An Infinite Hidden Markov Model With Local Transitions

Colin Reimer Dawson
Department of Mathematics
Oberlin College
Oberlin, OH 44074
cdawson@oberlin.edu

Clayton T. Morrison
School of Information
University of Arizona
claytonm@email.arizona.edu

Abstract

We describe a generalization of the Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) which is able to encode prior information that state transitions are more likely between "similar" states. This is accomplished by defining a similarity kernel on the state space, and scaling transition probabilities by pairwise similarities. This induces a global correlation structure over the transition probabilities based on the topology induced by the similarity kernel. We call this model the Hierarchical Dirichlet Process Hidden Markov Model with Local Transitions (HDP-HMM-LT). Unfortunately the conditional posterior of the transition distributions are no longer conjugate, due to the varying scale parameters, and so we present an alternative representation of this process as the marginalization of a Markov Jump Process in which: (1) some jump attempts fail, and (2) the probability of success is proportional to the similarity between the source and destination states. When holding times and failed transitions are reintroduced as auxiliary data, conditional conjugacy is restored, admitting exact Gibbs sampling. Even without the LT modification, conditioning on the holding times simplifies inference for the concentration parameters of the HDP, and allows immediate generalization to Semi-Markov dynamics without additional data augmentation. We evaluate the model and inference method on a collection of speaker diarization data sets in which speakers form conversational groups, so that there is prior dependence among chains, but most transitions involve only one or two chains at a time (transitions are local). Our model compares favorably to both the the HDP-H(S)MM and Binary Factorial HMM without suffering in performance when the data is generated directly from the comparison models.

1 Background and Related Work

The conventional Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) [1, 9] is a prior distribution on the transition matrix of a Hidden Markov Model with a countably infinite state space. The rows of the infinite matrix are coupled through their dependence on a common, discrete base measure, itself drawn from a Dirichlet Process (DP). The hierarchical structure ensures that, despite the infinite state space, a common set of destination states will be reachable with high probability from each source state. The generative process for the HDP-HMM is the following:

Each of a countably infinite set of states, indexed by j, has parameters, θ_j, drawn from a base measure, H. A top-level set of state weights, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots)$, is drawn from a stick-breaking process (GEM) with concentration parameter $\gamma > 0$.

$$\theta_j \stackrel{i.i.d.}{\sim} H \qquad \boldsymbol{\beta} \sim \mathsf{GEM}(\gamma) \tag{1}$$

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

The actual transition distribution, $\boldsymbol{\pi}_j$, from state j, is drawn from a DP with concentration α and base measure $\boldsymbol{\beta}$:

$$\boldsymbol{\pi}_j \overset{i.i.d}{\sim} DP(\alpha\boldsymbol{\beta}) \qquad j = 1, 2, \ldots \tag{2}$$

The hidden state sequence is then generated according to the π_j. Let z_t be the index of the chain's state at time t. Then we have

$$z_t \mid z_{t-1}, \boldsymbol{\pi}_{z_{t-1}} \sim \boldsymbol{\pi}_{z_{t-1}} \qquad t = 1, 2, \ldots, T \tag{3}$$

where T is the length of the data sequence. Finally, the emission distribution for state j is a function of θ_j, so that we have

$$y_t \mid z_t, \theta_{z_t} \sim F(\theta_{z_t}) \tag{4}$$

A shortcoming of this model is that the generative process does not take into account the fact that the set of source states is the same as the set of destination states: that is, the distribution $\boldsymbol{\pi}_j$ has an element which corresponds to state j. Put another way, there is no special treatment of the diagonal of the transition matrix, so that self-transitions are no more likely *a priori* than transitions to any other state. The Sticky HDP-HMM [3] addresses this issue by adding an extra mass at location j to the base measure of the DP that generates $\boldsymbol{\pi}_j$. That is, (2) is replaced by

$$\boldsymbol{\pi}_j \sim DP(\alpha\boldsymbol{\beta} + \kappa\delta_j). \tag{5}$$

An alternative approach that treats self-transitions as special is the HDP Hidden Semi-Markov Model (HDP-HSMM) [7], wherein state duration distributions are modeled separately, and ordinary self-transitions are ruled out. In both the Sticky HDP-HMM and the HDP-HSMM, auxiliary latent variables are introduced to simplify conditional posterior distributions and facilitate Gibbs sampling. However, while both of these models have the ability to privilege self-transitions, they contain no notion of similarity for pairs of states that are not identical: in both cases, when the transition matrix is integrated out, the prior probability of transitioning to state j' depends only on the top-level stick weight associated with state j', and not on the identity or parameters of the previous state j.

The present paper makes two main contributions: first we define the HDP-HMM with Local Transitions (HDP-HMM-LT), in which a similarity structure is defined on the state space, and "similar" states are *a priori* more likely to have transitions between them. This is accomplished by element-wise rescaling and then renormalizing the HDP transition matrix. A similar rescaling and renormalization approach is taken in the Discrete Infinite Logistic Normal (DILN) model [8] in the setting of topic modeling. There, however, the contexts and the mixture components (topics) are distinct sets, and there is no notion of temporal dependence. The second contribution of the present paper is an augmented data technique that allows Gibbs sampling to be used. Previous rescaled HDP models such as DILN have relied on variational inference.

The paper is structured as follows: In section 2 we define the HDP-HMM-LT. In section 3, we develop a straightforward Gibbs sampling algorithm based on an augmented data representation, the Markov Jump Process with Failed Transitions. In section 4 we test the model and inference algorithm on synthetic data generated from similar generative models, as well as on a speaker diarization task in which the speakers are inter-dependent. Finally, in section 5, we conclude and discuss the relationships between the HDP-HMM-LT and existing HMM variants.

2 An HDP-HMM With Local Transitions

We wish to add to the transition model the concept of a transition to a "nearby" state, where nearness of j and j' is a function of θ_j and $\theta_{j'}$. In order to accomplish this, we first consider an alternative construction of the transition distributions, based on the Normalized Gamma Process.

2.1 A Normalized Gamma Process representation of the HDP-HMM

We can define a random measure, $\mu = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}$, where

$$\pi_j \overset{ind}{\sim} \mathsf{Gamma}(w_j, 1) \qquad T = \sum_{j=1}^{\infty} \pi_j \qquad \tilde{\pi}_j = \frac{\pi_j}{T} \tag{6}$$

and subject to the constraint that $\sum_{j\geq 1} w_j < \infty$. It follows [8, 2] that μ is distributed as a Dirichlet Process with base measure $\mathbf{w} = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$. If we draw $\boldsymbol{\beta}$ from a stick-breaking process and then draw a series $\{\mu_m\}_{m=1}^{M}$ of i.i.d. random measures from the above process, setting $\mathbf{w} = \alpha\boldsymbol{\beta}$ for some $\alpha > 0$, then this defines a Hierarchical Dirichlet Process. If, moreover, there is one μ associated with every state j, then we obtain the transition prior for the HDP-HMM, where

$$p(z_t \mid z_{t-1}, \boldsymbol{\pi}) = \tilde{\pi}_{z_{t-1} z_t} \tag{7}$$

2.2 Promoting "Local" Transitions

In the preceding formulation, the θ_j and the $\pi_{jj'}$ are independent conditioned on the top-level measure. Our goal is to relax this assumption, in order to incorporate possible prior knowledge that certain "location" pairs, $(\theta_j, \theta_{j'})$, are more likely than others to produce large transition weights (i.e., states adjacent in time should tend to be similar). This can be accomplished by scaling the elements $\pi_{jj'}$ by a function of $(\theta_j, \theta_{j'})$ prior to normalization, or equivalently letting the Gamma distribution have a proximity-dependent rate parameter.

Let $\Phi : \Omega \times \Omega \to [0, \infty)$ represent a "similarity function", and define a collection of random variables $\{\phi_{jj'}\}_{j,j'\geq 1}$ according to $\phi_{jj'} = \phi(\theta_j, \theta'_j)$. We can then generalize (6) to

$$\pi_{jj'} \mid \boldsymbol{\beta}, \boldsymbol{\theta} \sim \mathsf{Gamma}(\alpha\beta_{j'}, \phi_{jj'}^{-1}) \qquad T_j = \sum_{j'=1}^{\infty} \pi_{jj'} \qquad \tilde{\pi}_{jj'} = \frac{\pi_{jj'}}{T_j} \tag{8}$$

so that the prior mean of $\pi_{jj'}$ is $\alpha\beta_{j'}\phi_{jj'}$. Since a similarity between one object and another should not exceed the similarity between an object and itself, and since a constant rescaling of the similarity will be absorbed in normalization, we will assume that $0 \leq \phi_{jj'} \leq 1$ for all j and j'.

2.3 The HDP-HMM-LT as the Marginalization of a Markov Jump Process with "Failed" Jumps

We can gain stronger intuition, as well as simplify posterior inference, by representing the HDP-HMM-LT described in the last section as a continuous time Markov jump process where holding times have been integrated out. In particular, suppose that some of the attempts to jump from one state to another fail, and the failure probability increases as a function of the "distance" between the states.

Let Φ be defined as in the last section, and let $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ be defined as in the Normalized Gamma Process representation of the ordinary HDP-HMM (so, $\pi_{jj'} \mid \boldsymbol{\beta}, \boldsymbol{\theta} \sim \mathsf{Gamma}(\alpha\beta_{j'}, 1)$). Now suppose that when the process is in state j, jumps to state j' are made at rate $\pi_{jj'}$. This defines a continuous-time Markov Process where the off-diagonal elements of the transition rate matrix are the off diagonal elements of $\boldsymbol{\pi}$. In addition, self-jumps are allowed, and occur with rate π_{jj}. If we only observe the jumps and not the durations between jumps, this is an ordinary Markov chain. If we do not observe the jumps themselves, but instead an observation is generated once per jump from a distribution that depends on the state being jumped to, then we have an ordinary HMM.

We modify this process as follows. Suppose that each jump attempt from state j to state j' has a chance of failing, which is an increasing function of the "distance" between the states. In particular, let the success probability be $\phi_{jj'}$ (recall that we assumed above that $0 \leq \phi_{jj'} \leq 1$ for all j, j'). Then, the rate of successful jumps from j to j' is $\pi_{jj'}\phi_{jj'}$, and the corresponding rate of unsuccessful jump attempts is $\pi_{jj'}(1 - \phi_{jj'})$. We denote the overall rate of successful jumps while in state j by $T_j := \sum_{j'} \pi_{jj'}\phi_{jj'}$. Given that the process is in state j at discretized time t (that is, $z_t = j$), the probability that the first successful jump is to state j' (that is, $z_{t+1} = j'$) is proportional to the rate of successful jump attempts to j', which is $\pi_{jj'}\phi_{jj'}$. The holding time, τ_{t-1}, is independent of z_{t+1} and is distributed $\mathsf{Exp}(T_j)$. The total time spent in state j given that it is visited n_j times, is then

$$u_j \mid \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta} \stackrel{ind}{\sim} \mathsf{Gamma}(n_{j\cdot}, T_j) \tag{9}$$

During this period there will be $q_{jj'}$ unsuccessful attempts to jump to state j', where $q_{jj'}$ is distributed $\mathsf{Pois}(u_j\pi_{jj'}(1-\phi_{jj'}))$. Incorporating $\mathbf{u} = \{u_j\}$ as augmented data simplifies the likelihood for the transition parameters, yielding

$$L(\boldsymbol{\pi},\boldsymbol{\phi}\,|\,\mathbf{z},\mathbf{u},\mathbf{Q}) = \left(\prod_{t=1}^{T} p(z_t\,|\,z_{t-1},\boldsymbol{\pi},\boldsymbol{\phi})\right)\prod_{j} p(u_j\,|\,\mathbf{z},\boldsymbol{\pi},\boldsymbol{\phi})\prod_{j'} p(q_{jj'}\,|\,u_j\pi_{jj'},\phi_{jj'})$$
$$\propto \prod_{j}\prod_{j'} \pi_{jj'}^{n_{jj'}+q_{jj'}}\phi_{jj'}^{n_{jj'}}(1-\phi_{jj'})^{q_{jj'}}e^{-\pi_{jj'}u_j} \tag{10}$$

2.4 An HDP-HSMM-LT modification

We note that it is trivial to modify the HDP-HMM-LT to allow for non-Geometric duration distributions, by simply fixing the diagonal elements of $\boldsymbol{\pi}$ to be zero, allowing D_t observations to be emitted $i.i.d.$ according to $F(\theta_{z_t})$ at jump t, where D_t is drawn from a state-specific duration distribution, and sampling the latent state sequence using a message passing algorithm suited for HSMMs [7]. Inference for the $\boldsymbol{\phi}$ matrix is not affected, since the diagonal elements are assumed to be 1. Unlike in the original representation of the HDP-HSMM, there is no need to introduce additional auxiliary variables as a result of this modification, due to the presence of the (continuous) durations, $\mathbf{u}$, which were already needed to account for the normalization of the $\boldsymbol{\pi}$.

3 Inference

We develop a Gibbs sampling algorithm based on the Markov Process with Failed Jumps representation, augmenting the data with the duration variables $\mathbf{u}$, the failed jump attempt count matrix, $\mathbf{Q}$, as well as additional auxiliary variables which we will define below. In this representation the transition matrix is not represented directly, but is a function of the unscaled transition matrix π and the similarity matrix $\boldsymbol{\phi}$. The full set of variables is partitioned into blocks: $\{\gamma,\alpha,\boldsymbol{\beta},\boldsymbol{\pi}\}$, $\{\mathbf{z},\mathbf{u},\mathbf{Q},\Lambda\}$, $\{\boldsymbol{\theta}\}$, and $\{\xi\}$, where Λ represents a set of auxiliary variables that will be introduced below, $\boldsymbol{\theta}$ represents the emission and state location parameters (which may be further factored depending on the specific choice of model), and ξ represents additional parameters such as any free parameters of the similarity function, Φ, and any hyperparameters of the emission distribution.

3.1 Sampling Transition Parameters and Hyperparameters

The joint posterior over γ, α, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ given the other variables will factor as

$$p(\gamma,\alpha,\boldsymbol{\beta},\boldsymbol{\pi}) = p(\gamma)p(\alpha)p(\boldsymbol{\beta}\,|\,\gamma)p(\boldsymbol{\pi}\,|\,\alpha,\boldsymbol{\beta}) \tag{11}$$

where we have omitted the dependence on the augmented data, $\mathcal{D} = (\mathbf{z},\mathbf{u},\mathbf{Q},\Lambda)$ for conciseness. We describe these four factors in reverse order.

Sampling $\boldsymbol{\pi}$ Having used data augmentation to simplify the likelihood for $\boldsymbol{\pi}$ to the factored conjugate form in (10), the individual $\pi_{jj'}$ are *a posteriori* independent Gamma distributed:

$$\pi_{jj'}\,|\,\alpha,\beta_{j'},\mathcal{D} \overset{ind}{\sim} \mathsf{Gamma}(\alpha\beta_{j'}+n_{jj'}+q_{jj'},1+u_j) \qquad j,j'\geq 1 \tag{12}$$

Sampling $\boldsymbol{\beta}$ To enable joint sampling of the latent state sequence, we employ a weak limit approximation to the HDP [7], approximating the stick-breaking process for $\boldsymbol{\beta}$ using a finite Dirichlet distribution with a finite number of components, J, which is larger than we expect to need. Due to the product-of-Gammas form, we can integrate out $\boldsymbol{\pi}$ analytically from $p(\boldsymbol{\pi},\mathcal{D}\,|\,\boldsymbol{\beta})$, to obtain the marginal likelihood for $\boldsymbol{\beta}$. Together, we have

$$p(\boldsymbol{\beta}\,|\,\gamma) = \frac{\Gamma(\gamma/J)^J}{\Gamma(\gamma)}\prod_j \beta_j^{\frac{\gamma}{J}-1} \qquad p(\mathcal{D}\,|\,\boldsymbol{\beta},\alpha) \propto \prod_{j=1}^{J}(1+u_j)^{-\alpha}\prod_{j'}\frac{\Gamma(\alpha\beta_{j'}+n_{jj'}+q_{jj'})}{\Gamma(\alpha\beta_{j'})} \tag{13}$$

where we have used the fact that the β_j sum to 1 to pull out terms of the form $(1+u_j)^{-\alpha\beta_{j'}}$ from the inner product in the likelihood. Following Teh et al. (2006), we can introduce auxiliary variables $\{m_{jj'}\}$, with

$$p(m_{jj'}\,|\,\beta_{j'},\alpha,\mathcal{D}) \overset{ind}{\propto} s(n_{jj'}+q_{jj'},m_{jj'})\alpha^{m_{jj'}}\beta_{j'}^{m_{jj'}} \tag{14}$$

for integer $m_{jj'}$ ranging between 0 and $n_{jj'} + q_{jj'}$, where $s(n, m)$ is an unsigned Stirling number of the first kind. The normalizing constant in this distribution cancels the ratio of Gamma functions in the $\boldsymbol{\beta}$ likelihood, so, letting $m_{\cdot j} = \sum_{j'} m_{j'j}$, we obtain simply

$$\boldsymbol{\beta} \mid \mathbf{M}, \gamma \sim \text{Dirichlet}(\frac{\gamma}{J} + m_{\cdot 1}, \dots, \frac{\gamma}{J} + m_{\cdot J}) \tag{15}$$

Sampling Concentration Parameters After incorporating $\mathbf{M}$ into $\mathcal{D}$, we can integrate out $\boldsymbol{\beta}$, from the joint likelihood $p(\boldsymbol{\beta}, \mathcal{D} \mid \gamma, \alpha)$:

$$p(\mathcal{D} \mid \alpha, \gamma) \propto \alpha^{m_{\cdot\cdot}} e^{-\sum_{j''} \log(1+u_{j''})\alpha} \frac{\Gamma(\gamma)}{\Gamma(\gamma + m_{\cdot\cdot})} \prod_j \frac{\Gamma(\frac{\gamma}{J} + m_{\cdot j})}{\Gamma(\frac{\gamma}{J})} \tag{16}$$

Assume that α and γ have Gamma priors. Then the update to alpha is conjugate,

$$\alpha \mid \mathcal{D} \sim \mathsf{Gamma}(a_\alpha + m_{\cdot\cdot}, b_\alpha + \sum_j \log(1 + u_j)), \tag{17}$$

and to simplify the likelihood for γ, we can introduce a final set of auxiliary variables, $\mathbf{r} = (r_1, \dots, r_J)$, $r_j \in \{0, \dots, m_{\cdot j}\}$ and $t \in (0, 1)$ with the following distributions:

$$p(r_j \mid m_{\cdot j}, \gamma) \propto s(m_{\cdot j}, r) \left(\frac{\gamma}{J}\right)^r \qquad p(t \mid m_{\cdot\cdot} \gamma) \propto t^{\gamma-1}(1-t)^{m_{\cdot\cdot}-1}. \tag{18}$$

The normalizing constants are ratios of Gamma functions, which cancel those in (16), so that

$$\gamma \mid \mathcal{D} \sim \mathsf{Gamma}(a_\gamma + r_\cdot, b_\gamma - \log(t)) \tag{19}$$

3.2 Sampling $\mathbf{z}$ and the auxiliary variables

We sample the hidden state sequence, $\mathbf{z}$, jointly with the auxiliary variables, which consist of $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{M}$, $\mathbf{r}$ and t. The joint conditional distribution of these variables is defined directly by the generative model:

$$p(\mathcal{D}) = p(\mathbf{z})p(\mathbf{u} \mid \mathbf{z})p(\mathbf{Q} \mid \mathbf{u})p(\mathbf{M} \mid \mathbf{z}, \mathbf{Q})p(\mathbf{r} \mid \mathbf{M})p(t \mid \mathbf{M}) \tag{20}$$

Since we are conditioning on the transition matrix, we can sample the entire sequence $\mathbf{z}$ at once with the forward-backward algorithm, as in an ordinary HMM, or its corresponding generalization if we are using the HSMM variant. Having done this, we can sample $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{M}$, $\mathbf{r}$ and t from their forward distributions.

3.3 Sampling state and emission parameters

Depending on the application, the similarities $\{\phi_{jj'}\}$ may be based directly on the emission distributions, or may be based on a separate set of variables. In the experiments described below we assume the former. For simplicity, we denote the collection of these variables by $\boldsymbol{\theta}$. We have two likelihood components:

$$p(\mathbf{z}, \mathbf{Q} \mid \boldsymbol{\theta}) \propto \prod_j \prod_{j'} \phi_{jj'}^{n_{jj'}} (1 - \phi_{jj'})^{q_{jj'}} \qquad p(\mathbf{Y} \mid \mathbf{z}, \boldsymbol{\theta}) = \prod_{t=1}^{T} f(\mathbf{y}_t; \theta_{z_t}) \tag{21}$$

where proportionality is with respect to variation in $\boldsymbol{\theta}$.

4 Experiments

The parameter space for the hidden states, the associated prior H on $\boldsymbol{\theta}$, and the similarity function Φ, is application-specific; we consider here the case where a state consists of a finite D-dimensional binary vector, η_j, the similarity function is a Laplacian kernel defined with respect to Hamming distance between pairs η_j and $\eta_{j'}$ with decay parameter λ, and the emission distribution is linear-Gaussian, with $D \times K$ weight matrix $\mathbf{W}$, so that each K-dimensional observation is $\mathcal{N}(\mathbf{W}\eta_j, \Sigma)$. For experiments discussed here, we will assume that Σ does not depend on j, but this assumption is easily relaxed if appropriate. For finite-length binary vector states, the set of possible states is finite, and so on its face it may seem that a nonparametric model is unnecessary. However, if D is reasonably large, it is likely that most of the 2^D possible states are vanishingly unlikely (and, in fact, the number of observations may well be less than 2^D), and so we would like a model that encourages the selection of a sparse set of states. Moreover, there could be more than one state with the same emission parameters, but with different transition dynamics. Before describing individual experiments, we describe the additional inference steps needed for these variables.

4.1 Additional Inference Steps

Sampling η We put independent Beta-Bernoulli priors on each coordinate of η. We Gibbs sample each coordinate η_{jd} conditioned on all the others and the coordinate-wise prior means, $\{\mu_d\}$, which we sample in turn conditioned on the ηs.

Sampling λ The Laplacian kernel Φ is defined as $\Phi(\eta_j, \eta_{j'}) = e^{-\lambda d(\eta_j, \eta_{j'})}$, where in our case d is Hamming distance. The parameter λ governs the connection between $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Writing (21) in terms of λ and the distance matrix $\boldsymbol{\Delta}$ gives the likelihood

$$p(\mathbf{z}, \mathbf{Q} \mid \lambda, \boldsymbol{\theta}) \propto \prod_j \prod_{j'} e^{-\lambda \Delta_{jj'} n_{jj'}} (1 - e^{-\lambda \Delta_{jj'}})^{q_{jj'}} \tag{22}$$

We put an $\mathsf{Exp}(b_\lambda)$ prior on λ, which yields a posterior density

$$p(\lambda \mid \mathbf{z}, \mathbf{Q}, \boldsymbol{\theta}) \propto e^{-(b_\lambda + \sum_j \sum_{j'} \Delta_{jj'} n_{jj'})\lambda} \prod_j \prod_{j'} (1 - e^{-\lambda \Delta_{jj'}})^{q_{jj'}} \tag{23}$$

This density is log-concave, and so we use Adaptive Rejection Sampling [6] to sample from it.

Sampling $\mathbf{W}$ and Σ Conditioned on the state matrix $\boldsymbol{\theta}$ and the data matrix $\mathbf{Y}$, the weight matrix $\mathbf{W}$ can be sampled as well using standard methods for Bayesian linear regression. We place a zero mean Normal prior on each element of $\mathbf{W}$ (including a row of intercept terms), resulting in a multivariate Normal posterior for each column. For the experiments reported below, we constrain Σ to be a diagonal matrix, and place an Inverse Gamma prior on the variances, resulting in conjugate updates.

4.2 "Cocktail Party" Data

To evaluate the model, we created synthetic data based on a speaker diarization task, with the property that speakers are grouped into conversations, and take turns speaking within conversation. In such a task, there are naively 2^S possible states, where each state indicates which of S speakers is speaking at a particular time step. However, due to the conversational grouping, if at most one speaker in a conversation can be speaking at any given time (a natural assumption of within-group turn-taking), the state space is constrained, with only $\prod_c (s_c + 1)$ states possible, where s_c is the number of speakers in conversation c.

For the cocktail party data, the turn sequence within conversations is generated using a Poisson HSMM with s_c states, with pauses with shorter Poisson duration inserted between each "sentence". The states within conversations were then mapped to a binary vector with s_c entries, where silence is represented by all zeroes, and speaker s speaking corresponds to a 1 in position s. The binary vectors were concatenated across conversations to yield latent states consisting of length S binary vectors. To simulate speakers being recorded by K microphones, weights from speakers to microphones were generated independently from a $U(0, 1)$ distribution, resulting in a $D \times K$ weight matrix, $\mathbf{W}$. A constant "background sound level" parameter for each microphone was added as well, also $U(0, 1)$, and independent $\mathcal{N}(0, \sigma_k^2)$ noise was added to each time step at microphone k.

We generated transition and emission parameters from conjugate priors to the Poisson HSMM. The data set consisted of four conversations of four speakers each, and 12 microphones, so that $D = 16$, and $K = 12$. There are therefore $2^{16} = 65536$ possible binary vector-valued states, but only $(4 + 1)^4 = 625$, less than 1%, can actually occur. The noise variance σ_k^2 was set to a constant of $1/10$ for all $k = 1, \ldots, K$.

We attempted to infer the states from the data using three models: (1) a binary-state Factorial HMM, in which the individual binary speaker sequences are modeled as independent a priori, (2) an ordinary HDP-HMM without local transitions, where the latent states are binary vectors, and (3) our

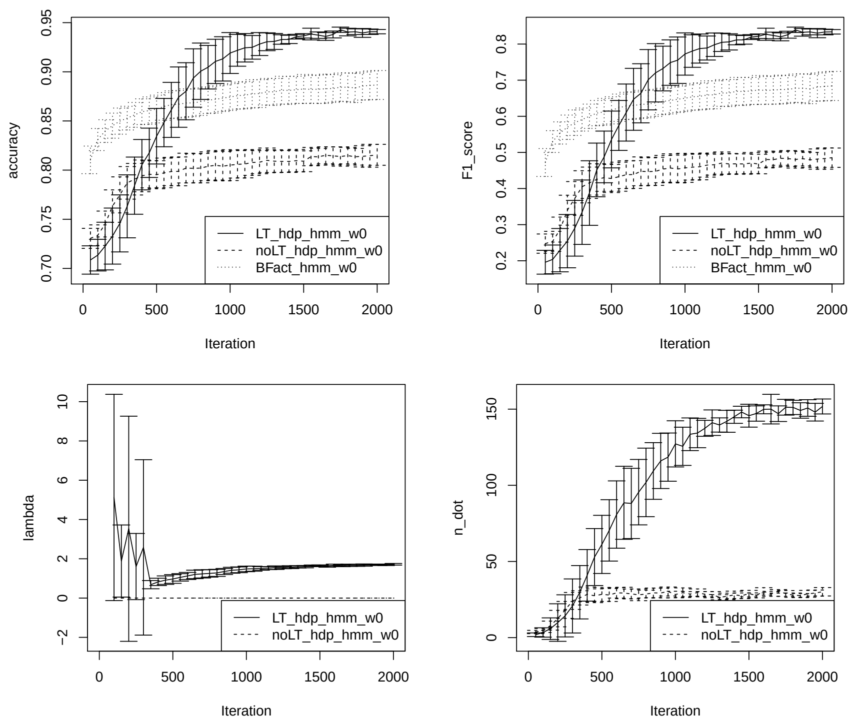

Figure 1: (a-b) Accuracy and F1 scores for the HDP-HMM-LT, standard HDP-HMM, and Binary Factorial HMM on the Cocktail Party Data. Metrics are averaged over 10 Gibbs runs on each model, with error bars representing a 99% confidence interval for the mean per iteration. The first 100 iterations are not shown. (c) Learned similarity parameter, λ, for the LT model, (d) Number of distinct states used by HDP-HMM and HDP-HMM-LT. The first 100 iterations are excluded.

HDP-HMM-LT model. To simplify interpretation of the results, the weight matrix was fixed to the true value (this makes the latent dimensions identifiable and makes distances between inferred and ground truth state matrices meaningful). We evaluated the models at each iteration using Hamming distance between inferred and ground truth state matrices and F1 score. The results for the three models are in Figure 1. The LT model outperforms the other two on all measures on all datasets. We also plot the inferred decay rate λ for the HDP-HMM-LT model. The LT model settles on a non-negligible λ value for this data, suggesting that the local transition structure explains the data well. It also uses more components than the non-LT model, perhaps owing to the fact that the weaker transition prior of the non-LT model is more likely to explain nearby similar observations as a single persisting state, whereas the LT model places a higher probability on transitioning to a new state with a similar latent vector.

4.3 Synthetic Data Without Local Transitions

We also generated data from an ordinary HDP-HMM, with no local transition property, in order to investigate the performance of our model in a case where the data did not have the key property that its prior equipped it to discover. The results are in Fig. 2. While the λ parameter is large, the LT model has worse performance than the non-LT model on this data, as its bias toward local transitions is not helpful; however, the λ parameter settles near zero as the model learns that local transitions are not more probable. When $\lambda = 0$, the HDP-HMM-LT is an ordinary HDP-HMM. Unlike on the cocktail party data, the LT model does not use more states when the data does not have the LT property.

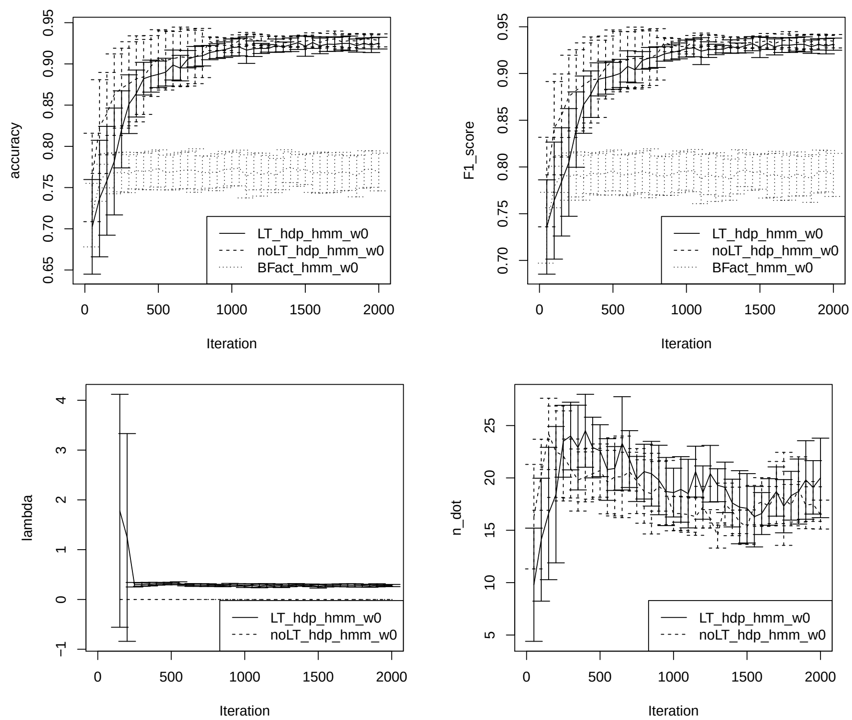

Figure 2: (a-b) Accuracy and F1 for the three models on data generated from an HDP-HMM without local transitions, (c) Learned similarity parameter, λ, for the LT model, (d) Number of states used by the HDP-HMM and HDP-HMM-LT. The first 100 iterations are omitted.

5 Discussion

We have defined a new probabilistic model, the Hierarchical Dirichlet Process Hidden Markov Model with Local Transitions (HDP-HMM-LT), which generalizes the HDP-HMM by encoding prior information about state space geometry via a similarity kernel, making transitions between "nearby" pairs of states more likely *a priori*. By introducing an augmented data representation in the form of a Markov Jump Process with Failed Transitions, we obtain a simple Gibbs sampling algorithm for both the LT and ordinary HDP-HMM. When multiple latent chains are interdependent, the HDP-HMM-LT model combines the HDP-HMM's capacity to discover a suitable number of joint states with the Factorial HMM's ability to encode the property that most transitions involve a small number of chains. The HDP-HMM-LT outperforms both on a speaker diarization task in which speakers perform conversational groups. Despite the addition of the similarity kernel, the HDP-HMM-LT is able to suppress its local transition prior when the data does not support it, achieving identical performance to the HDP-HMM on data generated directly from the latter.

Although the local transition property is particularly clear when changes in state occur at different times for different latent features, as with binary vector-valued states, it is worth noting that the model can be used with any state space equipped with a suitable similarity kernel (in fact, similarities need not be defined in terms of emission parameters as they are here; state "locations" could be represented and inferred separately as in the Discrete Infinite Logistic Normal model [8]). Furthermore, although for simplicity we have focused on fixed-dimension binary vectors, it would be relatively straightforward to add the LT property to a model in which the latent states consist (say) of infinite binary vectors, such as the Infinite Factorial Hidden Markov Model [4] and the Infinite Factorial Dynamic Model [10], both of which use the Indian Buffet Process [5] as the state prior. The similarity kernel used here could be employed there without any change: since all but finitely many coordinates are zero in the Indian Buffet Process, the distance between any two states is finite.

References

[1] Matthew J Beal, Zoubin Ghahramani, and Carl E Rasmussen. The infinite hidden Markov model. In *Advances in neural information processing systems*, pages 577–584, 2001.

[2] Stefano Favaro, Yee Whye Teh, et al. MCMC for normalized random measure mixture models. *Statistical Science*, 28(3):335–359, 2013.

[3] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. An HDP-HMM for systems with state persistence. In *Proceedings of the 25th international conference on Machine learning*, pages 312–319. ACM, 2008.

[4] Jurgen V Gael, Yee W Teh, and Zoubin Ghahramani. The infinite factorial hidden Markov model. In *Advances in Neural Information Processing Systems*, pages 1697–1704, 2009.

[5] Zoubin Ghahramani and Thomas L Griffiths. Infinite latent feature models and the Indian buffet process. In *Advances in neural information processing systems*, pages 475–482, 2005.

[6] Walter R Gilks and Pascal Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, pages 337–348, 1992.

[7] Matthew J Johnson and Alan S Willsky. Bayesian nonparametric hidden semi-Markov models. *The Journal of Machine Learning Research*, 14(1):673–701, 2013.

[8] John Paisley, Chong Wang, and David M Blei. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7(4):997–1034, 2012.

[9] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

[10] Isabel Valera, Francisco Ruiz, Lennart Svensson, and Fernando Perez-Cruz. Infinite factorial dynamical model. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.