Infinite Structured Hidden Markov Model

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present the infinite structured hidden Markov model (ISHMM). An ISHMM is an HMM that possesses an unbounded number of states, parameterizes state dwell-time distributions explicitly, and can constrain what kinds of state transitions are possible. We present two parameterizations of the ISHMM. The first is a novel construction for an infinite explicit duration HMM. The second is an entirely novel infinite left-to-right HMM. We provide inference algorithms for the ISHMM and show results from using the ISHMM to analyze both real and synthetic data.

1 Introduction

An ISHMM is an HMM [12] that possesses an unbounded number of states [1, 15, 16, 8], parameterizes state dwell-time distributions explicitly [10, 11, 17, 8], and can constrain state transitions. The infinite number of states allow model selection difficulties to be sidestepped while structured state transitions allow for an additional form of regularization as well as for the encoding of problem-specific knowledge.

The key difficulty in developing the ISHMM is constructing a set of dependent, infinite-dimensional transition distributions with structural zeros. For example, explicit duration HMMs disallow self-transitions; left-to-right HMMs disallow transitions to previously visited states. Infinite variants of both must do the same. Previous non-parametric HMMs have used hierarchical Dirichlet processes (HDPs) to construct the transition distributions, but HDPs cannot generatively produce transition distributions with structural zeros. We solve this problem by using spatial normalized gamma processes (SNΓPs)[13]. This construction is capable of generating infinite dimensional transition distributions that are structured in the sense that transitions to subsets of states can be constrained to have zero probability. This differs from the nonparametric hidden semi-Markov model proposed in a recent paper by Johnson & Willsky [8], which was based on the HDP infinite HMM (HDPHMM). The limitations of HDPs required the use of rejection sampling both to do inference and also to assert the existence of an underlying generative model with the required transition constraints.

We begin in Sec. 2 with a general description of the ISHMM. Sec. 2.1 reviews SNΓPs [13] and shows how they can be used to construct infinite dimensional distributions with structural zeros. Starting in Sec. 2.2 we complete the ISHMM description and develop infinite explicit duration and left-to-right HMM parameterizations. Markov chain Monte Carlo inference procedures for the ISHMM are developed in Sec. 3. Analyses of synthetic, nanoscale transistor random telegraph noise, morse code, and mine disaster datasets can be found in Sec. 4.

2 The Infinite Structured Hidden Markov Model (ISHMM)

The ISHMM is an explicit duration HMM with countable state cardinality generalized to allow a wide variety of state transition constraints. To specify an ISHMM one must choose an observation distribution and prior over the parameter space Θ (F_Θ and H_Θ respectively), a per-state dwell dis-

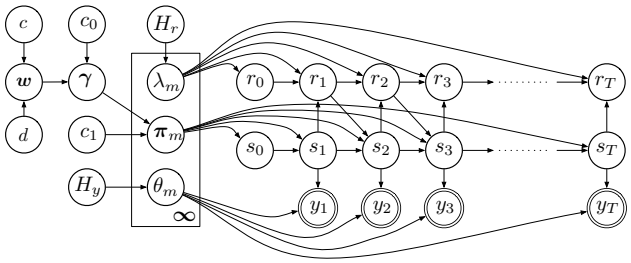


Figure 1: Graphical model for the ISHMM (auxiliary variables for slice-sampling not shown).

tribution and prior over state dwell durations $\{0, 1, 2, 3, \dots\}$ (F_r, H_r), and the set of states V_m to which each state m can transition (for $m \in \mathbb{N}$).

Let $\boldsymbol{s} = (s_1, \dots, s_T)$ be a sequence of latent states. Let $\boldsymbol{y} = (y_1, \dots, y_T)$ be the corresponding observation sequence and let $\boldsymbol{r} = (r_1, \dots, r_T)$ be the latent "remaining duration" sequence. Let $\boldsymbol{\pi}$ be the matrix of state transition distributions, where $\boldsymbol{\pi}_m = (\pi_{m1}, \pi_{m2}, \pi_{m3}, \dots)$ is the row vector corresponding to the transition probabilities out of state m, and $\boldsymbol{\pi}_0$ is the initial state vector. An explicit duration HMM transitions between states s_t and s_{t+1} in the following way: unless $r_t = 0$ the current remaining duration is decremented and the state does not change. If $r_t = 0$ then the HMM transitions to a state $m \neq s_t$ according to the distribution defined by $\boldsymbol{\pi}_{s_t}$. Note that this implies at least structural zeros $\pi_{mm} = 0$ for all m (given by $V_m = \mathbb{N} \backslash m$) The duration the HMM will remain in the new state m is drawn from a state-specific duration distribution $F_r(\lambda_m)$ governed by parameter λ_m. At each time the observation $y_t \sim F_\Theta(\theta_m)$ is drawn from a state-specific emission distribution parameterized by θ_m, which have a shared prior $\theta_m \sim H_\Theta$. To summarize

$$\theta_m \sim H_\Theta \qquad \lambda_m \sim H_r \qquad s_t | s_{t-1}, r_{t-1} \sim \begin{cases} \mathbb{I}(s_t = s_{t-1}), & r_{t-1} > 0 \\ \mathsf{Discrete}(\boldsymbol{\pi}_{s_{t-1}}), & r_{t-1} = 0 \end{cases}$$

$$y_t | s_t \sim F_\Theta(\theta_{s_t}) \qquad r_t | s_t, r_{t-1} \sim \begin{cases} \mathbb{I}(r_t = r_{t-1} - 1), & r_{t-1} > 0 \\ F_r(\lambda_{s_t}), & r_{t-1} = 0 \end{cases}$$

where $\mathbb{I}$ the indicator function that returns one when its argument evaluates to true. Figure 1 shows the corresponding graphical model.

2.1 Infinite Structured Transition Distributions via SNΓPs

SNΓPs can be used to construct dependent DPs [13]. An overview of how SNΓPs can be used to generate structurally constrained transition distributions for the ISHMM is as follows

- Draw an unnormalized marginal state occupancy distribution from a base ΓP defined over the entire set of states. Restrict this distribution according to V_m for each state. Normalize to produce a DP draw D_m for each state. This is the marginal occupancy distribution restricted to the states accessible from m and renormalized.
- Generate the transition distribution $\boldsymbol{\pi}_m$ from D_m.

We start the formal treatment of this procedure by defining a base measure for the base ΓP[1] of the form $\alpha = c_0 H_{\Theta \times \mathbb{N}}$, where c_0 is a concentration parameter and $H_{\Theta \times \mathbb{N}}$ is a probability measure on the joint space of parameters Θ and states $\mathbb{N}$ (the natural numbers). Construct $H_{\Theta \times \mathbb{N}}$ as follows. Let $\tilde{H}_\Theta \sim \mathcal{PY}(c, d, H_\Theta)$ be an atomic measure drawn from a Pitman-Yor process with base measure H_Θ, concentration parameter c, and discount parameter d. Thus $\tilde{H}_\Theta = \sum_{m=1}^{\infty} w_m \delta_{\theta_m}$, where $\boldsymbol{w} = (w_1, w_2, \dots) \sim \mathcal{SB}(c, d)$ is drawn from the stick-breaking prior [6] and $\theta_m \sim H_\Theta$. Let

$$\alpha(\tilde{\theta}, \tilde{M}) = c_o H_{\Theta \times \mathbb{N}}(\tilde{\theta}, \tilde{M}) = c_o \sum_{m=1}^{\infty} w_m \mathbb{I}(\theta_m \in \tilde{\theta}) \mathbb{I}(m \in \tilde{M}). \tag{1}$$

[1] A gamma process with base measure α, $\Gamma\mathrm{P}(\alpha)$, is called a gamma process with mean measure α when draws $G \sim \Gamma\mathrm{P}(\alpha)$ have the form $G = \sum_{n=1}^{\infty} \gamma_n \delta_{\theta_n}$ such for all $\tilde{\Theta} \in \Omega$, $\sum_{n=1}^{\infty} \gamma_n \mathbb{I}(\theta_n \in \tilde{\Theta}) \sim \mathsf{Gamma}(\alpha(\tilde{\Theta}), 1)$. [9]

This associates a single observation distribution parameter θ_m and weight w_m with each state m.

To generate transition distributions for each state that conform to our choice of V_m requires some organizing notation. Partition $\mathbb{N}$ into a collection of *disjoint* sets $\mathcal{A}$ such that every V_m can be constructed by taking the union of set in $\mathcal{A}$. Let $\mathcal{A}_m$ be the collection of sets in $\mathcal{A}$ whose union is V_m. Let R be an arbitrary element of $\mathcal{A}$. This partitioning is necessary to keep track of the independent DPs that will be mixed in (2). These conditions do not uniquely define a single partition. So while any partition that satisfies these conditions will work, some choices can result in more computationally efficient inference.

Let $G \sim \Gamma\mathrm{P}(\alpha)$. Now for each element $\mathsf{R} \in \mathcal{A}$ define the "restricted projection" [13] of α onto the set R as $\alpha_{\mathsf{R}}(\tilde{\Theta}) = c_0 H_{\Theta\times\mathbb{N}}(\tilde{\Theta}, \mathsf{R})$. Then with $G_{\mathsf{R}} \sim \Gamma\mathrm{P}(\alpha_{\mathsf{R}})$, $\sum_{\mathsf{R}\in\mathcal{A}} G_{\mathsf{R}}$ is equal in distribution to G, and G_{R} is independent of $G_{\mathsf{R}'}$ for $\mathsf{R}', \mathsf{R} \in \mathcal{A}$ and $\mathsf{R} \neq \mathsf{R}'$. In other words we can either think of restricting one draw G to produce the G_{R}'s or restricting the base measure α and then drawing the G_{R}'s independently.

Either way, normalizing these ΓP draws, $D_{\mathsf{R}} = G_{\mathsf{R}}/G_{\mathsf{R}}(\Theta)$ produces a set of draws from independent DPs $D_{\mathsf{R}} \sim \mathrm{DP}(\alpha_{\mathsf{R}})$. The key point we take from [13] is that these independent DP draws can be combined to construct a set of dependent DP draws D_m, one for each state

$$D_m = \frac{\sum_{\mathsf{R}\in\mathcal{A}_m} G_{\mathsf{R}}}{\sum_{\mathsf{R}'\in\mathcal{A}_m} G_{\mathsf{R}'}(\Theta)} = \sum_{\mathsf{R}\in\mathcal{A}_m} \frac{G_{\mathsf{R}}(\Theta)}{\sum_{\mathsf{R}'\in\mathcal{A}_m} G_{\mathsf{R}'}(\Theta)} D_{\mathsf{R}} = \sum_{\mathsf{R}\in\mathcal{A}_m} \frac{\gamma_R}{\sum_{\mathsf{R}'\in\mathcal{A}_m} \gamma_{R'}} D_{\mathsf{R}} \tag{2}$$

The D_m are dependent because they share atoms and the weights on those atoms are correlated. By construction D_m places mass only on states in the set V_m.

The D_m serve as base distributions for drawing the conditional state transition distributions $\tilde{D}_m \sim \mathrm{DP}(c_1 D_m)$, where $\tilde{D}_m := \sum_{k=1}^{\infty} \pi_{mk}\delta_{\theta_k}$. The resulting vectors $\boldsymbol{\pi}_m = (\pi_{m1}, \pi_{m2}, \pi_{m3}, \dots)$ are infinite state transition distributions with structural zeros. Drawing them in this way allows each state to flexibly deviate from the restricted, renormalized marginal state occupancy distribution D_m yet still share statistical strength in a hierarchical manner.

By choosing particular sets V_m ISHMMs can encode different kinds of transition structures into infinite HMMs. In the following two sections we describe choices that lead to infinite variants of two common finite HMM variants. The most important difference between the two models is the way in which the set of states is partitioned.

2.2 Infinite Explicit Duration HMM (IEDHMM)

An infinite explicit duration HMM must disallow self-transitions in order to support explicitly parameterized state dwell duration distributions. Therefore let the range of each state m be $V_m = \mathbb{N} \setminus \{m\}$. With this specific choice of range we can complete the generation of infinite transition distributions for an IEDHMM. Let the set of states used to generate some finite set of observations be $\mathcal{T} \subset \mathbb{N}$. Let $R_+ = \mathbb{N}\backslash\mathcal{T}$ be the states that were not used. Then, with $m \in \mathcal{T}, k \in \mathbb{N}$ and $w_+ := \sum_{k'\in R_+} w_{k'}$ we can derive from (2)

$$\gamma_m \sim \mathsf{Gamma}(c_0 w_m, 1) \qquad \gamma_+ \sim \mathsf{Gamma}(c_0 w_+, 1) \tag{3}$$

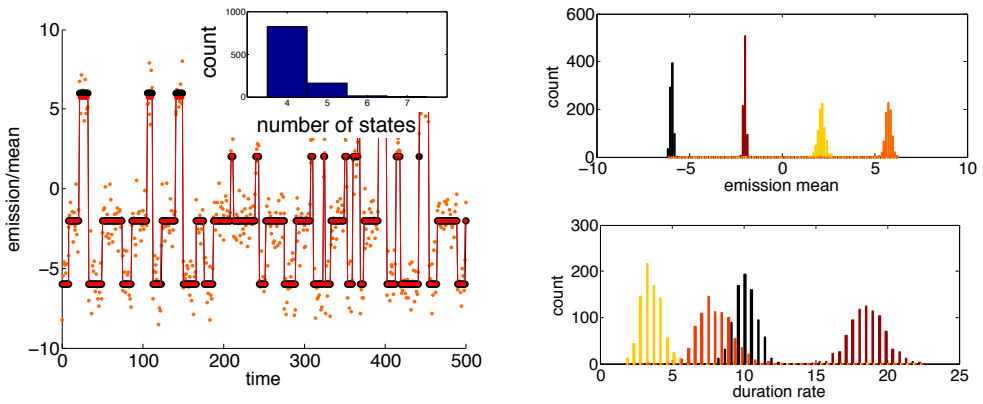
$$\beta_{mk} = \frac{\mathbb{I}(k \in \mathcal{T} \cap V_m)\gamma_k}{\gamma_+ + \sum_{k'\in\mathcal{T}\cap V_m} \gamma_{k'}} \qquad \beta_{m+} = \frac{\gamma_+}{\gamma_+ + \sum_{k'\in\mathcal{T}\cap V_m} \gamma_{k'}} \tag{4}$$

$$D_m = \sum_{k'\in\mathcal{T}\cap V_m} \beta_{mk'}\delta_{\theta_{k'}} + \beta_{m+} D_+ \qquad D_+ \sim \mathrm{DP}(\alpha_{R_+}) \tag{5}$$

If $\beta_{mk} \neq 0$ then state k is reachable from state m. Denoting $\boldsymbol{\beta}_m = (\beta_{m1}, \dots, \beta_{mM}, \beta_{m+})$ we draw the structured state transition distributions $\boldsymbol{\pi}_m = (\pi_{m1}, \dots, \pi_{mM}, \pi_{m+})$ from

$$\boldsymbol{\pi}_m \sim \mathsf{Dirichlet}(c_1 \boldsymbol{\beta}_m). \tag{6}$$

The conditional state transition probability row vector $\boldsymbol{\pi}_m$ is finite dimensional only because the probabilities of transitioning to unused states have been merged into $\pi_{m+} = \sum_{k'\in R_+}^{\infty} \pi_{mk'}$. During inference we will often need to perform inference about states that are currently merged into R_+.



(a) (b)

Figure 2: **(a)** Synthetic data generated from a 4 state EDHMM with MAP IEDHMM sample overlaid; means of the sampled states (red) are overlaid on the true data (orange) and true means (black). *Inset:* Posterior distribution of IEDHMM utilized state counts. **(b)** Posterior distributions of latent state parameters for data in (a). The true means were -6, -2, 2, and 6 and the true duration rates were 10, 20, 3, and 7.

2.3 Infinite left-to-right HMM

One useful set of constraints, particularly for recognition and change-point detection tasks, are those required by the left-to-right HMM: transitions to previously used states are forbidden, namely $\pi_{ij} = 0$ for all $j < i$. To the best of our knowledge, no nonparametric generalization of the left-to-right HMM has been defined before now. Enforcing this restriction in the ISHMM requires that we choose $V_m = \{m+1, m+2, \dots\}$. This choice of V_m requires increased notational complexity but no change to the generative model. Consider states $m \in \mathcal{T}, k \in \mathbb{N}$ and let m^* be the smallest $k' \in \mathcal{T}$ such that $k' > m$, or ∞ if no such k' exists. Let $R_{m+} = \{k' \in \mathbb{N} | m < k' < m^*\}$ be the region in between m and the next index that is in $\mathcal{T}$ (i.e. m^*). Then

$$\gamma_m \sim \mathsf{Gamma}(c_0 w_m, 1) \qquad \gamma_{m+} \sim \mathsf{Gamma}(c_0 \textstyle\sum_{k \in R_{m+}} w_k, 1) \tag{7}$$

$$\beta_{mk} = \frac{\mathbb{I}(k \in \mathcal{T} \cap V_m)\gamma_k}{\gamma_{m+} + \sum_{k' \in \mathcal{T} \cap V_m}(\gamma_{k'} + \gamma_{k'+})} \qquad \beta_{mk+} = \frac{\mathbb{I}(k \in (\mathcal{T} \cap V_m) \cup \{m\})\gamma_{k+}}{\gamma_{m+} + \sum_{k' \in \mathcal{T} \cap V_m}(\gamma_{k'} + \gamma_{k'+})} \tag{8}$$

$$D_m = \beta_{mm+} D_{m+} + \sum_{k \in \mathcal{T} \cap V_m} (\beta_{mk} \delta_{\theta_k} + \beta_{mk+} D_{k+}) \qquad D_{m+} \sim \mathrm{DP}(\alpha_{R_{m+}}) \tag{9}$$

The $\boldsymbol{\pi}_m$'s are drawn as they are in the IEDHMM.

3 Inference

Our approach to inference in ISHMMs is similar to standard approaches to inference in the HDP-HMM. In particular, we employ the forward filtering backward slice sampling approach of Van Gael et al. [16] for infinite HMMs and Dewar et al. [3] for explicit duration HMMs. In both the state and duration variables $\boldsymbol{s}$ and $\boldsymbol{r}$ are sampled conditioned on auxiliary slice variables $\boldsymbol{u}$. For space reasons only the distributions required for inference in the IEDHMM are shown. While the left-to-right infinite HMM is sampled in exactly the same way (and in fact, using the same code), the extra notational complexity requires too much space to include the exact sampling distributions. They are, however, easily derived in the same way as the following.

Sampling $\boldsymbol{s}$, $\boldsymbol{r}$, and $\boldsymbol{u}$ Define a "full" state random variable $z_t = (s_t, r_t)$ and define auxiliary random variables u_t distributed according to

$$p(u_t | z_t, z_{t-1}) = \mathbb{I}(u_t < p_t) p_t p_\beta(u_t / p_t; \alpha_u, \beta_u), \tag{10}$$

Figure 3: IEDHMM segmentation of morse code. *Left:* Spectrogram of Morse code audio. *Middle:* IEDHMM MAP state sequence *Right:* sticky HDP-HMM MAP state sequence. The IEDHMM correctly learns three states (off – black, dot – white, and dash – orange) even though the dot and dash states can only be distinguished by duration. Non-explicit-duration HMMs like the sticky HDP-HMM are only able to learn two because of this.

where $p_\beta(\cdot; \alpha_u, \beta_u)$ is the density for the beta distribution with parameters α_u and β_u and

$$\begin{aligned} p_t &:= p(z_t|z_{t-1}) = p((s_t, r_t)|(s_{t-1}, r_{t-1})) = \\ &= \begin{cases} r_{t-1} > 0, & \mathbb{I}(s_t = s_{t-1})\mathbb{I}(r_t = r_{t-1} - 1) \\ r_{t-1} = 0, & \pi_{s_{t-1}s_t} F_r(r_t; \lambda_{s_t}). \end{cases} \end{aligned}$$

It is straightforward to sample $\boldsymbol{u}$ according to equation (10). Sampling $\boldsymbol{z}$ from its conditional distribution given values of all other random variables in the model closely follows the beam sampling method in Van Gael et al. [16], namely a forward-backward procedure for block Gibbs sampling $\boldsymbol{z}$ given slice variables $\boldsymbol{u}$. To denote a portion of some vector, say $\boldsymbol{x}$, we use the notation $x_{s:t} = (x_s, \ldots, x_t)$. For notational clarity we omit the dependencies on π's, λ's, and θ's in what follows. Define the forward variables $\alpha_t(z_t)$ by

$$\begin{aligned} \alpha_t(z_t) &:= p(z_t|y_{1:t}, u_{1:t}) \\ &\propto F_\Theta(y_t|\theta_{s_t}) \sum_{z_{t-1}:u_t<p_t} p_\beta(u_t/p_t; \alpha_u, \beta_u)\alpha_{t-1}(z_{t-1}). \end{aligned} \tag{11}$$

Samples of the full latent states are taken during the backward pass by sampling from

$$\begin{aligned} p(z_T|y_{1:T}, u_{1:T}) &= \alpha_T(z_T) \\ p(z_t|z_{t+1}, y_{1:T}, u_{1:T}) &\propto \mathbb{I}(u_{t+1} < p_{t+1}) p_\beta(u_t/p_t; \alpha_u, \beta_u)\alpha_t(z_t). \end{aligned}$$

It is convenient to express α_u and β_u in terms of a single "temperature" parameter $\mathcal{K}$: $\alpha_u = 1/\mathcal{K}$ and $\beta_u = \mathcal{K}$. This temperature controls both the space of models the sampler can reach and how quickly it does so. For instance, letting $\mathcal{K} \to \infty$ fixes all the u_t's to zero, which results in an intractable infinite sum in equation (22) corresponding to a full forward-backward pass ignoring the complexity-controlling slice variables. Setting $\mathcal{K} = 1$ recovers the uniform distribution used by Van Gael et al. [16]. Values of $\mathcal{K}$ tending towards 0 send all the u_t's to 1, which limits the computational complexity of the sampler (fewer paths are explored on each sweep), but causes the chain to mix more slowly. Because of the deterministic state transitions introduced by using a remaining duration counter, sampling with a temperature greater than one was generally found to be beneficial in terms of faster mixing.

Sampling $\boldsymbol{\gamma}$ $\boldsymbol{\pi}$, and $\boldsymbol{w}$ Beam sampling the z's results in a sequence of "observed" transitions. Given these (and respective priors), $\boldsymbol{\gamma}$ and $\boldsymbol{\pi}$ are sampled. To do this, the observed transitions are sequentially seated in a Chinese restaurant franchise (CRF) [15] to instantiate counts for the dependent DPs. That is, with the observed counts from $\tilde{D}_m$ contained in the vector $\boldsymbol{C}_m$ (i.e. C_{mk} is the number of observed transitions from state m to state k), we denote the number of tables serving θ_k in the restaurant representation of $\tilde{D}_m$ by l_{mk}. This, by the inductive argument central to CRF sampling, is equal to the number of draws of θ_k from D_m. A Gibbs sample of the l_{mk}'s can be generated by running a Chinese restaurant process with customers $\boldsymbol{C}_m$ and keeping track of the number of tables generated

$$l_{mk}^{(1)} = 1 \qquad l_{mk}^{(j+1)} = l_{mk}^{(j)} + \mathbb{I}\left(X_j < \frac{c_1\beta_{mk}}{(c_1\beta_{mk} + j)}\right) \qquad l_{mk} = l_{mk}^{(C_{mk})} \tag{12}$$

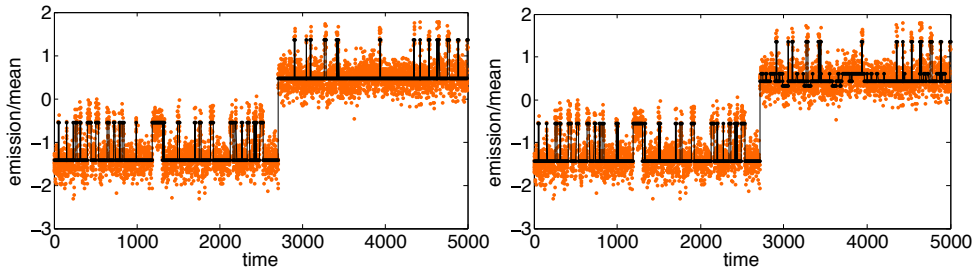


Figure 4: Rescaled random telegraph noise data (orange) and MAP estimate of the IEDHMM (black, left) and sticky HDP-HMM (black, right). Without state-specific, explicit, non-geometric duration distributions even a well-tuned sticky HDP-HMM tends to over-segment.

with $X_j \sim \mathsf{Uniform}(0,1)$. In order to sample the weights $\boldsymbol{\gamma}$ of the independent gamma processes, we follow [13] and introduce the auxiliary variables vector $\boldsymbol{L} = (L_1, \ldots, L_M)$ such that

$$L_j \sim \mathsf{Gamma}\left(l_{j\cdot}, \left(\gamma_+ + \textstyle\sum_{i\neq j}^{M} \gamma_i\right)^{-1}\right)$$

$$\gamma_m \sim \mathsf{Gamma}\left(l_{\cdot m} + c_0 w_m, \left(1 + L_{\backslash m}\right)^{-1}\right) \quad \gamma_+ \sim \mathsf{Gamma}\left(c_0 w_+, \left(1 + \textstyle\sum_{j=1}^{M} L_j\right)^{-1}\right),$$

where $L_{\backslash m} = \sum_{j\neq m}^{M} L_j$ and dots in subscripts indicate summation over that index. After sampling $\boldsymbol{\gamma}$, the matrix $\boldsymbol{\beta}$ is calculated deterministically using equation (4). The stick weights $\boldsymbol{w}$ were sampled using Metropolis Hastings updates and the rows of the transition matrix $\boldsymbol{\pi}$ are sampled according to

$$\boldsymbol{\pi}_m \sim \mathsf{Dirichlet}(\boldsymbol{C}_m + c_1 \boldsymbol{\beta}_m). \tag{13}$$

Sampling $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$, and hyperparameters Sampling of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ depends on the choice of prior distributions H_Θ and H_r, and data distributions F_Θ and F_r. For standard choices, straightforward MCMC sampling techniques can be employed. The concentration parameters c_0 and c_1 are sampled via Metropolis-Hastings.

3.1 Considerations During Forward Inference

When calculating the forward variables, if $\pi_{m+} > u_t$ for some m and t, then it is possible for a transition into one of the merged states to occur. In this case,merged states $M+1, \ldots$ must be re-instantiated on the fly until $\pi_{m+} < u_t$ for all m and t. To re-instantiate a state, note that γ_+ is the total weight for all unobserved states, i.e., the total weight for the draw G_+ from gamma process $\Gamma\mathrm{P}(\alpha_+)$. Thus, a state v_{M+1} will have weight $\gamma_{M+1} < \gamma_+$. Since the normalized weight γ_{M+1}/γ_+ is a weight from a DP, it can be sampled using the stick breaking construction

$$b_{M+1} \sim \mathsf{Beta}(1, \alpha_+(\Theta)) \tag{14}$$

$$\gamma_{M+1} = b_{M+1}\gamma_+ \tag{15}$$

$$\gamma_+ \leftarrow (1 - b_{M+1})\gamma_+. \tag{16}$$

The normalization terms for the $\boldsymbol{\beta}_m$'s do not change, but β_{mM+1} (and thus π_{mM+1}) must be added and β_{m+} (and thus π_{m+}) must be updated. The updates to $\boldsymbol{\pi}$ can be accomplished by noting that π_{mM+1} and π_{m+} are two components of a draw from $\mathsf{Dirichlet}(\ldots, c_1\beta_{mM+1}, c_1\beta_{m+})$, so a draw from $\mathsf{Dirichlet}(c_1\beta_{mM+1}, c_1\beta_{m+})$ (i.e. the beta distribution) gives the proportion of the old π_{m+} that stays on π_{m+} and the proportion that is broken off to form π_{mM+1}

$$\tilde{b}_{mM+1} \sim \mathsf{Beta}(c_1\beta_{mM+1}, c_1\beta_{m+}) \tag{17}$$

$$\pi_{mM+1} = \tilde{b}_{mM+1}\pi_{m+} \tag{18}$$

$$\pi_{m+} \leftarrow (1 - \tilde{b}_{mM+1})\pi_{m+}. \tag{19}$$

Finally, $\boldsymbol{\pi}_{M+1}$ is sampled according to equation (6). This procedure is repeated as many times as is necessary to ensure that $\pi_{m+} < u_t$ for all m and t. Also, it should be noted, this procedure allows for the development of incremental inference procedures for models in the ISHMM family.

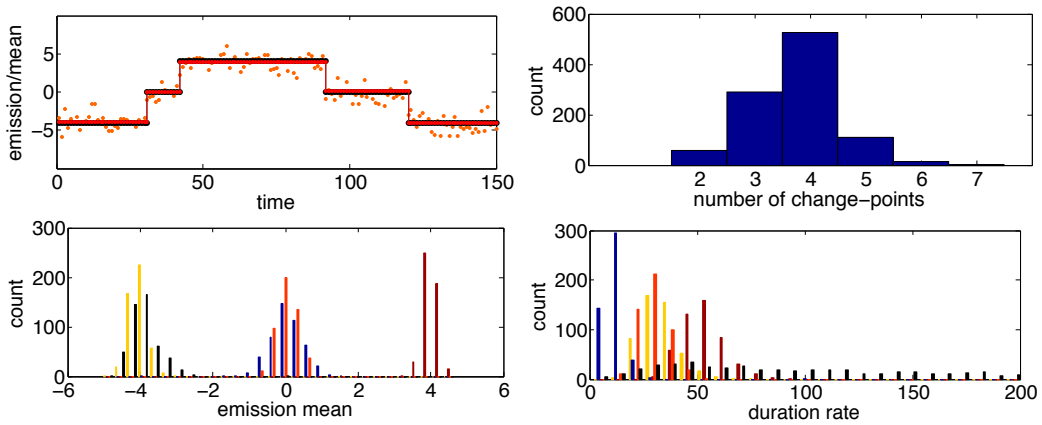


Figure 5: **Top left:** Synthetic data generated from a 5 state left-to-right HMM with MAP infinite left-to-right HMM sample overlaid; means of the sampled states (red) are overlaid on the true data (orange) and true means (black). **Top right:** Posterior distribution of infinite left-to-right HMM change points. **Bottom:** Posterior distributions of latent state parameters for generated data. The final rate is not well defined, hence the posterior has large support.

4 Experiments

4.1 IEDHMM

We illustrate IEDHMM learning on synthetic data. Five hundred datapoints were generated using a 4 state EDHMM with Poisson duration distributions (rates $\boldsymbol{\lambda} = (10, 20, 3, 7)$) and Gaussian emission distributions (means $\boldsymbol{\mu} = (-6, -2, 2, 6)$, all unit variance). For inference, the emission and duration distributions were given broad priors. The emission distributions were given Normal-scaled Inverse Gamma priors with $\mu_0 = 0, \nu_0 = .25, \alpha = 1$, and parameters for the Poisson duration distributions were given $\mathsf{Gamma}(1, 10^3)$ priors. The temperature was set to $\mathcal{K} = 3$.

One thousand samples were collected after a burn-in of 100 iterations. Figure 2a shows the highest scoring sample and (inset) a histogram of the state cardinalities of the models explored (82% of the samples had 4 states) The inferred posterior distribution of the duration distribution rates and means of the state observation distributions is shown in Fig. 2b. All contain the true values in regions of high posterior confidence.

4.1.1 Morse Code

Morse code consists of a sequence of short and long "on" tones. The frequency spectrum of a sequence of Morse code audio (8KHz., 9.46 sec.) is shown in Fig. 3. Following Johnson & Willsky [8], we segmented it to illustrate both the utility of explicitly parameterizing duration distributions and to illustrate the correctness of our ISHMM construction and sampling algorithms. Figure 3 also shows that, because each state has its own delayed geometric duration distribution $F_r(\lambda_{s_t}) = \mathsf{Geom}(q_{s_t}) + d_{s_t}$ (where $\lambda_m = (d_m, q_m), d_m \in \{0, \ldots, 30\}, q_m \in \mathbb{Z}^+$, and $H_r(d, q) = \frac{1}{30}$), the IEDHMM is able to distinguish short and long tones and assign a unique state identifier to each (using a Gaussian emission model for the first Mel-frequency Cepstrum coefficient). Non-explicit-duration HMMs such as the sticky HDP-HMM [4] can only infer the existence of two states because they cannot distinguish states by duration.

4.1.2 Nanoscale Transistor Noise

Random telegraph noise (RTN) is the name given to instantaneous temporal changes in modern-day nanoscale transistor current-driving characteristics due to quantum tunneling of charge carriers in and out of potential traps in the device oxide. In macro-scale electronic systems RTN can manifest itself as anything from an annoying flicker of a pixel to complete failure of the entire system. In order to quantify and mitigate the negative effects of RTN, the statistical characteristics of RTN must be

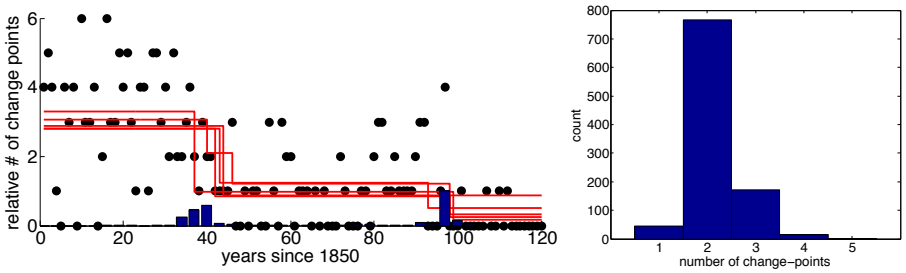


Figure 6: Results from using the infinite left-to-right HMM to model coal mining disaster data. *Left:* Number of disasters (black dots) and five inferred posterior sample paths (red). A sample path is the mean of the current state. Steps occur at inferred change-points. The histogram on the horizontal axis shows the relative number of change points inferred to have occurred at each time. *Right:* Marginal posterior change-point count sample distribution.

well understood [14]. IEDHMMs are well suited for this task since the duration of the temporal changes is random and of interest and the number of "error states" is not known a priori. Figure 4 shows the results of using the IEDHMM to a model RTN data in which the domain experts believe that there are four latent states with characteristic duration distributions. We find that the IEDHMM is able to learn a model in good correspondence with scientist expectation.

4.2 Infinite Left-to-Right HMM

We illustrate infinite left-to-right HMM learning on synthetic data. One hundred and fifty datapoints were generated using a 5 state left-to-right HMM with Poisson duration distributions (rates $\boldsymbol{\lambda} = (30, 10, 50, 30, -))$ and Gaussian emission distributions (means $\boldsymbol{\mu} = (-4, 0, 4, 0, -4)$, all unit variance). The last duration rate is undefined since there is no transition out of the 5th state. Hyperparameters were initialized in the same way as in the IEDHMM synthetic data experiment.

One thousand samples were collected after a burn-in of 100 iterations. Figure 5 shows the highest scoring sample and a histogram of the number of inferred change-points in the data. The inferred posterior distribution of the duration distribution rates and means of the state observation distributions are also shown in Fig. 5. All contain the true values in regions of high posterior confidence. The posterior duration rate for the final state has large support, which is consistent with the fact that its duration is not well defined.

4.2.1 Coal Mining Disasters

A well-studied change-point dataset is the number of major coal mining disaster in Britain between 1851 and 1962 [7]. In previous analyses, such as that by Chib [2], either one or two change-points (i.e. two or three states) were assumed. We used the infinite left-to-right HMM with Poisson emissions and durations (with gamma priors) to model the coal mining data. This allowed us to make no assumptions about the number of change-points. Using 1000 samples the model found two change points with high probability; however the model mixed over multiple interpretations of the data, considering anywhere from one to five change-points (Fig. 6, inset). Figure 6 shows the coal mining data and a representative set of posterior samples from the model. The locations of the change-points are well concentrated around 40 and 100 years. This is consistent with previous findings [5].

5 Discussion

The ISHMM parameterizes a large number of structured parametric and nonparametric Bayesian HMMs. This suggests that practitioners should find success in applying models in the ISHMM family to problems in domains where benefits are known to accrue from using HMMs with explicit, not-necessarily geometric state dwell distributions or with restricted state transition topologies: domains like speech segmentation, hand-writing recognition, change-point analysis, activity monitoring, and others. Thorough experimentation with applications of ISHMMs is of significant interest to us and suggests a great deal of future work.

References

[1] Beal, M J, Ghahramani, Z, and Rasmussen, C E. The Infinite Hidden Markov Model. In *Advances in Neural Information Processing Systems*, pp. 29–245, March 2002.

[2] Chib, S. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241, 1998.

[3] Dewar, M, Wiggins, C, and Wood, F. Inference in hidden Markov models with explicit state duration distributions. *IEEE Signal Processing Letters*, 19(4), 2012.

[4] Fox, E B, Sudderth, E B, Jordan, M I, and Willsky, A S. A Sticky HDP-HMM with Application to Speaker Diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

[5] Green, P J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711 – 732, 1995.

[6] Ishwaran, H and James, L F. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[7] Jarrett, R. G. A note on the intervals between coal-mining disasters. *Biometrika*, 66(1):191—193, 1979.

[8] Johnson, M and Willsky, A S. The hierarchical Dirichlet process hidden semi-Markov model. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pp. 252–259, 2010.

[9] Kingman, J F C. *Poisson Processes*. Oxford Studies in Probability. Oxford University Press, 1993.

[10] Mitchell, C, Harper, M, and Jamieson, L. On the complexity of explicit duration HMM's. *IEEE Transactions on Speech and Audio Processing*, 3(3):213–217, 1995.

[11] Murphy, K P. Hidden semi-markov models (HSMMs). Technical report, MIT, 2002.

[12] Rabiner, L R. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pp. 257–286, 1989.

[13] Rao, V and Teh, Y W Whye. Spatial Normalized Gamma Processes. In *Advances in Neural Information Processing Systems*, pp. 1554–1562, 2009.

[14] Realov, S and Shepard, K L. Random Telegraph Noise in 45-nm CMOS : Analysis Using an On-Chip Test and Measurement System. *Analysis*, (212):624–627, 2010.

[15] Teh, Y W, Jordan, M I, Beal, M J, and Blei, D M. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[16] Van Gael, J, Saatci, Y, Teh, Y W, and Ghahramani, Z. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1088–1095. ACM, 2008.

[17] Yu, S Z. Hidden semi-Markov models. *Artificial Intelligence*, 174(2):215–243, 2010.

A Supplementary Material

A.1 Derivation of forward variables

Define the forward variables $\alpha_t(z_t)$ by

$$\begin{aligned}
\alpha_t(z_t) &:= p(z_t|y_{1:t}, u_{1:t}) \\
&\propto p(z_t, u_t, y_t|y_{1:t-1}, u_{1:t-1}) \\
&= \sum_{z_{t-1}} p(y_t|z_t)p(u_t|z_t, z_{t-1})p_t p(z_{t-1}|y_{1:t}, u_{1:t}) \qquad (20) \\
&= p(y_t|s_t) \sum_{z_{t-1}} \mathbb{I}(u_t < p_t)p^\star_\beta \alpha_{t-1}(z_{t-1}) \qquad (21) \\
&= p(y_t|s_t) \sum_{z_{t-1}:u_t<p_t} p^\star_\beta \alpha_{t-1}(z_{t-1}), \qquad (22)
\end{aligned}$$

where $p^\star_\beta := p_\beta(u_t/p_t; \alpha_u, \beta_u)$ and $p(y_t|s_t) = F_\theta(y_t|\theta_{s_t})$. Samples of the full latent states are taken during the backward pass by sampling from

$$\begin{aligned}
p(z_T|y_{1:T}, u_{1:T}) &= \alpha_T(z_T) \\
p(z_t|z_{t+1}, y_{1:T}, u_{1:T}) &\propto p(z_t, z_{t+1}, y_{1:T}, u_{1:T}) \\
&= p(u_{t+1}|z_{t+1}, z_t)p(z_{t+1}|z_t)p(z_t|u_{1:t}, y_{1:t}) \\
&= \mathbb{I}(u_{t+1} < p_{t+1})p^\star_\beta \alpha_t(z_t).
\end{aligned}$$

where $p^*_\beta = p_\beta(u_t/p_t; \alpha_u, \beta_u)$ and $p(y_t|s_t) = F_\theta(y_t|\theta_{s_t})$

A.2 Derivation of backwards sampling pass

$$\begin{aligned}
p(z_T|y_{1:T}, u_{1:T}) &= \alpha_T(z_T) \\
p(z_t|z_{t+1}, y_{1:T}, u_{1:T}) &\propto p(z_t, z_{t+1}, y_{1:T}, u_{1:T}) \\
&= p(u_{t+1}|z_{t+1}, z_t)p(z_{t+1}|z_t)p(z_t|u_{1:t}, y_{1:t}) \\
&= \mathbb{I}(u_{t+1} < p_{t+1})p^\star_\beta \alpha_t(z_t).
\end{aligned}$$

A.3 The HDP-HMM and Bayesian HMM

One can recover the HDP-HMM and the standard Bayesian HMM as specific parameterizations of ISHMMs. For the former, if the duration prior is a delta function at zero, $H_r = \delta_0$, then $r_t = 0$ for all t and the implicit geometric state duration distribution is recovered. Next, let $c_0 \to \infty$ and fix $d = 0$, so $\boldsymbol{\gamma} = \boldsymbol{w}$. Also, because all possible state transitions are permitted in the HDP-HMM, the restricting sets on the auxiliary space should be chosen such that the dependent ΓPs are all equal, i.e. set $V_m = \mathbb{N}$. If all the dependent ΓPs are equal, draws from the dependent DPs are equal as well: $D_{V_m} = D_{V_{m'}}\ \forall m, m' \in \mathcal{T}$. In this case, all of the conditional state transition distributions $\tilde{D}_{V_m}$ are draws from DPs with the same base measure, exactly like the HDP-HMM. To recover a Bayesian HMM with M states, use the same parameter setup as for the HDP-HMM, but replace $\mathbb{N}$ everywhere by $\{1, \ldots, M\}$.