Maximum likelihood estimation from Uncertain Data in the Belief Function Framework

Thierry Denœux

Abstract—We consider the problem of parameter estimation in statistical models in the case where data are uncertain and represented as belief functions. The proposed method is based on the maximization of a generalized likelihood criterion, which can be interpreted as a degree of agreement between the statistical model and the uncertain observations. We propose a variant of the EM algorithm that iteratively maximizes this criterion. As an illustration, the method is applied to uncertain data clustering using finite mixture models, in the cases of categorical and continuous attributes.

Index Terms—Uncertain data Mining, Dempster-Shafer theory, Evidence theory, Clustering, EM Algorithm, Mixture Models.

1 Introduction

Recent years have seen a surge of interest in methods for managing and mining uncertain data [1], [2], [3]. As noted in [1], [4], [5], uncertain data arise in many applications due to limitations of the underlying equipment (e.g., unreliable sensors or sensor networks), use of imputation, interpolation or extrapolation techniques (to estimate, e.g., the position of moving objects), partial or uncertain responses in surveys, etc.

In recent work on uncertain data mining, probability theory has often been adopted as a formal framework for representing data uncertainty. Typically, an object is represented as a probability density function (pdf) over the attribute space, rather than as a single point as usually assumed when uncertainty is neglected. Mining techniques that have been proposed for such data include clustering algorithms [6], [7], [8], [9], density estimation techniques [10], outlier detection [11], support vector classification [12], decision trees [5], etc.

Beside this recent body of literature, a lot of work has been devoted to the analysis of interval-valued or fuzzy data, in which ill-known attributes are represented, respectively, by intervals [13] and possibility distributions [14], [15]. As examples of techniques developed for such data, we may mention principal component analysis [16], [17], [18], clustering [19], [20], [21], linear regression [22], [23], [24] and multidimensional scaling [25], [26], [27].

Probability distributions, intervals and possibility distributions may be seen as three instances of a more general model, in which data uncertainty is expressed by means of belief functions. The theory of belief functions, also known as Dempster-Shafer theory or Evidence theory, was developed by Dempster [28], [29] and Shafer [30], and was further elaborated by Smets [31], [32]. A belief function may be seen both as a generalized set and as a non-additive measure, i.e., a generalized probability distribution. The theory of belief functions thus includes extensions of set-theoretic operations, such as intersection and union, and extensions of probabilistic operations, such as conditioning and marginalization.

Until now, the use of belief functions for representing data uncertainty has been mostly confined to classification. In [33], a k-nearest neighbor rule based on Dempster-Shafer theory was introduced. This rule makes it possible to handle partially supervised data, in which uncertain class labels are represented by belief functions (see also [34], [35]). This rule was applied to regression problems with uncertain dependent variable in [36]. Methods for building decision trees from partially supervised data were proposed in [37], [38], [39]. An extension of the k-mode clustering algorithm to data with uncertain attributes was introduced in [40]. Following preliminary work presented in [41] and [42], Côme *et al.* [43] proposed another approach to the partially supervised learning problem based on mixture models and a variant of the EM algorithm [44] maximizing a generalized likelihood criterion. A similar method was used in [45] for partially supervised learning in hidden Markov models.

In this paper, we extend the approach introduced in [43], by allowing uncertainty to be expressed not only on class labels in classification problems, but on any continuous or discrete attribute, in any learning problem based on a parametric statistical model[1]. The contribution of this paper is threefold:

1) We propose an uncertain data model in which data uncertainty is represented by belief functions; this model encompasses probabilistic data, interval-valued data and fuzzy data as special cases;
2) We introduce an extension of the EM algorithm, called the evidential EM (E^2M) algorithm, allowing

T. Denœux is with Heudiasyc, Université de Technologie de Compiègne, CNRS, Compiègne, France.

1. A preliminary version of some of the ideas introduced here was presented in [46]. The present paper is a deeply revised and extended version of this work, with several new results.

us to estimate parameters in parametric statistical models based on uncertain data.

3) We demonstrate the application of this algorithm for handling partially supervised clustering problems with uncertain attributes using finite mixture models.

The rest of this paper is organized as follows. The main concepts of the theory of belief functions useful for our purpose are recalled in Section 2. The proposed data model is then presented in Section 3 and the E^{2}M algorithm is introduced in Section 4. Applications to clustering of discrete and continuous data using finite mixture models are presented in Section 5 and 6, respectively. Finally, Section 7 summarizes our main results and presents some research challenges and ongoing work.

2 Theory of Belief functions

This section recalls the necessary background notions related to Dempster-Shafer theory. Belief functions on discrete domains and Dempster's rule of combination are first recalled in Subsections 2.1 and 2.2, respectively. The notion of cognitive independence is then introduced in Subsection 2.3. Finally, some notions regarding the definition and manipulation of belief functions on the real line are recalled in Subsection 2.4.

2.1 Belief functions on discrete domains

Let X be a variable taking values in a finite domain Ω, called the *frame of discernment*. Uncertain information about X may be represented by a *mass function* m on Ω, defined as a function from the powerset of Ω, denoted as 2^Ω, to the interval $[0,1]$, such that

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{1}$$

Function m is said to be *normalized* if $m(\emptyset) = 0$, a condition that will be assumed in the rest of this paper. Any subset A of Ω such that $m(A) > 0$ is called a *focal element* of m. Two special cases are of interest:

1) If m has a single focal element A, it is said to be *categorical* and denoted as m_A. Such a mass function encodes a piece of evidence that tells us that $X \in A$, and nothing else. There is a one-to-one correspondence between subsets A of Ω and categorical mass functions m_A: categorical mass functions are thus equivalent to sets.
2) If all focal elements of m are singletons, then m is said to be *Bayesian*. There is a one-to-one correspondence between probability distributions $p : \Omega \to [0,1]$ and Bayesian mass functions m such that $m(\{\omega\}) = p(\omega)$, for all $\omega \in \Omega$: Bayesian mass functions are thus equivalent to probability distributions.

To each normalized mass function m, we may associate belief and plausibility functions from 2^Ω to $[0,1]$ defined as follows:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{2a}$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \tag{2b}$$

for all $A \subseteq \Omega$. These two functions are linked by the relation $Pl(A) = 1 - Bel(\overline{A})$, for all $A \subseteq \Omega$. Each quantity $Bel(A)$ may be interpreted as the degree to which the evidence *supports* A, while $Pl(A)$ can be interpreted as an upper bound on the degree of support that *could be* assigned to A if more specific information became available [32]. The following inequalities always hold: $Bel(A) \leq Pl(A)$, for all $A \subseteq \Omega$. The function $pl : \Omega \to [0,1]$ such that $pl(\omega) = Pl(\{\omega\})$ is called the *contour function* associated to m.

If m is Bayesian, then functions Bel is identical to Pl and it is a probability measure. Another special case of interest is that where m is *consonant*, i.e., its focal elements are nested. The following properties then hold [30]:

$$Pl(A \cup B) = \max\left(Pl(A), Pl(B)\right) \tag{3a}$$

and

$$Bel(A \cap B) = \min\left(Bel(A), Bel(B)\right), \tag{3b}$$

for all $A, B \subseteq \Omega$. The plausibility function is thus a *possibility measure*, and Bel is the dual *necessity measure* [14], [47]. The contour function pl is then the associated possibility distribution. Consequently, the theory of belief functions can be considered as having greater expressive power than possibility theory.

2.2 Dempster's rule

Let m_1 and m_2 be two mass functions induced by independent items of evidence. They can be combined using Dempster's rule [30] to form a new mass function defined as:

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C) \tag{4}$$

for all $A \subseteq \Omega$, $A \neq \emptyset$ and $(m_1 \oplus m_2)(\emptyset) = 0$, where

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{5}$$

is the *degree of conflict* between m_1 and m_2. If $\kappa = 1$, there is a logical contradiction between the two pieces of evidence and they cannot be combined. Dempster's rule is commutative, associative, and it admits as neutral element the *vacuous* mass function defined as $m(\Omega) = 1$.

Let us now assume that m_1 is Bayesian. Its contour function is a probability distribution p_1 defined by $p_1(\omega) = m_1(\{\omega\})$ for all $\omega \in \Omega$. Combining m_1 with an arbitrary mass function m_2 with contour function pl_2 yields a *Bayesian mass function* $m_1 \oplus m_2$ with contour function $p_1 \oplus pl_2$ defined by

$$(p_1 \oplus pl_2)(\omega) = \frac{p_1(\omega) pl_2(\omega)}{\sum_{\omega' \in \Omega} p_1(\omega') pl_2(\omega')}. \tag{6}$$

(We note that, without ambiguity, the same symbol $\oplus$ is used for mass functions and contour functions). The degree of conflict between p_1 and pl_2 is

$$\kappa = 1 - \sum_{\omega' \in \Omega} p_1(\omega') pl_2(\omega'). \tag{7}$$

It is equal to one minus the mathematical expectation of pl_2 with respect to p_1. Finally, we may also note that, if m_2 is categorical and such that $m_2(A) = 1$, then $p_1 \oplus pl_2$ is the probability distribution obtained by conditioning p_1 with respect to A.

2.3 Cognitive independence

Let X and Y be two variables defined on finite frames of discernment Ω_X and Ω_Y, and let m^{XY} be a mass function on the product frame $\Omega_X \times \Omega_Y$, expressing evidence on (X, Y). The *marginal* mass function on Ω_X is defined as

$$m^{XY\downarrow X}(A) = \sum_{C \downarrow \Omega_X = A} m^{XY}(C), \tag{8}$$

for all $A \subseteq \Omega_X$, where $C \downarrow \Omega_X$ denotes the projection of $C \subseteq \Omega_X \times \Omega_Y$ on Ω_X.

Let m^X and m^Y denote, respectively, the marginal mass functions on Ω_X and Ω_Y, and let pl^X and pl^Y denote the corresponding plausibility functions. Variables X and Y are said to be *cognitively independent* [30, page 149] with respect to m^{XY} if the following equalities hold:

$$Pl^{XY}(A \times B) = Pl^X(A) Pl^Y(B), \tag{9}$$

for all $A \subseteq \Omega_X$ and $B \subseteq \Omega_Y$. As shown by Shafer [30], this property means that new evidence on one variable does not affect our beliefs in the other variable. It is clear that cognitive independence reduces to stochastic independence when m^X and m^Y are Bayesian.

2.4 Belief functions on the real line

The theory of belief function can be extended to a continuous frame Ω by replacing the mass function m by a density function over a suitable family of subsets of Ω. We will only need to address the simplest case in which $\Omega = \mathbb{R}$ (see, e.g., [29], [48], [49]). In this case, a mass density function can be defined as a function m from the set of closed real intervals to $[0, +\infty)$ such that $m([u, v]) = f(u, v)$ for all $u \leq v$, where f is a two-dimensional probability density function (pdf) with support in $\{(u, v) \in \mathbb{R}^2 | u \leq v\}$. Intervals $[u, v]$ such that $m([u, v]) > 0$ are called focal intervals of m. The contour function pl corresponding to m is defined by the following integral:

$$pl(x) = \int_{-\infty}^{x} \int_{x}^{+\infty} f(u, v) dv du. \tag{10}$$

Two important special cases of continuous belief functions are Bayesian belief function, for which focal intervals are reduced to points, and consonant belief functions, for which focal intervals are nested. The two-dimensional pdf corresponding to a Bayesian belief function has the following form: $f(u, v) = p(u)\delta(u - v)$, where p is a univariate pdf and δ is the Dirac delta function. As in the discrete case, a consonant belief function on $\mathbb{R}$ corresponds to a possibility measure, the contour function being equal to the corresponding possibility distribution.

The reader is referred to [48] for a detailed exposition of different notions pertaining to belief functions on the real line. We will only recall here the following result that will be needed in the sequel. Let m_1 be a Bayesian mass density function defined by the univariate pdf p_1, and let m_2 be an arbitrary mass density function with contour function pl_2. The degree of conflict between m_1 and m_2 is:

$$\kappa = 1 - \int_{-\infty}^{+\infty} p_1(x) pl_2(x) dx. \tag{11}$$

Assuming that $\kappa < 1$, the combination of m_1 and m_2 is Bayesian with corresponding pdf:

$$(p_1 \oplus pl_2)(x) = \frac{p_1(x) pl_2(x)}{1 - \kappa}. \tag{12}$$

We note the similarity between (12) and (6)-(7).

Example 1 *Let $\phi(\cdot; \mu, \sigma^2)$ denote the univariate Gaussian pdf with mean μ and variance σ^2. Let us assume that $p_1(x) = \phi(x; \mu_1, \sigma_1^2)$ and that pl_2 is a normalized Gaussian, taking values in $[0, 1]$: $pl_2(x) = \phi(x; \mu_2, \sigma_2^2)\sigma_2\sqrt{2\pi}$. Application of Eq. (12) can be shown [50] to give*

$$p_1(x) pl_2(x) = c\phi(x; \mu_{12}, \sigma_{12}^2) \tag{13a}$$

with $\mu_{12} = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}$, $\sigma_{12}^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$ and

$$c = \phi(\mu_1; \mu_2, \sigma_1^2 + \sigma_2^2)\sigma_2\sqrt{2\pi}. \tag{13b}$$

Consequently, $p_1 \oplus pl_2$ is the Gaussian pdf $\phi\left(\cdot; \mu_{12}, \sigma_{12}^2\right)$ and the degree of conflict is

$$\kappa = 1 - \phi(\mu_1; \mu_2, \sigma_1^2 + \sigma_2^2)\sigma_2\sqrt{2\pi}. \tag{14}$$

3 Data model

The data model and the generalized likelihood criterion will now first be described in the discrete case in Subsection 3.1. The interpretation of the criterion will then be discussed in Subsection 3.2 and independence assumptions allowing us to simplify its expression will be introduced in Subsection 3.3. These notions will be illustrated using a simple example in Subsection 3.4. Finally, the continuous case will be addressed in Subsection 3.5.

3.1 Discrete case

To simplify the exposition, our data model will first be introduced in the discrete setting. The extension to the continuous setting will be postponed until Subsection 3.5.

Let $\mathbf{X}$ be a discrete random vector taking values in $\Omega_{\mathbf{X}}$, with probability mass function $p_{\mathbf{X}}(\cdot; \boldsymbol{\theta})$ depending on an unknown parameter $\boldsymbol{\theta} \in \Theta$. Let $\mathbf{x}$ denote a realization of

$\mathbf{X}$, referred to as the *complete data*. If $\mathbf{x}$ was perfectly observed, then the likelihood function given $\mathbf{x}$ would be defined as the function from Θ to $[0,1]$ such that:

$$L(\boldsymbol{\theta};\mathbf{x}) = p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \Theta. \tag{15}$$

Let us now assume that $\mathbf{x}$ is not precisely observed, but it is known for sure that $\mathbf{x} \in A$ for some $A \subseteq \Omega_X$. The likelihood function given such *imprecise data* is now:

$$L(\boldsymbol{\theta};A) = p_{\mathbf{X}}(A;\boldsymbol{\theta}) = \sum_{\mathbf{x}\in A} p_X(\mathbf{x};\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta} \in \Theta. \tag{16}$$

More generally, our knowledge of $\mathbf{x}$ may be not only imprecise, but also *uncertain*; it can then be described by a mass function m on $\Omega_{\mathbf{X}}$ with focal elements $A_1,\ldots,A_r$ and corresponding masses $m(A_1),\ldots,m(A_r)$. To extend the likelihood function (16) given such uncertain data, we may simply compute the weighted sum of the terms $L(\boldsymbol{\theta};A_i)$ with coefficients $m(A_i)$, which leads to the following expression:

$$L(\boldsymbol{\theta};m) = \sum_{i=1}^{r} m(A_i) L(\boldsymbol{\theta};A_i). \tag{17}$$

This extension of the classical likelihood criterion will be adopted in the rest of this paper. As will be shown, this criterion has a simple interpretation in terms of conflict between the model and the observations, and it admits a simple expression under independence assumptions.

Before discussing the interpretation of this criterion, it must be emphasized that, in our model, the probability mass function $p_{\mathbf{X}}(\cdot,\boldsymbol{\theta})$ and the Dempster-Shafer mass function m represent two different pieces of knowledge:

- $p_{\mathbf{X}}(\cdot,\boldsymbol{\theta})$ represents *generic* knowledge about the data generating process or, equivalently, about the underlying population; it corresponds to *random uncertainty*;
- m represents *specific* knowledge (or factual evidence[2]) about a given realization $\mathbf{x}$ of $\mathbf{X}$; this knowledge is only partial because the observation process is imperfect; function m captures *epistemic uncertainty*, i.e., uncertainty due to lack of knowledge.

The uncertain data m is thus not assumed to be produced by a random experiment, which is in sharp contrast with alternative approaches based on random set (see, e.g. [52], [53]) or fuzzy random variables (see, e.g., [54], [55], [24]).

3.2 Interpretation

To better understand the meaning of (17), let us rewrite $L(\boldsymbol{\theta};m)$ as:

$$\begin{aligned} L(\boldsymbol{\theta};m) &= \sum_{i=1}^{r} m(A_i) \sum_{\mathbf{x}\in A_i} p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) && \text{(18a)} \\ &= \sum_{\mathbf{x}\in\Omega_{\mathbf{X}}} p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) \sum_{A_i\ni\mathbf{x}} m(A_i) && \text{(18b)} \\ &= \sum_{\mathbf{x}\in\Omega_{\mathbf{X}}} p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) pl(\mathbf{x}). && \text{(18c)} \end{aligned}$$

2. See, e.g., [51] for discussion on the distinction between generic and specific information.

The likelihood $L(\boldsymbol{\theta};m)$ thus only depends on m through its associated contour function pl. For this reason, we will write indifferently $L(\boldsymbol{\theta};m)$ or $L(\boldsymbol{\theta};pl)$.

By comparing (18c) with (7), we can see that $L(\boldsymbol{\theta};m)$ equals *one minus the degree of conflict* (7) between $p_{\mathbf{X}}(\cdot;\boldsymbol{\theta})$ and m. Consequently, maximizing $L(\boldsymbol{\theta};m)$ with respect to $\boldsymbol{\theta}$ amounts to *minimizing the conflict* between the parametric model and the uncertain observations. Alternatively, $L(\boldsymbol{\theta};m)$ can be regarded as a *degree of agreement* [56] between the model and the observations.

We may also observe from (18c) that $L(\boldsymbol{\theta};m)$ can be alternatively defined as the mathematical expectation of $pl(\mathbf{X})$, given $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta};pl) = \mathbb{E}_{\boldsymbol{\theta}}\left[pl(\mathbf{X})\right] \tag{19}$$

The above expression suggests an interesting link with Zadeh's notion of probability of a fuzzy event [57]. Assume that m is consonant. Then, as recalled in Subsection 2.1, its contour function pl is a possibility distribution: it can thus be considered as *fuzzy data*. The generalized likelihood $L(\boldsymbol{\theta};m)$ is then the *probability of the fuzzy data*, given the parameter value $\boldsymbol{\theta}$ (see [58], [59]).

3.3 Independence assumptions

Let us assume that the random vector $\mathbf{X}$ can be written as $\mathbf{X} = (\mathbf{X}_1,\ldots,\mathbf{X}_n)$, where each $\mathbf{X}_i$ is a p-dimensional random vector taking values in $\Omega_{\mathbf{X}_i}$. Similarly, its realization can be written as $\mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_n) \in \Omega_{\mathbf{X}}$. Two different independence assumptions can then be made:

1) Under the *stochastic independence* of the random variables $\mathbf{X}_1,\ldots,\mathbf{X}_n$, the probability mass function $p_{\mathbf{X}}(\cdot;\boldsymbol{\theta})$ of $\mathbf{X}$ can be decomposed as:

$$p_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} p_{\mathbf{X}_i}(\mathbf{x}_i;\boldsymbol{\theta}), \quad \forall \mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_n) \in \Omega_{\mathbf{X}} \tag{20}$$

2) Under the *cognitive independence* of $\mathbf{x}_1,\ldots,\mathbf{x}_n$ with respect to m (see Subsection 2.3), we can write:

$$pl(\mathbf{x}) = \prod_{i=1}^{n} pl_i(\mathbf{x}_i), \quad \forall \mathbf{x} = (\mathbf{x}_1,\ldots,\mathbf{x}_n) \in \Omega_{\mathbf{X}}, \tag{21}$$

where pl_i is the contour function corresponding to the mass function m_i obtained by marginalizing m on $\Omega_{\mathbf{X}_i}$.

We can remark here that the two assumptions above are totally unrelated as they are of different natures: stochastic independence of the random variables $\mathbf{X}_i$ is an objective property of the random data generating process, whereas cognitive independence pertains to our state of knowledge about the unknown realization $\mathbf{x}$ of $\mathbf{X}$.

If both assumptions hold, the likelihood criterion (19) can be written as a product of n terms:

$$L(\boldsymbol{\theta};pl) = \prod_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}}\left[pl_i(\mathbf{X}_i)\right]. \tag{22}$$

3.4 Illustrative example

As an illustrative example, let us consider a medical study in which n patients are taken at random from some population, in which some disease is present in proportion θ. Let X_i be the Bernoulli variable such that $X_i = 1$ if patient i has the disease, $X_i = 0$ otherwise. Assuming the random sample $\mathbf{X} = (X_1, \ldots, X_n)$ to be independent and identically distributed (iid), we have

$$p_X(\mathbf{x}; \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}, \tag{23}$$

for all $\mathbf{x} = (x_1, \ldots, x_n) \in \Omega_{\mathbf{X}} = \{0, 1\}^n$.

After the n patients have been sampled from the population, the complete data consists in the vector $\mathbf{x} = (x_1, \ldots, x_n)$ describing the states of the n patients. However, assume that these states are only partially known from the opinions of a physician, who provides a mass function m on $\Omega_{\mathbf{X}}$. Let m_i be the marginal mass function concerning the state of patient i and pl_i the corresponding contour function. Using (22), we get the following expression for the log-likelihood given m under the cognitive independence assumption (21):

$$\log L(\theta; pl) = \sum_{i=1}^{n} \log \left[(1-\theta) pl_i(0) + \theta pl_i(1) \right]. \tag{24}$$

3.5 Continuous case

The above definitions can be straightforwardly transposed to the continuous case. Assume that $\mathbf{X}$ is a continuous random vector with pdf $p_{\mathbf{X}}(\cdot; \boldsymbol{\theta})$ and let $pl : \Omega_{\mathbf{X}} \to [0, 1]$ be the contour function of a continuous mass function m on $\Omega_{\mathbf{X}}$. Similarly to (18c), the likelihood function given pl can be defined in this case from (19) as

$$L(\boldsymbol{\theta}; pl) = \int_{\Omega_{\mathbf{X}}} p_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}) pl(\mathbf{x}) d\mathbf{x}, \tag{25}$$

assuming this integral to exist and to be nonzero. According to (11), the interpretation of $L(\boldsymbol{\theta}; pl)$ as a degree of agreement between $p_{\mathbf{X}}(\cdot; \boldsymbol{\theta})$ and m is still valid in this case. Equation (22) also holds under the assumptions of stochastic and cognitive independence.

4 Evidential EM algorithm

In this section, the classical EM algorithm [44], [60] will first be recalled in Subsection 4.1. A version of this algorithm, called the evidential EM (E^2M) algorithm, which maximizes the generalized criterion introduced in the previous section will then be described in Subsection 4.2. Finally, the application of this algorithm to solve the Bernoulli example of Subsection 3.4 will be presented in Subsection 4.3.

4.1 EM algorithm

The EM algorithm is a broadly applicable mechanism for computing maximum likelihood estimates (MLEs) from incomplete data, in situations where maximum likelihood estimation would be straightforward if complete data were available [44].

With the same notations as in Section 2.1, let us assume that we have a (discrete or continuous) random vector $\mathbf{X}$ with probability mass function or pdf $p_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta})$. A realization $\mathbf{x}$ has been drawn from $\mathbf{X}$, but it is incompletely observed. The observed data consists in a subset $A \subset \Omega_{\mathbf{X}}$ such that $\mathbf{x} \in A$. The likelihood function knowing that $\mathbf{X} \in A$ is given by (16).

The EM algorithm approaches the problem of maximizing the observed-data log likelihood $\log L(\boldsymbol{\theta}; A)$ by proceeding iteratively with the complete-data log-likelihood $\log L(\boldsymbol{\theta}; \mathbf{x}) = \log p_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta})$. 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

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \mathbb{E}_{\boldsymbol{\theta}^{(q)}} \left[\log L(\boldsymbol{\theta}; \mathbf{X}) \mid A \right], \tag{26}$$

where $\boldsymbol{\theta}^{(q)}$ denotes the current fit of $\boldsymbol{\theta}$ at iteration q and $\mathbb{E}_{\boldsymbol{\theta}^{(q)}}[\cdot|A]$ denotes expectation with respect to the conditional distribution of $\mathbf{X}$ given A, using the parameter vector $\boldsymbol{\theta}^{(q)}$.

The M-step then consists in maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ with respect to $\boldsymbol{\theta}$ over the parameter space Θ, i.e., finding $\boldsymbol{\theta}^{(q+1)}$ such that $Q(\boldsymbol{\theta}^{(q+1)}, \boldsymbol{\theta}^{(q)}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ for all $\boldsymbol{\theta} \in \Theta$. The E- and M-steps are iterated until the difference $L(\boldsymbol{\theta}^{(q+1)}; A) - L(\boldsymbol{\theta}^{(q)}; A)$ becomes smaller than some arbitrarily small amount.

It is proved in [44] that the observed-data likelihood $L(\boldsymbol{\theta}; A)$ is not decreased after an EM iteration, that is, $L(\boldsymbol{\theta}^{(q+1)}; A) \geq L(\boldsymbol{\theta}^{(q)}; A)$ for $q = 0, 1, 2, \ldots$. Hence, convergence to some value L^* is ensured as long as the sequence $L(\boldsymbol{\theta}^{(q)}; A)$ for $q = 0, 1, 2, \ldots$ is bounded from above. As noted in [60, page 85], L^* is, in most practical applications and except in pathological cases, a local maximum of the observed data log-likelihood $L(\boldsymbol{\theta}; A)$.

4.2 Extension to uncertain data

To maximize the likelihood function $L(\boldsymbol{\theta}; m)$ given uncertain data m, as introduced in Section 3, we propose to adapt the EM algorithm as follows.

In the E-step, the conditional expectation (26) of $\log L(\boldsymbol{\theta}; \mathbf{X})$ is now replaced by the expectation with respect to $p_{\mathbf{X}}(\cdot; \boldsymbol{\theta}^{(q)}) \oplus pl$, which will be denoted as $p_{\mathbf{X}}(\cdot|pl; \boldsymbol{\theta}^{(q)})$. We may remark that conditional expectation is recovered in the special case where m is a categorical mass function. Using (6), the probability mass function $p_{\mathbf{X}}(\cdot|pl; \boldsymbol{\theta}^{(q)})$ has the following expression:

$$p_{\mathbf{X}}(\mathbf{x}|pl; \boldsymbol{\theta}^{(q)}) = \frac{p_X(\mathbf{x}; \boldsymbol{\theta}^{(q)}) pl(\mathbf{x})}{L(\boldsymbol{\theta}^{(q)}; pl)}, \tag{27}$$

where $L(\boldsymbol{\theta}^{(q)}; pl)$ is given by (18c) in the discrete case, and (25) in the continuous case. At iteration q, the following function is thus computed:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \frac{\sum_{\mathbf{x} \in \Omega_{\mathbf{X}}} \log(L(\boldsymbol{\theta}; \mathbf{x})) p_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}^{(q)}) pl(\mathbf{x})}{L(\boldsymbol{\theta}^{(q)}; pl)} \quad (28)$$

in the discrete case, or

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \frac{\int_{\Omega_{\mathbf{X}}} \log(L(\boldsymbol{\theta}; \mathbf{x})) p_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}^{(q)}) pl(\mathbf{x}) d\mathbf{x}}{L(\boldsymbol{\theta}^{(q)}; pl)} \quad (29)$$

in the continuous case.

The M-step is unchanged and requires the maximization of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ with respect to $\boldsymbol{\theta}$. 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 E^{2}M algorithm inherits the monotonicity property of the EM algorithm, as shown by the following theorem.

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

$$L(\boldsymbol{\theta}^{(q+1)}; pl) \geq L(\boldsymbol{\theta}^{(q)}; pl) \quad (30)$$

for all q.

The proof of Theorem 1 is very similar to that given in [44]. It is given in Appendix A for completeness.

To conclude this section, we may note that the probability mass or density function $p_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) \oplus pl$ and, consequently, the E^{2}M algorithm depend only on the contour function pl and are unchanged if pl 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 [61], or if pl is viewed as a possibility distribution. However, we may observe that Dempster's rule, which plays a fundamental role in our approach, can only be justified within the theory of belief functions as a rule to combine independent items of evidence (see, e.g., [62] for a comparison between the Dempster-Shafer and Bayesian semantics for "probability judgements"). Consequently, there seems to be no easy way to cast our approach in a more restricted framework such as possibility theory or probability theory.

4.3 Example

Let us come back to the example introduced in Subsection 3.4. The oberved-data likelihood to be maximized is given by (24). The complete-data log-likelihood is obtained by taking the logarithm of (23):

$$\log L(\boldsymbol{\theta}; \mathbf{x}) = n \log(1 - \theta) + \log\left(\frac{\theta}{1 - \theta}\right) \sum_{i=1}^{n} x_i. \quad (31)$$

TABLE 1
Dataset for the Bernoulli example of Subsection 4.3.

i	1	2	3	4	5	6
$pl_i(0)$	1	1	1	α	0	0
$pl_i(1)$	0	0	0	$1-\alpha$	1	1

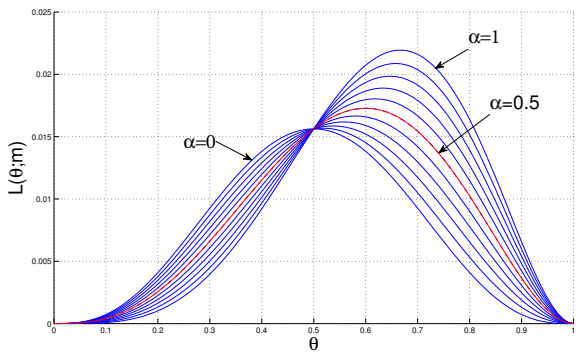

Fig. 1. Observed-data likelihood function for the data of Table 1 and ten equally spaced values of α ranging from 0 to 1.

As it is a linear function of the x_i, the expectation of this quantity with respect to $p_{\mathbf{X}}(\cdot; \theta^{(q)}) \oplus pl$ can be computed in the E-step at iteration q as

$$Q(\theta, \theta^{(q)}) = n \log(1 - \theta) + \log\left(\frac{\theta}{1 - \theta}\right) \sum_{i=1}^{n} \xi_i^{(q)}, \quad (32)$$

where $\xi_i^{(q)}$ is the expectation of X_i with respect to $p_X(\cdot|pl_i; \theta^{(q)}) = p_X(\cdot; \theta^{(q)}) \oplus pl_i$:

$$\xi_i^{(q)} = \mathbb{E}[X_i|pl_i; \theta^{(q)}] = \frac{\theta^{(q)} pl_i(1)}{(1 - \theta^{(q)}) pl_i(0) + \theta^{(q)} pl_i(1)}. \quad (33)$$

The maximum of (32) computed at the M-step is obtained for:

$$\theta^{(q+1)} = \frac{1}{n} \sum_{i=1}^{n} \xi_i^{(q)}, \quad (34)$$

which is the estimate of θ at iteration q.

The above algorithm was applied to the data shown in Table 1. This dataset is composed of $n = 6$ observations, one of which (for $i = 4$) is uncertain and depends on a coefficient α. In that special case it is assumed that $pl_4(0) + pl_4(1) = 1$, i.e., the corresponding mass function m_i is Bayesian. Figure 1 shows the observed-data likelihood function $L(\theta; m)$ for ten different values of α ranging from 0 to 1. In the special cases where $\alpha = 0$ and $\alpha = 1$ there is no data uncertainty and the MLEs are, respectively, $\widehat{\theta} = 0.5$ and $\widehat{\theta} = 2/3$. The E^{2}M algorithm was applied to the data with $\alpha = 0.5$. The results are shown in Table 2. The algorithm was stopped when the relative increase of the likelihood between two iterations was less than 10^{-6}. Starting from the initial value $\theta^{(0)} = 0.3$, this condition was met after 5 iterations. The final MLE is $\widehat{\theta} = 0.6$. This is the value of θ that minimizes the conflict between the uncertain data given in Table 1 and the parametric model (23).

TABLE 2
Intermediate and final results for the E^{2}M algorithm applied to the data of Table 1 with $\alpha = 0.5$.

q	$\theta^{(q)}$	$L(\theta^{(q)}; pl) \times 10^3$
0	0.3000	6.6150
1	0.5500	16.8455
2	0.5917	17.2676
3	0.5986	17.2797
4	0.5998	17.2800
5	0.6000	17.2800

5 Clustering of categorical data

In this section, we demonstrate the application of the E^{2}M algorithm to the clustering of uncertain categorical data based on a latent class model [63], [64]. The notations and the model will first be described in Subsection 5.1. The estimation algorithm for this problem will then be given in Subsection 5.2 and experimental results will be reported in Subsection 5.3.

5.1 Problem description and generative model

Let us consider a dataset composed of n objects described by p categorical attributes and a class attribute taking values in a set of g classes.

Let $\mathbf{w}_i = (w_i^1, \ldots, w_i^p)$ be the attribute vector for object i, and w_i^j the value of attribute j for object i. Each attribute j will be assumed to have r_j modalities and we will denote $w_i^{jh} = 1$ if attribute j takes modality h for object i and $w_i^{jh} = 0$ otherwise.

Similarly, the class of object i will be described by a vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{ig})$ of binary variables z_{ik} such that $z_{ik} = 1$ if object i belongs to class k, and $z_{ik} = 0$ otherwise.

The complete data $\mathbf{x} = \{(\mathbf{w}_1, \mathbf{z}_1), \ldots, (\mathbf{w}_n, \mathbf{z}_n)\}$ will be assumed to be a realization from an iid random sample $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, with $\mathbf{X}_i = (\mathbf{W}_i, \mathbf{Z}_i)$. We will further assume the attributes W_i^j to be conditionally independent given the class $\mathbf{Z}_i$. Denoting by α_k^{jh} the probability that attribute j takes modality h given class k and by π_k the prior probability of class k, we thus have:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{w}_i, \mathbf{z}_i; \boldsymbol{\theta}), \tag{35}$$

where $\boldsymbol{\theta} = (\{\alpha_k^{jh}\}, \{\pi_k\})$ is the vector of all parameters, and

$$\begin{aligned} p(\mathbf{w}_i, \mathbf{z}_i; \boldsymbol{\theta}) &= p(\mathbf{z}_i; \boldsymbol{\theta}) p(\mathbf{w}_i | \mathbf{z}_i; \boldsymbol{\theta}) && (36a)\\ &= \prod_{k=1}^{g} \pi_k^{z_{ik}} \prod_{k=1}^{g} p(\mathbf{w}_i | z_{ik} = 1; \boldsymbol{\theta})^{z_{ik}} && (36b)\\ &= \prod_{k=1}^{g} \left(\pi_k \prod_{j=1}^{p} \prod_{h=1}^{r_j} (\alpha_k^{jh})^{w_i^{jh}} \right)^{z_{ik}}. && (36c) \end{aligned}$$

The complete data $\mathbf{x}$ will be assumed to be imperfectly observed, and partial knowledge about $\mathbf{x}$ will be assumed to be described by a mass function m on the domain of $\mathbf{x}$. Assuming cognitive independence (21) between all variables according to m, the contour function associated to m is

$$\begin{aligned} pl(\mathbf{x}) &= \prod_{i=1}^{n} pl(\mathbf{z}_i) pl(\mathbf{w}_i) && (37a)\\ &= \prod_{i=1}^{n} \left(pl(\mathbf{z}_i) \prod_{j=1}^{p} pl(w_i^j) \right) && (37b)\\ &= \prod_{i=1}^{n} \left(\prod_{k=1}^{g} (pl_{ik})^{z_{ik}} \prod_{j=1}^{p} \prod_{h=1}^{r_j} (pl_i^{jh})^{w_i^{jh}} \right), && (37c) \end{aligned}$$

where pl_{ik} is the plausibility that object i belongs to class k, and pl_i^{jh} is the plausibility that attribute j takes modality h for object i.

The observed-data likelihood for this model has the following expression:

$$\begin{aligned} L(\boldsymbol{\theta}; pl) &= \prod_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}}[pl(\mathbf{X}_i)] && (38a)\\ &= \prod_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}}\left[pl(\mathbf{W}_i) pl(\mathbf{Z}_i)\right] && (38b)\\ &= \prod_{i=1}^{n} \sum_{k=1}^{g} \mathbb{E}_{\boldsymbol{\theta}}\left[pl(\mathbf{W}_i) pl(\mathbf{Z}_i) | Z_{ik} = 1\right] \pi_k && (38c)\\ &= \prod_{i=1}^{n} \sum_{k=1}^{g} \pi_k pl_{ik} \prod_{j=1}^{p} \sum_{h=1}^{r_j} pl_i^{jh} \alpha_k^{jh}. && (38d) \end{aligned}$$

5.2 Solution

The derivation of the E^{2}M algorithm for the above model[3] is detailed in Appendix B. Only the main equations are given here without proof.

E-step: The expectation of the complete-data log-likelihood $\log L(\boldsymbol{\theta}; \mathbf{x})$ with respect to $p(\cdot | pl; \boldsymbol{\theta}^{(q)}) = p(\cdot; \boldsymbol{\theta}^{(q)}) \oplus pl$, where $\boldsymbol{\theta}^{(q)}$ is the current estimate of $\boldsymbol{\theta}$ at iteration q, is:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \sum_{i,k} t_{ik}^{(q)} \log \pi_k + \sum_{i,j,k,h} \beta_{ik}^{jh(q)} \log \alpha_k^{jh}, \tag{39}$$

where $t_{ik}^{(q)}$ and $\beta_{ik}^{jh(q)}$ are, respectively, the expectations of Z_{ik} and $W_i^{jh} Z_{ik}$ with respect to $p(\cdot | pl; \boldsymbol{\theta}^{(q)})$. They are given by the following equations:

$$t_{ik}^{(q)} = \frac{\pi_k^{(q)} pl_{ik} \prod_j \sum_h \alpha_k^{jh(q)} pl_i^{jh}}{\sum_\ell \pi_\ell^{(q)} pl_{i\ell} \prod_j \sum_h \alpha_\ell^{jh(q)} pl_i^{jh}} \tag{40}$$

and

$$\beta_{ik}^{jh(q)} = \gamma_{ik}^{jh(q)} t_{ik}^{(q)} \tag{41}$$

with

$$\gamma_{ik}^{jh(q)} = \frac{\alpha_k^{jh(q)} pl_i^{jh}}{\sum_{h'} \alpha_k^{jh'(q)} pl_i^{jh'}}. \tag{42}$$

3. Matlab source code for all the algorithms described in this paper can be downloaded from the author's web page at `http://www.hds.utc.fr/~tdenoeux`.

TABLE 3
Conditional probability vectors for each attribute j given each class k for the simulated categorical data.

$\boldsymbol{\alpha}_k^j$	$k=1$	$k=2$	$k=3$
$j=1$	(.1, .1, .1, .7)	(.1, .1, .7, .1)	(.1, .7, .1, .1)
$j=2$	(.7, .1, .1, .1)	(.1, .1, .1, .7)	(.1, .7, .1, .1)
$j=3$	(.7, .1, .1, .1)	(.1, .7, .1, .1)	(.1, .1, .7, .1)

M-step: The maximum of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ is then obtained at the M-step for the following parameter estimates:

$$\pi_k^{(q+1)} = \frac{1}{n} \sum_{i=1}^{n} t_{ik}^{(q)} \tag{43}$$

and

$$\alpha_k^{jh(q+1)} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{jh(q)} t_{ik}^{(q)}}{\sum_{i=1}^{n} t_{ik}^{(q)}}. \tag{44}$$

5.3 Experimental results

The above approach was first applied to simulated data with $g = 3$ equiprobable classes and $p = 3$ attributes (with $r_j = 4$ modalities each). The conditional probability distributions of attributes given each class are shown in Table 3.

A dataset of size $n = 500$ was generated from this distribution. To simulate uncertainty on attributes and class labels, we proceeded as follows. For each object i and each attribute j, an error probability q_i^j was drawn randomly from a beta distribution with mean ρ and standard deviation 0.2. With probability q_i^j, the value w_i^j of attribute j for object i was then replaced by a completely random value $\widetilde{w}_i^j$ (with a uniform distribution over the four modalities). The plausibilities pl_i^{jh} were then determined as

$$pl_i^{jh} = P(w_i^{jh} = 1 | \widetilde{w}_i^{jh}) = \begin{cases} q_i^j/4 & \text{if } \widetilde{w}_i^{jh} = 0, \\ q_i^j/4 + 1 - q_i^j & \text{if } \widetilde{w}_i^{jh} = 1. \end{cases}$$

Similarly, class labels $\mathbf{z}_i$ were replaced with noisy versions $\widetilde{\mathbf{z}}_i$ with probabilities q_i drawn from the same distributions as the q_i^js, and plausibilities pl_{ik} were computed in a similar way.

The whole experiment (generation of initial and noisy data) was repeated 20 times. For each obtained data set, the E^{2}M algorithm was applied to

1) Uncertain labels pl_{ik} and uncertain attributes pl_i^{jh};
2) Noisy labels $\widetilde{z}_{ik}$ and noisy attributes $\widetilde{w}_i^{jh}$;
3) Uncertain attributes pl_i^{jh} and no information on class labels;
4) Noisy attributes $\widetilde{w}_i^{jh}$ and no information on class labels.

In each case, the E^{2}M algorithm was run 5 times with random initial conditions, and the best solution according to the observed-data likelihood was retained. Each object was then assigned to the class with the largest estimated posterior probability, and the obtained partition was compared to the true partition using the adjusted Rand index [65]. We recall that this commonly used clustering performance measure is a corrected-for-chance version of the Rand index, which equals 0 on average for a random partition, and 1 when comparing two identical partitions.

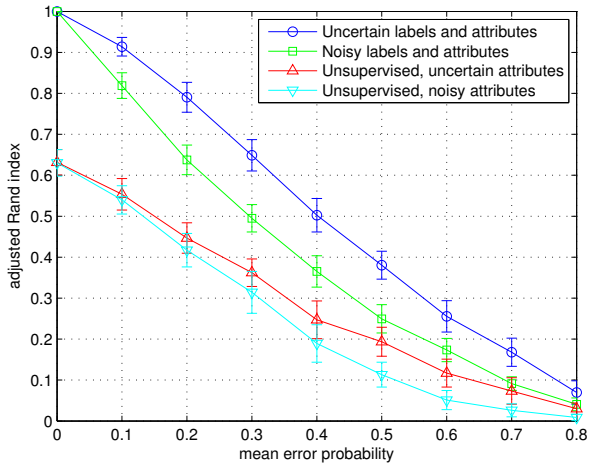

Fig. 2. Average values (plus and minus one standard deviation) of the adjusted Rand index over the 20 repetitions, as a function of the mean error probability ρ for the simulated categorical data.

The results are shown graphically in Figure 2. As expected, a degradation of the clustering performances is observed when the mean error probability ρ increases, with the adjusted rand index tending to zero as ρ tends to 1. More importantly, Figure 2 shows that the use of uncertain labels and/or uncertain attribute values (in the form of plausibilities) allows us to reach better clustering results than those obtained using noisy labels and/or attributes. These results show that our method is able to exploit additional information on data uncertainty, when such information is available.

We may remark here that, in the previous experiment, uncertain labels and attributes had to be generated by a random process. One could object that this random process, if known, could be incorporated in the generative model, and the standard EM algorithm could be used. It must be emphasized, however, that mass or plausibility functions in our model introduced in Section 3 represent epistemic uncertainty, which cannot be represented within a random generative model. They are only generated randomly here for the purpose of the experiment.

Our approach was also tested on the Breast Cancer dataset available from the UCI Machine Learning Repository[4]. This is a two-class dataset with 201 instances of one class and 85 instances of the other class. The instances are described by 9 discrete attributes treated as

4. This data set is available at `http://archive.ics.uci.edu/ml`. It was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia.

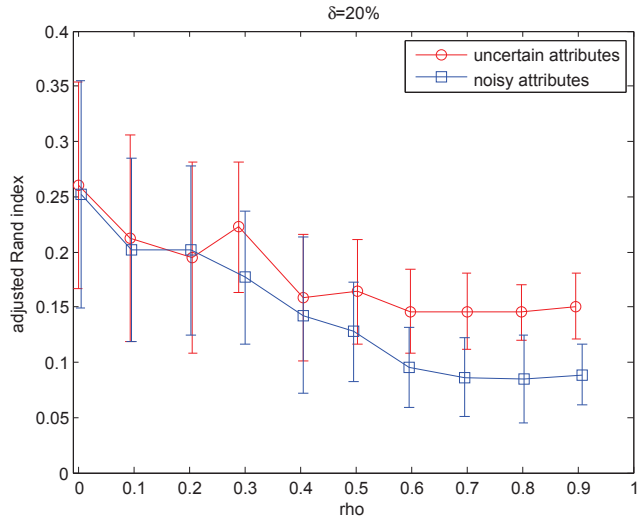

Fig. 3. Average values (plus and minus one standard deviation) of the adjusted Rand index over 20 repetitions, as a function of the mean error probability ρ for the Breast Cancer data with $\delta = 20\%$.

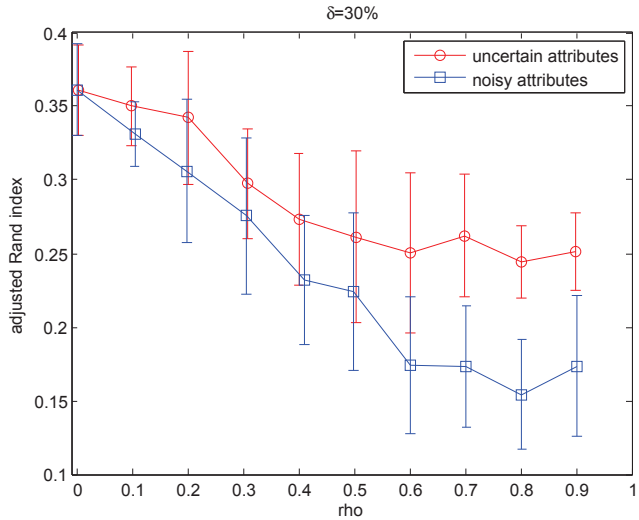

Fig. 4. Average values (plus and minus one standard deviation) of the adjusted Rand index over 20 repetitions, as a function of the mean error probability ρ for the Breast Cancer data with $\delta = 30\%$.

categorical. Noisy attributes $\widetilde{w}_i^{jh}$ and uncertain attributes pl_i^{jh} were generated in exactly the same way as in the previous experiment. We considered a semi-supervised learning task, in which a proportion δ of the objects is labeled. As before, we estimated the model parameters, assigned each unlabeled object to the class with the largest estimated posterior probability, and evaluated the result using the adjusted Rand index. The experiment was repeated 20 times. Figures 3 and 4 show the average values of the adjusted Rand index as a function of the mean error probability ρ, for noisy and uncertain attributes, with proportions of labeled data equal to $\delta = 20\%$ and $\delta = 30\%$, respectively. Once again, better results are obtained when taking into account attribute uncertainty, which demonstrates the usefulness of our approach.

6 Clustering of continuous data

Gaussian mixture models are commonly used for clustering continuous data (see, e.g., [66], [67]). In this section, we investigate the estimation of parameters in such models, when uncertainty on attributes is represented by belief functions with Gaussian contour functions, and partial information on class labels may also be available in the form of arbitrary mass functions. As in the previous section the model will first be introduced in Subsection 6.1. The estimation algorithm will then be described in Subsection 6.2 and simulation results will be presented in Subsection 6.3.

6.1 Problem description and generative model

The notations and assumptions are basically the same as those in Subsection 5.1, except that the attribute vector $\mathbf{W}_i = (W_i^1, \ldots, W_i^p)$ for object i is now assumed to be normally distributed conditionally on $z_{ik} = 1$, with mean $\boldsymbol{\mu}_k = (\mu_k^1, \ldots, \mu_k^p)$ and diagonal covariance matrix $\Sigma_k = \text{diag}[(\sigma_k^1)^2, \ldots, (\sigma_k^p)^2]$:

$$p(\mathbf{w}_i | z_{ik} = 1; \boldsymbol{\mu}_k, \Sigma_k) = \phi(\mathbf{w}_i; \boldsymbol{\mu}_k, \Sigma_k) = \prod_{j=1}^{p} \phi(w_i^j; \mu_k^j, (\sigma_k^j)^2). \quad (45)$$

Denoting by $\boldsymbol{\theta} = (\{\boldsymbol{\mu}_k\}, \{\Sigma_k\}, \{\pi_k\})$ the vector of all parameters, the complete data pdf is

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{z}_i; \boldsymbol{\theta}) p(\mathbf{w}_i | \mathbf{z}_i; \boldsymbol{\theta}) \quad (46a)$$

$$= \prod_{i=1}^{n} \prod_{k=1}^{g} (\pi_k \phi(\mathbf{w}_i; \boldsymbol{\mu}_k, \Sigma_k))^{z_{ik}}. \quad (46b)$$

As before, assuming cognitive independence (21), partial knowledge about $\mathbf{x}$ will be assumed to be represented by a mass function m with contour function

$$pl(\mathbf{x}) = \prod_{i=1}^{n} pl(\mathbf{z}_i) pl(\mathbf{w}_i) \quad (47a)$$

$$= \prod_{i=1}^{n} \left(\prod_{k=1}^{g} (pl_{ik})^{z_{ik}} \prod_{j=1}^{p} pl(w_i^j) \right). \quad (47b)$$

Assuming the marginal contour function $pl(w_i^j)$ on w_i^j to be a normalized Gaussian with mean m_i^j and standard deviation s_i^j: $pl(w_i^j) = \phi(w_i^j; m_i^j, (s_i^j)^2) s_i^j \sqrt{2\pi}$, we can

write

$$pl(\mathbf{x}_i) = (2\pi)^{p/2} \left(\prod_j s_i^j \right) \times \left(\prod_{k=1}^{g} (pl_{ik})^{z_{ik}} \right) \phi(\mathbf{w}_i; \mathbf{m}_i, S_i), \quad (48)$$

with $\mathbf{m}_i = (m_i^1, \ldots, m_i^p)$ and $S_i = \text{diag}[(s_i^1)^2, \ldots, (s_i^p)^2]$.

Using the same derivation as in (38a)-(38d), we get the following expression for the observed-data likelihood:

$$L(\boldsymbol{\theta}; pl) = \prod_{i=1}^{n} \sum_{k=1}^{g} \pi_k pl_{ik} \mathbb{E}_{\boldsymbol{\theta}} \left[pl(\mathbf{W}_i) | Z_{ik} = 1 \right], \quad (49)$$

with (using (13b) and the conditional independence of the attributes W_i^j given $Z_{ik} = 1$):

$$\mathbb{E}_{\boldsymbol{\theta}} \left[pl(\mathbf{W}_i) | Z_{ik} = 1 \right] = (2\pi)^{p/2} \prod_{j=1}^{p} \phi_{ik}^j s_i^j \quad (50)$$

and $\phi_{ik}^j = \phi(m_i^j; \mu_k^j, (\sigma_k^j)^2 + (s_i^j)^2)$.

6.2 Solution

A before, only the main equations of the E^{2}M algorithm for the above model will be given in this section. The derivation of these equations is detailed in Appendix C.

E-step: The expectation of the complete-data log-likelihood $\log L(\boldsymbol{\theta}; \mathbf{x})$ with respect to $p(\cdot; \boldsymbol{\theta}^{(q)}) \oplus pl$ is then

$$\begin{aligned} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}) = & \sum_{k=1}^{g} \left(\log \pi_k \sum_{i=1}^{n} t_{ik}^{(q)} \right) \\ & - \frac{np}{2} \log(2\pi) - \sum_{i,k} t_{ik}^{(q)} \sum_j \log \sigma_k^j \\ & - \frac{1}{2} \sum_{i,k} t_{ik}^{(q)} \sum_j \frac{\xi_{ik}^{j(q)} - 2\mu_{ik}^{j(q)} \mu_k^j + (\mu_k^j)^2}{(\sigma_k^j)^2}, \end{aligned} \quad (51)$$

where

$$\mu_{ik}^{j(q)} = \frac{\mu_k^{j(q)} (s_i^j)^2 + m_i^j (\sigma_k^{j(q)})^2}{(s_i^j)^2 + (\sigma_k^{j(q)})^2}, \quad (52)$$

$$(\sigma_{ik}^{j(q)})^2 = \frac{(\sigma_k^{j(q)})^2 (s_i^j)^2}{(s_i^j)^2 + (\sigma_k^{j(q)})^2}, \quad (53)$$

$$\xi_{ik}^{j(q)} = (\mu_{ik}^{j(q)})^2 + (\sigma_{ik}^{j(q)})^2 \quad (54)$$

and

$$t_{ik}^{j(q)} = \frac{\pi_k^{(q)} pl_{ik} \prod_j \phi_{ik}^{j(q)}}{\sum_\ell \pi_\ell^{(q)} pl_{i\ell} \prod_j \phi_{i\ell}^{j(q)}} \quad (55)$$

with

$$\phi_{ik}^{j(q)} = \phi(m_i^j; \mu_k^{j(q)}, (\sigma_k^{j(q)})^2 + (s_i^j)^2). \quad (56)$$

M-step: Maximizing (51) yields the following parameter estimates at iteration $q+1$:

$$\pi_k^{(q+1)} = \frac{1}{n} \sum_{i=1}^{n} t_{ik}^{j(q)}, \quad (57)$$

$$\mu_k^{j(q+1)} = \frac{\sum_{i=1}^{n} t_{ik}^{j(q)} \mu_{ik}^{j(q)}}{\sum_{i=1}^{n} t_{ik}^{j(q)}}, \quad (58)$$

$$(\sigma_k^{j(q+1)})^2 = \frac{\sum_{i=1}^{n} t_{ik}^{j(q)} (\xi_{ik}^{j(q)} - 2\mu_k^{j(q+1)} \mu_{ik}^{j(q)} + (\mu_k^{j(q+1)})^2)}{\sum_{i=1}^{n} t_{ik}^{j(q)}}. \quad (59)$$

6.3 Experimental results

The above algorithm was applied to simulated Gaussian data with $n = 200$ instances and the following parameter values:

$$\mu_1 = (3, 0, 0)', \quad \mu_2 = (0, 3, 0)', \quad \mu_3 = (0, 0, 3)',$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = I, \quad \pi_1 = \pi_2 = \pi_3 = 1/3.$$

We first considered the unsupervised learning situation in which no information on class labels is available, i.e., $pl_{ik} = 1$ for all i and k. To simulate uncertainty on attributes, we randomly generated for each object i and each attribute j a standard deviation s_i^j from the uniform distribution in $[0, s_{max}]$. The value w_i^j of attribute j for object i was then replaced by a random value $\widetilde{w}_i^j$ drawn from a Gaussian distribution with mean w_i^j and standard deviation s_i^j. We then set $m_i^j = \widetilde{w}_i^j$. The whole experiment (generation of initial and noisy data) was repeated 20 times. For each obtained dataset, the E^{2}M algorithm was applied to the noisy attributes $\widetilde{w}_i^j$ and to the uncertain attributes (m_i^j, s_i^j). In each case, the E^{2}M algorithm was run 5 times with random initial conditions, and the best solution according to the observed-data likelihood was retained. Each object was then assigned to the class with the largest estimated posterior probability, and the obtained partition was compared to the true partition using the adjusted Rand index. The results are shown in Figure 5. As we can see, the algorithm successfully exploits the additional information about attribute uncertainty, which allows us to better recover the true partition of the data.

To investigate the influence of class label uncertainty, we generated noisy class labels $\widetilde{z}_{ik}$ and plausibilities pl_{ik} as explained in Subsection 5.3. Attribute uncertainty was also simulated as above, with $s_{max} = 2$. Figure 6 shows the adjusted Rand index as a function of the mean error probability ρ on class labels, for the E^{2}M algorithm applied to data with uncertain and noisy labels, as well as to unsupervised data. Here again, uncertainty on class labels appears to be successfully exploited by the algorithm. Remarkably, the results with uncertain labels never get worse than those obtained without label

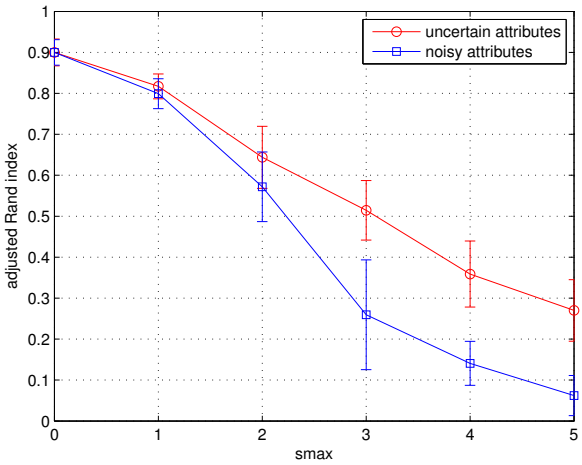

Fig. 5. Average values (plus and minus one standard deviation) of the adjusted Rand index over 20 repetitions, as a function of the maximal standard deviation s_{max} for the simulated Gaussian data, with no information on class labels.

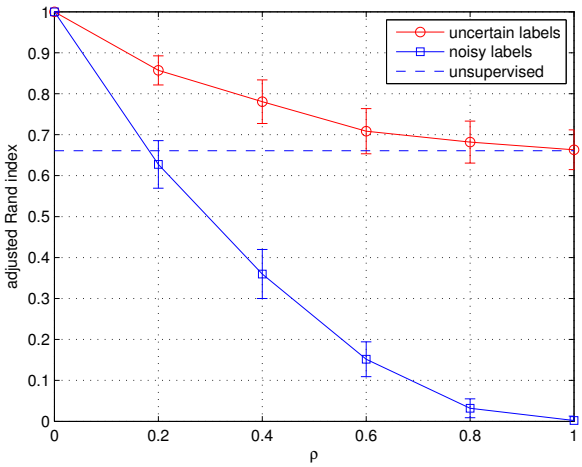

Fig. 6. Average values (plus and minus one standard deviation) of the adjusted Rand index over 20 repetitions, as a function of the error probability ρ on class labels for the simulated Gaussian data, with $s_{max} = 2$.

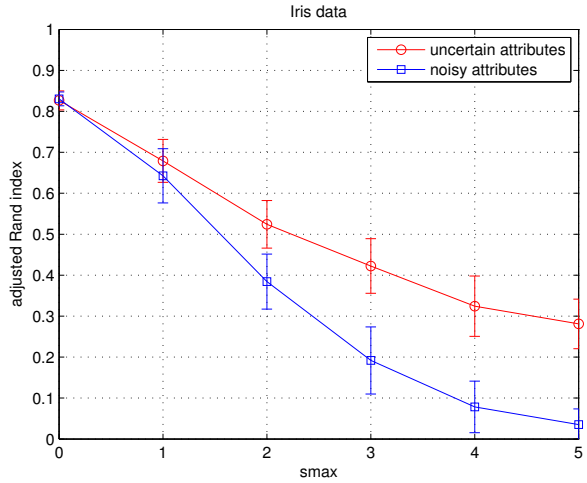

Fig. 7. Average values (plus and minus one standard deviation) of the adjusted Rand index over 20 repetitions, as a function of the maximal standard deviation s_{max} for the Iris data, with no information on class labels.

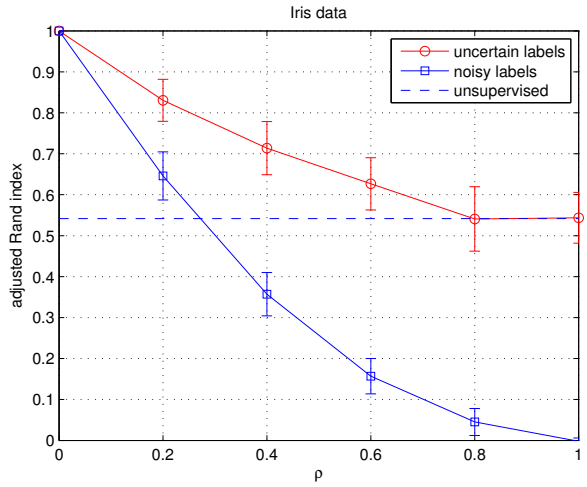

Fig. 8. Average values (plus and minus one standard deviation) of the adjusted Rand index over 20 repetitions, as a function of the error probability ρ on class labels for the Iris data, with $s_{max} = 2$.

information, even for error probabilities close to 1. This finding confirms a similar result reported in [43].

To corroborate the above results with real data, similar experiments were carried out with the well-known Iris dataset[5]. We recall that this dataset is composed of 150 4-dimensional attribute vectors partitioned in 3 classes, corresponding to three species of Iris. Unsupervised learning with attribute uncertainty generated as above was first carried out. The results, shown in Figure 7, are very similar to those obtained with simulated Gaussian data. Finally, results with both attribute uncertainty (with $s_{max} = 2$) and label uncertainty (generated as above) are shown in Figure 8. Again, these results are very similar to those obtained with simulated data, as shown in Figure 6.

5. This data set is available at `http://archive.ics.uci.edu/ml`.

7 Conclusion

A method for estimating parameters in statistical models in the case of uncertain observations has been introduced. The proposed formalism combines aleatory uncertainty captured by a parametric statistical model with epistemic uncertainty induced by an imperfect observation process and represented by belief functions. Our method then seeks the value of the unknown parameter that maximizes a generalized likelihood criterion, which can be interpreted as a degree of agreement between

the parametric model and the uncertain data. This is achieved using the evidential EM algorithm, which is a simple extension of the classical EM algorithm with proved convergence properties.

As an illustration, the method has been applied to clustering problems with partial knowledge of class labels and attributes, based on latent class and Gaussian mixture models. In these problems, our approach has been shown to successfully exploit the additional information about data uncertainty, resulting in improved performances in the clustering task.

More generally, the approach introduced in this paper is applicable to any uncertain data mining problem in which a parametric statistical model can be postulated and data uncertainty arises form an imperfect observation process. This includes a wide range of problems such as classification, regression, feature extraction and time series prediction.

APPENDIX A
PROOF OF THEOREM 1

Let $g(\mathbf{x}|pl;\boldsymbol{\theta})$ be defined using the following expression:

$$g(\mathbf{x}|pl;\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta};\mathbf{x})}{L(\boldsymbol{\theta};pl)} = \frac{p_{\mathbf{X}}(\cdot|pl;\boldsymbol{\theta})}{pl(\mathbf{x})}, \tag{60}$$

where the rightmost expression is derived from (27). We thus have

$$\log L(\boldsymbol{\theta};pl) = \log L(\boldsymbol{\theta};\mathbf{x}) - \log g(\mathbf{x}|pl;\boldsymbol{\theta}). \tag{61}$$

Taking the expectation of both sides with respect to $p_{\mathbf{X}}(\cdot|pl;\boldsymbol{\theta}^{(q)})$, we get

$$\begin{aligned}\log L(\boldsymbol{\theta};pl) &= \\ &\mathbb{E}_{\boldsymbol{\theta}^{(q)}}\left[\log L(\boldsymbol{\theta};\mathbf{X})|pl\right] - \mathbb{E}_{\boldsymbol{\theta}^{(q)}}\left[\log g(\mathbf{X}|pl;\boldsymbol{\theta})|pl\right] = \\ &Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(q)}) - H(\boldsymbol{\theta},\boldsymbol{\theta}^{(q)})\end{aligned} \tag{62}$$

with

$$H(\boldsymbol{\theta},\boldsymbol{\theta}^{(q)}) = \mathbb{E}_{\boldsymbol{\theta}^{(q)}}\left[\log g(\mathbf{X}|pl;\boldsymbol{\theta})|pl\right]. \tag{63}$$

We thus have

$$\begin{aligned}&\log L(\boldsymbol{\theta}^{(q+1)};pl) - \log L(\boldsymbol{\theta}^{(q)};pl) \\ &\quad = Q(\boldsymbol{\theta}^{(q+1)},\boldsymbol{\theta}^{(q)}) - Q(\boldsymbol{\theta}^{(q)},\boldsymbol{\theta}^{(q)}) \\ &\quad - \left(H(\boldsymbol{\theta}^{(q+1)},\boldsymbol{\theta}^{(q)}) - H(\boldsymbol{\theta}^{(q)},\boldsymbol{\theta}^{(q)})\right).\end{aligned} \tag{64}$$

The first difference on the right-hand side of (64) is nonnegative as $\boldsymbol{\theta}^{(q+1)}$ has been chosen to maximize $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(q)})$ with respect to $\boldsymbol{\theta}$. It thus remains to check that the second difference on the right-hand side of (64) is non-positive; that is, we need to verify that the following inequality holds: $H(\boldsymbol{\theta}^{(q+1)},\boldsymbol{\theta}^{(q)}) - H(\boldsymbol{\theta}^{(q)},\boldsymbol{\theta}^{(q)}) \leq 0$. Now for any $\boldsymbol{\theta}$,

$$\begin{aligned}&H(\boldsymbol{\theta},\boldsymbol{\theta}^{(q)}) - H(\boldsymbol{\theta}^{(q)},\boldsymbol{\theta}^{(q)}) = \\ &\qquad \mathbb{E}_{\boldsymbol{\theta}^{(q)}}\left[\log\frac{g(\mathbf{X}|pl;\boldsymbol{\theta})}{g(\mathbf{X}|pl;\boldsymbol{\theta}^{(q)})}|pl\right] \\ &\qquad \leq \log\mathbb{E}_{\boldsymbol{\theta}^{(q)}}\left[\frac{g(\mathbf{X}|pl;\boldsymbol{\theta})}{g(\mathbf{X}|pl;\boldsymbol{\theta}^{(q)})}|pl\right]\end{aligned} \tag{65}$$

and

$$\begin{aligned}&\log\mathbb{E}_{\boldsymbol{\theta}^{(q)}}\left[\frac{g(\mathbf{X}|pl;\boldsymbol{\theta})}{g(\mathbf{X}|pl;\boldsymbol{\theta}^{(q)})}|pl\right] = \\ &\quad \log\int\frac{g(\mathbf{x}|pl;\boldsymbol{\theta})}{g(\mathbf{x}|pl;\boldsymbol{\theta}^{(q)})}pl(\mathbf{x})g(\mathbf{x}|pl;\boldsymbol{\theta}^{(q)})d\mathbf{x} = \\ &\quad \log\int g(\mathbf{x}|pl;\boldsymbol{\theta})pl(\mathbf{x})d\mathbf{x} = \\ &\quad \log\int p_{\mathbf{X}}(\mathbf{x}|pl;\boldsymbol{\theta}^{(q)})d\mathbf{x} = 0,\end{aligned} \tag{66}$$

where the inequality in (65) is a consequence of Jensen's inequality. □

APPENDIX B
DERIVATION OF THE E^{2}M ALGORITHM FOR THE CATEGORICAL DATA CLUSTERING PROBLEM

From (35) and (36c), the complete-data log-likelihood is

$$\log L(\boldsymbol{\theta};\mathbf{x}) = \sum_{i,k} z_{ik}\log\pi_k + \sum_{i,j,k,h} w_i^{jh} z_{ik}\log\alpha_k^{jh}. \tag{67}$$

By taking its expectation with respect to $p(\cdot|pl;\boldsymbol{\theta}^{(q)}) = p(\cdot;\boldsymbol{\theta}^{(q)}) \oplus pl$, where $\boldsymbol{\theta}^{(q)}$ is the current estimate of $\boldsymbol{\theta}$ at iteration q, we get expression (39) for $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(q)})$. To compute this expression, we may first observe that, from (35) and (21):

$$p(\mathbf{x}|pl;\boldsymbol{\theta}^{(q)}) = \prod_{i=1}^{n} p(\mathbf{w}_i,\mathbf{z}_i|pl;\boldsymbol{\theta}^{(q)}). \tag{68}$$

Then, using (36c) and (37c), we get:

$$\begin{aligned}&p(\mathbf{w}_i, z_{ik}=1|pl;\boldsymbol{\theta}^{(q)}) = \\ &\qquad \frac{\pi_k^{(q)} pl_{ik}\prod_{j,h}(\alpha_k^{jh(q)} pl_i^{jh})^{w_i^{jh}}}{\sum_\ell \pi_\ell^{(q)} pl_{i\ell}\prod_j\sum_h \alpha_\ell^{jh(q)} pl_i^{jh}}.\end{aligned} \tag{69}$$

Marginalizing out $\mathbf{w}_i$ leads to the following expression for $p(Z_{ik}=1|pl;\boldsymbol{\theta}^{(q)})$:

$$p(Z_{ik}=1|pl;\boldsymbol{\theta}^{(q)}) = \frac{\pi_k^{(q)} pl_{ik}\prod_j\sum_h \alpha_k^{jh(q)} pl_i^{jh}}{\sum_\ell \pi_\ell^{(q)} pl_{i\ell}\prod_j\sum_h \alpha_\ell^{jh(q)} pl_i^{jh}}, \tag{70}$$

from which we get (40).

Conditioning (69) on $z_{ik}=1$, we get:

$$\begin{aligned}p(\mathbf{w}_i|z_{ik}=1,pl;\boldsymbol{\theta}^{(q)}) &= \frac{p(\mathbf{w}_i, z_{ik}=1|pl;\boldsymbol{\theta}^{(q)})}{p(z_{ik}=1|pl;\boldsymbol{\theta}^{(q)})} = \\ &\prod_j \frac{\prod_h\left(\alpha_k^{jh(q)} pl_i^{jh}\right)^{w_i^{jh}}}{\sum_h \alpha_k^{jh(q)} pl_i^{jh}}.\end{aligned} \tag{71}$$

Now, $\beta_{ik}^{jh(q)} = \mathbb{E}_{\boldsymbol{\theta}^{(q)}}[W_i^{jh} Z_{ik}|pl]$ can be written as

$$\beta_{ik}^{jh(q)} = p(W_i^{jh} = 1, Z_{ik} = 1|pl; \boldsymbol{\theta}^{(q)}) = p(W_i^{jh} = 1|Z_{ik} = 1, pl; \boldsymbol{\theta}^{(q)}) p(Z_{ik} = 1|pl; \boldsymbol{\theta}^{(q)}), \quad (72)$$

from which we obtain (41) and (42).

The update equations (43) and (44) at the M-step are obtained by maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ with respect to parameters π_k and α_k^{jh}, taking into account the equality constraints $\sum_k \pi_k = 1$ and $\sum_h \alpha_k^{jh} = 1$. The calculations are the same as in the classical latent class model without data uncertainty.

Appendix C
Derivation of the E²M algorithm for the continuous data clustering problem

Using (46b), the complete-data log-likelihood is obtained as:

$$\log L(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \log \left(\pi_k \prod_{j=1}^{p} \phi(w_i^j; \mu_k^j, (\sigma_k^j)^2) \right), \quad (73)$$

which can be developed as

$$\log L(\boldsymbol{\theta}; \mathbf{x}) = \sum_{k=1}^{g} \left(\log \pi_k \sum_{i=1}^{n} z_{ik} \right) + \sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \sum_{j=1}^{p} \log \phi(w_i^j; \mu_k^j, (\sigma_k^j)^2) \quad (74)$$

with

$$\log \phi(w_i^j; \mu_k^j, (\sigma_k^j)^2) = -\frac{1}{2} \log(2\pi) - \log \sigma_k^j - \frac{(w_i^j - \mu_k^j)^2}{2(\sigma_k^j)^2}. \quad (75)$$

We thus have

$$\log L(\boldsymbol{\theta}; \mathbf{x}) = \sum_{k=1}^{g} \left(\log \pi_k \sum_{i=1}^{n} z_{ik} \right) - \frac{np}{2} \log(2\pi) - \sum_{i,k} z_{ik} \sum_j \log \sigma_k^j - \frac{1}{2} \sum_{i,k} z_{ik} \sum_j \frac{(w_i^j)^2 - 2 w_i^j \mu_k^j + (\mu_k^j)^2}{(\sigma_k^j)^2}. \quad (76)$$

To compute the expectation $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)})$ of $\log L(\boldsymbol{\theta}; \mathbf{x})$ with respect to $p(\cdot|pl, \boldsymbol{\theta}^{(q)}) = p(\cdot; \boldsymbol{\theta}^{(q)}) \oplus pl$, where $\boldsymbol{\theta}^{(q)}$ is the current estimate of $\boldsymbol{\theta}$ at iteration q, we need to compute the expectations of Z_{ik}, $Z_{ik}(W_i^j)^2$ and $Z_{ik} W_i^j$. For that purpose, we may first observe that, similarly to the previous model:

$$p(\mathbf{x}|pl; \boldsymbol{\theta}^{(q)}) = \prod_{i=1}^{n} p(\mathbf{w}_i, \mathbf{z}_i|pl; \boldsymbol{\theta}^{(q)}). \quad (77)$$

Then, using (46b) and (48), we get:

$$p(\mathbf{w}_i, z_{ik} = 1|pl; \boldsymbol{\theta}^{(q)}) = \frac{\pi_k^{(q)} pl_{ik} \phi(\mathbf{w}_i; \boldsymbol{\mu}_k^{(q)}, \Sigma_k^{(q)}) \phi(\mathbf{w}_i; \mathbf{m}_i, S_i)}{\sum_\ell \pi_\ell^{(q)} pl_{i\ell} \int \phi(\mathbf{w}_i; \boldsymbol{\mu}_\ell^{(q)}, \Sigma_\ell^{(q)}) \phi(\mathbf{w}_i; \mathbf{m}_i, S_i) d\mathbf{w}_i}. \quad (78)$$

From (13a)-(13b), this expression may be simplified as

$$p(\mathbf{w}_i, z_{ik} = 1|pl; \boldsymbol{\theta}^{(q)}) = \frac{\pi_k^{(q)} pl_{ik} \prod_j \phi_{ik}^{j(q)} \phi(w_i^j; \mu_{ik}^{j(q)}, (\sigma_{ik}^{j(q)})^2)}{\sum_\ell \pi_\ell^{(q)} pl_{i\ell} \prod_j \phi_{i\ell}^{j(q)}}, \quad (79)$$

with $\mu_{ik}^{j(q)}$, $(\sigma_{ik}^{j(q)})^2$ and $\phi_{ik}^{j(q)}$ given, respectively, by (52), (53) and (56).

Marginalizing out $\mathbf{w}_i$ in (79), we obtain

$$P(Z_{ik} = 1|pl, \boldsymbol{\theta}^{(q)}) = \frac{\pi_k^{(q)} pl_{ik} \prod_j \phi_{ik}^{j(q)}}{\sum_\ell \pi_\ell^{(q)} pl_{i\ell} \prod_j \phi_{i\ell}^{j(q)}}. \quad (80)$$

from which we get the expression of $t_{ik}^{j(q)} = \mathbb{E}_{\boldsymbol{\theta}^{(q)}}[Z_{ik}|pl]$ as (55). Conditioning (79) by $z_{ik} = 1$ leads to the following conditional pdf:

$$p(\mathbf{w}_i|z_{ik} = 1, pl; \boldsymbol{\theta}^{(q)}) = \frac{\pi_k^{(q)} pl_{ik} \prod_j \phi_{ik}^{j(q)} \phi(w_i^j; \mu_{ik}^{j(q)}, (\sigma_{ik}^{j(q)})^2)}{\pi_k^{(q)} pl_{ik} \prod_j \phi_{ik}^{j(q)}} = \prod_j \phi(w_i^j; \mu_{ik}^{j(q)}, (\sigma_{ik}^{j(q)})^2). \quad (81)$$

Consequently,

$$\mathbb{E}_{\boldsymbol{\theta}^{(q)}}[Z_{ik}(W_i^j)^2|pl] = \mathbb{E}_{\boldsymbol{\theta}^{(q)}}[(W_i^j)^2|Z_{ik} = 1, pl] P(Z_{ik} = 1|pl; \boldsymbol{\theta}^{(q)}) = \xi_{ik}^{j(q)} t_{ik}^{j(q)} \quad (82)$$

with $\xi_{ik}^{j(q)}$ defined by (54) and

$$\mathbb{E}_{\boldsymbol{\theta}^{(q)}}[Z_{ik} W_i^j|pl] = \mathbb{E}_{\boldsymbol{\theta}^{(q)}}[W_i^j|Z_{ik} = 1, pl] P(Z_{ik} = 1|pl; \boldsymbol{\theta}^{(q)}) = \mu_{ik}^{j(q)} t_{ik}^{j(q)}. \quad (83)$$

The update equations (57)-(59) are obtained as in the standard Gaussian mixture model by maximizing $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)})$ with respect to parameters π_k, μ_k^j and σ_k^j under the equality constraint $\sum_k \pi_k = 1$.

References

[1] C. C. Aggarwal and P. S. Yu, "A survey of uncertain data algorithms and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 5, pp. 609–623, 2009.

[2] C. C. Aggarwal, *Managing and Mining Uncertain Data*, ser. Advances in database systems. Springer, 2009, no. 35.

[3] R. Cheng, M. Chau, M. Garofalakis, and J. X. Yu, "Guest editors' introduction: Special section on mining large uncertain and probabilistic databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 9, p. 1201, 2010.

[4] M. A. Cheema, X. Lin, W. Wang, W. Zhang, and J. Pei, "Probabilistic reverse nearest neighbor queries on uncertain data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 4, pp. 550–564, 2010.

[5] S. Tsang, B. Kao, K. Yip, W. Ho, and S. Lee, "Decision trees for uncertain data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 1, pp. 64–78, 2011.

[6] H.-P. Kriegel and M. Pfeifle, "Density-based clustering of uncertain data," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. Chicago, Illinois, USA: ACM, 2005, pp. 672 – 677.

[7] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip, "Efficient clustering of uncertain data," in *Sixth International Conference on Data Mining (ICDM '06)*, Hong Kong, 2006, pp. 436–445.

[8] B. Kao, S. D. Lee, F. Lee, D. Cheung, and W.-S. Ho, "Clustering uncertain data using voronoi diagrams and r-tree index," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 9, pp. 1219–1233, 2010.

[9] S. Günnemann, H. Kremer, and T. Seidl, "Subspace clustering for uncertain data," in *SIAM International Conference on Data Mining (SDM 2010)*, Columbus, Ohio, USA, 2010, pp. 385–396.

[10] C. C. Aggarwal, "On density based transforms for uncertain data mining," in *IEEE 23rd International Conference on Data Engineering (ICDE 2007)*, Istanbul, 2007, pp. 866–875.

[11] C. C. Aggarwal and P. S. Yu, "Outlier detection with uncertain data," in *Proceedings of the SIAM International Conference on Data Mining (SDM 2008)*, Atlanta, Georgia, USA, 2008, pp. 483–493.

[12] J. Bi and T. Zhang, "Support vector classification with input data uncertainty," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. Cambridge, MA: MIT Press, 2005, pp. 161–168.

[13] L. Billard and E. Diday, *Symbolic Data Analysis*. Chichester, England: Wiley, 2006.

[14] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy Sets and Systems*, vol. 1, pp. 3–28, 1978.

[15] J. Gebhardt, M. A. Gil, and R. Kruse, "Fuzzy set-theoretic methods in statistics," in *Fuzzy sets in decision analysis, operations research and statistics*, R. Slowinski, Ed. Boston: Kluwer Academic Publishers, 1998, pp. 311–347.

[16] P. Cazes, A. Chouakria, E. Diday, and Y. Schektman, "Extension de l'analyse en composantes principales à des données de type intervalle," *Revue de Statistique Appliquée*, vol. 14, no. 3, pp. 5–24, 1997.

[17] T. Denœux and M.-H. Masson, "Principal component analysis of fuzzy data using autoassociative neural networks," *IEEE Transactions on Fuzzy Systems*, vol. 12, no. 3, pp. 336–349, 2004.

[18] P. Giordani and H. A. L. Kiers, "A comparison of three methods for principal component analysis of fuzzy interval data," *Computational Statistics and Data Analysis*, vol. 51, no. 1, pp. 379–397, 2006.

[19] K. C. Gowda and E. Diday, "Symbolic clustering using a new similarity measure," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 2, pp. 368–378, 1992.

[20] P. D'Urso and P. Giordani, "A weighted fuzzy c-means clustering model for fuzzy data," *Computational Statistics and Data Analysis*, vol. 50, no. 6, pp. 1496–1523, 2006.

[21] F. D. T. D. Carvalho and Y. Lechevallier, "Partitional clustering algorithms for symbolic interval data based on single adaptive distances," *Pattern Recognition*, vol. 42, no. 7, pp. 1223–1236, 2009.

[22] H. Tanaka, "Fuzzy data analysis by possibilistic linear models," *Fuzzy Sets and Systems*, vol. 24, p. 363ñ375, 1987.

[23] R. Coppi, "Management of uncertainty in statistical reasoning: The case of regression," *International Journal of Approximate Reasoning*, vol. 47, no. 3, pp. 284–305, 2008.

[24] M. B. Ferraro, R. Coppi, G. G. Rodríguez, and A. Colubi, "A linear regression model for imprecise response," *International Journal of Approximate Reasoning*, vol. 51, no. 7, pp. 759–770, 2010.

[25] T. Denœux and M.-H. Masson, "Multidimensional scaling of interval-valued dissimilarity data," *Pattern Recognition Letters*, vol. 21, pp. 83–92, 2000.

[26] M.-H. Masson and T. Denœux, "Multidimensional scaling of fuzzy dissimilarity data," *Fuzzy Sets and Systems*, vol. 128, no. 3, pp. 339–352, 2002.

[27] P.-A. Hébert, M.-H. Masson, and T. Denœux, "Fuzzy multidimensional scaling," *Computational Statistics and Data Analysis*, vol. 51, no. 1, pp. 335–359, 2006.

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

[29] ——, "Upper and lower probabilities generated by a random closed interval," *Annals of Mathematical Statistics*, vol. 39, no. 3, pp. 957–966, 1968.

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

[31] P. Smets, "The combination of evidence in the Transferable Belief Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 447–458, 1990.

[32] P. Smets and R. Kennes, "The Transferable Belief Model," *Artificial Intelligence*, vol. 66, pp. 191–243, 1994.

[33] T. Denœux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 25, no. 05, pp. 804–813, 1995.

[34] T. Denœux and L. M. Zouhal, "Handling possibilistic labels in pattern classification using evidential reasoning," *Fuzzy Sets and Systems*, vol. 122, no. 3, pp. 47–62, 2001.

[35] 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*, vol. 36, no. 6, pp. 1395–1406, 2006.

[36] S. Petit-Renaud and T. Denœux, "Nonparametric regression analysis of uncertain and imprecise data using belief functions," *International Journal of Approximate Reasoning*, vol. 35, no. 1, pp. 1–28, 2004.

[37] T. Denœux and M. Skarstein-Bjanger, "Induction of decision trees for partially classified data," in *Proceedings of SMC'2000*. Nashville, TN: IEEE, October 2000, pp. 2923–2928.

[38] Z. Elouedi, K. Mellouli, and P. Smets, "Belief decision trees: Theoretical foundations," *International Journal of Approximate Reasoning*, vol. 28, pp. 91–124, 2001.

[39] S. Trabelsi, Z. Elouedi, and K. Mellouli, "Pruning belief decision tree methods in averaging and conjunctive approaches," *International Journal of Approximate Reasoning*, vol. 46, no. 3, pp. 568–595, 2007.

[40] S. Ben Hariz, Z. Elouedi, and K. Mellouli, "Clustering approach using belief function theory," in *Artificial Intelligence: Methodology, Systems, and Applications*, ser. Lecture Notes in Computer Science, J. Euzenat and J. Domingue, Eds. Springer Berlin / Heidelberg, 2006, vol. 4183, pp. 162–171.

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

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

[43] E. Côme, L. Oukhellou, T. Denœux, and P. Aknin, "Learning from partially supervised data using mixture models and belief functions," *Pattern Recognition*, vol. 42, no. 3, pp. 334–348, 2009.

[44] 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*, vol. B 39, pp. 1–38, 1977.

[45] E. Ramasso, "Contribution of belief functions to hidden Markov models with an application to fault diagnosis," in *IEEE International Worshop on Machine Learning for Signal Processing (MLSP'09)*. Grenoble, France: IEEE, 2009.

[46] T. Denœux, "Maximum likelihood from evidential data: an extension of the EM algorithm," in *Combining soft computing and statistical methods in data analysis (Proceedings of SMPS 2010)*, ser. Advances in Intelligent and Soft Computing, C. B. et al., Ed. Oviedeo, Spain: Springer, 2010, pp. 181–188.

[47] D. Dubois and H. Prade, *Possibility Theory: An approach to computerized processing of uncertainty*. New-York: Plenum Press, 1988.

[48] P. Smets, "Belief functions on real numbers," *International Journal of Approximate Reasoning*, vol. 40, no. 3, pp. 181–223, 2005.

[49] T. Denœux, "Extending stochastic ordering to belief functions on the real line," *Information Sciences*, vol. 179, pp. 1362–1376, 2009.

[50] J. O. Smith, *Spectral Audio Signal Processing, October 2008 Draft*. `https://ccrma.stanford.edu/~jos/sasp/`, accessed Feb. 2011, online book.

[51] D. Dubois and H. Prade, "Focusing vs. belief revision: A fundamental distinction when dealing with generic knowledge," in *Qualitative and quantitative practical reasoning*, ser. Lecture notes in computer science. Berlin, Germany: Springer-Verlag, 1997, vol. 1244, pp. 96–107.
[52] H. Nguyen, *An Introduction to Random Sets*. Boca Raton, Florida: Chapman and Hall/CRC Press, 2006.
[53] M. E. G. V. Cattaneo and A. Wiencierz, "Regression with imprecise data: A robust approach," in *Proceedings of the Seventh International Symposium on Imprecise Probabilities and Their Applications (ISIPTA'11)*, Innsbruck, Austria, 2011.
[54] M. A. Gil, M. López-Díaz, and D. A. Ralescu, "Overview on the development of fuzzy random variables," *Fuzzy Sets and Systems*, vol. 157, no. 19, pp. 2546–2557, 2006.
[55] G. González-Rodríguez, A. Colubi, P. D'Urso, and M. Montenegro, "Multi-sample test-based clustering for fuzzy random variables," *International Journal of Approximate Reasoning*, vol. 50, no. 5, pp. 721–731, 2009.
[56] G. Qi, W. Liu, and D. A. Bell, "Measuring conflict and agreement between two prioritized belief bases," in *Proceedings of the 19th international joint conference on Artificial intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2005, pp. 552–557.
[57] L. A. Zadeh, "Probability measures of fuzzy events," *J. Math. Analysis and Appl.*, vol. 10, pp. 421–427, 1968.
[58] T. Denœux, "Maximum likelihood estimation from fuzzy data using the fuzzy EM algorithm," *Fuzzy Sets and Systems*, vol. 183, no. 1, pp. 72–91, 2011.
[59] B. Quost and T. Denœux, "Clustering fuzzy data using the fuzzy EM algorithm," in *Proceedings of the 4th International Conference on Scalable Uncertainty Management (SUM2010)*, A. Deshpande and A. Hunter, Eds., no. LNAI-6379. Toulouse, France: Springer-Verlag, September 2010, pp. 333–346.
[60] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.
[61] 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*, vol. 41, no. 3, pp. 314–330, 2006.
[62] G. Shafer, "Constructive probability," *Synthese*, vol. 48, no. 1, pp. 1–60, 1981.
[63] L. A. Goodman, "Exploratory latent structure analysis using both identifiable and unidentifiable models," *Biometrika*, vol. 61, no. 2, pp. 215–231, 1974.
[64] G. Celeux and G. Govaert, "Clustering criteria for discrete data and latent class models," *Journal of Classification*, vol. 8, pp. 157–176, 1991.
[65] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–ñ218, 1985.
[66] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, vol. 49, pp. 803–821, 1993.
[67] G. Celeux and G. Govaert, "Gaussian parsimonious clustering models," *Pattern Recognition*, vol. 28, no. 5, pp. 781–793, 1995.

BIOGRAPHY

Thierry Denœux graduated in 1985 from the Ecole Nationale des Ponts et Chaussées in Paris, and received a doctorate from the same institution in 1989. Currently, he is a Full Professor with the Department of Information Processing Engineering at the Université de Technologie de Compiègne, France. His research interests concern belief functions theory, fuzzy data analysis and, more generally, the management of imprecision and uncertainty in data analysis, pattern recognition and information fusion. He is the Editor-in-Chief of the *International Journal of Approximate Reasoning*, and a member of the editorial board of *Fuzzy Sets and Systems*.