Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2017, Article ID 7318940, 26 pages
https://doi.org/10.1155/2017/7318940

Research Article

Efficient and Effective Learning of HMMs Based on Identification of Hidden States

Tingting Liu[1] and Jan Lemeire[1,2,3]

[1]*Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussels, Belgium*
[2]*Department of Industrial Sciences (INDI), Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussels, Belgium*
[3]*Department of Data Science, iMinds, Technologiepark 19, 9052 Zwijnaarde, Belgium*

Correspondence should be addressed to Tingting Liu; tliu@etrovub.be

Received 24 July 2016; Revised 21 December 2016; Accepted 29 December 2016; Published 23 February 2017

Academic Editor: Leonid Shaikhet

The predominant learning algorithm for Hidden Markov Models (HMMs) is local search heuristics, of which the Baum-Welch (BW) algorithm is mostly used. It is an iterative learning procedure starting with a predefined size of state spaces and randomly chosen initial parameters. However, wrongly chosen initial parameters may cause the risk of falling into a local optimum and a low convergence speed. To overcome these drawbacks, we propose to use a more suitable model initialization approach, a Segmentation-Clustering and Transient analysis (SCT) framework, to estimate the number of states and model parameters directly from the input data. Based on an analysis of the information flow through HMMs, we demystify the structure of models and show that high-impact states are directly identifiable from the properties of observation sequences. States having a high impact on the log-likelihood make HMMs highly specific. Experimental results show that even though the identification accuracy drops to 87.9% when random models are considered, the SCT method is around 50 to 260 times faster than the BW algorithm with 100% correct identification for highly specific models whose specificity is greater than 0.06.

1. Introduction

Hidden Markov Models (HMMs) [1] are one of the statistical modelling tools showing great success and have been widely used in diverse application fields such as speech processing [2], machine maintenance [3], acoustics [4], biosciences [5], handwriting and text recognition [6], and image processing [7]. Despite the merit of simplicity and learning capabilities, HMMs are still facing open problems such as learning effectiveness and efficiency.

There are two major problems in HMM learning: (1) choosing model size (number of hidden states); (2) estimating model parameters. Regarding the first problem, state-of-the-art approaches normally train multiple HMMs with different numbers of states and the best one is selected using specific criteria (e.g., the Akaike information criterion (AIC) [8], the Bayesian Information Criterion (BIC) [9]). In order to tackle the second problem, traditional learning algorithms such as the Baum-Welch (BW) algorithm are used to iteratively optimize model parameters starting from a, most often randomly chosen, initial set of parameters. Such iterative optimization heuristic approaches are prone to local optima. Therefore, multiple runs (typically, 10 [10, 11] or 20 [12, 13]) with several different initializations are performed and the optimal one of these is chosen. However, such iterative approaches with multiple trainings have significant drawbacks of time inefficiency and a high computational cost. Hsu et al. [14] introduced a noniterative method employing spectral-based algorithm for learning HMMs. It is simple and employs only a singular value decomposition and matrix multiplications. Nonetheless, it is evaluated in [15] and shown to be only applicable to identify systems when relatively few observations are available but fail completely for systems when the available observations are large. Fox et al. [8] proposed a sticky HDP-HMM which is a nonparametric, infinite-state model that automatically learns the size of state spaces and the smoothly varying dynamics robustly. However, this approach is computationally prohibitive when datasets are very large [9]. Therefore, in spite of the limitations, classical iterative approaches are still widely used to estimate model size and model parameters, for lack of alternatives.

The aim of this paper is to improve the effectiveness and efficiency in model learning compared to the conventional BW algorithm, in the sense of accurately and quickly finding the correct model. One of the HMM assumptions is that the observed data is only dependent on the hidden states given the model. Therefore, the observed data often reflects the structure and statistical properties of the model, which motivates us to introduce a data-driven preestimation procedure to estimate the number of states and choose proper initial model parameters.

We firstly provide insight into the essential features of an HMM model that help to improve the model's expressiveness as a stochastic process [16]. This is conducted by inspecting the role of each hidden state in generating observation distributions as well as providing information on the model structure. Hidden states with a large influence on observation sequences increase the value of a model more than those without or with low influence. By analysing how the information flows through the HMMs, we determine which cases make a state have a high impact. As discussed in Section 3, persistent and/or transient-cyclic states appear to be high-impact states. Moreover, a model with high-impact states is highly *specific* and will be easy to identify. We introduce the term *specificity* as the minimum model distance between a model and the best of HMMs with one state less. On the contrary, some HMMs are in principle unidentifiable which has been proved in [17] by linking the learning of HMMs to the nonlearnability results of finite automata. Furthermore, there are models in between the learnable and the unlearnable HMMs, which are hard to learn from observation sequences. Such HMMs contain complex parameter configurations with low specificity and low-impact states. Overall, experimental results show that a better number of states and proper initialization learned by the proposed method increase the learning speed and accuracy of highly specific HMMs compared to the traditional Baum-Welch algorithm.

The remainder of the paper is organized as follows: in Section 2, the preliminaries about HMMs and the Baum-Welch learning problems are briefly reviewed, followed by the concepts and definitions of model characteristics such as model identifiability, model equivalence, and the minimality of models. In Section 3, the impact of states on model specificity is studied through the information analysis. Followed by the approximate identification framework in Section 4, experiments and results are discussed in Section 5. Finally, conclusions are given in Section 6.

2. Preliminaries

An HMM [1] is a doubly stochastic process where the underlying process is characterized by a Markov chain and unobservable (hidden) but can be observed through another stochastic process which emits the sequence of observations. Let Q denote the number of states and M the number of observation symbols. Let $S = \{s_1, s_2, \ldots, s_Q\}$ and $V = \{v_1, v_2, \ldots, v_M\}$ denote the set of states and the set of observations, respectively. Using q_t and o_t to represent the state and the emitted observation at time t, respectively, the state and observation sequences are denoted by vectors $\mathbf{q}_{1:t} = q_1, q_2, \ldots, q_t$ and $\mathbf{o}_{1:t} = o_1, o_2, \ldots, o_t$, where $1 \le t \le T$, and T is the number of states or observations in the sequence. A discrete time HMM model can be characterized by the quintuple $\lambda = (\boldsymbol{\pi}, Q, M, \mathbf{A}, \mathbf{B})$ [1]: the initial state probability distribution is a column vector $\boldsymbol{\pi} = \{\pi_i\}$, where the ith element is

$$\pi_i = P\left(q_1 = s_i\right) \quad 1 \le i \le Q \tag{1}$$

the state transition probability distribution matrix is $\mathbf{A} = \{a_{ij}\} \in \mathbb{R}_+^{Q\times Q}$, where the i, jth element is

$$a_{ij} = P\left(q_{t+1} = s_j \mid q_t = s_i\right), \quad 1 \le i, j \le Q, \tag{2}$$

and the observation probability distribution matrix is $\mathbf{B} = \{b_{ik}\} \in \mathbb{R}_+^{Q\times M}$, where the i, kth element is

$$b_{ik} = P\left(o_t = v_k \mid q_t = s_i\right), \quad 1 \le i \le Q,\ 1 \le k \le M. \tag{3}$$

To note that the state transition probabilities of state s_r include both *incoming* and *outgoing* probabilities, the *incoming* state transition probabilities of s_r are the rth column vector of $\mathbf{A}$, denoted as

$$\mathbf{a}_{:r} = \{a_{1r}, a_{2r}, \ldots, a_{Qr}\} = \{a_{ir}\}_{i=1}^{Q} \in \mathbb{R}_+^{Q\times 1} \tag{4}$$

and the *outgoing* state transition probabilities of s_r is the rth row vector of $\mathbf{A}$, denoted as

$$\mathbf{a}_{r:} = \{a_{r1}, a_{r2}, \ldots, a_{rQ}\} = \{a_{ri}\}_{i=1}^{Q} \in \mathbb{R}_+^{1\times Q}, \tag{5}$$

where $1 \le r \le Q$ and $\mathbb{R}_+$ represents the set of nonnegative real numbers.

2.1. The Baum-Welch Learning Algorithm. One of the three basic problems for HMMs is the learning problem [1], which is often solved by an Expectation-Maximization (EM) algorithm [18], named the Baum-Welch algorithm [19, 20]. Starting with an initial guess of the model λ_0 at random, the model parameters are iteratively reestimated as long as the new model has an increased likelihood compared to the previous one; that is, $P_{\widetilde{\lambda}}(\mathbf{o}_{1:T}) \ge P_\lambda(\mathbf{o}_{1:T})$, where $P_\lambda(\mathbf{o}_{1:T})$ and $P_{\widetilde{\lambda}}(\mathbf{o}_{1:T})$ represent the likelihood values of an observation sequence $\mathbf{o}_{1:T}$ generated by the previous model λ and the newly updated model $\widetilde{\lambda}$, respectively. This procedure continues until the likelihood converges to a stationary point. However, the BW algorithm suffers from the problem of getting stuck at a local optimum if the initial model parameters are not well chosen, which inspires this study to search for a better estimation of the initial parameters.

For the analysis, we need to calculate the likelihood of observations given the model, that is, $P_\lambda(\mathbf{o}_{1:t})$. It can be written by the use of the projection operations; see, for instance, [16, p. 18]. Let $\boldsymbol{\alpha}_t = \{\alpha_t(i)\}_{i=1}^{Q} \in \mathbb{R}_+^{Q\times 1}$ and $\boldsymbol{\tau}_t = \{\tau_t(i)\}_{i=1}^{Q} \in \mathbb{R}_+^{Q\times 1}$, where

$$\tau_t(i) = \begin{cases} P_\lambda\left(q_t = s_i\right) & t = 1 \\ P_\lambda\left(q_t = s_i, \mathbf{o}_{1:t-1}\right) & 1 < t \le T, \end{cases} \tag{6}$$

$$\alpha_t(i) = P_\lambda\left(q_t = s_i, \mathbf{o}_{1:t}\right) \quad 1 \le t \le T;$$

thus

$$\begin{aligned} \boldsymbol{\alpha}_t &= \mathbf{B}_{o_t}\boldsymbol{\tau}_t, \\ \boldsymbol{\tau}_{t+1} &= \mathbf{A}\boldsymbol{\alpha}_t, \end{aligned} \tag{7}$$

where $o_t = v_k \in V$, $1 \le t \le T$ and $\mathbf{B}_{o_t} = \mathrm{diag}\{b_{1k}, b_{2k}, \ldots, b_{Qk}\} \in \mathbb{R}_+^{Q\times Q}$ which denotes the diagonal matrix of which the diagonal elements are the kth column of $\mathbf{B}$.

Therefore, the *likelihood* of the observations given the model can be expressed as

$$P_\lambda(\mathbf{o}_{1:t}) = P_\lambda(o_1)\prod_{t=2}^{T} P_\lambda(o_t \mid \mathbf{o}_{1:t-1}) = \boldsymbol{\tau}_t^{\mathrm{T}}\mathbf{B}_{o_t}\mathbf{e}, \tag{8}$$

where $\mathbf{e}$ is a column vector of length Q with all entries equal to 1; that is, $\mathbf{e} = [1\ 1\ \cdots\ 1]^{\mathrm{T}} = \mathbf{1}^{Q\times 1}$. For the convenience of calculations, the logarithm of likelihood *log-likelihood* (LL) is often used rather than the likelihood. Moreover, in this dissertation, we use *unit log-likelihood*, an averaged LL, to present the LL per single observation, that is, $(1/T)\log P_\lambda(\mathbf{o}_{1:T})$, where T is the number of observations. Within this paper, the term *log-likelihood* is used to represent *unit log-likelihood* for simplicity.

2.2. Definitions of Model Characteristics. In this paper, we determine the learnability of HMMs through model identifiability. If two models are equivalent, the true model cannot be uniquely identified. Hence we firstly introduce the definition for model equivalence. Note that the HMM learning can be considered as a probability distribution specific problem, where every HMM has to be identified from the observations generated according to its own likelihood distribution. Therefore, the equivalence of HMMs can be defined based on their observation likelihood distributions as follows.

Definition 1 (HMM equivalence). Two HMM models λ and $\widetilde{\lambda}$ are *equivalent* if and only if both models have the same observation emission probabilities (i.e., likelihood distribution over time series) for every observation sequence $\mathbf{o}_{1:t}$

$$P_\lambda(\mathbf{o}_{1:t}) = P_{\widetilde{\lambda}}(\mathbf{o}_{1:t}); \tag{9}$$

alternatively,

$$\begin{aligned} P_\lambda(o_1) &= P_{\widetilde{\lambda}}(o_1) \\ P_\lambda(o_t \mid \mathbf{o}_{1:t-1}) &= P_{\widetilde{\lambda}}(o_t \mid \mathbf{o}_{1:t-1}) \quad \forall t,\ 2 < t \le T. \end{aligned} \tag{10}$$

Note that the observation probabilities $P_\lambda(\mathbf{o}_{1:t})$ can remain the same by permuting the states of λ since the states can be arbitrarily labeled. The model $\widetilde{\lambda}$ with permuted states is called a *trivial equivalent* model of the original model λ as defined in [21]. We consider *trivial equivalent* models as the same model. In order to compare the models in later sections, we need to label states in a unique way such that corresponding states receive the same label. Therefore we define a process to normalize HMMs as follows.

Definition 2 (HMM normalization). For each state s_i, a score is calculated by $\varpi = \sum_{k=1}^{M} b_{ik}k$. Based on the score, we sort the states in ascending order.

Additionally, we can always construct an equivalent HMM with additional state numbers [22]; hence, in this paper, we consider HMM identifiability only when it is *minimal*, as defined below.

Definition 3 (HMM minimality). An HMM $\lambda = (\boldsymbol{\pi}, Q, M, \mathbf{A}, \mathbf{B})$ is *minimal* if and only if it has equal number of states to or fewer number of states than any other equivalent model $\widetilde{\lambda} = (\widetilde{\boldsymbol{\pi}}, \widetilde{Q}, \widetilde{M}, \widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$; that is, $Q \le \widetilde{Q}$. Model λ is called a *simpler* model of $\widetilde{\lambda}$ if they are equivalent and $Q < \widetilde{Q}$.

Definition 4 (HMM identifiability). An HMM λ is *identifiable* if and only if it is minimal and there does not exist any *nontrivially equivalent* model $\widetilde{\lambda}$ with an equal number of states; that is, $Q \ne \widetilde{Q}$.

Moreover, in this study we only address the identification of *stationary* (or *homogeneous*) HMMs where the prior probabilities can be eliminated in calculations. The initial state prior probability distribution $\boldsymbol{\pi}$ has an influence on learning only at the beginning of an observation sequence and its impact on large sequences vanishes over time and thus can be excluded for learning HMMs in practice. A stationary HMM is defined as follows.

Definition 5 (stationary HMM). An HMM is *stationary* if its state distribution remains the same at every time instant; that is, $\boldsymbol{\pi}(1) = \boldsymbol{\pi}(2) = \cdots = \boldsymbol{\pi}(t) = \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ equals the equilibrium state distribution; that is, $\mathbf{A}\mathbf{B}_{o_t}\boldsymbol{\pi} = \boldsymbol{\pi}$ [23, p. 4902].

The element $\boldsymbol{\pi}(t)$ is a column vector with $\boldsymbol{\pi}(t) = \{P(q_t = s_i)\}_{i=1}^{Q} \in \mathbb{R}_+^{Q\times 1}$, $1 \le t \le T$, and $\mathbf{e}^{\mathrm{T}}\boldsymbol{\pi}(t) = 1$. The element $\mathbf{A}\mathbf{B}_{o_t}$ represents the probability of going from state s_i to state s_j while emitting the observation v_k by state s_i, that is, $P(q_{t+1} = s_j,\ o_t = v_k \mid q_t = s_i)$.

Our proposed learning approach is based on the properties of observation sequences that make a state have a large impact on the model. To describe the degree of influence that a state can make on a model, we define a new term called *specificity* $\mathcal{S}(\lambda)$ as the *distance* between model λ and the *best* model with one state less. By *best*, we mean that it matches the most on observations generated by the original model λ among all the one-state-fewer models, which also means that it has the minimum model distance to the original model λ. A general definition of *model distance* is as follows.

Definition 6 (HMM model distance). A model distance between two HMMs λ_1 and λ_2 is the difference of the unit log-likelihood of an observation sequence $\mathbf{o}_{1:T}^{\lambda_2}$ [1, p. 271]:

$$\mathcal{D}(\lambda_1, \lambda_2) = \mathbb{E}\left[\frac{1}{T}\left(\log P_{\lambda_2}\left(\mathbf{o}_{1:T}^{\lambda_2}\right) - \log P_{\lambda_1}\left(\mathbf{o}_{1:T}^{\lambda_2}\right)\right)\right], \tag{11}$$

where $\mathbb{E}[\cdot]$ refers to the expectation operator, $\mathbf{o}_{1:T}^{\lambda_2}$ is an observation sequence generated by model λ_2, and T is the size of the sequence. Equation (11) is basically a measure of how well model λ_1 matches observations generated by model λ_2, in comparison with how well model λ_2 matches observations generated by itself [1]. The *specificity* of a model can be then defined as follows.

Definition 7 (HMM specificity). The *specificity* of an HMM λ with Q states is

$$\mathcal{S}(\lambda) = \min_{\widetilde{\lambda}\in\Lambda(Q-1)} \left(\mathbb{E}\left[\frac{1}{T} \left(\log P_{\lambda}\left(\mathbf{o}_{1:T}^{\lambda}\right) - \log P_{\widetilde{\lambda}}\left(\mathbf{o}_{1:T}^{\lambda}\right) \right) \right] \right), \tag{12}$$

where $\Lambda(Q-1)$ represents the set of all HMMs with $Q-1$ states and T is the length of an observation sequence $\mathbf{o}_{1:T}^{\lambda}$ generated by λ. We denote the optimal model $\widetilde{\lambda}$ with the minimum distance to model λ in (12) as $\overline{\lambda}_{Q-1}(\lambda_Q)$.

We have to note that, to use Definitions 6 and 7 in practice, we will calculate the expectation with a single generated observation sequence. We assume that this sequence is long enough such that it is a typical sequence and gives a stable value which comes close to the expected value and as such is independent of the exact sequence, as is done by Rabiner [1].

To use the above definitions on a limited set of observation sequences, we have to rely on an approximate equivalence approach. In order to compare the HMMs according to the likelihood probability $P_{\lambda}(\mathbf{o}_{1:T})$ given a set of observation sequences $\mathbf{o}_{1:T}$, we have to define a threshold on the model distance to decide whether two HMM models are equivalent or not.

Definition 8 (distance threshold of equivalent HMMs). The *distance threshold* is defined as

$$(-3\sigma, +3\sigma), \tag{13}$$

where $\mathcal{N}(\mu, \sigma^2)$ is the asympototic distribution of log-likelihood $(1/T)\log P_{\lambda}(\mathbf{o}_{1:T}^{(\lambda,i)})$ with $(T \to \infty)$, the element $\mathbf{o}_{1:T}^{(\lambda,i)}$, $(i = 1, 2, \ldots, n)$ represents randomly generated sequences by model λ, T is the length of an observation sequence, and n is the total number of observation sequences [24]. Duan et al. [24] prove that the distribution of the log-likelihood $(1/T)\log P_{\lambda}(\mathbf{o}_{1:T}^{(\lambda,i)})$ can be approximated by a normal distribution $\mathcal{N}(\mu, \sigma^2)$. According to the "three-sigma" rule, the interval $(\mu - 3\sigma, \mu + 3\sigma)$ contains 99% of the whole distribution. Thus a sequence $\mathbf{o}_{1:T}^{(\lambda,i)}$ has a 99% certainty of being generated by the model $\widetilde{\lambda}$ if its log-likelihood $(1/T)\log P_{\widetilde{\lambda}}(\mathbf{o}_{1:T}^{(\lambda,i)}) \in (\mu - 3\sigma, \mu + 3\sigma)$, $\forall i$. As defined in Definition 1, two models are equivalent if and only if both models have the same likelihood distribution on observations. Hence for any sequence $\mathbf{o}_{1:T}^{(\lambda,i)}$ generated by model λ, if $\widetilde{\lambda}$ has a log-likelihood within the interval, that is, $(1/T)\log P_{\widetilde{\lambda}}(\mathbf{o}_{1:T}^{(\lambda,i)}) \in (\mu - 3\sigma, \mu + 3\sigma)$, $\forall i$, we can say the two models are approximately equal. Therefore, the model distance threshold of equivalence is approximated as $(-3\sigma, +3\sigma)$ of the reference model for practical use.

As defined in Definition 3, a model is *minimal* if and only if it has equal number of states to or fewer number of states than any other equivalent models. In order to check model minimality in practice, we verify if there exists no one-state simpler model $\widetilde{\lambda}$ which is equivalent to model λ, in particular, to verify if the minimum distance between $\widetilde{\lambda}$ and λ (i.e., the specificity of $\widetilde{\lambda}$; see Definition 7) is outside the threshold of equivalent models defined in Definition 8. Therefore, the practical condition to check *model minimality* is defined as follows: a model λ can be approximately taken as *minimal* if the absolute value of its *specificity* is outside the *distance threshold* of 3-sigma; that is, $|\mathcal{S}(\lambda)| > 3\sigma$.

3. Impact of States on Observation Likelihood

We start the study through an information flow analysis as to see the impact of different types of states on model specificity.

3.1. Information Flow Analysis. Our aim is to understand which parameters make an HMM have a higher specificity. However, an analytical equation for the specificity function $\mathcal{S}(\lambda)$ requires us to know the optimal one-state-simpler model $\overline{\lambda}_{Q-1}(\lambda_Q)$, which is still an open problem. This leads us to an alternative approach by analysing *state properties* of models. In the following analysis, we will study which properties make up a *high-impact* state and which do not. A *high-impact* state makes itself more specific with a significant influence on $P_{\lambda}(\mathbf{o}_{1:t})$; thus it emits relatively unique patterns of observation sequences which can be distinguished from other states. Using this analysis, we will in this paper propose a framework to identify the *high-impact* states.

To study what influences the *specificity* of an HMM, we analyse the impact of a state on the likelihood $P_{\lambda}(\mathbf{o}_{1:t})$ and how it contributes to $\mathcal{S}(\lambda)$ as follows. Consider $P_{\lambda}(o_{t+1} \mid \mathbf{o}_{1:t})$ in (10). It can be seen as a probability used in predicting the future from the past and it represents the information flow from the past to the future. Hence we will analyse the contribution of a specific state to this probability. There are three cases whereby the probability of the state q_t plays a role in the information flow, as shown in Figure 1:

(a) The present state probability depends on the previous state probability and partly determines the observation probability $P_{\lambda}(o_t \mid \mathbf{o}_{1:t-1})$.

(b) The present state probability depends on the observations and determines the succeeding state probability. The observation probability $P_{\lambda}(o_{t+1} \mid \mathbf{o}_{1:t})$ depends on $P_{\lambda}(q_t \mid \mathbf{o}_{1:t})$ which is updated with the knowledge of o_t.

(c) The present state probability is determined by the past state probability and affects the future state probability.

3.2. High-Impact States. We now investigate the high-impact states on likelihood $P_{\lambda}(\mathbf{o}_{1:t})$, more specifically on the *specificity* $\mathcal{S}(\lambda)$. Such states should have a *high* and *unique* impact on the likelihood where *high* means a high information flow

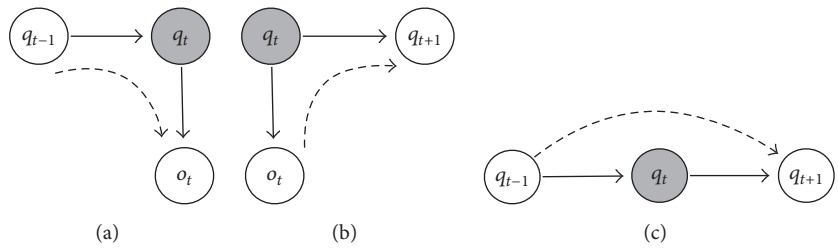


Figure 1: Role of a state in the likelihood.

passing from the past to the future states and *unique* ensures that no other states can fill in the same role, such that it cannot be mimicked by other states either with combined similar probabilities or emitting similar observation probabilities. For instance, a state with a probability of 0.5 can be mimicked by a combination of two states with probabilities of 0.1 and 0.9, respectively; or a state with observation emission probabilities of 0.5 is also not unique. Note that a relatively high or low probability is more difficult to be mimicked than 0.5 in the previous examples. Hence for the three cases outlined in Figure 1, the state plays an intermediate role in predicting the future based on the past; we can define the following conditions for high-impact state, respectively:

(a) (1) The incoming transition probabilities $\mathbf{a}_{\cdot r}$ (see (4)) of state s_r at time t are maximal or minimal; that is, $a_{ir} \gg a_{iq} \parallel a_{ir} \ll a_{iq},\ \forall q \neq r,\ 1 \leq i \leq Q$; (2) state s_r has a *dominant* observation v_k at time t, meaning the observation probability b_{rk} (see (3)) is maximal; that is, $b_{rk} \gg b_{qk},\ \forall q \neq r$.

(b) (1) The outgoing transition probabilities $\mathbf{a}_{r\cdot}$ (see (5)) of state s_r at time t are maximal or minimal; that is, $a_{ri} \gg a_{qi} \parallel a_{ri} \ll a_{qi},\ \forall q \neq r,\ 1 \leq i \leq Q$; (2) state s_r has a *dominant* observation v_k at time t; refer to condition a(2).

(c) Refer to conditions a(1) and b(1).

For high specificity, the above conditions should be met for all states of a model. Note that these conditions are based on state *transition* and *observation* probabilities. Regarding *transition* probabilities, a highly specific HMM should contain *persistent* and/or *transient-cyclic* states, as defined below:

(i) A *persistent* state is a state with a higher *self-transition* probability than the probabilities to transit to other states. When all states of an HMM are *persistent*, the HMM remains for a certain period in one state before changing into another state. Such HMM is called a *persistent* HMM.

(ii) A *transient* state, on the other hand, has a lower *self-transition* probability. A *transient-cyclic* state has one specific incoming transition probability which is high and dominant and one outgoing transition probability which is high. When all states of an HMM are *transient-cyclic*, the HMM flips from one state to another, mostly following a certain pattern (e.g., $s_1 \rightarrow s_2 \cdots \rightarrow s_Q \rightarrow s_1 \rightarrow s_2 \cdots$). Such HMM is called a *transient-cyclic* HMM. Otherwise, it is called a *transient-acyclic* HMM.

(iii) When an HMM contains both *persistent* and *transient-cyclic* states, we call it a *hybrid* HMM.

Secondly, regarding *observation* probabilities, a highly specific HMM should contain *privileged* states, which is defined as follows:

A *privileged* state is a state with at least one *dominant* observation probability.

HMMs containing only *privileged* states are called *privileged* HMMs. This is possible when the number of observations is larger than the number of states; that is, $M \geq Q$.

Considering both *transition* and *observation* probabilities, we define a *highly specific* HMM as an HMM containing only *persistent* states and/or *transient-cyclic* states, which will be shown as identifiable from observation sequences. Note that it is impossible to identify all minimal HMMs, especially when the influence of some states on a model is low, in the sense that such states can be neglected and the resultant simpler model is comparable to a complex one. In order to learn a *minimal identifiable* HMM, we propose in a later section an effective and efficient model approximation method which identifies *persistent* states with segmentation and clustering methods and *transient-cyclic* states with a transient analysis based on the following theorem.

Theorem 9. *The presence of transient-cyclic states with dominant observations can be identified as follows: for values of* $k, l, m \in [1, M],\ k \neq l \neq m$, *if* $\overline{P}(o_t = v_k,\ o_{t+1} = v_l) > \xi$ *and* $\overline{P}(o_{t+1} = v_l,\ o_{t+2} = v_m) > \xi$, *where* $0 \leq \xi \leq 1$, $\overline{P}(*)$ *represents the relative frequency (i.e., the ratio of the number of times) of event* $*$ *occurring in the observed sequence, which is also the predicted probability of the occurrence of event* $*$*; then for*

$$\hbar(k,l,m) = \frac{\overline{P}(o_t = v_k,\ o_{t+1} = v_l)\,\overline{P}(o_{t+1} = v_l,\ o_{t+2} = v_m)}{\overline{P}(o_t = v_k,\ o_{t+1} = v_l,\ o_{t+2} = v_m)}, \quad (14)$$

(a) *if* $\hbar \approx 1$, *that is,* $\hbar \in [1-\epsilon, 1+\epsilon]$, $\epsilon \approx 0$, *the triple* $\overline{P}(o_t = v_k,\ o_{t+1} = v_l,\ o_{t+2} = v_m)$ *does not reveal hidden transient-cyclic states and thus it can be modelled by a 1-order Markov model,*

(b) *if* $\hbar \not\approx 1$, *the triple* $\overline{P}(o_t = v_k,\ o_{t+1} = v_l,\ o_{t+2} = v_m)$ *reveals that hidden transient-cyclic states are present:*

(i) *If $\hbar < 1-\epsilon$, the triple reveals states with dominant observations.*

(ii) *If $\hbar > 1+\epsilon$, the triple reveals states with dominant observations and an extra mixing state.*

The proof is in Appendix A.

The definitions of a Markov model and a *mixing* state used in the theorem are given as follows:

(i) A *Markov* model is a stochastic process that is characterized by a Markov chain. It models the observed states with a random variable which satisfies the Markov property; that is, the distribution of the current state depends only on that of the previous state instead of the whole historical states. The state transition probability distribution and the initial state probability distribution are denoted by the same expressions as the HMM defined previously. The model can be written as $\lambda = (\boldsymbol{\pi}, \mathbf{A})$.

(ii) A *mixing* state is a state which outputs the same observation probabilities as a mixture of other states. HMM models containing mixing states are problematic, since one state has the same output distribution as a convex mixture of some other states' output distribution; therefore it is difficult to distinguish the ground truth state between a single state and a mixture of several states [14].

3.3. Equivalent States. Now we try to understand when a state has zero impact on the specificity such that in the extreme case a simpler HMM exists with the same distributions. Considering the information flow past $\rightarrow$ state $\rightarrow$ future, for the first arrow, the influence of a state is negligible when (1a) $P(\text{state} = s_r \mid \text{past})$ is close to zero; (1b) the state has an equal influence as another state if the probability equals that of another state; or (1c) the influence of the state can be mimicked by the other state if the probability is constant. Note that if it is neither constant nor the same as another state, the state probability will fluctuate which makes that its influence cannot be incorporated into that of other states. For the second arrow, the influence of the state can be incorporated into that of other states if (2a) $P(\text{future} \mid \text{state} = s_r)$ is the same as the probabilities of another state or (2b) the probability distribution is not dominant.

In case (1a) the state plays no role and can be removed, in cases (1b) and (2a) the state can be merged with a similar state, and in cases (1c) and (2b) the influence of the state can be "taken over" by some of the remaining states. This leads to the conditions for eliminating redundant (i.e., equivalent) states as shown in Table 1. Note that the difference between "removal" and "taken over" is that, by removing a state, its information is removed together with the state, while "taking over" a state means that even though the state is deleted, its information stays and is passed to other states instead.

Based on the conditions of equivalent states defined in Table 1, we now can formalize the results of our analysis in sufficient conditions for nonminimality HMMs as follows.

Table 1: Conditions on a state r $(1 \le r \le Q)$ to achieve minimality through the removal of the state, the merging with another state, or the adjustment of the probabilities by some other states such that the influence is "taken over."

State reduction	State conditions
Removal	(1a) $a_{ir} = 0,\ \forall i \in [1, Q]$
Merge	(1b) $a_{ir} = a_{iq},\ \forall q \neq r,\ \forall i \in [1, Q]$
	(2a) $a_{ri} = a_{qi},\ \forall q \neq r,\ \forall i \in [1, Q]$
Taken over	(1c) $a_{ir} = C,\ \forall i \in [1, Q]$, C is constant
	(2b) non dominant $b_{rk},\ \forall k \in [1, M]$

Theorem 10. *A stationary HMM is not minimal if one of the following conditions holds:*

(i) *The HMM contains a state r that has zero incoming state transition probabilities; that is, $a_{ir} = 0,\ \forall i \in [1, Q]$.*

(ii) *The HMM contains two states q and r that have the same state transition probabilities; that is, $a_{iq} = a_{ir}$ and $a_{qi} = a_{ri},\ \forall i \in [1, Q]$.*

(iii) *The HMM contains two states q and r that have the same observation probabilities $b_{qk} = b_{rk},\ \forall k \in [1, M]$ and meets one of the following conditions: (1) they have the same incoming state transition probabilities; that is, $a_{iq} = a_{ir},\ \forall i \in [1, Q]$; (2) they have the same outgoing state transition probabilities; that is, $a_{qi} = a_{ri},\ \forall i \in [1, Q]$; or (3) $a_{iq} = a_{ir}$ and $a_{qi} = a_{ri},\ \forall i \in [1, Q] \setminus \{q, r\}$.*

(iv) *The HMM has two observation values $(M = 2)$ and contains a state r that has constant incoming state transition probabilities; that is, $a_{ir} = C$ and for all k, r has nondominant observation probabilities; that is, $b_{rk} < b_{ik},\ \forall i \in [1, Q] \setminus r,\ \forall k \in [1, M]$.*

The proof is in Appendix B.

3.4. Low-Impact States. Unlike high-impact or equivalent (zero-impact) states, some states have larger-than-zero but very low impact, which makes them hard to learn. Such states are called low-impact states. HMMs containing these states are called *hard to learn* HMMs, as will be shown later.

Since low-impact states are in between high-impact and equivalent states, they meet a combination of partial conditions defined for both cases. As introduced in Section 3.2 for high-impact states, a learnable HMMs should contain only *persistent* and/or *transient-cyclic* states with *privileged* observations, while an unlearnable HMMs contains states which contains one or two states under conditions defined in Theorem 10. Therefore, combined partial conditions of both can be defined for *hard to learn* HMMs.

An HMM is *hard to learn* if it contains mostly *persistent* or *transient-cyclic* states with *privileged* states with dominant observations and is also under one of the following conditions:

(i) There exists a mixing state r whose observation distribution is a mixture of the observation distributions of two other states q and k; that is, $b_{rj} = (b_{qj} + b_{kj})/2$, where $q, r, k \in [1, Q],\ \forall j \in [1, M]$.

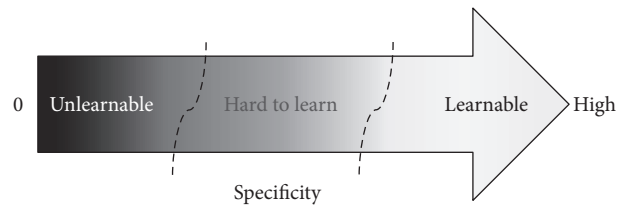


Figure 2: Learnability versus specificity of HMMs.

(ii) There exists a state r with constant incoming transitions, self-included; that is, $a_{ir} = 1/Q$, where $r \in [1, Q]$, $\forall i \in [1, Q]$.

(iii) There exists a state r with constant incoming transitions, self-excluded; that is, $a_{ir} = (1 - a_{rr})/(Q - 1)$, where $r \in [1, Q]$, $\forall i \in [1, Q] \setminus r$.

(iv) There exists a state r with constant outgoing transitions, self-included; that is, $a_{ri} = 1/Q$, where $r \in [1, Q]$, $\forall i \in [1, Q]$.

(v) There exists a state r with constant outgoing transitions, self-excluded; that is, $a_{ri} = (1 - a_{rr})/(Q - 1)$, where $r \in [1, Q]$, $\forall i \in [1, Q] \setminus r$.

(vi) There exist two states q and r with the same observation probabilities $b_{qj} = b_{rj}$, where $q, r \in [1, Q]$, $\forall j \in [1, M]$.

(vii) There exists a state r with constant (nondominant) observation emissions; that is, $b_{rj} \cong 1/M$, where $r \in [1, Q]$, $\forall j \in [1, M]$.

4. Approximate Identification Algorithm

An HMM is either identifiable or unidentifiable. In order to describe how hard it is to identify a model, we use the term *learnability*: for an identifiable HMM, it can be easy, moderate, or hard to learn. Thus, before presenting the approximate identification algorithm, we firstly explain our hypothesis on the correlations between model learnability and specificity as shown in Figure 2, which will be validated experimentally in Section 5. HMMs containing states with higher specificity have higher distances with less complex models and as shown later are easier to learn, and vice versa. Therefore, we classify HMMs into three identification categories based on their specificity: (1) *learnable* HMMs with relatively high specificity; (2) *hard to learn* HMMs with low specificity; and (3) *unlearnable* HMMs with almost zero specificity. Our focus is to identify learnable and highly specific models with high-impact states.

4.1. Algorithm Structure. Based on the previous analysis of the hidden states, we can construct an algorithm that identifies high-impact states directly from the observation sequences. Inspired by signal processing method such as Empirical Mode Decomposition- (EMD-) and wavelet-based denoising methods [25], which decompose the noisy signal into a number of components, filter each component, and finally reconstruct the denoised signal using the filtered components, here we reassemble the above procedures as follows: an unknown HMM is composed out of a number of hidden states. These states can be identified from observations and combined to form a reconstructed HMM*, as shown in Figure 3. In such manner, we decompose the model identification procedure into a combination of state identifications. The approximate state identification approach firstly recognizes persistent and transient states separately from observation sequences, then combines them into a set of identified states, and finally reduces or merges similar states into a new set of reconstructed states. The details of the identification framework will be explained as follows.

Models with high-impact states generate specific samples which are unique. Therefore, the states are identifiable through data analysis. The output of a *persistent* state is likely to stay in a period within which the behavior at each time step is "similar," which we call a *regime*. Thus a segmentation approach is advised splitting a signal sequence into regimes by identifying the specific behaviors within certain periods. In order to identify *transient-cyclic* states, our approach is to capture the changing of behaviors with a transient analysis based on Theorem 9.

Therefore, we propose a framework to identify most high-impact and minimal states: (1) persistent privileged states; (2) transient-cyclic privileged states; and (3) hybrid: persistent and transient-cyclic privileged states. A schema of the proposed algorithm is shown in Figure 4. We assume that both persistent and transient states exist in the model; therefore, both segmentation and clustering and transient analysis methods are applied on the data, followed by a reparameterization procedure to combine the parameters learned from both the previous methods. Finally, a model reduction step is conducted in the end to form a simplified minimal HMM model. We name our proposed method as Segmentation-Clustering and Transient analysis (SCT) framework.

4.2. Segmentation-Based Approach. The segmentation-based approach is defined using the following steps: Step 1: signals are split by segmentation techniques into different regimes with different signal behaviors; Step 2: the "similar" regimes of signals are grouped together by clustering techniques according to their similarities (the clusters are labeled and each cluster is a hidden state); Step 3: a clustering validation index is employed to determine the proper number of states; finally, Step 4: HMM parameters are estimated by calculating statistical occurrences of the observations and the estimated hidden states.

Step 1 (identification of persistent states by segmentation). Data sequences emitted by *persistent* states can be segmented into subsequences with constant behavior (observations are drawn from a stationary distribution). The transition from one state to another can be identified by detecting a difference in signal behavior. This is called a *change point*. In this paper, we propose a sliding window-based Bayesian segmentation based on the test of [26]. The Bayesian probability is calculated to determine whether two sequences have been generated by the same or by a different multinomial model.

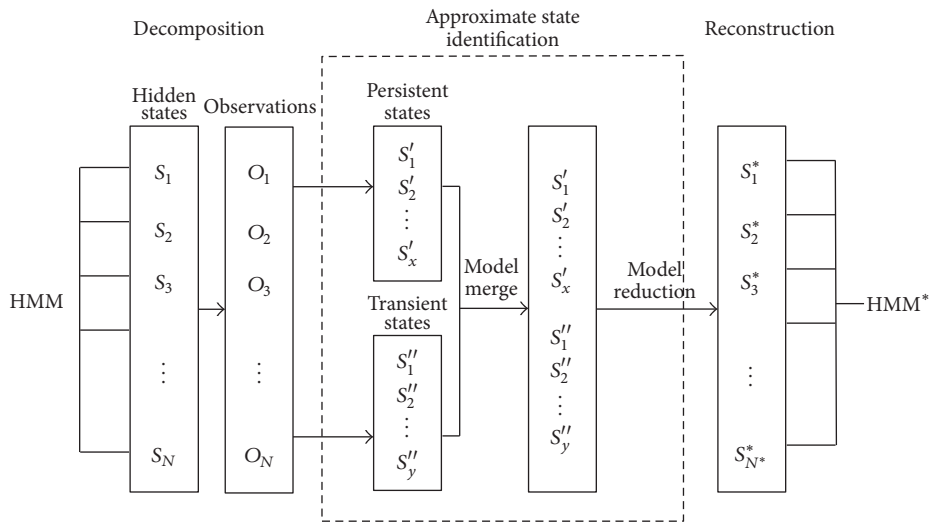


Figure 3: Identification of HMMs.

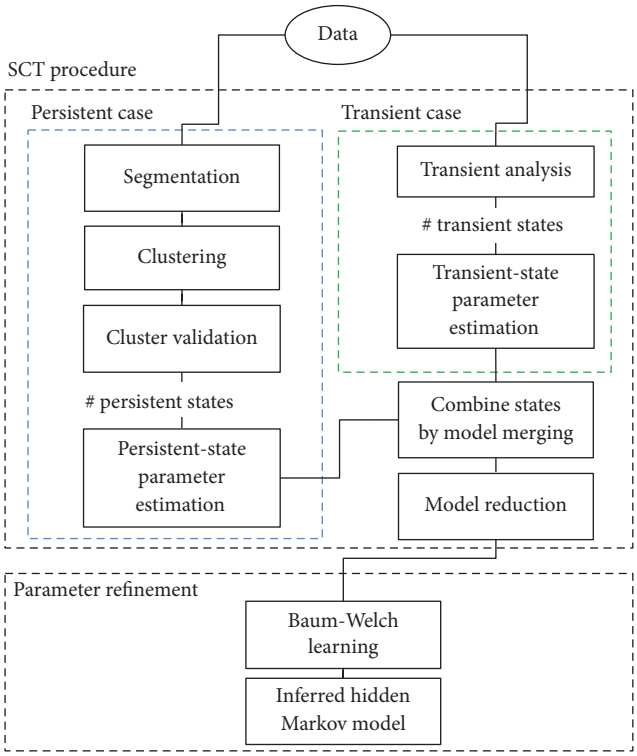


Figure 4: Scheme of the proposed approach.

A *multinomial* model is a stochastic process where the observations follow a multinomial distribution. It is sufficient to model observations instead of states, since the observations in a multinomial model can represent states correspondingly with absolute state knowledge. Each observed symbol o_t at time t is independent and falls into one of N categories with a fixed probability, denoted by $\mathscr{P} = \{p_1, p_2, \ldots, p_N\}$, where $\sum_{i=1}^{N} p_i = 1$. The observation probability distribution is denoted by $\boldsymbol{\pi} = P(o_t = s_i) \in \mathscr{P}$, where $s_i \in S$, $S = \{s_1, s_2, \ldots, s_N\}$, $1 \le t \le T$. The model can be represented with a compact notation $\lambda = (\boldsymbol{\pi})$.

The first sequence always starts from the last change point (the first point if at the beginning) and ends at the current time point; the second sequence is a fixed-length sliding window starting from the next time point. If the two successive sequences are very likely from different models, the point in between is marked as a change point. The procedure repeats until the end of the signal.

Step 2 (combination of states by clustering). With HMMs, segments corresponding to the same state will recur over time. Assuming that there is a finite number of states, segments with the same states are detected and clustered together. In this study, the classical *k-means* clustering approach [27, 28] is chosen to group and label segments. In our case, the *k-means* clustering algorithm tries to group the segments into k unique states based on the mean value of data features within each cluster, given by k. Because of the fact that *k-means* clustering encounters the problem of randomness in selecting initial parameters, we perform a preliminary step for selecting centroid starting locations. The selected properties in the segmentation step are a one-dimension sequence, which contains k subsequences with equal length. The median values of the subsequences are then used as initial centroid locations.

Step 3 (cluster validity). In order to select the optimal number of clusters, we propose a constraint-based clustering analysis considering both the cluster separation capabilities of hidden states and the simplicity of HMM models. *Constraint 1*: lower Davies-Bouldin index (DBI) [29] suggests that the clustering exhibited a better intracluster grouping and intercluster separation of each state. *Constraint 2*: instead of selecting the minimum DBI, an allowance with a threshold of 0.05 is given so that a smaller number of states will be selected if its DBI is within the range of min(DBI) + 0.05.

Suppose dataset X is partitioned into K disjoint nonempty clusters C_i and let $\{C_1, C_2, \ldots, C_K\}$ denote the obtained partitions, such that $C_i \cap C_j = \emptyset$ (empty set), $i \neq j$, $C_i \neq \emptyset$, and $X = \bigcup_{i=1}^{K} C_i$. The Davies-Bouldin index [29] is defined as

$$\mathrm{DBI} = \frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}\left\{\frac{\mathrm{diam}\left(C_i\right)+\mathrm{diam}\left(C_j\right)}{\mathrm{dist}\left(C_i, C_j\right)}\right\}, \quad (15)$$

where

$$\mathrm{diam}\left(C_i\right) = \max_{\mathbf{x}_m, \mathbf{x}_n \in C_i}\left\{d\left(\mathbf{x}_m, \mathbf{x}_n\right)\right\},$$

$$\mathrm{dist}\left(C_i, C_j\right) = \min_{\mathbf{x}_m \in C_i, \mathbf{x}_n \in C_j, i\neq j}\left\{d\left(\mathbf{x}_m, \mathbf{x}_n\right)\right\} \quad (16)$$

indicate the intracluster diameter and the intercluster distance, respectively. The partition with the minimum Davies-Bouldin index is considered as the optimal choice.

Step 4 (parameter estimations). Parameters of an HMM (i.e., probability matrices $\lambda_{\mathrm{pers}} = (\boldsymbol{\pi}_{\mathrm{pers}}, \mathbf{A}_{\mathrm{pers}}, \mathbf{B}_{\mathrm{pers}})$) can be calculated by simply counting the occurrence of the observed signal and the hidden states (i.e., labels retrieved from clustering), which is the same calculation as the reestimation step of the Baum-Welch algorithm [1]:

$$\begin{aligned}
\overline{\pi}_i &= \text{expected frequency (number of times) in state } s_i \text{ at time } t = 1,\\
\overline{a}_{ij} &= \frac{\text{expected number of transitions from } s_i \text{ to } s_j}{\text{expected number of transitions from } s_i},\\
\overline{b}_j(k) &= \frac{\text{expected number of times in } s_j \text{ observing } v_k}{\text{expected number of times in } s_j}.
\end{aligned} \quad (17)$$

4.3. Transition-Based Approach. In this section we present a transition-based approach in order to identify transient-cyclic states. In order to estimate the observation matrix, we apply Theorem 9 which is dedicated to identifying transient-cyclic privileged with or without mixing states. Firstly, the first-order transition probabilities can be identified by a Markov model assumption via counting the occurrence of the observation sequences:

$$\begin{aligned}
&\overline{P}\left(o_t = v_k,\ o_{t+1} = v_l\right)\\
&\quad = \frac{\text{expected number of transitions from } v_k \text{ to } v_l}{\text{expected total number of 1-order transitions}}.
\end{aligned} \quad (18)$$

Similarly, a second-order transition probability can be modelled by an HMM assumption and calculated by counting:

$$\begin{aligned}
&\overline{P}\left(o_t = v_k,\ o_{t+1} = v_l,\ o_{t+2} = v_m\right)\\
&\quad = \frac{\text{expected number of transitions from } v_k \text{ to } v_l \text{ to } v_m}{\text{expected total number of 2-order transitions}},
\end{aligned} \quad (19)$$

where $k, l, m \in [1, M]$, $k \neq l \neq m$. A threshold for dominant probabilities is calculated as

$$\overline{\xi} = \frac{1}{\text{expected total number of 1-order transitions}}. \quad (20)$$

If the two continuous first-order probabilities are dominant, that is, $\overline{P}(o_t = v_k,\ o_{t+1} = v_l) > \overline{\xi}$ and $\overline{P}(o_{t+1} = v_l,\ o_{t+2} = v_m) > \overline{\xi}$, where $0 \leq \overline{\xi} \leq 1$, then the division of the second-order transition probabilities calculated from a Markov chain and from an HMM assumption is

$$\overline{\hbar} = \frac{\overline{P}\left(o_t = v_k,\ o_{t+1} = v_l\right)\overline{P}\left(o_{t+1} = v_l,\ o_{t+2} = v_m\right)}{\overline{P}\left(o_t = v_k,\ o_{t+1} = v_l,\ o_{t+2} = v_m\right)}. \quad (21)$$

If $\overline{\hbar} = 1$, there is no transient-cyclic states. Otherwise, the first-order transition probabilities are taken as the dominant observation probabilities and used to build the observation matrix. If $\overline{\hbar} > 1$, a mixing state is present and one extra state is added to the observation matrix with a uniformly distributed probability of $1/M$. In the end, we map each observation value v_k, v_l, and $v_m$on a different state because we look for states with at least one dominant observation value. If a state has multiple observation values, they will be merged into one state. See Table 1 for conditions of model state reduction.

Take a simple 2-observation case as an example; we generate a 10-series sequence with length of 1000. The first-order observation transition probabilities are $\begin{bmatrix} P(o_t=v_1,\ o_{t+1}=v_1) & P(o_t=v_1,\ o_{t+1}=v_2) \\ P(o_t=v_2,\ o_{t+1}=v_1) & P(o_t=v_2,\ o_{t+1}=v_2) \end{bmatrix}$. If the calculated occurrence probabilities are $\begin{bmatrix} 0.4603 & 0.0420 \\ 0.0422 & 0.4555 \end{bmatrix}$, the dominant transitions larger than 1/4 are $P(o_t = v_1,\ o_{t+1} = v_1)$ and $P(o_t = v_2,\ o_{t+1} = v_2)$. Thus, the dominant second-order probabilities are $\begin{bmatrix} P(o_t=v_1,\ o_{t+1}=v_1,\ o_{t+2}=v_1) \\ P(o_t=v_2,\ o_{t+1}=v_2,\ o_{t+2}=v_2) \end{bmatrix}$, equal to $\begin{bmatrix} 0.4217 \\ 0.4170 \end{bmatrix}$ calculated by a Markov model, while being equal to $\begin{bmatrix} 0.4336 \\ 0.4269 \end{bmatrix}$ calculated by an HMM. Thus, the division of the two is $\hbar = \begin{bmatrix} 0.9725 \\ 0.9768 \end{bmatrix}$. Since both probabilities are smaller than 1, they are dominant states and there is no mixing state. Therefore, we map the two dominant probabilities to the observation probabilities of two states and the final observation matrix is $\mathbf{B}_{\mathrm{tran}} = \begin{bmatrix} 0.9725 & 0.0275 \\ 0.0232 & 0.9768 \end{bmatrix}$.

Furthermore, for calculating the prior and transition probabilities, we stick to the assumption that privileged state behaviors can be reflected by observation properties; therefore, we assume the number of states is the same as the number of observations and use a Markov model for learning state probabilities.

The prior and transition matrices can be calculated by counting the observation occurrences:

$$\begin{aligned}\overline{\pi}_i &= \text{expected frequency (number of times) in observation } v_i \text{ at time } t = 1,\\ \overline{a}_{ij} &= \frac{\text{expected number of transitions from } v_i \text{ to } v_j}{\text{expected number of transitions from } v_i}.\end{aligned} \tag{22}$$

Therefore, a model $\lambda_{\text{tran}} = (\boldsymbol{\pi}_{\text{tran}}, \mathbf{A}_{\text{tran}}, \mathbf{B}_{\text{tran}})$ is learned containing only transient-cyclic states.

4.4. Reparameterization. Parameters learned for both *persistent* and *transient-cyclic* states are combined together by a procedure called *reparameterization*. Let $\lambda_{\text{pers}} = (\boldsymbol{\pi}_{\text{pers}}, \mathbf{A}_{\text{pers}}, \mathbf{B}_{\text{pers}})$ be the parameters for *persistent* states and $\lambda_{\text{tran}} = (\boldsymbol{\pi}_{\text{tran}}, \mathbf{A}_{\text{tran}}, \mathbf{B}_{\text{tran}})$ be the parameters for the *transient-cyclic* states. Let Q_{pers} and Q_{tran} be the number of *persistent* and *transient-cyclic* states, respectively. Thus the combined number of states is $Q' = Q_{\text{pers}} + Q_{\text{tran}}$. The parameters of the combined model $\lambda'(\boldsymbol{\pi}', \mathbf{A}', \mathbf{B}')$ can be calculated as

$$\begin{aligned}\boldsymbol{\pi}' &= \text{Normalize}\left(\begin{bmatrix}\boldsymbol{\pi}_{\text{pers}}\\ \boldsymbol{\pi}_{\text{tran}}\end{bmatrix}_{N'\times 1}\right),\\ \mathbf{A}' &= \text{Stochastic}\left(\frac{1}{Q'} * \mathbf{1}_{Q'\times Q'} + \begin{bmatrix}\mathbf{A}_{\text{pers}} & \mathbf{0}_{Q_{\text{pers}}\times Q_{\text{tran}}}\\ \mathbf{0}_{Q_{\text{tran}}\times Q_{\text{pers}}} & \mathbf{A}_{\text{tran}}\end{bmatrix}_{Q'\times Q'}\right),\\ \mathbf{B}' &= \text{Stochastic}\left(\begin{bmatrix}\mathbf{B}_{\text{pers}}\\ \mathbf{B}_{\text{tran}}\end{bmatrix}_{Q'\times M}\right),\end{aligned} \tag{23}$$

where the function *Normalize* ensures that the sum of the given vectors equals 1 and the *Stochastic* function ensures that the sum of each row of the given matrix equals 1.

4.5. Model Reduction. After combining the persistent and transient-cyclic states, redundant states may occur. We introduce a model reduction procedure which removes redundant states to obtain minimal HMMs according to the conditions defined in Theorem 10. We relax the strict conditions given in the theorem via adding thresholds.

(i) *An HMM contains a state r that has zero incoming state transition probabilities; that is, $a_{ir} = 0,\ \forall i \in [1, Q]$.*

Instead of using a zero vector as a strict rule, a threshold is defined to allow near-zero cases such that if the sum of the incoming transition state probabilities of state r is lower than threshold θ_a, that is,

$$\sum_{i=1}^{Q}(a_{ir}) < \theta_a, \tag{24}$$

the state r can be removed.

(ii) *An HMM contains two states q and r that have the same state transition probabilities; that is, $a_{iq} = a_{ir}$ and $a_{qi} = a_{ri},\ \forall i \in [1, Q]$.*

We replace the equivalence condition by a subtraction calculation. If the maximum of the incoming and outgoing state transition probabilities of the two states q and r is below a threshold θ_b, that is,

$$\max\left(\max_{1\le i\le Q}\left(a_{iq} - a_{ir}\right), \max_{1\le i\le Q}\left(a_{qi} - a_{ri}\right)\right) < \theta_b, \tag{25}$$

the two states can be merged into one.

(iii) *An HMM contains two states q and r that have the same observation probabilities $b_{qk} = b_{rk},\ \forall k \in [1, M]$, and meet one of the following conditions: (1) they have the same incoming state transition probabilities; that is, $a_{iq} = a_{ir},\ \forall i \in [1, Q]$; (2) they have the same outgoing state transition probabilities; that is, $a_{qi} = a_{ri},\ \forall i \in [1, Q]$; or (3) $a_{iq} = a_{ir}$ and $a_{qi} = a_{ri},\ \forall i \in [1, Q] \setminus \{q, r\}$.*

Similar to (ii), we use subtraction instead of strict equivalence with added threshold θ_c. Moreover, the *AND* condition and the *OR* conditions can be represented by selecting the maximum and minimum values, respectively. Therefore, if

$$\begin{aligned}&\max\left(\max_{1\le k\le M}\left(b_{qk} - b_{rk}\right), \min\left(\max_{1\le i\le Q}\left(a_{iq} - a_{ir}\right),\right.\right.\\ &\left.\max_{1\le i\le Q}\left(a_{qi} - a_{ri}\right),\right.\\ &\left.\left.\max\left(\max_{\substack{1\le i\le Q\\ i\ne r\ne q}}\left(a_{iq} - a_{ir}\right), \max_{\substack{1\le i\le Q\\ i\ne r\ne q}}\left(a_{qi} - a_{ri}\right)\right)\right)\right)\\ &< \theta_c, \quad \forall i \in [1, Q],\end{aligned} \tag{26}$$

the two states q and r can be merged.

(iv) *An HMM has two observation values* ($M = 2$) *and contains a state* r *that has constant incoming state transition probabilities,* $a_{ir} = C$ *and* $\forall k$*, and* r *has nondominant observation probabilities,* $b_{rk} < b_{ik}$, $\forall i \in [1, Q] \setminus r$, $\forall k \in [1, M]$.

$$\max_{1 \le i \le Q} \left(a_{ir} - \overline{a}_{ir}\right) < \theta_d, \qquad b_{rk} = \min_{1 \le k \le M} \left(b_{ik}\right), \quad \forall i \in [1, Q], \tag{27}$$

where $\overline{a}_{ir}$ is the average of a_{ir} and the state r can be "taken over."

The selection of thresholds is conducted empirically since the correlation between the likelihood value and each condition is complex and is not our focus in this paper.

5. Experiments

Simulated data has been used to evaluate the effectiveness and efficiency of the proposed SCT inference framework. The simulated data were sampled from different classes of HMM models: nonminimal equivalent HMMs, identifiable (selected and random) minimal HMMs, and hard to learn HMMs.

5.1. Nonminimal Equivalent HMMs. Equivalent HMMs contains two cases: (1) two HMMs with the same number of states, where permutations of states apply to both models; (2) two HMMs with different numbers of states. This experiment focuses on case (2) and aims to test the model reduction conditions defined in Table 1, where a nonminimal HMM can remove, merge, or take over its redundant states to become an equivalent minimal HMM. One model λ is selected under each of the three reduction conditions and is used to construct an equivalent model $\widetilde{\lambda}$ by removing the redundant state (set as the last state here). The model parameters are listed hereafter, respectively:

With a *removable* state:

HMM λ_1:

$$\boldsymbol{\pi}_1 = \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix}, \quad \mathbf{A}_1 = \begin{bmatrix} 0.9 & 0.1 & 0 \\ 0.1 & 0.9 & 0 \\ 0.5 & 0.5 & 0 \end{bmatrix}, \quad \mathbf{B}_1 = \begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix} \tag{28}$$

HMM $\widetilde{\lambda}_1$:

$$\widetilde{\boldsymbol{\pi}}_1 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad \widetilde{\mathbf{A}}_1 = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}, \quad \widetilde{\mathbf{B}}_1 = \begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \end{bmatrix} \tag{29}$$

With a *mergeable* state:

HMM λ_2:

$$\boldsymbol{\pi}_2 = \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 0.45 & 0.45 & 0.1 \\ 0.45 & 0.45 & 0.1 \\ 0.05 & 0.05 & 0.9 \end{bmatrix}, \quad \mathbf{B}_2 = \begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix} \tag{30}$$

HMM $\widetilde{\lambda}_2$:

$$\widetilde{\boldsymbol{\pi}}_2 = \begin{bmatrix} 0.67 \\ 0.33 \end{bmatrix}, \quad \widetilde{\mathbf{A}}_2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}, \quad \widetilde{\mathbf{B}}_2 = \begin{bmatrix} 0.475 & 0.475 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix} \tag{31}$$

With a *taken-over* state:

HMM λ_3:

$$\boldsymbol{\pi}_3 = \begin{bmatrix} 0.33 \\ 0.33 \\ 0.33 \end{bmatrix}, \quad \mathbf{A}_3 = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.4 & 0.4 & 0.2 \end{bmatrix}, \quad \mathbf{B}_3 = \begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \\ 0.5 & 0.5 \end{bmatrix} \tag{32}$$

Table 2: Distance threshold test results for nonminimal equivalent models.

Cases	Remove		Merge		Take over	
Models	λ_1	$\widetilde{\lambda}_1$	λ_2	$\widetilde{\lambda}_2$	λ_3	$\widetilde{\lambda}_3$
μ	−0.92	−0.92	−1.24	−1.24	−0.97	−0.97
σ	0.012	0.012	0.014	0.014	0.003	0.003
$[\mu - 3\sigma, \mu + 3\sigma]$	[−0.955, −0.881]	[−0.955, −0.881]	[−1.279, −1.191]	[−1.279, −1.191]	[−0.978, −0.959]	[−0.977, −0.960]
Confidence of equivalence (%)	99.9		99.8		99.8	

HMM $\widetilde{\lambda}_3$:

$$\widetilde{\boldsymbol{\pi}}_3 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix},$$

$$\widetilde{\mathbf{A}}_3 = \begin{bmatrix} 0.708 & 0.292 \\ 0.292 & 0.708 \end{bmatrix}, \tag{33}$$

$$\widetilde{\mathbf{B}}_3 = \begin{bmatrix} 0.833 & 0.167 \\ 0.167 & 0.833 \end{bmatrix}$$

Each of the reference nonminimal models λ_i, $i = [1, 2, 3]$ is used to generate 1000 datasets of random observations containing $N = 20$ sequences of $T = 5000$ observation points. The datasets are used to determine log-likelihood distributions of the models and the distance threshold of equivalent models defined in Definition 8. By calculating the percentage of the log-likelihood values of the minimal model $\widetilde{\lambda}_i$ which fall inside the threshold of equivalence for the nonminimal model λ_i, we can obtain a confidence level. The results in Table 2 show that the two models are approximate equivalent models for all the three cases with high confidence levels. Moreover, the log-likelihood histograms are plotted in Figure 5. The highly overlapping histograms further demonstrate the model equivalence for the three examples.

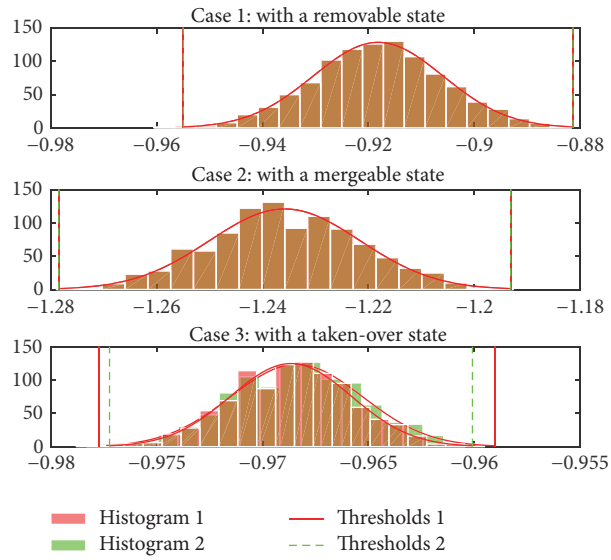


Figure 5: The histogram of log-likelihoods.

5.2. Highly Specific HMMs. As discussed previously, persistent and transient-cyclic HMMs with privileged observations are identifiable HMMs that have a high specificity. In this section, we compare the learning of such identifiable HMMs with the Baum-Welch (BW) algorithm and the proposed SCT method.

Firstly, we constructed 9 persistent and 9 transient-cyclic models as ground truth models with a fixed equal number of states and observations ($Q = M$) ranging from 2 to 10. These models can be expressed as follows:

Persistent λ_{pers}:

$$\boldsymbol{\pi}_{\text{pers}} = \begin{bmatrix} \frac{1}{Q} \\ \frac{1}{Q} \\ \vdots \\ \frac{1}{Q} \end{bmatrix}_{Q\times 1},$$

$$\mathbf{A}_{\text{pers}} = \begin{bmatrix} p_1 & \alpha_1 & & \alpha_1 \\ \alpha_1 & p_1 & & \alpha \\ & & \ddots & \\ \alpha_1 & \alpha_1 & & p_1 \end{bmatrix}_{Q\times Q}, \tag{34}$$

$$\mathbf{B}_{\text{pers}} = \begin{bmatrix} p_1 & \alpha_1 & & \alpha_1 \\ \alpha_1 & p_1 & & \alpha_1 \\ & & \ddots & \\ \alpha_1 & \alpha_1 & & p_1 \end{bmatrix}_{Q\times Q}$$

Transient-cyclic λ_{tran}:

$$\boldsymbol{\pi}_{\text{tran}} = \begin{bmatrix} \frac{1}{Q} \\ \frac{1}{Q} \\ \vdots \\ \frac{1}{Q} \end{bmatrix}_{Q\times 1},$$

$$\mathbf{A}_{\text{tran}} = \begin{bmatrix} \alpha_2 & p_2 & & \alpha_2 \\ & & \ddots & \\ \alpha_2 & \alpha_2 & & p_2 \\ p_2 & \alpha_2 & & \alpha_2 \end{bmatrix}_{Q\times Q}, \tag{35}$$

$$\mathbf{B}_{\text{tran}} = \begin{bmatrix} p_2 & \alpha_2 & & \alpha_2 \\ \alpha_2 & p_2 & & \alpha_2 \\ & & \ddots & \\ \alpha_2 & \alpha_2 & & p_2 \end{bmatrix}_{Q\times Q}$$

The state self-transition probabilities for persistent models and the transition probabilities to the next neighboring state for transient-cyclic models were both set to a value close to 1, noted as p_1 and p_2, respectively, where $p_1, p_2 \in [0.8, 0.99]$. The remaining transitions have equal probabilities; that is, $\alpha_i = (1 - p_i)/(Q - 1)$, $i = 1, 2$. For all the 18 models, the observation matrices are set the same as the transition matrices of persistent models in order to obtain privileged observations. The initial parameters are uniformly distributed.

We also generated 20 hybrid models as ground truth models containing both persistent and transient-cyclic states. The number of states q_1 and q_2 for both cases was randomly chosen from 2 to 3, amounting to a total number of states $Q = q_1 + q_2$ within a range of $[4, 6]$. The rest of the parameters were generated in the same manner as before. For simplicity, we use $[\boldsymbol{\alpha}_i]$ to represent a matrix containing only element α_i. Thus, a hybrid model can be represented as follows:

Hybrid λ_{hybr}:

$$\boldsymbol{\pi}_{\text{hybr}} = \begin{bmatrix} \frac{1}{Q} \\ \frac{1}{Q} \\ \vdots \\ \frac{1}{Q} \end{bmatrix}_{Q\times 1},$$

$$\mathbf{A}_{\text{hybr}} = \begin{bmatrix} [\mathbf{A}_{\text{pers}}]_{q_1\times q_1} & [\boldsymbol{\alpha}_1]_{q_1\times q_2} \\ [\boldsymbol{\alpha}_2]_{q_2\times q_1} & [\mathbf{A}_{\text{tran}}]_{q_2\times q_2} \end{bmatrix}_{Q\times Q},$$

$$\mathbf{B}_{\text{hybr}} = \begin{bmatrix} [\mathbf{B}_{\text{pers}}]_{q_1\times q_1} & [\boldsymbol{\alpha}_1]_{q_1\times q_2} \\ [\boldsymbol{\alpha}_2]_{q_2\times q_1} & [\mathbf{B}_{\text{tran}}]_{q_2\times q_2} \end{bmatrix}_{Q\times Q} \tag{36}$$

The experiments were carried out by using each of the constructed models as a reference model to generate a dataset of 10 series of 1000 observations. The first seven series were used as a training set and the last three series as a test set. The true number of states Q was assumed to be unknown and the learning methods have to select the number of states from a state pool of $[2, Q + 2]$. For the BW learning algorithm, $Q + 1$ models were generated with a number of states ranging from 2 to $Q+2$ and the model with the best Q is selected by the AIC criterion [30]. The learning of the BW algorithm was repeated 20 times to eliminate local optima and the one with the minimum AIC value is selected. In total, $20 * (Q + 1)$ models were generated to determine an optimal model. Moreover, for comparison purpose, we also train BW with a given number of states Q; therefore, a total of 20 models were generated and a best model is selected by the AIC criterion. On the other hand, for the proposed SCT method, the number of states is selected by the clustering validation method in Step 3 (See Section 4.2). Only one SCT model is trained and used for comparison the two best models selected by the BW method (with and without a given Q).

In order to use the 3-sigma rule to indicate if a true model is learned, 100 datasets of 10 series of 1000 observations were generated from each of the true models and used for calculating the log-likelihood distribution $\mathcal{N}(\mu, \sigma)$ (see Definition 8). If the log-likelihood difference between the ground truth model and the learned model is outside the distance threshold of $[-3\sigma, 3\sigma]$, we consider that the true model has not been found (i.e., a local optimum is learned). Moreover, for better understanding the log-likelihood results, we additionally calculated the log-likelihoods for the following models: (1) $\overline{\lambda}_{Q-1}(\lambda_Q)$: the best model with $Q - 1$ states selected from 100 randomly generated models and trained with BW; (2) $\overline{\lambda}_{Q-2}(\lambda_Q)$: the best model with $Q - 2$ states selected from 100 randomly generated models and trained with BW; (3) a multinomial model: the model assuming that there are no hidden states and the observations are the actual visible states. If an HMM model has similar log-likelihood as a multinomial model, the states have no impact on the model.

In addition to log-likelihoods, we define other performance indicators of accuracy as follows: (1) the *percentage of convergence* is the percentage of 20 BW learned models which did not fall into a local optimum; (2) the *percentage of identification* is the percentage of the best BW or the SCT learned models which did not fall into a local optimum; (3) the *parameter distance* is defined as the mean difference of the triples $(\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ between two HMM models. If the two models have different state space, values of 0s are filled into the probability matrices of the simpler model in order to have an equal number of states to the complex one. Moreover, all the permutations of the models are considered and the minimum distance is chosen as the *parameter distance*.

Table 3: Average characteristics of ground truth HMMs.

Type	Specificity	Minimal (%)	$\mu(\mathrm{LL}_{\mathrm{true}})$	$\sigma(\mathrm{LL}_{\mathrm{true}})$
Persistent	0.25	100	−1.00	0.018
Transient-cyclic	0.32	100	−1.25	0.017
Hybrid	0.17	100	−0.93	0.019

Note: $\mu(\mathrm{LL}_{\mathrm{true}})$: mean of the log-likelihood distribution of the true model; $\sigma(\mathrm{LL}_{\mathrm{true}})$: standard deviation of the log-likelihood distribution of the true model.

An average information of the ground truth models can be found in Table 3. The hybrid models are less specific than the persistent or transient-cyclic states only models, which make them harder to identify. Detailed learning results can be found in Table 4. The results show that the proposed SCT method learns much faster than the traditional BW algorithm, with a speedup around 180 to 260 times. The BW algorithm tends to overfit the model by using a larger number of states, resulting in a higher parameter distance. Even when the true number of states Q is given, for persistent cases, the BW still cannot learn correctly, which has a larger test-set log-likelihood difference and a lower convergence and identification rate.

Learning results for a 10-state persistent model, a 10-state transient-cyclic model, and a 6-state hybrid model are used as examples for visualization. For the persistent model, the iterative learning process is shown in Figures 6(a) and 6(b). In Figure 6(a), the BW training was conducted with an unknown number of states Q, while in Figure 6(b), Q was given. Similarly, results for the transient-cyclic model are shown in Figures 7(a) and 7(b). and the hybrid models are in Figures 8(a) and 8(b). The figures show that the proposed SCT method starts from a good initial model at the beginning and converges much faster than most of the 20 randomly initialized BW models which start from an almost equivalent level of the multinomial model for which hidden states play no roles. Moreover, although some of the best BW model converges in the end, the log-likelihood values are still not as good as the SCT method. We see that some of the repeated 20 models with Q states have been stuck in a local optimum with similar log-likelihood to model $\overline{\lambda}_{Q-1}(\lambda_Q)$ or $\overline{\lambda}_{Q-2}(\lambda_Q)$.

In order to compare the model parameters, heat maps of the original and inferred state transition and observation matrices are plotted in Figures 6(c), 7(c), and 8(c). A lighter color indicates a higher probability value close to one, while a darker color indicates a lower probability value close to zero. We notice that the BW method with an unknown Q learns a complex model with two more states than the true model for all the three cases, which is overfitting, while the SCT approach learns the state size correctly. Moreover, the SCT method has a one-to-one correspondence of the high probabilities (in white/light-yellow) between the transition and observation parameter matrices, meaning the SCT trained model is almost equivalent to the reference true model. However, for the BW method, especially when Q is unknown, there are no one-to-one relations in both transition and observation matrices, noticeable by some of the varied colors of heat maps from the true model. It means that some of the probabilities are wrongly learned.

5.3. Hard to Learn HMMs. For each of the seven hard to learn conditions defined in Section 3.4, we construct five ground truth models, resulting in a total of 35 models. For each model, persistent and transient-cyclic state numbers are randomly generated from a range of $[3, 5]$. The privileged state probability is set randomly within a range of $[0.85, 0.99]$. The remaining probabilities are uniformly distributed. For conditions (ii), (iii), (iv), and (v), one extra state is generated accordingly to the specified conditions. For condition (i), state 3 is defined as a mixing state of states 1 and 2. For condition (vi), the first two states are set to have the same observations. For condition (vii), state 2 is set to have a constant observation emission probabilities. The rest of the experiment is set the same as previous experiments designed for identifiable models. Results show that only 31% of the 35 ground truth models are specific with an average specificity of 0.04. A detailed comparison of the learning results is presented in Table 5.

From the results in Table 5 we can see that the BW algorithm is slower than the proposed SCT method with almost double learning convergence iterations. The SCT method is around 230 times faster than the BW algorithm. A positive average delta Q indicates that the BW method mostly overfits the true models with an average of 1.71 extra states, while the SCT has a negative delta Q indicating a slightly underfitting with an average of 0.66 fewer states. Even though the test-set log-likelihood difference of the SCT is higher than the BW method, the average parameter distance further proves that the BW algorithm tends to overfit the models in order to have a lower log-likelihood. Moreover, the percentage of convergence reveals that the number of repetitions (e.g., 20 times in this experiment) is still necessary for the BW method to learn effectively, even with the trade-off of longer learning time. Lastly, the SCT has a slightly lower but compatible identification percentage which has a significant learning speedup in return.

To visualize the results, we select two models under condition (i), a state being a mixing state, and condition (v), a state with constant self-excluded outgoing transaction, as examples shown in Figures 9 and 10, respectively.

Figures 9 and 10 show that the BW algorithm with an unknown Q overfits the truth models with two extra states while the SCT method underfits the model with one state fewer where both the mixing state and the state with the same outgoing transactions in the two examples are merged into other states because they are not specific enough to be identified.

5.4. Random HMMs. In this experiment, we generated 10000 random HMM models configured with a combination of random Q ($Q \in [3, 5]$) and random M ($M \in [3, 5]$). In order to guarantee that each HMM is minimal, we select models according to two criteria: (1) the model should have a higher test-set log-likelihood than the one of a multinomial model; (2) the model compared to the best $Q - 1$ state model should not satisfy the three-sigma rule for model equivalence criteria

Table 4: Identification results on identifiable HMMs.

Type	Method	# Iters.	Time (s)	Q select	ΔQ	ΔLL_{test}	Para. Dist.	Conv. (%)	Identi. (%)
Pers.	BW	15	1054	Min. AIC	1.67	0.028	0.15	45.56	88.89
				Correct Q	0.00	0.058	0.02	16.11	66.67
	SCT	4	4	DBI	0.00	0.012	0.00	—	100
Tran.	BW	22	1276	Min. AIC	1.00	0.014	0.08	100	100
				Correct Q	0.00	0.012	0.01	75.56	100
	SCT	8	5	DBI	0.00	0.011	0.01	—	100
Hybr.	BW	19	891	Min. AIC	1.75	0.008	0.19	88	100
				Correct Q	0.00	0.007	0.00	36.50	100
	SCT	10	5	DBI	0.00	0.007	0.00	—	100

Iters.: average number of iterations; Conv. (%): rate of convergence; Identi. (%): percentage of identification; ΔLL_{test}: unit log-likelihood difference between the true models and the learned model on test-sets; Para. Dist.: parameter distance; Pers.: Persistent; Tran.: Transient-cyclic; Hybr.: Hybrid. Note that, for the BW, when calculating ΔQ, ΔLL_{test}, and Para. Dist., the learned model is the best one selected from the $20(Q + 1)$ repeated random models.

Table 5: Identification results on hard to learn HMMs.

Method	# Iters.	Time (s)	Q select	ΔQ	ΔLL_{test}	Para. Dist.	Conv. (%)	Identi. (%)
BW	26	1835	Min. AIC	1.71	0.014	0.14	87.14	100
			Correct Q	0.00	0.013	0.03	52.00	100
SCT	15	8	DBI	−0.66	0.030	0.11	—	97.14

See abbreviations and notes in Table 4 for more details.

defined in Definition 8. A random HMM is discarded if it is not minimal. In the end, we obtain 149 specific minimal HMMs. The training procedure is conducted in the same way as in Section 5.2.

Experiment shows that the average specificity of the true models is 0.03, which is around 10 times less specific than the identifiable models used in the previous experiments in Section 5.2. Moreover, the mean of the log-likelihood distribution of the true model is 1.58, which is also much higher than the identifiable models. The above results indicate that random models are less specific and therefore less identifiable. A detailed comparison of the identification results is shown in Table 6.

The results show that the SCT method needs in average one more iteration than the BW algorithm and the identification results are less adequate because the models are not specific enough to be estimated correctly. However, the speedup of the SCT method shows an improvement *vis-a-vis* the Baum-Welch method, around 50 times. Both of the approaches overfit the models with an average of more than one state.

Figure 11(a) provides the dependence between true model *specificity* and test-set *log-likelihood difference* with the true models. When the specificity is too low, the SCT method identifies less correctly the models. Thus, the less specific the model is, the harder it becomes for the SCT method to learn. For the purpose of comparison, we plot the same figure in Figure 11(b) but for highly specific models which were generated in Section 5.2. The results further confirm that, for highly specific models, when the specificity is relatively low, the SCT method outperforms the BW method. The log-likelihood differences of the models learned by Baum-Welch have significantly increased indicating that completely wrong models are learned.

The results are expected because the SCT method is designed for highly specific models but not for random ones with less specificity. In order to see the influence of specificity for the SCT method to learn correctly, we plot the identification accuracy versus the specificity thresholds ranging from −0.01 to 0.2 with a step of 0.01 as shown in Figure 12. The models are selected when they have a specificity higher than a specificity threshold; then the percentage of correctly identified models within the selected models is used as the identification accuracy.

The identification accuracy of the SCT method starts with a low value of 87.9% and generally increases with an increase specificity threshold. When the specificity threshold is at 0.06, the identification percentage of the SCT drops to 93.8%. It is caused by a single case which is observable in Figure 11(a) with the highest log-likelihood difference. Such case cannot represent the dependency trend between specificity threshold and identification accuracy and thus can be ignored. When the threshold is higher than 0.06, the proposed SCT method converges to an identification of 100%.

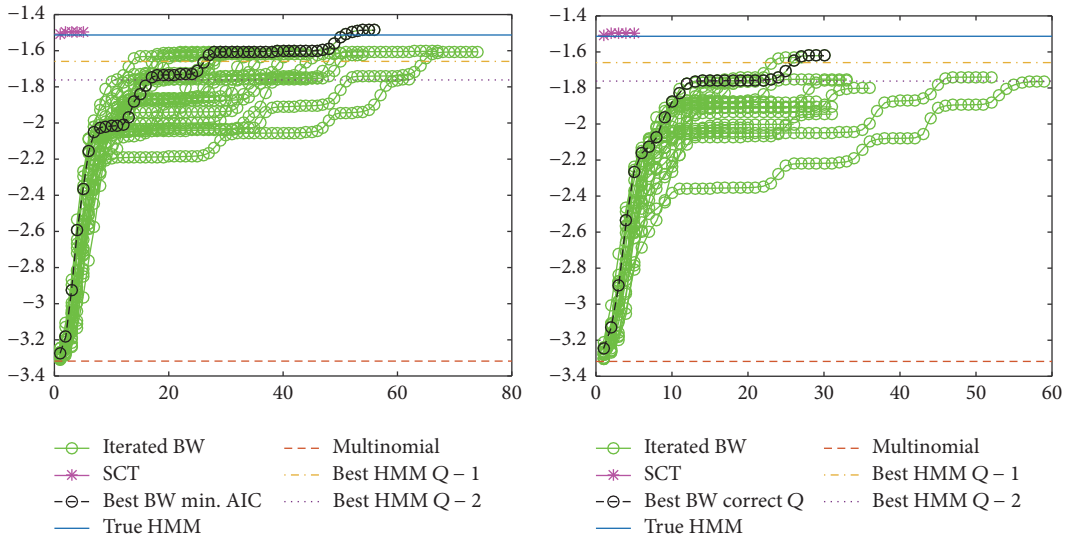
6. Conclusions

This paper studied the possibility of identifying HMMs from properties of the observation sequences directly. We conducted an analysis of the information flow throughout an HMM. Based on this analysis we were able to show that there are two types of states, namely, persistent and transient, that have a high impact on the observation likelihood. An HMM consisting of high-impact states is highly specific, in the sense that it differs substantially in observation likelihood from the best HMM with one state less.

Table 6: Identification results on random minimal HMMs.

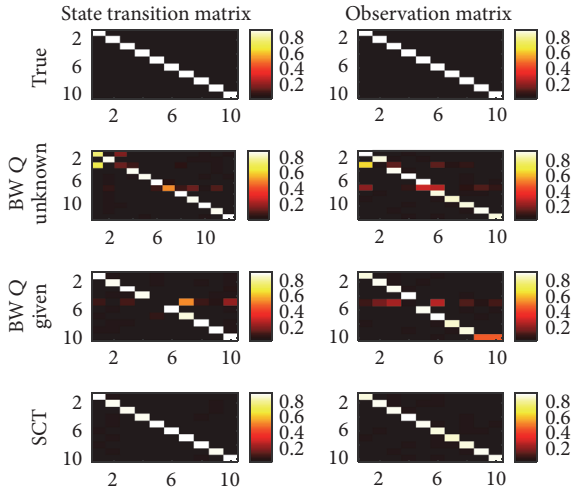
Method	# Iters.	Time (s)	Q select	ΔQ	ΔLL_{test}	Para. Dist.	Conv. (%)	Identi. (%)
BW	17	348	Min. AIC	1.16	0.0019	0.21	93.83	100
			Correct Q	0.00	0.0018	0.05	46.64	100
SCT	18	7	DBI	1.70	0.0051	0.22	—	87.92

See abbreviations and notes in Table 4 for more details.



(a) Log-likelihoods during iterative training, BW with unknown Q

(b) Log-likelihoods during iterative training, BW given correct Q



(c) Heatmap of HMM model parameters

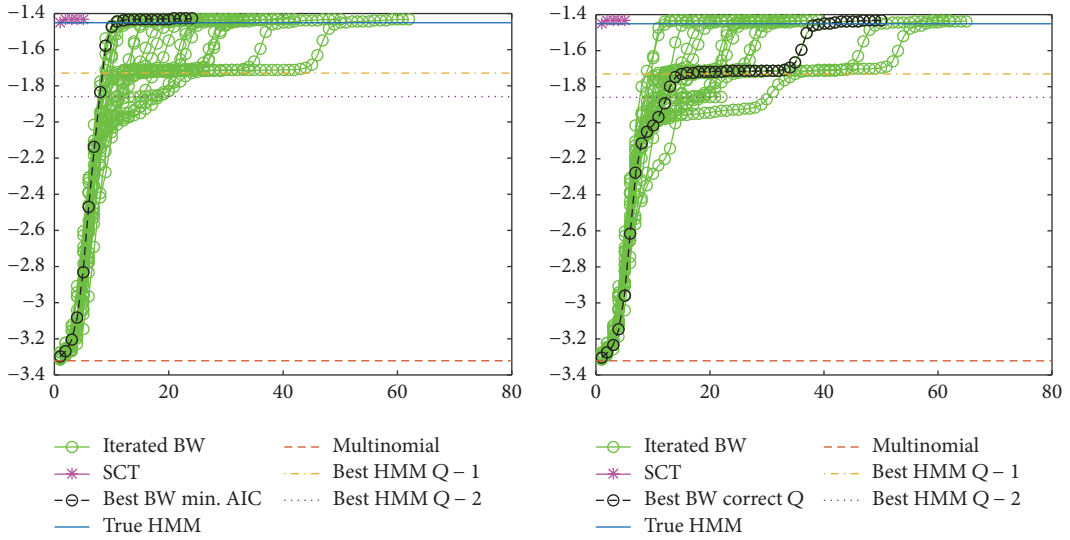
Figure 6: Identification performance for a 10-state discrete persistent HMM. (a) and (b) show a comparison of log-likelihoods during iterative trainings with different models: the 20 repetitive BW models with an unknown or known number of states Q, the SCT models, the selected best BW models without or with a given Q, the ground truth HMM, the multinomial model, the best one-state simpler model $\overline{\lambda}_{Q-1}(\lambda_Q)$, and the best two-state simpler model $\overline{\lambda}_{Q-2}(\lambda_Q)$. (c) shows a comparison of the model parameter heat maps for the ground truth HMM, the best BW models with an unknown or a given Q, and the SCT model.

A learning algorithm, called SCT, was constructed based on this analysis which correctly identifies highly specific models. But even for low-specific models, the identification accuracy is still around 88%. The algorithm is about two orders of magnitude faster than the traditional Baum-Welch algorithm.

Appendix

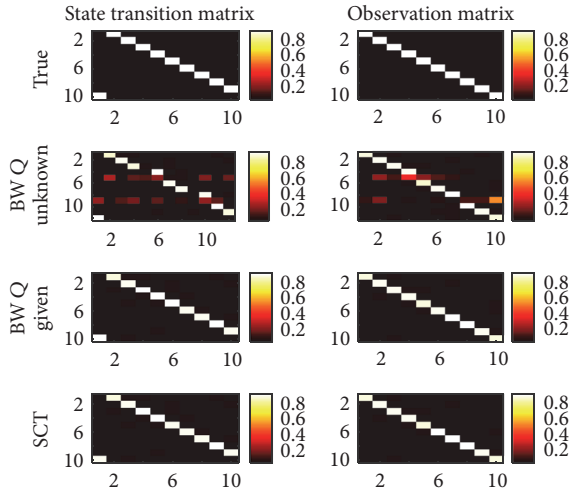
A. Proof of Theorem 9

We prove that the presence of transient-cyclic states with dominant observations can be identified through the division



(a) Log-likelihoods during iterative training, BW with unknown Q

(b) Log-likelihoods during iterative training, BW given correct Q



(c) Heatmap of HMM model parameters

FIGURE 7: Identification performance for a 10-state discrete transient-cyclic HMM. See caption of Figure 6 for more details.

$\hbar$ defined in (14) under the conditions as follows. Note that we consider that the relative frequency $\overline{P}$ is close to the true probability P such that the following derivations apply:

(i) *If* $\hbar < 1 - \epsilon, \epsilon \approx 0$, *there are only states with dominant observations.*

One type of HMM cases is the basic transient, cyclic model with dominant and privileged observation value. Without loss of generality, we assume v_k, v_l, and v_m are dominant and privileged observation values for states s_i, s_{i+1}, and s_{i+2} and that the transition cycle is $1 \rightarrow 2 \rightarrow 3, \ldots, \rightarrow Q \rightarrow 1$. So for the emission probabilities,

$$P\left(o_t = v_k \mid q_t = s_i\right) \gg P\left(o_t = v_l \mid q_t = s_i\right), \quad \forall j \neq i \quad \text{(A.1)}$$

or

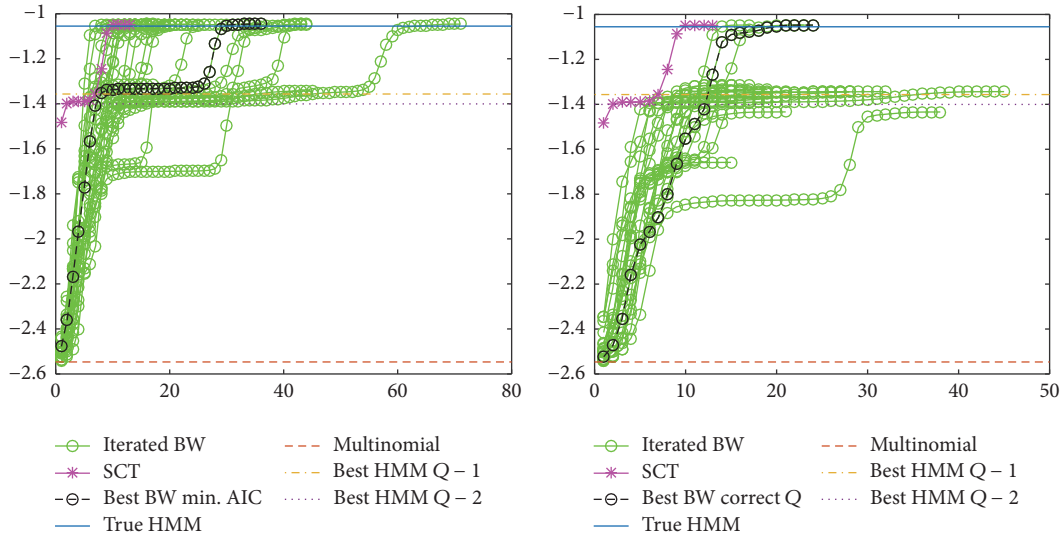
$$P\left(o_t = v_k \mid q_t = s_i\right) \gg P\left(o_t = v_k \mid q_t = s_j\right), \quad \forall j \neq i \quad \text{(A.2)}$$

and for the state transition probabilities,

$$P\left(q_{t+1} = s_{i+1} \mid q_t = s_i\right) \gg P\left(q_{t+1} = s_j \mid q_t = s_i\right), \quad \forall j \neq i+1. \quad \text{(A.3)}$$

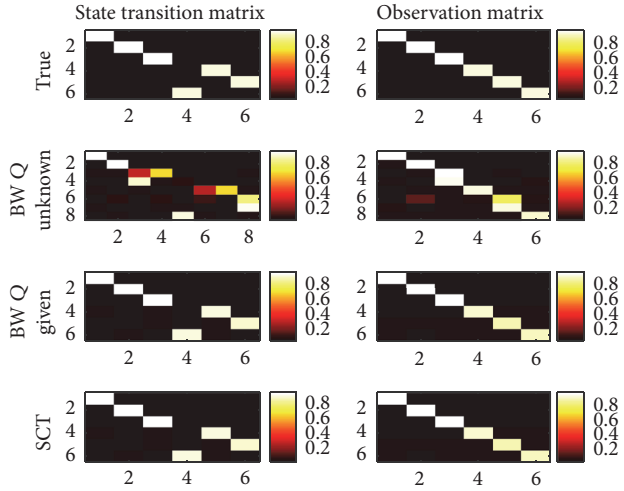
Note that if $i = Q$, we have $s_{i+1} = s_1$. For cyclic indices, operations are always followed by a modulo operation.

We assume that the probabilities (i.e., transition and observation probabilities) can be split into two



(a) Log-likelihoods during iterative training, BW with unknown Q

(b) Log-likelihoods during iterative training, BW given correct Q



(c) Heat map of HMM model parameters

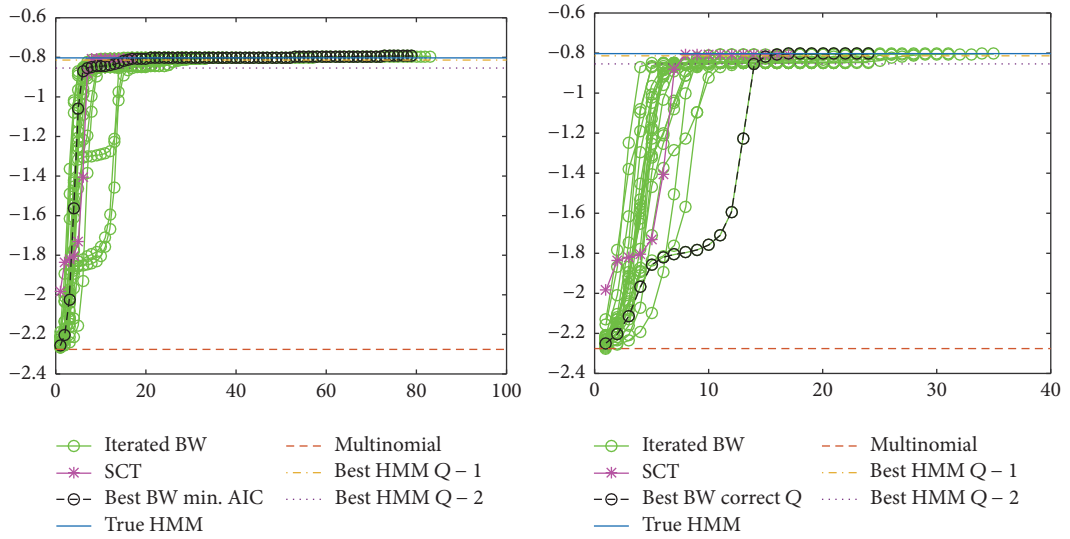
FIGURE 8: Identification performance for a 7-state discrete hybrid HMM. See the caption of Figure 6 for more details.

groups: large and small probabilities. There is a large deviation between both; that is, large probabilities are much higher than small probabilities. For instance, large and small probabilities are around 0.9 and 0.1, respectively. Thus for the case addressed previously, $P(q_{t+1} = s_{i+1} \mid q_t = s_i)$ and $P(o_t = v_k \mid q_t = s_i)$ are large probabilities, which we denote as a and c, respectively, for simplicity. Similarly, $P(q_{t+1} = s_j \mid q_t = s_i)$ and $P(o_t = v_l \mid q_t = s_i)$ are denoted as b and d, respectively. Thus $a \gg b$ and $c \gg d$. Moreover, we assume that $P(o_{t+1} = v_l) = P(q_t = s_j)$, $\forall 1 \le i \le Q$, $1 \le j \le M$ and with $P(q_t = s_j, o_{t+1} = v_l) = P(q_t = s_j \mid o_{t+1} = v_l)P(o_{t+1} = v_l) = P(o_{t+1} = v_l \mid q_t = s_j)P(q_t = s_j)$, we have

$$P\left(q_t = s_j \mid o_{t+1} = v_l\right) = P\left(o_{t+1} = v_l \mid q_t = s_j\right). \quad \text{(A.4)}$$

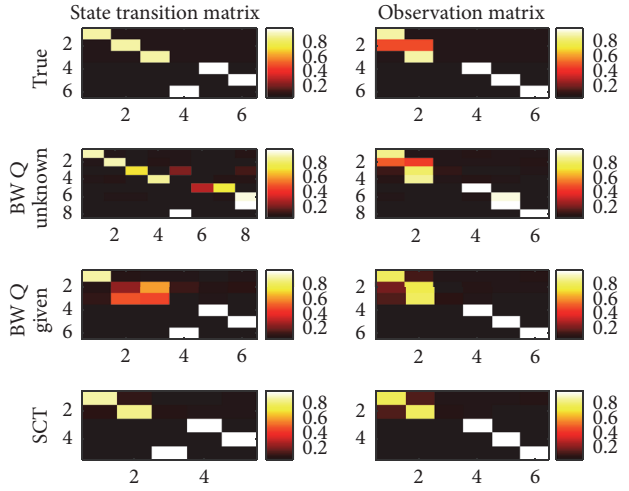
With (A.1)–(A.4), the following holds:

$$\begin{aligned} P\left(o_{t+1} = v_l \mid o_t = v_k\right) &= \sum_{i,j\neq i} P\left(o_t = v_k \mid q_t = s_i\right) \\ &\cdot P\left(q_{t+1} = s_j \mid q_t = s_i\right) P\left(o_{t+1} = v_l \mid q_{t+1} = s_j\right) \\ &= P\left(o_t = v_k \mid q_t = s_i\right) P\left(q_{t+1} = s_{i+1} \mid q_t = s_i\right) \\ &\cdot P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right) \\ &+ \sum_{i,\tilde{j}\neq i+1} P\left(o_t = v_k \mid q_t = s_i\right) P\left(q_{t+1} = s_{\tilde{j}} \mid q_t = s_i\right) \\ &\cdot P\left(o_{t+1} = v_l \mid q_{t+1} = s_{\tilde{j}}\right) \approx P\left(o_t = v_k \mid q_t = s_i\right) \\ &\cdot P\left(q_{t+1} = s_{i+1} \mid q_t = s_i\right) P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right) \\ &= ac^2, \end{aligned} \quad \text{(A.5)}$$



(a) Log-likelihoods during iterative training, BW with unknown Q

(b) Log-likelihoods during iterative training, BW given correct Q



(c) Heatmap of HMM model parameters

Figure 9: Identification performance for a 6-state hard to learn HMM under condition (i). See caption of Figure 6 for more details.

$$P(o_t = v_k,\ o_{t+1} = v_l) = P(o_t = v_k)\,P(o_{t+1} = v_l \mid o_t = v_k) = ac^2 P(o_t = v_k). \tag{A.6}$$

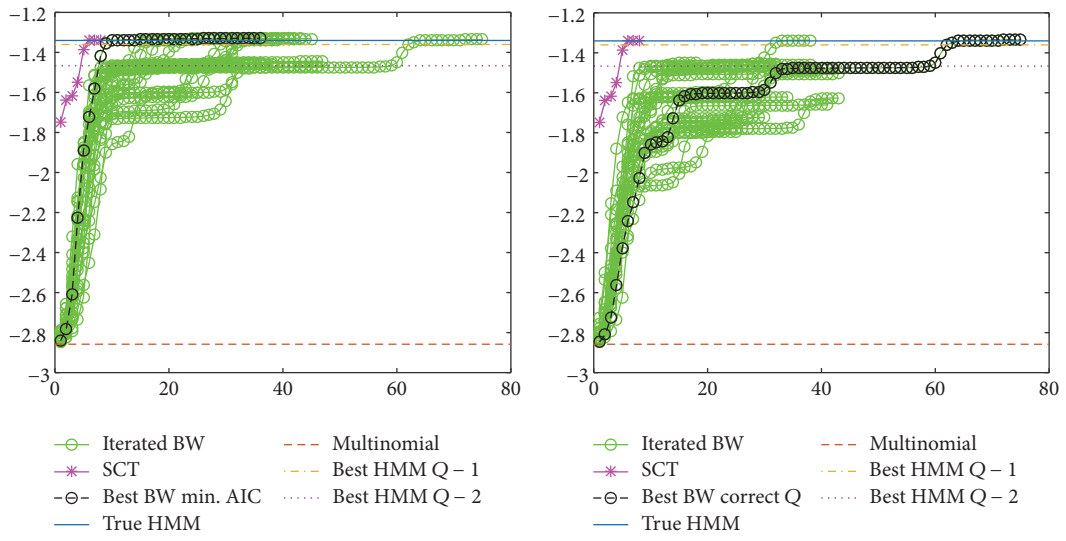
The approximation in (A.5) holds because the terms of the sum for $P(q_{t+1} = s_{\tilde{j}} \mid q_t = s_i)$ and $P(o_{t+1} = v_l \mid q_{t+1} = s_{\tilde{j}})$, $\forall \tilde{j} \neq i$ are small probability factors. Since $P(o_{t+1} = v_l \mid o_t = v_k)$ is two orders lower, it follows that $P(o_{t+1} = v_l \mid o_t = v_k) < P(q_{t+1} = s_{i+1} \mid q_t = s_i)$.

If we assume and train the observations with a first-order Markov model, then we have

$$P(o_t = v_k,\ o_{t+1} = v_l,\ o_{t+2} = v_m) = P(o_t = v_k) \cdot P(o_{t+1} = v_l \mid o_t = v_k)\,P(o_{t+2} = v_m \mid o_{t+1} = v_l) \tag{A.7}$$

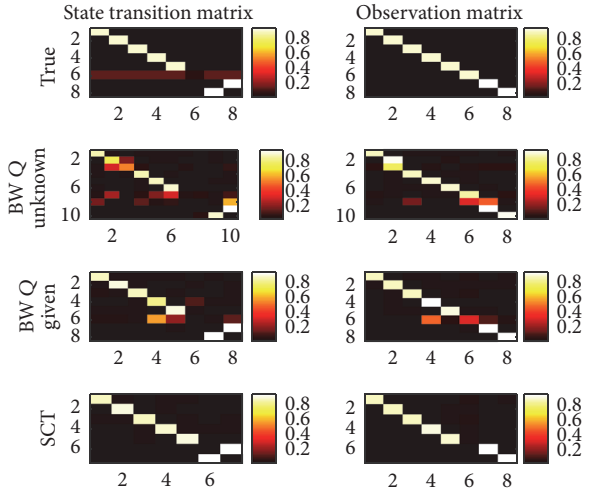
which can be approximated by a first-order HMM as shown in Figure 13(a); thus

$$P(o_t = v_k,\ o_{t+1} = v_l,\ o_{t+2} = v_m) \approx P(o_t = v_k) \cdot [P(o_t = v_k \mid q_t = s_i)\,P(q_{t+1} = s_{i+1} \mid q_t = s_i) \cdot P(o_{t+1} = v_l \mid q_{t+1} = s_{i+1})], \quad [P(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}) \cdot P(q_{t+2} = s_{i+2} \mid q_{t+1} = s_{i+1}) \cdot P(o_{t+2} = v_m \mid q_{t+2} = s_{i+2})] = P(o_t = v_k)(acc) \cdot (acc) = a^2c^4 P(o_t = v_k). \tag{A.8}$$



(a) Log-likelihoods during iterative training, BW with unknown Q

(b) Log-likelihoods during iterative training, BW given correct Q



(c) Heatmap of HMM model parameters

Figure 10: Identification performance for an 8-state hard to learn HMM under condition (v). See caption of Figure 6 for more details.

If we assume and train the observations with a second-order Markov model, the following holds:

$$
\begin{aligned}
&P\left(o_t = v_k,\ o_{t+1} = v_l,\ o_{t+2} = v_m\right) \\
&\quad = \sum_{i,j,k,i\neq j\neq k} P\left(q_t = s_i,\ q_{t+1} = s_j,\ q_{t+2} = s_k\right) \\
&\quad \cdot P\left(o_i = v_k \mid q_t = s_i\right), \\
&P\left(o_{t+1} = v_l \mid q_{t+1} = s_j\right) P\left(o_{t+2} = v_m \mid q_{t+2} = s_k\right) \\
&\quad = P\left(q_t = s_i\right) P\left(q_{t+1} = s_{i+1} \mid q_t = s_i\right) P\left(q_{t+2}\right. \\
&\quad \left. = s_{i+2} \mid q_{t+1} = s_{i+1}\right) P\left(o_t = v_k \mid q_t = s_i\right), \\
&P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right) P\left(o_{t+2} = v_m \mid q_{t+2} = s_{i+2}\right) \\
&\quad + \sum_{i,\tilde{j},\tilde{k},\tilde{j}\neq i+1,\tilde{k}\neq i+2} P\left(q_t = s_i,\ q_{t+1} = s_{\tilde{j}},\ q_{t+1} = s_{\tilde{k}}\right), \\
&P\left(o_t = v_k \mid q_t = s_i\right) P\left(o_{t+1} = v_l \mid q_{t+1} = s_{\tilde{j}}\right) P\left(o_{t+2}\right. \\
&\quad \left. = v_m \mid q_{t+2} = s_{\tilde{k}}\right)
\end{aligned}
\tag{A.9}
$$

which can be approximated by a second-order HMM as shown in Figure 13(b); thus

$$
\begin{aligned}
&P\left(o_t = v_k,\ o_{t+1} = v_l,\ o_{t+2} = v_m\right) \approx P\left(q_t = s_i\right) \\
&\quad \cdot P\left(q_{t+1} = s_{i+1} \mid q_t = s_i\right) \\
&\quad \cdot P\left(q_{t+2} = s_{i+2} \mid q_{t+1} = s_{i+1}\right)
\end{aligned}
$$

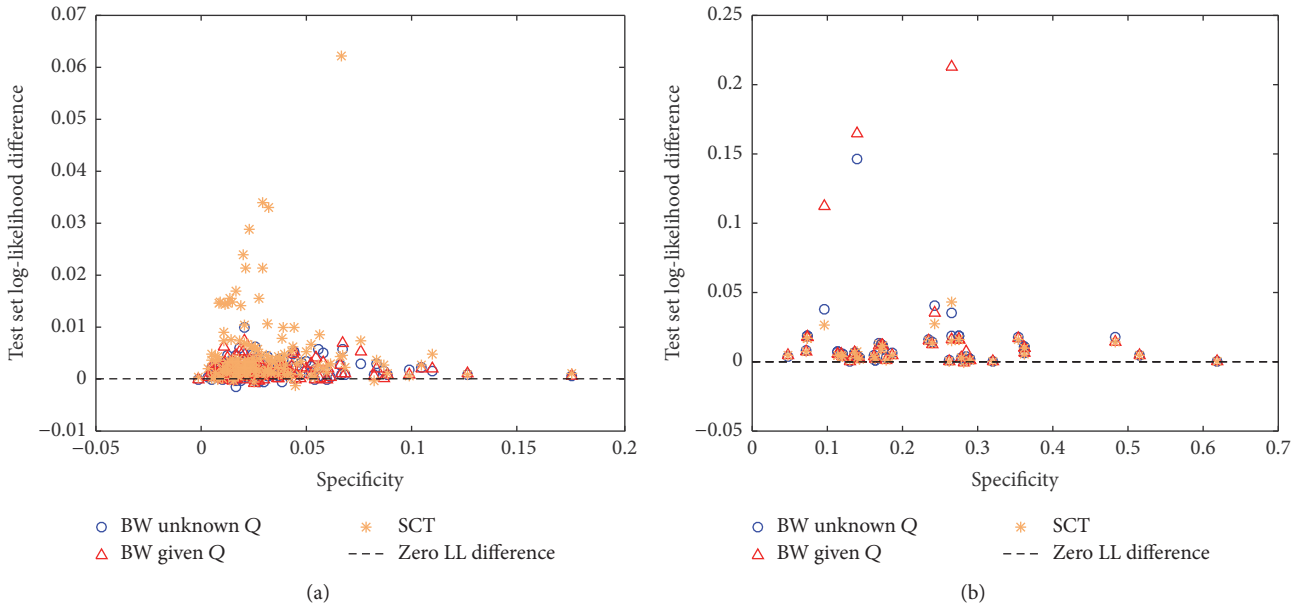


Figure 11: Specificity versus log-likelihood difference with ground truth model. (a) Random specific models generated in Section 5.4. (b) Highly specific models generated in Section 5.2.

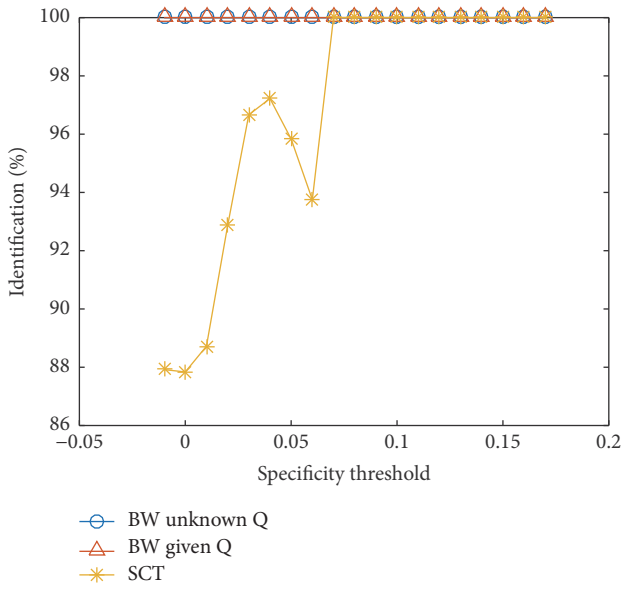


Figure 12: Correlation between specificity threshold and identification accuracy.

$$\cdot P\left(o_t = v_k \mid q_t = s_i\right),$$

$$P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right) P\left(o_{t+2} = v_m \mid q_{t+2} = s_{i+2}\right) = a^2 c^3 P\left(o_t = v_k\right). \tag{A.10}$$

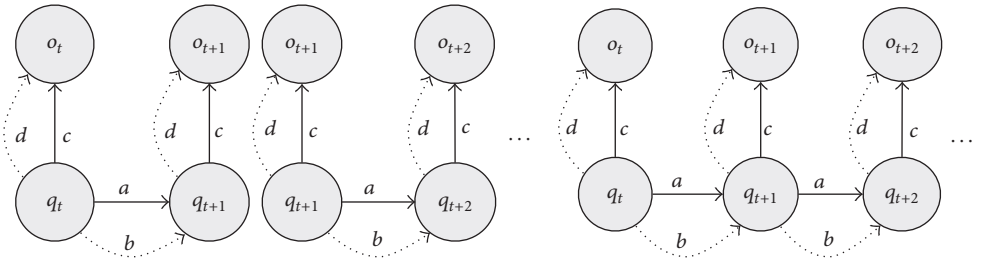
The first-order HMM assumption counts twice of emission probability $c = P(o_{t+1} = v_l \mid q_{t+1} = s_{i+1})$; that is, an larger probability factor of $c < 1$ is calculated in (A.8) than in (A.10). Thus the division $\hbar = c < 1$ and the calculated probability with a second-order HMM assumption in (A.10) is higher than that with a first-order HMM assumption in (A.8).

(ii) *If $\hbar > 1 + \epsilon, \epsilon \approx 0$, there are states with dominant observations and an extra mixing state.*

Another type of HMM cases is the basic transient, cyclic model with mostly dominant and privileged observation value, but also with mixing observations. For demonstration purpose, we assume there exists one mixing state s_{i+1} in the model, which emits observations v_k and v_l with equal probability x, where $x \leq 0.5$. We call x a medium probability because it is close to or equal to a probability of 0.5.

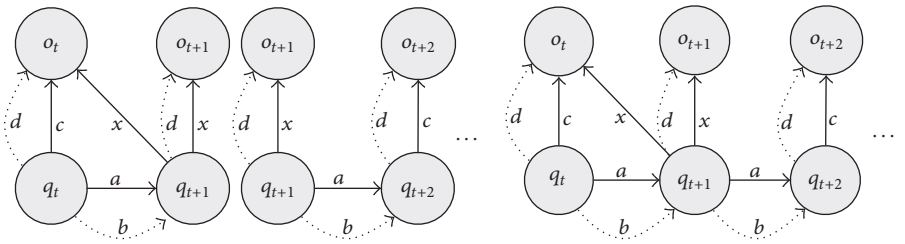
Thus the first-order Markov model assumption holds:

$$\begin{aligned} P\left(o_{t+1} = v_l \mid o_t = v_k\right) &= \sum_{i,j \neq i} P\left(o_t = v_k \mid q_t = s_i\right) \\ &\cdot P\left(q_{t+1} = s_j \mid q_t = s_i\right) P\left(o_{t+1} = v_l \mid q_{t+1} = s_j\right) \\ &= P\left(o_t = v_k \mid q_t = s_i\right) P\left(q_{t+1} = s_{i+1} \mid q_t = s_i\right) \\ &\cdot P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right) + P\left(o_t = v_k \mid q_{t+1}\right. \\ &= \left.s_{i+1}\right) P\left(q_{t+2} = s_{i+2} \mid q_{t+1} = s_{i+1}\right) P\left(o_{t+2}\right. \\ &= \left.v_m \mid q_{t+2} = s_{i+2}\right) \\ &+ \sum_{i,\tilde{j} \neq i+1, \tilde{j} \neq i+2} P\left(o_t = v_k \mid q_t = s_i\right) \\ &\cdot P\left(q_{t+1} = s_{\tilde{j}} \mid q_t = s_i\right) P\left(o_t = v_k \mid q_{t+1} = s_{i+1}\right), \end{aligned}$$



(a) Approximation by a first-order HMM assumption (b) Approximation by a second-order HMM assumption

Figure 13: Probability approximation with large probabilities without mixing states.



(a) Approximation by a first-order HMM assumption (b) Approximation by a second-order HMM assumption

Figure 14: Probability approximation with large probabilities with a mixing state.

$$\begin{aligned}
&P\left(o_t = v_k \mid q_{t+1} = s_{\tilde{j}}\right) P\left(o_{t+1} = v_l \mid q_{t+1} = s_{\tilde{j}}\right) \\
&\quad \approx P\left(o_t = v_k \mid q_t = s_i\right) P\left(q_{t+1} = s_{i+1} \mid q_t = s_i\right) \\
&\quad \cdot P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right) + P\left(o_t = v_k \mid q_{t+1}\right. \\
&\quad \left. = s_{i+1}\right) P\left(q_{t+2} = s_{i+2} \mid q_{t+1} = s_{i+1}\right) P\left(o_{t+2}\right. \\
&\quad \left. = v_m \mid q_{t+2} = s_{i+2}\right) = 2cax.
\end{aligned} \tag{A.11}$$

Similarly,

$$\begin{aligned}
&P\left(o_{t+2} = v_m \mid o_{t+1} = v_l\right) \\
&\quad = \sum_{i,j\neq i} P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right) \\
&\quad \cdot P\left(q_{t+2} = s_{j+1} \mid q_{t+1} = s_{i+1}\right) \\
&\quad \cdot P\left(o_{t+2} = v_m \mid q_{t+2} = s_{j+1}\right) = P\left(o_{t+1} = v_l \mid q_{t+1}\right. \\
&\quad \left. = s_{i+1}\right) P\left(q_{t+2} = s_{i+2} \mid q_{t+1} = s_{i+1}\right) P\left(o_{t+2}\right. \\
&\quad \left. = v_m \mid q_{t+2} = s_{i+2}\right) \\
&\quad + \sum_{i,\tilde{j}\neq i+2} P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right) \\
&\quad \cdot P\left(q_{t+2} = s_{\tilde{j}} \mid q_{t+1} = s_{i+1}\right) \\
&\quad \cdot P\left(o_{t+2} = v_m \mid q_{t+2} = s_{\tilde{j}}\right) \approx P\left(o_{t+1} = v_l \mid q_{t+1}\right. \\
&\quad \left. = s_{i+1}\right) P\left(q_{t+2} = s_{i+2} \mid q_{t+1} = s_{i+1}\right) P\left(o_{t+2}\right. \\
&\quad \left. = v_m \mid q_{t+2} = s_{i+2}\right) = cac.
\end{aligned} \tag{A.12}$$

If we train with a first-order Markov model with (A.11) and (A.12), it can be approximated by a first-order HMM as shown in Figure 14(a); then we have

$$\begin{aligned}
&P\left(o_t = v_k,\ o_{t+1} = v_l,\ o_{t+2} = v_m\right) = P\left(o_t = v_k\right) \\
&\quad \cdot P\left(o_{t+1} = v_l \mid o_t = v_k\right) P\left(o_{t+2} = v_m \mid o_{t+1} = v_l\right) \\
&\quad \approx P\left(o_t = v_k\right) \left[P\left(o_t = v_k \mid q_t = s_i\right)\right. \\
&\quad \cdot P\left(q_{t+1} = s_{i+1} \mid q_t = s_i\right) \\
&\quad \cdot P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right) \\
&\quad + P\left(o_t = v_k \mid q_{t+1} = s_{i+1}\right) \\
&\quad \cdot P\left(q_{t+2} = s_{i+2} \mid q_{t+1} = s_{i+1}\right) \\
&\quad \left. \cdot P\left(o_{t+2} = v_m \mid q_{t+2} = s_{i+2}\right)\right], \\
&\left[P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right)\right. \\
&\quad \cdot P\left(q_{t+2} = s_{i+2} \mid q_{t+1} = s_{i+1}\right) \\
&\quad \left. \cdot P\left(o_{t+2} = v_m \mid q_{t+2} = s_{i+2}\right)\right] = P\left(o_t = v_k\right) \\
&\quad \cdot (2cax)(xac) = 2xa^2c^3 P\left(o_t = v_k\right).
\end{aligned} \tag{A.13}$$

For a second-order HMM assumption, shown in Figure 14(b), it holds:

$$\begin{aligned}
&P\left(o_t = v_k,\ o_{t+1} = v_l,\ o_{t+2} = v_m\right)\\
&\quad = \sum_{i,j,k,i\neq j\neq k} P\left(q_t = s_i,\ q_{t+1} = s_j,\ q_{t+2} = s_k\right)\\
&\quad \cdot P\left(o_i = v_k \mid q_t = s_i\right) P\left(o_{t+1} = v_l \mid q_{t+1} = s_j\right),\\
&P\left(o_{t+2} = v_m \mid q_{t+2} = s_k\right) = P\left(q_t = s_i\right) P\left(q_{t+1}\right.\\
&\quad \left. = s_{i+1} \mid q_t = s_i\right) P\left(q_{t+2} = s_{i+2} \mid q_{t+1} = s_{i+1}\right) P\left(o_t\right.\\
&\quad \left. = v_k \mid q_t = s_i\right),\\
&P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right) P\left(o_{t+2} = v_m \mid q_{t+2} = s_{i+2}\right)\\
&\quad + \sum_{i,\tilde{j},\tilde{k},\tilde{j}\neq i+1,\tilde{k}\neq i+2} P\left(q_t = s_i,\ q_{t+1} = s_{\tilde{j}},\ q_{t+1} = s_{\tilde{k}}\right),\\
&P\left(o_t = v_k \mid q_t = s_i\right) P\left(o_{t+1} = v_l \mid q_{t+1} = s_{\tilde{j}}\right) P\left(o_{t+2}\right.\\
&\quad \left. = v_m \mid q_{t+2} = s_{\tilde{k}}\right) \approx P\left(q_t = s_i\right) P\left(q_{t+1} = s_{i+1} \mid q_t\right.\\
&\quad \left. = s_i\right) P\left(q_{t+2} = s_{i+2} \mid q_{t+1} = s_{i+1}\right) P\left(o_t = v_k \mid q_t\right.\\
&\quad \left. = s_i\right),\\
&P\left(o_{t+1} = v_l \mid q_{t+1} = s_{i+1}\right) P\left(o_{t+2} = v_m \mid q_{t+2} = s_{i+2}\right)\\
&\quad = P\left(o_t = v_k\right) aacxc = xa^2c^2 P\left(o_t = v_k\right).
\end{aligned} \tag{A.14}$$

Since a and c are both large probabilities, the division of (A.13) and (A.14) (i.e., the emission probability of the mixing case) is $\hbar = 2c$, greater than 1. Therefore, when there is a mixing case, calculations with a first-order HMM assumption are larger and thus can be used to distinguish cases with mixing cases to the cases without.

B. Proof of Theorem 10

We prove that a stationary HMM $\lambda = (\boldsymbol{\pi}, Q, M, \mathbf{A}, \mathbf{B})$ can be reduced to an equivalent simpler HMM $\widetilde{\lambda} = (\widetilde{\boldsymbol{\pi}}, \widetilde{Q}, \widetilde{M}, \widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$; that is, $P_\lambda(\mathbf{o}_{1:t}) = P_{\widetilde{\lambda}}(\mathbf{o}_{1:t})$, where $o_t = v_k \in V$ and $\widetilde{Q} = Q - 1$ if any of cases defined in Theorem 10 occurs.

(i) The state r has zero incoming probabilities; that is, $a_{ir} = 0,\ \forall i \in [1, Q]$, such that $\forall t \in [1, T],\ \tau_t(r) = 0,\ r \in [1, Q]$. The state r has no influence on $P_\lambda(\mathbf{o}_{1:t})$; thus r can be removed.

$$\begin{aligned}
P_\lambda\left(\mathbf{o}_{1:t}\right) &= \boldsymbol{\tau}_t^T \mathbf{B}_{o_t} \mathbf{e}, \quad \forall \tau_t(i),\ i \in [1, Q]\\
&= \widetilde{\boldsymbol{\tau}}_t^T \widetilde{\mathbf{B}}_{o_t} \mathbf{e} + \tau_t(r) b_{rk},\\
&\qquad \forall \widetilde{\tau}_t(j),\ j \in [1, Q] \setminus \{r\}\\
&= P_{\widetilde{\lambda}}\left(\mathbf{o}_{1:t}\right).
\end{aligned} \tag{B.1}$$

(ii) Suppose state r and q have equal incoming and outgoing transition probabilities; that is, $a_{ir} = a_{iq}$ and $a_{ri} = a_{qi}$, where $i \in [1, Q]$. With the same incoming probabilities, we also have

$$\tau_t(r) = \tau_t(q), \quad r \in [1, Q],\ \forall t \in [1, T]; \tag{B.2}$$

thus r and q can be merged into a single state l in $\widetilde{\lambda}$. The information flow of both states should remain equal after merging; thus for the merged state l

$$\begin{aligned}
&\widetilde{a}_{il} = a_{ir} + a_{iq} = 2a_{ir}, \quad \forall i \in [1, Q] \cup l \setminus \{r, q\},\\
&\widetilde{a}_{li} = a_{ri} = a_{qi}, \quad \forall i \in [1, Q] \setminus \{r, q\},\\
&\widetilde{b}_{lk} = \frac{b_{rk} + b_{qk}}{2},\\
&\widetilde{\tau}_t(l) = \tau_t(r) + \tau_t(q) = 2\tau_t(r).
\end{aligned} \tag{B.3}$$

For any state s_i which is not involved in the merging process, τ_i does not change; that is,

$$\widetilde{\tau}_t(i) = \tau_t(i), \quad \forall i \in [1, Q] \setminus \{l, r, q\}. \tag{B.4}$$

After merging, we have the following:

$$\begin{aligned}
\tau_{t+1}(i) &= a_{ri}\alpha_t(r) + a_{qi}\alpha_t(q) + \wp\\
&= a_{ri}\left(\alpha_t(r) + \alpha_t(q)\right) + \wp\\
&= a_{ri}\left(b_{rk} + b_{qk}\right)\tau_t(r) + \wp = \widetilde{a}_{li} 2\widetilde{b}_{lk} \frac{\widetilde{\tau}_t(l)}{2} + \wp,\\
&\qquad \forall i \in [1, Q] \setminus \{l, r, q\},
\end{aligned} \tag{B.5}$$

where $\wp$ represents the merging influence on the other states. From (B.4), $\wp$ remains the same after merging; that is,

$$\begin{aligned}
\wp = \sum_j a_{ij}\alpha_t(j) = \sum_j a_{ij} b_{jk} \tau_t(j) = \sum_j \widetilde{a}_{ij}\widetilde{b}_{jk}\widetilde{\tau}_t(j),\\
\forall i, j \in [1, Q] \setminus \{l, r, q\}.
\end{aligned} \tag{B.6}$$

With (B.5) and (B.6), we have

$$\tau_t(i) = \widetilde{\tau}_t(i), \quad \forall i \in [1, Q] \cup \{l\} \setminus \{r, q\}. \tag{B.7}$$

Thus we define $\mathfrak{I}$ as follows which also remains the same after merging:

$$\begin{aligned}
\mathfrak{I} = \sum_i b_{ik}\tau_t(i) = \sum_i \widetilde{b}_{ik}\widetilde{\tau}_t(i),\\
\forall i \in [1, Q] \cup \{l\} \setminus \{r, q\}.
\end{aligned} \tag{B.8}$$

Finally,

$$\begin{aligned} P_\lambda(\mathbf{o}_{1:t}) &= b_{rk}\tau_t(r) + b_{qk}\tau_t(q) + \mathfrak{F} \\ &= \left(b_{rk} + b_{qk}\right)\tau_t(r) + \mathfrak{F} = 2\widetilde{b}_{lk}\frac{\widetilde{\tau}_t(l)}{2} + \mathfrak{F} \\ &= P_{\widetilde{\lambda}}(\mathbf{o}_{1:t}). \end{aligned} \tag{B.9}$$

(iii) We prove that, similar to the previous condition, states r and q can be merged to form a simpler HMM which is equivalent. Since the observation probabilities are the same,

$$\begin{aligned} &b_{rk} = b_{qk}, \quad \forall r, q \in [1, Q], \; o_t = v_k \in V, \\ &P_\lambda(\mathbf{o}_{1:t}) = b_{rk}\tau_t(r) + b_{qk}\tau_t(q) + \mathfrak{F} \\ &\quad = b_{rk}\left(\tau_t(r) + \tau_t(q)\right) + \mathfrak{F} = \widetilde{b}_{lk}\widetilde{\tau}_t(l) + \mathfrak{F}. \end{aligned} \tag{B.10}$$

Next, we have to prove that $\mathfrak{F}$ remains the same by proving $\tau_t(i) = \widetilde{\tau}_t(i)$, $\forall i \in [1, Q] \cup \{l\} \setminus \{r, q\}$ after merging. If condition (1) in (iii) states r and q have the same incoming probabilities, we have (B.2); thus

$$\alpha_t(r) = b_{rk}\tau_t(r) = b_{qk}\tau_t(q) = \alpha_t(q) \tag{B.11}$$

such that

$$\begin{aligned} \tau_{t+1}(i) &= a_{ri}\alpha_t(r) + a_{qi}\alpha_t(q) + \wp = \widetilde{a}_{li}\widetilde{\alpha}_t(l) + \wp \\ &= \widetilde{\tau}_{t+1}(i), \quad \forall i \in [1, Q] \setminus \{l, r, q\}. \end{aligned} \tag{B.12}$$

If condition (2) in (iii) holds for states r and q, both states have the same outgoing probabilities and the effect on other states is the sum of both (see case (ii)). Finally, with condition (3) in (iii) the probability of the third state is affected by the sum of the equivalent states:

$$\begin{aligned} \tau_{t+1}(i) &= a_{ri}b_{rk}\tau_t(r) + a_{qi}b_{qk}\tau_t(q) + \sum a_{ji}b_{jk}\tau_t j \\ &= a_{ri}b_{rk}\left(\tau_t(r) + \tau_t(q)\right) + \sum a_{ji}b_{jk}\tau_t j, \\ &\qquad \forall i \in [1, Q] \setminus \{r, q\}; \end{aligned} \tag{B.13}$$

it follows that $\tau_{t+1}(i) = \widetilde{\tau}_{t+1}(i)$ for i different from r and q. Thus

$$P_\lambda(\mathbf{o}_{1:t}) = \sum_i \tau_{t+1}(i) = \sum_i \widetilde{\tau}_{t+1}(i) = P_{\widetilde{\lambda}}(\mathbf{o}_{1:t}); \tag{B.14}$$

it follows that $\tau_{t+1}(r) + \tau_{t+1}(q) = \widetilde{\tau}_{t+1}(l)$.

(iv) We define the state probabilities as $Q_t(i) = P_\lambda(q_t = s_i \mid \mathbf{o}_{1:t-1})$. We define $P_{o_{t-1}} = P_\lambda(\mathbf{o}_{1:t-1})$. It follows that $\tau_t(i) = Q_t(i)P_{o_{t-1}}$. Since the incoming state transition probabilities for state r are constant, $Q_t(r)$ is constant as well. We denote this state probability as q_r. Now we derive the equations that should hold for an HMM $\widetilde{\lambda}$ with $Q - 1$ states to be equivalent to the given model. To have $P_\lambda(\mathbf{o}_{1:t}) = P_{\widetilde{\lambda}}(\mathbf{o}_{1:t})$ and considering that $Q_t(i)$ will fluctuate depending on the observation sequence (for $i \neq r$), $P_\lambda(\mathbf{o}_{1:t}) = f(Q_t(i))$ with f a linear function. $P_{\widetilde{\lambda}}(\mathbf{o}_{1:t})$ should mimic this function by $P_{\widetilde{\lambda}}(\mathbf{o}_{1:t}) = \widetilde{f}(\widetilde{Q}_t(i))$; hence $\widetilde{Q}_t(i)$ should be a linear function of $Q_t(i)$:

$$\widetilde{Q}_t(i) = \mathbf{m}Q_t(i) + c_i \tag{B.15}$$

with $\mathbf{m}$, a vector, and c_i, a constant. It follows that

$$\widetilde{\tau}_t(i) = \mathbf{m}\tau_t(i) + c_i P_{o_{t-1}}. \tag{B.16}$$

For equivalence,

$$P_\lambda(\mathbf{o}_{1:t}) = P_{\widetilde{\lambda}}(\mathbf{o}_{1:t}) \Longleftrightarrow \tag{B.17}$$

$$\mathbf{B}_{o_k}\boldsymbol{\tau}_t = \widetilde{\mathbf{B}}_{o_k}\widetilde{\boldsymbol{\tau}}_t \Longleftrightarrow \tag{B.18}$$

$$\mathbf{B}_{o_k}\boldsymbol{\tau}_t = \widetilde{\mathbf{B}}_{o_k}\mathbf{m}\boldsymbol{\tau}_t + cP_{o_{t-1}}. \tag{B.19}$$

These conditions should hold for all probabilities $\tau_t(i)$, so the factors of each $\tau_t(i)$-term should sum up to zero, except that we have to consider that $\tau_t(r) = q_r P_{o_{t-1}}$ and $\sum \tau_t(i) = P_{o_{t-1}}$. For each observation value, we get $Q - 2$ equations for the independent $\tau_t(i)$-terms and one equation for the constant term. Note that all constant terms contain the factor $P_{o_{t-1}}$ such that the resulting equation is independent of this term. We end up with $M(Q - 1)$ conditions on the parameters. Next, the relation between $\boldsymbol{\tau}_t$ and $\widetilde{\boldsymbol{\tau}}_t$ given by (B.16) should hold in time (and therefore independent of the actual observations). This gives the following conditions on the parameters:

$$\begin{aligned} &\widetilde{\tau}_{t+1}(i) = \mathbf{m}\tau_{t+1}(i) + c_i P_{o_t} \Longleftrightarrow \\ &\widetilde{A}\widetilde{B}_{o_t}\widetilde{\tau}_t(i) = \mathbf{m}\left(\mathbf{A}\mathbf{B}_{o_t}\tau_t(i)\right) + c_i P_{o_t} \Longleftrightarrow \end{aligned} \tag{B.20}$$

$$\widetilde{A}\widetilde{B}_{o_t}\left(\mathbf{m}\tau_t(i) + c_i P_{o_{t-1}}\right) = \mathbf{m}\left(\mathbf{A}\mathbf{B}_{o_t}\tau_t(i)\right) + c_i P_{o_t}. \tag{B.21}$$

This results in $M(Q - 1)(Q - 2)$ equations. The equivalent HMM exists when all conditions can be met. $\widetilde{\lambda}$ has $(M - 1)(Q - 1) + (Q - 1)(Q - 2)$ free parameters. The linear transformation of the state probabilities (see (B.16)) contains $(Q - 2)(Q - 1)$ free

parameters. The conditions for equivalence given by Eq. (B.19) and Eq. (B.21) result in $M(Q-1)+M(Q-1)(Q-2)=(Q-1)(QM-M)$ equations on the free parameters.

For equivalence, there should be no more equations than free parameters:

$$\begin{aligned}&(M-1)(Q-1)+2(Q-1)(Q-2)+(Q-3)\\&\quad\geq (Q-1)(QM-M),\\&MQ-M-Q+1+2QQ-6Q+2+Q-3\\&\quad\geq QQM-MQ-MQ+M,\\&2QQ+(M-6)Q+2\geq MQQ-2MQ+M.\end{aligned}\tag{B.22}$$

We get equality for $M=2$. Larger values result in more equations than free parameters.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

Thanks are due to VUB-IRMO for awarding the Ph.D.-VUB scholarship and the Prognostics for Optimal Maintenance (POM) project (Grant no. 100031; https://pomsbo.wordpress.com/) for providing the application cases which is financially supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Flanders).

References

[1] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] M. J. F. Gales, "Cluster Adaptive Training of Hidden Markov Models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–427, 2000.

[3] J. Yu, "Health condition monitoring of machines based on hidden markov model and contribution analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 8, pp. 2200–2211, 2012.

[4] C. Raphael, "Automatic segmentation of acoustic musical signals using hidden Markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 4, pp. 360–370, 1999.

[5] R. J. Boys, D. A. Henderson, and D. J. Wilkinson, "Detecting homogeneous segments in DNA sequences by using hidden Markov models," *Journal of the Royal Statistical Society, Series C: Applied Statistics*, vol. 49, no. 2, pp. 269–285, 2000.

[6] A. Fischer, K. Riesen, and H. Bunke, "Graph similarity features for HMM-based handwriting recognition in historical documents," in *Proceedings of the 12th International Conference on Frontiers in Handwriting Recognition (ICFHR '10)*, pp. 253–258, November 2010.

[7] J. Li, A. Najmi, and R. M. Gray, "Image classification by a two-dimensional hidden Markov model," *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 517–533, 2002.

[8] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for systems with state persistence," in *Proceedings of the the 25th International Conference on Machine Learning (ICML '08)*, pp. 312–319, Helsinki, Finland, July 2008.

[9] L. Du, M. Chen, J. Lucas, and L. Carin, "Sticky hidden Markov modeling of comparative genomic hybridization," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5353–5368, 2010.

[10] M. Shashanka, "A fast algorithm for discrete hmm training using observed transitions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011.

[11] B. Frnay, G. de Lannoy, and M. Verleysen, "Label noise-tolerant hidden markov models for segmentation: application to ecgs," in *ECML/PKDD (1)*, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds., vol. 6911 of *Lecture Notes in Computer Science*, pp. 455–470, Springer, Berlin, Germany, 2011.

[12] P. Smyth, "Clustering sequences with hidden markov models," in *Advances in Neural Information Processing Systems*, pp. 648–654, MIT Press, 1997.

[13] F. Gelgi and H. Davulcu, "Baum-welch style em approach on simple bayesian models forweb data annotation," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI '07)*, pp. 736–742, IEEE, Fremont, Calif, USA, November 2007.

[14] D. Hsu, S. M. Kakade, and T. Zhang, "A spectral algorithm for learning hidden markov models," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1460–1480, 2012.

[15] B. W. Robert Mattila and R. Cristian, "Rojas, Evaluation of spectral learning for the identification of hidden markov models," in *Proceedings of the 17th IFAC Symposium on System Identification (SYSID '15)*, Beijing , China, 2015.

[16] V. Balasubramanian, *Equivalence and reduction of hidden markov models [Ph.D. thesis]*, Massachusetts Institute of Technology, 1993.

[17] S. Terwijn, "On the learnability of hidden Markov models," in *ICGI 2002*, vol. 2484 of *LNAI*, pp. 261–268, Springer, Berlin, Germany, 2002.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[19] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The Annals of Mathematical Statistics*, vol. 37, pp. 1554–1563, 1966.

[20] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics*, vol. 41, pp. 164–171, 1970.

[21] B. Vanluyten, J. C. Willems, and B. De Moor, "Equivalence of state representations for hidden Markov models," *Systems and Control Letters*, vol. 57, no. 5, pp. 410–419, 2008.

[22] L. Finesso, *Consistent estimation of the order for markov and hidden markov chains [Ph.D. thesis]*, University of Maryland, 1990.

[23] B. Vanluyten, J. C. Willems, and B. De Moor, "A new approach for the identification of hidden Markov models," in *Proceedings of the 46th IEEE Conference on Decision and Control (CDC '07)*, pp. 4901–4905, December 2007.

[24] J. Duan, J. Zeng, and D. Zhang, "A method for determination on HMM distance threshold," in *Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD '09)*, vol. 1, pp. 387–391, IEEE, Tianjin, China, August 2009.

[25] P. Wu, D. Jiang, and H. Sahli, "Physiological signal processing for emotional feature extraction," in *Proceedings of the International Conference on Physiological Computing Systems (PhyCS '14)*, pp. 40–47, January 2014.

[26] M. Johansson and T. Olofsson, "Bayesian model selection for Markov, hidden Markov, and multinomial models," *IEEE Signal Processing Letters*, vol. 14, no. 2, pp. 129–132, 2007.

[27] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, University of California Press, 1967.

[28] E. W. Forgy, "Cluster analysis of multivariate data: effciency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768–769, 1965.

[29] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 224–227, 1979.

[30] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Proceedings of the 2nd International Symposium on Information Theory, Akadémiai Kiado*, B. N. Petrov and F. Csaki, Eds., pp. 267–281, Akadémiai Kiado, 1973.