is equal to the exponential of the number of parameters: $e^{\sum_{i=1}^{m} R_i(K_i - 1)}$, taking logarithm and computing the non-dominated models is equivalent to computing the Bayesian network $(\hat{G}, \hat{\theta})$ maximizing the Akaike information criterion:

$$AIC(G, \theta_G) = \log(L(G, \theta)) = \sum_{i=1}^{m} \sum_{j=1}^{R_i} \sum_{k=1}^{K_i} n_{ijk} \log(\theta_{ijk}) - \sum_{i=1}^{m} R_i(K_i - 1) \tag{22}$$

This method implies that maximum likelihood estimation is employed to estimate the parameters, but in practice approaches are mixed and a Bayesian averaging solution is used. For example, considering Laplace

correction, i.e. estimating the probability of $X_i = x_k$ given the jth configuration of its parents as $\frac{n_{ijk}+1}{n_{ij}+K_i}$.

A credal network [8] is a generalization of a Bayesian network in which we have a graph with an imprecise set of probability distributions, all of which factorize according to the graph (they satisfy equation (19)), i.e. a graph G and a non-empty subset $H \subseteq \Theta_G$, instead of a single $\theta_G \in \Theta_G$. Masegosa and Moral [23] propose a generalization of this concept by allowing the graph to be imprecise too. So, a generalized credal network C is a non-empty set of pairs (G, θ_G), i.e. $C \subseteq \{(G, \theta_G) \ : \ G \in \mathcal{G}, \theta_G \in \Theta_G\}$.

Discounting the uniform prior information on Θ by ϵ implies that the result of learning is a set of graphs and parameters, i.e. a credal network. More specifically, we get:

- In the BDEu approach, all the pairs of $(G, \hat{\theta}_G)$ where $BDEu(G) \geq \frac{1-\epsilon}{1+\epsilon} \max_{G' \in \mathcal{G}} BDEu(G')$ and the parameters $\hat{\theta}_G$ for each graph are obtained by averaging as in Equation (21).

- In the Akaike information criterion we must compute the pair $(\hat{G}, \hat{\theta}_G)$ by maximizing it, and then all the pairs (G, θ_G) such that $AIC(G, \theta_G) \geq \log\left(\frac{1-\epsilon}{1+\epsilon}\right) + AIC(\hat{G}, \hat{\theta}_G)$.

In general, these are difficult computational problems, as the number of non-dominated networks can be very high. In this sense, we can cite the work of Liao et al. [28], proposing algorithms to compute all the Bayesian networks with a score within a giving factor of the optimal one.

But our approach offers other possibilities, like being initially imprecise in the global sample size s when using the BDEu approach. Usually, when learning a Bayesian network with this score, a somewhat arbitrary value of s is selected, but it is well known that this selection can have an important impact in the learned graph: Low values of s tend to produce sparse graphs, while with large values of s, dense graphs are obtained [29, 30]. In our approach, we can consider $\Theta = \{(G, s) \ : \ G \in \mathcal{G}, s \in [s_1, s_2]\}$, i.e. an interval $[s_1, s_2]$ is initially selected for the global sample size. B is the same as in the BDEu approach: $\{\theta_G \in \Theta_G \mid G \in \mathcal{G}\}$, and the rest of the model is also similar, with the difference that now the prior probability on B will depend on G and s, but it will be also a density concentrated in Θ_G according to which every parameter of every conditional probability θ_{ij} follow independent Dirichlet distributions with $\alpha_k = s/(R_i K_i)$. Now the BDEu score in Equation

(20) will be a function of G and s, $BDEu(G, s)$. The problem will be to determine the pair $(\hat{G}, \hat{s})$ maximizing $BDEu(G, s)$, and then all the pairs (G, s) with $BDEu(G, s) \geq \frac{1-\epsilon}{1+\epsilon} BDEu(\hat{G}, \hat{s})$. For a pair, (G, s) parameters are estimated by averaging (Equation (21)). IF $\epsilon = 0$ i.e. the uniform prior is not discounted, only $(\hat{G}, \hat{s})$ is selected and this is equivalent to using empirical Bayes to determine the global sample size.

7. Conclusions

This paper presents a general approach to learning with imprecise probabilities that combines Bayesian model averaging and model selection based on imprecise probabilities. It offers a theoretical foundation for many procedures presented in the literature and opens several computational problems to determine the set of non-dominated models using exact and approximate procedures. This will be an important task in our future research. Another pending task is extending the approach to consider other sets of decisions and loss functions in order to justify other methods for using the likelihood information as its transformation in a possibility measure [31] or computing conditional intervals [4]. Finally, as suggested by one of the reviewers, it would of interest to integrate our hierarchical model in a global decision problem in the set $\Theta \times B$, studying the loss functions that give rise to the same procedure than the approach proposed in the present paper.

Acknowledgements

We would like to thank the reviewers of this paper for their valuable and useful suggestions, which helped us to improve the quality of the article. This research was supported by the Spanish Ministry of Economy and Competitiveness under project TIN2016-77902-C3-2-P, and the European Regional Development Fund (FEDER).

References

[1] G. Claeskens, N. L. Hjort, et al., Model Selection and Model Averaging, Cambridge University Press, 2008.

[2] K. P. Burnham, D. R. Anderson, Multimodel inference: understanding aic and bic in model selection, Sociological methods & research 33 (2) (2004) 261–304.

[3] P. Gärdenfors, N.-E. Sahlin, Unreliable probabilities, risk taking, and decision making, Synthese 53 (3) (1982) 361–386.

[4] J. Cano, S. Moral, J. Verdegay-López, Combination of upper and lower probabilities, in: B. Ambrosio, P. Smets, P. Bonissone (Eds.), Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence, Morgan & Kaufmann, 1991, pp. 61–68.

[5] M. E. Cattaneo, A continuous updating rule for imprecise probabilities, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, 2014, pp. 426–435.

[6] Y. Pawitan, In All Likelihood: Statistical Modelling and Inference Using Likelihood, Oxford University Press, Oxford, 2001.

[7] P. Walley, Inferences from multinomial data: learning about a bag of marbles (with discussion), Journal of the Royal Statistical Society, Series B 58 (1996) 3–57.

[8] F. Cozman, Credal networks, Artificial Intelligence 120 (2000) 199–233.

[9] R. J. Hyndman, Computing and graphing highest density regions, The American Statistician 50 (2) (1996) 120–126.

[10] M. J. Schervish, Theory of statistics, Springer Science & Business Media, 2012.

[11] I. Couso, S. Moral, Sets of desirable gambles: conditioning, representation, and precise probabilities, International Journal of Approximate Reasoning 52 (7) (2011) 1034–1055.

[12] P. Walley, Towards a unified theory of imprecise probability, International Journal of Approximate Reasoning 24 (2000) 125–148.

[13] E. Quaeghebeur, Desirability, in: Introduction to Imprecise Probabilities, John Wiley & Sons, 2014, Ch. 1, pp. 1–27. doi:10.1002/9781118763117.ch1.

[14] G. Shafer, A Mathematical Theory of Evidence, Vol. 42, Princeton University Press, 1976.

[15] P. Huber, Robust Statistics, Wiley, New York, 1981.

[16] S. Moral, Discounting imprecise probabilities, in: E. Gil, E. Gil, J. Gil, M. Gil (Eds.), The Mathematics of the Uncertain, Springer, 2018, pp. 685–697.

[17] J. B. Kadane, N. A. Lazar, Methods and criteria for model selection, Journal of the American Statistical Association 99 (465) (2004) 279–290.

[18] J. De Bock, G. De Cooman, Conditioning, updating and lower probability zero, International Journal of Approximate Reasoning 67 (2015) 1–36.

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

[20] M. C. Troffaes, Decision making under uncertainty using imprecise probabilities., International Journal of Approximate Reasoning. 45 (1) (2007) 17–29.

[21] G. Casella, An introduction to empirical Bayes data analysis, The American Statistician 39 (2) (1985) 83–87.

[22] D. Heckerman, D. Geiger, D. M. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, Machine learning 20 (3) (1995) 197–243.

[23] A. R. Masegosa, S. Moral, Imprecise probability models for learning multinomial distributions from data. applications to learning credal networks, International Journal of Approximate Reasoning 55 (7) (2014) 1548–1569.

[24] A. Piatti, M. Zaffalon, F. Trojani, M. Hutter, Learning about a categorical latent variable under prior near-ignorance, in: G. de Cooman, J. Vejnarová, M. Zaffalon (Eds.), Proceedings of the Fifth International Symposium om Imprecise Probability: Theories and Applications (ISIPTA '07), 2007, pp. 357–364.

[25] H. Bozdogan, Model selection and Akaike's information criterion (aic): The general theory and its analytical extensions, Psychometrika 52 (3) (1987) 345–370.

[26] G. Schwarz, et al., Estimating the dimension of a model, The annals of statistics 6 (2) (1978) 461–464.

[27] J. Pearl, Probabilistic Reasoning with Intelligent Systems, Morgan & Kaufman, San Mateo, 1988.

[28] Z. A. Liao, C. Sharma, J. Cussens, P. van Beek, Finding all bayesian network structures within a factor of optimal, arXiv preprint arXiv:1811.05039.

[29] S. Moral, An empirical comparison of score measures for independence, in: Proceedings of the 10th IPMU international conference, 2004, pp. 1307–1314.

[30] T. Silander, P. Kontkanen, P. Myllymäki, On sensitivity of the map bayesian network structure to the equivalent sample size parameter, in: Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2007, pp. 360–367.

[31] D. Dubois, S. Moral, H. Prade, A semantics for possibility theory based on likelihoods, Journal of Mathematical analysis and applications 205 (2) (1997) 359–380.