ESTIMATION OF NONSTATIONARY HIDDEN MARKOV MODELS BY MCMC SAMPLING

Petar M. Djurić and Joon-Hwa Chun

Department of Electrical and Computer Engineering
State University of New York at Stony Brook
Stony Brook, NY 11794, USA
e-mail: djuric@ee.sunysb.edu and jchun@ee.sunysb.edu

ABSTRACT

Hidden Markov models are very important for analysis of signals and systems. In the past two decades they have been attracting the attention of the speech processing community, and recently they have become the favorite models of biologists. Major weakness of conventional hidden Markov models is their inflexibility in modeling state duration. In this paper, we analyze nonstationary hidden Markov models whose state transition probabilities are functions of time, thereby indirectly modeling state durations by a given probability mass function. The objective of our work is to estimate all the unknowns of the nonstationary hidden Markov model, its parameters and state sequence. To this end, we construct a Markov chain Monte Carlo sampling scheme in which all the posterior probability distributions of the unknowns are easy to sample from. Extensive simulation results show that the estimation procedure yields excellent results.

1. INTRODUCTION

Hidden Markov models (HMM's) have played a prominent role in many approaches to signal and system analysis. In speech processing they are the ultimate tool for various tasks of statistical modeling including speaker and speech recognition [6]. In modern biology with the emergence of molecular genetics and the advance of the Human Genome Project, an immense amount of data is being produced that require use of sequence analysis methods, and where the HMM's seem very well fitted for extracting information from the data [3].

The conventional HMM has a major structural weakness in that its state durations have fixed geometrical distributions, thereby limiting its range of applications [6]. Ferguson [4] introduced the variable duration HMM, where each state duration is modeled by a probability distribution, which is not necessarily geometric. This gives the HMM much more flexibility and widens the range of its applications. Most of the previous work on estimating variable duration HMM's is on extending the methods of the conventional HMM's, that is the dynamic programming algorithm and maximum likelihood estimators. Later, a different parametrization of variable state duration was introduced, where the state transition probabilities are explicitly modeled as functions of time [7], [8], and, thus, are referred to as nonstationary HMM's. It can be shown, however, that the variable duration and nonstationary HMM's are equivalent, and that the latter are more tractable for analysis.

Recently, a Markov chain Monte Carlo (MCMC) scheme has been applied to conventional HMM's [1]. Here we present an MCMC procedure for estimation of non-stationary HMM's. It is assumed that the observed sequence is modeled by a nonstationary HMM with known number of states, and that the state sequence and all the model parameters are unknown. We construct a Gibbs sampling scheme which converges quickly and has posterior distributions that are easy to sample from. From the samples of the posteriors drawn after convergence, the state sequence and parameter estimates of the model can straightforwardly be obtained. The simulation results of the method show excellent performance.

2. REVIEW OF CONVENTIONAL HMM'S

Here we provide a very brief review of conventional HMM's and outline the main modeling problems related to them. Consider a system that is described by a set of N distinct states, S_k, where $S_k \in \mathcal{S}$, and $\mathcal{S} = \{S_1, S_2, \ldots, S_N\}$. The system changes with time, and the state of the system at the time instants $t = 1, 2, \ldots, T$, is denoted by q_t, where $q_t \in \mathcal{S}$. Next, suppose that the dynamics of the system is described by a Markov chain, that is, whenever the system is in state S_i, there is a fixed probability that it will next be in state S_j. This is expressed by

$$\begin{aligned} &P(q_{t+1} = S_j | q_t = S_i, q_{t-1} = S_k, \ldots.) \\ &\quad = P(q_{t+1} = S_j | q_t = S_i) = a_{ij}. \end{aligned} \tag{1}$$

The complete description of the state transitions is given by the matrix $\mathbf{A} = \{a_{ij}\}$, where $a_{ij} \geq 0$, and $\sum_{j=1}^{N} a_{ij} = 1$.

In many modeling scenarios, it is assumed that the state sequence is not known, that is, it is hidden from the observer. Instead, at every time instant t, the system generates an observation y_t according to a probability distribution that depends on the state q_t. If the number of distinct observations is M, and the set of observation symbols is $\mathcal{V} = \{v_1, v_2, \ldots, v_M\}$, the probability distributions of observed symbols are given by an $N \times M$ matrix $\mathbf{B}$ whose

This work was supported by the National Science Foundation under Award No. MIP-9506743.

elements b_{jk} are known as emission probabilities and are defined according to

$$b_{jk} = P(v_k \text{ at } t | q_t = S_j), \quad 1 \le j \le N, \quad 1 \le k \le M \quad (2)$$

where $\sum_{k=1}^{M} b_{jk} = 1$.

Finally, to complete the specification of the model, one needs to provide the initial state distribution defined by $\pi_i = P(q_1 = S_i)$, $i = 1, 2, \ldots, N$, with $\sum_{i=1}^{N} \pi_i = 1$. The three probability distributions $\mathbf{A}$, $\mathbf{B}$, and π are in short denoted by λ, or $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$.

Typically, a common assumption for an observed sequence $\mathbf{y} = [y_1\ y_2 \ldots\ y_T]$ is that its joint probability mass function conditioned on the state sequence $\mathbf{q}^T = [q_1\ q_2 \ldots\ q_T]$ and the parameters λ is given by

$$P(\mathbf{y}|\mathbf{q}, \lambda) = \prod_{t=1}^{T} P(y_t|q_t, \lambda) \quad (3)$$

which means conditional independence of the observations. There are three basic problems related to HMM's [6], and in order of increasing complexity, they are:

1. Given a set of observations $\mathbf{y}^T = [y_1\ y_2\ \ldots\ y_T]$ and the model parameters λ, find the probability of the observed sequence $\mathbf{y}$, $P(\mathbf{y}|\lambda)$.

2. Given a set of observations $\mathbf{y}^T = [y_1\ y_2\ \ldots\ y_T]$ and the model parameters λ, find the corresponding state sequence $\mathbf{q}$.

3. Given a set of observations $\mathbf{y}^T = [y_1\ y_2\ \ldots\ y_T]$, find the state sequence $\mathbf{q}$ as well as the model parameters λ.

The solutions to these three problems are well known [6]. The first one can be solved efficiently by the forward-backward procedure, the second, by the Viterbi algorithm, and the third, by the iterative method of Baum-Welch.

3. NONSTATIONARY HMM'S

An important weakness of the conventional HMM is its inflexibility to model state durations. If d is the duration of a particular state, say S_k, then the probability of d is given by

$$P_k(d) = a_{kk}^{d-1}(1 - a_{kk}). \quad (4)$$

The distribution of d is thus geometric, and although in some practical cases it represents physical reality reasonably well, in many more applications, it is completely inappropriate.

One way of modifying the conventional HMM is by way of introducing state duration densities, $P_k(d)$, $k = 1, 2, \ldots, N$, [4], [6]. A state sequence according to this model can be generated as follows:

1. Generate q_1 from the initial state distribution π.
2. Set $t = 1$.
3. Obtain the duration of the state q_t, d, by sampling from $P_k(d)$, where $q_t = S_k$.
4. Set $t = t + d$.
5. Draw the next state q_t from the transition probabilities a_{ij}, where $q_{t-1} = S_i \neq S_j$.
6. If $t < T$, go back to 3; otherwise terminate the procedure.

It is interesting to note that the self-transition probabilities a_{ii} are not explicitly used in these models. The methods employed for solving the three basic problems of conventional HMM's can be extended to accommodate the variable state duration HMM's. The extensions, however, entail increased computational load.

A different parametrization of the state duration can be achieved by treating all the transition probabilities a_{ij} as functions of d, which we denote by $a_{ij}(d)$. They represent the probability that the system switches from state S_i to state S_j given that the system was in state S_i for d consecutive time units [7], [8].

It can be shown that the two models are equivalent, provided the first model gets the feature that a_{ij}, $i \neq j$, is a function of d. The direct relationship between the models can be deduced by observing that the self-transition probability $a_{ii}(d)$ may be expressed in terms of the cumulative distribution function, $F_i(d)$, of the state duration by

$$a_{ii}(d) = 1 - F_i(d). \quad (5)$$

The generation of states according to this model is summarized as follows:

1. Generate q_1 from the initial state distribution π, and set $t = 1$.
2. Record the duration of the current state d.
3. Draw the next state q_{t+1}, from $a_{ij}(d)$, where $q_t = S_i$, and $\sum_{j=1}^{N} a_{ij}(d) = 1$.
4. If $t < T$, set $t = t + 1$, and go back to 2; otherwise terminate the procedure.

We find the second representation more tractable for analysis. Also the implementation of MCMC sampling for estimation of the parameters and states of the model is then much easier. The models whose transitional probabilities are functions of time are called nonstationary HMM's.

4. ESTIMATION OF NONSTATIONARY HMM'S BY MCMC SAMPLING

MCMC methods are computational schemes used for drawing samples from complicated and high dimensional distributions. The samples are then used to summarize information about unknown quantities of a model or perform other tasks such as comparison of various models. Gibbs sampling is an important MCMC scheme where the transitional kernel is formed by the full conditional distributions. In other words, in many problems where drawing samples from complex distributions is intractable, the Gibbs sampler exploits the simplicity of sampling from the full conditionals.

In our problem, the joint distribution of the unknowns, the state sequence $\mathbf{q}$, the initial state probabilities π, the state transition probabilities $\mathbf{A}(d) = \{a_{ij}(d)\}$, and emission probabilities $\mathbf{B}$, is a complicated one, but we show that the full conditionals are quite simple to sample from. We make the assumption that the duration of the various states follows Poisson distributions with various parameters. The assumption is not restrictive by any means; the procedure that follows can be replicated with minor modifications with any probability mass function. If we do not want to assume any parametric distribution, the procedure is still applicable and its details will be presented elsewhere.

First we need to specify the prior distributions for all the unknowns.

1. The prior for the initial probabilities $\pi = [\pi_1\ \pi_2\ \ldots\ \pi_N]$ is the Dirichlet distribution, or $\pi_1, \pi_2, \ldots, \pi_{N-1} \sim \mathcal{D}(\alpha_1, \alpha_2, \ldots, \alpha_N)$, where $\alpha_i > 0$, $i = 1, 2, \ldots, N$.

2. The state durations are modeled by Poisson distributions, i.e., the probability that the duration of state S_i is d is given by

$$p_i(d) = \frac{\beta_i^{d-1} e^{-\beta_i}}{(d-1)!}, \quad d = 1, 2, \ldots \tag{6}$$

where β_i is the parameter of the Poisson distribution associated with the i-th state. All the β_i's have Gamma distributions, $\mathcal{G}(u, v)$, or

$$\beta_i \sim \frac{v^u}{\Gamma(u)} \beta_i^{u-1} e^{-v\beta_i}, \quad u > 0,\ v > 0. \tag{7}$$

Note that the self-transition probabilities are determined from (5). The outward state transition probabilities $a_{ij}(d)$, $i \neq j$, can be obtained from $w_{ij}(d)(1 - a_{ii}(d_i))$, where $w_{ij}(d)$ is the transition weight from state S_i to S_j, given that the duration of S_i has been d. For all i and all d, the weights must satisfy $\sum_{j=1}^{N} w_{ij}(d) = 1$, where $j \neq i$. For some models, the $w_{ij}(d)$'s do not have to be functions of d. In the sequel they are regarded as constant parameters, whose prior distribution is Dirichlet, or $\mathbf{w}_i \sim \mathcal{D}(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{i,N-1})$ where $\mathbf{w}_i^T = [w_{i1}\ w_{i2} \ldots w_{i,i-1}\ w_{i,i+1} \ldots w_{i,N-1}]$. We represent all the weights by the $N \times (N-1)$ matrix $\mathbf{W}$ defined by $\mathbf{W}^T = [\mathbf{w}_1\ \mathbf{w}_2 \ldots \mathbf{w}_N]$.

3. The emission parameters $\mathbf{B} = \{b_{jk}\}$ also have a Dirichlet prior $\mathbf{b}_i \sim \mathcal{D}(\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iM})$ and $\mathbf{b}_i = [b_{i1}\ b_{i2} \ldots b_{i,M-1}]$.

4. The priors of the states q_t are uniform, i.e., $p(q_t = S_i) = 1/N$.

5. GIBBS SAMPLING OF NONSTATIONARY HMM'S

With the chosen priors, the Gibbs sampler is very easy to implement. The steps are the following:

1. Draw π from

$$f(\pi|\beta^{(k-1)}, \mathbf{W}^{(k-1)}, \mathbf{B}^{(k-1)}, \mathbf{q}^{(k-1)}, \mathbf{y}) \propto \mathcal{D}(\alpha_1 + \delta_{q_1^{(k-1)} S_1}, \alpha_2 + \delta_{q_1^{(k-1)} S_2}, \ldots, \alpha_N + \delta_{q_1^{(k-1)} S_N}) \tag{8}$$

where $\delta_{q_1^{(k-1)} S_i}$ is the Kronecker delta function, or $\delta_{ij} = 1$ for $i = j$, and $\delta_{ij} = 0$ for $i \neq j$.

2. Draw $\beta_i^{(k)}$, for $i = 1, 2, \ldots, N$, from

$$f(\beta_i|\pi^{(k)}, \mathbf{W}^{(k-1)}, \mathbf{B}^{(k-1)}, \mathbf{q}^{(k-1)}, \mathbf{y}) \propto \mathcal{G}(u + \tilde{d}_i^{(k-1)}, v + m_i^{(k-1)}) \tag{9}$$

where $\tilde{d}_i^{(k-1)} = \sum_{t=1}^{T} I_{\{q_t^{(k-1)} = S_i\}}$, $I_{\{\cdot\}}$ is the indicator function, and $m_i^{(k-1)}$ is the number of segments in state S_i in iteration $k-1$.

3. Draw $w_{ij}^{(k)}$ from

$$f(w_{ij}|\pi^{(k)}, \beta^{(k)}, \mathbf{b}_j^{(k-1)}, \mathbf{q}^{(k-1)}, \mathbf{y}) \propto \mathcal{D}(\alpha_{i1} + n_{i1}^{(k-1)}, \alpha_{i2} + n_{i2}^{(k-1)}, \ldots, \alpha_{i,N-1} + n_{i,N-1}^{(k-1)}) \tag{10}$$

where $n_{ij}^{(k-1)}$ is the number of transitions from state S_i to state S_j.

4. The $a_{ii}^{(k)}(d)$'s are obtained from

$$a_{ii}^{(k)}(d) = 1 - \sum_{k=1}^{d} P(d = k|\beta_i^{(k)}) \tag{11}$$

where $P(\cdot)$ is the shifted Poisson probability mass function. The $a_{ij}^{(k)}(d)$'s, $i \neq j$, are found from $a_{ij}^{(k)}(d) = w_{ij}^{(k)}(1 - a_{ii}^{(k)}(d))$.

5. Draw $\mathbf{b}_i^{(k)}$ from

$$f(\mathbf{b}_i|\pi^{(k)}, \beta^{(k)}, \mathbf{W}^{(k)}, \mathbf{q}^{(k-1)}, \mathbf{y}) \propto \mathcal{D}(\gamma_{i1} + \tilde{m}_{i1}^{(k-1)}, \gamma_{i2} + \tilde{m}_{i2}^{(k-1)}, \ldots, \gamma_{iM} + \tilde{m}_{iM}^{(k-1)}) \tag{12}$$

where $\tilde{m}_{ij}^{(k-1)}$ is the number of symbols v_j in state S_i.

6. Draw $q_t^{(k)}$ from

$$P(q_t = S_i|\mathbf{q}_{-t}^{(k)}, \pi^{(k)}, \beta^{(k)}, \mathbf{W}^{(k)}, \mathbf{B}^{(k)}, \mathbf{y}) \propto P(q_t|q_{t-1}^{(k)}, d_-^{(k)}(q_{t-1})) P(q_{t+1}^{(k-1)}|q_t, d_+^{(k-1)}(q_{t+1})) \times P(y_t|q_t) \tag{13}$$

where $\mathbf{q}_{-t}^{(k)} = [q_1^{(k)}, \ldots, q_{t-1}^{(k)}, q_{t+1}^{(k-1)}, \ldots, q_T^{(k-1)}]$, $d_-^{(k)}(q_{t-1})$ and $d_+^{(k-1)}(q_{t+1})$ denote the durations of the states q_{t-1} and q_{t+1}, respectively. The duration of the former is measured from the first outward state transition left of $t-1$ to $t-1$, and the latter from $t+1$ to the first outward state transition right of $t+1$.

6. SIMULATION RESULTS

Experiments were made for the case of $N = 3$ states and $M = 5$ emission variables. The length of the tested sequence was $T = 400$. The parameters of the true model were

$$\pi = [0.8\ \ 0.1\ \ 0.1]$$

$$\beta = [10\ \ 20\ \ 35]$$

$$\mathbf{W} = \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$$

and

$$\mathbf{B} = \begin{bmatrix} 0.800 & 0.100 & 0.020 & 0.009 & 0.071 \\ 0.010 & 0.003 & 0.800 & 0.100 & 0.087 \\ 0.010 & 0.003 & 0.050 & 0.050 & 0.887 \end{bmatrix}.$$

The parameters of the Dirichlet distributions α's and γ's were all set to $1/2$. The Gamma distribution had parameters $u = 8$ and $v = 1.4$.

The initial state sequence was generated from uniform priors, and it is shown in Figure 1a (dotted line) together with the true state sequence (solid line). The MAP estimator was used for the state sequence estimation, and it is defined by

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}^{(k)}} \{P(\mathbf{q}^{(k)}|\mathbf{y}, \lambda^{(k)})\} \tag{14}$$

where k is the iteration number. In Figure 1b we have plotted the MAP estimate together with the true sequence. Out

of 400 samples, there were only 3 mismatches, although, as can be seen from Figure 1a, the initial chain was completely different from the true one.

The posterior probability of the estimated sequences is displayed in Figure 2 as a function of the iteration number. It is obvious that the chain needed about 700 iterations to converge. The convergence depends on the parameters of the nonstationary HMM.

Figure 3 shows the histograms of the samples drawn from the posterior of some of the model parameters. We chose to present the results for w_{21}, w_{23}, β_1, β_3, b_{35}, and b_{11}. The histograms were constructed from the samples obtained between iterations $k_1 = 2000$ and $k_2 = 2500$. We can see that all the histograms are concentrated around the true values of the model parameters.

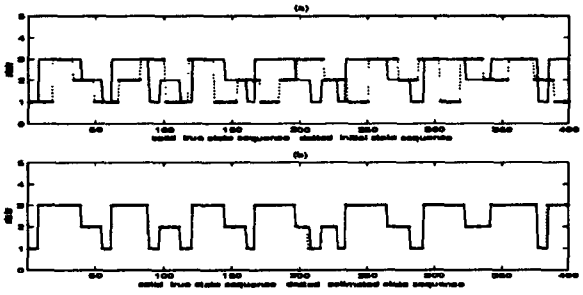

Figure 1: (a) The initial state sequence (dotted line) and the true state sequence (solid line) used in the simulation. (b) The MAP state sequence (dotted line) and the true state sequence(solid line).

7. CONCLUSIONS

We have presented a Gibbs sampling procedure for parameter estimation of nonstationary HMM's. All the parameters of the model except for the number of states were unknown. The scheme is easy to implement because it is straightforward to draw samples from the conditional distributions that define the scheme. The experiments showed quick convergence and very good accuracy of the estimated parameters.

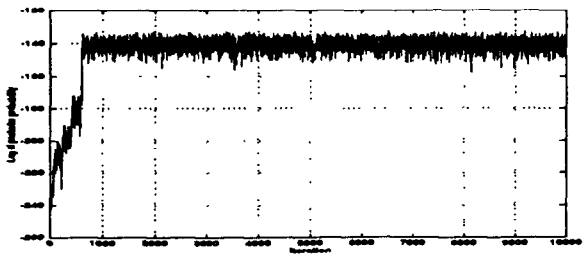

Figure 2: The log of the posterior probability of the estimated state sequence at each iteration.

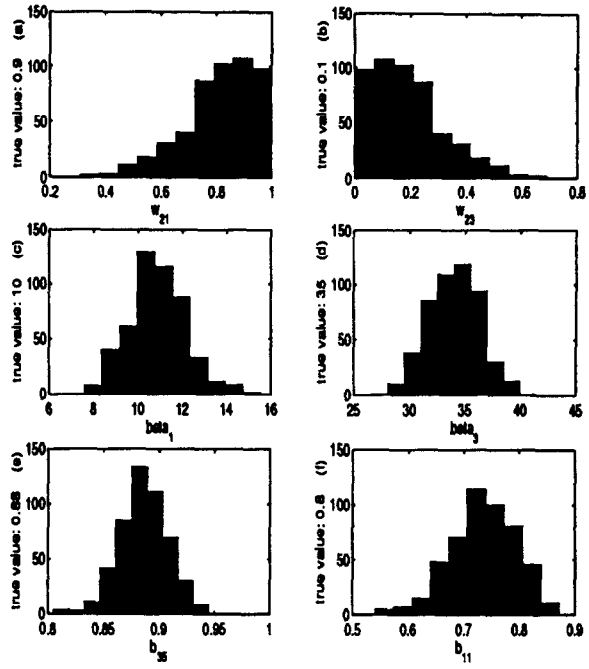

Figure 3: The histograms for several parameters. These were constructed from the samples of parameters between the iterations 2000 and 2500.

8. REFERENCES

[1] C. J. Andrieu, A. J. Doucet, P. Duvaut, W. J. Fitzgerald, and S. J. Godsill, "Simulation-based computational methods for statistical signal processing," Tutorial, ICASSP, Seattle, 1998.

[2] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory.* New York: John Wiley & Sons, 1994.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis – Probabilistic Models of Proteins and Nucleic Acids.* Cambridge, UK: Cambridge University Press, 1998.

[4] J. D. Ferguson, "Variable duration models for speech," Proc. Symp. on the Application of Hidden Markov Models to Text and Speech, Princeton, NJ, pp. 143-179, 1980.

[5] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice.* New York: Chapman Hill, 1996.

[6] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Processing," *Proceedings of IEEE*, vol. 77, pp. 257-286, 1989.

[7] B. Sin and J. H. Kim, "Nonstationary hidden Markov model," *Signal Processing*, vol. 46, pp. 31-46, 1995.

[8] S. V. Vaseghi, "State duration modelling in hidden Markov models," *Signal Processing*, vol. 41, pp. 31-41, 1995.