Selection of the number of states in hidden semi-Markov models

Filipa Rente
filipaoliveirarente@gmail.com
Instituto Superior Técnico (IST),
Universidade de Lisboa (ULisboa), Portugal

Zita Marinho
zam@priberam.pt
Instituto de Sistemas e Robótica, IST
Priberam Labs, Lisboa, Portugal

Mário Figueiredo
mario.figueiredo@tecnico.ulisboa.pt
Instituto de Telecomunicações,
IST, ULisboa, Portugal

Abstract

This paper addresses the order selection problem for Hidden semi-Markov models (HSMMs). We propose a novel approach to select the optimal number of states, as well as the state duration of an HSMM. One of the main contributions of this paper is a proof of an equivalence of models which allows us to focus only on the selection of the number of states. We propose a unique order selection criterion based on that proof. Furthermore, the optimal number of states is found through a sequential pruning strategy using a mixture minimum description length (MMDL) criterion, based on Figueiredo et al. [3], for mixture models, and Bicego et al. [2] for Hidden Markov models (HMMs). We demonstrate the effectiveness of the approach using synthetic experiments. Source code available at `https://github.com/filiparente/hsmm_mmdl`.

Keywords: Hidden semi-Markov models; Order selection problem; Minimum description length; Bayesian inference criterion; State pruning.

1 Introduction

HSMMs are a extension of Hidden Markov Models (HMMs) where the state duration is explicitly modelled. State duration distributions can be non-parametric (multinomial) or parametric. The duration in a k-state HMM is implicitly a geometric distribution with parameter $A_{ii}, \forall i \in [k]$, which corresponds to the diagonal entries of the transition probability matrix $A \in \mathbb{R}^{k\times k}$. This implicit modeling of state duration is not sufficient for many real problems. Consequently, for those problems a model such as HSMMs is more suitable to provide a higher expressiveness to the state duration modeling. One example include time series segmentation of human motion data into single actions, where the probability distribution used for the duration is extremely important in order to correctly cluster the different motion activities [4].

2 Hidden semi-Markov models

A hidden semi-Markov model is a probabilistic model in which a stochastic sequence of observations $\mathbf{O} = O_{1:T}$ is generated by an hidden random sequence $\mathbf{S} = S_{1:T}$, where O_t denotes the observed symbol at time t and S_t the state occupied by the system at time t. A k-state HSMM is completely defined by $\mu = (\mathbf{s}, A, \pi, B, D)$:

•A set $\mathbf{s} = \{s_1, ..., s_k\}$ of hidden states and a set $\{o_1, ..., o_m\}$ of observation symbols;

•A transition matrix $A \in \mathbb{R}^{k^2 \times {d_{max}}^2}$, whose element $A_{id',jd}$ represents a transition from state s_i with duration d' to state s_j with duration d,

$$A_{id',jd} = \mathbb{P}[S_{t+1} = s_j, \tau_{t+1} = d | S_t = s_i, \tau_t = d'], \tag{1}$$

where τ_t represents the residual time in state S_t and $i, j \in [k]$ and $d, d' \in [d_{max}]$. Here, the explicit duration HMM variant of HSMMs will be considered, which assumes that a state transition is not dependent on the duration of the previous state. Thus, eq. (1) simplifies to

$$A_{id',jd} = A_{i,jd} = A_{ij} p_j(d), \tag{2}$$

where A_{ij} is the transition probability matrix as in HMMs and $p_j(d)$ is the explicit state duration probability;

•An emission matrix $B \in \mathbb{R}^{k\times m}$ which represents

$$B_{ij} = \mathbb{P}[O_t = o_j | S_t = s_i] \tag{3}$$

the probability of observing symbol o_j while in state s_i. Here, we opted for a parametric Poisson emission distribution. That is,

$$\mathbb{P}[o_j | s_i] = \frac{\lambda_i^{o_j} e^{-\lambda_i}}{o_j!}, \tag{4}$$

where λ_i is the Poisson parameter associated with state s_i;

•An initial state probability $\pi \in \mathbb{R}^k$ with elements $\pi_i = \mathbb{P}[S_1 = s_i]$;

•A duration probability matrix $D \in \mathbb{R}^{k\times d_{max}}$ where $D_{jd} = p_j(d)$ is the probability of occupying state s_j with duration d. Here, we also considered a parametric distribution as a mixture of geometric distributions,

$$p_j(d) = \sum_{m=1}^{M_j} c_{jm} \text{Geom}(d|\theta_{jm}), \tag{5}$$

where $\text{Geom}(d|\theta_{jm})$ denotes a geometric density with parameter θ_{jm} for state s_j and component m. The durations in state s_j are modelled as samples from a geometric mixture with M_j components, with c_{jm} being the mixing weight of the m^{th} component in state s_j. One of the main contributions of this paper is a proof that an HSMM where the state duration distribution is a mixture of geometric distributions (for each state) is equivalent to another HSMM with only one geometric distribution $p_{jm}(d)$ per state s_{jm}, but more states. That is, the set of states $\mathbf{s}$ is augmented to $\mathbf{s}'$ in this equivalent model.[1] Using this result, we will consider only HSMMs with one geometric duration distribution for each state. That is,

$$p_{jm}(d) = \text{Geom}(d|\theta_{jm}) = (1-\theta_{jm})^{d-1}\theta_{jm}, \qquad 1 \le d \le d_{max}. \tag{6}$$

•Hyperparameters: the number of states k, and the maximum allowed duration d_{max}. d_{max} is used to truncate the forward-backward pass in the EM algorithm, and it represents the maximum number of observations within all states. The number of states is estimated using the sequential pruning strategy with the MMDL criterion, described in Section 3.

3 Order estimation for HSMMs

The order estimation problem in HSMMs consists of estimating the optimal number of states (k^*) and maximum allowed state duration (${d_{max}}^*$). Hence, the order estimation procedure is a multivariate optimization problem where we optimize both k and d_{max} w.r.t. the same criterion. One greedy approach would be to try every possible combination of k and d_{max} and assign the one with the highest model likelihood. Ideally, it would be better to guarantee that each state is connected to the set of possible durations $\{1, 2, ..., d_{max}\}$ in a one-by-one relationship. This result would allow the problem to be simplified to a univariate optimization problem, where the focus is only on the selection of k^*. If we consider a model with high statistical complexity, e.g., with non-parametric (multinomial) state duration distributions, then this one-by-one relationship is not present, because each state needs d_{max} "parameters" to specify its duration. Likewise, one can consider a model with lower statistical complexity, i.e., with parametric state duration distributions, where, in analogy with the multinomial case, each state has a mixture of parametric distributions for its duration probability (e.g. a mixture of geometric distributions). Nonetheless, the problem remains since each state has multiple parameters (as many as the number of components of the mixture of that state) to specify its duration. Fortunately, the equivalence of models (proof omitted due to space constraints) allows us to consider instead an equivalent model with more states (augmented states) but only one geometric distribution per state. Consequently, each state has one and only one parameter to specify its duration (the parameter of the geometric distribution), therefore guaranteeing the one-by-one relationship required. Using this equivalent model, d_{max} can be empirically estimated through the parameters of the geometric distributions. In fact, there is a correspondence between the parameter θ_{jm} of the geometric distribution (of state s_{jm}) and the durations d: a lower parameter is associated with higher durations. Therefore, d_{max} can be estimated through the following steps: (1) find the lowest parameter θ_{jm} between all states $\mathbf{s}'$; (2) find the duration for which the cumulative probability (easily derived from the PMF of eq. (6)), with parameter found in (1), is smaller than ε.[2] After determining the hyperparameter d_{max}, the state duration probability matrix $D \in \mathbb{R}^{k\times d_{max}}$ (used in the EM algorithm) can be filled with values from the geometric PMF (eq. (6)) evaluated at the corresponding durations $d = 1, ..., d_{max}$.

In the following sections, we will discuss the adaptation of two different selection criteria to the HSMM setting: Bayesian inference criterion (BIC) and mixture minimum description length (MMDL).

[1] $\mathbf{s}' = \{s'_1, ..., s'_{k'}\} = \{s_{11}, ..., s_{1M_1}, ..., s_{jm}, ..., s_{k1}, ..., s_{kM_k}\}$, where $k' = \sum_{j=1}^{k} M_j$ is the (augmented) number of states of the equivalent model.

[2] A reasonable value for ε is of the order of $0,01$.

3.1 Bayesian inference criterion (BIC)

The BIC criterion for k states is generally defined as

$$\mathrm{BIC}(k) = \log \mathbb{P}[\mathbf{O}|\hat{\mu}_k] - \frac{N_k}{2}\log(n), \quad (7)$$

where the first term is the log-likelihood of the observations ($\mathbf{O}$) given the maximum likelihood estimated parameters by the EM algorithm ($\hat{\mu}_k$), hereinafter referred to as LL. The second term penalizes the model complexity through the total number of free parameters N_k weighted by the total number of observations n. The optimal number of states is the one that maximizes the criterion in eq. (7), i.e., $\hat{k}_{\mathrm{BIC}} = \arg\max_k \mathrm{BIC}(k)$. For HSMMs, the BIC criterion is the one in eq. (7) with N_k decomposed in terms of the number of free parameters for each variable ${N_k}^A + {N_k}^\pi + {N_k}^B + {N_k}^D$. The number of free parameters for the transition matrix is ${N_k}^A = k(k-2)$ since: (1) it is a stochastic matrix and (2) the diagonal entries are zero, $A_{ii} = 0$, because in HSMMs no self-transitions are allowed due to the explicit specification of state durations. For the initial state distribution, ${N_k}^\pi = k-1$ and, finally, for the emission and state duration densities, since both are parametric distributions, the total number of free parameters for each is k, one per state. To sum up, the general BIC criterion defined in eq. (7) is extended to

$$\mathrm{BIC}_{\mathrm{HSMM}}(k) = \mathrm{LL} - \frac{k^2+k}{2}\log(n), \quad (8)$$

where we dropped all terms that do not depend on the number of states k.

3.2 Mixture minimum description length (MMDL)

This criterion, as the name suggests, is inspired by the Gaussian mixture models (GMMs), where the model complexity penalty term does not need to consider the whole data ($log(n)/2$) given that each component of the mixture is not estimated from all observations but only from those that were in fact generated by that component. Hence, the term $log(n)/2$ is replaced by a quantity that measures how much data was generated by a given component. For GMMs, this quantity is easily obtained as $log(nc_j)/2$, with c_j being the probability of the j^{th} component of the mixture. For HSMMs, finding this quantity is not so trivial. However, both the transition matrix and the initial state distribution are estimated from the whole data, so those are still weighted by the standard $log(n)/2$ term. On the contrary, the emission probability parameters (λ) and the state duration probability parameters (θ) are estimated using only the samples from the corresponding component. The quantity that specifies the state probabilities is given by the stationary distribution $p_\infty = [p_\infty(1), ..., p_\infty(k)]$[3], which represents the "average" occupation of each (steady) state after the semi-Markov chain has reached its equilibrium. Therefore, for an HSMM with k states, the MMDL criterion is given by

$$\mathrm{MMDL}_{\mathrm{HSMM}}(k) = \mathrm{LL} - \frac{{N_k}^A + {N_k}^\pi}{2}\log(n) - \frac{{N_1}^B + {N_1}^D}{2}\sum_{m=1}^{k} log(np_\infty(m)), \quad (9)$$

where ${N_1}^B$ and ${N_1}^D$ are the number of free parameters of the emission and state duration densities, respectively, of an HSMM with only one state. Since we assumed parametric distributions, ${N_1}^B = {N_1}^D = 1$. Replacing the free parameters in eq. (9), yields

$$\mathrm{MMDL}_{\mathrm{HSMM}}(k) = \mathrm{LL} - \frac{k^2-k}{2}\log(n) - \sum_{m=1}^{k} log(np_\infty(m)). \quad (10)$$

3.3 Stationary probability distribution for HSMMs

The stationary probability distribution for semi-Markov chains is not the same as for regular Markov chains. However, it can be found empirically. In [1], the authors proposed the following estimator

$$p_\infty(i) = \frac{\hat{\nu}_i(T)\hat{m}_i(T)}{\sum_{i=1}^{k}\hat{\nu}_i(T)\hat{m}_i(T)}, \quad \hat{\nu}_i(T) = \frac{N_i(T)}{N(T)}, \quad \forall i \in [k], \quad (11)$$

where $\hat{\nu}_i(T)$ is an estimator for the stationary probability distribution of the embedded[4] Markov chain. It is obtained from the ratio of the number of visits of the embedded Markov chain to state s_i up to time T ($N_i(T)$) and number of jumps of the embedded Markov chain in the time interval $(0,T]$ ($N(T)$), both available from the estimated state sequence after running the *Viterbi* algorithm. In eq. (11), $\hat{m}_i(T)$ is an estimator of the mean sojourn time in state i, considering all observations up to time T. It is obtained by averaging the EM estimate for the state duration probability matrix $\hat{D}$, weighted by the corresponding durations. As desired, the stationary probability distribution for semi-Markov chains p_∞, given by eq. (11), is a measure of the "average" state occupation.

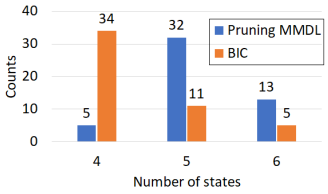
3.4 Sequential pruning strategy

The sequential pruning strategy entails the following steps:

1. Set the minimum and maximum allowed number of states: k_{min} and k_{max}, respectively;
2. Initialize the EM algorithm parameters randomly, using $k = k_{max}$ as the initial number of states;
3. While $k \geq k_{min}$: Run the EM algorithm until convergence and obtain the estimated parameters; Compute and store the value of the model selection criterion given by eq. (10); Find the least probable state (i.e., the smallest element of p_∞); Prune it by removing the corresponding rows/columns from the estimated parameters; Normalize the estimated parameters that after pruning need normalization[5]; Set these as the initial parameters of the EM algorithm and set $k = k-1$. Repeat, i.e., run EM again with one less state.
4. The final model is the one which corresponds to the maximum stored value for the model selection criterion, with the attached optimal number of states k_*.

4 Results with synthetic data

The synthetic data corresponds to 2 sequences of 1000 observations each sampled from an HSMM with random parameters μ and 5 states. According to the results shown in Fig.1, the standard BIC retrieves an expected number of states 4.42 on average over 50 runs, while the pruning MMDL criterion estimates a closer expected number of states 5.16. Furthermore, the pruning MMDL achieves a better performance in about 7 times less EM iterations. Besides the number of iterations, we can also report the elapsed time of the EM algorithm, which was 25.46 ± 22.56 seconds for the standard BIC criterion and 3.83 ± 1.91 seconds for the pruning MMDL criterion, again about 7 times less time consumed.



Method	N° states	N° iter.
BIC	4.4 ± 0.7	38.6 ± 34.2
MMDL	5.2 ± 0.6	5.8 ± 2.9

Figure 1: Left: Histograms of the selected number of states for the standard BIC and pruning MMDL strategies for a total of 50 runs of the algorithm when the true underlying data is generated from 5 states; Right: Expected number of estimated states using a BIC and MMDL criterion respectively.

5 Conclusions and future work

Experiments with synthetic data show that the pruning MMDL method for HSMMs provides a higher accuracy in the selection of the optimal number of states, when compared with BIC, and simultaneously a less demanding computational requirement. Future work concerns a more solid comparison with other methods such as the Akaike information criterion (AIC) and the method proposed in Yu [5]. Another important contribution would be to demonstrate the effectiveness in a more substantial application study, for example by using real data or by simulating synthetic data with some particularities that affect the learning process (e.g. observe the difference between parametric and non-parametric distributions, both in the state emissions and durations).

References

[1] Vlad Barbu, Jan Bulla, and Antonello Maruotti. Estimation of the stationary distribution of a semi-markov chain. *Journal of Reliability and Statistical Studies*, 5:15–26, 2012.

[2] Manuele Bicego, Vittorio Murino, and Mário AT Figueiredo. A sequential pruning strategy for the selection of the number of states in hidden markov models. *Pattern Recognition Letters*, 24(9-10):1395–1407, 2003.

[3] Mário AT Figueiredo, José MN Leitão, and Anil K Jain. On fitting mixture models. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 54–69. Springer, 1999.

[4] Tomoaki Nakamura, Takayuki Nagai, Daichi Mochihashi, Ichiro Kobayashi, Hideki Asoh, and Masahide Kaneko. Segmenting continuous motions with hidden semi-markov models and gaussian processes. *Frontiers in neurorobotics*, 11:67, 2017.

[5] Shun-Zheng Yu. *Hidden Semi-Markov models: theory, algorithms and applications*. Morgan Kaufmann, 2015.

[3] For HMMs this quantity is obtained from the left eigenvector of the transition matrix A associated with eigenvalue 1.

[4] The embedded Markov chain of a semi-Markov chain with state sequence $\mathbf{Q} = \{4,4,1,1,1,2,3,3,3\}$ is $\mathbf{Q'} = \{4,1,2,3\}$ with the corresponding jump times $\mathbf{J'} = \{2,5,6\}$.

[5] Set distributions to uniform if normalization is not possible (this problem can arise when considering a large amount of states).