Constructing Belief Functions from Sample Data Using Multinomial Confidence Regions

Thierry Denœux[1]

UMR CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529 - F-60205 Compiègne cedex - France

January 6, 2006

[1]Email: Thierry.Denoeux@hds.utc.fr. Fax: (33) 3 44 23 44 77.

Abstract

The Transferable Belief Model is a subjectivist model of uncertainty in which an agent's beliefs at a given time are modeled using the formalism of belief functions. Belief functions that enter the model are usually either elicited from experts, or must be constructed from observation data. There are, however, few simple and operational methods available for building belief functions from data. Such a method is proposed in this paper. More precisely, we tackle the problem of quantifying beliefs held by an agent about the realization of a discrete random variable X with unknown probability distribution $\mathbb{P}_X$, having observed a realization of an independent, identically distributed random sample with the same distribution. The solution is obtained using simultaneous confidence intervals for multinomial proportions, several of which have been proposed in the statistical literature. The proposed solution verifies two "reasonable" properties with respect to $\mathbb{P}_X$: it is less committed than $\mathbb{P}_X$ with some user-defined probability, and it converges towards $\mathbb{P}_X$ in probability as the size of the sample tends to infinity. A general formulation is given, and a useful approximation with a simple analytical expression is presented, in the important special case where the domain of X is ordered.

Keywords: Dempster-Shafer Theory, Evidence Theory, Transferable Belief Model, Multinomial Proportions, Confidence Intervals

1 Introduction

Since its foundation in the late 1960's and in the 1970's [6, 23, 29], the Dempster-Shafer theory of belief functions has been widely used as a conceptual framework for modeling partial knowledge and reasoning under uncertainty. Solving a real-world problem in this framework typically involves two steps: modeling each piece of information using a belief function on a suitable domain, and manipulating the resulting belief functions using such operations as marginalization, vacuous extension, and Dempster's rule of combination. Whereas many tools have been developed for the latter step (including, e.g., algorithms for propagation of evidence in belief function networks [26][27], and the Generalized Bayesian theorem (GBT) for inverting conditional beliefs [31]), modeling initial information using belief functions is still a challenge in many applications.

In practice, the two main sources of partial knowledge are human experts and observation data. Methods for the elicitation of belief functions from experts have been proposed by Wong and Lingras [41], Bryson and Mobolurin [4] and Dubois et al. [11], among others. In essence, these methods elicit weak information from the experts (such as preference relations [41], belief ratio intervals [4], or pignistic probabilities [32, 11]), and build a belief function that is consistent with this information.

In most applications, however, a significant part of the available information comes from statistical data, and it is crucial to be able to model such information in the belief functions framework. The first application of belief functions was indeed statistical inference about parametric models [6][7][8]. Shafer [24] describes several distinct approaches to this problem, among which the approach initially proposed by Dempster, based on pivotal quantities, the likelihood-based approach exposed in Shafer's book [23] (see also the discussions in [37] and [40]), and Smets' method based on the GBT. Principles of statistical inference within the theory of Hints, an interpretation of Dempster-Shafer theory closely related to Dempster's model, are exposed in [19, chapter 9].

The specific problem addressed in this paper is the following. We consider a population Ω, each element ω of which is described by a discrete observable characteristic $x \in \mathcal{X} = \{\xi_1, \ldots, \xi_K\}$. Individuals are randomly sampled from Ω according to some probability measure μ. The mapping $X : \omega \rightarrow x$ is thus a random variable, with

unknown probability distribution[1] $\mathbb{P}_X$ defined by $\mathbb{P}_X(A) = \mu(X^{-1}(A))$ for all $A \subseteq \mathcal{X}$. Having drawn n elements with replacement (or without replacement in the case of an infinite population), we have observed a realization $x_1, \ldots, x_n$ of an independent, identically distributed (iid) random sample $X_1, \ldots, X_n$ with parent distribution $\mathbb{P}_X$. We want to assess our degrees of belief concerning the realization of X that will be observed when we shortly draw an additional individual from Ω. A classical paradigm for this problem is that of an urn containing balls of different colors. Having observed the colors of n balls randomly taken from the urn with replacement, we want to assess our beliefs concerning the color of the next ball.

The conceptual framework adopted in this paper will be based on the Transferable Belief Model (TBM), a nonprobabilistic, subjectivist interpretation of the Dempster-Shafer theory of belief functions [35]. We shall thus assume the beliefs of a rational agent to be representable by a belief function, independently from any underlying probabilistic model (a major difference with Dempster's original model [30]). In particular, we shall not regard a belief function as the lower envelope of a family of probability distributions, a view that is known to be incompatible with Dempster's rule of combination and, consequently, with the TBM [25][30]. We shall, however, for the particular problem at hand, define a certain form of consistency between the belief function of interest and a lower probability measure, as will be shown below. Basic knowledge of the mathematics of belief functions and their interpretation in the TBM will be assumed in this paper. The reader is referred to Shafer's book [23] and to recent presentations of the TBM [33] for complete coverings of these topics.

The problem of inference from binomial and, more generally, multinomial data was originally addressed in the belief function framework by Dempster [6, 9, 8]. Dempster's solution was later recovered by Kohlas and Monney [19, page 261] in the Hint Theory framework (an interpretation of Dempster-Shafer theory close to Dempster's model), and in the TBM framework [32]. This solution will first be briefly recalled in Section 2, together with an alternative approach, the imprecise Dirichlet model, introduced by Walley [38] in the imprecise probability framework (Walley's solution happens to be a

[1] In this paper, the symbol $\mathbb{P}$ will be used to denote probability distributions of random variables and, more generally, probabilities interpreted as long-term frequencies of events in repeatable random experiments.

belief function). Our method will then be introduced in Section 3, and an approximate analytical solution for the case of ordered data will be presented in Section 4. Finally, Section 5 will conclude the paper.

2 Review of previous work

2.1 Dempster's approach

Binomial case

Belief functions were introduced by Dempster as part of a statistical inference framework proposed as an alternative to Bayesian methods and to Fisher's fiducial method [6]. One of the first applications of this new approach concerned binomial sampling with a continuous parameter p, and general multinomial sampling with a finite number of contemplated hypotheses [6]. The case of trinomial sampling was treated in a later paper [8], and some mathematical problems arising in the context of the general multinomial sampling model were studied in [10].

The main results concerning the binomial sampling model will first be summarized. Our presentation will be inspired from [1, chapter 9]. Let $X_1, \ldots, X_n$ be an iid sample with parent variable $X \in \mathcal{X} = \{0, 1\}$ following a Bernoulli distribution $\mathcal{B}(p)$ with parameter $p \in \mathcal{P} = [0, 1]$. A random variable W_i uniformly distributed on $\mathcal{W} = [0, 1]$ is supposed to underlie each observation X_i, with

$$X_i = 1 \Leftrightarrow W_i \leq p \ . \tag{1}$$

The uniform distribution of W_i can be thought of as modeling random sampling from an infinite population assimilated to the interval $[0, 1]$. Note also that this "trick" is commonly used to simulate binomial sampling using computer random number generators.

Equation (1) defines a multivalued mapping from $\mathcal{W}$ to $\mathcal{X} \times \mathcal{P}$, which maps any $w \in \mathcal{W}$ to $\{1\} \times [w, 1] \cup \{0\} \times [0, w]$. This mapping constrains the possible values of the triplet (X_i, W_i, p), and can alternatively be represented by a logical belief function (i.e., a belief function with a single focal set) m_i^{Θ} on the joint space $\Theta = \mathcal{X} \times \mathcal{W} \times \mathcal{P}$. Now, the uniform probability distribution of W_i defines a Bayesian belief function

$m_i^{\mathcal{W}}$. Belief functions m_i^{Θ} and $m_i^{\mathcal{W}}$, $i = 1, \ldots, n$ are the components of the graphical model shown in Figure 1. In this graph, variables are represented by circular nodes, belief functions on a single variable are represented by rectangular nodes, and belief functions on a product space are represented by diamond-shaped nodes; each belief function node is connected by an undirected edge to each variable node in its domain (see, e.g., [27]).

Having observed a realization x_i of each X_i, a belief function on $\mathcal{P}$ can be obtained by combining each belief function in the model using Dempster's rule, conditioning by $X_i = x_i$ for each $i = 1, \ldots, n$, and marginalizing on p. The result is a continuous belief function[2] on $\mathcal{P}$ with the following mass density function:

$$\begin{array}{rcll} m^{\mathcal{P}}(a,b) & = & \frac{n!}{(N-1)!(n-N-1)!} a^{N-1}(1-b)^{n-N-1} & 0 < N < n \\ m^{\mathcal{P}}(0,b) & = & n(1-b)^{n-1} & N = 0 \\ m^{\mathcal{P}}(a,1) & = & na^{n-1} & N = n, \end{array} \tag{2}$$

for all $a \leq b$, with $N = \sum_{i=1}^{n} x_i$.

The prediction problem can now be handled by defining a new variable X following a Bernoulli distribution $\mathcal{B}(p)$, with associated uniform random variable W. This defines the graphical model of Figure 2. The marginal bba induced about X may be shown [6] to be:

$$m^{\mathcal{X}}(\{1\}) = \frac{N}{n+1} = \frac{\widehat{p}}{1+1/n} \tag{3}$$

$$m^{\mathcal{X}}(\{0\}) = \frac{n-N}{n+1} = \frac{1-\widehat{p}}{1+1/n} \tag{4}$$

$$m^{\mathcal{X}}(\mathcal{X}) = \frac{1}{n+1}, \tag{5}$$

where $\widehat{p} = N/n$.

Multinomial case

The above approach can be extended to the general multinomial case as follows. Let us now assume that X is discrete variable with (unordered) values in $\mathcal{X} = \{\xi_1, \ldots, \xi_k\}$, and let $p_k = \mathbb{P}(X = \xi_k)$. We observe a realization of an iid sample $X_1, \ldots, X_n$ from X, and we want to make inference statements regarding the parameter vector $\mathbf{p} = (p_1, \ldots, p_K)$. In that case, the underlying population from which random sampling

[2] For a recent account of continuous belief functions, see [34].

takes place cannot be ordered as in the binomial case, because the modalities of X are no longer ordered: we have a "structure of the second kind" using the terminology introduced in [6]. The approach proposed by Dempster is then to assume uniform sampling from a $K-1$ dimensional simplex S_K. Using barycentric coordinates, the general point of such a simplex can be represented by a K-tuple of real numbers $(\alpha_1, \ldots, \alpha_K)$ where $\alpha_k \geq 0$ for $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \alpha_k = 1$. A random drawing of X may be created by drawing a random vector $\mathbf{W} = (W_1, \ldots, W_K)$ from a uniform distribution over S_K, and declaring that $X = \xi_k$ for some $k \in \{1, \ldots, K\}$ if

$$\frac{p_k}{W_k} \geq \frac{p_\ell}{W_\ell}, \quad \forall \ell \neq k.$$

Coming back to the iid sample $X_1, \ldots, X_n$, we can proceed exactly as above, and associate a random vector $\mathbf{W}_i$ to each X_i. Marginal belief functions about $\mathbf{p}$ and a new observation X may then theoretically be obtained as in the binomial case. However, the calculations are now much more complex. Dempster studied the trinomial case in [8] (without providing the equivalent of (3)-(5)), and he presented some results pertaining to the general case in [10]. However, the application of these results to compute the marginal belief function of X has proved, so far and to our knowledge, mathematically intractable.

Justification in the TBM

The introduction of the pivotal variables W_i may be argued to be artificial and somewhat arbitrary (see, e.g., the discussion in [8] and [1, page 252]). In [32], Smets attempted to solve the multinomial probability estimation problem by deducing the form of the belief function $m^{\mathcal{X}\times\mathcal{P}}$ on $\mathcal{X} \times \mathcal{P}$ from first principles, without resorting to pivotal quantities.

The main requirement imposed by Smets is the Hacking Frequency Principle [15], which equates the degree of belief of an event to its probability (long run frequency), when the latter is known. As shown in [32], this principle entails that the focal sets of $m^{\mathcal{X}\times\mathcal{P}}$ are of the form $\bigcup_{k=1}^{K} \{\xi_k\} \times A_k$, where $A_1, \ldots, A_K$ is a partition of $\mathcal{P}$.

In the binomial case, the focal sets of $m^{\mathcal{X}\times\mathcal{P}}$ are thus of the form $\{0\} \times A \cup \{0\} \times \overline{A}$ for some $A \subset [0, 1]$. Using a simple argument, Smets showed that, for any focal set, A

is of the form $[0, a)$ for some $a \in [0, 1]$. By applying once again the Hacking's principle, he then deduced directly the form of the belief density function $m^{\mathcal{P}}$ in (2), and the marginal belief function $m^{\mathcal{X}}$ in (3)-(5).

Using a similar line of reasoning, the form of the focal sets in the trinomial case ($K = 3$) could be obtained. However, the problem again quickly proved to be analytically intractable, and no formula such as (2) and (3)-(5) were given, even for the case $K = 3$.

2.2 The imprecise Dirichlet model

In this review of previous work, it is worth mentioning the imprecise Dirichlet model (IDM) introduced by Walley [38, 3]. This model was proposed in the imprecise probability framework [39], which is distinct from the TBM [30]. However, it turns out to yield a belief function on X when applied to our problem, which is the reason why it is mentioned here.

In short, the IDM extends Bayesian inference as follows. In the Bayesian setting, the conjugate prior probability distribution of parameter $\mathbf{p} = (p_1, \ldots, p_K)$ in the multinomial model is the Dirichlet $(s, \mathbf{t})$ distribution, where $\mathbf{t} = (t_1, \ldots, t_K)$ is an element of the interior of the unit simplex $S(1, K)$, and $s > 0$ is a hyperparameter determining the influence of the prior distribution on posterior probabilities. The predictive probability distribution on X, based on an iid random sample $X_1, \ldots, X_n$ and a prior Dirichlet $(s, \mathbf{t})$ distribution is:

$$p(\xi_k|\mathbf{N}, \mathbf{t}, s) = \frac{N_k + st_k}{n + s},$$

where $N_k = \sum_{i=1}^{n} 1_{\xi_k}(X_i)$ denote the number of observations in category ξ_k, and $\mathbf{N} = (N_1, \ldots, N_K)$. Assume now that we no longer take a single prior Dirichlet distribution, but the set of all Dirichlet $(s, \mathbf{t})$ distributions with $\mathbf{t} \in S(1, K)$. The family of all corresponding predictive distributions on X is then characterized by the following lower probability measure:

$$\underline{P}(A|\mathbf{N}, s) = \frac{N(A)}{n + s}, \quad \forall A \subseteq \mathcal{X},$$

with $N(A) = \sum_{\xi_k \in A} N_k$. It happens that $\underline{P}(\cdot|\mathbf{N}, s)$ is a belief function, with corre-

sponding bba:

$$\begin{aligned} m^{\mathcal{X}}(\{\xi_k\}|\mathbf{N}, s) &= \frac{N_k}{n+s} = \frac{\widehat{p}_k}{1+s/n}, \quad k = 1, \ldots, K, \\ m^{\mathcal{X}}(\mathcal{X}|\mathbf{N}, s) &= \frac{s}{s+n}, \end{aligned}$$

with $\widehat{p}_k = N_k/n$. We observe that, with $s = 1$, this solution is identical to Dempster's solution (3)-(5) in the binomial case.

This approach was extended by Utkin [36] to the case of imprecise observations.

2.3 Discussion

Dempster's approach outlined in Section 2.1 seems well founded theoretically. It provides a usable solution at least in the binomial case, and maybe for $K = 3$ (although this solution does not seem to have been fully worked out in that case). However, this approach seems to become quickly analytically intractable for larger K, essentially because of the difficulty to manipulate belief functions over continuous multidimensional spaces.

The IDM approach does lead to a simple belief function in the general multinomial case. However, whereas this approach is well founded in the imprecise probability setting, its justification in the TBM framework is unclear. It is based on the assumption that one's prior knowledge on the probability distribution of X is represented by a family of Dirichlet distributions, an assumption which can hardly be justified in the TBM, where each piece of knowledge is assumed to be represented by a belief function.

The approach proposed in this paper, which was also applied in a possibilistic framework [20], is fundamentally different. First of all, our objective will be more limited, in that we shall only attempt to build a belief function regarding a future observation X, given past observations $X_1, \ldots, X_n$, without expliciting our beliefs on $\mathcal{P}$. Hence, belief functions will not be used as a tool for parametric inference (for which frequentist confidence regions will be employed), but as a tool for prediction.

As mentioned aboved, another feature of our approach is that it will be essentially based on frequentist analysis. Given an iid random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ with parent probability distribution $\mathbb{P}_X$, we want to produce a belief function on $\mathcal{X}$, noted $bel^{\mathcal{X}}[\mathbf{X}_n]$, in such a way that the inequality $bel^{\mathcal{X}}[\mathbf{X}_n] \leq \mathbb{P}_X$ will hold in the long run

at least in $100(1-\alpha)$ % of cases (i.e., for a fraction $100(1-\alpha)$ of the samples). For a given realization $\mathbf{x}_n = (x_1, \ldots, x_n)$, we shall thus obtain a belief function $bel^{\mathcal{X}}[\mathbf{x}_n]$, which will be guaranteed to have been obtained by a method yielding a belief function less committed than the probability measure $\mathbb{P}_X$ in $100(1-\alpha)$ % of cases. As will be shown below, such a belief function can easily be computed from multinomial confidence regions, and it has a simple interpretation.

This method will be presented in detail in the rest of this paper.

3 Exploiting multinomial confidence regions

3.1 Basic principles

Let us assume that we have an urn with balls of different colors, noted $\mathcal{X} = \{\xi_1, \ldots, \xi_K\}$. The set $\mathcal{X}$ is given, but neither the number of balls, nor the proportions of balls of different colors are known. Let X denote the color of a ball taken randomly from the urn. As before, the probability distribution of X is noted $\mathbb{P}_X$. For each $A \subseteq \mathcal{X}$, $\mathbb{P}_X(A)$ represents the long run frequency of the event $X \in A$, which is simply equal in this example to the proportion of balls with color in A contained in the urn. This quantity is constant (it depends only on the experimental setting), but unknown.

Assume that we will shortly draw a ball from this urn, and we want to model our beliefs regarding its color by a belief function $bel^{\mathcal{X}}$. If we know the composition of the urn, and hence the underlying long run frequency distribution $\mathbb{P}_X$, it is reasonable to postulate $bel^{\mathcal{X}} = \mathbb{P}_X$. As remarked by Hacking [15], this "frequency principle" seems very natural.

Let us now assume that we do not know the composition of the urn, but we have drawn n balls with replacement. We have thus observed a realization of an iid random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$, with parent distribution $\mathbb{P}_X$. Let $bel^{\mathcal{X}}[\mathbf{X}_n]$ denote a belief function constructed using $\mathbf{X}_n$. Which properties should be satisfied by $bel^{\mathcal{X}}[\mathbf{X}_n]$?

First, it seems natural to impose that $bel^{\mathcal{X}}[\mathbf{X}_n]$ become closer to $\mathbb{P}_X$ as $n \to \infty$, which can be seen as a weak form of Hacking's frequency principle. Loosely speaking, a sample of infinite size is equivalent to knowledge of the distribution of X, hence the belief function should asymptotically become identical to $\mathbb{P}_X$. This translates to the

following requirement:

Requirement R_1:

$$\forall A \subseteq \mathcal{X}, \quad bel^{\mathcal{X}}[\mathbf{X}_n](A) \xrightarrow{P} \mathbb{P}_X(A), \text{ as } n \to \infty, \tag{6}$$

where $\xrightarrow{P}$ denotes convergence in probability.

For finite n, what kind of relationship should be imposed between $bel^{\mathcal{X}}[\mathbf{X}_n]$ and $\mathbb{P}_X$? Since we have less information than in the asymptotic case, it seems natural to impose that $bel^{\mathcal{X}}[\mathbf{X}_n]$ be *less committed* than $\mathbb{P}_X$, as a consequence of the Least Commitment Principle, which plays a central role in the TBM [31]. We should then have $bel^{\mathcal{X}}[\mathbf{X}_n] \leq \mathbb{P}_X$. This requirement, however, appears to be much too stringent. Having observed a positive count n_k for a certain value ξ_k of X, we can rule out 0 as a possible value for p_k. However, neither the total number of balls, nor an upper bound of it, are given. Consequently, any arbitrarily small value ϵ remains possible, unlikely as it may be. The above requirement would then lead to $bel^{\mathcal{X}}[\mathbf{X}_n](A) = 0$, for any strict subset A of $\mathcal{X}$, i.e., to the vacuous belief function.

As a less stringent requirement, we propose to impose that the inequality $bel^{\mathcal{X}}[\mathbf{X}_n] \leq \mathbb{P}_X$ be satisfied only "in most cases". Assuming that the random experiment that consists of drawing n balls from the urn is repeated indefinitely, we would like $bel^{\mathcal{X}}[\mathbf{X}_n]$ to be less committed than $\mathbb{P}$ "most of the time", i.e. with at least some prescribed long run frequency $1 - \alpha$, where $\alpha \in (0, 1)$ is an arbitrarily small positive number. More formally, this can be expressed by the following second requirement:

Requirement R_2:

$$\mathbb{P}(bel^{\mathcal{X}}[\mathbf{X}_n] \leq \mathbb{P}_X) \geq 1 - \alpha. \tag{7}$$

Equation (7) can alternatively be written:

$$\mathbb{P}\left(bel^{\mathcal{X}}[\mathbf{X}_n](A) \leq \mathbb{P}_X(A), \forall A \subset \mathcal{X}\right) \geq 1 - \alpha.$$

It should be quite clear that, in this expression, as in (6), $\mathbb{P}_X$ denotes the parent probability distribution of X, which is constant but unknown. The quantity $bel^{\mathcal{X}}[\mathbf{X}_n](A)$ is random, as it is a function of the random sample $\mathbf{X}_n$.

A belief function satisfying requirements R_1 and R_2 will be called a *predictive belief function at confidence level* $1-\alpha$.

In the following, we shall examine methods for deriving such belief functions from multinomial confidence regions. Some definitions and results regarding these confidence regions will first be recalled in the following section.

3.2 Multinomial confidence regions

The main building block of our approach to constructing belief functions is composed of methods for building confidence regions on multinomial parameters. Given an iid sample $X_1, \ldots, X_n$ of a discrete random variable X taking values in $\mathcal{X} = \{\xi_1, \ldots, \xi_K\}$, let $N_k = \sum_{i=1}^{n} 1_{\xi_k}(X_i)$ denote the number of observations in category ξ_k. The random vector $\mathbf{N} = (N_1, \ldots, N_K)$ has a multinomial distribution with parameters n and $\mathbf{p} = (p_1, \ldots, p_K)$, with $p_k = \mathbb{P}_X(\{\xi_k\})$.

Let $S(\mathbf{N})$ be a random subset of the parameter space $\mathcal{P} = \{\mathbf{p} = (p_1, \ldots, p_K) \in [0,1]^K \mid \sum_{k=1}^{K} p_k = 1\}$. $S(\mathbf{N})$ is said to be a confidence region for $\mathbf{p}$ at confidence level $1-\alpha$, if

$$\mathbb{P}(S(\mathbf{N}) \ni \mathbf{p}) \geq 1 - \alpha,$$

i.e., the random region $S(\mathbf{N})$ contains the constant parameter vector $\mathbf{p}$ with probability (long-run frequency) $1-\alpha$. It is an asymptotic confidence region if the above inequality only holds in the limit as $n \to \infty$.

The problem of finding confidence regions for multinomial proportions has received considerable attention in the statistical literature from the 1960's [22] [14] up to these days [12] [28] [21] [13] [17]. Of particular interest are simultaneous confidence intervals, i.e., regions defined as a Cartesian product of intervals:

$$S(\mathbf{N}) = [P_1^-, P_1^+] \times \ldots \times [P_K^-, P_K^+],$$

which have easy interpretation. Such asymptotic confidence regions were proposed by Quesenberry and Hurst [22], and Goodman [14]. The first solution is defined as:

$$P_k^- = \frac{a + 2N_k - \sqrt{\Delta_k}}{2(n+a)} \tag{8}$$

$$P_k^+ = \frac{a + 2N_k + \sqrt{\Delta_k}}{2(n+a)}, \tag{9}$$

where a is the quantile of order $1-\alpha$ of the chi-square distribution with one degree of freedom, and

$$\Delta_k = a\left(a + \frac{4N_k(n-N_k)}{n}\right).$$

It can easily be checked that the classical confidence interval on binomial p is recovered as a special case when $K = 2$. For $K > 2$, Goodman remarked that the above confidence region is too conservative, and showed that a could be replaced by b, the quantile of order $1-\alpha/K$ of the chi-square distribution with one degree of freedom. Note that we have $P_k^- \xrightarrow{P} p_k$ and $P_k^+ \xrightarrow{P} p_k$ as $n \to +\infty$, for $k = 1, \ldots, K$.

Other simple analytical expressions were suggested in [12], while more complex computer procedures were proposed in [28][17], and bootstrap methods were presented in [13, 18]. The quality of a confidence region may be measured by its volume and its coverage probability $\mathbb{P}(\mathbf{p} \in S(\mathbf{N}))$. An asymptotic confidence region is conservative if its coverage probability for finite n is greater than the prescribed confidence level. Among conservative confidence regions, it is desirable to find one with as small a volume as possible. Although bootstrap methods may yield smaller regions, particularly for small sample sizes, Goodman's intervals have been found to be good enough in most practical applications [21].

EXAMPLE 1 The following example is taken from [21]. A sample of 220 psychiatric patients were categorized as either neurotic, depressed, schizophrenic or having a personality disorder. The observed counts were $\mathbf{n} = (91, 49, 37, 43)$. The Goodman confidence intervals at confidence level $1-\alpha = 0.95$ are given in Table 1.

3.3 From multinomial confidence regions to lower probabilities

A confidence region $S(\mathbf{N})$ for multinomial proportions such as reviewed in Section 3.2 is usually interpreted as defining a set of plausible values for the vector parameter $\mathbf{p}$. However, since each value of $\mathbf{p}$ specifies a unique probability measure of $\mathcal{X}$, it is clear that $S(\mathbf{N})$ can equivalently be seen as defining a family of probability measures[3]. Such a family, obtained by bounding the probability of each singleton, is called a *set*

[3]To keep the notation as simple as possible, the same symbol $S(\mathbf{N})$ will be used to denote both the set of parameter values $\mathbf{p}$ and the set of probability measures.

of probability intervals in [5]. Note that we have [14]:

$$P_k^+ \leq 1 - \sum_{\ell \neq k} P_k^- \tag{10}$$

and

$$P_k^- \geq 1 - \sum_{\ell \neq k} P_k^+. \tag{11}$$

Consequently, this set of probability intervals is *reachable*, using the terminology introduced in [5].

Let P^- and P^+ denote, respectively, the lower and upper envelopes of $S(\mathbf{N})$, defined as $P^-(A) = \min_{P \in S(\mathbf{N})} P(A)$ and $P^+(A) = \max_{P \in S(\mathbf{N})} P(A)$. For all strict nonempty subset A of $\mathcal{X}$, we have [5]:

$$P^-(A) = \max \left(\sum_{\xi_k \in A} P_k^-, 1 - \sum_{\xi_k \notin A} P_k^+ \right) \tag{12}$$

$$P^+(A) = \min \left(\sum_{\xi_k \in A} P_k^+, 1 - \sum_{\xi_k \notin A} P_k^- \right). \tag{13}$$

Note that we have, as a direct consequence of the above formula:

$$P^+(A) = 1 - P^-(\overline{A}), \quad \forall A \subseteq \mathcal{X}.$$

Hence, the lower probability measure P^- is sufficient to characterize $S(\mathbf{N})$:

$$S(\mathbf{N}) = \{P \mid P^- \leq P\}.$$

By construction, we have

$$\mathbb{P}(\mathbb{P}_X \in S(\mathbf{N})) = \mathbb{P}(P^- \leq \mathbb{P}_X) \geq 1 - \alpha, \tag{14}$$

and it is clear that $P^-(A) \xrightarrow{P} \mathbb{P}_X(A)$ as $n \to \infty$, for all $A \subseteq \mathcal{X}$. Hence, P^- verifies our two requirements R_1 and R_2. Unfortunately, P^- is not, in general, a belief function, except for the cases $K = 2$ and $K = 3$ (see Section 3.4 below). This can be shown by the following counterexample.

Example 2 Let us return to the confidence region computed in Example 1. The corresponding lower probabilities are shown in Table 2. As shown by Shafer [23], a mapping $f : 2^{\mathcal{X}} \to [0, 1]$ is a belief function iff its Möbius inverse, defined as:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} f(B), \quad \forall A \subseteq \mathcal{X},$$

is a basic belief assignment (bba), i.e., if $m(A) \geq 0$ for all A, and $\sum_{A \subseteq \mathcal{X}} m(A) = 1$. The Möbius inverse of P^-, shown in Table 2, assigns a negative value to $\mathcal{X}$. Consequently, P^- is not a belief function.

Before proposing in the next section a way to construct a belief function from P^-, we shall conclude this section by noticing that P^-, although not a belief function, possesses a weaker property: as shown by Campos et al. [5], sets of probability intervals are Choquet capacities of order two, i.e., we have

$$P^-(A \cup B) \geq P^-(A) + P^-(B) - P^-(A \cap B), \quad \forall A, B \subseteq \mathcal{X}. \tag{15}$$

3.4 From lower probabilities to predictive belief functions

The case $K = 2$

When $K = 2$, the lower probability measure P^- defined above is actually a belief function. Its bba is simply equal to:

$$\begin{aligned} m^{\mathcal{X}}(\{\xi_1\}) &= P_1^- \\ m^{\mathcal{X}}(\{\xi_2\}) &= P_2^- \\ m^{\mathcal{X}}(\mathcal{X}) &= 1 - P_1^- - P_2^-, \end{aligned}$$

with P_1^- and P_2^- defined by (8). If we note $N = N_1$, we have the expressions:

$$\begin{aligned} m^{\mathcal{X}}(\{\xi_1\}) &= \frac{a + 2N - \sqrt{\Delta}}{2(n+a)} \\ m^{\mathcal{X}}(\{\xi_2\}) &= \frac{a + 2(n-N) - \sqrt{\Delta}}{2(n+a)} \\ m^{\mathcal{X}}(\mathcal{X}) &= \frac{2\sqrt{\Delta}}{2(n+a)}, \end{aligned}$$

where a is the quantile of order $1 - \alpha$ of the chi-square distribution with one degree of freedom (which is also equal to $u^2_{1-\alpha/2}$, the square of the normal quantile of order $1 - \alpha/2$), and

$$\Delta = a \left(a + \frac{4N(n-N)}{n} \right).$$

Using the classical approximation of binomial confidence intervals, it is easy to show that:

$$m^{\mathcal{X}}(\{\xi_1\}) \sim \widehat{p} - u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \quad (16)$$

$$m^{\mathcal{X}}(\{\xi_2\}) \sim 1 - \widehat{p} - u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \quad (17)$$

$$m^{\mathcal{X}}(\mathcal{X}) \sim 2u_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \quad (18)$$

where, as before, $\widehat{p} = N/n$, and $\sim$ denotes asymptotic equivalence. It is interesting to compare these expressions with (3)-(5). We can see that the mass $m^{\mathcal{X}}(\mathcal{X})$ tends towards 0 as $n^{-1/2}$ in our approach, whereas it has the higher convergence rate of n^{-1} in Dempster's solution. Our solution is thus more conservative (even for small n), which seems to be the price to pay to statisfy requirement R_2.

EXAMPLE 3 To illustrate the nature of these two solutions, we generated 100 realizations of a binomial random variable with $p = 0.3$ and $n = 30$. This simulates 100 repetitions of the random experiment consisting in drawing 30 balls, with replacement, from an urn containing 30 % of white balls. We thus obtained 100 predictive belief functions at confidence level $1 - \alpha = 0.95$, and 100 belief functions using Dempster's approach. These belief functions are plotted in Figure 3 in the three-dimensional probability simplex. Each belief function is represented as a point in an equilateral triangle using barycentric coordinates, with the lower left corner corresponding to $\{\xi_1\}$ (say, the elementary event consisting in drawing a white ball), the lower right corner corresponding to $\{\xi_2\}$, and the upper corner corresponding to $\mathcal{X}$. The orthogonal distance to the lower side of the triangle is thus proportional to $m^{\mathcal{X}}(\Omega)$, while the distances to the right-hand and left-hand sides are proportional to $m^{\mathcal{X}}(\{\xi_1\})$ and $m^{\mathcal{X}}(\{\xi_2\})$, respectively. The grey region corresponds to the set of belief functions $bel^{\mathcal{X}}$ less committed that $\mathbb{P}_X$, i.e., such that $bel^{\mathcal{X}}(\{\xi_1\}) \leq \mathbb{P}(\{\xi_1\}) = 0.3$ and $bel^{\mathcal{X}}(\{\xi_2\}) \leq \mathbb{P}(\{\xi_2\}) = 0.7$. We can verify that, out of the 100 predictive belief functions $m^{\mathcal{X}*}$, about 95% satisfy this property, which is an experimental verification of requirement R_2. Dempster's belief functions are more specific (they are closer to the lower side of the rectangle in the graphical representation), but most of them are more committed than $\mathbb{P}_X$: requirement R_2 is not satisfied in this approach. Figure 4 shows

the result of a similar numerical experiment repeated with $n = 100$. As expected, the belief functions computed by each of the two methods get closer to $\mathbb{P}_X$ as n increases, which is a consequence of requirement R_1 being satisfied by the two approaches.

The case $K = 3$

When $K = 3$, P^- is again a belief function. To prove this assertion, let us consider the Möbius inverse of P^- in this case. We have

$$m^{\mathcal{X}}(\{\xi_k\}) = P_k^-, \quad k = 1, 2, 3$$

$$\begin{aligned} m^{\mathcal{X}}(\{\xi_1, \xi_2\}) &= P^-(\{\xi_1, \xi_2\}) - P^-(\{\xi_1\}) - P^-(\{\xi_2\}) \\ &= 1 - P_3^+ - P_1^- - P_2^- \end{aligned}$$

$$\begin{aligned} m^{\mathcal{X}}(\{\xi_1, \xi_3\}) &= P^-(\{\xi_1, \xi_3\}) - P^-(\{\xi_1\}) - P^-(\{\xi_3\}) \\ &= 1 - P_2^+ - P_1^- - P_3^- \end{aligned}$$

$$\begin{aligned} m^{\mathcal{X}}(\{\xi_2, \xi_3\}) &= P^-(\{\xi_2, \xi_3\}) - P^-(\{\xi_2\}) - P^-(\{\xi_3\}) \\ &= 1 - P_1^+ - P_2^- - P_3^- \end{aligned}$$

$$\begin{aligned} m^{\mathcal{X}}(\mathcal{X}) &= 1 - \sum_{k=1}^{3} m(\{\xi_k\}) - \sum_{k \neq \ell} m(\{\xi_k, \xi_\ell\}) \\ &= 1 - \sum_{k=1}^{3} P_k^- - (1 - P_3^+ - P_1^- - P_2^-) - (1 - P_2^+ - P_1^- - P_3^-) \\ &\quad -(1 - P_1^+ - P_2^- - P_3^-) \\ &= \sum_{k=1}^{3} (P_k^+ + P_k^-) - 2 \\ &= \sum_{k=1}^{3} \frac{b + 2N_k}{n + b} - 2 = \frac{b}{n + b}, \end{aligned}$$

where b is, as before, the quantile of order $1 - \alpha/3$ of the chi-square distribution with one degree of freedom. The masses assigned to pairs $\{\xi_k, \xi_\ell\}$ are positive because of (10), and all other masses are obviously positive. Consequently, P^- is a belief function.

EXAMPLE 4 Let us consider an urn containing ball of three different colors, denoted ξ_1, ξ_2 and ξ_3 (say, black, white and red). We have drawn 100 balls with replacement, of which 20 were black, 30 were white and 50 were red. Let X denote the color of the next ball to be drawn from the urn. What is our belief regarding the value of X ?

We have the counts $N_1 = 20$, $N_2 = 30$ and $N_3 = 50$. The bounds of the Goodman confidence intervals, at confidence level 99 % are:

$$P_1^- = 0.1087 \qquad P_1^+ = 0.3389$$

$$P_2^- = 0.1858 \qquad P_2^+ = 0.4459$$

$$P_3^- = 0.3592 \qquad P_3^+ = 0.6408,$$

and we have $b = \chi^2_{1,1-0.01/3} = 8.6154$. We thus obtain the following bba:

$$m^{\mathcal{X}}(\{\xi_1\}) = 0.1087, \quad m^{\mathcal{X}}(\{\xi_2\}) = 0.1858, \quad m^{\mathcal{X}}(\{\xi_3\}) = 0.3592$$

$$m^{\mathcal{X}}(\{\xi_1, \xi_2\}) = 1 - 0.6408 - 0.1087 - 0.1858 = 0.0647$$

$$m^{\mathcal{X}}(\{\xi_1, \xi_3\}) = 1 - 0.4459 - 0.1087 - 0.3592 = 0.0862$$

$$m^{\mathcal{X}}(\{\xi_2, \xi_3\}) = 1 - 0.3389 - 0.1858 - 0.3592 = 0.1161$$

$$m^{\mathcal{X}}(\mathcal{X}) = \frac{8.6154}{100 + 8.6154} = 0.0793.$$

The case $K > 3$

When $K > 3$, P^- is no longer guaranteed to be a belief function, as shown by Example 2 above. We thus have to approximate P^- by a belief function satisfying requirements R_1 and R_2.

Let $\mathcal{B}^{\mathcal{X}}(P^-)$ denote the set of belief functions $bel^{\mathcal{X}}$ on $\mathcal{X}$ verifying $bel^{\mathcal{X}} \leq P^-$. As a consequence of (14), we have, for any $bel^{\mathcal{X}} \in \mathcal{B}^{\mathcal{X}}(P^-)$:

$$\mathbb{P}(bel^{\mathcal{X}} \leq \mathbb{P}_X) \geq \mathbb{P}(P^- \leq \mathbb{P}_X) \geq 1 - \alpha.$$

Every element of $\mathcal{B}$ thus complies with requirement R_2 expressed by (7). However, most elements of $\mathcal{B}^{\mathcal{X}}(P^-)$ (such as, e.g., the vacuous belief function) will generally not be very informative, and it seems natural to concentrate on the most committed elements of $\mathcal{B}^{\mathcal{X}}(P^-)$. Since there is not a single most specific element in $\mathcal{B}^{\mathcal{X}}(P^-)$, a

good solution can be found by maximizing a specificity criterion such as the sum of belief degrees[4] $bel^{\mathcal{X}}(A)$ for all $A \subseteq \mathcal{X}$, under the constraints $bel^{\mathcal{X}}(A) \leq \mathbb{P}_X(A)$, for all $A \subseteq \mathcal{X}$. Let $J(m^{\mathcal{X}})$ denote this criterion. We have

$$\begin{aligned} J(m^{\mathcal{X}}) &= \sum_{A \subseteq \mathcal{X}} bel^{\mathcal{X}}(A) && (19) \\ &= \sum_{A \subseteq \mathcal{X}} \sum_{B \subseteq A} m^{\mathcal{X}}(B) && (20) \\ &= \sum_{B \subseteq \mathcal{X}} m^{\mathcal{X}}(B) \, |\{A \subseteq \mathcal{X}, B \subseteq A\}| && (21) \\ &= 2^K \sum_{B \subseteq \mathcal{X}} 2^{-|B|} m^{\mathcal{X}}(B) \ , && (22) \end{aligned}$$

where $|\cdot|$ denotes cardinality. We then have to solve the following linear program:

$$\max_{m^{\mathcal{X}}} J(m^{\mathcal{X}}) \tag{23}$$

under the constraints:

$$\sum_{B \subseteq A} m^{\mathcal{X}}(B) \leq P^-(A), \quad \forall A \subset \mathcal{X}, \tag{24}$$

$$\sum_{A \subseteq \mathcal{X}} m^{\mathcal{X}}(A) = 1, \tag{25}$$

$$m^{\mathcal{X}}(A) \geq 0, \quad \forall A \subseteq \mathcal{X} \ . \tag{26}$$

Any belief function $bel_n^{\mathcal{X}*}$ solution to the above linear programming problem obviously satisfies requirement R_2. The following proposition states that it also satisfies R_1.

PROPOSITION 1 *Let $bel_n^{\mathcal{X}*}, n = 1, \ldots, \infty$ be a sequence of solutions of linear program (23)-(26). We have:*

$$bel_n^{\mathcal{X}*} \xrightarrow{P} \mathbb{P}_X \ \textit{as} \ n \to \infty.$$

Proof. See Appendix A.

Solving linear program (23)-(26) is thus a way to construct a predictive belief function, as illustrated by the next example. Note that the uniqueness of the solution is not guaranteed, which is not important in practice, since all the solutions may be regarded as equivalent.

[4] A similar criterion was proposed by Baroni and Vicig [2] and by Hall and Lawry [16] for approximating a lower probability measure by a belief function.

EXAMPLE 5 Table 3 shows optimal belief and mass functions, at confidence level 0.95, obtained for the data of Example 1 using a standard linear programming algorithm (we used the Matlab Optimization Toolbox). The value of the objective function for this solution is $J(m^{\mathcal{X}*}) = 6.4825$.

Note that the applicability of the method is obviously limited to moderate values of K (up to 10-15), since both the number of variables and the number of constraints grow exponentially with K. For large K or when computation speed is an issue, however, it may be sufficient to compute suboptimal solutions.

This can be done, for instance, using the Iterative Rescaling Method (IRM) described in [16], which heuristically transforms the Möbius inversion of P^- into a bba, by replacing each negative mass $m(A) < 0$ by zero, and rescaling masses assigned to relevant subsets of A. The result of this algorithm for the psychiatric data are shown in the last two columns of Table 3. In that particular case, the obtained solution happens to be optimal, since the value of J for this solution is the same as the one computed in Example 5. However, the IRM provides only an approximation to the optimum in the general case.

Although the IRM algorithm may allow to find good approximations for moderate values of K, its time and space complexity is still exponential as a function of K (it involves a loop over the subsets of $\mathcal{X}$). Much more drastic approximations may be obtained by limiting the search to a restricted parametrized family of belief functions $\mathcal{B}_0^{\mathcal{X}}(P^-) \subset \mathcal{B}^{\mathcal{X}}(P^-)$. The simplest such family is perhaps the set of belief functions whose focal elements are taken among the singletons and $\mathcal{X}$: in that case, the optimal solution is $m_n^{\mathcal{X}\circ}$ introduced in Appendix B, defined by:

$$m_n^{\mathcal{X}\circ}(\{\xi_k\}) = P_k^-, \quad k = 1, \ldots, K \tag{27}$$

$$m_n^{\mathcal{X}\circ}(\mathcal{X}) = 1 - \sum_{k=1}^{K} P_k^- \ , \tag{28}$$

which can easily be shown to statisfy requirements R_1 and R_2.

EXAMPLE 6 For the psychiatric data of Example 1, we have $m_n^{\mathcal{X}\circ}(\{\xi_1\}) = 0.33$, $m_n^{\mathcal{X}\circ}(\{\xi_2\}) = 0.16$, $m_n^{\mathcal{X}\circ}(\{\xi_3\}) = 0.11$, $m_n^{\mathcal{X}\circ}(\{\xi_4\}) = 0.14$, and $m_n^{\mathcal{X}\circ}(\mathcal{X}) = 0.26$. The value of the objective function is $J(m_n^{\mathcal{X}\circ}) = 6.2296$. This solution is thus not optimal,

but it can be considered as an approximation of $m^{\mathcal{X}*}$.

In general, richer families of belief functions could be considered (e.g., by constraining the size of the focal sets). When the elements of $\mathcal{X}$ are ordered, it is quite natural to consider belief functions whose focal elements are intervals, since the corresponding basic belief masses can easily be represented and interpreted. In that case, the optimal solution has a simple analytical expression, as will be shown in the next section.

4 Approximation in the case of ordered data

4.1 Definitions

We assume in this section that a meaningful ordering[5] has been defined among the elements of $\mathcal{X}$. By convention, we shall assume that $\xi_1 < \ldots < \xi_K$.

Let $A_{k,r}$ denote the subset $\{\xi_k, \ldots, \xi_r\}$, for $1 \leq k \leq r \leq K$ and let $\mathcal{I}$ denote the set of intervals of $\mathcal{X}$: $\mathcal{I} = \{A_{k,r}, 1 \leq k \leq r \leq K\}$. These intervals may be represented graphically as in Figure 5, in which each interval $A_{k,r}$ appears at the intersection of row k and column r of a two-dimensional table. In this representation, the singletons are located on the main diagonal, the intervals of length 2 on the second diagonal, etc.

By imposing that the focal sets of m be taken in $\mathcal{I}$, one reduces the number of basic belief numbers from $2^K - 1$ to $K(K+1)/2$. Let $m^{\mathcal{X}}$ denote such a bba, and

[5]One could argue that this assumption is, to some extent, contradictory with the use of the multinomial model, which does not assume any order among the outcomes. However, there are cases where a natural ordering exists and makes sense to the user, in particular for graphical representations, but this knowledge may not easily be incorporated into a statistical model. This happens, for instance, when a quantitative variable is discretized by defining a finite number of classes. The multinomial model is then still a common choice, as it makes minimal assumptions. This situation is considered in this section.

$bel^{\mathcal{X}}$ the corresponding belief function. We have:

$$m^{\mathcal{X}}(A_{k,r}) =$$
$$\begin{cases} bel^{\mathcal{X}}(\{\xi_k\}) & \text{if } k = r \\ bel^{\mathcal{X}}(A_{k,r}) - bel^{\mathcal{X}}(A_{k+1,r}) - bel^{\mathcal{X}}(A_{k,r-1}) & \text{if } r = k+1, \\ bel^{\mathcal{X}}(A_{k,r}) - bel^{\mathcal{X}}(A_{k+1,r}) - bel^{\mathcal{X}}(A_{k,r-1}) + bel^{\mathcal{X}}(A_{k+1,r-1}) & \text{if } r > k+1, \end{cases} \tag{29}$$

If there exists a bba $m^{\mathcal{X}*}$ verifying $bel^{\mathcal{X}*}(A_{k,r}) = P^-(A_{k,r})$ for all $A_{k,r} \in \mathcal{I}$, we then have necessarily:

$$m^{\mathcal{X}*}(A_{k,r}) = \begin{cases} P_k^- & \text{if } k = r \\ P^-(A_{k,r}) - P^-(A_{k+1,r}) - P^-(A_{k,r-1}) & \text{if } r = k+1, \\ P^-(A_{k,r}) - P^-(A_{k+1,r}) - P^-(A_{k,r-1}) + P^-(A_{k+1,r-1}) & \text{if } r > k+1, \end{cases} \tag{30}$$

$$m^{\mathcal{X}*}(B) = 0, \quad \forall B \notin \mathcal{I} \ . \tag{31}$$

In the following, we will show that $m^{\mathcal{X}*}$ defined by (30)-(31) is indeed a valid bba (i.e., it defines a belief function), and that it is optimal according to criterion J defined by (19), in the set of bbas with focal elements in $\mathcal{I}$.

4.2 Properties of $m^{\mathcal{X}*}$

PROPOSITION 2 *The function defined by (30)-(31) is a valid bba.*

Proof. As a consequence of (15), we have $m^{\mathcal{X}*}(A_{k,r}) \geq 0$, $\forall r \geq k$. We then have to prove that

$$\sum_{k=1}^{K} \sum_{r=k}^{K} m^{\mathcal{X}*}(A_{k,r}) = 1 \ .$$

In this sum, each term is a linear combination of lower probabilities $P^-(A_{k,r})$. Each value $P^-(A_{k,r})$ appears:

- with a + sign in $m^{\mathcal{X}*}(A_{k,r})$;
- with a − sign in $m^{\mathcal{X}*}(A_{k-1,r})$ if $k > 1$;
- with a − sign in $m^{\mathcal{X}*}(A_{k,r+1})$ if $r < K$;
- with a + sign in $m^{\mathcal{X}*}(A_{k-1,r+1})$ if $k > 1$ and $r < K$.

The sum of these terms is equal to 0, except for $k = 1$ and $r = K$. Consequently, we have

$$\sum_{k=1}^{K} \sum_{r=k}^{K} m^{\mathcal{X}*}(A_{k,r}) = P^-(A_{1,K}) = 1 \ .$$

□

We will now show that $m^{\mathcal{X}*}$ defined above belongs to $\mathcal{B}^{\mathcal{X}}(P^-)$, and that the associated belief function $bel^{\mathcal{X}*}$ coincides with P^- on the intervals of $\mathcal{X}$.

PROPOSITION 3 *Let $bel^{\mathcal{X}*}$ be the belief function associated with $m^{\mathcal{X}*}$ defined by (30)-(31). We have:*

$$bel^{\mathcal{X}*}(A_{k,r}) = P^-(A_{k,r}), \quad \forall A_{k,r} \in \mathcal{I}, \tag{32}$$

$$bel^{\mathcal{X}*}(A) \leq P^-(A), \quad \forall A \in 2^{\mathcal{X}} \ . \tag{33}$$

Proof. We first prove (32) by induction on the length of the interval $\ell = r - k + 1$. Obviously, (32) is true for $\ell = 1$, since, by definition $bel^{\mathcal{X}*}(A_{k,k}) = m(\{\xi_k\}) = P_k^-$. It is also true for $\ell = 2$, since

$$\begin{aligned} bel^{\mathcal{X}*}(A_{k,k+1}) &= m(\{\xi_k\}) + m(\{\xi_{k+1}\}) + m(A_{k,k+1}) \\ &= P_k^- + P_{k+1}^- + P^-(A_{k,k+1}) - P_k^- - P_{k+1}^- \\ &= P^-(A_{k,k+1}). \end{aligned}$$

Let us now consider two indices k and r such that $r - k + 1 \geq 3$, and let us assume that (32) is true for all $\ell < r - k + 1$. It is easy to see that

$$bel^{\mathcal{X}*}(A_{k,r}) = bel^{\mathcal{X}*}(A_{k,r-1}) + bel^{\mathcal{X}*}(A_{k+1,r}) - bel^{\mathcal{X}*}(A_{k+1,r-1}) + m^{\mathcal{X}*}(A_{k,r}) \tag{34}$$

Because (32) is true for all intervals smaller than $A_{k,r}$, we can replace $bel^{\mathcal{X}*}$ by P^- in the first three terms in the right-hand side of (34). Replacing the last term by its definition using (30), we have

$$\begin{aligned} bel^{\mathcal{X}*}(A_{k,r}) = P^-(A_{k,r-1}) + P^-(A_{k+1,r}) - P^-(A_{k+1,r-1}) + \\ \big(P^-(A_{k,r}) - P^-(A_{k+1,r}) - P^-(A_{k,r-1}) + P^-(A_{k+1,r-1})\big) = P^-(A_{k,r}), \end{aligned} \tag{35}$$

which completes the proof of (32).

To show (33), we remark that any arbitrary nonempty subset A of $\mathcal{X}$ may be written as the union of Q disjoint intervals:

$$A = \bigcup_{q=1}^{Q} A_{k_q,r_q} \ .$$

Since the focal sets of $m^{\mathcal{X}*}$ are intervals, all focal sets included in A are included in one of the A_{k_q,r_q}. Consequently, we have

$$bel^{\mathcal{X}*}(A) = \sum_{q=1}^{Q} bel^{\mathcal{X}*}(A_{k_q,r_q}),$$

which implies that

$$bel^{\mathcal{X}*}(A) = \sum_{q=1}^{Q} P^-(A_{k_q,r_q})$$

Now, since the A_{k_q,r_q} are disjoint, we have, as a consequence of (15):

$$P^-(A) \geq \sum_{q=1}^{Q} P^-(A_{k_q,r_q}).$$

Hence, $bel^{\mathcal{X}*}(A) \leq P^-(A)$, which establishes the result. □

Finally, the following proposition states that $m^{\mathcal{X}*}$ is optimal (according to criterion J defined by (19)), in the set of all bbas with focal elements in $\mathcal{I}$.

PROPOSITION 4 *$m^{\mathcal{X}*}$ defined by (30)-(31) is the unique solution to the linear program (23)-(26), under the additional constraints*

$$m^{\mathcal{X}}(A) = 0, \quad \forall A \notin \mathcal{I} \ .$$

Proof. We have seen in the proof of Proposition 3 that any $A \subseteq \mathcal{X}$ is either an interval, or a union of disjoint intervals, and $bel^{\mathcal{X}}(A)$ can then be written as the sum of the beliefs given to the disjoint intervals (assuming that the focal elements of $bel^{\mathcal{X}}$ are intervals). Consequently, $J(m^{\mathcal{X}})$ can be written as a linear combination of $bel^{\mathcal{X}}(A_{k,r})$ for all $A_{k,r} \in \mathcal{I}$:

$$J(m^{\mathcal{X}}) = \sum_{k=1}^{K} \sum_{r=k}^{K} \alpha_{k,r} bel^{\mathcal{X}}(A_{k,r}),$$

where the $\alpha_{k,r}$ are positive coefficients. Since $bel^{\mathcal{X}}(A_{k,r}) \leq P^-(A_{k,r})$ for all k, r, it is clear that J is maximum for $m^{\mathcal{X}} = m^{\mathcal{X}*}$, which is the only bba satisfying $bel^{\mathcal{X}*}(A_{k,r}) = P^-(A_{k,r})$ for all k, r. □

EXAMPLE 7 Table 4 shows categorized data[6] concerning January precipitation in Arizona (in inches), recorded during the period 1895-2004, together with the estimated probabilities of each class, and Goodman simultaneous confidence intervals at confidence level 0.95. Based on this data, what is our belief that the precipitation in Arizona next january will exceed, say, 2.25 inches? The masses $m^{\mathcal{X}*}(A_{k,r})$ are given numerically in Table 5 and graphically in Figure 6 (where each mass is represented by a circle with proportional area), using the representation of Figure 5. The same information is depicted differently in Figure 7, showing on the y-axis the masses given to intervals whose bounds are read on the x-axis, together with the plausibility contour function $\xi \to pl^{\mathcal{X}*}(\{\xi\})$, and the upper bounds of the multinomial confidence intervals (by construction, the lower bounds p_k^- are equal to the masses $m^{\mathcal{X}*}(\{\xi_k\})$ given to the singletons). It may be noted that the plausibility values are significantly higher than the upper bounds of the multinomial confidence intervals, which reflects a loss of information due to the approximation of a set of probability intervals by a belief function. Using the data in Table 5, the answer to the above question can easily be computed; we have:

$$\begin{aligned} bel(X \geq 2.25) &= bel^{\mathcal{X}*}(\{\xi_5, \xi_6\}) \\ &= m^{\mathcal{X}*}(\{\xi_5\}) + m^{\mathcal{X}*}(\{\xi_6\}) + m^{\mathcal{X}*}(\{\xi_5, \xi_6\}) \\ &= 0.020 + 0.035 + 0 = 0.055, \end{aligned}$$

and

$$pl(X \geq 2.25) = pl^{\mathcal{X}*}(\{\xi_5, \xi_6\}) = 0.020 + 0.035 + 0.012 + 0.14 + 0.11 = 0.317 \ .$$

5 Conclusion

We have proposed a method for quantifying, in the belief functions framework, the uncertainty concerning a discrete random variable X with unknown probability distribution $\mathbb{P}_X$, based on a realization of an iid sample from the same distribution. The

[6]The Arizona precipitation data have been obtained from the web page of the National Climatic Data Center, National Oceanic and Atmospheric Administration (NOAA), at the following address: http://www.ncdc.noaa.gov/oa/ncdc.html.

proposed solution verifies two "reasonable" properties with respect to $\mathbb{P}_X$: it is less committed than $\mathbb{P}_X$ with some user-defined probability, and it converges towards $\mathbb{P}_X$ in probability as the size of the sample tends to infinity.

This solution is obtained by searching for the most committed belief function that is less committed than the lower probability measure induced by simultaneous confidence intervals on multinomial parameters, at a given confidence level. This can be formalized as a linear programming problem, which can be solved using standard iterative procedures. However, an analytic expression has been given in the case of ordered data, under the additional constraint that all focal elements are intervals.

Although the resulting belief function is deduced from a lower probability measure, their semantics are different: the lower probability defines a set of "plausible" values for $\mathbb{P}_X$, given the data, whereas the belief function is interpreted as quantifying beliefs held by a rational agent, as assumed in the TBM framework. It might be argued that the imprecise probability measure induced by the confidence intervals is an equally good characterization of the uncertainty on X. However, this lower probability is not a belief function; consequently, it cannot be combined with other pieces of information expressed in the belief function framework. Its transformation into a belief function is thus needed if ones adopts the TBM as a model of uncertain reasoning.

In this paper, only the case of a discrete random variable X has been considered. The method can be applied to the continuous case by discretizing the sample values (which is a form of coarsening), and vacuously extending the obtained belief function. A specific method designed for the continuous case is under study.

References

[1] R. G. Almond. *Graphical belief models*. Chapman and Hall, London, 1995.

[2] P. Baroni and P. Vicig. An uncertainty interchange format with imprecise probabilities. *International Journal of Approximate Reasoning*, 40(3):147–180, 2005.

[3] J.-M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2–3):123–150, 2005.

[4] N. Bryson and A. Mobolurin. A process for generating quantitative belief functions. *European Journal of Operational Research*, 115:624–633, 1999.

[5] L. M. de Campos, J. F. Huete, and S. Moral. Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(2):167–196, 1994.

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

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

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

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

[10] A. P. Dempster. A class of random convex polytopes. *Annals of Mathematical Statistics*, 43(1):260–272, 1972.

[11] D. Dubois, H. Prade, and Ph. Smets. New semantics for quantitative possibility theory. In *2nd International Symposium on Imprecise Probabilities and their Applications*, Ithaca, NY, 2001.

[12] S. Fitzpatrick and A. Scott. Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association*, 82(399):875–878, 1987.

[13] J. Glaz and C. P. Sison. Simultaneous confidence intervals for multinomial proportions. *Journal of Statistical Planning and Management*, 82:251–262, 1999.

[14] L. A. Goodman. On simultaneous confidence intervals for multinomial proportions. *Technometrics*, 7(2):247–254, 1965.

[15] I. Hacking. *Logic of Statistical Inference*. Cambridge University Press, Cambridge, 1965.

[16] J. W. Hall and J. Lawry. Generation, combination and extension of random set approximations to coherent lower and upper probabilities. *Reliability Engineering and System Safety*, (85):89–101, 2004.

[17] C.-D. Hou, J. Chiang, and J. J. Tai. A family of simultaneous confidence intervals for multinomial proportions. *Computational Statistics and Data Analysis*, 43:23–45, 2003.

[18] M. Jhun and H.-C. Jeong. Application of bootstrap methods for categorical data analysis. *Computational Statistics and Data Analysis*, 35:83–91, 2000.

[19] J. Kohlas and P.-A. Monney. *A Mathematical Theory of Hints. An Approach to the Dempster-Shafer Theory of Evidence*. Springer-Verlag, Berlin, 1995.

[20] M. Masson and T. Denœux. Inferring a possibility distribution from empirical data. *Fuzzy Sets and Systems*, 157(3):319–340, 2006.

[21] W. L. May and W. D. Johnson. A SAS macro for constructing simultaneous confidence intervals for multinomial proportions. *Computer methods and Programs in Biomedicine*, 53:153–162, 1997.

[22] C. P. Quesenberry and D. C. Hurst. Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics*, 6(2):191–195, 1964.

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

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

[25] G. Shafer. Perspectives in the theory and practice of belief functions. *International Journal of Approximate Reasoning*, 4:323–362, 1990.

[26] G. Shafer, P. P. Shenoy, and K. Mellouli. Propagating belief functions in qualitative Markov trees. *International Journal of Approximate Reasoning*, 1:349–400, 1987.

[27] P. P. Shenoy. Binary joint trees for computing marginals in the Shenoy-Shafer architecture. *International Journal of Approximate Reasoning*, 17:239–263, 1997.

[28] C. P. Sison and J. Glaz. Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association*, 90(429):366–369, 1995.

[29] Ph. Smets. *Un modèle mathématico-statistique simulant le processus du diagnostic médical*. PhD thesis, Université Libre de Bruxelles, Brussels, Belgium, 1978. (in French).

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

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

[32] Ph. Smets. Belief induced by the partial knowledge of the probabilities. In D. Heckerman *et al.*, editor, *Uncertainty in AI'94*, pages 523–530. Morgan Kaufmann, San Mateo, 1994.

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

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

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

[36] L. V. Utkin. Extensions of belief functions and possibility distributions by using the imprecise Dirichlet model. *Fuzzy Sets and Systems*, 154(3):413–431, 2005.

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

[38] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *J. R. Statist. Soc. B*, 58(1):3–57, 1996.

[39] P. Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2–3):125–148, 2000.

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

[41] S. K. M. Wong and P. Lingras. Representation of qualitative user preference by quantitative belief functions. *IEEE Transactions on Knowledge and Data Engineering*, 6(1):72–78, 1994.

A Proof of Proposition 1

Let us introduce the basic belief assignment $m_n^{\mathcal{X}\circ}$ defined as

$$m_n^{\mathcal{X}\circ}(\{\xi_k\}) = P_k^-, \quad k = 1, \ldots, K \tag{36}$$

$$m_n^{\mathcal{X}\circ}(\mathcal{X}) = 1 - \sum_{k=1}^{K} P_k^-, \tag{37}$$

and let $bel_n^{\mathcal{X}\circ}$ denote the corresponding belief function ($m_n^{\mathcal{X}\circ}$ is a valid bba, since $\sum_{k=1}^{K} P_k^- \leq 1$). We have

$$bel_n^{\mathcal{X}\circ}(\{\xi_k\}) = P^-(\{\xi_k\}), \quad k = 1, \ldots, K,$$

and

$$bel_n^{\mathcal{X}\circ}(A) = \sum_{\xi_k \in A} P_k^- \leq P^-(A).$$

Hence, $bel_n^{\mathcal{X}\circ} \leq P^-$. Since $m_n^{\mathcal{X}*}$ maximizes J, we thus have $J(m_n^{\mathcal{X}\circ}) \leq J(m_n^{\mathcal{X}*})$.

Additionally, it is clear that $m_n^{\mathcal{X}\circ}(\{\xi_k\}) \xrightarrow{P} p_k$ for all k and, consequently, $m_n^{\mathcal{X}\circ}(\mathcal{X}) \xrightarrow{P} 0$. Hence, $J(m_n^{\mathcal{X}\circ}) \xrightarrow{P} 2^{K-1}$. Now, the unconstrained maximum of J is obtained when the mass is distributed to singletons, hence $J(m_n^{\mathcal{X}*}) \leq 2^{K-1}$. We thus have $J(m_n^{\mathcal{X}\circ}) \leq J(m_n^{\mathcal{X}*}) \leq 2^{K-1}$ and, consequently, $J(m_n^{\mathcal{X}*}) \xrightarrow{P} 2^{K-1}$. From this, it is easy to see that

$$\sum_{k=1}^{K} m_n^{\mathcal{X}*}(\{\xi_k\}) \xrightarrow{P} 1.$$

Now, we have also

$$\sum_{k=1}^{K} P_k^- \xrightarrow{P} 1.$$

Hence,

$$\sum_{k=1}^{K} (P_k^- - m_n^{\mathcal{X}*}(\{\xi_k\})) \xrightarrow{P} 0.$$

Since $P_k^- \geq m_n^{\mathcal{X}*}(\{\xi_k\})$, this implies that $(P_k^- - m_n^{\mathcal{X}*}(\{\xi_k\})) \xrightarrow{P} 0$ for all k, or, equivalently, $m_n^{\mathcal{X}*}(\{\xi_k\}) \xrightarrow{P} P_k^-$, for all k. Since $P_k^- \xrightarrow{P} p_k$, we thus have $m_n^{\mathcal{X}*}(\{\xi_k\}) \xrightarrow{P} p_k$, for all k, which implies that $m_n^{\mathcal{X}*}(A) \xrightarrow{P} \mathbb{P}_X(A)$ for all $A \subseteq \mathcal{X}$. □

Tables

Table 1: Goodman simultaneous confidence intervals for Example 1, at confidence level $1 - \alpha = 0.95$.

Diagnosis	N_k/n	P_k^-	P_k^+
Neurotic	0.41	0.33	0.50
Depressed	0.22	0.16	0.30
Schizophrenic	0.17	0.11	0.24
Personality disorder	0.20	0.14	0.27

Table 2: Lower probabilities induced by the confidence intervals of Table 1, and corresponding Möbius inverse. The ξ_k are the four mental diseases, in the order in which they appear in Table 1.

A	$P^-(A)$	$m^-(A)$
$\{\xi_1\}$	0.33	0.33
$\{\xi_2\}$	0.16	0.16
$\{\xi_1, \xi_2\}$	0.50	0
$\{\xi_3\}$	0.11	0.11
$\{\xi_1, \xi_3\}$	0.45	0
$\{\xi_2, \xi_3\}$	0.28	0
$\{\xi_1, \xi_2, \xi_3\}$	0.73	0.12
$\{\xi_4\}$	0.14	0.14
$\{\xi_1, \xi_4\}$	0.47	0
$\{\xi_2, \xi_4\}$	0.30	0
$\{\xi_1, \xi_2, \xi_4\}$	0.76	0.13
$\{\xi_3, \xi_4\}$	0.25	0
$\{\xi_1, \xi_3, \xi_4\}$	0.70	0.11
$\{\xi_2, \xi_3, \xi_4\}$	0.50	0.090
$\mathcal{X}$	1	-0.20

Table 3: Belief and mass functions, at confidence level 0.95, for the data of Example 1. The solution $(m^{\mathcal{X}*}(A), bel^{\mathcal{X}*}(A))$ was obtained using a linear program solver. The solution $(m^{\mathcal{X}\dagger}(A), bel^{\mathcal{X}\dagger}(A))$ was computed using the IRM algorithm.

A	$P^-(A)$	$bel^{\mathcal{X}*}(A)$	$m^{\mathcal{X}*}(A)$	$bel^{\mathcal{X}\dagger}(A)$	$m^{\mathcal{X}\dagger}(A)$
$\{\xi_1\}$	0.33	0.33	0.33	0.33	0.33
$\{\xi_2\}$	0.16	0.14	0.14	0.16	0.16
$\{\xi_1, \xi_2\}$	0.50	0.50	0.021	0.50	0
$\{\xi_3\}$	0.11	0.097	0.097	0.11	0.11
$\{\xi_1, \xi_3\}$	0.45	0.45	0.020	0.45	0
$\{\xi_2, \xi_3\}$	0.28	0.28	0.036	0.28	0
$\{\xi_1, \xi_2, \xi_3\}$	0.73	0.69	0.040	0.68	0.067
$\{\xi_4\}$	0.14	0.12	0.12	0.14	0.14
$\{\xi_1, \xi_4\}$	0.47	0.47	0.02	0.47	0
$\{\xi_2, \xi_4\}$	0.30	0.30	0.035	0.30	0
$\{\xi_1, \xi_2, \xi_4\}$	0.76	0.72	0.045	0.70	0.072
$\{\xi_3, \xi_4\}$	0.25	0.25	0.035	0.25	0
$\{\xi_1, \xi_3, \xi_4\}$	0.70	0.66	0.038	0.65	0.064
$\{\xi_2, \xi_3, \xi_4\}$	0.50	0.48	0.019	0.46	0.050
$\mathcal{X}$	1	1	0	1	0

Table 4: Arizona January precipitation data, with simultaneous 95 % confidence intervals. The bounds of the class intervals are in inches.

class ξ_k	N_k	N_k/n	P_k^-	P_k^+
< 0.75	48	0.44	0.32	0.56
$[0.75, 1.25)$	17	0.15	0.085	0.27
$[1.25, 1.75)$	19	0.17	0.098	0.29
$[1.75, 2.25)$	11	0.10	0.047	0.20
$[2.25, 2.75)$	6	0.055	0.020	0.14
≥ 2.75	9	0.082	0.035	0.18

Table 5: Basic belief masses for the precipitation data, using the representation explained in Figure 5. Masses are given to intervals $A_{k,r} = \{\xi_k, \ldots, \xi_r\}$ with $r \geq k$. Each cell at the intersection of row k and colums r contains $m(A_{k,r})$.

	1	2	3	4	5	6
1	0.32	0	0	0.13	0.11	0
2	-	0.085	0	0	0.012	0.14
3	-	-	0.098	0	0	0
4	-	-	-	0.047	0	0
5	-	-	-	-	0.020	0
6	-	-	-	-	-	0.035

Figures

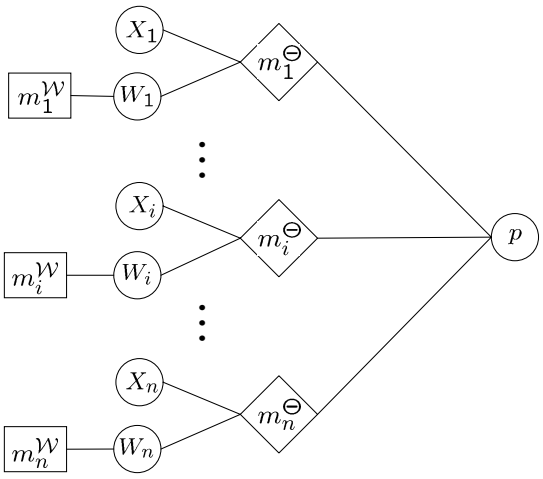

Figure 1: Graphical model for binomial inference in Dempster's approach.

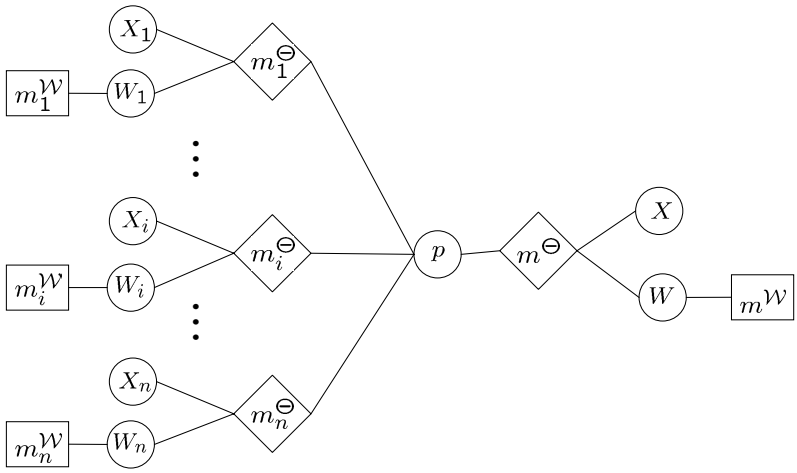

Figure 2: Graphical model for binomial inference in Dempster's approach, with one additional variable X to be predicted.

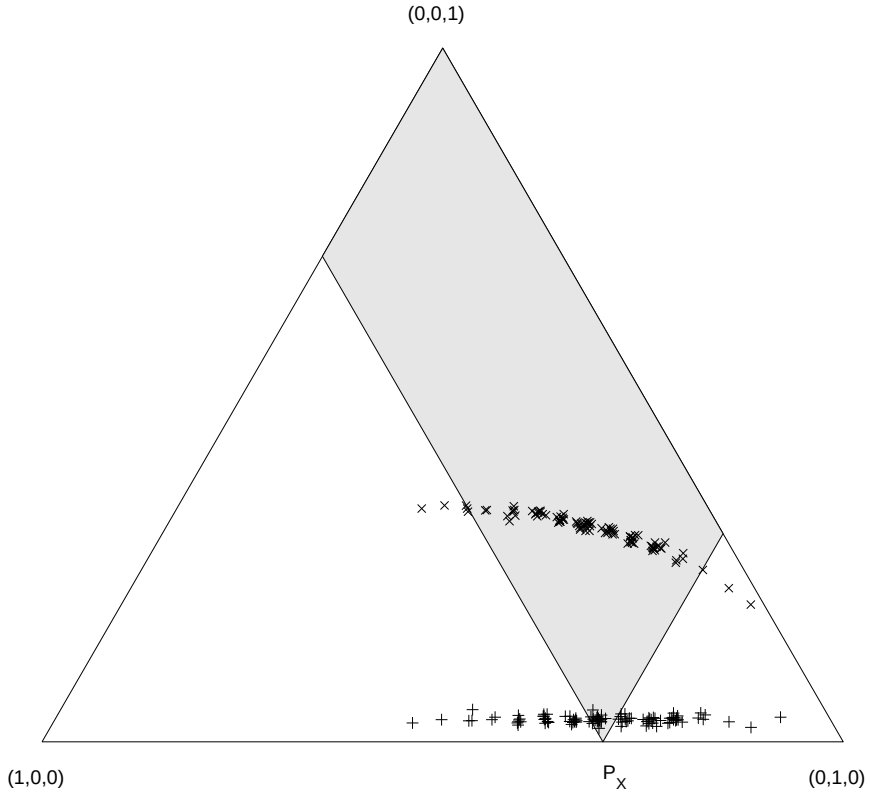

Figure 3: Results of the experiment of Example 3 with $n = 30$: predictive belief functions at confidence level $1 - \alpha = 0.95$ ($\times$), and belief functions computed using Dempster's method ($+$). Each belief function is represented as a point in barycentric coordinates, with the lower left corner corresponding to $\{\xi_1\}$, the lower right corner corresponding to $\{\xi_2\}$, and the upper corner corresponding to $\mathcal{X}$. Some random noise was added to avoid the superposition of points corresponding to the same value of X.

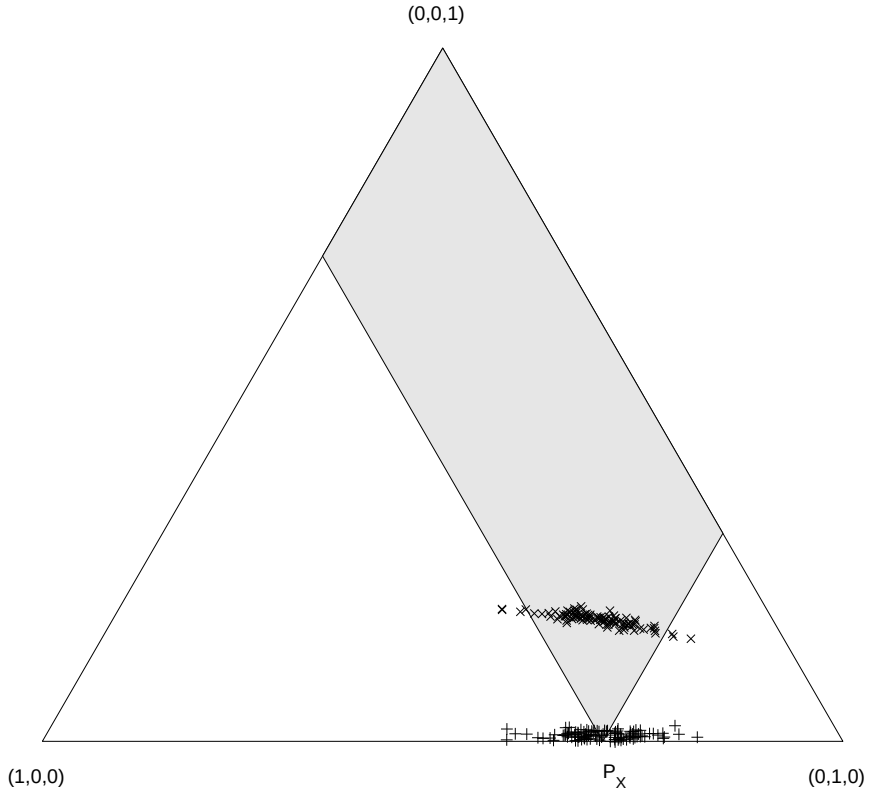

Figure 4: Results of the experiment of Example 3 with $n = 100$: predictive belief functions at confidence level $1 - \alpha = 0.95$ ($\times$), and belief functions computed using Dempster's method ($+$). Each belief function is represented as a point in barycentric coordinates, with the lower left corner corresponding to $\{\xi_1\}$, the lower right corner corresponding to $\{\xi_2\}$, and the upper corner corresponding to $\mathcal{X}$. Some random noise was added to avoid the superposition of points corresponding to the same value of X.

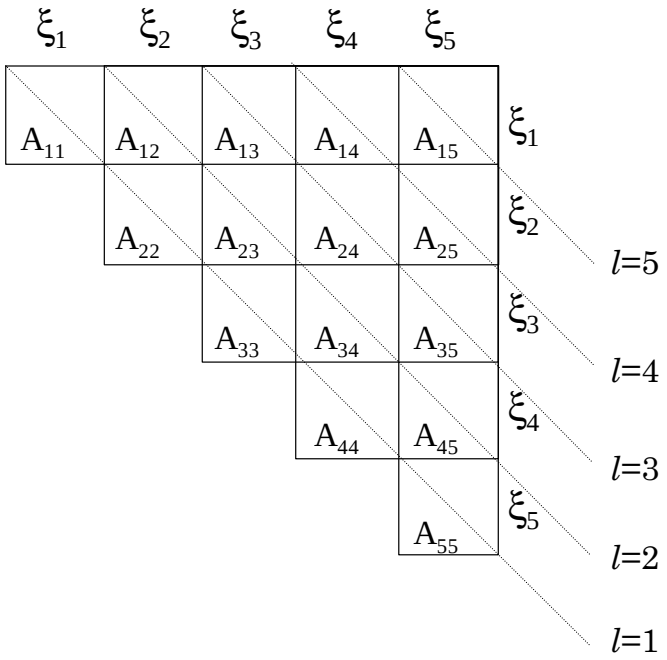

Figure 5: Representation of intervals of $\mathcal{X} = \{\xi_1, \ldots, \xi_K\}$, with $K = 5$. Each cell at the intersection of row k and column r corresponds to interval $A_{k,r} = \{\xi_k, \ldots, \xi_r\}$. The singletons are located on the main diagonal, the intervals of length 2 on the second upper diagonal, etc. The frame $A_{1,K} = \mathcal{X}$ corresponds to the upper right corner.

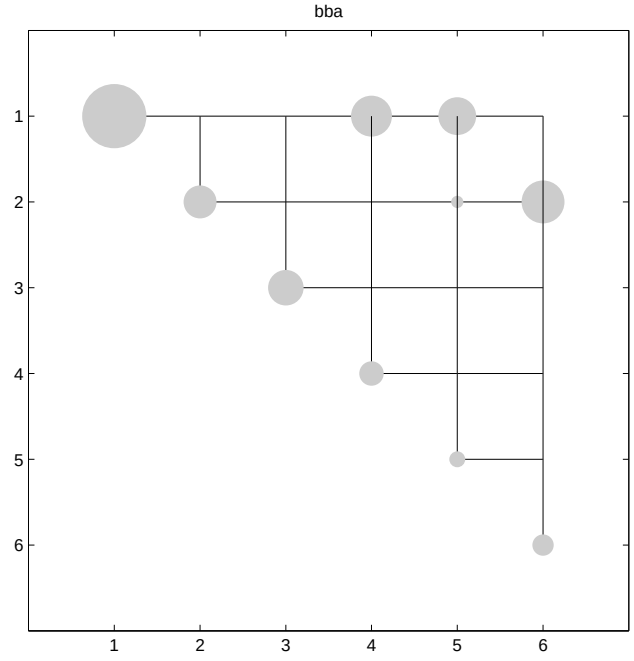

Figure 6: Graphical representation of the basic belief assignment given in Table 5. Each mass is proportional to the area of the circle.

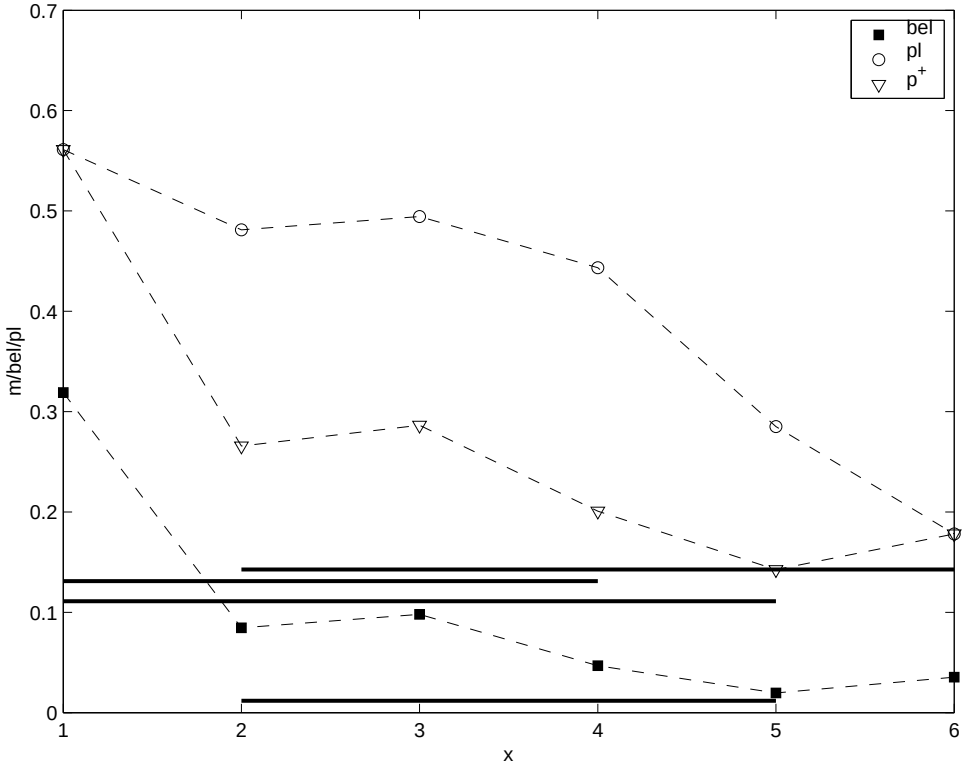

Figure 7: Alternative representation of the basic belief assignment given in Table 5 and in Figure 6. Each mass given to a singleton is represented by a filled square, and each mass given to an interval $A_{k,\ell}$ is represented by a horizontal line ranging from ξ_k to ξ_ℓ. The circles and the triangles represent, respectively, the plausibilities and the upper probabilities of the singletons.