A Maximum Likelihood Approach to Inference Under Coarse Data Based on Minimax Regret

Romain Guillaume, Didier Dubois

▶ **To cite this version:**

Romain Guillaume, Didier Dubois. A Maximum Likelihood Approach to Inference Under Coarse Data Based on Minimax Regret. International Conference on Soft Methods in Probability and Statistics (SMPS 2018), Sep 2018, Compiègne, France. pp.99-106, 10.1007/978-3-319-97547-4_14 . hal-02181935

HAL Id: hal-02181935
https://hal.science/hal-02181935

Submitted on 12 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A maximum likelihood approach to inference under coarse data based on minimax regret

Romain Guillaume, Didier Dubois

IRIT, CNRS and Université de Toulouse, 118 Route de Narbonne, F-31062 Toulouse Cedex 9,France
Romain.Guillaume@irit.fr,dubois@irit.fr

Abstract. Various methods have been proposed to express and solve maximum likelihood problems with incomplete data. In some of these approaches, the idea is that incompleteness makes the likelihood func-tion imprecise. Two proposals can be found to cope with this situation: maximize the maximal likelihood induced by precise datasets compat-ible with the incomplete observations, or maximize the minimal such likelihood. These approaches prove to be extremist, the maximax ap-proach having a tendency to disambiguate the data, while the maximin approach favors uniform distributions. In this paper we propose an al-ternative approach consisting in minimax relative regret criterion with respect to maximal likelihood solutions obtained for all precise datasets compatible with the coarse data. It uses relative likelihood and seems to achieve a trade-off between the maximax and the maximin methods.

1 Introduction

Maximum likelihood is a standard approach to finding a probabilistic model based on data. It maximizes the probability of obtaining the observations (supposed to belong to a set of mutually exclusive outcomes) [3]. When observations are coarse and may overlap, it is not clear how to define the likelihood function: several options exist, according to whether we take into account the measurement process governing the incompleteness or not. In this paper, we focus on optimizing the likelihood function that we should have observed, had observations been precise. Due to incomplete observations, this likelihood function is imprecise, since there are several possible precise datasets compatible with the coarse observations. Two approaches have been proposed: one considers an optimistic point of view aiming to disambiguate the data, by maximizing the maximum likelihood value across candidate datasets [5]. Another more cautious one tries to maximize the minimum likelihood value across candidate datasets thus adopting a robust optimization approach [6]. Both approaches have their weaknesses and can be criticized as being extreme ones, yielding either too deterministic or too dispersed distribution functions.

In this paper we propose an alternative criterion that tries to define a trade-off between the two previous approaches, and can be seen as minimizing a maximal regret criterion. We provide the results of some preliminary investigations

of this approach, first by considering examples in the discrete setting, such as ill-observed coin flipping. Optimizing this criterion seems to pose challenging computational problems.

2 Definition of the problem

We consider the setting of a random experiment where a variable X is incompletely observed via a measurement device providing values of a variable Y, namely [1]:

- $X : \Omega \to \mathcal{X}$ represents the outcome of a certain random experiment. In the finite case we assume $\mathcal{X} = \{a_1, \ldots, a_m\}$.
- $Y : \Omega \to \mathcal{Y} \subseteq \wp(\mathcal{X})$ that models the reports of the measurement device, where $\mathcal{Y} = \{A_1, \ldots, A_r\}$, $\wp(\mathcal{X})$ is the set of subsets of $\mathcal{X}$, and $A_i \subseteq \mathcal{X}$.

The information about the joint probability distribution $P(X, Y)$ on $\mathcal{X} \times \mathcal{Y}$ of the random vector (X, Y) can be obtained by modeling the random variable X ($P(X)$) and its measurement process ($P(Y|X)$). However, in this paper, we shall just ignore the measurement process and try to figure out what is the best choice for $P(X)$ based on the available data. In general, this probability function depends on a model parameter θ, and we write it $P(X|\theta)$.

Let $\mathbf{y} = (G_1, \ldots G_N)$ be a sample of coarse observations, where $G_j \in \mathcal{Y}$ denote the results of observing X through the measurement device $Y = G_j$ means that $x_j \in G_j$, where x_j is the jth (unobserved) outcome of in the sample $\mathbf{x} = (x_1, \ldots x_N)$ of the random process governing X. Let $\mathcal{X}^{\mathbf{y}}$ be the set of samples of X compatible with $\mathbf{y}$. We may consider three different likelihood functions depending on whether we refer to [1]:

1. the observed sample in $\mathcal{Y}$: $P(\mathbf{y}|\theta) = \prod_{i=1}^{N} p(y_i|\theta)$.
2. the hidden sample in $\mathcal{X}$: $P(\mathbf{x}|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$.
3. the complete sample in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$: $L^{\mathbf{z}}(\theta) = \prod_{i=1}^{N} p(z_i|\theta)$, where $z_i = (x_i, G_i), \mathbf{z} = ((x_1, G_1), \ldots (x_N, G_N)), x_j \in G_j, \forall j = 1, \ldots, N$.

In the following we focus on the likelihood function based on the hidden sample $\mathbf{x}$. Clearly this likelihood function is ill-known and belongs to the set $\{P(\mathbf{x}|\theta) : \mathbf{x} \in \mathcal{X}^{\mathbf{y}}\}$. There are two strategies of likelihood maximization, based on a sequence of imprecise observations $\mathbf{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$:

1. The maximax strategy : it aims at finding $(\mathbf{x}^*, \theta^*)$ that maximizes $\overline{\Lambda}(\theta) = \max_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}} P(\mathbf{x}|\theta)$.
2. The maximin strategy : it aims at finding $\theta_* \in \Theta$ that maximizes $\underline{\Lambda}(\theta) = \min_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}} P(\mathbf{x}|\theta)$.

The above setting for inference with incomplete data differs from the older, more standard approach by Heijan and Rubin [8], which relies on an extensive use of partitions and a different view of the measurement process: an observation G_j is viewed the unique element, such that $x_j \in G_j$, of a partition of $\mathcal{X}$ that is

selected at random. Here we rather adopt the framework of Dempster [2] where the defect in the measurement process is modelled by means of a multimapping from the sample space to the outcome space $\mathcal{X}$.

These optimization problems take the following form in the discrete case where $\mathcal{X}$ is finite [7], assuming exchangeability. Let n_k be the number of appearances of value a_k in the virtual sample $\mathbf{x}$, and q_j be the number of observations of A_j in the observed sample $\mathbf{y}$, and let n_{ik} be the number of times (a_i, A_k) is obtained in the complete sample $\mathbf{z}$; we have that $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$ if and only:

$$\begin{cases} (a) & \sum_{k=1}^{r} n_k = \sum_{j=1}^{r} q_j = N \\ (b) & n_k = \sum_{j=1}^{r} n_{kj}, \forall k = 1, \ldots, m \\ (c) & q_j = \sum_{k=1}^{m} n_{kj}, \forall j = 1, \ldots, r. \\ (d) & n_{kj} = 0 \text{ if } a_k \notin A_j, \forall k, j. \end{cases} \tag{1}$$

Let $\mathcal{N}^{\mathbf{y}}$ be the set of tuples $\mathbf{n} = (n_1, \ldots, n_m)$ that are compatible with $\mathbf{y}$, namely they satisfy (1). The maximax, resp. maximin, strategy then takes the following form, using log-likelihoods, and defining $p_k^\theta = P(X = a_k|\theta)$:

$$\begin{array}{ll} & \max_{\mathbf{p}} \max_{\mathbf{n}} \sum_{k=1}^{m} n_k \cdot \log p_k^\theta \\ & \text{or} \\ & \max_{\mathbf{p}} \min_{\mathbf{n}} \sum_{k=1}^{m} n_k \cdot \log p_k^\theta \\ & \text{s.t. constraints (1) hold and} \\ (e) & \sum_{k=1}^{m} p_k^\theta = 1 \\ (f) & n_k, n_{jk} \in \mathbb{N}, p_k^\theta > 0, \qquad \forall k = 1, \ldots, m, \end{array}$$

In the finite case, when all discrete distributions are possible ($\theta = (p_1, \ldots p_{m-1})$), the set of probabilistic models is the credal set associated to the belief function defined by the mass assignment $m(A_j) = q_j/N$, for $j = 1, \ldots, r$. Then the optimal solution to the maximin likelihood problem is the distribution with maximal entropy, namely the solution to: $\max_{\mathbf{n}} -\sum_{k=1}^{m} \frac{n_k}{N} \cdot \log \frac{n_k}{N}$ under conditions (a, b, c), and $n_k \in \mathbb{R}^+$, i.e. $\mathbf{n}$ in the convex hull of $\mathcal{N}^{\mathbf{y}}$ [7, 1].

In contrast, the optimal solution to the maximax likelihood problem (2) is the solution with minimal entropy, namely the solution to: $\min_{\mathbf{n}} -\sum_{k=1}^{m} \frac{n_k}{N} \cdot \log \frac{n_k}{N}$ under conditions (a, b, c) [7, 1].

As a consequence the two approaches to handling incomplete observations look somewhat extreme. On the one hand, the max-max approach tends to strongly disambiguate the data, yielding Dirac distributions consistent with the coarse data when they are feasible. If the measured quantity is deterministic in nature, this is natural. However it becomes questionable if the measured quantity is tainted with variability (like the ill-known overlapping observed outcomes of tossing a die). On the other hand, the max-min approach yields distributions with high variances interpreting incomplete information as the result of extreme randomness: the less information the larger the variance, which is not fully satisfactory either.

3 A criterion based on a likelihood ratio trade-off

In this section we propose a new criterion which tries to maximize the confidence level in the fact that the parameters are acceptable for all possible realizations. It is well-known that the likelihood function represents the plausibility of parameter θ in view of some precise results $\mathbf{x}$ of observing a variable X, in relative value only. Namely the likelihood of θ is proportional to $P(\mathbf{x}|\theta)$ but the proportionality coefficient is arbitrary. We cannot evaluate the extent to which data $\mathbf{x}$ supports θ more that data $\mathbf{x}'$ supports θ' by comparing $P(\mathbf{x}|\theta)$ and $P'(\mathbf{x}'|\theta')$ We must compare the likelihood ratio $\frac{P(\mathbf{x}|\theta)}{P(\mathbf{x}|\theta')}$ with $\frac{P(\mathbf{x}'|\theta)}{P(\mathbf{x}'|\theta')}$. In other words, only likelihood ratios can be used to choose the right parameter value.

The new criterion we propose consists in comparing the likelihood ratios $\frac{P(\mathbf{x}|\theta)}{P(\mathbf{x}|\hat{\theta}_\mathbf{x})}$ across observations $\mathbf{x} \in \mathcal{X}^\mathbf{y}$, where $\hat{\theta}_\mathbf{x}$ is the maximum likelihood estimate of the distribution parameter θ for observation $\mathbf{x}$, and adopt a variant of the minmax regret approach (we could call minmax *relative* regret) of the form

$$\max_{\theta\in\Theta} \min_{\mathbf{x}\in\mathcal{X}^\mathbf{y}} \frac{P(\mathbf{x}|\theta)}{P(\mathbf{x}|\hat{\theta}_\mathbf{x})}. \tag{2}$$

The idea is that since the ideal parameter value θ for observation $\mathbf{x}$ is $\hat{\theta}_\mathbf{x}$, and X is ill-observed, we try to find the value of θ that reaches a best compromise between the various ideal values $\hat{\theta}_\mathbf{x}$ for all $\mathbf{x}$ in agreement with the incomplete data $\mathbf{y}$. The value $\frac{P(\mathbf{x}|\theta)}{P(\mathbf{x}|\hat{\theta}_\mathbf{x})}$ lies in $[0, 1]$, and can be viewed as the degree of membership of θ to the fuzzy set $\tilde{\theta}_\mathbf{x}$ of good parameter estimates based on observing $\mathbf{x}$. Then the problem (2) is a standard fuzzy multicriteria optimisation problem (finding θ with the maximal membership values in the intersection of fuzzy sets $\tilde{\theta}_\mathbf{x}$). Note that the problem (2) is very similar to the one induced by the maximin strategy. The latter can be seen as a fuzzy optimisation problem, albeit with non normalized fuzzy sets.

Using the log-likelihood $L(\mathbf{x}|\theta) = \log P(\mathbf{x}|\theta)$, the above problem can be formulated as one of minimizing the maximal regret:

$$\min_{\theta\in\Theta} \max_{\mathbf{x}\in\mathcal{X}^\mathbf{y}} L(\mathbf{x}|\hat{\theta}_\mathbf{x}) - L(\mathbf{x}|\theta) \tag{3}$$

Or yet, in the finite case ($\mathcal{X}^\mathbf{y} = \{\mathbf{x}^1, \ldots, \mathbf{x}^h\}$), it can be expressed as a goal programming problem, minimizing the L^∞ distance $\max_{i=1}^h |L(\mathbf{x}^i|\hat{\theta}_{\mathbf{x}^i}) - L(\mathbf{x}^i|\theta)|$ between the vector $(\theta, \ldots, \theta)$ with length h and the ideal point $(\hat{\theta}_{\mathbf{x}^1}, \ldots, \hat{\theta}_{\mathbf{x}^h})$.

Remark An alternative choice of probabilistic model in the presence of coarse data consists in choosing the pignistic probability of Smets [10]. The observation vector $(q_1, \ldots, q_r)$ on $\mathcal{Y}$ can be viewed as a Dempster-Shafer mass assignment m on $\wp(\mathcal{X})$ [9], letting $m(A_j) = q_j/N$ for $j = 1, \ldots r$ inducing lower probabilities in the sense of [2] in the form of a belief function $Bel(A) = \sum_{E\subset A} m(A)$. This Dempster-Shafer mass assignment defines a convex set $\{P_X : \bar{P}_X(A) \geq Bel(A), \forall A \subseteq \mathcal{X}\}$ of probabilities on $\mathcal{X}$, hence of joint probabilities on $\mathcal{X} \times \mathcal{Y}$ with known marginals $\hat{q}_j = q_j/N$ for $j = 1, \ldots, r$ on $\mathcal{Y}$. Knowing the distribution

on $\mathcal{Y}$ (which can be estimated from $\mathbf{y}$), the pignistic probability induced by m can be obtained as:

$$p_k^P = \sum_{j:A_j \ni a_k} \frac{q_j}{N \cdot |A_j|}.$$

where $|A_j|$ is the cardinality of A_j, which yields a marginal distribution on $\mathcal{X}$.

4 Resolution method

In this section we propose a mathematical programming approach to solving problem (3) in the discrete case. Let $c(\mathcal{N}^{\mathbf{y}})$ be the convex hull of $\mathcal{N}^{\mathbf{y}}$. Elements of $c(\mathcal{N}^y)$ are denoted by $\mathbf{w} = (w_1, \dots w_m)$. To build this model, we show (see Prop.1) that the maximization part of Problem (3) can be reduced to maximization over the set of vertices of $c(\mathcal{N}^y)$, denoted by $V(\mathcal{N}^{\mathbf{y}})$, that actually lies in $\mathcal{N}^{\mathbf{y}}$. The efficiency of the approach presupposes $V(\mathcal{N}^{\mathbf{y}})$ is small enough. Note that $L(\mathbf{x}|\theta) = \sum_{k=1}^m n_k \cdot \log p_k^\theta$, which can be extended to $c(\mathcal{N}^y)$, in the form $L(\mathbf{w}|\theta) = \sum_{k=1}^m w_k \cdot \log p_k^\theta$, where $\mathbf{w}$ is a vector of non-negative reals.

Proposition 1 $\max_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}}} L(\mathbf{x}|\hat{\theta}_{\mathbf{x}}) - L(\mathbf{x}|\theta) = \max_{\mathbf{w}\in c(\mathcal{N}^{\mathbf{y}})} L(\mathbf{w}|\hat{\theta}_{\mathbf{w}}) - L(\mathbf{w}|\theta) = \max_{\boldsymbol{n}\in V(\mathcal{N}^{\mathbf{y}})} L(\boldsymbol{n}|\hat{\theta}_{\boldsymbol{n}}) - L(\boldsymbol{n}|\theta)$.

Proof: $L(\mathbf{x}|\hat{\theta}_{\mathbf{x}}) - L(\mathbf{x}|\theta) = \sum_{k=1}^m n_k \cdot \log n_k - \sum_{k=1}^m n_k \cdot \log p_k^\theta$ with $(n_1, \dots n_m) \in \mathcal{N}^{\mathbf{y}}$, the term $\log p_k^\theta$'s being constant. Consider maximizing $L(\mathbf{w}|\hat{\theta}_{\mathbf{x}}) - L(\mathbf{w}|\theta)$ for $\mathbf{w} \in c(\mathcal{N}^y)$ instead. The function $\sum_{k=1}^m w_k \cdot \log w_k$ is convex, $\sum_{k=1}^m w_k \cdot \log p_k^\theta$ is convex, and clearly $L(\mathbf{w}|\hat{\theta}_{\mathbf{x}}) - L(\mathbf{w}|\theta)$ is convex too. We also know that $c(\mathcal{N}^{\mathbf{y}})$ is a convex polyhedron with integer vertices, so the maximal value of $L(\mathbf{w}|\hat{\theta}_{\mathbf{x}}) - L(\mathbf{w}|\theta)$ is reached at one of these vertices, i.e., for some $\mathbf{w} = \mathbf{n} \in V(\mathcal{N}^{\mathbf{y}}) \subseteq \mathcal{N}^{\mathbf{y}}$ corresponding to some sample $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$. □

From Proposition 1, we can solve the maximin relative regret using the following mathematical programming model, introducing an additional variable $\alpha \geq 0$:

$$\min_{\mathbf{p}} \alpha \tag{4}$$

s.t.

(a) $\sum_{k=1}^m n_k \cdot \log n_k - \sum_{k=1}^m n_k \cdot \log p_k^\theta \leq \alpha, \forall \mathbf{x} \in V(\mathcal{X}^{\mathbf{y}})$

(b) $\sum_{k=1}^m p_k^\theta = 1$

(c) $p_k^\theta > 0, \qquad \forall k = 1, \dots, m,$

This model has an exponential number of constraints (a). As perspective, we intend to develop an iterative algorithm to compute the optimal probability distribution.

5 Examples

In this section, we discuss the differences between the maximin and maximax approaches and the new one proposed in this paper by means of examples.

Let us suppose that you want to estimate the probability to observe heads or tails in a coin flipping experiment, where $\mathcal{X} = \{h, t\}$. But the only precise observations made yielded heads and we could not see the other outcomes: so we have $\mathcal{Y} = \{\{h\}, \{h, t\}\}$. Suppose that we observed 30 times $\{h\}$ and 30 times nothing ($\{h, t\}$). To estimate the minmax regret probability distribution on $\mathcal{X}$ (noted p^R) we solved the mathematical formulation (4) given in the previous section using the solver SQP of software Octave.[1] To discuss the result, let us compare it with the probability distribution obtained using the maximax approach (denoted by p^M) and the maximin approach (denoted by p^m) and Smets' pignistic probability distribution [10] (denoted by p^P) induced by the belief function whose family of focal sets is $\mathcal{Y}$. The results are given in table 1.

$\mathcal{X}$	$\{h\}$	$\{t\}$
$p^M(X = a_i) =$	1	0
$p^m(X = a_i) =$	0.5	0.5
$p^R(X = a_i) =$	0.8	0.2
$p^P(X = a_i) =$	0.75	0.25

Table 1. Probability distribution with 30 times $\{h\}$ and 30 times $\{h, t\}$

Firstly we can see that the solution p^R lies between the maximax and the maximin solutions. The maximax one assumes that the observation $\{h, t\}$ are $\{h\}$, in contrast with the maximin solution, which assumes that the outcomes behind $\{h, t\}$ are $\{t\}$. The criterion proposed in this paper achieves a trade-off between the two extreme samples S1: (30 $\{h\}$, 30 $\{t\}$) and S2: (60 $\{h\}$, 0 $\{t\}$) compatible with the observations. It selects parameter values which are the least incompatible with both samples, here $p_h = 0.8$ and $p_t = 0.2$. Noted that it is different from the pignistic distribution here, $p_h^P = 0.75$ and $p_t^P = 0.25$.

number of $\{h\}$	60	50	40	30	20	10	0
number of $\{h, t\}$	0	10	20	30	40	50	60
$p^M(X = h) =$	1	1	1	1	1	1	1 or 0
$p^m(X = h) =$	1	$\frac{5}{6}$	$\frac{4}{6}$	0.5	0.5	0.5	0.5
$p^R(X = h) \approx$	1	0.937	0.87	0.8	0.72	0.63	0.5
$p^P(X = h) \approx$	1	0.92	0.83	0.75	0.66	0.58	0.5

Table 2. Impact of imprecision on the optimal parameter

[1] https://www.gnu.org/software/octave/

Let us now study, on this simple example, the impact of the imprecise data on the optimal parameter found by the 4 methods. This is provided on Table 2. The extreme behavior of maximin and maximax methods is patent. The maximax approach disambiguates the data assuming the coin always yield heads. If it is known that the measured process is random (a regular coin flipping experiment), this approach sounds totally counter-intuitive. However if the (ill-)observed process is known to be deterministic (e.g., it consists in repetitively reading h on the visible side of a coin that is not flipped), concluding h with probability 1, if it has been observed at least once, is natural. In contrast, the maximin approach (like the other ones) interprets the lack of precise observations as pure randomness, which may sound questionable both when the underlying phenomenon is known to be deterministic, and when it is not. When no outcome can be observed, the maximax approach expresses pure ignorance, while the uniform distribution found by other approaches is an instance of Laplace principle of insufficient reason.

Observe that the maximin regret approach yields a smoother variation of the optimal value, in terms of the amount of dataset imprecision than the maximin approach (the latter produces a uniform distribution as soon as it can). The same smooth behavior is observed for the pignistic probability. However, the latter is the center of gravity of the set of probability assignments $\mathbf{p} = (p_1, \ldots, p_m)$ such that $N\mathbf{p} \in c(\mathcal{N}^{\mathbf{x}})$, hence based on Euclidean distance, while the minmax regret uses an L^∞ distance.

Let us analyze the value of likelihood ratio in (2) for optimal parameters p^R obtained by model (4), as provided on Table 3. This ratio measures our confidence in the obtained model, i.e., the extent to which parameter $\theta = (p_h^R, p_t^R)$ is acceptable for all possible samples in $\mathcal{N}^{\mathbf{x}}$.

number of $\{h\}$	60	50	40	30	20	10	0
number of $\{h,t\}$	0	10	20	30	40	50	60
θ	(1,0)	(0.937,0.063)	(0.87,0.13)	(0.8,0.2)	(0.72,0.28)	(0.63,0.37)	(0.5,0.5)
$\min_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}}} \frac{P(\mathbf{x}\|\theta)}{P(\mathbf{x}\|\hat{\theta}_{\mathbf{x}})}$	1	0.68	0.44	0.26	0.14	0.06	0.015

Table 3. Confidence on the parameter θ

We can see that when the imprecision of the dataset increases, the extent to which the optimal parameter is acceptable for all possible samples decreases. In other words, when all observations are imprecise we are bound to choose $(0.5, 0.5)$ but we know that our confidence on this choice is low.

6 Conclusion

This paper is a preliminary step towards a more robust approach to statistical estimation in the presence of coarse data. This approach is more compatible with the usual view of likelihood functions as defined up to multiplicative constants,

while the maximax and maximin approaches compare absolute likelihood values, albeit on equal datasets. Rather than selecting a single probabilistic model we could also use our criterion to build a small range of suboptimal parameter values that guarantee a given confidence level. Another issue is to compare our approach with more traditional ones to coarse data using partitions [8]. In these approaches, it is assumed that the imprecision generation process stems from the random choice of a coarsening of $\mathcal{X}$ (a partition) such that if $x = a$ occurs, the corresponding element of the random partition is observed. In contrast, our setting, based on [1], relies on the multimapping representation of incomplete information due to Dempster [2]. The latter framework sounds more natural, while the partition-based framework seems to require the specification of more parameters (there are more partitions of a finite set of size m than non-empty subsets thereof, when m is large enough). In other words, if no specific knowledge is available on the measurement process, there are more parameters to be set in the partition-based framework, than the number of conditional probabilities defining $P(Y|X)$.

References

1. I. Couso, D. Dubois: A general framework for maximizing likelihood under incomplete data. Int. J. Approx. Reasoning 93: 238-260 (2018)
2. A. P. Dempster, Upper and Lower Probabilities Induced by a Multivalued Mapping, Ann. Math. Statist. 38: 325-339 (1967).
3. A.W.F. Edwards, Likelihood, Cambridge University Press (1972).
4. D.F. Heitjan, D. B. Rubin, Ignorability and coarse data, Annals of Statistics 19: 2244-2253 (1991).
5. E. Hüllermeier, Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization, Int. J. Approx. Reasoning 55: 1519-1534 (2014).
6. R. Guillaume, D. Dubois, Robust parameter estimation of density functions under fuzzy interval observations, 9th ISIPTA Symposium, Pescara, Italy, 147-156 (2015).
7. R. Guillaume, I. Couso, D. Dubois: Maximum Likelihood with Coarse Data based on Robust Optimisation, 10th ISIPTA Symposium, Lugano, Switzerland, 169-180 (2017)
8. G. Shafer, A Mathematical Theory of Evidence, Princeton University Press (1976).
9. P. Smets. Constructing the pignistic probability function in a context of uncertainty, Uncertainty in Artificial Intelligence 5 (Henrion M. et al., Eds.), North-Holland, Amsterdam, 29-39 (1990).