Sharing Features among Dynamical Systems with Beta Processes

Emily B. Fox
Electrical Engineering & Computer Science, Massachusetts Institute of Technology
ebfox@mit.edu

Erik B. Sudderth
Computer Science, Brown University
sudderth@cs.brown.edu

Michael I. Jordan
Electrical Engineering & Computer Science and Statistics, University of California, Berkeley
jordan@cs.berkeley.edu

Alan S. Willsky
Electrical Engineering & Computer Science, Massachusetts Institute of Technology
willsky@mit.edu

Abstract

We propose a Bayesian nonparametric approach to the problem of modeling related time series. Using a beta process prior, our approach is based on the discovery of a set of latent dynamical behaviors that are shared among multiple time series. The size of the set and the sharing pattern are both inferred from data. We develop an efficient Markov chain Monte Carlo inference method that is based on the Indian buffet process representation of the predictive distribution of the beta process. In particular, our approach uses the sum-product algorithm to efficiently compute Metropolis-Hastings acceptance probabilities, and explores new dynamical behaviors via birth/death proposals. We validate our sampling algorithm using several synthetic datasets, and also demonstrate promising results on unsupervised segmentation of visual motion capture data.

1 Introduction

In many applications, one would like to discover and model dynamical behaviors which are shared among several related time series. For example, consider video or motion capture data depicting multiple people performing a number of related tasks. By jointly modeling such sequences, we may more robustly estimate representative dynamic models, and also uncover interesting relationships among activities. We specifically focus on time series where behaviors can be individually modeled via temporally independent or linear dynamical systems, and where transitions between behaviors are approximately Markovian. Examples of such *Markov jump processes* include the hidden Markov model (HMM), switching vector autoregressive (VAR) process, and switching linear dynamical system (SLDS). These models have proven useful in such diverse fields as speech recognition, econometrics, remote target tracking, and human motion capture. Our approach envisions a large *library* of behaviors, and each time series or *object* exhibits a subset of these behaviors. We then seek a framework for discovering the set of dynamic behaviors that each object exhibits. We particularly aim to allow flexibility in the number of total and sequence-specific behaviors, and encourage objects to share similar subsets of the large set of possible behaviors.

One can represent the set of behaviors an object exhibits via an associated list of *features*. A standard featural representation for N objects, with a library of K features, employs an $N \times K$ binary matrix $F = \{f_{ik}\}$. Setting $f_{ik} = 1$ implies that object i exhibits feature k. Our desiderata motivate a Bayesian nonparametric approach based on the *beta process* [10, 22], allowing for infinitely many

potential features. Integrating over the latent beta process induces a predictive distribution on features known as the *Indian buffet process* (IBP) [9]. Given a feature set sampled from the IBP, our model reduces to a collection of Bayesian HMMs (or SLDS) with partially shared parameters.

Other recent approaches to Bayesian nonparametric representations of time series include the HDP-HMM [2, 4, 5, 21] and the infinite factorial HMM [24]. These models are quite different from our framework: the HDP-HMM does not select a subset of behaviors for a given time series, but assumes that all time series share the same set of behaviors and switch among them in exactly the same manner. The infinite factorial HMM models a single time-series with emissions dependent on a potentially infinite dimensional feature that evolves with independent Markov dynamics. Our work focuses on modeling multiple time series and on capturing dynamical modes that are shared among the series.

Our results are obtained via an efficient and exact Markov chain Monte Carlo (MCMC) inference algorithm. In particular, we exploit the finite dynamical system induced by a fixed set of features to efficiently compute acceptance probabilities, and reversible jump birth and death proposals to explore new features. We validate our sampling algorithm using several synthetic datasets, and also demonstrate promising unsupervised segmentation of data from the CMU motion capture database [23].

2 Binary Features and Beta Processes

The beta process is a *completely random measure* [12]: draws are discrete with probability one, and realizations on disjoint sets are independent random variables. Consider a probability space Θ, and let B_0 denote a finite *base measure* on Θ with total mass $B_0(\Theta) = \alpha$. Assuming B_0 is absolutely continuous, we define the following *Lévy measure* on the product space $[0,1] \times \Theta$:

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

Here, $c > 0$ is a *concentration parameter*; we denote such a beta process by $\text{BP}(c, B_0)$. A draw $B \sim \text{BP}(c, B_0)$ is then described by

$$B = \sum_{k=1}^{\infty} \omega_k \delta_{\theta_k}, \tag{2}$$

where $(\omega_1, \theta_1), (\omega_2, \theta_2), \ldots$ are the set of atoms in a realization of a nonhomogeneous Poisson process with rate measure ν. If there are atoms in B_0, then these are treated separately; see [22]. The beta process is conjugate to a class of *Bernoulli processes* [22], denoted by $\text{BeP}(B)$, which provide our sought-for featural representation. A realization $X_i \sim \text{BeP}(B)$, with B an atomic measure, is a collection of unit mass atoms on Θ located at some subset of the atoms in B. In particular, $f_{ik} \sim \text{Bernoulli}(\omega_k)$ is sampled independently for each atom θ_k in Eq. (2), and then $X_i = \sum_k f_{ik} \delta_{\theta_k}$.

In many applications, we interpret the atom locations θ_k as a shared set of global features. A Bernoulli process realization X_i then determines the subset of features allocated to object i:

$$\begin{aligned} B \mid B_0, c &\sim \text{BP}(c, B_0) \\ X_i \mid B &\sim \text{BeP}(B), \quad i = 1, \ldots, N. \end{aligned} \tag{3}$$

Because beta process priors are conjugate to the Bernoulli process [22], the posterior distribution given N samples $X_i \sim \text{BeP}(B)$ is a beta process with updated parameters:

$$B \mid X_1, \ldots, X_N, B_0, c \sim \text{BP}\left(c + N, \frac{c}{c+N} B_0 + \sum_{k=1}^{K_+} \frac{m_k}{c+N} \delta_{\theta_k}\right). \tag{4}$$

Here, m_k denotes the number of objects X_i which select the k^{th} feature θ_k. For simplicity, we have reordered the feature indices to list the K_+ features used by at least one object first.

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. As shown by Thibaux and Jordan [22], marginalizing over the beta process measure B, and taking $c = 1$, provides a predictive distribution on indicators known as the Indian buffet process (IBP) Griffiths and Ghahramani [9]. The IBP is a culinary metaphor inspired by the Chinese restaurant process, which is itself the predictive distribution on partitions induced by the Dirichlet process [21]. 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.

3 Describing Multiple Time Series with Beta Processes

Assume we have a set of N objects, each of whose dynamics is described by a switching vector autoregressive (VAR) process, with switches occurring according to a discrete-time Markov process. Such autoregressive HMMs (AR-HMMs) provide a simpler, but often equally effective, alternative to SLDS [17]. Let $\mathbf{y}_t^{(i)}$ represent the observation vector of the i^{th} object at time t, and $z_t^{(i)}$ the latent dynamical mode. Assuming an order r switching VAR process, denoted by VAR(r), we have

$$z_t^{(i)} \sim \pi^{(i)}_{z_{t-1}^{(i)}} \tag{5}$$

$$\mathbf{y}_t^{(i)} = \sum_{j=1}^{r} A_{j,z_t^{(i)}} \mathbf{y}_{t-j}^{(i)} + \mathbf{e}_t^{(i)}(z_t^{(i)}) \triangleq \boldsymbol{A}_{z_t^{(i)}} \tilde{\mathbf{y}}_t^{(i)} + \mathbf{e}_t^{(i)}(z_t^{(i)}), \tag{6}$$

where $\mathbf{e}_t^{(i)}(k) \sim \mathcal{N}(0, \Sigma_k)$, $\boldsymbol{A}_k = [A_{1,k} \quad \dots \quad A_{r,k}]$, and $\tilde{\mathbf{y}}_t^{(i)} = [\mathbf{y}_{t-1}^{(i)^T} \quad \dots \quad \mathbf{y}_{t-r}^{(i)^T}]^T$. The standard HMM with Gaussian emissions arises as a special case of this model when $\boldsymbol{A}_k = \mathbf{0}$ for all k. We refer to these VAR processes, with parameters $\theta_k = \{\boldsymbol{A}_k, \Sigma_k\}$, as *behaviors*, and use a beta process prior to couple the dynamic behaviors exhibited by different objects or sequences.

As in Sec. 2, let $\boldsymbol{f}_i$ be a vector of binary indicator variables, where f_{ik} denotes whether object i exhibits behavior k for some $t \in \{1, \dots, T_i\}$. Given $\boldsymbol{f}_i$, we define a *feature-constrained transition distribution* $\boldsymbol{\pi}^{(i)} = \{\pi_k^{(i)}\}$, which governs the i^{th} object's Markov transitions among its set of dynamic behaviors. In particular, motivated by the fact that a Dirichlet-distributed probability mass function can be interpreted as a normalized collection of gamma-distributed random variables, for each object i we define a doubly infinite collection of random variables:

$$\eta_{jk}^{(i)} \mid \gamma, \kappa \sim \text{Gamma}(\gamma + \kappa\delta(j,k), 1), \tag{7}$$

where $\delta(j,k)$ indicates the Kronecker delta function. We denote this collection of *transition variables* by $\boldsymbol{\eta}^{(i)}$, and use them to define object-specific, feature-constrained transition distributions:

$$\pi_j^{(i)} = \frac{\begin{bmatrix}\eta_{j1}^{(i)} & \eta_{j2}^{(i)} & \dots\end{bmatrix} \otimes \boldsymbol{f}_i}{\sum_{k|f_{ik}=1} \eta_{jk}^{(i)}}. \tag{8}$$

Here, $\otimes$ denotes the element-wise vector product. 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 $\pi_j^{(i)}$:

$$\tilde{\pi}_j^{(i)} \mid \boldsymbol{f}_i, \gamma, \kappa \sim \text{Dir}([\gamma, \dots, \gamma, \gamma + \kappa, \gamma, \dots \gamma]). \tag{9}$$

The κ 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 Fox et al. [4]. We refer to this model, which is summarized in Fig. 1, as the *beta process autoregressive HMM* (BP-AR-HMM).

4 MCMC Methods for Posterior Inference

We have developed an MCMC method which alternates between resampling binary feature assignments given observations and dynamical parameters, and dynamical parameters given observations and features. The sampler interleaves Metropolis-Hastings (MH) and Gibbs sampling updates, which are sometimes simplified by appropriate auxiliary variables. We leverage the fact that fixed feature assignments instantiate a set of *finite* AR-HMMs, for which dynamic programming can be used to efficiently compute marginal likelihoods. Our novel approach to resampling the potentially infinite set of object-specific features employs incremental "birth" and "death" proposals, improving on previous exact samplers for IBP models with non-conjugate likelihoods.

4.1 Sampling binary feature assignments

Let $\boldsymbol{F}^{-ik}$ denote the set of all binary feature indicators excluding f_{ik}, and K_+^{-i} be the number of behaviors currently instantiated by objects other than i. For notational simplicity, we assume that

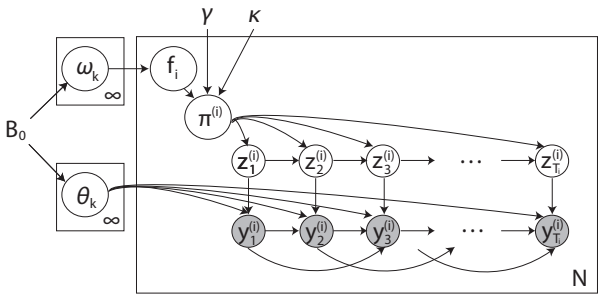


Figure 1: Graphical model of the BP-AR-HMM. The beta process distributed measure $B \mid B_0 \sim \text{BP}(1, B_0)$ is represented by its masses ω_k and locations θ_k, as in Eq. (2). The features are then conditionally independent draws $f_{ik} \mid \omega_k \sim \text{Bernoulli}(\omega_k)$, and are used to define feature-constrained transition distributions $\pi_j^{(i)} \mid \boldsymbol{f}_i, \gamma, \kappa \sim \text{Dir}([\gamma, \ldots, \gamma, \gamma + \kappa, \gamma, \ldots] \otimes \boldsymbol{f}_i)$. The switching VAR dynamics are as in Eq. (6).

these behaviors are indexed by $\{1, \ldots, K_+^{-i}\}$. Given the i^{th} object's observation sequence $\mathbf{y}_{1:T_i}^{(i)}$, transition variables $\boldsymbol{\eta}^{(i)} = \eta^{(i)}_{1:K_+^{-i},1:K_+^{-i}}$, and shared dynamic parameters $\theta_{1:K_+^{-i}}$, feature indicators f_{ik} for currently used features $k \in \{1, \ldots, K_+^{-i}\}$ have the following posterior distribution:

$$p(f_{ik} \mid \boldsymbol{F}^{-ik}, \mathbf{y}_{1:T_i}^{(i)}, \boldsymbol{\eta}^{(i)}, \theta_{1:K_+^{-i}}, \alpha) \propto p(f_{ik} \mid \boldsymbol{F}^{-ik}, \alpha) p(\mathbf{y}_{1:T_i}^{(i)} \mid \boldsymbol{f}_i, \boldsymbol{\eta}^{(i)}, \theta_{1:K_+^{-i}}). \tag{10}$$

Here, the IBP prior implies that $p(f_{ik} = 1 \mid \boldsymbol{F}^{-ik}, \alpha) = m_k^{-i}/N$, where m_k^{-i} denotes the number of objects *other* than object i that exhibit behavior k. In evaluating this expression, we have exploited the exchangeability of the IBP [9], which follows directly from the beta process construction [22].

For binary random variables, MH proposals can mix faster [6] and have greater statistical efficiency [14] than standard Gibbs samplers. To update f_{ik} given $\boldsymbol{F}^{-ik}$, we thus use the posterior of Eq. (10) to evaluate a MH proposal which flips f_{ik} to the complement $\bar{f}$ of its current value f:

$$\begin{aligned} f_{ik} &\sim \rho(\bar{f} \mid f)\delta(f_{ik}, \bar{f}) + (1 - \rho(\bar{f} \mid f))\delta(f_{ik}, f) \\ \rho(\bar{f} \mid f) &= \min\left\{ \frac{p(f_{ik} = \bar{f} \mid \boldsymbol{F}^{-ik}, \mathbf{y}_{1:T_i}^{(i)}, \boldsymbol{\eta}^{(i)}, \theta_{1:K_+^{-i}}, \alpha)}{p(f_{ik} = f \mid \boldsymbol{F}^{-ik}, \mathbf{y}_{1:T_i}^{(i)}, \boldsymbol{\eta}^{(i)}, \theta_{1:K_+^{-i}}, \alpha)}, 1 \right\}. \end{aligned} \tag{11}$$

To compute likelihoods, we combine $\boldsymbol{f}_i$ and $\boldsymbol{\eta}^{(i)}$ to construct feature-constrained transition distributions $\pi_j^{(i)}$ as in Eq. (8), and apply the sum-product message passing algorithm [19].

An alternative approach is needed to resample the Poisson(α/N) "unique" features associated only with object i. Let $K_+ = K_+^{-i} + n_i$, where n_i is the number of features unique to object i, and define $\boldsymbol{f}_{-i} = f_{i,1:K_+^{-i}}$ and $\boldsymbol{f}_{+i} = f_{i,K_+^{-i}+1:K_+}$. The posterior distribution over n_i is then given by

$$\begin{aligned} &p(n_i \mid \boldsymbol{f}_i, \mathbf{y}_{1:T_i}^{(i)}, \boldsymbol{\eta}^{(i)}, \theta_{1:K_+^{-i}}, \alpha) \\ &\quad \propto \frac{(\frac{\alpha}{N})^{n_i} e^{-\frac{\alpha}{N}}}{n_i!} \iint p(\mathbf{y}_{1:T_i}^{(i)} \mid \boldsymbol{f}_{-i}, \boldsymbol{f}_{+i} = \mathbf{1}, \boldsymbol{\eta}^{(i)}, \boldsymbol{\eta}_+, \theta_{1:K_+^{-i}}, \boldsymbol{\theta}_+)\, dB_0(\boldsymbol{\theta}_+) dH(\boldsymbol{\eta}_+), \end{aligned} \tag{12}$$

where H is the gamma prior on transition variables, $\boldsymbol{\theta}_+ = \theta_{K_+^{-i}+1:K_+}$ are the parameters of unique features, and $\boldsymbol{\eta}_+$ are transition parameters $\eta_{jk}^{(i)}$ to or from unique features $j, k \in \{K_+^{-i} + 1 : K_+\}$. Exact evaluation of this integral is intractable due to dependencies induced by the AR-HMMs.

One early approach to approximate Gibbs sampling in non-conjugate IBP models relies on a finite truncation [7]. Meeds et al. [15] instead consider independent Metropolis proposals which replace the existing unique features by $n_i' \sim$ Poisson(α/N) new features, with corresponding parameters $\boldsymbol{\theta}_+'$ drawn from the prior. For high-dimensional models like that considered in this paper, however, moves proposing large numbers of unique features have low acceptance rates. Thus, mixing rates are greatly affected by the beta process hyperparameter α. We instead develop a "birth and death" reversible jump MCMC (RJMCMC) sampler [8], which proposes to either add a single new feature,

or eliminate one of the existing features in $\boldsymbol{f}_{+i}$. Some previous work has applied RJMCMC to finite binary feature models [3, 27], but not to the IBP. Our proposal distribution factors as follows:

$$q(\boldsymbol{f}'_{+i}, \boldsymbol{\theta}'_{+}, \boldsymbol{\eta}'_{+} \mid \boldsymbol{f}_{+i}, \boldsymbol{\theta}_{+}, \boldsymbol{\eta}_{+}) = q_f(\boldsymbol{f}'_{+i} \mid \boldsymbol{f}_{+i}) q_\theta(\boldsymbol{\theta}'_{+} \mid \boldsymbol{f}'_{+i}, \boldsymbol{f}_{+i}, \boldsymbol{\theta}_{+}) q_\eta(\boldsymbol{\eta}'_{+} \mid \boldsymbol{f}'_{+i}, \boldsymbol{f}_{+i}, \boldsymbol{\eta}_{+}). \quad (13)$$

Let $n_i = \sum_k f_{+ik}$. The feature proposal $q_f(\cdot \mid \cdot)$ encodes the probabilities of birth and death moves: a new feature is created with probability 0.5, and each of the n_i existing features is deleted with probability $0.5/n_i$. For parameters, we define our proposal using the generative model:

$$q_\theta(\boldsymbol{\theta}'_{+} \mid \boldsymbol{f}'_{+i}, \boldsymbol{f}_{+i}, \boldsymbol{\theta}_{+}) = \begin{cases} b_0(\theta'_{+,n_i+1}) \prod_{k=1}^{n_i} \delta_{\theta_{+k}}(\theta'_{+k}), & \text{birth of feature } n_i + 1; \\ \prod_{k \neq \ell} \delta_{\theta_{+k}}(\theta'_{+k}), & \text{death of feature } \ell, \end{cases} \quad (14)$$

where b_0 is the density associated with $\alpha^{-1} B_0$. The distribution $q_\eta(\cdot \mid \cdot)$ is defined similarly, but using the gamma prior on transition variables of Eq. (7). The MH acceptance probability is then

$$\rho(\boldsymbol{f}'_{+i}, \boldsymbol{\theta}'_{+}, \boldsymbol{\eta}'_{+} \mid \boldsymbol{f}_{+i}, \boldsymbol{\theta}_{+}, \boldsymbol{\eta}_{+}) = \min\{r(\boldsymbol{f}'_{+i}, \boldsymbol{\theta}'_{+}, \boldsymbol{\eta}'_{+} \mid \boldsymbol{f}_{+i}, \boldsymbol{\theta}_{+}, \boldsymbol{\eta}_{+}), 1\}. \quad (15)$$

Canceling parameter proposals with corresponding prior terms, the acceptance ratio $r(\cdot \mid \cdot)$ equals

$$\frac{p(\mathbf{y}^{(i)}_{1:T_i} \mid [\boldsymbol{f}_{-i}\, \boldsymbol{f}'_{+i}], \theta_{1:K_+^{-i}}, \boldsymbol{\theta}'_{+}, \boldsymbol{\eta}^{(i)}, \boldsymbol{\eta}'_{+}) \text{ Poisson}(n'_i \mid \alpha/N)\, q_f(\boldsymbol{f}_{+i} \mid \boldsymbol{f}'_{+i})}{p(\mathbf{y}^{(i)}_{1:T_i} \mid [\boldsymbol{f}_{-i}\, \boldsymbol{f}_{+i}], \theta_{1:K_+^{-i}}, \boldsymbol{\theta}_{+}, \boldsymbol{\eta}^{(i)}, \boldsymbol{\eta}_{+}) \text{ Poisson}(n_i \mid \alpha/N)\, q_f(\boldsymbol{f}'_{+i} \mid \boldsymbol{f}_{+i})}, \quad (16)$$

with $n'_i = \sum_k f'_{+ik}$. Because our birth and death proposals do not modify the values of existing parameters, the Jacobian term normally arising in RJMCMC algorithms simply equals one.

4.2 Sampling dynamic parameters and transition variables

Posterior updates to transition variables $\boldsymbol{\eta}^{(i)}$ and shared dynamic parameters θ_k are greatly simplified if we instantiate the mode sequences $z^{(i)}_{1:T_i}$ for each object i. We treat these mode sequences as *auxiliary variables*: they are sampled given the current MCMC state, conditioned on when resampling model parameters, and then discarded for subsequent updates of feature assignments $\boldsymbol{f}_i$.

Given feature-constrained transition distributions $\boldsymbol{\pi}^{(i)}$ and dynamic parameters $\{\theta_k\}$, along with the observation sequence $\mathbf{y}^{(i)}_{1:T_i}$, we *jointly* sample the mode sequence $z^{(i)}_{1:T_i}$ by computing backward messages $m_{t+1,t}(z^{(i)}_t) \propto p(\mathbf{y}^{(i)}_{t+1:T_i} \mid z^{(i)}_t, \tilde{\mathbf{y}}^{(i)}_t, \boldsymbol{\pi}^{(i)}, \{\theta_k\})$, and then recursively sampling each $z^{(i)}_t$:

$$z^{(i)}_t \mid z^{(i)}_{t-1}, \mathbf{y}^{(i)}_{1:T_i}, \boldsymbol{\pi}^{(i)}, \{\theta_k\} \sim \pi^{(i)}_{z^{(i)}_{t-1}}(z^{(i)}_t) \mathcal{N}\big(\mathbf{y}^{(i)}_t; \boldsymbol{A}_{z^{(i)}_t} \tilde{\mathbf{y}}^{(i)}_t, \Sigma_{z^{(i)}_t}\big) m_{t+1,t}(z^{(i)}_t). \quad (17)$$

Because Dirichlet priors are conjugate to multinomial observations $z^{(i)}_{1:T}$, the posterior of $\pi^{(i)}_j$ is

$$\pi^{(i)}_j \mid \boldsymbol{f}_i, z^{(i)}_{1:T}, \gamma, \kappa \sim \text{Dir}([\gamma + n^{(i)}_{j1}, \ldots, \gamma + n^{(i)}_{jj-1}, \gamma + \kappa + n^{(i)}_{jj}, \gamma + n^{(i)}_{jj+1}, \ldots] \otimes \boldsymbol{f}_i). \quad (18)$$

Here, $n^{(i)}_{jk}$ are the number of transitions from mode j to k in $z^{(i)}_{1:T}$. Since the mode sequence $z^{(i)}_{1:T}$ is generated from feature-constrained transition distributions, $n^{(i)}_{jk}$ is zero for any k such that $f_{ik} = 0$. Thus, to arrive at the posterior of Eq. (18), we only update $\eta^{(i)}_{jk}$ for instantiated features:

$$\eta^{(i)}_{jk} \mid z^{(i)}_{1:T}, \gamma, \kappa \sim \text{Gamma}(\gamma + \kappa\delta(j,k) + n^{(i)}_{jk}, 1), \quad k \in \{\ell \mid f_{i\ell} = 1\}. \quad (19)$$

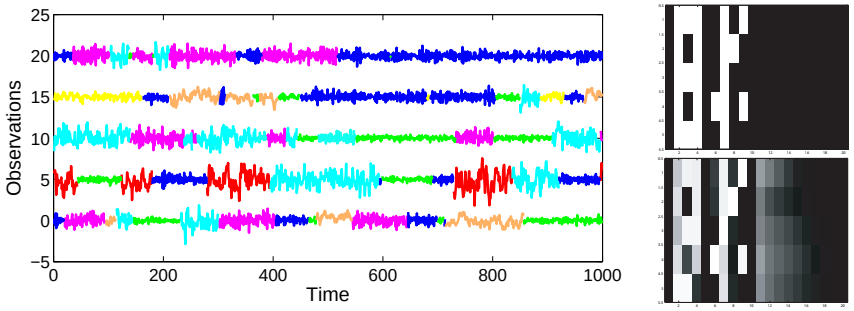
We now turn to posterior updates for dynamic parameters. We place a conjugate matrix-normal inverse-Wishart (MNIW) prior [26] on $\{\boldsymbol{A}_k, \Sigma_k\}$, comprised of an inverse-Wishart prior $\text{IW}(S_0, n_0)$ on Σ_k and a matrix-normal prior $\mathcal{MN}(\boldsymbol{A}_k; M, \Sigma_k, K)$ on $\boldsymbol{A}_k$ given Σ_k. We consider the following sufficient statistics based on the sets $\boldsymbol{Y}_k = \{\mathbf{y}^{(i)}_t \mid z^{(i)}_t = k\}$ and $\tilde{\boldsymbol{Y}}_k = \{\tilde{\mathbf{y}}^{(i)}_t \mid z^{(i)}_t = k\}$ of observations and lagged observations, respectively, associated with behavior k:

$$S^{(k)}_{\tilde{y}\tilde{y}} = \sum_{(t,i) \mid z^{(i)}_t = k} \tilde{\mathbf{y}}^{(i)}_t \tilde{\mathbf{y}}^{(i)^T}_t + \boldsymbol{K} \qquad S^{(k)}_{y\tilde{y}} = \sum_{(t,i) \mid z^{(i)}_t = k} \mathbf{y}^{(i)}_t \tilde{\mathbf{y}}^{(i)^T}_t + \boldsymbol{M}\boldsymbol{K}$$

$$S^{(k)}_{yy} = \sum_{(t,i) \mid z^{(i)}_t = k} \mathbf{y}^{(i)}_t \mathbf{y}^{(i)^T}_t + \boldsymbol{M}\boldsymbol{K}\boldsymbol{M}^T \qquad S^{(k)}_{y|\tilde{y}} = S^{(k)}_{yy} - S^{(k)}_{y\tilde{y}} S^{-(k)}_{\tilde{y}\tilde{y}} S^{(k)^T}_{\tilde{y}\tilde{y}}.$$

Following Fox et al. [5], the posterior can then be shown to equal

$$\boldsymbol{A}_k \mid \Sigma_k, \boldsymbol{Y}_k \sim \mathcal{MN}\left(\boldsymbol{A}_k; S^{(k)}_{y\tilde{y}} S^{-(k)}_{\tilde{y}\tilde{y}}, \Sigma_k, S^{(k)}_{\tilde{y}\tilde{y}}\right), \quad \Sigma_k \mid \boldsymbol{Y}_k \sim \text{IW}\left(S^{(k)}_{y|\tilde{y}} + S_0, |\boldsymbol{Y}_k| + n_0\right).$$



(a) (b)

Figure 2: (a) Observation sequences for each of 5 switching AR(1) time series colored by true mode sequence, and offset for clarity. (b) True feature matrix (top) of the five objects and estimated feature matrix (bottom) averaged over 10,000 MCMC samples taken from 100 trials every 10th sample. White indicates active features. The estimated feature matrices are produced from mode sequences mapped to the ground truth labels according to the minimum Hamming distance metric, and selecting modes with more than 2% of the object's observations.

4.3 Sampling the beta process and Dirichlet transition hyperparameters

We additionally place priors on the Dirichlet hyperparameters γ and κ, as well as the beta process parameter α. Let $\boldsymbol{F} = \{\boldsymbol{f}_i\}$. As derived in [9], $p(\boldsymbol{F} \mid \alpha)$ can be expressed as

$$p(\boldsymbol{F} \mid \alpha) \propto \alpha^{K_+} \exp\left(-\alpha \sum_{n=1}^{N} \frac{1}{n} \right), \tag{20}$$

where, as before, K_+ is the number of unique features activated in $\boldsymbol{F}$. As in [7], we place a conjugate Gamma(a_α, b_α) prior on α, which leads to the following posterior distribution:

$$p(\alpha \mid \boldsymbol{F}, a_\alpha, b_\alpha) \propto p(\boldsymbol{F} \mid \alpha) p(\alpha \mid a_\alpha, b_\alpha) \propto \text{Gamma}\left(a_\alpha + K_+, b_\alpha + \sum_{n=1}^{N} \frac{1}{n} \right). \tag{21}$$

Transition hyperparameters are assigned similar priors $\gamma \sim \text{Gamma}(a_\gamma, b_\gamma)$, $\kappa \sim \text{Gamma}(a_\kappa, b_\kappa)$. Because the generative process of Eq. (7) is non-conjugate, we rely on MH steps which iteratively resample γ given κ, and κ given γ. Each sub-step uses a gamma proposal distribution $q(\cdot \mid \cdot)$ with fixed variance σ_γ^2 or σ_κ^2, and mean equal to the current hyperparameter value. To update γ given κ, the acceptance probability is $\min\{r(\gamma' \mid \gamma), 1\}$, where $r(\gamma' \mid \gamma)$ is defined to equal

$$\frac{p(\gamma' \mid \kappa, \boldsymbol{\pi}, \boldsymbol{F}) q(\gamma \mid \gamma')}{p(\gamma \mid \kappa, \boldsymbol{\pi}, \boldsymbol{F}) q(\gamma' \mid \gamma)} = \frac{p(\boldsymbol{\pi} \mid \gamma', \kappa, \boldsymbol{F}) p(\gamma') q(\gamma \mid \gamma')}{p(\boldsymbol{\pi} \mid \gamma, \kappa, \boldsymbol{F}) p(\gamma) q(\gamma' \mid \gamma)} = \frac{f(\gamma') \Gamma(\vartheta) e^{-\gamma' b_\gamma} \gamma^{\vartheta' - \vartheta - a_\gamma} \sigma_\gamma^{2\vartheta}}{f(\gamma) \Gamma(\vartheta') e^{-\gamma b_\gamma} \gamma'^{\vartheta - \vartheta' - a_\gamma} \sigma_\gamma^{2\vartheta'}}.$$

Here, $\vartheta = \gamma^2/\sigma_\gamma^2$, $\vartheta' = \gamma'^2/\sigma_\gamma^2$, and $f(\gamma) = \prod_i \frac{\Gamma(\gamma K_i + \kappa)^{K_i}}{\Gamma(\gamma)^{K_i^2 - K_i} \Gamma(\gamma + \kappa)^{K_i}} \prod_{(j,k)=1}^{K_i} \pi_{kj}^{(i)^{\gamma + \kappa \delta(k,j) - 1}}$. The MH sub-step for resampling κ given γ is similar, but with an appropriately redefined $f(\kappa)$.

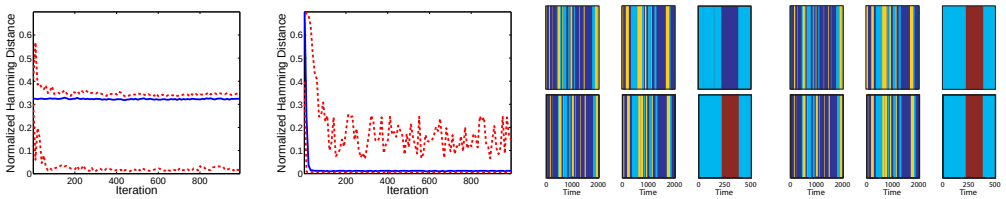
5 Synthetic Experiments

To test the ability of BP-AR-HMM to discover shared dynamics, we generated five time series that switched between AR(1) models

$$y_t^{(i)} = a_{z_t^{(i)}} y_{t-1}^{(i)} + e_t^{(i)}(z_t^{(i)}) \tag{22}$$

with $a_k \in \{-0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8\}$ and process noise covariance Σ_k drawn from an IW$(0.5, 3)$ prior. The object-specific features, shown in Fig. 2(b), were sampled from a truncated IBP [9] using $\alpha = 10$ and then used to generate the observation sequences of Fig. 2(a). The resulting feature matrix estimated over 10,000 MCMC samples is shown in Fig. 2. Comparing to the true feature matrix, we see that our model is indeed able to discover most of the underlying latent structure of the time series despite the challenging setting defined by the close AR coefficients.

One might propose, as an alternative to the BP-AR-HMM, using an architecture based on the hierarchical Dirichlet process of [21]; specifically we could use the HDP-AR-HMMs of [5] tied together with a shared set of transition and dynamic parameters. To demonstrate the difference between these models, we generated data for three switching AR(1) processes. The first two objects, with four times the data points of the third, switched between dynamical modes defined



(a) (b) (c) (d)

Figure 3: (a)-(b) The 10th, 50th, and 90th Hamming distance quantiles for object 3 over 1000 trials for the HDP-AR-HMMs and BP-AR-HMM, respectively. (c)-(d) Examples of typical segmentations into behavior modes for the three objects at Gibbs iteration 1000 for the two models (top = estimate, bottom = truth).

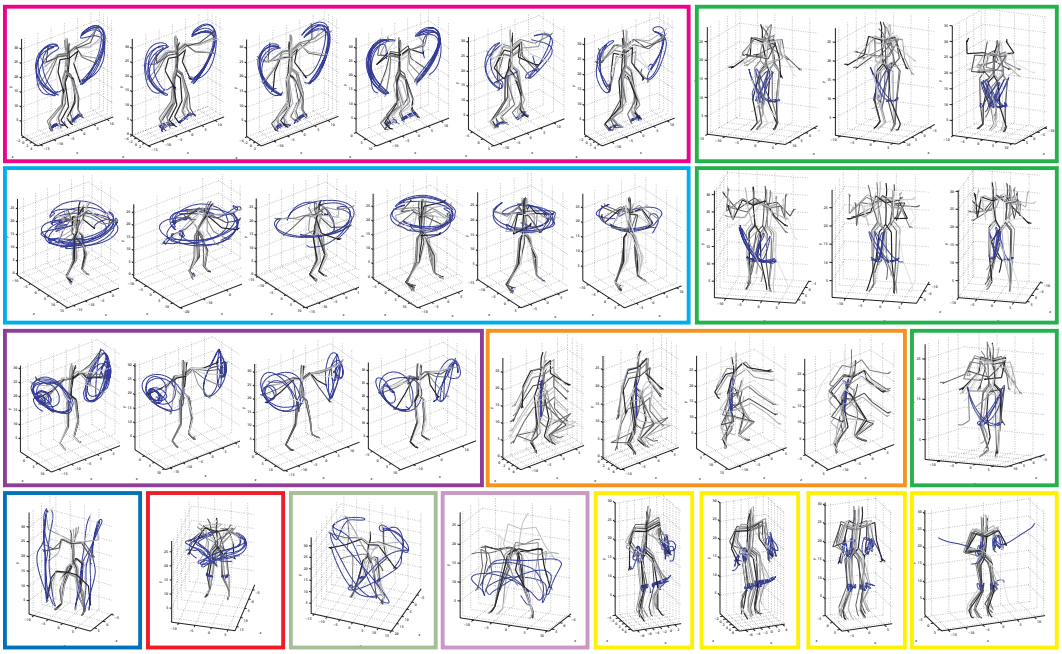


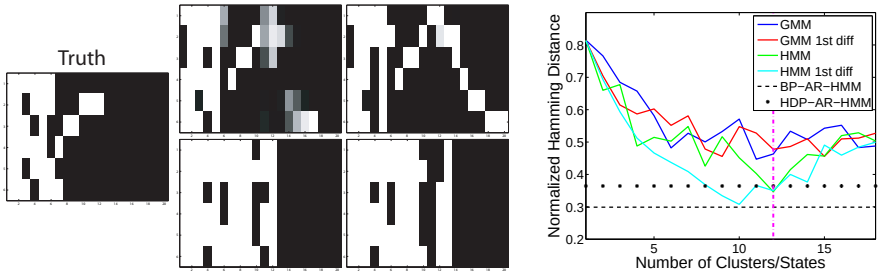
Figure 4: Each skeleton plot displays the trajectory of a learned contiguous segment of more than 2 seconds. To reduce the number of plots, we preprocessed the data to bridge segments separated by fewer than 300 msec. The boxes group segments categorized under the same feature label, with the color indicating the true feature label. Skeleton rendering done by modifications to Neil Lawrence's Matlab MoCap toolbox [13].

by $a_k \in \{-0.8, -0.4, 0.8\}$ and the third object used $a_k \in \{-0.3, 0.8\}$. The results shown in Fig. 3 indicate that the multiple HDP-AR-HMM model typically describes the third object using $a_k \in \{-0.4, 0.8\}$ since this assignment better matches the parameters defined by the other (lengthy) time series. These results reiterate that the feature model emphasizes choosing behaviors rather than assuming all objects are performing minor variations of the same dynamics.

For the experiments above, we placed a Gamma$(1, 1)$ prior on α and γ, and a Gamma$(100, 1)$ prior on κ. The gamma proposals used $\sigma_\gamma^2 = 1$ and $\sigma_\kappa^2 = 100$ while the MNIW prior was given $M = 0$, $K = 0.1 * I_d$, $n_0 = d + 2$, and S_0 set to 0.75 times the empirical variance of the joint set of first difference observations. At initialization, each time series was segmented into five contiguous blocks, with feature labels unique to that sequence.

6 Motion Capture Experiments

The linear dynamical system is a common model for describing simple human motion [11], and the more complicated SLDS has been successfully applied to the problem of human motion synthesis, classification, and visual tracking [17, 18]. Other approaches develop non-linear dynamical models using Gaussian processes [25] or based on a collection of binary latent features [20]. However, there has been little effort in jointly segmenting and identifying common dynamic behaviors amongst a set of *multiple* motion capture (MoCap) recordings of people performing various tasks. The BP-AR-HMM provides an ideal way of handling this problem. One benefit of the proposed model, versus the standard SLDS, is that it does not rely on manually specifying the set of possible behaviors.



(a) (b)

Figure 5: (a) MoCap feature matrices associated with BP-AR-HMM (top-left) and HDP-AR-HMM (top-right) estimated sequences over iterations 15,000 to 20,000, and MAP assignment of the GMM (bottom-left) and HMM (bottom-right) using first-difference observations and 12 clusters/states. (b) Hamming distance versus number of GMM clusters / HMM states on raw observations (blue/green) and first-difference observations (red/cyan), with the BP- and HDP- AR-HMM segmentations (black) and true feature count (magenta) shown for comparison. Results are for the most-likely of 10 EM initializations using Murphy's HMM Matlab toolbox [16].

As an illustrative example, we examined a set of six CMU MoCap exercise routines [23], three from Subject 13 and three from Subject 14. 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.

From the set of 62 position and joint angles, we selected 12 measurements deemed most informative for the gross motor behaviors we wish to capture: one body torso position, two waist angles, one neck angle, one set of right and left (R/L) shoulder angles, the R/L elbow angles, one set of R/L hip angles, and one set of R/L ankle angles. The MoCap data are recorded at 120 fps, and we block-average the data using non-overlapping windows of 12 frames. Using these measurements, the prior distributions were set exactly as in the synthetic data experiments except the scale matrix, S_0, of the MNIW prior which was set to 5 times the empirical covariance of the first difference observations. This allows more variability in the observed behaviors. We ran 25 chains of the sampler for 20,000 iterations and then examined the chain whose segmentation minimized the expected Hamming distance to the set of segmentations from all chains over iterations 15,000 to 20,000. Future work includes developing split-merge proposals to further improve mixing rates in high dimensions.

The resulting MCMC sample is displayed in Fig. 4 and in the supplemental video available online. Although some behaviors are merged or split, the overall performance shows a clear ability to find common motions. The split behaviors shown in green and yellow can be attributed to the two subjects performing the same motion in a distinct manner (e.g., knee raises in combination with upper body motion or not, running with hands in or out of sync with knees, etc.). We compare our performance both to the HDP-AR-HMM and to the Gaussian mixture model (GMM) method of Barbič et al. [1] using EM initialized with k-means. Barbič et al. [1] also present an approach based on probabilistic PCA, but this method focuses primarily on change-point detection rather than behavior clustering. As further comparisons, we look at a GMM on first difference observations, and an HMM on both data sets. The results of Fig. 5(b) demonstrate that the BP-AR-HMM provides more accurate frame labels than any of these alternative approaches over a wide range of mixture model settings. In Fig. 5(a), we additionally see that the BP-AR-HMM provides a superior ability to discover the shared feature structure.

7 Discussion

Utilizing the beta process, we developed a coherent Bayesian nonparametric framework for discovering dynamical features common to multiple time series. This formulation allows for object-specific variability in how the dynamical behaviors are used. We additionally developed a novel exact sampling algorithm for non-conjugate beta process models. The utility of our BP-AR-HMM was demonstrated both on synthetic data, and on a set of MoCap sequences where we showed performance exceeding that of alternative methods. Although we focused on switching VAR processes, our approach could be equally well applied to a wide range of other switching dynamical systems.

Acknowledgments

This work was supported in part by MURIs funded through AFOSR Grant FA9550-06-1-0324 and ARO Grant W911NF-06-1-0076.

References

[1] J. Barbič, A. Safonova, J.-Y. Pan, C. Faloutsos, J.K. Hodgins, and N.S. Pollard. Segmenting motion capture data into distinct behaviors. In *Proc. Graphics Interface*, pages 185–194, 2004.

[2] M.J. Beal, Z. Ghahramani, and C.E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 14, pages 577–584, 2002.

[3] A.C. Courville, N. Daw, G.J. Gordon, and D.S. Touretzky. Model uncertainty in classical conditioning. In *Advances in Neural Information Processing Systems*, volume 16, pages 977–984, 2004.

[4] E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky. An HDP-HMM for systems with state persistence. In *Proc. International Conference on Machine Learning*, July 2008.

[5] E.B. Fox, E.B. Sudderth, M.I. Jordan, and A.S. Willsky. Nonparametric Bayesian learning of switching dynamical systems. In *Advances in Neural Information Processing Systems*, volume 21, pages 457–464, 2009.

[6] A. Frigessi, P. Di Stefano, C.R. Hwang, and S.J. Sheu. Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics. *Journal of the Royal Statistical Society, Series B*, pages 205–219, 1993.

[7] D. Görür, F. Jäkel, and C.E. Rasmussen. A choice model with infinitely many latent features. In *Proc. International Conference on Machine learning*, June 2006.

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

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

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

[11] E. Hsu, K. Pulli, and J. Popović. Style translation for human motion. In *SIGGRAPH*, pages 1082–1089, 2005.

[12] J. F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

[13] N. Lawrence. MATLAB motion capture toolbox. *http://www.cs.man.ac.uk/ neill/mocap/*.

[14] J.S. Liu. Peskun's theorem and a modified discrete-state Gibbs sampler. *Biometrika*, 83(3):681–682, 1996.

[15] E. Meeds, Z. Ghahramani, R.M. Neal, and S.T. Roweis. Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems*, volume 19, pages 977–984, 2007.

[16] K.P. Murphy. Hidden Markov model (HMM) toolbox for MATLAB. *http://www.cs.ubc.ca/ murphyk/Software/HMM/hmm.html*.

[17] V. Pavlović, J.M. Rehg, T.J. Cham, and K.P. Murphy. A dynamic Bayesian network approach to figure tracking using learned dynamic models. In *Proc. International Conference on Computer Vision*, September 1999.

[18] V. Pavlović, J.M. Rehg, and J. MacCormick. Learning switching linear models of human motion. In *Advances in Neural Information Processing Systems*, volume 13, pages 981–987, 2001.

[19] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[20] G.W. Taylor, G.E. Hinton, and S.T. Roweis. Modeling human motion using binary latent variables. In *Advances in Neural Information Processing Systems*, volume 19, pages 1345–1352, 2007.

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

[22] R. Thibaux and M.I. Jordan. Hierarchical beta processes and the Indian buffet process. In *Proc. International Conference on Artificial Intelligence and Statistics*, volume 11, 2007.

[23] Carnegie Mellon University. Graphics lab motion capture database. *http://mocap.cs.cmu.edu/*.

[24] J. Van Gael, Y.W. Teh, and Z. Ghahramani. The infinite factorial hidden Markov model. In *Advances in Neural Information Processing Systems*, volume 21, pages 1697–1704, 2009.

[25] J.M. Wang, D.J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):283–298, 2008.

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

[27] F. Wood, T. L. Griffiths, and Z. Ghahramani. A non-parametric Bayesian method for inferring hidden causes. In *Proc. Conference on Uncertainty in Artificial Intelligence*, volume 22, 2006.