Robust Parameter Estimation of Density Functions under Fuzzy Interval Observations

Romain Guillaume
IRIT - Université de Toulouse, France
guillaum@irit.fr

Didier Dubois
IRIT, CNRS & Université de Toulouse, France
dubois@irit.fr

Abstract

This paper deals with the derivation of a probabilistic parametric model from interval or fuzzy data using the maximum likelihood principle. In contrast with classical techniques such as the EM algorithm, that define a precise likelihood function by averaging inside each imprecise observations, 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. We define an interval-valued likelihood function and apply robust optimisation methods to find a safe plausible estimate of the statistical parameters. The resulting density has a standard deviation that is large enough to cover the imprecision of the observations, making a pessimistic assumption on dispersion. This approach is extended to fuzzy data by optimizing the average of lower likelihoods over a collection of data sets obtained from cuts of the fuzzy intervals, as a trade off between optimistic and pessimistic interpretations of fuzzy data. The principles of this method are compared with those of other existing approaches to handle incompleteness of observations, especially the EM technique.

Keywords. Possibility theory, fuzzy intervals, maximum likelihood, robust optimisation, epistemic uncertainty

1 Introduction

Interval observations, and more generally, set-valued ones, do not always refer to the same situation [1]. Intervals may either represent exact observations of items taking the form of intervals (for instance, the daily min-max temperature ranges across one year), or, on the contrary, imprecise observations of precise quantities. In the first situation, interval data are a special kind of functional data where observations lie in a space of characteristic functions equipped with the suitable metric structure, enabling precise statistical parameters to be derived [18]. In this paper we are interested in the statistical analysis of data when observations are imprecise, more specifically, when we only know that the precise values of observations are restricted by intervals or fuzzy intervals. This kind of fuzzy interval is an epistemic set [1] which attaches to each value the possibility that it is the true observed value (unreachable for the observer). Under the epistemic approach, the expected value and the variance of a set of fuzzy intervals are fuzzy intervals [2].

This paper presents a general iterative approach to compute estimates of the parameters of a density function under imprecise observations, where the lack of precision is an epistemic rather than an aleatory phenomenon. To estimate the quality of parameters of the underlying precise random process, we use the maximum likelihood principle. Nevertheless, under imprecise observations, the likelihood function itself becomes imprecisely appraised too and is thus interval-valued. In this paper we adopt a pessimistic point of view and maximize the lower bound of the likelihood function, with a view to obtain a robust probability density whose standard variation accounts for potentially extreme variability across imprecision intervals.

The paper is organized as follows. In Section 2, we propose an algorithm that evaluates minimal and maximal bounds for the likelihood function. Then, we formulate the estimation problem for interval data as a robust optimization problem, which consists in maximizing the minimal expected likelihood. We study the cases of unimodal and Gaussian distributions. In Section 3, we define an extension of this approach to fuzzy interval data. Especially we discuss how to define a likelihood function for fuzzy interval data. In the literature, a classical approach to handling incomplete data in estimation is the famous EM algorithm [3]. It considers that the likelihood function is a precise function, even if observations are imprecise. In Section 4 we briefly discuss the difference between the two approaches, as well as the optimistic counterpart of ours.

2 Interval Uncertainty

Before solving the problem with fuzzy intervals, we focus on the problem with classical intervals. Firstly, we present a general framework for handing uncertainty on observations whatever the parametrized family of distributions. Secondly, we present an algorithm to solve the problem for unimodal density distributions. Finally, we study the case of normal density distributions.

2.1 General Framework

Let $\{x_i : i \in N\}$ with $|N| = n$, be a set of precise observations. To evaluate the quality of the parameters of the distribution that represents these observations, the usual approach is to define a likelihood measure $f(x_i|\theta)$ for each piece of data. Note that $f(x_i|\theta)$ can be understood as the possibility that the generation process for x_i is based on the parameter value θ [4]. The density function with vector of parameters θ and independent observations $\{x_i : i \in N\}$ takes the form of a product of likelihood functions:

$$L = \prod_{i \in N} f(x_i|\theta) \tag{1}$$

A standard criterion to define the parameters of the density function is the maximization of this likelihood function

$$\max_{\theta} \prod_{i \in N} f(x_i|\theta) \tag{2}$$

Under uncertainty, observations are of limited precision, and take the form of intervals $x_i \in [\underline{x_i}, \overline{x_i}], \forall i \in N$. Let $\Gamma = [\underline{x_1}, \overline{x_1}] \times ... \times [\underline{x_n}, \overline{x_n}]$ be the set of possible n-tuples of observations, we call *selections*. Namely, the selection $X \in \Gamma$ with $X = (x_1, ..., x_n)$ is a possible realization of the imprecise observation Γ. Fixing the parameter θ, one may argue that, in the spirit of [1], if observations are imprecise, the likelihood evaluation should become imprecise too, that is, $L(\theta) \in [\underline{L}; \overline{L}]$ with $\underline{L}$ and $\overline{L}$ respectively defined by:

- Lower likelihood

$$\underline{L}(\theta) = \min_{X \in \Gamma} \prod_{i \in N} f(x_i|\theta). \tag{3}$$

- Upper likelihood

$$\overline{L}(\theta) = \max_{X \in \Gamma} \prod_{i \in N} f(x_i|\theta). \tag{4}$$

To find robust solutions that cover potential variability, we can determine the parameter value (denoted by θ^{Rob}) which maximizes the *lower* likelihood. It can be formulated as a robust optimization problem:

$$\max_{\theta} \min_{X \in \Gamma} \prod_{i \in N} f(x_i|\theta) \tag{5}$$

This is equivalent to the log-likelihood problem:

$$\max_{\theta} \min_{X \in \Gamma} \sum_{i \in N} ln(f(x_i|\theta)) \tag{6}$$

2.2 Resolution Method

In this section, we propose an algorithm that evaluates the lower and upper bounds of the likelihood function for given parameters θ for a density function under the form (1). For a given data vector $X^* = (x_1^*, ..., x_n^*)$, the log-likelihood function $\sum_{i \in N} ln(f(x_i^*|\theta))$ is supposed to be convex with θ and to have a derivative.

Assumption 1 *$\exists x^m \in \mathbb{R}$ such that $f(x_i^*|\theta)$ is an increasing function on $]-\infty, x^m]$ and decreasing on $[x^m, +\infty[$.*

If the distribution is unimodal, x^m is the mode of the distribution.

2.2.1 Determining Upper and Lower Likelihood Functions

Note that the upper and lower likelihoods are of the form $f(X|\theta)$ for some $X \in \Gamma$. Moreover, from Property 1, we know that for a given parameter value θ, the minimum of function $f(x_i|\theta)$, where $x_i \in [\underline{x_i}, \overline{x_i}]$, is attained at the boundary of the domain ($x_i = \underline{x_i}$ or $x_i = \overline{x_i}$). It is called a *worst case selection*. This is not true for the *best case selection* obtained from the upper likelihood. Since the observations are assumed to be independent, the solution of problems (3) (worst case X^w for $\underline{L}(\theta)$) and (4) (best case X^b for $\overline{L}(\theta)$) can be computed using the following rules:

$$\text{if } x^m \in]-\infty; \underline{x_i}[\text{ then } \begin{cases} x_i^w = \overline{x_i} \\ x_i^b = \underline{x_i} \end{cases} \tag{7}$$

$$\text{if } x^m \in]\overline{x_i}; \infty[\text{ then } \begin{cases} x_i^w = \overline{x_i} \\ x_i^b = \underline{x_i} \end{cases} \tag{8}$$

$$\text{if } x^m \in [\underline{x_i}; \overline{x_i}] \text{ then } \begin{cases} x_i^b = x^m \\ x_i^w = \underline{x_i} \text{ if } f(\underline{x_i}|\theta) > f(\overline{x_i}|\theta), \\ \overline{x_i} \text{ otherwise.} \end{cases} \tag{9}$$

2.2.2 Computing Robust Parameters

The worst case selection $X^w(\theta) \in \Gamma$ is the one that minimizes the lower likelihood ($\underline{L}$) with parameter θ. If the density function is unimodal, it follows that the maximum likelihood problem comes down to discrete optimisation, that is we can restrict the selections of observations to extreme selections $X^w \in \Gamma_{\text{dis}}$, with $\Gamma_{\text{dis}} = \{\underline{x_1}, \overline{x_1}\} \times ... \times \{\underline{x_n}, \overline{x_n}\}$, the set of extreme assignments of x_i. Using Lagrange relaxation, problem (5) can be transformed into the following problem:

$$h^* = \max_{\theta} \sum_{X \in \Gamma_{\text{dis}}} \lambda_X \times (\sum_{i \in N} ln(f(x_i|\theta))) \tag{10}$$

where the Lagrange coefficients λ_X respect the conditions

$$\forall X \in \Gamma_{\text{dis}}, \lambda_X = \begin{cases} 1 \text{ if } h^{min} = \sum\limits_{i \in N} ln(f(x_i|\theta)) \\ 0 \text{ otherwise} \end{cases} \tag{11}$$

h^{min} being the minimal value of the log-likelihood over the selections:

$$h^{min} = \min_{X \in \Gamma_{\text{dis}}} \sum_{i \in N} ln(f(x_i|\theta)) \tag{12}$$

Proposition 1 *Expression (10) gives the optimal solution of the problem (5) if the Lagrange coefficients $\lambda_X, \forall X \in \Gamma_{dis}$ satisfy the conditions (11).*

Proof: Note that if the Lagrange coefficients respect the conditions (11), the expression (10) is equivalent to $h^* = \max\limits_{\theta} k \times (\min\limits_{X \in \Gamma_{\text{dis}}} \sum\limits_{i \in N} ln(f(x_i|\theta)))$ where k is the number of functions $\sum\limits_{i \in N} ln(f(x_i|\theta))$ that intersect at the maximum; it is the number of Lagrange coefficients $\lambda_X = 1$. Hence, the optimal solution of the previous expression (10) is the same as the optimal solution of problem $h^* = \max\limits_{\theta} \min\limits_{X \in \Gamma_{\text{dis}}} \sum\limits_{i \in N} ln(f(x_i|\theta)))$ which is equivalent to the problem (5) □

To solve problem (10), we use an iterative algorithm (Algorithm 1), which is an adaptation of the Uzawa method [5] to our problem.

Nevertheless the number of extreme selections is equal to 2^n. So we construct an iterative algorithm for solving problem (5) based on iterative relaxation scheme for min-max problems proposed in [6] and developed for min-max regret linear programming problems with an interval objective function [7, 8] coupled with Uzawa method.

Let RX-ROB be the problem (10) with a given set of assignments $\Gamma^*_{\text{dis}} \subseteq \Gamma_{\text{dis}}$. Obviously, the maximal cost h^* of problem RX-ROB over the discrete assignment set Γ^*_{dis} is an upper bound on the maximal cost of problem (6). Our algorithm (Algorithm 2) starts with zero upper bound $UB = 0$ and initial parameters θ^* (for instance the optimal parameter for the assignment of the mid-points of intervals) and empty discrete scenario set, $\Gamma^*_{\text{dis}} = \emptyset$. At each iteration, a worst case assignment X^w for θ^* is computed using rules (7, 8) and (9). Clearly, $\underline{L}(\theta^*)$ is an upper bound of $\underline{L}(\theta^{Rob})$. If a termination criterion is fulfilled (usually $\underline{L}(\theta^*) \leq UB - \epsilon$, $\epsilon > 0$ is a given constant) then the algorithm stops with an optimal robust parameter θ^*. Otherwise the worst case selection X^w is added to Γ^*_{dis}. Next the updated problem (RX-ROB) is solved to obtain a better candidate θ^* for an optimal solution to (5) and a new upper bound $UB = h^{min}$, based on Γ^*_{dis}. Since set Γ^*_{dis} is updated during the course of the algorithm, the computed values are upper bounds that form a nonincreasing sequence of values. Then, a new iteration is started.

Algorithm 1: A robust solution under a set of discrete scenarios

Input: Initial parameters $k = 0$, λ^0_X, the set of selections Γ_{dis}, and a convergence tolerance parameter $\rho > 0$.
Output: An optimal solution θ^{Rob},h^{Rob}
Step 1. Compute θ^k the optimal solution of problem (10) using $\lambda^k_X, X \in \Gamma_{\text{dis}}$
Step 2. If $\forall X \in \Gamma_{\text{dis}}$ the condition (11) is satisfied, then output θ^k, h^{min} and STOP.
Step 3. Compute the λ^{k+1}_X:
if $h^{min} = \sum\limits_{i \in N} ln(f(x_i|\theta^{k+1})$ then $\lambda^{k+1}_X = 1$ else
decrease the Lagrange parameter using
$\lambda^{k+1}_X = \max(0, \lambda^k_X - \rho \times (\sum\limits_{i \in N} ln(f(x_i|\theta^{k+1}) - h^{min}))$
Step 4. $k := k + 1$, and go to Step 1.

Algorithm 2: Finding optimal robust parameters.

Input: Observations $x_i = [\underline{x_i}; \overline{x_i}], \forall i \in N$, initial parameters θ^*, a convergence tolerance parameter $\epsilon > 0$.
Output: An optimal robust parameter θ^{Rob}
Step 0. $k := 0$, $UB := 0$, $\Gamma^*_{\text{dis}} := \emptyset$.
Step 1. $\theta^k := \theta^*$.
Step 2. Compute a worst case selection X^w for θ^k by solving problem (3) using rules (7), (8), (9). Then let $h = \sum\limits_{i \in N} ln(f(x^w_i|\theta))$
Step 3. If $(h \leq UB - \epsilon)$ then output θ^k and STOP.
Step 4. $k := k + 1$.
Step 5. $X^k := X^w$, $\Gamma^*_{\text{dis}} := \Gamma^*_{\text{dis}} \cup \{X^k\}$
Step 6. Compute an optimal solution θ^* by Algorithm 1, using Γ^*_{dis}; then set $UB = h^{min}$ and go to Step 1.

2.3 The Case of Normal Distributions

We suppose that the random variable follows a normal distribution:

$$f(x|(\mu,\sigma)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (13)$$

Upper and lower likelihoods can be reformulated into

$$\underline{L}(\mu,\sigma) = \min_{X\in\Gamma} \prod_{i\in N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}. \quad (14)$$

$$\overline{L}(\mu,\sigma) = \max_{X\in\Gamma} \prod_{i\in N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}. \quad (15)$$

2.3.1 Determining the Upper and Lower Likelihoods

The lower log-likelihood in the case of normal distributions becomes:

$$ln(\underline{L}(\mu,\sigma)) = -(\tfrac{Nln(\sigma^2)}{2} + \frac{1}{\sigma^2} \max_{X\in\Gamma} \sum_{i\in N} (x_i-\mu)^2)$$

Likewise, the upper log-likelihood in the case of normal distributions becomes:

$$ln(\overline{L}(\mu,\sigma)) = -(\tfrac{Nln(\sigma^2)}{2} + \frac{1}{\sigma^2} \min_{X\in\Gamma} \sum_{i\in N} (x_i-\mu)^2)$$

In this case the mode $x^m = \mu^*$ is the mean, and since the normal distribution is symmetric, the general equations (7, 8) and (9) that compute the worst and the best case selections become respectively:

$$\text{if } \mu^* \in]-\infty; \underline{x_i}[\text{ then } \begin{cases} x_i^w = \overline{x_i} \\ x_i^b = \underline{x_i} \end{cases} \quad (16)$$

$$\text{if } \mu^* \in]\overline{x_i}; \infty[\text{ then } \begin{cases} x_i^b = \overline{x_i} \\ x_i^w = \underline{x_i} \end{cases} \quad (17)$$

$$\text{if } \mu^* \in [\underline{x_i}; \overline{x_i}] \text{ then}$$

$$\begin{cases} x_i^b = \mu^* \\ x_i^w = \underline{x_i} \text{ if } (\underline{x_i}-\mu^*)^2 > (\overline{x_i}-\mu^*)^2, \\ \overline{x_i} \text{ otherwise} \end{cases} \quad (18)$$

It follows that the complexity for evaluating the lower and the upper likelihoods is $O(n)$.

2.3.2 Computing Robust Parameters

We can further decompose the problem of finding robust parameters into a sequence of two problems:

- first find the robust optimal μ^{rob}, solving the problem

$$ROB_{N,\mu}: \quad \min_{\mu} \max_{X\in\Gamma} \sum_{i\in N} (x_i-\mu)^2 \quad (19)$$

- and then compute the robust optimal σ^{rob}. We get the variance around μ^{rob} using the optimal selection X^w obtained at the previous step:

$$ROB_{N,\sigma}: \quad \sigma^{rob} = \sqrt{\frac{\sum_{i\in N} (x_i^w-\mu^{rob})^2}{n}} \quad (20)$$

Let us now focus on the problem $ROB_{N,\mu}$. Let $\underline{\mu} = \frac{1}{n}\sum_{i\in N} \underline{x_i}$ and $\overline{\mu} = \frac{1}{n}\sum_{i\in N} \overline{x_i}$.

Proposition 2 *The optimal solution μ^{rob} of the problem ROB_N lies in $[\underline{\mu}, \overline{\mu}]$.*

Proof Suppose $\exists \mu^{rob} < \underline{\mu}$. We have two cases. The first one is: the selection X^w associated to μ^{rob} is the same as the one for $\underline{\mu}$. We also know that, if $\mu^{rob} < \underline{\mu}$ then $\sum_{i\in N} (x_i-\mu^{rob})^2 > \sum_{i\in N} (x_i-\underline{\mu})^2$, since $\forall X \in \Gamma$, the optimal value $\mu^{op} \in [\underline{\mu}, \overline{\mu}]$, that contradicts the assumption that μ^{rob} is the optimal robust solution.

The second case is $X^w = X^w_{\underline{\mu}} + \delta$ where $X^w_{\underline{\mu}}$ is the worst case selection induced by $\underline{\mu}$ and δ is a vector of non-negative values. So $\sum_{i\in N} (x_i^w-\mu^{rob})^2 > \sum_{i\in N} (y_i-\mu^{rob})^2$ and $y_i = (X^w_{\underline{\mu}})_i$. We know that if $\mu^{rob} < \underline{\mu}$ then $\sum_{i\in N} (x_i-\mu^{rob})^2 > \sum_{i\in N} (x_i-\underline{\mu})^2$. Hence, $\sum_{i\in N} (x_i^w-\mu^{rob})^2 > \sum_{i\in N} (y_i-\underline{\mu})^2$, which contradicts the assumption that μ^{rob} is the optimal robust solution. The proof for the upper bound is similar. □

In the following Algorithm 3, we use the derivative

$$\frac{d(\sum_{i\in N} (x_i-\mu)^2)}{d\mu} = 2n\mu - 2\sum_{i\in N} x_i$$

.

Theorem 1 *Algorithm 3 finds the optimal robust parameter μ^{rob}.*

Proposition 3 *The complexity of computing the optimal robust solution μ^{rob} and σ^{rob} is $O(n.ln(|\mu|)$*

Algorithm 3: Finding optimal robust parameters for normal distribution.

Input: Observations $[\underline{x_i}; \overline{x_i}], \forall i \in N$, a convergence tolerance parameter $\epsilon > 0$.
Output: An optimal robust parameter μ^{Rob}
Step 0. $k := 0$, $a = \underline{\mu}$, $b = \overline{\mu}$.
Step 1. Compute a worst case selection X_c^w for the value $c = \frac{1}{2}(a + b)$.
Step 2. Compute the value $D = 2n\mu - 2\sum_{i \in N} x_i$ for the worst case selection X_c^w
Step 3. If $D < 0$ then $a := c$, else $b := c$.
Step 4. If $a - b > \epsilon$ then go to Step 1 else return $\frac{1}{2}(a + b)$ and STOP.

Proof: The major part of one iteration of the dichotomy algorithm is spent computing the worst case selection, which is $O(n)$. Since the dichotomy algorithm is $O(ln(|\mu|))$ where $|\mu|$ depends on the width of the interval $[\underline{\mu}, \overline{\mu}]$ and the precision parameter ϵ, the complexity of Algorithm 3 is $O(n.ln(|\mu|))$. And since σ^{rob} is directly computed from μ^{rob}, the complexity of computing the optimal robust solution μ^{rob} and σ^{rob} is $O(n.ln(|\mu|))$. □

2.3.3 Robust Solution vs. Maximal Variance

The robust solution can be understood as the parameter μ that minimizes the maximal possible variance under uncertainty (across all scenarios compatible with the interval data). Note that the problem $ROB_{N,\mu}$ is a relaxation of the problem of maximization of the variance of interval data [9]:

$$\max_{X \in \Gamma} \sum_{i \in N} (x_i - \sum_{i \in N} x_i/n)^2 \tag{21}$$

since, in the latter, $\mu = \sum_{i \in N} x_i/n$, while in problem $ROB_{N,\mu}$, μ is an independent variable. Let $(\sigma^{\max})^2$ be the maximal variance in problem (21). An imprecise probability solution to the estimation problem could be the set of normal distributions with $\mu \in [\underline{\mu}, \overline{\mu}]$ and $\sigma = \sigma^{\max}$. However the robust solution has the following property:

Proposition 4 $\sigma^{rob} \geq \sigma^{\max}$

Proof: It is enough to notice that

$$\begin{aligned} n(\sigma^{\max})^2 &= \max_{X \in \Gamma} \min_{\mu} \sum_{i \in N} (x_i - \mu)^2 \\ &\leq \min_{\mu} \max_{X \in \Gamma} \sum_{i \in N} (x_i - \mu)^2 = n(\sigma^{rob})^2 \quad \square \end{aligned}$$

Assume there is a single worst case solution X^w in problem $ROB_{N,\mu}$. In that case, the minimum is attained for the mean value $\mu^{rob} = \sum_{i \in N} x_i^w/n$, hence $\sigma^{rob} = \sigma^{\max}$. The maximal variance solution is then robust. However if there are several worst case solutions $X_j^w, j = 1, \ldots k$ in problem $ROB_{N,\mu}$, μ^{rob} is the intersection point of k parabolas

$$f_j(\mu) = \sum_{i \in N} (x_{ji}^w - \mu)^2,$$

while the maximal variance corresponds to the maximal ordinate of the minima of each parabola whose abscissa is

$$\mu^j = \sum_{i \in N} x_{ji}^w/n.$$

So the robust solution is a kind of compromise between extreme data selections, and is more pessimistic than the maximal variance solution.

3 Fuzzy Interval Uncertainty

We now use a more refined modeling of uncertainty pervading the observations. They are modeled by fuzzy intervals $\widetilde{x}_i, \forall i \in N$.

3.1 Selected Notions of Possibility Theory

A *fuzzy interval* $\tilde{A}$ is a fuzzy set in $\mathbb{R}$ whose membership function $\mu_{\tilde{A}}$ is normal, quasi concave and upper semicontinuous. Usually, it is assumed that the support of a fuzzy interval is compact. The main property of a fuzzy interval is the fact that all its α-cuts, that is, the sets $\tilde{A}^{[\alpha]} = \{x : \mu_{\tilde{A}}(x) \geq \alpha\}$, $\alpha \in (0, 1]$, are closed intervals. We will assume that $\tilde{A}^{[0]}$ is the smallest closed set containing the support of $\tilde{A}$. So, every fuzzy interval $\tilde{A}$ can be represented as a family of closed intervals $\tilde{A}^{[\alpha]} = [\underline{a}^{[\alpha]}, \overline{a}^{[\alpha]}]$, parametrized by the value of $\alpha \in [0, 1]$.

Let us now recall the possibilistic interpretation of fuzzy intervals. *Possibility theory* [10] is an approach to handle incomplete information and it relies on two dual measures: *possibility* and *necessity*, which express plausibility and certainty of events. Both measures are built from a *possibility distribution*. Let a fuzzy interval $\widetilde{A}$ be attached with a single-valued variable a (an uncertain real quantity). The membership function $\mu_{\widetilde{A}}$ is understood as a possibility distribution, $\pi_a = \mu_{\widetilde{A}}$, which describes the set of more or less plausible, mutually exclusive values of the variable a. It can encode a family of probability functions [11]. In particular, a degree of possibility can be viewed as the upper bound of a degree of probability [11]. The value of $\pi_a(v)$ represents the possibility degree of the assignment $a = v$, i.e. $\Pi(a = v) = \pi_a(v) = \mu_{\widetilde{A}}(v)$, where $\Pi(a = v)$ is the possibility of the event that a will take the value of v. In particular, $\pi_a(v) = 0$ means

that $a = v$ is impossible and $\pi_a(v) = 1$ means that $a = v$ is fully plausible. Equivalently, it means that the value of a belongs to an α-cut $\widetilde{A}^{[\alpha]}$ with confidence (or degree of necessity) $1 - \alpha$. It can be viewed as a random set defined by a multi-mapping from the unit interval equipped with Lebesgue measure to intervals consisting of cuts $\tilde{A}^{[\alpha]}$ [14]. Discrete approximations of π can also be viewed as random sets $(m, F)_\pi$, with nested focal sets E_i and masses $m(E_i)$, such that:

$$\begin{cases} E_i = \{x \in \mathbb{R} | \pi(x) \geq \alpha_i\} \\ m(E_i) = \alpha_i - \alpha_{i-1} \end{cases} \tag{22}$$

The possibility distribution is then approximated by: $\pi'(x) = \sum_{x \in E_i} m(E_i)$ [12].

3.2 Fuzzy Interval Datasets

A fuzzy interval data set is a collection of fuzzy intervals $\widetilde{x}_i, i = 1 \ldots, N$ whose membership functions are regarded as possibility distributions π_i restricting the values of the x_i's. The x_i's are stochastically independent but their uncertainties are non-interactive. We have thus extended the scenario set Γ from intervals (see Section 2) to the fuzzy case and now $\widetilde{\Gamma}$ is a fuzzy set of scenarios with membership function $\mu_{\widetilde{\Gamma}}(X) = \pi(X)$, $X \in \mathbb{R}^n$. The value of $\pi(X)$ stands for the possibility of the event that scenario $X \in \mathbb{R}^n$ has occurred. Hence, the possibility distributions associated with the observations x_i, forming the vector X, induce the following possibility distribution over all assignments in $X \in \mathbb{R}^n$ (see [13]):

$$\pi(X) \quad = \quad \min_{i=1,\ldots,n} \pi_i(x_i). \tag{23}$$

We see at once that the α-cuts of $\widetilde{\Gamma}$ for every $\alpha \in [0, 1]$ are such that: $\widetilde{\Gamma}^{[\alpha]} = \{X : \pi(X) \geq \alpha\} = [x_1^{-[\alpha]}, x_1^{+[\alpha]}] \times \cdots \times [x_T^{-[\alpha]}, x_T^{+[\alpha]}]$, from (23) and the definition of α-cut. Notice that $\widetilde{\Gamma}^\alpha$, $\alpha \in [0, 1]$, is the Cartesian product containing all selections (scenarios) whose possibility of occurrence is not less than α.

3.3 Formulations of Likelihood under Fuzzy Observations

In this section, we extend the definition of interval likelihood to the case of fuzzy intervals. There are several ideas that can be implemented to bring the fuzzy interval maximal likelihood problem back to a standard interval problem:

1. The simplest one is to turn fuzzy intervals into intervals by taking the interval mean [14], the Aumann integral $I(\widetilde{x}_i) = \int_0^1 [x_i^{-[\alpha]}, x_i^{+[\alpha]}] d\alpha$. However, one may then wonder why to start with fuzzy intervals in the first place.

2. Alternatively, we can solve the interval maximum likelihood problem for each α-cut, which would provide a set of possible solutions. If we remember that the fuzzy interval can also be interpreted in terms of subjective uncertainty, whereby $1 - \alpha$ is the degree of certainty that $[x^{-[\alpha]}, x^{+[\alpha]}]$ contains the actual observation x, the optimal parameter θ^*_α obtained from applying the interval approach to the α-cuts $\{[x_i^{-[\alpha]}, x_i^{+[\alpha]}] : i = 1, \ldots, n\}$ can be interpreted as the robust value of the model parameter corresponding to certainty $1-\alpha$, which can be viewed as a degree of pessimism of the solution. Indeed, if $\alpha = 1$ we take an optimistic view on the precision of the data, while if $\alpha = 0$, we assume the data is very imprecise and we try to be robust in the face of large interval uncertainty.

3. Yet another approach consists in considering all cuts of all fuzzy data $\widetilde{x}_i$ namely,

   $$\{[x_i^{-[\alpha]}, x_i^{+[\alpha]}] : i = 1, \ldots, n, \alpha \in [0, 1]\}$$

   as an equivalent set of interval data. In practice, this data set can be approximated using a finite set of cuts using equation (22). This approach considers the set of fuzzy data as a convex set of probabilities, induced by a random set in the spirit of Couso and Sanchez [15]. Indeed, the fuzzy data set is viewed as equivalent to a set of intervals generated as follows: Picking i at random in $\{1, \ldots, n\}$ and picking an α-cut at random ($[0, 1]$ is equipped with Lebesgue measure), obtaining the interval $[x^{-[\alpha]}, x^{+[\alpha]}]$.

All above approaches are amenable to a solution via the above proposed algorithms. These methods can be considered as somewhat extreme, as the first one does away with gradual membership, the second is difficult to use in practice (how to choose the best cut), and the third one considers two cuts of the same fuzzy observations as equivalent to two cuts each from a different observation, or in other words, fuzzy observations are the result of grouping together nested interval observations. Our next approach is a kind of trade-off between these views. Here we rely not on the mean interval of each fuzzy interval separately, but on the average of interval likelihoods obtained from all data sets $\widetilde{\Gamma}^\alpha$, $\alpha \in [0, 1]$.

3.4 The Average Robust Estimation Problem

We define a mean interval likelihood as follows:

Definition 1 *The mean interval likelihood under fuzzy observations is* $[\int_{\alpha\in[0,1]} \underline{L}^{\alpha} d\alpha, \int_{\alpha\in[0,1]} \overline{L}^{\alpha} d\alpha]$.

It can be approximated using a finite set of cuts using equation (22) and the average likelihood can then be expressed as:

$$[\sum_{j=1}^{k} m(\widetilde{\Gamma}_j^{\alpha})\underline{L}^{\alpha_j}, \sum_{j=1}^{k} m(\widetilde{\Gamma}_j^{\alpha})\overline{L}^{\alpha_j}] \quad (24)$$

The estimation of minimal and maximal likelihood under fuzzy observations can then be computed using the formulae (7, 8) and (9) $\forall i \in N, \forall j = 1, .., k$.

This average interval likelihood approach can be viewed as a balanced solution between working with the cores of the fuzzy intervals and their supports, while not letting cuts of a single fuzzy interval play the same role as cuts of different fuzzy intervals.

In the context of fuzzy information, the average robust problem can be formulated as follows:

$$\underline{L}^{rob} = \max_{\theta} \sum_{j=1}^{k} m(\widetilde{\Gamma}_j^{\alpha}) \min_{X\in\widetilde{\Gamma}_j^{\alpha}} \sum_{i\in N} ln(f(x_i|\theta)) \quad (25)$$

The reader may object to this formulation, as it seems that we give up our pessimistic point of view on interval data. However, it is not straightforward to define pessimism in a simple way in the face of fuzzy intervals. Indeed, fuzzy intervals carry two dimensions of pessimism, horizontal and vertical. On the one hand, the vertical dimension pertains to the choice of a cut of a fuzzy interval. Taking a cut at level 1, is optimistic in the sense that it is a narrow plausible range. Taking the support is safe but perhaps yields too imprecise an interval. On the other hand, the horizontal dimension (which end of the cut to choose ?) is the one at work in our approach to interval data, leading to take a pessimistic view on variance in the presence of imprecision. The approach proposed here achieves a global trade-off between vertical optimism and pessimism, and retains a pessimistic horizontal view.

3.4.1 The General Case

For simplicity we assume the discretisation of the membership set is such that $\forall j = 1, .., k, m(\widetilde{\Gamma}_j^{\alpha}) = 1/k$ (equidistant cuts).

Proposition 5 *The problem (25) can be reformulated as follows,*

$$h^* = \max_{\theta} \sum_{j=1}^{k} \sum_{X\in\widetilde{\Gamma}_{dis}^{\alpha_j}} \lambda_X^{\alpha_j} \times (\sum_{i\in N} ln(f(x_i|\theta))) \quad (26)$$

under the conditions: $\forall X \in \Gamma_{dis}$,

$$\lambda_X^{\alpha_j} = \begin{cases} \frac{1}{n_{\alpha_j}} \text{ if } h^* = \sum\limits_{i\in N} ln(f(x_i|\theta)), \\ 0, \text{ otherwise,} \end{cases} \quad (27)$$

where n_{α_j} *is the number of non-zero coefficients* $\lambda_X^{\alpha_j}$.

In fact, n_{α_j} is the number of functions $\sum\limits_{i\in N} ln(f(x_i^{[\alpha]}|\theta))$ that intersect at the maximum.

Proof: Note that if the Lagrange coefficients respect the conditions (27), the expression (26) is equivalent to $h^* = \max\limits_{\theta} \sum\limits_{j=1}^{k} \frac{n_{\alpha_j}}{n_{\alpha_j}} \times (\min\limits_{X\in\Gamma_{\text{dis}}} \sum\limits_{i\in N} ln(f(x_i|\theta)))$. Hence, the optimal solution of the previous expression (26) is the same as the optimal solution of problem (25). □

We can also generalize Algorithms 1 and 2 to the case of fuzzy observations by modifying Step 3 of Algorithm 1. This step becomes, for all $X \in \widetilde{\Gamma}_{\text{dis}}^{\alpha_j}$:

- if $h_j^{min} = \sum\limits_{i\in N} ln(f(x_i|\theta))$, then $\lambda_X^{k+1,\alpha_j} = \frac{1}{n_{\alpha_j}}$
- else $\lambda_X^{k+1,\alpha_j} = \max(0, \min(\lambda_X^{k,\alpha_j}, \frac{1}{n_{\alpha_j}}) - \rho(\sum\limits_{i\in N} ln(f(x_i|\theta^{k+1}) - h_j^{min}))$.

And Step 2 of Algorithm 2 must be used to find the worst case selection for each $j = 1, .., k$.

3.4.2 The Case of Normal Distributions

In the case of fuzzy observations, the optimal mean μ^* belongs to the set of means μ for scenarios with $\alpha = 0$.

Proposition 6 *The optimal value of the mean* μ *of the problem* ROB_N *is* $\mu^{rob} \in [\underline{\mu}^{[0]}, \overline{\mu}^{[0]}]$ *with* $\underline{\mu}^{[0]} = \frac{1}{n}\sum\limits_{i\in N} \underline{x_i}^{[0]}$ *and* $\overline{\mu}^{[0]} = \frac{1}{n}\sum\limits_{i\in N} \overline{x_i}^{[0]}$

To generalize Algorithm 3 to fuzzy observations, Step 1 becomes: Compute the worst case selection $X_j^w, \forall j = 1, .., k$ for the value $c = \frac{1}{2}(a + b)$. And the derivative of the likelihood function becomes $2nk\mu - 2\sum\limits_{j=1}^{k}\sum\limits_{i\in N} x_i^k$

Proposition 7 *The complexity of computing the optimal robust solution* μ^{rob} *and* σ^{rob} *is* $O(n.k.ln(|\mu^{[0]}|))$.

It is the same complexity as in the interval case but increased by a factor k (the number of cuts of the fuzzy intervals used in the algorithm).

4 Related Works

In the literature, a definition of likelihood under incomplete observations have been proposed by Dempster et al. [3]. This is the basis of the classical EM algorithm. Applied to our interval observations, it comes down to

Definition 2 *The likelihood of θ under one imprecise observation ($x \in [\underline{x};\overline{x}]$) is $L(\theta,[\underline{x};\overline{x}]) = P([\underline{x};\overline{x}]|\theta) = \int_{x\in[\underline{x};\overline{x}]} f(x|\theta)\,\mathrm{d}x$.*

The problem (5) is replaced by

$$\max_\theta P(\Gamma|\theta) = \max_\theta \prod_{i\in N} P_i([\underline{x_i},\overline{x_i}]|\theta) \qquad (28)$$

The EM method relies on the choice of an initial probability density, then compute averages over the intervals $[\underline{x_i},\overline{x_i}]$, which provides a precise dataset from which another density is obtained via maximal likelihood, and the process starts again until convergence.

However, this approach is often presented as handling latent unobserved variables, or missing values, not especially interval-valued observations. Namely, the authors using the EM algorithm rather present the framework as one with two kinds of variables: $\mathcal{X}$, a set of observed variables with precise realizations $\mathbf{x}$ and $\mathcal{Z}$ a set of non-observed variables, while here we consider a set of incomplete observations of the same variable. In the setting of the EM algorithm, an incomplete observation is thus of the form $\mathbf{x}\times Dom(\mathcal{Z})$, where Dom is short for domain. In other words, observations are set-valued, but form a partition of $Dom(\mathcal{X}\cup\mathcal{Z})$ into disjoint classes. So, moving from $\mathcal{X}$ to $\mathcal{X}\cup\mathcal{Z}$ corresponds to a change in granularity, whereby the second space is finer and the first can be viewed as a partition of the second. In such a situation, it sounds natural to consider that the likelihood $f(\mathbf{x}|\theta)$ is equated to the integral $\int_{Dom(\mathcal{Z})} f(\mathbf{x}\times\mathbf{z}|\theta)d\mathbf{z}$ (because the data $\mathcal{Z}$ is supposed to be missing at random, i.e., $f(\mathbf{x}|\theta) = f(\mathbf{x}|\mathbf{z},\theta)$). Insofar as one is only interested in events in the algebra formed by the coarse partition, the formulation of the likelihood function as in the EM approach is justified.

Recently, the EM algorithm has been generalized by Denoeux [16] to the case of data taking the form of mass functions of belief functions, and he also uses a scalar likelihood function defined as a weighted average of the EM likelihood:

Definition 3 *The likelihood of of a single imprecise observation x_i described by a belief function with mass function m_i bearing on intervals $[\underline{x}_j;\overline{x}_j]$ is $L(\theta,m) = \sum_j m([\underline{x}_j;\overline{x}_j])L(\theta,[\underline{x}_j;\overline{x}_j])$.*

Note that this definition also applies to fuzzy interval data viewed as consonant belief functions, and thus leads to yet another extension of maximum likelihood estimation to fuzzy data, studied by Denoeux [17].

In the case of interval-valued observations viewed as imprecise data (dealt with in our paper), the formulation of likelihood after the EM algorithm looks questionable.

On the one hand, there seems to be no point averaging the probabilities $f(x_i|\theta)$ over the interval $[\underline{x_i};\overline{x_i}]$, as if computing the frequency of this event from a sample space. Indeed, each observation $[\underline{x_i};\overline{x}_i]$ is a disjunctive set, one value of which is the real (unique) realization x_i. This defect in observing the x_i's leads to possibilistic uncertainty about the actual probability of their realizations (knowing θ), better expressed by the interval likelihood function $[\underline{L};\overline{L}]$. It contrasts with the problem of computing a frequency $P([\underline{x};\overline{x}]|\theta)$ based on a collection of precise observations. In the latter case, all observations inside $[\underline{x};\overline{x}]$ have been observed (it is a conjunctive set). In contrast, each interval $[\underline{x}_i,\overline{x}_i]$ is the incomplete description of a single precise observation x_i. The EM approach seems to interpret the equal possibility of all values in $[\underline{x};\overline{x}]$ as being an equal probability, or at least, it seems to admit the existence of a random process generating precise values inside this interval. In the case of latent or unobserved variables, it makes sense if they are indeed driven by an unobserved random process. But this is not our assumption in the case of interval uncertainty.

On the other hand, the overlapping nature of the interval valued data makes it hard to assume the existence of auxiliary random processes inside each interval, while if the incomplete observations partition the sample space, this assumption may look more natural.

Besides, Definition 2 does not generalize the definition of likelihood in the context of perfect observations, since under this definition, the likelihood of precise values is 0.

Here we view the intervals as describing epistemic uncertainty bearing on the observations that stem from a random process generating precise (even if grossly observed) data. Another option could be to assume that there is a second random process generating the intervals surrounding the outcomes of the first one. This purely aleatory view of imprecise observations is also at work in the trend on fuzzy random variables after Puri and Ralescu where they are interpreted as standard random variables whose images are functions [18]. However in this case, there would be three random variables, say $x, u > 0, v > 0$, such that the observed intervals are realizations of the form $[x-u, x+v]$. Then

we could rightfully define the likelihood function:

$$P([\underline{x}_i, \overline{x}_i]|\theta) = \int_{\underline{x}_i = x-u, \overline{x}_i = x+v} P(x, u, v|\theta) dx du dv$$

and apply the EM algorithm. However, here we consider the uncertainty pervading the observations x_i is not aleatory at all, it is just sheer lack of information due to the coarseness of the observation tool, and the width of the interval surrounding the x_i's is supposedly not generated by a random process.

More recently, Hüllermeier [19] proposed an approach similar to ours for simultaneously optimizing a model and disambiguating the (interval-valued) data. He presents the approach in terms of minimizing a loss function, and applies it to regression and classification problems. However, his idea comes down to maximizing the product of upper likelihoods in our setting, while our proposal, more in the spirit of robust optimisation, takes a pessimistic view on imprecise observations. Note that our approach also leads to disambiguating the data, albeit taking an opposite view, covering potential dispersion of the actual data, as testified by the use of extreme selections induced by rules (7, 8, 9).

Let us compare the two approaches on a simple case with two interval observations $x_1 \in [20, 30]$ and $x_2 \in [30, 40]$. The result of minimizing the loss function (maximizing the upper likelihood) is $x_1 = 30, x_2 = 30$ so $\mu = 30$ and $\sigma = 0$. In contrast, the robust approach applied to normal distributions will select the values $x_1 = 20, x_2 = 40$ so $\mu = 30$ and $\sigma = 10$. The Hüllermeier approach can be understood as the fusion of information items $x \in [20, 30]$ and $x \in [30, 40]$, privileging the common parts, which is optimistic, while our approach tends to assume the information could be very dissonant, with a variance equal to 100. See [20] for more comments along this line on the optimistic approach. Note that the EM approach on this case would maximise the product

$$L_{EM}(\theta) = P([20, 30]|\theta) \cdot P([30, 40]|\theta).$$

Using normal distributions, it would lead to the optimistic solution of Hüllermeier (a Dirac measure at 30) to ensure $L_{EM} = 1$. Note that the algorithm, assuming the initial distribution is fixed through the choice of θ_0, will compute the expectations $\hat{x}_1$ and $\hat{x}_2$ inside their intervals, and perform a likelihood maximization using these precise expectations as new data, based on the product of densities, getting a new value θ_1, and so on. This process will tend to shrink the expected interval $[\hat{x}_1, \hat{x}_2]$. However if the support of the current symmetric distribution lies inside $[20, 40]$, $L_{EM}(\theta)$ will remain constant (0.25) and jumps to 1 for the Dirac function on 30.

A natural issue is whether in the case of missing data, one may replace them by the whole range of the random variable, say an interval $[a, b]$, or not. This is immaterial for the EM algorithm applied to our setting as $P([a, b]|\theta) = 1$ in any case. So, in our setting, the EM algorithm would just neglect missing observations, whether unsuccessful experiments are carried out or not. On the contrary, in our approach, it makes a difference, as can be seen in the next example.

Consider the case when the range of x is [0, 200], 10 precise observations at 100 were made and the result of one experiment could not be properly observed, so its value lies in [0,200]. The mean value is $\mu = 100$ for the three approaches including the robust one, but the resulting distribution is the Dirac for the optimistic solution of Hüllermeier and the EM approach while the σ value of the robust approach is around 30 (not excluding a maximal deviation from 100 in the failed experiment). If now we have 10 precise observations at 100 and 10 completely imprecise ones modelled by [0,200]. The solution returned by the optimistic approach of Hüllermeier and the EM approach will still be a Dirac measure at 100 while the σ value of the robust approach becomes close to 70. In our approach, the imprecision of observations directly impacts the variance of the identified density. So, unsuccessful observations are not treated as observations not yet carried out. Whether this distinction is meaningful or not in all situations is a matter of debate.

More generally, in the case (perhaps unlikely in practice) where the dataset consists of overlapping intervals, it is clear that any density function with support inside the intersection of the intervals will ensure that the EM likelihood function $L_{EM} = 1$ in (28) since each term has probability 1 in the product (the same remark applies if one maximizes the upper likelihood). However our method will give a density whose standard deviation reflects the width of the uncertainty intervals. In this case, though, using a possibility measure to represent the data may sound more appropriate than a density that turns incompleteness into variability.

5 Conclusion

In this paper we propose to propagate the epistemic interval uncertainty pervading a data set over to the estimation of the likelihood. Then we propose an iterative algorithm which finds parameter values that maximize the lower likelihood values among all data sets compatible with the interval observations, under not too restrictive conditions on the density function. We have studied the case of normal distribution and have shown that the computation of optimal mean and variance can be achieved efficiently. As perspectives,

first we plan to compute robust parameter estimations for other classical distributions. In particular, the algorithm that finds optimal solutions can be improved taking into account the specificities of density functions (as for the normal distribution in this paper). Another perspective is the study of robust linear regression under imprecise observations. Finally, an experimental validation step will be useful to compare our results to those obtained by optimizing upper likelihoods, and methods in the style of the EM algorithm. This approach will be applied to the determination of robust production plans under ill-known demand modelled by fuzzy intervals, in the production engineering environment.

Acknowledgements

The authors are grateful to referees for interesting thought-provoking comments that led us to improve the paper while confirming our intuitions.

References

[1] I. Couso and D. Dubois. Statistical reasoning with set-valued information: Ontic vs. epistemic views. *Int. J. Approx. Reasoning*, 55(7):1502–1518, 2014.

[2] R. Kruse and K. Meyer. *Statistics with Vague Data*. D. Reidel, Dordrecht, 1987.

[3] A. Dempster, N. M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. of the Royal Stat. Soc., series B*, 39:1–38, 1977.

[4] D. Dubois, S. Moral, and H. Prade. A semantics for possibility theory based on likelihoods. *J. of Math. Anal. and Appl.*, 205:359–380, 1997.

[5] H. Uzawa. Iterative methods for concave programming. In K. J. Arrow, L. Hurwicz, and H. Uzawa, editors, *Studies in linear and nonlinear programming*. Stanford University Press, 1958.

[6] K. Shimizu and E. Aiyoshi. Necessary Conditions for Min-Max Problems and Algorithms by a Relaxation Procedure. *IEEE Trans. on Automatic Control*, 25:62–66, 1980.

[7] M. Inuiguchi and M. Sakawa. Minimax regret solution to linear programming problems with an interval objective function. *Eur. J. of Operational Research*, 86:526–536, 1995.

[8] H. E. Mausser and M. Laguna. A heuristic to minimax absolute regret for linear programs with interval objective function coefficients. *Eur. J. of Operational Research*, 117:157–174, 1999.

[9] V. Kreinovich, G. Xiang, and S. Ferson. Computing mean and variance under dempster-shafer uncertainty: Towards faster algorithms. *Int. J. Approx. Reasoning*, 42(3):212–227, 2006.

[10] D. Dubois and H. Prade. *Possibility theory: an approach to computerized processing of uncertainty*. Plenum Press, New York, 1988.

[11] D. Dubois and H. Prade. When upper probabilities are possibility measures. *Fuzzy Sets and Systems*, 49:65–74, 1992.

[12] D. Dubois and H. Prade. On several representations of an uncertain body of evidence. In M.M. Gupta and E. Sanchez, editors, *Fuzzy Information and Decision Processes*, pages 167–181. North-Holland, Amsterdam, 1982.

[13] D. Dubois, H. Fargier, and V. Galvagnon. On latest starting times and floats in activity networks with ill-known durations. *European Journal of Operational Research*, 147:266–280, 2003.

[14] D. Dubois and H. Prade. The mean value of a fuzzy number. *Fuzzy Sets and Systems*, 24:279–300, 1987.

[15] I. Couso and L. Sánchez. Upper and lower probabilities induced by a fuzzy random variable. *Fuzzy Sets and Systems*, 165(1):1–23, 2011.

[16] T. Denoeux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Trans. Knowl. Data Eng.*, 25(1):119–130, 2013.

[17] T. Denoeux. Maximum likelihood estimation from fuzzy data using the EM algorithm. *Fuzzy Sets and Systems*, 183(1):72–91, 2011.

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

[19] E. Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *Int. J. Approx. Reasoning*, 55(7):1519–1534, 2014.

[20] D. Dubois. On various ways of tackling incomplete information in statistics. *Int. J. Approx. Reasoning*, 55(7):1570–1574, 2014.