An Infinite Hidden Markov Model With Similarity-Biased Transitions

Colin Reimer Dawson [1] **Chaofan Huang** [1] **Clayton T. Morrison** [2]

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 "nearby" states. This is accomplished by defining a similarity function on the state space and scaling transition probabilities by pairwise similarities, thereby inducing correlations among the transition distributions. 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. This augmentation restores conditional conjugacy and admits a simple Gibbs sampler. 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 hierarchical Dirichlet process hidden Markov model (HDP-HMM) (Beal et al., 2001; Teh et al., 2006) is a Bayesian model for time series data that generalizes the conventional hidden Markov Model to allow a countably infinite state space. The hierarchical structure ensures that, despite the infinite state space, a common set of destination states will be reachable with positive probability from each source state. The HDP-HMM can be characterized by the following generative process.

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 (Ewens, 1990; Sethuraman, 1994) for Griffiths, Engen and McCloskey. We thus have a discrete probability measure, G_0, with weights β_j at locations θ_j, $j = 1, 2, \dots$, defined by

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

G_0 drawn in this way is a Dirichlet Process (DP) random measure with concentration γ and base measure H.

The actual transition distribution, $\boldsymbol{\pi}_j$, from state j, is drawn from another DP with concentration α and base measure G_0:

$$\boldsymbol{\pi}_j \overset{\text{i.i.d.}}{\sim} \mathsf{DP}(\alpha G_0) \qquad j = 0, 1, 2, \dots \tag{2}$$

where $\boldsymbol{\pi}_0$ represents the initial distribution. The hidden state sequence, $z_1, z_2, \dots z_T$ is then generated according to $z_1 \mid \boldsymbol{\pi}_0 \sim \mathsf{Cat}(\boldsymbol{\pi}_0)$, and

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

Finally, the emission distribution for state j is a function of θ_j, so that observation y_t is drawn according to

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

A shortcoming of the HDP prior on the transition matrix is that it does not use the fact that the source and destination states are the same set: that is, each $\boldsymbol{\pi}_j$ has a special element which corresponds to a self-transition. In the HDP-HMM, however, 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 κ at location j to the base measure of the DP that generates $\boldsymbol{\pi}_j$. That is, (2) is replaced by

$$\boldsymbol{\pi}_j \sim \mathsf{DP}(\alpha G_0 + \kappa \delta_{\theta_j}). \tag{5}$$

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. However, while both of these

[1] Oberlin College, Oberlin, OH, USA [2] The University of Arizona, Tucson, AZ, USA. Correspondence to: Colin Reimer Dawson <cdawson@oberlin.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

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 two main contributions of this paper are (1) a generalization of the HDP-HMM, which we call the HDP-HMM with local transitions (HDP-HMM-LT) that allows for a geometric structure to be defined on the latent state space, so that "nearby" states are *a priori* more likely to have transitions between them, and (2) a simple Gibbs sampling algorithm for this model. The "LT" property is introduced by elementwise rescaling and then renormalizing of the HDP transition matrix. 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, 2016), 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. Zhu et al. (2016) developed an HMM based directly on the DILN model[1]. Both Paisley et al. and Zhu et al. employ variational approximations, whereas we present a Gibbs sampler, which converges asymptotically to the true posterior. We discuss additional differences between our model and the DILN-HMM in Sec. 2.2.

One class of application in which it is useful to incorporate a notion of locality occurs when the latent state sequence consists of several parallel chains, so that the global state changes incrementally, but where these increments are not independent across chains. Factorial HMMs (Ghahramani et al., 1997) are commonly used in this setting, but this ignores dependence among chains, and hence may do poorly when some combinations of states are much more probable than suggested by the chain-wise dynamics.

Another setting where the LT property is useful is when there is a notion of state geometry that licenses syllogisms: e.g., if A frequently leads to B and C and B frequently leads to D and E, then it may be sensible to infer that A and C may lead to D and E as well. This property is arguably present in musical harmony, where consecutive chords are often (near-)neighbors in the "circle of fifths", and small steps along the circle are more common than large ones.

The paper is structured as follows: In section 2 we define the model. In section 3, we develop a 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. Code and additional details are available at `http://colindawson.net/hdp-hmm-lt/`

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), that can be defined as $G = \sum_{k=1}^{\infty} \tilde{\pi}_k \delta_{\theta_k}$, where

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

δ_θ is a measure assigning 1 to sets if they contain θ and 0 otherwise, and subject to the constraint that $\sum_{k\geq 1} \beta_k = 1$ and $0 < \alpha < \infty$. It has 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. sequence of random measures, $\{G_j\}, j = 1, 2, \dots$, from the above process, 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

[1] We thank an anonymous ICML reviewer for bringing this paper to our attention.

with the jth random measure, and T_j is the sum of row j, then we obtain the prior for the HDP-HMM, where

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

2.2. Promoting "Local" Transitions

In the HDP prior, the rows of the transition matrix are conditionally independent. We wish to relax this assumption, to incorporate possible prior knowledge that certain pairs of states are "nearby" in some sense and thus more likely than others to produce large transition weights between them (in both directions); that is, transitions are likely to be "local". We accomplish this by associating each latent state j with a location ℓ_j in some space Ω, introducing a "similarity function" $\phi : \Omega \times \Omega \to (0, 1]$, and scaling each element $\pi_{jj'}$ by $\phi_{jj'} = \phi(\ell_j, \ell_{j'})$. For example, we might wish to define a (possibly asymmetric) divergence function $d : \Omega \times \Omega \to [0, \infty)$ and set $\phi(\ell_j, \ell_j) = \exp\{-d(\ell_j, \ell_{j'})\}$ so that transitions are less likely the farther apart two states are. By setting $\phi \equiv 1$, we obtain the standard HDP-HMM. The DILN-HMM (Zhu et al., 2016), employs a similar rescaling of transition probabilities via an exponentiated Gaussian Process, following (Paisley et al., 2012), but the scaling function must be positive semi-definite, and in particular symmetric, whereas in the HDP-HMM-LT, ϕ need only take values in $(0, 1]$. Moreover, the DILN-HMM does not allow the scales to be tied to other state parameters, and hence encode an independent notion of similarity.

Letting $\boldsymbol{\ell} = (\ell_1, \ell_2, \dots)$, we can replace (6) for $j \geq 1$ by

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

Since the $\phi_{jj'}$ are positive and bounded above by 1,

$$0 < \pi_{j1}\phi_{j1} \leq T_j \leq \sum_{j'} \pi_{jj'} < \infty \tag{9}$$

almost surely, where the last inequality carries over from 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

$$\begin{aligned} p(\mathbf{z} \mid \boldsymbol{\pi}, \boldsymbol{\ell}) &= \prod_{t=1}^{T} \pi_{z_{t-1}z_t}\phi_{z_{t-1}z_t} T_{z_{t-1}}^{-n_{z_{t-1}\cdot}} \\ &= \prod_{j=1}^{\infty} T_j^{-1} \prod_{j'=1}^{\infty} \pi_{jj'}^{n_{jj'}} \phi_{jj'}^{n_{jj'}} \end{aligned} \tag{10}$$

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. Since 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 $p(z_{t+1} = j' \mid z_t = j) \propto \pi_{jj'}$, 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 directly, 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 each jump attempt from state j to state j' has probability $(1 - \phi_{jj'})$ of failing, in which case no transition occurs and no observation is generated. Assuming independent failures, the rates of successful and failed jumps from j to j' are $\pi_{jj'}\phi_{jj'}$ and $\pi_{jj'}(1 - \phi_{jj'})$, respectively. 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 as $\mathsf{Exp}(T_{z_t})$. We denote the total time spent in state j by $u_j = \sum_{t:z_t=j} \tilde{u}_t$, where, as the sum of i.i.d. Exponentials,

$$u_j \mid \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta} \overset{\text{ind.}}{\sim} \mathsf{Gamma}(n_{j\cdot}, T_j) \tag{11}$$

During this period there will be $q_{jj'}$ failed attempts to jump to state j', where $q_{jj'} \sim \mathsf{Poisson}(u_j\pi_{jj'}(1 - \phi_{jj'}))$ are independent. This data augmentation bears some conceptual similarity to the Geometrically distributed ρ auxiliary variables introduced to the HDP-HSMM (Johnson & Willsky,

2013) to restore conditional conjugacy. However, there are key differences: first, ρ measure how many steps the chain would have remained in state j under Markovian dynamics, whereas our u represents putative continuous holding times between each transition, and second ρ allows for the restoration of a zeroed out entry in each row, whereas u allows us to work with unnormalized π entries, avoiding the need to restore zeroed out entries in the HSMM-LT

Incorporating $\mathbf{u} = \{u_j\}$ and $\mathbf{Q} = \{q_{jj'}\}$ as augmented data simplifies the likelihood for $\boldsymbol{\pi}$, yielding

$$p(\mathbf{z}, \mathbf{u}, \mathbf{Q} \,|\, \boldsymbol{\pi}) = p(\mathbf{z} \,|\, \boldsymbol{\pi}) p(\mathbf{u} \,|\, \mathbf{z}, \boldsymbol{\pi}) p(\mathbf{Q} \,|\, \mathbf{u}, \boldsymbol{\pi}) \qquad (12)$$

where dependence on $\boldsymbol{\ell}$ has been omitted for conciseness. After grouping terms and omitting terms that do not depend on $\boldsymbol{\pi}$, this proportional (as a function of $\boldsymbol{\pi}$) to

$$\prod_j \prod_{j'} \pi_{jj'}^{n_{jj'} + q_{jj'}} \phi_{jj'}^{n_{jj'}} (1 - \phi_{jj'})^{q_{jj'}} e^{-\pi_{jj'} u_j} \qquad (13)$$

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 the local transition property of the HDP-HMM-LT can be combined 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), to add additional prior weight 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 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 suitable semi-Markov message passing algorithm (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. Obtaining the Factorial HMM as a Limiting Case

One setting in which a local transition property is desirable is the case where the latent states encode multiple hidden features at time t as a vector of categories. Such problems are often modeled using factorial HMMs (Ghahramani et al., 1997). In fact, the HDP-HMM-LT yields the factorial HMM in the limit as $\alpha, \gamma \to \infty$, fixing each row of π to be uniform with probability 1, so the dynamics are controlled entirely by ϕ. If $\mathbf{A}^{(d)}$ is the transition matrix for chain d, then setting $\phi(\boldsymbol{\ell}_j, \boldsymbol{\ell}_{j'}) = \exp - d(\boldsymbol{\ell}_j, \boldsymbol{\ell}_{j'})$ with asymmetric "divergences" $d(\boldsymbol{\ell}_j, \boldsymbol{\ell}_{j'}) = -\sum_d \log(\mathbf{A}^{(d)}_{\ell_{jd}, \ell_{j'd}})$ yields the factorial transition model.

2.6. An Infinite Factorial HDP-HMM-LT

Nonparametric extensions of the factorial HMM, 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 augmented data $\mathcal{D} = (\mathbf{z}, \mathbf{u}, \mathbf{Q}, \Lambda)$ will factor as

$$\begin{aligned} p(\gamma, \alpha, \boldsymbol{\beta}, \boldsymbol{\pi} \,|\, \mathcal{D}) \\ = p(\gamma \,|\, \mathcal{D}) p(\alpha \,|\, \mathcal{D}) p(\beta \,|\, \gamma, \mathcal{D}) p(\boldsymbol{\pi} \,|\, \alpha, \beta, \mathcal{D}) \end{aligned} \qquad (14)$$

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 (13), 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 $\mathbf{z}$, 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 J components, where J 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:

$$p(\beta \,|\, \gamma) = \frac{\Gamma(\gamma/J)^J}{\Gamma(\gamma)} \prod_j \beta_j^{\frac{\gamma}{J}-1} \tag{15}$$

$$p(\mathcal{D} \,|\, \beta, \alpha) \propto \prod_{j=1}^{J} (1+u_j)^{-\alpha} \prod_{j'} \frac{\Gamma(\alpha\beta_{j'} + n_{jj'} + q_{jj'})}{\Gamma(\alpha\beta_{j'})}$$

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

$$p(m_{jj'} \,|\, \beta_{j'}, \alpha, \mathcal{D}) \overset{ind}{\propto} s_{n_{jj'}+q_{jj'}, m_{jj'}} \alpha^{m_{jj'}} \beta_{j'}^{m_{jj'}} \tag{16}$$

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 M into $\mathcal{D}$, we can integrate out $\boldsymbol{\beta}$ to obtain

$$p(\mathcal{D} \,|\, \alpha, \gamma) \propto \alpha^{m_{\cdot\cdot}} e^{-\sum_{j''} \log(1+u_{j''})\alpha} \\ \frac{\Gamma(\gamma)}{\Gamma(\gamma + m_{\cdot\cdot})} \times \prod_j \frac{\Gamma(\frac{\gamma}{J} + m_{\cdot j})}{\Gamma(\frac{\gamma}{J})} \tag{17}$$

Assuming that α and γ have Gamma priors with shape and rate parameters a_α, b_α and a_γ, b_γ, then

$$\alpha \,|\, \mathcal{D} \sim \mathsf{Gamma}(a_\alpha + m_{\cdot\cdot}, b_\alpha + \sum_j \log(1+u_j)). \tag{18}$$

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 $w \in (0,1)$ with the following distributions:

$$p(r_{j'} = r \,|\, m_{\cdot j'}, \gamma) \propto s(m_{\cdot j'}, r) \left(\frac{\gamma}{J}\right)^r \tag{19}$$

$$p(w \,|\, m_{\cdot\cdot}\gamma) \propto w^{\gamma-1}(1-w)^{m_{\cdot\cdot}-1} \tag{20}$$

The normalizing constants are ratios of Gamma functions, which cancel those in (17), so that

$$\gamma \,|\, \mathcal{D}, \mathbf{r}, w \sim \mathsf{Gamma}(a_\gamma + r_{\cdot}, b_\gamma - \log(w)) \tag{21}$$

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 w. The joint conditional distribution of these variables is defined directly by the generative model:

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

Since we are conditioning on the transition matrix, we can sample the entire sequence $\mathbf{z}$ jointly with the forward-backward algorithm, as in an ordinary HMM. Since we are sampling the labels jointly, this step requires $\mathcal{O}(TJ^2)$ computation per iteration, which is the bottleneck of the inference algorithm for reasonably large T or J (other updates are constant in T or in J). Having done this, we can sample $\mathbf{u}$, $\mathbf{Q}$, $\mathbf{M}$, $\mathbf{r}$ and w from their forward distributions. It is also possible to employ a variant on beam sampling (Van Gael et al., 2008) to speed up each iteration, at the cost of slower mixing, but we did not use this variant here.

3.3. Sampling state and emission parameters

Depending on the application, the locations $\boldsymbol{\ell}$ may or may not depend on the emission parameters, $\boldsymbol{\theta}$. If not, sampling $\boldsymbol{\theta}$ conditional on $\mathbf{z}$ is unchanged from the HDP-HMM. There is no general-purpose method for sampling $\boldsymbol{\ell}$, or for sampling $\boldsymbol{\theta}$ in the dependent case, due to the dependence on the form of ϕ and on the emission model, but specific instances are illustrated in the experiments below.

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 The data was constructed using audio signals collected from the PASCAL 1st Speech Separation Challenge[2]. The underlying signal consisted of $D = 16$ speaker channels recorded at each of $T = 2000$ time steps, with the

[2] http://laslab.org/SpeechSeparationChallenge/

resulting $T \times D$ signal matrix, denoted by $\boldsymbol{\theta}^*$, mapped to $K = 12$ microphone channels via a weight matrix, $\mathbf{W}$. The 16 speakers were grouped into 4 conversational groups of 4, where speakers within a conversation took turns speaking (see Fig. 2). 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 a total of 625 possible states).

Each "turn" within a conversation consisted of a single sentence (average duration $\sim$ 3s) and turn orders within a conversation were randomly generated, with random pauses distributed as $\mathcal{N}(1/4s, (1/4s)^2)$ inserted between sentences. Every time a speaker has a turn, the sentence is drawn randomly from the 500 sentences uttered by that speaker in the data. The conversations continued for 40s, and the signal was down-sampled to length 2000. The 'on' portions of each speaker's signal were normalized to have amplitudes with mean 1 and standard deviation 0.5. An additional column of 1s was added to the speaker signal matrix, $\boldsymbol{\theta}^*$, representing background noise. The resulting signal matrix, denoted $\boldsymbol{\theta}^*$, was thus 2000×17 and the weight matrix, $\mathbf{W}$, 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 $\mathcal{N}(0, 0.3^2)$ noise was added to each entry of the observation matrix.

The Model The latent states, $\boldsymbol{\theta}_j$, are the D-dimensional binary vectors whose dth entry indicates whether or not speaker d is speaking. The locations $\boldsymbol{\ell}_j$ are identified with the binary vectors, $\boldsymbol{\ell}_j := \boldsymbol{\theta}_j$. We use a Laplacian similarity function on Hamming distance, d_0, so that $\phi_{jj'} := \exp(-\lambda d_0(\boldsymbol{\ell}_j, \boldsymbol{\ell}_{j'})), \lambda \geq 0$. The emission model is linear-Gaussian as in the data, with $(D + 1) \times K$ weight matrix $\mathbf{W}$, and $T \times (D + 1)$ signal matrix $\boldsymbol{\theta}^*$ whose t^{th} row is $\boldsymbol{\theta}_t := (1, \boldsymbol{\theta}_{z_t})$, so that $\mathbf{y}_t \mid \mathbf{z} \sim \mathcal{N}(\mathbf{W}^\top \boldsymbol{\theta}^*_t, \boldsymbol{\Sigma})$. For the experiments discussed here, we assume that $\boldsymbol{\Sigma}$ is independent of j, but this assumption is easily relaxed if appropriate.

For finite-length binary vector states, the set of possible states is finite, and so it may seem that a nonparametric model is unnecessary. However, if D is reasonably large, likely 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 to encourage 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. Next we describe the additional inference steps needed for this version of the model.

Sampling $\boldsymbol{\theta}$ / $\boldsymbol{\ell}$ Since $\boldsymbol{\theta}_j$ and $\boldsymbol{\ell}_j$ are identified, influencing both the transition matrix and the emission distributions, both the state sequence $\mathbf{z}$ and the observation matrix $\mathbf{Y}$ are used in the update. We put independent Beta-Bernoulli priors on each coordinate of $\boldsymbol{\theta}$, and Gibbs sample each coordinate θ_{jd} conditioned on all the others and the coordinate-wise prior means, $\{\mu_d\}$, which we sample in turn conditioned on $\boldsymbol{\theta}$. Details are in the supplement.

Sampling λ The λ parameter of the similarity function governs the connection between $\boldsymbol{\ell}$ and $\boldsymbol{\phi}$. Substituting the definition of $\boldsymbol{\phi}$ into (13) yields

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

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

$$\begin{aligned} p(\lambda \mid \mathbf{z}, \mathbf{Q}, \boldsymbol{\ell}) \propto\ & e^{-(b_\lambda + \sum_j \sum_{j'} d_{jj'} n_{jj'})\lambda} \quad (23) \\ & \times \prod_j \prod_{j'} (1 - e^{-\lambda d_{jj'}})^{q_{jj'}} \end{aligned}$$

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 $\mathbf{Y}$ and $\boldsymbol{\theta}^*$, $\mathbf{W}$ and $\boldsymbol{\Sigma}$ can be sampled as in Bayesian linear regression. If each column of $\mathbf{W}$ has a multivariate Normal prior, then the columns are *a posteriori* independent multivariate Normals. For the experiments reported here, we fix $\mathbf{W}$ to its ground truth value so that $\boldsymbol{\theta}^*$ can be compared directly with the ground truth signal matrix, and we constrain $\boldsymbol{\Sigma}$ to be diagonal, with Inverse Gamma priors on the variances, resulting in conjugate updates.

Results We attempted to infer the binary speaker matrices using five models: (1) a binary-state Factorial HMM (Ghahramani et al., 1997), where individual binary speaker sequences are modeled as independent, (2) an ordinary HDP-HMM without local transitions (Teh et al., 2006), where the latent states are binary vectors, (3) a Sticky HDP-HMM (Fox et al., 2008), (4) our HDP-HMM-LT model, and (5) a model that combines the Sticky and LT properties[3]. For all models, all concentration and noise precision parameters are given $\mathsf{Gamma}(0.1, 0.1)$ priors. For the Sticky models, the ratio $\frac{\kappa}{\alpha+\kappa}$ is given a $\mathsf{Unif}(0, 1)$ prior. We evaluated the models at each iteration using both the Hamming distance between inferred and ground truth state matrices and F1 score. We also plot the inferred decay rate λ, and the number of states used by the LT and Sticky-LT models. The results for the five models are in Fig. 1. In

[3] We attempted to add a comparison to the DILN-HMM (Zhu et al., 2016) as well, but code could not be obtained, and the paper did not provide enough detail to reproduce their inference algorithm.

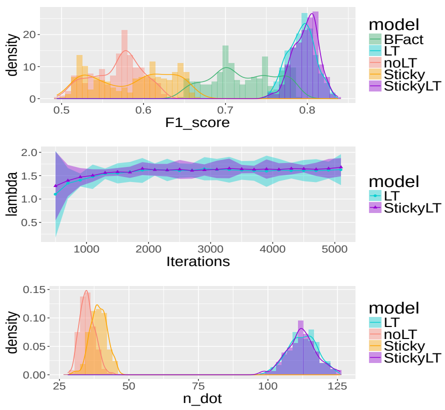

Figure 1. Top: F1 score for inferred relative to ground truth binary speaker matrices on cocktail party data, evaluated every 50th Gibbs iteration after the first 2000, aggregating across 5 runs of each model. Middle: Inferred λ, for the LT and Sticky-LT models by Gibbs iteration, averaged over 5 runs. Bottom: Number of states used, $n_{\cdot}$, by each model in the training set. Error bands are 99% confidence interval of the mean per iteration.

Fig. 2, we plot the ground truth state matrix against the average state matrix, $\boldsymbol{\eta}^*$, averaged over runs and post-burn-in iterations.

The LT and Sticky-LT models outperform the others, while the regular Sticky model exhibits only a small advantage over the vanilla HDP-HMM. Both converge on a non-negligible λ value of about 1.6 (see Fig. 1), suggesting that the local transition structure explains the data well. The LT models also use more states 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 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. 3. 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. The LT model does not make entirely the same inferences as the non-LT model, however; in particular, the α concentration parameter is larger. To some extent, α and λ trade off: sparsity of the transition matrix can be achieved either by beginning with a sparse rate matrix prior to rescaling (α small), or by beginning with a less sparse rate matrix which becomes sparser through rescaling (larger α and non-zero λ).

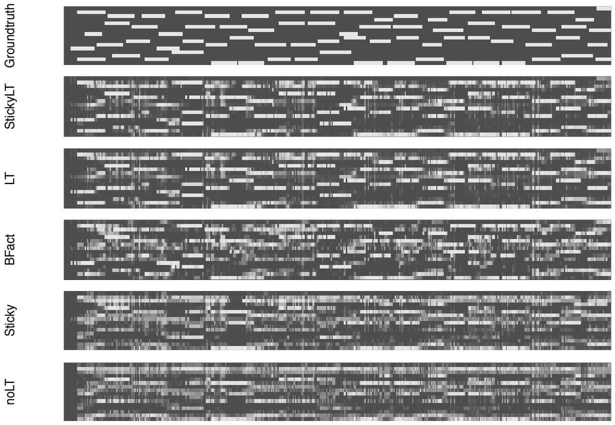

Figure 2. Binary speaker matrices for the cocktail data, with time on the horizontal axis and speaker on the vertical axis. White is 1, black is 0. The ground truth matrix is at the top, followed by the inferred speaker matrix for the Sticky HDP-HMM-LT, HDP-HMM-LT, binary factorial, Sticky-HDP-HMM, and "vanilla" HDP-HMM. All inferred matrices are averaged over 5 runs of 5000 Gibbs iterations each, with the first 2000 iterations discarded as burn-in.

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 from music21[4], 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 C-major, 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 chord types that were unique to the test set.

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

In this experiment, the locations, $\boldsymbol{\ell}_j$, are independent of the $\boldsymbol{\theta}_j$, with $\mathcal{N}(0, \mathbf{I})$ priors. We use a Gaussian similarity function, $\phi_{jj} := \exp\{-\lambda d_2(\boldsymbol{\ell}_j, \boldsymbol{\ell}_{j'})^2\}$ where d_2 is Euclidean distance. Since the latent states are continuous, we use

[4] http://web.mit.edu/music21

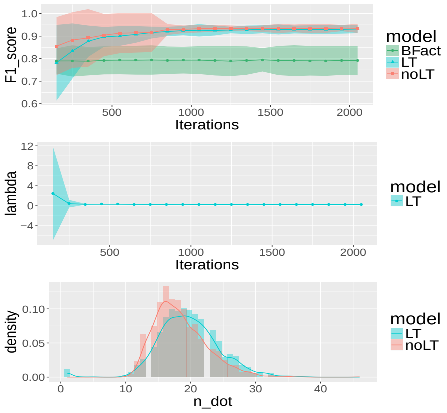

Figure 3. Top: F1 score for inferred relative to ground truth binary speaker matrices on synthetic data generated from the vanilla HDP-HMM model. Middle: Learned similarity parameter, λ, for the LT model by Gibbs iteration, averaged over 5 runs. Bottom: Number of states used, $n_{\cdot}$, by each model in the training set. Error bands are 99% confidence interval of the mean per iteration. The first 100 iterations are omitted.

a Hamiltonian Monte Carlo (HMC) update (Duane et al., 1987; Neal et al., 2011) to update the $\boldsymbol{\ell}_j$ simultaneously, conditioned on $\mathbf{z}$ and $\boldsymbol{\pi}$ (see the supplement for details).

Results We ran 5 Gibbs chains for 10,000 iterations each using the HDP-HMM-LT, Sticky-HDP-HMM-LT, HDP-HMM and Sticky-HDP-HMM models on the 200 training chorales, which were modeled as conditionally independent of one another. We evaluated the marginal log likelihood on the 17 test chorales (integrating out $\mathbf{z}$) at every 50th iteration. The training and test log likelihoods are in Fig. 4. Although the LT model does not achieve as close a fit to the training data, its generalization performance is better, suggesting that the vanilla HDP-HMM is overfitting. This is perhaps counterintuitive, since the LT model is more flexible, and might be expected to be more prone to overfitting. However, the similarity bias induces greater information sharing across parameters, as in a hierarchical model: instead of each entry of the transition matrix being informed mainly by transitions directly involving the corresponding states, it is informed to some extent by *all* transitions, as they all inform the similarity structure.

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 allowing state space geometry to be represented via a similarity kernel, making transitions between "nearby" pairs of states ("local" transitions), more likely *a priori*. By introducing an augmented data representation, which we call the Markov Jump Process with Failed Transitions (MJP-FT), we obtain a Gibbs sampling algorithm that simplifies inference in both the LT and ordinary HDP-HMM. When multiple latent chains are interdependent, as in speaker diarization, the HDP-HMM-LT model combines the HDP-HMM's capacity to discover a small set 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, as well as outperforming the Sticky-HDP-HMM, on a speaker diarization task in which speakers form 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.

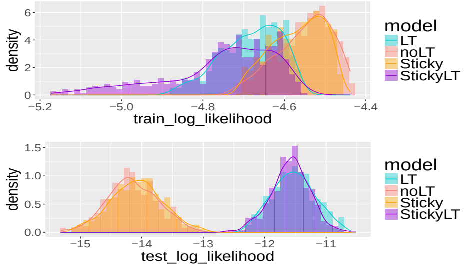

Figure 4. Training set and test set log marginal likelihoods for Bach chorale data on the four HDP-based models: HDP-HMM-LT, HDP-HMM, Sticky HMM, and Sticky HDP-HMM-LT.

The local transition property is particularly clear when transitions occur at different times for different latent features, as with binary vector-valued states in the cocktail party setting, but the model can be used with any state space equipped with a suitable similarity kernel. Similarities need not be defined in terms of emission parameters; state "locations" can be represented and inferred separately, which we demonstrate using Bach chorale data. There, the LT model achieves better predictive performance on a held-out test set, while the ordinary HDP-HMM overfits the training set: the LT property here acts to encourage a concise harmonic representation where chord contexts are arranged in bidirectional functional relationships.

We focused on fixed-dimension binary vectors for the cocktail party and synthetic data experiments, but it would be straightforward to add the LT property to a model with nonparametric latent states, such as the iFHMM (Gael et al., 2009) and the infinite factorial dynamic model (Valera et al., 2015), both of which use the Indian Buffet Process (IBP) (Ghahramani & Griffiths, 2005) as a state prior. The similarity function used here could be employed without changes: since only finitely many coordinates are non-zero in the IBP, the distance between any two states is finite.

ACKNOWLEDGMENTS

This work was funded in part by DARPA grant W911NF-14-1-0395 under the Big Mechanism Program and DARPA grant W911NF-16-1-0567 under the Communicating with Computers Program.

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.

Ewens, Warren John. Population genetics theory – the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory*, pp. 177–227. Springer, 1990.

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*, 2016.

Sethuraman, Jayaram. A constructive definition of Dirichlet processes. *Statistica Sinica*, 4:639–650, 1994.

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.

Van Gael, Jurgen, Saatci, Yunus, Teh, Yee Whye, and Ghahramani, Zoubin. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1088–1095. ACM, 2008.

Zhu, Hao, Hu, Jinsong, and Leung, Henry. Hidden Markov models with discrete infinite logistic normal distribution priors. In *Information Fusion (FUSION), 2016 19th International Conference on*, pp. 429–433. IEEE, 2016.