Hidden Markov Models with Stick Breaking Priors

John Paisley and Lawrence Carin
Department of Electrical and Computer Engineering
Duke University, Durham, NC 27708
{jwp4,lcarin}@ece.duke.edu

Abstract

The number of states in a hidden Markov model is an important parameter that has a critical impact on the inferred model. Bayesian approaches to addressing this issue include the nonparametric hierarchical Dirichlet process, which does not extend to a variational Bayesian solution. We present a fully conjugate, Bayesian approach to determining the number of states in a hidden Markov model, which does have a variational solution. The infinite-state hidden Markov model presented here utilizes a stick-breaking construction for each row of the state transition matrix, which allows for a sparse utilization of the same subset of observation parameters by all states. In addition to our variational solution, we discuss retrospective and collapsed Gibbs sampling methods for MCMC inference. We demonstrate our model on a music recommendation problem containing 2,250 pieces of music from the classical, jazz and rock genres.

I. Introduction

The hidden Markov model (HMM) [28] is an effective means for statistically representing sequential data, and has been used in a variety of applications, including the modeling of music and speech [4][5], target identification [10][29] and bioinformatics [15]. These models are often trained using the maximum-likelihood Baum-Welch algorithm [28][14], where the number of states is preset and a point estimate of the parameters is returned. However, this can lead to over- or under-fitting if the underlying state structure is not modeled correctly.

Bayesian approaches to automatically inferring the number of states in an HMM have been investigated in the MCMC setting using reversible jumps [9], as well as a nonparametric, infinite-state model that utilizes the hierarchical Dirichlet process (HDP) [31]. This latter method has proven effective [27], but a lack of conjugacy between the two levels of Dirichlet processes prohibits fast variational inference [6], making this approach computationally prohibitive when modeling very large data sets.

To address this issue, we propose the *stick-breaking hidden Markov model* (SB-HMM), a fully conjugate prior for an infinite-state HMM that does have a variational solution. In this model, each row of the transition matrix is given an infinite dimensional stick-breaking prior [18][19], which has the nice property of sparseness on the same subset of locations. Therefore, for an ordered set of states, as we have with an HMM, a stick-breaking prior on the transition probabilities encourages sparse usage of the same subset of states.

To review, the general definition of a stick-breaking construction of a probability mass function [19], $\boldsymbol{p} = (p_1, \ldots, p_{d+1})$, is as follows,

$$\begin{aligned} p_i &= V_i \prod_{j=1}^{i-1} (1 - V_j) \\ p_{d+1} &= 1 - \sum_{i=1}^{d} p_i \\ V_i &\sim Beta(\upsilon_i, \omega_i) \end{aligned} \tag{1}$$

for $i = 1, \ldots, d$ and a set of non-negative, real parameters $\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_d)$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)$. In this definition, d can be finite or infinite, with the finite case also known as the generalized Dirichlet distribution [12][34]. When infinite, different settings of $\boldsymbol{\upsilon}$ and $\boldsymbol{\omega}$ result in a variety of well-known priors reviewed in [19]. For the model presented here, we are interested in using the stick-breaking prior on the ordered set of states of an HMM. The stick-breaking process derives its name from an analogy to iteratively breaking the proportion V_i from the remainder of a unit-length stick, $\prod_{j=1}^{i-1}(1 - V_j)$. As with the standard Dirichlet distribution discussed below, this prior distribution is conjugate to the multinomial distribution parameterized by $\boldsymbol{p}$ and therefore can be used in variational inference [6].

Another prior on a discrete probability vector is the finite symmetric Dirichlet distribution [24][20], which we call the standard Dirichlet prior. This distribution is denoted as $Dir(\boldsymbol{p}; \frac{\alpha}{L}, \ldots, \frac{\alpha}{L})$, where α is a non-negative, real parameter and L is the dimensionality of the probability vector, $\boldsymbol{p}$. This distribution can be viewed as the standard prior for the weights of a mixture model (the relevance of the mixture model to the HMM is discussed in Section 3), as it is the more popular of the two conjugate priors to the multinomial distribution. However, for variational inference, the parameter α must be set due to conjugacy issues, which can impact the inferred model structure. In contrast, for the case where $\upsilon_i = 1$, the model in (1) allows for a gamma prior on ω_i, removing the need to select this value *a priori*.

This nonparametric uncovering of the underlying state structure will be shown to improve model inference, as demonstrated on a music recommendation problem, where we analyze 2,250 pieces of music from the classical, jazz and rock genres. We will show that our model has comparable or improved performance over other approaches, including the standard finite Dirichlet distribution prior, while using substantially fewer states. This performance is quantified via measures of the quality of the top ten recommendations for all songs using the log-likelihood ratio between models as a distance metric.

We motivate our desire to simplify the underlying state structure on practical grounds. The computing of likelihood ratios with hidden Markov models via the forward algorithm requires the integrating out of the underlying state transitions and is $O(n^2)$ in an n-state model. For use in an environment where many likelihood ratios are to be calculated, it is therefore undesirable to utilize hidden Markov models with superfluous state structure. The calculation of 2250^2 likelihoods in our music recommendation problem will show a substantial time-savings. Furthermore, when many HMMs are to be built in a short period of time, a fast inference algorithm, such as variational inference may be sought. For example, when analyzing a large and evolving database of digital music, it may be desirable to perform HMM design quickly on each piece. When computing the N^2 relationships between the pieces, computational efficiency manifested by a parsimonious HMM state representation is of interest.

The remainder of the paper is organized as follows: We review the hidden Markov model in Section 2. Section 3 provides a review of hidden Markov models with finite-Dirichlet and Dirichlet process priors. In Section 4 we review the stick-breaking process and its finite representation as a generalized Dirichlet distribution, followed by the presentation of the stick-breaking HMM formulation. Section 5 discusses two MCMC inference methods and variational inference for the SB-HMM. For MCMC in particular, we discuss retrospective and collapsed Gibbs samplers, which serve as benchmarks to validate the accuracy of the simpler and much faster variational inference algorithm. Finally, we demonstrate our model on synthesized and music data in Section 6, and conclude in Section 7.

II. The Hidden Markov Model

The hidden Markov model [28] is a generative statistical representation of sequential data, where an underlying state transition process operating as a Markov chain selects state-dependent distributions from which observations are drawn. That is, for a sequence of length T, a state sequence $\boldsymbol{S} = (s_1, s_2, \dots, s_T)$ is drawn from $p(s_t|s_{t-1}, \dots, s_1) = p(s_t|s_{t-1})$, as well as an observation sequence, $\boldsymbol{X} = (x_1, x_2, \dots, x_T)$, from some distribution $p(x_t|\theta_{s_t})$, where θ_{s_t} is the set of parameters for the distribution indexed by s_t. An initial-state distribution, $p(s_1)$, is also defined to begin the sequence.

For an HMM, the state transitions are discrete and therefore have multinomial distributions, as does the initial-state probability mass function (pmf). We assume the model is ergodic, or that any state can reach any other state in a finite number of steps. Therefore, states can be revisited in a single sequence and the observation distributions reused. For concreteness, we define the observation distributions to be discrete, multinomial of dimension M. Traditionally, the number of states associated with an HMM is initialized and fixed [28]; this is a nuisance parameter that can lead to over-fitting if set too high or under-fitting if set too low. However, to complete our initial definition we include this variable, D, with the understanding that it is not fixed, but to be inferred from the data.

With this, we define the parameters of a discrete HMM:

A = $D \times D$ matrix where entry a_{ij} represents the transition probability from state i to j

B = $D \times M$ matrix where entry b_{ij} represents the probability of observing j from state i

$\boldsymbol{\pi}$ = D dimensional probability mass function for the initial state.

The data generating process may be represented as

$$\begin{aligned} x_t &\sim Mult(b_{s_t,1}, b_{s_t,2}, \ldots, b_{s_t,M}) \\ s_t &\sim Mult(a_{s_{t-1},1}, a_{s_{t-1},2}, \ldots, a_{s_{t-1},D}), \quad t \geq 2 \\ s_1 &\sim Mult(\pi_1, \pi_2, \ldots, \pi_D) \end{aligned}$$

The observed and hidden data, $\boldsymbol{X}$ and $\boldsymbol{S}$ respectively, combine to form the "complete data," for which we may write the complete-data likelihood,

$$p(\boldsymbol{X}, \boldsymbol{S}|\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}) = \pi_{s_1} \prod_{t=1}^{T-1} a_{s_t,s_{t+1}} \prod_{t=1}^{T} b_{s_t,x_t}$$

with the data likelihood, $p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$, obtained by integrating over the states using the forward algorithm [28].

III. The Hidden Markov Model with Dirichlet Priors

It is useful to analyze the properties of prior distributions to assess what can be expected in the inference process. For hidden Markov models, the priors typically given to all multinomial parameters are standard Dirichlet distributions. Below, we review these common Dirichlet priors for finite and infinite-state hidden Markov models. We also note that the priors we are concerned with are for the state transition probabilities, or the rows of the **A** matrix. We assume separate M-dimensional standard Dirichlet priors for all rows of the **B** matrix throughout this paper (since the M-dimensional observation alphabet is assumed known, as is typical), though this may be generalized to other data-generating modalities.

A. *The Dirichlet Distribution*

The standard Dirichlet distribution is conjugate to the multinomial distribution and is written as

$$f(p_1, \ldots, p_D | \beta_1, \ldots, \beta_D) = \frac{\Gamma(\sum_i \beta_i)}{\prod_i \Gamma(\beta_i)} \prod_{i=1}^{D} p_i^{\beta_i - 1} \tag{2}$$

with the mean and variance of an element, p_i, being,

$$\mathbb{E}[p_i] = \frac{\beta_i}{\sum_i \beta_i}, \qquad \mathbb{V}[p_i] = \frac{\beta_i(\sum_i \beta_i - \beta_i)}{(\sum_i \beta_i)^2(\sum_i \beta_i + 1)} \tag{3}$$

This distribution can be viewed as a continuous density on the $D-1$ dimensional simplex in $\mathbb{R}^D$, since $0 \leq p_i \leq 1$ for all i and $\sum_{i=1}^{D} p_i = 1$. When $\sum_i \beta_i \ll D$, draws of $\boldsymbol{p}$ are sparse, meaning that most of the probability mass is located on a subset of $\boldsymbol{p}$. When seeking a sparse state representation, one might find this prior appealing and suggest that the dimensionality, D, be made large (much larger than the anticipated number of states), allowing the natural sparseness of the prior to uncover the correct state number. However, we discuss the issues with this approach in Section III-C.

B. *The Dirichlet Process*

The Dirichlet process [17] is the extension of the standard Dirichlet distribution to a measurable space and takes two inputs: a scalar, non-negative precision, α, and a base distribution, G_0, defined over the measurable space $(\Theta, \mathcal{B})$. For a partition of Θ into disjoint sets, $B = \{B_1, \ldots, B_D\}$, where $\bigcup_i B_i = \Theta$, the Dirichlet process is defined as

$$(G(B_1), \ldots, G(B_D)) \sim Dir(\alpha G_0(B_1), \ldots, \alpha G_0(B_D)) \tag{4}$$

Because G_0 is a continuous distribution, the Dirichlet process is infinite-dimensional, i.e. $D \to \infty$. A draw, G, from a Dirichlet process is expressed $G \sim DP(\alpha G_0)$ and can be written in the form

$$G = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$$

The location, θ_i, associated with a measure, p_i, is drawn from G_0. In the context of an HMM, each θ_i represents the parameters for the distribution associated with state i from which an observation is drawn. The key difference between the distribution of Section III-A and the Dirichlet process is the infinite extent of the Dirichlet process. Indeed, the finite-dimensional Dirichlet distribution is a good approximation to the Dirichlet process under certain parameterizations [20].

To link the notation of (2) with (4), we note that $\alpha = \sum_i \beta_i$. When these β_i are uniformly parameterized, choosing $\alpha < \infty$ means that $\beta_i \to 0$ for all i as $D \to \infty$. Therefore, the expectation as well as the variance of $p_i \to 0$. Ferguson [17] showed that this distribution remains well defined in this extreme circumstance. Later, Sethuraman [30] showed that we can still draw explicitly from this distribution.

1) Constructing a Draw from the Dirichlet Process: Sethuraman's constructive definition of a Dirichlet prior allows one to draw directly from the infinite-dimensional Dirichlet process and is one instance of a stick-breaking prior included in [19]. It is written as,

$$G = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}, \qquad p_i = V_i \prod_{j=1}^{i-1} (1 - V_j) \tag{5}$$

$$V_i \sim Beta(1, \alpha), \qquad \theta_i \sim G_0 \tag{6}$$

From Sethuraman's definition, the effect of α on a draw from the Dirichlet process is clear. When $\alpha \to 0$, the entire stick is broken and allocated to one component, resulting in a degenerate measure at a random component with location drawn from G_0. Conversely, as $\alpha \to \infty$, the breaks become infinitesimally small and G converges to the empirical distribution of the individual draws from G_0, reproducing G_0 itself.

When the space of interest for the DP prior is not the data itself, but rather a parameter, θ, for a distribution, $p(x|\theta)$, from which the data is drawn, a Dirichlet process mixture model results [2] with the following generative process,

$$x_i|\theta_i \sim p(x|\theta_i), \qquad \theta_i|G \sim G, \qquad G|\alpha G_0 \sim DP(\alpha G_0) \tag{7}$$

C. The Dirichlet Process Hidden Markov Model

Dirichlet process mixture modeling is relevant because hidden Markov models may be viewed as a special case of state-dependent mixture modeling where each mixture shares the same support, but has different mixing weights. These mixture models are linked in that the component selected at time t from which we draw observation x_t also indexes the mixture model to be used at time $t+1$. Using the notation of Section II, and defining $\theta_i \equiv (b_{i1}, \dots, b_{iM})$, we can write the state-dependent mixture model of a discrete HMM as

$$x_t|\theta_{s_t} \sim Mult(\theta_{s_t}), \qquad \theta_{s_t}|s_{t-1} \sim G_{s_{t-1}}, \qquad G_i = \sum_{j=1}^{D} a_{ij} \delta_{\theta_j} \tag{8}$$

where it is assumed that the initial state has been selected from the pmf, $\boldsymbol{\pi}$. From this formulation, we see that we are justified in modeling each state as an independent mixture model with the condition that respective states share the same observation parameters, θ_j. For unbounded state models, one might look to model each transition as a Dirichlet process. However, a significant problem arises when this is done. Specifically, in Sethuraman's representation of the Dirichlet process, each row, i, of the infinite state transition matrix would be drawn as follows,

$$G_i = \sum_{j=1}^{\infty} a_{ij}\delta_{\theta_{ij}}, \qquad a_{ij} = V_{ij}\prod_{k=1}^{j-1}(1 - V_{ik}), \qquad V_{ij} \sim Beta(1, \alpha), \qquad \theta_{ij} \sim G_0 \tag{9}$$

where a_{ij} is the j^{th} component of the infinite vector $\boldsymbol{a}_i$. However, with each θ_{ij} indexing a state, it becomes clear that since $p(\theta_m = \theta_n) = 0$ for $m \neq n$ when G_0 is continuous, the probability of a transition to a previously visited state is zero. Because G_0 is continuous, this approach for solving an infinite-state HMM is therefore impractical.

1) The Hierarchical Dirichlet Process Hidden Markov Model: To resolve this issue, though developed for more general applications than the HMM, Teh *et al.* [31] proposed the hierarchical Dirichlet process (HDP), in which the base distribution, G_0, over Θ is itself drawn from a Dirichlet process, making G_0 almost surely discrete. The formal notation is as follows,

$$G_m \sim DP(\beta G_0), \qquad G_0 \sim DP(\alpha H) \tag{10}$$

The HDP is a two-level approach, where the distribution on the points in Θ is shifted from the continuous H to the discrete, but infinite G_0. In this case, multiple draws for G_m will have substantial weight on the same set of atoms (i.e., states). We can show this by writing the top-level DP in stick-breaking form and truncating at K, allowing us to write the second-level DP explicitly,

$$G_0 = \sum_{i=1}^{K} p_i\delta_{\theta_i}, \qquad p_i = V_i\prod_{j=1}^{i-1}(1 - V_j), \qquad V_i \sim Beta(1, \alpha), \qquad \theta_i \sim H \tag{11}$$

$$(G_m(\theta_1), G_m(\theta_2), \ldots, G_m(\theta_K)) \sim Dir(\beta p_1, \beta p_2, \ldots, \beta p_K) \tag{12}$$

where $G(\theta_i)$ is a probability measure at location θ_i. Therefore, the HDP formulation effectively selects the number of states and their observation parameters via the top-level DP and uses the mixing weights as the prior for a second-level Dirichlet distribution from which the transition probabilities are drawn. The lack of conjugacy between these two levels means that a truly variational solution [6] does not exist.

IV. The Hidden Markov Model with Stick Breaking Priors

A more general class of priors for a discrete probability distribution is the stick-breaking prior [19]. As discussed above, these are priors of the form

$$\begin{aligned} p_i &= V_i \prod_{j=1}^{i-1}(1-V_j) \\ V_i &\sim Beta(\upsilon_i, \omega_i) \end{aligned} \tag{13}$$

for $i = 1, 2, \dots$ and the parameters $\boldsymbol{\upsilon} = (\upsilon_1, \upsilon_2, \dots)$ and $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots)$. This stick-breaking process is more general than that for drawing from the Dirichlet process, in which both the $Beta(1, \alpha)$-distributed random variables and the atoms associated with the resulting weights are drawn simultaneously. The representation of (13) is more flexible in its parametrization, and in that the weights have been effectively detached from the locations in this process. As written in (13), the process is infinite, however, when $\boldsymbol{\upsilon}$ and $\boldsymbol{\omega}$ terminate at some finite number, d, with $p_{d+1} \equiv 1 - \sum_{i=1}^{d} p_i$, the result is a draw from a generalized Dirichlet distribution (GDD) [12][34]. Since the necessary truncation of the variational model means that we are using a GDD prior for the state transitions, we briefly review this prior below.

For the GDD, the density function of $\boldsymbol{V} = (V_1, \dots, V_d)$ is

$$f(\boldsymbol{V}) = \prod_{i=1}^{d} f(V_i) = \prod_{i=1}^{d} \frac{\Gamma(\upsilon_i + \omega_i)}{\Gamma(\upsilon_i)\Gamma(\omega_i)} V_i^{\upsilon_i - 1}(1 - V_i)^{\omega_i - 1} \tag{14}$$

Following a change of variables from $\boldsymbol{V}$ to $\boldsymbol{p}$, the density of $\boldsymbol{p}$ is written as,

$$f(\boldsymbol{p}) = \prod_{i=1}^{d} \left(\frac{\Gamma(\upsilon_i + \omega_i)}{\Gamma(\upsilon_i)\Gamma(\omega_i)} p_i^{\upsilon_i - 1} \right) p_{d+1}^{\omega_d - 1} (1 - P_1)^{\omega_1 - (\upsilon_2 + \omega_2)} \times \cdots \times (1 - P_{d-1})^{\omega_{d-1} - (\upsilon_{d-1} + \omega_{d-1})} \tag{15}$$

which has a mean and variance for an element, p_i,

$$\mathbb{E}[p_i] = \frac{\upsilon_j \prod_{\ell=1}^{j-1} \omega_\ell}{\prod_{\ell=1}^{j} (\upsilon_\ell + \omega_\ell)}, \qquad \mathbb{V}[p_i] = \frac{\upsilon_j(\upsilon_j + 1) \prod_{\ell=1}^{j-1} \omega_\ell(\omega_\ell + 1)}{\prod_{\ell=1}^{j} (\upsilon_\ell + \omega_\ell)(\upsilon_\ell + \omega_\ell + 1)} \tag{16}$$

When $\omega_i = \sum_{j=i+1}^{d} \upsilon_j$ for $i < d$, and leaving $\omega_d = \omega_d$, the GDD reverts to the standard Dirichlet distribution of (2).

We will refer to a construction of $\boldsymbol{p}$ from the infinite process of (13) as $\boldsymbol{p} \sim SB(\boldsymbol{\upsilon}, \boldsymbol{\omega})$ and from (15) as $\boldsymbol{p} \sim GDD(\boldsymbol{\upsilon}, \boldsymbol{\omega})$. For a set of N integer-valued observations, $X_n \overset{iid}{\sim} Mult(\boldsymbol{p})$, the posterior of the respective priors are parameterized by $\boldsymbol{\upsilon}'$ and $\boldsymbol{\omega}'$, where $\upsilon_i' = \upsilon_i + \sum_{n=1}^{N} \mathbf{1}(X_n = i)$ and $\omega_i' = \sum_{j>i} \sum_{n=1}^{N} \mathbf{1}(X_n = j)$, where $\mathbf{1}(\cdot)$ is an indicator function that equals one when the argument is true and zero otherwise and is used to count the number of times the random variables are equal to values of interest. This posterior calculation will be used in deriving update equations for MCMC and variational inference.

A. The Stick Breaking Hidden Markov Model

For the stick-breaking hidden Markov model (SB-HMM), we model each row of the infinite state transition matrix, **A**, with a stick-breaking prior of the form of (13), with the state-dependent parameters drawn iid from a base measure, G_0.

$$\begin{aligned} G_i &= \sum_{j=1}^{\infty} a_{ij}\delta_{\theta_j} \\ \boldsymbol{a}_i &\overset{iid}{\sim} SB(\boldsymbol{\upsilon}, \boldsymbol{\omega}) \\ \theta_j &\overset{iid}{\sim} G_0 \end{aligned} \tag{17}$$

for $i = 1, 2, \dots$ and $j = 1, 2, \dots$ and where $\boldsymbol{\upsilon} = (\upsilon_{i1}, \upsilon_{i2}, \dots)$ and $\boldsymbol{\omega} = (\omega_{i1}, \omega_{i2}, \dots)$. The initial state pmf, $\boldsymbol{\pi}$, is also constructed according to an infinite stick-breaking construction. The key difference in this formulation is that the state transition pmf, $\boldsymbol{a}_i$, for state i has an infinite stick-breaking prior on the domain of the positive integers, which index the states. That is, the random variable, $V_{ij} \sim Beta(\upsilon_{ij}, \omega_{ij})$, is defined to correspond to the portion broken from the remainder of the unit length stick belonging to state i, which defines the transition probability from state i to state j. The corresponding state-dependent parameters, θ_j, are then drawn separately, effectively detaching the construction of **B** from the construction of **A**. This is in contrast with Dirichlet process priors, where these two matrices are linked in their construction, thus requiring an HDP solution as previously discussed.

We observe that this distribution requires an infinite parametrization. However, to simplify this we propose the following generative process,

$$\begin{aligned} G_i &= \sum_{j=1}^{\infty} a_{ij}\delta_{\theta_j} \\ a_{ij} &= V_{ij}\prod_{k=1}^{j-1}(1 - V_{ik}) \\ V_{ij} &\sim Beta(1, \alpha_{ij}) \\ \theta_j &\overset{iid}{\sim} G_0 \\ \alpha_{ij} &\overset{iid}{\sim} Gamma(c, d) \end{aligned} \tag{18}$$

where we have fixed $\upsilon_i = 1$ for all i and changed ω to α to emphasize the similar function of this variable in our model as that found in the finite Dirichlet model. Doing this, we can exploit the conjugacy of the gamma distribution to the $Beta(1, \alpha)$ distribution to provide further flexibility of the model. As in the Dirichlet process, the value of α has a significant impact in our model, as does the setting of the gamma hyperparameters. For example, the posterior of a single α_{ij} is

$$p(\alpha_{ij}|V_{ij}, c, d) = Gamma(c + 1, d - \ln(1 - V_{ij})) \tag{19}$$

Allowing $c, d \to 0$, the posterior expectation is $\mathbb{E}[\alpha_{ij}|V_{ij}] = -1/\ln(1 - V_{ij})$. We see that when V_{ij} is near one, a large portion of the remaining stick is broken, which drives the value of α_{ij} down, with the opposite occurring when V_{ij} is near zero. Therefore, one can think of the gamma prior on each α_{ij} as acting as a faucet that is opened and closed in the inference process.

Therefore, the value of the hyperparameter d is important to the model. If $d \to 0$, then it is possible for $\mathbb{E}[\alpha_{ij}|V_{ij}]$ to grow very large, effectively turning off a state transition, which may encourage sparseness in a particular transition pmf, $\boldsymbol{a}_i$, but not over the entire state structure of the model. With $c \to 0$, the value of d has the effect of ensuring that, under the conditional posterior of α_{ij},

$$\mathbb{E}[\alpha_{ij}|V_{ij}] < \frac{1}{d} \tag{20}$$

which will mitigate this effect. The stick-breaking hidden Markov model has therefore been designed to accommodate an infinite number of states, with the statistical property that only a subset will be used with high probability.

B. The Marginal Stick-Breaking Construction

To further analyze the properties of this prior, we discuss the process of integrating out the state transition probabilities by integrating out the random variables, $\boldsymbol{V}$, of the stick-breaking construction. This will be used in the collapsed Gibbs sampler discussed in the next section. To illustrate this idea with respect to the Dirichlet process, we recall that marginalizing G yields what is called the Chinese restaurant process (CRP) [1]. In sampling from this process, observation θ_{N+1} is drawn according to the distribution

$$\theta_{N+1}|\theta_N, \ldots, \theta_1 \sim \sum_{i=1}^{k} \frac{n_i}{\alpha + N}\delta_{\theta_i} + \frac{\alpha}{\alpha + N}G_0 \tag{21}$$

where n_i is the number of previous observations that used parameter θ_i. If the α component is selected, a new parameter is drawn from G_0. In the CRP analogy, n_i represents the number of customers sitting at the table θ_i. A new customer sits at this table with probability proportional to n_i, or a new table with probability proportional to α. The empirical distribution of (21) converges to G as $N \to \infty$. Below, we detail a similar process for drawing samples from a single, marginalized stick-breaking construction, with this naturally extended to the state transitions of the HMM.

In (13), we discussed the process for constructing $\boldsymbol{p}$ from the infinite stick-breaking construction using beta-distributed random variables, $\boldsymbol{V} = \{V_1, V_2, \dots\}$. The process of sampling from $\boldsymbol{p}$ is also related to $\boldsymbol{V}$ in the following way: Instead of drawing integer-valued samples, $C \sim \boldsymbol{p}$, one can alternatively sample from $\boldsymbol{V}$ by drawing random variables $Z_j \in \{0, 1\}$, from $Bernoulli(V_j)$ with $p(Z_j = 1) = V_j$ and setting $C = \arg\min_j Z_j = 1$. That is, by sampling sequentially from $V_1, V_2, \dots$, and terminating when a one is drawn, setting C to that respective index. Figure 1(a) contains a visualization of this process using a tree structure, where we sample paths until terminating at a leaf node.

We now consider sampling from the marginalized version of this distribution by sequentially sampling from a set of two-dimensional Pólya urn models. As with the urn model for the DP, we can sample observation Z_{N+1} from a marginalized beta distribution having prior hyperparameters υ and ω and conditioned on observations $Z_1, \dots, Z_N$ as follows,

$$Z_{N+1}|Z_N, \dots, Z_1 \sim \frac{\upsilon + n_1}{\upsilon + \omega + N}\delta_1 + \frac{\omega + n_0}{\upsilon + \omega + N}\delta_0 \tag{22}$$

where n_k is the number of previous samples taking value k, or $n_k = \sum_{n=1}^{N} \mathbf{1}(Z_n = k)$ and $n_1 + n_0 = N$. The function δ_1 indicates that $Z_{N+1} = 1$ with probability $(\upsilon + n_1)/(\upsilon + \omega + N)$ and the corresponding probability for δ_0. This equation arises by performing the following integration of the random variable, $\mu \sim Beta(\upsilon, \omega)$,

$$P(Z_{N+1}|Z_N, \dots, Z_1, \upsilon, \omega) = \int_0^1 P(Z_{N+1}|\mu)P(\mu|Z_N, \dots, Z_1, \upsilon, \omega)\ d\mu \tag{23}$$

where $P(Z_{N+1}|\mu) = \mu$ if $Z_{N+1} = 1$ and $P(Z_{N+1}|\mu) = 1 - \mu$ if $Z_{N+1} = 0$.

As with the CRP, this converges to a sample from the beta distribution as $N \to \infty$. Figure 1(b) contains the extension of this to the stick-breaking process. As shown, each point of intersection of the tree follows a local marginalized beta distribution, with paths along this tree selected according to these distributions. The final observation is the leaf node at which this process terminates. In keeping with other restaurant analogies, we tell the following intuitive story to accompany this picture.

Person N walks down a street containing many Chinese restaurants. As he approaches each restaurant, he looks inside to see how many people are present. He also hears the noise coming from other restaurants down the road and is able to perfectly estimate the total number of people who have chosen to eat at one of the restaurants he has not yet seen. He chooses to enter restaurant i with probability $(\upsilon_i + n_i)/(\upsilon_i + \omega_i + \sum_{j=i}^{\infty} n_j)$ and to continue walking down the road with probability $(\omega_i + \sum_{j=i+1}^{\infty} n_j)/(\upsilon_i + \omega_i + \sum_{j=i}^{\infty} n_j)$, which are probabilities nearly proportional to the number of people in restaurant i and in restaurants still down the road, respectively. If he reaches the end of populated restaurants, he continues to make decisions according to the prior shared by all people. This analogy can perhaps be called the "Chinese restaurant district."

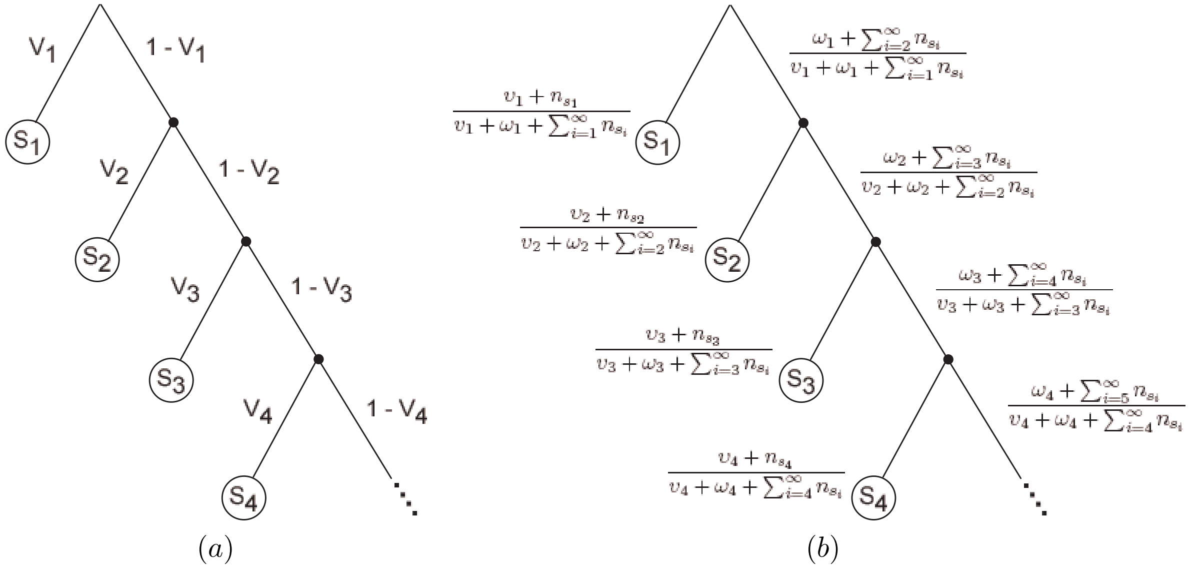

V_1
$1 - V_1$
S_1
V_2
$1 - V_2$
S_2
V_3
$1 - V_3$
S_3
V_4
$1 - V_4$
S_4
(a)
$\frac{\omega_1 + \sum_{i=2}^{\infty} n_{s_i}}{\upsilon_1 + \omega_1 + \sum_{i=1}^{\infty} n_{s_i}}$
$\frac{\upsilon_1 + n_{s_1}}{\upsilon_1 + \omega_1 + \sum_{i=1}^{\infty} n_{s_i}}$
S_1
$\frac{\omega_2 + \sum_{i=3}^{\infty} n_{s_i}}{\upsilon_2 + \omega_2 + \sum_{i=2}^{\infty} n_{s_i}}$
$\frac{\upsilon_2 + n_{s_2}}{\upsilon_2 + \omega_2 + \sum_{i=2}^{\infty} n_{s_i}}$
S_2
$\frac{\omega_3 + \sum_{i=4}^{\infty} n_{s_i}}{\upsilon_3 + \omega_3 + \sum_{i=3}^{\infty} n_{s_i}}$
$\frac{\upsilon_3 + n_{s_3}}{\upsilon_3 + \omega_3 + \sum_{i=3}^{\infty} n_{s_i}}$
S_3
$\frac{\omega_4 + \sum_{i=5}^{\infty} n_{s_i}}{\upsilon_4 + \omega_4 + \sum_{i=4}^{\infty} n_{s_i}}$
$\frac{\upsilon_4 + n_{s_4}}{\upsilon_4 + \omega_4 + \sum_{i=4}^{\infty} n_{s_i}}$
S_4
(b)

Fig. 1. The tree structure for drawing samples from (a) an infinite stick-breaking construction and (b) the marginalized stick-breaking construction. The parameters for (a) are drawn exactly and fixed. The parameters of (b) evolve with the data and converge as $\sum_{i=1}^{\infty} n_{s_i} \rightarrow \infty$. Beginning at the top of the tree, paths are sequentially chosen according to a biased coin flip, terminating at a leaf node.

To extend this idea to the HMM, we note that each leaf node represents a state, and therefore sampling a state at time t according to this process also indexes the parameters that are to be used for sampling the subsequent state. Therefore, each of the trees in Figure 1 is indexed by a state, with its own count statistics. This illustration provides a good indication of the natural sparseness of the stick-breaking prior, as these binary sequences are more likely to terminate on a lower indexed state, with the probability of a leaf node decreasing as the state index increases.

V. Inference for the Stick Breaking Hidden Markov Model

In this section, we discuss Markov chain Monte Carlo (MCMC) and variational Bayes (VB) inference methods for the SB-HMM. For Gibbs sampling MCMC, we derive a retrospective sampler that is similar to [25] in that it allows the state number to grow and shrink as needed. We also derive a collapsed inference method [22], where we marginalize over the state transition probabilities. This process is similar to other urn models [21], including that for the Dirichlet process [11].

A. MCMC Inference for the SB-HMM: A Retrospective Gibbs Sampler

We define D to be the total number of used states following an iteration. Our sampling methods require that only the D utilized states be monitored for any given iteration, along with state $D+1$ drawn from the base distribution, a method which accelerates inference.

The full posterior of our model can be written as $p(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}, \{\boldsymbol{S}\}_1^N, \boldsymbol{\alpha}, \alpha_\pi | \boldsymbol{X})$, where $\{\boldsymbol{S}\}_1^N$ represents N sequences of length T. Following Bayes' rule, we sample each parameter from its full conditional posterior as follows:

(1) Sample α_{ij} from its respective gamma-distributed posterior

$$\alpha_{ij} \sim Ga\left(c+1, d - \ln(1 - V_{ij})\right) \tag{24}$$

Draw new values for α_{iD} and the new row of $\mathbf{A}$ from the $Gamma(c, d)$ prior.

(2) Construct the state transition probabilities of $\mathbf{A}$ from their stick-breaking conditional posteriors:

$$V_{ij} \sim Beta\left(1 + \sum_{n=1}^{N} u_{ij}^{A}(n), \alpha_{ij} + \sum_{n=1}^{N} \sum_{k=j+1}^{D} u_{ik}^{A}(n)\right), \qquad a_{i1} = V_{i1}, \qquad a_{ij} = V_{ij} \prod_{k=1}^{j-1}(1 - V_{ik}) \tag{25}$$

where for sequence n, $u_{ij}^{A}(n) \equiv \sum_{t=1}^{T-1} \mathbf{1}(s_t = i, s_{t+1} = j)$, the count of transitions from state i to state j. Set $a_{i,D+1} = 1 - \sum_{j=1}^{D} a_{ij}$ and draw the $D+1$ dimensional probability vector from the GDD prior.

(3) Sample the innovation parameters, α_π^j, for $\boldsymbol{\pi}$ from their gamma-distributed posteriors

$$\alpha_\pi^j \sim Ga\left(\tau_{\pi 1} + 1, \tau_{\pi 2} - \ln(1 - V_j^\pi)\right) \tag{26}$$

(4) Sample $\boldsymbol{\pi}$ from its stick-breaking conditional posterior:

$$V_j^\pi \sim Beta\left(1 + u_j^\pi, \alpha_\pi^j + \sum_{n=1}^{N} \mathbf{1}(s_1^n > j)\right), \qquad \pi_1 = V_1^\pi, \qquad \pi_j = V_j^\pi \prod_{k=1}^{j-1}(1 - V_k^\pi) \tag{27}$$

for $j = 1, 2, \ldots, D$ with $u_j^\pi \equiv \sum_{n=1}^{N} \mathbf{1}(s_1^n = j)$, the number of times a sequence begins in state j. Set

$\pi_{D+1} = 1 - \sum_{j=1}^{D} \pi_j$.

(5) Sample each row, $i = 1, 2, \ldots, D$ of the observation matrix from its Dirichlet conditional posterior:

$$(b_{i1}, \ldots, b_{iM}) \sim Dir\left(\beta_1 + \sum_{n=1}^{N} u_{i1}^{B}(n), \ldots, \beta_M + \sum_{n=1}^{N} u_{iM}^{B}(n)\right) \tag{28}$$

where for sequence n, $u_{ij}^{B}(n) \equiv \sum_{t=1}^{T} \mathbf{1}(s_t = i, x_t = m)$, the number of times $x = m$ is observed while in state i. Draw a new atom from the base distribution $(b_{D+1,1}, \ldots, b_{D+1,M}) \sim Dir(\beta_1, \ldots, \beta_M)$.

(6) Sample each new state sequence from the conditional posterior, defined for a given sequence as:

$$\begin{aligned} p(s_1|x_1, s_2, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}) &\propto p(s_1)p(x_1|s_1)p(s_2|s_1) \\ p(s_t|x_t, s_{t-1}, s_{t+1}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}) &\propto p(s_t|s_{t-1})p(x_t|s_t)p(s_{t+1}|s_t) \quad 2 \leq t \leq T-1 \\ p(s_T|x_T, s_{t-1}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}) &\propto p(x_T|s_T)p(s_T|s_{T-1}) \end{aligned} \tag{29}$$

(7) Set D equal to the number of unique states drawn in (6) and prune away any unused states.

This process iterates until convergence [3], at which point each uncorrelated sample of (1-6) can be considered a sample from the full posterior. In practice, we find that it is best to initialize with a large number of states and let the algorithm prune away, rather than allow for the state number to grow from a small number, a process that can be very time consuming. We also mention that, as the transition weights of the model are ordered in expectation, a label switching procedure may be employed [26][25], where the indices are reordered based on the state membership counts calculated from u^A such that these counts are decreasing with an increasing index number.

B. MCMC Inference for the SB-HMM: A Collapsed Gibbs Sampler

It is sometimes useful to integrate out parameters in a model, thus collapsing the structure and reducing the amount of randomness in the model [22]. This process can lead to faster convergence to the stationary posterior distribution. In this section, we discuss integrating out the infinite state transition probabilities, $\boldsymbol{a}_i$, which removes the randomness in the construction of the A matrix. These equations are a result of the discussion in Section IV-B. The following modifications can be made to the MCMC sampling method of Section V-A to perform collapsed Gibbs sampling:

(1) To sample α_{ij}, sample V_{ij} as in Step (2) of Section V-A using the previous values for α_{ij}. This is only for the purpose of resampling α_{ij} as in Step (1) of Section V-A. See [16] for a similar process for marginalized DP mixtures.

(2) The transition probability, a_{ij}, is now constructed using the marginalized beta probabilities

$$a_{ij} = \frac{1 + n_{ij}}{1 + \alpha_{ij} + \sum_{k=j}^{\infty} n_{ik}} \prod_{\ell=1}^{j-1} \frac{\alpha_{i\ell} + \sum_{m=\ell+1}^{\infty} n_{im}}{1 + \alpha_{i\ell} + \sum_{m=\ell}^{\infty} n_{im}}$$

A similar process must be undertaken for the construction of $\boldsymbol{\pi}$. Furthermore, we mention that the state sequences can also be integrated out using the forward-backward algorithm for additional collapsing of the model structure.

C. Variational Bayes Inference for the SB-HMM

Variational Bayesian inference [6][32] is motivated by the equality

$$\int_{\theta} Q(\theta) \ln \frac{Q(\theta)}{P(\theta|X)P(X)} d\theta = \int_{\theta} Q(\theta) \ln \frac{Q(\theta)}{P(X|\theta)P(\theta)} d\theta \tag{30}$$

which can be rewritten as

$$\ln P(X) = \mathcal{L}(Q) + KL(Q||P) \tag{31}$$

where θ represents the model parameters and hidden data, X the observed data, $Q(\theta)$ an approximating density to be determined and

$$\mathcal{L}(Q) = \int_{\theta} Q(\theta) \ln \frac{P(X|\theta)P(\theta)}{Q(\theta)} d\theta, \qquad KL(Q||P) = \int_{\theta} Q(\theta) \ln \frac{Q(\theta)}{P(\theta|X)} d\theta \tag{32}$$

The goal is to best approximate the true posterior $p(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}, \{\boldsymbol{S}\}_1^N, \boldsymbol{\alpha}, \alpha_\pi | \boldsymbol{X})$ by minimizing $KL(Q||P)$. Due to the fact that $KL(Q||P) \geq 0$, this can be done by maximizing $\mathcal{L}(Q)$. This requires a factorization of the Q distributions, or

$$Q\left(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}, \{\boldsymbol{S}\}_1^N, \boldsymbol{\alpha}, \alpha_\pi\right) = Q(\boldsymbol{A})Q(\boldsymbol{B})Q(\boldsymbol{\pi})Q(\{\boldsymbol{S}\}_1^N)Q(\boldsymbol{\alpha})Q(\alpha_\pi) \tag{33}$$

A general method for performing variational inference for conjugate-exponential Bayesian networks outlined in [32] is as follows: For a given node in a graph, write out the posterior as though everything were known, take the natural logarithm, the expectation with respect to all unknown parameters and exponentiate the result. Since it requires the computational resources comparable to the expectation-maximization algorithm, variational inference is fast. However, the deterministic nature of the algorithm requires that we truncate to a fixed state number, K. As will be seen in the posterior, however, only a subset of these states will contain substantial weight. We call this truncated version the truncated stick-breaking HMM (TSB-HMM).

1) VB-E Step: For the VB-E step, we calculate the variational expectation with respect to all unknown parameters. For a given sequence, we can write

$$Q(\boldsymbol{S}) \propto \tilde{\pi}_{s_1} \prod_{t=1}^{T-1} \tilde{a}_{s_t s_{t+1}} \prod_{t=1}^{T} \tilde{b}_{s_t,x_t} \tag{34}$$

To aid in the cleanliness of the notation, we first provide the general variational equations for drawing from the generalized Dirichlet distribution, which we recall is the distribution that results following the necessary truncation of the variational model. Consider a truncation to K-dimensions and let m_i be the expected number of observations from component i for a given iteration. The variational equations can be written as

$$\begin{aligned}
\langle \ln V_i \rangle &= \psi(1+m_i) - \psi\left(1+\alpha_i+\sum_{j=i}^{K} m_j\right) \\
\langle \ln(1-V_i) \rangle &= \psi\left(\alpha_i+\sum_{j=i+1}^{K} m_j\right) - \psi\left(1+\alpha_i+\sum_{j=i}^{K} m_j\right)
\end{aligned}$$

$$\begin{aligned}
\langle \ln p_1 \rangle &= \langle \ln V_i \rangle \\
\langle \ln p_k \rangle &= \langle \ln V_k \rangle + \sum_{j=1}^{k-1} \langle \ln(1-V_j) \rangle \qquad 2 \leq k < K \\
\langle \ln p_K \rangle &= \sum_{j=1}^{K-1} \langle \ln(1-V_j) \rangle
\end{aligned} \tag{35}$$

Where $\psi(\cdot)$ represents the digamma function. See [8] for further discussion. We use the above steps with the appropriate count values, $\langle \pi_i \rangle$ or $\langle n_{ij} \rangle$, inserted for m_i to calculate the variational expectations for $\tilde{\pi}$ and $\tilde{\boldsymbol{a}}_i$

$$\tilde{\pi} = \exp\left[\langle \ln \pi \rangle\right], \qquad \tilde{\boldsymbol{a}}_i = \exp\left[\langle \ln \boldsymbol{a}_i \rangle\right]$$

Where $\tilde{\boldsymbol{a}}_i$ is the i^{th} row of the transition matrix. This requires use of the expectation $\langle \alpha_{ij} \rangle = \frac{c'_{ij}}{d'_{ij}}$, where c'_{ij} and d'_{ij} are the posterior parameters of α_{ij}. The variational equation for $\tilde{b}_{ij}$ is

$$\tilde{b}_{ij} = \exp\left[\psi(\beta_j + \langle o_{ij} \rangle) - \psi\left(\sum_{j=1}^{M} \beta_j + \langle o_{ij} \rangle\right)\right] \tag{36}$$

The values for $\langle \pi_i \rangle$, $\langle n_{ij} \rangle$ and $\langle o_{ij} \rangle$ are outputs of the forward-backward algorithm from the previous iteration. Given these values for $\tilde{\pi}$, $\tilde{a}_i$ and $\tilde{b}_{ij}$, the forward-backward algorithm is employed as usual.

2) VB-M Step: Updating of the variational posteriors in the VB-M step is a simple updating of the sufficient statistics obtained from the VB-E step. They are as follows

$$\begin{aligned} Q(A) &= \prod_{i=1}^{K} GDD(\boldsymbol{\upsilon}'_i, \boldsymbol{\omega}'_i) \\ Q(B) &= \prod_{i=1}^{K} Dir(\beta_1 + \langle o_{i1} \rangle, \ldots, \beta_M + \langle o_{iM} \rangle) \\ Q(\boldsymbol{\pi}) &= GDD\left(\boldsymbol{\upsilon}'_\pi, \boldsymbol{\omega}'_\pi\right) \\ Q(\boldsymbol{\alpha}) &= \prod_{i=1}^{K} \prod_{j=1}^{K-1} Ga\left(c+1, d - \langle \ln(1 - V_{ij}) \rangle\right) \\ Q(\alpha_\pi) &= \prod_{i=1}^{K-1} Ga\left(\tau_{\pi 1} + 1, \tau_{\pi 2} - \langle \ln(1 - V_{\pi i}) \rangle\right) \end{aligned} \tag{37}$$

where $\boldsymbol{\upsilon}'$ and $\boldsymbol{\omega}'$ are the respective posterior parameters calculated using the appropriate $\langle n \rangle$ and $\langle \alpha \rangle$ values. This process iterates until convergence to a local optimal solution.

VI. Experimental Results

We demonstrate our model on both synthetic and digital music data. For synthetic data, this is done on a simple HMM using MCMC and variational Bayes inference. We then provide a comparison of our inference methods on a small-scale music problem, which is done to help motivate our choice of variational Bayes inference for a large-scale music recommendation problem, where we analyze 2,250 pieces of music from the classical, jazz and rock genres. For this problem, our model will demonstrate comparable or better performance than several algorithms while using substantially fewer states than the finite Dirichlet HMM approach.

A. MCMC Results on Synthesized Data

We synthesize data from the following HMM to demonstrate the effectiveness of our model in uncovering the underlying state structure:

$$\mathbf{A} = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \quad \boldsymbol{\pi} = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \tag{38}$$

From this model were generated $N = 100$ sequences of length $T = 15$. We placed $Gamma(10^{-6}, 0.1)$ priors on each α_{ij} and $Dir(2/3, 2/3, 2/3)$ priors for the observation matrix. We select $c = 10^{-6}$ as an arbitrarily small number. The setting of $d = 0.1$ is a value that requires some care, and we were motivated by setting a bound of $\alpha_{ij} < 10$ as discussed in Section IV-A. We ran 10,000 iterations using the Gibbs sampling methods outlined in Sections V-A and V-B and plot the results in Figures 2 and 3. Because we are not sampling from this chain, but are only interested in the inferred state number, we do not distinguish between burn-in and collection iterations.

It is observed that the state value does not converge exactly to the true number, but continually tries to innovate around the true state number. To give a sense of the underlying data structure, a threshold is set, and the minimum number of states containing at least 99% of the data is plotted in red. This value is calculated for each iteration by counting the state memberships from the sampled sequences, sorting in decreasing order and finding the resulting distribution. This is intended to emphasize that almost all of the data clusters into the correct number of states, whereas if only one of the 1,500 observations selects to innovate, this is included in the total state number of the blue line. As can be seen, the red line converges more tightly to the correct state number. In Figure 3, we see that collapsing the model structure shows even tighter convergence with less innovation [22].

We mention that the HDP solution also identifies the correct number of states for this problem, but as we are more interested in the variational Bayes solution, we do not present these results here. Our main interest here is in validating our model using MCMC inference.

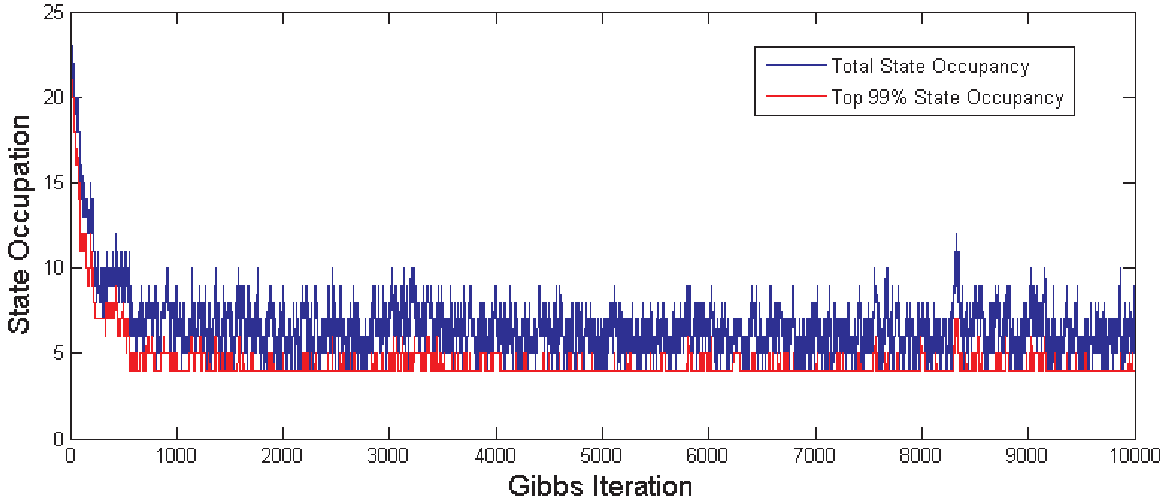


Fig. 2. Synthesized Data - State occupancy as a function of Gibbs sample number. A threshold set to 99% indicates that most of the data resides in the true number of states, while a subset attempts to innovate.

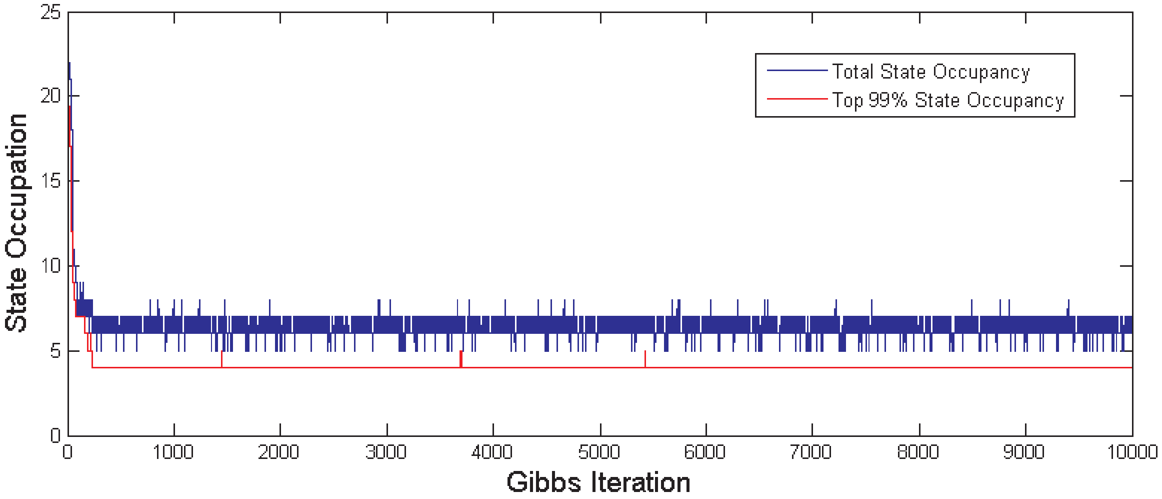


Fig. 3. Synthesized Data - State occupancy as function of sample for collapsed Gibbs sampling. The majority of the data again resides in the true state number, with less innovation.

B. Variational Results on Synthesized Data

Using the same sequences and parametrization as above, we then built 500 TSB-HMM models, truncating to 30 states and defining convergence to be when the fractional change in the lower bound falls below 10^{-5}. In Figure 4, we show the number of states as a function of iteration number averaged over 500 runs for the TSB-HMM compared with the standard Dirichlet HMM models with $\alpha = 1$.

For this toy problem, both models converge to the same state number, but the TSB-HMM converges noticeably faster. We also observe that the variational solution does not converge exactly to the true state number on average, but slightly overestimates the number of states. This may be due to the local optimal nature of the solution. In the large-scale example, we will see that these two models do not converge to the same state numbers; future research is required to discover why this may be.

To emphasize the impact that the inferred state structure can have on the quality of the model, we compare our results with the EM-HMM [7]. For each state initialization shown in Figure 5, we built 100 models using the two approaches and compared the quality of the resulting models by calculating the negative log-likelihood ratio between the inferred model and the ground truth model using a data sequence of length 5,000 generated from the true model. The more similar two HMMs are statistically, the smaller this value should be. As Figure 5 shows, for the EM algorithm this “distance” increases with an increasingly incorrect state initialization, indicating poorer quality models due to overfitting. Our model is shown to be relatively invariant to this initialization.

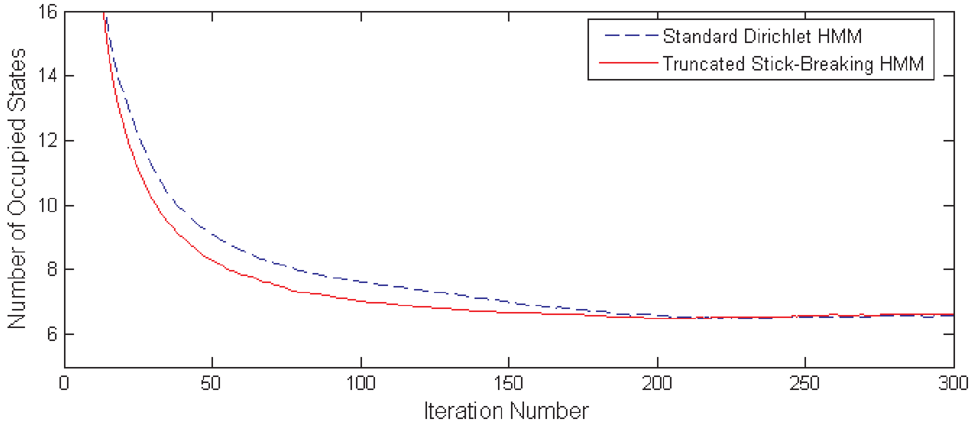


Fig. 4. The number of occupied states as a function of iteration number averaged over 500 runs. As can be seen, the truncated stick-breaking HMM converges more quickly to the simplified model.

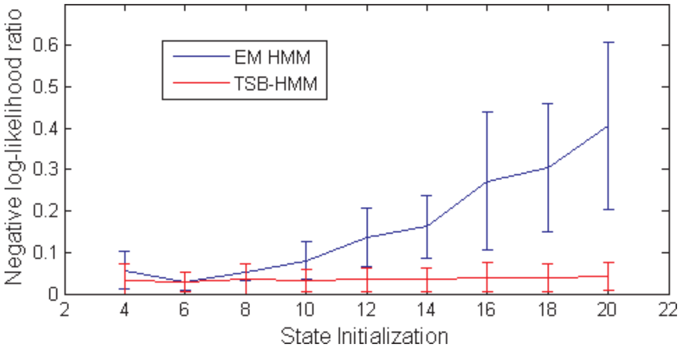


Fig. 5. The negative log-likelihood ratio between the inferred HMM and the true HMM displayed as a per-observation average. The TSB-HMM is shown to be invariant to state initialization, while the EM-HMM experiences a degradation in quality as the number of states deviates from the true number.

C. A Comparison of MCMC and Variational Inference Methods on Music Data

To assess the relative quality of our MCMC and variational Bayes approaches, we consider a small-scale music recommendation problem. We select eight pieces each from the classical, jazz and rock genres for a total of 24 pieces (see Table 1) that are intended to cluster by genre. Furthermore, the first and last four pieces within each genre are also selected to cluster together, though not as distinctly.

From each piece of music, we first extracted 20-dimensional MFCC features [13], ten per second. Using these features, we constructed a global codebook of size $M = 50$ using k-means clustering with which we quantized each piece of music. For each inference method, we placed $Gamma(10^{-6}, 0.1)$ priors on each α_{ij} and adaptively set the prior on the observation statistics to be $Dir(g_0^i + 10^{-6})$, where g_0^i is the empirical pmf of observations for the i^{th} piece of music. For the retrospective and collapsed

Classical	Piece	Jazz	Piece	Rock	Piece
Joseph Haydn	String Quartet 64-5 I	John Coltrane	Giant Steps	Jimi Hendrix	Purple Haze
Joseph Haydn	String Quartet 74-3 I	John Coltrane	Cousin Mary	Jimi Hendrix	Love or Confusion
Joseph Haydn	String Quartet 76-2 I	John Coltrane	Spiral	Jimi Hendrix	Foxy Lady
Joseph Haydn	String Quartet 77-1 I	John Coltrane	Mr. P.C.	Jimi Hendrix	Can You See Me
Joseph Haydn	Symphony 101 Mvt III	Miles Davis	Walkin’	Neil Young	Tell Me Why
Joseph Haydn	Symphony 102 Mvt III	Miles Davis	Blue ’n’ Boogie	Neil Young	Only Love Can Break Your Heart
Joseph Haydn	Symphony 103 Mvt III	Miles Davis	Seven Steps to Heaven	Neil Young	Harvest Moon
Joseph Haydn	Symphony 104 Mvt III	Miles Davis	Well, You Needn’t	Neil Young	Heart of Gold

TABLE I

A list of music pieces used by genre. The data is intended to cluster by genre, as well as by subgenre groups of size four.

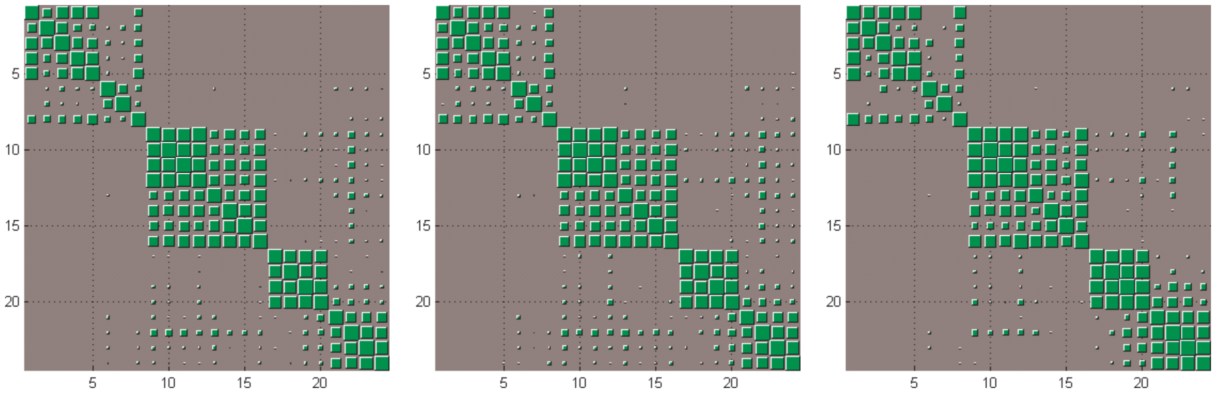
MCMC methods, we used 7,000 burn-in iterations to ensure proper convergence [3] and 3,000 collection iterations, sampling every 300 iterations. Also, ten TSB-HMM models are built for each piece of music using the same convergence measure as in the previous section.

To assess quality, we consider the problem of music recommendation where for a particular piece of music, other pieces are recommended based on the sorting of a distance metric. In this case, we use a distance based on the log-likelihood ratio between two HMMs, which is obtained by using data generated from each model as follows

$$D(HMM_1, HMM_2) = -\frac{1}{2}\ln\frac{p(x_{HMM_1}|HMM_2)}{p(x_{HMM_1}|HMM_1)} - \frac{1}{2}\ln\frac{p(x_{HMM_2}|HMM_1)}{p(x_{HMM_2}|HMM_2)} \tag{39}$$

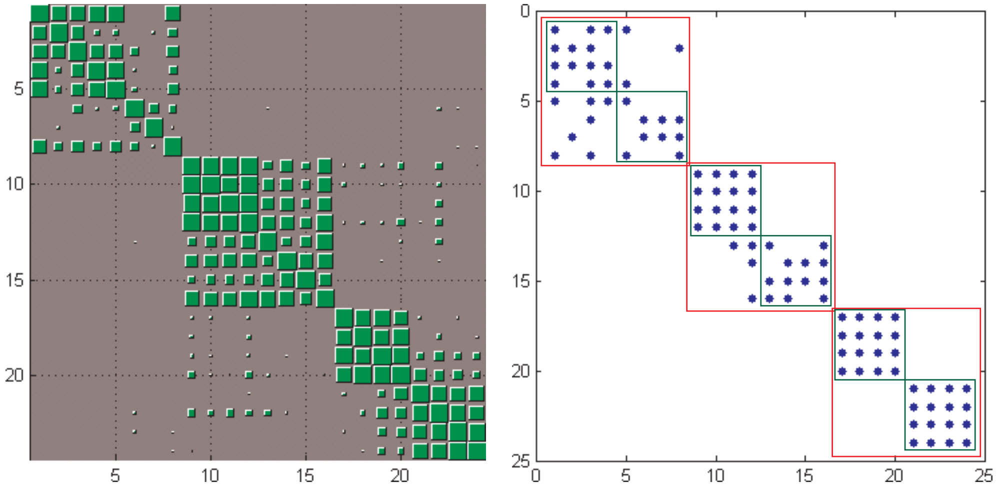
where x_{HMM} is a set of sequences generated from the indicated HMM; in this paper, we use the original signal to calculate this distance and the expectation of the posterior to represent the models. For this small problem, the likelihood was averaged over the multiple HMMs before calculating the ratio.

Figure 6 displays the kernel maps of these distances, where we use the radial basis function with a kernel width set to the 10% quantile of all distances to represent proximity, meaning larger boxes indicate greater similarity. We observe that the performance is consistent for all inference methods and uniformly good. The genres cluster clearly and the subgenres cluster as a whole, with greater similarity between other pieces in the same genre. In Figure 7, we show the kernel map for one variational Bayes run, indicating consistency with the average. We also show a plot of the top four recommendations for each piece of music for this variational Bayes run showing with more clarity the clustering by subgenre. From these results, we believe that we can be confident in the results provided in the next section, where we only consider variational Bayes inference and build only one HMM per piece of music. We also mention that this same experiment was performed with the HDP-iHMM of Section III-C1 and produced similar results using MCMC sampling methods. However, as previously discussed, we cannot consider this method for fast variational inference.



(a) (b) (c)

Fig. 6. Kernel maps for (a) retrospective MCMC (b) collapsed MCMC (c) variational Bayes (averaged over 10 runs). The larger the box, the larger the similarity between any two pieces of music. The variational Bayes results are very consistent with the MCMC results.



(a) (b)

Fig. 7. (a) Kernel map for one typical variational Bayes run, indicating good consistency with the average variational run. (b) A graphic of the four closest pieces of music. The performance indicates an ability to group by genre as well as subgenre.

D. Experimental Results on a 2,250 Piece Music Database

For our large-scale analysis, we used a personal database of 2,250 pieces of music, 750 each in the classical, jazz and rock genres. Using two minutes selected from each piece of music, we extracted ten, 20 dimensional MFCC feature vectors per second and quantized using a global codebook of size 100, again constructing this codebook using the k-means algorithm on a sampled subset of the feature vectors. We built a TSB-HMM on each quantized sequence with a truncation level of 50 and using the prior settings of the previous section. For inference, we devised the following method to significantly accelerate convergence: Following each iteration, we check the expected number of observations from each state and prune those states that have smaller than one expected observation. Doing so, we were able to significantly reduce inference times for both the TSB-HMMs and the finite Dirichlet HMMs.

As previously mentioned, the parameter α has a significant impact on inference for the finite Dirichlet HMM (here also initialized to 50 states). This is seen in the box plots of Figure 8, where the number of utilized states increases with an increasing α. We also observe in the histogram of Figure 9 that the state usage of our TSB-HMM reduces to levels unreachable in the finite Dirichlet model. Therefore, provided that the reduced complexity of our inferred models does not degrade the quality of the HMMs, our model is able to provide a more compact representation of the sequential properties of the data.

This compactness can be important in the following way: As noted in the introduction, the calculation of distances using the approach of the previous section is $O(n^2)$, where n is the number of states. Therefore, for very large databases of music, superfluous states can waste significant processing power when recommending music in this way. For example, the distance calculations for our problem took 14 hours, 22 minutes using the finite Dirichlet HMM with $\alpha = 1$ and 13 hours for our TSB-HMM; much of this time was due to the same computational overhead. We believe that this time savings can be a significant benefit for very large scale problems, including those beyond music modeling.

Below, we compare the performance of our model with several other methods. We compare with the traditional VB-HMM [24] where $\alpha = 1$ and $\alpha = 20$, as well as the maximum-likelihood EM-HMM [7] with a state initialization of 50. We also compare with a simple calculation of the KL-divergence between the empirical codebook usage of two pieces of music - equivalent to building an HMM with one state. Another comparison is with the Gaussian mixture model using variational Bayes inference [33], termed the VB-GMM, where we use a 50-dimensional Dirichlet prior with $\alpha = 1$ on the mixing weights. This model is built on the original feature vectors and an equivalent distance metric is used.

To assess quality, in Tables 2-4, we ask a series of questions regarding the top 10 recommendations for each piece of music and tabulate the probability of a recommendation meeting these criteria. We believe that these questions give a good overview of the quality of the recommendations on both large and fine scales. For example, the KL-divergence is able to recommend music well by genre, but on the subgenre level the performance is noticeably worse; the VB-GMM also performs significantly worse. A possible explanation for these inferior results is the lack of sequential modeling of the data. We also notice that the performance of the finite Dirichlet model is not consistent for various α settings. As α increases, we see a degradation of performance, which we attribute to overfitting due to the increased state usage. This overfitting is most clearly highlighted in the EM implementation of the HMM with the state number

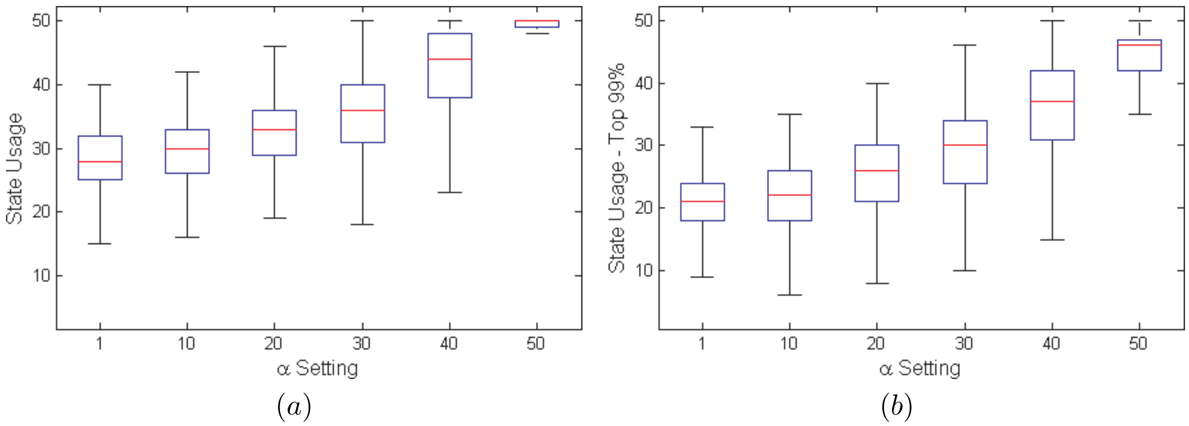


Fig. 8. (a) A box plot of the state usage for the finite Dirichlet HMMs of 2,250 pieces of music as a function of parameter α with 50 initialized states. (b) Box plots of the smallest number of states containing at least 99% of the data as a function of α As can be seen, the state usage for the standard variational HMM is very sensitive to this parameter.

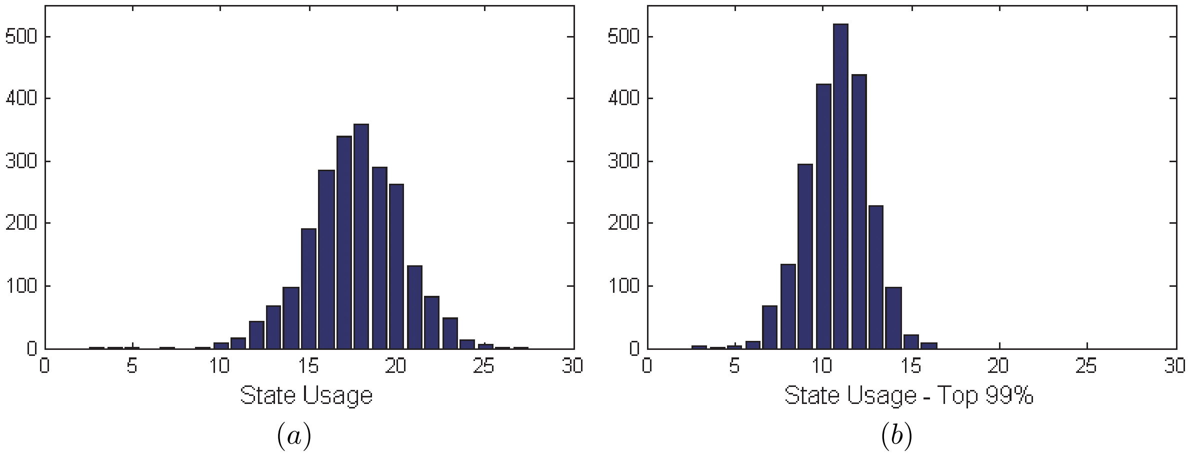


Fig. 9. (a) A histogram of the number of occupied states using a $Gamma(10^{-6}, 0.1)$ prior on α_{ij} and state initialization 50 as well as (b) the smallest number of states containing at least 99% of the data. The reduction in states is greater than that for any parametrization of the Dirichlet model.

initialized to 50.

We emphasize that we do not conclude on the basis of our experiments that our model is superior to the finite Dirichlet HMM, but rather that our prior provides the ability to infer a greater range in the underlying state structure than possible with the finite Dirichlet approach, which in certain cases, such as our music recommendation problem, may be desirable. Additionally, though we infer α_{ij} through the use of $Gamma(c, d)$ priors, the hyperparameter settings of those priors, specifically d, may still need to be tuned to the problem at hand. Though it can be argued that a wider range in state structure can be obtained by a smaller truncation of the Dirichlet priors used on $\mathbf{A}$, we believe that this is a less principled approach as it prohibits the data from controlling the state clustering that naturally arises under the given prior. Doing so would also not be in the spirit of nonparametric inference, which is our motivation here. Rather, we could see our model being potentially useful in verifying whether a simpler state structure is or is not more appropriate than what the Dirichlet approach naturally infers in the variational Bayes setting, while still allowing for this state structure to be inferred nonparametrically.

VII. Conclusion

We have presented an infinite-state hidden Markov model that utilizes the stick-breaking construction to simplify model complexity by reducing state usage to the amount necessary to properly model the data. This was aided by the use of gamma priors on the α parameters of the beta distributions used for each break, which acts as a faucet allowing data to pass to higher state indices. The efficacy of our model was demonstrated in the MCMC and variational Bayes settings on synthesized data, as well as on a music recommendation problem, where we showed that our model performs as well, or better, than a variety of other algorithms.

We mention that this usage of gamma priors can also be applied to infinite mixture modeling as well. The stick-breaking construction of Sethuraman, which uses a $Beta(1, \alpha)$ prior on the stick-breaking proportions, only allows for a single gamma prior to be placed on the shared α. This is necessary to be theoretically consistent with the Dirichlet process, but can be too restrictive in the inference process due to the "left-bias" of the prior. If this were to be relaxed and separate gamma priors were to be placed on each α_i for each break V_i, this problem would be remedied, though the resulting model would no longer be a Dirichlet process.

References

[1] D. Aldous (1985). Exchangeability and related topics. *École d'ete de probabilités de Saint-Flour XIII-1983* 1-198 Springer, Berlin.

[2] C.E. Antoniak (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2:1152-1174.

[3] K.B. Athreya, H. Doss and J. Sethuraman (1996). On the convergence of the Markov chain simulation method. *Annals of Statistics*, 24:69-100.

[4] J.J. Aucouturier and M. Sandler (2001). Segmentation of musical signals using hidden markov models. *In Proceedings of the 110th Convention of the Audio Engineering Society.*

[5] L.R. Bahl, F. Jelinek, and R. L. Mercer (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 2, pp. 179190.

[6] M.J. Beal (2003). *Variational Algorithms for Approximate Bayesian Inference* PhD thesis, Gatsby Computational Neuroscience Unit, University College London.

[7] J.A. Bilmes (1998). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report 97-021, UC-Berkeley.

[8] D.M. Blei and M.I. Jordan (2006). Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, vol. 1, no. 1, pp. 121-144.

[9] R.J. Boys, and D.A. Henderson (2001). A comparison of reversible jump MCMC algorithms for DNA sequence segmentation using hidden Markov models. *Computing Science and Statistics* 33:3549.

[10] P.K. Bharadwaj, P.R. Runkle, and L. Carin (1999). Target identification with wave-based matched pursuits and hidden Markov models. *IEEE Transactions on Antennas and Propagation* vol. 47, pp. 1543-1554.

[11] D. Blackwell and J.B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics,* 1:353-355.

[12] R.J. Connor and J.E. Mosimann (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association,* 64:194-206.

[13] S. Davis and P. Mermelstein (1980). Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust.Speech Signal Proc.,* 28: 357-366.

[14] A. Dempster, N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, vol. 39, no. 1, pp. 138.

[15] R. Durbin, S.R. Eddy, A. Krogh and G. Mitchison (1999). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press.

[16] M.D. Escobar and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association.* vol. 90, no. 430, pp. 577-588.

[17] T. Ferguson (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics,* 1:209-230.

[18] P. Halmos (1944). Random Alms. *The Annals of Mathematical Statistics,* 15:182-189.

[19] H. Ishwaran and L.F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association,* 96:161-173,.

[20] H. Ishwaran and M. Zarepour (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica,* 12:941-963.

[21] N. Johnson and S. Kotz (1977). *Urn Models and Their Applications*. Wiley Series in Probability and Mathematical Statistics.

[22] J.S. Liu (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association,* 89:958-966.

[23] S. MacEachern and P. Mueller (2000). "Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichelt Process Mixture Models," *in Robust Bayesian Analysis*, F. Ruggeri and D. Rios Insua (eds.), Springer-Verlag.

[24] D.J.C. MacKay (1997). Ensemble learning for hidden Markov models. Technical report, Cavendish Laboratory, University of Cambridge.

[25] O. Papaspiliopoulos and G.O. Roberts (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika,* 95,1, pp. 169-186.

[26] I. Porteous, A. Ihler, P. Smyth and M. Welling (2006). Gibbs sampling for (coupled) infinite mixture models in the stick-breaking representation. *Proc. of the Conf. on Uncertainty in Artificial Intelligence*, Pittsburgh, PA.

[27] K. Ni, Y. Qi and L. Carin (2007). Multi-aspect target detection with the infinite hidden Markov model. *Journal of the Acoustical Society of America,* To appear.

[28] L.R. Rabiner (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE,* Vol. 77, No 2.

[29] P.R. Runkle, P.K. Bharadwaj, L. Couchman, and L. Carin (1999). Hidden Markov models for multiaspect target classification. *IEEE Transactions on Signal Processing* vol. 47, pp. 2035-2040.

[30] J. Sethuraman (1994). A constructive definition of Dirichlet priors. *Statistica Sinica,* 4:639-650.

[31] Y.W. Teh, M.I. Jordan, M.J. Beal and D.M. Blei (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association,* 101(476):1566-1581.

[32] J. Winn and C.M. Bishop (2005). Variational message passing. *Journal of Machine Learning Research,* 6:661-694.

[33] J. Winn (2004). *Variational Message Passing and its Applications* PhD thesis Inference Group, Cavendish Laboratory, University of Cambridge.

[34] T.T. Wong (1998). Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation,* 97:165-181.

Music Genre	Classical	Jazz	Rock	Overall
TSB-HMM	0.9581	0.9269	0.9121	0.9324
DD-HMM, α = 1	0.9627	0.9220	0.9161	0.9336
DD-HMM, α = 20	0.9633	0.9399	0.8225	0.9086
Direct KL	0.9388	0.9213	0.9117	0.9240
EM-HMM, state = 50	0.8597	0.8749	0.9444	0.8930
VB-GMM	0.8933	0.8116	0.8933	0.8661

TABLE II
THE PROBABILITY THAT A RECOMMENDATION IN THE TOP 10 IS OF THE SAME GENRE.

Classical Genre	Chamber	Orchestral	Piano	Overall
TSB-HMM	0.7496	0.6968	0.9696	0.8053
DD-HMM, α = 1	0.7380	0.7148	0.9748	0.8092
DD-HMM, α = 20	0.7948	0.6808	0.9128	0.7961
Direct KL	0.6832	0.6608	0.9556	0.7665
EM-HMM, state = 50	0.5660	0.5692	0.8908	0.6753
VB-GMM	0.7124	0.3916	0.8988	0.6676

TABLE III
CLASSICAL - THE PROBABILITY THAT A RECOMMENDATION IN THE TOP 10 IS OF THE SAME SUBGENRE.

Jazz/Rock Genres	Saxophone	Hard Rock	The Beatles
TSB-HMM	0.6951	0.5710	0.4884
DD-HMM, α = 1	0.7092	0.5735	0.4953
DD-HMM, α = 20	0.5919	0.4548	0.4675
Direct KL	0.5968	0.4600	0.3978
EM-HMM, state = 50	0.6311	0.5019	0.3549
VB-GMM	0.5095	0.5310	0.3310

TABLE IV
JAZZ - THE PROBABILITY THAT A TOP 10 RECOMMENDATION FOR SAXOPHONISTS JOHN COLTRANE, CHARLIE PARKER AND SONNY ROLLINS IS FROM ONE OF THESE SAME THREE ARTISTS. HARD ROCK - THE PROBABILITY THAT A TOP 10 RECOMMENDATION FOR HARD ROCK GROUPS JIMI HENDRIX AND LED ZEPPELIN IS FROM ONE OF THESE SAME TWO ARTISTS. THE BEATLES - THE PROBABILITY THAT A TOP 10 BEATLES RECOMMENDATION IS ANOTHER BEATLES SONG.