An MCMC Sampling Approach to Estimation of Nonstationary Hidden Markov Models

Petar M. Djurić, *Senior Member, IEEE*, and Joon-Hwa Chun

***Abstract*—Hidden Markov models (HMMs) represent a very important tool for analysis of signals and systems. In the past two decades, HMMs have attracted the attention of various research communities, including the ones in statistics, engineering, and mathematics. Their extensive use in signal processing and, in particular, speech processing is well documented. A major weakness of conventional HMMs is their inflexibility in modeling state durations. This weakness can be avoided by adopting a more complicated class of HMMs known as nonstationary HMMs. In this paper, we analyze nonstationary HMMs whose state transition probabilities are functions of time that indirectly model state durations by a given probability mass function and whose observation spaces are discrete. The objective of our work is to estimate all the unknowns of a nonstationary HMM, which include its parameters and the state sequence. To that end, we construct a Markov chain Monte Carlo (MCMC) sampling scheme, where sampling from all the posterior probability distributions is very easy. The proposed MCMC sampling scheme has been tested in extensive computer simulations on finite discrete-valued observed data, and some of the simulation results are presented in the paper.**

***Index Terms*—Gibbs sampling, hidden Markov models, Markov chain Monte Carlo, nonstationary.**

I. Introduction

HMMs have played a prominent role in many approaches to statistical analysis of signals and systems. For example, in speech processing, they are the ultimate tool for various tasks including speaker and speech recognition [8], [16], [21], [24]. They have been exploited in communications for suppressing narrowband interference of code division multiple access (CDMA)-spread spectrum signals [17] and in blind equalization for noisy IIR channels [20]. HMMs have also been used in target identification [3], [27], [32], where, for example, the HMMs perform classification using spatiotemporal sequences of radar range profiles [32]. In [23], HMMs have been used to process electroencephalogram data and, in [15], to conduct ion-channel analysis from patch-clamp recordings. In modern biology, with the emergence of molecular genetics and the work on the Human Genome Project, an immense amount of data is being produced that require use of sequence analysis methods and where the HMMs seem very well fitted for extracting information from the data [9], [10]. In addition, current methods for automatic classification of protein sequences into structure/function groups and DNA sequence multiple alignment are carried out by HMMs [18], [19]. Other areas of application include econometrics [7] and finance [28].

A major structural weakness of the conventional HMMs is its inflexibility to model state durations. Their state durations have fixed geometric distributions, and they imply a limited range of applications of the HMMs [24]. In [11], Ferguson introduced the variable duration HMM (VDHMM), whose state durations are modeled by various types of probability distributions. These models are more complex for analysis than the conventional ones, but they are also considerably more flexible in modeling signals, which significantly widens the range of their applications.

Most of the previous work on estimating the unknowns of VDHMMs is on extending the methods of the conventional HMMs, that is, on dynamic programming-based algorithms and maximum likelihood estimators. Later, a different parameterization of the variable state durations was introduced, where the state transition probabilities are explicitly modeled as functions of time [29], [31] and, therefore, are referred to as nonstationary HMMs (NSHMMs). A more recent article that addresses NSHMMs is [4]. We show, however, that the VDHMMs and the NSHMMs are equivalent but that the latter are often more tractable for use. Recently, a Markov chain Monte Carlo (MCMC) scheme has been applied for estimation of conventional HMMs [1], [14], [26]. Here, we propose an MCMC procedure for estimation of NSHMMs. It is assumed that the observed sequence can be modeled by a NSHMM with known number of states and that the state sequence and all the model parameters are unknown. Although the models in this paper are with discrete observation spaces, it should be noted that the proposed methodology can be extended to models with continuous observations spaces. We construct a Gibbs sampling scheme where all the posterior distributions of the unknowns 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. Convergence is assessed by running several parallel Markov chains and by computing the *scale reduction* of each estimated unknown based on the between- and within-sequence variances of the estimates. Simulation results of various experiments for testing the performance of the approach are presented. The main contributions of the paper are the extension of the MCMC-based methods to estimation of NSHMMs (this includes a variety of original details such as the block-wise

Manuscript received July 17, 2000; revised January 22, 2002. This work was supported by the National Science Foundation under Award CCR-9903120. The associate editor coordinating the review of this paper and approving it for publication was Dr. Kristine L. Bell.

P. M. Djurić is with the Department of Electrical and Computer Engineering, State University of New York at Stony Brook, Stony Brook, NY 11794 USA (e-mail: djuric@ece.sunysb.edu).

J.-H. Chun was with the Department of Electrical and Computer Engineering, State University of New York at Stony Brook, NY 11794 USA. He is now with SandBridge Technologies Inc., White Plains, NY 10601 USA.

Publisher Item Identifier S 1053-587X(02)03283-X.

Gibbs sampling, which we found empirically to improve the convergence of the Gibbs sampler) and the establishment of equivalence of the VDHMMs and NSHMMs.

II. Review of Conventional HMMs

First, we provide a very brief review of conventional HMMs 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 states of the system may change with time, and at the time instants $t = 1, 2, \ldots$, they are denoted by q_t, where $q_t \in \mathcal{S}$. The dynamics of the system is described by a Markov chain, that is, when at time t the system is in state S_i, there is a fixed probability that at time $t+1$, it will be in state S_j, where the probability depends only on the state at time t. This is expressed by

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

The complete description of the state transitions is given by the matrix $\mathbf{A} = \{a_{ij}\}$, where

$$a_{ij} \geq 0, \quad \text{and} \quad \sum_{j=1}^{N} a_{ij} = 1.$$

In many modeling scenarios, it is assumed that the state sequence is not known, i.e., 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 elements b_{jk} are known as emission probabilities and are defined according to

$$b_{jk} = P(y_t = v_k \,|\, q_t = S_j), \quad 1 \leq j \leq N, \quad 1 \leq k \leq M$$

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 $\boldsymbol{\pi} = (\pi_1\, \pi_2\, \cdots\, \pi_N)$, where $\pi_i = P(q_1 = S_i), i = 1, 2, \ldots, N$, with $\sum_{i=1}^{N} \pi_i = 1$. The three probability distributions described by $\mathbf{A}, \mathbf{B}$, and $\boldsymbol{\pi}$ are, in short, denoted by λ, or

$$\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}).$$

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

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

which means conditional independence of the observations. A graphical representation of a conventional HMM is presented in Fig. 1.

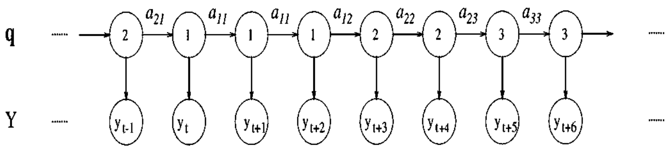

Fig. 1. Representation of a conventional HMM.

There are three basic problems related to HMMs [24], and in order of increasing complexity, they are the following.

1) Given a set of observations $\mathbf{y}^{\top} = [y_1\, y_2\, \cdots\, y_T]$ and the model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$, the objective is to find the probability of the observed sequence $\mathbf{y}, P(\mathbf{y} \,|\, \lambda)$.
2) Given a set of observations $\mathbf{y}^{\top} = [y_1\, y_2\, \cdots\, y_T]$ and the model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$, the objective is to find the corresponding state sequence $\mathbf{q}$.
3) Given a set of observations $\mathbf{y}^{\top} = [y_1\, y_2\, \cdots\, y_T]$, the objective is to find the state sequence $\mathbf{q}$ as well as the model parameters λ.

The solutions to these three problems are well known [24]. The first problem 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.

III. Nonstationary HMMs

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

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

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

One way of modifying the conventional HMM is by way of introducing state duration probability mass functions $P_k(d)$, where $k = 1, 2, \ldots, N$, and $d = 1, 2, \ldots$, [11], [24]. These models are known as variable duration HMMs (VDHMMs), and their state sequences are generated along the following steps.

1) Generate q_1 from the initial state distribution $\boldsymbol{\pi}$.
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 $q_l = S_k$ for $l = t+1, \ldots, t+d-1$ for as long as $l \leq T$.
5) Set $t = t + d$.
6) If $t \leq T$, draw the next state q_t from the transition probabilities $a'_{kj} = a_{kj}/(1 - a_{kk})$, where $q_{t-1} = S_k \neq S_j$, and go back to 3; otherwise, terminate the procedure.

A graph that represents the generation of a VDHMM is given in Fig. 2.

It is important to note that the methods used for solving the three basic problems of conventional HMMs can be extended to accommodate VDHMMs. The extensions, however, entail work

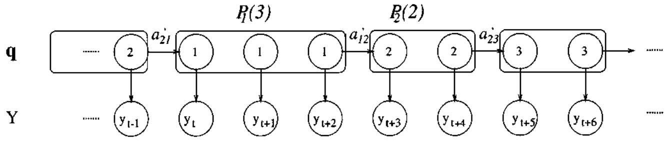

Fig. 2. Representation of a VDHMM.

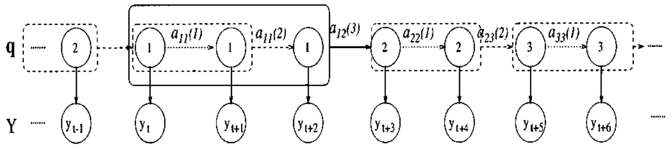

Fig. 3. Representation of an NSHMM.

with a more complicated model and with larger number of unknowns.

A different parameterization of the state duration can be achieved by allowing all the transition probabilities a_{ij} to be functions of d, which we denote by $a_{ij}(d)$. More specifically, $a_{ij}(d)$ is the probability that the system will switch from S_i to S_j, given that the system has already been in the state S_i for d consecutive time units [29], [31], that is

$$a_{ij}(d) = P(q_t = S_j \,|\, q_{t-1} = q_{t-2} = \cdots = q_{t-d} = S_i).$$

The transitional probabilities are thus functions of time, and therefore, we refer to these HMMs as nonstationary HMMs (NSHMMs).

The generation of states according to the NSHMM proceeds in a different way, and it can be summarized as follows.

1) Generate q_1 from the initial state distribution $\boldsymbol{\pi}$, 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.

A graphic representation of the generation of an NSHMM is shown in Fig. 3.

Proposition 1: The relationship between the duration probability mass functions $P_i(d)$ and the self-transition probabilities $a_{ii}(d)$ is given by

$$a_{ii}(d) = \begin{cases} 1 - P_i(d), & d = 1 \\ \frac{1-\sum_{k=1}^{d} P_i(k)}{1-\sum_{l=1}^{d-1} P_i(l)}, & d > 1. \end{cases} \tag{1}$$

Proof: It is straightforward to write

$$P_i(d) = \begin{cases} 1 - a_{ii}(d), & d = 1 \\ \prod_{k=1}^{d-1}(1 - a_{ii}(d))a_{ii}(k), & d > 1. \end{cases} \tag{2}$$

Since, in (2), $P_i(d)$ is represented by the duration specific $a_{ii}(d)$ for each d, the probabilities $a_{ii}(d)$ can be expressed

$$\begin{aligned} a_{ii}(1) &= 1 - P_i(1) \\ a_{ii}(2) &= 1 - \frac{P_i(2)}{a_{ii}(1)} = 1 - \frac{P_i(2)}{1 - P_i(1)} \\ &= \frac{1 - P_i(1) - P_i(2)}{1 - P_i(1)} \\ a_{ii}(3) &= 1 - \frac{P_i(3)}{a_{ii}(1)a_{ii}(2)} = 1 - \frac{P_i(3)}{1 - P_i(1) - P_i(2)} \\ &= \frac{1 - P_i(1) - P_i(2) - P_i(3)}{1 - P_i(1) - P_i(2)} \\ &\vdots \end{aligned}$$

or in general

$$\begin{aligned} a_{ii}(d) &= 1 - \frac{P_i(d)}{\prod_{k=1}^{d-1} a_{ii}(k)} = 1 - \frac{P_i(d)}{1 - \sum_{l=1}^{d-1} P_i(l)} \\ &= \frac{1 - \sum_{k=1}^{d} P_i(k)}{1 - \sum_{l=1}^{d-1} P_i(l)}. \end{aligned} \qquad \diamondsuit$$

Here, we point out that the self-transition probability $a_{ii}(d)$ is defined as the ratio of probabilities of two events: the probability that the state duration is greater than d and the probability that the state duration is greater than $d - 1$ or

$$a_{ii}(d) = \frac{P(\text{duration of } S_i > d)}{P(\text{duration of } S_i > d - 1)}.$$

In [31], the $a_{ii}(d)$s were expressed in terms of the cumulative distribution function $F_i(d) = \sum_{k=1}^{d} P_i(k)$ of the state duration only, or

$$a_{ii}(d) = 1 - F_i(d) = 1 - \sum_{k=1}^{d} P_i(k) \tag{3}$$

which leads to biased state durations. To verify this, we performed a simple experiment of generating states whose duration distribution is Poisson with mean 15. For the self-transition probabilities, we used (1) and (3). The obtained durations are represented by the histograms given in Fig. 4(b) and (c), respectively, which clearly show that (3) should not be used.

The outward state transition probabilities $a_{ij}(d), i \neq j$ can be obtained from $w_{ij}(d)(1 - a_{ii}(d))$, where $w_{ij}(d)$ is the transition weight for state j from i, given that the duration of i has been d. For all i and all d, the weights have to satisfy

$$\sum_{\substack{j=1 \\ i \neq j}}^{N} w_{ij}(d) = 1.$$

The $w_{ij}(d)$s do not necessarily have to be functions of d. Of course, there is a tradeoff between using time varying transition weights $w_{ij}(d)$ and constant weights w_{ij}. With time-varying weights, one can capture more subtle features of the hidden stochastic process, but the estimation of these weights is much more tedious than that of the constant weights. In this paper, the transition weights are regarded as constant parameters, and therefore, we write

$$a_{ij}(d) = \begin{cases} w_{ij} P_i(d), & d = 1, \\ w_{ij} \frac{P_i(d)}{1-\sum_{l=1}^{d-1} P_i(l)}, & d > 1, \end{cases} \quad i \neq j. \tag{4}$$

Proposition 2: The NSHMM with constant state transition weights is equivalent to Ferguson's [11] type VDHMM.

The proof is omitted because it is straightforward, and instead, a simple example is provided. Suppose we have a state sequence $\mathbf{q} = (1\,1\,1\,2\,2\,3\cdots)$. Its joint probability obtained by the VDHMM is

$$P(\mathbf{q} \,|\, \lambda_V) = \pi_1 P_1(3) a'_{12} P_2(2) a'_{23} \ldots \tag{5}$$

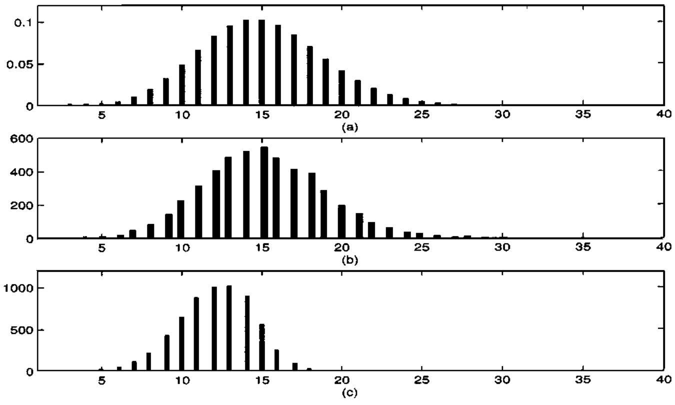

Fig. 4. (a) Poisson probability mass function of state duration with mean 15. (b) Histogram of generated durations by using (1). (c) Histogram of generated durations by using (15).

where λ_V denotes the set of parameters of the VDHMM. The joint probability found by the NSHMM is

$$\tilde{P}(\mathbf{q} \,|\, \lambda_N) = \pi_1 a_{11}(1) a_{11}(2) a_{12}(3) a_{22}(1) a_{23}(2) \ldots$$

where λ_N are the parameters of the NSHMM. Using (1) and (4), we get

$$\tilde{P}(\mathbf{q} \,|\, \lambda_N) = \pi_1 (1 - P_1(1)) \frac{1 - \sum_{k=1}^{2} P_1(k)}{1 - P_1(1)} \times w_{12} \frac{P_1(3)}{1 - \sum_{k=1}^{2} P_1(k)} (1 - P_2(1)) w_{23} \frac{P_2(2)}{1 - P_2(1)} \cdots.$$

After cancellation, we can see that

$$\tilde{P}(\mathbf{q} \,|\, \lambda_N) = \pi_1 P_1(3) w_{12} P_2(2) w_{23} \ldots$$

which is the same as $P(\mathbf{q} \,|\, \lambda_V)$ in (5) if we set the w_{ij}s equal to the a'_{ij}s.

In general, although the VDHMMs and NSHMMs are equivalent models, we found that the NSHMMs are more convenient for use because the description of their data generating seems more “natural,” and they are more tractable for analysis. In addition, the implementation of MCMC sampling for estimation of the parameters and states of the models is then much easier.

IV. Estimation of NSHMMs by MCMC Sampling

MCMC-based methods are procedures that, in the 1990s, attracted great interest among researchers in the Bayesian community [13]. Their advantage over alternative approaches is in their capacity to work with high-dimensional and complex models. In brief, MCMC sampling is a methodology for generating samples from a desired probability distribution function, which is usually referred to as a target distribution. The sample generation proceeds by an evolving Markov chain, and the obtained samples are used for various types of inference. Here, we do not provide a general description of MCMC procedures, but we are referred to some excellent textbooks such as [12] and [13], or, from a signal processing perspective, to [1] and [22]. Of the several known MCMC schemes, we use the Gibbs sampler, which effectively reduces the sampling from high-dimensional distributions to sampling from a series of low-dimensional distributions.

Let the state sequence $\mathbf{q}^\mathsf{T} = [q_1\, q_2\, \cdots\, q_T]$ be a discrete time Markov chain, where $q_t \in \mathcal{S} = \{S_1, S_2, \ldots, S_N\}$, and $\mathbf{y}^\mathsf{T} = [y_1\, y_2\, \cdots\, y_T]$ is a sequence of observations whose alphabet is $\mathcal{V} = \{v_1, v_2, \ldots, v_M\}$. The unknowns are the state sequence $\mathbf{q}$, the initial state probabilities $\boldsymbol{\pi}$, the state transition weights $\mathbf{W}$, the emission probabilities $\mathbf{B}$, and the parameters of the models that describe the state durations $\boldsymbol{\epsilon}$. Thus, λ now is defined by $\lambda = (\mathbf{W}, \mathbf{B}, \boldsymbol{\pi}, \boldsymbol{\epsilon})$. We make the assumption that the durations of the various states follow truncated Poisson distributions with different parameters. The truncated Poisson is used to allow for use of additional information about the state durations. This assumption, however, 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.

The objective is to estimate the unknowns, which are $\mathbf{q}$ and $\lambda = (\mathbf{W}, \mathbf{B}, \boldsymbol{\pi}, \boldsymbol{\epsilon})$, and for that purpose, we use the Bayesian methodology applied via MCMC sampling. It should be noted that all the information about the unknowns is contained in the posteriors of the unknowns. The MCMC sampling methods draw samples of the unknowns from their posteriors so that once the sampling is completed, the posteriors can be approximated using these samples. Moreover, the samples, can be used to obtain different types of estimates of the unknowns.

A. Specification of the Priors

Before we proceed with the Gibbs sampling scheme, we need to specify the prior distributions of all the unknowns. The likelihoods of the initial state probabilities, the state transition weights, and the emission parameters are modeled by multinomial distributions. A standard prior when the likelihood is a multinomial distribution is the multivariate Dirichlet distribution [2]; therefore, we model the priors of the initial state probabilities, the state transition weights, and the emission parameters by multivariate Dirichlet distributions. In particular, if $\boldsymbol{\pi}$ is distributed according to the Dirichlet distribution, where $\sum_{l=1}^{N} \pi_l = 1$, and $0 < \pi_l < 1$, we write

$$\mathcal{D}i(\alpha_1, \alpha_2, \ldots, \alpha_N) = c \left(1 - \sum_{l=1}^{N-1} \pi_l \right)^{\alpha_N - 1} \prod_{l=1}^{N-1} \pi_l^{\alpha_l - 1}$$

where $\alpha_i > 0, i = 1, 2, \ldots, N$ are the parameters of the distribution, and c is the normalizing constant given by

$$c = \frac{\Gamma\left(\sum_{l=1}^{N} \alpha_l\right)}{\prod_{l=1}^{N} \Gamma(\alpha_l)}.$$

In addition, we assume that the durations of the various states follow truncated Poisson distributions with different parameters, and for the priors of these parameters, we choose Gamma distributions. The hyperparameters of all the priors should be selected using prior knowledge about the problem at hand. In our simulations, we found that our results were not sensitive to the choice of the hyperparameters.

More specifically, the priors are defined as follows.

1) The prior of the initial state probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ is the multivariate Dirichlet distribution

of dimension $N-1$, or

$$\boldsymbol{\pi} \sim \mathcal{D}i(\alpha_1, \alpha_2, \ldots, \alpha_N).$$

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

$$P_i(d) \propto \frac{\epsilon_i^d e^{-\epsilon_i}}{d!}, \quad d = 1, 2, \ldots, D$$

where ϵ_i is a parameter that identifies the Poisson distribution of the ith state, and $\epsilon_i > 0$. All the ϵ_is have Gamma distributions $\mathcal{G}a(u, v)$

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

The Gamma distributions are convenient because the posterior of ϵ_i given d is also Gamma distributed. Note that the self-transition probabilities $a_{ii}(d)$ and the outward state transition probabilities $a_{ij}(d)$ can be obtained from (1) and (4).

3) For a given i, the prior distribution of the state transition weights w_{ij} is a multivariate Dirichlet distribution of dimension $N-2$. We use the notation $\mathbf{w}_i^{\mathsf{T}} = [w_{i1}\ w_{i2}\ w_{i,i-1}\ w_{i,i+1}\ \cdots\ w_{iN}]$ and write

$$\mathbf{w}_i \sim \mathcal{D}i(\eta_{i1}, \eta_{i2}, \ldots, \eta_{i,N-1}).$$

We represent all the weights by the matrix $\mathbf{W}$, whose size is $N \times N-1$ and which is defined by $\mathbf{W}^{\mathsf{T}} = [\mathbf{w}_1\ \mathbf{w}_2\ \cdots\ \mathbf{w}_N]$.

4) The emission parameters $\mathbf{B} = \{b_{ik}\}$ also have multivariate Dirichlet priors ($(M-1)$-dimensional), i.e.,

$$\mathbf{b}_i \sim \mathcal{D}i(\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iM})$$

where $\mathbf{b}_i = (b_{i1}b_{i2}\ldots, b_{i,M}), i = 1, 2, \ldots, N$.

B. Gibbs Sampling Procedure

With the chosen priors, the Gibbs sampling procedure is rather simple, and the steps of its implementation are carried out as follows. At iteration k, do the following.

1) Draw $\boldsymbol{\pi}^{(k)}$ from the $(N-1)$-dimensional multivariate Dirichlet distribution according to

$$\begin{aligned}\boldsymbol{\pi} &\sim p\left(\boldsymbol{\pi} \,|\, \boldsymbol{\epsilon}^{(k-1)}, \mathbf{W}^{(k-1)}, \mathbf{B}^{(k-1)}, \mathbf{q}^{(k-1)}, \mathbf{y}\right)\\ &= p\left(\boldsymbol{\pi} \,\middle|\, \mathbf{q}^{(k-1)}\right)\\ &= \mathcal{D}i\left(\alpha_1 + \delta_{q_1,1}^{(k-1)}, \alpha_2 + \delta_{q_1,2}^{(k-1)}, \ldots, \alpha_N + \delta_{q_1,N}^{(k-1)}\right) \quad (6)\end{aligned}$$

where $p(\,\cdot\,)$ denotes density, and $\delta_{q_1,i}$ is the Kronecker delta function

$$\delta_{q_1,i}^{(k-1)} = \begin{cases} 1, & q_1^{(k-1)} = S_i \\ 0, & \text{otherwise.} \end{cases}$$

In (6) and in the sequel, the superscripts in brackets denote iteration number.

2) Draw $\epsilon_i^{(k)}$, for $i = 1, 2, \ldots, N$, using

$$\begin{aligned}\epsilon_i &\sim p\left(\epsilon_i \,\middle|\, \boldsymbol{\pi}^{(k)}, \mathbf{W}^{(k-1)}, \mathbf{B}^{(k-1)}, \mathbf{q}^{(k-1)}, \mathbf{y}\right)\\ &= p\left(\epsilon_i \,\middle|\, \mathbf{q}^{(k-1)}\right)\\ &= \mathcal{G}a\left(u + \tilde{d}_i^{(k-1)}, v + m_i^{(k-1)}\right) \quad (7)\end{aligned}$$

where $\tilde{d}_i^{(k-1)} = \sum_{t=1}^{T} \delta_{q_t^{(k-1)}, S_i}$, and $m_i^{(k-1)}$ is the number of segments in state S_i at iteration $k-1$. (A segment is a subsequence of equal states, where the first and the last states of the sequence are preceded and followed, respectively, by different states.)

3) For $i = 1, 2, \ldots, N$, draw $\mathbf{w}_i$ from a multivariate Dirichlet distribution, i.e.,

$$\begin{aligned}\mathbf{w}_i &\sim p\left(\mathbf{w}_i \,\middle|\, \boldsymbol{\pi}^{(k)}, \boldsymbol{\epsilon}^{(k)}, \mathbf{B}^{(k-1)}, \mathbf{q}^{(k-1)}, \mathbf{y}\right)\\ &= p\left(\mathbf{w}_i \,\middle|\, \mathbf{q}^{(k-1)}\right)\\ &= \mathcal{D}i\left(\eta_{i1} + n_{i1}^{(k-1)}, \eta_{i2} + n_{i2}^{(k-1)}, \ldots, \eta_{i,N-1} + n_{i,N-1}^{(k-1)}\right)\end{aligned}$$

where $n_{ij}^{(k-1)}$ is the number of transitions from state i to state j at iteration $k-1$.

4) The $a_{ii}^{(k)}(d)$s are obtained from (1), where $P_i(d)$ is a truncated Poisson distribution with parameter $\epsilon_i^{(k)}$. Note that we do not sample from the truncated Poisson distribution but, rather, use it to compute the probabilities $a_{ii}(d)$.

5) For $i = 1, 2, \ldots, N$, draw $\mathbf{b}_i$ from a multivariate Dirichlet distribution, that is

$$\begin{aligned}\mathbf{b}_i &\sim p\left(\mathbf{b}_i \,\middle|\, \boldsymbol{\pi}^{(k)}, \boldsymbol{\epsilon}^{(k)}, \mathbf{W}^{(k)}, \mathbf{q}^{(k-1)}, \mathbf{y}\right)\\ &= p(\mathbf{b}_i \,|\, \mathbf{q}^{(k-1)}, \mathbf{y})\\ &= \mathcal{D}i\left(\gamma_{i1} + \tilde{m}_{i1}^{(k-1)}, \gamma_{i2} + \tilde{m}_{i2}^{(k-1)}, \ldots, \gamma_{iM} + \tilde{m}_{iM}^{(k-1)}\right)\end{aligned}$$

where $\tilde{m}_{ij}^{(k-1)}$ is the number of symbols v_j in state S_i. Note that the emission probabilities can be integrated out instead of being sampled.

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

$$\begin{aligned}q_t &\sim P\Big(q_t \,\Big|\, q_1^{(k)}, \ldots, q_{t-1}^{(k)}, q_{t+1}^{(k-1)}, \ldots, q_T^{(k-1)}\\ &\qquad \boldsymbol{\pi}^{(k)}, \boldsymbol{\epsilon}^{(k)}, \mathbf{W}^{(k)}, \mathbf{B}^{(k)}, \mathbf{y}\Big)\\ &= P\Big(q_t \,\Big|\, q_{t-1}^{(k)}, d\left(q_{t-1}^{(k)}\right), q_{t+1}^{(k-1)}, \ldots, q_{t+\tau}^{(k-1)}\\ &\qquad \boldsymbol{\pi}^{(k)}, \boldsymbol{\epsilon}^{(k)}, \mathbf{W}^{(k)}, \mathbf{B}^{(k)}, y_t\Big)\\ &\propto P\left(q_t \,\middle|\, q_{t-1}^{(k)}, d\left(q_{t-1}^{(k)}\right)\right) P\left(q_{t+1}^{(k-1)} \,\middle|\, q_t, d(q_t)\right)\\ &\quad \times P\left(q_{t+2}^{(k-1)} \,\middle|\, q_{t+1}^{(k-1)}, d\left(q_{t+1}^{(k-1)}\right)\right)\\ &\quad \times P\left(q_{t+3}^{(k-1)} \,\middle|\, q_{t+2}^{(k-1)}, d\left(q_{t+2}^{(k-1)}\right)\right) \cdots\\ &\qquad P\left(q_{t+\tau}^{(k-1)} \,\middle|\, q_{t+\tau-1}^{(k-1)}, d\left(q_{t+\tau-1}^{(k-1)}\right)\right) P(y_t \,|\, q_t)\\ &= a_{q_{t-1}^{(k)}, q_t}\left(d\left(q_{t-1}^{(k)}\right)\right) a_{q_t, q_{t+1}^{(k-1)}}(d(q_t))\\ &\quad \times a_{q_{t+1}^{(k-1)}, q_{t+2}^{(k-1)}}\left(d\left(q_{t+1}^{(k-1)}\right)\right)\\ &\quad \times a_{q_{t+2}^{(k-1)}, q_{t+3}^{(k-1)}}\left(d\left(q_{t+2}^{(k-1)}\right)\right) \cdots\\ &\quad \times a_{q_{t+\tau-1}^{(k-1)}, q_{t+\tau}^{(k-1)}}\left(d\left(q_{t+\tau-1}^{(k-1)}\right)\right) b_{y_t, q_t}\end{aligned}$$

where $q_{t+\tau}^{(k-1)} \neq q_{t+\tau-1}^{(k-1)}$, $q_{t+\tau-1}^{(k-1)} = q_{t+1}^{(k-1)}$, and $2 \leq \tau$.

C. Block-Wise Gibbs Sampling

We have found that with the inclusion of an additional block-wise sampling step, the speed of convergence of our Gibbs sampler improves considerably. To explain the block-wise Gibbs sampling, we define another index set for $\mathbf{q}, \{1, \ldots, R\}$, where R is the total number of segments. Let $\tilde{q}_r$ be the rth segment of $\mathbf{q}$ with boundaries t_r and $t_{r+1} - 1$, where $1 \leq t \leq T$ and $1 \leq r \leq R$. Now, with two types of indices t and r, the state sequence can be described by either the collection of states $\mathbf{q} = \{q_1, \ldots, q_T\}$ or the collection of segments $\tilde{\mathbf{q}} = \{\tilde{q}_1, \ldots, \tilde{q}_R\}$ with associated boundaries.

The revised Gibbs sampling procedure now consists of three main steps

- sampling of λ;
- sampling of $\{q_t\}_{t=1}^T$;
- sampling of $\{\tilde{q}_r\}_{r=1}^R$.

The implementation of the steps is carried out sequentially, where the sampling of λ and $\{q_t\}_{t=1}^T$ follows the procedure described in the previous section. The $\tilde{q}_r$s are sampled according to

$$\tilde{q}_r \sim p\left(\tilde{q}_r \mid q_{t_r-1}, q_{t_{r+1}}, d(q_{t_r-1}), \lambda, y_{t_r}, \ldots, y_{t_{r+1}-1}\right)$$
$$\propto \prod_{t=t_r}^{t_{r+1}} P(q_t \mid q_{t-1}, d(q_{t-1})) \prod_{t=t_r}^{t_{r+1}-1} P(y_t \mid q_t).$$

The first factor in the last expression accounts for the state sequence probability of the segment and the second factor for the emission probabilities of the segment. Therefore, for example, if the collection of segments after the second step is $\tilde{q}_1, \tilde{q}_2, \tilde{q}_3, \ldots$, and we want to sample $\tilde{q}_2$, we do so by drawing the second segment from the probability mass function $P(\tilde{q}_2 \mid q_{t_1-1}, q_{t_2}, d(q_{t_1-1}), y_{t_1}, \ldots, y_{t_2-1})$, where t_1 and t_2 are the instants that denote the first samples of the second and third segments, respectively. Note that the block-wise sampling does not increase the number of segments; it may only remove some of them, which often induces faster convergence of the chain.

D. Problem of Label-Switching

In our problem, the states of the HMM were labeled as $S_1, S_2, \ldots, S_N$. It is obvious that this labeling is arbitrary and that the joint posterior distribution of the unknowns has $N!$ modes. The problem is that the likelihood function is the same for all permutations of the states and their parameters. If the prior is symmetric for all permutations of the parameters, the posterior is also symmetric, which creates problems in summarizing joint posterior distributions by marginal distributions and estimating unknowns by their posterior means. The MCMC or any other iterative procedure would usually explore only one of the modes of the posterior. This implies that a postprocessing step would be needed in order to interpret the obtained results. The postprocessing step would, in general, be problem dependent as the arbitrarily labeled states would have some specific physical meaning. Here, we do not address this problem. For recent results on label switching, see [6] and [30].

V. Assessment of Convergence

An important issue in the use of MCMC-based methods is the assessment of convergence to the target distributions of the used chains. Namely, it is important to diagnose how long a Markov chain must be run before the generated samples can be considered drawn (approximately) from the stationary distribution of the chain. There are various procedures for assessing convergence, and they can be classified in various ways. For example, some procedures use one chain to assess convergence and others use multiple chains, or some can diagnose convergence of full joint densities, and others cannot, or some methods are computationally intensive and others are not [5].

In our work, we have adopted a method based on multiple chains and the use of between-sequence and within sequence variances. Here, we briefly describe the procedure [12].

Let θ be a parameter that is being simulated with J different chains, and let $\theta_j^{(k)}$ be the kth sample in the jth chain, where $j = 1, 2, \ldots, J$, and $k = 1, 2, \ldots, K$. First, we compute the between- and within- sequence variances $\hat{\sigma}_B^2$ and $\hat{\sigma}_W^2$, where the between-sequence variance is obtained from

$$\hat{\sigma}_B^2 = \frac{K}{J-1} \sum_{j=1}^{J} (\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot})^2$$

with

$$\bar{\theta}_{\cdot j} = \frac{1}{K} \sum_{k=1}^{K} \theta_j^{(k)}$$

and

$$\bar{\theta}_{\cdot\cdot} = \frac{1}{J} \sum_{j=1}^{J} \bar{\theta}_{\cdot j}$$

and the within-sequence variance from

$$\hat{\sigma}_W^2 = \frac{1}{J(K-1)} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\theta_j^{(k)} - \bar{\theta}_{\cdot j}\right)^2.$$

Then, the marginal posterior variance of $\hat{\theta}$ is evaluated using

$$\hat{\sigma}^2 = \frac{K-1}{K} \hat{\sigma}_W^2 + \frac{1}{K} \hat{\sigma}_B^2.$$

Finally, we estimate the *potential scale reduction* of $\hat{\theta}$, $\sqrt{\hat{R}}$ by

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{\sigma}^2}{\hat{\sigma}_W^2}}.$$

Note that as $K \to \infty$, $\sqrt{\hat{R}}$ declines to 1. A recommendation for convergence assessment from [12], which is adopted here, is that the potential scale reduction is computed for every parameter and that $\sqrt{\hat{R}}$s for each of them is below 1.2.

VI. Simulation Results

We performed two experiments where the number of states was $N = 3$ and the number of emission variables was $L = 7$. The length of the tested sequence was $T = 500$. The parameters of the true model in the first experiment were

$$\pi_1 = 0.8, \quad \pi_2 = 0.1, \quad \pi_3 = 0.1$$

the Poisson parameters were

$$\epsilon_1 = 10, \quad \epsilon_2 = 20, \quad \epsilon_3 = 35$$

the transition weights were

$$w_{12} = 0.2, \quad w_{13} = 0.8$$
$$w_{21} = 0.8, \quad w_{23} = 0.2$$
$$w_{31} = 0.2, \quad w_{32} = 0.8$$

and the emission parameters were set to

$$\begin{aligned}
&b_{11} = 0.9, \quad b_{21} = 0.01, \quad b_{31} = 0.01, \quad b_{41} = 0.01\\
&b_{51} = 0.01, \quad b_{61} = 0.01, \quad b_{71} = 0.05\\
&b_{12} = 0.01, \quad b_{22} = 0.9, \quad b_{32} = 0.01, \quad b_{42} = 0.01\\
&b_{52} = 0.01, \quad b_{62} = 0.01, \quad b_{72} = 0.05\\
&b_{13} = 0.01, \quad b_{23} = 0.01, \quad b_{33} = 0.9, \quad b_{43} = 0.01\\
&b_{53} = 0.01, \quad b_{63} = 0.01, \quad b_{73} = 0.05.
\end{aligned}$$

In the second experiment, we changed the emission parameters only to make the problem much more challenging. We used the same state sequence as in the first experiment and generated the observations using the following emission parameters:

$$\begin{aligned}
&b_{11} = 0.6, \quad b_{21} = 0.0667, \quad b_{31} = 0.0667, \quad b_{41} = 0.0667\\
&b_{51} = 0.0667, \quad b_{61} = 0.0667, \quad b_{71} = 0.0667\\
&b_{12} = 0.0667 \quad b_{22} = 0.6, \quad b_{32} = 0.0667, \quad b_{42} = 0.0667\\
&b_{52} = 0.0667, \quad b_{62} = 0.0667, \quad b_{72} = 0.0667\\
&b_{13} = 0.0667, \quad b_{23} = 0.0667, \quad b_{33} = 0.6, \quad b_{43} = 0.0667\\
&b_{53} = 0.0667, \quad b_{63} = 0.0667, \quad b_{73} = 0.0667.
\end{aligned}$$

In the two experiments, the parameters of all the Dirichlet distributions, the αs, γs, and ηs were all set to 1, and the Gamma distribution in (7) had parameters $u = 2$ and $v = 1$. The parameter D of the truncated Poisson distribution was $D = 100$. The initial state sequence was generated from uniform priors, and it is shown in Fig. 5(a) (dotted line) together with the true state sequence (solid line). As already mentioned, we used the same initial state sequence in the second experiment, and it is shown in Fig. 6(a). Note that the Dirichlet priors are noninformative. The Gamma priors of the ϵs have two hyperparameters, and the final result is robust to the choice of these parameters. For example, the true values of ϵ_1, ϵ_2, and ϵ_3 are 10, 20, and 35, respectively, and since the Gamma distributions have parameters $u = 2$ and $v = 1$, the true values of the ϵs are in the tails of the prior.

For estimation of the state sequence, we used the MAP estimator, which is defined by

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}^{(k)}} \left\{ P(\mathbf{q}^{(k)}, \lambda^{(k)} \,\middle|\, \mathbf{y}) \right\}$$

where

$$P\left(\mathbf{q}^{(k)}, \lambda^{(k)} \,\middle|\, \mathbf{y}\right) \propto P\left(\mathbf{y} \,\middle|\, \mathbf{q}^{(k)}, \lambda^{(k)}\right) \times P\left(\mathbf{q}^{(k)} \,\middle|\, \lambda^{(k)}\right) p\left(\lambda^{(k)}\right) \qquad (8)$$

where $p(\lambda^{(k)})$ is the prior of $\lambda^{(k)}$ evaluated at $\lambda^{(k)}$, and k is the iteration number. The first two factors on the right of the proportionality sign in (8) are computed according to

$$P\left(\mathbf{y} \,\middle|\, \mathbf{q}^{(k)}, \lambda^{(k)}\right) = \prod_{t=1}^{T} b_{y_t q_t}{}^{(k)}$$

and

$$P\left(\mathbf{q}^{(k)} \,\middle|\, \lambda^{(k)}\right) = \pi_{q_1}^{(k)} \prod_{t=2}^{T} a_{q_{t-1}^{(k)} q_t^{(k)}} \left(d\left(q_{t-1}^{(k)}\right)\right).$$

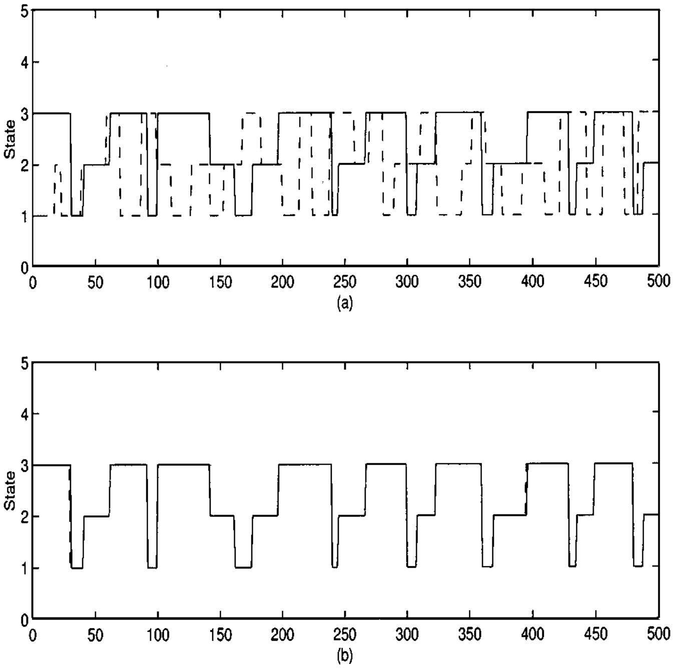

Fig. 5. Experiment 1. (a) Initial state sequence (dotted line) and the true state sequence (solid line) used in the simulation. (b) MAP state sequence (dotted line) and the true state sequence (solid line).

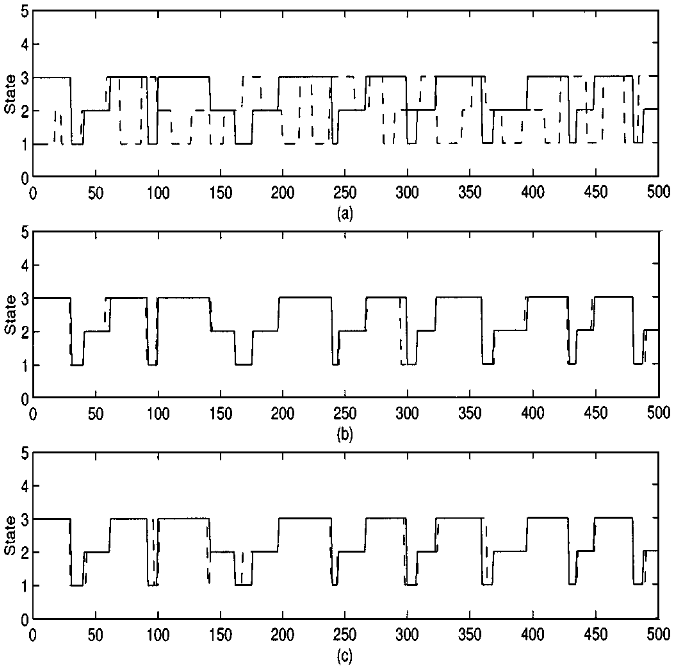

Fig. 6. Experiment 2. (a) Initial state sequence (dotted line) and the true state sequence (solid line) used in the simulation. (b) MAP state sequence (dotted line) and the true state sequence (solid line). (c) Estimate of the state sequence (dotted line) obtained by the EM algorithm combined with the Viterbi algorithm and the true state sequence (solid line).

In Figs. 5(b) and 6(b), we have plotted the MAP estimates together with the true sequences. In the first experiment, out of 500 samples, there were only two mismatches, although, as can be seen from Fig. 5(a), the starting hidden chain was completely different from the true one. In the second experiment, we observed 23 mismatches out of 500 samples, and this increased number is due to the more "ambiguous" values of the emission parameters in the second experiment. In Fig. 6(c), we see the results obtained by using the expectation-maximization

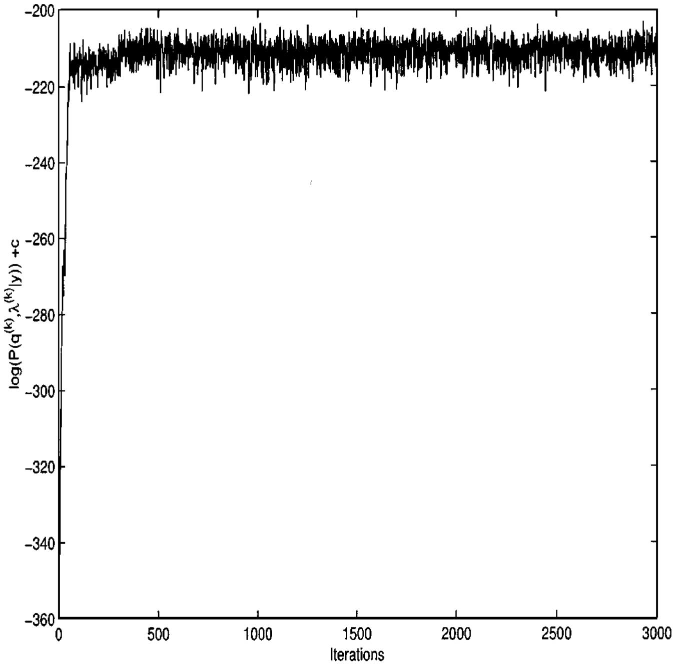

Fig. 7. Experiment 1. Logarithm of the posterior probability of the estimated state sequence at each iteration.

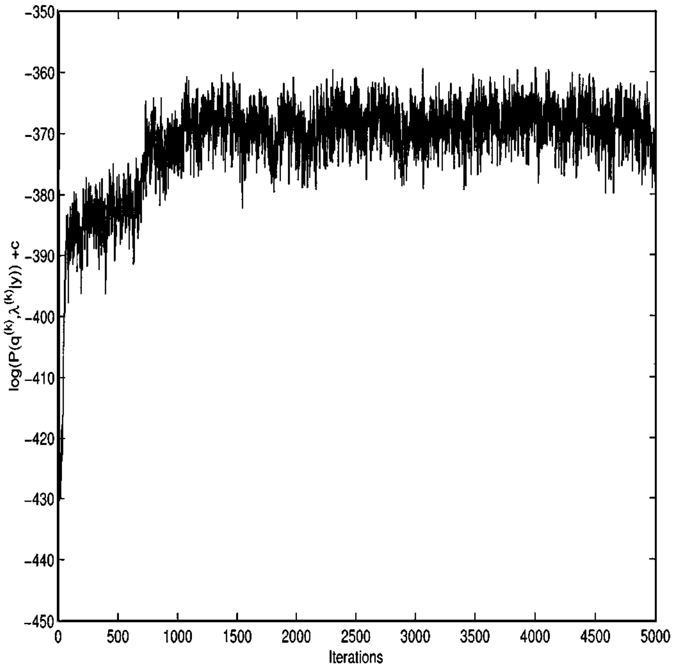

Fig. 8. Experiment 2. Logarithm of the posterior probability of the estimated state sequence at each iteration.

method combined with the Viterbi (EM+Viterbi) algorithm [24]. When the method was started with the same initial state sequence as in Fig. 6(a), the method could not converge to a solution close to the true sequence. After several trials with modified initial sequences, it converged to the solid line in Fig. 6(c), and it had 34 mismatches. In general, while running the experiments, we observed the following: 1) The results obtained by the EM+Viterbi algorithm are not as accurate as the results of the MCMC method, and 2) the EM+Viterbi algorithm is sensitive to initializations.

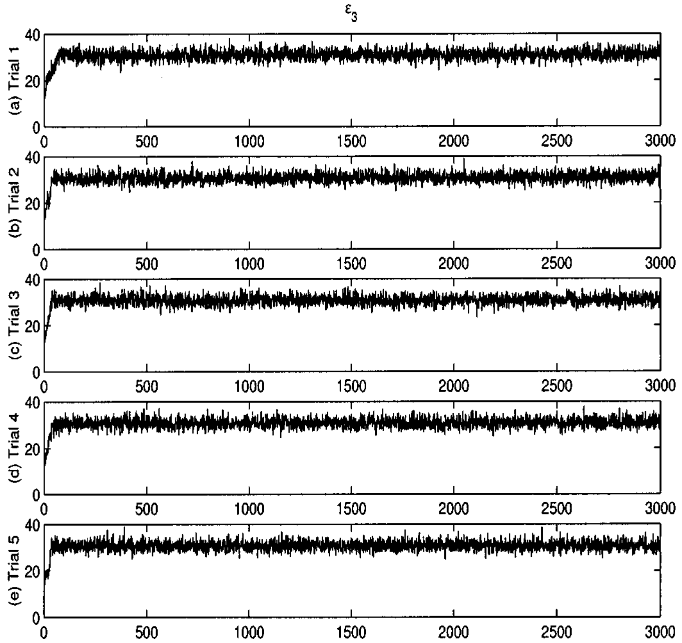

Fig. 9. Experiment 1. Convergence assessment with five different Markov chains. The estimated $\sqrt{\hat{R}}$ from the last 1500 samples was 1.0026.

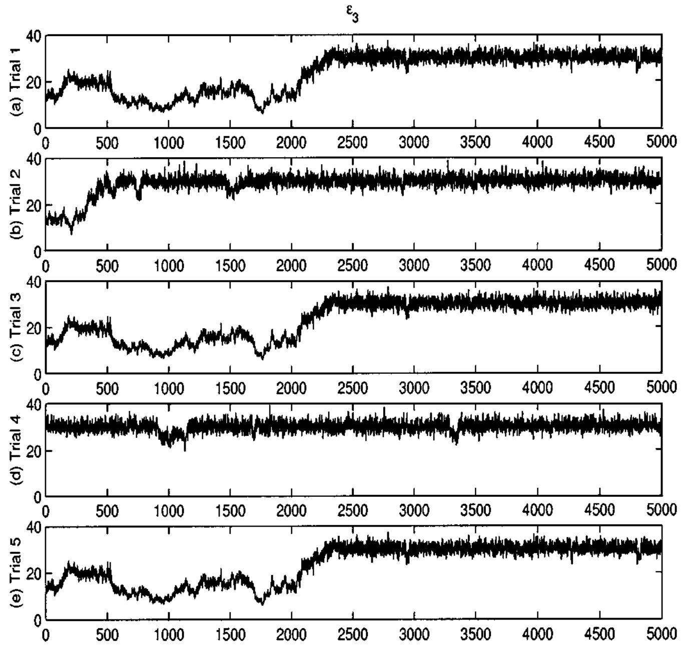

Fig. 10. Experiment 2. Convergence assessment with five different Markov chains. The estimated $\sqrt{\hat{R}}$ from the last 2500 samples was 1.0013.

The posterior probabilities of the estimated sequences of the two experiments as functions of the iteration number are displayed in Figs. 7 and 8. It is observed that in the first experiment, the chain needed about 150 iterations to converge, whereas in the second, it needed more than 2000 iterations. Again, the convergence depends on the parameters of the nonstationary HMM.

As explained before, the convergence of the MCMC samples was assessed by using multiple chains. In Figs. 9 and 10, we show five chains of ϵ_3 for each of the two experiments. The estimated $\sqrt{\hat{R}}$s of ϵ_3 were 1.0026 and 1.0013, respectively. In the

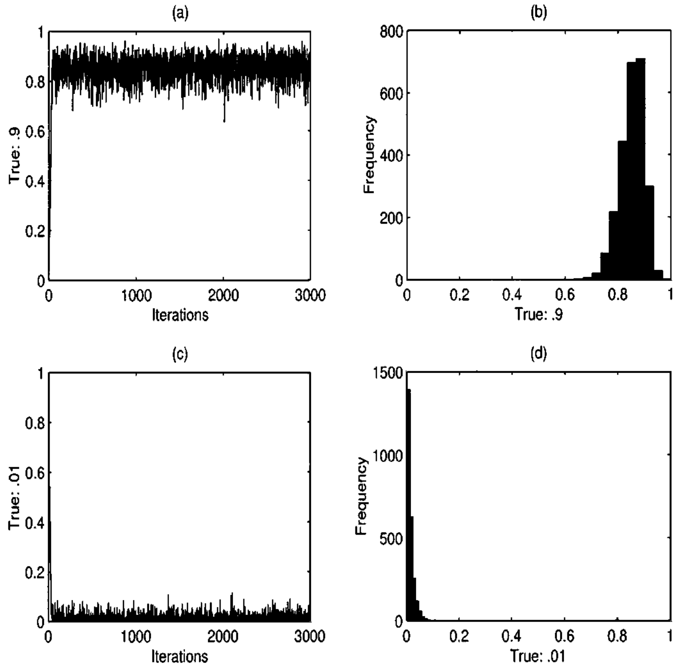

Fig. 11. Experiment 1. (a) Samples of b_{11} at each iteration. (b) Histogram of $b_{11}^{(k)}$. (c) Samples of b_{31} at each iteration. (d) Histogram of $b_{31}^{(k)}$.

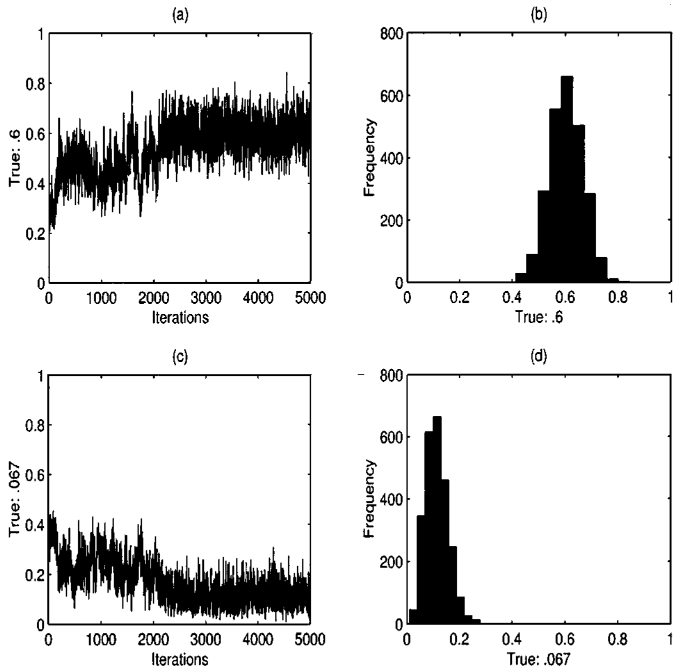

Fig. 12. Experiment 2. (a) Samples of b_{11} at each iteration. (b) Histogram of $b_{11}^{(k)}$. (c) Samples of b_{31} at each iteration. (d) Histogram of $b_{31}^{(k)}$.

first experiment, the last 1500 samples were used in computing $\sqrt{\hat{R}}$ and, in the second, the last 2500 samples.

Some of the parameter estimates are shown in Figs. 11–13. The histograms were constructed from drawn samples after convergence was assessed.

We made two additional experiments under identical conditions, except that this time, we did not use block-wise sampling. Some results are shown in Figs. 14 and 15. In the top figures, we see the logarithms of the posterior probabilities of the estimated

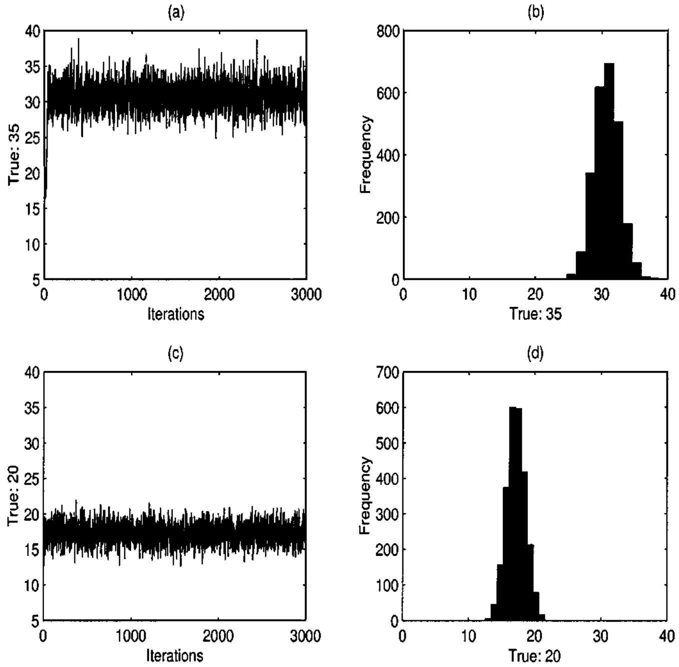

Fig. 13. Experiment 1. (a) Samples of ϵ_3 at each iteration. (b) Histogram of $\epsilon_3^{(k)}$. (c) Samples of ϵ_2 at each iteration. (d) Histogram of $\epsilon_2^{(k)}$.

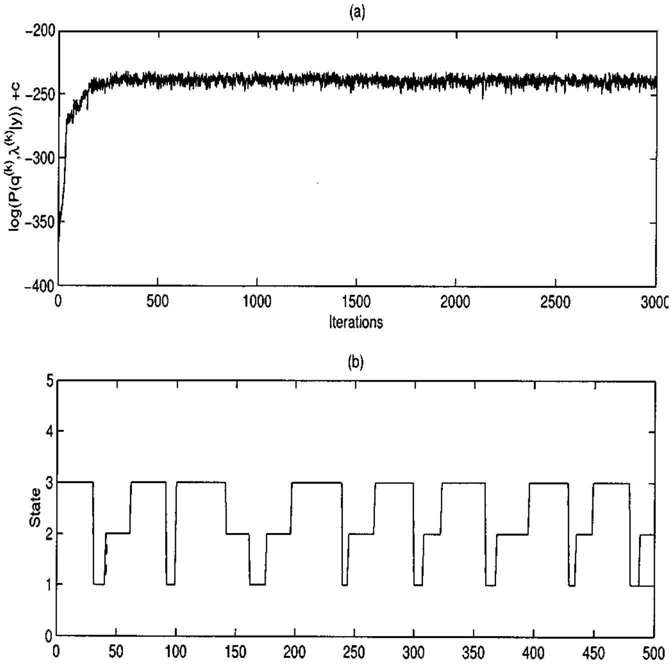

Fig. 14. Experiment 1 (without block-wise sampling). (a) Log of the posterior probability of the estimated state sequence at each iteration. (b) MAP state sequence (dotted line) and the true state sequence (solid line).

state sequences. In the bottom figures, we observe the results of the sequence estimates. This time, in the first experiment, there were 14 mismatches [Fig. 14(b)], and in the second, there were 29 mismatches [Fig. 15(b)].

VII. Conclusion

We have presented a Gibbs sampling procedure for parameter estimation of NSHMMs. All the parameters of the model,

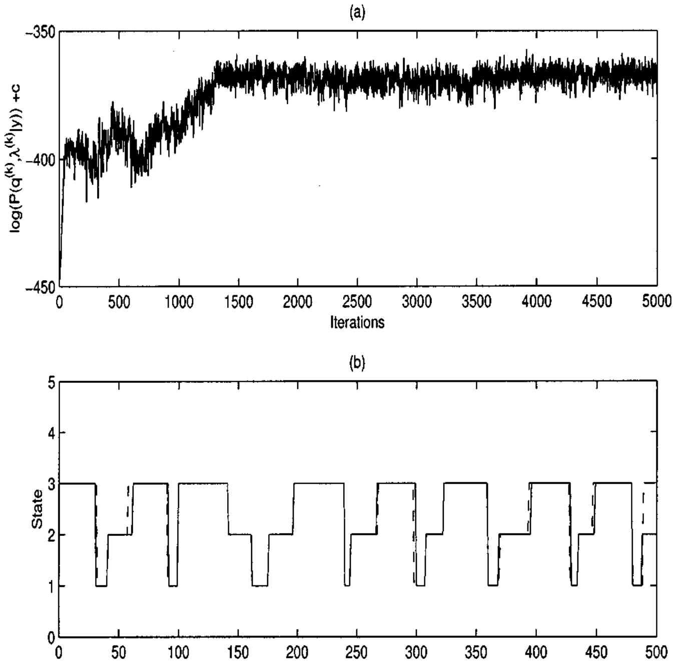

Fig. 15. Experiment 2 (without block-wise sampling). (a) Log of the posterior probability of the estimated state sequence at each iteration. (b) MAP state sequence (dotted line) and the true state sequence (solid line).

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. To improve the convergence to the target distribution of the scheme, a block-wise Gibbs sampling step was added. The experiments showed quick convergence and very good accuracy of the estimated states and model parameters.

References

[1] C. Andrieu, A. Doucet, and W. Fitzgerald, “An introduction to Monte Carlo methods for Bayesian data analysis,” in *Nonlinear Dynamics and Statistics*, A. Mees and R. L. Smith, Eds. Boston, MA: Birkhauser, 2000.

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

[3] P. K. Bharadwaj, P. R. Runkle, and L. Carin, “Target identification with wave-based matched pursuits and hidden Markov models,” *IEEE Trans. Antennas Propagat.*, vol. 47, pp. 1543–1554, 1999.

[4] D. R. Brillinger, P. A. Morettin, R. A. Irizarry, and C. Chiann, “Some wavelet-based analyzes of Markov chain data,” *Signal Process.*, vol. 80, pp. 1607–1627, 2000.

[5] S. P. Brooks and G. O. Roberts, “Convergence assessment techniques foe Markov chain Monte Carlo,” *Statist. Comput.*, vol. 8, pp. 319–335, 1998.

[6] G. Celleux, M. Hurn, and C. P. Robert, “Computational and inferential difficulties withy mixture posterior distributions,” *J. Amer. Statist. Assoc.*, vol. 95, no. 451, 2000.

[7] S. Chib, “Calculating posterior distributions and model estimates in Markov mixture models,” *J. Econometr.*, vol. 75, pp. 79–97, 1996.

[8] J. T. Chien, “Online hierarchical transformation of hidden Markov models for speech recognition,” *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 656–667, June 1999.

[9] G. A. Churchill and B. Lazareva, “Bayesian restoration of a hidden Markov chain with applications to DNA sequencing,” *J. Comput. Biol.*, vol. 6, no. 2, pp. 261–277, 1999.

[10] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis—Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[11] J. D. Ferguson, “Variable duration models for speech,” in *Proc. Symp. Appl. Hidden Markov Models Text Speech*, Princeton, NJ, 1980, pp. 143–179.

[12] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. New York: Chapman & Hall/CRC, 1995.

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

[14] H. Gu, C. Tseng, and L. Lee, “Blind restoration of linearly degraded discrete signals by Gibbs sampling,” *IEEE Trans. Signal Processing*, vol. 43, pp. 2410–2413, Oct. 1995.

[15] M. E. A. Hodgson and P. J. Green, “Bayesian choice among Markov models of ion channels using Markov chain Monte Carlo,” *Proc. R. Soc. Lond. Ser. A—Math. Phys. Eng. Sci.*, vol. 455, pp. 3425–3448, 1999.

[16] T. Holter and T. Svendsen, “Maximum likelihood modeling of pronunciation variation,” *Speech Commun.*, vol. 29, pp. 177–191, 1999.

[17] L. Johnston and V. Krishnamurthy, “Hidden Markov model algorithms for narrowband interference suppression in CDMA spread spectrum systems,” *Signal Process.*, vol. 79, pp. 315–324, 1999.

[18] K. Karplus, C. Barrett, M. Cline, M. Diekhans, L. Grate, and R. Hughey, “Predicting protein structure using only sequence information,” *Proteins-Struct. Funct. Genetics*, pp. 121–125, Suppl. 3, 1999.

[19] M. Korenberg, J. E. Solomon, and M. E. Regelson, “Parallel cascade identification as a means for automatically classifying protein sequences into structure/function groups,” *Biol. Cybern.*, vol. 82, pp. 15–21, 2000.

[20] V. Krishnamurthy, S. Dey, and J. P. LeBlanc, “Blind equalization of IIR channels using hidden Markov models and extended least squares,” *IEEE Trans. Signal Processing*, vol. 43, pp. 2994–3006, Oct. 1995.

[21] J. Muller and H. Stahl, “Speech understanding and speech translation by maximum a-posteriori semantic decoding,” *Artif. Intell. Eng.*, vol. 13, no. 4, pp. 373–384, 1999.

[22] J. J. K. O Ruanaidh and W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*. New York: Springer-Verlag, 1996.

[23] W. D. Penny and S. J. Roberts, “Dynamic models for nonstationary signal segmentation,” *Comput. Biomed. Res.*, vol. 32, pp. 483–502, 1999.

[24] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech processing,” *Proc. IEEE*, vol. 77, pp. 257–286, 1989.

[25] Bayesian inference in hidden Markov models through jump Markov chain, C. P. Robert, T. Ryden, and D. M. Titterington. (1999). [Online]. Available: http://www.mcs.surrey.ac.uk/Personal/S.Brooks/MCMC/pages/listnz.html.

[26] C. P. Robert, G. Celeux, and J. Diebolt, “Bayesian estimation of hidden Markov chains: A stochastic implementation,” *Statist. Probab. Lett.*, vol. 16, pp. 77–83, 1993.

[27] P. R. Runkle, P. K. Bharadwaj, L. Couchman, and L. Carin, “Hidden Markov models for multiaspect target classification,” *IEEE Trans. Signal Processing*, vol. 47, pp. 2035–2040, July 1999.

[28] T. Rydén, T. Terasvirta, and S. Asbrink, “Stylized facts of daily return series and the hidden Markov model,” *J. Appl. Econometr.*, vol. 13, no. 3, pp. 217–244, 1998.

[29] B. Sin and J. H. Kim, “Nonstationary hidden Markov model,” *Signal Process.*, vol. 46, pp. 31–46, 1995.

[30] M. Stephens, “Dealing with label switching in mixture models,” *J. R. Statist. Soc.*, ser. B, pt. 4, vol. 62, pp. 795–809, 2000.

[31] S. V. Vaseghi, “State duration modeling in hidden Markov models,” *Signal Process.*, vol. 41, pp. 31–41, 1995.

[32] D. Q. Zhou, G. S. Liu, and J. X. Wang, “Spatio-temporal target identification method of high-range resolution radar,” *Pattern Recognit.*, vol. 33, no. 1, 2000.

Petar M. Djurić (SM’99) received the B.S. and M.S. degrees in electrical engineering from the University of Belgrade, Belgrade, Yugoslavia, in 1981 and 1986, respectively, and the Ph.D. degree in electrical engineering from the University of Rhode Island, Kingston, in 1990.

From 1981 to 1986, he was Research Associate with the Institute of Nuclear Sciences, Vinča, Belgrade. Since 1990, he has been with the State University of New York at Stony Brook, where he is Professor with the Department of Electrical and Computer Engineering. He works in the area of statistical signal processing, and his primary interests are in the theory of modeling, detection, estimation, and time series analysis and its application to a wide variety of disciplines, including telecommunications, biomedicine, and power engineering.

Prof. Djurić has served on numerous Technical Committees for the IEEE and SPIE and has been invited to lecture at universities in the United States and overseas. He was Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and currently, he is the Treasurer of the IEEE Signal Processing Conference Board. He is also Vice Chair of the IEEE Signal Processing Society Committee on Signal Processing—Theory and Methods, and a Member of the American Statistical Association and the International Society for Bayesian Analysis.

Joon-Hwa Chun was born in Jeon-Ju, Korea. He received the B.S. degree from SungKyunKwan University, Seoul, Korea, in 1990, the M.S. degree from the Syracuse University, Syracuse, NY, in 1994, and the Ph.D. degree from the State University of New York at Stony Brook, in 2000, all in electrical engineering.

From 1990 to 1991, he was with the LG Anyang Reasearch Laboratory, Anyang, Korea, where he was a Member of Technical Staff involved in TFT-LCD development. He is currently with the wireless application group of SandBridge Technologies, Inc., White Plains, NY, where he is a Senior Engineer, involved in system design of third-generation WCDMA FDD. His areas of interest include statistical signal processing with emphasis on spread spectrum systems, joint detection, and parameter estimation.