# Inference and learning in evidential discrete latent markov models

Emmanuel Ramasso

# Inference and learning in evidential discrete latent Markov models

Emmanuel Ramasso
FEMTO-ST Institute, UMR CNRS 6174 - UBFC / ENSMM / UTBM,
Applied Mechanics Department, Automatic Control and Micro-Mechatronic Systems Department
25000, Besançon, France

**Abstract**

We present the Evidential Hidden Markov Model (EvHMM), an extension of standard HMM for time-series modelling where conditional belief functions are used in place of probabilities to manage uncertainty on discrete latent variables. Inference and learning mechanisms are described and allow to solve the three problems initially defined for HMM, namely: the classification problem (find the most plausible model), the decoding problem (finding the best sequence of hidden states) and the learning problem based on incomplete and uncertain data (estimate the parameters). Exact inference mechanisms based on the Generalized Bayesian Theorem are proposed which allows one to recover standard HMM when probabilities are considered. An EM-like procedure is developed for parameter learning, relying on some approximations suggested to make the solutions tractable. Relationships are discussed with both the learning criterion conjectured by Vannoorenberghe and Smets and the formulation of Evidential Markov Chains by Pieczynski et al. A comparison with standard HMM on simulated data confirms the interest of considering random disjuctive sets to represent data incompleteness in evidential temporal graphical models.

**Index Terms**

Evidential Hidden Markov Models, Dempster-Shafer theory of Belief functions, Expectation-Maximization, Generalized Bayesian Theorem, evidential temporal graphical models

## NOMENCLATURE

| | |
|---|---|
| ARI | Adjusted Rand Index |
| BBA | Basic Belief Assignment |
| CBF | Conditional Belief Functions |
| DEVN | Directed Evidential Network |
| DRC | Disjunctive Rule of Combination |
| EFB | Evidential Forward-Backward algorithm |
| EMC | Evidential Markov Chain |
| EM | Expectation-Maximization algorithm |
| EvHMM | Evidential HMM |
| EvHMM-CT | EvHMM with conditional form of transition |
| EvHMM-JT | EvHMM with joint form of transition |
| GBT | Generalized Bayesian Theorem |
| GMM | Gaussian Mixture Model |
| HMM | Hidden Markov model |
| TSDEVN | Time-sliced DEVN |
| TWD | Theory of Weighted Distributions |

## I. INTRODUCTION

### A. Latent variable models for time-series

The statistical treatment of temporal multivariate measurements originating from complex dynamical systems is of paramount interest in many application fields as diverse as biological sequences alignment, computer vision and image understanding, speech recognition and synthesis, or diagnostics and prognostics of industrial equipments.

In real-world applications, the sequences of high-dimensional observations are generated from a number of possible *sources* which are generally partially unknown and unobservable. One important challenge is to recognize which source is active at a time (latent event), discover which source leads to particular sensor measurements (causal structure) and determine the relationship between latent source activity and measurements (observation model) [1], [2].

More generally, latent variable models are convenient statistical tools used in machine learning in order to represent unobserved factors. Those variables allow to develop state models which divide the feature space of time-series into meaningful areas that can be used to characterize sources [3]. Estimating the parameters of those models as well as making inference of the latent structure from temporal observations is thus of key importance.

### B. Time series modeling using HMM

Hidden Markov Models (HMM) [4], [5] belong to the family of latent variable models dedicated to time-series analysis. It considers multivariate time-series $\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \dots \mathbf{x}_T]$ of length $T$ in $D$ dimensions with $\mathbf{x}_t = (x_1, \dots, x_D)'$ as being generated by a doubly random process (Figure 1(a)). The first process is a discrete latent variable Markov model which supposes that the system is governed by a Markov chain and can switch back and forth between discrete states through time. The states can be hidden (not observed) [5] or partially hidden [6] according to the amount of prior knowledge available during learning and inference. The states are represented by discrete random variables $z_1, z_2 \dots z_t$ taking values in a finite set $\Omega_z = \{s_1, s_2, \dots, s_K\}$. Starting from a state at $t = 1$, with a probability $\pi_j = p(z_1 = s_j)$ with $\mathbf{\Pi} = [\pi_j], j = 1 \dots K$, the system switches between states from $t$ to $t + 1$ with a probability $a_{jk} = p(z_t = s_k | z_{t-1} = s_j)$ with $\mathbb{A} = [a_{jk}], j = 1 \dots K, k = 1 \dots K$ called the transition matrix. The second process is a mixture model which represents the distribution of the data given each possible state. It practically means that the time-series is represented as consecutive segments, each characterized by a particular distribution over input data. It is supposed that the observations distribution is a mixture of multivariate Gaussians in each state $s_k$ (Fig. 1(b)):

$$p(\mathbf{x}_t | z_t = s_k) = \sum_{j=1}^{M} c_{jk} \mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_{jk}, \mathbf{\Sigma}_{jk}), k = 1 \dots K \tag{1}$$

where $c_{jk}$ is the mixing coefficient for the $j$-th component in the $k$-th state and $\mathbf{c} = [c_{jk}], j = 1 \dots M, k = 1 \dots K$ the set of coefficients for all states and components. In the sequel, the parameters of an HMM are denoted as $\theta = \{\mathbf{\Pi}, \mathbb{A}, \boldsymbol{\mu}, \mathbf{c}, \mathbf{\Sigma}\}$.

There are three main efficient algorithms especially dedicated to HMM that have been proposed over the years [5], accounting for their applicability in many disciplines:

- The first algorithm solves a first problem (P1) that is the computation of the likelihood of a HMM model given the parameters and a sequence of observations. The practical interest of this problem is to select the best HMM model within a library that best fits the observations.
- The second algorithm is focused on a second problem (P2) that is the estimation of the best sequence of hidden states given the parameters and a sequence of observations. It practically aims at finding the hidden structure of the data (for exploration) and is used to determine whether a particular state has been reached (for detection or forecasting purposes).
- Finally, the third problem (P3) is the estimation of the parameters given a sequence of observations solved by an iterative algorithm called Expectation-Maximization (EM) [7].

Problems P1 and P2 rely on efficient inference mechanisms based on the forward-backward algorithm and on dynamic programming, while problem P3 is a learning problem that makes use of the latter mechanisms within EM [5], [8, Chap. 13], [9], [10].

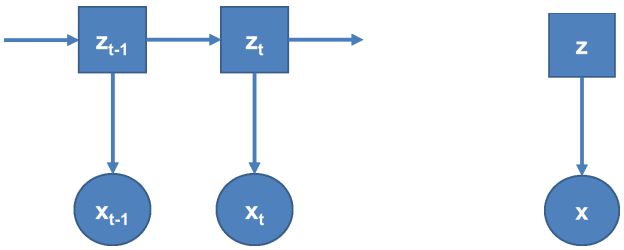



Fig. 1. Graphical representation of hidden Markov (a) and mixture models (b), $\mathbf{z}$ is discrete, $\mathbf{x}$ continuous. [11].

### C. Problem statement and related work

In standard HMM, a latent variable is represented as a $K$-dimensional binary random variable $\mathbf{z}$. It has the particularity that one element $z_k$ is equal to 1 while all other elements are equal to 0 and the values of $z_k$ satisfy $z_k \in \{0, 1\}$ with $\sum_k z_k = 1$. Therefore, there are $K$ possible states for the vector $\mathbf{z}$ according to which element is nonzero. Inference mechanisms defined in HMM aims at estimating the probability that a particular state occurs at $t$ given the observations.

Consider now a situation where the system can gradually evolve through random disjunctive sets (Figure 2). It means that, for a given data vector $\mathbf{x}_t$, and in addition to uncertainty, a doubt may explicitly exist about the membership of this data vector to a state (or a component). The uncertainty on states is encoded by belief functions [12], [13] so that a Basic Belief Assignment (BBA) $m^{\Omega_z}(S)$ at time $t$ represents the amount of probability to be shared among subsets in $S$ made of an union of states and without being assigned to a smaller subset in $S$ by *lack of knowledge*. This is in agreement with the idea of incomplete data [14, Chap. 5]. A BBA is defined as:

$$\begin{array}{lll} m: & 2^{\Omega_z} & \mapsto [0, 1] \\ & S & \to m^{\Omega_z}(S) \text{ s.c. } \sum_S m^{\Omega_z}(S) = 1, m^{\Omega_z}(S) \geq 0 \end{array} \tag{2}$$

For the sake of convenience, we use capital letter (e.g. $S \subseteq \Omega_z$) to indicate that general subsets are considered and small letter (e.g. $s \in \Omega_z$) for singletons. When $m^{\Omega_z}(S) > 0$, $A$ is called a *focal set*, and if $m^{\Omega_z}(\emptyset) = 0$ then the BBA is said *normal* (a particular case).

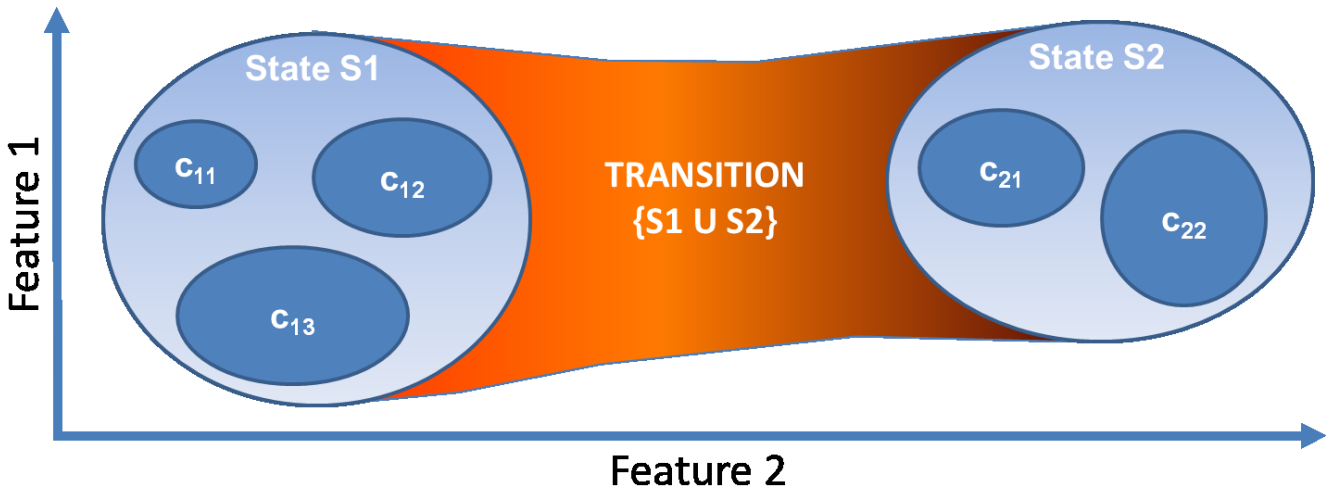



Fig. 2. Flow of belief from singleton state $s_1$ (made of 3 components) to a singleton state $s_2$ (with 2 components) through an imprecise state $s_1 \cup s_2$ which explicitly represents a doubt between both singletons without the possibility to select a particular state due to lack of knowledge.

The question posed in this paper is the following: Can we extend the procedures for inference and parameters learning defined initially for HMM to the case where the Markov chain is neither governed by probabilities but by belief functions? Pros and cons of using such functions in place of probabilities has been discussed in many publications since the advent of Belief Networks [15]–[20].

The latent variable can thus be represented as a $2^K$-dimensional binary random variable $\mathbf{z}_t$ which equals 1 for a particular element and 0 for all others, satisfying $z_{tk} \in \{0, 1\}$ and $\sum_k z_{tk} = 1$. The set of vectors $\mathbf{z}_t$ for $t = 1 \dots T$ are gathered into a matrix $\mathbf{S}$, likewise to the data matrix $\mathbf{X}$. It is important to make a distinction between the $K$ *singleton states* and the other $2^K - K$ values which do *not* represent "states" but *subsets* of states (without considering additivity). Therefore, the knowledge about a subset $A$ of states at time $t$ may "flow" [21] towards subsets $B$ at time $t + 1$ if $A \cap B \neq \emptyset$ (for instance if more information is available through new observations). Given the temporal graphical model (Fig. 1(a)), the flow of belief towards subsets can be managed by the Generalized Bayesian Theorem (GBT) [17, Theorem 4] on one hand, and the Total Plausibility Theorem (TPT) [22] together with the Disjunctive Rule of Combination (DRC) [17, Theorem 3] on the other hand, defined as:

$$\text{GBT: } pl^{\Omega_z}(S \mid \mathbf{x}_t, \theta) = pl^{\Omega_x}(\mathbf{x}_t \mid S, \theta) \tag{3a}$$

$$\text{DRC: } pl^{\Omega_x}(\mathbf{x}_t \mid S, \theta) = 1 - \prod_{s_k \in S} \left(1 - pl^{\Omega_x}(\mathbf{x}_t \mid s_k, \theta)\right) \tag{3b}$$

$$\text{TPT: } f_{1 \textcircled{\scriptsize{$\cap$}} 2}^{\Omega_x}(\mathbf{x}_t) = \sum_{S \subseteq \Omega_z} m_1^{\Omega_z}(S) \cdot f_2^{\Omega_x}(\mathbf{x}_t \mid S) \tag{3c}$$

The GBT here assumes no prior on $\Omega_z$. The DRC allows to compute a plausibility conditionally to subsets given only plausibilities on singletons. In the TPT, $f_2$ represents the causal link between both variables $\mathbf{X}$ and $\mathbf{S}$. Only a BBA can be used to weigh $f_2$ and the result is always of the same type as $f_2$ such as a BBA, a belief function, a plausibility or a commonality. Some of those functions are defined subsequently, the reader may also find details in [17], [21] and all those functions have one-to-one correspondence [12], [13]. For the sequel, the plausibility is of interest and defined as

$$pl^{\Omega_z}(B) = \sum_{C \cap B \neq \emptyset} m^{\Omega_z}(C) \tag{4}$$

in particular

$$pl^{\Omega_z}(\Omega_z) = 1 - m^{\Omega_z}(\emptyset) \tag{5}$$

The latter expression means that the plausibility of "everything" (for instance of the observed data and after evaluating all possible states) is equal to one minus the degree of "nothing". "Nothing" here represents a situation where the model used to quantify the uncertainty on subsets is in *conflict* with the observed data. This quantity has been used to evaluate the quality of

a model in several publications since the advent of the Transferable Belief Model which allows to consider that a mass can be assigned on the empty set [13]. Of particular interest, a connection between conflict (Eq. 5) and the likelihood has been formally established in [23, Sect. 3.2] and suggested in some previous papers of the author [24, Eq. 18]. Shafer [12, Chap. 3, Section 4] already pointed out the potential importance of this measure.

The consideration of belief propagation is a fundamental difference with the approaches developed by Pieczynski et al. [10], [25]–[27] in which $2^K$ sets of parameters are estimated, one for each subset: It conceptually means that the elements in $\mathbf{z}_t$ are considered as different, one to each other. The propagation used in the latter work involves probabilities on singletons using the standard forward-backward algorithm and therefore the uncertainty can not flow through set-intersections operations as considered in the belief functions theory.

### *D. Contribution*

The problem tackled in this paper is thus the development of mechanisms for the inference of hidden states and parameter learning based on imcomplete data in HMM when the uncertainty is no more represented by probabilities but by belief functions [12], [13], [28]. The proposed model is called Evidential Hidden Markov Model (EvHMM) and can be viewed as a "Time-Sliced Directed Evidential Networks" (TSDEVN) [29] that is the counterpart of Dynamic Bayesian Networks based on the Probability Theory [11] extended to belief functions.

In order to solve the problems P1 and P2 in EvHMM, some elements have been proposed by the author [24], [29] which are summarized in the sequel, in particular the extension of the Forward-Backward (FB) algorithm to belief functions (called EFB: Evidential FB) which is of key importance for the learning phase (P3). This algorithm is based and Shenoy's [30] and Smets' [17] works on belief propagation and "Directed Evidential Networks" (DEVN). We consider Conditional Belief Functions (CBF) [17] that allow one to store $|\Omega_x| \times 2^{|\Omega_y|}$ belief functions to express the relationship between both variable $x$ and $y$ (by a conditional distribution of $y$ given $x$) tacking values in $\Omega_x$ and $\Omega_y$ respectively, instead of using joint BF which may be prohibitive in practice when the frames of discernment are large. It is demonstrated that the forward propagation extended to belief functions allows to get the plausibility (equivalent to the likelihood in standard HMM) of an EvHMM. Concerning the learning problem (P3), a new approach is proposed which is discussed with respect to the pioneering work of Vannoorenberghe and Smets [31].

In order to develop the solution to the learning problem in EvHMM, we first start by considering a simpler case: Evidential Markov Chain. We then proceed with the EvHMM and some illustrations.

## II. Markov Chain revisited with conditional belief functions

Consider a system described at time $t$ as being in one state in $\Omega_z$. Its evolution is managed by a first-order probabilistic Markov chain such that $p(s_t \mid s_{t-1}, s_{t-2} \dots s_1) = p(s_t \mid s_{t-1})$. A particular sequence of singleton states $\mathbf{S} = (s_1, s_2, \dots s_t \dots s_T), s_t \in \Omega_z$ has the probability $p_\pi(s_1) \prod_{t=2}^{T} p(s_t \mid s_{t-1})$ where $p_\pi$ is the initial prior. If the chain takes its values in subsets of $\Omega_z$, then what is the total support in terms of plausibility given to a particular sequence?

### *A. Support to a sequence made of subsets*

Suppose that the transition matrix is made of BBAs $m_a^{\Omega_z}(\cdot \mid S_{t-1}), S_{t-1} \subseteq \Omega_z$. A sequence $\mathbf{S} = (S_1, S_2, \dots S_t \dots S_T), S_t \subseteq \Omega_z$ starting at $S_1$ requires to considering that $S_1$ is true at $t = 1$, $S_2$ is true at $t = 2$ and so on. Consider that the initial BBA representing one's belief in the first states is *vacuous*: $m_\pi^{\Omega_z}(\Omega_z) = 1$. At $t = 1$, $m_\pi^{\Omega_z}$ is combined by Dempster's rule [28] with a categorical BBA, i.e. made of one focal set on $S_1$ (with mass equal to 1). Given two BBAs $m_1^{\Omega_z}$ and $m_2^{\Omega_z}$, Dempster's rule is

$$m_{1\textcircled{}2}^{\Omega_z}(A) = \sum_{C \cap D = A} m_1^{\Omega_z}(C) \cdot m_2^{\Omega_z}(D) \tag{6}$$

which becomes a *conditioning rule* if one of the BBAs is categorical. Given the definition of the plausibility (Eq. 4 and Eq. 5), conditioning a mass on a subset $S$ generates a BBA such that the sum of masses on all focal sets except the empty set represents the plausibility of $S$.

Therefore, after conditioning $m_\pi$ on $S_1$, we get a BBA with one focal set equal to $S_1$ since $m_\pi$ is vacuous. Now, if the second state of the chain is $S_2$, then the transfer of mass from $t = 1$ to $t = 2$ is driven by the TPT (Eq. 3c). The BBA resulting from the transfer has to be combined by Dempster's rule with a BBA made of one focal set on $S_2$. The BBA obtained is made of several focal sets such that the sum of masses on all focal sets except the empty set represents the plausibility of $S_2$ given $S_1$ which is simply given by an element of the matrix $pl_a$. The same reasoning can be applied for $t = 3 \dots T$ yielding the following result.

*Proposition 1: The total support assigned to a sequence $\mathbf{S} = (S_1, S_2, \dots S_t \dots S_T), S_t \subseteq \Omega_z$ can be quantified by the plausibility on $\Omega_z^T = \Omega_z \times \Omega_z \times \dots \Omega_z$ ($T$ times) after conditioning on the sequence. Given a vacuous BBA on initial states, the total support is given by:*

$$pl^{\Omega_z^T}(\mathbf{S}) = \prod_{t=2}^{T} pl_a^{\Omega_z}(S_t \mid S_{t-1}) \tag{7}$$

It defines an Evidential Markov Chain (EMC). Note that the solution is different if the prior is not vacuous: In that case, one may use a forward propagation as presented subsequently. The solution is also different from the result proposed in [10, Def. 4.1] which is based on the product of BBAs.

*Definition 1 (proposed in [10]):*

$$m^{\Omega_z^T}(\mathbf{S}) = \prod_{t=2}^{T} m_a^{\Omega_z}(S_t \mid S_{t-1}) \tag{8}$$

Note that for the sake of convenience, in particular for subsequent developments, no prior is considered. This relation means that the BBA especially assigned to $S$ is considered while omitting all other BBAs that can be assigned to subsets of $S$. The support computed with the plausibilities has a larger value than the support with the BBAs and the difference collapses for Bayesian BBA (focal sets are singletons).

### *B. Learning the evidential transition matrix from incomplete data*

Following the notation proposed in [8], let $z_{t,k} = 1$ if the true state is $S_k$ at time $t$, 0 otherwise. Then it follows

$$pl^{\Omega_z^T}(\mathbf{S} \mid \mathbb{A}) = \prod_{t=2}^{T} pl_a^{\Omega_z}(S_t \mid S_{t-1}, \mathbb{A})^{z_{t,j} z_{t-1,i}} \tag{9}$$

where $\mathbb{A}$ allows to emphasize that $pl_a$ is parameterized. It yields

$$\log pl^{\Omega_z^T}(\mathbf{S} \mid \mathbb{A}) = \sum_{t=2}^{T} (z_{t,j} z_{t-1,i}) \log pl_a^{\Omega_z}(S_t \mid S_{t-1}, \mathbb{A}) \tag{10}$$

Taking the expectation of this expression similarly to HMM [8, page 616] (since the states and thus $\mathbf{z}_t$ are unknown) requires to define a BBA $m^{\Omega_z \times \Omega_z}_{\xi(t,t-1)}$ defined on two consecutive slices that represents the probability mass of observing the binary variable $(z_{t,j} z_{t-1,i})$ we get:

$$\begin{aligned} &\mathbb{E}\left[\log pl^{\Omega_z^T}(\mathbf{S} \mid \mathbb{A})\right]_{m_\xi} = \dots \\ &\sum_{t=2}^{T} \sum_{S_j \subseteq \Omega_z} \sum_{S_i \subseteq \Omega_z} m^{\Omega_z \times \Omega_z}_{\xi(t,t-1)}(S_i, S_j) \log pl_a^{\Omega_z}(S_j \mid S_i, \mathbb{A}) \end{aligned} \tag{11}$$

This defines a cross-entropy between a BBA and a plausibility. Cross-entropy maximization is at a basis of EM where, at iteration $(q)$, the step "E" estimates $m_\xi$ given the data using inference procedures while the step "M" finds the parameters that allows the function within the logarithm (here $pl$) to get closer to the target function (here $m_\xi$) that weigh the logarithm. Trying to maximize such a criterion in an EM-like algorithm with respect to the parameters in the plausibility function (for instance the transition $\mathbb{A}$ for a Markov chain) would lead to reestimation formula based on both BBA and plausibilities which seems inconsistent. The constraints are actually expressed on $m_a$ with $\sum_A m_a^{\Omega_z}(A \mid S_{t-1}) = 1, \forall S_{t-1} \subseteq \Omega_z$ and the link with $pl_a$ should be made by a Moebius transform [21].

The approximation of Eq. 7 by Eq. 8 is thus of practical interest yielding reestimation formula based only on BBAs. Consider an EM-like algorithm (described subsequently) where the parameters $\mathbb{A}$ (conditional mass functions in the present case) have to be estimated given the previous estimates $\mathbb{A}^{(q)}$ in an iterative manner. Using the approximation based on BBAs in the expectation, and based on the fact that $\sum_A m(A) \log pl(A) \geq \sum_A m(A) \log m(A)$, the criterion can be modified as follows using BBAs:

$$\mathbb{E}\left[\log pl^{\Omega_z^T}(\mathbf{S} \mid \mathbb{A})\right]_{m_\xi} \geq \mathcal{Q}^a_{m,m}(\mathbb{A}^{(q)}, \mathbb{A}) \tag{12}$$

with

$$\begin{aligned} \mathcal{Q}^a_{m,m}(\mathbb{A}^{(q)}, \mathbb{A}) \quad &= \sum_{t=2}^{T} \sum_{S_j \subseteq \Omega_z} \sum_{S_i \subseteq \Omega_z} \dots \\ &m^{\Omega_z \times \Omega_z}_{\xi(t,t-1)}(S_i, S_j \mid \mathbb{A}^{(q)}) \log m_a^{\Omega_z}(S_j \mid S_i, \mathbb{A}) \end{aligned} \tag{13}$$

The maximization of $\mathcal{Q}^a_{m,m}$ with respect to $m_a$ at iteration $(q)$ requires to take the derivative of $\mathcal{Q}^a_{m,m}$ and using appropriate Lagrange multipliers (ensuring that $\sum_B m_a^{\Omega_z}(B \mid S_{t-1}) = 1, \forall S_{t-1} \subseteq \Omega_z$) yielding:

$$\begin{aligned} m_a^{(q+1)}(S_{j,t} \mid S_{i,t-1}) = \quad &\dots \\ &\frac{\sum_{t=2}^{T} m^{\Omega_z \times \Omega_z}_{\xi(t,t-1)}(S_i, S_j \mid \mathbb{A}^{(q)})}{\sum_{t=2}^{T} \sum_{\emptyset \neq S_l \subseteq \Omega_z} m^{\Omega_z \times \Omega_z}_{\xi(t,t-1)}(S_i, S_l \mid \mathbb{A}^{(q)})} \end{aligned} \tag{14}$$

By assuming that the BBAs defined conditionally to subsets are computed by the DRC based only on BBAs defined conditionally to singletons, it follows that Eq. 14 allows to estimate $|\Omega_z| \times 2^{|\Omega_z|}$ parameters. This procedure supposes a similar conjecture as

in [31] where the authors suggested an EM-like procedure for parameters learning in GMM using belief functions on mixing variables.

*Conjecture 1: Similarly to the auxiliary function in EM [4, Theorem 2.1], the maximization of the lower bound of the support $\mathcal{Q}_{m,m}$ does not decrease the total support.*

It is shown subsequently that it can be practically feasible to check whether the conjecture holds or not when applied on datasets using the evidential forward algorithm. This algorithm allows to estimate the *exact* total support which can not be directly maximized as in standard HMM.

### C. Incorporating evidential prior to adjust the posterior BBA

We can observe that the target BBA $m_{\xi}^{(q)}$ computed in the E-step, which gives information about the hidden variables, is used in the reestimation formula of $m_a$. In cases of model's misspecification (choice of $\mathbb{A}$ for instance) or biases induced by the data collection process, this BBA may eventually lead to bad parameter estimates. This problem is well known in statistics and two solutions can be mentionned.

The first solution was proposed in [31]. It considers that the prior knowledge on hidden variables are encoded by a set of $T$ belief functions. In the case of EMC described above, the prior may be ideally defined on $\Omega_z \times \Omega_z$, denoted as $m^{\Omega_z \times \Omega_z}_{\text{prior}(t,t-1)}, t = 2 \dots T$. It may also be defined on $\Omega_z$ with a BBA $m^{\Omega_z}_{\text{prior}(t)}$. In that case, it has to be extended on $\Omega_z \times \Omega_z$ using the vacuous extension defined as:

$$m^{\Omega_z \uparrow \Omega_z \times \Omega_z}_{\text{prior}(t,t-1)}(C) = \begin{cases} m^{\Omega_z}_{\text{prior}(t)}(B) & \text{if } C = B \times \Omega_z \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Note that if nothing is known about the hidden variables, then $\forall t, m^{\Omega_z \times \Omega_z}_{\text{prior}(t,t-1)}(\Omega_z \times \Omega_z) = 1$. Those priors can then be incorporated into the computation of the mathematical expectation (Eq. 20a) by Dempster's rule (Eq. 6):

$$m_{\xi}^{(q)} \leftarrow m_{\xi}^{(q)} \textcircled{\cap} m_{\text{prior}(t,t-1)} \tag{16}$$

The second solution relies on the Theory of Weighted Distributions (TWD) [32] which allows to incorporate prior knowledge on expectations by means of (positive) weights with the aim to "adjust" the posterior distribution. It has been used in EM in [33].

The next section now considers that the states are hidden and only some measurements are available. A model is thus necessary to build a relationship between those observations and the latent variables, while an EMC is supposed to still govern the latent variables (Figure 1(a)).

## III. Learning parameters in Evidential Hidden Markov Models

EM-based learning in standard probabilistic HMM is based on the maximization of the cross entropy between both a posterior distribution over latent variables (computed using the parameters at the previous iteration) and a joint distribution over observed and latent variables (given unknown parameters). The strong advantage of standard HMM is the possibility to build a joint distribution using only products [5, Section 3.A]. However, it is not the case when considering belief functions. This is illustrated in the application of the TPT (Eq. 3c) for the transfer between two consecutive time-slices where a BBA appears. To expect a tractable solution while maintaining a connection with standard HMM, some approximations are required.

### A. The criterion

The goal is to find a criterion that reflects the quality of a model (such as Fig. 1) given that belief functions are used to quantify uncertainty on discrete latent variables. As explained previously, the quality of a model can be quantified by minimizing the *amount of conflict* (Eq. 5) between the model and the data or equivalently maximizing the total support. Likewise to EMC, it seems relevant to express the plausibility on $\Omega_z^T$ after observing $\mathbf{X}$. More specifically, finding the parameters $\theta^*$ in a latent variable model when uncertainty is managed by belief functions can be turned into the maximization of the potential support assigned to the subset $(\mathbf{x}_1, \Omega_z), (\mathbf{x}_2, \Omega_z) \dots, (\mathbf{x}_T, \Omega_z)$ after observing all data vectors:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ pl^{\mathbb{R}^T \times \Omega_z^T}\left((\mathbf{x}_1, \Omega_z), (\mathbf{x}_2, \Omega_z) \dots (\mathbf{x}_T, \Omega_z) \mid \theta\right) \tag{17}$$

where $\mathbb{R}$ is the domain of $\mathbf{x}_t$. For short, this criterion is rewritten as $\operatorname{argmax}_{\theta} pl(\mathbf{X}, \boldsymbol{\Omega_z} \mid \theta)$. This plausibility can be computed by summing the belief masses assigned to all configurations of the hidden variables $\mathbf{S}$ (Eq. 4):

$$pl(\mathbf{X}, \boldsymbol{\Omega_z} \mid \theta) = \sum_{\mathbf{S} \not\supseteq \emptyset} m(\mathbf{X}, \mathbf{S} \mid \theta) \tag{18}$$

which can be rewritten as

$$pl(\mathbf{X}, \boldsymbol{\Omega_z} \mid \theta) = \sum_{\mathbf{S} \not\supseteq \emptyset} R(\mathbf{S}) \frac{m(\mathbf{X}, \mathbf{S} \mid \theta)}{R(\mathbf{S})} \tag{19}$$

where $R$ is a distribution such that $\sum_A R(A) = 1$. This constraint allows Jensen's inequality to be applied [34, Eq. 5] since the logarithm is a concave function. It leads to a *lower bound* of the logarithm of the quantity of interest:

$$\log pl(\mathbf{X}, \mathbf{\Omega_z} \mid \theta) \geq \mathcal{Q}_{m,m}(\theta^{(q)}, \theta) - H_{m,m}(\theta^{(q)}, \theta^{(q)}) \tag{20a}$$

$$\mathcal{Q}_{m,m}(\theta^{(q)}, \theta) = \sum_{\mathbf{S} \not\ni \emptyset} R(\mathbf{S}, \theta^{(q)}) \log m(\mathbf{X}, \mathbf{S} \mid \theta) \tag{20b}$$

$$H_{m,m}(\theta^{(q)}, \theta^{(q)}) = \sum_{\mathbf{S} \not\ni \emptyset} R(\mathbf{S}, \theta^{(q)}) \log R(\mathbf{S}, \theta^{(q)}) \tag{20c}$$

$$s.t. \sum_{\mathbf{S}} R(\mathbf{S}, \theta^{(q)}) = 1 \tag{20d}$$

where $H_{m,m}$ depends only on previous estimates $\theta^{(q)}$. The criterion expectedly presents similarities with (13). In particular, the introduction of $H_{m,m}$ allows to underline that when the function in the logarithm ideally evolves towards the target $R$ (which can change at each iteration) then $\mathcal{Q}_{m,m} - \mathcal{H}_{m,m} \to 0$.

Since $R(\mathbf{S})$ must sum up to 1 (due to Jensen's inequality), it follows that a rational choice for $R$ is a BBA denoted as $m_\gamma$ subsequently. By considering only the part dependent on the parameters $\theta$, the criterion $\mathcal{Q}_{m,m} = \mathbb{E}_{m_\gamma}[\log m(\mathbf{X}, \mathbf{S} \mid \theta)]$ is thus an expectation taken with respect to $m_\gamma$. Moreover, as presented in Section II-C, evidential prior can be incorporated either by applying Dempster's rule with $R$ or using the TWD.

An EM-like procedure can thus be applied. At iteration $q$, the E-step aims at maximizing the expectation 20b given fixed parameters $\theta^{(q)}$. We can cancel its derivative with repect to $R$ using appropriate Lagrangian multipliers (integrating the aforementionned constraint on $R$) to get the maximizer $m_\gamma^{(q)}$:

$$\text{E-step: } \Rightarrow m_\gamma^{(q)} = \frac{m(\mathbf{X}, \mathbf{S} \mid \theta^{(q)})}{\sum_{\mathbf{S}' \not\ni \emptyset} m(\mathbf{X}, \mathbf{S}' \mid \theta^{(q)})} \equiv m(\mathbf{S} \mid \mathbf{X}, \theta^{(q)}) \tag{21}$$

The denominator of this expression is the sum of the masses on all subsets $\mathbf{S}$ which is equal to the opposite of the belief mass on the emptyset. $m_\gamma^{(q)}(\cdot \mid \mathbf{X})$ is the posterior BBA on states given observations. The posterior is then used in the M-step to find the best estimate $\theta^{(q+1)}$ for the next iteration so that it maximizes the expectation under $m_\gamma^{(q)}$:

$$\text{M-step: } \Rightarrow \theta^{(q+1)} = \underset{\theta}{\operatorname{argmax}}\ \mathbb{E}_{m_\gamma^{(q)}}[\log m(\mathbf{X}, \mathbf{S} \mid \theta)] \tag{22}$$

The algorithm iterates likewise to standard EM until the relative increase of the support $pl(\mathbf{X}, \mathbf{\Omega_z})$ between two consecutive iterations remains below a threshold.

*Property 1: Since $R$ is a BBA, then Jensen's inequality holds so that this algorithm is guaranteed to converge.*

*Proof 1: Let $\mathcal{Q}_{m,m}(\theta^{(q)}, \theta) = \sum_{\mathbf{S} \not\ni \emptyset} R(\mathbf{S}, \theta^{(q)}) \log m(\mathbf{X}, \mathbf{S} \mid \theta)$, $\mathcal{Q}_{m,m}(\theta^{(q)}, \theta^{(q)}) = \sum_{\mathbf{S} \not\ni \emptyset} R(\mathbf{S}, \theta^{(q)}) \log m(\mathbf{X}, \mathbf{S} \mid \theta^{(q)})$, $H_{m,m}(\theta^{(q)}, \theta) = -\sum_{\mathbf{S} \not\ni \emptyset} R(\mathbf{S}, \theta^{(q)}) \log R(\mathbf{S}, \theta)$ and $H_{m,m}(\theta^{(q)}, \theta^{(q)}) = -\sum_{\mathbf{S} \not\ni \emptyset} R(\mathbf{S}, \theta^{(q)}) \log R(\mathbf{S}, \theta^{(q)})$ defining the criterion at the current ($\theta$) and the previous ($\theta^{(q)}$) iteration. Then $\mathcal{Q}_{m,m}(\theta^{(q)}, \theta) - \mathcal{Q}_{m,m}(\theta^{(q)}, \theta^{(q)}) \geq 0$ due to the maximization step, and $H_{m,m}(\theta^{(q)}, \theta) - H_{m,m}(\theta^{(q)}, \theta^{(q)}) \leq 0$ since $\sum_{\mathbf{S} \not\ni \emptyset} R(\mathbf{S}, \theta^{(q)}) \log \frac{R(\mathbf{S}, \theta)}{R(\mathbf{S}, \theta^{(q)})} \leq \log \sum_{\mathbf{S} \not\ni \emptyset} R(\mathbf{S}, \theta^{(q)}) \frac{R(\mathbf{S}, \theta)}{R(\mathbf{S}, \theta^{(q)})} \equiv 0$, where the inequality comes from Jensen's inequality [34] and assuming that $R$ is normalized. Therefore, by defining the criterion as $\mathcal{Q}_{m,m} - \mathcal{H}_{m,m}$ as proposed above, it follows that the difference of the criterion between two iterations $\left[\mathcal{Q}_{m,m}(\theta^{(q)}, \theta) - \mathcal{H}_{m,m}(\theta^{(q)}, \theta)\right] - \left[\mathcal{Q}_{m,m}(\theta^{(q)}, \theta^{(q)}) - \mathcal{H}_{m,m}(\theta^{(q)}, \theta^{(q)})\right]$ is positive (or equal to zero) which completes the proof.*

The problem with Eq. 22 is that the joint BBA in the logarithm can not be expressed using only products which makes the M-step untractable.

*Assumption 1: It is possible to decouple both the maximization of $Q^a$ (Eq. 13) concerning the Markov chain and the maximization of the criterion $\mathcal{Q}^b_{m,m}$ related to observations given latent variables.*

This decoupling appears *naturally* in standard HMM due to factorisation [8, Chap. 13]. Therefore, the whole criterion becomes $\mathcal{Q}_{m,m} = \mathcal{Q}^a_{m,m} + \mathcal{Q}^b_{m,m}$ and the reestimation formula obtained in Eq. 14 can be applied for EvHMM at each iteration of EM. It thus remains to find the parameters of the observation model, supposed to be a GMM (Eq. 1).

### *B. M step*

In [31], the authors suggested an approach (EM-like) to estimate the parameters in a GMM using belief functions to represent uncertainty on mixing (discrete latent) variables. The criterion $\mathcal{Q}_{m,pl}$ used the plausibility in the logarithm in place of the BBA:

$$\mathcal{Q}_{m,pl}(\Phi^{(q)}, \Phi) = \sum_{\mathbf{S}} m_\gamma^{\Omega_z}(\mathbf{S} \mid \mathbf{X}, \Phi^{(q)}) \log pl(\mathbf{X}, \mathbf{S} \mid \Phi) \tag{23}$$

with $\Phi = \{\mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. This criterion relies on the same conjecture as suggested above (its maximization does not decrease the support to the model). Note that the expression of the support was not explicitly provided by the authors [31]. A proposition

of reestimation formulas can be found in [35] and are based on approximations using the pignistic transform [13] that allows to get a probability distribution from a BBA.

As for EMC, the use of plausibilities creates inconsistency with the BBAs for reestimation formulas and therefore, a criterion such as Eq. 20b is more appropriate (using only BBAs). Considering only the observation model, we aim at maximizing the support $\log pl^{\mathbb{R}^T}(\mathbf{X})$ approximated by:

$$Q^b_{m,m}(\theta^{(q)},\theta) = \sum_{t=1}^{T}\sum_{S\subseteq\Omega_z}\dots \\ m^{\Omega_z}_{\gamma,t}(S \mid \mathbf{X}, \mathbb{A}^{(q)}, \Phi^{(q)}) \log m^{\Omega_z}_b(S \nmid \mathbf{x}_t, \Phi) \tag{24}$$

where it is important to remark that $m_\gamma$ is made dependent not only on the current parameters of the observation model ($\Phi^{(q)}$) but also on the EMC ($\mathbb{A}^{(q)}$). Indeed, the Evidential FB algorithm is shown subsequently to be able to compute this quantity, which is related to Eq. 14 by a marginal operation likewise to standard HMM. The GBT allows to deduce the BBA $m^{\Omega_z}_b(S \nmid \mathbf{x}_t, \Phi)$ given plausibilities conditional to singleton $pl^{\Omega_z}(\mathbf{x}_t \nmid S_t, \Phi)$ [36]:

$$m^{\Omega_z}_b(S \nmid \mathbf{x}_t, \Phi) = \prod_{s_k\in S} pl^{\mathbb{R}}(\mathbf{x}_t \nmid s_k, \Phi)\dots \\ \prod_{s_k\notin S}\left(1 - pl^{\mathbb{R}}\left(\mathbf{x}_t \nmid s_k, \Phi\right)\right) \tag{25}$$

where $pl^{\mathbb{R}}(\mathbf{x}_t \nmid s_k, \theta), \forall s_k \in \Omega_z$ is given by 1.

*Remark 1: The link between plausibility defined conditionally to hypotheses and likelihood has been discussed in several papers, see for instance [17], [36], [37].*

Making use of Eq. 25, the criterion $\mathcal{Q}^b_{m,m}$ can be rewritten as:

$$\begin{aligned}\mathcal{Q}^b_{m,m}(\theta^{(q)},\theta) &= \sum_{t=1}^{T}\sum_{S_k\subseteq\Omega_z} m^{\Omega_z}_{\gamma,t}(S_k \mid \theta^{(q)})\Big\{\dots \\ &\sum_{s_l\in S_k}\log pl(\mathbf{x}_t \nmid s_l,\theta) + \sum_{s_{l'}\notin S_k}\log\left(1 - pl(\mathbf{x}_t \nmid s_{l'},\theta)\right)\Big\} \\ &= \sum_{t=1}^{T}\sum_{s_l\in\Omega_z}\Big\{pl^{\Omega_z}_{\gamma,t}(s_l \mid \theta^{(q)})\log pl(\mathbf{x}_t \nmid s_l,\theta) + \dots \\ &bel_{\gamma,t}(\overline{s_l} \mid \theta^{(q)})\log\left(1 - pl(\mathbf{x}_t \nmid s_l,\theta)\right)\Big\}\end{aligned} \tag{26}$$

where $bel_{\gamma,t}(A) = \sum_{\emptyset\neq B\subseteq A} m_{\gamma,t}(B)$ is the belief function. We can see that the logarithm in the right-hand side of Eq. 26 can make the optimization untractable.

*Assumption 2: The contribution of "$bel_{\gamma,t}(\overline{s_l} \mid \theta^{(q)})\log\left(1 - pl(\mathbf{x}_t \nmid s_l,\theta)\right)$" is negligible compared to the left-hand side expression $pl^{\Omega_z}_{\gamma,t}(s_l \mid \theta^{(q)})\log pl(\mathbf{x}_t \nmid s_l,\theta)$.*

This assumption does not narrow the expression down to a probabilistic formulation because the weight $pl^{\Omega_z}_{\gamma,t}(s_l \mid \theta^{(q)})$ makes use of the information held by all subsets that contain $s_l$.

It means that the weight $bel_{\gamma,t}(\overline{s_l} \mid \theta^{(q)})$ is low for data points located close to $\mu_l$ whereas, conversely, $pl^{\Omega_z}_{\gamma,t}(s_l \mid \theta^{(q)})$ is high. Since for any belief function and associated plausibility we have $pl(A) = bel(\Omega_z) - bel(\overline{A})$, it follows that if $pl_\gamma(s_l)$ is high and assuming that $R \equiv m_\gamma$ is normalized (M-step, Proof 1), then $bel(\overline{s_l})$ is necessarily low. Now, if $pl_\gamma(s_l)$ is low, then $bel(\overline{s_l})$ becomes high and thus the contribution of states potentially far from $\mu_l$ may contribute to $s_l$ which may justify the approximation.

*Example 1: Figure 3 depicts the simplification (left-hand side) and illustrates it on a real case (right-hand side) with an 1D data set made of two normal distributions centered on $\mu_1 = 2$ and $\mu_2 = 4$ with standard deviations $\sigma_1 = 2$ and $\sigma_2 = 1$.*

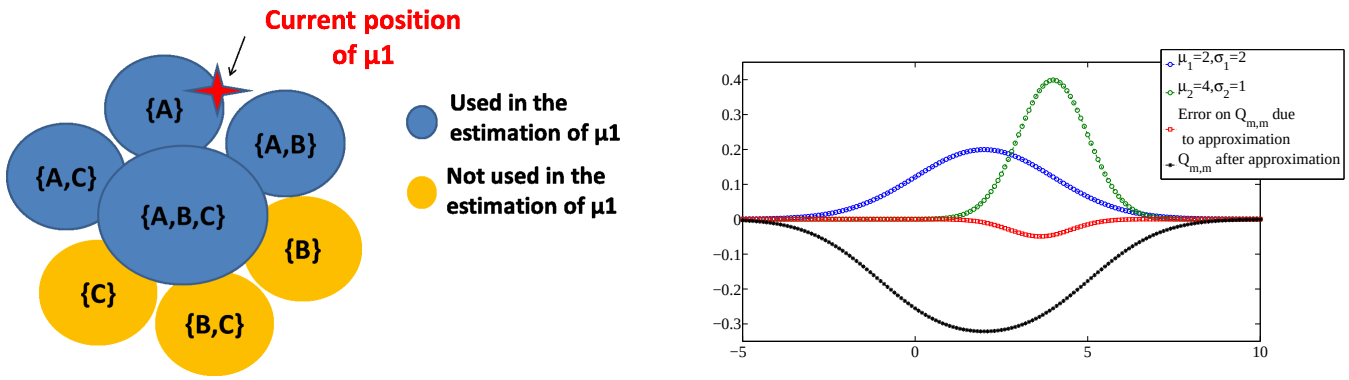
For illustration purpose, we consider one Gaussian component for each singleton state. The criterion can thus be approximated as:

$$\mathcal{Q}^b_{m,m}(\theta^{(q)},\theta) \approx \sum_{t=1}^{T}\sum_{s_l\in\Omega_z} pl^{\Omega_z}_{\gamma,t}(s_l \mid \theta^{(q)})\log pl(\mathbf{x}_t \mid s_l,\theta) \tag{27}$$

Therefore, the means $\mu_k, k = 1\dots K$ for the next iteration are obtained by taking the derivative of Eq. 27 with respect to $\mu_j$:

$$\frac{\partial\mathcal{Q}^b_{m,m}}{\partial\mu_j} = \sum_t pl^{(q)}_{\gamma,t}(s_j)\cdot\Sigma_j^{-1}\left(\mathbf{x}_t - \mu_j\right) \equiv 0 \tag{28a}$$

$$\Rightarrow \mu_j^{(q+1)} = \frac{\sum_t pl^{(q)}_{\gamma,t}(s_j)\,\mathbf{x}_t}{\sum_t pl^{(q)}_{\gamma,t}(s_j)} \tag{28b}$$




(a) Concept behind the approximation

(b) Impact of the approximation

Fig. 3. Impact of the approximation on the estimation of the mean. Left: The use of the plausibilities allows to take into account the points within several subsets to estimate the mean attached to singleton $A$. The information brought by the other subsets (with empty intersection with $A$) are supposed to be negligible. Right: Plot of $\mathcal{Q}^b_{m,m}$ before summing over $t$ in Eq. 27 against the neglected term in Eq. 26.

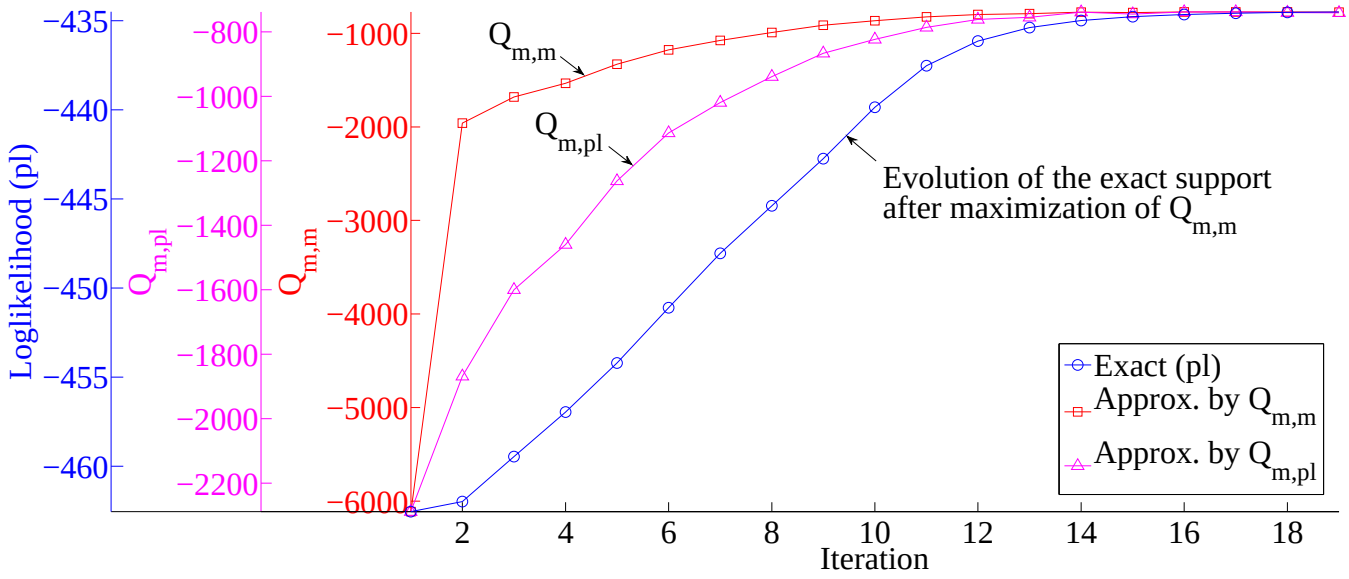



Fig. 4. Evolution of $Q_{m,m}$, $Q_{m,pl}$ and the exact plausibility of the model.

.

The covariances can be obtained with a similar approach as explained in [38] yielding:

$$\Sigma_j^{(q+1)} = \frac{\sum_t pl_{\gamma,t}^{(q)}(s_j) \cdot (\mathbf{x}_t - \mu_j)\,(\mathbf{x}_t - \mu_j)'}{\sum_t pl_{\gamma,t}^{(q)}(s_j)} \tag{29}$$

and the weight of the $j$-th component is given by:

$$\pi_j^{(q+1)} = \frac{\sum_t pl_{\gamma,t}^{(q)}(s_j)}{\sum_t \sum_l pl_{\gamma,t}^{(q)}(s_l)} \tag{30}$$

*Example 2: Figure 4 represents the evolution of $Q_{m,m} = \mathcal{Q}^a_{m,m} + \mathcal{Q}^b_{m,m}$ (proposed criterion), $Q_{m,pl}$ [31] and the exact support (likelihood) of the model computed by Proposition 2 (Eq. 43) for an EvHMM applied to a simulated dataset described in Section V-A. This figure shows that the assumption holds for this dataset and that both approximations are rather correlated to the exact support to the model (Eq. 18).*

### C. E-step

$m_{\gamma,t}$ represents the knowledge on subsets of states after observing $\mathbf{X}$ which is obtained by the evidential forward-backward algorithm [24]. This algorithm can be written using commonality functions which allows point-wise multiplication. The notation $\mathbf{x}_{1:t}$ means that observations from $t=1$ to $t$ were taken into account $(\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_t)$, and $\mathbf{X} \equiv \mathbf{x}_{1:T}$. The forward-backward passes are given by the following sequence of operations. First the forward initialisation:

$$q_{\alpha,1}^{\Omega_z}(S_j, \mathbf{x}_1) = q_{\pi}^{\Omega_z}(S_j) \cdot q_{b,1}^{\Omega_z}(\mathbf{x}_1 \nmid S_j) \tag{31}$$

the forward induction:

$$q_{\alpha,t}^{\Omega_z}(S_j, \mathbf{x}_{1:t}) = \hat{q}_{\alpha,t}^{\Omega_z}(S_j, \mathbf{x}_{1:t-1}) \cdot q_{b,t}^{\Omega_z}(\mathbf{x}_t \nmid S_j) \tag{32}$$

with $\hat{q}_{\alpha,t}^{\Omega_z}(S_j, \mathbf{x}_{1:t-1})$ given by

$$\sum_{S_i \subseteq \Omega_z} m_{\alpha,t-1}^{\Omega_z}(S_i, \mathbf{x}_{1:t-1}) \cdot \overrightarrow{q}_{a,t|t-1}^{\Omega_z}(S_j \mid S_i) \tag{33}$$

and

$$\overrightarrow{pl}_{a,t|t-1}^{\Omega_z}(S_j \mid S_i) = 1 - \prod_{s \in S_i} (1 - pl_{a,t|t-1}^{\Omega_z}(S_j \mid s)) \tag{34}$$

The commonalities $\overrightarrow{q}_a$ are computed from $\overrightarrow{pl}_a$ by a fast Moebius transform [21]. Similarly, the backward initialisation is given by

$$q_{\beta,T}^{\Omega_z} = 1 \tag{35}$$

representing full ignorance, and followed by the backward induction

$$\begin{aligned} q_{\beta,t}^{\Omega_z}(\mathbf{x}_{t+1:T} \mid S_i) \quad &= \dots \\ &\sum_{S_j \subseteq \Omega_z} m_{\beta \textcircled{\cap} b,t+1}^{\Omega_z}(\mathbf{x}_{t+1:T} \nmid S_j) \cdot \overleftarrow{q}_{a,t|t+1}^{\Omega_z}(S_i \mid S_j) \end{aligned} \tag{36}$$

with $q_{\beta \textcircled{\cap} b,t+1}^{\Omega_z}(\mathbf{x}_{t+1:T} \nmid S_j)$ given by

$$q_{\beta,t+1}^{\Omega_z}(\mathbf{x}_{t+2:T} \nmid S_j) \cdot q_{b,t+1}^{\Omega_z}(\mathbf{x}_{t+1} \nmid S_j) \tag{37}$$

The commonalities $\overleftarrow{q}_a$ are computed from $pl_a$ by the GBT:

$$\overleftarrow{q}_{a,t|t+1}^{\Omega_z}(S_i \mid S_j) = \prod_{s \in S_j} pl_{a,t+1|t}^{\Omega_z}(S_j \mid s) \tag{38}$$

The posterior commonalities on states are then given by:

$$q_{\gamma,t}^{\Omega_z}(S_j \mid \mathbf{x}_{1:T}) = q_{\alpha,t}^{\Omega_z}(S_j, \mathbf{x}_{1:t}) \cdot q_{\beta,t}^{\Omega_z}(\mathbf{x}_{t+1:T} \mid S_j) \tag{39}$$

$m_{\xi(t-1,t)}^{\Omega_z \times \Omega_z}$ represents the knowledge on subsets of states at $t$ and $t-1$ after observing $\mathbf{X}$ obtained by the forward-backward algorithm [24]:

$$m_{\alpha,t-1}^{\uparrow \Omega_z \times \Omega_z} \textcircled{\cap} m_{\beta,t}^{\uparrow \Omega_z \times \Omega_z} \textcircled{\cap} m_{a,t,t-1}^{\Uparrow \Omega_z \times \Omega_z} \textcircled{\cap} m_{b,t}^{\uparrow \Omega_z \times \Omega_z} \tag{40}$$

where $\Uparrow$ is the ballooning extension [17] with:

$$\begin{aligned} &m_{a,t,t-1}^{\Uparrow \Omega_z \times \Omega_z}(B) = \prod_{s \in \Omega_z} m_{a,t|t-1}^{\Omega_z}(A \mid s_{t-1}), B \subseteq \Omega_z \times \Omega_z \\ &A = ((s_{t-1} \times \Omega_z) \cap B)) \downarrow \Omega_z \end{aligned} \tag{41}$$

and where $\downarrow$ is a marginalization defined as:

$$\begin{aligned} m^{\Theta \times \Omega \downarrow \Omega}(B) \quad &= \quad \sum_{\substack{C \subseteq \Theta \times \Omega \\ \text{Proj}(C \downarrow \Omega) = B}} m^{\Theta \times \Omega}(C), \quad \forall B \subseteq \Omega \end{aligned} \tag{42}$$

with $\text{Proj}(C \downarrow \Omega)$ the projection of $C$ onto $\Omega$. Symbol $\uparrow$ represents a vacuous extension defined in Eq. 15.

One must make a distinction between the transitions $\overrightarrow{q}_a$ and $\overleftarrow{q}_a$ used in the forward pass and in the backward pass respectively.

*Property 2: Conversely to probabilistic HMM, the transition matrix can not be transposed blindly, unless the following conditions are satisfied:*

- *Condition 1: The elements of the transition matrix represent plausibilities:* $pl_{a,t|t-1}^{\Omega_z}(S_j \mid S_i), \forall S_j \subseteq \Omega_z, \forall S_i \subseteq \Omega_z$ *;*
- *Condition 2: The plausibilities defined conditionally to a subset, i.e.* $pl_{a,t|t-1}^{\Omega_z}(\cdot \mid S_i), S_i \subseteq \Omega_z, s.c.\ |S_i| > 1$*, can be computed from the plausibilities defined on singletons, i.e. from* $pl_{a,t|t-1}^{\Omega_z}(\cdot \mid s_i), s_i \in \Omega_z, s.c.\ |s_i| = 1$*, with the DRC (Eq. 3b).*

The origin of these two conditions comes from the use of the GBT [17]. If these conditions are not satisfied, one must work on the joint space $\Omega_z \times \Omega_z$ to perform the forward and the backward passes which may be practically prohibitive.

An important consequence of these conditions concerns the expression of the reestimation formula for transitions found by Eq. 14 which directly provides the BBA defined conditionally to subsets. For a coherence with the GBT, we advise to keep the BBA defined conditionally to singletons and to apply the DRC (Eq. 3b) to get the ones defined on subsets. In addition, this allows one to decrease memory consumption by storing $|\Omega_z| \times 2^{\Omega_z}$ elements of transition instead of $2^{\Omega_z} \times 2^{\Omega_z}$. The main disadvantage is that the transitions found in this way may partly maximize the criterion since only the transitions defined conditionally to singletons are ensured to maximize Eq. 14.

The "three problems" solved by standard HMM and defined by Rabiner [5] can now be solved for EvHMM as follows.

## IV. Solving the "three problems"

### A. Problem 1 – The classification problem, or how to estimate the likelihood of an EvHMM given some observed data $\mathbf{X}$?

*Proposition 2: The log-likelihood* $\log pl(\theta; \mathbf{X})$ *of an EvHMM model specified by a set of parameters* $\theta$ *and given an observation sequence* $\mathbf{X} \equiv \mathbf{x}_{1:T}$ *is given by:*

$$\log pl(\theta; \mathbf{X}) \equiv \log \Big(1 - m_{\alpha,T}^{\Omega_z}(\emptyset, \mathbf{x}_{1:T} \mid \theta)\Big) \tag{43a}$$

$$\equiv \log pl_{\alpha,T}^{\Omega_z}(\Omega_z, \mathbf{x}_{1:T} \mid \theta) \tag{43b}$$

$$= \sum_{t=1}^{T} \log \Big(1 - m_{\alpha,t}^{\Omega_z}(\emptyset, \mathbf{x}_{1:t} \mid \theta)\Big) \tag{43c}$$

$$= \sum_{t=1}^{T} \log pl_{\alpha,t}^{\Omega_z}(\Omega_z, \mathbf{x}_{1:t} \mid \theta) \tag{43d}$$

*which means that, on the basis of the observations* $\mathbf{X}$*, the total degree of support given to a model is given by summing the opposite of the amount of conflict between the model and the data computed at each time step of the forward propagation.*

*Proof 2:* We start by rewritting the forward pass using the BBA:

$$m_{\alpha,t}^{\Omega_z,*}(S_j, X_{1:t}) = c_t \cdot \Big(\hat{m}_{\alpha,t}^{\Omega_z} \textcircled{\cap} m_{b,t}^{\Omega_z}\Big)(S_j, X_{1:t}) \tag{44}$$

where $m_{\alpha,t}^{\Omega_z,*}$ is a normalized BBA such that $m_{\alpha,t}^{\Omega_z,*}(\emptyset, X_{1:t}) = 0$, $\sum_{\emptyset \neq S_j \subseteq \Omega_z} m_{\alpha,t}^{\Omega_z,*}(S_j, X_{1:t}) = 1$ and is given by:

$$m_{\alpha,t}^{\Omega_z,*}(S_j, X_{1:t}) = c_t \cdot m_{\alpha,t}^{\Omega_z}(S_j, X_{1:t}) \tag{45}$$

with

$$c_t^{-1} = 1 - m_{\alpha,t}^{\Omega_z}(\emptyset, X_{1:t}) = pl_{\alpha,t}^{\Omega_z}(\Omega_z, X_{1:t}) \tag{46}$$

If the normalization is applied from $t = 1$, then $\forall t > 1$, and especially at $T$, the prediction phase can be rewritten as:

$$\begin{aligned} \hat{m}_{\alpha,T}^{\Omega_z}(S_j, \mathbf{x}_{1:T-1}) \quad &= \prod_{t=1}^{T-1} c_t \times \dots \\ \sum_{\emptyset \neq S_i \subseteq \Omega_z} & m_{\alpha,T-1}^{\Omega_z}(S_i, \mathbf{x}_{1:T-1}) \cdot m_{a,T|T-1}^{\Omega_z}(S_j \mid S_i) \end{aligned} \tag{47}$$

leading to

$$m_{\alpha,T}^{\Omega_z,*}(S_j, \mathbf{x}_{1:T}) = \prod_{t=1}^{T} c_t \cdot m_{\alpha,T}^{\Omega_z}(S_j, \mathbf{x}_{1:T}) \tag{48}$$

so that:

$$1 = \sum_{\emptyset \neq S_j \subseteq \Omega_z} m_{\alpha,T}^{\Omega_z,*}(S_j, \mathbf{x}_{1:T}) \tag{49}$$

$$= \prod_{t=1}^{T} c_t \cdot \sum_{\emptyset \neq S_j \subseteq \Omega_z} m_{\alpha,T}^{\Omega_z}(S_j, \mathbf{x}_{1:T}) \tag{50}$$

yielding

$$1 - m_{\alpha,T}^{\Omega_z}(\emptyset, \mathbf{x}_{1:T}) = pl_{\alpha,T}^{\Omega_z}(\Omega_z, \mathbf{x}_{1:T}) = \frac{1}{\prod_{t=1}^{T} c_t} \tag{51}$$

Applying the logarithm on both sides completes the proof.

Algorithm 1 summarizes the steps to follow to get the evidential likelihood that can be used to compare different models.

**Algorithm 1** Algorithm EvHMM Classification

**Require:** model $\theta = \{q_a, \psi, \Pi\}$
**Ensure:** Evidential likelihood $\mathcal{L}_e$
**Ensure:** Evidential forward variable $m_\gamma$
1: Prepare transitions $\overrightarrow{q}_a$ with Eq. 34
2: Apply forward propagation with Eq. 31-34 to get $m_{\alpha,t}$, $t = 1 \dots T$
3: Normalize it at each time step as in Eq. 46
4: Compute the likelihood $\mathcal{L}_e$ by Eq. 43

### *B. Problem 2 – The decoding problem, or how to find the best sequence of hidden states given some observed data* $\mathbf{X}$ *and an EvHMM?*

The problem is to find the sequence of singleton states within a set of observations $\mathbf{X}$ with the highest degree of support to an EvHMM model specified by its parameters $\theta$ or, equivalently, the sequence which presents the minimum amount of conflict with the model.

The important point is to observe that only singleton states are considered. Moreover, it is known that a conjunctive combination between a BBA $m_1$ together with a categorical BBA $m_2$ focused on a singleton state, i.e. $m_2(s_j) = 1, s_j \in \Omega_z$, results in a BBA with one focal set for which the belief mass is equal to the plausibility of the singleton state, i.e. to $pl_1(s_j)$, and the remaining of the mass is assigned to the empty set.

Let's now consider a sequence $S = \{s_1, s_2 \dots s_T\}$. What is its degree of support to the model specified by $\theta$? This sequence can be represented by a set of $T$ categorical BBAs $m_t^{\Omega_z}$ with only one focal sets such that $m_t^{\Omega_z}(s_t) = 1$. Those prior BBAs can be combined conjunctively with the posterior BBA $m_{\gamma,t}$ yielding a first estimate of the best state at $t$ in terms of a plausibility function on singletons:

$$s_t^* = \underset{s_j \in \Omega_z}{\operatorname{argmax}}\ pl_{\gamma,t}^{\Omega_z}(s_j) \tag{52}$$

As for the probabilistic HMM [5], this solution may be sufficient in some cases but it does not take the occurrence of sequences of states into account.

Another solution is to find the sequence $S = \{s_1, s_2 \dots s_T\}$ such that

$$S : \max_{s_1, s_2 \dots s_T} pl_{\delta,T}^{\Omega_z}\left((s_1, \mathbf{x}_1), (s_2, \mathbf{x}_2) \dots (s_T, \mathbf{x}_T) \mid \theta\right) \tag{53}$$

*Proposition 3: The best sequence of singleton states (with the highest degree of support or the minimum amount of conflict) is given by the Viterbi algorithm defined as for probabilistic HMM but using plausibilities on singletons ($\theta$ is implicit):*

- *Viterbi initialisation:*

$$pl_{\delta,1}^{\Omega_z}(s_j, \mathbf{x}_1) = pl_{\pi}^{\Omega_z}(s_j) \cdot pl_{b,1}^{\Omega_z}(\mathbf{x}_1 \mid s_j) \tag{54}$$

- *Viterbi induction, $\forall t = 2 \dots T$:*

$$\begin{aligned} pl_{\delta,t}^{\Omega_z}(s_j, \mathbf{x}_{1:t}) \quad &= pl_b(\mathbf{x}_t \mid s_j) \times \dots \\ &\max_{s_i \in \Omega_z} \left[ pl_{\delta,t-1}^{\Omega_z}(s_j, \mathbf{x}_{1:t-1}) pl_a(s_j | s_i) \right] \end{aligned} \tag{55}$$

$$\begin{aligned} &\psi(i,t) = \dots \\ &\operatorname{argmax}_{\ s_i \in \Omega_z} \left[ pl_{\delta,t-1}^{\Omega_z}(s_j, \mathbf{x}_{1:t-1}) \cdot pl_a(s_j | s_i) \right] \end{aligned} \tag{56}$$

- *Viterbi backtracking, initialized with*

$$s_T^* = \underset{s_j \in \Omega_z}{\operatorname{argmax}}\ pl_{\delta,T}^{\Omega_z}(s_j, \mathbf{x}_{1:T}) \tag{57}$$

*and then $\forall t = T-1, T-2 \dots 2, 1$:*

$$s_{t+1}^* = \underset{s_j}{\operatorname{argmax}}\ \psi(s_{t+1}^*, t+1) \tag{58}$$

*Proof 3:* One can observe that Prop. 3 reduces to a similar expression as for the standard HMM except that probabilities are replaced by plausibilities on singletons. Therefore, we can proceed as for the classical Viterbi algorithm with those plausibilities.

### *C. Problem 3 – The learning problem, or how to estimate the parameters given observed data* $\mathbf{X}$*?*

The solution has been tackled in details in the previous paragraphs and the algorithm 2 summarizes the steps. Two simplifications are proposed.

*a) Concerning the estimation of transitions:* The expression (14) makes use of Eq. 40 which can be difficult to compute due to memory and time consumption since it involves both a ballooning extension, which potentially creates many subsets on $\Omega_z \times \Omega_z$ with high cardinality, and several conjunctive rules between BBAs defined on a joint space. We suggest an approximation by making use of the fact that $m_{\gamma,t}$ is the marginal of $m_\xi$ on $\Omega_z$ at $t-1$. We thus propose to simplify the estimation of transitions by approximating the joint BBA $m_\xi$ by the conjunctive combination with vacuous extension of $m_{\gamma,t}$ and $m_{\gamma,t+1}$:

$$m^{(q+1)}_{\xi(t-1,t)} \approx m^{\Omega_z \uparrow \Omega_z \times \Omega_z}_{\gamma,t-1} \textcircled{\cap} m^{\Omega_z \uparrow \Omega_z \times \Omega_z}_{\gamma,t} \tag{59}$$

That greatly simplifies the problem to estimate transitions since there are less conjunctive combinations with less number of focal elements. Eq. 59 leads to two kinds of transitions:

- The joint form given by Eq. 14 for all subsets $S_{t-1,k} \subseteq \Omega_z$ and subsets $S_{tl} \subseteq \Omega_z$;
- The conditional form given by Eq. 14 for singletons $s_{t-1,k} \in \Omega_z$ and subsets $S_{tl} \subseteq \Omega_z$. The conditional form on subsets are then computed by the DRC (Eq. 3b).

As discussed previously (Section III-C) and shown in the experiments (Section V), the conditional form may be more suitable in practice.

*b) Concerning the Gaussian mixing weights:* The amount of overlap among the mixture components and their relative size may have a negative impact on the convergence of EM which becomes slower and may lead to component annealing [39]. This effect is expected to be amplified by the use of the plausibilities in parameter optimization (Eq. 27) since it shares belief from subsets involving several singletons. In practice, we suggest to not consider the weights in the mixture. This simplification decreases the number of parameters to be estimated. The quality of the model may not be degraded since other additional parameters are considered: Larger number of transitions by considering subsets, and doubt between components of the mixtures through plausibilities.

**Algorithm 2** Find parameters

**Require:** Initial model $\Psi^{(0)} = \{q_a^{(0)}, \theta^{(0)}, \Pi^{(0)}\}$ ($\theta$ given by Eq. 1)
**Ensure:** Best parameter estimates $\Psi^* = \{q_a^*, \theta^*, \Pi^*\}$ and evidential likelihood $\mathcal{L}_e$

{Initialisation}
Prepare transitions $\overrightarrow{q}_a$ with Eq. 34
Prepare transitions $\overleftarrow{q}_a$ with Eq. 38
$\mathcal{L}_e^{(0)} \leftarrow -\infty$
**repeat**
  $q \leftarrow q+1$
  {E-STEP, with $\Psi^{(q)}$}
  Apply forward propagation with Eq. 31-34 to get $m_{\alpha,t}$, $t = 1 \dots T$
  Normalize $m_{\alpha,t}$ at each time step as in Eq. 46 to get $c_t$
  Compute the likelihood $\mathcal{L}_e^{(q)}$ as in Eq. 43 using $c_t$
  Apply the backward propagation with Eq. 35-38 to get $m_{\beta,t}$, $t = 1 \dots T$
  Apply Eq. 39 to get the posterior BBA $m_{\gamma,t}$ using $m_{\beta,t}$ and $m_{\alpha,t}$

  {M-STEP, to get $\Psi^{(q+1)}$}
  Get the plausibilities on singletons $pl_{\gamma,t}$ with Eq. 4
  Compute new means and covariances with Eq. 28 and 29
  Compute new transitions with Eq. 40 using the approximation in Eq. 59
**until** $\frac{|\mathcal{L}_e^{(q)} - \mathcal{L}_e^{(q-1)}|}{\mathcal{L}_e^{(q-1)}} < \epsilon$

*Remark 2: The evidential forward-backward algorithm allows to retrieve the results of standard HMM since there is no approximation (when using the joint BBA for transitions in Eq. 40). A Bayesian BBA is indeed easily encoded using commonalities which are used in most of equations. The same remark holds for the learning procedure, the only difference holds in the fact that the prior on states at $t = 1$ is not considered.*

## V. Results

The goal of this section is to illustrate the inference and learning procedures in EvHMM using synthetic datasets. Some complementary results obtained on benchmarks used in prognostics and health management are presented in [40].

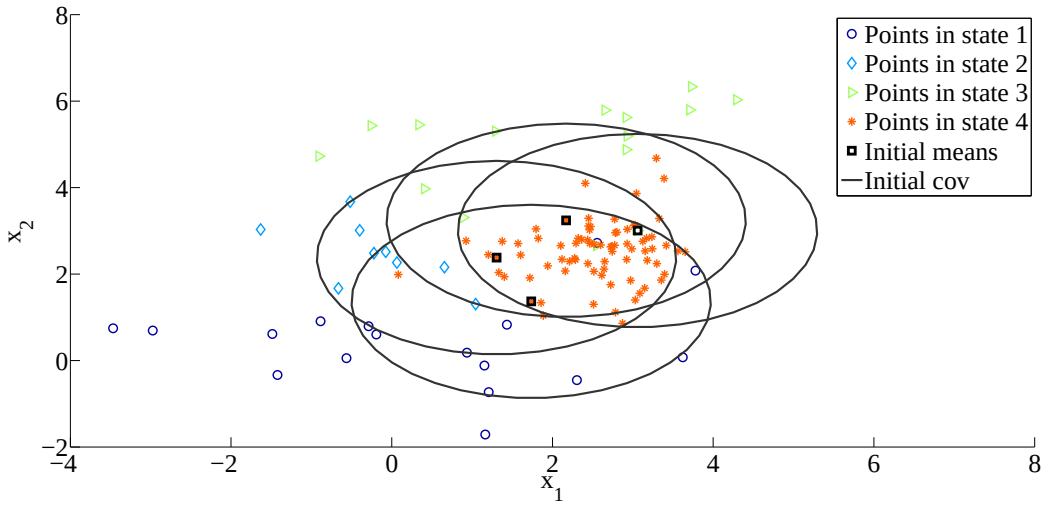



Fig. 5. One simulated dataset. Initial position of means corresponds to random selection of data points, initial covariances are set to $diag([5,5])$

.

### A. Datasets

Two-dimensional synthetic datasets are generated from a probabilistic HMM with $Q = 4$ states with the following parameters:

- Transition probabilities are set to $p_a(S_j \mid S_i) = \frac{a_{ij}}{\sum_{j'} a_{ij'}}$ where elements $a_{ij}$ are drawn from an uniform distribution such that the elements of the diagonal are greater than the off-diagonal ones: $a_{ij} \sim \mathbb{1}_{i=j} + \mathcal{U}_{[0,1]}$;
- Prior probabilities on the first state are drawn from an uniform distribution: $p_\pi(S_j) = \frac{\pi_j}{\sum_{j'} \pi_{j'}}$ with $\pi_j \sim \mathcal{U}_{[0,1]}$.
- The means of the Gaussian emission model are set to: $\boldsymbol{\mu} = \begin{bmatrix} 0 & 0 & 2.5 & 2.5 \\ 0 & 2.5 & 5 & 2.5 \end{bmatrix}$
- The covariance matrix $\Sigma_j$ for state $j$ is calculated as $\boldsymbol{\Sigma}_j = \nu_j \, \mathrm{cov}(\boldsymbol{D}_j)$ where $\boldsymbol{D}_j = \begin{bmatrix} \mathbf{x}_1 \, \mathbf{x}_2 \, \ldots \, \mathbf{x}_t \, \ldots \, \mathbf{x}_T \end{bmatrix}'$ is a set of data with length $T_j = 120$ (number of data points per state) sampled as $\mathbf{x}_t \sim \mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix})$ and using a scattering controlled by $\boldsymbol{\nu} = \begin{bmatrix} 2 & 0.5 & 2 & 0.5 \end{bmatrix}$.

This probabilistic HMM is used as a generative model to obtain 200 datasets, each with 480 data points. The HMM and EvHMM are then run onto each dataset with the aim to recover the hidden structure of the data. For each dataset, the sequence of hidden states is known (by sampling from the HMM) and corresponds to a ground truth to which the results of both HMM and EvHMM are compared using the Adjusted Rand Index (ARI) [41]. The ARI is a measure of agreement between two partitions in clustering, and in the present case between two sequences of states (the estimated one by the EvHMM or HMM, and the ground truth), and tends to 1 if the sequence estimated is equal to the ground truth. For comparison purposes, both the EvHMM and HMM are initialised with the same means and covariance matrices for each dataset.

### B. Comparison with HMM

*1) Conditional form for transitions in EvHMM (EvHMM-CT):* Figure 6 represents the box plots of the relative performance of EvHMM-CT (EvHMM with conditional form of transition, Section IV-C) and HMM with respect to the $n$ best results, where $n$ is computed by the percentiles (so here $n = 20, 40 \ldots 200$). The EvHMM performance is computed as:

$$\mathcal{P}^{\text{rel}}_{\text{EvHMM}} = 100 \frac{\mid ARI_{\text{EvHMM}} - ARI_{\text{HMM}} \mid}{ARI_{\text{HMM}}} \tag{60}$$

that is *relative* to the performance of the probabilistic HMM which should expectedly provide similar or better results compared to EvHMM due to the data generation process which is made by a standard HMM. This is the case for $n = 200$ for which results are rather similar on average. This result means that the EM procedure for the EvHMM works but, similarly to HMM, is quite sensitive to the initialisation. Remembering that the datasets have been generated by a probabilistic HMM, an advantage can be attributed to the EvHMM with the 75-th percentile equal to $+26\%$. Note also that we do not make use of prior on latent variables (Sections II-C and III) which has been shown to be interesting for HMM [6].

Now, consider lower $n$. The trend of the median against $n$ evolves exponentially with $n$: When considering the 20 best results, an improvement of $117.8\%$ (75-th percentile) can be obtained (with median equal to $103\%$) in favor of the EvHMM. These results show that, with appropriate initialisation, the EvHMM has the highest potential for state recognition on this

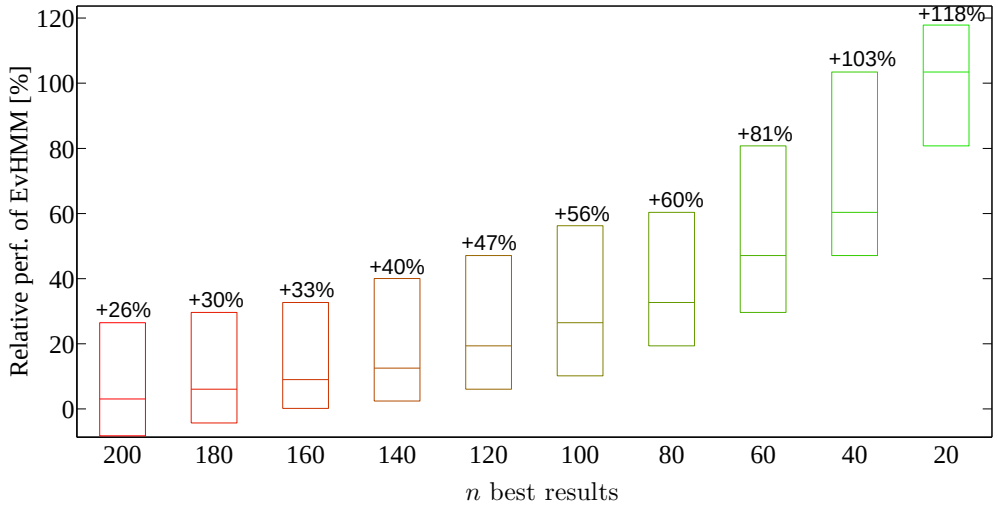



Fig. 6. Using conditional form of transitions (EvHMM-CT): Relative performance ($n$ best results among 200) of EvHMM taking the HMM as a reference.

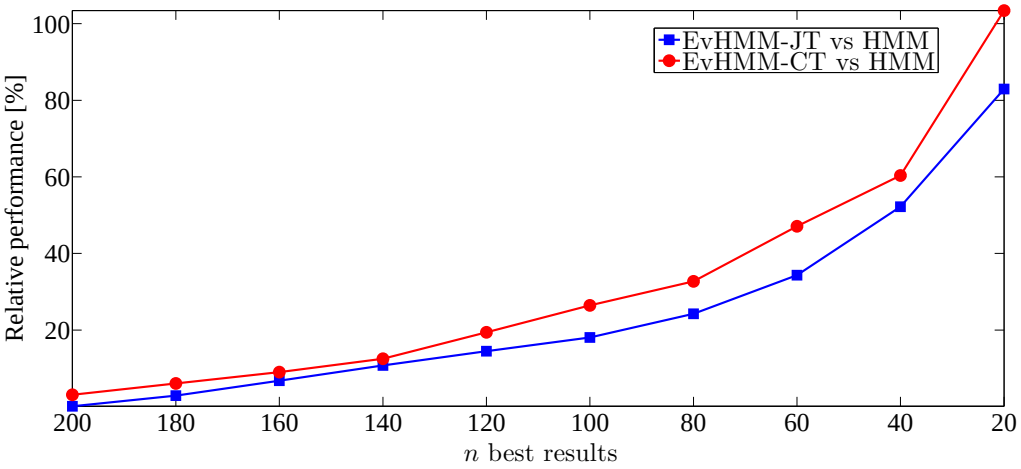



Fig. 7. Performance when using joint form of transitions (EvHMM-JT) against EvHMM-CT ($n$ best results among 200, taking the HMM as a reference). Only trends in median are reported (variability of EvHMM-JT was similar to the previous case with EvHMM-CT).

dataset. Note that the initial conditions have been the same for both HMM and EvHMM. Therefore, those results also mean that both models seem complementary.

*2) Comparison with joint form for transitions in EvHMM (EvHMM-JT):* The results obtained with the conditional form of transitions (EvHMM-CT) are compared to the previous case in Figure 7. EvHMM-JT leads to better performance than HMM as previously but are are lower than with EvHMM-CT.

The EvHMM-CT seems thus to be more suitable in terms of performance for these datasets. As discussed previously, this model is more parcimonious than the EvHMM-JT so it requires less memory (coding only belief functions conditionally to singletons in transitions) and is coherent with the GBT.

## VI. Conclusion

This paper provides the formulations for inference and learning in Evidential HMM with conditional belief functions, which are an equivalent model of probabilistic HMM where the uncertainty on states is managed by belief functions. We have shown that the "three problems" defined for HMM can be solved for the EvHMM. In particular, the likelihood can be computed exactly in EvHMM. The learning problem has been solved using some assumptions to make the solutions tractable:

- We suggested to use conditional belief functions for transitions between states for a coherence with the Generalized Bayesian Theorem and to decrease time and memory consumption. Results on simulated data show a better performance with the conditional form compared to the joint form.

- We did not consider mixing weights in the GMM due to component annealing observed in practice which was also reported for probabilistic models [39].
- It was shown that the support assigned to a Markov chain can be computed exactly using plausibilities. It was approximated by a product of BBAs as proposed in [10] to get simple reestimation formula.
- The criterion used for the optimization is approximated by considering only the terms based on plausibilities (and therefore by neglecting some subsets).
- The criterion used for learning relies on the same conjecture as in [31] which states that its maximization does not degrade the support of the model. It was shown that this conjecture can be easily controled in EvHMM at each iteration using the forward algorithm.

Despite those assumptions, results on simulated datasets show a high potential compared to standard HMM. Moreover, replacing BBAs by probabilities in all algorithms allows to recover standard HMM.

The adjustment of the posterior BBA to compensate model's misspecification [33] was not studied in experiments, this could be studied in the future since it has been shown for HMM that this procedure could be of great interest in partially supervised learning [6]. Some preliminary results have been presented in [40] where it is shown that the EvHMM with partial knowledge on hidden states behaves similarly in presence of uncertainty but differently, in a better way, with noisy labels. More experiments on diverse applications, such as related to time-series classification, forecasting and systems' prognostics [42], is necessary to confirm those results.

Comparison of the methodology of evidential HMM against other imprecise and fuzzy uncertainty theories can also be performed. Mohamed and Gader [43] suggested a Choquet-based formulation using conditional fuzzy measure with application to handwritten word images analysis. Soubaras [44] considered a similar extension to propose risk measures with application in crisis management. The relation with HMM formulated with imprecise probabilities [45], [46] can also be of interest. Those models are able to generate bounds around expectations on states or on parameters which may be interesting for datasets with limited knowledge or for decision-making in critical applications. Since the plausibility and belief on states can be generated during the forward-backward propagations, the use of the EvHMM for such aims could be a path to follow.

The learning procedure in EvHMM was sensitive to the initialisation phase, likewise to standard HMM due to the use of EM. It could be of interest to consider EM's alternatives such as Iterative Conditional Estimation and Stochastic EM [47] to make initialisation more robust. Still on learning, the relationships with entropy and divergence measures in the belief functions framework should be more studied. On this subject the reader can be interested in the recent survey paper by Jirousek and Shenoy [48].

A fourth long-term perspective concerns the application of the proposed learning concepts to more general Time-Sliced/Dynamic/Temporal Evidential Networks considering for instance the formalism of Pairwise and Triplet Markov Chain [9], [10].

## ACKNOWLEDGMENT

The author would like to express his gratitude to Michèle Rombaut, Denis Pellerin and Thierry Denoeux for discussions around inference in EvHMM and EM-based learning in HMM.

This work has been carried out in various projects managed at the Department of Applied Mechanics, at the Department of Automatic Control and Micro-Mechatronic Systems, in particular the CNRS-PEPS project "EVIPRO" and the "SMART COMPOSITES" project (FRI2) supported by the Région Franche-Comté and *Bpifrance financement*. It also gets support from the Laboratory of Excellence "ACTION" through the program *Investments for the future* managed by the National Agency for Research (references ANR-11-LABX-01-01).

## REFERENCES

[1] D. Wingate, N. D. Goodman, D. M. Roy, and J. B. Tenenbaum, "The infinite latent events model," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 607–614.
[2] A. Anandkumar, D. Hsu, and S. M. Kakade, "A method of moments for mixture models and hidden Markov models," in *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, 2012, pp. 33.1–33.34.
[3] S. L. Scott, G. M. James, and C. A. Sugar, "Hidden Markov models for longitudinal comparisons," *Journal of the American Statistical Association*, vol. 100, no. 470, 2005.
[4] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The annals of mathematical statistics*, pp. 164–171, 1970.
[5] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, 1989.
[6] E. Ramasso and T. Denoeux, "Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions," *Fuzzy Systems, IEEE Transactions on*, vol. 22, no. 2, pp. 395–405, 2014.
[7] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.
[8] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[9] W. Pieczynski, "Pairwise markov chains," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 634–639, 2003.
[10] ——, "Multisensor triplet Markov chains and theory of evidence," *International Journal of Approximate Reasoning*, vol. 45, no. 1, pp. 1–16, 2007.
[11] K. Murphy, "Dynamic Bayesian Networks: Representation, inference and learning," Ph.D. dissertation, UC Berkeley, Computer Science Division, 2002.
[12] G. Shafer, *A mathematical theory of Evidence*. Princeton University Press, 1976.

[13] P. Smets and R. Kennes, "The Transferable Belief Model," *Artificial Intelligence*, vol. 66, pp. 191–234, 1994.

[14] T. Augustin, F. P. A. Coolen, G. D. Cooman, and M. C. M. Troffaes, Eds., *Introduction to Imprecise Probabilities*, ser. Wiley Series in Probability and Statistics. Wiley, 2014.

[15] J. Pearl, "Reasoning with belief functions: An analysis of compatibility," *Int. Jour. of Approximate Reasoning*, vol. 4, no. 5-6, pp. 363–389, 1990.

[16] B. Cobb and P. Shenoy, "A comparison of Bayesian and belief function reasoning," *Information Systems Frontiers*, vol. 5, no. 4, pp. 345–358, 2003.

[17] P. Smets, "Beliefs functions: The disjunctive rule of combination and the generalized Bayesian theorem," *IJAR*, vol. 9, pp. 1–35, 1993.

[18] ——, "Advances in the dempster-shafer theory of evidence," R. R. Yager, J. Kacprzyk, and M. Fedrizzi, Eds. New York, NY, USA: John Wiley & Sons, Inc., 1994, ch. What is Dempster-Shafer's Model?, pp. 5–34.

[19] G. Shafer, "Perspectives on the theory and practice of belief functions," *International Journal of Approximate Reasoning*, vol. 4, pp. 323–362, 1990.

[20] E. Pollard, M. Rombaut, and B. Pannetier, "Bayesian networks vs. evidential networks. an application to convoy detection," in *IPMU conference*, 2010.

[21] P. Smets, "The application of the matrix calculus to belief functions," *Int. Jour. of Approximate Reasoning*, vol. 31, no. 1-2, pp. 1–30, 2002.

[22] D. Dubois and H. Prade, "On the unicity of Dempster rule of combination," *International Journal of Intelligent Systems*, vol. 1, no. 2, pp. 133–142, 1986.

[23] T. Denoeux, "Maximum likelihood estimation from uncertain data in the belief function framework," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 1, pp. 119–130, 2013.

[24] E. Ramasso, M. Rombaut, and D. Pellerin, "Forward-backward-viterbi procedures in TBM for state sequence analysis using belief functions," in *ECSQARU*, 2007, pp. 405–417.

[25] A. Bendjebbour, Y. Delignon, F. Fouque, V. Samson, and W. Pieczynski, "Multisensor image segmentation using Dempster-Shafer fusion in Markov fields context," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 8, pp. 1789–1798, 2001.

[26] P. Lanchantin and W. Pieczynski, "Unsupervised restoration of hidden nonstationary Markov chains using evidential priors," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3091–3098, 2005.

[27] M. E. Y. Boudaren, E. Monfrini, W. Pieczynski, and A. Aissani, "Dempster-Shafer fusion of multisensor signals in nonstationary Markovian context," *EURASIP J. Adv. Sig. Proc.*, vol. 2012, no. 134, pp. 1–13, 2012.

[28] A. Dempster, "A generalization of Bayesian inference," *Journal of the RSS*, vol. 30, pp. 205–247, 1968.

[29] L. Serir, E. Ramasso, and N. Zerhouni, "Time-sliced temporal evidential networks: The case of evidential HMM with application to dynamical system analysis," in *Prognostics and Health Management (PHM), 2011 IEEE Conference on*, June 2011, pp. 1–10.

[30] P. Shenoy, "Valuation-based systems: A framework for managing uncertainty in expert systems," *Fuzzy Logic for theManagement of Uncertainty*, pp. 83–104, 1992.

[31] P. Vannoorenberghe and P. Smets, "Partially supervised learning by a credal EM approach," in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, ser. Lecture Notes in Computer Science, L. Godo, Ed. Springer Berlin Heidelberg, 2005, vol. 3571, pp. 956–967.

[32] G. Patil, *Weighted distributions*, A. H. El-Shaarawi and W. W. Piegorsch, Eds. John Wiley & Sons, Ltd, Chichester, 2002, vol. 4, pp. 2369-2377.

[33] P. Cano and E. Ramasso, "Ascertainment-adjusted parameter estimation approach to improve robustness against misspecification of health monitoring methods," *Mechanical Systems and Signal Processing*, 2016, submitted, revision 2.

[34] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, vol. 30, no. 1, pp. 175–193, 1906.

[35] P. Vannoorenberghe, "Estimation de modèles de mélanges finis par un algorithme em crédibiliste," *Traitement du signal*, vol. 24, pp. 103–113, 2006.

[36] F. Delmotte and P. Smets, "Target identification based on the Transferable Belief Model interpretation of Dempster-Shafer model," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 34, no. 4, pp. 457–471, 2004.

[37] T. Denoeux, "Likelihood-based belief function: justification and some extensions to low-quality data," *International Journal of Approximate Reasoning*, vol. 55, no. 7, pp. 1535–1547, 2014.

[38] J. Bilmes, "A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden Markov models," University of Berkeley, Technical Report ICSI-TR-97-021, 1997.

[39] I. Naim and D. Gildea, "Convergence of the EM algorithm for gaussian mixtures with unbalanced mixing coefficients," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, J. Langford and J. Pineau, Eds. New York, NY, USA: ACM, 2012, pp. 1655–1662.

[40] E. Ramasso, "A solution for the learning problem in evidential (partially) hidden markov models based on conditional belief functions and EM," in *16th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems*, ser. Communications in Computer and Information Science (CCIS). Eindhoven, The Netherlands: Springer, 20-24 June 2016.

[41] N. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clustering comparison: Is a correction for chance necessary?" in *Proc. of the 26th Annual Int. Conf. on Machine Learning*, 2009, pp. 1073–1080.

[42] E. Ramasso and A. Saxena, "Performance benchmarking and analysis of prognostic methods for CMAPSS datasets," *International Journal on Prognostics and Health Management*, vol. 5, no. 2, pp. 1–15, 2014.

[43] M. Mohamed and P. Gader, "Generalized hidden markov models - Part I: Theoretical frameworks," *IEEE Trans. on Fuzzy Systems*, vol. 8, no. 1, pp. 67–81, 2000.

[44] H. Soubaras, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 10th European Conference, ECSQARU 2009, Verona, Italy, July 1-3, 2009. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, ch. An Evidential Measure of Risk in Evidential Markov Chains, pp. 863–874.

[45] J. D. Bock and G. de Cooman, "An efficient algorithm for estimating state sequences in imprecise hidden markov models," *Journal of Artificial Intelligence Research*, vol. 50, pp. 189–233, 2014.

[46] A. Antonucci, R. D. Rosa, A. Giusti, and F. Cuzzolin, "Robust classification of multivariate time series by imprecise hidden Markov models," *International Journal of Approximate Reasoning*, vol. 56, no. B, pp. 249 – 263, 2015.

[47] A. Peng and W. Pieczynski, "Adaptive mixture estimation and unsupervised local bayesian image segmentation," *CVGIP: Graphical Models and Image Processing*, vol. 57, no. 5, pp. 389–399, 1995.

[48] R. Jirousek and P. P. Shenoy, "A new definition of entropy of belief functions in the dempster-shafer theory," University of Kansas School of Busines, Lawrence, KS 66045, USA, Tech. Rep. 330, 2016.