Bayesian Nonparametric Inference of Switching Dynamic Linear Models

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1109/tsp.2010.2102756">http://dx.doi.org/10.1109/tsp.2010.2102756</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Institute of Electrical and Electronics Engineers (IEEE)</td>
</tr>
<tr>
<td>Version</td>
<td>Author's final manuscript</td>
</tr>
<tr>
<td>Accessed</td>
<td>Thu Jan 10 02:37:47 EST 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/80811">http://hdl.handle.net/1721.1/80811</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution-Noncommercial-Share Alike 3.0</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc-sa/3.0/">http://creativecommons.org/licenses/by-nc-sa/3.0/</a></td>
</tr>
</tbody>
</table>
Bayesian Nonparametric Inference of
Switching Dynamic Linear Models

Emily Fox, Member, IEEE, Erik Sudderth, Member, IEEE, Michael Jordan, Fellow, IEEE,
and Alan Willsky, Fellow, IEEE

Abstract

Many complex dynamical phenomena can be effectively modeled by a system that switches among a set of conditionally linear dynamical modes. We consider two such models: the switching linear dynamical system (SLDS) and the switching vector autoregressive (VAR) process. Our Bayesian nonparametric approach utilizes a hierarchical Dirichlet process prior to learn an unknown number of persistent, smooth dynamical modes. We additionally employ automatic relevance determination to infer a sparse set of dynamic dependencies allowing us to learn SLDS with varying state dimension or switching VAR processes with varying autoregressive order. We develop a sampling algorithm that combines a truncated approximation to the Dirichlet process with efficient joint sampling of the mode and state sequences. The utility and flexibility of our model are demonstrated on synthetic data, sequences of dancing honey bees, the IBOVESPA stock index, and a maneuvering target tracking application.

I. INTRODUCTION

Linear dynamical systems (LDSs) are useful in describing dynamical phenomena as diverse as human motion [3], [4], financial time-series [5]–[7], maneuvering targets [8], [9], and the dance of honey bees [10]. However, such phenomena often exhibit structural changes over time, and the LDS models which describe them must also change. For example, a ballistic missile makes an evasive maneuver; a country experiences a recession, a central bank intervention, or some national or global event; a honey bee changes from a waggle to a turn right dance. Some of these changes will appear frequently, while others are only rarely observed. In addition, there is always the possibility of a new, previously unseen
dynamical behavior. These considerations motivate us to develop a Bayesian nonparametric approach for learning switching LDS (SLDS) models. We also consider a special case of the SLDS—the switching vector autoregressive (VAR) model—in which direct observations of the underlying dynamical process are assumed available.

One can view the SLDS, and the simpler switching VAR process, as an extension of hidden Markov models (HMMs) in which each HMM state, or mode, is associated with a linear dynamical process. Within the signal processing community, such HMM-based models have received considerable attention and proven useful in modeling the complex time evolution of signals. Specifically, HMMs have a long history of signal processing applications, with major success stories in speech processing (see the early influential tutorial by Rabiner [11]). While the HMM makes a strong Markovian assumption that observations are conditionally independent given the mode, the SLDS and switching VAR processes are able to capture more complex temporal dependencies often present in real data. Applications of switching linear dynamical processes, with roots in the control and econometrics literature, have recently become more prevalent within signal processing [10], [12]–[14]. However, most existing methods for learning SLDS and switching VAR processes rely on either fixing the number of HMM modes, such as in the preceding papers, or considering a change-point detection formulation where each inferred change is to a new, previously unseen dynamical mode, such as in [15]. There is growing interest in expanding the modeling framework to remove the purely parametric assumption of these previous formulations. In this paper we show how one can, in a seamless manner, remain agnostic about the number of dynamical modes while still allowing for returns to previously exhibited dynamical behaviors.

The rapidly developing field of Bayesian nonparametrics provides a new direction for analyzing HMMs with unknown state space cardinality. In particular, it has been shown that the hierarchical Dirichlet process (HDP) provides a useful prior on the HMM parameters [16], [17]. An alternative formulation of a Bayesian nonparametric HMM with application to music analysis has been presented in [18], though without the shared sparsity induced by the HDP. Another application of Bayesian nonparametrics to music analysis was presented in [19], where the authors propose Dirichlet process clustering of fixed-length segments of a time series, with each cluster modeling the dynamics of the given segments via a different finite HMM. See also [20] for a signal processing application of Dirichlet processes, specifically nonparametric modeling of excitations to a switching dynamical process. In this paper we make use of a variant of the HDP-HMM—the sticky HDP-HMM of [21]—to obtain improved control over the number of modes inferred; such control is crucial for the problems we examine. Our Bayesian nonparametric approach for learning switching dynamical processes extends the sticky HDP-HMM formulation to infer an unknown
number of persistent dynamical modes and thereby capture a wider range of temporal dependencies. We then explore a method for learning which components of the underlying state vector contribute to the dynamics of each mode by employing automatic relevance determination (ARD) [22]–[24]. The resulting model allows for learning realizations of SLDS that switch between an unknown number of dynamical modes with possibly varying state dimensions, or switching VAR processes with varying autoregressive orders.

A. Previous System Identification Techniques

Paoletti et. al. [25] provide a survey of recent approaches to identification of switching dynamical models. The most general formulation of the problem involves learning: (i) the number of dynamical modes, (ii) the model order, and (iii) the associated dynamic parameters. For noiseless switching VAR processes, Vidal et. al. [26] present an exact algebraic approach, though relying on fixing a maximal mode space cardinality and autoregressive order. Psaradakis and Spagnolo [27] alternatively consider a penalized likelihood approach to identification of stochastic switching VAR processes.

For SLDS, identification is significantly more challenging, and methods typically rely on simplifying assumptions such as deterministic dynamics or knowledge of the mode space. Huang et. al. [28] present an approach that assumes deterministic dynamics and embeds the input/output data in a higher-dimensional space and finds the switching times by segmenting the data into distinct subspaces [29]. Kotsalis et. al. [30] develop a balanced truncation algorithm for SLDS assuming the mode switches are i.i.d. within a fixed, finite set; the authors also present a method for model-order reduction of HMMs (see also [31]). In [32], a realization theory is presented for generalized jump-Markov linear systems in which the dynamic matrix depends both on the previous mode and current mode. Finally, when the number of dynamical modes is assumed known, Ghahramani and Hinton [33] present a variational approach to segmenting the data from a mixture of experts SLDS into the linear dynamical regimes and learning the associated dynamic parameters. For questions on observability and identifiability of SLDS in the absence of noise, see [34].

In the Bayesian approach that we adopt, we coherently incorporate noisy dynamics and uncertainty in the mode space cardinality. Our choice of prior penalizes more complicated models, both in terms of the number of modes and the state dimension describing each mode, allowing us to distinguish between the set of equivalent models described in [34]. Thus, instead of placing hard constraints on the model, we simply increase the posterior probability of simpler explanations of the data. As opposed to a penalized likelihood approach using Akaike’s information criterion (AIC) [35] or the Bayesian information criterion (BIC) [36], our approach provides a model complexity penalty in a purely Bayesian manner.
In Sec. II, we provide background on the switching linear dynamical systems we consider herein, and previous Bayesian nonparametric methods of learning HMMs. Our Bayesian nonparametric switching linear dynamical systems are described in Sec. III. We proceed by analyzing a conjugate prior on the dynamic parameters, and a sparsity-inducing prior that allows for variable-order switching processes. The section concludes by outlining a Gibbs sampler for the proposed models. In Sec. IV we present results on synthetic and real datasets, and in Sec. V we analyze a set of alternative formulations that are commonly found in the maneuvering target tracking and econometrics literature.

II. BACKGROUND

A. Switching Linear Dynamic Systems

A state space (SS) model consists of an underlying state, \( x_t \in \mathbb{R}^n \), with dynamics observed via \( y_t \in \mathbb{R}^d \). A linear time-invariant (LTI) SS model is given by

\[
\begin{align*}
    x_t &= A x_{t-1} + e_t \\
    y_t &= C x_t + w_t,
\end{align*}
\]

where \( e_t \) and \( w_t \) are independent Gaussian noise processes with covariances \( \Sigma \) and \( R \), respectively.

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

\[
    y_t = \sum_{i=1}^{r} A_i y_{t-i} + e_t \quad e_t \sim \mathcal{N}(0, \Sigma).
\]

Every \( \text{VAR}(r) \) process can be described in SS form, though not every SS model may be expressed as a \( \text{VAR}(r) \) process for finite \( r \) [37].

The dynamical phenomena we examine in this paper exhibit behaviors better modeled as switches between a set of linear dynamical models. We define a switching linear dynamical system (SLDS) by

\[
\begin{align*}
    z_t | z_{t-1} &\sim \pi_{z_{t-1}} \\
    x_t &= A^{(z_t)} x_{t-1} + e_{t}^{(z_t)} \\
    y_t &= C x_t + w_t.
\end{align*}
\]

The first-order Markov process \( z_t \) with transition distributions \( \{\pi_j\} \) indexes the mode-specific LDS at time \( t \), which is driven by Gaussian noise \( e_{t}^{(z_t)} \sim \mathcal{N}(0, \Sigma^{(z_t)}) \). One can view the SLDS as an extension of the classical hidden Markov model (HMM) [11], which has the same mode evolution, but conditionally independent observations:

\[
\begin{align*}
    z_t | z_{t-1} &\sim \pi_{z_{t-1}} \\
    y_t | z_t &\sim F(\theta_z)
\end{align*}
\]

for an indexed family of distributions \( F(\cdot) \) where \( \theta_z \) are the emission parameters for mode \( i \).
We similarly define a switching VAR(r) process by

\[ z_t \mid z_{t-1} \sim \pi_{z_{t-1}} \]
\[ y_t = \sum_{i=1}^{r} A_i(z_i) y_{t-i} + e_i^{(z_i)}. \]  

(5)

B. Dirichlet Processes and the Sticky HDP-HMM

To examine a Bayesian nonparametric SLDS, and thus relax the assumption that the number of dynamical modes is known and fixed, it is useful to first analyze such methods for the simpler HMM. One can equivalently represent the finite HMM of Eq. (4) via a set of transition probability measures

\[ G_j = \sum_{k=1}^{K} \pi_{jk} \delta_{\theta_k}, \]

where \( \delta_{\theta} \) is a mass concentrated at \( \theta \). We then operate directly in the parameter space \( \Theta \) and transition between emission parameters with probabilities given by \( \{G_j\} \). That is,

\[ \theta'_t \mid \theta'_{t-1} \sim G_j ; \theta'_{t-1} = \theta_j \]
\[ y_t \mid \theta'_t \sim F(\theta'_t). \]  

(6)

Here, \( \theta'_t \in \{\theta_1, \ldots, \theta_K\} \) and is equivalent to \( \theta_{z_t} \) of Eq. (4). A Bayesian nonparametric HMM takes \( G_j \) to be random with an infinite collection of atoms corresponding to the infinite HMM mode space.

The Dirichlet process (DP), denoted by \( \text{DP}(\gamma, H) \), provides a distribution over discrete probability measures with an infinite collection of atoms

\[ G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} ; \theta_k \sim H, \]  

(7)

on a parameter space \( \Theta \) that is endowed with a base measure \( H \). The weights are sampled via a stick-breaking construction [38]:

\[ \beta_k = \nu_k \prod_{\ell=1}^{k-1} (1 - \nu_\ell) \quad \nu_k \sim \text{Beta}(1, \gamma). \]  

(8)

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 \( \nu_k \) of the remaining stick after the previous \( (k - 1) \) weights have been defined. Letting \( \beta = [\beta_1 \beta_2 \ldots] \), we denote this distribution by \( \beta \sim \text{GEM}(\gamma) \).

The DP has proven useful in many applications due to its clustering properties, which are clearly seen by examining the predictive distribution of draws \( \theta'_t \sim G_0 \). Because probability measures drawn from a DP are discrete, there is a strictly positive probability of multiple observations \( \theta'_t \) taking identical values within the set \( \{\theta_k\} \), with \( \theta_k \) defined as in Eq. (7). For each value \( \theta'_t \), let \( z_t \) be an indicator random variable

1Formally, 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 random variable for each \( A \in \mathcal{A} \).
Namely, the mode-specific transition distributions are

\[ \theta_{\cdot t} \circ \pi_{\cdot k} \]

an urn representation of the \( \theta_{\cdot t} \). Blackwell and MacQueen [39] introduced a Polya urn representation of the \( \theta_{\cdot t} \):

\[
\theta_{i} | \theta_{1}', \ldots, \theta_{i-1}' \sim \frac{\gamma}{\gamma + i - 1} H + \sum_{j=1}^{i-1} \frac{1}{\gamma + i - 1} \delta_{\theta_{j}} = \frac{\gamma}{\gamma + i - 1} H + \sum_{k=1}^{K} \frac{n_{k}}{\gamma + i - 1} \delta_{\theta_{k}}. \tag{9}
\]

Here, \( n_{k} \) is the number of observations \( \theta_{i}' \) taking the value \( \theta_{k} \). From Eq. (9), and the discrete nature of \( G_{0} \), we see a reinforcement property of the DP that induces sparsity in the number of inferred mixture components.

A hierarchical extension of the DP, the hierarchical Dirichlet process (HDP) [16], has proven useful in defining a prior on the set of HMM transition probability measures \( G_{j} \). The HDP defines a collection of probability measures \( \{ G_{j} \} \) on the same support points \( \{ \theta_{1}, \theta_{2}, \ldots \} \) by assuming that each discrete measure \( G_{j} \) is a variation on a global discrete measure \( G_{0} \). Specifically, the Bayesian hierarchical specification takes \( G_{j} \sim DP(\alpha, G_{0}) \), with \( G_{0} \) itself a draw from a \( DP(\gamma, H) \). Through this construction, one can show that the probability measures are described as

\[
G_{0} = \sum_{k=1}^{\infty} \beta_{k} \delta_{\theta_{k}} \quad \beta | \gamma \sim \text{GEM}(\gamma) \quad \theta_{k} | H \sim H. \tag{10}
\]

Here, we use the notation \( \pi_{j} = [\pi_{j1} \pi_{j2} \ldots] \). Applying the HDP prior to the HMM, we obtain the \textit{HDP-HMM} of Teh et. al. [16]. This corresponds to the model in Fig. 1(a), but without the edges between the observations.

By defining \( \pi_{j} \sim DP(\alpha, \beta) \), the HDP prior encourages modes to have similar transition distributions. Namely, the mode-specific transition distributions are \textit{identical} in expectation:

\[
E[\pi_{jk} | \beta] = \beta_{k}. \tag{11}
\]
However, it does not differentiate self–transitions from moves between modes. When modeling dynamical processes with mode persistence, the flexible nature of the HDP-HMM prior allows for mode sequences with unrealistically fast dynamics to have large posterior probability. Recently, it has been shown [21] that one may mitigate this problem by instead considering a sticky HDP-HMM where $\pi_j$ is distributed as follows:

$$
\pi_j | \beta, \alpha, \kappa \sim \text{DP} \left( \alpha + \kappa, \frac{\alpha \beta + \kappa \delta_j}{\alpha + \kappa} \right).
$$

(12)

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

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

(13)

Here, $\delta(j, k)$ denotes the discrete Kronecker delta. When $\kappa = 0$ the original HDP-HMM of Teh et al. [16] is recovered. We place a prior on $\kappa$ and learn the self-transition bias from the data. See [21] for details.

III. THE HDP-SLDS AND HDP-AR-HMM

We now consider a significant extension of the sticky HDP-HMM for both SLDS and VAR modeling, capturing dynamic structure underlying the observations by allowing switches among an unknown number dynamical modes. Our proposed Bayesian nonparametric approach aims to capture these uncertainties. Additionally, the methodology allows both learning the number of modes and estimating the dimensionality and associated parameterization of the system state process. Fig. 1(b) illustrates the HDP-SLDS model, while Fig. 1(a) illustrates the HDP-AR-HMM model (for the case of VAR(2)). The generative processes for these two models are summarized in Table I.

The prior on the underlying discrete-valued Markov process $\{z_t\}$ is just as in the sticky HDP-HMM. The question now is in determining an appropriate base measure $H$ for the model parameters $\theta_k$. For
the HDP-SLDS, we place priors on the dynamic parameters \( \{A^{(k)}, \Sigma^{(k)}\} \) and on the measurement noise covariance \( R \) and infer their posterior from the data. Note that we assume the dynamics of the latent state process are 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 \( w^{(z_t)}_t \sim \mathcal{N}(0, R^{(z_t)}) \). However, such a choice is not always necessary nor appropriate for certain applications, and can have implications on the identifiability of the model. Based on a shared measurement matrix \( C \), we fix \( C = [I_d \ 0] \) without loss of generality, implying that it is the first \( d \) components of the state that are measured. Our choice of the state dimension \( n \) is, in essence, a choice of model order, and an issue we address in Sec. III-A2. For the HDP-AR-HMM, we similarly place a prior on the dynamic parameters, which in this case consist of \( \{A_1^{(k)}, \ldots, A_r^{(k)}, \Sigma^{(k)}\} \). Our specific choice of priors is discussed in Sec. III-A.

A Gibbs sampling inference scheme for our models is derived in Sec. III-B. There is, of course, a difference between the steps required for the SLDS-based model (in which there is an unobserved continuous-valued state \( x_t \)) and the AR-based model. In particular, for the HDP-SLDS the algorithm iterates among the following steps:

1) Sample the state sequence \( x_{1:T} \) given the mode sequence \( z_{1:T} \) and SLDS parameters \( \{A^{(k)}, \Sigma^{(k)}, R\} \).
2) Sample the mode sequence \( z_{1:T} \) given the state sequence \( x_{1:T} \), HMM parameters \( \{\pi_k\} \), and dynamic parameters \( \{A^{(k)}, \Sigma^{(k)}\} \).
3) Sample the HMM parameters \( \{\pi_k\} \) and SLDS parameters \( \{A^{(k)}, \Sigma^{(k)}, R\} \) given the sequences \( z_{1:T}, x_{1:T}, \) and \( y_{1:T} \).

For the HDP-AR-HMM, step (1) does not exist. Step (2) then involves sampling the mode sequence \( z_{1:T} \) given the observations \( y_{1:T} \) (rather than \( x_{1:T} \)), and step (3) involves conditioning solely on the sequences \( z_{1:T} \) and \( y_{1:T} \) (not \( x_{1:T} \)). Also, we note that step (2) involves a fairly straightforward extension of the sampling method developed in [21] for the simpler HDP-HMM model; the other steps, however, involve new constructs, as they require capturing and dealing with the temporal dynamics of the underlying continuous state models. Sec. III-A provides the structure of the posteriors needed to develop these steps.

### A. Priors and Posteriors of Dynamic Parameters

We begin by developing a prior to regularize the learning of the dynamic parameters (and measurement noise) conditioned on a fixed mode assignment \( z_{1:T} \). To make the connections between the samplers for the HDP-SLDS and HDP-AR-HMM explicit, we introduce the concept of pseudo-observations \( \psi_{1:T} \) and rewrite the dynamic equation for both the HDP-SLDS and HDP-AR-HMM generically as

\[
\psi_t = A^{(k)} \psi_{t-1} + e^{(k)}_t,
\]  

(14)
where we utilize the definitions outlined in Table II.

For the HDP-AR-HMM, we have simply written the dynamic equation in Table I in matrix form by concatenating the lag matrices into a single matrix $A^{(k)}$ and forming a lag observation vector $\tilde{\psi}_t$ comprised of a series of previous observation vectors. For this section (for the HDP-SLDS), we assume a sample of the state sequence $x_{1:T}$ (and hence $\{\psi_t, \bar{\psi}_t\}$) is available so that Eq. (14) applies equally well to both the HDP-SLDS and the HDP-AR-HMM. Methods for resampling this state sequence are discussed in Sec. III-B.

Conditioned on the mode sequence, one may partition this dynamic sequence into $K$ different linear regression problems, where $K = |\{z_1, \ldots, z_T\}|$. That is, for each mode $k$, we may form a matrix $\Psi^{(k)}$ with $n_k$ columns consisting of the $\psi_t$ with $z_t = k$. Then,

$$\Psi^{(k)} = A^{(k)} \bar{\Psi}^{(k)} + E^{(k)},$$

(15)

where $\bar{\Psi}^{(k)}$ is a matrix of the associated $\bar{\psi}_{t-1}$, and $E^{(k)}$ the associated noise vectors.

1) Conjugate Prior on $\{A^{(k)}, \Sigma^{(k)}\}$: The matrix-normal inverse-Wishart (MNIW) prior [40] is conjugate to the likelihood model defined in Eq. (15) for the parameter set $\{A^{(k)}, \Sigma^{(k)}\}$. Although this prior is typically used for inferring the parameters of a single linear regression problem, it is equally applicable to our scenario since the linear regression problems of Eq. (15) are independent conditioned on the mode sequence $z_{1:T}$. We note that while the MNIW prior does not enforce stability constraints on each mode, this prior is still a reasonable choice since each mode need not have stable dynamics for the SLDS to be stable [41], and conditioned on data from a stable mode, the posterior distribution will likely be sharply peaked around stable dynamic matrices.

Let $D^{(k)} = \{\Psi^{(k)}, \bar{\Psi}^{(k)}\}$. The posterior distribution of the dynamic parameters for the $k^{th}$ mode decomposes as

$$p(A^{(k)}, \Sigma^{(k)} \mid D^{(k)}) = p(A^{(k)} \mid \Sigma^{(k)}, D^{(k)}) p(\Sigma^{(k)} \mid D^{(k)}).$$

(16)

The resulting posterior of $A^{(k)}$ is straightforwardly derived to be (see [42])

$$p(A^{(k)} \mid \Sigma^{(k)}, D^{(k)}) = MN( S_{\psi\psi}^{-1}(k), \Sigma^{(k)} S_{\psi\psi}^{-1}(k) ),$$

(17)
with $\mathbf{B}^{-(k)}$ denoting $(\mathbf{B}^{(k)})^{-1}$ for a given matrix $\mathbf{B}$, $\mathcal{MN}(M, V, K)$ denoting a matrix-normal prior\(^2\) for $\mathbf{A}^{(k)}$ with mean matrix $M$ and left and right covariances $K^{-1}$ and $V$, and

$$
\mathbf{S}^{(k)}_{\psi \psi} = \bar{\Psi}^{(k)} \bar{\Psi}^{(k)T} + K \quad \mathbf{S}^{(k)}_{\psi \psi} = \bar{\Psi}^{(k)} \bar{\Psi}^{(k)T} + MK \quad \mathbf{S}^{(k)}_{\psi \psi} = \bar{\Psi}^{(k)} \bar{\Psi}^{(k)T} + MKM^T.
$$

The marginal posterior of $\Sigma^{(k)}$ is

$$
p(\Sigma^{(k)} | \mathbf{D}^{(k)}) = \text{IW} \left(n_k + n_0, \mathbf{S}^{(k)}_{\psi \psi} + S_0\right),
$$

where $\text{IW}(n_0, S_0)$ denotes an inverse-Wishart prior for $\Sigma^{(k)}$ with $n_0$ degrees of freedom and scale matrix $S_0$, and is updated by data terms $\mathbf{S}^{(k)}_{\psi \psi} = \mathbf{S}^{(k)} - \mathbf{S}^{(k)}_{-\psi \psi} \bar{\Psi}^{(k)T}$ and $n_k = |\{t \mid z_t = k, t = 1, \ldots, T\}|$.

2) Alternative Prior — Automatic Relevance Determination: The MNIW prior leads to full $\mathbf{A}^{(k)}$ matrices, which (i) becomes problematic as the model order grows in the presence of limited data; and (ii) does not provide a method for identifying irrelevant model components (i.e. state components in the case of the HDP-SLDS or lag components in the case of the HDP-AR-HMM.) To jointly address these issues, we alternatively consider automatic relevance determination (ARD) [22]–[24], which encourages driving components of the model parameters to zero if their presence is not supported by the data.

For the HDP-SLDS, we harness the concepts of ARD by placing independent, zero-mean, spherically symmetric Gaussian priors on the columns of the dynamic matrix $\mathbf{A}^{(k)}$:

$$
p(\mathbf{A}^{(k)} | \alpha^{(k)}) = \prod_{j=1}^n \mathcal{N} \left(\alpha_j^{(k)}; 0, \alpha_j^{-(k)} I_n\right).
$$

Each precision parameter $\alpha_j^{(k)}$ is given a Gamma$(a, b)$ prior. The zero-mean Gaussian prior penalizes non-zero columns of the dynamic matrix by an amount determined by the precision parameters. Iterative estimation of these hyperparameters $\alpha_j^{(k)}$ and the dynamic matrix $\mathbf{A}^{(k)}$ leads to $\alpha_j^{(k)}$ becoming large for columns whose evidence in the data is insufficient for overcoming the penalty induced by the prior. Having $\alpha_j^{(k)} \rightarrow \infty$ drives $\alpha_j^{(k)} \rightarrow 0$, implying that the $j^{th}$ state component does not contribute to the dynamics of the $k^{th}$ mode. Thus, examining the set of large $\alpha_j^{(k)}$ provides insight into the order of that mode. Looking at the $k^{th}$ dynamical mode alone, having $\alpha_j^{(k)} = 0$ implies that the realization of that mode is not minimal since the associated Hankel matrix

$$
\mathcal{H} = \left[\mathbf{C}^T \quad \mathbf{C} \mathbf{A}^T \quad \cdots \quad \mathbf{C} \mathbf{A}^{d-1} \right]^T \begin{bmatrix} \mathbf{G} & \mathbf{A} \mathbf{G} & \cdots & \mathbf{A}^{d-1} \mathbf{G} \end{bmatrix} \equiv \mathcal{O} \mathcal{R}
$$

has reduced rank. However, the overall SLDS realization may still be minimal.

For our use of the ARD prior, we restrict attention to models satisfying the property that the state components that are observed are relevant to all modes of the dynamics:

\(^2\)If $\mathbf{A} \sim \mathcal{MN}(M, V, K)$, then $\text{vec}(\mathbf{A}) \sim \mathcal{N}(\text{vec}(M), K^{-1} \otimes V)$, with $\otimes$ denoting the Kronecker product.
Criterion 3.1: If for some realization $R$ a mode $k$ has $a_j^{(k)} = 0$, then that realization must have $c_j = 0$, where $c_j$ is the $j^{th}$ column of $C$. Here we assume, without loss of generality, that the observed states are the first components of the state vector.

This assumption implies that our choice of $C = [I_d \ 0]$ does not interfere with learning a sparse realization. We could avoid restricting our attention to models satisfying Criterion 3.1 by considering a more general model where the measurement equation is mode-specific and we place a prior on $A^{(k)}$ instead of fixing this matrix. However, this model leads to identifiability issues that are considerably less pronounced in the above case.

The ARD prior may also be used to learn variable-order switching VAR processes. Here, the goal is to “turn off” entire lag blocks $A_i^{(k)}$ (whereas in the HDP-SLDS we were interested in eliminating columns of the dynamic matrix.) Instead of placing independent Gaussian priors on each column of $A^{(k)}$ as we did in Eq. (20), we decompose the prior over the lag blocks $A_i^{(k)}$:

$$p(A^{(k)} | \alpha^{(k)}) = \prod_{i=1}^{r} \mathcal{N}\left(\text{vec}(A_i^{(k)}); 0, \alpha_i^{−1(k)} I_d \right).$$

(22)

Since each element of a given lag block $A_i^{(k)}$ is distributed according to the same precision parameter $\alpha_i^{(k)}$, if that parameter becomes large the entire lag block will tend to zero.

In order to examine the posterior distribution on the dynamic matrix $A^{(k)}$, it is useful to consider the Gaussian induced by Eq. (20) and Eq. (22) on a vectorization of $A^{(k)}$. Our ARD prior on $A^{(k)}$ is equivalent to a $\mathcal{N}(0, \Sigma_0^{(k)})$ prior on vec($A^{(k)}$), where

$$\Sigma_0^{(k)} = \text{diag}\left(\alpha_1^{(k)}, \ldots, \alpha_1^{(k)}, \ldots, \alpha_m^{(k)}, \ldots, \alpha_m^{(k)}\right)^{-1}.$$

(23)

Here, $m = n$ for the HDP-SLDS with $n$ replicates of each $\alpha_i^{(k)}$, and $m = r$ for the HDP-AR-HMM with $d^2$ replicates of $\alpha_i^{(k)}$. (Recall that $n$ is the dimension of the HDP-SLDS state vector $x_t$, $r$ the autoregressive order of the HDP-AR-HMM, and $d$ the dimension of the observations $y_t$.) To examine the posterior distribution of $A^{(k)}$, we note that we may rewrite the state equation as,

$$\psi_{t+1} = \begin{bmatrix} \tilde{T}_{t,1}I_\ell & \tilde{T}_{t,2}I_\ell & \cdots & \tilde{T}_{t,r}I_\ell \end{bmatrix} \text{vec}(A^{(k)}) + e_{t+1}^{(k)} \quad \forall t | z_t = k$$

$$\Delta \tilde{T}_t \text{vec}(A^{(k)}) + e_{t+1}^{(k)},$$

(24)

where $\ell = n$ for the HDP-SLDS and $\ell = d$ for the HDP-AR-HMM. Using Eq. (24), we derive the posterior distribution as

$$p(\text{vec}(A^{(k)}) | D^{(k)}, \Sigma^{(k)}, \Sigma_0^{(k)}) = \mathcal{N}^{-1}\left(\sum_{t | z_t = k} \tilde{T}_{t-1}^{T} \Sigma_0^{−1} \tilde{T}_{t-1} \psi_t, \Sigma_0^{−1} + \sum_{t | z_t = k} \tilde{T}_{t-1}^{T} \Sigma^{−1} \tilde{T}_{t-1} \right).$$

(25)
See [42] for a detailed derivation. Here, \( \mathcal{N}^{-1}(\vartheta, \Lambda) \) represents a Gaussian \( \mathcal{N}(\mu, \Sigma) \) with information parameters \( \vartheta = \Sigma^{-1}\mu \) and \( \Lambda = \Sigma^{-1} \). Given \( A^{(k)} \), and recalling that each precision parameter is gamma distributed, the posterior of \( \alpha^{(k)}_\ell \) is given by

\[
p(\alpha^{(k)}_\ell \mid A^{(k)}) = \text{Gamma} \left( a + \frac{|S_\ell|}{2}, b + \frac{\sum_{(i,j) \in S_\ell} a^{(k)}_{ij}^2}{2} \right).
\]

(26)

The set \( S_\ell \) contains the indices for which \( a^{(k)}_{ij} \) has prior precision \( \alpha^{(k)}_\ell \). Note that in this model, regardless of the number of observations \( y_t \), the size of \( S_\ell \) (i.e., the number of \( a^{(k)}_{ij} \) used to inform the posterior distribution) remains the same. Thus, the gamma prior is an informative prior and the choice of \( a \) and \( b \) should depend upon the cardinality of \( S_\ell \) (see Sec. IV-B for an example). For the HDP-SLDS, this cardinality is given by the maximal state dimension \( n \), and for the HDP-AR-HMM, by the square of the observation dimensionality \( d^2 \).

We then place an inverse-Wishart prior \( \text{IW}(n_0, S_0) \) on \( \Sigma^{(k)} \) and look at the posterior given \( A^{(k)} \):

\[
p(\Sigma^{(k)} \mid D^{(k)}, A^{(k)}) = \text{IW} \left( n_k + n_0, S^{(k)}_{\psi|\bar{\psi}} + S_0 \right),
\]

(27)

where here, as opposed to in Eq. (19), we define

\[
S^{(k)}_{\psi|\bar{\psi}} = \sum_{t|z_t=k} (\psi_t - A^{(k)}\bar{\psi}_{t-1})(\psi_t - A^{(k)}\bar{\psi}_{t-1})^T.
\]

(28)

3) Measurement Noise Posterior: For the HDP-SLDS, we additionally place an \( \text{IW}(r_0, R_0) \) prior on the measurement noise covariance \( R \). The posterior distribution is given by

\[
p(R \mid y_{1:T}, x_{1:T}) = \text{IW}(T + r_0, S_R + R_0),
\]

(29)

where \( S_R = \sum_{t=1}^T (y_t - Cx_t)(y_t - Cx_t)^T \). Here, we assume that \( R \) is shared between modes. The extension to mode-specific measurement noise is straightforward.

B. Gibbs Sampler

For inference in the HDP-AR-HMM, we use a Gibbs sampler that iterates between sampling the mode sequence, \( z_{1:T} \), and the set of dynamic and sticky HDP-HMM parameters. The sampler for the HDP-SLDS is identical with the additional step of sampling the state sequence, \( x_{1:T} \), and conditioning on this sequence when resampling dynamic parameters and the mode sequence. Periodically, we interleave a step that sequentially samples the mode sequence \( z_{1:T} \) marginalizing over the state sequence \( x_{1:T} \) in a similar vein to that of Carter and Kohn [43]. We describe the sampler in terms of the pseudo-observations \( \psi_t \), as defined by Eq. (14), in order to clearly specify the sections of the sampler shared by both the HDP-AR-HMM and HDP-SLDS.
1) Sampling Dynamic Parameters \(\{A^{(k)}, \Sigma^{(k)}\}\): Conditioned on the mode sequence, \(z_{1:T}\), and the pseudo-observations, \(\psi_{1:T}\), we can sample the dynamic parameters \(\theta = \{A^{(k)}, \Sigma^{(k)}\}\) from the posterior densities of Sec. III-A. For the ARD prior, we then sample \(\alpha^{(k)}\) given \(A^{(k)}\). In practice we iterate multiple times between sampling \(\alpha^{(k)}\) given \(A^{(k)}\) and \(A^{(k)}\) given \(\alpha^{(k)}\) before moving to the next sampling stage.

2) Sampling Measurement Noise \(R\) (HDP-SLDS only): For the HDP-SLDS, we additionally sample the measurement noise covariance \(R\) conditioned on the sampled state sequence \(x_{1:T}\).

3) Block Sampling \(z_{1:T}\): As shown in [21], the mixing rate of the Gibbs sampler for the HDP-HMM can be dramatically improved by using a truncated approximation to the HDP and jointly sampling the mode sequence using a variant of the forward-backward algorithm. In the case of our switching dynamical systems, we must account for the direct correlations in the observations in our likelihood computation. The variant of the forward-backward algorithm we use here then involves computing backward messages \(m_{t+1,t}(z_t) \propto p(\psi_{t+1:T}^1 | z_t, \bar{\psi}_{t}, \pi, \theta) \) for each \(z_t \in \{1, \ldots, L\}\) with \(L\) the chosen truncation level, followed by recursively sampling each \(z_t\) conditioned on \(z_{t-1}\) from

\[
p(z_t | z_{t-1}, \psi_{1:T}, \pi, \theta) \propto \pi_{z_{t-1}, z_t} p(\psi_t | \bar{\psi}_{t-1}, A^{(z_t)}, \Sigma^{(z_t)}) m_{t+1,t}(z_t).
\]

Joint sampling of the mode sequence is especially important when the observations are directly correlated via a dynamical process since this correlation further slows the mixing rate of the sequential sampler of Teh et. al. [16]. Note that using an order \(L\) weak limit approximation to the HDP still encourages the use of a sparse subset of the \(L\) possible dynamical modes.

4) Block Sampling \(x_{1:T}\) (HDP-SLDS only): Conditioned on the mode sequence \(z_{1:T}\) and the set of SLDS parameters \(\theta = \{A^{(k)}, \Sigma^{(k)}, R\}\), our dynamical process simplifies to a time-varying linear dynamical system. We can then block sample \(x_{1:T}\) by first running a backward Kalman filter to compute \(m_{t+1,t}(x_t) \propto p(y_{t+1:T}^1 | x_t, z_{t+1:T}, \theta)\) and then recursively sampling each \(x_t\) conditioned on \(x_{t-1}\) from

\[
p(x_t | x_{t-1}, y_{1:T}^1, z_{1:T}, \theta) \propto p(x_t | x_{t-1}, A^{(z_t)}, \Sigma^{(z_t)}) p(y_t | x_t, R) m_{t+1,t}(x_t).
\]

The messages are given in information form by \(m_{t,t-1}(x_{t-1}) \propto \mathcal{N}^{-1}(\theta_{t,t-1}, \Lambda_{t,t-1})\), where the information parameters are recursively defined as

\[
\begin{align*}
\theta_{t,t-1} &= A^{(z_t)^T} \Sigma^{-1}(z_t) (\Sigma^{-1}(z_t) + A_{tt}^b)^{-1} \phi_{tt}^b \\
\Lambda_{t,t-1} &= A^{(z_t)^T} \Sigma^{-1}(z_t) A^{(z_t)} - A^{(z_t)^T} \Sigma^{-1}(z_t) (\Sigma^{-1}(z_t) + A_{tt}^b)^{-1} \Sigma^{-1}(z_t) A^{(z_t)}.
\end{align*}
\]

\(\phi_{tt}^b\) is the pseudo-observation for \(t\) given \(\phi_{tt}^b\), and \(\Lambda_{tt}^b\) is the pseudo-observation noise covariance.
The standard \( \vartheta^b_{t \mid t} \) and \( \Lambda^b_{t \mid t} \) updated information parameters for a backward running Kalman filter are given by
\[
\Lambda^b_{t \mid t} = C^T R^{-1} C + \Lambda_{t+1 \mid t} \\
\vartheta^b_{t \mid t} = C^T R^{-1} y_t + \vartheta_{t+1 \mid t}.
\] (33)

See [42] for a derivation and for a more numerically stable version of this recursion.

5) Sequentially Sampling \( z_{1:T} \) (HDP-SLDS only): For the HDP-SLDS, iterating between the previous sampling stages can lead to slow mixing rates since the mode sequence is sampled conditioned on a sample of the state sequence. For high-dimensional state spaces \( \mathbb{R}^n \), this problem is exacerbated. Instead, one can analytically marginalize the state sequence and sequentially sample the mode sequence from \( p(z_t \mid z_{1:t}, y_{1:T}, \pi, \theta) \). This marginalization is accomplished by once again harnessing the fact that conditioned on the mode sequence, our model reduces to a time-varying linear dynamical system. When sampling \( z_t \) and conditioning on the mode sequence at all \textit{other} time steps, we can run a forward Kalman filter to marginalize the state sequence \( x_{1:t−2} \) producing \( p(x_{t−1} \mid y_{1:t−1}, z_{1:t−1}, \theta) \), and a backward filter to marginalize \( x_{t+1:T} \) producing \( p(y_{t+1:T} \mid x_t, z_{t+1:T}, \theta) \). Then, for each possible value of \( z_t \), we combine these forward and backward messages with the local likelihood \( p(y_t \mid x_t) \) and local dynamic \( p(x_t \mid x_{t−1}, \theta, z_t = k) \) and marginalize over \( x_t \) and \( x_{t−1} \) resulting in the likelihood of the observation sequence \( y_{1:T} \) as a function of \( z_t \). This likelihood is combined with the prior probability of transitioning from \( z_{t−1} \) to \( z_t = k \) and from \( z_t = k \) to \( z_{t+1} \). The resulting distribution is given by:

\[
p(z_t = k \mid z_{1:t}, y_{1:T}, \pi, \theta) \propto \pi_{z_{t−1}, k} \pi_{k, z_{t+1}} \\
\frac{|\Lambda^b_{t \mid t}|^{1/2}}{|\Lambda_t^{(k)} + \Lambda^b_{t \mid t}|^{1/2}} \exp \left( -\frac{1}{2} \vartheta_t^T \Lambda_t^{−(k)} \vartheta_t + \frac{1}{2} \vartheta_t^T (\Lambda_t^{(k)} + \Lambda^b_{t \mid t})^{-1} \vartheta_t + \vartheta_t^T (\Lambda^b_{t \mid t})^{-1} \vartheta_t \right)
\] (34)

with
\[
\Lambda_t^{(k)} = (\Sigma^{(k)} + \boldsymbol{A}^{(z_t)} \Lambda_{t−1 \mid t−1}^{−f} \boldsymbol{A}^{(z_t)^T})^{-1} \\
\vartheta_t^{(k)} = (\Sigma^{(k)} + \boldsymbol{A}^{(z_t)} \Lambda_{t−1 \mid t−1}^{−f} \boldsymbol{A}^{(z_t)^T})^{-1} \boldsymbol{A}^{(z_t)} \Lambda_{t−1 \mid t−1}^{−f} \vartheta_{t−1 \mid t−1}.
\] (35)

See [42] for full derivations. Here, \( \vartheta^f_{t \mid t} \) and \( \Lambda^f_{t \mid t} \) are the updated information parameters for a forward running Kalman filter, defined recursively as
\[
\Lambda^f_{t \mid t} = C^T R^{-1} C + \Sigma^{−(z_t)} - \Sigma^{−(z_t)} \boldsymbol{A}^{(z_t)} (\boldsymbol{A}^{(z_t)^T} \Sigma^{−(z_t)} \boldsymbol{A}^{(z_t)} + \Lambda^f_{t−1 \mid t−1})^{-1} \boldsymbol{A}^{(z_t)^T} \Sigma^{−(z_t)} \\
\vartheta^f_{t \mid t} = C^T R^{-1} y_t + \Sigma^{−(z_t)} \boldsymbol{A}^{(z_t)} (\boldsymbol{A}^{(z_t)^T} \Sigma^{−(z_t)} \boldsymbol{A}^{(z_t)} + \Lambda^f_{t−1 \mid t−1})^{-1} \vartheta^f_{t−1 \mid t−1}.
\] (36)

Note that a sequential node ordering for this sampling step allows for efficient updates to the recursively defined filter parameters. However, this sequential sampling is still computationally intensive, so our
Gibbs sampler iterates between blocked sampling of the state and mode sequences many times before interleaving a sequential mode sequence sampling step.

The resulting Gibbs sampler is outlined in Algorithm 1.

IV. RESULTS

A. MNIW prior

We begin by examining a set of three synthetic datasets displayed in Fig. 2(a) in order to analyze the relative modeling power of the HDP-VAR(1)-HMM, HDP-VAR(2)-HMM, and HDP-SLDS using the MNIW prior. Here, we use the notation HDP-VAR(r)-HMM to explicitly denote an order r HDP-AR-HMM with vector observations. We compare to a baseline sticky HDP-HMM using first difference observations, imitating a HDP-VAR(1)-HMM with $A^{(k)} = I$ for all $k$. In Fig. 2(b)-(e) we display Hamming distance errors that are calculated by choosing the optimal mapping of indices maximizing overlap between the true and estimated mode sequences.

We place a Gamma($a, b$) prior on the sticky HDP-HMM concentration parameters $\alpha + \kappa$ and $\gamma$, and a Beta($c, d$) prior on the self-transition proportion parameter $\rho = \kappa / (\alpha + \kappa)$. We choose the weakly informative setting of $a = 1$, $b = 0.01$, $c = 10$, and $d = 1$. The details on setting the MNIW hyperparameters from statistics of the data are discussed in the Appendix.

For the first scenario (Fig. 2 (top)), the data were generated from a five-mode switching VAR(1) process with a 0.98 probability of self-transition and equally likely transitions to the other modes. The same mode-transition structure was used in the subsequent two scenarios, as well. The three switching linear dynamical models provide comparable performance since both the HDP-VAR(2)-HMM and HDP-SLDS with $C = I_3$ contain the class of HDP-VAR(1)-HMMs. In the second scenario (Fig. 2 (middle)), the data were generated from a 3-mode switching AR(2) process. The HDP-AR(2)-HMM has significantly better performance than the HDP-AR(1)-HMM while the performance of the HDP-SLDS with $C = [1 \ 0]$ performs similarly, but has greater posterior variability because the HDP-AR(2)-HMM model family is smaller. Note that the HDP-SLDS sampler is slower to mix since the hidden, continuous state is also sampled. The data in the third scenario (Fig. 2 (bottom)) were generated from a three-mode SLDS model with $C = I_3$. Here, we clearly see that neither the HDP-VAR(1)-HMM nor HDP-VAR(2)-HMM is equivalent to the HDP-SLDS. Note that all of the switching models yielded significant improvements relative to the baseline sticky HDP-HMM. This input representation is more effective than using raw observations for HDP-HMM learning, but still much less effective than richer models which switch among learned LDS. Together, these results demonstrate both the differences between our models as well as the models’ ability to learn switching processes with varying numbers of modes.
Given a previous set of mode-specific transition probabilities $\pi^{(n-1)}$, the global transition distribution $\beta^{(n-1)}$, and dynamic parameters $\theta^{(n-1)}$:

1) Set $\pi = \pi^{(n-1)}$, $\beta = \beta^{(n-1)}$, and $\theta = \theta^{(n-1)}$.

2) If HDP-SLDS,
   a) For each $t \in \{1, \ldots, T\}$, compute $\{\varphi_t f_t^t, \Lambda_t f_t^t\}$ as in Eq. (36).
   b) For each $t \in \{T, \ldots, 1\}$,
      i) Compute $\{\varphi_t b_t^t, \Lambda_t b_t^t\}$ as in Eq. (33).
      ii) For each $k \in \{1, \ldots, L\}$, compute $\{\varphi_t (k)^t, \Lambda_t (k)^t\}$ as in Eq. (35) and set
          
          \[
          f_k(y_{1:T}) = \frac{1}{2} \rho_t^{(k)} \Lambda_t^{(k)} - 1 \rho_t^{(k)} T (\Lambda_t^{(k)} + \Lambda_t b_t^t)^{-1} \rho_t^{(k)} + \frac{1}{2} \rho_t^{(k)} (\Lambda_t^{(k)} + \Lambda_t b_t^t)^{-1} (\rho_t^{(k)} + \varphi_t b_t^t) \] 
          
      iii) Sample a mode assignment
          
          \[
          z_t \sim \sum_{k=1}^L \pi_{z_{t-1},k} \pi_{k,z_{t+1}} f_k(y_{1:T}) \delta(z_t, k). \]
   c) Working sequentially forward in time sample
          
          \[
          x_t \sim N((\Sigma^{-(z_t)} + \Lambda_t b_t^t)^{-1} (\Sigma^{-(z_t)} A^{(z_t)} x_{t-1} + \varphi_t b_t^t), (\Sigma^{-(z_t)} + \Lambda_t b_t^t)^{-1}). \]
   d) Set pseudo-observations $\psi_{1:T} = x_{1:T}$.

3) If HDP-AR-HMM, set pseudo-observations $\psi_{1:T} = y_{1:T}$.

4) Block sample $z_{1:T}$ given transition distributions $\pi$, dynamic parameters $\theta$, and pseudo-observations $\psi_{1:T}$ as in Algorithm 2.

5) Update the global transition distribution $\beta$ (utilizing auxiliary variables $m$, $w$, and $\bar{m}$), mode-specific transition distributions $\pi_k$, and hyperparameters $\alpha$, $\gamma$, and $\kappa$ as in [21].

6) For each $k \in \{1, \ldots, L\}$, sample dynamic parameters $\{A^{(k)}, \Sigma^{(k)}\}$ given the pseudo-observations $\psi_{1:T}$ and mode sequence $z_{1:T}$ as in Algorithm 3 for the MNIW prior and Algorithm 4 for the ARD prior.

7) If HDP-SLDS, also sample the measurement noise covariance

          \[
          R \sim IW(T + r_0, \sum_{t=1}^T (y_t - C x_t)(y_t - C x_t)^T + R_0) \] 

8) Fix $\pi^{(n)} = \pi$, $\beta^{(n)} = \beta$, and $\theta^{(n)} = \theta$.

\textbf{Algorithm 1:} HDP-SLDS and HDP-AR-HMM Gibbs sampler.
Given mode-specific transition probabilities $\pi$, dynamic parameters $\theta$, and pseudo-observations $\psi_{1:T}$:

1) Calculate messages $m_{t,t-1}(k)$, initialized to $m_{T+1,T}(k) = 1$, and the sample mode sequence $z_{1:T}$:
   a) For each $t \in \{T, \ldots, 1\}$ and $k \in \{1, \ldots, L\}$, compute
   $$m_{t,t-1}(k) = \sum_{j=1}^{L} \pi_{kj} \mathcal{N}\left(\psi_t; \sum_{i=1}^{r} A_{i}^{(j)} \psi_{t-i}, \Sigma_{i}^{(j)}\right) m_{t+1,t}(j)$$
   b) Working sequentially forward in time, starting with transitions counts $n_{jk} = 0$:
      i) For each $k \in \{1, \ldots, L\}$, compute the probability
         $$f_k(\psi_t) = \mathcal{N}\left(y_t; \sum_{i=1}^{r} A_{i}^{(k)} \psi_{t-i}, \Sigma_{i}^{(k)}\right) m_{t+1,t}(k)$$
      ii) Sample a mode assignment $z_t$ as follows and increment $n_{z_{t-1}z_t}$:
          $$z_t \sim \sum_{k=1}^{L} \pi_{z_{t-1},k} f_k(\psi_t) \delta(z_t, k)$$
   Note that the likelihoods can be precomputed for each $k \in \{1, \ldots, L\}$.

Algorithm 2: Blocked mode-sequence sampler for HDP-AR-HMM or HDP-SLDS.

Given pseudo-observations $\psi_{1:T}$ and mode sequence $z_{1:T}$, for each $k \in \{1, \ldots, K\}$:

1) Construct $\Psi^{(k)}$ and $\bar{\Psi}^{(k)}$ as in Eq. (15).
2) Compute sufficient statistics using pseudo-observations $\psi_t$ associated with $z_t = k$:
   $$S_{\psi\psi}^{(k)} = \bar{\Psi}^{(k)} \bar{\Psi}^{(k)^T} + K \quad S_{\psi\psi}^{(k)} = \Psi^{(k)} \bar{\Psi}^{(k)^T} + MK \quad S_{\psi\psi}^{(k)} = \Psi^{(k)} \Psi^{(k)^T} + MKM^T$$
3) Sample dynamic parameters:
   $$\Sigma^{(k)} \sim \text{IW}\left(n_k + n_0, S_{\psi\psi}^{(k)} + S_0\right) \quad A^{(k)} | \Sigma^{(k)} \sim \mathcal{M}\mathcal{N}\left(S_{\psi\psi}^{(k)} S_{\psi\psi}^{-1} \Sigma^{(k)}, S_{\psi\psi}^{(k)}\right)$$

Algorithm 3: Parameter sampling using MNIW prior.

B. ARD prior

We now compare the utility of the ARD prior to the MNIW prior using the HDP-SLDS model when the true underlying dynamical modes have sparse dependencies relative to the assumed model order. That is, the HDP-SLDS may have dynamical regimes reliant on lower state dimensions, or the HDP-AR-HMM may have modes described by lower order VAR processes. We generated data from a two-mode SLDS
Given pseudo-observations $\psi_{1:T}$, mode sequence $z_{1:T}$, and a previous set of dynamic parameters $(A^{(k)}, \Sigma^{(k)}, \alpha^{(k)})$, for each $k \in \{1, \ldots, K\}$:

1) Construct $\tilde{\Psi}_t$ as in Eq. (24).

2) Iterate multiple times between the following steps:
   
   a) Construct $\Sigma_0^{(k)}$ given $\alpha^{(k)}$ as in Eq. (23) and sample the dynamic matrix:
   
   $$\text{vec}(A^{(k)}) | \Sigma^{(k)}, \alpha^{(k)} \sim \mathcal{N}^{-1} \left( \sum_{t \mid z_t = k} \tilde{\Psi}_{t-1}^T \Sigma^{-(k)} \psi_t, \Sigma^{-(k)} \psi_0 + \sum_{t \mid z_t = k} \tilde{\Psi}_{t-1}^T \Sigma^{-(k)} \tilde{\Psi}_{t-1} \right).$$

   b) For each $\ell \in \{1, \ldots, m\}$, with $m = n$ for the SLDS and $m = r$ for the switching V AR, sample ARD precision parameters:
   $$\alpha^{(k)}_{\ell} | A^{(k)} \sim \text{Gamma} \left( a + \frac{|S_\ell|}{2}, b - \frac{\sum_{(i,j) \in S_\ell} a^{(k)}_{ij}^2}{2} \right).$$

   c) Compute sufficient statistic:
   $$S^{(k)}_{\psi|\psi} = \sum_{t \mid z_t = k} (\psi_t - A^{(k)} \tilde{\psi}_{t-1})(\psi_t - A^{(k)} \tilde{\psi}_{t-1})^T$$

   and sample process noise covariance:
   $$\Sigma^{(k)} | A^{(k)} \sim \text{IW} \left( n_k + n_0, S^{(k)}_{\psi|\psi} + S_0 \right).$$

**Algorithm 4:** Parameter sampling using ARD prior.

with 0.98 probability of self-transition and

$$A^{(1)} = \begin{bmatrix} 0.8 & -0.2 & 0 \\ -0.2 & 0.8 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A^{(2)} = \begin{bmatrix} -0.2 & 0 & 0.8 \\ 0.8 & 0 & -0.2 \\ 0 & 0 & 0 \end{bmatrix},$$

with $C = [I_2 \ 0]$, $\Sigma^{(1)} = \Sigma^{(2)} = I_3$, and $R = I_2$. The first dynamical process can be equivalently described by just the first and second state components since the third component is simply white noise that does not contribute to the state dynamics and is not directly (or indirectly) observed. For the second dynamical process, the third state component is once again a white noise process, but does contribute to the dynamics of the first and second state components. However, we can equivalently represent the
Fig. 2. (a) Observation sequence (blue, green, red) and associated mode sequence (magenta) for a 5-mode switching VAR(1) process (top), 3-mode switching AR(2) process (middle), and 3-mode SLDS (bottom). The components of the observation vector are offset for clarity. The associated 10th, 50th, and 90th Hamming distance quantiles over 100 trials are shown for the (b) HDP-VAR(1)-HMM, (c) HDP-VAR(2)-HMM, (d) HDP-SLDS with $C = I$ (top and bottom) and $C = [1 \ 0]$ (middle), and (e) sticky HDP-HMM using first difference observations.

dynamics of this mode as
\[
\begin{align*}
x_{1,t} &= -0.2 x_{1,t-1} + \tilde{\epsilon}_{1,t} \\
x_{2,t} &= 0.8 x_{1,t-1} + \tilde{\epsilon}_{2,t}
\end{align*}
\]
where $\tilde{\epsilon}_t$ is a white noise term defined by the original process noise combined with $x_{3,t}$, and $\tilde{A}^{(2)}$ is the dynamical matrix associated with this equivalent representation of the second dynamical mode. Notice that this SLDS does not satisfy Criterion 3.1 since the second column of $A^{(2)}$ is zero while the second column of $C$ is not. Nevertheless, because the realization is in our canonical form with $C = [I_2 \ 0]$, we still expect to recover the $a_2^{(2)} = a_3^{(2)} = 0$ sparsity structure. We set the parameters of the Gamma($\alpha, \beta$) prior on the ARD precisions as $\alpha = |S_\ell|$ and $\beta = \alpha/1000$, where we recall the definition of $S_\ell$ from Eq. (26). This specification fixes the mean of the prior to 1000 while aiming to provide a prior that is roughly equally informative for various choices of model order (i.e., sizes $|S_\ell|$).

In Fig. 3, we see that even in this low-dimensional example, the ARD provides superior mode-sequence
Fig. 3. (a) Observation sequence (green, blue) and mode sequence (magenta) of a 2-mode SLDS, where the first mode can be realized by the first two state components and the second mode solely by the first. The associated 10th, 50th, and 90th Hamming distance quantiles over 100 trials are shown for the (b) MNIW and (c) ARD prior. (d) Box plots of inferred ARD precisions associated with the first and second dynamical modes at the 5000th Gibbs iteration. The center line indicates the median, edges the 25th and 75th quantiles, and whiskers the range of data excluding outliers which are separately marked. Larger ARD precision values correspond to non-dynamical components.

estimates, as well as a mechanism for identifying non-dynamical state components. The box plots of the inferred $\alpha^{(k)}$ are shown in Fig. 3(d). From the clear separation between the sampled dynamic range of $\alpha_3^{(1)}$ and $(\alpha_1^{(1)}, \alpha_2^{(1)})$, and between that of $(\alpha_2^{(2)}, \alpha_3^{(2)})$ and $\alpha_1^{(2)}$, we see that we are able to correctly identify dynamical systems with $\alpha_3^{(1)} = 0$ and $\alpha_2^{(2)} = \alpha_3^{(2)} = 0$.

C. Dancing Honey Bees

Honey bees perform a set of dances within the beehive in order to communicate the location of food sources. Specifically, they switch between a set of waggle, turn-right, and turn-left dances. During the waggle dance, the bee walks roughly in a straight line while rapidly shaking its body from left to right. The turning dances simply involve the bee turning in a clockwise or counterclockwise direction. We display six such sequences of honey bee dances in Fig. 4. The data consist of measurements $y_t = [\cos(\theta_t) \sin(\theta_t) \ x_t \ y_t]^T$, where $(x_t, y_t)$ denotes the 2D coordinates of the bee’s body and $\theta_t$ its head angle. Both Oh et al. [10] and Xuan and Murphy [15] used switching dynamical models to analyze these honey bee dances. We wish to analyze the performance of our Bayesian nonparametric variants of these models in segmenting the six sequences into the dance labels displayed in Fig. 4.

**MNIW Prior — Unsupervised:** We start by testing the HDP-VAR(1)-HMM using a MNIW prior. (Note that we did not see performance gains by considering the HDP-SLDS, so we omit showing results for that architecture.) We set the prior distributions on the dynamic parameters and hyperparameters as in Sec. IV-A for the synthetic data examples, with the MNIW prior based on a pre-processed observation sequence. The pre-processing involves centering the position observations around 0 and scaling each component of $y_t$ to be within the same dynamic range. We compare our results to those of Xuan and
Fig. 4. Top: Trajectories of the dancing honey bees for sequences 1 to 6, colored by waggle (red), turn right (blue), and turn left (green) dances. Bottom: Sine of the bee’s head angle measurements colored by ground truth labels for 400 frames of each sequence. The data are available at http://www.cc.gatech.edu/borg/ijcv_psslds/.

Murphy [15], who used a change-point detection technique for inference on this dataset. As shown in Fig. 6(a)-(b), our model achieves a superior segmentation compared to the change-point formulation in almost all cases, while also identifying modes which reoccur over time. Example segmentations are shown in Fig. 5. Oh et. al. [10] also presented an analysis of the honey bee data, using an SLDS with a fixed number of modes. Unfortunately, that analysis is not directly comparable to ours, because Oh et. al. [10] used their SLDS in a supervised formulation in which the ground truth labels for all but one of the sequences are employed in the inference of the labels for the remaining held-out sequence, and in which the kernels used in the MCMC procedure depend on the ground truth labels. (The authors also considered a “parameterized segmental SLDS (PS-SLDS),” which makes use of domain knowledge specific to honey bee dancing and requires additional supervision during the learning process.) Nonetheless, in Table III we report the performance of these methods as well as the median performance (over 100 trials) of the unsupervised HDP-VAR(1)-HMM in order to provide a sense of the level of performance achievable without detailed, manual supervision. As seen in Table III, the HDP-VAR(1)-HMM yields very good performance on sequences 4 to 6 in terms of the learned segmentation and number of modes (see Fig. 5); the performance approaches that of the supervised method.

For sequences 1 to 3—which are much less regular than sequences 4 to 6—the performance of the unsupervised procedure is substantially worse. In Fig. 4, we see the extreme variation in head angle
Fig. 5. Estimated mode sequences at the 1000th Gibbs iteration representing the median error over 100 trials. For each sequence we plot the true labels (top), labels from the partially supervised HDP-VAR-HMM (middle), and unsupervised HDP-VAR-HMM (bottom). Colors are as in Fig. 4.

Fig. 6. (a)-(b) ROC curves for the unsupervised HDP-VAR-HMM, partially supervised HDP-VAR-HMM, and change-point formulation of [15] using the Viterbi sequence for segmenting datasets 1-3 and 4-6, respectively. (c)-(e) The 10th, 50th, and 90th Hamming distance quantiles over 100 trials are shown for sequences 4, 5, and 6, respectively.

during the waggle dances of sequences 1 to 3. As noted by Oh, the tracking results based on the vision-based tracker are noisier for these sequences and the patterns of switching between dance modes is more irregular. This dramatically affects our performance since we do not use domain-specific information. For sequence 2 in particular, our learned segmentations often create new, sequence-specific waggle dance
TABLE III
Median label accuracy of the HDP-VAR(1)-HMM using unsupervised and partially supervised Gibbs sampling, compared to accuracy of the supervised SLDS data-driven MCMC (DD-MCMC) MAP segmentation and PS-SLDS approximate Viterbi segmentation procedures of Oh et. al. [10].

<table>
<thead>
<tr>
<th>Sequence</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDP-VAR(1)-HMM unsupervised</td>
<td>45.0</td>
<td>42.7</td>
<td>47.3</td>
<td>88.1</td>
<td>92.5</td>
<td>88.2</td>
</tr>
<tr>
<td>HDP-VAR(1)-HMM partially supervised</td>
<td>55.0</td>
<td>86.3</td>
<td>81.7</td>
<td>89.0</td>
<td>92.4</td>
<td>89.6</td>
</tr>
<tr>
<td>SLDS DD-MCMC</td>
<td>74.0</td>
<td>86.1</td>
<td>81.3</td>
<td>93.4</td>
<td>90.4</td>
<td>90.2</td>
</tr>
<tr>
<td>PS-SLDS approx. Viterbi</td>
<td>75.9</td>
<td>92.4</td>
<td>83.1</td>
<td>93.4</td>
<td>91.0</td>
<td>90.4</td>
</tr>
</tbody>
</table>

modes contributing to our calculated Hamming distance errors on this sequence. Overall, however, we are able to achieve reasonably good segmentations without having to manually input domain-specific knowledge.

**MNIW Prior — Partially Supervised:** The discrepancy in performance between our results and the supervised approach of Oh et. al. [10] motivated us to also consider a partially supervised variant of the HDP-VAR(1)-HMM in which we fix the ground truth mode sequences for five out of six of the sequences, and jointly infer both a combined set of dynamic parameters and the left-out mode sequence. This is equivalent to informing the prior distributions with the data from the five fixed sequences, and using these updated posterior distributions as the prior distributions for the held-out sequence. As we see in Table III and the segmentations of Fig. 5, this partially supervised approach considerably improves performance for these three sequences, especially sequences 2 and 3. In this analysis, we hand-aligned sequences so that the waggle dances tended to have head angle measurements centered about $\pi/2$ radians. Aligning the waggle dances is possible by looking at the high frequency portions of the head angle measurements. Additionally, the pre-processing of the unsupervised approach is not appropriate here as the scalings and shiftings are dance-specific, and such transformations modify the associated switching VAR(1) model. Instead, to account for the varying frames of reference (i.e., point of origin for each bee body) we allowed for a mean $\mu(k)$ on the process noise, and placed an independent $\mathcal{N}(0, \Sigma_0)$ prior on this parameter. See the Appendix for details on how the hyperparameters of these prior distributions are set.

**ARD Prior:** Using the cleaner sequences 4 to 6, we investigate the affects of the sparsity-inducing ARD prior by assuming a higher order switching VAR model and computing the likelihood of the second half of each dance sequence based on parameters inferred from Gibbs sampling using the data from the first half of each sequence. In Fig. 7, we specifically compare the performance of an HDP-VAR($r$)-HMM with a conjugate MNIW prior for $r = 1, 2, 7$ to that of an HDP-VAR(7)-HMM with an ARD prior. We use the same approach to setting the hyperparameters as in Sec. IV-B. We see that assuming a higher
order model improves the predictive likelihood performance, but only when combined with a regularizing prior (e.g., the ARD) that avoids over-fitting in the presence of limited data. Although not depicted here (see instead [42]), the ARD prior also informs us of the variable-order nature of this switching dynamical process. When considering an HDP-VAR(2)-HMM with an ARD prior, the posterior distribution of the ARD hyperparameters for the first and second order lag components associated with each of the three dominant inferred dances clearly indicates that two of the dances (turning dances) simply rely on the first lag component while the other dance (waggle dance) relies on both lag components. To verify these results, we provided the data and ground truth labels to MATLAB’s \texttt{lpc} implementation of Levinson’s algorithm, which indicated that the turning dances are well approximated by an order 1 process, while the waggle dance relies on an order 2 model. Thus, our learned orders for the three dances match what is indicated by Levinson’s algorithm on ground-truth segmented data.

V. MODEL VARIANTS

There are many variants of the general SLDS and switching VAR models that are pervasive in the literature. One important example is when the dynamic matrix is shared between modes; here, the dynamics are instead distinguished based on a switching mean, such as the Markov switching stochastic volatility (MSSV) model. In the maneuvering target tracking community, it is often further assumed that the dynamic matrix is shared and known (due to the understood physics of the target). We explore both of these variants in the following sections.

A. Shared Dynamic Matrix, Switching Driving Noise

In many applications, the dynamics of the switching process can be described by a shared linear dynamical system matrix $A$; the dynamics within a given mode are then determined by some external...
force acting upon this LDS, and it is how this force is exerted that is mode-specific. The general form for such an SLDS is given by

\begin{equation}
\begin{align*}
 z_t | z_{t-1} & \sim \pi_{z_{t-1}} \\
 x_t &= A x_{t-1} + \epsilon_t^{(z_t)} \\
 y_t &= C x_t + w_t,
\end{align*}
\end{equation}

with process and measurement noise \( \epsilon_t^{(k)} \sim \mathcal{N}(\mu^{(k)}, \Sigma^{(k)}) \) and \( w_t \sim \mathcal{N}(0, R) \), respectively. In this scenario, the data are generated from one dynamic matrix, \( A \), and multiple process noise covariance matrices, \( \Sigma^{(k)} \). Thus, one cannot place a MNIW prior jointly on these parameters (conditioned on \( \mu^{(k)} \)) due to the coupling of the parameters in this prior. We instead consider independent priors on \( A, \Sigma^{(k)} \), and \( \mu^{(k)} \). We will refer to the choice of a normal prior on \( A \), inverse-Wishart prior on \( \Sigma^{(k)} \), and normal prior on \( \mu^{(k)} \) as the \( N-IW-N \) prior. See [42] for details on deriving the resulting posterior distributions given these independent priors.

**Stochastic Volatility:** An example of an SLDS in a similar form to that of Eq. (37) is the Markov switching stochastic volatility (MSSV) model [5], [6], [44]. The MSSV assumes that the log-volatilities follow an AR(1) process with a Markov switching mean. This underlying process is observed via conditionally independent and normally distributed daily returns. Specifically, let \( y_t \) represent, for example, the daily returns of a stock index. The state \( x_t \) is then given the interpretation of log-volatilities and the resulting state space model is given by [7]

\begin{equation}
\begin{align*}
 z_t | z_{t-1} & \sim \pi_{z_{t-1}} \\
 x_t &= a x_{t-1} + \epsilon_t^{(z_t)} \\
 y_t &= u_t(x_t),
\end{align*}
\end{equation}

with \( \epsilon_t^{(k)} \sim \mathcal{N}(\mu^{(k)}, \sigma^2) \) and \( u_t(x_t) \sim \mathcal{N}(0, \exp(x_t)) \). Here, only the mean of the process noise is mode-specific. Note, however, that the measurement equation is non-linear in the state \( x_t \). Carvalho and Lopes [7] employ a particle filtering approach to cope with these non-linearities. In [6], the MSSV is instead modeled in the log-squared-daily-returns domain such that

\begin{equation}
\log(y_t^2) = x_t + w_t,
\end{equation}

where \( w_t \) is additive, non-Gaussian noise. This noise is sometimes approximated by a moment-matched Gaussian [45], while So et al. [6] use a mixture of Gaussians approximation. The MSSV is then typically bestowed a fixed set of two or three regimes of volatility.

We examine the IBOVESPA stock index (Sao Paulo Stock Exchange) over the period of 01/03/1997 to 01/16/2001, during which ten key world events are cited in [7] as affecting the emerging Brazilian market. The key world events are summarized in Table IV and shown in the plots of Fig. 8. Use of this dataset was motivated by the work of Carvalho and Lopes [7], in which a two-mode MSSV model is
We consider a variant of the HDP-SLDS to match the MSSV model of Eq. (38). Specifically, we examine log-squared daily returns, as in Eq. (39), and use a DP mixture of Gaussians to model the measurement noise:

$$e_t^{(k)} \sim N(\mu^{(k)}, \Sigma^{(k)})$$

$$w_t \sim \sum_{\ell=1}^{\infty} \omega_\ell \mathcal{N}(0, R_\ell) \quad \omega \sim \text{GEM}(\sigma_\omega) \quad R_\ell \sim \text{IW}(n_\ell, S_\ell). \quad (40)$$

We truncate the measurement noise DP mixture to 10 components. For the HDP concentration hyperparameters, $\alpha$, $\gamma$, and $\kappa$, we use the same prior distributions as in Sec. IV-A-IV-C. For the dynamic parameters, we rely on the N-IW-N prior described in Sec. V-A and once again set the hyperparameters of this prior from statistics of the data as described in the Appendix. Since we allow for a mean on the process noise and examine log-squared daily returns, we do not preprocess the data.

The posterior probability of an HDP-SLDS inferred change point is shown in Fig. 8(a), and in Fig. 8(b) we display the corresponding plot for a non-sticky variant (i.e., with $\kappa = 0$ so that there is no bias towards mode self-transitions.) The HDP-SLDS is able to infer very similar change points to those presented in [7]. Without the sticky extension, the non-sticky model variant over-segments the data and rapidly switches between redundant states leading to many inferred change points that do not align with any world event.

As a quantitative comparison of the inferred change points, we compute a Hamming distance metric as follows. For the cited event dates, we form a “true” label sequence with labels that increment at each event. Then, for each inferred set of change points we form a separate label sequence in an analogous manner (i.e., with incrementing label numbers at each inferred change point.) We then compute the Hamming distance between the true and estimated label sequences after an optimal mapping between

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>07/02/1997</td>
<td>Thailand devalues the Baht by as much as 20%</td>
</tr>
<tr>
<td>08/11/1997</td>
<td>IMF and Thailand set a rescue agreement</td>
</tr>
<tr>
<td>10/23/1997</td>
<td>Hong Kong’s stock index falls 10.4%. South Korea won starts to weaken</td>
</tr>
<tr>
<td>12/02/1997</td>
<td>IMF and South Korea set a bailout agreement</td>
</tr>
<tr>
<td>06/01/1998</td>
<td>Russia’s stock market crashes</td>
</tr>
<tr>
<td>06/20/1998</td>
<td>IMF gives final approval to a loan package to Russia</td>
</tr>
<tr>
<td>08/19/1998</td>
<td>Russia officially falls into default</td>
</tr>
<tr>
<td>10/09/1998</td>
<td>IMF and World Bank joint meeting to discuss global economic crisis. The Fed cuts interest rates</td>
</tr>
<tr>
<td>01/15/1999</td>
<td>The Brazilian government allows its currency, the Real, to float freely by lifting exchange controls</td>
</tr>
<tr>
<td>02/02/1999</td>
<td>Arminio Fraga is named President of Brazil’s Central Bank</td>
</tr>
</tbody>
</table>

TABLE IV
TABLE OF 10 KEY WORLD EVENTS AFFECTING THE IBOVESPA STOCK INDEX (SAO PAULO STOCK EXCHANGE) OVER THE PERIOD OF 01/03/1997 TO 01/16/2001, AS CITED BY CARVALHO AND LOPES [7].

assumed.
Fig. 8. (a) Plot of the estimated probability of a change point on each day using 3,000 Gibbs samples for a MSSV variant of the HDP-SLDS using a shared dynamic matrix and allowing a mean on the mode-specific process noise and a mixture of Gaussian measurement noise model. The observations are log-squared daily return measurements, and the 10 key events are indicated with red lines. (b) Similar plot for the non-sticky HDP-SLDS with no bias towards self-transitions. (e)-(h) Analogous plots for the HDP-SLDS of Table I, a non-sticky variant of the HDP-SLDS, an HDP-AR(1)-HMM, and the switching AR(1) product partition model (PPM) of Xuan and Murphy [15], all using raw daily return measurements. The Viterbi change points provided by the formulation of [15] are shown with green triangles. (c) For each of the compared MSSV models, box plot of the normalized Hamming distance between a label sequence associated with the true event dates and that formed for each of the Gibbs sampled change points. (d) Analogous plot for the models analyzed in (e)-(h).

We also analyzed the performance of an HDP-SLDS as defined in Table I. We used raw daily-return observations, and first pre-processed the data in the same manner as the honey bee data by centering the observations around 0 and scaling the data to be roughly within a $[-10, 10]$ dynamic range. We then took a MNIW prior on the dynamic parameters, as outlined in the Appendix. Overall, although the state of this HDP-SLDS does not have the interpretation of log-volatilities, we see are still able to capture regime-changes in the dynamics of this stock index and find change points that align better with the true world events than in the MSSV HDP-SLDS model. See Fig. 8(d)-(h), which also provides a comparison with the change points inferred by an HDP-AR(1)-HMM and a switching AR(1) product partition model (PPM) of Xuan and Murphy [15]. The PPM inferred change points align well with those of the HDP-AR(1)-HMM—we expect this similar performance in such low-dimensional, long time series where the

---

4We do not compare to an HDP-AR(1)-HMM for the MSSV formulation since there is no adequate way to capture the complex MSSV observation model with an autoregressive process.
penalty incurred (in terms of quality of parameter estimates) by not revisiting modes is minimal.

B. Fixed Dynamic Matrix, Switching Driving Noise

There are some cases in which the dynamical model is well-defined through knowledge of the physics of the system being observed, such as simple kinematic motion. More complicated motions can typically be modeled using the same fixed dynamical model, but with a more complex description of the driving force. A generic LDS driven by an unknown control input $u_t$ can be represented as

$$
x_t = Ax_{t-1} + Bu_t + v_t \quad y_t = Cx_t + Du_t + w_t,
$$

where $v_t \sim N(0,Q)$ and $w_t \sim N(0,R)$. It is often appropriate to assume $D = 0$, as we do herein.

Maneuvering Target Tracking: Target tracking provides an application domain in which one often assumes that the dynamical model is known. One method of describing a maneuvering target is to consider the control input as a random process [46]. For example, a jump-mean Markov process [47] yields dynamics described as

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

$$
x_t = Ax_{t-1} + Bu_t(z_t) + v_t \quad y_t = Cx_t + w_t
$$

$$
u_t^{(k)} \sim N(\mu^{(k)}, \Sigma^{(k)}) \quad v_t \sim N(0,Q) \quad w_t \sim N(0,R).
$$

Classical approaches rely on defining a fixed set of dynamical modes and associated transition distributions. The state dynamics of Eq. (42) can be equivalently described as

$$
x_t = Ax_{t-1} + e_t^{(z_t)}
$$

$$
e_t^{(k)} \sim N(B\mu^{(k)}, B\Sigma^{(k)}B^T + Q).
$$

This model can be captured by our HDP-SLDS formulation of Eq. (37) with a fixed dynamic matrix (e.g., constant velocity or constant acceleration models [46]) and mode-specific, non-zero mean process noise. Such a formulation was explored in [9] along with experiments that compare the performance to that of standard multiple model techniques, demonstrating the flexibility of the Bayesian nonparametric approach. Fox et. al. [9] also present an alternative sampling scheme that harnesses the fact that the control input may be much lower-dimensional than the state and sequentially block-samples $(z_t, u_t)$ analytically marginalizing over the state sequence $x_{1:T}$. Note that this variant of the HDP-SLDS can be viewed as an extension of the work by Caron et. al. [20] in which the exogenous input is modeled as an independent noise process (i.e., no Markov structure on $z_t$) generated from a DP mixture model.
VI. CONCLUSION

In this paper, we have addressed the problem of learning switching linear dynamical models with an unknown number of modes for describing complex dynamical phenomena. We presented a Bayesian nonparametric approach and demonstrated both the utility and versatility of the developed HDP-SLDS and HDP-AR-HMM on real applications. Using the same parameter settings, although different model choices, in one case we are able to learn changes in the volatility of the IBOVESPA stock exchange while in another case we learn segmentations of data into waggle, turn-right, and turn-left honey bee dances. We also described a method of applying automatic relevance determination (ARD) as a sparsity-inducing prior, leading to flexible and scalable dynamical models that allow for identification of variable order structure. We concluded by considering adaptations of the HDP-SLDS to specific forms often examined in the literature such as the Markov switching stochastic volatility model and a standard multiple model target tracking formulation.

The batch processing of the Gibbs samplers derived herein may be impractical and offline-training online-tracking infeasible for certain applications. Due both to the nonlinear dynamics and uncertainty in model parameters, exact recursive estimation is infeasible. One could leverage the *conditionally linear* dynamics and use *Rao-Blackwellized particle filtering* (RBPF) [48]. However, one challenge is that such particle filters can suffer from a progressively impoverished particle representation. A possible direction of future research is to consider building on the recent work of [49] and embedding a RBPF within an MCMC algorithm. Another interesting avenue of research is to analyze high-dimensional time series. Although there is nothing fundamentally different in considering such datasets, based on experiments in related models [21] we expect to run into mixing rate issues with the Gibbs sampler since the parameter associated with each new considered dynamical mode is a sample from the (high-dimensional) prior. Developing split-merge algorithms similar to those developed in [50] for the DP mixture model could be useful in ameliorating these issues.

Overall, the formulation we developed herein represents a flexible, Bayesian nonparametric model for describing complex dynamical phenomena and discovering simple underlying temporal structures.

**APPENDIX**

*a) MNIW General Method:* For the experiments of Sec. IV-A, we set \( M = 0 \) and \( K = I_m \). This choice centers the mass of the prior around stable dynamic matrices while allowing for considerable variability. The inverse-Wishart portion is given \( n_0 = m + 2 \) degrees of freedom. For the HDP-AR-HMM, the scale matrix \( S_0 = 0.75 \Sigma \), where \( \Sigma = \frac{1}{T} \sum(y_t - \bar{y})(y_t - \bar{y})^T \). Setting the prior directly
from the data can help move the mass of the distribution to reasonable values of the parameter space. Since each new considered dynamical mode is associated with a set of parameters sampled from the prior distribution, and this dynamical mode is compared against others that have already been informed by the data, setting the base measure in this manner can improve mixing rates over a non-informative setting. For an HDP-SLDS with $x_t \in \mathbb{R}^n$ and $y_t \in \mathbb{R}^d$ and $n = d$, we set $S_0 = 0.675\bar{\Sigma}$. We then set the inverse-Wishart prior on the measurement noise, $R$, to have $r_0 = d + 2$ and $R_0 = 0.075\bar{\Sigma}$. For $n > d$, see [42].

b) Partially Supervised Honey Bee Experiments: For the partially supervised experiments of Sec. IV-C, we set $\Sigma_0 = 0.75S_0$. Since we are not shifting and scaling the observations, we set $S_0$ to 0.75 times the empirical covariance of the first difference observations. We also use $n_0 = 10$, making the distribution tighter than in the unsupervised case. Examining first differences is appropriate since the bee’s dynamics are better approximated as a random walk than as i.i.d. observations. Using raw observations in the unsupervised approach creates a larger expected covariance matrix making the prior on the dynamic matrix less informative, which is useful in the absence of other labeled data.

c) IBOVESPA Stock Index Experiments: For the HDP-SLDS variant of the MSSV model of Eq. (38), we rely on the N-IW-N prior described in Sec. V-A. For the dynamic parameter $a$ and process noise mean $\mu^{(k)}$, we use $\mathcal{N}(0, 0.75\bar{\Sigma})$ priors. The IW prior on $\Sigma^{(k)}$ was given 3 degrees of freedom and an expected value of $0.75\bar{\Sigma}$. Finally, each component of the mixture-of-Gaussian measurement noise was given an IW prior with 3 degrees of freedom and an expected value of $5 \pi^2$, which matches with the moment-matching technique of Harvey et. al. [45].

For the HDP-SLDS comparison using the model of Table I, we use a MNIW prior with $M = 0$, $K = 1$, $n_0 = 3$, and $S_0 = 0.75\bar{\Sigma}$. The IW prior on $R$ was given $r_0 = 100$ and an expected covariance of 25. Our sampler initializes parameters from the prior, and we found it useful to set the prior around large values of $R$ in order to avoid initial samples chattering between dynamical regimes caused by the state sequence having to account for the noise in the observations. After accounting for the residuals of the data in the posterior distribution, we typically learned $R \approx 10$.

REFERENCES


Emily B. Fox (S’06, M’09) received the S.B., M.Eng., E.E., and Ph.D. degrees from the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology in 2004, 2005, 2008, and 2009, respectively. She is currently a postdoctoral scholar in the Department of Statistical Science at Duke University. She is a recipient of the National Defense Science and Engineering Graduate (NDSEG) and National Science Foundation (NSF) Graduate Research fellowships, and currently holds an NSF Mathematical Sciences Postdoctoral Research Fellowship. She has also been awarded the 2009 Leonard J. Savage Thesis Award in Applied Methodology, the 2009 MIT EECS Jin-Au Kong Outstanding Doctoral Thesis Prize, the 2005 Chorafas Award for superior contributions in research, and the 2005 MIT EECS David Adler Memorial 2nd Place Master’s Thesis Prize. Her research interests are in Bayesian and Bayesian nonparametric approaches to multivariate time series analysis.

Erik B. Sudderth is an Assistant Professor in the Brown University Department of Computer Science. He received the Bachelor’s degree (summa cum laude) in Electrical Engineering from the University of California, San Diego, and the Master’s and Ph.D. in EECS from the Massachusetts Institute of Technology. His research interests include probabilistic graphical models; nonparametric Bayesian analysis; and applications of statistical machine learning in computer vision, signal processing, artificial intelligence, and geoscience. He was awarded a National Defense Science and Engineering Graduate Fellowship (1999), an Intel Foundation Doctoral Fellowship (2004), and in 2008 was named one of “AI’s 10 to Watch” by IEEE Intelligent Systems Magazine.

Michael I. Jordan is the Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley. His research in recent years has focused on Bayesian nonparametric analysis, probabilistic graphical models, spectral methods, kernel machines and applications to problems in signal processing, statistical genetics, computational biology, information retrieval and natural language processing. Prof. Jordan was named to the National Academy of Sciences and the National Academy of Engineering in 2010. He is a Fellow of the American Association for the Advancement of Science. He has been named a Neyman Lecturer and a Medallion Lecturer by Institute of Mathematical Statistics. He is a Fellow of the IMS, the IEE, the ACM, the AAAI and the ASA.

Alan S. Willsky (S’70, M’73, SM’82, F’86) joined M.I.T. in 1973 and is the Edwin Sibley Webster Professor of Electrical Engineering and Director of the Laboratory for Information and Decision Systems. He was a founder of Alphatech, Inc. and Chief Scientific Consultant, a role in which he continues at BAE Systems Advanced Information Technologies. From 1998-2002 he served on the US Air Force Scientific Advisory Board. He has received a number of awards including the 1975 American Automatic Control Council Donald P. Eckman Award, the 1979 ASCE Alfred Noble Prize, the 1980 IEEE Browder J. Thompson Memorial Award, the IEEE Control Systems Society Distinguished Member Award in 1988, the 2004 IEEE Donald G. Fink Prize Paper Award, Doctorat Honoris Causa from Université de Rennes in 2005, and the 2009 Technical Achievement Award from the IEEE Signal Processing Society. In 2010 Dr. Willsky was elected to the National Academy of Engineering. He and his students, have also received a variety of Best Paper Awards at various conferences and for papers in journals, including the 2001 IEEE Conference on Computer Vision and Pattern Recognition, the 2003 Spring Meeting of the American Geophysical Union, the 2004 Neural Information Processing Symposium, Fusion 2005, and the 2008 award from the journal Signal Processing for the outstanding paper in the year 2007. Dr. Willsky has delivered numerous keynote addresses and is co-author of the text Signals and Systems. His research interests are in the development and application of advanced methods of estimation, machine learning, and statistical signal and image processing.