Maximum Likelihood from Evidential Data: An Extension of the EM Algorithm

Thierry Denœux

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. The method is illustrated using the problem of univariate normal mean and variance estimation from uncertain data.

Keywords: Belief functions, Dempster-Shafer theory, Statistical inference, Uncertain data.

1 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 [4, 8] 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 $\mathscr{X}$ 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

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

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 181–188.
springerlink.com

assessment of x (a process known as *labeling*). This assessment may take the form of a subset $A \subseteq \mathscr{X}$, a probability distribution p on $\mathscr{X}$ or, more generally, a mass function m on $\mathscr{X}$, i.e., a function $m : 2^{\mathscr{X}} \rightarrow [0,1]$. 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, and 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 [5, 6], 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, 7]. A rigorous solution to this problem, which is a special case of the general method presented in this paper, was introduced in [2].

The rest of the paper is organized as follows. The EM algorithm will first be recalled in Section 2. The extension of the likelihood function and the E^2M algorithm will then be introduced in Sections 3 and 4, respectively. Section 5 will demonstrate the application of this algorithm to the problem of univariate normal mean and variance estimation using uncertain data.

2 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 [4]. Formally, we assume the existence of two sample spaces $\mathscr{X}$ and $\mathscr{Y}$, and a many-to-one mapping $\boldsymbol{\varphi}$ from $\mathscr{X}$ to $\mathscr{Y}$. The observed (incomplete) data $\mathbf{y}$ are a realization from $\mathscr{Y}$, while the corresponding $\mathbf{x}$ in $\mathscr{X}$ is not observed and is only known to lie in the set

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

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 likelihood $L(\boldsymbol{\Psi})$ is related to $g_c(\mathbf{x}; \boldsymbol{\Psi})$ by

$$L(\boldsymbol{\Psi}) = \int_{\mathscr{X}(\mathbf{y})} g_c(\mathbf{x};\boldsymbol{\Psi}) d\mathbf{x}. \quad (1)$$

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

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

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

The M-step then consists in maximizing $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})$ with respect to $\boldsymbol{\Psi}$ over the parameter space $\boldsymbol{\Omega}$. The E- and M-steps are iterated until the difference $L(\boldsymbol{\Psi}^{(q+1)}) - L(\boldsymbol{\Psi}^{(q)})$ becomes smaller than some arbitrarily small amount.

3 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 $\mathscr{X}$. 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 Pr 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 $\mathscr{X}(\mathbf{y}, \theta) \subseteq \mathscr{X}$. Having observed $\mathbf{y}$, the probability measure Pr is carried to $2^{\mathscr{X}}$ by the mapping $\theta \to \mathscr{X}(\mathbf{y}, \theta)$, which defines a Dempster-Shafer mass function m on $\mathscr{X}$. 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 $\mathscr{X}_k = \mathscr{X}(\mathbf{y}, \theta_k)$ and masses $m_k = m(\mathscr{X}_k) = \Pr(\{\theta_k\})$ for $k = 1, \ldots, K$.

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

$$L(\boldsymbol{\Psi}) = \sum_{k=1}^{K} m_k \int_{\mathscr{X}_k} g_c(\mathbf{x};\boldsymbol{\Psi}) d\mathbf{x} = \int_{\mathscr{X}} g_c(\mathbf{x};\boldsymbol{\Psi}) \left(\sum_{k=1}^{K} m_k 1_{\mathscr{X}_k}(\mathbf{x}) \right) d\mathbf{x}$$
$$= \int_{\mathscr{X}} g_c(\mathbf{x};\boldsymbol{\Psi}) pl(\mathbf{x}) d\mathbf{x} = \mathbb{E}_{\boldsymbol{\Psi}} [pl(\mathbf{X})], \quad (2)$$

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

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

probability of that fuzzy subset, according to Zadeh's definition of the probability of a fuzzy event [11].

In the more general setting of belief functions, $L(\boldsymbol{\Psi})$ has another interpretation that will now be explained. Let $g_c(\cdot|m;\boldsymbol{\Psi}) = m \oplus g_c(\cdot;\boldsymbol{\Psi})$ denote the p.d.f. obtained by combining m with the complete data p.d.f. $g_c(\cdot;\boldsymbol{\Psi})$ using Dempster's rule [3, 9]:

$$g_c(\mathbf{x}|m;\boldsymbol{\Psi}) = \frac{g_c(\mathbf{x};\boldsymbol{\Psi})pl(\mathbf{x})}{\int_{\mathscr{X}} g_c(\mathbf{u};\boldsymbol{\Psi})pl(\mathbf{u})d\mathbf{u}} = \frac{g_c(\mathbf{x};\boldsymbol{\Psi})pl(\mathbf{x})}{L(\boldsymbol{\Psi})}. \tag{3}$$

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

The expression of the observed data likelihood (2) can often be simplified by making 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. $X_1,\ldots,X_n$:

$$g_c(\mathbf{u};\boldsymbol{\Psi}) = \prod_{i=1}^{n} g_c(u_i;\boldsymbol{\Psi}), \quad \forall \mathbf{u} = (u_1,\ldots,u_n) \in \mathscr{X}.$$

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

$$pl(\mathbf{u}) = \prod_{i=1}^{n} pl_i(u_i), \quad \forall \mathbf{u} = (u_1,\ldots,u_n) \in \mathscr{X},$$

where pl_i is the contour function corresponding to the marginal mass function m_i on x_i.

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 [9].

Under Assumptions A1 and A2, the observed data log likelihood can be written as a sum of n terms:

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

4 The Evidential EM Algorithm

To maximize function $L(\boldsymbol{\Psi})$ defined by (2), we propose to adapt the EM algorithm as follows. Let the E-step now consist in the calculation of the expectation of $\log L_c(\boldsymbol{\Psi})$ with respect to $g_c(\cdot|m;\boldsymbol{\Psi}^{(q)})$ defined by (3):

$$Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) = \mathbb{E}_{\boldsymbol{\Psi}^{(q)}} [\log L_c(\boldsymbol{\Psi}) | m] = \frac{\int \log(L_c(\boldsymbol{\Psi})) g_c(\mathbf{x}; \boldsymbol{\Psi}^{(q)}) pl(\mathbf{x}) d\mathbf{x}}{L(\boldsymbol{\Psi}^{(q)})}. \quad (4)$$

The M-step is unchanged and requires the maximization of $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})$ with respect to $\boldsymbol{\Psi}$. 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 $L(\boldsymbol{\Psi}^{(q)})$ for $q = 0, 1, 2, \ldots$ of likelihood values obtained using the E^{2}M algorithm is non decreasing, i.e., it verifies*

$$L(\boldsymbol{\Psi}^{(q+1)}) \geq L(\boldsymbol{\Psi}^{(q)}), \quad \forall q. \quad (5)$$

Proof. The proof is similar to that of Dempster et al. [4]. □

To conclude this section, we may note that the p.d.f. $g_c(\mathbf{x}|m; \boldsymbol{\Psi})$ and, consequently, the E^{2}M algorithm depend only on the contour function $pl(\mathbf{x})$ and they 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.

5 Normal Mean and Variance Estimation

To illustrate the above algorithm, let us assume that the complete data $\mathbf{x} = (x_1, \ldots, x_n) \in \mathscr{X} = \mathbb{R}^n$ is a realization from an i.i.d. random sample from a univariate normal distribution $\mathscr{N}(\mu, \sigma^2)$. The parameter vector is thus $\boldsymbol{\Psi} = (\mu, \sigma)$. The observed data has the form $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ with $\mathbf{y}_i = (w_i, \alpha_i)$, where w_i is an estimate of x_i (provided, e.g., by a sensor), and $\alpha_i \in [0, 1]$ is a degree of confidence in that estimation. For each $\mathbf{y}_i$, there are two interpretations θ_{i1} and θ_{i2}. Under interpretation θ_{i1}, we admit that $x_i = w_i$; under interpretation θ_{i2}, we know only that $\mathbf{x}_i \in \mathbb{R}$. 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 defined by

$$m_i(\{w_i\}) = \alpha_i, \quad m_i(\mathbb{R}) = 1 - \alpha_i.$$

The corresponding contour function is defined by

$$pl_i(x) = \alpha_i \delta(x - w_i) + 1 - \alpha_i$$

for all $x \in \mathbb{R}$, where $\delta(\cdot)$ is the Dirac Delta function.

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

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

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,

$$Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)}) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\beta_i^{(q)} - 2\mu\sum_{i=1}^{n}\gamma_i^{(q)} + \mu^2\right), \quad (6)$$

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

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

defined by

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

The following equalities thus hold:

$$\gamma_i^{(q)} = \frac{\alpha_i g_c(w_i;\boldsymbol{\Psi}^{(q)}) w_i + (1-\alpha_i)\mu^{(q)}}{\alpha_i g_c(w_i;\boldsymbol{\Psi}^{(q)}) + 1 - \alpha_i} \quad (7)$$

and

$$\beta_i^{(q)} = \frac{\alpha_i g_c(w_i;\boldsymbol{\Psi}^{(q)}) w_i^2 + (1-\alpha_i)\left[\left(\mu^{(q)}\right)^2 + \left(\sigma^{(q)}\right)^2\right]}{\alpha_i g_c(w_i;\boldsymbol{\Psi}^{(q)}) + 1 - \alpha_i}. \quad (8)$$

The maximum of $Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)})$ defined by (6) is obtained for the following values of μ and σ:

$$\mu^{(q+1)} = \frac{1}{n}\sum_{i=1}^{n}\gamma_i^{(q)} \quad (9)$$

and

$$\sigma^{(q+1)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \beta_i^{(q)} - \left(\mu^{(q+1)}\right)^2}. \tag{10}$$

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 (7) and (8), respectively. The M-step then updates the estimates of μ and σ using (9) and (10). The algorithm stops when the relative increase of the observed data likelihood becomes less than some threshold ε.

Example 1. To illustrate the application of the above algorithm to a situation where data are unreliable, we considered the following experiments. Random samples of size $n = 100$ were drawn from a standard normal distribution. For each realization x_i, a number α_i was drawn from the uniform distribution $\mathscr{U}_{[0,1]}$. With probability α_i, w_i was defined as x_i, and with probability $1-\alpha_i$ it was set to $x_i+\varepsilon_i$, with $\varepsilon_i \sim \mathscr{N}(0,s^2)$. Parameters μ and σ were estimated using the E^{2}M algorithm based on the data (w_i,α_i), $i=1,\ldots,n$. The experiment was repeated $N = 100$ times and mean squared errors on μ and σ were computed. The results are shown in Figure 1. Our approach was compared with the simple strategy that consists in estimating μ and σ by the sample mean and standard deviation of the w_i for all i such that $\alpha_i \geq c$, for different choices of c. We can see that the E^{2}M algorithm is much more robust than this simple reference method. Further experiments involving comparisons with more sophisticated alternative estimators are under way.

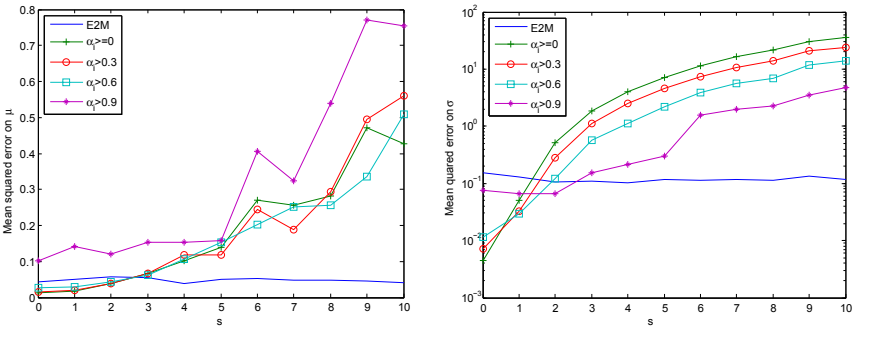

Fig. 1 Mean squared errors on μ (left) and σ (right, logarithmic y scale) as functions of the noise standard deviation s for the E^{2}M algorithm and alternative methods (see details in text).

6 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. Aggarwal, C.C., Yu, P.S.: A survey of uncertain data algorithms and applications. IEEE Trans. Knowl. Data Eng. 21(5), 609–623 (2009)
2. Côme, E., Oukhellou, L., Denœux, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. Pattern Recogn. 42(3), 334–348 (2009)
3. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Statist. 38, 325–339 (1967)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39, 1–38 (1977)
5. Denœux, T., Masson, M.-H., Hébert, P.-A.: Nonparametric rank-based statistics and significance tests for fuzzy data. Fuzzy Sets Syst. 153, 1–28 (2005)
6. Gebhardt, J., Gil, M.A., Kruse, R.: Fuzzy set-theoretic methods in statistics. In: Slowinski, R. (ed.) Fuzzy sets in decision analysis, operations research and statistics, pp. 311–347. Kluwer Academic Publishers, Boston (1998)
7. Jraidi, I., Elouedi, Z.: Belief classification approach based on generalized credal EM. In: Mellouli, K. (ed.) ECSQARU 2007. LNCS (LNAI), vol. 4724, pp. 524–535. Springer, Heidelberg (2007)
8. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. Wiley, New York (1997)
9. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton (1976)
10. Vannoorenberghe, P., Smets, P.: Partially supervised learning by a credal EM approach. In: Godó, L. (ed.) ECSQARU 2005. LNCS (LNAI), vol. 3571, pp. 956–967. Springer, Heidelberg (2005)
11. Zadeh, L.A.: Probability measures of fuzzy events. J. Math. Anal. Appl. 10, 421–427 (1968)