Maximum Likelihood Under Incomplete Information: Toward a Comparison of Criteria

Inés Couso and Didier Dubois

Abstract Maximum likelihood is a standard approach to computing a probability distribution that best fits a given dataset. However, when datasets are incomplete or contain imprecise data, depending on the purpose, a major issue is to properly define the likelihood function to be maximized. This paper compares several proposals in terms of their intuitive appeal, showing their anomalous behavior on examples.

1 Introduction

Edwards ([6], p. 9) defines a likelihood function as being proportional to the probability of obtaining results given a hypothesis, according to a probability model. A fundamental axiom is that the probability of obtaining at least one among two results is the sum of the probabilities of obtaining each of these results. In particular, a *result* in the sense of Edwards is not any kind of event, it is an elementary event. Only elementary events can be observed. For instance, when tossing a die, and seeing the outcome, you cannot observe the event "odd", you can only see 1, 3 or 5. If this point of view is accepted, what becomes of the likelihood function under incomplete or imprecise observations? To properly answer this question, one must understand what is a result in this context. Namely, if we are interested in a certain random phenomenon, observations we get in this case do not directly inform us about the underlying random variables. Due to the interference with an imperfect measurement process, observations will be set-valued. So, in order to properly exploit such incomplete information, we must first decide what to model:

1. the random phenomenon *through* its measurement process;
2. or the random phenomenon *despite* its measurement process.

I. Couso (✉)
Department of Statistics, Universidad de Oviedo, Oviedo, Spain
e-mail: couso@uniovi.es

D. Dubois
IRIT, Université Paul Sabatier, Toulouse, France
e-mail: dubois@irit.fr

M.B. Ferraro et al. (eds.), *Soft Methods for Data Science*, Advances in Intelligent Systems and Computing 456, DOI 10.1007/978-3-319-42972-4_18

In the first case, imprecise observations are considered as results, and we can construct the likelihood function of a random set, whose realizations contain precise but ill-known realizations of the random variable of interest. Actually, most authors are interested in the other point of view, consider that outcomes are the precise, although ill-observed, realizations of the random phenomenon. However in this case there are as many likelihood functions as precise datasets in agreement with the imprecise observations. Authors have proposed several ways of addressing this issue. The most traditional approach is based on the EM algorithm and it comes down to constructing a fake sample of the ill-observed random variable in agreement with the imprecise data, and maximizing likelihood wrt this sample. In this paper we analyze this methodology in the light of the epistemic approach to statistical reasoning outlined in [1] and compare it with several recent proposals by Denoeux [5], Hüllermeier [8], and Guillaume [7]. Note that in this paper we do not consider the issue of imprecision due to too small a number of precise observations (see for instance Serrurier and Prade [10]).

2 The Random Phenomenon and Its Measurement Process

Let the random variable $X : \Omega \to \mathcal{X}$ represent the outcome of a certain random experiment. For the sake of simplicity, let us assume that $\mathcal{X} = \{a_1, \ldots, a_m\}$ is finite. Suppose that there is a measurement tool that provides an incomplete report of observations. Namely, the measurement tool reports information items $\Gamma(\omega) = B \in 2^{\mathcal{X}}$, for some multimapping $\Gamma : \Omega \to 2^{\mathcal{X}}$, which represents our (imprecise) perception of X, in the sense that we assume that X is a selection of Γ, i.e. $X(\omega) \in \Gamma(\omega),\ \forall\, \omega \in \Omega$ [3]. Let $\mathcal{G} = Im(\Gamma) = \{A_1, \ldots, A_r\}$ denote the image of Γ (the collection of possible outcomes).

We overview below two different ways to represent the information about the joint distribution of the random vector (X, Γ).

The imprecision generation standpoint. Here, we emphasize the outcome of the experiment X and the "imprecisiation" process that leads us to just get imprecise observations of X, Let us consider the following matrix: $(M|\mathbf{p})$, where M is called the mixing matrix with terms:

- $a_{jk} = p_{j|k.} = P(\Gamma = A_j | X = a_k)$ denotes the (conditional) probability of observing A_j if the true outcome is a_k and
- $p_{k.} = P(X = a_k)$ denotes the probability that the true outcome is a_k.

Such a matrix determines the joint probability distribution modeling the underlying generating process plus the connection between true realizations and incomplete observations. Some examples and their characterizing matrices are as follows:

- **Partition** [4]. Suppose that $Im(\Gamma) = \{A_1, \ldots, A_r\}$ forms a partition of $\mathcal{X}$. Therefore, we can easily observe that the probabilities $P(\Gamma = A_j | X = a_k) = 1$ if $a_k \in A_j$ and 0 otherwise, forall j, k.

- **Superset assumption** [9]. $Im(\Gamma)$ coincides with $2^{\mathcal{X}} \setminus \{\emptyset\}$. For each $k = 1, \ldots, m$ there is a constant c_k such that $P(\Gamma = A_j | X = a_k) = c_k$, if $A_j \ni a_k$ ($P(\Gamma = A_j | X = a_k)) = 0$, otherwise.) Furthermore, for every $k \in \{1, \ldots, m\}$ there are 2^{m-1} subsets of $\mathcal{X}$ that contain it. Therefore the constant is equal to $1/2^{m-1}$, i.e.: $P(\Gamma = A_j | X = a_k) = \begin{cases} 1/2^{m-1} & if A_j \ni a_k \\ 0 & otherwise. \end{cases}$ This is a kind of missing-at-random assumption, whereby the imprecisiation process is completely random. It is often presented as capturing the idea of "lack of information" about this process, which sounds questionable.

The disambiguation standpoint. We can alternatively characterize the joint probability distribution of (X, Γ) by means of the marginal distribution of Γ (the mass assignment $m(A_j) = P(\Gamma = A_j) = p_{\cdot j},\ j = 1, \ldots, r$ of a belief function describing imprecise observations [3]) and the conditional probability of each result $X = a_k$, knowing that the observation was $\Gamma(\omega) = A_j$, for every $j = 1, \ldots, r$. The new matrix $(M'|\mathbf{p}')$ can be written as follows:

- $b_{kj} = p_{k.|j} = P(X = a_k | \Gamma = A_j|)$ denotes the (conditional) probability that the true value of X is a_k if we have been reported that it belongs to A_j
- $p_{\cdot j} = P(Y = A_j) = P(\Gamma = A_j)$ denotes the probability that the generation plus the measurement processes lead us to observe A_j.

Such a matrix determines the joint probability distribution modeling the underlying generating process plus the connection between true outcomes and incomplete observations. (More specifically, the vector $(p_{\cdot 1}, \ldots, p_{\cdot r})^T$ characterizes the observation process while the matrix $B = (p_{k.|j})_{k=1,\ldots,m;j=1,\ldots,r}$ represents the conditional probability of X (true outcome) given Γ (observation). Here is an example:

- **Uniform conditional distribution** Under the uniform conditional distribution, the (marginal) probability P_X induced by X is the pignistic transform [11] of the belief measure associated to the mass assignment m. The conditional distribution is given by: $p_{k.|j} = \frac{1}{\#A_j}$, if $a_k \in A_j$ and 0 otherwise. And the marginal distribution is: $p_{k.} = \sum_{j:A_j \ni a_k} \frac{1}{\#A_j} p_{\cdot j}$.

3 Different Likelihood Functions

Both matrices $M = (A|\mathbf{p})$ and $M' = (B|\mathbf{p}')$ univocally characterize the joint distribution of (X, Γ). For each pair $(k, j) \in \{1, \ldots, m\} \times \{1, \ldots, r\}$, let p_{kj} denote the joint probability $p_{kj} = P(X = a_k, \Gamma = A_j)$. According to the nomenclature used in the preceding subsections, the respective marginals on $\mathcal{X}$ and $\mathcal{G}$ are denoted as follows:

- $p_{\cdot j} = \sum_{k=1}^{m} p_{kj}$ will denote the mass of $\Gamma = A_j$, for each $j = 1, \ldots, r$, and
- $p_{k.} = P(X = a_k) = \sum_{j=1}^{r} p_{kj}$ will denote the mass of $X = a_k$, for every k.

Now, let us assume that the above joint distribution is characterized by means of a (vector of) parameter(s) $\theta \in \Theta$ (in the sense that M and M' can be written as functions of θ). We naturally assume that the number of components of θ is less than or equal to the dimension of both matrices, i.e., it is less than or equal to the minimum $\min\{m \times (r+1), r(m+1)\}$. In other words, the approach uses a parametric model such that a value of θ determines a joint distribution on $\mathcal{X} \times Im(\Gamma)$.

For a sequence of N iid copies of $Z = (X, \Gamma)$, $\mathbf{Z} = ((X_1, \Gamma_1), \ldots, (X_N, \Gamma_N))$, we denote by $\mathbf{z} = ((x_1, G_1), \ldots, (x_N, G_N)) \in (\mathcal{X} \times \mathcal{G})^N$ a specific sample of the vector (X, Γ). Thus, $\mathbf{G} = (G_1, \ldots, G_N)$ will denote the observed sample (an observation of the set-valued vector $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_n)$), and $\mathbf{x} = (x_1, \ldots, x_N)$ will denote an arbitrary artificial sample from $\mathcal{X}$ for the unobservable latent variable X, that we shall vary in $\mathcal{X}^N$. The samples $\mathbf{x}$ are chosen such that the number of repetitions n^{kj} of each pair $(a_k, A_j) \in \mathcal{X} \times \mathcal{G}$ in the sample are in agreement with the numbers $n_{.j}$ of observations A_j. We denote by $\mathcal{X}^{\mathbf{G}}$ (resp. $\mathcal{Z}^{\mathbf{G}}$), the set of samples $\mathbf{x}$ (resp. complete joint samples $\mathbf{z}$) respecting this condition.We may consider three different log-likelihood functions depending on whether we refer to

- the observed sample: $L^{\mathbf{G}}(\theta) = \log \mathbf{p}(\mathbf{G}; \theta) = \log \prod_{i=1}^{N} p(G_i; \theta)$. It also writes $= \sum_{j=1}^{r} n_{.j} \log p_{.j}^{\theta}$. where $n_{.j}$ denotes the number of repetitions of A_j in the sample of size N
- the (ill-observed) sample of outcomes: $L^{\mathbf{x}}(\theta) = \log \mathbf{p}(\mathbf{x}, \theta)$. It also writes $\log \prod_{i=1}^{N} p(x_i; \theta) = \sum_{k=1}^{m} n_{k.} \log p_{k.}^{\theta}$, where $n_{k.}$ denotes the number of occurrences of a_k in the sample $\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}^{\mathbf{G}}$.
- the complete sample: $L^{\mathbf{z}}(\theta) = \log \mathbf{p}(\mathbf{z}, \theta) = \log \prod_{i=1}^{N} p(z_i; \theta)$. It also writes $\sum_{k=1}^{m} \sum_{j=1}^{r} n_{kj} \log p_{kj}^{\theta}$ where $n_{kj} = \sum_{i=1}^{N} 1_{\{(a_k, A_j)\}}(x_i, G_i)$ denotes the number of repetitions of the pair (a_k, A_j) in the sample (i.e., $\mathbf{z} \in \mathcal{Z}^{\mathbf{G}}$).

In the sequel, we compare some existing strategies of likelihood maximization, based on a sequence of imprecise observations $\mathbf{G} = (G_1, \ldots, G_N) \in \mathcal{G}^N$:

- The standard maximum likelihood estimation (MLE) : it computes the argument of the maximum of $L^{\mathbf{G}}$ considered as a mapping defined on Θ, i.e.: $\hat{\theta} = \arg\max_{\theta \in \Theta} L^{\mathbf{G}}(\theta) = \arg\max_{\theta \in \Theta} \prod_{j=1}^{r} (p_{.j}^{\theta})^{n_{.j}}$. The result is a mass assignment on $2^{\mathcal{X}}$. For instance, the EM algorithm [4] is an iterative technique using a latent variable X to achieve a local maximum of $L^{\mathbf{G}}$.
- The maximax strategy [8]: it aims at finding the pair $(\mathbf{x}^*, \theta^*) \in \mathcal{X}^{\mathbf{G}} \times \Omega$ that maximizes $L^{\mathbf{z}}(\theta)$, i.e.: $(\mathbf{x}^*, \theta^*) = \arg\max_{\mathbf{x} \in \mathcal{X}^{\mathbf{G}}, \theta \in \Theta} L^{\mathbf{z}}(\theta)$, i.e., $\arg\max_{\mathbf{x} \in \mathcal{X}^{\mathbf{G}}, \theta \in \Theta} \prod_{k=1}^{m} \prod_{j=1}^{r} (p_{kj}^{\theta})^{n_{kj}}$.
- The maximin strategy [7]: it aims at finding $\theta_* \in \Theta$ that maximizes $L^{-}(\theta) = \min_{\mathbf{x} \in \mathcal{X}^{\mathbf{G}}} L^{\mathbf{z}}(\theta) = \min_{\mathbf{x} \in \mathcal{X}^{\mathbf{G}}} \sum_{k=1}^{m} \sum_{j=1}^{r} n_{kj} \log p_{kj}^{\theta}$. It is a robust approach that also identifies a fake optimal sample $\mathbf{x}_*$.
- The Evidential EM strategy [5]: It assumes that the data set is uncertain and defined by a mass-function over $2^{\mathcal{X}^N}$. Under the particular situation where it has a single focal element $B \subset \mathcal{X}^N$, with mass $m(B) = 1$, the EEM approach considers the following expression as a likelihood function, given such imprecise data (see

Eq. 16 in [5]): $\mathbf{p}(B;\theta) = P((X_1,\ldots,X_N) \in B;\theta)$. The Evidential EM algorithm is viewed as a variation of the classical EM algorithm in order to select a value of θ that maximizes the "likelihood" $\mathbf{p}(B;\theta)$. In particular, if we assume that B is a Cartesian product of the sets in the collection $\{A_1,\ldots,A_r\}$ the criterion can be alternatively written as follows: $\mathbf{p}(B;\theta) = \prod_{j=1}^{m} P_\theta(X \in A_j)^{n_j}$. The EEM procedure may not coincide with a maximum likelihood estimation since this criterion is not always in the spirit of a likelihood function, as seen later on. The EM algorithm uses it when the imprecise data forms a partition.

Under some particular conditions about the matrices M and M', some of the above likelihood maximization procedures may coincide or not. In the rest of the paper we provide some examples, focusing on the optimal samples $\mathbf{z} \in \mathcal{Z}^{\mathbf{G}}$ or $\mathbf{x} \in \mathcal{X}^{\mathbf{G}}$ computed by the methods and that are supposed to disambiguate the imprecise data. Indeed most existing techniques end up with computing a probability distribution on $\mathcal{X}$ or a fake sample achieving an imputation of X.

4 A Comparison of Estimation-Disambiguation Methods

Let us to compare the potentials and limitations of these approaches. Here we just give a few hints by means of examples.

EM-based approaches. Let $\mathcal{P}^{\mathcal{X}^N}$ be the set of all probability measures P we can define on the measurable space $(\mathcal{X}^N, \wp(\mathcal{X}^N))$. The EM algorithm [4] tries to maximize the function $F : \mathcal{P}^{\mathcal{X}^N} \times \Theta \to \mathbb{R}$: $F(\mathbf{P},\theta) = L^{\mathbf{G}}(\theta) - D(\mathbf{P},\theta)$, $\forall P \in \mathcal{P}^{\mathcal{X}^N}$, $\theta \in \Theta$, where $\mathbf{p}(\mathbf{x}|\mathbf{G};\theta) = \frac{\mathbf{p}(\mathbf{x},\mathbf{G};\theta)}{\mathbf{p}(\mathbf{G};\theta)}$, whenever $\mathbf{p}(\mathbf{G};\theta) > 0$. Moreover, $D(\mathbf{P},\mathbf{P}')$ is the Kullback-Leibler divergence from $\mathbf{P}'$ to $\mathbf{P}$, $\sum_{\mathbf{x}\in\mathcal{X}^N} \mathbf{p}(\mathbf{x}) \log[\frac{\mathbf{p}(\mathbf{x})}{\mathbf{p}'(\mathbf{x})}]$, where $\mathbf{p}$ is the mass function associated to $\mathbf{P}$. It is then clear that $L^{\mathbf{G}}(\theta) \geq F(\mathbf{P},\theta)$ and that if $\mathbf{P} = \mathbf{P}(\cdot|\mathbf{G};\theta)$, then $F(\mathbf{P},\theta) = L^{\mathbf{G}}(\theta)$. Given a value $\theta^{(n-1)}$ obtained at the $n-1$ M-step, the E-step actually computes $\mathbf{P}(\cdot|\mathbf{G};\theta^{(n-1)})$ (which is basically like determining a fake sample $\mathbf{z} \in \mathcal{Z}^{\mathbf{G}}$), and the next M step finds a value of θ that maximizes $F(\mathbf{P}(\cdot|\mathbf{G};\theta^{(n-1)}),\theta)$, i.e. $L^{\mathbf{G}}(\theta)$ based on the fake sample $\mathbf{z}$. In fact, the EM algorithm iteratively finds a parametric probability model $\mathbf{P}^\theta$ and a probability distribution $\mathbf{P}(\cdot|\mathbf{G};\theta)$ on $\mathcal{X}$, that is in agreement with the data $\mathbf{G}$, such that the divergence from $\mathbf{P}^\theta$ to $\mathbf{P}(\cdot|\mathbf{G};\theta)$ is minimal [2]. $\mathbf{P}^\theta$ is an MLE for the fake sample $\mathbf{z} \in \mathcal{Z}^{\mathbf{G}}$ in agreement with $\mathbf{P}(\cdot|\mathbf{G};\theta)$, which yields the best imputation of X in this sense. There are situations where the result of the EM algorithm will be questionable [2].

Example 1 Suppose that a dice is tossed and let X pertain to the result of the trial. The probability distribution of X is a vector $(p_1,\ldots,p_6) \in [0,1]^6$, with $\sum_{i=1}^{6} p_i = 1$. Suppose after each trial we are told either that the result has been less than or equal to 3 (A_1) or greater than or equal to 3 (A_2). After each toss, when the actual result (X) is 3, the reporter needs to decide A_1 or A_2. Assume the conditional probability $P(G_n = A_1|X_n = 3)$ is a fixed number $\alpha \in [0,1]$ for every trial, $n = 1,\ldots,N$. Suppose that

we toss the dice $N = 1000$ times and the report tells us $n_{.1} = 300$ times that the result was less than or equal to 3. Let θ denote the vector $(p_1, p_2, p_3, p_4, p_5; \alpha)$. The likelihood function based on the observed sample $\mathbf{G}$ can be written as: $L^{\mathbf{G}}(\theta) = (p_1 + p_2 + \alpha p_3)^{300} \cdot [1 - (p_1 + p_2 + \alpha p_3)]^{700}$. Such a function is maximized for any vector θ satisfying the constraint $p_1 + p_2 + \alpha p_3 = 0.3$. If we use the EM algorithm, we get a vector θ satisfying the above constraints after the first iteration of the M algorithm. We will get a different vector $\theta^{(1)}$, depending on the initial point $\theta^{(0)}$. If we start from $\theta^{(0)} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}; \frac{1}{2})$, we get $\theta^{(1)} = (0.12, 0.12, 0.16, 0.2, 0.2; \frac{3}{8})$. It is also the MLE of θ based on a (fake) sample of 1000 tosses of the dice where the number of repetitions of each of the six facets has been respectively 120, 120, 160, 200, 200, 200. But this is not the only MLE based on the observed sample.

Evidential EM Algorithm. We can distinguish the following cases:

- The case where $Im(\Gamma)$ forms a partition of $\mathcal{X}$. In this case, $P(X \in A_j) = P(Y = A_j) = p_{.j}$, $\forall j = 1, \ldots, r$, and therefore $\prod_{j=1}^{r} P(X \in A_j; \theta)^{n_{.j}}$ coincides with the likelihood $\mathbf{p}(\mathbf{G}; \theta)$.
- The case where the sets $A_1, \ldots, A_r$ do overlap. In this case, $\mathbf{p}(\mathbf{G}; \theta)$ and $\mathbf{p}(B; \theta)$ do not necessarily coincide, as shown in the following example.

Example 2 Let us take a sample of N tosses of the dice in Example 1 and let us assume that the reporter has told us that n_1 of the times the result was less than or equal to 3, and $n_2 = N - n_1$ otherwise. The EEM likelihood is $\mathbf{p}(B; \theta) = (p_1 + p_2 + p_3)^{n_1} \cdot (p_3 + p_4 + p_5 + p_6)^{n_2}$ with $\sum_{i=1}^{6} p_i = 1$. We can easily observe that it reaches its maximum ($\mathbf{p}(B; \theta) = 1$) for any vector θ satisfying the constraint $p_3 = 1$. But such a prediction of θ would not be a reasonable estimate for θ.

The maximax approach. The parametric estimation based on the maximax approach does not coincide in general with the MLE. Furthermore, it may lead to questionable imputations of X.

Example 3 Let us suppose that a dice is tossed $N = 10$ times, and that Peter reports 4 heads, 2 tails and he does not tell whether there was heads or tails for the remaining 4 times. Let us consider the parameter $\theta = (p, \alpha, \beta)$, where $p = P(X = h)$, $\alpha = P(\Gamma = \{h, t\} | X = h)$ and $\beta = P(\Gamma = \{h, t\} | X = t)$. It determines the following joint probability distribution induced by (X, Γ): $P(h, \{h\}) = (1 - \alpha)p$; $P(h, \{h, t\}) = \alpha p$; $P(t, \{t\}) = (1 - \beta)p$; $P(t, \{h, t\}) = \beta p$; and 0 otherwise.

The MLE of θ is not unique. It corresponds to all the vectors $\theta = (p, \alpha, \beta) \in [0, 1]^3$ satisfying the constraints: $(1 - \alpha)p = 0.4$ and $(1 - \beta)(1 - p) = 0.2$, indicating the marginal probabilities $P(\Gamma = \{h\})$ and $P(\Gamma = \{t\})$ respectively.

In contrast, the maximax strategy seeks for a pair $(\theta^*; \mathbf{x}^*) = (p^*, \alpha^*, \beta^*; \mathbf{x}^*)$ that maximizes $L^{\mathbf{z}}(\theta)$. It can be checked that the tuple that maximizes $L^{\mathbf{z}}(\theta)$ is unique. It corresponds to the vector of parameters $\theta^* = (p^*, \alpha^*, \beta^*) = (0.8, 0.5, 0)$ and the sample where all the unknown outcomes are heads. In words, the maximax strategy assumes that all the ill-observed results correspond to the most frequent observed outcome ("heads"). Accordingly, the estimation of the probability of heads is the

corresponding frequency (0.8). According to this strategy, and without having any insight about the behaviour of Peter, we predict that each time he refused to report, the result was in fact "heads".

Example 4 Let us now consider the situation about the coin described in Example 3, and let us suppose in addition that the following conditions hold: $\alpha = 1 - \alpha = 0.5$ and $\beta = 1 - \beta = 0.5$. In words, no matter what the true outcome is (heads or tails) Peter refuses to give any information about it with probability 0.5 (the behavior of Peter does not depend on the true outcome). This is the "superset assumption" [8] already mentioned. Under this additional constraint, the MLE of $\theta = (p, 0.5, 0.5)$ is reached at $\hat{p} = 4/6 = 2/3$. The maximum likelihood estimator provides the same estimation as if we had just tossed the coin six times, since, as a consequence of the superset assumption here, the four remaining tosses play no role in the statistics. As a result, the conditional probability $P(X = h|\Gamma = \{h, t\})$ is assumed to coincide with $P(X = h|\Gamma \neq \{h, t\})$ and with $P(X = h) = p$. Such a probability is estimated from the six observed outcomes, where four of them were "heads" and the rest were "tails". In contrast, the maximax strategy without the superset assumption leads us to take into account the unobserved tosses as matching the most frequent observed outcome, hence the imputation of X is compatible with a data set containing 8 heads and only 2 tails.

The maximin approach. Consider again Example 3. The maximin approach consists of considering all log-likelihood functions $L_k^{\mathbf{x}}(p) = (4 + k)\log p + (6 - k)\log(1 - p)$ with $0 \leq k \leq 4$. The approach consists in finding for each value of p the complete data that minimizes $L^{\mathbf{x}}(p)$. Since $L_k^{\mathbf{x}}(p)$ is of the form $k \log \frac{p}{(1-p)} + a$, it is easy to see that if $p < 1/2$, the minimum $L^-(p)$ is reached for $k = 4$, and if $p > 1/2$, it is reached for $k = 0$. So, it is $8 \log p + 2 \log(1 - p)$ if $p < 1/2$ and $4 \log p + 6 \log(1 - p)$ otherwise. So $L^-(p)$ is increasing when $p < 1/2$ and decreasing when $p > 1/2$. It reaches its maximum for $p = 1/2$. So the maximin approach is cautious in the sense of maximizing entropy in the coin-tossing experiment. It yields the uniform distribution, i.e., an imputation of 5 heads and 5 tails, in agreement with the observations.

5 Conclusion

This paper suggests that it is not trivial to extend MLE methods to incomplete data despite the existence of several proposals. In particular, it is very questionable to reconstruct distributions for unobserved variables when parameters of distributions that generate them are not closely connected to parameters of distributions that govern observed ones. In contrast, the famous EM article [4] deals with imprecise observations forming a partition and starts with an example in which a single parameter determines the joint distribution of X and Γ. However, it is not straightforward to adapt the EM procedure to incomplete overlapping data. In the general case, either

one applies standard MLE to observed imprecise data only (yielding a mass function) or one has to add an assumption that comes down to selecting a single probability measure in the credal set induced by this mass function. Each approach to imprecise data MLE proposes its own assumption. As can be seen from the examples, it is easy to find cases where these methods lead to debatable solutions: the solution to the EM algorithm [4] depends on the initial parameter value, the EEM approach [5] seems to optimize a criterion that sometimes does not qualify as a genuine likelihood function, the maximax approach [8] may select a very unbalanced distribution for the hidden variable, while the maximin robust MLE [7] favors uninformative distributions. More work is needed to characterize classes of problems where one estimation method is justified and the other method fails.

Acknowledgments This work is partially supported by ANR-11-LABX-0040-CIMI within the program ANR-11-IDEX-0002-02, by TIN2014-56967-R (Spanish Ministry of Science and Innovation) and FC-15-GRUPIN14-073 (Regional Ministry of the Principality of Asturias).

References

1. Couso I, Dubois D (2014) Statistical reasoning with set-valued information: Ontic vs. epistemic views. Int J Approximate Reasoning 55(7):1502–1518
2. Couso I, Dubois D (2016) Belief revision and the EM algorithm. Proc, IPMU
3. Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 38:325–339
4. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc B 39:1–38
5. Denoeux T (2013) Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Trans Knowl Data Eng 26:119–130
6. Edwards AWF (1972) Likelihood. Cambridge University Press
7. Guillaume R, Dubois D (2015) Robust parameter estimation of density functions under fuzzy interval observations, 9th ISIPTA Symposium. Pescara, Italy
8. Hüllermeier E (2014) Learning from imprecise and fuzzy observations. Int J Approximate Reasoning 55(7):1519–1534
9. Hüllermeier E, Cheng W (2015) Superset learning based on generalized loss minimization. ECML/PKDD 2:260–275
10. Serrurier M, Prade H (2013) An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data. Int J Approximate Reasoning 54(7):919–933
11. Smets P (2005) Decision making in the TBM: the necessity of the pignistic transformation. Int J Approximate Reasoning 38:133–147