A general framework for maximizing likelihood under incomplete data

Inés Couso*

Dep. of Statistics and O.R., University of Oviedo, Spain

Didier Dubois

IRIT, Université Paul Sabatier, France

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, a major issue is to properly define the likelihood function to be maximized. This paper highlights the fact that there are several possible likelihood functions to be considered, depending on the purpose to be addressed, namely whether the behavior of the imperfect measurement process causing incompleteness should be included or not in the model, and what are the assumptions we can make or the knowledge we have about this measurement process. Various possible approaches, that differ by the choice of the likelihood function and/or the attitude of the analyst in front of imprecise information are comparatively discussed on examples, and some light is shed on the nature of the corresponding solutions.

Keywords Random sets, maximum likelihood, incomplete information, entropy.

1. Introduction

The key role of likelihood functions in statistical inference was first highlighted by Fisher [17] with the maximum likelihood principle. In his seminal book, Edwards

*Corresponding author

Email addresses: couso@uniovi.es (Inés Couso), dubois@irit.fr (Didier Dubois)

A preliminary version of this paper [7] was presented at the 8th Conference on Soft Methods in Probability and Statistics (SMPS) in Roma, September, 12-14, 2016.

([16], p. 9) defines a likelihood function as being proportional to the probability of obtaining results given a hypothesis, according to a probability model:

> Let $P(R|H)$ be the probability of obtaining results R given the hypothesis H, according to the probability model ... The likelihood of the hypothesis H given data R, and a specific model, is proportional to $P(R|H)$, the constant of proportionality being arbitrary.

Edwards mentions that "this probability is defined for any member of the set of possible results given any one hypothesis ... As such its mathematical properties are well-known. A fundamental axiom is that if R_1 and R_2 are two of the possible results, mutually exclusive, then $P(R_1 or R_2|H) = P(R_1|H) + P(R_2|H)$".

In other words, 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. So, a likelihood function is proportional to the conditional probability of an elementary event (the observed sample), where the condition part (the hypothesis) is a value of some model parameter. For instance, the conditional probability of the sure event cannot be viewed as the likelihood of the hypothesis given the sure event.

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 modelled by a random variable, observations we get in this case may not directly inform us about this random variable. Due to the interference with an imperfect measurement process, observations will be set-valued [4, 5]. So, in order to properly exploit such incomplete information (called *coarse data* in the literature [22]), we must first decide what to model:

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

In the first case, imprecise observations are considered as results, and we can construct the likelihood function of a random set, whose realizations are sets. These sets contain precise but ill-known realizations of the random variable of interest, to which we have no direct access. We say that this unreachable random variable is *latent*. Actually, most authors are interested in the other point of view. They consider that outcomes are the precise, although ill-observed, realizations of the random

phenomenon, and wish to reconstruct a distribution for the latent variable. However in this case there are as many potential 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 [11, 31, 13], which is an iterative procedure for efficient maximization of the likelihood of observed data. It constructs a distribution on the latent variable that minimizes divergence from the parametric model in agreement with the available data. It can also serve to reconstruct a sample of the latent variable.

In this paper, we propose a formal setting for the modelling of imprecisely observed random experiments, and define the three likelihood functions that can be built in this framework. Apart from the likelihood function based on available observations, there is the likelihood function based on outcomes of the latent random variable that was imprecisely observed, and the likelihood function based on the joint probability induced by pairs of outcomes and their measurement. The two latter likelihood functions are imprecisely known and we compare several alternatives to the maximization of the likelihood of imprecise observations, such as the maximax approach, and the robust approach to incomplete data. It includes more recent proposals by Hüllermeier [23], or Guillaume and Dubois [19], or Plass et al. [37]. We also discuss the use of assumptions on the measurement process such as the coarsening-at-random [22] and the superset assumptions, that help relating the various likelihood functions. 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, Masson and Denœux [34], or Serrurier and Prade [46]). We assume that the cause of imprecision lies in the incomplete description of the random experiment outcomes, not in the scarcity of observations.

2. 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 its range $\mathcal{X} = \{a_1, \ldots, a_m\}$ is finite. Suppose that observations of X are imprecise, namely let $\Gamma : \Omega \to \wp(\mathcal{X})$ denote the (observable) multi-valued mapping representing our (imprecise) perception of X. So, if ω occurs then all we know is that $X(\omega) \in \Gamma(\omega) \subseteq \mathcal{X}$. In other words, we assume that X is a selection of Γ, i.e. $X(\omega) \in \Gamma(\omega),\ \forall \omega \in \Omega$. This setting is very close to the one of Dempster [10] who introduces a special case of upper and lower probabilities, based on random sets, later interpreted by Shafer [47] as belief and plausibility functions. The issue of set-valued data has been discussed in Ref. [5] from the point of view of descriptive statistics. In this paper we start addressing inferential statistics.

Let $Im(\Gamma) = \{A_1, \ldots, A_r\} \in \wp(\mathcal{X})$ denote the image of Γ (the collection of possible set-valued outcomes). We can equivalently suppose that the imperfect measurement process is driven by another random variable Y, with finite range $\mathcal{Y} = \{b_1, \ldots, b_r\}$, that provides incomplete reports of observations of X. Namely, $Y(\omega) = b_j$ means that the measurement tool reports $\Gamma(\omega) = A_j$. The cardinality of the image of $Im(Y) = \mathcal{Y} = \{b_1, \ldots, b_r\}$ thus coincides with that of $Im(\Gamma)$ and let us assume that there is a bijection between Γ and Y as follows:

$$Y(\omega) = b_j \text{ iff } \Gamma(\omega) = A_j,\ j = 1, \ldots, r,$$

or yet we can assume that $b_j = A_j$. Let $P(X, Y)$ be the joint probability describing X and its measurement.

In some applications, the variable X is made of two components X_o and X_u respectively corresponding to observed and unobserved variables with respective domains $\mathcal{X}_o$ and $\mathcal{X}_u$, and Γ is of the form $\{X_o\} \times \Gamma_u$, i.e., $Y = \{\{X_0\} \times \Gamma_u\}$. The observed variable Y can then be identified with the random vector $Y = (X_o, Y_u)$, where $Y_u = \Gamma_u$.

This framework highlights the difference between the outcome $X = a_k$ (its probability is $P(X = a_k)$), the fact that event A_j occurs whenever the outcome $X = a_k$ belongs to A_j (its probability is $P(X \in A_j)$), and observing the result A_j via the measurement process (its probability is $P(Y = A_j)$). The latter is always a singleton, even when it corresponds to an imprecise observation of $X = a_k$.

In the paper we assume the results of experiments are available in the form of relative frequencies $\hat{p}_{.j} = \frac{n_{.j}}{n}$, where $n_{.j}$ denotes the number of observations of $b_j = A_j$ in the sample, and n is the sample size. The probability distribution $(\hat{p}_{.1}, \ldots, \hat{p}_{.r})$ on $\mathcal{Y}$ can also be viewed as a Dempster-Shafer mass assignment m on $\wp(\mathcal{X})$ [47], letting $m(A_j) = \hat{p}_{.j}$ for $j = 1, \ldots r$ inducing lower probabilities in the sense of [10] in the form of a belief function $Bel(A) = \sum_{E \subseteq A} m(A)$. This Dempster-Shafer mass assignment defines a convex set $\{P_X : P_X(A) \geq Bel(A), \forall A \subseteq \mathcal{X}\}$ of probabilities on $\mathcal{X}$, hence of joint probabilities on $\mathcal{X} \times \mathcal{Y}$ with known marginals $\hat{p}_{.j}$ for $j = 1, \ldots, r$ on $\mathcal{Y}$.

An alternative way of modeling the generation of coarse data consists in using so-called *coarsening variables* [21]. It supposes the existence of a random variable C valued on a finite space $\mathcal{C}$, and a function $F : \mathcal{X} \times \mathcal{C} \to \wp(X) \setminus \{\emptyset\}$ such that $Y = F(X, C)$.

We overview below two different ways to represent the information about the joint distribution of the random vector (X, Y). Subsection 2.1 will refer to the outcome

of the experiment X and the "coarsening" or "imprecisiation" process[2] that leads us to just get imprecise observations of X, described by Y. Subsection 2.2 will represent the joint distribution of (X, Y) the other way around, by means of the marginal probability of the observations (Y) and the conditional probability of X given Y. The "imprecisiation" or "disambiguation" views respectively correspond to what Little [29] calls selection models and pattern mixture models, albeit expressed in the framework of missing data using coarsening variables.

2.1. Generation and imprecisiation processes

Let us consider the following matrix: $(M|\mathbf{p})$:

$$\left(\begin{array}{ccc|c} p_{.1|1.} & \dots & p_{.r|1.} & p_{1.} \\ \dots & \dots & \dots & \dots \\ p_{.1|m.} & \dots & p_{.r|m.} & p_{m.} \end{array}\right)$$

where

- $p_{.j|k.} = P(Y = A_j|X = a_k)$ denotes the (conditional) probability of observing $b_j = A_j$ 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 $P(X, Y)$ modeling the underlying generating process plus the connection between true outcomes and incomplete observations. (More specifically, the vector $(p_{1.}, \dots, p_{m.})^T$ characterizes the underlying generating random process while the matrix $M = (p_{.j|k.})_{k=1,\dots,m;j=1,\dots,r}$ is the so-called mixing matrix ([48]) that represents the imprecisiation process. In the setting of Dempster's upper and lower probabilities [10], nothing is assumed about the matrix M and $(p_{1.}, \dots, p_{m.})^T$ is unknown. This is not the case in more recent works whose aim is to retrieve information about X from information about Y, using

[2]The term "coarsening" is commonly used in the literature of statistics with incomplete data. Notwithstanding, the idea of coarsening has been also linked to the idea of partition and indiscernibility. For instance, in Rough Set Theory [36] it means "change of granularity". Also, in Shafer's Theory of Evidence [47], it is linked to the idea of defining a partition of indiscernible elements. Indeed in some cases, within the literature of coarse statistics the notion of coarsening variable comes down to choose a partition, one element of which is the imprecise observation. So the term "coarsening process" seems to often underly this partition-based modeling of imprecise observation generation. However in this paper this is modeled by a mere multi-mapping in the style of Dempster [10]. In the rest of the paper, we will use the term "imprecisiation", that does not presuppose coarse data to be generated through partitioning.

a model of the measurement process, by means of some assumption on the mixing matrix M.

Some particular settings and their characteristic matrices:

- **Partition.** Suppose that $\{A_1, \ldots, A_r\}$ forms a partition of $\mathcal{X}$. Therefore, we can easily observe that the probabilities $P(Y = A_j | X = a_k) = 1$ if $a_k \in A_j$, and 0 otherwise, $\forall\, j, k$. Then, we can divide the m elements of $\mathcal{X}$ into r categories of respectively $k_1, \ldots, k_r$ elements each. We can denote $\mathcal{X} = \{a_{11}, \ldots, a_{1k_1}, \ldots, a_{r1}, \ldots, a_{rk_r}\}$ and particularize the above matrix as follows:

$$\left(\begin{array}{ccc|c} 1 & \ldots & 0 & p_{11.} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & \ldots & 0 & p_{1k_1.} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & p_{r1.} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 1 & p_{rk_r.} \end{array}\right)$$

  where $p_{il.} = P(X = a_{il})$ denotes the (marginal) probability that X takes the value a_{il}. Y is then a function of X, $Y = f(X)$. In this case, the joint distribution of (X, Y) is determined by the marginal distribution of X. This procedure determines an equivalence relation over $\mathcal{X}$:

$$a_i R a_j \Leftrightarrow f(a_i) = f(a_j),$$

  and therefore, a collection of equivalence classes, $\Pi = \{A_1, \ldots, A_r\}$, determining a partition of $\mathcal{X}$. It is clear that in this case $P(X, Y)$ is generated by a coarsening variable reduced to a constant ($f(X) = F(X, c)$). This setting is the one proposed by Dempster *et al.* [11] in their famous paper on the EM algorithm, presented as an approach to obtaining a maximum likelihood estimate (MLE) under incomplete information.[3]

- **Miss-or-observe setting.** In this case, we assume that either the value of X is observed precisely or that it is not observed at all. Then, $r = m + 1$ and $\{A_1, \ldots, A_r\} = \{\mathcal{X}, \{a_1\}, \ldots, \{a_m\}\}$.

[3]Curiously, the connection between this approach and the more general setting for incomplete information of [10] is not made by Dempster et al. [11]. It may be surprising as Dempster is also famous for his work on the foundations of evidence theory.

Let $P(\Gamma = \mathcal{X}|X = a_k) = \alpha_k, P(\Gamma = \{a_k\}|X = a_k) = 1 - \alpha_k,\ k = 1, \ldots, m$. The mixing matrix M is therefore of the form:

$$\left(\begin{array}{c|cccc} out.\backslash obs. & \mathcal{X} & \{a_1\} & \ldots & \{a_m\} \\ \hline a_1 & \alpha_1 & 1-\alpha_1 & \ldots & 0 \\ a_i & \alpha_i & \ldots & 1-\alpha_i & \\ a_m & \alpha_m & \ldots & \ldots & 1-\alpha_m \end{array}\right)$$

This is the situation of missing data [41]. In this case, there is a coarsening variable C with range $\{0, 1\}$, such that $F(X, C)$ is a singleton in $\mathcal{X}$ if $C = 0$ and is $\mathcal{X}$ otherwise. We can see that $P(C = 0|X = a_i) = 1 - \alpha_i = P(Y = \mathcal{X}|X = a_i)$. An important particular case is when the probabilities α_i's are constant, that is, the probability of missing data does not depend on the outcome of the latent variable. It is also a particular case of the missing-completely-at random (MCAR) assumption known in the literature [30]. If the latent variable X is made of two components X_o and X_u respectively corresponding to observed and unobserved variables, the MCAR assumption then reads $P(\Gamma_u = \mathcal{X}_u|X_o, X_u) = P(\Gamma_u = \mathcal{X}_u) = \alpha$. Another usual assumption is the missing-at-random one (MAR), that reads $P(\Gamma_u = \mathcal{X}|X_o, X_u) = P(\Gamma_u = \mathcal{X}_u|X_o)$.

- **Coarsening at random assumption (CAR).** This notion was introduced by Heitjan and Rubin (see [22]). According to this assumption, the underlying data do not affect the observations, i.e., an observation $Y = A_j$ is not influenced by the specific value taken by the random variable X inside the set A_j. Mathematically, this condition is expressed as follows:

$$P(Y = A_j|X = a_k) = P(Y = A_j|X = a_{k'}),\ \forall\, k, k',\ \text{with } a_k, a_{k'} \in A_j,\ \forall\, j,$$

or equivalently,

$$P(Y = A_j|X = a_k) = P(Y = A_j|X \in A_j),\ j = 1, \ldots, r.$$

A generalization of this notion has been recently considered in a machine learning context by J. Plass et al. in References [37, 38], where X involves observed and unobserved variables, respectively X_o and X_u. Under the assumption CAR, which generalizes MCAR, set-valued observations Γ_u are assumed to be independent from the true value of X_u. Let us notice that this generalized assumption collapses into the original CAR definition whenever X_o is a constant. The authors of [37, 38] also introduced a kind of "orthogonal" assumption in this machine learning context called "subgroup independence (SI)". Under this

assumption, the set-valued observations Γ_u are assumed not to be influenced by the value of X_o. Testability of this assumption is studied by the authors, a problem that falls out of the scope of our manuscript.

- **Coarsening Completely at Random Assumption (CCAR) ([41]).** The natural definition would be that of imprecise observations Y independent of actual outcomes for X, i.e., $P(X,Y) = P(X)P(Y)$. However Jaeger [26] points out that this definition is problematic because it allows for joint outcomes (x, A) such that $x \notin A$. Enforcing the consistency between set-valued observations ($x(\omega) \in \Gamma(\omega)$), the CCAR assumption must refer to a coarsening variable. Jaeger [26] proposes that $P(X,Y)$ is CCAR with respect to a coarsening variable C if $P(C = c|X = a_k) = P(C = c|X = a_{k'})$ for every pair $a_k, a_{k'} \in \mathcal{X}$ of elements with positive probability, i.e., included in the set $\{x \in \mathcal{X} : P(X = x) > 0\}$.

- **Superset assumption ([24]).** Again we consider observed and unobserved variables with respective domains $\mathcal{X}_o$ and $\mathcal{X}_u$. We assume that when $X = x$ is fixed, the conditional $P(Y = B|X_u = x)$ does not depend on B, whenever $x \in B$ and is 0 otherwise. For every $x \in \mathcal{X}_u$ the number of subsets of $\mathcal{X}$ that contain it is the same. Therefore $P(Y = B|X_u = x) = 1/2^{\#\mathcal{X}_u - 1}$.

  This assumption is dual to the missing-at-random assumption in the sense that in the latter, $Y = B$ is fixed and $P(Y = B|X_u = x)$ does not depend on the choice of x inside B, while in the superset assumption, the set-valued observation induced by $X = x$ can be any superset of x with equal probability. This assumption is often presented as capturing the idea of "lack of information" about the measurement process. In fact, the uniform distribution is used to reflect a balance among outcomes when no information is available. This modeling of ignorance has been questioned in the context of non additive representations of belief (e.g., Shafer belief functions [47], or Walley's imprecise probability theory [52], where a complete lack of information is usually represented by means of a vacuous possibility distribution).

  The superset assumption can be particularised to the case where X_o is a constant. It is then to the original general superset assumption what the original CAR assumption is to the generalized version considered by J. Plass et al. [37, 38] The next example illustrates both assumptions.

Example 1. *A coin is tossed. The random variable $X : \Omega \to \mathcal{X}$, where $\mathcal{X} = \{h, t\}$, represents the result of the toss. We do not directly observe the outcome,*

that is reported by Peter, who sometimes decides not to tell us the result. The rest of the time, the information he provides about the outcome is faithful. Let Y denote the information provided by this person about the result. It takes the "values" $\{h\}$, $\{t\}$ and $\{h,t\}$.

This example corresponds to the following matrix $(M|\mathbf{p})$ where $a_{kj} = p_{.j|k.}, k = 1,2; j = 1,2,3$:

$$\left(\begin{array}{ccc|c} 1-\alpha & 0 & \alpha & p \\ 0 & 1-\beta & \beta & 1-p \end{array}\right)$$

The marginal distribution of X (outcome of the experiment) is given as

- $p_{1.} = P(X = h) = p$,
- $p_{2.} = P(X = t) = 1 - p$.

The joint probability distribution of (X, Y) is therefore determined by:

$$\left(\begin{array}{c|ccc} X\backslash Y & \{h\} & \{t\} & \{h,t\} \\ \hline h & (1-\alpha)p & 0 & \alpha p \\ t & 0 & (1-\beta)(1-p) & \beta(1-p) \end{array}\right)$$

The marginal distribution of Y (information provided by Peter) is thus:

- $p_{.1} = P(Y = \{h\}) = P(Y = \{h\}, X = h) + P(Y = \{h\}, X = t) = (1-\alpha)\cdot p + 0 = (1-\alpha)\cdot p$,
- $p_{.2} = P(Y = \{t\}) = P(Y = \{t\}, X = h) + P(Y = \{t\}, X = t) = 0 + (1-\beta)\cdot(1-p) = (1-\beta)\cdot(1-p)$,
- $p_{.3} = P(Y = \{h,t\}) = P(Y = \{h,t\}, X = h) + P(Y = \{h,t\}, X = t) = \alpha\cdot p + \beta\cdot(1-p)$.

Under the CAR assumption, we have that $\alpha = \beta$, i.e.,

$$P(Y = \{h,t\}|X = h) = P(Y = \{h,t\}|X = t).$$

Let us now consider the (binary) random variable C that takes the value $C = 0$ when $Y = \{h,t\}$ and $C = 1$ otherwise. Let F be defined as $F(x,0) = \{h,t\}$ and $F(x,1) = \{x\}$. The CCAR assumption wrt the coarsening variable C is equivalent to the above mentioned CAR condition.

The superset assumption is more restrictive and assumes that

– $\alpha = P(Y = \{h, t\}|X = h) = P(Y = \{h\}|X = h) = 0.5$ *and*
– $\beta = P(Y = \{h, t\}|X = t) = P(Y = \{t\}|X = t) = 0.5$

and therefore $P(Y = \{h, t\}) = 0.5$. *In words, no matter what the true outcome is (heads or tails) Peter does not give any information about it 50% of the time. The only remaining parameter is* p.

This example demonstrates that the superset assumption represents significant knowledge on the classification process. This is the price paid if we need to provide a stochastic model of the measurement process, as the uniform distribution over all supersets of $\{x\}$ is the least prejudiced probabilistic assumption we can make. The superset assumption is in fact stronger than the CAR assumption: since under the former assumption, $P(Y = B|X_u = x)$ only depends on the number of subsets of $\mathcal{X}$ that contain x, and this number is the same for all x, it follows that $P(Y = B|X_u = x) = P(Y = B|X_u = x')$ for $x, x' \in B$, which is CAR.

2.2. Imprecise observations and their disambiguations

We can alternatively characterize the joint probability distribution of (X, Y) by means of the marginal distribution of Y (observations) and the conditional probability of each result $X = a_k$, knowing that the observation was $Y = b_j$ (or equivalently $\Gamma = A_j$), for every $j = 1, \ldots, r$.

The new matrix $(M'|\mathbf{p}')$ can be written as follows:

$$\left(\begin{array}{ccc|c} p_{1.|.1} & \cdots & p_{m.|.1} & p_{.1} \\ \cdots & \cdots & \cdots & \cdots \\ p_{1.|.r} & \cdots & p_{m.|.r} & p_{.r} \end{array}\right)$$

where

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

Such a matrix also determines the joint probability distribution modeling the underlying generating process plus the connection between true outcomes and incomplete observations. (More specifically, the vector $(p_{.1}, \ldots, p_{.r})^T$ characterizes the observation process while the matrix $M' = (p_{k.|.j})_{k=1,\ldots,m;j=1,\ldots,r}$ represents the conditional probability of X (true outcome) given Y (observation).

Some recent studies on "partial identification" (see [32], for instance) can be somehow related with this framework. They consider situations where parameters of interest are partially identified (see [33]). For instance, when the parameter is real-valued, the so-called "identification region" is a subset of the real line. Imbens and Manski [25] propose the construction of confidence intervals that cover every element in the region with a specific confidence level whose bounds can be computed from sample data.

In this regard, the marginal distribution on $\mathcal{X}$ $(p_{1.}, \ldots, p_{m.})$ is sometimes partially identified, on the basis of our knowledge of the marginal distribution on $\mathcal{Y}$. In fact, let us notice that, according to the total probability theorem, we can write:

$$p_{k.} = \sum_j p_{.j} \cdot p_{k.|.j}, \ \forall\, k = 1, \ldots, m.$$

Furthermore, partial information about the matrix M' is sometimes available. As consequence, confidence regions for any parameter inducing the marginal distribution $(p_{1.}, \ldots, p_{m.})$ on $\mathcal{X}$ can be derived from the above information, on the basis of observable frequencies, that may allow us to provide confidence estimations for the marginal distribution on $\mathcal{Y}$.

Consider for instance a miss-or-observe problem, where Y takes $r = m+1$ values of the form $b_j = \{a_j\}$, $j = 1, \ldots, m$ and $b_{m+1} = \{\mathcal{X}\}$. The equality $P(X = a_k | Y = \{a_k\}) = 1$ holds, for every $k = 1, \ldots, m$. Furthermore, $P(X = a_k | Y = \mathcal{X})$ is known to be included in the unit interval. Based on the above information and on observable marginal frequencies, we can compute set-valued confidence estimations for every $(p_{1.}, \ldots, p_{m.})$ or alternatively, for a parameter determining it.

One example of assumption in this setting is the *Uniform Conditional Distribution Assumption*. In this case we assume that if A_j is observed, all the possible outcomes $a_k \in A_j$ are equally probable, due to a symmetry argument such as the insufficient reason principle. The conditional distribution is then given by:

$$p_{k.|.j} = \begin{cases} \frac{1}{\#A_j}, & \text{if } a_k \in A_j \\ 0 & \text{otherwise.} \end{cases}$$

Knowing the distribution on $\mathcal{Y}$ (which can be estimated from $\mathbf{y}$), the marginal distribution on $\mathcal{X}$ can be estimated as well since:

$$p_{k.} = \sum_{j=1}^{r} p_{k.|.j} \cdot p_{.j} = \sum_{j : A_j \ni a_k} \frac{1}{\#A_j} p_{.j}.$$

Note that this assumption is similar to the superset assumption, exchanging the roles of subsets and elements of $\mathcal{X}$. In the coin-tossing example, it comes down to

assuming

$$P(X = h|Y = \{h, t\}) = P(X = t|Y = \{h, t\}) = 0.5$$

instead of $P(Y = \{h, t\}|X = h)) = P(Y = \{h, t\})|X = t) = 0.5$.

Viewing a probability distribution on $\mathcal{Y}$ as a Dempster-Shafer mass assignment m on $\wp(\mathcal{X})$ as mentioned at the beginning of this section, P_X, as defined above, is the pignistic transform of the belief function induced by the following mass assignment:

$$m(A_j) = p_{.j}, \ \forall\, j = 1, \ldots, r.$$

[49], or yet its Shapley value. More generally, fixing a mixing matrix M' comes down to picking a probability distribution on $\mathcal{X}$ from the convex credal set $\{P_X : P_X(A) \geq Bel(A), \forall A \subseteq \mathcal{X}\}$.

3. Maximum likelihood strategies under incomplete information

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

- $p_{.j} = \sum_{k=1}^{m} p_{kj}$ will denote the mass of $Y = A_j$, 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 entries in 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 $\min\{m \times (r+1), r(m+1)\}$. In other words, the approach uses a parametric model such that a value of θ determines a joint distribution on $\mathcal{X} \times \mathcal{Y}$. When the joint probability measure is the parametric distribution associated to the (vector of) value(s) θ of the parameter, we will respectively use the nomenclature

- $p^{\theta}_{kj} = P(X = a_k, Y = A_j; \theta)$,
- $p^{\theta}_{k.} = \sum_{j=1}^{r} p^{\theta}_{kj} = P(X = a_k; \theta)$ and
- $p^{\theta}_{.j} = \sum_{k=1}^{m} p^{\theta}_{kj} = P(Y = A_j; \theta)$.

Let us consider a sequence $\mathbf{Z} = ((X_1, Y_1), \ldots, (X_N, Y_N))$ of N iid random variables that are "copies" of $Z = (X, Y)$. We will use the nomenclature $\mathbf{z} = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$ to represent 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 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 $\mathcal{X}^N$. Let $(G_1, \ldots, G_N)$ be the sequence of subsets of $\mathcal{X}$ that corresponds to the observed sample $\mathbf{y}$ (namely if $Y_j = y_j$ it corresponds to $X_j \in G_j$).

We can describe any sample $\mathbf{z}$ in frequentist terms assuming exchangeability:

- $n_{kj} = \sum_{i=1}^{N} 1_{\{(a_k, b_j)\}}(x_i, y_i)$ is the number of repetitions of (a_k, b_j) in sample $\mathbf{z}$;
- $\sum_{k=1}^{m} n_{kj} = n_{\cdot j}$ be the number of observations of $b_j = A_j$ in $\mathbf{y}$;
- $\sum_{j=1}^{r} n_{kj} = n_{k\cdot}$ be the number of appearances of a_j in $\mathbf{x}$.

Clearly, $\sum_{k=1}^{m} n_{k\cdot} = \sum_{j=1}^{r} n_{\cdot j} = N$. Let the reader notice that, once a specific sample $\mathbf{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$ has been observed, the number of n_{kj} repetitions of each pair $(a_k, b_j) \in \mathcal{X} \times \mathcal{Y}$ in the sample, can be expressed as a function of $\mathbf{x} = (x_1, \ldots, x_N)$. Moreover, in the following $\mathcal{X}^{\mathbf{y}}$ denote the collection of feasible marginal samples $(x_1, \ldots, x_N)$ of X, in accordance with the observation $\mathbf{y}$:

$$\mathcal{X}^{\mathbf{y}} = \{\mathbf{x} \in \mathcal{X}^N \,:\, x_i \in G_i,\ i = 1, \ldots, N\}$$

and likewise $\mathcal{Z}^{\mathbf{z}}$ denote the collection of feasible (joint) samples $(z_1, \ldots, z_N)$ of Z, in accordance with the observation $\mathbf{y}$:

$$\mathcal{Z}^{\mathbf{y}} = \{\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^N \,:\, z_i = (x_i, y_i) \text{ and } x_i \in G_i,\ i = 1, \ldots, N\}.$$

3.1. Which likelihood function?

We may consider three different likelihood functions (and their respective logarithms), depending on whether we refer to the observed sample $\mathbf{y} = (y_1, \ldots, y_N)$, the sample of (ill-observed) outcomes $\mathbf{x} = (x_1, \ldots, x_N)$, or the complete sample $\mathbf{z}$, and a fourth expression that interprets imprecise observations as events. We will use the following nomenclature to distinguish them from each other:

Visible likelihood function. $\mathbf{p}(\mathbf{y}; \theta) = \prod_{i=1}^{N} p(y_i; \theta)$ denotes the probability of observing $\mathbf{y} \in \mathcal{Y}^N$, assuming that the value of the parameter is θ. It can be alternatively expressed as $\mathbf{p}(\mathbf{y}; \theta) = \prod_{j=1}^{r} (p_{\cdot j}^{\theta})^{n_{\cdot j}}$, where $n_{\cdot j}$ denotes the number of repetitions of $b_j = A_j$ in the sample of size N (the number of times that the reporter says that the

outcome of the experiment belongs to A_j.) The logarithm of this likelihood function will be denoted by

$$L^{\mathbf{y}}(\theta) = \log \mathbf{p}(\mathbf{y};\theta) = \sum_{i=1}^{N} \log p(y_i;\theta) = \sum_{j=1}^{r} n_{\cdot j} \log p^{\theta}_{\cdot j}.$$

We call $\mathbf{p}(\mathbf{y};\theta)$ the *visible likelihood function*, because we can compute it based on the available data only, that is the observed sample $\mathbf{y}$. It is also sometimes called the *marginal likelihood of the observed data* in the EM literature, not to be confused with the *marginal likelihood* in a Bayesian context (see [3], for instance).

Face likelihood function. Note that $L^{\mathbf{y}}(\theta)$ differs from the quantity

$$\lambda(\mathbf{y};\theta) = \prod_{j=1}^{r} P(X \in A_j;\theta)^{n_{\cdot j}},$$

called the "face likelihood" in Ref. [9, 27]. The latter quantity does not refer to the observation process, and replaces the probability of reporting A_j as the result of an observation (i.e. $P(Y = A_j)$) by the probability that the true outcome falls inside the set A_j, $P(X \in A_j)$. In particular, the occurrence of event "$X \in A_j$" is a consequence of, but does not necessarily coincide with the outcome "$Y = A_j$". In our context, $\mathbf{p}(\mathbf{y};\theta)$ represents the probability of occurrence of the result "$(y_1, \ldots, y_N) = \mathbf{y}$", given the hypothesis θ. Therefore given two arbitrary different samples $\mathbf{y} \neq \mathbf{y}'$ the respective events $(y_1, \ldots, y_N) = \mathbf{y}$ and "$(y_1, \ldots, y_N) = \mathbf{y}'$" are mutually exclusive. In contrast, $\lambda(\mathbf{y};\theta)$ denotes the probability of occurrence of the event $(X_1, \ldots, X_N) \in G_1 \times \ldots \times G_N$, where $G_j = A_j$ if $Y_j = A_j$. Events of this form may overlap, in the sense that, given two different samples $\mathbf{y} \neq \mathbf{y}'$, the corresponding events $(X_1, \ldots, X_N) \in G_1 \times \ldots \times G_N$ and $(X_1, \ldots, X_N) \in G'_1 \times \ldots \times G'_N$ are not necessarily mutually exclusive. Therefore $\lambda(\mathbf{y};\theta)$ cannot be regarded as a likelihood in the sense of Edwards ([16]).

However, under the CAR assumption and the assumption of distinctness of parameters, maximizing the face likelihood is the same as maximizing the visible likelihood. Jaeger ([26], Th. 2.18) points out that under CAR, the following equality obtains

$$P(Y = A|X = x) = P(Y = A|X \in A) = \frac{P(Y = A)}{P(X \in A)}.$$

This is easy to see noticing that since from CAR, $P(Y = A|X = x) = k_A$ does not depend on $x \in A$, the equality

$$P(X = x|Y = A) \cdot P(Y = A) = P(Y = A|X = x) \cdot P(X = x)$$

implies

$$\sum_{x \in A} P(X = x | Y = A) \cdot P(Y = A) = k_A \cdot \sum_{x \in A} P(X = x).$$

Hence $P(X \in A | Y = A) \cdot P(Y = A) = k_A \cdot P(X \in A)$, but $P(X \in A | Y = A) = 1$. So, $P(Y = A)$ only depends on the probability of event A on $\mathcal{X}$. We can deduce from this fact that, under the assumption of separability with respect to M (also referred to as distinctness of the parameters) the arguments of the maxima of the visible and the face likelihoods do coincide. Further comments about this equivalence will be provided in Subsection 4.5.

Latent likelihood function. $\mathbf{p}(\mathbf{x}, \theta) = \prod_{i=1}^{N} p(x_i; \theta) = \prod_{i=1}^{m} (p^{\theta}_{k.})^{n_{k\cdot}}$, where $n_{k.}$ denotes the number of occurrences of a_k in the sample $\mathbf{x} = (x_1, \ldots, x_N)$. This sample is never observed directly, by assumption. However it must be in agreement with the observed sample $\mathbf{y}$. Each virtual sample of X in $\mathcal{X}^{\mathbf{y}}$ yields a possible likelihood function $\mathbf{p}(\mathbf{x}, \theta)$ in agreement with the actual observations $\mathbf{y}$. The logarithm of this likelihood function will be denoted by

$$L^{\mathbf{x}}(\theta) = \log \mathbf{p}(\mathbf{x}; \theta) = \sum_{i=1}^{N} \log p(x_i; \theta) = \sum_{k=1}^{m} n_{k.} \log p^{\theta}_{k.},$$

where $n_{k.} = \sum_{j=1}^{r} n_{kj}$ and n_{kj} is such that $\sum_{k=1}^{m} n_{kj} = n_{\cdot j}$, the number of times $Y = A_j$ has been observed. Note that the definition of $n_{k.}$ is in agreement with $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$. We call $\mathbf{p}(\mathbf{x}; \theta)$ the *latent likelihood function*, because $\mathbf{x}$ is not actually observed, nor are the $n_{k.}$'s, since only the $n_{\cdot j}$'s are.

Total likelihood function. $\mathbf{p}(\mathbf{z}, \theta) = \prod_{i=1}^{N} p(z_i; \theta) = \prod_{k=1}^{m} \prod_{j=1}^{r} (p^{\theta}_{kj})^{n_{kj}}$ is the likelihood function induced by the whole artificial sample $\mathbf{z}$, we call the *total likelihood*. Again, it must be in agreement with the observed sample $\mathbf{y} = (y_1, \ldots, y_N)$, which is fixed by assumption. Each virtual sample of Z in $\mathcal{Z}^{\mathbf{y}}$ yields a possible likelihood function $\mathbf{p}(\mathbf{z}, \theta)$ in agreement with the actual observations $\mathbf{y}$. We will denote its logarithm by

$$L^{\mathbf{z}}(\theta) = \log \mathbf{p}(\mathbf{z}; \theta) = \sum_{i=1}^{N} \log p(z_i; \theta) = \sum_{k=1}^{m} \sum_{j=1}^{r} n_{kj} \log p^{\theta}_{kj}.$$

Maximizing $\mathbf{p}(\mathbf{z}, \theta)$ allows us to introduce assumptions on the measurement process, either in terms of imprecisiation or disambiguation. Namely, the conditional probabilities $p_{\cdot j | k.}$ may be known because, for instance, the superset assumption is made. Alternatively, probabilities $p_{\cdot k | \cdot j}$ could be known, which, along with a particular distribution on $\mathcal{Y}$ (to be estimated from observations $\mathbf{y}$), is enough to derive a

concrete distribution on $\mathcal{Z}$ and therefore on $\mathcal{X}$. More generally, there may be some dependence between the process driving the latent variable X and the measurement process driving the actual observations $\mathbf{y}$. In this case, maximizing $L^{\mathbf{z}}(\theta)$ enables this kind of additional information to be accounted for.

Remark 3.1. *In the above expressions, we use the convention* $0^0 = 1$. *In other words, the expression* $\prod_{k=1}^{m}\prod_{j=1}^{r} p_{kj}^{n_{kj}}$ *replaces the formally correct expression*

$$\prod_{(k,j)\in\{1,\dots,m\}\times\{1,\dots,r\}\,:\,n_{kj}\neq 0} p_{kj}^{n_{kj}}.$$

Example 2. *Consider again Example 1, i.e. the coin tossing experiment, assuming for 10 tosses that Peter reports 4 times Heads, 2 times Tails and 4 times nothing. Let us write the four likelihood functions.*

- *Visible likelihood:*
  $\mathbf{p}(\mathbf{y},\theta) = P(\{\{h\}\})^4 \cdot P(\{\{t\}\})^2 \cdot P(\{\{h,t\}\})^4 = [[(1-\alpha)p]^4[(1-\alpha)(1-p)]^2\alpha^4]$
  using the parameters introduced earlier. Note that $P(\{\{h\}\}) + P(\{\{t\}\}) + P(\{\{h,t\}\}) = 1$ *as it is a probability distribution on* $2^{\{h,t\}}$.

- *Face likelihood:* $\mathbf{p}(\mathbf{y},\theta) = P(\{h\})^4 \cdot P(\{t\})^2 = p^4(1-p)^2$ *since* $P(\{h,t\}) = 1$. *Optimizing it comes down to forgetting the missing information.*

- *Hidden Likelihood:* $\mathbf{p}(\mathbf{x},\theta) = p^{4+n_{13}} \cdot (1-p)^{6-n_{13}}$, *where* n_{13} *is the unknown number of times Peter does not report, while the result is Head.*

- *Total likelihood:* $\mathbf{p}(\mathbf{z},\theta) = [[(1-\alpha)p]^4[(1-\alpha)(1-p)]^2(\alpha p)^{n_{13}}(\alpha(1-p))^{4-n_{13}}]$. *The four outcomes obtained by describing both the toss result and the report are evaluated.*

3.2. Maximum likelihood strategies

In this paper we will compare different existing strategies of likelihood maximization, based on a sequence of observations $\mathbf{y} = (y_1,\dots,y_N) \in \mathcal{Y}^N$:

- **Maximizing** $L^{\mathbf{y}}(\theta)$**.** The argument of the maximum of $L^{\mathbf{y}}$ considered as a mapping defined on Θ is called maximum likelihood estimator (MLE) i.e.:

$$\hat{\theta} = \arg\max_{\theta\in\Theta} L^{\mathbf{y}}(\theta) = \arg\max_{\theta\in\Theta} \prod_{j=1}^{r} (p_{\cdot j}^{\theta})^{n_{\cdot j}}.$$

Note that this maximization process does not need any reference to the non-observed variable X. From optimizing $L^{\mathbf{y}}(\theta)$, what is obtained is a probability distribution on $\mathcal{Y}$, which, as already suggested can also be viewed as a Dempster-Shafer mass assignment m_θ on $\wp(\mathcal{X})$, letting $m_\theta(A_j) = p^\theta_{\cdot j}$ for $j = 1, \ldots r$. But a concrete choice of $\theta \in \Theta$ also leads us to select a specific joint distribution on $\mathcal{X} \times \mathcal{Y}$ $(p^\theta_{ij})_{i=1,\ldots,m;j=1,\ldots r.}$ When the argument of the maximum of the log-likelihood function $L^{\mathbf{y}}$ is not unique, the MLE determines a collection of joint distributions on $\mathcal{X} \times \mathcal{Y}$. Under some circumstances, this collection of joint distributions coincides with the credal set associated to a mass function, and such a mass function determines a unique distribution on $\mathcal{Y}$. We will provide a brief discussion about such a situation in Example 3. The EM algorithm [31] is an iterative technique that uses a latent variable X in order to reach a local maximum of $L^{\mathbf{y}}$ when its optimization is tricky. In this case, we also obtain a precise imputation $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$. The latent variable is sometimes fictitious, as in the case of learning a mixture of normal distributions [11].

- **Maximizing** $L^{\mathbf{x}}(\theta)$. This is the genuine goal if one is interested to find the MLE of X despite the imprecise data. However, since the precise sample $\mathbf{x}$ is not available, there is a subset $L^{\mathcal{X}^{\mathbf{y}}}(\theta) = \{L^{\mathbf{x}}(\theta) : \mathbf{x} \in \mathcal{X}^{\mathbf{y}}\}$ of possible likelihood functions [19]. So we must find not only an optimal value of θ, but also an optimal sample $\mathbf{x}$, according to some strategy. There are two obvious strategies that come to mind:

  1. *Maximax strategy*: find a pair $(\mathbf{x}^{**}, \theta^{**}) \in \mathcal{X}^N \times \Theta$ satisfying

  $$(\mathbf{x}^{**}, \theta^{**}) = \arg \max_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}, \theta \in \Theta} L^{\mathbf{x}}(\theta) = \arg \max_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}, \theta \in \Theta} \prod_{k=1}^{m} (p^\theta_{k\cdot})^{n_{k\cdot}}.$$

  It comes down to maximizing an upper log-likelihood function $\overline{L}^{\mathbf{x}}(\theta) = \max\{L^{\mathbf{x}}(\theta) : \mathbf{x} \in \mathcal{X}^{\mathbf{y}}\}$, which can be viewed as an optimistic strategy; it tends to favor distributions with small entropy, under certain conditions, as we shall see later. The maximax technique has been proposed by E. Hüllermeier ([23]) using more general loss functions. His paper makes the point that the choice of an optimal pair $(\mathbf{x}^{**}, \theta^{**})$ leads to a simultaneous selection of a best model together with a disambiguation of the imprecise observations.

  2. *Maximin strategy*: find a pair $(\mathbf{x}_{**}, \theta_{**}) \in \mathcal{X}^N \times \Theta$ satisfying:

  $$(\mathbf{x}_{**}, \theta_{**}) = \arg \max_{\theta \in \Theta} \min_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}} L^{\mathbf{x}}(\theta) = \arg \max_{\theta \in \Theta} \min_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}} \prod_{k=1}^{m} (p^\theta_{k\cdot})^{n_{k\cdot}},$$

where $n_{k\cdot} = \sum_{i=1}^{N} 1_{\{a_k\}}(x_i)$ denotes the number of repetitions of a_k in the sample $\mathbf{x}$. It comes down to maximizing a lower log-likelihood function $\underline{L}^{\mathbf{x}}(\theta) = \min\{L^{\mathbf{x}}(\theta) : \mathbf{x} \in \mathcal{X}^{\mathbf{y}}\}$, which can be viewed as a robust strategy, that copes with the imprecision of the likelihood function; it tends to favor distributions with large dispersions, as we shall see later. The maximin technique has been proposed in Ref. [19].

Note that one might object to these approaches. First, considering $\overline{L}^{\mathbf{x}}(\theta)$ or $\underline{L}^{\mathbf{x}}(\theta)$ requires the comparison of values of $L^{\mathbf{x}}(\theta)$ for several samples $\mathbf{x}$, which maximal likelihood advocates will strongly question. Following them, one cannot compare likelihood functions coming from distinct data sets [16]. However one may reply to it that in the case of our imprecise information setting, two samples $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^{\mathbf{y}}$ not only have the same size, but represent the same data set. This data set is unique but ill-known, and the considered samples are in agreement with the same body of observations $\mathbf{y}$. It is then sure that the true likelihood function lies in the interval $[\underline{L}^{\mathbf{x}}(\theta), \overline{L}^{\mathbf{x}}(\theta)]$. Maximizing one of its bounds is a usual strategy in the face of intervals. We may also use any other strategy that compares intervals, for example the safe but very demanding (if not impossible to reach) partial interval ordering requirement $\underline{L}^{\mathbf{x}}(\theta^\star) > \overline{L}^{\mathbf{x}}(\theta), \forall \theta \neq \theta^\star$.

Hüllermeier [23] justifies the maximax approach by saying that if $\overline{L}^{\mathbf{x}_1}(\theta_1) > \overline{L}^{\mathbf{x}_2}(\theta_2)$ then the sample $\mathbf{x}_1$ is arguably more plausible than the the sample $\mathbf{x}_2$, simply because the first instantiation allows for a much better fit to the model based on θ_1 than the second one based on θ_2. This philosophy leads to disambiguating the imprecise data through the choice of the best model. However, this line of reasoning makes sense if we are sure that the random process generating X follows a model in the prescribed class parameterized by θ. Then it is natural to consider that the most plausible values compatible with the imprecise observations are those which enable a best fit with the class of parameterized model, so that one may select at the same time the best model and the best sample that justifies it. In that case, the resulting disambiguation is a form of data reconciliation [14]. In constrast, if the set of parameterized models is chosen for its computational simplicity and is known to be an approximation of the real phenomenon, the disambiguation rationale of the maximax approach is then not so strong.

- **Maximizing** $L^{\mathbf{z}}(\theta)$. As said earlier this is the natural way to go, if some information regarding the dependence between the latent variable X and its

measurement process is available, for instance the superset or the CAR assumption is made. We can again adopt maximax or maximin strategies, since the full sample $\mathbf{z}$ is not available, and only the observations $\mathbf{y}$ are. There is also an iterative strategy that exploits the links between X and Y, such as the EM algorithm, which maximizes $L^{\mathbf{y}}(\theta)$ via the production of a fake sample $\mathbf{z}$.

1. The maximax strategy aims at finding the pair $(\mathbf{z}^*, \theta^*) \in \mathcal{Z}^N \times \Omega$ that maximizes the function $L^{\mathbf{z}}(\theta)$:

$$(\mathbf{z}^*, \theta^*) = \arg \max_{\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}, \theta \in \Theta} L^{\mathbf{z}}(\theta) = \arg \max_{\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}, \theta \in \Theta} \prod_{k=1}^{m} \prod_{j=1}^{r} (p_{kj}^{\theta})^{n_{kj}}.$$

   It comes down to maximizing an upper log-likelihood function $\overline{L}^{\mathcal{Z}^{\mathbf{y}}}(\theta) = \max\{L^{\mathbf{z}}(\theta) : \mathbf{z} \in \mathcal{Z}^{\mathbf{y}}\}$. The complete sample $\mathbf{z}^*$ also yields an optimal sample $\mathbf{x}^* \in \mathcal{X}^{\mathbf{y}}$ since $L^{\mathbf{z}}(\theta)$ can be viewed as a function $f_{\mathbf{y}} : \mathcal{X}^N \times \Theta \to \mathbb{R}$ that only depends on $\mathbf{x}$. This maximization procedure has been considered in [24] under the superset assumption.
2. It is clear we can similarly envisage the corresponding maximin strategy and find the pair $(\mathbf{x}_*, \theta_*) \in \mathcal{X}^N \times \Omega$ induced by the pair $(\mathbf{z}_*, \theta_*)$ that maximizes the lower log-likelihood function $\underline{L}^{\mathcal{Z}^{\mathbf{y}}}(\theta) = \min\{L^{\mathbf{z}}(\theta) : \mathbf{z} \in \mathcal{Z}^{\mathbf{y}}\}$.

$$(\mathbf{z}^*, \theta^*) = \arg \max_{\theta \in \Theta} \min_{\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}} L^{\mathbf{z}}(\theta) = \arg \max_{\theta \in \Theta} \min_{\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}} \prod_{k=1}^{m} \prod_{j=1}^{r} (p_{kj}^{\theta})^{n_{kj}}.$$

3. Guess an initial value of θ, which enables to construct a fictitious sample $\mathbf{z}$; then, an MLE of θ for this sample can be found, and this process is iterated till convergence. This kind of strategy is adopted by the EM algorithm and tries to find a probability model p^{θ} as close as possible to the empirical distribution of a fake sample $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$, in the sense of Kullback-Leibler divergence [31] (see also [6]).

In this paper, we focus on the maximax and maximin strategies for maximizing $L^{\mathbf{x}}$ and $L^{\mathbf{z}}$.

3.3. Connections between MLE strategies

Under some particular conditions about the matrices M and M', some of the above maximization procedures may coincide. Below, some results are provided. There are two kinds of results: some that relate the total likelihood function $\mathbf{p}(\mathbf{z}; \theta)$

and the latent one $\mathbf{p}(\mathbf{x};\theta)$ under suitable assumptions, and those that relate the total likelihood function $\mathbf{p}(\mathbf{z};\theta)$ and the visible one $\mathbf{p}(\mathbf{y};\theta)$. This is done by introducing assumptions about the incomplete data or the conditional distributions describing the measurement process. It comes down to some information about the matrices M and M'.

A first issue concerns the parameter θ, which so far is used in the three likelihood functions as driving the joint distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, hence the respective marginals on $\mathcal{X}$ and $\mathcal{Y}$. In some situations, X and Y are driven by distinct parameters θ_1, θ_2.

The following result concerns the disambiguation point of view and involves matrix M'.

Definition 1. *We say that the parameter $\theta \in \Theta$ is* separable *with respect to the matrix $(M'|\mathbf{p}')$ if it can be "separated" into two (maybe multidimensional) components $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$ such that $\Theta = \Theta_1 \times \Theta_2$, where $p^{\theta}_{k.|.j}$ and $p^{\theta}_{.j}$ can be respectively written as functions of θ_1 and θ_2.*

Proposition 1. *$\cup_{\mathbf{z}\in\mathcal{Z}^{\mathbf{y}}} \arg\max_{\theta\in\Theta} L^{\mathbf{z}}(\theta) \subseteq \arg\max_{\theta\in\Theta} L^{\mathbf{y}}(\theta)$ provided that θ is separable wrt $(M'|\mathbf{p}')$.*[4]

Proof: Let $\mathbf{y} \in \mathcal{Y}^N$ denote the observed sample. Let us select an arbitrary complete sample $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$. $\mathbf{p}(\mathbf{z};\theta) = \prod_{j=1}^{r}\prod_{k=1}^{m}(p^{\theta}_{kj})^{n_{kj}} = \prod_{j=1}^{r}\prod_{k=1}^{m}(p^{\theta}_{k.|.j}\cdot p^{\theta}_{.j})^{n_{kj}} = \prod_{j=1}^{r}(p^{\theta}_{.j})^{\sum_{k=1}^{m} n_{kj}} \cdot \prod_{j=1}^{r}\prod_{k=1}^{m}(p^{\theta}_{k.|.j})^{n_{kj}} = \prod_{j=1}^{r}(p^{\theta}_{.j})^{n_{.j}} \cdot \prod_{j=1}^{r}\prod_{k=1}^{m}(p^{\theta}_{k.|.j})^{n_{kj}}$. Thus, if θ is separable wrt $(M'|\mathbf{p}')$ we can write:

$$\mathbf{p}(\mathbf{z};\theta) = \prod_{j=1}^{r}(p^{\theta_1}_{.j})^{n_{.j}} \prod_{j=1}^{r}\prod_{k=1}^{m}(p^{\theta_2}_{k.|.j})^{n_{kj}}.$$

Thus, if θ^* is an optimal parameter such that $L^{\mathbf{z}}(\theta^*) = \max_{\theta\in\Theta} L^{\mathbf{z}}(\theta)$, then its projection on Θ_1, $\theta_1^* \in \Theta_1$ must necessarily satisfy the equality $L^{\mathbf{y}}(\theta_1^*) = \max_{\theta_1\in\Theta_1} L^{\mathbf{y}}(\theta_1)$. □

Now let us check the consequence of the uniform conditional distribution assumption.

[4]Remember that $\arg\max_{\theta\in\Theta} L^{\mathbf{z}}(\theta)$ is the set of θ's that maximize $L^{\mathbf{z}}(\theta)$. It may be a singleton, but not an element of Θ.

Proposition 2. *Let* $\mathbf{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$ *denote the observed sample. Let us suppose that* $\prod_{j=1}^{r} \prod_{k:n_{kj}\neq 0} p_{k.|.j}^{n_{kj}}$ *is a value* c *that does not depend on the particular choice of* $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$*, nor on* θ*. Then for every* $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ *we have* $\mathbf{p}(\mathbf{z};\theta) = c\mathbf{p}(\mathbf{y};\theta)$ *and therefore* $\arg\max_{(\mathbf{x},\theta)\in\mathcal{X}^{\mathbf{y}}\times\Theta} \mathbf{p}(\mathbf{z};\theta) = \arg\max_{\theta\in\Theta} \arg\min_{\mathbf{z}\in\mathcal{Z}^{\mathbf{y}}} \mathbf{p}(\mathbf{z};\theta) = \arg\max_{\theta\in\Theta} \mathbf{p}(\mathbf{y};\theta)$.

Proof: Using the previous proof, we already have that

$$\mathbf{p}(\mathbf{z};\theta) = \prod_{j=1}^{r} p_{.j}^{n_{.j}} \prod_{k:n_{kj}\neq 0} (p_{k.|.j})^{n_{kj}} = c\mathbf{p}(\mathbf{y};\theta).$$

□

Proposition 3. *Let* $\mathbf{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$ *denote the observed sample. Let us consider the uniform conditional distribution assumption. Then* $\prod_{j=1}^{r} \prod_{k:n_{kj}\neq 0} p_{.j|k.}^{n_{kj}}$ *is a value* c *that does not depend on the particular choice of* $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$*, nor on* θ*.*

Proof: Under the uniform conditional distribution assumption we have:

$$p_{k.|.j} = \begin{cases} \frac{1}{\#A_j} & \text{if } a_k \in A_j \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\prod_{j=1}^{r} \prod_{k:n_{kj}\neq 0} p_{k.|.j}^{n_{kj}} = \prod_{j=1}^{r} \frac{1}{\#A_j}^{\sum_{k:n_{kj}\neq 0} n_{kj}} = \prod_{j=1}^{r} \left(\frac{1}{\#A_j}\right)^{n_{.j}}.$$

□

Corollary 4. *If the uniform conditional distribution assumption holds then, for all* $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$*,* $\mathbf{p}(\mathbf{z};\theta) = c\mathbf{p}(\mathbf{y};\theta)$*, where* c *does not depend either on the particular* $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ *nor on* θ *and therefore*

$$\arg \max_{(\mathbf{x},\theta)\in\mathcal{X}^{\mathbf{y}}\times\Theta} \mathbf{p}(\mathbf{z};\theta) = \arg \max_{\theta\in\Theta} \arg \min_{\mathbf{z}\in\mathcal{Z}^{\mathbf{y}}} \mathbf{p}(\mathbf{z};\theta) = \arg \max_{\theta\in\Theta} \mathbf{p}(\mathbf{y};\theta).$$

The next results concern the imprecisiation process and involve matrix M:

Definition 2. *We say that the parameter* $\theta \in \Theta$ *is* separable *with respect to the matrix* $(M|\mathbf{p})$ *if it can be "separated" into two (maybe multidimensional) components* $\theta_3 \in \Theta_3$*,* $\theta_4 \in \Theta_4$ *such that* $\Theta = \Theta_3 \times \Theta_4$ *and* $p_{.j|k.}^{\theta}$ *and* $p_{k.}^{\theta}$ *can be respectively written as functions of* θ_3 *and* θ_4*.*

This type of separability corresponds to the notion of "distinct parameters" in the sense of Heitjan and Rubin ([22]) in the context of coarse data, and Little and Rubin ([30]) in the context of missing data.

Proposition 5. *If θ is separable wrt $(M|\mathbf{p})$ then, given a specific sample $\mathbf{x} \in \mathcal{X}^N$ and the corresponding $\mathbf{z} \in (\mathcal{X} \times \mathcal{Y})^N$ induced by $\mathbf{x}$ and $\mathbf{y}$, $\arg\max_{\theta\in\Theta} L^{\mathbf{z}}(\theta) \subseteq \arg\max_{\theta\in\Theta} L^{\mathbf{x}}(\theta)$.*

Proof: The proof of this result is similar to the one given in Proposition 1. □

Remark 3.2. *Proposition 5 assumes a fixed sample $\mathbf{x} \in \mathcal{X}^N$. Let us notice that the separability wrt M does not imply that the respective solutions of both maximax problems, $\theta^{**} = \arg\max_{\theta\in\Theta} \overline{L}^{\mathcal{X}^y}(\theta)$ and $\theta^* = \arg\max_{\theta\in\Theta} \overline{L}^{\mathcal{Z}^y}(\theta)$, coincide. They can be attained for different samples in $\mathcal{X}^{\mathbf{y}}$.*

Proposition 6. *Let $\mathbf{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$ denote the observed sample. Let us suppose that $\prod_{k=1}^m \prod_{j:n_{kj}\neq 0} p_{.j|k.}^{n_{kj}}$ is a value c that does not depend on the particular choice of $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$, nor on θ. Then, for all $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$ and the corresponding $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ we have $\mathbf{p}(\mathbf{z};\theta) = c\mathbf{p}(\mathbf{x};\theta)$ and therefore $\arg\max_{(\mathbf{x},\theta)\in\mathcal{X}^{\mathbf{y}}\times\Theta} \mathbf{p}(\mathbf{x};\theta) = \arg\max_{(\mathbf{x},\theta)\in\mathcal{X}^{\mathbf{y}}\times\Theta} \mathbf{p}(\mathbf{z};\theta)$.*

Proof:

$$\mathbf{p}(\mathbf{z};\theta) = \prod_{k=1}^{m} \prod_{j:n_{kj}\neq 0} p_{kj}^{n_{kj}} = \prod_{k=1}^{m} \prod_{j:n_{kj}\neq 0} (p_{.j|k.} \cdot p_{k.})^{n_{kj}} =$$

$$\prod_{k=1}^{m} \prod_{j:n_{kj}\neq 0} p_{.j|k.}^{n_{kj}} \cdot p_{k.}^{n_{kj}} = \prod_{k=1}^{m} p_{k.}^{\sum_{j=1}^{r} n_{kj}} \cdot \prod_{k=1}^{m} \prod_{j:n_{kj}\neq 0} (p_{.j|k.})^{n_{kj}} =$$

$$\prod_{k=1}^{m} p_{k.}^{n_{k.}} \cdot c = p(\mathbf{x};\theta) \cdot c.$$

□

Proposition 7. *Let $\mathbf{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$ denote the observed sample. Let us suppose that $\{A_1, \ldots, A_r\}$ forms a partition of $\mathcal{X}$ or that the superset assumption is satisfied. Then $\prod_{k=1}^m \prod_{j:n_{kj}\neq 0} p_{.j|k.}^{n_{kj}}$ is a constant c.*

Proof:

On one hand, we can easily check that, if $\{A_1, \ldots, A_r\}$ forms a partition of $\mathcal{X}$ then $\prod_{k=1}^m \prod_{j:n_{kj}\neq 0} p_{.j|k.}^{n_{kj}} = 1$. Now, let us check that the above condition holds under the superset assumption. Under the superset assumption we have already shown that $p_{.j|k.} = 2^{m-1}$ if $a_k \in A_j$, and 0 otherwise. Therefore, $\prod_{k=1}^m \prod_{j:n_{kj}\neq 0} p_{.j|k.}^{n_{kj}} = \prod_{k=1}^m \left(\frac{1}{2^{m-1}}\right)^{\sum_{j:n_{kj}\neq 0} n_{kj}} = \prod_{k=1}^m \left(\frac{1}{2^{m-1}}\right)^{n_{k.}} = \left(\frac{1}{2^{m-1}}\right)^N$. □

Corollary 8. *If any of the following conditions is satisfied:*

- *$\{A_1, \ldots, A_r\}$ forms a partition of $\mathcal{X}$*

- *The superset assumption holds*

then $\mathbf{p}(\mathbf{z};\theta) = c\mathbf{p}(\mathbf{x};\theta)$ *and therefore* $\arg\max_{(\mathbf{x},\theta)\in\mathcal{X}^N\times\Theta} p(\mathbf{x};\theta) = \arg\max_{(\mathbf{x},\theta)\in\mathcal{X}^N\times\Theta} L^z(\theta)$. *Furthermore* $c = 1$ *in the first case.*

Most approaches in statistical inference insist on the necessity to have a statistical model of the observation process. In that case the natural likelihood function to maximize is $\mathbf{p}(\mathbf{z},\theta)$. In contrast, if we maximise the latent likelihood function $\mathbf{p}(\mathbf{x},\theta)$ directly with respect to both $\mathbf{x}$ and θ, we in some way give up the idea of providing a statistical model for the observation (imprecisiation) process. The superset assumption can be used to justify the use of $\mathbf{p}(\mathbf{x},\theta)$ by providing such a statistical model, since in that case maximizing $\mathbf{p}(\mathbf{z},\theta)$ is the same as maximizing $\mathbf{p}(\mathbf{x},\theta)$ (this is the message apparently carried by the authors of [24], for instance). When no information about the measurement process is available, it is to us an open question whether one should maximize $\mathbf{p}(\mathbf{x},\theta)$ or $\mathbf{p}(\mathbf{z},\theta)$ with respect to both $\mathbf{x}$ and θ. Indeed, the two corresponding MLE may differ as indicated on examples in the following.

4. Comparing the maximum likelihood strategies on examples

Based on some of the results provided in Subsection 3.3, we can compare the various choices of likelihood functions (maximization of $L^{\mathbf{y}}(\theta), L^{\mathbf{x}}(\theta), L^{\mathbf{z}}(\theta)$) in an imprecise environment, on the basis of the acceptability of results obtained by their maximization on a number of prototypical examples. These examples, shed light on the nature of the maximin and the maximax strategies, as opposed to maximizing the visible likelihood function. We consider several settings where imprecise data occur in various ways: the possible observations may form a partition of space $\mathcal{X}$, the data is either precise or missing, and the general case where imprecise data may overlap.

4.1. Observations forming a partition: separable case

The first example illustrates the situation where imprecise data form a partition of $\mathcal{X}$ and the parameter θ is separable.

Example 3. *Consider the random experiment that consists on rolling a dice. We do not know whether the dice is fair or not. Suppose that the person that rolls it just tells us whether the outcome is even or odd. Let* X *be the random variable denoting the actual outcome of the dice roll (*$a_i = i$ *for* $i = 1,\ldots,6$*) and let* Y *be a binary variable taking the values* b_1 *(odd) and* b_2 *(even). So, the collection* $\{A_1 = \{1,3,5\}, A_2 = \{2,4,6\}\}$ *determines a partition of the whole set of possible*

outcomes $\mathcal{X} = \{1, \ldots, 6\}$. *Let the 6-dimensional vector* $(p_{1.}, \ldots, p_{6.})$ *represent the actual (unknown) probability distribution of* X, *where* $p_{i.} = P(X = a_i), i = 1, \ldots, 6$ *and* $p_{6.} = 1 - \sum_{i=1}^{5} p_{i.}$.

Let $p_{.2} = \pi = p_{2.} + p_{4.} + p_{6.}$ *denote the probability of getting an even number. The six-dimensional vector* $(p_{1.}, \ldots, p_{6.})$ *can be alternatively expressed as*

$$((1-\pi)\alpha_1, \pi\beta_1, (1-\pi)\alpha_2, \pi\beta_2, (1-\pi)(1-\alpha_1-\alpha_2), \pi(1-\beta_1-\beta_2))$$

where $\alpha = (\alpha_1, \alpha_2)$ *and* $\beta = (\beta_1, \beta_2)$ *respectively denote* $\alpha_i = \frac{p_{(2i-1).}}{1-\pi}$ *and* $\beta_i = \frac{p_{2i.}}{\pi}$. *Therefore, we can identify* $(p_{1.}, \ldots, p_{6.})$ *with the parameter* $\theta = (\theta_1, \theta_2) = (\pi, (\alpha, \beta))$ *and see that it is separable with respect to the matrix* $(M'|\mathbf{p}')$ *given as*

$$\left(\begin{array}{cccccc|c} \alpha_1 & 0 & \alpha_2 & 0 & 1-\alpha_1-\alpha_2 & 0 & 1-\pi \\ 0 & \beta_1 & 0 & \beta_2 & 0 & 1-\beta_1-\beta_2 & \pi \end{array}\right)$$

since (α, β) *determines the conditional distribution of* X *given* Y *while* π *determines the marginal distribution of* Y.

Maximizing $L^{\mathbf{y}}(\theta)$**.** *Based on a sample of* $n_{.1}$ *occurrences of* b_1 *and* $n_{.2}$ *occurrences of* b_2 *in a sample of* $N = n_{.1} + n_{.2}$ *trials, any parameter* $\hat{\theta}$ *of the form* $(\hat{\theta}_1, \theta_2) = \left(\frac{n_{.2}}{N}, (\alpha, \beta)\right)$, *for an arbitrary pair* $(\alpha, \beta))((\alpha_1, \alpha_2), (\beta_1, \beta_2))$ *is an MLE from* $\mathbf{y}$. *Such an MLE determines a single marginal distribution on* $\mathcal{Y}$, $(\hat{p}_{.1}, \hat{p}_{.2}) = (\frac{n_{.1}}{N}, \frac{n_{.2}}{N})$. *Due to the separability, such an MLE does not constrain any of the conditional distributions* $X|Y = b_1$ *or* $X|Y = b_2$. *Furthermore, neither does the parameter* $\theta_2 = (\alpha, \beta)$ *constrain them.Therefore, the collection of marginal distributions on* $\mathcal{X}$ *associated to the collection of MLE from* $\mathbf{y}$ *is the credal set associated to the mass function* $m_{\hat{\theta}_1}(\{1,3,5\}) = \frac{n_{.1}}{N}$, $m_{\hat{\theta}_1}(\{2,4,6\}) = \frac{n_{.2}}{N}$.

Maximax strategy. *Assume a fixed probability vector* $(p_{1.}, \ldots, p_{6.})$ *and let* i^* *be the index of the greatest of* $p_{1.}, p_{3.}, p_{5.}$, j^* *be the index of the greatest of* $p_{2.}, p_{4.}, p_{6.}$. *The maximum of*

$$\mathbf{p}(\mathbf{z}; \theta) = \mathbf{p}(\mathbf{x}; \theta) = p_{1.}^{n_{11}} \cdot p_{2.}^{n_{22}} \cdot p_{3.}^{n_{31}} \cdot p_{4.}^{n_{42}} \cdot p_{5.}^{n_{51}} \cdot p_{6.}^{n_{62}},$$

where $n_1 = n_{11} + n_{13} + n_{15}$, $n_2 = n_{22} + n_{24} + n_{26}$ *is reached for any selection* $\mathbf{x}^* \in \mathcal{X}^N$ *where* $n_{.1}$ identical *outcomes corresponding to odd number* i^* *and the remaining ones,* $n_{.2} = N - n_{.1}$, identical *outcomes corresponding to even number* j^* *are obtained. Now, maximizing wrt* $(p_{1.}, \ldots, p_{6.})$ *leads to let* $p_{i^*} = p_{j^*} = 1$. *This means any choice* $\theta^* = (\frac{n_{.2}}{N}, \alpha^*, \beta^*) \in \Theta$, *with* $\alpha^* \in \{(1,0), (0,1), (0,0)\}$ *and* $\beta^* \in \{(1,0), (0,1), (0,0)\}$ *is optimal. In words, the maximum of* $L^{\mathbf{x}}(\theta) = L^{\mathbf{z}}(\theta)$ *is attained for all situations*

where only two facets of the dice, one of them corresponding to an even number and the other one representing an odd one, are possible. The parameter θ^* determines the probability distribution of X where the probabilities of those two numbers are respectively π and $1-\pi$. The sample $\mathbf{x}^*$ is the only sample compatible with $\mathbf{y}$ where there are $n_{.2}$ repetitions of the even number and $n_{.1}$ repetitions of the odd one. Taking into account the separability of θ wrt M' and according to Proposition 1, the collection of optimal selections of the parameter:

$$\left\{\left(\frac{n_{.2}}{N}, \alpha^*, \beta^*\right), \text{ with } \alpha^* \in \{(1,0),(0,1),(0,0)\} \text{ and } \beta^* \in \{(1,0),(0,1),(0,0)\}\right\}$$

is included in the set of maximum likelihood estimators for $L^{\mathbf{y}}$,

$$\left\{\left(\frac{n_{.2}}{N}, (\alpha_1,\alpha_2), (\beta_1,\beta_2)\right) \,:\, \alpha_1,\alpha_2,\beta_1,\beta_2 \in [0,1]\right\}.$$

Maximin strategy. *Assume a fixed $(p_{1.},\ldots,p_{6.})$ and let i^* be the index of the least of $p_{1.}, p_{3.}, p_{5.}$ j^* be the index of the least of $p_{2.}, p_{4.}, p_{6.}$. Since, $L^{\mathbf{x}}(p_{1.},\ldots,p_{6.}) = \sum_{i=1}^{6} n_i \log p_{i.}$, we easily observe that $L^{\mathbf{x}}(\theta)$ reaches its minimum when $n_{i^*.} = n_{1.} + n_{3.} + n_{5.} = n_{.1}$ and $n_{j^*} = n_{2.} + n_{4.} + n_{6.} = n_{.2}$ and therefore, $\underline{L}^{\mathbf{x}}(p_{1.},\ldots,p_{6.}) = n_{.1}\log p_{i^*} + n_{.2}\log p_{j^*.}$. Now let us maximise $\underline{L}^{\mathbf{x}}(\theta)$ with respect to $(p_{1.},\ldots,p_{6.})$. For a fixed value of $\pi = p_{2.}+p_{4.}+p_{6.}$; this maximum is reached at $p_{i^*} = p_{1.} = p_{3.} = p_{5.} = \frac{1-\pi}{3}$ and $p_{j^*} = p_{2.} = p_{4.} = p_{6.} = \frac{\pi}{3}$ and therefore*

$$\max_{(p_{1.},\ldots,p_{6.})} \underline{L}^{\mathbf{x}}(p_{1.},\ldots,p_{6.}) = \max_{\pi\in[0,1]} n_{.1}\log\left(\frac{\pi}{3}\right) + n_{.2}\log\left(\frac{1-\pi}{3}\right).$$

The argument of the maximum is therefore the vector $(p_{1^*},\ldots,p_{6^*})$ where $p_{i^*} = \frac{n_{.1}}{3N}$ $i = 1,3,5$ and $p_{i^*} = \frac{n_{.2}}{3N}$ $i = 2,4,6$. In words, the maximin approach assigns the same probability to all odd outcomes, and also to the even outcomes. Note that it does not suggest a best imputation of $(n_{1.},\ldots,n_{6.})$, as all corresponding samples $\mathbf{x}$ in agreement with the $n_{.1}$ observations of odds and $n_{.2}$ observations of evens are equiprobable.

The above example shows that the maximax strategy supplies, as most plausible imputations, one odd and one even number only (any pair can be chosen). More generally, the same reasoning applies to the general case where $\{A_1,\ldots A_r\}$ forms a partition and $\mathcal{P}_{\mathbf{y}} \subseteq \mathcal{P}_\Theta$, where $\mathcal{P}_\Theta$ denotes the set of parameterized joint distributions (dependent on θ) and $\mathcal{P}_{\mathbf{y}}$ the set of joint distributions whose marginals on $\mathcal{Y}$ agree with the empirical distribution induced by $\mathbf{y}$. Namely:

Proposition 9. *If the imprecise observations* $\{A_1, \dots A_r\}$ *form a partition of* $\mathcal{X}$ *and* $\mathcal{P}_{\mathbf{y}} \subseteq \mathcal{P}_\Theta$ *then the maximax strategy applied to* $L^{\mathbf{x}}(\theta)$ *with respect to* $\mathbf{x}$ *and* θ *is reached for any sample of* X *consisting of* $n_{\cdot j}$ *copies of the same element* $a_k \in A_j$ *for each* A_j *(i.e.,* $\frac{n_{k\cdot}}{N} = \frac{n_{kj}}{N} = \frac{n_{\cdot j}}{N}$*), the corresponding parameterized distribution on* X *is the one that coincides with the empirical one induced by this sample, i.e.,* $\hat{p}_{k\cdot} = \frac{n_{\cdot j}}{N}$ *for* $a_k \in A_j$*, and* $\hat{p}_{k'\cdot} = 0$ *for* $a_{k'} \neq a_k$*,* $a_{k'} \in A_j$*.*

Proof: The proof is the same as in the case when $r = 2$ given in the above example. □

In other words the maximax solution presupposes that each occurrence of $Y = A_j$ underlies the same outcome $X = a_k \in A_j$ (for instance, we may assume the midpoint of A_j is the most likely, if A_j is an interval). If the number of observations $Y = A_j$ is large, it becomes a very strong assumption, whereby constant imprecise observations hide constant outcomes. One may argue on the contrary, that the coarse sensor is insensitive to the variability of the outcomes, which explains its apparent constant measurements. An interesting consequence follows:

Corollary 10. *If the imprecise observations* $\{A_1, \dots A_r\}$ *form a partition of* $\mathcal{X}$ *and* $\mathcal{P}_{\mathbf{y}} \subseteq \mathcal{P}_\Theta$*, then the maximax strategy applied to* $L^{\mathbf{x}}(\theta)$ *yields a distribution on* $\mathcal{X}$ *minimizing the entropy among distributions whose marginals on* $\mathcal{Y}$ *coincide with the empirical distribution induced by the observed sample* $\mathbf{y}$*.*

We have a companion result for the maximin strategy.

Proposition 11. *If the imprecise observations* $\{A_1, \dots A_r\}$ *form a partition of* $\mathcal{X}$ *and* $\mathcal{P}_{\mathbf{y}} \subseteq \mathcal{P}_\Theta$*, then the maximin strategy applied to* $L^{\mathbf{x}}(\theta)$ *with respect to* $\mathbf{x}$ *and* θ *is reached for a parameterized distribution whose restriction to each* A_j *is uniform and such that* $p^\theta_{k\cdot} = \frac{n_{\cdot j}}{N \# A_j}$ *if* $a_k \in A_j$*. All samples* $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$ *are equally probable.*

Proof: The proof is the same as in the case when $r = 2$ given in the above example. □

An interesting consequence, opposite to the maximax case, follows:

Corollary 12. *If the imprecise observations* $\{A_1, \dots A_r\}$ *form a partition of* $\mathcal{X}$ *and* $\mathcal{P}_{\mathbf{y}} \subseteq \mathcal{P}_\Theta$*, then the maximin strategy applied to* $L^{\mathbf{x}}(\theta)$ *yields a distribution on* $\mathcal{X}$ *maximizing the entropy among distributions whose marginals on* $\mathcal{Y}$ *coincide with the empirical distribution induced by the observed sample* $\mathbf{y}$*.*

In the terminology of belief functions, the estimated probability distribution on $\mathcal{X}$ is the maximum entropy distribution among the probabilities dominating the belief function, that here coincides with the pignistic transform of the belief function.

The maximin strategy in the partition case seems to presuppose incomplete information hides pure randomness corresponding to maximal entropy. The idea can be that the sensor reporting an element of the partition is too coarse to perceive the variability of X inside the $n_{\cdot j}$ actual outcomes in A_j. In other words, an interval-valued observation reflects a set of precise observations inside each interval. The maximin strategy tries to propose a model that covers all these situations. On the contrary, the maximax strategy explains the $n_{\cdot j}$ observations A_j as resulting from the occurrence of the same outcome a_k $n_{\cdot j}$ times. Clearly, these two strategies are in some sense extreme.

The above results will be generalized in Propositions 13 and 14 to the case where the imprecise observations $\{A_1, \dots A_r\}$ do not necessarily form a partition of $\mathcal{X}$.

4.2. Observations forming a partition: non-separable case

The next (well-known) example is a case where the MLE for $L^{\mathbf{y}}$ is not the one obtained from $L^{\mathbf{x}}$ via any of the above strategies, and the parameter θ is not separable. Namely, the visible likelihood $L^{\mathbf{y}}$ determines the actual one $L^{\mathbf{x}}$ to a large extent.

Example 4. *Let us consider the following example by Dempster et al. in Reference [11] under the light of our analysis. It is based on a former example by Rao. There is a sample of 197 animals distributed into four categories, so that the observed data consist of:*

$$n_{\cdot 1} = 125, n_{\cdot 2} = 18, n_{\cdot 3} = 20, n_{\cdot 4} = 34.$$

Suppose that the first category is in fact a mixture of two sub-categories, but we do not have information about the number of individuals observed from each of them. On the other hand, a genetic model for the population specifies the following restrictions about the five categories: $p_{11} = 0.5$, $p_{12} = p_{\cdot 4}$, $p_{\cdot 2} = p_{\cdot 3}$. *If we use the notation:* $p_{12} = 0.25\pi = p_{\cdot 4}$ *and* $p_{\cdot 2} = 0.25(1-\pi) = p_{\cdot 3}$, *the corresponding matrix* $(M'|\mathbf{p}')$ *is given as*

$$\left(\begin{array}{ccccc|c} \frac{0.5}{0.5+0.25\pi} & \frac{0.25\pi}{0.5+0.25\pi} & 0 & 0 & 0 & 0.5+0.25\pi \\ 0 & 0 & 1 & 0 & 0 & 0.25(1-\pi) \\ 0 & 0 & 0 & 1 & 0 & 0.25(1-\pi) \\ 0 & 0 & 0 & 0 & 1 & 0.25\pi \end{array}\right)$$

and only depends on a single parameter. Dempster et al. ([11]) maximize $L^{\mathbf{y}}(\pi) = \log\left[(0.5+0.25\pi)^{125}(0.25(1-\pi))^{18}(0.25(1-\pi))^{20}(0.25\pi)^{34}\right]$ *by making its derivative vanish. The maximum likelihood estimator of* π *with respect to* $L^{\mathbf{y}}$ *is found to be* $\hat{\pi} = 0.6268214980$. *They show that using the EM algorithm, this value can be retrieved as the limit of a sequence of values obtained by a iterative procedure.*

The likelihood associated to a sample $\mathbf{x}$ *compatible with the observed sample* $\mathbf{y}$ *will be of the form:*

$$L^{\mathbf{x}}(\theta) = \log\left[0.5^{n_{11}}(0.25\pi)^{n_{12}}(0.25(1-\pi))^{18}(0.25(1-\pi))^{20}(0.25\pi)^{34}\right],$$

where $n_{11} + n_{12} = 125$. *There, the collection* $\{A_{11}, A_{12}, A_2, A_3, A_4\}$ *forms a partition of* $\mathcal{X}$ *(the set representing five categories of animals). Therefore, according to Corollary 8,* $L^{\mathbf{x}}(\pi)$ *and* $L^{\mathbf{z}}(\pi)$ *do coincide for every* π. *Thus, the strategies based on their respective maximizations lead to the same optimal pairs. Consider the maximax strategy. It yields* $(\mathbf{x}^*, \pi^*)$, *where* $\pi^* = 0.4722222$. *The corresponding optimal sample* $\mathbf{x}^*$ *is the (only) sample inducing the empirical distribution* $(\frac{125}{197}, 0, \frac{18}{197}, \frac{20}{197}, \frac{34}{197})$. *Thus, according to this criterion, all the 125 observed outcomes from group one are assumed to belong to the first subgroup, again a questionable choice. The maximin strategy leads to the opposite choice, that is* $\mathbf{x}_* = (0, \frac{125}{197}, \frac{18}{197}, \frac{20}{197}, \frac{34}{197})$ *since* $0.25\pi < 0.5$. *The maximin optimal* π_* *is* $159/197 = 0.8071$. *Besides, the parameter* π *is not separable wrt* $(M'|\mathbf{p}')$. *Consequently, none of the values* $\pi^* = 0.4722222$, $\pi_* = 0.8071$ *coincides with the maximum likelihood estimator of* π *with respect to* $L^{\mathbf{y}}$, $\hat{\pi} = 0.6268214980$. *The latter yields another imputation that is not as extreme as the samples obtained with the other strategies.*

This example reveals again the extreme nature of the maximin and maximax strategies. The empirical distribution associated to the samples $\mathbf{x}^*$ and $\mathbf{x}_*$ are respectively the least and the most entropic in the set $\mathcal{P}^{\mathbf{y}}$.The connection with entropy maximization and minimization of the corresponding optimal parametric distributions is less clear because the likelihood functions are constrained by a single parameter. As $p_{11} > p_{12}, \forall\pi$, the maximin strategy leads to grant the maximal number of outcomes to the least probable situation.

4.3. Miss-or-observe case with or without the superset assumption

In the following example, either the outcome is observed or the observation is just missing. So, the collection of imprecise observations does not form a partition, but it contains the singletons in $\mathcal{X}$ and $\mathcal{X}$ itself. We compare the CAR and superset assumptions with the case when no information on the measurement process is made. For simplicity, we assume $\mathcal{X}$ has only two elements, so that the "superset assumption" can be made. Moreover the parameter θ is separable.

Example 5 (The CAR and the superset assumptions). *Let us take up the situation of Example 1. A coin is tossed. The random variable $X : \Omega \to \mathcal{X}$, where $\mathcal{X} = \{h, t\}$, represents the result of the toss. We do not directly observe the outcome. The outcome is reported by Peter. Let us suppose that the coin is tossed $N = 10$ times, and that Peter faithfully reports 4 heads, 2 tails and he does not report anything for the remaining 4 times. Let us consider again the parameter $\theta = (p, \alpha, \beta)$, where $p = P(X = h)$, $\alpha = P(Y = \{h, t\}|X = h)$ and $\beta = P(Y = \{h, t\}|X = t)$. This is clearly a case where θ is separable. As seen in Example 1, the CAR assumption yields $\alpha = \beta$ and the superset assumption yields $\alpha = \beta = 0.5$. In words, no matter what the true outcome is (heads or tails) Peter does not give any information about it 50% of the time. The only remaining parameter is p.*

The set-valued mapping Γ underlying variable Y takes the "values" $\{h\}$, $\{t\}$ and $\{h, t\}$ with respective probabilities $(1 - \alpha)p$, $(1 - \beta)(1 - p)$ and $\alpha p + \beta(1 - p)$. Under the CAR assumption, these three probabilities are $(1 - \alpha)p$, $(1 - \alpha)(1 - p)$ and α. Furthermore, under the superset assumption, they are $0.5p$, $0.5(1 - p)$ and 0.5, respectively.

Maximizing $L^{\boldsymbol{y}}(\theta)$. *Let us maximize the visible log-likelihood under each of these assumptions.*

- The CAR assumption. *We have to find the argument of the maximum of*

$$L^{\boldsymbol{y}}(\alpha, p) = \log\Big[[(1 - \alpha)p]^4[(1 - \alpha)(1 - p)]^2\alpha^4\Big] = \log[(1 - \alpha)^6\alpha^4p^4(1 - p)^2].$$

The maximum is attained for $(\hat{\alpha}, \hat{p}) = (0.4, 2/3)$. According to this criterion (MLE based on the observed sample) those tosses where Peter does not report the result are not taken into account. This idea fits the initial assumptions about the conditional probability of Y with respect to X. In fact, according to the CAR assumption, $P(Y = \{h, t\}|X = h) = P(Y = \{h, t\}|X = t) = P(Y = \{h, t\})$ or, in other words, $Y = \{h, t\}$ and $X = h$ are independent events. Therefore, we deduce that $X = h$ is also independent from the complementary of $Y = \{h, t\}$, $Y \in \{\{h\}, \{t\}\}$ and therefore:

$$P(X = h) = P(X = h|Y \in \{\{h\}, \{t\}\}) = \frac{P(X = h, Y \in \{\{h\}, \{t\}\})}{P(Y \in \{\{h\}, \{t\}\})} =$$

$$\frac{P(X = h, Y = \{h\})}{P(Y \in \{\{h\}, \{t\}\})} = \frac{P(Y = \{h\})}{P(Y = \{h\}) + P(Y = \{t\})}.$$

The above two equalities mean that, according to the CAR assumption, the ill-informed tosses are assumed to correspond to each of the possible outcomes

of the coin (heads or tails) in the same proportion as the observed ones. So, $P(X = h)$ coincides, under the CAR assumption, with the expected proportion of times when Peter says "heads" divided by the expected proportion of times when he provides us information about the result, whatever it is, heads or tails. Thus, it makes sense to estimate the probability of heads $P(X = h)$ by means of the quotient $\frac{0.4}{0.4+0.2} = 2/3$, representing the quotient of the respective frequencies of the events $Y = \{h\}$ and $Y \in \{\{h\}, \{t\}\}$. In this case, the statistical model obtained on $\mathcal{Y}$ (such that $p_{.1} = 2(1-\alpha)/3 = 0.4$ since $\alpha = 0.4$, $p_{.2} = 0.2, p_{.3} = 0.4$) is in agreement with the empirical distribution.

- The superset assumption. *Under this assumption, the visible log-likelihood function is:*

$$L^{\mathbf{y}}(p) = \log[(0.5p)^4 \cdot [0.5(1-p)]^2 \cdot 0.5^4].$$

As $L^{\mathbf{y}}(p)$ is equal to $f(p) = \log[p^4(1-p)^2]$ up to an additive constant, it reaches its maximum at $\hat{p} = 2/3$. The restrictions on the set of statistical models due to the superset assumption imply that the distribution on $\mathcal{Y}$ induced by the MLE distribution on $\mathcal{X}$ ($p = 2/3$) is not in agreement with the empirical distribution on $\mathcal{Y}$. The former is such that $p_{.1} = 2(1-\alpha)/3 = 1/3$ since $\alpha = 0.5$; likewise $p_{.2} = 1/6, p_{.3} = 0.5$.

Maximax strategy on $L^{\mathbf{x}}(\theta)$. *Let us compare the above estimation of θ with the maximax strategy-based estimation: Given a sample $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$, we can easily observe that*

$$L^{\mathbf{x}}(\theta) = n_{1.} \log p + (N - n_{1.}) \log(1-p) = (4 + n_{13}) \log p + (6 - n_{13}) \log(1-p),$$

(where n_{13} denotes the number of tosses where the result is "heads" and Peter did not report.) Thus, according to this strategy, we calculate the argument

$$\max_{\mathbf{x} \in \mathcal{X}^{\mathbf{y}}, \theta \in \Theta} L^{\mathbf{x}}(\theta) = \max_{n_{13} \in \{1,2,3,4\}, p \in [0,1]} (4 + n_{13}) \log p + (6 - n_{13}) \log(1-p).$$

For a specific $n_{13} \in \{0, 1, 2, 3, 4\}$, the maximum of $f_{n_{13}}(p) = (4 + n_{13}) \log p + (6 - n_{13}) \log(1-p)$ is attained at $p^*(n_{13}) = \frac{4+n_{13}}{10}$. Therefore, the above maximum coincides with the maximum: $\max_{n_{13}=1,2,3,4} f\left(\frac{4+n_{13}}{10}\right)$. We can check that it is attained for $n_{13}^* = 4$ and $p^* = \frac{4+n_{13}^*}{10} = 0.8$. According to this procedure, the optimal estimate of p is 0.8 and the optimal sample is the one that assigns the result "heads" to all those tosses where Peter has not reported the result. It is clear that the distribution on $\mathcal{Y}$ induced by the maximax strategy ($p = 0.8$ on $\mathcal{X}$) is not in agreement with the empirical distribution on $\mathcal{Y}$, whatever the assumption made. Under CAR, the

former is of the form $(0.8(1-\alpha), 0.2(1-\alpha), \alpha)$ *which cannot match the empirical* $(0.4, 0.2, 0.4)$ *whatever the choice of* α*. Under the superset assumption* $\alpha = 0.5$ *and the induced distribution on* $\mathcal{Y}$ *is* $p_{.1} = 0.4, p_{.2} = 0.1, p_{.3} = 0.5$*.*

Maximax strategy on $L^{\mathbf{z}}(\theta)$. *Under the CAR assumption, the expression of the log-likelihood* $L^{\mathbf{z}}$ *for a specific* $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ *is as follows:*

$$L^{\mathbf{z}}(\alpha, p) = \log\left[[(1-\alpha)p]^4[(1-\alpha)(1-p)]^2(\alpha p)^{n_{13}}[\alpha(1-p)]^{4-n_{13}}\right] =$$

$$\log\left[(1-\alpha)^6\alpha^4 \cdot p^{4+n_{13}}(1-p)^{6+n_{13}}\right],$$

which is maximized at $(n_{13}, \alpha, p) = (4, 0.4, 0.8)$*. Again, all non-informed outcomes are understood as heads as for the maximax strategy on* $L^{\mathbf{x}}$ *under CAR.The result is slightly different from the former case, as this time* α *is set to 0.4, while it remains free under the maximisation of* $L^{\mathbf{x}}$*.*

Under the superset assumption, following Corollary 8, $L^{\mathbf{x}}(\theta) - L^{\mathbf{z}}(\theta)$ *is a constant. The maximax procedure aiming to maximize* $L^{\mathbf{z}}(\theta)$ *will give the same result as maximizing* $\arg\max L^{\mathbf{x}}(\theta)$*, with* $\theta = p$ *that does not depend on* (α, β) *(conditional probabilities are fixed under the superset assumption).*

Maximin strategy on $L^{\mathbf{x}}(\theta)$. *The maximin approach consists of considering all log-likelihood functions* $L^{\mathbf{x}}_k(p) = (4+n_{13})\log p + (6-n_{13})\log(1-p)$ *with* $0 \leq n_{13} \leq 4$*. The approach consists in finding for each value of* p *the complete data that minimizes* $L^{\mathbf{x}}(p)$*. Since* $L^{\mathbf{x}}_{n_{13}}(p)$ *is of the form* $n_{13}\log\frac{p}{(1-p)} + a$*, it is easy to see that if* $p < 1/2$*, the minimum* $\underline{L}(p)$ *is reached for* $n_{13} = 4$*, and if* $p > 1/2$*, it is reached for* $n_{13} = 0$*. So, it is* $8\log p + 2\log(1-p)$ *if* $p < 1/2$ *and* $4\log p + 6\log(1-p)$ *otherwise. So* $\underline{L}(p)$ *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. However, the corresponding imputation* $(n_{1.}, n_{2.}) = (4+n_{13}, 6-n_{13})$ *is totally free, as all sequences of outcomes of* N *tosses in agreement with* $\mathbf{y}$ *are equally probable in this case. Under CAR, none of the possible distributions on* $\mathcal{Y}$ *in agreement with* $p = 0.5$ *is in agreement with the empirical distribution (since it predicts the same number of heads and tails). It implies that under the superset assumption, the distribution on* $\mathcal{Y}$ *induced by the maximin strategy (using* $p = 0.5$ *on* $\mathcal{X}$ *and* $\alpha = 0.5$*) is not in agreement with the empirical distribution on* $\mathcal{Y}$*. The former is such that* $p_{.1} = 0.25, p_{.2} = 0.25, p_{.3} = 0.5$*.*

Maximin strategy on $L^{\mathbf{z}}(\theta)$. *As we have pointed out above, under the CAR assumption, the expression of the log-likelihood* $L^{\mathbf{z}}$ *for a specific* $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ *is as follows:*

$$L^{\mathbf{z}}(\alpha, p) = \log\left[[(1-\alpha)p]^4[(1-\alpha)(1-p)]^2(\alpha p)^{n_{13}}[\alpha(1-p)]^{4-n_{13}}\right] =$$

$$\log\left[(1-\alpha)^6\alpha^4 \cdot p^{4+n_{13}}(1-p)^{6+n_{13}}\right].$$

Now, taking into account the above course of reasoning about the maximin strategy, our estimation under this strategy is $(\alpha, p) = (0.4, 0.5)$ *and the corresponding imputation would be completely free. Note that this time, the induced distribution on* $\mathcal{Y}$ *is in agreement with the empirical distributions, as it is precisely the unique such one among the possible distributions obtained from the maxmin solution on* $L^{\mathbf{x}}$ *under CAR. Under the superset assumption, following Corollary 8,* $L^{\mathbf{x}}(\theta) - L^{\mathbf{z}}(\theta)$ *is a constant, and therefore this strategy coincides with the maximin approach from last paragraph.*

The obtained results in this example are summarized on Table 1. It shows that under the superset assumption, maximizing $L^{\mathbf{y}}(\theta)$ *sometimes comes down to doing away with missing information, while maximizing* $L^{\mathbf{x}}(\theta)$ *over* $\mathbf{x}$ *and* θ *(or equivalently* $L^{\mathbf{x}}(\theta)$*) with the maximax strategy implicitly relies on the sample where the unobserved outcomes correspond to the most probable among the visible ones. On the other hand, the maximin approach maximizing* $\underline{L}^{\mathbf{x}}(\theta)$ *in this example, comes down again to maximizing entropy among distributions compatible with the missing information, thus turning ignorance into maximal randomness. Clearly the three strategies here yield different results.*

Table 1: Comparison of maximum likelihood strategies

Method	(α, β, p)	on $\mathcal{X} = \{h, t\}$	on $\mathcal{Y} = \{\{h\}, \{t\}, \{h, t\}\}$	n_{13}
$L^{\mathbf{y}}$ CAR	$(0.4, 0.4, 2/3)$	$(2/3, 1/3)$	$(0.4, 0.2, 0.4)$	free
$L^{\mathbf{y}}$ superset	$(0.5, 0.5, 2/3)$	$(2/3, 1/3)$	$(1/3, 1/6, 0.5)$	free
$L^{\mathbf{x}}$ max-max CAR	$(\alpha, \alpha, 0.8)$	$(0.8, 0.2)$	$(0.8(1-\alpha), 0.2(1-\alpha), \alpha)$	4
$L^{\mathbf{x}}$ max-max superset	$(0.5, 0.5, 0.8)$	$(0.8, 0.2)$	$(0.4, 0.1, 0.5)$	4
$L^{\mathbf{z}}$ max-max CAR	$(0.4, 0.4, 0.8)$	$(0.8, 0.2)$	$(0.48, 0.12, 0.5)$	4
$L^{\mathbf{z}}$ max-max superset	$(0.5, 0.5, 0.8)$	$(0.8, 0.2)$	$(0.4, 0.1, 0.5)$	4
$L^{\mathbf{x}}$ max-min CAR	$(\alpha, \alpha, 0.5)$	$(0.5, 0.5)$	$(0.5(1-\alpha), 0.5(1-\alpha), \alpha)$	free
$L^{\mathbf{x}}$ max-min superset	$(0.5, 0.5, 0.5)$	$(0.5, 0.5)$	$(0.25, 0.25, 0.5)$	free
$L^{\mathbf{z}}$ max-min CAR	$(0.5, 0.5, 0.5)$	$(0.5, 0.5)$	$(0.25, 0.25, 0.5)$	free
$L^{\mathbf{y}}$	(α, β, p)	$(p, 1-p)$	$(0.4, 0.2, 0.4)$	free
$L^{\mathbf{z}}$ max-max	$(0.5, 0.0, 0.8)$	$(0.8, 0.2)$	$(0.4, 0.2, 0.4)$	4
$L^{\mathbf{x}}$ max-max	$(\alpha, \beta, 0.8)$	$(0.8, 0.2)$	$(0.8(1-\alpha), 0.2(1-\beta), 0.8\alpha + 0.2\beta)$	4
$L^{\mathbf{x}}$ max-min	$(\alpha, \beta, 0.5)$	$(0.5, 0.5)$	$(0.5(1-\alpha), 0.5(1-\beta), 0.5(\alpha+\beta))$	free

Remark. The results of the maximin strategy on $L^{\mathbf{x}}(\theta)$ can be extended to the case of N coin tosses, only N_o of which are observed, and k among the latter are heads. Then $L^{\mathbf{x}}(\theta) = (k + n_{13}) \log p + (N - k - n_{13}) \log(1 - p)$, where $0 \leq n_{13} \leq N_o$ is the number of unobserved heads. The same analysis as above would lead to estimate $p = \frac{k+N-N_o}{N}$ if $k > N - k$, using the maximax strategy, and the corresponding imputation being $(n_{1.}, n_{2.}) = (k + N - N_o, N_o - k)$. Under the maximin strategy, we again get $p = 0.5$, and no preferred imputation. If $k = N - k$ then the maximax strategy yields maximal likelihood by interpreting all N_o unobserved outcomes as heads or as tails, leading to any of the two estimates $p = \frac{k}{N}$ or $p = \frac{k+N-N_o}{N}$.

Example 6 (No assumption on the measurement process). *Let us consider the same Example 1 without the superset assumption. Now, we do not assume anything about the proportion of times where Peter does not inform us about the results of the tosses. Such a proportion does exist, but it is not necessarily equal to* 0.5 *and furthermore it may depend on the true result of the coin. In other words, we do not put any assumption about the "imprecisiation process". Formally speaking, the conditional probability* $\alpha = P(Y = \{h, t\}|X = h)$ *may differ from* $\beta = P(Y = \{h, t\}|X = t)$ *and each of them can be any number in the unit interval.*

Maximization of $L^{\mathbf{y}}$. *Taking into account the marginal distribution of* Y *calculated at the beginning of this section, we can write:*

$$L^{\mathbf{y}}(\theta) = \log([(1 - \alpha)p]^4 \cdot [(1 - \beta)(1 - p)]^2 \cdot [\alpha p + \beta(1 - p)]^4)$$

where $\theta = (p, \alpha, \beta)$*. The maximum is attained for any* θ *satisfying the constraints:*

- $(1 - \alpha)p = 0.4$ *and*
- $(1 - \beta)(1 - p) = 0.2$,

i.e., the set of tuples for which the respective probabilities of $Y = \{h\}$*,* $Y = \{t\}$ *and* $Y = \{h, t\}$ *are 0.4, 0.2 and 0.4. In other words, in this general setting, the MLE on* $\mathcal{X}$ *constrained by* $\mathbf{y}$ *is not unique, contrary to the case when the superset assumption is made, as in the previous example where we fixed the conditional probability of* Y *given* X*. Here the collection of MLE obtained from* $\mathbf{y}$ *corresponds to the set of all the joint distributions whose marginals on* $\mathcal{Y}$ *coincide with the empirical distribution* $(0.4, 0.2, 0.4)$*. It corresponds all probability distributions compatible with the mass assignment* $m(\{h\}) = 0.4, m(\{t\}) = 0.2, m(\{h, t\}) = 0.4$ *on* $\{h, t\}$*. Clearly, in this case, the set of statistical models defined by the three parameters is compatible with the empirical distributions on observations.*

Maximization of $L^{\mathbf{z}}$. *On the other hand, the likelihood based on* $\mathbf{z}$ *is:*

$$L^{\mathbf{z}}(p,\alpha,\beta) = \log([(1-\alpha)p]^4(\alpha p)^{n_{13}}[(1-\beta)(1-p)]^2[\beta(1-p)]^{4-n_{13}})$$

where $0 \leq n_{13} \leq n_{\cdot 3} = 4$. *Regrouping terms, we can alternatively express it as:*

$$L^{\mathbf{z}}(p,\alpha,\beta) = \log([p^{4+n_{13}}(1-p)^{6-n_{13}}]) + \log(\cdot[\alpha^{n_{13}}(1-\alpha)^4]) + \log([\beta^{4-n_{13}}(1-\beta)^2]).$$

Thus, given a specific sample $\mathbf{z}$, *the optimal parameter* $\theta^*(n_{13})$ *is obtained as:*

$$p^*(n_{13}) = \frac{4+n_{13}}{10}; \quad \alpha^*(n_{13}) = \frac{n_{13}}{n_{13}+4}; \quad \beta^*(n_{13}) = \frac{4-n_{13}}{6-n_{13}},$$

where n_{13} *denotes the number of times where the result of the coin was "heads" in the true sample* $\mathbf{z}$, *but Peter did not report it. The optimal pair* (z^*, θ^*) *is attained for the sample* $\mathbf{z}^* \in \mathcal{Z}^{\mathbf{y}}$ *satisfying the equality* $n_{13} = 4$ *(the sample where all those tosses of the coin where Peter provided no information corresponds to "heads"). The optimal* $\theta^* = (p^*, \alpha^*, \beta^*)$ *is therefore* $p^* = 0.8$, $\alpha^* = 0.5$, $\beta^* = 0$. *Let the reader notice that, for every* $n_{13} \in \{0,1,2,3,4\}$, *vector* $(p^*(n_{13}), \alpha^*(n_{13}), \beta^*(n_{13}))$ *is an MLE based on* $\mathbf{y}$, *i.e., it satisfies the constraints:*

- $(1-\alpha^*(n_{13}))p^*(n_{13}) = 0.4$ *and*
- $(1-\beta^*(n_{13}))(1-p^*(n_{13})) = 0.2$.

According to the maximax strategy, the optimal estimation of the probability of heads is 0.8. *On the other hand, according to this maximization criterion, it is postulated that Peter decides 50% of the times to inform us about the result when it is "heads", but he always decides us to inform us about it when it is "tails". This is the same distribution on* $\mathcal{X}$ *as under the superset assumption, however, with different parameters driving the joint distributions, as here* $\alpha^* = 0.5$, $\beta^* = 0$, *so that the empirical and theoretical distributions on* $\mathcal{Y}$ *are the same (see Table 1).*

Finding the lower distribution $\underline{L}^{\mathbf{z}}$ *and the maximin estimate of the parameter* θ *is trickier.*

Maximization of $L^{\mathbf{x}}$. *Finally, as we already pointed out in Example 5,* $L^{\mathbf{x}}(\theta) = (4+n_{13})\log p + (6-n_{13})\log(1-p)$ *depends neither on* α *nor on* β. *According to this, we easily deduce that the argument of the maximum of* $L^{\mathbf{x}}(\theta)$ *is the same as in the last example. Thus given a sample* $\mathbf{x}$ *inducing the frequency* n_{13} *of times where the result was heads and Peter did not report it, the optimal value of* p *is* $p^*(n_{13}) = \frac{4+n_{13}}{10}$. *The argument of the maximum* $\max_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}},\theta\in\Theta} L^{\mathbf{x}}(\theta)$ *therefore corresponds to* $n_{13}^* = 4$

and $p^*(n_{13}^*) = 0.8$ *and leaves the conditional probabilities* α *and* β *move freely in the unit interval. According to the maximax strategy on* $L^{\mathbf{x}}$*, the probability of "heads" is again estimated to be* 0.8*. Nevertheless, nothing is postulated about the probability that Peter does not inform us about the result, and about the relation between his decision and the true result. However, enforcing the identity between the empirical and parametric distributions on* $\mathcal{Y}$ *enforce* $\alpha = 0.5$ *and* $\beta = 0$*.*

The maximin strategy on $L^{\mathbf{x}}$ *also yields the same solution (*$p = 0.5$*) as when the superset assumption holds, this time with freedom on the choice of* α *and* β *(as these parameters are not involved in the optimisation). Enforcing the identity between the empirical and parametric distributions on* $\mathcal{Y}$ *enforces* $\alpha = 0.2$ *and* $\beta = 0.6$*.*

These three optimal solutions (each of them attending a particular criterion) match the results given in Propositions 1 and 5. We can easily observe that the parameter is separable with respect to both matrices $(M|\mathbf{p})$ and $(M'|\mathbf{p}')$. According to Proposition 1,

$$\{\arg\max_{\theta\in\Theta} L^{\mathbf{z}}(\theta) \,:\, \mathbf{z} \in \mathcal{Z}^{\mathbf{y}}\} \subseteq \{\arg\max_{\theta\in\Theta} L^{\mathbf{y}}(\theta)\}.$$

Furthermore, according to Proposition 5, given a specific sample $\mathbf{x} \in \mathcal{X}^N$, the set $\arg\max_{\theta\in\Theta} L^{\mathbf{z}}(\theta)$ is included in $\arg\max_{\theta\in\Theta} L^{\mathbf{x}}(\theta)$.

4.4. Overlapping observations

Finally consider an example with pure overlap of imprecise observations.

Example 7. *Consider the same dice setting as in Example 3, but now take a sample of* N *tosses of the dice and assume that the reporter has told us* n_1 *of the times that the result was less than or equal to* 3 *and the remaining* $n_2 = N - n_1$ *tosses, he told us that it was greater than or equal to 3. After each toss, when the actual result* (X) *is 3, the reporter needs to make a decision. Let us assume that such a decision does not depend on the previous trials of the dice. In other words, let us assume that* Y_n *is independent from* $(X_1, Y_1), \ldots, (X_{n-1}, Y_{n-1})$*- Furthermore, let us assume that the conditional probability* $P(Y_n = y_1 | X_n = 3)$ *is a fixed number* $\alpha \in [0, 1]$ *for every trial,* $n = 1, \ldots, N$*. According to this information, we can easily check that* $((X_1, Y_1), \ldots, (X_N, Y_N))$ *is a sequence of* N *i.i.d. vectors. The joint distribution of* (X, Y) *can be written as a function of* $(p_1, \ldots, p_6)$ *and* α*, since it is determined by the following matrix:* $(M|\mathbf{p})$:

$$\left(\begin{array}{cc|c} 1 & 0 & p_1 \\ 1 & 0 & p_2 \\ \alpha & 1-\alpha & p_3 \\ 0 & 1 & p_4 \\ 0 & 1 & p_5 \\ 0 & 1 & p_6 \end{array}\right)$$

corresponding to the joint probability

Y, X	*1*	*2*	*3*	*4*	*5*	*6*
y_1	p_1	p_2	$\alpha\, p_3$	*0*	*0*	*0*
y_2	*0*	*0*	$(1-\alpha)\, p_3$	p_4	p_5	p_6

Suppose that we toss the dice $N = 1000$ *times and the reporter tells us* $n_{.1} = 300$ *times that the result was less than or equal to 3. The remaining* $n_{.2} = 700$ *times he tells us that the result was greater than or equal to 3. Let* θ *denote the vector* $(p_1, p_2, p_3, p_4, p_5, p_6, \alpha)$*. The likelihood function based on the observed sample* $\mathbf{y}$ *can be written as*[5]*:*

$$L^{\mathbf{y}}(\theta) = (p_1 + p_2 + \alpha p_3)^{300} \cdot ((1-\alpha)p_3 + p_4 + p_5 + p_6)^{700}.$$

Such a function is maximized for any vector θ *satisfying the constraints:*

$$p_1 + p_2 + \alpha p_3 = 0.3 \text{ and } (1-\alpha)p_3 + p_4 + p_5 + p_6 = 0.7.$$

The other likelihood functions take the form

$$L^{\mathbf{x}}(\theta) = \prod_{i \neq 3} p_{i.}^{n_{i.}} \cdot p_3^{n_{31}+n_{32}} \text{ and } L^{\mathbf{z}}(\theta) = \prod_{i \neq 3} p_{i.}^{n_{i.}} \cdot p_{31}^{n_{31}} \cdot p_{32}^{n_{32}}$$

with $n_{1.} = n_{11}, n_{2.} = n_{21}, n_{4.} = n_{24}, n_{5.} = n_{25}, n_{6.} = n_{26}$*. The maximax value of* $L^{\mathbf{x}}$ *is clearly obtained (with* $\mathbf{p}(\mathbf{x}^*; \theta^*) = 1$*) for* $p_3 = 1$ *and* $n_{31} = n_{.1}$ *,* $n_{32} = n_{.2}$*, while there are many maximax fake samples* $\mathbf{z}$ *consisting in* n_1 *times any* $i \in \{1, 2, 3\}$*, and* n_2 *times any* $k \in \{3, 4, 5, 6\}$*, the corresponding parameters being* $p_{i1} = \frac{n_1}{N}$ *and* $p_{k2} = \frac{n_1}{N}$*. It includes the maximax solution for* $L^{\mathbf{x}}$ *since it is obtained when* $i = k = 3$*.*

[5]And not, as pointed out in Section 3.1,

$$P(A_1)^{300} \cdot P(A_2)^{700} = (p_1 + p_2 + p_3)^{300} \cdot (p_3 + p_4 + p_5 + p_6)^{700},$$

whose maximum is reached for $p_3 = 1$.

This example confirms the trend of the maximax solution to disambiguate the data assuming the same imprecise observations to underlie the same outcome. Namely, when all observations A_j overlap, it is clear that the maximax strategy on $L^{\mathbf{x}}$ comes down to a disambiguation that selects an element in the intersection and a maximum probability assignment to it. In the example the result of the maximax strategy looks debatable because it comes down to assuming a deterministic dice. It would make sense if as part of our knowledge, the underlying process was not die tossing, but a single number hidden in a box and that the reporter opens each time. From the obtained reports it is then clear that the most likely value is 3. It again corresponds to the idea of minimizing entropy as shown by the following result.

Let $\mathcal{P}_\Theta$ again denote the set of parameterized joint distributions (dependent on θ) and $\mathcal{P}_{\mathbf{y}}$ the set of joint distributions whose marginals on $\mathcal{Y}$ agree with the empirical distribution induced by $\mathbf{y}$. Let $\mathcal{P}_\Theta^{\mathcal{X}}$ and $\mathcal{P}_{\mathbf{y}}^{\mathcal{X}}$ denote the respective collections of their marginals on $\mathcal{X}$.

Proposition 13. *If $\mathcal{P}_{\mathbf{y}}^{\mathcal{X}} \subseteq \mathcal{P}_\Theta^{\mathcal{X}}$ then the argument of the maximax strategy with respect to $L^{\mathbf{x}}$ is the collection of pairs in the form $(\mathbf{x}^*, \theta^*)$ where:*

- *The empirical distribution induced by $\mathbf{x}^*$ on $\mathcal{X}$ attains the minimum entropy (among the set of all the feasible samples $\mathcal{X}^{\mathbf{y}}$).*
- *The marginal of P_{θ^*} on $\mathcal{X}$ coincides with such an empirical distribution.*

Proof: First notice that the set $\arg\max_{(\mathbf{x},\theta)\in\mathcal{X}^{\mathbf{y}}\times\Theta} L^{\mathbf{x}}(\theta)$ coincides with the set $\arg\max_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}}} \arg\max_{\theta\in\Theta} L^{\mathbf{x}}(\theta)$. Now, for a fixed $\mathbf{x} \in \mathcal{X}^y$ inducing the empirical distribution $(\frac{n_{1.}}{N}, \ldots, \frac{n_{M.}}{N})$ we can write $L^{\mathbf{x}}(\theta) = \sum_{i=1}^m \frac{n_{i.}}{N} \log p_{i.}^{\theta}$. According to Gibbs' inequality, we therefore get $L^{\mathbf{x}}(\theta) \leq \sum_{i=1}^m \frac{n_{i.}}{N} \log \frac{n_{i.}}{N}$. Furthermore, according to the hypotheses, the empirical distribution $(\frac{n_{1.}}{N}, \ldots, \frac{n_{M.}}{N})$ belongs to $\mathcal{P}_\Theta^{\mathcal{X}}$. Thus, given $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$ inducing such an empirical distribution, $\arg\max_{\theta\in\Theta} L^{\mathbf{x}}(\theta) = \theta^*(\mathbf{x})$, where $p_{i.}^{\theta^*(\mathbf{x})} = \frac{n_{i.}}{N},\ \forall\, i = 1, \ldots, m$. Now, we need to maximize the expression $L^{\mathbf{x}}(\theta^*(\mathbf{x})) = \sum_{i=1}^m \frac{n_{i.}}{N} \log \frac{n_{i.}}{N}$ with respect to the first argument, which consists in finding a sample $\mathbf{x}^* \in \mathcal{X}^{\mathbf{y}}$ inducing an empirical distribution on $\mathcal{X}$ with the minimum possible entropy. □

A similar result can be obtained, using the same kind of proof, in relation to the maximax strategy associated to $L^{\mathbf{z}}$:

Proposition 14. *If $\mathcal{P}_{\mathbf{y}} \subseteq \mathcal{P}_\Theta$ then the argument of the maximax strategy with respect to $L^{\mathbf{z}}$ is the collection of pairs in the form $(\mathbf{z}^{**}, \theta^{**})$ where:*

- *The empirical distribution induced by* $\mathbf{z}^{**}$ *on* $\mathcal{X} \times \mathcal{Y}$ *attains the minimum entropy (among the set of all the feasible samples* $\mathcal{Z}^{\mathbf{y}}$*).*
- *The distribution* $P_{\theta^{**}}$ *on* $\mathcal{X} \times \mathcal{Y}$ *coincides with such an empirical distribution.*

The problem of computing the minimal entropy distribution in a convex set of probabilities has been addressed in Ref. [1], and can be applied to find the maximax solution for the total or latent likelihood function optimization.

The maximin solution is less easy to obtain in this case, but one can show it does maximize entropy on $\mathcal{P}_{\mathbf{y}}$. Problems of the form $\max_{\theta\in\Theta} \min_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}}} L^{\mathbf{x}}(\theta)$ are well-known in the framework of game theory [51] and lead to the notion of Nash equilibrium. The major issue is to check whether conditions under which this expression is equal to $\min_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}}} \max_{\theta\in\Theta} L^{\mathbf{x}}(\theta)$ are satisfied. The inequality $\max_{\theta\in\Theta} \min_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}}} L^{\mathbf{x}}(\theta) \leq \min_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}}} \max_{\theta\in\Theta} L^{\mathbf{x}}(\theta)$ always holds. That the other inequality holds was recently proved in [20].

The maxmin strategy on the hidden likelihood function considers the problem $\max_{\theta} \min_{\mathbf{n}\in\mathcal{X}^{\mathbf{y}}} f(\mathbf{n}, \mathbf{p})$, where function f has the form: $f(\mathbf{n}, \mathbf{p}) = \sum_{k=1}^{m} n_{k\cdot} \log p_{k\cdot}$, and $\mathbf{p}$ is the unknown distribution on $\mathcal{X}$, $\mathbf{n}$ is the count vector of a sample on $\mathcal{X}$ with components $n_{k\cdot}$ compatible with $\mathbf{y}$ ($\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$ for the corresponding sample). It is shown in [20] that $\max_{\theta} \min_{\mathbf{n}\in\mathcal{X}^{\mathbf{y}}} f(\mathbf{n}, \mathbf{p}) = \min_{\mathbf{n}\in\mathcal{X}^{\mathbf{y}}} \max_{\theta} f(\mathbf{n}, \mathbf{p})$ and the following result is obtained:

Proposition 15. *The optimal solution to the maxmin optimisation of the hidden likelihood has maximal entropy, namely it is the solution to:* $\max_{\boldsymbol{n}} -\sum_{k=1,\ldots,m} \frac{n_{k\cdot}}{N} \cdot \log \frac{n_{k\cdot}}{N}$*. under conditions that* $\boldsymbol{n}$ *corresponds to a sample* $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$*.*

So the maxmin optimisation of the hidden likelihood comes down to finding the maximum entropy probability in the credal set $\mathcal{P}_{\mathbf{y}}$ a problem already addressed in the past by [1, 2]. Similar considerations can be devised about the maxmin strategy for the total likelihood function. The connection between maximal entropy, game theory, and minimizing worst-case expected loss (of which maximum likelihood is a special case) has been studied more generally by Grünwald and Dawid [18].

4.5. Maximizing the face likelihood

Finally let us reconsider the face likelihood of $\lambda(\mathbf{y}; \theta) = \prod_{j=1}^{r} P(X \in A_j; \theta)^{n_{\cdot j}}$, after Jaeger [27]. As said earlier, in this criterion, the set A_j is not an outcome for Y, it is a more general event for X. Then the face likelihood is not always a likelihood for the space $\mathcal{X}$ stricto sensu. If we consider A_j as an outcome, we should consider it as a singleton in the set of possible results $2^{\mathcal{X}}$. Note that even if $A_i \cap A_j \neq \emptyset$, we do have that $\{b_i\} \cap \{b_j\} = \emptyset$ (distinct outcomes are always mutually exclusive,

following Edwards [16]). Nevertheless, the maximization of the face likelihood will come down to standard likelihood maximization in three cases.

- The case where the range $\{A_1, \ldots, A_r\}$ of Γ 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 the face likelihood $\prod_{j=1}^{r} P(X \in A_j; \theta)^{n_{.j}}$ coincides with the visible likelihood $\mathbf{p}(\mathbf{y}; \theta)$.

- The miss-or-observe situation when the the range of Γ contains either singletons from $\mathcal{X}$ or $\mathcal{X}$ itself. Then $\lambda(\mathbf{y}; \theta) = (\prod_{i=1}^{k} P(Y = \{a_i\}; \theta)^{n_{.i}}) \cdot P(X \in \mathcal{X}; \theta)^{n_{.k+1}}) = \prod_{i=1}^{k} P(X = a_i; \theta)^{n_{.i}}$ since $P(X \in \mathcal{X}; \theta) = 1$. In other words, the face likelihood is the standard likelihood on X based only on precise observations, dropping the missed ones.

- The case where there is CAR and distinctness of the parameters (separability with respect to $(M|p)$). Heitjan and Rubin ([22]) prove that, under this assumptions, the optimal argument of the face likelihood coincides with that of the visible likelihood function. In fact, taking into account the fact that $Y = A_j$ implies $X \in A_i$, i.e., $P(Y = A_i | X \notin A_i) = 0$, the visible likelihood can be decomposed as follows under those assumptions:

$$\mathbf{p}(\mathbf{y}; \theta) = \prod_{j=1}^{r} P(Y = \{A_j\}; (\theta_3, \theta_4))^{n_{.j}} =$$

$$\left(\prod_{j=1}^{r} P(Y = A_j | X \in A_j; \theta_3) \cdot P(X \in A_j; \theta_4) \right)^{n_{.j}} =$$

$$\prod_{j=1}^{r} P(Y = A_j | X \in A_j; \theta_3)^{n_{.j}} \cdot \prod_{j=1}^{r} P(X \in A_j; \theta_4)^{n_{.j}} =$$

$$\prod_{j=1}^{r} P(Y = A_j | X \in A_j; \theta_3)^{n_{.j}} \cdot \lambda(\mathbf{y}; \theta_4).$$

  Therefore,

$$\arg\max_{\theta_4 \in \Theta_4} \left(\arg\max_{\theta_3 \in \Theta_3} \mathbf{p}(\mathbf{y}; \theta_3, \theta_4) \right) = \arg\max_{\theta_4 \in \Theta_4} \lambda(\mathbf{y}; \theta_4).$$

In the case where the sets $A_1, \ldots, A_r$ do overlap the face likelihood and the visible one do not necessarily coincide, as shown in the following example.

Example 8. *Let us consider again the situation described in Example 7, where on a sample of N tosses of the dice, 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 face likelihood is $\lambda(\mathbf{y};\theta) = (p_1+p_2+p_3)^{n_1} \cdot (p_3+p_4+p_5+p_6)^{n_2}$ with $\sum_{i=1}^{6} p_i = 1$. We can easily observe that, like in the case of the maximax strategy on $L^{\mathbf{x}}(\theta)$, the face likelihood reaches its maximum ($\lambda(\mathbf{y};\theta) = 1$) for any vector θ satisfying the constraint $p_3 = 1$. But such a prediction of θ is not a reasonable estimate for θ. The visible likelihood function is of the form $\mathbf{p}^{\mathbf{y}}(\theta) = (p_1 + p_2 + \alpha p_3)^{n_1} \cdot [\alpha p_3 + p_4 + p_5 + p_6)]^{n_2}$, and optimizing it yields different results (see Example 7).*

More generally, in the extreme situation where the imprecise observations completely overlap ($\cap_{i=1}^{r} A_i \neq \emptyset$) then the face likelihood is maximal for any distribution on $\mathcal{X}$ such that $P(\cap_{i=1}^{r} A_i) = 1$.

5. Conclusion

This paper is a preliminary investigation of the problem of likelihood maximisation under incomplete information. The contribution of this paper lies in the proposal of new formulations of maximum likelihood problems in the presence of incomplete information and the strategies for solving them. The main message is that there is not a single way to define a likelihood function, hence not a single solution to the problem of statistical inference in this situation. All depends on the problem one is interested to solve and the available information. The maximization of the visible likelihood function $\mathbf{p}(\mathbf{y};\theta)$ has been the topic of a large literature, especially the EM algorithm, for which the underlying variable X has been often introduced artificially as a trick to facilitate likelihood maximization via an iterative alternating optimisation method [31, 13, 42].

The EM algorithm has been proposed as a general approach to handling incomplete data as well [11], at least in the case when the data partitions the range of the random variable X of interest. However, imprecise data does not necessarily form a partition, but it generally corresponds to a random set in the sense of the other well-known paper by Dempster [11]; moreover the EM algorithm is mainly used to retrieve the unobserved sample of the latent random variable and to provide an estimate of the underlying probability distribution over $\mathcal{X}$ based on it, thanks to the existing relations between the parameters of the measured quantity Y and the latent variable X. Thus, when there is a unique maximum likelihood estimate based on the observed sample $\mathbf{y}$, this method selects a single marginal distribution over $\mathcal{X}$. Alternatively, when the parameter $\theta = (\theta_1, \theta_2)$ is separable with

respect to $(M'|p')$, the MLE from $\mathbf{y}$ only depends on θ_1 (the component that determines the marginal distribution on $\mathcal{Y}$) and therefore the MLE is not unique. But, when there is a single $\hat{\theta}_1$ such that $\arg\max L^{\mathbf{y}}(\theta) = \{\hat{\theta}_1\} \times \Theta_2$ and furthermore the parametrization does not impose any additional restriction on the conditional distributions, then this maximization procedure yields a belief function on $\mathcal{X}$, as we have illustrated in Example 3. Nevertheless, this does not happen in general. To give an example, we may consider the case where $X = Y$ and $\mathcal{P}_\Theta$ represents the collection of uniform distributions of the form $U(\theta, \theta + 1), \theta \in \mathcal{Y} = \mathbb{R}$. In that case, the collection of marginal distributions over $\mathcal{X}$ based on the maximization of $L^{\mathbf{y}}$ is $\{U(\hat{\theta}, \hat{\theta} + 1) \,:\, \theta \in [\min\{y_1, \ldots, y_n\}, \max\{y_1, \ldots, y_n\} - 1]\}$, that does not coincide with the credal set of any belief measure.

In this paper, we address the case when we are interested in directly maximizing the likelihood function of the precise data that would have been obtained, had the observations been complete and precise. We introduce the latent and total likelihood functions $\mathbf{p}(\mathbf{x}; \theta)$ and $\mathbf{p}(\mathbf{z}; \theta)$ that are possible candidates for maximization. The latter is instrumental if some knowledge of the measurement process is available in the form of conditional probabilities of observations, while the former does away with the idea of modeling the measurement process. Such likelihood functions are imprecisely known due to the lack of precision of the observed data. Two optimization strategies, maximin and maximax, have been studied to cope with such imprecision. Their behavior has been observed via examples and compared to the strategy based on the maximization of $\mathbf{p}(\mathbf{y}; \theta)$. The parametric marginal distributions over $\mathcal{X}$ determined from those strategies do not necessarily coincide with the parametric marginal distribution over $\mathcal{X}$ determined from the maximization of $\mathbf{p}(\mathbf{y}; \theta)$.

Further studies would be needed in order to select the most appropriate strategy in each practical problem. In particular one may use other strategies different from maximin and maximax to compare the interval-valued latent and total likelihood functions, by means of some interval order, or choosing another reference point like the midpoint of the interval. A preliminary discussion about the necessity to find a satisfactory interval order was developed in Ref. [43] in the context of machine learning. There, empirical risk values were considered instead of likelihood values. The connection between both notions, under the corresponding assumptions regarding the underlying distribution of X, is well known in Statistics. Besides, the natural partial ordering between intervals (where $[a, b]$ is said to be less than of equal to $[c, d]$ if and only if $b \leq c$) is considered in Ref. [44]. A collection of "non-dominated" solutions for the optimisation of the latent likelihood is derived from this comparison criterion. The comparison between the mid-points of the interval-valued latent likelihoods has been also used as a criterion to select the best classifier in [35]. A

more exhaustive collection of interval orders is considered in [8].

We have pointed out that the face likelihood discussed by Jaeger [27] also differs from the three other ones studied in more details here. The maximisation of the face likelihood leads to results different from but quite related to the ones of the max-average criterion, as explained by E. Hüllermeier in [[23], Section 5]. This criterion has been generalized to uncertain data and exploited in the Evidential EM algorithm of Denœux [12]. This extension of EM has been successfully used in some applications (see [39, 40] and references therein). A variant of the maximization procedure considered in Ref. [12], where the proposed generalization of $\lambda(\mathbf{y};\theta)$ to uncertain data would be replaced by the corresponding generalization of $p(\mathbf{y};\theta)$ seems to be a promising alternative in the context of parametric estimation. A comparison between both procedures in practice is worth carrying out.

Finally, it has been shown that the maximax strategy favors estimated distributions concentrated on precise selections of the imprecise observations (minimizing entropy), while the maximin strategy favors estimated distributions with large entropy. These conjectures must be studied in further research. Moreover, the comparison of solutions for $\mathbf{p}(\mathbf{x};\theta)$ and $\mathbf{p}(\mathbf{z};\theta)$ needs to be better understood, especially conditions where maximax or maximin solutions for $\mathbf{p}(\mathbf{x};\theta)$ are distinct from solutions for $\mathbf{p}(\mathbf{z};\theta)$. Finally, these maximum likelihood schemes should be studied in the light of existing imputation methods for incomplete data [45, 50].

Acknowledgements. We wish to thank Thierry Denœux and Eyke Hüllermeier for their helpful comments on a preliminary version of this manuscript, as well as Romain Guillaume for discussions on the maximin strategy. The referee reports were useful to clarify some explanations and point out related literature on ignorability under coarse or missing data. This work is partially supported by ANR-11-LABX-0040-CIMI within the program ANR-11-IDEX-0002-02, by the Spanish Ministry of Economy, Industry and Competitivity (projects TIN2014-56967-R and TIN2017-84804-R) and by the Regional Ministry of the Principality of Asturias (under project FC-15-GRUPIN14-073).

References

[1] J. Abellán, S. Moral, Range of entropy for credal sets. In M. López-Díaz et al., Eds.: Soft Methodology and Random Information Systems (Proc. SMPS 2004), Advances in Soft Computing Series, Springer, 157-164 (2004).

[2] J. Abellán, S. Moral, Upper entropy of credal sets. Applications to credal classification Int. J. of Approximate Reasoning, 39: 235-255 (2005).

[3] S. Chib, Marginal likelihood from the Gibbs output, Journal of the American Statistical Association 90 (1995) 1313-1321.

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

[5] I. Couso, D. Dubois, L. Sánchez, Random sets and random fuzzy sets as ill-perceived random variables, SpringerBriefs in Computational Intelligence, Springer (2014).

[6] I. Couso, D. Dubois, Belief revision and the EM algorithm. In: J. P. Carvalho et al. (Eds), IPMU 2016, Part II, CCIS 611, 279-290 (2016).

[7] I. Couso, D. Dubois, Maximum likelihood under incomplete information: toward a comparison of criteria, Soft Methods for Data Science (SMPS 2016), Intelligent Systems and Computing Series, Vol. 456, Springer, 141-148 (2016).

[8] I. Couso, L. Sánchez, Machine learning models, epistemic set-valued data and generalized loss functions: An encompassing approach, Information Sciences, 358-359: 129-150 (2016).

[9] A. P. Dawid, J. M. Dickey, Likelihood and Bayesian inference from selectively reported data, J. Amer. Statist. Assoc. 72: 845-850 (1977).

[10] A. P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Statist. 38: 325-339 (1967).

[11] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. B 39: 1-38 (1977).

[12] T. Denœux, Maximum likelihood estimation from uncertain data in the belief function framework, IEEE Transactions on Knowledge and Data Engineering 26: 119-130 (2013).

[13] C.B. Do, S Batzoglou, What is the expectation maximization algorithm? Nature Biotechnology, 26(8): 897-899 (2009).

[14] D. Dubois, H. Fargier, M. Ababou, D. Guyonnet. A fuzzy constraint-based approach to data reconciliation in material flow analysis. International Journal of General Systems, 43: 787-809, (2014).

[15] Z. Domotor, Probability kinematics - Conditional and entropy principles, Synthese, 63: 74-115 (1985).

[16] A.W.F. Edwards, Likelihood, Cambridge University Press (1972).

[17] R.A. Fisher, On the mathematical foundations of theoretical statistics". Philosophical Transactions of the Royal Society A. 222 (594-604): 309-368 (1922).

[18] P. D. Grünwald, A. P. Dawid, Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. The Annals of Statistics, 32(4): 1367-1433.

[19] R. Guillaume, D. Dubois, Robust parameter estimation of density functions under fuzzy interval observations, 9th ISIPTA Symposium, Pescara, Italy, 147-156 (2015).

[20] R. Guillaume, I. Couso, D. Dubois, Maximum likelihood and robust optimisation on coarse data, 10th ISIPTA Symposium, Lugano, Switzerland, 147-156 (2017).

[21] D. F. Heitjan, Ignorability, sufficiency and ancillarity. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59 (2): 375-381 (1997)

[22] D.F. Heitjan, D. B. Rubin, Ignorability and coarse data, Annals of Statistics 19: 2244-2253 (1991).

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

[24] E. Hüllermeier, W. Cheng, Superset learning based on generalized loss minimization, In: Machine Learning and Knowledge Discovery in Databases; LNCS 9825, European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, Proceedings, Part II, 260-275 (2015).

[25] G, W. Imbens, C. F. Manski, Confidence Intervals for Partially Identified Parameters, Econometrica 72 1845-1857 (2004).

[26] M. Jaeger Ignorability in Statistical and Probabilistic Inference. J. Artif. Intell. Res. (JAIR) 24: 889-917 (2005)

[27] M. Jaeger, The AI&M Procedure for Learning from Incomplete Data, Proc. Uncertainty In Artificial Intelligence Conference (UAI-06), 225-232 (2006).

[28] R. C. Jeffrey, The logic of decision, McGraw-Hill, New York, 1965. 2nd edition: University of Chicago Press, Chicago, IL, 1983. Paperback Edition (1990).

[29] R. J. A. Little, Pattern-mixture models for multivariate incomplete data, Journal of the American Statistical Association, 88(421):125-134 (1993)

[30] R. J. A. Little, and D. B. Rubin, Statistical analysis with missing data. Brisbane, John Wiley & Sons. (1987).

[31] G. McLachlan, T. Krishnan, The EM algorithm and extensions, Wiley (2007).

[32] C.F. Masnki, Partial identification with missing data: concepts and findings, International Journal of Approximate Reasoning 39 (2005) 151-165.

[33] C.F. Manski, Partial identification of probability distributions. New York: Springer-Verlag, 2013.

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

[35] A.M. Palacios, L. Sánchez, I. Couso, Diagnosis of dyslexia with low quality data with genetic fuzzy systems, Int. J. Approx. Reason. 51: 993-1009 (2010).

[36] Z.Pawlak, Rough sets, International Journal of Parallel Programming, 11: 341-356 (1982).

[37] J. Plass, M. Cattaneo, G. Schollmeyer, T. Augustin, Testing of coarsening mechanisms: Coarsening at random versus subgroup independence, Soft Methods for Data Science (SMPS 2016), Intelligent Systems and Computing Series, Vol. 456, 415-422, Springer (2016).

[38] J. Plass, T. Augustin, M. Cattaneo, G. Schollmeyer, Statistical modelling under epistemic data imprecision: Some Results on Estimating Multinomial Distributions and Logistic Regression for Coarse Categorical Data, Proc. ISIPTA'2015, 247-256. SIPTA, 2015.

[39] B. Quost, T. Denœux, Clustering and classification of fuzzy data using the fuzzy EM algorithm; Fuzzy Sets and Systems, 286: 134-156 (2016).

[40] E. Ramasso, T. Denœux, Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions. IEEE Transactions on Fuzzy Systems, 22(2): 395-405 (2014).

[41] D. B. Rubin. Inference and missing data. Biometrika 63(3): 581-592 (1976)

[42] S. Russell, The EM algorithm, Lecture notes CS 281, Computer Science Department University of California, Berkeley (1998).

[43] L. Sánchez, I. Couso, Advocating the use of imprecisely observed data in genetic fuzzy systems, IEEE Trans. Fuzzy Syst. 15(4): 551-562, (2007).

[44] L. Sánchez, I. Couso, J. Casillas, Genetic learning of fuzzy rules based on low quality data, Fuzzy Sets Syst. 160: 2524-2552 (2009).

[45] J.L. Schafer, Multiple imputation: a primer, Statistical Methods in Medical Research 8: 3-15 (1999).

[46] M. Serrurier, H. Prade, An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data. Int. J. Approx. Reasoning 54(7): 919-933 (2013).

[47] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press (1976).

[48] J. Sid-Sueiro. Proper losses for learning from partial labels. In: Proc. Neural Information Processing Systems Conference (NIPS2012), Lake Tahoe, Nevada, USA, (2012).

[49] P. Smets. Constructing the pignistic probability function in a context of uncertainty, Uncertainty in Artificial Intelligence 5 (Henrion M. et al., Eds.), North-Holland, Amsterdam, 29-39 (1990).

[50] A. Ting, H. Lin, Comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data, Quality & Quant. 44: 277-287 (2010)

[51] J. Von Neumann. Zur Theorie der Gesellschaftsspiele, Math. Annalen., 100:295-320 (1928)

[52] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall (1991).

[53] P. Williams, Bayesian conditionalization and the principle of minimum information, British J. for the Philosophy of Sciences, 31: 131–144 (1980)