The Block Diagonal Infinite Hidden Markov Model

Thomas Stepleton[†] **Zoubin Ghahramani**[‡†] **Geoffrey Gordon**[†] **Tai Sing Lee**[†]

[†]School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[‡]Department of Engineering
University of Cambridge
Cambridge CB2 1PZ, UK

Abstract

The Infinite Hidden Markov Model (IHMM) extends hidden Markov models to have a countably infinite number of hidden states (Beal et al., 2002; Teh et al., 2006). We present a generalization of this framework that introduces nearly block-diagonal structure in the transitions between the hidden states, where blocks correspond to "sub-behaviors" exhibited by data sequences. In identifying such structure, the model classifies, or partitions, sequence data according to these sub-behaviors in an unsupervised way. We present an application of this model to artificial data, a video gesture classification task, and a musical theme labeling task, and show that components of the model can also be applied to graph segmentation.

1 Introduction

Ordinary hidden Markov models (HMMs) characterize data sequences as sequences of stochastic observations, each of which depends on the concurrent state of a Markov chain operating over a finite set of unobserved states. HMMs are ubiquitous in time series modeling; however, they impose relatively little structure on the dynamics of systems they model.

Many process comprise several "sub-processes", such as a musical composition with several recurring motifs, a dance with repeated gestures, even human speech with common words and phrases. An HMM transition matrix describing the dynamics of these processes will exhibit nearly-block diagonal structure, since transitions between states in the same sub-process will usually be more likely than transitions between states in different behavioral regimes. However, ordinary HMM learning methods cannot bias the transition matrix to be block diagonal and may not infer these important relationships between states.

We present an unsupervised method that learns HMMs with block-diagonal dynamic structure from time series data. Neither the number of states, nor the number of blocks into which states organize, is specified beforehand. This technique is a generalization of the HMM learning technique presented in (Beal et al., 2002) and (Teh et al., 2006), and it has important new capabilities. First, it can isolate distinct behavioral regimes in the dynamics of temporal processes: the "sub-behaviors" mentioned above. Second, it can partition, or classify, data sequences into segments corresponding to the times when these different sub-behaviors are executed. Finally, components of the model offer a useful framework for related inference tasks, including partitioning non-negative integer-weighed graphs.

The technique we generalize, the Infinite Hidden Markov Model (IHMM), described in (Beal et al., 2002) and further formalized in (Teh et al., 2006) (where it is called the HDP-HMM), extends HMMs to Markov chains operating over a countably infinite set of hidden states. The IHMM exhibits a characteristic behavior in generated hidden state sequences, in that the number of visited states always increases with time, but a smaller collection of states receives a large proportion of repeat visits throughout the sequence. This behavior arises because the IHMM expresses a hierarchical Dirichlet process (HDP) prior on the infinitely large transition matrix governing transition behavior between states. Practically, this means that a finite data sequence of length T will usually have come from a smaller collection of $M \ll T$ states, and that the IHMM posterior conditioned on the sequence can exhibit meaningful transition dynamics over these M states while still retaining flexibility over their exact number.

Our generalization, the Block-Diagonal Infinite Hidden Markov Model (BD-IHMM), involves partitioning the infinite set of hidden states into an infinite number of blocks, then modifying the Dirichlet process prior for each hidden state such that transitions between states in the same block are usually more likely than transitions between states in different blocks. Since finite data sequences will usually visit $K \ll M$ blocks, a BD-IHMM posterior can flexibly isolate sub-behaviors in the data sequence by harnessing the model's tendency to group hidden states.

A number of extensions to the IHMM have been proposed recently, including a "tempered HDP-HMM" that exhibits a configurable bias for self-transitions in the hidden states (Fox et al., 2008), a hierarchical model that uses an IHMM to identify system subregimes that are modeled by Kalman filters (Fox et al., 2007), and a model that shares a library of hidden states across a collection of IHMMs that model separate processes (Ni & Dunson, 2007). To our knowledge, there has been no effort to extend the IHMM to express a prior that induces "block-diagonal" behavior in the hidden state dynamics, though the dynamics of our model will bear similarities to those of (Fox et al., 2008) as the number of blocks $K \to M$. More broadly, a literature analyzing block-diagonal HMMs exists (Pekergin et al., 2005), though most of its efforts presume the transition matrix is known *a priori*.

2 The model

Like many characterizations of Dirichlet process-based models, our account of the IHMM and the BD-IHMM, depicted graphically in Figure 1, invokes the "stick-breaking process" of (Sethuraman, 1994). The stick-breaking process is a partitioning of the unit interval into an infinite set of sub-intervals or proportions, akin to snapping partial lengths off the end of a stick. Given a positive *concentration parameter* γ, β_n, the length of interval n, is drawn via the following scheme:

$$\beta'_n \sim \mathrm{Beta}(1, \gamma) \qquad \beta_n = \beta'_n \textstyle\prod_{k=1}^{i-1}(1 - \beta'_i), \tag{1}$$

where metaphorically β'_n is the fraction of the remaining stick to snap off. When the proportions β_n are paired with outcomes θ_n drawn IID from a finite measure H, or *atoms*, the resulting discrete probability distribution over the countably infinite set of atoms is said to be drawn from the Dirichlet process $\mathrm{DP}(\gamma, H)$. In the hierarchical Dirichlet process, the sampled distribution is itself "plugged into" subordinate Dirichlet processes in the place of the measure H. Samples from these subordinate DPs are discrete distributions over the same set of atoms, albeit with varying probabilities. Equivalently, reflecting the generative processes in Figure 1, it is shown in (Teh et al., 2006) that it is also possible to start with the original intervals $\boldsymbol{\beta}$ and draw subordinate collections of intervals $\boldsymbol{\pi}_m$ via stick-breaking as

$$\begin{aligned} &\pi'_{mn} \sim \mathrm{Beta}\left(\alpha_0 \beta_n, \alpha_0(1 - \textstyle\sum_{i=1}^{n} \beta_i)\right) \\ &\pi_{mn} = \pi'_{mn} \textstyle\prod_{i=1}^{n-1}(1 - \pi'_{mi}), \end{aligned} \tag{2}$$

where α_0 is the concentration parameter for the subordinate DPs. Elements in each set of proportions $\boldsymbol{\pi}_m$ are then paired with the same set of atoms drawn from H, in the same order, to generate the subordinate Dirichlet process samples. Note that both the proportions β_n and $\boldsymbol{\pi}_{mn}$ tend to grow smaller as n increases, making it likely that a finite set of T draws from either will result in $M \ll T$ unique outcomes.

The generative process behind the IHMM can now be characterized as follows:

$$\begin{aligned} &\boldsymbol{\beta} \,|\, \gamma \sim \mathrm{SBP1}(\gamma) \\ &\boldsymbol{\pi}_m \,|\, \alpha_0, \boldsymbol{\beta} \sim \mathrm{SBP2}(\alpha_0, \boldsymbol{\beta}) \qquad && v_t \,|\, v_{t-1}, \boldsymbol{\pi} \sim \boldsymbol{\pi}_{v_{t-1}} \\ &\theta_m \,|\, H \sim H && y_t \,|\, v_t, \boldsymbol{\theta} \sim F(\theta_{v_t}), \end{aligned} \tag{3}$$

where SBP1 and SBP2 indicate the stick-breaking processes in Equations 1 and 2 respectively. Here, the transition matrix $\boldsymbol{\pi}$ comprises rows of transition probabilities sampled from a prior set of proportions $\boldsymbol{\beta}$; sequences of hidden states $v_1, v_2, \ldots$ are drawn according to these probabilities as in an ordinary HMM. Emission model parameters for states $\boldsymbol{\theta}$ are drawn from H and generate successive observations y_t from the density $F(\theta_{v_t})$. Note that pairing transition probabilities in $\boldsymbol{\pi}_i$ with corresponding emission model parameter atoms in $\boldsymbol{\theta}$ yields a draw from a hierarchical Dirichlet process, as characterized above.

The BD-IHMM uses an additional infinite set of proportions $\boldsymbol{\rho}$, governed by the concentration parameter ζ, to partition the countably infinite set of hidden states into "blocks", as indicated by per-state block labels z_m. For each state m, the stick-breaking process that samples $\boldsymbol{\pi}_m$ uses a modified prior set of proportions $\boldsymbol{\beta}^*_m$ in which elements β^*_{mn} are scaled to favor relatively higher probabilities for transitions between states in the same block and lower probabilities for transitions between states in different blocks:

$$\begin{aligned} &\boldsymbol{\rho} \,|\, \zeta \sim \mathrm{SBP1}(\zeta) \\ &\boldsymbol{\beta} \,|\, \gamma \sim \mathrm{SBP1}(\gamma) \qquad && z_m \,|\, \boldsymbol{\rho} \sim \boldsymbol{\rho} \\[1ex] &\xi^*_m = 1 + \xi / \left(\textstyle\sum_k \beta_k \cdot \delta(z_m \!=\! z_k)\right) \\ &\beta^*_{mn} = \tfrac{1}{1+\xi} \beta_n {\xi^*_m}^{\delta(z_m = z_n)} \\[1ex] &\boldsymbol{\pi}_m \,|\, \alpha_0, \boldsymbol{\beta}^*_m \sim \mathrm{SBP2}(\alpha_0, \boldsymbol{\beta}^*_m) \\[1ex] & && v_t \,|\, v_{t-1}, \boldsymbol{\pi} \sim \boldsymbol{\pi}_{v_{t-1}} \\ &\theta_m \,|\, H \sim H && y_t \,|\, v_t, \boldsymbol{\theta} \sim F(\theta_{v_t}). \end{aligned} \tag{4}$$

Note that $\sum_n \beta^*_{mn} = 1$, and that states with the same block label have identical corresponding $\boldsymbol{\beta}^*_m$. Here, ξ is a non-negative hyperparameter controlling the amount of prior bias for within-block transitions. Setting $\xi = 0$ yields the original IHMM, while giving it larger values makes transitions between different blocks increasingly improbable. Figure 3 depicts the generation of transition probabilities $\boldsymbol{\pi}_m$ described by Equation 4 graphi-

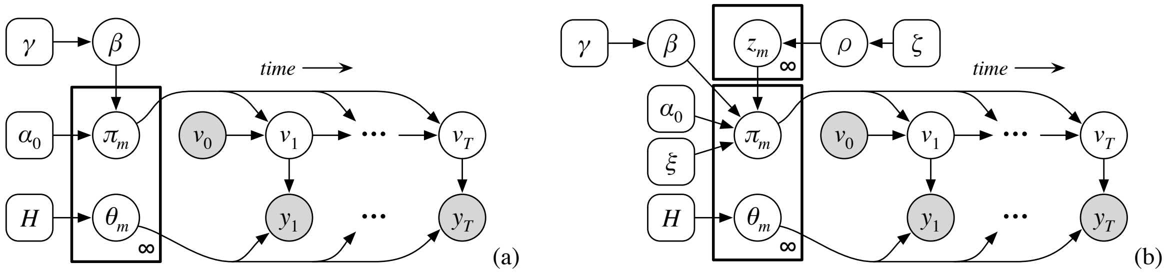
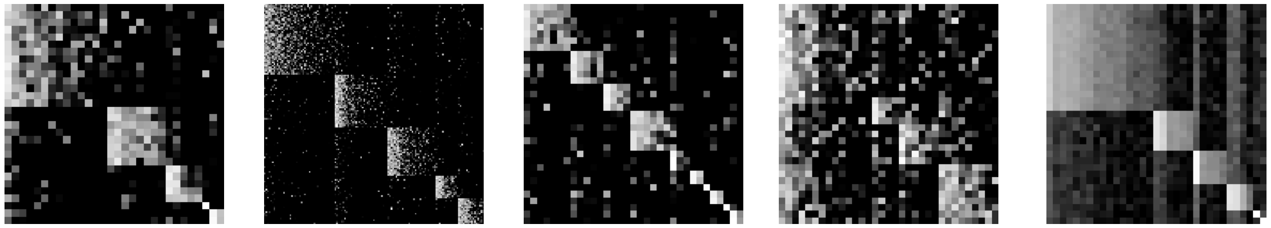


Figure 1: Graphical model depictions of (a) the IHMM as described in (Teh et al., 2006) and (b) the BD-IHMM.



$\zeta = 1$ $\gamma = 10$ $\alpha_0 = 10$ $\xi = 10$

$\zeta = 1$ $\gamma = 50$ $\alpha_0 = 10$ $\xi = 10$

$\zeta = 5$ $\gamma = 10$ $\alpha_0 = 10$ $\xi = 10$

$\zeta = 1$ $\gamma = 10$ $\alpha_0 = 1$ $\xi = 1$

$\zeta = 10$ $\gamma = 10$ $\alpha_0 = 1000$ $\xi = 10$

Figure 2: Truncated Markov transition matrices (right stochastic) sampled from the BD-IHMM prior with various fixed hyperparameter values; highlighted hyperparameters yield the chief observable difference from the leftmost matrix. The second matrix has more states; the third more blocks; the fourth stronger transitions between blocks, and the fifth decreased variability in transition probabilities.

cally, while Figure 2 shows some $\boldsymbol{\pi}$ transition probability matrices sampled from truncated (finite) versions of the BD-IHMM for fixed γ, α_0, ζ, and ξ hyperparameters.

3 Inference

Our inference strategy for the BD-IHMM elaborates on the "direct assignment" method for HDPs presented in (Teh et al., 2006). Broadly, the technique may be characterized as a Gibbs sampling procedure that iterates over draws from posteriors for observation assignments to hidden states $\boldsymbol{v}$, the shared transition probabilities prior $\boldsymbol{\beta}$, hidden state block assignments $\boldsymbol{z}$, and the hyperparameters ζ, γ, α_0, and ξ.

3.1 Observation assignments to hidden states

The direct assignment sampler for IHMM inference samples assignments of observations to hidden states $\boldsymbol{v}$ by integrating the per-state transition probabilities $\boldsymbol{\pi}_m$ out of the conditional distribution of $\boldsymbol{v}$ while conditioning on an instantiated sample of $\boldsymbol{\beta}$. Since the BD-IHMM specifies sums over infinitely large partitions of $\boldsymbol{\beta}$ to compute the $\boldsymbol{\beta}^*_m$, we employ a high-fidelity approximation via truncating $\boldsymbol{\beta}$ once its sum becomes very close to 1, as proposed in (Ishwaran & James, 2002). With these steps, the posterior for a given v_t hidden state assignment invokes a truncated analog to the familiar Chinese Restaurant process for Dirichlet process inference twice, once to account for the transition to state v_t, and once to account for the transition to the next state:

$$P(v_t = m \,|\, \boldsymbol{v}_{\backslash t}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\theta}, y_t, \alpha_0, \xi) \propto \\ p(y_t \,|\, \theta_m) \left(c_{v_{t-1}m} + \alpha_0 \beta^*_{v_{t-1}m} \right) \\ \cdot \frac{\left(\begin{array}{r} c_{mv_{t+1}} + \alpha_0 \beta^*_{mv_{t+1}} \\ + \delta(v_{t-1}=m)\delta(m=v_{t+1}) \end{array} \right)}{c_{m\cdot} + \alpha_0 + \delta(v_{t-1}=m)}, \quad (5)$$

where c_{mn} is the count of inferred transitions between hidden states m and n in $\boldsymbol{v}_{\backslash t}$, and the $\cdot$ index in $c_{m\cdot}$ expands that term to $\sum_{n=1}^{M} c_{mn}$. The delta functions increment the transition counts in cases where self transitions are considered. Note: for simplicity of notation, we will reuse c_{mn} for the count of $m \to n$ transitions throughout $\boldsymbol{v}$ in all forthcoming expressions.

In cases where H is conjugate to $P(y_t \,|\, \theta_m)$, explicit instantiation of the θ_m emission model parameters is not necessary. For brevity, this paper omits further discussion of the emission model inference, which is no different from ordinary hierarchical mixture model inference.

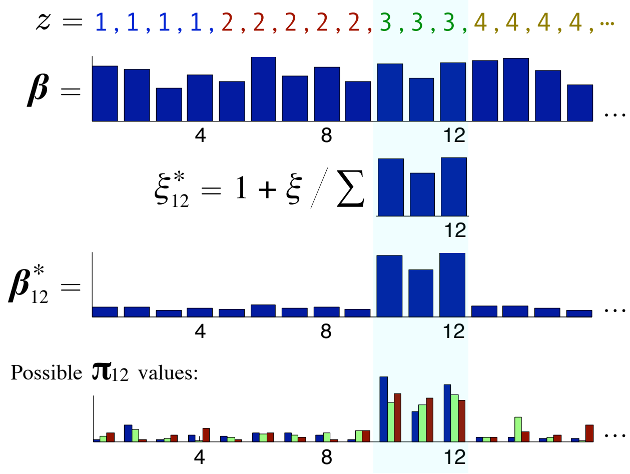


Figure 3: A depiction of the BD-IHMM's process for generating transition probabilities—here, for transitions out of hidden state 12. This state has been assigned block label $z_{12} = 3$. (Note: hidden state labels $\boldsymbol{z}$ are sorted here for clarity, and in the actual model there are a countably infinite number of states assigned to each block.) The shared transition probabilities prior parameter $\boldsymbol{\beta}$ is modified by multiplying proportions assigned to states in block 3 by the scalar value ξ^*_{12}, then renormalizing, yielding a new block-specific parameter $\boldsymbol{\beta}^*_{12}$. Next, probabilities for transitions out of state 12 are drawn conditioned on $\boldsymbol{\beta}^*_{12}$ and the concentration parameter α_0; some possible outcomes are like-colored bars in the bottom graph. Note transitions to states in block 3 are clearly favored.

3.2 Shared transition probabilities prior

For convenient bookkeeping, the "direct assignment" inference method does not keep some useful information. In particular, in order to sample the posterior for $\boldsymbol{\beta}$, we first need to know how many times q_{mn} a particular transition $m \to n$ was selected due to the v_t sampler "landing on" the original $\alpha_0\beta^*_{mn}$ mass allocated to that transition in (5) rather than transition count mass accumulated subsequently. In (Teh et al., 2006), referencing (Antoniak, 1974), this is shown to be distributed as

$$P(q_{mn}|c_{mn},\boldsymbol{z},\boldsymbol{\beta}^*,\alpha_0) = \\ s(c_{mn},q_{mn})(\alpha_0\beta^*_{mn})^{q_{mn}}\frac{\Gamma(\alpha_0\beta^*_{mn})}{\Gamma(c_{mn}+\alpha_0\beta^*_{mn})}, \quad (6)$$

where $s(c_{mn},q_{mn})$ is the unsigned Stirling number of the first kind.

We also require a partial instantiation of $\boldsymbol{\rho}$, the discrete distribution yielding block assignments for hidden states. If $w_k = \sum_{m=1}^{M}\delta(z_m = k)$ when summing only over the M unique states visited during the hidden state sequence $\boldsymbol{v}$, then ρ_k, the probability of drawing a particular block label k out of the K unique block labels belonging to those states, as well as ρ_{new}, the probability of drawing any novel block label, is Dirichlet distributed as:

$$(\rho_1,\ldots,\rho_K,\rho_{\text{new}}) \sim \text{Dir}(w_1,\ldots,w_k,\zeta). \quad (7)$$

After drawing q_{mn} and $\boldsymbol{\rho}$, we sample the posterior for both (a) the $\boldsymbol{\beta}$ terms corresponding to the M visited states and (b) β_{new}, the sum of $\boldsymbol{\beta}$ terms for all hitherto unvisited states. Let K be the number of unique blocks visited in $\boldsymbol{v}$ and $r_k = \sum_{m=1,n=1}^{M} q_{mn}\delta(z_m = k)\delta(z_n = k)$, a sum of within-block transitions. If we now compute the ξ^*_m as $1+\xi/(\rho_{z_m}\beta_{\text{new}}+\sum_{n=1}^{M}\beta_n\delta(z_n\!=\!z_m))$, thereby marginalizing over block assignments for the unvisited states, our posterior can be formulated as:

$$P(\beta_1,\ldots,\beta_M,\beta_{\text{new}}\,|\,\boldsymbol{q},\boldsymbol{z},\xi,\gamma) \propto \\ \text{Dir}(\beta_1,\ldots,\beta_M,\beta_{\text{new}};q_{\cdot 1},\ldots,q_{\cdot M},\gamma) \\ \cdot\prod_{k=1}^{K}\left[1+\xi\Bigg/\left(\rho_k\beta_{\text{new}}+\sum_{n=1}^{M}\beta_n\delta(z_n\!=\!k)\right)\right]^{r_k}. \quad (8)$$

Sampling this posterior is simplified if the $\beta_1,\ldots,\beta_M,\beta_{\text{new}}$ are transformed to a new set of variables $G_1,\ldots,G_K,g_1,\ldots,g_M,\beta_{\text{new}}$, where

$$G_k = \textstyle\sum_{n=1}^{M}\beta_n\delta(z_n\!=\!k), \qquad g_m = \frac{\beta_m}{G_{z_m}}; \quad (9)$$

a block-wise sum of and within-block proportions of $\boldsymbol{\beta}$ elements respectively. It can be shown that the g_m belonging to a single block are Dirichlet distributed with corresponding parameters $q_{\cdot m}$, and that the G_k and β_{new} have a density proportional to

$$\beta_{\text{new}}^{\gamma-1}\prod_{k=1}^{K}G_k^{-1+\sum_{n=1}^{M}q_{\cdot n}\delta(z_n=k)}\left(1+\frac{\xi}{\rho_k\beta_{\text{new}}+G_k}\right)^{r_k}. \quad (10)$$

We sample the above multidimensional density on the K-simplex with the Multiple-Try Metropolis algorithm (Liu et al., 2000), although for large $q_{\cdot n}$, scaling G_k to be proportional to $\sum_{n=1}^{M}q_{\cdot n}$ appears to yield a very close approximation to estimates of its mean.

Once sampled, the G_k and g_m variables are transformed back into β_m proportions, and β_{new} is subdivided into several additional β_m proportions for unvisited states via a truncated version of the stick-breaking process in Equation 1.

3.3 Hidden state block assignments

The posterior over assigning one of the M visited hidden states m to one of the blocks,

$$P(z_m\,|\,\boldsymbol{\rho},\boldsymbol{v},\boldsymbol{\beta},\alpha_0,\xi) \propto P(z_m\,|\,\boldsymbol{\rho})\,P(\boldsymbol{v}\,|\,\boldsymbol{z},\boldsymbol{\beta},\alpha_0,\xi),$$

has two components. The left term is the prior probability of the block assignment, which may be sampled via a truncated stick-breaking expansion of the $\boldsymbol{\rho}$ proportions computed in the prior section. The right term is the likelihood of the sequence of data assignments to hidden states $\boldsymbol{v}$. For smaller problems, the evaluation of the second term can be shown to be $\mathcal{O}(M^2)$ as

$$P(\boldsymbol{v} \,|\, \boldsymbol{z}, \boldsymbol{\beta}, \alpha_0, \xi) = \prod_{m=1}^{M} \frac{\prod_{n=1}^{M} \frac{\Gamma(c_{mn}+\alpha_0\beta^*_{mn})}{\Gamma(\alpha_0\beta^*_{mn})}}{\frac{\Gamma(c_{m\cdot}+\alpha_0)}{\Gamma(\alpha_0)}}. \quad (11)$$

Note that the values of the β^*_{mn} change with different block assignments z_m, as detailed in (4). For larger problems, a different strategy for sampling the $\boldsymbol{z}$ posterior, inspired by the Swendsen-Wang algorithm for MCMC on Potts models (Edwards & Sokal, 1988), changes labels for multiple hidden states simultaneously, resulting in vastly faster mixing. Space constraints permit only a sketch of this bookkeeping-intensive technique: for each pair of states, we sample an auxiliary variable indicating whether the bias for transitions between the same block was responsible for any of the transitions between both states. If this is the case, both states and any other states so connected to them must have the same label—in other words, their new labels can be resampled simultaneously.

3.4 Hyperparameters

The hyperparameters α_0, ξ, ζ, and γ are positive quantities for which we specify Gamma priors. For α_0, the sampling method described in (Teh et al., 2006) for the analogous IHMM parameter is directly applicable. In the case of ξ, the likelihood term is again (11), which does not permit straightforward sampling. Fortunately, its posterior is amenable to Metropolis-Hastings sampling and appears to exhibit rapid mixing. Meanwhile, conditioned on the number of visited blocks K, the ζ posterior may be sampled via the standard techniques of (Escobar & West, 1995) or (Rasmussen, 2000) (see also (Teh et al., 2006)).

In traditional hierarchical Dirichlet processes, inference for the parameter analogous to the BD-IHMM's γ relies on marginalizing away $\boldsymbol{\beta}$. The required integration is complicated in the BD-IHMM by the summations used in computing $\boldsymbol{\beta}^*_m$ in (4). We use numerical methods to capture the γ likelihood instead. When the summed counts $q_{\cdot n}$ described earlier tend to be large, the auxiliary variable-based γ sampling scheme for ordinary HDPs described in (Teh et al., 2006) can be shown to represent a reasonable approximation of the BD-IHMM's data generating process, at least with respect to the effects of γ. Because large $q_{\cdot n}$ are not always present, though, especially for shorter sequences, we achieve a finer approximation by applying a multiplicative adjustment to the γ value used in this method's likelihood density. This factor, derived from tens of thousands of runs of the BD-IHMM's generative process with varying hyperparameters, is a function of ζ and M. The use of a multiplicative adjustment allows the kind of sampling techniques applied to ζ and α_0 to be used with minimal modification.

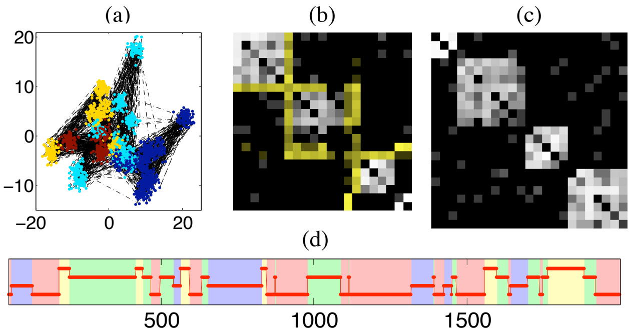


Figure 4: Results on one 2000-step 2-D synthetic dataset (a) exhibiting four sub-behaviors: dots (colored by sub-behavior) are observations, black dotted lines connect them in sequence. The matrix of training-set transition counts inferred by the IHMM (b) shows conflation of at least three pairs of hidden states (yellow shading); the BD-IHMM learns the correct block structure (c). In (d), the sequence of sub-behaviors executed in the training data: inferred (red line) and ground truth (background shading).

4 Experiments

4.1 Artificial data

We compare the results of IHMM and BD-IHMM inference on 100 randomly-generated 2-D datasets. Each is generated by sampling the number of "sub-behaviors" the data should exhibit, then the numbers of states in each sub-behavior. These ranged from 2-4 and 3-9 respectively. Means of 2-D spherical Gaussian emission models ($\sigma = 1$) were drawn from a larger, sub-behavior specific Gaussian ($\sigma = 6$), whose mean in turn was drawn from another Gaussian ($\sigma = 4$). Transition probabilities between hidden states favored within-block transitions at a rate of 98%, giving hidden state sequences within-block "dwelling half lives" of around 34 time steps. Each dataset had 2000 steps of training data and 2000 steps of test data. Figure 4(a) shows one set of sampled observations.

On each dataset we performed IHMM and BD-IHMM inference with vague Gamma priors on all hyperparameters. Emission models used by both the IHMM and BD-IHMM were the same spherical Gaussians used to sample the data.

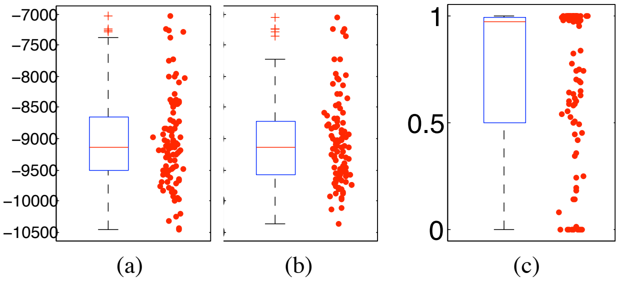


Figure 5: Test-set log probabilities for the 100 synthetic datasets for models learned by (a) the BD-IHMM and (b) the IHMM; these were not significantly different for this task (two sample t-test, $p = 0.8$). In (c), adjusted Rand index scores comparing sub-behavior labels inferred for the training data with ground truth (c.f. Fig. 4(d)); over half have scores greater than 0.95. Scores of 0 typically correspond to complete undersegmentations, i.e. all data associated with just one cluster.

Well after sufficient Gibbs sampling iterations for both models to converge, we simply stopped the inference and selected the last-sampled models for evaluation. We “froze” these models by computing the maximum likelihood transition probabilities and emission model parameters from these draws. We then applied standard HMM techniques to these to compute the likelihood of test-set process data, conditioned on the restriction that inferred trajectories could not visit states that were not visited in the inferred training-set trajectories. IHMM evaluation in (Beal et al., 2002) is more elaborate: it allows the IHMM to continue learning about new data encountered during testing. We chose our simpler scheme because we consider it adequate to reveal whether both models have learned useful dynamics, and because our approach facilitates rapid evaluation on many randomly-generated datasets. Figure 5 shows that, overall, both the BD-IHMM and the IHMM are able to model the dynamics of this data equally well. Nevertheless, as shown in Figure 4(b), the IHMM could be “tricked” into conflating states belonging to separate sub-behaviors, while the BD-IHMM inferred the proper structure.

Given the hidden state labels $v_1, \ldots, v_T$ inferred for data sequences, we can assign block labels to each observation as $z_{v_1}, \ldots, z_{v_T}$. If each block corresponds to a different “sub-behavior” that the generative process executes, this new labeling is an inferred classification or partitioning of the data by the behavior that created it. These partitions may be compared with the true pattern of behaviors known to generate the data. We employ the adjusted Rand index, a partition comparing technique, in this task (Hubert & Arabie, 1985).

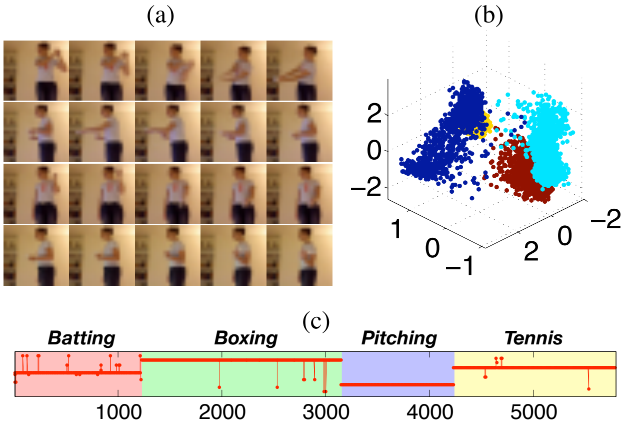


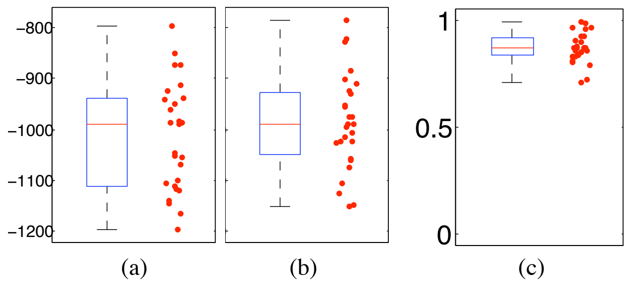
Figure 6: (a) Selected downscaled video frames for (top to bottom) batting, boxing, pitching and tennis swing gestures. (b) First three dimensions of video frames’ PCA projections, colored by gesture type. (c) Sequences of sub-behaviors executed in one set of training data: inferred by one of the model runs (red line) and ground truth (background shading). Each sub-behavior actually comprises numerous video clips of the same gesture.

The numerous index scores near 1 in Figure 5 indicate frequent close matches between the inferred classification and the ground truth.

4.2 Video gesture classification

We collected multiple video clips of a person executing four different gestures for a motion-activated video game (Figure 6). After downscaling the color video frames to 21 × 19 pixels, we projected the frames onto their first four principal components to create data for the IHMM and BD-IHMM algorithms.

For inference, parameters were similar to the artificial data experiment, except here the emission models were 4-D spherical Gaussians ($\sigma = 0.275$). We repeated a 9-way cross-validation scheme three times to collect results over multiple trials; training sets contained around 6,000 observations. Subjecting these results to the same analysis as the artificial data reveals similar compared test-set likelihoods and favorable training-set sub-behavior labeling performance (Figure 7). Both models allocated around 45 hidden states to describe the training data (combined mean: 44.5, $\sigma = 5.0$). We note that since both the BD-IHMM and the IHMM use multiple states to describe each gesture, inferred hidden state trajectories do not usefully identify separate sub-behaviors in the data: adjusted Rand indices comparing the IHMM’s inferred trajectory labeling to ground truth sub-behavior labeling are poor ($\mu = 0.28$, $\sigma = 0.036$).



(a) (b) (c)

Figure 7: Video gesture dataset log probabilities for models learned by (a) the BD-IHMM and (b) the IHMM; these were not significantly different (two sample t-test, $p = 0.3$). In (c), adjusted Rand index scores comparing sub-behavior labels inferred for the training data with ground truth (c.f. Fig. 6(c)). Most errors in labeling were attributable to one of the four sub-behaviors being split into two.

4.3 Musical theme labeling

To test the limits of the BD-IHMM's sub-behavior identification ability, we used the model to identify musical themes in a rock song (Collective Soul, 1993). Apart from the challenge of modeling acoustic information, this dataset is difficult because the themes are lengthy, and no theme is repeated more than four times contiguously; some sections are only played once. It is useful to learn whether the BD-IHMM can recognize separate sub-behaviors in real-world data with such limited repetition.

We prepared a representation of the musical data by computing the cepstrums of 113ms windows (yielding 2714 observations) of a monophonic signal created by combining both stereo channels (Childers et al., 1977). We isolated a distinctive quefrency band between 0.9ms and 3.2ms, smoothed the band across time and quefrency with a 2-D Gaussian kernel, then binned the quefrencies in each window into 20-bin histograms, which, finally, we normalized to sum to 1. The resulting signal appears in Figure 8.

For the hidden-state emission model for the normalized histograms, we selected 20-D Dirichlet distributions with a shared, narrow fixed precision and means limited to a fixed collection of 500 means generated by running K-means on the quefrency histograms. Forsaking actual Dirichlet mean estimation for this somewhat ad-hoc, discrete prior enables simple and rapid marginalization of the emission models.

We performed BD-IHMM inference on the dataset 29 times, achieving sub-behavior labelings like those shown in Figure 8. Fluctuations in the labeling can be attributed in part to vocal variation in the music. We compared the labelings with human-generated ground truth and achieved middling adjusted Rand index scores ($\mu = 0.38$, $\sigma = 0.056$) due mainly to undersegmentation; whether the representation exhibited meaningful variation for different themes may also be an issue. Nevertheless, a qualitative evaluation of the labelings consistently reveals the discovery of considerable theme structure, as exemplified in Figure 8. We conclude that the BD-IHMM is capable of isolating sub-behaviors in datasets with limited repetition, though a task like this one may approach the minimum threshold for this capability.

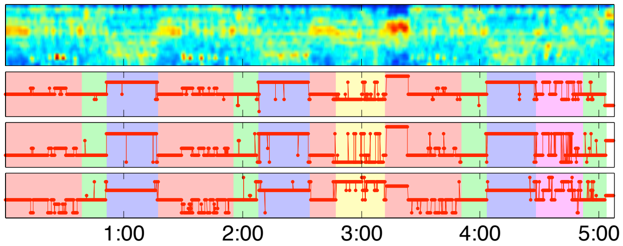


Figure 8: Data and results for the musical theme labeling task. At top, the time series used as input for the model: normalized quefrency histograms (columns). Below, three sub-behavior (i.e. musical theme) labelings for the data; inferred labels (red lines) are plotted atop ground truth (background shading).

4.4 Non-negative integer-weighted graph partitioning

Due to the structure of the BD-IHMM, the counts of transitions c_{mn} are a sufficient statistic for inferring the hidden state block labels $\boldsymbol{z}$; the observations and the actual ordering of the transitions in the process's hidden state trajectory are irrelevant. These counts may be represented as non-negative integer edge weights on a directed graph whose vertices correspond to the hidden states, and analogously block label inference may be understood as a partitioning of this graph. With this analogy, we show how parts of the BD-IHMM and its inference machinery may be applied to a different problem domain.

Consider the U.S. Census Bureau's 2000 dataset on the daily commuting habits of Pennsylvania residents (United States Census Bureau, 2000). This dataset takes the form of a matrix of counts c_{mn} of the number of people commuting each day between municipalities indexed by m and n. We may assume that this matrix has a block-diagonal structure, since people are more likely to commute between areas of shared economic interest. By finding block labels for the municipalities based on this data, we can identify these regions.

We apply the Swendsen-Wang derived label sampler

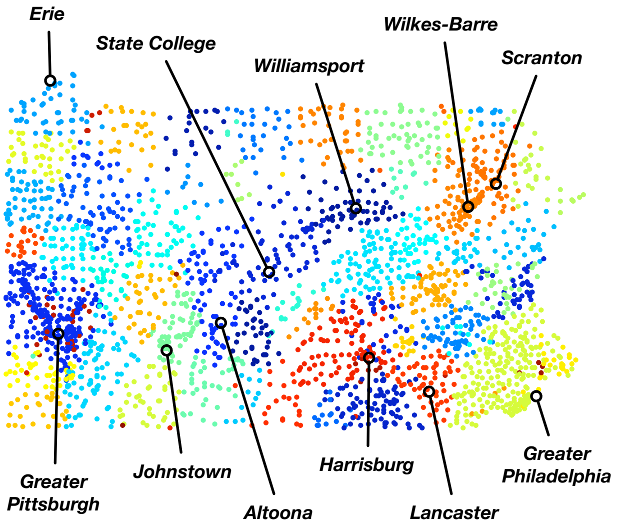


Figure 9: "Segmentsylvania" map showing locations of Pennsylvania municipalities in the U.S. Census Bureau's 2000 daily commuter dataset (United States Census Bureau, 2000). Colors show vertex groupings achieved by applying BD-IHMM hidden state block assignment inference to the commute graph. Nearly all groupings are spatially localized; some colors are reused or similar due to the large number of clusters.

to the 2580-vertex commute graph, achieving the labeling result in Figure 9. 97% of the municipalities are in blocks with 10 or more members, of which there are 55. We offer no quantitative evaluation of this result here, but qualitatively the blocks are nearly all geographically localized and centered around middle- and large-sized urban centers, despite the algorithm's having no access to geographic information.

5 Conclusion

We have demonstrated a generalization of the Infinite HMM of Beal et al. (Beal et al., 2002) (also (Teh et al., 2006)) whose prior induces a clustering effect in the hidden state transition dynamics. Posterior samples of this model may be used to identify characteristic "sub-behaviors" within a data sequence and to partition input sequences according to these behaviors. We have also shown that components of the BD-IHMM may be used to partition non-negative integer-weighted graphs. Future work may explore online learning as well as applications of BD-IHMM components to relational data analysis.

Acknowledgements

We thank Laura Halderman and Reed Taylor for playing video games on camera. TSL was supported by NSF IIS-073206; TS by NSF IGERT fellowship DGE0549352.

References

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, *2*, 1152–1174.

Beal, M., Ghahramani, Z., & Rasmussen, C. (2002). The infinite hidden Markov model. *NIPS 14*.

Childers, D., Skinner, D., & Kemerait, R. (1977). The cepstrum: A guide to processing. *Proceedings of the IEEE*, *65*, 1428–1443.

Collective Soul (1993). "Shine". *Hints Allegations and Things Left Unsaid*. CD. Atlantic.

Edwards, R., & Sokal, A. (1988). Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Review D*, *38*, 2009.

Escobar, M., & West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Stat. Assoc.*, *90*, 577–588.

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2008). An HDP-HMM for systems with state persistence. *Proceedings of the 25th International Conference on Machine Learning*.

Fox, E., Sudderth, E., & Willsky, A. (2007). Hierarchical Dirichlet processes for tracking maneuvering targets. *Proceedings of the 2007 International Conference on Information Fusion*.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Ishwaran, H., & James, L. (2002). Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information. *Journal of Computational and Graphical Statistics*, *11*, 508–532.

Liu, J., Liang, F., & Wong, W. (2000). The multiple-try method and local optimization in Metropolis sampling. *J. Amer. Stat. Assoc.*, *96*, 561–573.

Ni, K., & Dunson, D. (2007). Multi-task learning for sequential data via iHMMs and the nested Dirichlet process. *Proceedings of the 24th International Conference on Machine Learning*.

Pekergin, N., Dayar, T., & Alparslan, D. (2005). Componentwise bounds for nearly completely decomposable Markov chains using stochastic comparison and reordering. *European Journal of Operational Research*, *165*, 810–825.

Rasmussen, C. (2000). The infinite Gaussian mixture model. *NIPS 12*.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*, 639–650.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical Dirichlet processes. *J. Amer. Stat. Assoc.*, *101*, 1566–1581.

United States Census Bureau (2000). MCD/county-to-MCD/county worker flow files. Online, `http://www.census.gov/population/www/cen2000/commuting/mcdworkerflow.html`. Retrieved 26 September 2008.