Belief Revision and the EM Algorithm

Inés Couso[1(✉)] and Didier Dubois[2]

[1] Universidad Oviedo, Gijon, Spain
couso@uniovi.es
[2] IRIT, CNRS and University of Toulouse, Toulouse, France

Abstract. This paper provides a natural interpretation of the EM algorithm as a succession of revision steps that try to find a probability distribution in a parametric family of models in agreement with frequentist observations over a partition of a domain. Each step of the algorithm corresponds to a revision operation that respects a form of minimal change. In particular, the so-called expectation step actually applies Jeffrey's revision rule to the current best parametric model so as to respect the frequencies in the available data. We also indicate that in the presence of incomplete data, one must be careful in the definition of the likelihood function in the maximization step, which may differ according to whether one is interested by the precise modeling of the underlying random phenomenon together with the imperfect observation process, or by the modeling of the underlying random phenomenon alone, despite imprecision.

Keywords: Maximum likelihood · Belief revision · Incomplete information · Expectation-Maximization

1 Introduction

The EM (Expectation-Maximization) algorithm is an iterative technique aiming to find a parameterized model achieving a local maximum of the likelihood function when there is no closed-form solution for the maximum likelihood estimator. Another case where EM is repeatedly used is when there are missing data (unsupervised or semi-supervised learning). In order to do so, a latent (unobserved) variable (artificial, in the first case, meaningful in the second case) is used, whose distribution depends on the same parameter as the observed one. The procedure starts with the assessment of an initial value (or vector of values) for the parameter. Each iteration alternates two steps, the "expectation" (E) step and the "maximization" (M) step. The expectation step postulates an empirical distribution for the unobserved variable that agrees with the observed data. During the maximization step, the maximum likelihood estimator based on the joint empirical distribution of both the latent and the observed variable is determined. The process iterates until some stability is reached. The procedure is known to provide an increasing sequence of values for the likelihood function. It converges to a local maximum when some additional conditions are satisfied [12].

J.P. Carvalho et al. (Eds.): IPMU 2016, Part II, CCIS 611, pp. 279–290, 2016.
DOI: 10.1007/978-3-319-40581-0_23

In the following we study the EM algorithm for likelihood-based estimation, where an observed random variable Y goes along with a latent variable X with range $\mathcal{X}$, and where, based on a sequence of precise observations $y_1, \ldots, y_N$, a likelihood function of the form $\prod_{i=1}^{N} P_Y(y_i;\theta)$, understood as a likelihood function on the joint space $\prod_{i=1}^{N} P_{(X,Y)}(\mathcal{X}\times\{y_i\};\theta)$, is maximized. The EM algorithm proceeds based on an alternating optimisation scheme, where at each step, a fictitious precise data set pertaining to (X, Y) and agreeing with the observed result on Y is generated in agreement with the optimal probabilistic model obtained at the previous step from the previous fictitious data set pertaining to (X, Y) and agreeing with the observed result on Y.

The aim of the paper is to better understand the nature of the solution provided by the EM algorithm on the range of (X, Y). We provide an interpretation of the EM algorithm in terms of a sequence of revision steps. More specifically, the E step consists in determining the sample that minimises Kullback divergence with respect to the parametrical distribution postulated during the M step of the last iteration, while respecting the constraints imposed by the data. We show that it corresponds to a natural use of Jeffrey's rule of revision, that comes down to an imputation of sample values for the latent variable. This result enables a better understanding of what the EM algorithm actually aims to. To the best of our knowledge the relationship between the EM algorithm and Jeffrey's rule has not been previously pointed out.

Moreover, we provide an example-based preliminary discussion on cases of incompletely informed data where the EM algorithm should not be used without caution, either because the collection of postulated parametrized distributions is large enough in order to contain all the joint distributions in agreement with the empirical one, or because, in case of overlapping incomplete observations, the definition of the proper likelihood function is a delicate issue.

The paper is organized as follows: Sect. 2 proposes an original introduction to the EM algorithm where the basic steps are formally justified. In Sect. 3, we recall Jeffrey's rule of revision, the properties it satisfies and its connection with the minimization of divergence. We then reinterpret the EM algorithm as a succession of revision steps. Finally, in Sect. 4, we give some examples of anomalies due to an inefficient or incautious usage of the EM algorithm.

2 Introduction to the EM Algorithm

Let X be a random variable, namely a mapping from a sample space $(\Omega, \mathcal{A}, P)$ to the range of X. For simplicity, we assume that $\mathcal{X}$ is finite, and P_X, the probability function attached to X depends on a parameter θ, i.e. $P_X(\cdot;\theta)$ is a model of the random process driving X. We suppose that instead of observing X, another random quantity Y is observed, also driven by parameter θ. Y incompletely informs about the realization of X, in the sense that if $Y = b \in \mathcal{Y} = \{b_1, \ldots, b_n\}$, we only know that $X \in \Gamma(b) \subseteq \mathcal{X}$, for some multimapping Γ [3]. Dempster et al. [4] give a version of the EM algorithm when the observations y_i are viewed as incomplete perceptions of a latent variable X, assuming that the observations bear on a partition of the whole state space. So, the range of Y is of the form $\{\{A_1\}, \ldots, \{A_r\}\}$, where the A_i's form a partition of $\mathcal{X}$.

Let us consider a sequence of N iid 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 latent variable X, that we shall vary in $\mathcal{X}^N$. Let us also use the nomenclature $L^{\mathbf{y}}(\theta) = \log \mathbf{p}(\mathbf{y}; \theta)$ for the log-likelihood function, where $\mathbf{p}(\mathbf{y}; \theta) = \prod_{i=1}^N p(y_i; \theta)$ denotes the probability of observing $\mathbf{y} \in \mathcal{Y}^N$, assuming that the value of the parameter is θ. The final goal of EM is to find a value of the parameter θ that is a (maybe local) maximum of $L^{\mathbf{y}}(\theta)$.

We are interested in modelling the likelihood function associated to the result of the random process driving the random variable X *despite* imprecision. Namely, behind the measurement report $(y_1, \ldots, y_N)$ there exists a sequence of precise outcomes for X, $(x_1^*, \ldots, x_N^*)$ that would have been observed, had the measurement device been accurate (had Γ been a one-to-one function).

2.1 From the Likelihood Function to the EM Criterion

Let $\mathcal{P}^{\mathcal{X}^N}$ be the set of all probability measures $\mathbf{P}$ we can define on the measurable space $(\mathcal{X}^N, \wp(\mathcal{X}^N))$.When the optimisation of the log-likelihood $L^{\mathbf{y}}(\theta) = \log \sum_{\mathbf{x} \in \mathcal{X}^N} \mathbf{p}(\mathbf{x}, \mathbf{y}; \theta)$ is too difficult, a trick is to optimize a lower bound $F(\mathbf{P}, \theta)$ of it that is simpler to optimize. This is allowed by the introduction of arbitrary latent or fake variables[1] and the use of Jensen inequality. Haas [8] proposes the simple following derivation of the functional F:

$$\begin{aligned}
L^{\mathbf{y}}(\theta) &= \log \sum_{\mathbf{x} \in \mathcal{X}^N} \mathbf{p}(\mathbf{x}, \mathbf{y}; \theta) = \log \sum_{\mathbf{x} \in \mathcal{X}^N} \frac{\mathbf{p}(\mathbf{x})\mathbf{p}(\mathbf{x}, \mathbf{y}; \theta)}{\mathbf{p}(\mathbf{x})} \\
&\geq \sum_{\mathbf{x} \in \mathcal{X}^N} \mathbf{p}(\mathbf{x}) \log \left[\frac{\mathbf{p}(\mathbf{x}, \mathbf{y}; \theta)}{\mathbf{p}(\mathbf{x})} \right] \text{ (Jensen's inequality)} \\
&= \sum_{\mathbf{x} \in \mathcal{X}^N} \mathbf{p}(\mathbf{x}) \log \left[\frac{\mathbf{p}(\mathbf{x}|\mathbf{y}; \theta)\mathbf{p}(\mathbf{y}; \theta)}{\mathbf{p}(\mathbf{x})} \right] \\
&= \sum_{\mathbf{x} \in \mathcal{X}^N} \mathbf{p}(\mathbf{x}) \log \mathbf{p}(\mathbf{y}; \theta) + \sum_{\mathbf{x} \in \mathcal{X}^N} \mathbf{p}(\mathbf{x}) \log \left[\frac{\mathbf{p}(\mathbf{x}|\mathbf{y}; \theta)}{\mathbf{p}(\mathbf{x})} \right] \\
&= L^{\mathbf{y}}(\theta) - D(\mathbf{P}, \mathbf{P}(\cdot|\mathbf{y}; \theta)) = F(\mathbf{P}, \theta).
\end{aligned}$$

where $D(\mathbf{P}, \mathbf{P}') = \sum_{\mathbf{x} \in \mathcal{X}^{\mathbf{N}}} \mathbf{p}(\mathbf{x}) \log[\frac{\mathbf{p}(\mathbf{x})}{\mathbf{p}'(\mathbf{x})}]$ is the Kullback-Leibler divergence of $\mathbf{P}'$ from $\mathbf{P}$, and $\mathbf{p}$ is the mass function associated to $\mathbf{P}$.[2]

[1] In some cases, they are not artificial, and are naturally present in the problem.

[2] In the expression in line 2 of the above derivation, $F(\mathbf{P}, \theta)$ could be, with some abuse of notation, written $-D(\mathbf{P}, \mathbf{P}(\cdot, \mathbf{y}; \theta))$ as it is a kind of divergence from $\mathbf{P}(\cdot, \mathbf{y}; \theta))$. However the sum on $\mathcal{X}^N$ of the latter quantities is not 1 (it is $\mathbf{p}(\mathbf{y}; \theta)$) and this pseudo-divergence can be negative.

Some authors use the nomenclature $\ell(\theta|\theta^{(n-1)}) = F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)});\theta)$. According to the definition of F, the properties of logarithms, we can alternatively express $\ell(\theta|\theta^{(n-1)})$ as follows:

$$\ell(\theta|\theta^{(n-1)}) = \log(\mathbf{p}(\mathbf{y};\theta)) + \sum_{\mathbf{x}\in\mathcal{X}^N} \mathbf{p}(\mathbf{x}|\mathbf{y};\theta^{(n-1)}) \log \frac{\mathbf{p}(\mathbf{x}|\mathbf{y};\theta)}{\mathbf{p}(\mathbf{x}|\mathbf{y};\theta^{(n-1)})}.$$

Moreover, taking into account the fact that $\mathbf{p}(\cdot|\mathbf{y};\theta^{(n-1)}) : \mathcal{X}^N \to [0,1]$ is a mass function (the sum of the masses is equal to 1), $\ell(\theta|\theta^{(n-1)})$ also reads

$$\sum_{\mathbf{x}\in\mathcal{X}^N} \mathbf{p}(\mathbf{x}|\mathbf{y};\theta^{(n-1)}) \log \frac{\mathbf{p}(\mathbf{x}|\mathbf{y};\theta)\mathbf{p}(\mathbf{y};\theta)}{\mathbf{p}(\mathbf{x}|\mathbf{y};\theta^{(n-1)})} = \sum_{\mathbf{x}\in\mathcal{X}^N} \mathbf{p}(\mathbf{x}|\mathbf{y};\theta^{(n-1)}) \log \frac{\mathbf{p}(\mathbf{x},\mathbf{y};\theta)}{\mathbf{p}(\mathbf{x}|\mathbf{y};\theta^{(n-1)})}. \tag{1}$$

since $\mathbf{p}(\mathbf{x},\mathbf{y};\theta) = \mathbf{p}(\mathbf{x}|\mathbf{y};\theta)\mathbf{p}(\mathbf{y};\theta)$. We can therefore express $\ell(\theta;\theta^{(n-1)})$ as the sum of an entropy and a term that takes the form of an expectation:

$$\ell(\theta|\theta^{(n-1)}) = H(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}) + E_{\cdot|\mathbf{y};\theta^{(n-1)}}[\log \mathbf{p}(\mathbf{X},\mathbf{y};\theta)]. \tag{2}$$

The last term represents indeed the expectation of a function of the random variable $\mathbf{X}$ taking the value $\log \mathbf{p}(\mathbf{x},\mathbf{y};\theta)$ with probability $\mathbf{p}(\mathbf{x}|\mathbf{y};\theta^{(n-1)})$ for every $\mathbf{x} \in \mathcal{X}^N$.

The main structure of the EM algorithm is then as follows. We first provide an initial value for the parameter, $\theta^{(0)} \in \Theta$. Each iteration of the algorithm, $n \geq 1$ consists of two steps, respectively called "expectation" (E) and "maximization" (M). According to [13], they can be described as follows:

- *Expectation step*: We compute the expectation $E_{\cdot|\mathbf{y};\theta^{(n-1)}}[\log \mathbf{p}(\mathbf{X},\mathbf{y};\theta)]$.
- *Maximization step*: We maximize $\ell(\theta|\theta^{(n-1)})$ wrt θ. According to Eq. (1), this is equivalent to minimizing the divergence $D(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)});\mathbf{P}(\cdot|\mathbf{y};\theta))$.

2.2 The EM Algorithm as a Succession of Revision Steps

Computing $E_{\cdot|\mathbf{y};\theta^{(n-1)}}[\log \mathbf{p}(\mathbf{X},\mathbf{y};\theta)]$ requires the determination of the conditional distribution $\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)})$. The algorithm can then be alternatively described as follows:

- *"Expectation" step*: We compute the first argument of F as the probability measure determined by the mass function $\mathbf{p}(\cdot|\mathbf{y};\theta^{(n-1)}) : \mathcal{X}^N \to [0,1]$. In other words, we find the value of the first argument of the function F in order to fulfill the equality $F(\mathbf{P},\theta^{(n-1)}) = L^{\mathbf{y}}(\theta^{(n-1)})$.
- *Maximization step*: We determine $\theta^{(n)} = \arg\max_{\theta\in\Theta} F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta)$.

Note that in this presentation, the E-step is no longer, strictly speaking, computing an expectation, as it yields a mass function on $\mathcal{X}^N$. In this case, the computation of the expectation proper takes place when determining $F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta)$.

With these two steps, it is easy to guarantee that the sequence $(L^{\mathbf{y}}(\theta^{(n)}))_{n\in\mathbb{N}}$ is increasing. Namely as noticed above, we have that

$F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta^{(n-1)}) = L^{\mathbf{y}}(\theta^{(n-1)})$, for an arbitrary n. Now, since $\theta^{(n)} = \arg\max_{\theta\in\Theta} F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta)$, we have that $F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta^{(n)}) \geq F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta^{(n-1)}) = L^{\mathbf{y}}(\theta^{(n-1)})$. Taking into account the non-negativity of Kullback-Leibler's divergence (due to Jensen's inequality), we can deduce that $L^{\mathbf{y}}(\theta^{(n)}) \geq F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta^{(n)})$, and therefore that $L^{\mathbf{y}}(\theta^{(n)}) \geq L^{\mathbf{y}}(\theta^{(n-1)})$.

Some authors also describe the EM algorithm as a maximization-maximization procedure, since both steps refer to the maximization of the function F:

- *Expectation step*: We maximize $F(\mathbf{P},\theta^{(n-1)})$ with respect to $\mathbf{P}$; we get $\mathbf{P} = \mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)})$.
- *Maximization step*: maximize $F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta)$ with respect to θ; we get $\theta = \theta^{(n)}$.

3 The EM Algorithm from a Belief Revision Perspective

In this section, we shall prove that the E-step is an example of application of Jeffrey's revision rule governed by the minimal change principle. As the M-step also implements a form of minimal change, we thus show that the EM algorithm tries to iteratively find a statistical model that is as close as possible to a distribution of latent variables that is compatible with the observed incomplete data, oscillating from one distribution to the other.

3.1 Jeffrey's Revision Rule

In probability theory, there is a natural method for revising a prior probability P on a set S of mutually exclusive alternatives, in the presence of new probabilistic information I: a distribution $\rho_1,\ldots,\rho_r$ on elements of a partition $\{A_1,\ldots,A_r\}$ of S. The coefficients ρ_i sum to 1 and act as constraints on the posterior probability of elements A_i of the partition. Such an updating rule is proposed by Jeffrey [11]. Jeffrey's rule provides an effective means to revise a prior probability distribution P to a posterior P', given input I. Some axioms guide the revision process:

$$P'(A_i) = \rho_i. \tag{3}$$

This axiom clearly expresses that P' should respect the input information which is of the same nature as the prior probability, with priority given to the input.

Jeffrey's method also relies on the assumption that, while the probability on a prescribed subalgebra of events is enforced by the input information, the probability of any event $B \subseteq S$ conditional to any uncertain event A_i in this subalgebra is the same in the original and the revised distributions. Namely,

$$\forall A_i, \forall B, P(B|A_i) = P'(B|A_i). \tag{4}$$

The underlying interpretation of minimal change implied by the constraint of Eq. (4) is that the revised probability measure P' must preserve the conditional probability degree of any event B knowing event A_i has occurred. Jeffrey's rule of conditioning yields the unique distribution that satisfies (3) and (4) and takes the following form:

$$P'(B) = \sum_{i=1}^{r} \rho_i \cdot \frac{P(B \cap A_i)}{P(A_i)}. \tag{5}$$

Jeffrey's rule respects the probability kinematics principle, whose objective is to minimize change, usually in the sense of an informational distance between probability distributions [1]: The posterior probability P' minimizes the Kullback-Leibler divergence $D(P, P') = \sum_{s \in S} p'(s) \log[\frac{p'(s)}{p(s)}]$ with respect to the original distribution under the probabilistic constraints (3) defined by the input I (as explained in [16]).

3.2 The EM Algorithm from the Standpoint of Joint Distributions: E-step

Since we have assumed that $\mathbf{z}$ represents a sequence of N i.i.d. copies of (X, Y), we can decompose the probability mass $\mathbf{p}(\cdot|\mathbf{y}; \theta) : \mathcal{X}^N \to [0, 1]$ into a product of N mass functions, each one determining a distribution on $\mathcal{X}$. Let us now denote by n_{kj} the number of times that the pair (a_k, b_j) appears in the sample $\mathbf{z}$. Now, in order to denote the product mass function, we will use the nomenclature

$$\mathbf{p}(\mathbf{x}|\mathbf{y}; \theta) = \prod_{i=1}^{N} p(x_i|y_i; \theta) = \prod_{k=1}^{m} \prod_{j=1}^{r} p(a_k|b_j; \theta)^{n_{kj}}, \tag{6}$$

where $p(\cdot|b_j; \theta)$ denotes the mass function associated to the j-th marginal distribution:

$$p(a_k|b_j; \theta) = \frac{p^{\theta}_{kj}}{p^{\theta}_{.j}},\ \forall\, j = 1, \ldots, r.$$

At the expectation step of the nth iteration of the EM algorithm, we compute the conditional probabilities $p(\cdot|b_j; \theta^{(n-1)}), \forall\, j = 1, \ldots, r$. If we consider the joint probability that results from combining those conditional probabilities with the marginal distribution on $(\mathcal{Y}, \wp(\mathcal{Y}))$ determined by the empirical distribution associated to the observed sample $\mathbf{y}$, $(\frac{n_{.1}}{N}, \ldots, \frac{n_{.r}}{N})$, where $n_{.j} = \sum_{k=1}^{m} n_{kj}$ is the number of times b_j appears in the observed sample, we will get the following joint mass distribution on $(\mathcal{X} \times \mathcal{Y}, \wp(\mathcal{X}) \times \wp(\mathcal{Y}))$:

$$\hat{p}^{(n-1)}(a_k, b_j) := \frac{n_{.j}}{N} \cdot p(a_k|b_j; \theta^{(n-1)}) = \frac{n_{.j}}{N} \cdot \frac{p^{\theta^{(n-1)}}_{kj}}{p^{\theta^{(n-1)}}_{.j}} \tag{7}$$

The E-step thus leads to a joint probability measure $\hat{P}^{(n-1)}$, on $\mathcal{X} \times \mathcal{Y}$ that, if the terms $n_{.j} \cdot \frac{p^{\theta^{(n-1)}}_{kj}}{p^{\theta^{(n-1)}}_{.j}}$ are integers, corresponds to an artificial sample $\mathbf{z}^{(n-1)} \in$

$(\mathcal{X}\times\mathcal{Y})^N$ involving the latent variable X, that is in agreement with the observed sample $\mathbf{y}$. Let us denote by $\mathcal{P}_{\mathbf{y}}$, the set of such joint probability measures on $(\mathcal{X}\times\mathcal{Y}, \wp(\mathcal{X})\times\wp(\mathcal{Y}))$ whose marginal distribution on $\mathcal{Y}$ coincides with the empirical distribution $(p_{.1},\ldots,p_{.r}) = (\frac{n_{.1}}{N},\ldots,\frac{n_{.r}}{N})$, associated to the sample $\mathbf{y}$.

Proposition 1. *The result* $\hat{p}^{(n-1)}$ *of the E-step is the posterior probability distribution generated by Jeffrey's rule of conditioning where the input information is given by the observed sample probabilities.*

Proof: Compare Eqs. (5) and (7). In the above Eq. (7), let $S = \mathcal{X}\times\mathcal{Y}$, the prior probability P is the parametric one with mass function $p(a_k, b_j;\theta^{(n-1)})$, the input comes from the observable sample $\mathbf{y}$, in the sense that $A_j = \mathcal{X}\times\{b_j\}$, with probabilities $\rho_j = \frac{n_{.j}}{N}$.

According to the result provided in [16] by P.M. Williams, if we consider the collection, $\mathcal{P}_{\mathbf{y}}$, of joint probability measures on $(\mathcal{X}\times\mathcal{Y}, \wp(\mathcal{X})\times\wp(\mathcal{Y}))$ whose marginal distribution on $\mathcal{Y}$ coincides with the empirical distribution associated to the sample $\mathbf{y}$, $(p_{.1},\ldots,p_{.r}) = (\frac{n_{.1}}{N},\ldots,\frac{n_{.r}}{N})$, the above joint probability measure, $\hat{P}^{(n)}$, is, among all of them, the one that minimizes Kullback-Leibler's divergence with respect to the joint distribution $p(\cdot,\cdot;\theta^{(n-1)}) : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ obtained in the maximization step of the previous iteration.

3.3 The EM Algorithm from the Standpoint of Joint Distributions: M-step

Next we will check that the M step aims at looking for the Maximum Likelihood Estimate (MLE) of θ, given the joint empirical distribution proposed in Eq. (7). The criterion to be optimised at the nth M- step is

$$F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta) = L^{\mathbf{y}}(\theta) - D\left(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}), \mathbf{P}(\cdot|\mathbf{y};\theta)\right).$$

Let us also notice that:

$$\begin{aligned} D(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\mathbf{P}(\cdot|\mathbf{y};\theta)) &= \sum_{i=1}^{N} D\left(P(\cdot|y_i;\theta^{(n-1)}), P(\cdot|y_i;\theta)\right) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{m} p(a_k|y_i;\theta^{(n-1)})\cdot\log\frac{p(a_k|y_i;\theta^{(n-1)})}{p(a_k|y_i;\theta)}. \end{aligned}$$

On the other hand, due to the properties of the logarithmic function, we can write $L^{\mathbf{y}}(\theta) = \sum_{i=1}^{N}\log p(y_i;\theta)$. Moreover, taking into account the fact that $p(\cdot|y_i;\theta^{(n-1)}) : \mathcal{X}\to[0,1]$ is a mass function ($\sum_{k=1}^{m} p(a_k|y_i;\theta^{(n-1)}) = 1$), we can equivalently write:

$$L^{\mathbf{y}}(\theta) = \sum_{i=1}^{N}\log p(y_i;\theta) = \sum_{i=1}^{N}\sum_{k=1}^{m} p(a_k|y_i;\theta^{(n-1)})\log p(y_i;\theta).$$

Again, taking into account the properties of logarithm and also the commutativity of the sum, we can write:

$$\begin{aligned}F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta) &= -\sum_{i=1}^{N}\sum_{k=1}^{m} p(a_k|y_i;\theta^{(n-1)})\cdot\log\frac{p(a_k|y_i;\theta^{(n-1)})}{(p(a_k|y_i;\theta)\cdot p(y_i;\theta))}\\ &= -\sum_{i=1}^{N}\sum_{k=1}^{m} p(a_k|y_i;\theta^{(n-1)})\cdot\log\frac{p(a_k|y_i;\theta^{(n-1)})}{p(a_k,y_i;\theta)}\end{aligned}$$

$$=\sum_{i=1}^{N} H(P(\cdot|y_i;\theta^{(n-1)})) + \sum_{i=1}^{N}\sum_{k=1}^{m} p(a_k|y_i;\theta^{(n-1)})\cdot\log p(a_k,y_i;\theta),$$

where H stands for Shannon entropy. For each $j=1,\ldots,r$, recall that $n_{.j}$ is the number of occurrences of $b_j\in\mathcal{Y}$ in the observed sample $\mathbf{y}=(y_1,\ldots,y_N)$. Then we can rewrite the above expression of $F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta)$ as follows:

$$-\sum_{j=1}^{r} n_{.j}H(P(\cdot|b_j;\theta^{(n-1)})) + \sum_{j=1}^{r}\sum_{k=1}^{m} n_{.j}\cdot p(a_k|b_j;\theta^{(n-1)})\cdot\log p(a_k,b_j;\theta).$$

And due to the properties of logarithm, we can rewrite $F(\mathbf{P}(\cdot|\mathbf{y};\theta^{(n-1)}),\theta)$ as:

$$-\sum_{j=1}^{r} n_{.j}H(P(\cdot|b_j;\theta^{(n-1)})) + \log\left(\prod_{j=1}^{r}\prod_{k=1}^{m} p(a_k,b_j;\theta)^{n_{.j}p(a_k|b_j;\theta^{(n-1)})}\right). \qquad (8)$$

According to the nomenclature established in (Eq. (7)), the above exponent $n_{.j}p(a_k|b_j;\theta^{(n-1)})$ coincides with $N\cdot\hat{p}^{(n)}(a_k,b_j)$. Such an exponent can be seen as the number of occurrences of (a_k,b_j) in an artificial sample inducing the empirical distribution determined by $\hat{p}^{(n)}$ (the joint distribution characterised by the mass function displayed in Eq. (7)). Moreover the entropy term in (8) does not depend on θ. Therefore, maximizing the above expression with respect to θ is equivalent to finding the maximum likelihood estimator associated to such an artificial sample on $\mathcal{X}\times\mathcal{Y}$.

In a nutshell, the M step at iteration n actually finds the MLE associated to the fake sample (the joint distribution) determined by Eq. (7). If the algorithm stops at iteration n^*, we have determined the collection of maximum likelihood estimators associated to all the joint artificial samples on $\mathcal{X}\times\mathcal{Y}$ constructed for the first n^* iterations (the n^* samples inducing the empirical distributions determined by the collection of joint mass functions $\{\hat{p}^{(n)}: n=1,\ldots,n^*\}$). Let the reader notice that, for a specific iteration n, the exponent $n_{.j}\cdot p(a_k|b_j;\theta^{(n-1)}) = N\cdot\hat{p}^{(n)}(a_k,b_j)$ may not be an integer necessarily, and therefore such an empirical joint distribution is not necessarily in total correspondence with some feasible joint sample. In some papers, the fake sample is interpreted as a probability distribution over possible imputations (see e.g., the short paper by Do and Batzoglou [5]), over which the expectation is then taken. This makes the fact that the fake sample could be unobservable much less problematic from an interpretation standpoint.

4 Some Difficulties with the EM Algorithm for Handling Incomplete Information

In this section we show that when the set $\mathcal{P}_\theta$ of parameterized joint distributions includes the set $\mathcal{P}_{\mathbf{y}}$ of joint distributions whose marginals on $\mathcal{Y}$ agree with the empirical distribution induced by $\mathbf{y}$, the EM algorithm cannot be properly used. Moreover in the case of overlapping pieces of incomplete data, a non-careful definition of the likelihood function leads to anomalous results.

Case of Imprecise Data Forming a Partition. As above, we consider the situation where each observation $y = b_i$ on $\mathcal{Y}$ is interpreted as a report providing an element A_i of a partition of $\mathcal{X}$. The maximum likelihood estimator of θ based on the observed sample $\mathbf{y}$ will be the value of the argument for which the likelihood of $\mathbf{y}$ is maximal among all the maximum likelihood estimators associated to all the joint empirical distributions compatible with $\mathbf{y}$. If the probabilities of elements of $\mathcal{X}$ and $\mathcal{Y}$ are not related to each other via enough constraints, there will generally be several MLE distributions on $\mathcal{X}$ in agreement with the observed sample on $\mathcal{Y}$. Moreover, the collection of n^* joint distributions determined by the n^* E-steps of the algorithm are, in general, just a fraction of this collection of compatible joint distributions.

Example 1. Consider the random experiment that consists of rolling a dice. We do not know whether the dice is fair or not. Suppose we only get reports on whether the outcomes are even or odd. Let X be the random variable denoting the actual outcome of the dice roll (from $a_1 = 1$ to $a_6 = 6$) and let Y be a binary variable taking the values b_1 (odd) and b_2 (even). Let the 6-dimensional vector $\theta = (p_1, \ldots, p_6)$ represent the actual (unknown) probability distribution of Z, with $p_6 = 1 - \sum_{i=1}^{5} p_i$. Let $\pi = p_2 + p_4 + p_6$ and $1 - \pi = p_1 + p_3 + p_5$ respectively denote the probabilities of getting an even or an odd number. 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, the maximum likelihood estimator of π would be $\hat{\pi} = \frac{n_{.2}}{N}$. Also, we can easily check that any vector $(\hat{p}_1, \ldots, \hat{p}_6)$ satisfying the constraints $\hat{p}_2 + \hat{p}_4 + \hat{p}_6 = \frac{n_{.2}}{N}$ is a maximum likelihood estimator of θ given the observed sample. Now, let us suppose that we use the EM algorithm in order to find such an MLE. We first initialize the vector θ, by means of selecting some $\theta^{(0)} = (p_1^{(0)}, \ldots, p_6^{(0)})$. Then, we have to apply the E-step, that is, Jeffrey's rule with $\rho_1 = \frac{n_{.1}}{N}, A_1 = \{1, 3, 5\}, \rho_2 = \frac{n_{.2}}{N}, A_2 = \{2, 4, 6\}$. We get $(p_1^{(1)}, \ldots, p_6^{(1)})$, where:

$$p_i^{(1)} = \frac{n_{.1}}{N} \frac{p_i^{(0)}}{p_1^{(0)} + p_3^{(0)} + p_5^{(0)}}, i = 1, 3, 5; \quad p_i^{(1)} = \frac{n_{.2}}{N} \frac{p_i^{(0)}}{p_4^{(0)} + p_6^{(0)} + p_6^{(0)}}, i = 2, 4, 6.$$

For instance, if we take the starting point $\left(p_1^{(0)}, \ldots, p_6^{(0)}\right) = (\frac{1}{6}, \ldots, \frac{1}{6})$ then, we will get $p_i^{(1)} = \frac{n_{.1}}{3N}, i = 1, 3, 5$ and $p_i^{(1)} = \frac{n_{.2}}{3N}, i = 2, 4, 6$. Such a vector is also the maximum likelihood estimator of θ based on a fake sample with equal numbers of 1, 3, 5's and equal numbers of 2, 4, 6's. This vector is thus the optimum of the first M step based on this fake sample. A different postulated initial vector would be identified with a different imputation $\left(p_1^{(1)}, \ldots, p_6^{(1)}\right)$.

The previous example illustrates a case where the MLE based on the observed sample $\mathbf{y}$ is not unique, and an MLE is reached after the first iteration of the EM algorithm. Whatever the starting point $\theta^{(0)}$, the estimate based on the subsequent iteration of the algorithm, $\theta^{(1)}$ is an MLE of θ based on $\mathbf{y}$, which completely depends on $\theta^{(0)}$. Using EM in this situation sounds questionable. In cases where the probabilities on $\mathcal{X}$ are tightly constrained, the MLE for $\mathbf{y}$ can be unique and is asymptotically reached after several iterations of the EM algorithm, independently of the initial choice of the parameter (see the first example in the paper by Dempster et al. [4]).

Anomalies when Imprecise Observations Overlap. When the elements of the range of the observed variable $\mathcal{Y}$ correspond to elements of a partition of $\mathcal{X}$, the likelihood function of Y takes the form $\prod_{j=1}^r P(X \in A_j)^{n_{\cdot j}}$, with $\sum_{j=1}^r n_{\cdot j} = N$. Suppose now that the images $\{A_1, \ldots, A_r\}$ of $\varGamma$ do not form a partition of $\mathcal{X}$. In other words, if $x \in X$ there may be several $A_i's$ enclosing outcome x_j. Maximizing the product $\prod_{j=1}^r P(X \in A_j)^{n_{\cdot j}}$ instead of $L^{\mathbf{y}}(\theta) = \prod_{j=1}^r P(Y = \{A_j\})^{n_{\cdot j}}$ leads to counter-intuitive results, as we show in the following example.

Example 2. Suppose that a dice is tossed, as in the previous example. Suppose we are told either that the result has been less than or equal to 3 or that it has been greater than or equal to 3. Then $A_1 = \{1, 2, 3\}$ and $A_2 = \{3, 4, 5, 6\}$. Let us denote both responses by y_1 and y_2, respectively. After each toss, when the actual result (X) is 3, the reporter says y_1 or y_2 but we do not know how it is chosen. Let us take a sample of N tosses of the dice and let us assume that we have been told us n_1 times "less than or equal to 3" and $n_2 = N - n_1$ times "greater than or equal to 3". Suppose we take as a likelihood function

$$h(\theta) = P(Z \in A_1)^{n_1} \cdot P(Z \in A_2)^{n_2} = (p_1 + p_2 + p_3)^{n_1} \cdot (p_3 + p_4 + p_5 + p_6)^{n_2},$$

where $\theta = (p_1, \ldots, p_6) \in [0, 1]^6$ such that $\sum_{i=1}^6 p_i = 1$. We can easily observe that it reaches its maximum $(h(\theta) = 1)$ for any vector θ satisfying the constraint $p_3 = 1$. But such a prediction of θ would not be a reasonable estimate for θ. Worse, the EM algorithm applied to this case would also stop after the first iteration and fail to reach this maximum, for the same reason as in the previous example.

The difficulty comes from the fact that, with overlapping pieces of data, the function $h(\theta)$ is arguably not a likelihood function. Edwards ([6], p. 9) defines a likelihood function as follows:

> Let $P(R|\theta)$ be the probability of obtaining results R given the hypothesis θ, according to the probability model ... The likelihood of the hypothesis θ given data R, and a specific model, is proportional to $P(R|\theta)$, 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 $P(R_1 \text{ or } R_2|\theta) = P(R_1|\theta) + P(R_2|\theta)$".

The key point in our problem with overlapping imprecise observations is what we understand by "a result". Actually, an imprecise result taking the form of a subset A_i of $\mathcal{X}$ should be modelled by a singleton $R_i = \{A_i\}$ of the power set of $\mathcal{X}$ in order to satisfy the requirements of Edwards. In other words, if the possible observable results are $\{\{A_1\}, \ldots, \{A_r\}\}$ then $\sum_{i=1}^{r} P(\{A_i\}|\theta) = 1$. In our case, a result is not an event A_i, it is an elementary event $\{A_i\}$ (a report carrying imprecise information about X). Only elementary events can be observed. For instance, when tossing a die, you cannot observe the event "odd". What you see is 1, 3 or 5. But some source may report "{odd}". So, a likelihood function is proportional to $P(\{A_i\}|\theta)$ where R is an elementary event. For instance, $P(\mathcal{X}|\theta) = 1$ cannot be viewed as the likelihood of θ given the sure event.

In order to properly apply the EM algorithm to find the distribution of X, in the case of overlapping observations A_i, we have to introduce a parametric model describing which A_i is chosen by the reporter when the outcome of X is x_j, say a conditional probability $P_\theta(\{A_i\} \mid x_j)$ and let the likelihood function $L^y(\theta)$ account for it, e.g. $P(\{A_i\}|\theta) = \sum_{i=1,m} P_\theta(\{A_i\} \mid x_j)P(x_j \mid \theta)$. Generally, $P_\theta(\{A_i\} \mid x_j) > 0$ only if $x_j \in A_i$. For instance, the superset assumption [10] considers $P_\theta(\{A_i\} \mid x_j)$ to be constant over all supersets of x_j.

In the above example, suppose we model the measurement device by assuming $P(\Gamma = \{1, 2, 3\}|Z = 3) = \alpha$. If m denotes the mass function associated to the imprecise observations, we have that $m(A_1) = P(Y = y_1) = p_1 + p_2 + \alpha\, p_3,\ m(A_2) = (1 - \alpha)p_3 + p_4 + p_5 + p_6$.

Notice that in this case $P(X \in A)$ does not coincide with $m(A) = P(Y = \{A\})$. It would, only under the special situations where $\alpha = 1$ or $p_3 = 0$. Moreover, the difficulty due to the inclusion $\mathcal{P}_\theta \supseteq \mathcal{P}_{\mathbf{y}}$, making the EM algorithm inefficient, remains.

5 Conclusion

What our results show is that the EM algorithm oscillates between the set $\mathcal{P}_\theta$ of parameterized joint distributions and the set $\mathcal{P}_{\mathbf{y}}$ of joint distributions whose marginals on $\mathcal{Y}$ agree with the empirical distribution induced by $\mathbf{y}$: In the initial step a probability measure in $\mathcal{P}_\theta$ is chosen, and then it is updated in the E-step into a probability measure $\hat{P}^{(0)}$ in $\mathcal{P}_{\mathbf{y}}$ using Jeffrey's rule of revision, thus producing an artificial sample on $\mathcal{X} \times \mathcal{Y}$; on this basis, an MLE estimate $\theta^{(1)}$, hence a probability measure $P(\cdot; \theta^{(1)}) \in \mathcal{P}_\theta$ is computed in the M-step, based on the artificial sample underlying $\hat{P}^{(0)}$, and so on, until convergence of the $\theta^{(n)}$ sequence. At each stage n, the log-likelihood function $L^{\mathbf{y}}(\theta)$ increases. However, we have shown cases of incomplete information management where this method does not seem to work properly. In future works we shall propose a more systematic analysis of situations when the EM algorithm stops at the first iteration under the partition assumption, and explore alternative ways of posing the problem of maximum likelihood estimation under incomplete overlapping data [2,7,9,10]. Another issue is to investigate the cogency of the fake sample found by the EM algorithm viewed as an imputation method [14,15].

Acknowledgements. This work is partially supported by ANR-11-LABX-0040-CIMI (Centre International de Mathématiques et d'Informatique) within the program ANR-11-IDEX-0002-02, while Inés Couso was a visiting expert scientist at CIMI, Toulouse, and by TIN2014-56967-R (Spanish Ministry of Science and Innovation) and FC-15-GRUPIN14-073 (Regional Ministry of the Principality of Asturias).

References

1. Domotor, Z.: Probability kinematics - conditional and entropy principles. Synthese **63**, 74–115 (1985)
2. Couso, I., Dubois, D.: Statistical reasoning with set-valued information: ontic vs. epistemic views. Int. J. Approximate Reasoning **55**(7), 1502–1518 (2014)
3. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat. **38**, 325–339 (1967)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Stat. Soc. B **39**, 1–38 (1977)
5. Do, C.B., Batzoglou, S.: What is the expectation maximization algorithm? Nat. Biotechnol. **26**(8), 897–899 (2009)
6. Edwards, A.W.F.: Likelihood. Cambridge University Press, Cambridge (1972)
7. Guillaume, R., Dubois, D.: Robust parameter estimation of density functions under fuzzy interval observations. In: 9th ISIPTA Symposium, Pescara, Italy, pp. 147–156 (2015)
8. Haas, S.E.: The Expectation-Maximization and Alternating Minimization Algorithms. MIT, Cambridge (2002)
9. Hüllermeier, E.: Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization. Int. J. Approximate Reasoning **55**(7), 1519–1534 (2014)
10. Hüllermeier, E., Cheng, W.: Superset learning based on generalized loss minimization. In: Appice, A., Rodrigues, P.P., Santos Costa, V., Gama, J., Jorge, A., Soares, C. (eds.) ECML PKDD 2015. LNCS, vol. 9285, pp. 260–275. Springer, Heidelberg (2015)
11. Jeffrey, R.C.: The Logic of Decision. McGraw-Hill, New York (1965), 2nd edn.: University of Chicago Press, Chicago, (1983). Paperback Edition (1990)
12. McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions. Wiley, New York (2007)
13. Russell, S.: The EM algorithm, Lecture notes CS 281, Computer Science Department University of California, Berkeley (1998)
14. Schafer, J.L.: Multiple imputation: a primer. Stat. Methods Med. Res. **8**, 3–15 (1999)
15. Ting, A., Lin, H.: Comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. Qual. Quant. **44**, 277–287 (2010)
16. Williams, P.: Bayesian conditionalization and the principle of minimum information. Brit. J. Philos. Sci. **31**, 131–144 (1980)