Maximum likelihood estimation from evidential data

Thierry Denœux
Heudiasyc
Université de Technologie de Compiègne, CNRS
Compiègne, France
Email: tdenoeux@hds.utc.fr

***Abstract*—We consider the problem of statistical parameter estimation when the data are uncertain and described by belief functions. An extension of the Expectation-Maximization (EM) algorithm, called the Evidential EM (E^2M) algorithm, is described and shown to maximize a generalized likelihood function. This general procedure provides a simple mechanism for estimating the parameters in statistical models when observed data are uncertain. Two simple application examples are demonstrated.**

Keywords: Statistical inference, point estimation, EM algorithm, uncertain data.

I. Introduction

In statistics, observations of random quantities are usually assumed to be either precise or imprecise, i.e., set-valued. The latter situation occurs, e.g., in the case of censored data, where an observation is only known to belong to a set, usually an interval. The Expectation-Maximization (EM) algorithm [5], [12] has proved to be a powerful mechanism for performing maximum likelihood parameter estimation from such incomplete data.

There are situations, however, where the observations are not only imprecise, but also *uncertain*, i.e., partially reliable [1]. Consider, e.g., a classification problem in which objects in a population belong to one and only one group. Let $\mathcal{X}$ be the finite set of groups, and X be the group of an object randomly drawn from the population. In some applications, realizations x of X are not known with certainty. Rather, an expert provides a subjective assessment of x (a process known as *labeling* [7], [8]). This assessment may take the form of a subset $A \subseteq \mathcal{X}$, a probability distribution p on $\mathcal{X}$ or, more generally, a mass function m on $\mathcal{X}$, i.e., a function $m : 2^{\mathcal{X}} \rightarrow [0, 1]$. It must be stressed that, in this example, the data generation process can be decomposed into two components:

- a random component, which generates a realization x from X;
- a non random component, which produces a mass function m that models the expert's partial knowledge of x.

If this process is repeated n times independently, the data takes the form of n mass functions $m_1, \ldots, m_n$, considered as a partial specification of an unknown realization $x_1, \ldots, x_n$ of an i.i.d. random sample $X_1, \ldots, X_n$. We will refer to such data as *evidential data*. If a parametric model is postulated for X, how can the method of maximum likelihood be extended to handle such data? This is the problem considered in this paper. A generalization of the likelihood function will be proposed, and an extension of the EM algorithm, called the evidential EM (E^2M) algorithm, will be introduced for its maximization.

We may note that, in the special case where each mass functions m_i is consonant, the data can be equivalently represented as n possibility distribution $\widetilde{x}_1, \ldots, \widetilde{x}_n$, which constitutes a *fuzzy random sample*. The problem of statistical inference from fuzzy data, which has received a lot of attention in the past few years [6], [9], [11], [16], is thus a special case of the problem considered here.

Early attempts to adapt the EM algorithm to evidential data, in the special case of mixture models with evidential class labels, were presented in [10], [14], [15]. A rigorous solution to this problem, which is a special case of the general method presented in this paper, was introduced in [3].

The rest of the paper is organized as follows. The EM algorithm will first be recalled in Section II. The extension of the likelihood function and the E^2M algorithm will then be introduced in Sections III and IV, respectively. The application of this algorithm will then be illustrated through two statistical inference problems based on evidential data: binomial probability estimation (Section V) and univariate normal mean and variance estimation (Section VI).

II. The EM algorithm

The EM algorithm is a broadly applicable mechanism for computing MLEs from incomplete data, in situations where ML estimation would be straightforward if complete data were available [5].

Formally, we assume the existence of two sample spaces $\mathcal{X}$ and $\mathcal{Y}$, and a many-to-one mapping φ from $\mathcal{X}$ to $\mathcal{Y}$ (see Figure 1). The observed (incomplete) data $\mathbf{y}$ are a realization from $\mathcal{Y}$, while the corresponding $\mathbf{x}$ in $\mathcal{X}$ is not observed and is only known to lie in the set

$$\mathcal{X}(\mathbf{y}) = \varphi^{-1}(\mathbf{y}) = \{\mathbf{x} \in \mathcal{X} | \varphi(\mathbf{x}) = \mathbf{y}\}.$$

Vector $\mathbf{x}$ is referred to as the *complete data* vector. It is a realization from a random vector $\mathbf{X}$ with p.d.f. $g_c(\mathbf{x}; \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_d)'$ is a vector of unknown parameters with parameter space $\boldsymbol{\Omega}$. The observed data p.d.f. $g(\mathbf{y}; \boldsymbol{\Psi})$ is related to $g_c(\mathbf{x}; \boldsymbol{\Psi})$ by

$$g(\mathbf{y}; \boldsymbol{\Psi}) = \int_{\mathcal{X}(\mathbf{y})} g_c(\mathbf{x}; \boldsymbol{\Psi}) d\mathbf{x}. \tag{1}$$

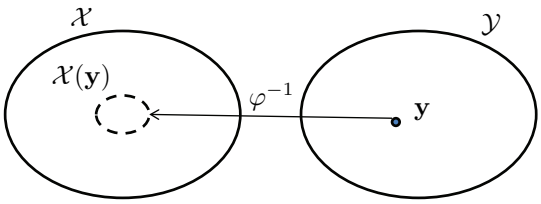

Figure 1. Complete and observed sample spaces.

The EM algorithm approaches the problem of maximizing the observed-data log likelihood $\log L(\mathbf{\Psi}) = \log g(\mathbf{y}; \mathbf{\Psi})$ by proceeding iteratively with the complete-data log likelihood $\log L_c(\mathbf{\Psi}) = \log g_c(\mathbf{x}; \mathbf{\Psi})$. Each iteration of the algorithm involves two steps called the expectation step (E-step) and the maximization step (M-step).

The E-step requires the calculation of

$$Q(\mathbf{\Psi}, \mathbf{\Psi}^{(q)}) = \mathbb{E}_{\mathbf{\Psi}^{(q)}} \left[\log L_c(\mathbf{\Psi}) | \mathbf{y} \right],$$

where $\mathbf{\Psi}^{(q)}$ denotes the current fit of $\mathbf{\Psi}$ at iteration q, and $\mathbb{E}_{\mathbf{\Psi}^{(q)}}$ denotes expectation using the parameter vector $\mathbf{\Psi}^{(q)}$.

The M-step then consists in maximizing $Q(\mathbf{\Psi}, \mathbf{\Psi}^{(q)})$ with respect to $\mathbf{\Psi}$ over the parameter space $\mathbf{\Omega}$, i.e., finding $\mathbf{\Psi}^{(q+1)}$ such that

$$Q(\mathbf{\Psi}^{(q+1)}, \mathbf{\Psi}^{(q)}) \geq Q(\mathbf{\Psi}, \mathbf{\Psi}^{(q)})$$

for all $\mathbf{\Psi} \in \mathbf{\Omega}$. The E- and M-steps are iterated until the difference $L(\mathbf{\Psi}^{(q+1)}) - L(\mathbf{\Psi}^{(q)})$ becomes smaller than some arbitrarily small amount.

It is proved in [5] that the observed-data likelihood $L(\mathbf{\Psi})$ is not decreased after an EM iteration, that is,

$$L(\mathbf{\Psi}^{(q+1)}) \geq L(\mathbf{\Psi}^{(q)})$$

for $q = 0, 1, 2, \ldots$. Hence, convergence to some value L^* is ensured as long as the sequence $L(\mathbf{\Psi}^{(q)})$ for $q = 0, 1, 2, \ldots$ is bounded from above. As noted in [12, page 85], L^* is, in most practical applications and except in pathological cases, a local maximum of the incomplete data likelihood $L(\mathbf{\Psi})$.

III. Generalized likelihood function

Let us now consider the more complex situation where the relationship between the observed and complete spaces is uncertain, so that observed data $\mathbf{y}$ can no longer be associated with certainty to a unique subset of $\mathcal{X}$. This situation will be formalized as follows.

Let us assume the existence of a set Θ of interpretations, one and only one of which holds, and a probability measure $\mathbb{P}$ on Θ. If $\mathbf{y}$ has been observed and $\theta \in \Theta$ is the true interpretation, then the complete data $\mathbf{x}$ is known to belong to $\mathcal{X}(\mathbf{y}, \theta) \subseteq \mathcal{X}$ (see Figure 2). Having observed $\mathbf{y}$, the probability measure $\mathbb{P}$ is carried to $2^{\mathcal{X}}$ by the mapping $\theta \rightarrow \mathcal{X}(\mathbf{y}, \theta)$, which defines a Dempster-Shafer mass function m on $\mathcal{X}$. For simplicity, we will assume from now on that Θ is finite: $\Theta = \{\theta_1, \ldots, \theta_K\}$, in which case m is a discrete mass function with focal sets $\mathcal{X}_k = \mathcal{X}(\mathbf{y}, \theta_k)$ and masses $m_k = m(\mathcal{X}_k)$ for $k = 1, \ldots, K$.

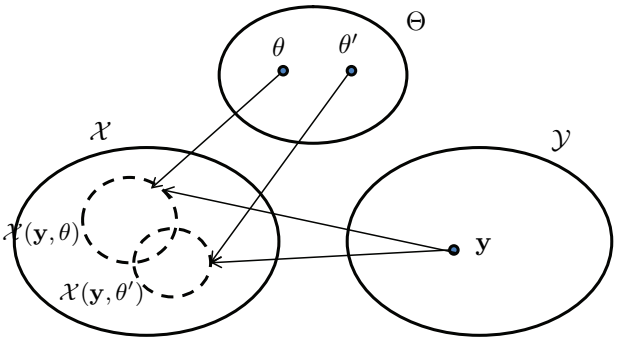

Figure 2. Uncertain relationship between the complete and observed data spaces.

With the same notations as in the previous section, the observed data likelihood may now be defined as:

$$\begin{aligned}
L(\mathbf{\Psi}) &= \sum_{k=1}^{K} g(\mathbf{y}|\theta_k; \mathbf{\Psi}) \mathbb{P}(\{\theta_k\}) && (2)\\
&= \sum_{k=1}^{K} m_k \int_{\mathcal{X}_k} g_c(\mathbf{x}; \mathbf{\Psi}) d\mathbf{x} && (3)\\
&= \int_{\mathcal{X}} g_c(\mathbf{x}; \mathbf{\Psi}) \left(\sum_{k=1}^{K} m_k 1_{\mathcal{X}_k}(\mathbf{x}) \right) d\mathbf{x} && (4)\\
&= \int_{\mathcal{X}} g_c(\mathbf{x}; \mathbf{\Psi}) pl(\mathbf{x}) d\mathbf{x}, && (5)\\
&= \mathbb{E}_{\mathbf{\Psi}} \left[pl(\mathbf{X}) \right], && (6)
\end{aligned}$$

where $pl : \mathcal{X} \rightarrow [0, 1]$ is the plausibility contour function associated to m.

The generalized likelihood of $\mathbf{\Psi}$ is thus equal to the expectation of the plausibility contour function, with respect to the probability distribution $g_c(\mathbf{x}; \mathbf{\Psi})$. We can remark that, when m is consonant, the plausibility contour function can be seen as the membership function of a fuzzy subset of $\mathcal{X}$: $L(\mathbf{\Psi})$ is then the probability of that fuzzy subset, according to Zadeh's definition of the probability of a fuzzy event [17].

In the more general setting of belief functions, $L(\mathbf{\Psi})$ has another interpretation that will now be explained. Let $g_c(\cdot|m; \mathbf{\Psi}) = m \oplus g_c(\cdot; \mathbf{\Psi})$ denote the p.d.f. obtained by combining m with the complete data p.d.f. $g_c(\cdot; \mathbf{\Psi})$ using Dempster's rule [4], [13]:

$$\begin{aligned}
g_c(\mathbf{x}|m; \mathbf{\Psi}) &= \frac{g_c(\mathbf{x}; \mathbf{\Psi}) pl(\mathbf{x})}{\int_{\mathcal{X}} g_c(\mathbf{u}; \mathbf{\Psi}) pl(\mathbf{u}) d\mathbf{u}} && (7)\\
&= \frac{g_c(\mathbf{x}; \mathbf{\Psi}) pl(\mathbf{x})}{L(\mathbf{\Psi})}. && (8)
\end{aligned}$$

The quantity $L(\mathbf{\Psi})$ is thus equal to one minus the degree of conflict between m and $g_c(\mathbf{x}; \mathbf{\Psi})$. Consequently, maximizing $L(\mathbf{\Psi})$ amounts to *minimizing the conflict* between the observations (represented by m) and the parametric model $g_c(\cdot; \mathbf{\Psi})$.

Independence assumptions

Let us assume that the observed data $\mathbf{x} = (x_1, \ldots, x_n)$ is a realization from a random vector $\mathbf{X} = (X_1, \ldots, X_n)$. In many

applications, we can make the following assumptions:

A1: Stochastic independence of the r.v. $X_1, \ldots, X_n$; we can thus write, for all $\mathbf{u} = (u_1, \ldots, u_n) \in \mathcal{X}$:

$$g_c(\mathbf{u}; \mathbf{\Psi}) = \prod_{i=1}^{n} g_c(u_i; \mathbf{\Psi}).$$

A2: The plausibility contour function $pl(\mathbf{x})$ can be written as

$$pl(\mathbf{u}) = \prod_{i=1}^{n} pl_i(u_i),$$

for all $\mathbf{u} = (u_1, \ldots, u_n) \in \mathcal{X}$.

It should be noted that Assumption A2 is totally unrelated to A1: it is not a property of the random variables $X_1 \ldots, X_n$, but of the uncertain observation process. It is actually a weaker form of the *cognitive independence* assumption, as defined by Shafer [13].

Under Assumptions A1 and A2, the expression of the observed data likelihood (6) can be simplified as:

$$L(\mathbf{\Psi}) = \prod_{i=1}^{n} \mathbb{E}_{\mathbf{\Psi}} \left[pl_i(X_i) \right],$$

and the observed data log likelihood can be written as a sum of n terms:

$$\log L(\mathbf{\Psi}) = \sum_{i=1}^{n} \log \mathbb{E}_{\mathbf{\Psi}} \left[pl_i(X_i) \right].$$

IV. The evidential EM algorithm

To maximize function $L(\mathbf{\Psi})$ defined by (2)-(6), we propose to adapt the EM algorithm as follows. Let the E-step now consist in the calculation of the expectation of $\log L_c(\mathbf{\Psi})$ with respect to $g_c(\cdot|m; \mathbf{\Psi}^{(q)})$ defined by (7):

$$Q(\mathbf{\Psi}, \mathbf{\Psi}^{(q)}) = \mathbb{E}_{\mathbf{\Psi}^{(q)}} \left[\log L_c(\mathbf{\Psi}) | m \right] \quad (9)$$

$$= \frac{\int \log(L_c(\mathbf{\Psi})) g_c(\mathbf{x}; \mathbf{\Psi}^{(q)}) pl(\mathbf{x}) d\mathbf{x}}{L(\mathbf{\Psi}^{(q)})}. \quad (10)$$

The M-step is unchanged and requires the maximization of $Q(\mathbf{\Psi}, \mathbf{\Psi}^{(q)})$ with respect to $\mathbf{\Psi}$. The E^{2}M algorithm alternately repeats the E- and M-steps above until the increase of observed-data likelihood becomes smaller than some threshold. The following theorem shows that E^{2}M algorithm inherits the monotonicity property of the EM algorithm, which ensures convergence provided the sequence of incomplete-data likelihood values remains bounded from above.

Theorem 1: Any sequence $L(\mathbf{\Psi}^{(q)})$ for $q = 0, 1, 2, \ldots$ of likelihood values obtained using the E^{2}M algorithm is non decreasing, i.e., it verifies

$$L(\mathbf{\Psi}^{(q+1)}) \geq L(\mathbf{\Psi}^{(q)}) \quad (11)$$

for all q.

Proof. The proof is similar to that of Dempster et al. [5]. We give it here for completeness. Let $k(\mathbf{x}|m; \mathbf{\Psi})$ be defined using the following expression:

$$k(\mathbf{x}|m; \mathbf{\Psi}) = \frac{L_c(\mathbf{\Psi})}{L(\mathbf{\Psi})}.$$

We thus have

$$\log L(\mathbf{\Psi}) = \log L_c(\mathbf{\Psi}) - \log k(\mathbf{x}|m; \mathbf{\Psi}).$$

Taking the expectation of both sides with respect to the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$, using the fit $\mathbf{\Psi}^{(q)}$ for $\mathbf{\Psi}$, we get

$$\log L(\mathbf{\Psi}) = \mathbb{E}_{\mathbf{\Psi}^{(q)}} \left[\log L_c(\mathbf{\Psi}) | m \right] - \mathbb{E}_{\mathbf{\Psi}^{(q)}} \left[\log k(\mathbf{X}|m; \mathbf{\Psi}) | m \right] = Q(\mathbf{\Psi}, \mathbf{\Psi}^{(q)}) - H(\mathbf{\Psi}, \mathbf{\Psi}^{(q)})$$

with

$$H(\mathbf{\Psi}, \mathbf{\Psi}^{(q)}) = \mathbb{E}_{\mathbf{\Psi}^{(q)}} \left[\log k(\mathbf{X}|m; \mathbf{\Psi}) | m \right].$$

We thus have

$$\log L(\mathbf{\Psi}^{(q+1)}) - \log L(\mathbf{\Psi}^{(q)}) = Q(\mathbf{\Psi}^{(q+1)}, \mathbf{\Psi}^{(q)}) - Q(\mathbf{\Psi}^{(q)}, \mathbf{\Psi}^{(q)}) - \left(H(\mathbf{\Psi}^{(q+1)}, \mathbf{\Psi}^{(q)}) - H(\mathbf{\Psi}^{(q)}, \mathbf{\Psi}^{(q)}) \right). \quad (12)$$

The first difference on the right-hand side of (12) is nonnegative as $\mathbf{\Psi}^{(q+1)}$ has been chosen to maximize $Q(\mathbf{\Psi}, \mathbf{\Psi}^{(q)})$ with respect to $\mathbf{\Psi}$. It thus remains to check that the second difference on the right-hand side of (12) is non-positive; that is, we need to verify that the following inequality holds:

$$H(\mathbf{\Psi}^{(q+1)}, \mathbf{\Psi}^{(q)}) - H(\mathbf{\Psi}^{(q)}, \mathbf{\Psi}^{(q)}) \leq 0.$$

Now for any $\mathbf{\Psi}$,

$$H(\mathbf{\Psi}, \mathbf{\Psi}^{(q)}) - H(\mathbf{\Psi}^{(q)}, \mathbf{\Psi}^{(q)}) = \mathbb{E}_{\mathbf{\Psi}^{(q)}} \left[\log \frac{k(\mathbf{X}|m; \mathbf{\Psi})}{k(\mathbf{X}|m; \mathbf{\Psi}^{(q)})} | m \right] \leq \log \mathbb{E}_{\mathbf{\Psi}^{(q)}} \left[\frac{k(\mathbf{X}|m; \mathbf{\Psi})}{k(\mathbf{X}|m; \mathbf{\Psi}^{(q)})} | m \right] \quad (13)$$

and

$$\log \mathbb{E}_{\mathbf{\Psi}^{(q)}} \left[\frac{k(\mathbf{X}|\mathbf{y}; \mathbf{\Psi})}{k(\mathbf{X}|m; \mathbf{\Psi}^{(q)})} | m \right] = \log \int \frac{k(\mathbf{x}|m; \mathbf{\Psi})}{k(\mathbf{x}|m; \mathbf{\Psi}^{(q)})} pl(\mathbf{x}) k(\mathbf{x}|m; \mathbf{\Psi}^{(q)}) d\mathbf{x} = \log \int k(\mathbf{x}|m; \mathbf{\Psi}) pl(\mathbf{x}) d\mathbf{x} = \log \int g_c(\mathbf{x}|m; \mathbf{\Psi}^{(q)}) d\mathbf{x} = 0,$$

where the inequality in (13) is a consequence of Jensen's inequality. □

To conclude this section, we may note that the p.d.f. $g_c(\mathbf{x}|m; \mathbf{\Psi})$ and, consequently, the E^{2}M algorithm depend only on the contour function $pl(\mathbf{x})$ are unchanged if $pl(\mathbf{x})$ is multiplied by a constant. Consequently, the results are unchanged if m is converted into a probability distribution by normalizing the contour function, as suggested in [2].

V. Application to probability estimation

Let us assume that the complete data $\mathbf{x} = (x_1, \ldots, x_n)$ is a realization from an i.i.d. sample $X_1, \ldots, X_n$ from a Bernoulli distribution $\mathcal{B}(\theta)$ with $\theta \in [0, 1]$, and the observed data has the following form: $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ with $\mathbf{y}_i = (p_i, q_i)$, where $p_i = pl_i(0)$ is the degree of plausibility that $x_i = 0$, and $q_i = pl_i(1)$ is the degree of plausibility that $x_i = 1$. Assuming cognitive independence, we get

$$pl(\mathbf{u}) = \prod_{i=1}^{n} pl_i(u_i),$$

for all $\mathbf{u} = (u_1, \ldots, u_n) \in \{0, 1\}^n$.

With these assumptions and notations, the observed data log likelihood can be written as

$$\log L(\theta) = \sum_{i=1}^{n} \log \left[(1 - \theta) p_i + \theta q_i \right].$$

This is the quantity we want to maximize with respect to θ.

To perform the E-step of the E^{2}M algorithm, we observe that the complete data log likelihood is a linear combination of the x_i:

$$\log L_c(\theta) = \ln \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1 - x_i} = n \log(1 - \theta) + \log \left(\frac{\theta}{1 - \theta} \right) \sum_{i=1}^{n} x_i. \quad (14)$$

Consequently,

$$Q(\theta, \theta^{(q)}) = n \log(1 - \theta) + \log \left(\frac{\theta}{1 - \theta} \right) \sum_{i=1}^{n} \xi_i^{(q)} \quad (15)$$

with

$$\xi_i^{(q)} = \mathbb{E}_{\theta^{(q)}} [X_i | m] = \frac{\theta^{(q)} q_i}{(1 - \theta^{(q)}) p_i + \theta^{(q)} q_i}. \quad (16)$$

The M-step consists in maximizing $Q(\theta, \theta^{(q)})$ defined by (15) with respect to θ. The maximum is obtained for

$$\theta^{(q+1)} = \frac{1}{n} \sum_{i=1}^{n} \xi_i^{(q)}. \quad (17)$$

Starting with the an initial guess $\theta^{(0)}$, the E^{2}M algorithm for this problem thus alternates the computation of the $\xi_i^{(q)}$ using (16) and the computation of a new estimate for θ using (17), until the relative difference

$$\frac{L(\theta^{(q+1)}) - L(\theta^{(q)})}{|L(\theta^{(q)})|}$$

becomes less than some threshold ϵ.

Example 1: The above algorithm was applied to the data given in Table I. The observed data likelihood function is displayed in Figure 3. Starting from the randomly chosen initial value $\theta^{(0)} = 0.4628$, the maximum likelihood estimate $\widehat{\theta} = 0.7614$ was found in 27 iterations, with $\epsilon = 10^{-6}$.

Table I
Data of Example 1.

i	1	2	3	4	5
p_i	0.5968	0.6478	0.5962	1.0000	1.0000
q_i	1.0000	1.0000	1.0000	0.3289	0.1892
i	6	7	8	9	10
p_i	0.3752	0.0698	0.6630	1.0000	0.2139
q_i	1.0000	1.0000	1.0000	0.4179	1.0000

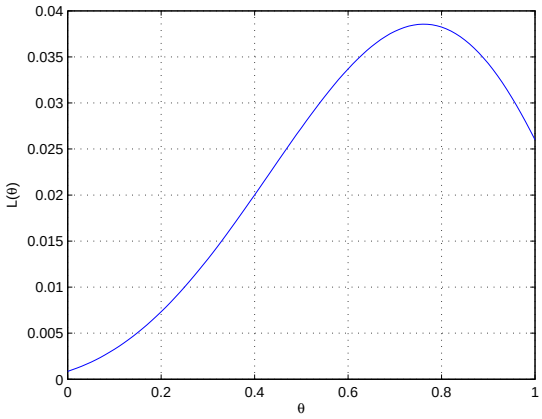

Figure 3. Observed data likelihood function in Example 1.

VI. Normal mean and variance estimation

Let us now assume that the complete data $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X} = \mathbb{R}^n$ is a realization from an i.i.d. random sample from a univariate normal distribution $\mathcal{N}(\mu, \sigma^2)$. The parameter vector is this $\mathbf{\Psi} = (\mu, \sigma)$. The observed data has the form $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ with $\mathbf{y}_i = (w_i, \alpha_i)$. For each $\mathbf{y}_i$, there are two interpretations θ_{i1} and θ_{i2}. Under interpretation θ_{i1}, $x_i = w_i$; under interpretation θ_{i2}, $\mathbf{x}_i \in \mathbb{R}$. The probability for interpretation θ_{i1} to be correct is α_i, which can thus be interpreted as a degree of reliability of the piece of information $\mathbf{y}_i$. The induced mass function m_i on $\mathbb{R}$ is

$$m_i = \alpha_i \delta_{w_i} + (1 - \alpha_i) m_{\mathbb{R}},$$

where δ_{w_i} is the Dirac distribution centered as w_i and $m_{\mathbb{R}}$ is the vacuous mass function defined by $m_{\mathbb{R}}(\mathbb{R}) = 1$. The corresponding plausibility contour function is defined by

$$pl_i(x) = \alpha_i \delta_{w_i}(x) + (1 - \alpha_i)$$

for all $x \in \mathbb{R}$.

Let $g_c(\cdot; \mu, \sigma)$ denote the normal p.d.f. with mean μ and standard deviation σ. The observed data log likelihood is

$$\begin{aligned} \log L(\mu, \sigma) &= \sum_{i=1}^{n} \log \left(\int_{-\infty}^{\infty} g_c(x; \mu, \sigma) pl_i(x) dx \right) \\ &= \sum_{i=1}^{n} \log \left(\alpha_i g_c(w_i; \mu, \sigma) + 1 - \alpha_i \right), \end{aligned}$$

which is to be maximized with respect to μ and σ.

The complete data log likelihood is

$$\log L_c(\mu,\sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 =$$

$$-\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + \mu^2\right).$$

Consequently,

$$Q(\mathbf{\Psi},\mathbf{\Psi}^{(q)}) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\beta_i^{(q)} - 2\mu\sum_{i=1}^{n}\gamma_i^{(q)} + \mu^2\right), \quad (18)$$

where $\gamma_i^{(q)}$ and $\beta_i^{(q)}$ denote, respectively, the expectations of X and X^2 with respect to the conditional probability distribution

$$g_c(\cdot|m_i;\mathbf{\Psi}^{(q)}) = g_c(\cdot;\mu^{(q)},\sigma^{(q)}) \oplus m_i$$

defined by

$$g_c(x|m_i;\mathbf{\Psi}^{(q)}) = \frac{g_c(x;\mathbf{\Psi}^{(q)})pl_i(x)}{\int_{-\infty}^{+\infty} g_c(u;\mathbf{\Psi}^{(q)})pl_i(u)du} = \frac{g_c(x;\mathbf{\Psi}^{(q)})\left[\delta_{w_i}(x) + (1-\alpha_i)\right]}{\alpha_i g_c(w_i;\mathbf{\Psi}^{(q)}) + 1 - \alpha_i}.$$

The following equalities thus hold:

$$\gamma_i^{(q)} = \frac{\alpha_i g_c(w_i;\mathbf{\Psi}^{(q)})w_i + (1-\alpha_i)\mu^{(q)}}{\alpha_i g_c(w_i;\mathbf{\Psi}^{(q)}) + 1 - \alpha_i} \quad (19)$$

and

$$\beta_i^{(q)} = \frac{\alpha_i g_c(w_i;\mathbf{\Psi}^{(q)})w_i^2 + (1-\alpha_i)\left[\left(\mu^{(q)}\right)^2 + \left(\sigma^{(q)}\right)^2\right]}{\alpha_i g_c(w_i;\mathbf{\Psi}^{(q)}) + 1 - \alpha_i}. \quad (20)$$

The maximum of $Q(\mathbf{\Psi},\mathbf{\Psi}^{(q)})$ defined by (18) is obtained for the following values of μ and σ:

$$\mu^{(q+1)} = \frac{1}{n}\sum_{i=1}^{n}\gamma_i^{(q)} \quad (21)$$

and

$$\sigma^{(q+1)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\beta_i^{(q)} - \left(\mu^{(q+1)}\right)^2}. \quad (22)$$

In E-step of the E^{2}M algorithm for this problem thus consists in the calculation of $\gamma_i^{(q)}$ and $\beta_i^{(q)}$ for all i using (19) and (20), respectively. The M-step then updates the estimates of μ and σ using (21) and (22). The algorithm stops when the relative increase of the observed data likelihood becomes less than some threshold ϵ.

Example 2: The above algorithm was applied to the data shown in Table II. The observed data log likelihood function is displayed in Figure 4. Starting from the initial value $(\mu^{(0)},\sigma^{(0)}) = (0, 0.8)$, the maximum likelihood estimate $(\widehat{\mu},\widehat{\sigma}) = (0.3487, 0.0273)$ was found in 30 iterations, with $\epsilon = 10^{-6}$.

Table II
DATA OF EXAMPLE 2.

i	1	2	3	4	5
w_i	0.7845	0.3722	0.3113	0.3521	-0.2376
α_i	0.4966	0.8998	0.8216	0.6449	0.8180
i	6	7	8	9	10
w_i	0.3830	0.5592	0.6574	1.2218	0.3245
α_i	0.6602	0.3420	0.2897	0.3412	0.5341

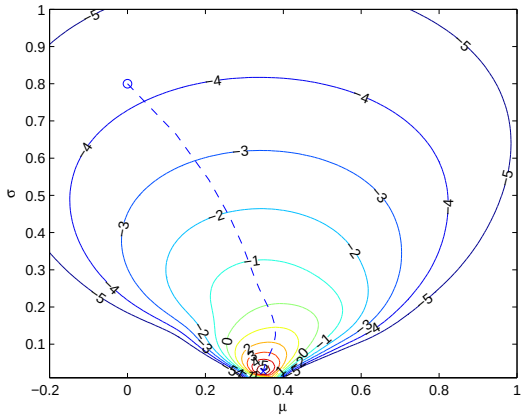

Figure 4. Contour plot of the observed data log likelihood function and trajectory in parameter space (broken line) from the initial parameter values $(\mu^{(0)},\sigma^{(0)})$ (o) to the final MLE $(\widehat{\mu},\widehat{\sigma})$ (x), for the data of Example 2.

VII. CONCLUSION

An iterative procedure for estimating the parameters in a statistical model using evidential data has been proposed. This procedure, which generalizes the EM algorithm, minimizes the degree of conflict between the uncertain observations and the parametric model. It provides a general mechanism for statistical inference when the observed data are uncertain. It remains an open problem to determine the conditions under which the obtained estimator is consistent. This is the topic of on-going research.

REFERENCES

[1] C. C. Aggarwal and P. S. Yu. A survey of uncertain data algorithms and applications. *IEEE Transactions on Knowledge and Data Engineering*, 21(5):609–623, 2009.

[2] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. *International Journal of Approximate Reasoning*, 41(3):314–330, 2006.

[3] E. Côme, L. Oukhellou, T. Denœux, and P. Aknin. Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition*, 42(3):334–348, 2009.

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

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38, 1977.

[6] T. Denœux, M. Masson, and P.-A. Hébert. Nonparametric rank-based statistics and significance tests for fuzzy data. *Fuzzy Sets and Systems*, 153:1–28, 2005.

[7] T. Denœux and P. Smets. Classification using belief functions: the relationship between the case-based and model-based approaches. *IEEE Transactions on Systems, Man and Cybernetics B*, 36(6):1395–1406, 2006.

[8] T. Denœux and L. M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, 122(3):47–62, 2001.

[9] J. Gebhardt, M. A. Gil, and R. Kruse. Fuzzy set-theoretic methods in statistics. In R. Slowinski, editor, *Fuzzy sets in decision analysis, operations research and statistics*, pages 311–347. Kluwer Academic Publishers, Boston, 1998.

[10] I. Jraidi and Z. Elouedi. Belief classification approach based on generalized credal EM. In K. Mellouli, editor, *9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU '07)*, pages 524–535, Hammamet, Tunisia, October/November 2007. Springer.

[11] R. Kruse and K. D. Meyer. *Statistics with vague data*. Kluwer, Dordrecht, 1987.

[12] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

[13] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.

[14] P. Vannoorenberghe. Estimation de modèles de mélanges finis par un algorithm EM crédibiliste. *Traitement du Signal*, 24(2):103–113, 2007.

[15] P. Vannoorenberghe and P. Smets. Partially supervised learning by a credal EM approach. In L. Godo, editor, *Proceedings of the 8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU '05)*, pages 956–967, Barcelona, Spain, 2005. Springer.

[16] R. Viertl. Univariate statistical analysis with fuzzy data. *Computational Statistics & Data Analysis*, 51(1):133–147, 2006.

[17] L. A. Zadeh. Probability measures of fuzzy events. *J. Math. Analysis and Appl.*, 10:421–427, 1968.