Maximum Likelihood with Coarse Data based on Robust Optimisation

Romain Guillaume ROMAIN.GUILLAUME@IRIT.FR
IRIT, University of Toulouse, Toulouse (France)

Inés Couso COUSO@UNIOVI.ES
Universidad de Oviedo, Gijon (Spain)

Didier Dubois DUBOIS@IRIT.FR
IRIT, CNRS and University of Toulouse, Toulouse (France)

Abstract

This paper deals with the problem of probability estimation in the context of coarse data. Probabilities are estimated using the maximum likelihood principle. Our approach presupposes that each imprecise observation underlies a precise one, and that the uncertainty that pervades its observation is epistemic, rather than representing noise. As a consequence, the likelihood function of the ill-observed sample is set-valued. In this paper, we apply a robust optimization method to find a safe plausible estimate of the probabilities of elementary events on finite state spaces. More precisely we use a maximin criterion on the imprecise likelihood function. We show that there is a close connection between the robust maximum likelihood strategy and the maximization of entropy among empirical distributions compatible with the incomplete data. A mathematical model in terms of maximal flow on graphs, based on duality theory, is proposed. It results in a linear objective function and convex constraints. This result is somewhat surprizing since maximum entropy problems are known to be complex due to the maximization of a concave function on a convex set.

Keywords: maximum likelihood; incomplete information; robust optimization; entropy.

1. Introduction

Interval observations, and more generally, set-valued ones, do not always reflect the same phenomenon (Couso and Dubois, 2014). Sets, e.g. intervals, may either represent exact observations of items taking the form of sets (for instance, the daily min-max temperature ranges across one year), or, on the contrary, imprecise observations of precise quantities. In the later case, we speak of coarse data (Heitjan and Rubin, 1991). In the first situation, set data are a special kind of functional data where observations lie in a space of characteristic functions equipped with a suitable metric structure, enabling precise statistical parameters to be derived, e.g., (González-Rodríguez et al., 2012). In this paper we are interested in the statistical analysis of data when observations are imprecise, or coarse, more specifically, when we only know that the precise values of observations are restricted by sets of possible outcomes of a random variable of interest. In this kind of representation, sets model epistemic states (or states of knowledge) in the sense that no value outside the set is possibly the true observed value (unreachable for the observer). Under the epistemic approach, the expected value and the variance of a collection of intervals are themselves intervals (Kruse and Meyer, 2012).

This paper addresses the problem of statistical inference in the presence of epistemic set-valued data using the maximum likelihood principle. Under imprecise observations, the likelihood function itself becomes imprecisely appraised too and is thus set-valued. There are several possible ways of defining a scalar likelihood function in this situation (Couso and Dubois, 2016a). In this paper we adopt a robust optimisation point of view and maximize the lower bound of the imprecise likelihood

function, with a view to obtain a probability density that accounts for the potential variability of the random variable observed via sets of possible outcomes. We give an interpretation of the robust solution in terms of entropy maximization, and propose algorithms for computing robust maximum likelihood distributions in the discrete (finite) case of coarse nominal data, based on a maximal flow approach.

The paper is organized as follows. In Section 2, we recall a general framework for maximum likelihood estimation under coarse data due to Couso and Dubois (2016a), and situate our robust optimization strategy in this framework. It consists in maximizing the minimal likelihood function in agreement with the coarse data. We discuss the difference between our approach and the optimistic maximax strategy. Section 3 proposes a methodology for solving the robust optimisation problem in the discrete case, based on max-flow formulation and duality. Section 4 shows that the optimal estimate corresponds to maximizing entropy among empirical distributions of all possible samples in agreement with the coarse data. In section 5, we propose a new method for solving the maximin likelihood estimation problem and discuss an illustrative example.

2. General framework

A likelihood function is proportional to the probability of obtaining the observed data given a hypothesis, according to a probability model. Observed data are considered as outcomes, i.e., elementary events. If this point of view is accepted, what becomes of the likelihood function under coarse observations? If coarse observations are considered as results, we can construct the likelihood function for set-valued outcomes, and compute a random set. However, coarse observations being set-valued, they do not directly inform us about the underlying random variable. In order to properly exploit such incomplete information, we must first decide what to model (Couso and Dubois, 2016a): (1) the random phenomenon *despite* the deficiencies its measurement process; or (2) the random phenomenon *as known via* its measurement process.

In the first case, authors have proposed several ways of restoring a distribution for the underlying random phenomenon. The most traditional approach constructs a virtual sample of the ill-observed random variable in agreement with the imprecise data, by minimizing divergence from a parametric model, and maximizing likelihood wrt this sample, so as to update this parametric model. This is often carried out by means of EM algorithm (Dempster et al., 1977). The problem with this approach is that there may be several optimal distributions, hence virtual samples, especially when the connection between the hidden random variable and its observation process is loose (Couso and Dubois, 2016b). The result of an iterative algorithm such as EM may depend on the initial parameter value. Moreover the EM algorithm assumes that observed coarse data form a partition of the domain of the random variable of interest (see the introduction of (Dempster et al., 1977)).

In this paper, we take the other point of view, the one of ill-observed outcomes. Then, there are as many likelihood functions as precise datasets in agreement with the coarse observations, and it is not clear which one to maximize. We pursue our study of a methodology based on a robust maximin optimisation approach applied to a set-valued likelihood Guillaume and Dubois (2015). Note that here, we do not consider the issue of modelling imprecision due to too small a number of precise observations (see for instance (Serrurier and Prade, 2013)). Let us recall the formal setting for statistics with coarse data proposed by Couso and Dubois (2016a), then we study the meaning of the solution to the maximin approach, and finally propose an algorithm to solve it in the case of nominal outcome sets.

2.1 The random phenomenon and its measurement process

Let a 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 driven by a random variable Y that provides an incomplete report of observations. Namely, there is a set-valued random variable $Y : \Omega \to \wp(\mathcal{X})$ that models the reports of a measurement device, where $\wp(\mathcal{X})$ is the set of subsets of $\mathcal{X}$. Y is thus a multimapping which represents our (imprecise) perception of X, in the sense that we assume that X is a selection of Y, i.e. $X(\omega) \in Y(\omega)$, $\forall \omega \in \Omega$, in agreement with the setting of imprecise probabilities proposed by Dempster (1967). X is often called the latent variable. Let $\mathcal{Y} = \{A_1, \ldots, A_r\}$ denote the set of possible set-valued outcomes of Y, where $A_j \in \wp(\mathcal{X})$.

The information about the joint distribution of the random vector (X, Y) modeling the random variable X and its measurement process can be represented by a joint probability on $\mathcal{X} \times \mathcal{Y}$ defined by means of $m \times r$ coefficients $p_{ij} = P(X = a_i, Y = A_j)$. Some knowledge may be available about this probability matrix. For instance, in the case when $\mathcal{Y}$ is a partition of $\mathcal{X}$, we have

$$p_{ij} = P(Y = A_j | X = a_i) \cdot P(X = a_i) = \begin{cases} P(X = a_i) \text{ if } a_i \in A_j \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

Sometimes assumptions are made about the conditional probability $P(Y = A_j | X = a_i)$ describing the imprecise measurement process, like the superset assumption (Hüllermeier and Cheng, 2015) considering that $\mathcal{Y} = \wp(\mathcal{X})$ and stating that this probability is a constant c_i over all sets containing a_i, i.e. $c_i = 1/2^{m-1}$ that does not depend on i. Another less restrictive assumption is called "coarse at random" (CAR) whereby $P(Y = A_j | X = a_i)$ does not depend on the value $a_i \in A_j$ (Heitjan and Rubin, 1991). In this paper, we shall just ignore the measurement process.

2.2 Different likelihood functions

The respective marginals on $\mathcal{X}$ and $\mathcal{Y}$ are denoted as follows:

- $p_{\cdot j} = \sum_{k=1}^{m} p_{kj}$ denotes the mass of $Y = A_j$, $j = 1, \ldots, r$, and
- $p_{k\cdot} = \sum_{j=1}^{r} p_{kj}$ denotes the mass of $X = a_k$, $k = 1, \ldots, m$.

Now, let us assume that the above joint distribution is characterized by means of a (vector of) parameter(s) $\theta \in \Theta$ that determines a joint distribution on $\mathcal{X} \times \mathcal{Y}$.

For a sequence of N iid copies of $Z = (X, Y)$, $\mathbf{Z} = ((X_1, Y_1), \ldots, (X_N, Y_N))$, we denote by $\mathbf{z} = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$ a specific sample of the vector (X, Y). Thus, $\mathbf{y} = (y_1, \ldots, y_N)$ will denote the observed sample (an observation of the set-valued vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$, 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 $(a_k, A_j) \in \mathcal{X} \times \mathcal{Y}$ in the sample are in agreement with the number q_j of actual observations A_j. We denote by $\mathcal{X}^{\mathbf{y}}$ (resp. $\mathcal{Z}^{\mathbf{y}}$), the set of samples $\mathbf{x}$ (resp. complete joint samples $\mathbf{z}$) respecting this condition. We assume that the measurements are reliable in the sense that, observing $y = G \subseteq \mathcal{X}$, we can be sure that the actual outcome $X = x \in G$. If we let n_k be the number of appearances of a_k in the virtual sample $\mathbf{x}$, we have that any $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$

satisfies:

$$\begin{cases} \sum_{k=1,...,r} n_k = \sum_{j=1,...,r} q_j = N \\ n_k = \sum_{j=1}^{r} n_{kj}, \forall k = 1, ..., m \\ q_j = \sum_{k=1}^{m} n_{kj} \forall j = 1, ..., r. \\ n_{kj} = 0 \text{ if } a_k \notin A_j, \forall k, j. \end{cases} \tag{2}$$

For a complete sample $\mathbf{z}$ to be compatible with the observation $\mathbf{y}$, we have that any $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ satisfies:

$$\begin{cases} \sum_{k=1,...,r} \sum_{j=1,...,r} n_{kj} = N \\ q_j = \sum_{k=1}^{m} n_{kj}, \forall j = 1, ..., r. \\ n_{kj} = 0 \text{ if } a_k \notin A_j, \forall k, j. \end{cases} \tag{3}$$

As pointed out by Couso and Dubois (2016a), we may consider three different log-likelihood functions depending on whether we refer to

1. *the observed sample in* Y: $L^{\mathbf{y}}(\theta) = \log \prod_{i=1}^{N} p(y_i; \theta) = \sum_{j=1}^{r} q_j \log p^{\theta}_{.j}$.
2. *the hidden sample in* $\mathcal{X}$: $L^{\mathbf{x}}(\theta) = \log \prod_{i=1}^{N} p(x_i; \theta) = \sum_{k=1}^{m} n_{k.} \log p^{\theta}_{k.}$.
3. *the complete sample in* $\mathcal{X} \times \mathcal{Y}$: $L^{\mathbf{z}}(\theta) = \log \prod_{i=1}^{N} p(z_i; \theta) = \sum_{k=1}^{m} \sum_{j=1}^{r} n_{kj} \log p^{\theta}_{kj}$

The two last ones are ill-known. The choice of one likelihood function vs. another depends upon what problem we are interested to solve. Maximizing $L^{\mathbf{y}}(\theta)$ means that we are interested in modeling our perception of the random variable only. It is the standard maximum likelihood estimation (MLE) that computes the argument of the maximum of $L^{\mathbf{y}}$ considered as a mapping defined on Θ, i.e.: $\hat{\theta} = \arg\max_{\theta \in \Theta} L^{\mathbf{y}}(\theta) = \arg\max_{\theta \in \Theta} \prod_{j=1}^{r} (p^{\theta}_{.j})^{q_j}$. The result is a mass assignment on $2^{\mathcal{X}}$ if there is no constraint relating the distributions of X and Y via the parameter θ. It computes a belief function on $\mathcal{X}$ with focal sets in $\mathcal{Y}$.

The EM algorithm (Dempster et al., 1977) is an iterative technique maximizing this likelihood function via the use of the latent variable X and a virtual sample for X in order to achieve a local maximum of $L^{\mathbf{y}}$. This procedure makes sense if the observed sample $\mathbf{y}$ provides enough information on X (via suitable assumptions on the model parameters θ) to guarantee the convergence of the iterative procedure to a solution that minimizes the distance between the empirical distribution of the final virtual sample in agreement with $\mathbf{y}$, and the resulting parametric distribution on X (Couso and Dubois, 2016a).

Maximizing $L^{\mathbf{z}}(\theta)$ enables to take into account the knowledge we may have about the measurement process, and allows for a fine-grained modeling. On the contrary, maximizing $L^{\mathbf{x}}(\theta)$ means that we completely give up modeling the measurement process and try to extract information about X based on information about Y, assuming complete ignorance about the measurement process. The difficulty with $L^{\mathbf{z}}(\theta)$ and $L^{\mathbf{x}}(\theta)$ is that they are ill-known, namely we must consider for all values of θ, the sets $\mathbb{L}^{\mathbf{z}}(\theta) = \{L^{\mathbf{z}}(\theta) : \mathbf{z} \in \mathcal{Z}^{\mathbf{y}}\}$ and $\mathbb{L}^{\mathbf{x}}(\theta) = \{L^{\mathbf{x}}(\theta) : \mathbf{x} \in \mathcal{X}^{\mathbf{y}}\}$, respectively. In the paper we shall deal with $L^{\mathbf{x}}(\theta)$, i.e., try to find results independently of the measurement process.

Applying the maximum likelihood principle when the likelihood function is ill-known requires the choice of a representative likelihood function from $\mathbb{L}^{\mathbf{x}}(\theta)$. Obvious natural choices are $\overline{L}(\theta) = \max_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}} \mathbb{L}^{\mathbf{x}}(\theta)$ and $\underline{L}(\theta) = \min_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}} \mathbb{L}^{\mathbf{x}}(\theta)$.

On this basis, 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* (Hüllermeier, 2014): it aims at finding the pair $(\mathbf{x}^*, \theta^*) \in \mathcal{X}^{\mathbf{y}} \times \Omega$ that maximizes $L^{\mathbf{x}}(\theta)$. In other words, compute $(\mathbf{x}^*, \theta^*) = \arg\max_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}, \theta \in \Theta} L^{\mathbf{x}}(\theta)$.
2. *The maximin strategy* (Guillaume and Dubois, 2015): it aims at finding $\theta_* \in \Theta$ that maximizes $\underline{L}(\theta) = \min_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}} L^{\mathbf{x}}(\theta)$. It is a robust optimization approach that takes a pessimistic view on likelihood maximization.

The maximax strategy tries to disambiguate the coarse data by choosing a virtual sample $\mathbf{x}$ that makes the parametric model maximally in agreement with the data. In the case of the maximin strategy, it is pessimistic in the sense that it tends to select distributions with large variability as we shall show.

3. The robust approach to discrete probability estimation with coarse date

In this section, we try to estimate the probability $P(X = a_k), k = 1, ..., m$ when the reports of N observations of X are imperfect and take the form of an imprecise sample $\mathbf{y}$ containing q_j copies of subsets $A_j \in \mathcal{Y}$ of values of X, for $j = 1 \ldots r$. To determine the parameter we adopt the usual approach based on likelihood maximization, which in the case of precise observations takes the form:

$$maximize : L^{\mathbf{x}}(\theta) = \log p(\mathbf{x}; \theta) = \sum_{i=1}^{m} n_k \log p(X = a_k | \theta) \tag{4}$$

Note that in our context the numbers $n_k, k = 1, ..., m$ are ill-known, because we did not fully observe the outcomes. All we know is that $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$. Hence, the vector $\mathbf{n} = (n_k)_{k=1,...,m} \in \mathcal{N}^{\mathbf{y}}$, where $\mathcal{N}^{\mathbf{y}}$ is the set of possible statistics in agreement with the imprecise observations $\mathbf{y}$, that is, respecting equation (2). We call an assignment $\mathbf{n} \in \mathcal{N}^{\mathbf{y}}$ a virtual sample. To manage the uncertainty on $\mathbf{n}$ we use the pessimistic maxmin strategy. Namely, we will maximize the minimal value of likelihood function for the hidden sample $\mathbf{x}$:

$$\max_{\theta} \min_{\mathbf{n} \in \mathcal{N}^{\mathbf{y}}} \sum_{i=1}^{m} n_k \log p(X = a_k | \theta) \tag{5}$$

Note that by using the likelihood based on the hidden sample $\mathbf{x}$, we make no assumption on the measurement process that from observing $X \in \mathcal{X}$, yields a subset of $\mathcal{X}$. We only know that if $Y = A_j$ is observed, some $x_k \in A_j$ has been produced. One can see that Equation (5) is then equivalent to the more explicit mathematical formulation:

$$\max_{\mathbf{p}} \min_{\mathbf{n}} \textstyle\sum_{k=1,...,m} n_k \cdot \log p_k \tag{6}$$

$$\begin{aligned}
&s.t. \\
&(a) \quad \textstyle\sum_{k=1,...,m} n_k = \sum_{j=1,...,r} q_j = N \\
&(b) \quad n_k = \textstyle\sum_{j:(j,k) \in \mathbb{E}^{\mathbf{y}}} n_{j,k}, \qquad \forall k = 1, ..., m \\
&(c) \quad q_j = \textstyle\sum_{k:(j,k) \in \mathbb{E}^{\mathbf{y}}} n_{j,k}, \qquad \forall j = 1, ..., r \\
&(d) \quad \textstyle\sum_{k=1,...,m} p_k = 1 \\
&(e) \quad n_k, n_{j,k} \in \mathbb{N}^+, p_k > 0, \qquad \forall k = 1, ..., m,
\end{aligned}$$

where

- the value $q_j, j = 1, ..., r$ is the number of actual observations of Y of the form A_j,
- the decision variables $(p_k)_{k=1,...,m}$ stand for the ill-known model probabilities $p(X = a_k|\theta)$, $k = 1, ..., m$ on $\mathcal{X}$; in the loosest situation, there is no constraint relating the p_k via an explicit parameter θ.
- $(n_k)_{k=1,...,m}$ are the ill-known numbers of occurrences of values $a_k, k = 1, ..., m$ of X,
- $\mathbb{E}^{\mathbf{y}} = \{(j, k) : a_k \in A_j, \forall k = 1, ..., m\}$. Indeed, since coarse observations are supposed to be faithful, $n_{j,k} = 0$ if $a_k \notin A_j$.

The constraints (6(a)) guarantee that all observations are taken into account. The constraints (6(b)) and (6(c)) guarantee that the number of virtual samples $\mathbf{n} \in \mathcal{N}^{\mathbf{y}}$ is in agreement with observations. Equation (6(d)) is a normalisation constraint. Moreover we add constraints (6(e)) since the observation is integer and $\log(x)$ is not defined for $x = 0$. Constraint (6(d)) expresses the reliability of imprecise observations. In particular, the set $\mathcal{N}$ of feasible statistics $\mathbf{n}$ for X is thus defined by the set of m-tuples of integers verifying constraints (6(a, b, c)), and such that $(j, k) \in \mathbb{E}^{\mathbf{y}}$.

Remark 1 *Note that the maximal size of $\mathcal{Y}$ is a linear function of the number of observations and not exponential of the form $2^{|\mathcal{X}|}$. More precisely, it is $\min(2^{|\mathcal{X}|}, \sum_{k=1}^{r} q_r)$. In fact, in the case where $2^{|\mathcal{X}|} > \sum_{k=1}^{r} q_r$, observations could be different from one another, i.e., $q_k = 1, k = 1, ..., r$.*

4. The maxmin strategy maximizes entropy

Problems of the form (6) are well-known in the framework of game theory. The major issue is to find conditions under which the expression $\max_u \min_v f(u, v)$ is equal to $\min_u \max_v f(u, v)$ for (u, v) lying in a compact convex subset of $\mathbb{R}^2$. In the general case, the following inequality always holds:

$$\max_u \min_v f(u, v) \leq \min_v \max_u f(u, v).$$

When there is a saddle point, that is a pair (u^*, v^*) such that

$$f(u^*, v) \leq f(u^*, v^*) \leq f(u, v^*), \forall\, u, v,$$

then the equality holds, and corresponds to the notion of Nash equilibrium in game theory. This is the case when the function f is convex-concave and continuous, that is when f is convex in u and concave in v (Von Neumann, 1928; Sion, 1958; Komiya, 1988).

Problem (6) can be written as $\max_\theta \min_{\mathbf{n} \in \mathcal{N}^{\mathbf{y}}} f(\mathbf{n}, \theta)$, where function f has the form: $f(\mathbf{n}, \theta) = \sum_{i=1}^{m} n_k \log p(X = a_k|\theta)$. Provisionally, let us drop the assumption that $\mathbf{n}$ is a vector of integers, and assume it is a set of reals obeying (6(a, f)). It is easy to see that $f(\mathbf{n}, \theta)$ is increasing and linear in $\mathbf{n}$, while is concave and continuous with respect to $\theta = (p_k)_{k=1,...,m}$. The optimisation domain is clearly a compact and convex set. So, f is convex concave, and the above known result then applies:

Proposition 1 *Assuming $\boldsymbol{n}$ is not restricted to being integer-valued, the equality $\max_{\boldsymbol{p}} \min_{\boldsymbol{n}} \sum_{k=1,...,m} n_k \cdot \log p_k = \min_{\boldsymbol{n}} \max_{\boldsymbol{p}} \sum_{k=1,...,m} n_k \cdot \log p_k$ holds.*

The solution to the minmax problem is easier to find. Indeed the problem $\max_{\mathbf{p}} \sum_{k=1,...,m} n_k \cdot \log p_k$ is a standard maximum likelihood problem with a fixed vector $\mathbf{n}$. The optimal solution is given by $p_k = n_k/N, k = 1, ..., m$. Now we are led to find $\mathbf{n}$ that maximizes an expression of the form $-n_k \cdot \log(n_k/N)$ which, divided by N, is clearly the entropy of $(n_1/N, \ldots, n_m/N)$. We can thus conclude that:

Corollary 1 *The optimal solution to the maxmin likelihood problem (6) is the solution with maximal entropy, namely the solution to:* $\max_{\boldsymbol{n}} -\sum_{k=1,...,m} \frac{n_k}{N} \cdot \log \frac{n_k}{N}$ *under conditions (6(a, b, c)), and* $n_k \in \mathbb{R}^+$*, i.e.* $\boldsymbol{n}$ *in the convex hull of* $\mathcal{N}^{\mathbf{y}}$.

In fact, it is easy to see that the observed data $(q_j, A_j), j = 1 \ldots r$ defines a belief function Bel with mass function $\mu(A_j) = \frac{q_j}{N}, j = 1 \ldots r$, and that the convex set of probabilities $\mathcal{P} = \{P : P \geq Bel\}$ is nothing but the credal set defined by the set of probability assignments $\mathbf{p} = (n_1/N, \ldots, n_m/N)$ (Zaffalon, 2002), where $\sum_{k=1,...,m} n_k = \sum_{j=1,...,r} q_j = N$ plus conditions (6(b, c)) and $n_i \geq 0$ are reals. So the maxmin likelihood problem (6) comes down to a finding the maximum entropy probability in the credal set $\mathcal{P}$, a problem already addressed in the past by Abellán and Moral (2003).

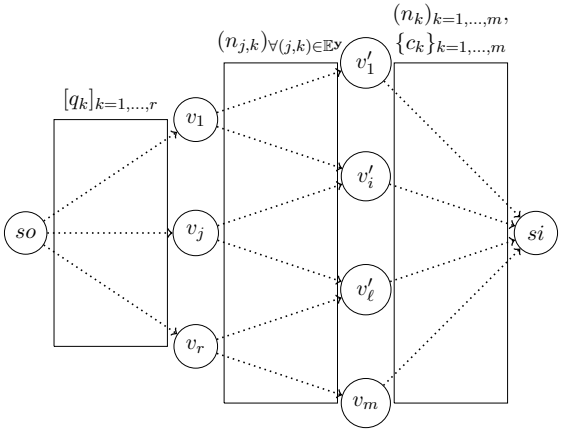

Figure 1: Graph representation of the problem

It remains to be checked whether the optimal solution $\mathbf{n}^*$ is integer or not. To this end, we focus on the problem of minimizing $\sum_{k=1,...,m} n_k \cdot \log p_k$ for a given probability distribution $\mathbf{p}$. The decision variables form the vector $\mathbf{n}$ in the convex hull of $\mathcal{N}^{\mathbf{y}}$. The problem considered is :

$$\min_{\mathbf{n}} \sum_{k=1,...,m} n_k c_k \tag{7}$$

$$\begin{aligned}
&s.t.\\
(a)\quad & \textstyle\sum_{k=1,...,m} n_k = \sum_{k=1,...,r} q_k = N,\\
(b)\quad & n_k - \textstyle\sum_{j:(j,k)\in\mathbb{E}^{\mathbf{y}}} n_{j,k} = 0, && \forall k = 1, ..., m\\
(c)\quad & \textstyle\sum_{k:(j,k)\in\mathbb{E}^{\mathbf{y}}} n_{j,k} = q_j, && \forall j = 1, ..., r\\
(d)\quad & n_k \in \mathbb{N}^+, && \forall k = 1, ..., m
\end{aligned}$$

where $c_k = \log p_k, k = 1, ..., m$ are constant. The problem (7) can be modeled by a bipartite transportation graph, as done by Zaffalon (2002). The graph is $(V, \mathcal{E})$ where the vertices V include

a source node so related to r vertices corresponding to elements of $\mathcal{Y}$, themselves related to m nodes corresponding to the elements of $\mathcal{X}$, and finally a sink node si. Edges in $\mathcal{E}$ are of the form (so, v_j), (v_j, v'_k) if $(j,k) \in \mathbb{E}^{\mathbf{y}}$, and (v'_k, si) (see Fig.1). The values in brackets provide the flow along these edges.

Proposition 2 *The problem (7) is a maximum flow minimum cost problem.*

Proof: The constraint (7(a)) is the equality constraint between the source flow and the sink flow. The constraints (7(b)) and (7(c)) are flow conservation constraints. In our case, the maximum flow is equal to $\sum_{k=1,\ldots,r} q_k$. □

From Proposition 2, we know that this problem has a totally unimodular structure, i.e., it is a linear problem with a totally unimodular constraint matrix. Therefore, the linear program relaxation of the model (7), letting $n_k \in \mathbb{R}^+$ yields an integral solution, which is thus the one of problem (7). So, the maximal entropy solution we have defined above for the maximization of the lower hidden likelihood is indeed of the form $(n_1/N, \ldots, n_m/N)$ for integer values of n_k.

Remark 2 *The solution of the maximization of the upper hidden likelihood, that is* $\max_{\boldsymbol{p}} \max_{\boldsymbol{n}} \sum_{k=1,\ldots,m} n_k \cdot \log p_k$ *under constraints (7(a-d)), is trivially equivalent to* $\max_{\boldsymbol{n}} \max_{\boldsymbol{p}} \sum_{k=1,\ldots,m} n_k \cdot \log p_k$. *It corresponds to minimizing the entropy of the vector* $(n_1/N, \ldots, n_m/N) \in \mathcal{N}^{\mathbf{y}}$ *in the credal set induced by* $\mathbf{y}$*, i.e. looking for the minimally uncertain frequency tuples compatible with observations, which corresponds to the idea of disambiguation put forward by Hüllermeier (2014).*

The above results shed light on the significance of the maximin and the maximax strategies and are useful to understand when to apply one or the other.

- The maximin strategy makes sense if we know that the process generating the variable X is genuinely non-deterministic, and that the imprecision of the observation may hide some variability (for instance the pace of variability of X is higher than the one of the observation process, so that X may vary during the making of one observation). Consider the case of reporting daily the temperature of the outside air based on a device that records the temperature variation within each day. This information is representative of the "average daily temperature", which may lead to their modelling as epistemic intervals containing this average value. Then it is reasonable to interpret the coarseness of A_i in terms of underlying variability and to go for a maximal entropy solution to the maximum likelihood problem.

- The maximax strategy makes sense if it is assumed that the underlying phenomenon is deterministic but the observations are noisy and coarse. If we try to learn a best model taken from a class of models and we have some good reason to think that the phenomenon under study can be represented by one of these models, then it is natural to try and choose one of them. In particular, it is clear that if $A = \cap_{j=1} A_j \neq \emptyset$, then the maximax strategy yields any Dirac function on $\mathcal{X}$ such that $P(A) = 1$ (it picks any element in A). For instance, consider a linear regression problem with interval observations, an example from Hüllermeier (2014). If the studied phenomenon is known to be affine, then one may choose the straight line that achieves a best fit with respect to the intervals. Especially any linear model that would be consistent with all interval observations will be preferred. The maximin strategy clearly yields a very different result due to the link with maximal entropy laid bare above.

5. Resolution method and an example

In this section, we propose a mathematical programming approach to solving problem (6), which comes down to optimizing a linear objective function under convex constraints. From the duality theorem, we know that the cost value of an optimal solution of the original (primal) model is equal to the cost value of the optimal solution of its dual. Let α be the dual variable associated to constraint (7(a)), $\beta_k, k = 1, ..., m$ the dual variables for constraints (7(b)) and $\gamma_k, k = 1, ..., r$ the dual variables for constraints (7(c)). The dual form of problem (7) is:

$$\begin{array}{lll} \max_{\alpha,\beta,\gamma} -(\alpha \sum_{k=1}^{r} q_k + \sum_{k=1}^{r} \gamma_k q_k) & & (8) \\ s.t. & \alpha + \beta_k \geq -\log(p_k), & \forall k = 1, ..., m \\ & -\beta_j + \gamma_k \geq 0, & \forall (k, j) \in \mathbb{E}^{\mathbf{y}} \\ & \alpha, \beta_j, \gamma_k \in \mathbb{R}, & \forall j = 1, ..., r, k = 1, ..., m \end{array}$$

Let us now return to the initial problem (eq.6) where the probability distribution is a decision variable. Its dual problem can be now written as a maximax problem:

$$\begin{array}{lll} \max_{\mathbf{p}} \max_{\alpha,\beta,\gamma} -(\alpha \sum_{k=1}^{r} q_k + \sum_{k=1}^{r} \gamma_k q_k) & & (9) \\ s.t. & & \\ (a) & \alpha + \beta_k \geq -\log(p_k), & \forall k = 1, ..., m \\ (b) & -\beta_j + \gamma_k \geq 0, & \forall (j, k) \in \mathbb{E}^{\mathbf{y}} \\ (c) & \sum_{k=1,...,m} p_k = 1 & \\ (d) & p_k > 0, & \forall k = 1, ..., m \\ (e) & \alpha, \beta_j, \gamma_k \in \mathbb{R}, & \forall j = 1, ..., r, k = 1, ..., m \end{array}$$

One can reformulate the problem (9) as follows with $\epsilon \to 0$:

$$\begin{array}{lll} \min_{P,\alpha,\beta,\gamma} \alpha \sum_{k=1}^{r} q_k + \sum_{k=1}^{r} \gamma_k q_k & & (10) \\ s.t. & & \\ (a) & \alpha + \beta_k \geq -\log(p_k), & \forall k = 1, ..., m \\ (b) & -\beta_j + \gamma_k \geq 0, & \forall (j, k) \in \mathbb{E}^{\mathbf{y}} \\ (c) & \sum_{k=1,...,m} p_k = 1 & \\ (d) & p_k + \epsilon \geq 0, & \forall k = 1, ..., m \\ (e) & p_k, \alpha, \beta_j, \gamma_k \in \mathbb{R}, & \forall j = 1, ..., r, k = 1, ..., m \end{array}$$

The problem (10) has a linear objective function to minimize, m convex constraints 10.(a) plus linear constraints. Hence this problem can be efficiently solved using a nonlinear solver.

Example

We want to estimate the probability that a type of car is present in some parking lot. The custodian provides some characteristics of cars (color and the number of doors) in a data base. For simplicity we consider three colors: red (r), blue(b),grey (g) and two situations for doors: 3 doors (3) and 5 doors (5).There are 6 possible types of cars:$\{r3, r5, b3, b5, g3, g5\}$. The information reported by the custodian can be both the color and the number of doors or only the color or only the number

$\mathcal{Y}$	$\{r3\}$	$\{r5\}$	$\{b3\}$	$\{b5\}$	$\{g3\}$	$\{g5\}$
q	9	167	120	199	164	188
$\mathcal{Y}$	$\{r3, b3, g3\}$	$\{r5, b5, g5\}$	$\{r3, r5\}$	$\{b3, b5\}$	$\{g3, g5\}$	
q	80	80	18	107	100	

Table 1: Distribution of Coarse Observations

of doors. So, we have $\mathcal{Y} = \{\{r3\}, \{r5\}, \{b3\}, \{b5\}, \{g3\}, \{g5\}, \{r3, b3, g3\}, \{r5, b5, g5\}, \{r3, r5\}, \{b3, b5\}, \{g3, g5\}\}$. Table 1 provides the coarse dataset.

To estimate the maximin probability distribution on $\mathcal{X}$ (noted p^{Mm}) we solve the mathematical formulation given in section 5 using the solver SQP of software Octave.[1] To discuss the result, we compare it with the probability distribution obtained using a maximax approach (noted p^{MM}). The results are given in table 2. Firstly, the maximin allow us to conclude that $\{r3\}$ is the least probable,

$\mathcal{X}$	$\{r3\}$	$\{r5\}$	$\{b3\}$	$\{b5\}$	$\{g3\}$	$\{g5\}$
$p^{Mm}(X = a_i) \approx$	0.141	0.171	0.171	0.171	0.173	0.173
$p^{MM}(X = a_i) \approx$	0.007	0.150	0.098	0.313	0.279	0.153

Table 2: Estimations of probability distributions on the latent variable

$\{r5\}$, $\{b3\}$, and $\{b5\}$ have the same probability to be present in this parking. Finally $\{g5\}$ and $\{g3\}$ are the most expected ones in this parking. The uncertainty on data prevents us from differentiating between $\{r5\}$, $\{b3\}$ or $\{b5\}$. In the same way, it is not possible to differentiate the probabilities of a car of types $\{g3\}$ or $\{g5\}$.

Let us compare both approaches on the resulting distributions pictured on Table 2. Both the maximin as the maximax approaches suggest that the cars of type $r3$ have the least probability to appear. But its probability in the maxmin approach is around two times the probability obtained by the maximax approach. In fact, in the maximax approach the observations $\{r3, b3, g3\}$ and $\{r3, r5\}$ are respectively interpreted as $\{g3\}$ and $\{r5\}$. It supposes that when the custodian just writes the characteristic "3 doors" in data base, the car is supposed to be grey. And when the custodian only writes the characteristic "red", the car has 3 doors. One can see that the probability of $\{r5\}$, $\{b3\}$, and $\{b5\}$ are very different, like probability $\{g5\}$ and $\{g3\}$.

We focus now on the probability of $\{b3\}$, and $\{b5\}$. In the maximin approach they were equal but in the maximax approach the probability $\{b3\}$ is the second less probable while $\{b5\}$ is the most probable type of car. But one can see that around half of observations concerning $\{b3\}$ or $\{b5\}$ are imprecise. It is clear that the maximin approach favors uniform distributions over outcomes while the maximax approach tends to put more weights on some specific cars, namely those which have been already most often observed precisely (such as $\{b5\}$).

Let us swap observations $\{b5\}$ and $\{b3\}$,i.e., suppose there are 120 observations for $\{b5\}$ and 199 observations for $\{b3\}$. The probability distribution of maximin approach does not change since the number of imprecise observations $\{b3, b5\}$ is too high to separate the probabilities of $\{b3\}$ and $\{b5\}$. But the probability distribution of the maximax approach is very sensitive to this exchange (see Table3). Of course, the probability of $\{b3\}$ becomes higher than that of $\{b5\}$. We point out to that the probability of $\{g3\}$ and $\{g5\}$ changes a lot. In fact, now the observations $\{r3, b3, g3\}$ are

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

interpreted as $\{b3\}$ and not $\{g3\}$ while the observations $\{g3, g5\}$ are interpreted as $\{g3\}$ and not $\{g5\}$.

$\mathcal{X}$	$\{r3\}$	$\{r5\}$	$\{b3\}$	$\{b5\}$	$\{g3\}$	$\{g5\}$
$p^{MM}(X = a_i) \approx$	0.008	0.150	0.313	0.097	0.133	0.299

Table 3: Maximax probability distribution with the modified dataset

In this example we show that the maximin approach is cautious compared to the maximax approach. More precisely, a high number of very coarse observations tends to equalize the probabilities of elementary outcomes while the maximax approach tends to select a best outcome consistent to coarse observations and this result can be completely altered by slightly changing the number of observations of each kind, which may lead to very different results.

6. Conclusion

This paper is a contribution to the study of maximum likelihood methods when data are coarse. The most popular approaches often assume some knowledge about the measurement process (as witnessed by the use of the superset of the CAR assumptions). These assumptions are strong and lead to work with the likelihood function of the complete joint sample involving both the observed and the latent variables. In our approach, we ignore the measurement process, and adopt a cautious approach involving robust optimisation and graph-theoretic methods. This approach, introduced previously (Guillaume and Dubois, 2015) for continuous parametric distributions and interval data, is here studied for finite sets of outcomes. The close connections between maximax and maximin strategies with entropy optimization shed light on the significance of each approach: the intuitive character of the resulting distribution depends on whether the observed phenomenon is genuinely random, or if it is deterministic, with a known class of models, and randomness comes from the measurement tool that is both imprecise and noisy: only in the latter case does the disambiguation strategy sound natural. On the contrary, the maximin approach interprets imprecision as the effect of the variability of the real outcomes. Moreover, we have proposed an efficient solving technique that can use existing non-linear optimization software. Further work is needed to test the approach on real data, and compare obtained results with other approaches that use belief functions (Denoeux, 2013), and also recent possibilistic maximum likelihood methods, which yield possibility distributions with fixed levels of specificity (Haddad et al., 2016).

References

J. Abellán and S. Moral. Maximum of entropy for credal sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(5):587–598, 2003.

I. Couso and D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *International Journal of Approximate Reasoning*, 55(7):1502–1518, 2014.

I. Couso and D. Dubois. Maximum likelihood under incomplete information: Toward a comparison of criteria. In *Soft Methods for Data Science*, pages 141–148. Springer, 2016a.

I. Couso and D. Dubois. Belief revision and the em algorithm. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 279–290. Springer, 2016b.

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

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

T. Denoeux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Transactions on knowledge and data engineering*, 25(1):119–130, 2013.

G. González-Rodríguez, A. Colubi, and M. Á. Gil. Fuzzy data treated as functional data: A one-way anova test approach. *Computational Statistics & Data Analysis*, 56(4):943–955, 2012.

R. Guillaume and D. Dubois. Robust parameter estimation of density functions under fuzzy interval observations. In *9th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA'15)*, pages 147–156, 2015.

M. Haddad, P. Leray, and N. B. Amor. Possibilistic networks: Parameters learning from imprecise data and evaluation strategy. *CoRR*, abs/1607.03705, 2016.

D. F. Heitjan and D. B. Rubin. Ignorability and coarse data. *The annals of statistics*, pages 2244–2253, 1991.

E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014.

E. Hüllermeier and W. Cheng. Superset learning based on generalized loss minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 260–275. Springer, 2015.

H. Komiya. Elementary proof for sion's minimax theorem. *Kodai mathematical journal*, 11(1):5–7, 1988.

R. Kruse and K. D. Meyer. *Statistics with vague data*, volume 6. Springer Science & Business Media, 2012.

M. Serrurier and H. Prade. An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data. *International Journal of Approximate Reasoning*, 54(7):919–933, 2013.

M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

J. Von Neumann. Zur theorie der gesellschaftsspiele. *Math. Annalen.*, 100:295–320, 1928.

M. Zaffalon. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105:105–122, 2002.