# Bayesian Nonparametric Methods for Learning Markov Switching Processes

Emily B. Fox, *Member, IEEE,* Erik B. Sudderth, *Member, IEEE,*

Michael I. Jordan, *Fellow, IEEE,* and Alan S. Willsky, *Fellow, IEEE*

## Abstract

Markov switching processes, such as the hidden Markov model (HMM) and switching linear dynamical system (SLDS), are often used to describe rich dynamical phenomena. They describe complex behavior via repeated returns to a set of simpler models: imagine a person alternating between *walking*, *running*, and *jumping* behaviors, or a stock index switching between regimes of high and low volatility. Classical approaches to identification and estimation of these models assume a fixed, pre-specified number of dynamical models. We instead examine Bayesian nonparametric approaches that define a prior on Markov switching processes with an unbounded number of potential model parameters (i.e., Markov modes). By leveraging stochastic processes such as the beta and Dirichlet process, these methods allow the data to drive the complexity of the learned model, while still permitting efficient inference algorithms. They also lead to generalizations which discover and model dynamical behaviors shared among multiple related time series.

## Index Terms

Bayesian nonparametric methods, hidden Markov model, Markov jump linear system, time series.

## I. INTRODUCTION

A core problem in statistical signal processing is the partitioning of temporal data into segments, each of which permits a relatively simple statistical description. This segmentation problem arises in a

E. Fox is with the Department of Statistical Science, Duke University, Durham, NC, 27708 USA e-mail: fox@stat.duke.edu. E. Sudderth is with the Department of Computer Science, Brown University, Providence, RI, 02912 USA e-mail: sudderth@cs.brown.edu. M. Jordan is with the Department of Electrical Engineering and Computer Science, and Department of Statistics, University of California, Berkeley, CA, 94720 USA e-mail: jordan@eecs.berkeley.edu. A. Willsky is with the Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 02139 USA e-mail: willsky@mit.edu.

variety of applications, in areas as diverse as speech recognition, computational biology, natural language processing, finance, and cryptanalysis. While in some cases the problem is merely that of detecting temporal changes—in which case the problem can be viewed as one of changepoint detection—in many other cases the temporal segments have a natural meaning in the domain and the problem is to recognize recurrence of a meaningful entity (e.g., a particular speaker, a gene, a part of speech, or a market condition). This leads naturally to state-space models, where the entities that are to be detected are encoded in the state.

The classical example of a state-space model for segmentation is the hidden Markov model (HMM) [1]. The HMM is based on a discrete state variable and on the probabilistic assertions that the state transitions are Markovian and that the observations are conditionally independent given the state. In this model, temporal segments are equated with states, a natural assumption in some problems but a limiting assumption in many others. Consider, for example, the dance of a honey bee as it switches between a set of *turn right*, *turn left*, and *waggle* dances. It is natural to view these dances as the temporal segments, as they "permit a relatively simple statistical description." But each dance is not a set of conditionally independent draws from a fixed distribution as required by the HMM—there are additional dynamics to account for. Moreover, it is natural to model these dynamics using continuous variables. These desiderata can be accommodated within the richer framework of *Markov switching processes*, specific examples of which include the *switching vector autoregressive* (VAR) process and the *switching linear dynamical system* (SLDS). These models distinguish between the discrete component of the state, which is referred to as a "mode," and the continuous component of the state, which captures the continuous dynamics associated with each mode. These models have become increasingly prominent in applications in recent years [2]–[8].

While Markov switching processes address one key limitation of the HMM, they inherit various other limitations of the HMM. In particular, the discrete component of the state in the Markov switching process has no topological structure (beyond the trivial discrete topology). Thus it is not easy to compare state spaces of different cardinality and it is not possible to use the state space to encode a notion of similarity between modes. More broadly, many problems involve a collection of state-space models (either HMMs or Markov switching processes), and within the classical framework there is no natural way to talk about overlap between models. A particular instance of this problem arises when there are multiple time series, and where we wish to use overlapping subsets of the modes to describe the different time series. In Sec. IV we discuss a concrete example of this problem where the time series are motion-capture videos of humans engaging in exercise routines, and where the modes are specific exercises, such as "jumping

jacks," or "squats." We aim to capture the notion that two different people can engage in the same exercise (e.g., jumping jacks) during their routine.

To address these problems we need to move beyond the simple discrete Markov chain as a description of temporal segmentation. In this article, we describe a richer class of stochastic processes known as *combinatorial stochastic processes* that provide a useful foundation for the design of flexible models for temporal segmentation. Combinatorial stochastic processes have been studied for several decades in probability theory (see, e.g., [9]), and they have begun to play a role in statistics as well, most notably in the area of Bayesian nonparametric statistics where they yield Bayesian approaches to clustering and survival analysis (see, e.g., [10]). The work that we present here extends these efforts into the time series domain. As we aim to show, there is a natural interplay between combinatorial stochastic processes and state-space descriptions of dynamical systems.

Our primary focus is on two specific stochastic processes—the *Dirichlet process* and the *beta process*—and their role in describing modes in dynamical systems. The Dirichlet process provides a simple description of a clustering process where the number of clusters is not fixed a priori. Suitably extended to a *hierarchical Dirichlet process* (HDP), this stochastic process provides a foundation for the design of state-space models in which the number of modes is random and inferred from the data. In Sec. II, we discuss the HDP and its connection to the HMM. Building on this connection, Sec. III shows how the HDP can be used in the context of Markov switching processes with conditionally linear dynamical modes. Finally, in Sec. IV we discuss the beta process and show how it can be used to capture notions of similarity among sets of modes in modeling multiple time series.

## II. HIDDEN MARKOV MODELS

The hidden Markov model, or *HMM*, generates a sequence of latent modes via a discrete-valued Markov chain [1]. Conditioned on this mode sequence, the model assumes that the observations, which may be discrete or continuous valued, are independent. The HMM is the most basic example of a Markov switching process, and forms the building block for more complicated processes examined later.

### A. Finite HMM

Let $z_t$ denote the *mode* of the Markov chain at time $t$, and $\pi_j$ the mode-specific *transition distribution* for mode $j$. Given the mode $z_t$, the observation $y_t$ is conditionally independent of the observations and
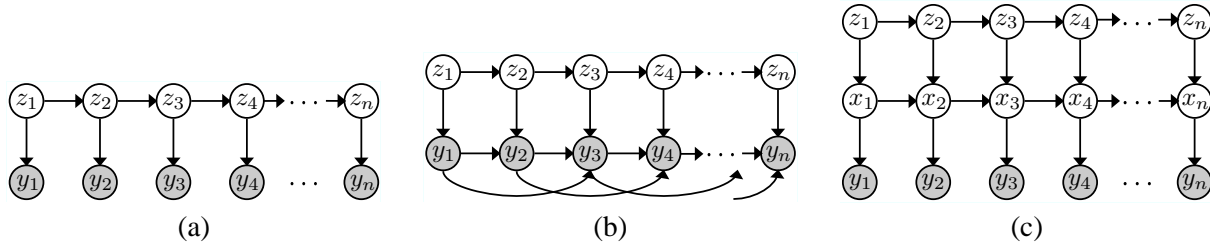
Fig. 1. Graphical representations of three Markov switching processes: (a) hidden Markov Model (HMM), (b) order 2 switching vector autoregressive (VAR) process, and (c) switching linear dynamical system (SLDS). For all models, a discrete-valued Markov process $z_t$ evolves as $z_{t+1}|\{\pi_k\}_{k=1}^{K}, z_t \sim \pi_{z_t}$. For the HMM, observations are generated as $y_t|\{\theta_k\}_{k=1}^{K}, z_t \sim F(\theta_{z_t})$, whereas the switching VAR(2) process assumes $y_t \sim \mathcal{N}(A_1^{(z_t)}y_{t-1} + A_2^{(z_t)}y_{t-2}, \Sigma^{(z_t)})$. The SLDS instead relies on a latent, continuous-valued Markov state $x_t$ to capture the history of the dynamical process as specified in Eq. (15).

modes at other time steps. The generative process can be described as[1]

$$z_t \mid z_{t-1} \sim \pi_{z_{t-1}}$$

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

(1)

for an indexed family of distributions $F(\cdot)$ (e.g., multinomial for discrete data or multivariate Gaussian for real, vector-valued data), where $\theta_i$ are the *emission parameters* for mode $i$. The directed graphical model associated with the HMM is shown in Fig. 1(a).

One can equivalently represent the HMM via a set of *transition probability measures* $G_j = \sum_{k=1}^{K} \pi_{jk}\delta_{\theta_k}$, where $\delta_\theta$ is a unit mass concentrated at $\theta$. Instead of employing *transition distributions* on the set of integers (i.e., modes) which index into the collection of emission parameters, we operate directly in the parameter space $\Theta$, and transition between emission parameters with probabilities $\{G_j\}$. Specifically, let $j_{t-1}$ be the unique emission parameter index $j$ such that $\theta'_{t-1} = \theta_j$. Then,

$$\theta'_t \mid \theta'_{t-1} \sim G_{j_{t-1}}$$

$$y_t \mid \theta'_t \sim F(\theta'_t).$$

(2)

Here, $\theta'_t \in \{\theta_1, \ldots, \theta_K\}$ takes the place of $\theta_{z_t}$ in Eq. (1). A visualization of this process is shown by the trellis diagram of Fig. 2.

One can consider a *Bayesian HMM* by treating the transition probability measures $G_j$ as *random*[2], and endowing them with a prior. Since the probability measures are solely distinguished by their weights

---

[1]The notation $x \sim F$ indicates that the random variable $x$ is drawn from a distribution $F$. We use bar notation $x \mid F \sim F$ to specify conditioned upon random variables, such as a random distribution.

[2]Formally, a random measure on a measurable space $\Theta$, with sigma algebra $\mathcal{A}$, is defined as a stochastic process whose index set is $\mathcal{A}$. That is, $G(A)$ is a non-negative random variable for each $A \in \mathcal{A}$.
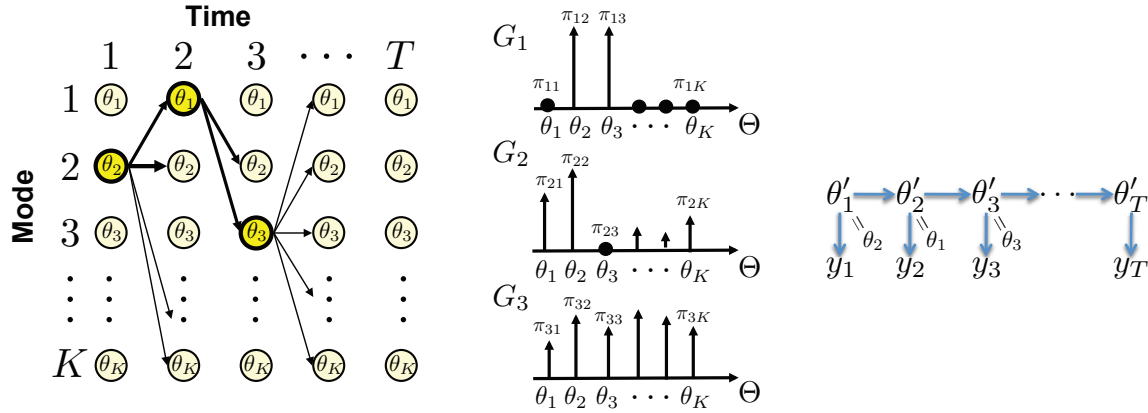
Fig. 2.  *Left:* Trellis representation of an HMM. Each circle represents one of the $K$ possible HMM emission parameters at various time steps. The highlighted circles indicate the selected emission parameter $\theta_t'$ at time $t$, and the arrows represent the set of possible transitions from that HMM mode to each of the $K$ possible next modes. The weights of these arrows indicate the relative probability of the transitions encoded by the mode-specific transition probability measures $G_j$. *Middle:* Transition probability measures $G_1$, $G_2$, and $G_3$ corresponding to the example trellis diagram. *Right:* A representation of the HMM observations $y_t$, which are drawn from emission distributions parameterized by the highlighted nodes.

on the shared set of emission parameters $\{\theta_1, \ldots, \theta_K\}$, we consider a prior that independently handles these components. Specifically, take the weights $\pi_j = [\pi_{j1} \ldots \pi_{jK}]$ (i.e., transition distributions) to be independent draws from a $K$-dimensional Dirichlet distribution,

$$\pi_j \sim \text{Dir}(\alpha_1, \ldots, \alpha_K) \qquad j = 1, \ldots, K, \tag{3}$$

implying that $\sum_k \pi_{jk} = 1$, as desired. Then, assume that the atoms are drawn as $\theta_j \sim H$ for some base measure $H$ on the parameter space $\Theta$. Depending on the form of the emission distribution, various choices of $H$ lead to computational efficiencies via conjugate analysis.

*B. Sticky HDP-HMM*

In the Bayesian HMM of the previous section, we assumed that the number of HMM modes $K$ is known. But what if this is not the case? For example, in the speaker diarization task described in Sec. I, determining the number of speakers involved in the meeting is one of the primary inference *goals*. Moreover, even when a model adequately describes previous observations, it can be desirable to allow new modes to be added as more data are observed. For example, what if more speakers enter the meeting? To avoid restrictions on the size of the mode space, such scenarios naturally lead to priors on probability measures $G_j$ that have an unbounded collection of support points $\theta_k$.

The *Dirichlet process* (DP), denoted by $\mathrm{DP}(\gamma, H)$, provides a distribution over countably infinite probability measures

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad \theta_k \sim H \tag{4}$$

on a parameter space $\Theta$. The weights are sampled via a *stick-breaking construction* [11]:

$$\beta_k = \nu_k \prod_{\ell=1}^{k-1} (1 - \nu_\ell) \qquad \nu_k \sim \mathrm{Beta}(1, \gamma). \tag{5}$$

In effect, we have divided a unit-length stick into lengths given by the weights $\beta_k$: the $k^{th}$ weight is a random proportion $v_k$ of the remaining stick after the first $(k-1)$ weights have been chosen. We denote this distribution by $\beta \sim \mathrm{GEM}(\gamma)$. See Fig. 3 for a pictorial representation of this process.

The Dirichlet process has proven useful in many applications due to its clustering properties, which are clearly seen by examining the *predictive distribution* of draws $\theta_i' \sim G_0$. Because probability measures drawn from a Dirichlet process are discrete, there is a strictly positive probability of multiple observations $\theta_i'$ taking identical values within the set $\{\theta_k\}$, with $\theta_k$ defined as in Eq. (4). For each value $\theta_i'$, let $z_i$ be



Fig. 3. Pictorial representation of the stick-breaking construction of the Dirichlet process.

an indicator random variable that picks out the unique value $\theta_k$ such that $\theta_i' = \theta_{z_i}$. Blackwell and MacQueen [12] derived a Pólya urn representation of the $\theta_i'$:

$$\theta_i' \mid \theta_1', \ldots, \theta_{i-1}' \sim \frac{\gamma}{\gamma + i - 1} H + \sum_{j=1}^{i-1} \frac{1}{\gamma + i - 1} \delta_{\theta_j'}$$

$$\sim \frac{\gamma}{\gamma + i - 1} H + \sum_{k=1}^{K} \frac{n_k}{\gamma + i - 1} \delta_{\theta_k}. \tag{6}$$

This implies the following predictive distribution on the indicator assignment variables:

$$p(z_{N+1} = z \mid z_1, \ldots, z_N, \gamma) = \frac{\gamma}{N + \gamma} \delta(z, K+1) + \frac{1}{N + \gamma} \sum_{k=1}^{K} n_k \delta(z, k). \tag{7}$$

Here, $n_k = \sum_{i=1}^{N} \delta(z_i, k)$ is the number of indicator random variables taking the value $k$, and $K + 1$ is a previously unseen value. The discrete Kronecker delta $\delta(z, k) = 1$ if $z = k$, and $0$ otherwise. The distribution on partitions induced by the sequence of conditional distributions in Eq. (7) is commonly referred to as the *Chinese restaurant process*. Take $i$ to be a customer entering a restaurant with infinitely many tables, each serving a unique dish $\theta_k$. Each arriving customer chooses a table, indicated by $z_i$,
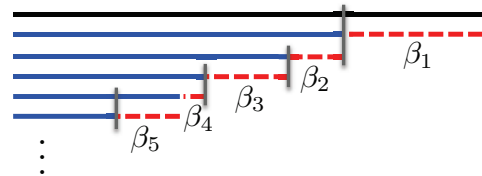
in proportion to how many customers are currently sitting at that table. With some positive probability proportional to $\gamma$, the customer starts a new, previously unoccupied table $K + 1$. From the Chinese restaurant process, we see that the Dirichlet process has a reinforcement property that leads to a clustering at the values $\theta_k$. This representation also provides a means of sampling observations from a Dirichlet process without explicitly constructing the infinite probability measure $G_0 \sim \text{DP}(\gamma, H)$.

One could imagine using the Dirichlet process to define a prior on the set of HMM transition probability measures $G_j$. Taking each transition measure $G_j$ as an independent draw from $\text{DP}(\gamma, H)$ implies that these probability measures are of the form $\sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_{jk}}$, with weights $\pi_j \sim \text{GEM}(\gamma)$ and atoms $\theta_{jk} \sim H$. Assuming $H$ is absolutely continuous (e.g., a Gaussian distribution), this construction leads to transition measures with non-overlapping support (i.e., $\theta_{jk} \neq \theta_{\ell k'}$ with probability one.) Based on such a construction, we would move from one infinite collection of HMM modes to an entirely new collection at each transition step, implying that previously visited modes would *never* be revisited. This is clearly not what we intended. Instead, consider the *hierarchical Dirichlet process* (HDP) [13], which defines a collection of probability measures $\{G_j\}$ on the same support points $\{\theta_1, \theta_2, \dots\}$ by assuming that each discrete measure $G_j$ is a variation on a global discrete measure $G_0$. Specifically, the Bayesian hierarchical specification takes $G_j \sim \text{DP}(\alpha, G_0)$, with $G_0$ itself a draw from a Dirichlet process $\text{DP}(\gamma, H)$. Through this construction, one can show that the probability measures are described as

$$
\begin{aligned}
G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} & \beta \mid \gamma &\sim \text{GEM}(\gamma) & \theta_k \mid H &\sim H. \\
G_j &= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} & \pi_j \mid \alpha, \beta &\sim \text{DP}(\alpha, \beta)
\end{aligned}
\tag{8}
$$

Applying the HDP prior to the HMM, we obtain the *HDP-HMM* of Teh et. al. [13].

Extending the Chinese restaurant process analogy of the Dirichlet process, one can examine a Chinese restaurant *franchise* that describes the partitions induced by the HDP. In the context of the HDP-HMM, the infinite collection of emission parameters $\theta_k$ determine a global menu of shared dishes. Each of these emission parameters is also associated with a single, unique restaurant in the franchise. The HDP-HMM assigns a customer $\theta_t'$ to a restaurant based on the previous customer $\theta_{t-1}'$, since it is this parameter that determines the distribution of $\theta_t'$ (see Eq. (2)). Upon entering the restaurant determined by $\theta_{t-1}'$, customer $\theta_t'$ chooses a table with probability proportional to the current occupancy, just as in the Dirichlet process. The dishes for the tables are then chosen from the global menu $G_0$ based on their popularity throughout the *entire* franchise, and it is through this pooling of dish selections that the HDP induces a *shared* sparse subset of model parameters.

By defining $\pi_j \sim \text{DP}(\alpha, \beta)$, the HDP prior encourages modes to have similar transition distributions.
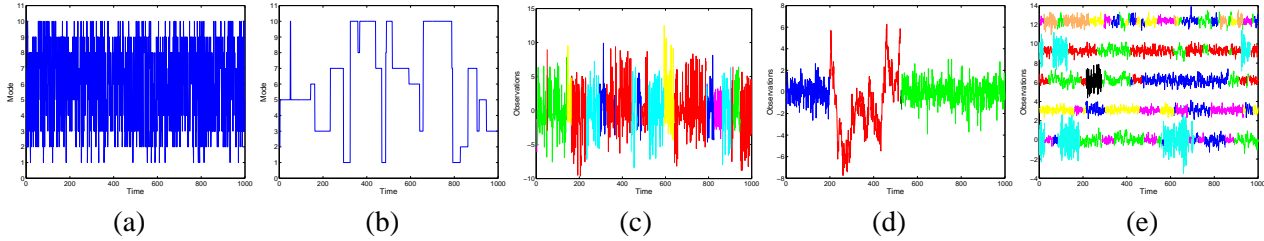
Fig. 4.    (a)-(b) Mode sequences drawn from the HDP-HMM and sticky HDP-HMM priors, respectively. (c)-(e) Observation sequences corresponding to draws from a sticky HDP-HMM, an order 2 HDP-AR-HMM, and an order 1 BP-AR-HMM with five time series offset for clarity. The observation sequences are colored by the underlying mode sequences.

In particular, the mode-specific transition distributions are *identical* in expectation:

$$\mathbb{E}[\pi_{jk} \mid \beta] = \beta_k. \tag{9}$$

Although it is through this construction that a shared sparse mode space is induced, we see from Eq. (9) that the HDP-HMM does not differentiate self–transitions from moves between different modes. When modeling data with mode persistence, the flexible nature of the HDP-HMM prior allows for mode sequences with unrealistically fast dynamics to have large posterior probability, thus impeding identification of a compact dynamical model which best explains the observations. See an example mode sequence in Fig. 4(a). The HDP-HMM places only a small prior penalty on creating an extra mode, but no penalty on that mode having a similar emission parameter to another mode, nor on the time series rapidly switching between two modes. The increase in likelihood by fine-tuning the parameters to the data can counteract the prior penalty on creating this extra mode, leading to significant posterior uncertainty. These rapid dynamics are unrealistic for many real datasets. For example, in the speaker diarization task, it is very unlikely in that two speakers are rapidly switching who is speaking. Such fast-switching dynamics can harm the predictive performance of the learned model since parameters are informed by fewer data points. Additionally, in some applications one cares about the accuracy of the inferred label sequence instead of just doing model averaging. In such applications, one would like to be able to incorporate prior knowledge that slow, smoothly varying dynamics are more likely.

To address these issues, Fox et. al. [14] proposed to instead sample transition distributions $\pi_j$ as:

$$\beta \mid \gamma \sim \mathrm{GEM}(\gamma)$$
$$\pi_j \mid \alpha, \kappa, \beta \sim \mathrm{DP}\left(\alpha + \kappa, \frac{\alpha\beta + \kappa\delta_j}{\alpha + \kappa}\right). \tag{10}$$

Here, $(\alpha\beta + \kappa\delta_j)$ indicates that an amount $\kappa > 0$ is added to the $j^{th}$ component of $\alpha\beta$. This construction increases the expected probability of self-transition by an amount proportional to $\kappa$. Specifically, the
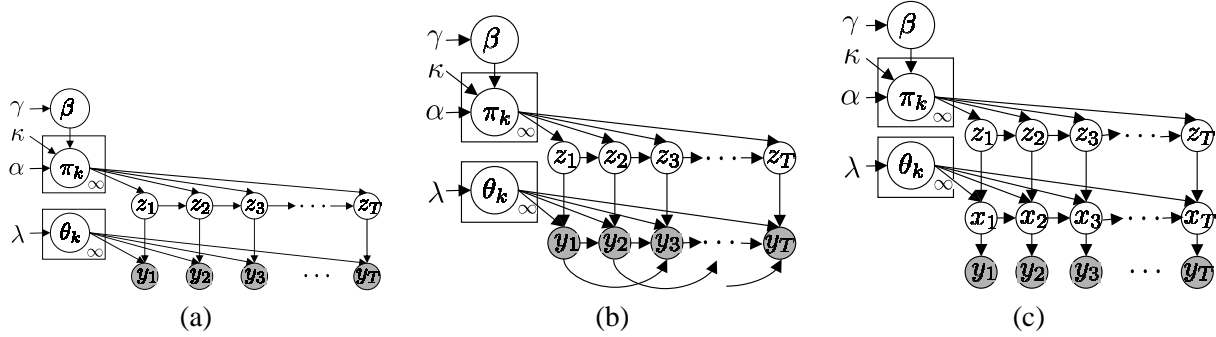
Fig. 5. Graphical representations of three Bayesian nonparametric variants of the Markov switching processes shown in Fig. 1. The *plate* notation is used to compactly represent replication of nodes in the graph [13], with the number of replicates indicated by the number in the corner of the plate. (a) The sticky HDP-HMM. The mode evolves as $z_{t+1}|\{\pi_k\}_{k=1}^{\infty}, z_t \sim \pi_{z_t}$, where $\pi_k|\alpha, \kappa, \beta \sim \mathrm{DP}(\alpha + \kappa, (\alpha\beta + \kappa\delta_k)/(\alpha + \kappa))$ and $\beta|\gamma \sim \mathrm{GEM}(\gamma)$, and observations are generated as $y_t|\{\theta_k\}_{k=1}^{\infty}, z_t \sim F(\theta_{z_t})$. The HDP-HMM of [13] has $\kappa = 0$. (b) A switching VAR(2) process and (c) SLDS, each with a sticky HDP-HMM prior on the Markov switching process. The dynamical equations are as in Eq. (17).

expected set of weights for transition distribution $\pi_j$ is a convex combination of those defined by $\beta$ and mode-specific weight defined by $\kappa$:

$$\mathbb{E}[\pi_{jk} \mid \beta, \kappa] = \frac{\alpha}{\alpha + \kappa}\beta_k + \frac{\kappa}{\alpha + \kappa}\delta(j, k). \quad (11)$$

When $\kappa = 0$ the original HDP-HMM of Teh et. al. [13] is recovered. Because positive $\kappa$ values increase the prior probability $\mathbb{E}[\pi_{jj} \mid \beta]$ of self–transitions, this model is referred to as the *sticky* HDP-HMM. See the graphical model of Fig. 5(a), and the mode sequence sample path of Fig. 4(b). The $\kappa$ parameter is reminiscent of the self-transition bias parameter of the infinite HMM [15], a precursor of the HDP-HMM. However, that paper relied on heuristic, approximate inference methods. The full connection between the infinite HMM and the HDP, as well as the development of a globally consistent inference algorithm, was made in [13], but without a treatment of a self-transition parameter. Note that in the formulation of Fox et. al. [14], the HDP concentration parameters $\alpha$ and $\gamma$, and the sticky parameter $\kappa$, are given priors and are thus learned from the data as well. The flexibility garnered by incorporating learning of the sticky parameter within a cohesive Bayesian framework allows the model to additionally capture fast mode-switching if such dynamics are present in the data.

In [14], the sticky HDP-HMM was applied to the speaker diarization task, which we recall involves segmenting an audio recording into speaker-homogeneous regions, while simultaneously identifying the number of speakers. The data for the experiments consisted of a standard benchmark data set distributed by NIST as part of the Rich Transcription 2004-2007 meeting recognition evaluations [16], with the

observations taken to be the first 19 Mel Frequency Cepstral Coefficients (MFCCs) computed over short, overlapping windows. Because the speaker-specific emissions are not well-approximated by a single Gaussian, a Dirichlet process mixture of Gaussian extension of the sticky HDP-HMM was considered. The sticky parameter proved pivotal in learning such multimodal emission distributions. Combining both the mode-persistence captured by the sticky HDP-HMM, along with a model allowing multimodal emissions, state-of-the-art speaker diarization performance was achieved.

## III. MARKOV JUMP LINEAR SYSTEMS

The HMM's assumption of conditionally independent observations is often insufficient in capturing the temporal dependencies present in many real datasets. Recall the example of the dancing honey bees from Sec. I. In such cases, one can consider more complex Markov switching processes, namely the class of *Markov jump-linear systems* (MJLS), in which each dynamical mode is modeled via a linear dynamical process. Just as with the HMM, the switching mechanism is based on an underlying discrete-valued Markov mode sequence. Switched affine and piecewise affine models, which we do not consider in this paper, instead allow mode transitions to depend on the continuous state of the dynamical system [17].

### A. State Space Models, VAR Processes, and Finite Markov Jump Linear Systems

A state space model provides a general framework for analyzing many dynamical phenomena. The model consists of an underlying state, $\boldsymbol{x}_t \in \mathbb{R}^n$, with linear dynamics observed via $\boldsymbol{y}_t \in \mathbb{R}^d$. A linear time-invariant state space model, in which the dynamics do not depend on time, is given by

$$\boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + \boldsymbol{e}_t \qquad \boldsymbol{y}_t = C\boldsymbol{x}_t + \boldsymbol{w}_t, \tag{12}$$

where $e_t$ and $w_t$ are independent, zero-mean Gaussian noise processes with covariances $\Sigma$ and $R$, respectively. The graphical model for this process is equivalent to that of the hidden Markov model depicted in Fig. 1(a), replacing $z_t$ with $x_t$.

An order $r$ VAR process, denoted by VAR($r$), with observations $\boldsymbol{y}_t \in \mathbb{R}^d$, can be defined as

$$\boldsymbol{y}_t = \sum_{i=1}^{r} A_i \boldsymbol{y}_{t-i} + \boldsymbol{e}_t \qquad \boldsymbol{e}_t \sim \mathcal{N}(0, \Sigma). \tag{13}$$

Here, the observations depend linearly on the previous $r$ observation vectors. Every VAR($r$) process can

be described in state space form by, for example, the following transformation:

$$\boldsymbol{x}_t = \begin{bmatrix} A_1 & A_2 & \ldots & A_r \\ I & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & I & 0 \end{bmatrix} \boldsymbol{x}_{t-1} + \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} \boldsymbol{e}_t \qquad \boldsymbol{y}_t = \begin{bmatrix} I & 0 & \ldots & 0 \end{bmatrix} \boldsymbol{x}_t. \qquad (14)$$

On the other hand, not every state space model may be expressed as a VAR($r$) process for finite $r$ [18].

Building on the HMM of Sec. II-A, we define a *switching linear dynamical system* (SLDS) by

$$z_t \sim \pi_{z_{t-1}}$$
$$\boldsymbol{x}_t = A^{(z_t)} \boldsymbol{x}_{t-1} + \boldsymbol{e}_t(z_t) \quad \boldsymbol{y}_t = C \boldsymbol{x}_t + \boldsymbol{w}_t, \qquad (15)$$

Here, we assume the process noise $\boldsymbol{e}_t(z_t) \sim \mathcal{N}(0, \Sigma^{(z_t)})$ is mode-specific, while the measurement mechanism is not. This assumption could be modified to allow for both a mode-specific measurement matrix $C^{(z_t)}$ and noise $\boldsymbol{w}_t(z_t) \sim \mathcal{N}(0, R^{(z_t)})$. However, such a choice is not always necessary nor appropriate for certain applications, and can have implications on the identifiability of the model. We similarly define a *switching* VAR($r$) process by

$$z_t \sim \pi_{z_{t-1}}$$
$$\boldsymbol{y}_t = \sum_{i=1}^{r} A_i^{(z_t)} \boldsymbol{y}_{t-i} + \boldsymbol{e}_t(z_t). \qquad (16)$$

Both the SLDS and the switching VAR process are contained within the class of MJLS, with graphical model representations shown in Fig. 1(b)-(c). Compare to that of the HMM in Fig. 1(a).

### B. HDP-AR-HMM and HDP-SLDS

In the formulations of the MJLS of Sec. III, it was assumed that the number of dynamical modes was known. However, it is often desirable to relax this assumption in order to provide more modeling flexibility. It has been shown that in such cases, the sticky HDP-HMM can be extended as a Bayesian nonparametric approach to learning both SLDS and switching VAR processes [19], [20]. Specifically, the transition distributions are defined just as in the sticky HDP-HMM. However, instead of independent observations, each mode now has conditionally linear dynamics. The generative processes for the resulting *HDP-AR-HMM* and *HDP-SLDS* are summarized as follows, with an example HDP-AR-HMM observation sequence depicted in Fig. 4(d):

|  | HDP-AR-HMM | HDP-SLDS |
|---|---|---|
| Mode dynamics | $z_t \sim \pi_{z_{t-1}}$ | $z_t \sim \pi_{z_{t-1}}$ |
| Observation dynamics | $\boldsymbol{y}_t = \sum_{i=1}^{r} A_i^{(z_t)} \boldsymbol{y}_{t-i} + \boldsymbol{e}_t(z_t)$ | $\boldsymbol{x}_t = A^{(z_t)} \boldsymbol{x}_{t-1} + \boldsymbol{e}_t(z_t)$ |
|  |  | $\boldsymbol{y}_t = C \boldsymbol{x}_t + \boldsymbol{w}_t$ |

$$(17)$$

Here, $\pi_j$ is as defined in Eq. (10). The issue, then, is in determining an appropriate prior on the dynamic parameters. In [19], a conjugate matrix-normal inverse-Wishart (MNIW) prior [21] was proposed for the dynamic parameters $\{A^{(k)}, \Sigma^{(k)}\}$ in the case of the HDP-SLDS, and $\{A_1^{(k)}, \ldots, A_r^{(k)}, \Sigma^{(k)}\}$ for the HDP-AR-HMM. The HDP-SLDS additionally assumes an inverse-Wishart prior on the measurement noise $R$; however, the measurement matrix, $C$, is fixed for reasons of identifiability. The MNIW prior assumes knowledge of either the autoregressive order $r$, or the underlying state dimension $n$, of the switching VAR process or SLDS, respectively. Alternatively, Fox et. al. [20] explore an automatic relevance determination (ARD) sparsity-inducing prior [22]–[24] as a means of learning MJLS with variable order structure. The ARD prior penalizes non-zero components of the model through a zero-mean Gaussian prior with gamma-distributed precision. In the context of the HDP-AR-HMM and HDP-SLDS, a maximal autoregressive order or SLDS state dimension is assumed. Then, a structured version of the ARD prior is employed so as to drive entire VAR lag blocks $A_i^{(k)}$ or columns of the SLDS matrix $A^{(k)}$ to zero, allowing one to infer non-dynamical components of a given dynamical mode.

The previous work of Fox et al. [8] considered a related, yet simpler formulation for modeling a maneuvering target as a fixed LDS driven by a switching exogenous input. Since the number of maneuver modes was assumed unknown, the exogenous input was taken to be the emissions of a HDP-HMM. This work can be viewed as an extension of the work by Caron et. al. [25] in which the exogenous input was an independent noise process generated from a DP mixture model. The HDP-SLDS of [19] is a departure from these works since the dynamic parameters themselves change with the mode, providing a much more expressive model.

In [19], the utility of the HDP-SLDS and HDP-AR-HMM was demonstrated on two different problems: detecting changes in the volatility of the IBOVESPA stock index, and segmenting sequences of honey bee dances. The dynamics underlying both of these data sets appear to be quite complex, yet can be described by repeated returns to simpler dynamical models, and as such have been modeled with Markov switching processes [26], [27]. Without pre-specifying domain-specific knowledge, and instead simply relying on a set of observations along with weakly informative hyperprior settings, the HDP-SLDS and HDP-AR-HMM were able to discover the underlying structure of the data with performance competitive with these alternative methods, and consistent with domain expert analysis.

## IV. MULTIPLE RELATED TIME SERIES

In many applications, one would like to discover and model dynamical behaviors which are shared among several related time series. By jointly modeling such time series, one can improve parameter estimates, especially in the case of limited data, and find interesting structure in the relationships between the time series. Assuming that each of these time series is modeled via a Markov switching process, our Bayesian nonparametric approach envisions a large *library* of behaviors, with each time series or *object* exhibiting a subset of these behaviors. We aim to allow flexibility in the number of total and sequence-specific behaviors, while encouraging objects to share similar subsets of the behavior library. Additionally, a key aspect of a flexible model for relating time series is to allow the objects to switch between behaviors in different manners (e.g., even if two people both exhibit *running* and *walking* behaviors, they might alternate between these dynamical modes at different frequencies).

One could imagine a Bayesian nonparametric approach based on tying together multiple time series under the HDP prior outlined in Sec. II-B. However, such a formulation assumes that all time series share the same set of behaviors, and switch among them in exactly the same manner. Alternatively, Fox et. al. [28] consider a featural representation, and show the utility of an alternative family of priors based on the *beta process* [29], [30].

### A. Finite Feature Models of Markov Switching Processes

Assume we have a finite collection of *behaviors* $\{\theta_1, \ldots, \theta_K\}$ that are shared in an unknown manner among $N$ objects. One can represent the set of behaviors each object exhibits via an associated list of *features*. A standard featural representation for describing the $N$ objects employs an $N \times K$ binary matrix $F = \{f_{ik}\}$. Setting $f_{ik} = 1$ implies that object $i$ exhibits feature $k$ for some $t \in \{1, \ldots, T_i\}$, where $T_i$ is the length of the $i^{th}$ time series. To discover the structure of behavior sharing (i.e., the feature matrix), one takes the feature vector $\boldsymbol{f}_i = [f_{i1}, \ldots, f_{iK}]$ to be *random*. Assuming each feature is treated independently, this necessitates defining a feature inclusion probability $\omega_k$ for each feature $k$. Within a Bayesian framework, these probabilities are given a prior that is then informed by the data to provide a posterior distribution on feature inclusion probabilities. For example, one could consider the finite Bayesian feature model of [31] that assumes

$$\omega_k \sim \text{Beta}\left(\frac{\alpha}{K}, 1\right)$$

$$f_{ik} \mid \omega_k \sim \text{Bernoulli}(\omega_k).$$

(18)

Beta random variables $\omega_k \in (0, 1)$, and can thus be thought of as defining coin-tossing probabilities. The resulting biased coin is then tossed to define whether $f_{ik}$ is 0 or 1 (i.e., the outcome of a Bernoulli trial). Because each feature is generated independently, and a Beta$(a, b)$ random variable has mean $a/(a + b)$, the expected number of active features in an $N \times K$ matrix is $N\alpha/(\alpha/K + 1) < N\alpha$.

A hierarchical Bayesian featural model also requires priors for behavior parameters $\theta_k$, and the process by which each object switches among its selected behaviors. In the case of Markov switching processes, this switching mechanism is governed by the transition distributions of object $i$, $\pi_j^{(i)}$. As an example of such a model, imagine that each of the $N$ objects is described by a switching VAR process (see Sec. III) that moves among some subset of $K$ possible dynamical modes. Each of the VAR process parameters $\theta_k = \{\boldsymbol{A}_k, \Sigma_k\}$ describes a unique behavior. The feature vector $\boldsymbol{f}_i$ constrains the transitions of object $i$ to solely be between the selected subset of the $K$ possible VAR processes by forcing $\pi_{jk}^{(i)} = 0$ for all $k$ such that $f_{ik} = 0$. One natural construction places Dirichlet priors on the transition distributions, and some prior $H$ (e.g., a MNIW) on the behavior parameters. Then,

$$\pi_j^{(i)} \mid \boldsymbol{f}_i, \gamma, \kappa \sim \text{Dir}([\gamma, \ldots, \gamma, \gamma + \kappa, \gamma, \ldots, \gamma] \otimes \boldsymbol{f}_i) \qquad \theta_k \sim H, \tag{19}$$

where $\otimes$ denotes the element-wise, or Hadamard, vector product. Let $\mathbf{y}_t^{(i)}$ represent the observation vector of the $i^{th}$ object at time $t$, and $z_t^{(i)}$ the latent behavior mode. Assuming an order $r$ switching VAR process, the dynamics of the $i^{th}$ object are described by the following generative process:

$$z_t^{(i)} \mid z_{t-1}^{(i)} \sim \pi_{z_{t-1}^{(i)}}^{(i)} \tag{20}$$

$$\mathbf{y}_t^{(i)} = \sum_{j=1}^{r} A_{j,z_t^{(i)}} \mathbf{y}_{t-j}^{(i)} + \mathbf{e}_t^{(i)}(z_t^{(i)}) \triangleq \boldsymbol{A}_{z_t^{(i)}} \tilde{\mathbf{y}}_t^{(i)} + \mathbf{e}_t^{(i)}(z_t^{(i)}), \tag{21}$$

where $\mathbf{e}_t^{(i)}(k) \sim \mathcal{N}(0, \Sigma_k)$, $\boldsymbol{A}_k = \begin{bmatrix} A_{1,k} & \ldots & A_{r,k} \end{bmatrix}$, and $\tilde{\mathbf{y}}_t^{(i)} = \begin{bmatrix} \mathbf{y}_{t-1}^{(i)T} & \ldots & \mathbf{y}_{t-r}^{(i)T} \end{bmatrix}^T$. The standard HMM with Gaussian emissions arises as a special case of this model when $\boldsymbol{A}_k = \mathbf{0}$ for all $k$.

## B. A Bayesian Nonparametric Featural Model Utilizing Beta and Bernoulli Processes

Following the theme of Sec. II-B and Sec. III-B, it is often desirable to consider a Bayesian nonparametric featural model that relaxes the assumption that the number of features is known or bounded. Such a featural model seeks to allow for infinitely many features, while encouraging a sparse, finite representation. Just as the Dirichlet process provides a useful Bayesian nonparametric prior in clustering applications (i.e., when each observation is associated with a single parameter $\theta_k$), it has been shown that a stochastic process known as the *beta process* is useful in Bayesian nonparametric featural models (i.e., when each observation is associated with a subset of parameters) [30].

The beta process is a special case of a general class of stochastic processes known as *completely random measures* [32]. A completely random measure $G$ is defined such that for any disjoint sets $A_1$ and $A_2$, the corresponding random measures $G(A_1)$ and $G(A_2)$ are independent. This idea generalizes the family of *independent increments processes* on the real line. All completely random measures (up to a deterministic component) can be constructed from realizations of a nonhomogenous Poisson process [32]. Specifically, a Poisson rate measure $\eta$ is defined on a product space $\Theta \otimes \mathbb{R}$, and a draw from the specified Poisson process yields a collection of points $\{\theta_j, \omega_j\}$ that can be used to define a completely random measure:

$$G = \sum_{j=1}^{\infty} \omega_j \delta_{\theta_j}. \tag{22}$$

This construction assumes $\eta$ has infinite mass, yielding the countably infinite collection of points from the Poisson process. From Eq. (22), we see that completely random measures are discrete. Letting the rate measure be defined as a product of a base measure $G_0$ and an improper gamma distribution:

$$\eta(d\theta, d\omega) = cp^{-1}e^{-cp}dpG_0(d\theta), \tag{23}$$

with $c > 0$, gives rise to completely random measures $G \sim \mathrm{GP}(c, G_0)$, where GP denotes a *gamma process*. Normalizing $G$ yields draws from a Dirichlet process $\mathrm{DP}(\alpha, G_0/\alpha)$, with $\alpha = G_0(\Theta)$[3].

Now consider a rate measure defined as the product of a base measure $B_0$, with total mass $B_0(\Theta) = \alpha$, and an improper beta distribution on the product space $\Theta \otimes [0, 1]$:

$$\nu(d\omega, d\theta) = c\omega^{-1}(1 - \omega)^{c-1}d\omega B_0(d\theta). \tag{24}$$

where, once again, $c > 0$. The resulting completely random measure is known as the *beta process* with draws denoted by $B \sim \mathrm{BP}(c, B_0)$. Note that using this construction, the weights $\omega_k$ of the atoms in $B$ lie in the interval $(0, 1)$. Since $\eta$ is $\sigma$-finite, Campbell's theorem [33] guarantees that for $\alpha$ finite, $B$ has finite expected measure. The characteristics of this process define desirable traits for a Bayesian nonparametric featural model: we have a countably infinite collection of coin-tossing probabilities (one for each of our infinite number of features), but only a sparse, finite subset are active in any realization.

The beta process is conjugate to a class of *Bernoulli processes* [30], denoted by $\mathrm{BeP}(B)$, which provide our sought-for featural representation. A realization $X_i \sim \mathrm{BeP}(B)$, with $B$ an atomic measure,

---

[3]Random *probability* measures $G$ are necessarily not completely random since the random variables $G(A_1)$ and $G(A_2)$ for disjoint sets $A_1$ and $A_2$ are dependent due to the normalization constraint.

is a collection of unit mass atoms on $\Theta$ located at some subset of the atoms in $B$. In particular,

$$f_{ik} \sim \text{Bernoulli}(\omega_k) \qquad (25)$$

is sampled independently for each atom $\theta_k$ in $B$, and then $X_i = \sum_k f_{ik} \delta_{\theta_k}$. In many applications, we interpret the atom locations $\theta_k$ as a shared set of global features. A Bernoulli process realization $X_i$ then determines the subset of features allocated to object $i$:

$$B \mid B_0, c \sim \text{BP}(c, B_0)$$

$$X_i \mid B \sim \text{BeP}(B), \quad i = 1, \ldots, N. \qquad (26)$$

Computationally, Bernoulli process realizations $X_i$ are often summarized by an infinite vector of binary indicator variables $\boldsymbol{f}_i = [f_{i1}, f_{i2}, \ldots]$, where $f_{ik} = 1$ if and only if object $i$ exhibits feature $k$. Using the beta process measure $B$ to tie together the feature vectors encourages them to share similar features while allowing object-specific variability.

As shown by Thibaux and Jordan [30], marginalizing over the latent beta process measure $B$, and taking $c = 1$, induces a predictive distribution on feature indicators known as the Indian buffet process (IBP) [31]. The IBP is a culinary metaphor inspired by the Chinese restaurant process of Eq. (7), which is itself the predictive distribution on partitions induced by the Dirichlet process. The Indian buffet consists of an infinitely long buffet line of dishes, or features. The first arriving customer, or object, chooses $\text{Poisson}(\alpha)$ dishes. Each subsequent customer $i$ selects a previously tasted dish $k$ with probability $m_k/i$ proportional to the number of previous customers $m_k$ to sample it, and also samples $\text{Poisson}(\alpha/i)$ new dishes. The feature matrix associated with a realization from an Indian buffet process is shown in Fig. 6(b).

### C. BP-AR-HMM

Recall the model of Sec. IV-A, in which the binary feature indicator variables $f_{ik}$ denote whether object $i$ exhibits dynamical behavior $k$ for some $t \in \{1, \ldots, T_i\}$. Now, however, take $\boldsymbol{f}_i$ to be an infinite dimensional vector of feature indicators realized from the beta process featural model of Sec. IV-B. Continuing our focus on switching VAR processes, we define a *beta process autoregressive HMM* (BP-AR-HMM) [28], in which the features indicate which behaviors are utilized by each object or sequence. Considering the *feature space* (i.e., set of autoregressive parameters) and the *temporal dynamics* (i.e., set of transition distributions) as separate dimensions, one can think of the BP-AR-HMM as a spatio-temporal process comprised of a (continuous) beta process in space and discrete-time Markovian dynamics in time.

Given $\boldsymbol{f}_i$, the $i^{th}$ object's Markov transitions among its set of dynamic behaviors are governed by a set of *feature-constrained transition distributions* $\boldsymbol{\pi}^{(i)} = \{\pi_k^{(i)}\}$. In particular, motivated by the fact
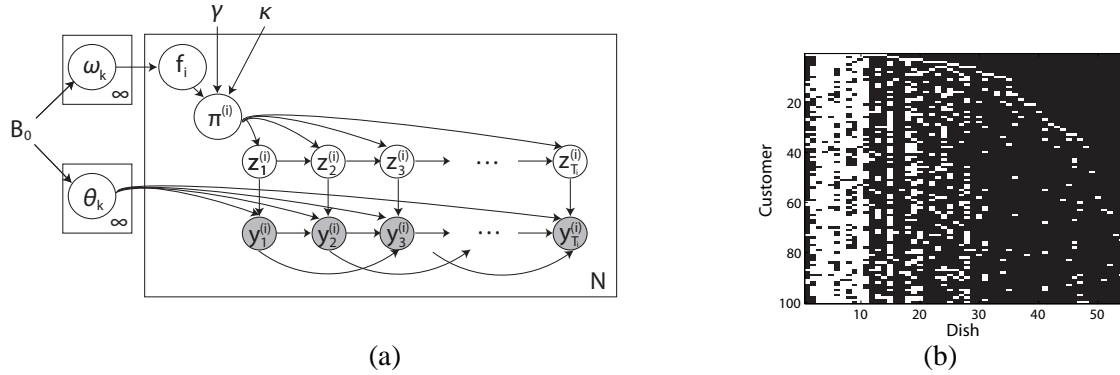
(a)                                                                                                           (b)

Fig. 6.   (a) Graphical model of the BP-AR-HMM. The beta process distributed measure $B \mid B_0 \sim \mathrm{BP}(1, B_0)$ is represented by its masses $\omega_k$ and locations $\theta_k$, as in Eq. (22). The features are then conditionally independent draws $f_{ik} \mid \omega_k \sim \mathrm{Bernoulli}(\omega_k)$, and are used to define feature-constrained transition distributions $\pi_j^{(i)} \mid \boldsymbol{f}_i, \gamma, \kappa \sim \mathrm{Dir}([\gamma, \dots, \gamma, \gamma + \kappa, \gamma, \dots] \otimes \boldsymbol{f}_i)$. The switching VAR dynamics are as in Eq. (21). (b) An example feature matrix $F$ with elements $f_{ik}$ for 100 objects drawn from the Indian buffet predictive distribution using $B_0(\Theta) = \alpha = 10$.

that Dirichlet-distributed probability mass functions can be generated via normalized gamma random variables, for each object $i$ we define a doubly infinite collection of random variables:

$$\eta_{jk}^{(i)} \mid \gamma, \kappa \sim \mathrm{Gamma}(\gamma + \kappa\delta(j,k), 1), \tag{27}$$

Using this collection of *transition variables*, denoted by $\boldsymbol{\eta}^{(i)}$, one can define object-specific, feature-constrained transition distributions:

$$\pi_j^{(i)} = \frac{\left[\eta_{j1}^{(i)} \quad \eta_{j2}^{(i)} \quad \dots\right] \otimes \boldsymbol{f}_i}{\sum_{k \mid f_{ik}=1} \eta_{jk}^{(i)}}. \tag{28}$$

This construction defines $\pi_j^{(i)}$ over the full set of positive integers, but assigns positive mass only at indices $k$ where $f_{ik} = 1$.

The preceding generative process can be equivalently represented via a sample $\tilde{\pi}_j^{(i)}$ from a finite Dirichlet distribution of dimension $K_i = \sum_k f_{ik}$, containing the non-zero entries of $\pi_j^{(i)}$:

$$\tilde{\pi}_j^{(i)} \mid \boldsymbol{f}_i, \gamma, \kappa \sim \mathrm{Dir}([\gamma, \dots, \gamma, \gamma + \kappa, \gamma, \dots \gamma]). \tag{29}$$

The $\kappa$ hyperparameter places extra expected mass on the component of $\tilde{\pi}_j^{(i)}$ corresponding to a self-transition $\pi_{jj}^{(i)}$, analogously to the sticky hyperparameter of Sec. II-B. To complete the Bayesian model specification, a conjugate matrix-normal inverse-Wishart (MNIW) prior is placed on the shared collection of dynamic parameters $\theta_k = \{\boldsymbol{A}_k, \Sigma_k\}$. Since the dynamic parameters are shared by all time series, posterior inference of each parameter set $\theta_k$ relies on pooling data amongst the time series that have

$f_{ik} = 1$. It is through this pooling of data that one may achieve more robust parameter estimates than from considering each time series individually. The resulting model is depicted in Fig. 6(a), and a collection of observation sequences is shown in Fig. 4(e).

The ability of the BP-AR-HMM to find common behaviors amongst a collection of time series was demonstrated on data from the CMU motion capture database [34]. As an illustrative example, a set of six exercise routines were examined, where each of these routines used some combination of the following motion categories: running in place, jumping jacks, arm circles, side twists, knee raises, squats, punching, up and down, two variants of toe touches, arch over, and a reach out stretch. The overall performance of the BP-AR-HMM showed a clear ability to find common motions, and provided more accurate movie frame labels than previously considered approaches [35]. Most significantly, the BP-AR-HMM provided a superior ability to discover the shared feature structure, while allowing objects to exhibit unique features.

## V. CONCLUSION

In this paper, we explored a Bayesian nonparametric approach to learning Markov switching processes. This framework requires one to make fewer assumptions about the underlying dynamics, and thereby allows the data to drive the complexity of the inferred model. We began by examining a Bayesian nonparametric HMM, the sticky HDP-HMM, that uses a hierarchical Dirichlet process prior to regularize an unbounded mode space. We then considered extensions to Markov switching processes with richer, conditionally linear dynamics, including the HDP-AR-HMM and HDP-SLDS. We concluded by considering methods for transferring knowledge among multiple related time series. We argued that a featural representation is more appropriate than a rigid global clustering, as it encourages sharing of behaviors among objects while still allowing sequence-specific variability. In this context, the beta process provides an appealing alternative to the Dirichlet process.

The models presented herein, while representing a flexible alternative to their parametric counterparts in terms of defining the set of dynamical modes, still maintain a number of limitations. First, the models assume Markovian dynamics with observations on a discrete, evenly-spaced temporal grid. Extensions to semi-Markov formulations and non-uniform grids are interesting directions for future research. Second, there is still the question of which dynamical model is appropriate for a given dataset: HMM, AR-HMM, SLDS? The fact that the models are nested (i.e., HMM $\subset$ AR-HMM $\subset$ SLDS) aids in this decision process—choose the simplest formulation that does not egregiously break the model assumptions. For example, the honey bee observations are clearly not independent given the dance mode so choosing an HMM is likely not going to provide desirable performance. Typically, it is useful to have domain-

specific knowledge of at least one example of a time-series segment that can be used to design the structure of individual modes in a model. Overall, however, this issue of model selection in the Bayesian nonparametric setting is an open area of research. Finally, given the Bayesian framework, the models that we have presented necessitate a choice of prior. We have found in practice that the models are relatively robust to the hyperprior settings for the concentration parameters. On the other hand, the choice of base measure tends to affect results significantly, which is typical of simpler Bayesian nonparametric models such as Dirichlet process mixtures. We have found that quasi-empirical Bayes approaches for setting the base measure tend to help push the mass of the distribution into reasonable ranges (see [36] for details).

Our focus in this paper has been the advantages of various hierarchical, nonparametric Bayesian models; detailed algorithms for learning and inference were omitted. One major advantage of the particular Bayesian nonparametric approaches explored in this paper is that they lead to *computationally efficient* methods for learning Markov switching models of unknown order. We point the interested reader to [13], [14], [19], [28] for detailed presentations of Markov chain Monte Carlo (MCMC) algorithms for inference and learning.

## REFERENCES

[1] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] V. Pavlović, J. Rehg, and J. MacCormick, "Learning switching linear models of human motion," in *Advances in Neural Information Processing Systems*, vol. 13, 2001, pp. 981–987.

[3] L. Ren, A. Patrick, A. Efros, J. Hodgins, and J. Rehg, "A data-driven approach to quantifying natural human motion," in *SIGGRAPH*, August 2005.

[4] C.-J. Kim, "Dynamic linear models with Markov-switching," *Journal of Econometrics*, vol. 60, pp. 1–22, 1994.

[5] M. So, K. Lam, and W. Li, "A stochastic volatility model with Markov switching," *Journal of Business & Economic Statistics*, vol. 16, no. 2, pp. 244–253, 1998.

[6] C. Carvalho and H. Lopes, "Simulation-based sequential analysis of Markov switching stochastic volatility models," *Computational Statistics & Data Analysis*, vol. 51, pp. 4526–4542, 9 2007.

[7] X. Rong Li and V. Jilkov, "Survey of maneuvering target tracking. Part V: Multiple-model methods," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 41, no. 4, pp. 1255–1321, 2005.

[8] E. Fox, E. Sudderth, and A. Willsky, "Hierarchical Dirichlet processes for tracking maneuvering targets," in *Proc. International Conference on Information Fusion*, July 2007.

[9] J. Pitman, "Combinatorial stochastic processes," U.C. Berkeley Department of Statistics, Tech. Rep. 621, 2002.

[10] N. Hjort, C. Holmes, P. Müller, and S. Walker, Eds., *Bayesian Nonparametrics: Principles and Practice*. Cambridge, UK: Cambridge University Press, 2010.

[11] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[12] D. Blackwell and J. MacQueen, "Ferguson distributions via Polya urn schemes," *The Annals of Statistics*, vol. 1, no. 2, pp. 353–355, 1973.

[13] Y. Teh, M. Jordan, M. Beal, and D. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[14] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "An HDP-HMM for systems with state persistence," in *Proc. International Conference on Machine Learning*, July 2008.

[15] M. Beal, Z. Ghahramani, and C. Rasmussen, "The infinite hidden Markov model," in *Advances in Neural Information Processing Systems*, vol. 14, 2002, pp. 577–584.

[16] NIST, "Rich transcriptions database," *http://www.nist.gov/speech/tests/rt/*, 2007.

[17] S. Paoletti, A. Juloski, G. Ferrari-Trecate, and R. Vidal, "Identification of hybrid systems: A tutorial." *European Journal of Control*, vol. 2–3, pp. 242–260, 2007.

[18] M. Aoki and A. Havenner, "State space modeling of multiple time series," *Econometric Reviews*, vol. 10, no. 1, pp. 1–59, 1991.

[19] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "Nonparametric Bayesian learning of switching dynamical systems," in *Advances in Neural Information Processing Systems*, vol. 21, 2009, pp. 457–464.

[20] ——, "Nonparametric Bayesian identification of jump systems with sparse dependencies," in *Proc. 15th IFAC Symposium on System Identification*, July 2009.

[21] M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models*. Springer, 1997.

[22] D. MacKay, *Bayesian methods for backprop networks*, ser. Models of Neural Networks, III. Springer, 1994, ch. 6, pp. 211–254.

[23] R. Neal, Ed., *Bayesian Learning for Neural Networks*, ser. Lecture Notes in Statistics. Springer, 1996, vol. 118.

[24] M. Beal, "Variational algorithms for approximate bayesian inference," Ph.D. Thesis, University College London, London, UK, 2003.

[25] F. Caron, M. Davy, A. Doucet, E. Duflos, and P. Vanheeghe, "Bayesian inference for dynamic models with Dirichlet process mixtures," in *Proc. International Conference on Information Fusion*, July 2006.

[26] C. Carvalho and M. West, "Dynamic matrix-variate graphical models," *Bayesian Analysis 2*, pp. 69–98, 2007.

[27] S. Oh, J. Rehg, T. Balch, and F. Dellaert, "Learning and inferring motion patterns using parametric segmental switching linear dynamic systems," *International Journal of Computer Vision*, vol. 77, no. 1–3, pp. 103–124, 2008.

[28] E. Fox, E. Sudderth, M. Jordan, and A. Willsky, "Sharing features among dynamical systems with beta processes," in *Advances in Neural Information Processing Systems*, vol. 22, 2010.

[29] N. Hjort, "Nonparametric Bayes estimators based on beta processes in models for life history data," *The Annals of Statistics*, pp. 1259–1294, 1990.

[30] R. Thibaux and M. Jordan, "Hierarchical beta processes and the Indian buffet process," in *Proc. International Conference on Artificial Intelligence and Statistics*, vol. 11, 2007.

[31] T. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," *Gatsby Computational Neuroscience Unit, Technical Report #2005-001*, 2005.

[32] J. F. C. Kingman, "Completely random measures," *Pacific Journal of Mathematics*, vol. 21, no. 1, pp. 59–78, 1967.

[33] J. Kingman, *Poisson Processes*. Oxford University Press, 1993.

[34] C. M. University, "Graphics lab motion capture database," *http://mocap.cs.cmu.edu/*.

[35] J. Barbič, A. Safonova, J.-Y. Pan, C. Faloutsos, J. Hodgins, and N. Pollard, "Segmenting motion capture data into distinct behaviors," in *Proc. Graphics Interface*, 2004, pp. 185–194.

[36] E. Fox, "Bayesian nonparametric learning of complex dynamical phenomena," Ph.D. dissertation, MIT, July 2009.