STAIRS 2014
U. Endriss and J. Leite (Eds.)

doi:10.3233/978-1-61499-421-3-171

Effective and Efficient Identification of Persistent-state Hidden (semi-) Markov Models

Tingting LIU [a,1], and Jan LEMEIRE [a,b]
[a] *Vrije Universiteit Brussel, ETRO Dept., Pleinlaan 2, B-1050 Brussels, Belgium*
[b] *iMinds, Dept. of Multimedia Technologies (MMT), Gaston Crommenlaan 8 (Box 102), B-9050 Ghent, Belgium*

Abstract. The predominant learning strategy for H(S)MMs is local search heuristics, of which the Baum-Welch/ expectation maximization (EM) algorithm is mostly used. It is an iterative learning procedure starting with a predefined topology and randomly-chosen initial parameters. However, state-of-the-art approaches based on arbitrarily defined state numbers and parameters can cause the risk of falling into a local optima and a low convergence speed with enormous number of iterations in learning which is computationally expensive. For models with *persistent* states, i.e. states with high self-transition probabilities, we propose a segmentation-based identification approach used as a pre-identification step to approximately estimate parameters based on segmentation and clustering techniques. The identified parameters serve as input of the Baum-Welch algorithm. Moreover, the proposed approach identifies automatically the state numbers. Experimental results conducted on both synthetic and real data show that the segmentation-based identification approach can identify H(S)MMs more accurately and faster than the current Baum-Welch algorithm.

Keywords. hidden Markov models (HMMs), hidden semi-Markov models (HSMMs), Baum-Welch, local optima, model identification

1. Introduction

Hidden Markov Models (HMMs) [1] and its extension hidden semi-Markov Models (HSMMs) [2] are one of the statistical modeling tools with great success and widely used in a vast range of application fields such as audio-visual speech processing [3], machine maintenance [4], acoustics [5], biosciences [6], handwriting and text recognition [7] and image processing [8].

Classical iterative approaches (e.g., the Baum-Welch algorithm [9,10] and the gradient descent algorithm[11]) are the most commonly used methods when one wants to estimate H(S)MM parameters. However, they require a predefined number of states, which does not necessarily match the real life cases. In spite of this limitation, classical iterative approaches are still widely used to estimate H(S)MM parameters, for lack of alternatives.

[1] Corresponding Author: Tingting Liu, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium; E-mail: tliu@etro.vub.ac.be

The careless adoption of a previously known model state numbers may give misleading results. In order to solve this problem, state-of-the-art approaches decide an optimal state number either using specific criteria (e.g., the Akaike information criterion (AIC) [12], the Bayesian Information Criterion (BIC) [13]), or by structure evolving methods (e.g., the state splitting/ merging approach [14], the genetic approach [15]). However, learning H(S)MMs iteratively using the heuristic approaches above is computationally hard and often produces local optima issues. With respect to this problem, Hsu et al. [16] introduce a spectral-based algorithm for learning HMMs, which employs only a singular value decomposition and matrix multiplications, nonetheless, makes restrictive and problematic assumptions that the transition and emission matrices are full rank and the initial state vector is positive in all coordinates.

In this paper, we address the problem of model initializations and focus on models with *persistent* states (i.e., "sticky transitions"). Fox et al. [17] propose a sticky HDP-HMM which is a non-parametric, infinite-state model that automatically learns the size of state spaces and the smoothly varying dynamics robustly. However, this approach is computationally prohibitive when data sets are very large [18]. In this paper, a segmentation-based identification approach is proposed for models with *persistent* states, based on the segmentation of the observed data. Specifically, a pre-estimation step is conducted to decide the number of states and the initial model parameters approximately. This approximate estimation is served as an effective starting point of the Baum-Welch algorithm which refines the initial parameters. Consequently, both the number of iterations needed and the chance of falling into a local optimum are reduced. The improvements in effectiveness and efficiency of the proposed approach are confirmed experimentally using both simulated and real data.

The remainder of the paper is organized as follows: in Section 2, the preliminaries about HMMs and HSMMs are briefly reviewed, followed by the classifications of hidden states. Section 3 discusses the methodology of the proposed method. Experiments conducted on both synthetic and real data are described and discussed in Section 4. Finally, conclusions are given in Section 5.

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 N denote the number of states and M the number of observation symbols. Let $\mathbf{S} = \{s_1, s_2, \ldots, s_N\}$ and $\mathbf{O} = \{v_1, v_2, \ldots, v_M\}$ denote the set of states and the set of observations, respectively. Using q_t to represent the state and o_t the observation at time t, an HMM model can be characterized as below with the notation in [1]: the state transition probability matrix is $\mathbf{A} = \{a_{ij}\}$, where

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i), 1 \leq i, j \leq N \tag{1}$$

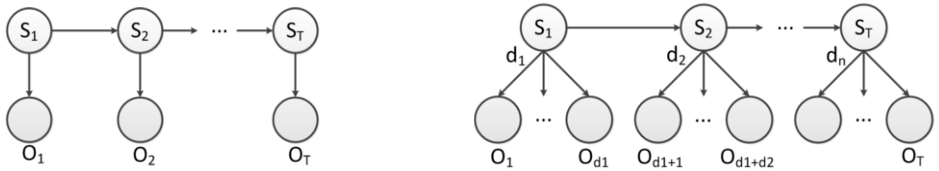
The observation probability matrix is $\mathbf{B} = \{b_j(k)\}$, where

$$b_j(k) = P(o_t = v_k | q_t = s_j), 1 \leq j \leq N, 1 \leq k \leq M \tag{2}$$

The initial state probability distribution

$$\pi_i = P(q_1 = s_i), 1 \leq i \leq N \tag{3}$$

where $s_i \in \mathbf{S}$. An HMM can be expressed with the abbreviation $\lambda = (\pi, \mathbf{A}, \mathbf{B})$. An example is shown in Figure 1a.



(a) An HMM

(b) An HSMM

Figure 1. Example of an HMM and an HSMM

Hidden semi-Markov model (HSMMs) [2] is an extension of the HMMs, of which the underlying stochastic process is a semi-Markov chain instead of a Markov chain as in the HMMs. Each state has an explicit state duration variable d, which is associated the number of observations being emitted while in the state [2]. Let D denote the maximum allowed duration in a state and the state duration set as $\mathbf{D} = \{1, 2, \ldots, D\}$. For each observation sequence $o_{1:T}$, the corresponding state sequence is denoted as $s_{[1:d_1]} = i_1, s_{[d_1+1:d_1+d_2]} = i_2, \ldots, s_{[d_1+\cdots+d_{n-1}+1:d_1+\cdots+d_n]} = i_n$ and the state transitions are $(i_m, d_m) \rightarrow (i_{m+1}, d_{m+1})$, for $m = 1, \ldots, n-1$, where $\sum_{m=1}^{n} d_m = T, i_1, \ldots, i_n \in \mathbf{S}$ and $d_1, \ldots, d_n \in \mathbf{D}$. The state transition probability is defined as

$$a_{(i,d')(j,d)} = P(s_{[t+1:t+d]} = j | s_{[t-d'+1:t]} = i) \tag{4}$$

subject to $\sum_{j \in \mathbf{S} \setminus \{i\}} \sum_{d \in \mathbf{D}} a_{(i,d')(j,d)} = 1$ with zero self-transition probabilities $a_{(i,d')(j,d)} = 0$, where $i, j \in \mathbf{S}$ and $d, d' \in \mathbf{D}$. The observation probability of d observations $o_{t+1:t+d}$ being emitted in state j can be written as

$$b_{j,d}(o_{t+1:t+d}) = P(o_{[t+1:t+d]} | s_{[t+1:t+d]} = j) \tag{5}$$

The initial state probability is denoted by

$$\pi_{j,d} = P(s_{[t-d+1:t]} = j), t \leq 0, d \in \mathbf{D}. \tag{6}$$

An HSMM can be abbreviated by $\lambda = (a_{(i,d')(j,d)}, b_{j,d}(v_{k_1:k_d}), \pi_{i,d})$, where $i, j \in \mathbf{S}, d, d' \in \mathbf{D}$, and $v_{k_1:k_d}$ represents $v_{k_1}, \ldots, v_{k_d} \in \mathbf{O} \times \cdots \times \mathbf{O}$. An example is shown in Figure 1b.

A *persistent* state is a state with a high self-transition probability, i.e. the rate of remaining at the same state is high while the rates of going to other states are low. A *transient* state, on the other hand, is very likely to move to other states instead of staying at the same state. Hence the self-transition probability a_{ii} of state $i, 1 \leq i \leq N$ is used as an indicator to distinguish between persistent and transient state, i.e. if $a_{ii} > 1/N$, it is persistent, otherwise transient. This paper focuses on H(S)MMs with *persistent* states.

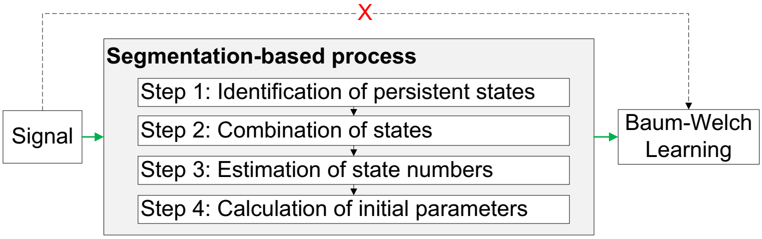

X
Segmentation-based process
Step 1: Identification of persistent states
Step 2: Combination of states
Step 3: Estimation of state numbers
Step 4: Calculation of initial parameters
Signal
Baum-Welch Learning

Figure 2. Scheme of the proposed approach

3. Methodology of model identification

The application we will consider is industrial machinery system maintenance which is suitable of being modeled with *persistent-state* H(S)MMs. As stated in [19], the reason of using H(S)MMs in machine maintenance and decision making is that machine operation condition can be classified into a number of meaningful states, such as “Good”, “OK”, “Minor defects only”, “Maintenance required”, “Unserviceable”, so that the state definition is closer to what is used in industry and thus easy to interpret. As states determine the behavior of a system, persistence of states implies that the system will exhibit the same behavior for a certain period. Such period is called a *regime*, i.e., a time period in which the state of the system does not change, meaning the observation probabilities are constant. The assumption of state persistence is reasonable in industrial machinery systems since machine condition opt to stay in a stable and persistent state for a certain period before jumping to another state if nothing goes wrong. For instance, a machine in a “Good” condition at the current time is more likely to remain “Good” at the next time step instead of going into an “OK” condition unless the machine already degrades over a certain time period (i.e., a regime). Our algorithm is based on identifying the regimes of a state through segmentation and clustering.

The segmentation-based identification approach contains four steps: firstly, signals are split into different regimes based on different signal behaviors. Secondly, the ‘similar’ regimes of signal are grouped together by clustering techniques according to their similarities. The achieved labeled regimes are assumed to correspond to hidden states. Thirdly, a clustering validation index is employed to determine the number of states. Finally, H(S)MM parameters are estimated by calculating statistical occurrences of the observed signal and the estimated hidden states, then used as initial input of the standard Baum-Welch algorithm. The scheme of the methodology is shown in Fig. 2.

3.1. Step 1: Identification of persistent states by segmentation

Data sequences emitted by *persistent* states can be segmented into sub-sequences 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 for splitting discrete signals by employing the test of [20]. The test calculates the Bayesian probability that two sequences have been generated by the same or by a different multinomial model. 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 test indicates that it is very likely (with a confidence level, for example 90%) that the two sequences are from a different model, the current time point is marked as a change point. The procedure repeats until the end of the signal. An example is shown in Figure 3.

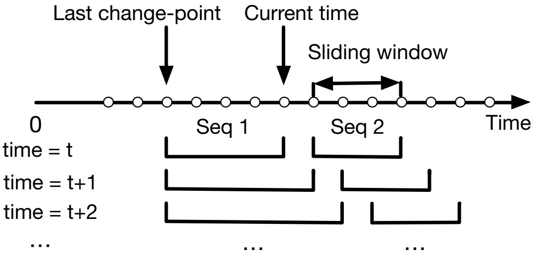


Figure 3. Sliding-window based segmentation

3.2. Step 2: Combination of states by clustering

Regimes corresponding to the same states will recur over time. Assuming 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 [21,22] is used to combine and label each segment, described as below: 1) feature points are obtained by averaging the data in each segment; 2) the feature points are divided into k subsequences with equal length; 3) the median values of each subsequence are used as initial starting centroids for k means clustering. Notably, 2) and 3) are the preliminary steps designed to avoid the problem of randomness in initializations of *k-means* clustering.

3.3. Step 3: Estimation of state numbers by cluster validity

In order to select the optimal number of clusters, a robust index, called Davies-Bouldin index (DBI) [23], is applied in this paper.

Suppose dataset X is partitioned into K disjoint non-empty 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 [23] is defined as:

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{i \neq j} \{ \frac{diam(C_i) + diam(C_j)}{dist(C_i, C_j)} \} \tag{7}$$

where $diam(C_i) = \max_{\mathbf{x}_m, \mathbf{x}_n \in C_i} \{d(\mathbf{x}_m, \mathbf{x}_n)\}$ and $dist(C_i, C_j) = \min_{\mathbf{x}_m \in C_i, \mathbf{x}_n \in C_j, i \neq j} \{d(\mathbf{x}_m, \mathbf{x}_n)\}$ denote the intra-cluster diameter and the inter-cluster distance, respectively. Apparently, the partition with the minimum Davies-Bouldin index is considered as the optimal choice.

3.4. Step 4: Estimation of initial parameters

The underlying assumption of our method is that segmentation of the observed signal allows us to identify quite accurately the regimes of the true model. If the regimes belonging to the same state are grouped correctly, these regimes offer us good insight into the behavior of the states, i.e. the observation and transition probabilities as well as the duration distribution. The probabilities are estimated based on the observed frequencies.

Parameters of an HMM (i.e., probability matrices) can be calculated by simple counting the occurrence of the observed signal and the hidden states (i.e., labels retrieved from clustering), which are computed as below [1,9,10]:

$$\bar{\pi}_i = \text{ frequency in state } s_i \text{ at time } t = 1 \tag{8}$$

$$\bar{a}_{ij} = \frac{\text{\# of trans. from } s_i \text{ to } s_j}{\text{\# of trans. from } s_i} \tag{9}$$

$$\bar{b}_j(k) = \frac{\text{\# of times in } s_j \text{ observing } v_k}{\text{\# of times in } s_j} \tag{10}$$

where *trans.* is the abbreviation for transition. Note that Baum-Welch uses the same equations in (re)-estimating model parameters. Similarly, the parameters of an HSMM can be computed as below:

$$\bar{\pi}_{i,d} = \text{ frequency in state } s_i \text{ at time } t = 1\text{, with dur. } d \tag{11}$$

$$\bar{a}_{(i,d')(j,d)} = \frac{\text{\# of trans. from } s_i \text{ with dur. } d' \text{ to } s_j \text{ with dur. } d}{\text{\# of trans. from } s_i \text{ with dur. } d'} \tag{12}$$

$$\bar{b}_{j,d}(o_{t+1:t+d}) = \frac{\text{\# of times } o_{t+1:t+d} \text{ emitted in } s_j}{\text{\# of times in } s_j} \tag{13}$$

where *dur.* is the abbreviation for duration. The distribution of the duration d for each state can be calculated by the kernel density estimation (KDE) based on a normal kernel function [24,25]:

$$\bar{f}_h(d) = \frac{1}{\gamma h} \sum_{i=1}^{\gamma} K_N\left(\frac{d - d_i}{h}\right) \tag{14}$$

where $(d_1, d_2, \ldots, d_\gamma)$ is a duration sample drawn from a distribution with density f, K_N represents a normal kernel and h is bandwidth for the smoothing purpose, which is set as the optimal for normal densities.

4. Experimental Validation

Experiments on both synthetic and real case datasets are performed to evaluate the accuracy and efficiency of the proposed method.

4.1. Synthetic datasets

Simulated datasets are generated by 50 randomly created *persistent-state* HMMs with number of states Q (from 2 to 6), number of observations O (from 2 to 6), each combination of Q and O is repeated 2 times. Each HMM is served as a reference model and then used to generate a dataset with 4000 samples (20×200, number of observation sequences × length of sequence). The first $4/5$ observation sequences are selected as training samples and the remaining $1/5$ are used as test samples. Similar parameters are set for HSMMs with a maximum duration of $D = 30$ and datasets are generated with $T = 1000$ time steps. To identify each reference model, we train the model with both the proposed method and the standard Baum-Welch approach for comparisons. The state numbers are selected from a state pool of $[2, 2Q]$ by the Baum-Welch with the AIC criterion [12], and the proposed method with DBI cluster validation, respectively. As a result, $2Q - 1$ times of the BW learning is required for each learning task. On the contrary, the proposed method starts with a pre-learned initial parameters, hence requires only 1 time. In the segmentation step of the proposed method, the window size and the confidence level can be adjusted according to different applications, which here are set to 20 and 0.9 empirically.

The comparisons are conducted on two aspects: learning accuracy and speed. The accuracy in model identification is evaluated by comparing the log-likelihoods (LL) difference with the reference model on test samples. The LL of the observation $o_{1:T}$ given the model λ measures how well the model fits the data $(\log(P(o_{1:T}|\lambda))$. If the difference between the likelihoods is below a certain threshold (5% in this paper), the model is considered as correctly learned; otherwise the learned model is assumed to be trapped in a local optimum. The learning speed is compared on the total time of learning (measured in seconds) and the number of iterations to converge.

Table 1. Performance on synthetic data for standard Baum-Welch algorithm and proposed method

Models		HMM		HSMM	
Criteria		Random BW + AIC	Proposed method	Random BW + AIC	Proposed method
Accuracy	Test-set LL difference (%)	18.7	2.6	46.27	39.37
	Test-set local optima (%)	39.4	14.0	90.20	72.00
Speed	Average learning time (seconds)	13.32	2.62	2.44	1.55
	Average number of iterations (#)	25.94	8.70	4.32	5.56

Experiment results in Table 1 show an obvious improvement of the proposed method compared to the Baum-Welch algorithm for HMMs: the model distance with the reference model and the number of local optima are lowered and with a faster learning speed and fewer number of iterations to converge. For the HSMMs, improvements can be seen marginally.

4.2. Bearing dataset

The proposed method is applied to a bearing dataset for machine maintenance provided by the POM project[2]. The set-up consists of steel cord production machines located in the production plant in China. These machines were continuously monitored for bearing degradation using accelerometers and temperature sensors. The temperature are logged regularly by the temperature sensors on both sides, i.e., ‘input’ and ‘output’ side. A temperature overshoot protection is implemented in the machine’s controllers in order to avoid catastrophic failures of bearings. When temperature exceeds the temperature threshold for more than predefined observation time, the machine stops and restarts when the temperature decreases below the temperature threshold.

The temperature evolution is predicted by one of the POM project partners in a run-by-run way, on which regression is applied taking context feature into account by Dynamic Time Warping. However, each run is learned separately without considering the run-dependence. To address this issue, we use the H(S)MM models to study the dynamic nature and correct the prediction errors of the regression method by the inferred model. In the experiment, the dataset is the prediction errors of the ‘input’ temperature signal from one of the machines at the end of each run, which contains 3160 data points. The state number selection of the BW and the proposed methods is conducted on a state pool of [2, 8] via the AIC criterion and the DBI index, respectively. To lower the side affect of the randomness in the BW initialization, the learning is repeated 5 times for the BW and the averaged performance is used to compare with the proposed method.

Figure 4a shows the learning results of the proposed method: segmentation via change-point detection (above) and the labeled hidden states by the K-means clustering (below). The clustering validation by the DBI index is shown in Figure 4b of which the one with 8 clusters (states) with the minimum DBI value is selected as an optimal choice. Results suggest the feasibility of combining the states with ‘similar’ behaviors by clustering. Performances of both approaches are shown in Table 2. The averaged log-likelihood values of the Baum-Welch method are lower and the speed is dramatically

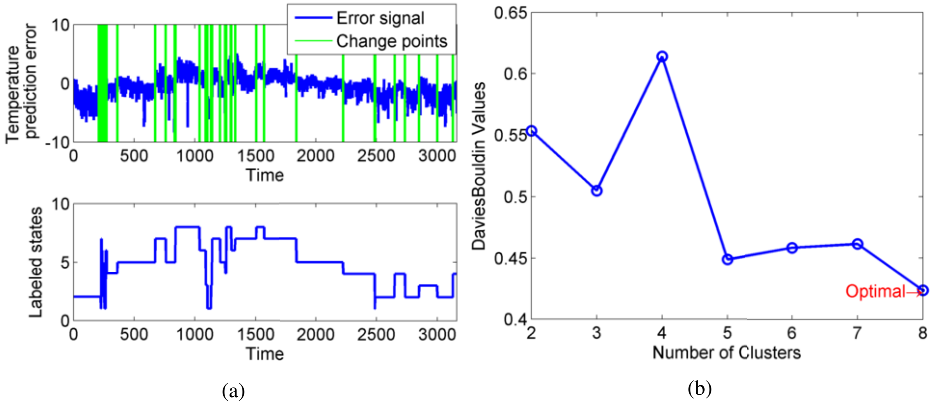


Figure 4. Pre-estimation results of the proposed method. (a) Segmentation via change-point detection and labeling via clustering. (b) The DBI indexes.

[2] www.pom-sbo.org

Table 2. Comparisons of the average learning performance over repetitions on the bearing data

Models	HMM		HSMM	
Criteria	Random BW + AIC	Proposed method	Random BW + AIC	Proposed method
Number of states	8	8	7	8
Average log-likelihood	-2.0284	-2.0175	-2.0523	-2.0149
Average learning time (seconds)	205.3406	9.0711	163.6585	7.1672

slowed down compared to the proposed method. The reasons for the speed gain are explicit: both methods select the state numbers from a state pool from 2 to 8. However, the proposed method uses a clustering validation index with only one run of Baum-Welch learning, instead of 7 runs for the traditional BW method; moreover, even the average duration of runs in traditional BW, is still much longer than proposed method, because a rather 'accurate' initialization of the proposed method requires not only fewer iterations, but also less time to converge.

5. Conclusions

This paper introduces an extension to the current algorithm for H(S)MMs identification based on segmentation and clustering techniques. Both state number selection and parameters initialization are addressed. Enhancement in the accuracy of H(S)MMs identification and the learning converge speed are achieved through the development of a pre-estimation step which avoids the local optimal problem. The effectiveness and efficiency of the proposed method are confirmed through experiments on both synthetic and real signals. Future work will improve the proposed method and extend it to more types of models (e.g., the ones with *non-persistent* states), as well as consider an application difference for a better quantification of the model parameters estimation.

References

[1] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, pp. 257–286, 1989.

[2] S. zheng Yu, "Hidden semi-markov models," *Artificial Intelligence*, 2010.

[3] A. Verma, N. Rajput, and L. Subramaniam, "Using viseme based acoustic models for speech driven lip synthesis," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 5, pp. V–720, IEEE, 2003.

[4] J. bo Yu, "Health condition monitoring of machines based on hidden markov model and contribution analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, pp. 2200 – 2211, 2012.

[5] B. Logan and P. Moreno, "Factorial hmms for acoustic modeling," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1998.

[6] 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.

[7] A. Fischer, K. Riesen, and H. Bunke, "Graph similarity features for hmm-based handwriting recognition in historical documents," in *2010 12th International Conference on Frontiers in Handwriting Recognition*, vol. 0, pp. 253–258, IEEE, Nov. 2010.

[8] J. Li, A. Najmi, and R. M. Gray, "Image classification by a two dimensional hidden markov model," *IEEE Transactions on Signal Processing*, vol. 48, 2000.

[9] 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.

[10] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Annals of Mathematical Statistics*, vol. 41, pp. 164–171, 1970.

[11] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a markov proces to automatic speech recognition," 1983.

[12] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki, eds.), pp. 267–281, Akadémiai Kiado, 1973.

[13] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[14] M. Ostendorf and H. Singer, "HMM topology design using maximum likelihood successive state splitting," *Computer Speech and Language*, vol. 11, pp. 17–41, 1997.

[15] J. Goh, L. Tang, and L. A. turk, "Evolving the structure of Hidden Markov models for micro aneurysms detection," in *UK Workshop on Computational Intelligence*, 2010.

[16] D. Hsu, S. M. Kakade, and T. Zhang, "A Spectral Algorithm for Learning Hidden Markov Models," *Computing Research Repository*, vol. abs/0811.4, 2008.

[17] 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 25th International Conference on Machine Learning*, ICML '08, (New York, NY, USA), pp. 312–319, ACM, 2008.

[18] L. Du, M. Chen, J. Lucas, and L. Carin, "Sticky hidden markov modeling of comparative genomic hybridization," *Signal Processing, IEEE Transactions on*, vol. 58, pp. 5353–5368, Oct 2010.

[19] X.-S. Si, W. Wang, C.-H. Hu, and D.-H. Zhou, "Remaining useful life estimation - A review on the statistical data driven approaches," *European Journal of Operational Research*, vol. 213, pp. 1–14, August 2011.

[20] M. Johansson and T. Olofsson, "Bayesian Model Selection for Markov, Hidden Markov, and Multinomial Models," *IEEE Signal Processing Letters*, vol. 14, pp. 129–132, 2007.

[21] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, vol. 1, pp. 281–297, Univ. of Calif. Press, 1967.

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

[23] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[24] M. Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, vol. 27, pp. 832–837, Sept. 1956.

[25] E. Parzen, "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.