A new method for learning imprecise hidden Markov models

Arthur Van Camp and Gert de Cooman

Ghent University, SYSTeMS Research Group
Technologiepark–Zwijnaarde 914, 9052 Zwijnaarde, Belgium
{arthur.vancamp,gert.decooman}@UGent.be

Abstract We present a method for learning imprecise local uncertainty models in stationary hidden Markov models. If there is enough data to justify precise local uncertainty models, then existing learning algorithms, such as the Baum–Welch algorithm, can be used. When there is not enough evidence to justify precise models, the method we suggest here has a number of interesting features.

Keywords: Hidden Markov model, learning, expected counts, imprecise Dirichlet model.

1 Introduction

In practical applications of reasoning with hidden Markov models, or HMMs, an important problem is the assessment of the local uncertainty models. In many applications, the amount of data available for learning the local models is limited. This may be due to the costs of data acquisition, lack of expert knowledge, time limitations, and so on [4,9]. In this case, we believe using precise(-probabilistic) local uncertainty models is hard to justify. This leads us to use imprecise(-probabilistic) local uncertainty models, turning the HMM into an imprecise hidden Markov model (iHMM).

Convenient imprecise probability models are coherent lower previsions, see [6] for a detailed exposition. In this paper we develop a method for learning imprecise local models, in the form of coherent lower previsions, in iHMMs.

Learning of iHMMs has been explored earlier [1,5]. However, these papers deal with learning transition models and do not consider learning emission models. In this paper, we want to extend this to learning all the local models of an iHMM.

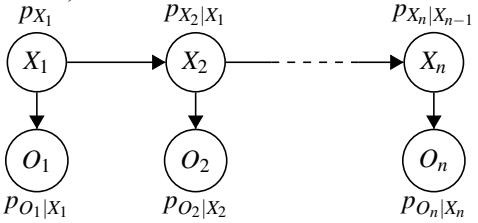
We start with a short introduction of the relevant aspects of HMMs and iHMMs in Section 2. In Section 3, we show how to learn imprecise local models—first if the state sequence is supposed to be known, and finally for hidden state sequences. In Section 4, we recall basic aspects of the Baum–Welch algorithm, relevant to our purpose. In Section 5, we apply our method to a problem of predicting future earthquake rates.

2 Hidden Markov models and basic notions

2.1 Precise hidden Markov models

An HMM with length n has n state variables X_t that are hidden or unobservable, and n observation variables O_t that are observable. The figure below shows a graphical

representation of a HMM, with the local uncertainty model (characterised by a mass function in the precise case) for each variable shown next to the corresponding node.



Each state variable X_t, with t in $\{1,\ldots,n\}$, takes one of the m possible values in the finite set $\mathscr{X}_t = \mathscr{X} := \{\xi_1,\ldots,\xi_m\}$. Each observation variable O_t, with t in $\{1,\ldots,n\}$, takes one of the p possible values in the finite set $\mathscr{O}_t = \mathscr{O} := \{\omega_1,\ldots,\omega_p\}$. We denote by x_t a generic value that X_t takes in $\mathscr{X}$, and by o_t a generic value that O_t takes in $\mathscr{O}$.

The local uncertainty model $p_{X_t|X_{t-1}}$ describes probabilistic knowledge about state variable X_t, conditional on the previous state variable X_{t-1}, with t in $\{2,\ldots,n\}$, and is called a *precise transition model*. The probability that state variable X_t takes value x_t, conditional on $X_{t-1} = x_{t-1}$, is written as $p_{X_t|X_{t-1}}(x_t|x_{t-1})$.

The local uncertainty model $p_{O_t|X_t}$ describes probabilistic knowledge about observation variable O_t, conditional on the corresponding state variable X_t, with t in $\{1,\ldots,n\}$, and is called a *precise emission model*. The probability that observation variable O_t takes value o_t, conditional on $X_t = x_t$, is written as $p_{O_t|X_t}(o_t|x_t)$.

The only variable we have not paid attention to so far is the first state variable X_1. The local uncertainty model p_{X_1} describes probabilistic knowledge about the first state variable X_1, and is not conditional. It is called a *precise marginal model*. The probability that state variable $X_1 = x_1$ is written as $p_{X_1}(x_1)$.

We write the state sequence as $X_{1:n} = x_{1:n}$ and the observation sequence as $O_{1:n} = o_{1:n}$. We use notations like $A_{p:n} := (A_p,\ldots,A_n)$ if $p \leq n$ and $A_{p:n} := \emptyset$ if $p > n$. For notational convenience, we also use another way of denoting state and observation sequences. There is a unique $l_{1:n} \in \times_{i=1}^{n} \mathscr{X}_i$ such that the state sequence $X_{1:n} = x_{1:n}$ can be written as $X_{1:n} = (\xi_{l_1},\ldots,\xi_{l_n})$, and a unique $h_{1:n} \in \times_{i=1}^{n} \mathscr{O}_i$ such that the observation sequence $O_{1:n} = o_{1:n}$ can be written as $O_{1:n} = (\omega_{h_1},\ldots,\omega_{h_n})$. We will use these unique letters l_i and h_i throughout.

We assume each HMM to be *stationary*, meaning that $p_{O_t|X_t} = p_{O|X}$ for all t in $\{1,\ldots,n\}$ and $p_{X_t|X_{t-1}} = p_{X_i|X_{i-1}}$ for all t,i in $\{2,\ldots,n\}$. The probability $p_{O|X}(\omega_h|\xi_l)$, with h in $\{1,\ldots,p\}$ and l in $\{1,\ldots,m\}$, of a state variable that takes value ξ_l emitting value ω_h is also denoted as E_{hl}. Furthermore, the probability $p_{X_t|X_{t-1}}(\xi_h|\xi_l)$, with l,h in $\{1,\ldots,m\}$ (this probability does not depend on t since the HMM is stationary), of a transition from a state variable taking value ξ_l to a state variable taking value ξ_h is also denoted as T_{lh}. Finally, the probability $p_{X_1}(\xi_l)$ that the first state variable X_1 assumes the value ξ_l is also denoted by p_l.

The model parameter $\boldsymbol{\theta}$ is the vector with all parameters of the marginal, transition and emission models. It has $m(p+m+1)$ elements, and is explicitly defined as:

$$\boldsymbol{\theta} := \left[p_1 \cdots p_m \; T_{11} \cdots T_{mm} \; E_{11} \cdots E_{pm}\right].$$

We write models that depend on (components of) $\boldsymbol{\theta}$ as models conditional on $\boldsymbol{\theta}$.

In our HMMs, we impose the usual *Markov condition* for Bayesian networks: for any variable, conditional on its mother variable, the non-parent non-descendent variables are independent of it (and its descendants). Here, the Markov condition reduces to the following conditional independence conditions. For each t in $\{1,\ldots,n\}$:

$$p_{X_t|X_{1:t-1},O_{1:t-1}}(x_t|x_{1:t-1},o_{1:t-1},\boldsymbol{\theta}) = p_{X_t|X_{t-1}}(x_t|x_{t-1}),$$
$$p_{O_t|X_{1:n},O_{1:t-1},O_{t+1:n}}(o_t|x_{1:n},o_{1:t-1},o_{t+1:n},\boldsymbol{\theta}) = p_{O_t|X_t}(o_t|x_t).$$

2.2 Imprecise hidden Markov models

An iHMM has the same graphical structure as an HMM, with the precise-probabilistic local models replaced by imprecise-probabilistic variants. Convenient imprecise probability models are coherent lower previsions.

A coherent lower prevision (or lower expectation functional) $\underline{P}$ is a real-valued functional defined on real-valued functions—called *gambles*—of variables. We denote the set of all gambles on the variable X by $\mathscr{L}(\mathscr{X})$. A gamble is interpreted as an uncertain award or penalty: it yields $f(x)$ if X takes value x in $\mathscr{X}$. A *coherent lower prevision* $\underline{P}$ defined on $\mathscr{L}(\mathscr{X})$ satisfies the following requirements:

C1. $\underline{P}(f) \geq \min_{x\in\mathscr{X}} f(x)$ for all f in $\mathscr{L}(\mathscr{X})$;
C2. $\underline{P}(\mu f) = \mu\underline{P}(f)$ for all real $\mu \geq 0$ and all f in $\mathscr{L}(\mathscr{X})$;
C3. $\underline{P}(f+g) \geq \underline{P}(f)+\underline{P}(g)$ for all f,g in $\mathscr{L}(\mathscr{X})$.

With a coherent lower prevision, we can associate a conjugate coherent upper prevision $\overline{P}$ as follows: $\overline{P}(f) := -\underline{P}(-f)$ for all gambles f. The interpretation of coherent lower and upper previsions is as follows. $\underline{P}(f)$ is a subject's supremum buying price for the gamble f, and consequently $\overline{P}(f)$ is his infimum selling price for f. For more information, see for instance [6].

The lower and upper probability of an event $A \subseteq \mathscr{X}$ are defined as $\underline{P}(A) := \underline{P}(\mathbb{I}_A)$ and $\overline{P}(A) := \overline{P}(\mathbb{I}_A)$, where $\mathbb{I}_A$ is the indicator (gamble) of a set A, which assumes the value 1 on A and 0 elsewhere.

We denote the *imprecise marginal model* by $\underline{Q}_1$, defined on $\mathscr{L}(\mathscr{X}_1)$. The *imprecise transition model* for state variable X_t, for t in $\{2,\ldots,n\}$, is denoted by $\underline{Q}_t(\cdot|X_{t-1})$, defined on $\mathscr{L}(\mathscr{X}_t)$ and the *imprecise emission model* for observation variable O_t, for t in $\{1,\ldots,n\}$, is denoted by $\underline{S}_t(\cdot|X_t)$, defined on $\mathscr{L}(\mathscr{O}_t)$. We assume our iHMM also to be stationary, meaning that the local models do not depend on t.

In an iHMM, the Markov condition turns into the *conditional irrelevance assessment*, meaning that, conditional on its mother variable, the non-parent non-descendant variables are assumed to be epistemically irrelevant to the variable and her descendants (see [3]). With this conditional irrelevance assessment, the following recursion relations hold for the joint lower prevision $\underline{P}_t(\cdot|X_{t-1})$ on $\mathscr{L}(\times_{i=t}^{n}(\mathscr{X}_i\times\mathscr{O}_i))$:

$$\underline{P}_t(\cdot|X_{t-1}) = \underline{Q}_t(\underline{E}_t(\cdot|X_t)|X_{t-1}) \qquad \text{for } t = n,\ldots,2,$$
$$\begin{cases} \underline{E}_n(\cdot|X_n) = \underline{S}_n(\cdot|X_n) \\ \underline{E}_t(\cdot|X_t) = \underline{P}_t(\cdot|X_{t-1})\otimes\underline{S}_{t-1}(\cdot|X_{t-1}) \qquad \text{for } t = n-1,\ldots,1, \end{cases}$$

The joint lower prevision $\underline{P}$ defined on $\mathscr{L}(\times_{i=1}^{n}(\mathscr{X}_i \times \mathscr{O}_i))$ of all the variables is given by $\underline{P}_f(\cdot) = \underline{Q}_1(\underline{E}_t(\cdot|X_t))$.

In the next section, we start by presenting a method for learning the imprecise local uncertainty models of an iHMM, if both the observation sequence and the state sequence is given. Since the state sequence is actually unobservable, in Section 3.3 we present a method to estimate the relevant quantities from only an observation sequence.

3 Learning imprecise local uncertainty models

Since transitions between state variables and emissions of observation variables can be seen as instances of IID processes, whose behaviour is usefully summarised by multinomial processes, a convenient model to describe uncertainty about transition and emission probabilities are the conjugate Dirichlet models. One important imprecise-probabilistic variant of these is the imprecise Dirichlet model (IDM) [7].

3.1 Imprecise Dirichlet model

Without going into too much detail, let us briefly recall the relevant ideas about the IDM. If $n(A)$ is the number of occurrences of an event A in N experiments, then the lower and upper probability of A according to an IDM are defined as

$$\underline{P}(A) = \frac{n(A)}{N+s} \text{ and } \overline{P}(A) = \frac{n(A)+s}{N+s},$$

where s is a hyperparameter called the number of pseudo-counts. This is a non-negative real number on which the imprecision $\Delta(P(A)) := \overline{P}(A) - \underline{P}(A) = s/N+s$ depends. The larger s, the more imprecise the inferences. If $s = 0$, the resulting precise model returns the relative frequency $\overline{P}(A) = \underline{P}(A) = n(A)/N$ of the occurrence of A.

Once we have chosen a value for s, we can use the IDM to infer interval estimates for the probability of A from observations. The choice of s is, however, fairly arbitrary; see also [7], where it is argued that for example $s = 2$ might be a good choice.

3.2 Known state sequence

Our aim is to learn local models, based on a known observation sequence $O_{1:n} = (\omega_{h_1}, \ldots, \omega_{h_n})$. Assume for the time being the state sequence $X_{1:n} = x_{1:n} = (\xi_{l_1}, \ldots, \xi_{l_n})$ to be also known, then we can build imprecise estimates for the local uncertainty models as follows.

We first define for each i and g in $\{1, \ldots, m\}$ the following numbers (or rather functions of the state sequence $x_{1:n}$) n_{ξ_i} and n_{ξ_i,ξ_g} as:

$$n_{\xi_i}(x_{1:n}) := \sum_{t=1}^{n} \mathbb{I}_{\{\xi_i\}}(x_t) \text{ and } n_{\xi_i,\xi_g}(x_{1:n}) := \sum_{t=2}^{n} \mathbb{I}_{\{(\xi_i,\xi_g)\}}(x_{t-1}, x_t).$$

The interpretation of these numbers is immediate: n_{ξ_i} is the number of times the value ξ_i is reached in the whole state sequence $x_{1:n}$ and n_{ξ_i,ξ_g} is the number of times that a state transition from value ξ_i to value ξ_g takes place in the whole state sequence $x_{1:n}$.

Imprecise transition model: The event of interest here is the transition from a state variable taking value ξ_i in $\mathscr{X}$ to the subsequent state variable taking value ξ_g in $\mathscr{X}$. This event occurs n_{ξ_i,ξ_g} times. The number of "experiments" N is the number of times $\sum_{g=1}^{m} n_{\xi_i,\xi_g}$ that a transition from value ξ_i takes place. The IDM leads to the following imprecise transition model (in terms of lower and upper transition probabilities):

$$\underline{Q}(\{\xi_g\}|\xi_i) = \frac{n_{\xi_i,\xi_g}}{s+\sum_{g=1}^{m} n_{\xi_i,\xi_g}} \text{ and } \overline{Q}(\{\xi_g\}|\xi_i) = \frac{s+n_{\xi_i,\xi_g}}{s+\sum_{g=1}^{m} n_{\xi_i,\xi_g}}.$$

Since here and in what follows, the IDM produces a linear-vacuous model [6,7] for the probabilities, these lower and upper probabilities determine the imprecise model.

Imprecise emission model: The event of interest here is the emission of observation o in $\mathscr{O}$ from corresponding state variable taking value ξ_i in $\mathscr{X}$. This event occurs $\sum_{\{t:\omega_{h_t}=o\}} \mathbb{I}_{\{\xi_i\}}(x_t)$ times. The total number of times an emission from value ξ_i takes place, is n_{ξ_i}. The IDM then leads to the following imprecise emission model:

$$\underline{S}(\{o\}|\xi_i) = \frac{\sum_{\{t:\omega_{h_t}=o\}} \mathbb{I}_{\{\xi_i\}}(x_t)}{s+n_{\xi_i}} \text{ and } \overline{S}(\{o\}|\xi_i) = \frac{s+\sum_{\{t:\omega_{h_t}=o\}} \mathbb{I}_{\{\xi_i\}}(x_t)}{s+n_{\xi_i}}.$$

Imprecise marginal model: The event of interest here is the state variable X_1 taking value ξ_i in $\mathscr{X}$. The number of times this event occurs is $\mathbb{I}_{\{\xi_i\}}(x_1)$. The total number of times state variable X_1 takes any value is of course 1. The IDM then leads to the following imprecise marginal model:

$$\underline{Q}_1(\{\xi_i\}) = \frac{\mathbb{I}_{\{\xi_i\}}(x_1)}{s+1} \text{ and } \overline{Q}_1(\{\xi_i\}) = \frac{s+\mathbb{I}_{\{\xi_i\}}(x_1)}{s+1}.$$

3.3 Unknown state sequence

Since in an HMM the state sequence $X_{1:n}$ is unobservable (hidden), the numbers n_{ξ_i} and n_{ξ_i,ξ_g} are actually random variables N_{ξ_i} and N_{ξ_i,ξ_g}: functions of the hidden state sequence $X_{1:n}$. This means we can no longer use them directly to learn the imprecise local models. Instead of using these random variables N_{ξ_i} and N_{ξ_i,ξ_g}, we will rather use their expected values, conditional on the known observation sequence $o_{1:n}$ and the model parameter $\boldsymbol{\theta}^*$. Here $\boldsymbol{\theta}^*$ is a local maximum of the likelihood, obtained by the Baum–Welch algorithm (see Section 4). We define the expected counts $\hat{n}_{\xi_i}$ an $\hat{n}_{\xi_i,\xi_g}$ as

$$\begin{aligned}
\hat{n}_{\xi_i} &:= E\left(N_{\xi_i}\middle|o_{1:n},\boldsymbol{\theta}^*\right) = \sum_{t=1}^{n} E\left(\mathbb{I}_{\{\xi_i\}}(x_t)\middle|o_{1:n},\boldsymbol{\theta}^*\right) = \sum_{t=1}^{n} p_{X_t|O_{1:n}}(\xi_i|o_{1:n},\boldsymbol{\theta}^*) \\
\hat{n}_{\xi_i,\xi_g} &:= E\left(N_{\xi_i,\xi_g}\middle|o_{1:n},\boldsymbol{\theta}^*\right) = \sum_{t=2}^{n} E\left(\mathbb{I}_{\{(\xi_i,\xi_g)\}}(x_{t-1},x_t)\middle|o_{1:n},\boldsymbol{\theta}^*\right) \\
&= \sum_{t=2}^{n} p_{X_{t-1:t}|O_{1:n}}(\xi_i,\xi_g|o_{1:n},\boldsymbol{\theta}^*).
\end{aligned}$$

We can calculate $\boldsymbol{\theta}^*$, and from this $p_{X_t|O_{1:n}}(\xi_i|o_{1:n},\boldsymbol{\theta}^*)$ and $p_{X_{t-1:t}|O_{1:n}}(\xi_i,\xi_g|o_{1:n},\boldsymbol{\theta}^*)$, efficiently with the Baum–Welch algorithm and forward and backward probabilities. Instead of using real counts of transitions and emissions, we use the expected number of occurrences of these events to build the imprecise local models. These expected numbers of occurrences are non-negative real numbers instead of non-negative integers. The estimated imprecise transition model is given by

$$\underline{Q}(\{\xi_g\}|\xi_i) = \frac{\hat{n}_{\xi_i,\xi_g}}{s+\sum_{g=1}^m \hat{n}_{\xi_i,\xi_g}} \text{ and } \overline{Q}(\{\xi_g\}|\xi_i) = \frac{s+\hat{n}_{\xi_i,\xi_g}}{s+\sum_{g=1}^m \hat{n}_{\xi_i,\xi_g}},$$

the estimated imprecise emission model by

$$\underline{S}(\{o\}|\xi_i) = \frac{\sum_{\{t:\omega_{h_t}=o\}} p_{X_t|O_{1:n}}(\xi_i|o_{1:n},\boldsymbol{\theta}^*)}{s+n_{\xi_i}}$$
$$\text{and } \overline{S}(\{o\}|\xi_i) = \frac{s+\sum_{\{t:\omega_{h_t}=o\}} p_{X_t|O_{1:n}}(\xi_i|o_{1:n},\boldsymbol{\theta}^*)}{s+n_{\xi_i}},$$

and the estimated imprecise marginal model by

$$\underline{Q}_1(\{\xi_i\}) = \frac{p_{X_1|O_{1:n}}(\xi_i|o_{1:n},\boldsymbol{\theta}^*)}{s+1} \text{ and } \overline{Q}_1(\{\xi_i\}) = \frac{s+p_{X_1|O_{1:n}}(\xi_i|o_{1:n},\boldsymbol{\theta}^*)}{s+1}.$$

3.4 Imprecision of the imprecise local uncertainty models

The imprecision $\Delta(Q(\{\xi_h\}|\xi_i))$ of the imprecise transition model and the imprecision $\Delta(S(\{o\}|\xi_i))$ of the imprecise emission model satisfy interesting properties.

Proposition 1. *The harmonic mean $H_{\Delta(Q)}$ of the set $\{\Delta(Q(\{\xi_h\}|\xi_i)) : i \in \{1,\ldots,m\}\}$ is given by $H_{\Delta(Q)} = {ms}/{ms+n-1}$ and the harmonic mean $H_{\Delta(S)}$ of the set $\{\Delta(S(\{o\}|\xi_i)) : i \in \{1,\ldots,m\}\}$ is given by $H_{\Delta(S)} = {ms}/{ms+n}$.*

Proof. The harmonic mean $H_{\Delta(Q)}$ of $\{\Delta(Q(\{\xi_h\}|\xi_i)) : i \in \{1,\ldots,m\}\}$ is given by

$$\begin{aligned} H_{\Delta(Q)} &= \frac{m}{\sum_{i=1}^m \frac{1}{\Delta(Q(\{\xi_h\}|\xi_i))}} = \frac{ms}{\sum_{i=1}^m \left(s+\sum_{g=1}^m \hat{n}_{\xi_i,\xi_g}\right)} \\ &= \frac{ms}{ms+\sum_{t=1}^{n-1}\sum_{i=1}^m E\left(\mathbb{I}_{\{\xi_i\}}(X_t)\middle|o_{1:n},\boldsymbol{\theta}^*\right)} = \frac{ms}{ms+\sum_{t=1}^{n-1} 1} = \frac{ms}{ms+n-1}. \end{aligned}$$

The harmonic mean $H_{\Delta(S)}$ of $\{\Delta(S(\{o\}|\xi_i)) : i \in \{1,\ldots,m\}\}$ is given by

$$H_{\Delta(S)} = \frac{m}{\sum_{i=1}^m \frac{1}{\Delta(S(\{o\}|\xi_i))}} = \frac{ms}{\sum_{i=1}^m \left(s+n_{\xi_i}\right)} = \frac{ms}{ms+n}. \qquad \square$$

$H_{\Delta(Q)}$ increases with m (if $n > 1$) and decreases with n, and $H_{\Delta(S)}$ increases with m and decreases with n. The IDM yields more precise estimates as the number of relevant observations (of transitions or emissions) increases: the more relevant data, the more

precise the estimates. For a fixed number of data (observation sequence length n), the precision tends to decrease as the number of possible state values m increases. Notably in cases where states are useful fictions (as in the earthquake example discussed further on), there is a cost to increasing the number of states. The increase of the imprecision with increasing m is, obviously, not present in precise HMM estimation. When making inferences based on precise HMM estimation, for example using the Viterbi algorithm for state sequence estimation, all results seem equally reliable, regardless of the number of possible state values m. But when making inferences in iHMMs, based on the model estimates provided by our method, for example using the EstiHMM algorithm [2], this is not the case: for smaller m, inferences will be more precise (or decisive); and if m is fairly large, inferences about state sequences will tend to become more imprecise. Lumping states together will increase the predictive power (for a given observation sequence), refining states will reduce it: there is a certain limit on what can be inferred using an iHMM estimated from a given information sequence, which is not there if we use a precise HMM estimation. Using precise HMM estimation, the coarseness of the state space representation has no influence on the precision, irrespective of the amount of data available. We believe this is a weakness rather than a strength of precise models.

4 The Baum–Welch algorithm

We give a brief overview of how to find the model parameter $\boldsymbol{\theta}^*$ using the Baum–Welch algorithm. It is an EM algorithm specifically for learning HMMs (see, e.g., [9]). It iteratively finds a (local) maximum $\boldsymbol{\theta}^*$ of the likelihood, which we define presently.

4.1 Likelihood in hidden Markov models

The complete likelihood $L_{o_{1:n},x_{1:n}}(\boldsymbol{\theta})$ in an HMM, with the observation sequence $O_{1:n} = o_{1:n}$ as data, an arbitrary state sequence $X_{1:n} = x_{1:n}$ and model parameter $\boldsymbol{\theta}$, is defined as $p_{O_{1:n},X_{1:n}}(o_{1:n},x_{1:n}|\boldsymbol{\theta})$. By the Markov condition, this can be written as $L_{o_{1:n},x_{1:n}}(\boldsymbol{\theta}) = p_{l_1}\prod_{t=2}^{n} T_{l_{t-1}l_t}\prod_{t=1}^{n} E_{h_t l_t}$. Although we are interested in the likelihood for the observation sequence $L_{o_{1:n}}(\boldsymbol{\theta}) \coloneqq p_{O_{1:n}}(o_{1:n}|\boldsymbol{\theta})$, the Baum–Welch algorithm finds a maximum $\boldsymbol{\theta}^*$ for the complete likelihood. Welch proves [8] that the Baum–Welch algorithm also locally maximises the likelihood for the observations.

A $\boldsymbol{\theta}^*$ that maximises $L_{o_{1:n},x_{1:n}}(\boldsymbol{\theta})$ also maximises $\ln L_{o_{1:n},x_{1:n}}(\boldsymbol{\theta})$, given by:

$$\ln L_{o_{1:n},x_{1:n}}(\boldsymbol{\theta}) = \sum_{z=1}^{m} \mathbb{I}_{\xi_z}(x_1)\ln p_z + \sum_{i=1}^{m}\sum_{g=1}^{m} n_{\xi_i,\xi_g}\ln T_{ig} + \sum_{t=1}^{n}\sum_{z_t=1}^{m} \mathbb{I}_{\xi_{z_t}}(x_t)\ln E_{h_t z_t}. \qquad (1)$$

The Baum–Welch algorithm consists in executing two steps—the *Expectation (E) step* and the *Maximisation (M) step*—iteratively until some convergence is achieved.

4.2 Expectation step

In the E step we calculate the expectation of the complete log-likelihood conditional on the observations $o_{1:n}$ (and of course the model parameter $\boldsymbol{\theta}$). We call this expecation $\ln \hat{L}_{o_{1:n}}(\boldsymbol{\theta}) \coloneqq E\,(\ln L_{o_{1:n},X_{1:n}}(\boldsymbol{\theta})|o_{1:n},\boldsymbol{\theta})$. It is given by the right-hand side of (1), but with the indicators and the n_{ξ_i,ξ_g} replaced by their expectations, as in Section 3.3.

4.3 Maximisation step

In this step we search the argument $\boldsymbol{\theta}^*$ that maximises the expectation of the complete log-likelihood.

Lemma 1. *The argument $\boldsymbol{\theta}^*$ that maximises the expected complete log-likelihood of a HMM with observation sequence $\omega_{h_1:h_n}$ is given by, for all $i,g \in \{1,\ldots,m\}$ and all $h \in \{1\ldots,p\}$:*

$$p_i^* = p_{X_1|o_{1:n}}(\xi_i|o_{1:n},\boldsymbol{\theta}^*), T_{ig}^* = \frac{\hat{n}_{\xi_i,\xi_g}}{\sum_{g=1}^m \hat{n}_{\xi_i,\xi_g}}, \text{ and } E_{hi}^* = \frac{\sum_{\{t:h_t=h\}} p_{X_t|o_{1:n}}(\xi_i|o_{1:n},\boldsymbol{\theta}^*)}{\hat{n}_{\xi_i}}.$$

By repeatedly performing the E step followed by the M step (with in the E step $\boldsymbol{\theta}$ taken as $\boldsymbol{\theta}^*$), we eventually reach a stable value of $\boldsymbol{\theta}^*$, guaranteed to be also a local maximum of the likelihood for the observation sequence.

Incidentally, Lemma 1 guarantees that our method, with the choice for the pseudo-counts $s = 0$, gives local models that maximise the likelihood for the observation sequence.

5 Predicting the Earth's earthquake rate

5.1 Introduction

We apply our method to a problem where we are interested in using HMMs to predict earthquake rates in future years. To do this, we will see that we need to learn a transition model. To this end, we use data of counted annual numbers of major earthquakes (with magnitude 7 and higher).

We assume that the earth can be in m different seismic states $\lambda_1,\ldots,\lambda_m$ and that the occurrence of earthquakes in a year depends on the seismic state λ of the Earth in that year. We assume that the Earth, being in a seismic state λ, "emits" a number of earthquakes o governed by a Poisson distribution with parameter λ: $p_{O|X}(o|\lambda) = {e^{-\lambda}\lambda^o}/{o!}$.

The data are (yearly) earthquake counts over 107 subsequent years, from 1900 to 2006. It is freely available on `http://neic.usgs.gov/neis/eqlists`.

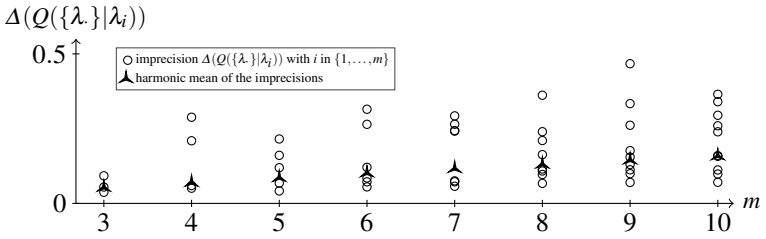
We model this problem as an iHMM of length 107, in which each observation variable O_i corresponds to one of the 107 yearly earthquake counts. The states correspond to the seismic states Earth can be in. The set of seismic states $\{\lambda_1,\ldots,\lambda_m\}$ defines the possibility space $\mathscr{X}$ of the state variables in the HMM.

5.2 Results

Imprecise transition model Since there is only 107 years of data, we believe that a precise local transition model is not justified, so we decided to try an imprecise estimation for the transition model. The emission model is kept precise for simplicity, due to its relation to a Poisson process.

To show how the imprecision changes with changing number of possible state values m, we plot the learned transition model for varying m. The figure below shows, as a

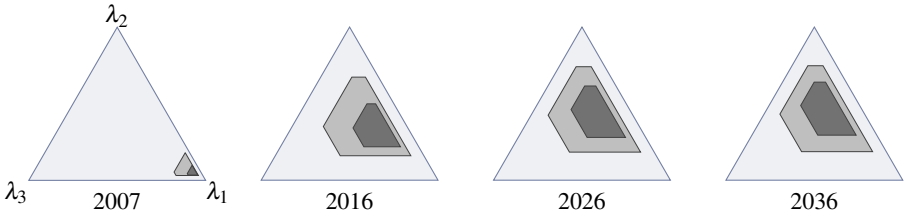
function of m (ranging from 3 to 10), the imprecision $\Delta(Q(\{\lambda_\cdot\}|\lambda_1)),\ldots,\Delta(Q(\{\lambda_\cdot\}|\lambda_m))$ of the transition probabilities of going from state λ_i to state $\lambda_\cdot$, for $s=2$ (this imprecision depends on the state λ_i, but not on the state $\lambda_\cdot$ the transition goes to).



The harmonic mean of the imprecisions increases with m, as predicted by Proposition 1.

Predicting the earthquake rate With the learned transition model (with $m=3$), we make predictions of the earthquake rate in future years. We do this in order to validate our learned model. We want to make inferences about the years 2007, 2016, 2026 and 2036: we are interested in the model describing the state variables of these years, updated using the observation sequence. We can use this updated model to get some idea of the future earthquake rate. To perform such updating, we can use the MePiCTIr algorithm [3].

The figure below shows conservative approximations (the smallest hexagons with vertices parallel with the borders of the simplex) of such updated models describing future state variables. In the dark grey credal sets, we have used the transition model estimates for $s=2$, and in the light grey ones the estimated transition models for $s=5$.



The precision of the inferences decreases as we move forward in time. For 2007, we can be fairly confident that the local seismic rate of the earth will be close to λ_1, while for 2036, we can only make very imprecise inferences about the seismic rate. This is a property that predictions with precise HMMs do not have.

6 Conclusion

We have presented a new method for learning imprecise local uncertainty models in stationary hidden Markov models. In contradistinction with the classical EM learning algorithm, our approach allows the local models to be imprecise, which is useful if there is insufficient data to warrant precision. We have studied some of the properties of our learned local models, especially with respect to their imprecision.

We conclude by proposing some avenues for further research. We have based the present discussion on the maximum likelihood approach of learning in Bayesian networks. The epistemic nature of imprecise probability theory however suggests that a Bayesian learning approach would be more appropriate, and we intend to investigate this in the near future.

Acknowledgements. Arthur Van Camp is a Ph.D. Fellow of the Ghent University Special Research Fund (BOF) and wishes to acknowledge its financial support. We are indebted to Jasper De Bock and Marco Zaffalon for suggesting and discussing the idea of using expected counts in an IDM, and for suggesting to use the Baum–Welch algorithm.

References

1. Alessandro Antonucci, Rocco de Rosa, and Alessandro Giusti. Action recognition by imprecise hidden Markov models. In *Proceedings of the 2011 International Conference on Image Processing, Computer Vision and Pattern Recognition (IPCV 2011)*, pages 474–478. CSREA Press, 2011.
2. Jasper De Bock and Gert de Cooman. State sequence prediction in imprecise hidden Markov models. In Frank Coolen, Gert de Cooman, Thomas Fetz, and Michael Oberguggenberger, editors, *ISIPTA'11: Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, pages 159–168, Innsbruck, 2011. SIPTA.
3. Gert de Cooman, Filip Hermans, Alessandro Antonucci, and Marco Zaffalon. Epistemic irrelevance in credal nets: The case of imprecise markov trees. *International Journal of Approximate Reasoning*, 51(9), 2010.
4. Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.
5. Arthur Van Camp, Gert de Cooman, Jasper De Bock, Erik Quaeghebeur, and Filip Hermans. Learning imprecise hidden Markov models, Poster at ISIPTA '11.
6. P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.
7. P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–57, 1996. With discussion.
8. L. Welch. Hidden Markov models and the Baum–Welch algorithm. *IEEE Information Theory Society Newsletter*, pages 1, 10 – 13, 2003.
9. W. Zucchini and I. L. MacDonald. *Hidden Markov models for time series: an introduction using R*. Chapman and Hall, 2009.