An Infinite Hidden Markov Model With Similarity-Biased Transitions

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 function on the state space, and scaling transition probabilities by pairwise similarities. This induces a global correlation structure over the transition probabilities based on the geometry induced by the similarity function. We call this model the Hierarchical Dirichlet Process Hidden Markov Model with Local Transitions (HDP-HMM-LT). Since the conditional posterior of the transition distributions is no longer conjugate, we present an augmented data representation of the model as 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 during inference, 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 speaker diarization task and a "harmonic parsing" task using four-part chorale data, as well as on several synthetic datasets, achieving favorable comparisons to existing models.

1. Introduction and Background

The conventional Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) (Beal et al., 2001; Teh et al., 2006) is a generalization of the conventional finite-state Hidden Markov Model to allow a countably infinite state space. The rows of the infinite transition matrix are coupled by the prior through a dependence on a common, discrete base measure, which is 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 state, indexed by j, has parameters, θ_j, drawn from a base measure, H. A top-level sequence of state weights, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots)$, is drawn by iteratively breaking a "stick" off of the remaining weight according to a $\mathsf{Beta}\,(1, \gamma)$ distribution. The parameter $\gamma > 0$ is known as the concentration parameter and governs how quickly the weights tend to decay, with large γ corresponding to slow decay, and hence more weights needed before a given cumulative weight is reached. This stick-breaking process is denoted by GEM. We thus have

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

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

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

The hidden state sequence is then generated according to the $\boldsymbol{\pi}_j$. Let z_t be the index of the chain's state at time t. Then we have

$$z_t \,|\, z_{t-1}, \boldsymbol{\pi}_{z_{t-1}} \sim \mathsf{Cat}(\boldsymbol{\pi}_{z_{t-1}}) \qquad t = 1, 2, \dots, T \tag{3}$$

where T is the length of the data sequence. Finally, the emission distribution for state j is a function of θ_j, so that observation $\mathbf{y}_t$ is drawn according to

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

A shortcoming of the HDP prior on the transition matrix is that it 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 (Fox et al., 2008) 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

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

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

An alternative approach that treats self-transitions as special is the HDP Hidden Semi-Markov Model (HDP-HSMM; Johnson & Willsky (2013)), 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 main contribution of this paper is a generalization of the HDP-HMM that allows for a similarity structure to be defined on the latent state space, so that "similar" states are *a priori* more likely to have transitions between them. This is accomplished by elementwise rescaling and then renormalizing the HDP transition matrix. We call this model the HDP-HMM with Local Transitions (HDP-HMM-LT). Two versions of the similarity structure are illustrated: in one case, two states are similar to the extent that their emission distributions are similar. In another, the similarity structure is inferred separately. In both cases, we give augmented data representations that restore conditional conjugacy and thus allow a simple Gibbs sampling algorithm to be used for inference.

A rescaling and renormalization approach similar to the one used in the HDP-HMM-LT is used by Paisley et al. (2012) to define their Discrete Infinite Logistic Normal (DILN) model, an instance of a correlated random measure (Ranganath & Blei, in press), 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. In addition, Paisley et al. (2012) employ variational inference, whereas we present a Gibbs sampler.

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, which we call the Markov Jump Process with Failed Transitions (MJP-FT). In section 4 we test two versions of the model: one on a speaker diarization task in which the speakers are inter-dependent, and another on a four-part chorale corpus, demonstrating performance improvements over state-of-the-art models when "local transitions" are more common in the data. Using sythetic data from an HDP-HMM, we show that the LT variant can learn not to use its similarity bias when the data does not support it. 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 transitions between states j and j' are more likely *a priori* to the extent that they are "nearby" in some similarity space. In order to accomplish this, we first consider an alternative construction of the transition distributions, based on the Normalized Gamma Process representation of the DP (Ishwaran & Zarepour, 2002; Ferguson, 1973).

2.1. A Normalized Gamma Process representation of the HDP-HMM

The Dirichlet Process is an instance of a normalized completely random measure (Kingman, 1967; Ferguson, 1973), which can be defined as $G = \sum_{k=1}^{\infty} \tilde{\pi}_k \delta_{\theta_k}$, where

$$\pi_k \overset{ind}{\sim} \mathsf{Gamma}(\alpha\beta_k, 1) \quad T = \sum_{k=1}^{\infty} \pi_k \quad \tilde{\pi}_k = \frac{\pi_k}{T}, \quad (6)$$

δ_θ is a measure assigning 1 to any set containing θ and 0 to any other set, and subject to the constraint that $\sum_{k\geq 1} \beta_k = 1$ and $0 < \alpha < \infty$. It has previously been shown (Ferguson, 1973; Paisley et al., 2012; Favaro et al., 2013) that the normalization constant T is positive and finite almost surely, and that G is distributed as a DP with base measure $G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k}$. If we draw $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots)$ from the $\mathsf{GEM}(\gamma)$ stick-breaking process, draw an i.i.d. sequence of θ_k from a base measure H, and then draw an i.i.d. series of random measures, $\{G_j\}, j = 1, 2, \dots$ from the above process then this defines a Hierarchical Dirichlet Process (HDP). If each G_j is associated with the hidden states of an HMM, $\boldsymbol{\pi}$ is the infinite matrix where entry $\pi_{jj'}$ is the j'th mass associated with the jth random measure, and T_j is the sum of row j, we obtain the prior for the HDP-HMM, where

$$p(z_t \mid z_{t-1}, \boldsymbol{\pi}) = \tilde{\pi}_{z_{t-1} z_t} = \pi_{jj'}/T_j \quad (7)$$

2.2. Promoting "Local" Transitions

In the preceding formulation, the rows of the transition matrix are independent conditioned on the top-level measure. Our goal is to relax this assumption, in order to incorporate possible prior knowledge that certain pairs of states are more likely than others to produce large transition weights between them (in both directions). We accomplish this by associating each latent state, j, with a location, ℓ_j in some space Ω, introducing a "similarity function", $\phi : \Omega \times \Omega \to (0, 1]$, which depending on the application

may or may not depend on the other state parameters, θ_j, and scaling each element, $\pi_{jj'}$, by $\phi_{jj'} = \phi(\ell_j, \ell_{j'})$. Letting $\boldsymbol{\ell} = (\ell_1, \ell_2, \dots)$, we can then replace (6) by

$$\begin{aligned} \pi_{jj'} \,|\, \boldsymbol{\beta}, \boldsymbol{\ell} &\sim \mathsf{Gamma}(\alpha\beta_{j'}, 1) \quad T_j = \sum_{j'=1}^{\infty} \pi_{jj'}\phi_{jj'} \\ \tilde{\pi}_{jj'} &= \frac{\pi_{jj'}\phi_{jj'}}{T_j} \quad p(z_t \,|\, z_{t-1}, \boldsymbol{\pi}, \boldsymbol{\ell}) = \tilde{\pi}_{z_{t-1}z_t}. \end{aligned} \tag{8}$$

Since the $\phi_{jj'}$ are assumed positive and bounded above, each T_j is bounded below by π_{j1} and above by $\sum_{j'} \pi_{jj'}$, both of which are positive and finite $a.s.$ (as in the original HDP). The prior means of the unnormalized transition distributions, $\boldsymbol{\pi}_j$ are then proportional (for each j) to $\alpha\boldsymbol{\beta}\boldsymbol{\phi}_j$ where $\boldsymbol{\phi}_j = (\phi_{j1}, \phi_{j2}, \dots)$.

The distribution of the latent state sequence $\mathbf{z}$ given $\boldsymbol{\pi}$ and $\boldsymbol{\ell}$ is now

$$\begin{aligned} p(\mathbf{z} \,|\, \boldsymbol{\pi}, \boldsymbol{\ell}) &= \prod_{t=1}^{T} \pi_{z_{t-1}z_t}\phi_{z_{t-1}z_t} T_{z_{t-1}}^{-n_{j\cdot}} \\ &= \prod_{j=1}^{\infty} T_j^{-1} \prod_{j'=1}^{\infty} \pi_{jj'}^{n_{jj'}} \phi_{jj'}^{n_{jj'}} \end{aligned} \tag{9}$$

where $n_{jj'} = \sum_{t=1}^{T} I(z_{t-1} = j, z_t = j')$ is the number of transitions from state j to state j' in the sequence $\mathbf{z}$ and $n_{j\cdot} = \sum_{j'} n_{jj'}$ is the total number of visits to state j. Due to the fact that T_j is a sum over products of $\pi_{jj'}$ and $\phi_{jj'}$ terms, the posterior for $\boldsymbol{\pi}$ is no longer a DP. However, conditional conjugacy can be restored by a data-augmentation process with a natural interpretation, which is described next.

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

In this section, we define a stochastic process that we call the Markov Jump Process with Failed Transitions (MJP-FT), from which we obtain the HDP-HMM-LT by marginalizing over some of the variables. By reinstating these auxiliary variables, we obtain a simple Gibbs sampling algorithm over the full MJP-FT, which can be used to sample from the marginal posterior of the variables used by the HDP-HMM-LT.

Let $\boldsymbol{\beta}$, $\boldsymbol{\pi}$, $\boldsymbol{\ell}$ and $T_j, j = 1, 2, \dots$ be defined as in the last section. Consider a continuous-time Markov Process over the states $j = 1, 2, \dots$, and suppose that if the process makes a jump to state z_t at time τ_t, the next jump, which is to state z_{t+1}, occurs at time $\tau_t + \tilde{u}_t$, where $\tilde{u}_t \sim \mathsf{Exp}(\sum_{j'} \pi_{jj'})$, and $p(z_{t+1} = j' \,|\, z_t = j) \propto \pi_{jj'}$, independent of $\tilde{u}_t$. Note that in this formulation, unlike in standard formulations of Markov Jump Processes, we are assuming that self-jumps are possible.

If we only observe the jump sequence $\mathbf{z}$ and not the holding times, $\tilde{u}_t$, this is an ordinary Markov chain with transition matrix row-proportional to $\boldsymbol{\pi}$. 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 whose transition matrix is obtained by normalizing $\boldsymbol{\pi}$; that is, we have the HDP-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 $\phi_{jj'}$. Assuming successes are independent, then the rate of successful jumps from j to j' is $\pi_{jj'}\phi_{jj'}$, and the corresponding rate of unsuccessful jump attempts is $\pi_{jj'}(1 - \phi_{jj'})$ (see Supplementary Materials for a proof). The probability that the first successful jump is to state j' (that is, that $z_{t+1} = j'$) is proportional to the rate of successful jump attempts to j', which is $\pi_{jj'}\phi_{jj'}$. Conditioned on z_t, the holding time, $\tilde{u}_t$, is independent of z_{t+1} and is distributed $\mathsf{Exp}(T_{z_t})$. The total time spent in state j given that it is visited n_j times, is then

$$u_j \,|\, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta} \overset{ind}{\sim} \mathsf{Gamma}(n_{j\cdot}, T_j) \tag{10}$$

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

$$p(\mathbf{z}, \mathbf{u}, \mathbf{Q} \,|\, \boldsymbol{\pi}) = p(\mathbf{z} \,|\, \boldsymbol{\pi}) p(\mathbf{u} \,|\, \mathbf{z}, \boldsymbol{\pi}) p(\mathbf{Q} \,|\, \mathbf{u}, \boldsymbol{\pi}) \tag{11}$$

where dependence on $\boldsymbol{\ell}$ has been omitted for conciseness. After grouping terms, this is

$$\prod_{j} \prod_{j'} \pi_{jj'}^{n_{jj'} + q_{jj'}} \phi_{jj'}^{n_{jj'}} (1 - \phi_{jj'})^{q_{jj'}} e^{-\pi_{jj'}u_j} \tag{12}$$

Conveniently, the T_j have canceled, and the exponential terms involving $\pi_{jj'}$ and $\phi_{jj'}$ in the Gamma and Poisson distributions of u_j and $q_{jj'}$ combine to cause $\phi_{jj'}$ to vanish.

2.4. Sticky and Semi-Markov Generalizations

We note that it is trivial to employ the local transition property of the HDP-HMM-LT together with the Sticky property of the Sticky HDP-HMM (Fox et al., 2008), or the non-geometric duration distributions of the HDP-HSMM (Johnson & Willsky, 2013), should an application call for additional prior weight to be placed on self-transitions. In the former case, no changes to inference are needed; one can simply add the the extra mass κ to the shape parameter of the Gamma prior on the π_{jj}, and employ the same auxiliary variable method used by Fox et al. to distinguish "Sticky"

from "regular" self-transitions. For the semi-Markov case, we can simply fix the diagonal elements of $\boldsymbol{\pi}$ to zero, and allow D_t observations to be emitted $i.i.d.$ according to a state-specific duration distribution, and sample the latent state sequence using a message passing algorithm suited for HSMMs (Johnson & Willsky, 2013). 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, no further data-augmentation is needed, as the (continuous) durations $\mathbf{u}$ already account for the normalization of the $\boldsymbol{\pi}$.

2.5. An Infinite Factorial HDP-HMM-LT

One setting in which a local transition property is desireable is the case where the latent states indicate which in a set of hidden features is "active" at time t; that is, when the latent state is represented by a binary vector. A parametric example of such a model is the Factorial HMM (Ghahramani et al., 1997), nonparametric extensions of which, such as the infinite factorial hidden Markov Model (Gael et al., 2009) and the infinite factorial dynamic model (Valera et al., 2015), have been developed in recent years by making use of the Indian Buffet Process (Ghahramani & Griffiths, 2005) as a state prior. It would be conceptually straightforward to combine the IBP state prior with the similarity bias of the LT model, provided the chosen similarity function is uniformly bounded above on the space of infinite length binary vectors (for example, take $\phi(u, v)$ to be the exponentiated negative Hamming distance between u and v). Since the number of differences between two draws from the IBP is finite with probability 1, this yields a reasonable similarity metric.

3. Inference

We develop a Gibbs sampling algorithm based on the MJP-FT representation described in Sec. 2.3, 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 deterministic function of the unscaled transition "rate" matrix, $\boldsymbol{\pi}$, and the similarity matrix, $\boldsymbol{\phi}$. The full set of variables is partitioned into blocks: $\{\gamma, \alpha, \beta, \boldsymbol{\pi}\}$, $\{\mathbf{z}, \mathbf{u}, \mathbf{Q}, \Lambda\}$, $\{\theta, \boldsymbol{\ell}\}$, and $\{\xi\}$, where Λ represents a set of auxiliary variables that will be introduced below, θ represents the emission parameters (which may be further blocked 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

$$p(\gamma, \alpha, \boldsymbol{\beta}, \boldsymbol{\pi}) = p(\gamma)p(\alpha)p(\beta \,|\, \gamma)p(\boldsymbol{\pi} \,|\, \alpha, \beta) \tag{13}$$

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 (12), the individual $\pi_{jj'}$ are *a posteriori* independent $\mathsf{Gamma}(\alpha\beta_{j'} + n_{jj'} + q_{jj'}, 1 + u_j)$ distributed.

Sampling $\boldsymbol{\beta}$ To enable joint sampling of the latent state sequence, we employ a weak limit approximation to the HDP (Johnson & Willsky, 2013), 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 to obtain the marginal likelihood:

$$\begin{aligned} p(\beta \,|\, \gamma) &= \frac{\Gamma(\gamma/J)^J}{\Gamma(\gamma)} \prod_j \beta_j^{\frac{\gamma}{J}-1} \\ p(\mathcal{D} \,|\, \beta, \alpha) &\propto \prod_{j=1}^{J} (1+u_j)^{-\alpha} \prod_{j'} \frac{\Gamma(\alpha\beta_{j'} + n_{jj'} + q_{jj'})}{\Gamma(\alpha\beta_{j'})} \end{aligned} \tag{14}$$

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 $\mathbf{M} = \{m_{jj'}\}$, with

$$p(m_{jj'} \,|\, \beta_{j'}, \alpha, \mathcal{D}) \overset{ind}{\propto} s(n_{jj'} + q_{jj'}, m_{jj'}) \alpha^{m_{jj'}} \beta_{j'}^{m_{jj'}} \tag{15}$$

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 $m_{\cdot j'} = \sum_j m_{jj'}$ and $m_{\cdot\cdot} = \sum_{j'} m_{\cdot j'}$, the posterior for (the truncated) $\boldsymbol{\beta}$ is a Dirichlet whose jth mass parameter is $\frac{\gamma}{J} + m_{\cdot j}$.

Sampling Concentration Parameters Incorporating $\mathbf{M}$ into $\mathcal{D}$, we can integrate out $\boldsymbol{\beta}$ to obtain

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

Assume that α and γ have Gamma priors with shape and rate parameters a_α, b_α and a_γ, b_γ. Then

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

To simplify the likelihood for γ, we can introduce a final set of auxiliary variables, $\mathbf{r} = (r_1, \ldots, r_J)$, $r_{j'} \in \{0, \ldots, m_{\cdot j'}\}$ and $t \in (0, 1)$ with the following distributions:

$$p(r_{j'} = r \,|\, m_{\cdot j'}, \gamma) \propto s(m_{\cdot j'}, r) \left(\frac{\gamma}{J}\right)^r \quad (18)$$

$$p(t \,|\, m_{\cdot\cdot} \gamma) \propto t^{\gamma - 1}(1 - t)^{m_{\cdot\cdot} - 1} \quad (19)$$

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

$$\gamma \,|\, \mathcal{D}, \mathbf{r}, t \sim \mathsf{Gamma}(a_\gamma + r_{\cdot}, b_\gamma - \log(t)) \quad (20)$$

3.2. Sampling 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:

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

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 locations $\boldsymbol{\ell}$ may be based on the emission parameters, $\boldsymbol{\theta}$, or may be independent. The experiments described below illustrate both cases. In general, we have

$$p(\mathbf{z}, \mathbf{Q} \,|\, \boldsymbol{\ell}) \propto \prod_j \prod_{j'} \phi_{jj'}^{n_{jj'}} (1 - \phi_{jj'})^{q_{jj'}} \quad (21)$$

$$p(\mathbf{Y} \,|\, \mathbf{z}, \boldsymbol{\theta}) = \prod_{t=1}^{T} f(\mathbf{y}_t; \theta_{z_t}) \quad (22)$$

where $\mathbf{Y}$ is the matrix with rows $\mathbf{y}_1, \ldots, \mathbf{y}_T$, and proportionality in the first equation is as a function of $\boldsymbol{\ell}$.

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 two cases. The first is a speaker-diarization task, where each state consists of a finite D-dimensional binary vector whose entries indicate which speakers are currently speaking. In this experiment, the state vectors both determine the pairwise similarities and partially determine the emission distributions via a linear-Gaussian model. In the second experiment, the data consists of Bach chorales, and the latent states can be thought of as harmonic contexts. There, the components of the states that govern similarities are modeled as independent of the emission distributions, which are categorical distributions over four-voice chords.

4.1. Cocktail Party

The data in this experiment consists of signals generated by D speakers being picked up by K microphones over each of T discrete time windows. Each latent state, $\boldsymbol{\eta}_j$, is the D-dimensional binary vector whose dth entry indicates whether or not speaker d is speaking. The locations ℓ_j are identified with the binary vectors, $\ell_j := \boldsymbol{\eta}_j$. We use a Laplacian similarity defined using Hamming distance, d_0: $\phi_{jj'} = \exp(-\lambda d_0(\boldsymbol{\eta}_j, \boldsymbol{\eta}_{j'})), \lambda \geq 0$. The emission distribution is linear-Gaussian, with $D \times K$ weight matrix $\mathbf{W}$, so that $\mathbf{y}_t \,|\, z_t \sim \mathcal{N}(\mathbf{W}^\mathsf{T} \boldsymbol{\eta}_{z_t}, \boldsymbol{\Sigma})$. For the experiments discussed here, we will assume that $\boldsymbol{\Sigma}$ 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 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.

The Data The data was constructed using audio signals from the PASCAL CHiME Speech Separation and Recognition Challenge. The underlying signal consisted of 16 simultaneous speakers, with the signal matrix linearly mapped to 12 microphone channels. The 16 speakers were grouped into 4 conversational groups of 4 speakers each, where speakers took turns speaking within conversation. In such a task, there are naively 2^D possible states (here, 65536). However, due to the conversational grouping, if at most one speaker in a conversation is speaking at any given time, 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 (in this case $s_c \equiv 4$ for all c, for a total of 625).

Each "turn" within a conversation consisted of a single sentence (average duration 2s) and turn orders within a conversation were randomly generated, with random pauses inserted between sentences. Each conversation consisted of a minimum of 20 sentences, and the signal was

down-sampled to length 2000. The 'on' portions of each speaker's signal were normalized to have a mean of 1 and a standard deviation of 0.5. An additional column of 1s was added to the speaker signal matrix, representing background noise. The resulting signal matrix was thus 2000×17 and the weight matrix was 17×12. Following Gael et al. (2009) and Valera et al. (2015), the weights were drawn independently from a $\mathsf{Unif}(0, 1)$ distribution, and independent $\mathcal{N}(0, 0.09)$ noise was added to each entry of the observation matrix.

Sampling $\boldsymbol{\eta}$ We put independent Beta-Bernoulli priors on each coordinate of $\boldsymbol{\eta}$, the matrix whose jth row is $\boldsymbol{\eta}_j$. We Gibbs sample each binary coordinate η_{jd} conditioned on all the others and the coordinate-wise prior means, $\{\mu_d\}$, which we sample in turn conditioned on $\boldsymbol{\eta}$. Details are in the supplement.

Sampling λ The λ parameter of the similarity function governs the connection between $\boldsymbol{\ell}$ and $\boldsymbol{\phi}$. Writing (21) in terms of λ and the pairwise distances, $d_{jj'} = d_0(\eta_j, \eta_{j'})$ gives the likelihood

$$p(\mathbf{z}, \mathbf{Q} \,|\, \boldsymbol{\eta}, \lambda) \propto \prod_j \prod_{j'} e^{-\lambda d_{jj'} n_{jj'}} (1 - e^{-\lambda d_{jj'}})^{q_{jj'}} \quad (23)$$

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

$$p(\lambda \,|\, \mathbf{z}, \mathbf{Q}, \boldsymbol{\eta}) \propto e^{-(b_\lambda + \sum_j \sum_{j'} d_{jj'} n_{jj'})\lambda} \times \prod_j \prod_{j'} (1 - e^{-\lambda d_{jj'}})^{q_{jj'}} \quad (24)$$

This density is log-concave, and so we use Adaptive Rejection Sampling (Gilks & Wild, 1992) to sample from it.

Sampling $\mathbf{W}$ and $\boldsymbol{\Sigma}$ Conditioned on the state matrix $\boldsymbol{\eta}$, and the weight matrix $\mathbf{W}$, $\mathbf{Y}$ is $\mathcal{N}(\mathbf{W}^\mathsf{T}\boldsymbol{\eta}^*, \boldsymbol{\Sigma})$ (where $\boldsymbol{\eta}^*$ is the $T \times D$ matrix whose tth row is $\boldsymbol{\eta}_{z_t}$), and so conditioned on $\mathbf{Y}$, and $\boldsymbol{\eta}^*$, $\mathbf{W}$ and $\boldsymbol{\Sigma}$ can be sampled using standard methods for Bayesian linear regression. If a Normal prior is used for each column of $\mathbf{W}$, then the posterior is Normal. For the experiments reported below, we fix $\mathbf{W}$ to the ground truth matrix so that the entries of $\boldsymbol{\eta}$ are identifiable with the ground truth matrix, and we constrain $\boldsymbol{\Sigma}$ to be diagonal, with an Inverse Gamma prior on the variances, resulting in conjugate updates.

Results We attempted to infer the states from the data using five 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, (3) a Sticky HDP-HMM, (4) our HDP-HMM-LT model, and (5) a model that combines the Sticky and LT properties. We evaluated the models at each iteration using both the Hamming distance between inferred and ground truth state matrices and F1 score. The results for the three models are in Figure 4.1. For models (2)-(5) we also computed marginal likelihoods using both the training sequence and a second test sequence, where $\mathbf{z}$ was integrated out. The LT and Sticky-LT models outperform the others on all measures, while the regular Sticky model exhibits only a small advantage over the vanilla HDP-HMM. We also plot the inferred decay rate λ and the number of states used by the LT and Sticky-LT models. Both settle on a non-negligible λ values, suggesting that the local transition structure explains the data well.

In Fig 4.1, we also plot the ground truth state matrix against the average state matrix, $\boldsymbol{\eta}^*$, averaged over runs and iterations after the first 1000.

These models also uses more components than the non-LT models, 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.

4.2. Synthetic Data Without Local Transitions

We also generated data directly from the ordinary HDP-HMM used in the cocktail experiment as a sanity check, to examine the performance of the LT model in the absence of a similarity bias. The results are in Fig. 4.2. When the λ parameter is large, the LT model has worse performance than the non-LT model on this data; 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.

4.3. Bach Chorales

To test a version of the HDP-HMM-LT model in which the components of the latent state governing similarity are unrelated to the emission distributions, we used our model to do unsupervised "grammar" learning from a corpus of Bach chorales. The data was a corpus of 217 four-voice major key chorales by J.S. Bach, 200 of which were randomly selected as a training set, with the other 17 used as a test set to evaluate surprisal (marginal log likelihood per observation) by the trained models. All chorales were transposed to a common key, and each distinct four-voice chord (with voices ordered) was encoded as a single integer. In total there were 3307 distinct chord types and 20401 chord tokens in the 217 chorales, with 3165 types and 18818 tokens in the 200 training chorales, and 143

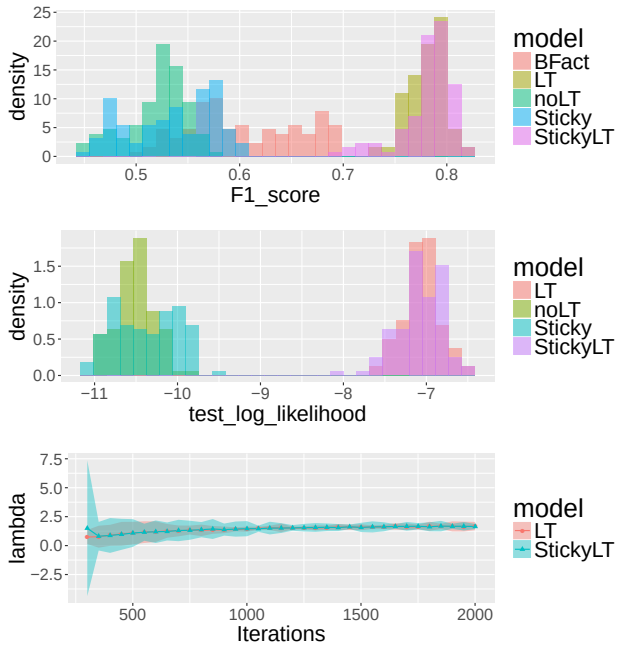

Figure 1. (Top) F1 score for inferred relative to ground truth binary speaker matrices on cocktail party data, using every 50th Gibbs iteration between 1000-2000, aggregating across 5 runs of each model. (Middle) Log likelihood on a held out test set. (Bottom) Learned similarity parameter, λ, for the LT and Sticky-LT models by Gibbs iteration, averaged over 5 runs. Error bands are 99% confidence interval of the mean.

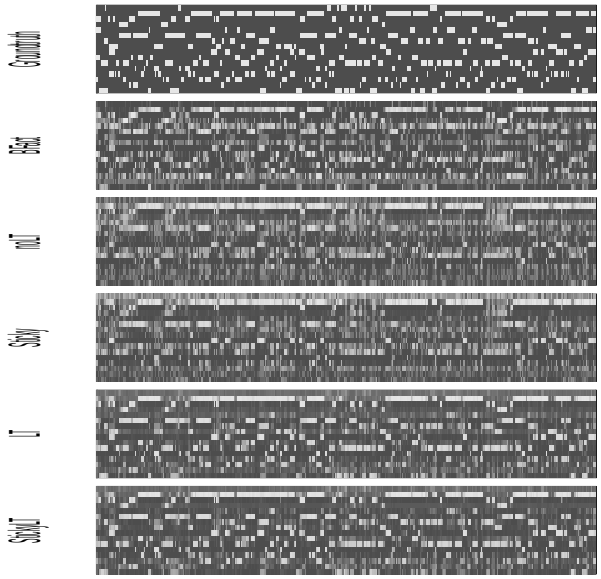

Figure 2. The ground binary speaker matrix for the cocktail party data is at the top, followed by the inferred speaker matrix for the binary factorial, "vanilla" HDP-HMM, Sticky-HDP-HMM, HDP-HMM-LT, and Sticky HDP-HMM-LT. All inferred matrices are averaged over 5 runs and every 50 iteration after the first 1000.

chord types that were unique to the test set.

Since the chords were encoded as integers, the emission distribution for each state is $\mathsf{Cat}(\boldsymbol{\theta}_j)$. We use a 3307-dimensional symmetric Dirichlet prior for each $\boldsymbol{\theta}_j$, resulting in conjugate updates to $\boldsymbol{\theta}$ conditioned on the latent state sequence, $\mathbf{z}$.

Modifications to Model and Inference In this experiment, the locations, $\boldsymbol{\ell}_j$, are abstract and independent of the $\boldsymbol{\theta}_j$. We use a bivariate $\mathcal{N}(0, \mathbf{I})$ prior on each $\boldsymbol{\ell}_j$, and a squared exponential similarity function based on Euclidean distance: $\phi_{jj} = \exp\{-\lambda d_2(\boldsymbol{\ell}_j, \boldsymbol{\ell}_{j'})^2\}$ where d_2 is Euclidean distance. Since the latent states are now continuous, we use a Hamiltonian Monte Carlo (HMC) update (Duane et al., 1987; Neal et al., 2011) to update of the $\boldsymbol{\ell}_j$ simultaneously, conditioned on $\mathbf{z}$ and $\boldsymbol{\pi}$. HMC is a variation on Metropolis-Hastings algorithm which is designed to more efficiently explore a high-dimensional continuous distribution by adopting a proposal distribution which incorporates an auxiliary "momentum" variable to make it more likely that proposals will go in useful directions and improve mixing compared to naive movement. The HMC update requires the gradient of the log (conditional) posterior. The relevant prior and likelihood are

$$\begin{aligned} p(\boldsymbol{\ell}_j) &\propto -\frac{1}{2}\,||\ell_j||^2 \\ p(\mathbf{z}, \mathbf{u}, \mathbf{Q} \,|\, \boldsymbol{\ell}) &\propto \prod_j \prod_{j'} \phi_{jj'}^{n_{jj'}} (1-\phi_{jj'})^{q_{jj'}}. \end{aligned} \tag{25}$$

The coordinate of the gradient of the log likelihood corresponding to dimension d in state j is

$$\frac{\partial \log L}{\partial \boldsymbol{\ell}_{jd}} = -\lambda \sum_{(j,j'):j\neq j'} (\ell_{jd} - \ell_{j'd}) \left(n_{jj'} - q_{jj'} \frac{\phi_{jj'}}{1-\phi_{jj'}} \right)$$

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

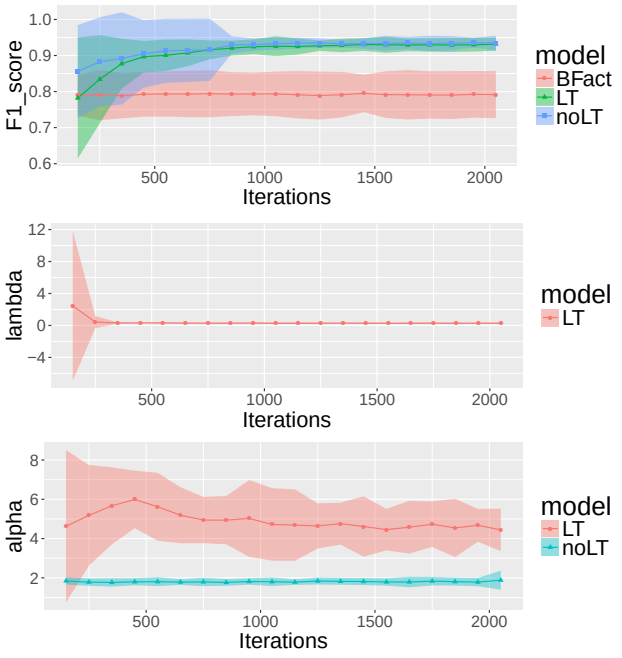

Figure 3. (Top) F1 score for inferred relative to ground truth binary speaker matrices on synthetic data generated from the vanilla HDP-HMM model. (Middle) Learned similarity parameter, λ, for the LT model by Gibbs iteration, averaged over 5 runs. Error bands are 99% confidence interval of the mean. (Bottom) Concentration parameter α for the HDP-HMM-LT and HDP-HMM models.

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 (Paisley et al., 2012)). 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 (Gael et al., 2009) and the Infinite Factorial Dynamic Model (Valera et al., 2015), both of which use the Indian Buffet Process (Ghahramani & Griffiths, 2005) 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

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

Duane, Simon, Kennedy, Anthony D, Pendleton, Brian J, and Roweth, Duncan. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

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

Ferguson, Thomas S. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, pp. 209–230, 1973.

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

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

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

Ghahramani, Zoubin, Jordan, Michael I, and Smyth, Padhraic. Factorial hidden Markov models. *Machine learning*, 29(2-3):245–273, 1997.

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

Ishwaran, Hemant and Zarepour, Mahmoud. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

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

Kingman, John. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

Neal, Radford M et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

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

Ranganath, Rajesh and Blei, David M. Correlated random measures. *Journal of the American Statistical Association*, in press.

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

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