4-2011

Bayesian Nonparametric Inference of Switching Linear Dynamical Systems

Emily B. Fox
University of Pennsylvania

Erik B. Sudderth

Michael I. Jordan

Alan Willsky

Recommended Citation

Fox, E. B., Sudderth, E. B., Jordan, M. I., & Willsky, A. (2011). Bayesian Nonparametric Inference of Switching Linear Dynamical Systems. *IEEE Transactions on Signal Processing, 59* (4), 1569-1585. http://dx.doi.org/10.1109/TSP.2010.2102756

Bayesian Nonparametric Inference of Switching Linear Dynamical Systems

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 lineardynamical system (SLDS) and the switching vector autoregressive (VAR) process. Our Bayesiannonparametric 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 orswitching 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.

Keywords

Bayes methods, autoregressive processes, inference mechanisms, linear systems, nonparametric statistics, sampling methods, target tracking, time-varying systems, Bayesian nonparametric inference, IBOVESPA stock index, complex dynamical phenomena, conditionally linear dynamical mode, dancing honey bee, hierarchical Dirichlet process, state sequence, switching dynamic linear model, target tracking sampling algorithm, vector autoregressive process, autoregressive processes, Bayesian methods, hidden Markov models, state-space methods, time series analysis, unsupervised learning

Disciplines

Computer Sciences | Statistics and Probability

Bayesian Nonparametric Inference of Switching Linear Dynamical Systems

Emily Fox, Erik Sudderth, Michael Jordan, and Alan Willsky

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.

Index Terms

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

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. 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. Most existing methods for learning SLDS and switching VAR processes rely on either fixing the number of HMM modes, such as in [10], or considering a change-point detection formulation where each inferred change is to a new, previously unseen dynamical mode, such as in [11]. In this paper we show how one can remain agnostic about the number of dynamical modes while still allowing for returns to previously exhibited dynamical behaviors.

Hierarchical Dirichlet processes (HDP) can be used as a prior on the parameters of HMMs with unknown mode space cardinality [12], [13]. In this paper we use a variant of the HDP-HMM—the *sticky HDP-HMM* of [14]—that provides 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 learn 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) [15]–[17]. The

E. Fox is with the Department of Statistical Science, Duke University, Durham, NC, 27708 USA e-mail: fox@stat.duke.edu. E. Sudderth is with the Department of Computer Science, Brown University, Providence, RI, 02912 USA e-mail: sudderth@cs.brown.edu. M. Jordan is with the Department of Electrical Engineering and Computer Science, and Department of Statistics, University of California, Berkeley, CA, 94720 USA e-mail: jordan@eecs.berkeley.edu. A. Willsky is with the Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 02139 USA e-mail: willsky@mit.edu. This work was supported in part by MURIs funded through AFOSR Grant FA9550-06-1-0324 and ARO Grant W911NF-06-1-0076. Preliminary versions (without detailed development or analysis) of this work have been presented at two conferences [1], [2].

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. [18] 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. [19] present an exact algebraic approach, though relying on fixing a maximal mode space cardinality and autoregressive order. Psaradakis and Spagnolog [20] 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. [21] 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 [22]. Kotsalis et. al. [23] 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[1]. In [25], a realization theory is presented for *generalized jump-Markov linear systems* (GJMLS) 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 [26] present a variational approach to segmenting the data into the linear dynamical regimes and learning the associated dynamic parameters[2]. For questions of observability and identifiability of SLDS in the absence of noise, see [27].

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 [27]. 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) [28] or the *Bayesian information criterion* (BIC) [29], 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, $\boldsymbol{x}_t \in \mathbb{R}^n$, with dynamics observed via $\boldsymbol{y}_t \in \mathbb{R}^d$. A linear time-invariant (LTI) SS model is given by

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

where $\boldsymbol{e}_t$ and $\boldsymbol{w}_t$ are independent Gaussian noise processes with covariances Σ and R, respectively.

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

$$\boldsymbol{y}_t = \sum_{i=1}^{r} A_i \boldsymbol{y}_{t-i} + \boldsymbol{e}_t \qquad \boldsymbol{e}_t \sim \mathcal{N}(0, \Sigma). \tag{2}$$

[1]The problem of identification of HMMs is thoroughly analyzed in [24].

[2]This formulation uses a *mixture of experts* SLDS in which M different continuous-valued state sequences evolve independently with linear dynamics and the Markovian dynamical mode selects which state sequence is observed at a given time.

Every VAR(r) process can be described in SS form, though not every SS model may be expressed as a VAR(r) process for finite r [30].

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

$$\begin{aligned} z_t \mid z_{t-1} &\sim \pi_{z_{t-1}} \\ \boldsymbol{x}_t = A^{(z_t)}\boldsymbol{x}_{t-1} + \boldsymbol{e}_t(z_t) \qquad \boldsymbol{y}_t &= C\boldsymbol{x}_t + \boldsymbol{w}_t. \end{aligned} \tag{3}$$

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 $\boldsymbol{e}_t(z_t) \sim \mathcal{N}(0, \Sigma^{(z_t)})$. One can view the SLDS as an extension of the classical hidden Markov model (HMM) [31], which has the same mode evolution, but conditionally *independent* observations:

$$\begin{aligned} z_t \mid z_{t-1} &\sim \pi_{z_{t-1}} \\ y_t \mid z_t &\sim F(\theta_{z_t}) \end{aligned} \tag{4}$$

for an indexed family of distributions $F(\cdot)$ where θ_i are the *emission parameters* for mode i.

We similarly define a *switching* VAR(r) process by

$$\begin{aligned} z_t \mid z_{t-1} &\sim \pi_{z_{t-1}} \\ \boldsymbol{y}_t &= \sum_{i=1}^{r} A_i^{(z_t)} \boldsymbol{y}_{t-i} + \boldsymbol{e}_t(z_t). \end{aligned} \tag{5}$$

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 δ_θ is a mass concentrated at θ. We then operate directly in the parameter space Θ and transition between emission parameters with probabilities given by $\{G_j\}$. That is,

$$\begin{aligned} \theta'_t \mid \theta'_{t-1} &\sim G_{j:\theta'_{t-1}=\theta_j} \\ y_t \mid \theta'_t &\sim F(\theta'_t). \end{aligned} \tag{6}$$

Here, $\theta'_t \in \{\theta_1, \ldots, \theta_K\}$ and is equivalent to θ_{z_t} of Eq. (4). A Bayesian nonparametric HMM takes G_j to be *random*[3] with an infinite collection of atoms corresponding to the infinite HMM mode space.

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

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad \theta_k \sim H, \tag{7}$$

on a parameter space Θ. The weights are sampled via a *stick-breaking construction* [32]:

$$\beta_k = \nu_k \prod_{\ell=1}^{k-1} (1 - \nu_\ell) \qquad \nu_k \sim \mathrm{Beta}(1, \gamma). \tag{8}$$

In effect, we have divided a unit-length stick into lengths given by the weights β_k: the k^{th} weight is a random proportion ν_k of the remaining stick after the previous $(k-1)$ weights have been defined. We denote this distribution by $\beta \sim \mathrm{GEM}(\gamma)$.

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

[3]Formally, a random measure on a measurable space Θ 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 $A \in \mathcal{A}$.

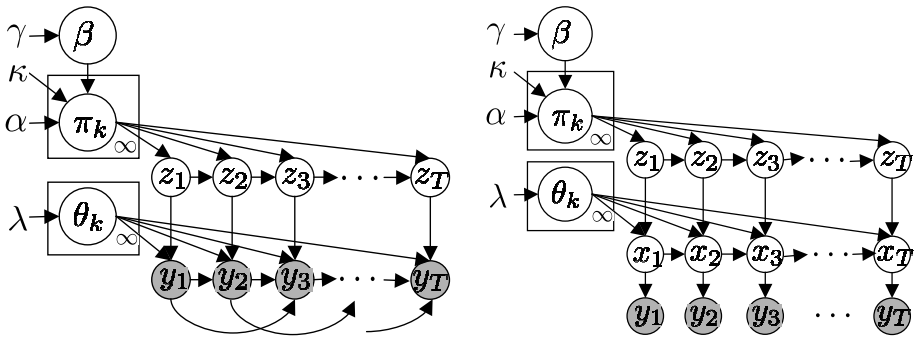


Fig. 1. Sticky HDP-HMM prior on (a) switching VAR(2) and (b) SLDS processes with the mode evolving as $z_{t+1}|\{\pi_k\}_{k=1}^{\infty}, z_t \sim \pi_{z_t}$ for $\pi_k|\alpha,\kappa,\beta \sim \mathrm{DP}(\alpha+\kappa,(\alpha\beta+\kappa\delta_k)/(\alpha+\kappa))$. Here, $\beta \mid \gamma \sim \mathrm{GEM}(\gamma)$ and $\theta_k \mid H,\lambda \sim H(\lambda)$. The dynamical processes are as in Table I.

that picks out the unique value θ_k such that $\theta'_i = \theta_{z_i}$. Blackwell and MacQueen [33] introduced a Pólya urn representation of the θ'_i:

$$\theta'_i \mid \theta'_1,\ldots,\theta'_{i-1} \sim \frac{\gamma}{\gamma+i-1}H + \sum_{j=1}^{i-1}\frac{1}{\gamma+i-1}\delta_{\theta'_j} = \frac{\gamma}{\gamma+i-1}H + \sum_{k=1}^{K}\frac{n_k}{\gamma+i-1}\delta_{\theta_k}. \tag{9}$$

Here, n_k is the number of observations θ'_i taking the value θ_k. From Eq. (9), and the discrete nature of G_0, we see a reinforcement property of the Dirichlet process that induces sparsity in the number of inferred mixture components.

A hierarchical extension of the Dirichlet process, the hierarchical Dirichlet process (HDP) [12], 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 $G_j \sim \mathrm{DP}(\alpha,G_0)$, with G_0 itself a draw from a Dirichlet process $\mathrm{DP}(\gamma,H)$. Through this construction, one can show that the probability measures are described as

$$\begin{array}{ll} G_0 = \sum_{k=1}^{\infty}\beta_k\delta_{\theta_k} & \beta \mid \gamma \sim \mathrm{GEM}(\gamma) \\ G_j = \sum_{k=1}^{\infty}\pi_{jk}\delta_{\theta_k} & \pi_j \mid \alpha,\beta \sim \mathrm{DP}(\alpha,\beta) \end{array} \qquad \theta_k \mid H \sim H. \tag{10}$$

Applying the HDP prior to the HMM, we obtain the *HDP-HMM* of Teh et. al. [12]. This corresponds to the model in Fig. 1(a), but without the edges between the observations.

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

$$\mathbb{E}[\pi_{jk} \mid \beta] = \beta_k. \tag{11}$$

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 [14] that one may mitigate this problem by instead considering a *sticky* HDP-HMM where π_j is distributed as follows:

$$\pi_j \mid \beta,\alpha,\kappa \sim \mathrm{DP}\left(\alpha+\kappa, \frac{\alpha\beta+\kappa\delta_j}{\alpha+\kappa}\right). \tag{12}$$

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 κ. Specifically, the expected set of weights for transition distribution π_j is a convex combination of those defined by β and mode-specific weight defined by κ:

$$\mathbb{E}[\pi_{jk} \mid \beta,\alpha,\kappa] = \frac{\alpha}{\alpha+\kappa}\beta_k + \frac{\kappa}{\alpha+\kappa}\delta(j,k). \tag{13}$$

When $\kappa = 0$ the original HDP-HMM of Teh et. al. [12] is recovered. We place a prior on κ and learn the self-transition bias from the data.

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

TABLE I

DYNAMIC EQUATIONS FOR THE HDP-AR-HMM AND HDP-SLDS. HERE, π_j IS AS DEFINED IN EQ. (12) FOR THE STICKY HDP-HMM. THE ADDITIVE NOISE PROCESSES ARE DISTRIBUTED AS $\boldsymbol{e}_t(k) \sim \mathcal{N}(0, \Sigma^{(k)})$ AND $\boldsymbol{w}_t \sim \mathcal{N}(0, R)$.

	HDP-AR-HMM	HDP-SLDS
Dynamic matrix	$\mathbf{A}^{(k)} = [A_1^{(k)} \dots A_r^{(k)}] \in \mathbb{R}^{d \times (d*r)}$	$\mathbf{A}^{(k)} = A^{(k)} \in \mathbb{R}^{n \times n}$
Pseudo-observations	$\boldsymbol{\psi_t} = \boldsymbol{y}_t$	$\boldsymbol{\psi_t} = \boldsymbol{x}_t$
Lag pseudo-observations	$\bar{\boldsymbol{\psi}}_t = [\boldsymbol{y}_{t-1}^T \dots \boldsymbol{y}_{t-r}^T]^T$	$\bar{\boldsymbol{\psi}}_t = \boldsymbol{x}_{t-1}$.

TABLE II

NOTATIONAL CONVENIENCES USED IN DESCRIBING THE GIBBS SAMPLER FOR THE HDP-AR-HMM AND HDP-SLDS.

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 switching among unknown number of unknown dynamics using Bayesian nonparametric methods to capture these uncertainties (and to allow both learning the number of modes and estimating system state). 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.

For the HDP-SLDS, we place priors on the *dynamic parameters* $\{A^{(k)}, \Sigma^{(k)}\}$ and on measurement noise R and infer their posterior from the data. However, without loss of generality[4], we fix the measurement matrix to $C = [I_d \ 0]$ 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 $\{A_1^{(k)}, \dots, A_r^{(k)}, \Sigma^{(k)}\}$.

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

1) Sample the state sequence $\boldsymbol{x}_{1:T}$ 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 $\boldsymbol{x}_{1:T}$, 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, $z_{1:T}$, $\boldsymbol{x}_{1:T}$, and $\boldsymbol{y}_{1:T}$.

For the HDP-AR-HMM, step (1) does not exist. Step (2) then involves sampling the mode sequence $z_{1:T}$ given the observations $\boldsymbol{y}_{1:T}$ (rather than $\boldsymbol{x}_{1:T}$), and step (3) involves conditioning solely on the sequences $z_{1:T}$ and $\boldsymbol{y}_{1:T}$ (not $\boldsymbol{x}_{1:T}$). Also, we note that step (2) involves a fairly straightforward extension of the sampling method developed in [14] for the simpler HDP-HMM model; the other steps, however, involve new constructs, as they involve capturing and dealing with the temporal dynamics of the underlying continuous state models. Sec. III-A provides the necessary priors and 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* $\boldsymbol{\psi}_{1:T}$ and rewrite the dynamic equation for both the HDP-SLDS and HDP-AR-HMM generically as

$$\boldsymbol{\psi}_t = \mathbf{A}^{(k)} \bar{\boldsymbol{\psi}}_{t-1} + \boldsymbol{e_t}, \tag{14}$$

where we utilize the definitions outlined in Table II.

[4]This is, in essence, an issue of choosing a similarity transformation for the state of a minimal system, exploiting the fact that the measurement matrix is shared by all modes of the HDP-SLDS so that the same transformation can be used for all modes.

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 $\mathbf{A}^{(k)}$ and forming a *lag observation vector* $\bar{\boldsymbol{\psi}}_t$ comprised of a series of previous observation vectors. For this section (for the HDP-SLDS), we assume such a sample of the state sequence $\boldsymbol{x}_{1:T}$ (and hence $\{\boldsymbol{\psi}_t, \bar{\boldsymbol{\psi}}_t\}$) 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 $\boldsymbol{\Psi}^{(k)}$ with n_k columns consisting of the $\boldsymbol{\psi}_t$ with $z_t = k$. Then,

$$\boldsymbol{\Psi}^{(k)} = \mathbf{A}^{(k)}\bar{\boldsymbol{\Psi}}^{(k)} + \mathbf{E}^{(k)}, \tag{15}$$

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

1) Conjugate Prior on $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$: The *matrix-normal inverse-Wishart* (MNIW) prior [34] is conjugate to the likelihood model defined in Eq. (15) for the parameter set $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$. 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 although 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 [35], and conditioned on data from a stable mode, the posterior distribution will likely be sharply peaked around stable dynamic matrices.

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

$$p(\mathbf{A}^{(k)}, \Sigma^{(k)} \mid \mathbf{D}^{(k)}) = p(\mathbf{A}^{(k)} \mid \Sigma^{(k)}, \mathbf{D}^{(k)}) p(\Sigma^{(k)} \mid \mathbf{D}^{(k)}). \tag{16}$$

The resulting posterior of $\mathbf{A}^{(k)}$ is straightforwardly derived to be (see [36])

$$p(\mathbf{A}^{(k)} \mid \Sigma^{(k)}, \mathbf{D}^{(k)}) = \mathcal{MN}\left(\mathbf{A}^{(k)}; \mathbf{S}^{(k)}_{\psi\bar{\psi}}\mathbf{S}^{-(k)}_{\bar{\psi}\bar{\psi}}, \Sigma^{(k)}, \mathbf{S}^{(k)}_{\bar{\psi}\bar{\psi}}\right), \tag{17}$$

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

$$\mathbf{S}^{(k)}_{\bar{\psi}\bar{\psi}} = \bar{\boldsymbol{\Psi}}^{(k)}\bar{\boldsymbol{\Psi}}^{(k)^T} + K \qquad \mathbf{S}^{(k)}_{\psi\bar{\psi}} = \boldsymbol{\Psi}^{(k)}\bar{\boldsymbol{\Psi}}^{(k)^T} + MK \qquad \mathbf{S}^{(k)}_{\psi\psi} = \boldsymbol{\Psi}^{(k)}\boldsymbol{\Psi}^{(k)^T} + MKM^T. \tag{18}$$

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

$$p(\Sigma^{(k)} \mid \mathbf{D}^{(k)}) = \mathrm{IW}\left(n_k + n_0, \mathbf{S}^{(k)}_{\psi|\bar{\psi}} + S_0\right), \tag{19}$$

where $\mathrm{IW}(n_0, S_0)$ denotes an inverse-Wishart prior with n_0 degrees of freedom and scale matrix S_0, and is updated by data terms $\mathbf{S}^{(k)}_{\psi|\bar{\psi}} = \mathbf{S}^{(k)}_{\psi\psi} - \mathbf{S}^{(k)}_{\psi\bar{\psi}}\mathbf{S}^{-(k)}_{\bar{\psi}\bar{\psi}}\mathbf{S}^{(k)^T}_{\psi\bar{\psi}}$ 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) [15]–[17], 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 $\mathbf{A}^{(k)}$:

$$p(\mathbf{A}^{(k)}|\alpha^{(k)}) = \prod_{j=1}^{n} \mathcal{N}\left(\boldsymbol{a}_j^{(k)}; 0, \alpha_j^{-(k)} I_n\right). \tag{20}$$

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 $\alpha_j^{(k)} \to \infty$ drives $\boldsymbol{a}_j^{(k)} \to 0$, 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 $\boldsymbol{a}_j^{(k)} = 0$ implies that the realization of *that mode* is not minimal since the associated Hankel matrix

$$\mathcal{H} = \begin{bmatrix} C^T & CA^T & \cdots & (CA^{d-1})^T \end{bmatrix}^T \begin{bmatrix} G & AG & \cdots & A^{d-1}G \end{bmatrix} \equiv \mathcal{O}\mathcal{R} \tag{21}$$

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:

Criterion 3.1: If for some realization $\mathcal{R}$ a mode k has $\boldsymbol{a}_j^{(k)} = 0$, then that realization must have $\boldsymbol{c}_j = 0$, where $\boldsymbol{c}_j$ 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[5].

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 $\mathbf{A}^{(k)}$ as we did in Eq. (20), we decompose the prior over the lag blocks $A_i^{(k)}$:

$$p(\mathbf{A}^{(k)}|\boldsymbol{\alpha}^{(k)}) = \prod_{i=1}^{r} \mathcal{N}\left(\text{vec}(A_i^{(k)}); 0, \alpha_i^{-(k)} I_{d^2}\right). \tag{22}$$

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

$$\Sigma_0^{(k)} = \text{diag}\left(\alpha_1^{(k)}, \ldots, \alpha_1^{(k)}, \ldots, \alpha_m^{(k)}, \ldots, \alpha_m^{(k)}\right)^{-1}. \tag{23}$$

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 $\boldsymbol{x}_t$, r the autoregressive order of the HDP-AR-HMM, and d the dimension of the observations $\boldsymbol{y}_t$.) To examine the posterior distribution of $\mathbf{A}^{(k)}$, we note that we may rewrite the state equation as,

$$\begin{aligned} \boldsymbol{\psi}_{t+1} &= \begin{bmatrix} \bar{\boldsymbol{\psi}}_{t,1} I_\ell & \bar{\boldsymbol{\psi}}_{t,2} I_\ell & \cdots & \bar{\boldsymbol{\psi}}_{t,\ell * r} I_\ell \end{bmatrix} \text{vec}(\mathbf{A}^{(k)}) + \boldsymbol{e}_{t+1}(k) \qquad \forall t | z_t = k \\ &\triangleq \tilde{\Psi}_t \text{vec}(\mathbf{A}^{(k)}) + \boldsymbol{e}_{t+1}(k), \end{aligned} \tag{24}$$

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

$$p(\text{vec}(\mathbf{A}^{(k)}) \mid \mathbf{D}^{(k)}, \Sigma^{(k)}, \boldsymbol{\alpha}^{(k)}) = \mathcal{N}^{-1}\left(\sum_{t|z_t=k} \tilde{\Psi}_{t-1}^T \Sigma^{-(k)} \boldsymbol{\psi_t}, \Sigma_0^{-(k)} + \sum_{t|z_t=k} \tilde{\Psi}_{t-1}^T \Sigma^{-(k)} \tilde{\Psi}_{t-1} \right). \tag{25}$$

See [36] 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 $\mathbf{A}^{(k)}$, and recalling that each precision parameter is gamma distributed, the posterior of $\alpha_\ell^{(k)}$ is given by

$$p(\alpha_\ell^{(k)} \mid \mathbf{A}^{(k)}) = \text{Gamma}\left(a + \frac{|\mathcal{S}_\ell|}{2}, b + \frac{\sum_{(i,j)\in\mathcal{S}_\ell} a_{ij}^{(k)^2}}{2} \right). \tag{26}$$

The set $\mathcal{S}_\ell$ contains the indices for which $a_{ij}^{(k)}$ has prior precision $\alpha_\ell^{(k)}$. Note that in this model, regardless of the number of observations $\boldsymbol{y}_t$, the size of $\mathcal{S}_\ell$ (i.e., the number of $a_{ij}^{(k)}$ used to inform the posterior distribution)

[5]If there does not exist a realization $\mathcal{R}$ satisfying Criterion 3.1, we may instead consider a more general model where the measurement equation is mode-specific and we place a prior on $C^{(k)}$ instead of fixing this matrix. However, this model leads to identifiability issues that are considerably less pronounced in the above case.

remains the same. Thus, the gamma prior is an informative prior and the choice of a and b should depend upon the cardinality of $\mathcal{S}_\ell$. 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 $\mathrm{IW}(n_0, S_0)$ on $\Sigma^{(k)}$ and look at the posterior given $\mathbf{A}^{(k)}$:

$$p(\Sigma^{(k)} \mid \mathbf{D}^{(k)}, \mathbf{A}^{(k)}) = \mathrm{IW}\left(n_k + n_0, \mathbf{S}^{(k)}_{\psi|\bar{\psi}} + S_0\right), \tag{27}$$

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

$$\mathbf{S}^{(k)}_{\psi|\bar{\psi}} = \sum_{t|z_t=k} (\boldsymbol{\psi}_t - \mathbf{A}^{(k)}\bar{\boldsymbol{\psi}}_{t-1})(\boldsymbol{\psi}_t - \mathbf{A}^{(k)}\bar{\boldsymbol{\psi}}_{t-1})^T. \tag{28}$$

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

$$p(R \mid \boldsymbol{y}_{1:T}, \boldsymbol{x}_{1:T}) = \mathrm{IW}(T + r_0, S_R + R_0), \tag{29}$$

where $S_R = \sum_{t=1}^{T} (\boldsymbol{y}_t - C\boldsymbol{x}_t)(\boldsymbol{y}_t - C\boldsymbol{x}_t)^T$. 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, $\boldsymbol{x}_{1:T}$, 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 $\boldsymbol{x}_{1:T}$ in a similar vein to that of Carter and Kohn [37]. We describe the sampler in terms of the pseudo-observations $\boldsymbol{\psi}_t$, 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 $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$: Conditioned on the mode sequence, $z_{1:T}$, and the pseudo-observations, $\boldsymbol{\psi}_{1:T}$, we can sample the dynamic parameters $\boldsymbol{\theta} = \{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$ from the posterior densities of Sec. III-A. For the ARD prior, we then sample $\boldsymbol{\alpha}^{(k)}$ given $\mathbf{A}^{(k)}$. In practice we iterate multiple times between sampling $\boldsymbol{\alpha}^{(k)}$ given $\mathbf{A}^{(k)}$ and $\mathbf{A}^{(k)}$ given $\boldsymbol{\alpha}^{(k)}$ 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 $\boldsymbol{x}_{1:T}$.

3) Block Sampling $z_{1:T}$: As shown in [14], 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 $m_{t+1,t}(z_t) \propto p(\boldsymbol{\psi}_{t+1:T}|z_t, \bar{\boldsymbol{\psi}}_t, \boldsymbol{\pi}, \boldsymbol{\theta})$ 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

$$p(z_t \mid z_{t-1}, \boldsymbol{\psi}_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto p(z_t \mid \pi_{z_{t-1}}) p(\boldsymbol{\psi}_t \mid \bar{\boldsymbol{\psi}}_{t-1}, \mathbf{A}^{(z_t)}, \Sigma^{(z_t)}) m_{t+1,t}(z_t). \tag{30}$$

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

$$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) \propto p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, A^{(z_t)}, \Sigma^{(z_t)}) p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, R) m_{t+1,t}(\boldsymbol{x}_t). \tag{31}$$

The messages are given in information form by $m_{t,t-1}(\boldsymbol{x}_{t-1}) \propto \mathcal{N}^{-1}(\boldsymbol{x}_{t-1}; \vartheta_{t,t-1}, \Lambda_{t,t-1})$, where the information parameters are recursively defined as

$$\begin{aligned}\vartheta_{t,t-1} &= A^{(z_t)^T}\Sigma^{-(z_t)}\tilde{\Lambda}_t(C^T R^{-1}\boldsymbol{y}_t + \vartheta_{t+1,t}) \\ \Lambda_{t,t-1} &= A^{(z_t)^T}\Sigma^{-(z_t)}A^{(z_t)} - A^{(z_t)^T}\Sigma^{-(z_t)}\tilde{\Lambda}_t\Sigma^{-(z_t)}A^{(z_t)},\end{aligned} \tag{32}$$

with $\tilde{\Lambda}_t = (\Sigma^{-(z_t)} + C^T R^{-1} C + \Lambda_{t+1,t})^{-1}$. The standard $\vartheta^b_{t|t}$ and $\Lambda^b_{t|t}$ updated information parameters for a backward running Kalman filter are given by

$$\begin{aligned}\Lambda^b_{t|t} &= C^T R^{-1} C + \Lambda_{t+1,t} \\ \vartheta^b_{t|t} &= C^T R^{-1} y_t + \vartheta_{t+1,t}.\end{aligned} \tag{33}$$

See [36] 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 $p(z_t \mid z_{\setminus t}, \boldsymbol{y}_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta})$. 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 *other* time steps, we can run a forward Kalman filter to marginalize the state sequence $\boldsymbol{x}_{1:t-2}$ producing $p(\boldsymbol{x}_{t-1} \mid y_{1:t-1}, z_{1:t-1}, \boldsymbol{\theta})$, and a backward filter to marginalize $\boldsymbol{x}_{t+1:T}$ producing $p(\boldsymbol{y}_{t+1:T} \mid x_t, z_{t+1:T}, \boldsymbol{\theta})$. Then, for each possible value of z_t, we combine these forward and backward messages with the local likelihood $p(\boldsymbol{y}_t \mid \boldsymbol{x}_t)$ and local dynamic $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}, z_t = k)$ and marginalize over $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ resulting in the likelihood of the observation sequence $\boldsymbol{y}_{1:T}$ as a function of $\boldsymbol{z}_t$. 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:

$$\begin{aligned}p(z_t = k \mid z_{\setminus t}, \boldsymbol{y}_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) &\propto \pi_{z_{t-1}}(k)\pi_k(z_{t+1}) \\ &\frac{|\Lambda^{(k)}_t|^{1/2}}{|\Lambda^{(k)}_t + \Lambda^b_{t|t}|^{1/2}} \exp\left(-\frac{1}{2}\vartheta^{(k)^T}_t \Lambda^{-(k)}_t \vartheta^{(k)}_t + \frac{1}{2}(\vartheta^{(k)}_t + \vartheta^b_{t|t})^T(\Lambda^{(k)}_t + \Lambda^b_{t|t})^{-1}(\vartheta^{(k)}_t + \vartheta^b_{t|t})\right)\end{aligned} \tag{34}$$

with

$$\begin{aligned}\Lambda^{(k)}_t &= (\Sigma^{(k)} + \mathbf{A}^{(z_t)}\Lambda^{-f}_{t-1|t-1}\mathbf{A}^{(z_t)^T})^{-1} \\ \vartheta^{(k)}_t &= (\Sigma^{(k)} + \mathbf{A}^{(z_t)}\Lambda^{-f}_{t-1|t-1}\mathbf{A}^{(z_t)^T})^{-1}\mathbf{A}^{(z_t)}\Lambda^{-f}_{t-1|t-1}\vartheta^f_{t-1|t-1}.\end{aligned} \tag{35}$$

See [36] for full derivations. Here, $\vartheta^f_{t|t}$ and $\Lambda^f_{t|t}$ are the updated information parameters for a forward running Kalman filter, defined recursively as

$$\begin{aligned}\Lambda^f_{t|t} &= C^T R^{-1} C + \Sigma^{-(z_t)} - \Sigma^{-(z_t)}\mathbf{A}^{(z_t)}(\mathbf{A}^{(z_t)^T}\Sigma^{-(z_t)}\mathbf{A}^{(z_t)} + \Lambda^f_{t-1|t-1})^{-1}\mathbf{A}^{(z_t)^T}\Sigma^{-(z_t)} \\ \vartheta^f_{t|t} &= C^T R^{-1} y_t + \Sigma^{-(z_t)}\mathbf{A}^{(z_t)}(\mathbf{A}^{(z_t)^T}\Sigma^{-(z_t)}\mathbf{A}^{(z_t)} + \Lambda^f_{t-1|t-1})^{-1}\vartheta^f_{t-1|t-1}.\end{aligned} \tag{36}$$

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[6], HDP-VAR(2)-HMM, and HDP-SLDS using the MNIW prior. We compare to a baseline sticky HDP-HMM using first difference observations, imitating a HDP-VAR(1)-HMM with

[6]We use the notation HDP-VAR(r)-HMM to refer to an order r HDP-AR-HMM with vector observations.

Given a previous set of mode-specific transition probabilities $\boldsymbol{\pi}^{(n-1)}$, the global transition distribution $\beta^{(n-1)}$, and dynamic parameters $\boldsymbol{\theta}^{(n-1)}$:

1) Set $\boldsymbol{\pi} = \boldsymbol{\pi}^{(n-1)}$, $\boldsymbol{\beta} = \boldsymbol{\beta}^{(n-1)}$, and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(n-1)}$.
2) If HDP-SLDS,
   a) For each $t \in \{1, \ldots, T\}$, compute $\{\vartheta^f_{t|t}, \Lambda^f_{t|t}\}$ as in Eq. (36).
   b) For each $t \in \{T, \ldots, 1\}$,
      i) Compute $\{\vartheta^b_{t|t}, \Lambda^b_{t|t}\}$ as in Eq. (33).
      ii) For each $k \in \{1, \ldots, L\}$, compute $\{\vartheta^{(k)}_t, \Lambda^{(k)}_t\}$ as in Eq. (35) and set
      $$f_k(\boldsymbol{y}_{1:T}) = |\Lambda^{(k)}_t|^{1/2}|\Lambda^{(k)}_t + \Lambda^b_{t|t}|^{-1/2} \exp\left(-\frac{1}{2}\vartheta^{(k)T}_t \Lambda^{-(k)}_t \vartheta^{(k)}_t + \frac{1}{2}(\vartheta^{(k)}_t + \vartheta^b_{t|t})^T(\Lambda^{(k)}_t + \Lambda^b_{t|t})^{-1}(\vartheta^{(k)}_t + \vartheta^b_{t|t})\right).$$
      iii) Sample a mode assignment
      $$z_t \sim \sum_{k=1}^{L} \pi_{z_{t-1}}(k)\pi_k(z_{t+1}) f_k(\boldsymbol{y}_{1:T})\delta(z_t, k).$$
   c) Working sequentially forward in time sample
   $$\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t; (\Sigma^{-(z_t)} + \Lambda^b_{t|t})^{-1}(\Sigma^{-(z_t)} A^{(z_t)} \boldsymbol{x}_{t-1} + \vartheta^b_{t|t}), (\Sigma^{-(z_t)} + \Lambda^b_{t|t})^{-1}).$$
   d) Set pseudo-observations $\boldsymbol{\psi}_{1:T} = \boldsymbol{x}_{1:T}$.
3) If HDP-AR-HMM, set pseudo-observations $\boldsymbol{\psi}_{1:T} = \boldsymbol{y}_{1:T}$.
4) Block sample $z_{1:T}$ given transition distributions $\boldsymbol{\pi}$, dynamic parameters $\boldsymbol{\theta}$, and pseudo-observations $\boldsymbol{\psi}_{1:T}$ as in Algorithm 2.
5) Update the global transition distribution β (utilizing auxiliary variables $\boldsymbol{m}$, $\boldsymbol{w}$, and $\bar{\boldsymbol{m}}$), mode-specific transition distributions π_k, and hyperparameters α, γ, and κ as in [14].
6) For each $k \in \{1, \ldots, L\}$, sample dynamic parameters $(\mathbf{A}^{(k)}, \Sigma^{(k)})$ given the pseudo-observations $\boldsymbol{\psi}_{1:T}$ 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
$$R \sim \text{IW}\left(T + r_0, \sum_{t=1}^{T}(\boldsymbol{y}_t - C\boldsymbol{x}_t)(\boldsymbol{y}_t - C\boldsymbol{x}_t)^T + R_0\right).$$
8) Fix $\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}$, $\beta^{(n)} = \beta$, and $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}$.

Algorithm 1: HDP-SLDS and HDP-AR-HMM Gibbs sampler.

$A^{(k)} = I$ 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 γ, 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

Given mode-specific transition probabilities $\boldsymbol{\pi}$, dynamic parameters $\boldsymbol{\theta}$, and pseudo-observations $\boldsymbol{\psi}_{1:T}$:

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 $t \in \{T, \dots, 1\}$ and $k \in \{1, \dots, L\}$, compute

$$m_{t,t-1}(k) = \sum_{j=1}^{L} \pi_k(j) \mathcal{N}\left(\boldsymbol{\psi}_t; \sum_{i=1}^{r} A_i^{(j)} \boldsymbol{\psi}_{t-i}, \Sigma^{(j)}\right) m_{t+1,t}(j)$$

   b) Working sequentially forward in time, starting with transitions counts $n_{jk} = 0$:
      i) For each $k \in \{1, \dots, L\}$, compute the probability

$$f_k(\boldsymbol{\psi}_t) = \mathcal{N}\left(\boldsymbol{y}_t; \sum_{i=1}^{r} A_i^{(k)} \boldsymbol{\psi}_{t-i}, \Sigma^{(k)}\right) m_{t+1,t}(k)$$

      ii) Sample a mode assignment z_t as follows and increment $n_{z_{t-1} z_t}$:

$$z_t \sim \sum_{k=1}^{L} \pi_{z_{t-1}}(k) f_k(\boldsymbol{\psi}_t) \delta(z_t, k)$$

   Note that the likelihoods can be precomputed for each $k \in \{1, \dots, L\}$.

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

Given pseudo-observations $\boldsymbol{\psi}_{1:T}$ and mode sequence $z_{1:T}$, for each $k \in \{1, \dots, K\}$:

1) Construct $\boldsymbol{\Psi}^{(k)}$ and $\bar{\boldsymbol{\Psi}}^{(k)}$ as in Eq. (15).
2) Compute sufficient statistics using pseudo-observations $\boldsymbol{\psi}_t$ associated with $z_t = k$:

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

3) Sample dynamic parameters:

$$\Sigma^{(k)} \sim \text{IW}\left(n_k + n_0, \mathbf{S}_{\psi|\bar{\psi}}^{(k)} + S_0\right) \qquad \mathbf{A}^{(k)} \mid \Sigma^{(k)} \sim \mathcal{MN}\left(\mathbf{A}^{(k)}; \mathbf{S}_{\psi\bar{\psi}}^{(k)} \mathbf{S}_{\bar{\psi}\bar{\psi}}^{-(k)}, \Sigma^{(k)}, \mathbf{S}_{\bar{\psi}\bar{\psi}}^{(k)}\right).$$

Algorithm 3: Parameter sampling using MNIW prior.

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.

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[7]. We generated data from a two-mode SLDS with 0.98 probability of self-transition and

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

[7]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.

Given pseudo-observations $\boldsymbol{\psi}_{1:T}$, mode sequence $z_{1:T}$, and a previous set of dynamic parameters $(\mathbf{A}^{(k)}, \Sigma^{(k)}, \boldsymbol{\alpha}^{(k)})$, 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 $\boldsymbol{\alpha}^{(k)}$ as in Eq. (23) and sample the dynamic matrix:

$$\mathrm{vec}(\mathbf{A}^{(k)}) \mid \Sigma^{(k)}, \boldsymbol{\alpha}^{(k)} \sim \mathcal{N}^{-1}\left(\sum_{t|z_t=k} \tilde{\Psi}_{t-1}^T \Sigma^{-(k)} \boldsymbol{\psi_t}, \Sigma_0^{-(k)} + \sum_{t|z_t=k} \tilde{\Psi}_{t-1}^T \Sigma^{-(k)} \tilde{\Psi}_{t-1} \right).$$

   b) For each $\ell \in \{1, \ldots, m\}$, with $m = n$ for the SLDS and $m = r$ for the switching VAR, sample ARD precision parameters:

$$\alpha_\ell^{(k)} \mid \mathbf{A}^{(k)} \sim \text{Gamma}\left(a + \frac{|\mathcal{S}_\ell|}{2}, b + \frac{\sum_{(i,j)\in\mathcal{S}_\ell} a_{ij}^{(k)2}}{2} \right).$$

   c) Compute sufficient statistic:

$$\mathbf{S}_{\psi|\bar{\psi}}^{(k)} = \sum_{t|z_t=k} (\boldsymbol{\psi}_t - \mathbf{A}^{(k)} \bar{\boldsymbol{\psi}}_{t-1})(\boldsymbol{\psi}_t - \mathbf{A}^{(k)} \bar{\boldsymbol{\psi}}_{t-1})^T$$

   and sample process noise covariance:

$$\Sigma^{(k)} \mid \mathbf{A}^{(k)} \sim \text{IW}\left(n_k + n_0, \mathbf{S}_{\psi|\bar{\psi}}^{(k)} + S_0 \right).$$

Algorithm 4: Parameter sampling using ARD prior.

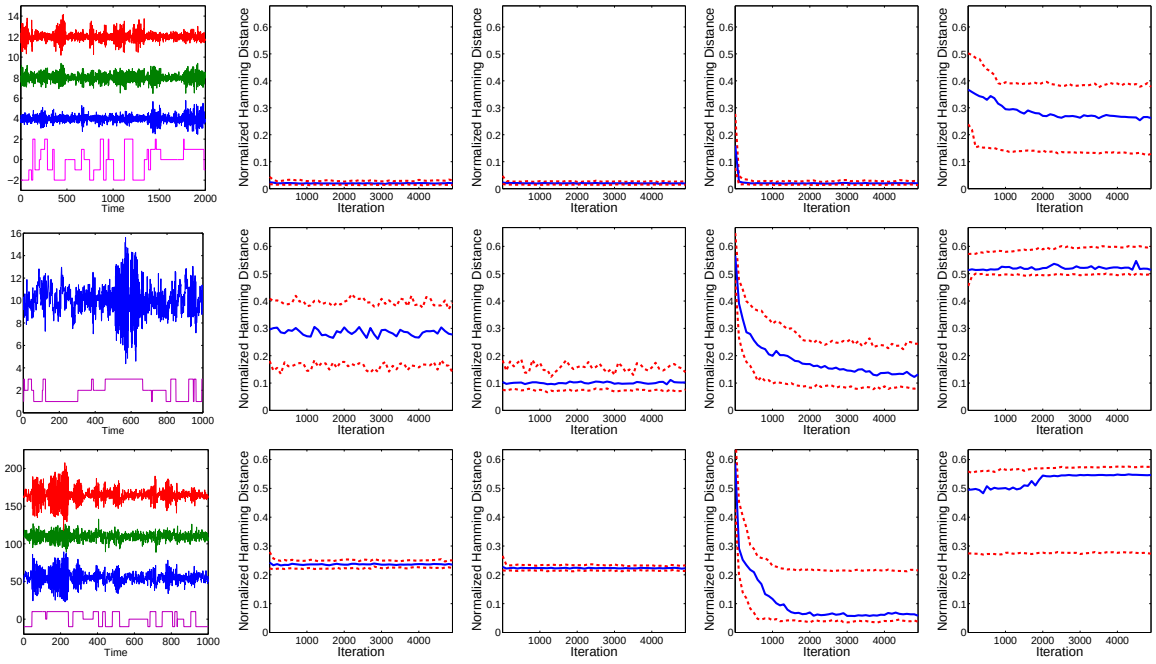
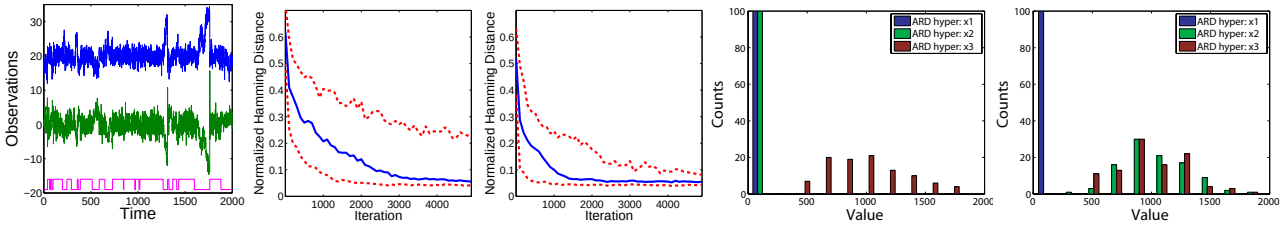


Fig. 2. (a) Observation sequence (blue, green, red) and associated mode sequence (magenta) for a 5-mode switching VAR(1) process (top), 3-mode switching AR(2) process (middle), and 3-mode SLDS (bottom). The associated 10th, 50th, and 90th Hamming distance quantiles over 100 trials are shown for the (b) 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.

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



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

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)-(e) Histograms of inferred ARD precisions associated with the first and second dynamical modes, respectively, at the 5000th Gibbs iteration. Larger values correspond to non-dynamical components.

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 dynamics of this mode as

$$x_{1,t} = -0.2x_{1,t-1} + \tilde{e}_{1,t} \qquad \tilde{\mathbf{A}}^{(2)} = \begin{bmatrix} -0.2 & 0 & 0 \\ 0.8 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$
$$x_{2,t} = 0.8x_{1,t-1} + \tilde{e}_{2,t}$$

where $\tilde{\boldsymbol{e}}_t$ is a white noise term defined by the original process noise combined with $x_{3,t}$, and $\tilde{\mathbf{A}}^{(2)}$ 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 $\mathbf{A}^{(2)}$ 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 $\boldsymbol{a}_2^{(2)} = \boldsymbol{a}_3^{(2)} = 0$ sparsity structure. We set the parameters of the Gamma(a, b) prior on the ARD precisions as $a = |\mathcal{S}_\ell|$ and $b = a/1000$, where we recall the definition of $\mathcal{S}_\ell$ from Eq. (26). This specification fixes the mean of the prior to 1000 while aiming to provide a prior that is equally informative for various choices of model order (i.e., sizes $|\mathcal{S}_\ell|$).

In Fig. 3, we see that even in this low-dimensional example, the ARD provides superior mode-sequence estimates, as well as a mechanism for identifying non-dynamical state components. The histograms of the inferred $\boldsymbol{\alpha}^{(k)}$ are shown in Fig. 3(d)-(e). 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 $\boldsymbol{a}_3^{(1)} = 0$ and $\boldsymbol{a}_2^{(2)} = \boldsymbol{a}_3^{(2)} = 0$.

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 $\boldsymbol{y}_t = [\cos(\theta_t) \quad \sin(\theta_t) \quad x_t \quad y_t]^T$, where (x_t, y_t) denotes the 2D coordinates of the bee's body and θ_t its head angle[8]. Both Oh et. al. [10] and Xuan and Murphy [11] 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 $\boldsymbol{y}_t$ to be within the same dynamic range. We compare our results to those of Xuan and Murphy [11], who used a change-point detection technique for inference on this dataset. As shown in Fig. 5(d) and (h), our model achieves a superior segmentation compared to

[8]The data are available at http://www.cc.gatech.edu/ borg/ijcv_pssIds/.

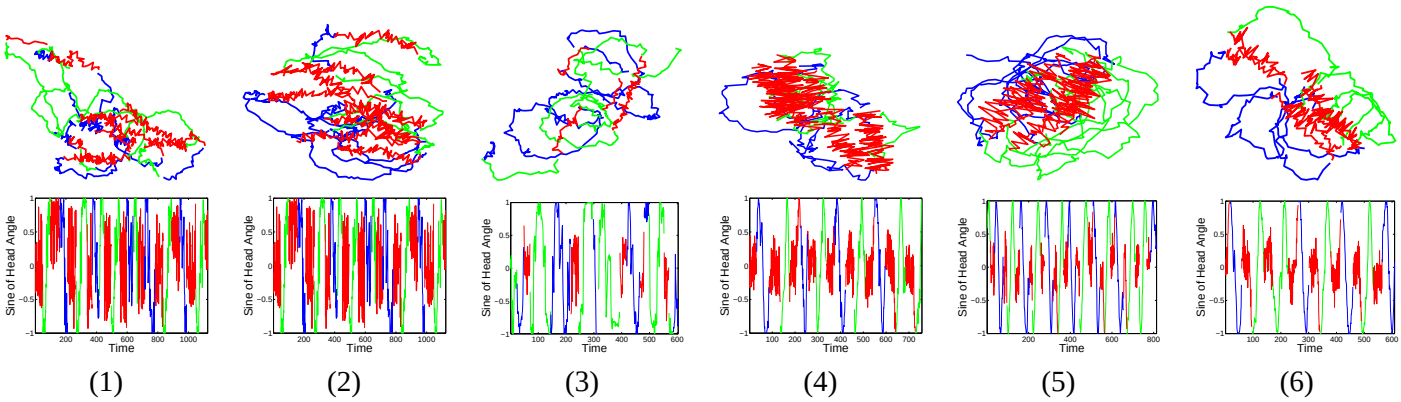


Fig. 4. *Top:* Trajectories of the dancing honey bees for sequences 1 to 6, colored by *waggle* (red), *turn right* (blue), and *turn left* (green) dances. *Bottom:* Sine of the bee's head angle measurements colored by ground truth labels.

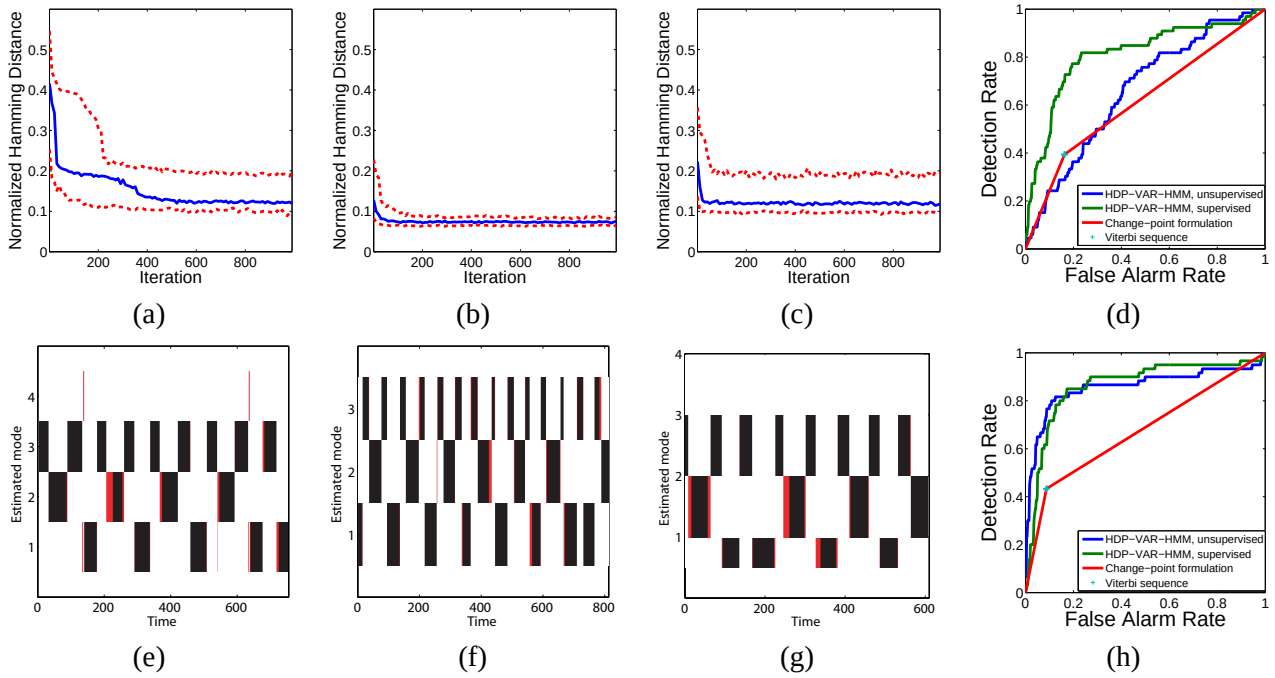


Fig. 5. (a)-(c) The 10th, 50th, and 90th Hamming distance quantiles over 100 trials are shown for sequences 4, 5, and 6, respectively. (e)-(g) Estimated mode sequences representing the median error for sequences 4, 5, and 6 at the 200th Gibbs iteration, with errors indicated in red. (d) and (h) ROC curves for the unsupervised HDP-VAR-HMM, partially supervised HDP-VAR-HMM, and change-point formulation of [11] using the Viterbi sequence for segmenting datasets 1-3 and 4-6, respectively.

the change-point formulation in almost all cases, while also identifying modes which reoccur over time. 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

Sequence	1	2	3	4	5	6
HDP-VAR(1)-HMM unsupervised	45.0	42.7	47.3	88.1	92.5	88.2
HDP-VAR(1)-HMM partially supervised	55.0	86.3	81.7	89.0	92.4	89.6
SLDS DD-MCMC	74.0	86.1	81.3	93.4	90.2	90.4
PS-SLDS DD-MCMC	75.9	92.4	83.1	93.4	90.4	91.0

TABLE III
MEDIAN LABEL ACCURACY OF THE HDP-VAR(1)-HMM USING UNSUPERVISED AND PARTIALLY SUPERVISED GIBBS SAMPLING, COMPARED TO ACCURACY OF THE SUPERVISED PS-SLDS AND SLDS PROCEDURES, WHERE THE LATTER ALGORITHMS WERE BASED ON A SUPERVISED MCMC PROCEDURE (DD-MCMC) [10].

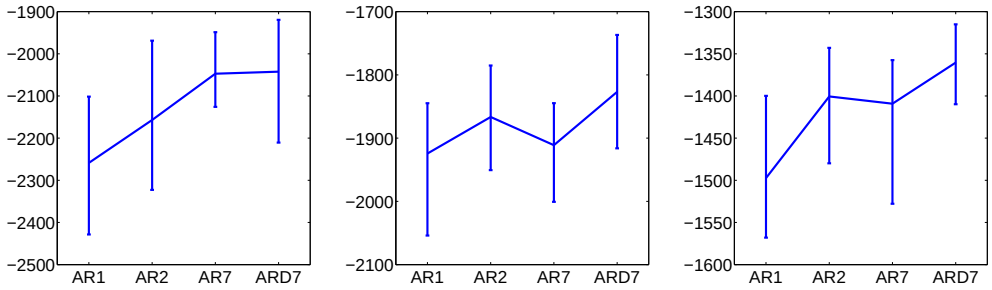


Fig. 6. For an order 1, 2, and 7 HDP-AR-HMM with a MNIW prior and an order 7 HDP-AR-HMM with an ARD prior, we plot the shortest intervals containing 95% of the held-out log-likelihoods calculated based on a set of Gibbs samples taken at iteration 1000 from 100 chains. (a) Log-likelihood of the second half of honey bee dance sequence 4 based on model parameters inferred from the first half of the sequence. (b)-(c) Similarly for sequences 5 and 6, respectively.

in head angle during the waggle dances of sequences 1 to 3.[9] 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. Indeed, our learned segmentations consistently identify turn-right and turn-left modes, but often create a new, sequence-specific waggle dance mode. Many of our errors can be attributed to creating multiple waggle dance modes within a 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, this partially supervised approach considerably improves performance for these three sequences, especially sequences 2 and 3. Here, 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 $\boldsymbol{\mu}^{(k)}$ 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. 6, 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

[9]From Fig. 4, we also see that even in sequences 4 to 6, the ground truth labeling appear to be inaccurate at times. Specifically, certain time steps are labeled as waggle dances (red) that look more typical of a turning dance (green, blue).

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 [36]), 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 turning dances simply rely on the first lag component while the other dance relies on both lag components. To verify these results, we provided the data and ground truth labels to MATLAB's `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

$$\begin{aligned} z_t \mid z_{t-1} &\sim \pi_{z_{t-1}} \\ \boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + \boldsymbol{e}_t(z_t) \quad & \boldsymbol{y}_t = C\boldsymbol{x}_t + \boldsymbol{w}_t, \end{aligned} \tag{37}$$

with process and measurement noise $\boldsymbol{e}_t(k) \sim \mathcal{N}(\boldsymbol{\mu}^{(k)}, \Sigma^{(k)})$ and $\boldsymbol{w}_t \sim \mathcal{N}(0, R)$, 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 $\boldsymbol{\mu}^{(k)}$) due to the coupling of the parameters in this prior. We instead consider independent priors on A, $\Sigma^{(k)}$, and $\boldsymbol{\mu}^{(k)}$. We will refer to the choice of a normal prior on A, inverse-Wishart prior on $\Sigma^{(k)}$, and normal prior on $\boldsymbol{\mu}^{(k)}$ as the *N-IW-N* prior. See [36] 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], [38]. 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 is given by [7]

$$\begin{aligned} z_t \mid z_{t-1} &\sim \pi_{z_{t-1}} \\ x_t = ax_{t-1} + e_t(z_t) \quad & y_t = u_t(x_t), \end{aligned} \tag{38}$$

with $e_t(k) \sim \mathcal{N}(\mu^{(k)}, \sigma^2)$ 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

$$\log(y_t^2) = x_t + w_t, \tag{39}$$

where w_t is additive, non-Gaussian noise. This noise is sometimes approximated by a moment-matched Gaussian [39], 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 during this time period. The key world events are summarized in Table IV and shown in the plots of Fig. 7. Use of this dataset was motivated

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

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].

by the work of Carvalho and Lopes [7], in which a two-mode MSSV model is assumed. 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:

$$\begin{aligned} e_t(k) &\sim \mathcal{N}(\mu^{(k)}, \Sigma^{(k)}) \\ w_t \sim \textstyle\sum_{\ell=1}^{\infty} \omega_\ell \mathcal{N}(0, R_\ell) \quad \boldsymbol{\omega} &\sim \text{GEM}(\sigma_r), \quad R_\ell \sim \text{IW}(n_r, S_r). \end{aligned} \tag{40}$$

We truncate the measurement noise DP mixture to 10 components. For the HDP concentration hyperparameters, α, γ, and κ, 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. 7(a), and in Fig. 7(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. In Fig. 7(c), the overall change-point detection performance of the HDP-SLDS is compared to that of the HDP-AR(1)-HMM, HDP-AR(2)-HMM, and non-sticky HDP-SLDS. The ROC curves shown are calculated by windowing the time axis and taking the maximum probability of a change point in each window. These probabilities are then used as the confidence of a change point in that window. From this plot, we clearly see the advantage of using an SLDS model combined with the sticky HDP-HMM prior on the mode sequence.

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 changepoints that align better with the true world events than in the MSSV HDP-SLDS model.

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 using a more complex description of the driving force. A generic LDS driven by an unknown control input $\boldsymbol{u}_t$ can be represented as

$$\boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + B\boldsymbol{u}_t + \boldsymbol{v}_t \quad \boldsymbol{y}_t = C\boldsymbol{x}_t + D\boldsymbol{u}_t + \boldsymbol{w}_t, \tag{41}$$

where $\boldsymbol{v}_t \sim \mathcal{N}(0, Q)$ and $\boldsymbol{w}_t \sim \mathcal{N}(0, R)$. It is often appropriate to assume $D = 0$, as we do herein.

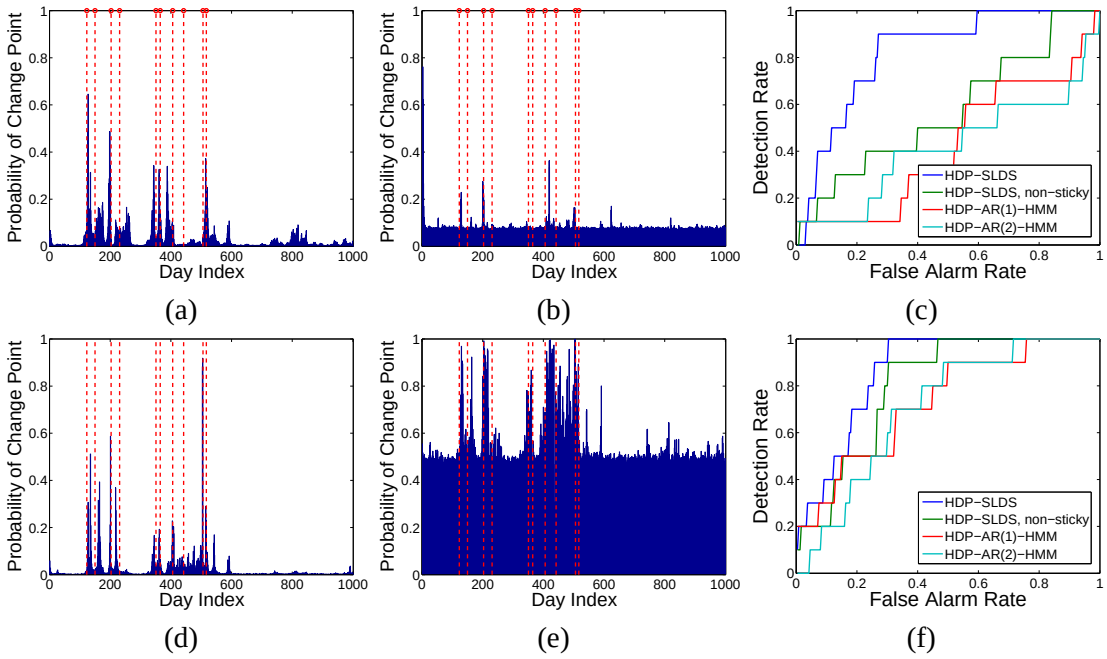


Fig. 7. (a) Plot of the estimated probability of a change point on each day using 3,000 Gibbs samples for a MSSV variant of the HDP-SLDS using a shared dynamic matrix and allowing a mean on the mode-specific process noise and a mixture of Gaussian measurement noise model. The observations are log-squared dialy 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. (c) ROC curves for the HDP-SLDS, non-sticky HDP-SLDS, HDP-AR(1)-HMM, and HDP-AR(2)-HMM. (d)-(f) Analogous plots for the HDP-SLDS of Table I using raw daily return measurements.

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 [40]. For example, a *jump-mean* Markov process [41] yields dynamics described as

$$\begin{aligned} z_t \mid z_{t-1} &\sim \pi_{z_{t-1}} \\ \boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + B\boldsymbol{u}_t(z_t) + \boldsymbol{v}_t \quad & \boldsymbol{y}_t = C\boldsymbol{x}_t + \boldsymbol{w}_t \\ \boldsymbol{u}_t(k) \sim \mathcal{N}(\boldsymbol{\mu}^{(k)}, \Sigma^{(k)}) \quad \boldsymbol{v}_t \sim \mathcal{N}(0, Q) \quad & \boldsymbol{w}_t \sim \mathcal{N}(0, R). \end{aligned} \tag{42}$$

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

$$\boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + \boldsymbol{e}_t(z_t) \tag{43}$$

$$\boldsymbol{e}_t(k) \sim \mathcal{N}(B\boldsymbol{\mu}^{(k)}, B\Sigma^{(k)}B^T + Q). \tag{44}$$

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 [40]) 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 $(z_t, \boldsymbol{u}_t)$ analytically marginalizing over the state sequence $\boldsymbol{x}_{1:T}$. Note that this variant of the HDP-SLDS can be viewed as an extension of the work by Caron et. al. [42] 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) [43]. However, one challenge is that such particle filters can suffer from a progressively impoverished particle representation.

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 $M = \mathbf{0}$ 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 $S_0 = 0.75\bar{\Sigma}$, where $\bar{\Sigma} = \frac{1}{T}\sum(\boldsymbol{y}_t - \bar{\boldsymbol{y}})(\boldsymbol{y}_t - \bar{\boldsymbol{y}})^T$. Setting the prior directly from the data can help move the mass of the distribution to reasonable values of the parameter space. For an HDP-SLDS with $\boldsymbol{x}_t \in \mathbb{R}^n$ and $\boldsymbol{y}_t \in \mathbb{R}^d$ 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 [36].

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. [39]. For the HDP-AR(r)-HMM's to which we compare in Fig. 7, we place a zero-mean normal prior on the dynamic parameter a with covariance set to the expected noise covariance, which in this case is equal to 0.75 times the empirical covariance plus $5 * \pi^2$. The mean parameter $\mu^{(k)}$ is defined as in the HDP-SLDS.

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

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

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

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

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

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

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

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

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

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

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

[11] X. Xuan and K. Murphy, "Modeling changing dependency structure in multivariate time series," in *Proc. International Conference on Machine Learning*, June 2007.

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

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

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

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

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

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

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

[19] R. Vidal, S. Soatto, Y. Ma, and S. Sastry, "An algebraic geometric approach to the identification of a class of linear hybrid systems." in *Proc. IEEE Conference on Decision and Control*, December 2003.

[20] Z. Psaradakis and N. Spagnolo, "Joint determination of the state dimension and autoregressive order for models with Markov regime switching," *Journal of Time Series Analysis*, vol. 27, pp. 753–766, 2006.

[21] K. Huang, A. Wagner, and Y. Ma, "Identification of hybrid linear time-invariant systems via subspace embedding and segmentation SES." in *Proc. IEEE Conference on Decision and Control*, December 2004.

[22] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA): Subspace clustering by polynomial factorization, differentiation, and division." *UC Berkeley, Technical Report UCB/ERL.*, August 2003.

[23] G. Kotsalis, A. Megretski, and M. Dahleh, "Model reduction of discrete-time Markov jump linear systems," in *Proc. American Control Conference*, June 2006.

[24] B. Anderson, "The realization problem for hidden Markov models," *Mathematics of Control, Signals, and Systems*, vol. 12, pp. 80–120, 1999.

[25] M. Petreczky and R. Vidal, "Realization theory of stochastic jump-Markov linear systems," in *Proc. IEEE Conference on Decision and Control*, December 2007.

[26] Z. Ghahramani and G. Hinton, "Variational learning for switching state-space models," *Neural Computation*, vol. 12, no. 4, pp. 831–864, 2000.

[27] R. Vidal, A. Chiuso, , and S. Soatto, "Observability and identifiability of jump linear systems," in *Proc. IEEE Conference on Decision and Control*, December 2002.

[28] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[29] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, pp. 461–464, 1978.

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

[31] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the*

IEEE, vol. 77, no. 2, pp. 257–286, 1989.

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

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

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

[35] O. Costa, M. Fragoso, and R. Marques, *Discrete-Time Markov Jump Linear Systems*. Springer, 2005.

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

[37] C. Carter and R. Kohn, "Markov chain Monte Carlo in conditionally Gaussian state space models," *Biometrika*, vol. 83, pp. 589–601, 3 1996.

[38] J. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," *Econometrica*, vol. 57, no. 2, pp. 357–384, 1989.

[39] A. Harvey, E. Ruiz, and N. Shephard, "Multivariate stochastic variance models," *Review of Economic Studies*, vol. 61, pp. 247–264, 1994.

[40] X. Rong Li and V. Jilkov, "Survey of maneuvering target tracking. Part I: Dynamic models," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1333–1364, 2003.

[41] R. Moose, H. VanLandingham, and D. McCabe, "Modeling and estimation of tracking maneuvering targets," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 15, no. 3, pp. 448–456, 1979.

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

[43] A. Doucet, N. de Freitas, K. Murphy, and S. Russell.

APPENDIX A
DYNAMIC PARAMETER POSTERIORS

In this appendix, we derive the posterior distribution over the dynamic parameters of a switching VAR(r) process defined as follows:

$$\boldsymbol{y}_t = \sum_{i=1}^{r} A_i^{(z_t)} \boldsymbol{y}_{t-i} + \boldsymbol{e}_t(z_t) \quad \boldsymbol{e}_t(k) \sim \mathcal{N}(\boldsymbol{\mu}^{(k)}, \Sigma^{(k)}), \tag{45}$$

where z_t indexes the mode-specific VAR(r) process at time t. Assume that the mode sequence $\{z_1, \ldots, z_T\}$ is known and we wish to compute the posterior distribution of the k^{th} mode's VAR(r) parameters $A_i^{(k)}$ for $i = 1, \ldots, r$ and $\Sigma^{(k)}$. Let $\{t_1, \ldots, t_{n_k}\} = \{t | z_t = k\}$. Then, we may write

$$\begin{bmatrix} \boldsymbol{y}_{t_1} & \boldsymbol{y}_{t_2} & \cdots & \boldsymbol{y}_{t_{n_k}} \end{bmatrix} = \begin{bmatrix} A_1^{(k)} & A_2^{(k)} & \ldots & A_r^{(k)} \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_{t_1-1} & \boldsymbol{y}_{t_2-1} & \cdots & \boldsymbol{y}_{t_{n_k}-1} \\ \boldsymbol{y}_{t_1-2} & \boldsymbol{y}_{t_2-2} & \cdots & \boldsymbol{y}_{t_{n_k}-2} \\ \vdots & & & \\ \boldsymbol{y}_{t_1-r} & \boldsymbol{y}_{t_2-r} & \cdots & \boldsymbol{y}_{t_{n_k}-r} \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_{t_1} & \boldsymbol{e}_{t_2} & \ldots & \boldsymbol{e}_{t_{n_k}} \end{bmatrix}. \tag{46}$$

We define the following notation for Eq. (46):

$$\mathbf{Y}^{(k)} = \mathbf{A}^{(k)} \bar{\mathbf{Y}}^{(k)} + \mathbf{E}^{(k)}, \tag{47}$$

and let $\mathbf{D}^{(k)} = \{\mathbf{Y}^{(k)}, \bar{\mathbf{Y}}^{(k)}\}$. In the following sections, we consider two possible priors on the dynamic parameter. In Appendix A- A, we assume that $\boldsymbol{\mu}^{(k)}$ is 0 for all k and consider the conjugate matrix-normal inverse-Wishart (MNIW) prior for $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$. In Appendix A- B, we consider the more general form of Eq. (45) and take independent priors on $\mathbf{A}^{(k)}$, $\Sigma^{(k)}$, and $\boldsymbol{\mu}^{(k)}$.

A. Conjugate Prior — MNIW

To show conjugacy, we place a MNIW prior on the dynamic parameters $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$ and show that the posterior remains MNIW given a set of data from the model of Eq. (45) (assuming $\boldsymbol{\mu}^{(k)} = 0$). The MNIW prior is given by placing a matrix-normal prior $\mathcal{MN}\left(\mathbf{A}^{(k)}; M, \Sigma^{(k)}, K\right)$ on $\mathbf{A}^{(k)}$ given $\Sigma^{(k)}$:

$$p(\mathbf{A}^{(k)} \mid \Sigma^{(k)}) = \frac{|K|^{d/2}}{|2\pi\Sigma^{(k)}|^{m/2}} \exp\left(-\frac{1}{2} tr((\boldsymbol{A} - M)^T \Sigma^{-(k)} (\boldsymbol{A} - M) K)\right) \tag{48}$$

and an inverse-Wishart prior IW(n_0, S_0) on $\Sigma^{(k)}$:

$$p(\Sigma^{(k)}) = \frac{|S_0|^{n_0/2} |\Sigma^{(k)}|^{-(d+n_0+1)/2}}{2^{n_0 d/2} \Gamma_d(n_0/2)} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)} S_0)\right) \tag{49}$$

where $\Gamma_d(\cdot)$ is the multivariate gamma function and $\boldsymbol{B}^{-(k)}$ denotes $(\boldsymbol{B}^{(k)})^{-1}$ for some matrix $\boldsymbol{B}$.

We first analyze the likelihood of the data, $\mathbf{D}^{(k)}$, given the k^{th} mode's dynamic parameters, $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$. Starting with the fact that each observation vector, $\boldsymbol{y}_t$, is conditionally Gaussian given the lag observations, $\bar{\boldsymbol{y}}_t = [\boldsymbol{y}_{t-1}^T \ldots \boldsymbol{y}_{t-r}^T]^T$, we have

$$\begin{aligned} p(\mathbf{D}^{(k)} | \mathbf{A}^{(k)}, \Sigma^{(k)}) &= \frac{1}{|2\pi\Sigma^{(k)}|^{n_k/2}} \exp\left(-\frac{1}{2} \sum_i (\boldsymbol{y}_{t_i} - \mathbf{A}^{(k)} \bar{\boldsymbol{y}}_{t_i})^T \Sigma^{-(k)} (\boldsymbol{y}_{t_i} - \mathbf{A}^{(k)} \bar{\boldsymbol{y}}_{t_i})\right) \\ &= \frac{1}{|2\pi\Sigma^{(k)}|^{n_k/2}} \exp\left(-\frac{1}{2} tr((\mathbf{Y}^{(k)} - \mathbf{A}^{(k)} \bar{\mathbf{Y}}^{(k)})^T \Sigma^{-(k)} (\mathbf{Y}^{(k)} - \mathbf{A}^{(k)} \bar{\mathbf{Y}}^{(k)}) \boldsymbol{I})\right) \\ &= \mathcal{MN}\left(\mathbf{Y}^{(k)}; \mathbf{A}^{(k)} \bar{\mathbf{Y}}^{(k)}, \Sigma^{(k)}, \boldsymbol{I}\right). \end{aligned} \tag{50}$$

To derive the posterior of the dynamic parameters, it is useful to first compute

$$p(\mathbf{D}^{(k)}, \mathbf{A}^{(k)} \mid \Sigma^{(k)}) = p(\mathbf{D}^{(k)} \mid \mathbf{A}^{(k)}, \Sigma^{(k)}) p(\mathbf{A}^{(k)} \mid \Sigma^{(k)}). \tag{51}$$

Using the fact that both the likelihood $p(\mathbf{D}^{(k)} \mid \mathbf{A}^{(k)}, \Sigma^{(k)})$ and the prior $p(\mathbf{A}^{(k)} \mid \Sigma^{(k)})$ are matrix-normally distributed sharing a common parameter $\Sigma^{(k)}$, we have

$$\begin{aligned}
&\log p(\mathbf{D}^{(k)}, \mathbf{A}^{(k)} \mid \Sigma^{(k)}) + C \\
&\quad = -\frac{1}{2} tr((\mathbf{Y}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{Y}}^{(k)})^T \Sigma^{-(k)} (\mathbf{Y}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{Y}}^{(k)}) + (\mathbf{A}^{(k)} - M)^T \Sigma^{-(k)} (\mathbf{A}^{(k)} - M) K) \\
&\quad = -\frac{1}{2} tr(\Sigma^{-(k)} \{\mathbf{A}^{(k)} \mathbf{S}_{\bar{y}\bar{y}}^{(k)} \mathbf{A}^{(k)^T} - 2\mathbf{S}_{y\bar{y}}^{(k)} \mathbf{A}^{(k)^T} + \mathbf{S}_{yy}^{(k)}\}) \\
&\quad = -\frac{1}{2} tr(\Sigma^{-(k)} \{(\mathbf{A}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)} \mathbf{S}_{\bar{y}\bar{y}}^{-(k)}) \mathbf{S}_{\bar{y}\bar{y}}^{(k)} (\mathbf{A}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)} \mathbf{S}_{\bar{y}\bar{y}}^{-(k)})^T + \mathbf{S}_{y|\bar{y}}^{(k)}\}), \qquad (52)
\end{aligned}$$

where we have used the definitions:

$$C = -\log \frac{1}{|2\pi \mathbf{\Sigma}^{(k)}|^{n_k/2}} \frac{|K|^{d/2}}{|2\pi \mathbf{\Sigma}^{(k)}|^{r n_k/2}} \qquad \mathbf{S}_{y|\bar{y}}^{(k)} = \mathbf{S}_{yy}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)} \mathbf{S}_{\bar{y}\bar{y}}^{-(k)} \mathbf{S}_{y\bar{y}}^{(k)^T},$$

$$\mathbf{S}_{\bar{y}\bar{y}}^{(k)} = \bar{\mathbf{Y}}^{(k)} \bar{\mathbf{Y}}^{(k)^T} + K \qquad \mathbf{S}_{y\bar{y}}^{(k)} = \mathbf{Y}^{(k)} \bar{\mathbf{Y}}^{(k)^T} + MK \qquad \mathbf{S}_{yy}^{(k)} = \mathbf{Y}^{(k)} \mathbf{Y}^{(k)^T} + MKM^T.$$

Conditioning on the noise covariance $\Sigma^{(k)}$, we see that the dynamic matrix posterior is given by:

$$\begin{aligned}
p(\mathbf{A}^{(k)} \mid \mathbf{D}^{(k)}, \Sigma^{(k)}) &\propto \exp\left(-\frac{1}{2} tr((\mathbf{A}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)} \mathbf{S}_{\bar{y}\bar{y}}^{-(k)})^T \Sigma^{-(k)} (\mathbf{A}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)} \mathbf{S}_{\bar{y}\bar{y}}^{-(k)}) \mathbf{S}_{\bar{y}\bar{y}}^{(k)})\right) \\
&= \mathcal{MN}\left(\mathbf{A}^{(k)}; \mathbf{S}_{y\bar{y}}^{(k)} \mathbf{S}_{\bar{y}\bar{y}}^{-(k)}, \Sigma^{(k)}, \mathbf{S}_{\bar{y}\bar{y}}^{(k)}\right). \qquad (53)
\end{aligned}$$

Marginalizing Eq. (52) over the dynamic matrix $\mathbf{A}^{(k)}$, we derive

$$\begin{aligned}
p(\mathbf{D}^{(k)} \mid \Sigma^{(k)}) &= \int_{\mathbf{A}^{(k)}} p(\mathbf{D}^{(k)}, \mathbf{A}^{(k)} \mid \Sigma^{(k)}) d\mathbf{A}^{(k)} \\
&= \frac{|K|^{d/2}}{|2\pi\Sigma^{(k)}|^{n_k/2}} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)} \mathbf{S}_{y|\bar{y}}^{(k)})\right) \int_{\mathbf{A}^{(k)}} \frac{1}{|\mathbf{S}_{\bar{y}\bar{y}}^{(k)}|^{d/2}} \mathcal{MN}\left(\mathbf{A}^{(k)}; \mathbf{S}_{y\bar{y}}^{(k)} \mathbf{S}_{\bar{y}\bar{y}}^{-(k)}, \Sigma^{(k)}, \mathbf{S}_{\bar{y}\bar{y}}^{(k)}\right) d\mathbf{A}^{(k)} \\
&= \frac{|K|^{d/2}}{|2\pi\Sigma^{(k)}|^{n_k/2} |\mathbf{S}_{\bar{y}\bar{y}}^{(k)}|^{d/2}} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)} \mathbf{S}_{y|\bar{y}}^{(k)})\right), \qquad (54)
\end{aligned}$$

which leads us to our final result of the covariance having an inverse-Wishart marginal posterior distribution:

$$\begin{aligned}
p(\Sigma^{(k)} \mid \mathbf{D}^{(k)}) &\propto p(\mathbf{D}^{(k)} \mid \Sigma^{(k)}) p(\Sigma^{(k)}) \\
&\propto \frac{|K|^{d/2}}{|2\pi\Sigma^{(k)}|^{n_k/2} |\mathbf{S}_{\bar{y}\bar{y}}^{(k)}|^{d/2}} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)} \mathbf{S}_{y|\bar{y}}^{(k)})\right) |\Sigma^{(k)}|^{-(d+n_0+1)/2} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)} S_0)\right) \\
&\propto |\Sigma^{(k)}|^{-(d+n_k+n_0+1)/2} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)} (\mathbf{S}_{y|\bar{y}}^{(k)} + S_0))\right) \\
&= \text{IW}(n_k + n_0, \mathbf{S}_{y|\bar{y}}^{(k)} + S_0). \qquad (55)
\end{aligned}$$

B. Non-Conjugate Independent Priors on $\mathbf{A}^{(k)}$, $\Sigma^{(k)}$, and $\boldsymbol{\mu}^{(k)}$

In this section, we provide the derivations for the posterior distributions of $\mathbf{A}^{(k)}$, $\Sigma^{(k)}$, and $\boldsymbol{\mu}^{(k)}$ when each of these parameters is given an independent prior. One example of a non-conjugate prior is our proposed ARD sparsity-inducing prior.

a) Normal Prior on $\mathbf{A}^{(k)}$: Assume we place a Gaussian prior, $\mathcal{N}(\boldsymbol{\mu}_A, \Sigma_A)$, on the vectorization of the matrix $\mathbf{A}^{(k)}$, which we denote by $\text{vec}(\mathbf{A}^{(k)})$. To examine the posterior distribution, we first aim to write the data as a linear function of $\text{vec}(\mathbf{A}^{(k)})$. We may rewrite Eq. (45) as

$$\begin{aligned}\boldsymbol{y}_t &= \mathbf{A}^{(k)} \begin{bmatrix} \boldsymbol{y}_{t-1}^T & \boldsymbol{y}_{t-2}^T & \cdots & \boldsymbol{y}_{t-r}^T \end{bmatrix}^T + \boldsymbol{e}_t \qquad \forall t | z_t = k \\ &\triangleq \mathbf{A}^{(k)} \bar{\boldsymbol{y}}_t + \boldsymbol{e}_t(k). \end{aligned} \tag{56}$$

Recalling that r is the autoregressive order and d the dimension of the observation vector $\boldsymbol{y}_t$, we can equivalently represent the above as

$$\begin{aligned}\boldsymbol{y}_t &= \begin{bmatrix} \bar{y}_{t,1} & \bar{y}_{t,2} & \cdots & \bar{y}_{t,d*r} & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \bar{y}_{t,1} & \bar{y}_{t,2} & \cdots & \bar{y}_{t,d*r} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \bar{y}_{t,1} & \bar{y}_{t,2} & \cdots & \bar{y}_{t,d*r} \end{bmatrix} \begin{bmatrix} a_{1,1}^{(k)} \\ a_{1,2}^{(k)} \\ \vdots \\ a_{1,d*r}^{(k)} \\ a_{2,1}^{(k)} \\ a_{2,2}^{(k)} \\ \vdots \\ a_{d,d*r}^{(k)} \end{bmatrix} + \boldsymbol{e}_t(k) \\ &= \begin{bmatrix} \bar{y}_{t,1} I_d & \bar{y}_{t,2} I_d & \cdots & \bar{y}_{t,d*r} I_d \end{bmatrix} \text{vec}(\mathbf{A}^{(k)}) + \boldsymbol{e}_t(k) \triangleq \bar{\boldsymbol{Y}}_t \text{vec}(A) + \boldsymbol{e}_t(k). \end{aligned} \tag{57}$$

Here, the columns of $\bar{\boldsymbol{y}}_t$ are permutations of those of the matrix in the first line such that we may write $\boldsymbol{y}_t$ as a function of $\text{vec}(\mathbf{A}^{(k)})$. Noting that $\boldsymbol{e}_t(k) \sim \mathcal{N}(\boldsymbol{\mu}^{(k)}, \Sigma^{(k)})$,

$$\begin{aligned}&\log p(\mathbf{D}^{(k)}, \mathbf{A}^{(k)} \mid \Sigma^{(k)}, \boldsymbol{\mu}^{(k)}) \\ &\qquad = C - \frac{1}{2} \sum_{t|z_t=k} (\boldsymbol{y}_t - \boldsymbol{\mu}^{(k)} - \bar{\boldsymbol{Y}}_t \text{vec}(\mathbf{A}^{(k)}))^T \Sigma^{-(k)} (\boldsymbol{y}_t - \boldsymbol{\mu}^{(k)} - \bar{\boldsymbol{Y}}_t \text{vec}(\mathbf{A}^{(k)})) \\ &\qquad\quad - \frac{1}{2} (\text{vec}(\mathbf{A}^{(k)}) - \boldsymbol{m}_A)^T \Sigma_A^{-1} (\text{vec}(\mathbf{A}^{(k)}) - \boldsymbol{m}_A), \end{aligned} \tag{58}$$

which can be rewritten as,

$$\begin{aligned}\log p(\mathbf{D}^{(k)}, \mathbf{A}^{(k)} \mid \Sigma^{(k)}, \boldsymbol{\mu}^{(k)}) = C &- \frac{1}{2} \text{vec}(\mathbf{A}^{(k)})^T \left(\Sigma_A^{-1} + \sum_{t|z_t=k} \bar{\boldsymbol{Y}}_t^T \Sigma^{-(k)} \bar{\boldsymbol{Y}}_t \right) \text{vec}(\mathbf{A}^{(k)}) \\ &+ \text{vec}(\mathbf{A}^{(k)})^T \left(\Sigma_A^{-1} \boldsymbol{m}_A + \sum_{t|z_t=k} \bar{\boldsymbol{Y}}_t^T \Sigma^{-(k)} (\boldsymbol{y}_t - \boldsymbol{\mu}^{(k)}) \right) \\ &- \frac{1}{2} \boldsymbol{m}_A^T \Sigma_A^{-1} \boldsymbol{m}_A - \frac{1}{2} \sum_{t|z_t=k} (\boldsymbol{y}_t - \boldsymbol{\mu}^{(k)})^T \Sigma^{-(k)} (\boldsymbol{y}_t - \boldsymbol{\mu}^{(k)}) \end{aligned} \tag{59}$$

Conditioning on the data, we arrive at the desired posterior distribution

$$\begin{aligned}\log p(\mathbf{A}^{(k)} \mid \mathbf{D}^{(k)}, \Sigma^{(k)}, \boldsymbol{\mu}^{(k)}) &= C - \frac{1}{2} \Bigg(\text{vec}(\mathbf{A}^{(k)})^T (\Sigma_A^{-1} + \sum_{t|z_t=k} \bar{\boldsymbol{Y}}_t^T \Sigma^{-(k)} \bar{\boldsymbol{Y}}_t) \text{vec}(\mathbf{A}^{(k)}) \\ &\qquad - 2 \text{vec}(\mathbf{A}^{(k)})^T (\Sigma_A^{-1} \boldsymbol{m}_A + \sum_{t|z_t=k} \bar{\boldsymbol{Y}}_t^T \Sigma^{-(k)} (\boldsymbol{y}_t - \boldsymbol{\mu}^{(k)})) \Bigg) \\ &= \mathcal{N}^{-1} \left(\Sigma_A^{-1} \boldsymbol{m}_A + \sum_{t|z_t=k} \bar{\boldsymbol{Y}}_t^T \Sigma^{-(k)} (\boldsymbol{y}_t - \boldsymbol{\mu}^{(k)}), \Sigma_A^{-1} + \sum_{t|z_t=k} \bar{\boldsymbol{Y}}_t^T \Sigma^{-(k)} \bar{\boldsymbol{Y}}_t \right) \end{aligned} \tag{60}$$

b) Inverse Wishart Prior on $\Sigma^{(k)}$: We place an inverse-Wishart prior, $\mathrm{IW}(n_0, S_0)$, on $\Sigma^{(k)}$. Let $n_k = |\{t | z_t = k, t = 1, 2, \ldots, T\}|$. Conditioned on $\mathbf{A}^{(k)}$ and $\boldsymbol{\mu}^{(k)}$, standard conjugacy results imply that the posterior of $\Sigma^{(k)}$ is:

$$p(\Sigma^{(k)} \mid \mathbf{D}^{(k)}, \mathbf{A}^{(k)}, \boldsymbol{\mu}^{(k)}) = \mathrm{IW}\left(n_k + n_0, S + \sum_{t|z_t=k} (\boldsymbol{y}_t - \mathbf{A}^{(k)}\bar{\boldsymbol{y}}_t - \boldsymbol{\mu}^{(k)})(\boldsymbol{y}_t - \mathbf{A}^{(k)}\bar{\boldsymbol{y}}_t - \boldsymbol{\mu}^{(k)})^T\right). \quad (61)$$

c) Normal Prior on $\boldsymbol{\mu}^{(k)}$: Finally, we place a Gaussian prior, $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$, on $\boldsymbol{\mu}^{(k)}$. Conditioned on $\mathbf{A}^{(k)}$ and $\Sigma^{(k)}$, the posterior of $\boldsymbol{\mu}^{(k)}$ is:

$$p(\boldsymbol{\mu}^{(k)} \mid \mathbf{D}^{(k)}, \mathbf{A}^{(k)}, \Sigma^{(k)}) = \mathcal{N}^{-1}\left(\boldsymbol{\mu}^{(k)}; \Sigma_0^{-1}\boldsymbol{\mu}_0 + \Sigma^{-(k)} \sum_{t|z_t=k} (\boldsymbol{y}_t - \mathbf{A}^{(k)}\bar{\boldsymbol{y}}_t), \Sigma_0^{-1} + n_k\Sigma^{-(k)}\right). \quad (62)$$

We iterate between sampling $\mathbf{A}^{(k)}$, $\Sigma^{(k)}$, and $\boldsymbol{\mu}^{(k)}$ many times before moving on to the next step of the Gibbs sampler.

APPENDIX B
SPARSITY-INDUCING PRIORS FOR INFERRING VARIABLE ORDER MODELS

Recall Sec. III-A2 and the proposed automatic relevance determination (ARD) prior for inferring non-dynamical components of the state vector in the case of the HDP-SLDS or lag components in the HDP-AR-HMM by shrinking components of the model parameters to zero. However, if we would like to ensure that our choice of $C = [I_d\ 0]$ does not interfere with learning a sparse realization if one exists, we must restrict ourselves to considered a constrained class of dynamical phenomenon. For example, imagine a realization of an LDS with

$$\tilde{A} = \begin{bmatrix} 0.8 & 0 \\ 0.2 & 0 \end{bmatrix}, \qquad \tilde{C} = \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

Then, the transformation to $C = [1\ 0]$ leads to

$$A = T^{-1}\tilde{A}T = \begin{bmatrix} 0.5 & 1 \\ 0.15 & 0.3 \end{bmatrix}, \qquad \text{for} \qquad T = \begin{bmatrix} 0.5 & 1 \\ 0.5 & -1 \end{bmatrix}.$$

So, for this example, fixing $C = [1\ 0]$ would not lead to learning a sparse dynamical matrix A. Criterion 3.1 provides a set of sufficient, though not necessary, conditions for maintaining the sparsity within each $\mathbf{A}^{(k)}$ when transforming to the realization with $C = [I_d\ 0]$. That is, given there exists a realization $\mathcal{R}_1$ of our dynamical phenomena that satisfies Criterion 3.1, the transformation T to an equivalent realization $\mathcal{R}_2$ with $C = [I_d\ 0]$ will maintain the sparsity structure seen in $\mathcal{R}_1$, which we aim to infer with the ARD prior. Criterion 3.1, which states that the observed state vector components are a subset of those relevant to *all* modes, is reasonable for many applications: we often have observations only of components of the state vector that are essential to *all* modes while *some* modes may have additional components that affect the dynamics, but are not directly observed.

To clarify the conditions of Criterion 3.1, consider a 3-mode SLDS realization $\mathcal{R}$ with

$$\mathbf{A}^{(1)} = \begin{bmatrix} \boldsymbol{a}_1^{(1)} & \boldsymbol{a}_2^{(1)} & \boldsymbol{a}_3^{(1)} & 0 & 0 \end{bmatrix} \quad \mathbf{A}^{(2)} = \begin{bmatrix} \boldsymbol{a}_1^{(2)} & \boldsymbol{a}_2^{(2)} & 0 & \boldsymbol{a}_4^{(2)} & 0 \end{bmatrix}$$
$$\mathbf{A}^{(3)} = \begin{bmatrix} \boldsymbol{a}_1^{(3)} & \boldsymbol{a}_2^{(3)} & \boldsymbol{a}_3^{(3)} & 0 & \boldsymbol{a}_5^{(3)} \end{bmatrix}, \quad (63)$$

then the observation matrix must be of the form $C = \begin{bmatrix} \boldsymbol{c}_1 & \boldsymbol{c}_2 & 0 & 0 & 0 \end{bmatrix}$ to satisfy Criterion 3.1.

APPENDIX C
HDP-SLDS AND HDP-AR-HMM MESSAGE PASSING

In this appendix, we explore the computation of the backwards message passing and forward sampling scheme used for generating samples of the mode sequence $z_{1:T}$ and state sequence $\boldsymbol{x}_{1:T}$.

A. Mode Sequence Message Passing for Blocked Sampling

Consider a switching VAR(r) process. To derive the forward-backward procedure for jointly sampling the mode sequence $z_{1:T}$ given observations $\boldsymbol{y}_{1:T}$, plus r initial observations $\boldsymbol{y}_{1-r:0}$, we first note that the chain rule and Markov structure allows us to decompose the joint distribution as follows:

$$p(z_{1:T} \mid \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) = p(z_T \mid z_{T-1}, \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(z_{T-1} \mid z_{T-2}, \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \cdots p(z_2 \mid z_1, \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(z_1 \mid \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta}). \quad (64)$$

Thus, we may first sample z_1 from $p(z_1 \mid \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta})$, then condition on this value to sample z_2 from $p(z_2 \mid z_1, \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta})$, and so on. The conditional distribution of z_1 is derived as:

$$\begin{aligned} p(z_1 \mid \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) &\propto p(z_1) p(\boldsymbol{y}_1 \mid \theta_{z_1}, \boldsymbol{y}_{1-r:0}) \sum_{z_{2:T}} \prod_t p(z_t \mid \pi_{z_{t-1}}) p(\boldsymbol{y}_t \mid \theta_{z_t}, \boldsymbol{y}_{t-r:t-1}) \\ &\propto p(z_1) p(\boldsymbol{y}_1 \mid \theta_{z_1}, \boldsymbol{y}_{1-r:0}) \sum_{z_2} p(z_2 \mid \pi_{z_1}) p(\boldsymbol{y}_2 \mid \theta_{z_2}, \boldsymbol{y}_{2-r:1}) m_{3,2}(z_2) \\ &\propto p(z_1) p(\boldsymbol{y}_1 \mid \theta_{z_1}, \boldsymbol{y}_{1-r:0}) m_{2,1}(z_1), \end{aligned} \quad (65)$$

where $m_{t,t-1}(z_{t-1})$ is the backward message passed from z_t to z_{t-1} and is recursively defined by:

$$m_{t,t-1}(z_{t-1}) \propto \begin{cases} \sum_{z_t} p(z_t \mid \pi_{z_{t-1}}) p(\boldsymbol{y}_t \mid \theta_{z_t}, \boldsymbol{y}_{t-r:t-1}) m_{t+1,t}(z_t), & t \leq T; \\ 1, & t = T+1. \end{cases} \quad (66)$$

The general conditional distribution of z_t is:

$$p(z_t \mid z_{t-1}, \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto p(z_t \mid \pi_{z_{t-1}}) p(\boldsymbol{y}_t \mid \theta_{z_t}, \boldsymbol{y}_{t-r:t-1}) m_{t+1,t}(z_t). \quad (67)$$

For the HDP-AR-HMM, these distributions are given by:

$$\begin{aligned} p(z_t = k \mid z_{t-1}, \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) &\propto \pi_{z_{t-1}}(k) \mathcal{N}(\boldsymbol{y}_t; \sum_{i=1}^r A_i^{(k)} \boldsymbol{y}_{t-i}, \Sigma^{(k)}) m_{t+1,t}(k) \\ m_{t+1,t}(k) &= \sum_{j=1}^L \pi_k(j) \mathcal{N}(\boldsymbol{y}_{t+1}; \sum_{i=1}^r A_i^{(j)} \boldsymbol{y}_{t-i}, \Sigma^{(j)}) m_{t+2,t+1}(j) \\ m_{T+1,T}(k) &= 1 \quad k = 1, \ldots, L. \end{aligned} \quad (68)$$

B. State Sequence Message Passing for Blocked Sampling

A similar sampling scheme is used for generating samples of the state sequence $\boldsymbol{x}_{1:T}$. Although we now have a continuous state space, the computation of the backwards messages $m_{t+1,t}(\boldsymbol{x}_t)$ is still analytically feasible since we are working with Gaussian densities. Assume, $m_{t+1,t}(\boldsymbol{x}_t) \propto \mathcal{N}^{-1}(\boldsymbol{x}_t; \theta_{t+1,t}, \Lambda_{t+1,t})$, where $\mathcal{N}^{-1}(x; \theta, \Lambda)$ denotes a Gaussian distribution on x in information form with mean $\mu = \Lambda^{-1}\theta$ and covariance $\Sigma = \Lambda^{-1}$. Given a fixed mode sequence $z_{1:T}$, we simply have a time-varying linear dynamic system. The backwards messages for the HDP-SLDS can be recursively defined by

$$m_{t,t-1}(\boldsymbol{x}_{t-1}) \propto \int_{\mathcal{X}_t} p(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, z_t) p(\boldsymbol{y}_t | \boldsymbol{x}_t) m_{t+1,t}(\boldsymbol{x}_t) d\boldsymbol{x}_t. \quad (69)$$

For this model, the state transition density of Eq. (69) can be expressed as

$$\begin{aligned} p(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, z_t) &\propto \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_t - A^{(z_t)} \boldsymbol{x}_{t-1} - \boldsymbol{\mu}^{(z_t)})^T \Sigma^{-(z_t)} (\boldsymbol{x}_t - A^{(z_t)} \boldsymbol{x}_{t-1} - \boldsymbol{\mu}^{(z_t)}) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} A^{(z_t)^T} \Sigma^{-(z_t)} A^{(z_t)} & -A^{(z_t)^T} \Sigma^{-(z_t)} \\ -\Sigma^{-(z_t)} A^{(z_t)} & \Sigma^{-(z_t)} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix} \right. \\ &\qquad \left. + \begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} -A^{(z_t)^T} \Sigma^{-(z_t)} \boldsymbol{\mu}^{(z_t)} \\ \Sigma^{-(z_t)} \boldsymbol{\mu}^{(z_t)} \end{bmatrix} \right\}. \end{aligned} \quad (70)$$

We can similarly write the likelihood in exponentiated quadratic form

$$\begin{aligned} p(\boldsymbol{y}_t|\boldsymbol{x}_t) &\propto \exp\left\{ -\frac{1}{2}(\boldsymbol{y}_t - C\boldsymbol{x}_t)^T R^{-1}(\boldsymbol{y}_t - C\boldsymbol{x}_t)\right\} \\ &\propto \exp\left\{ -\frac{1}{2}\begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} 0 & 0 \\ 0 & C^T R^{-1} C \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} 0 \\ C^T R^{-1}\boldsymbol{y}_t \end{bmatrix} \right\}, \end{aligned} \tag{71}$$

as well as the messages

$$\begin{aligned} m_{t+1,t}(\boldsymbol{x}_t) &\propto \exp\left\{ -\frac{1}{2}\boldsymbol{x}_t^T \Lambda_{t+1,t}\boldsymbol{x}_t + \boldsymbol{x}_t^T \theta_{t+1,t}\right\} \\ &\propto \exp\left\{ -\frac{1}{2}\begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} 0 & 0 \\ 0 & \Lambda_{t+1,t} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} 0 \\ \theta_{t+1,t} \end{bmatrix} \right\}. \end{aligned} \tag{72}$$

The product of these quadratics is given by:

$$\begin{aligned} &p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, z_t)p(\boldsymbol{y}_t|\boldsymbol{x}_t)m_{t+1,t}(\boldsymbol{x}_t) \propto \\ &\exp\left\{ -\frac{1}{2}\begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} A^{(z_t)^T}\Sigma^{-(z_t)}A & -A^{(z_t)^T}\Sigma^{-(z_t)} \\ -\Sigma^{-(z_t)}A^{(z_t)} & \Sigma^{-(z_t)} + C^T R^{-1} C + \Lambda_{t+1,t} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix} \right. \\ &\qquad \left. + \begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} -A^{(z_t)^T}\Sigma^{-(z_t)}\boldsymbol{\mu}^{(z_t)} \\ C^T R^{-1}\boldsymbol{y}_t + \Sigma^{-(z_t)}\boldsymbol{\mu}^{(z_t)} + \theta_{t+1,t} \end{bmatrix} \right\} \end{aligned} \tag{73}$$

Using standard Gaussian marginalization identities we integrate over $\boldsymbol{x}_t$ to get,

$$m_{t,t-1}(\boldsymbol{x}_{t-1}) \propto \mathcal{N}^{-1}(\boldsymbol{x}_{t-1}; \theta_{t,t-1}, \Lambda_{t,t-1}), \tag{74}$$

where,

$$\begin{aligned} \theta_{t,t-1} &= -A^{(z_t)^T}\Sigma^{-(z_t)}\boldsymbol{\mu}^{(z_t)} + A^{(z_t)^T}\Sigma^{-(z_t)}(\Sigma^{-(z_t)} + C^T R^{-1} C + \Lambda_{t+1,t})^{-1}(C^T R^{-1}\boldsymbol{y}_t + \Sigma^{-(z_t)}\boldsymbol{\mu}^{(z_t)} + \theta_{t+1,t}) \\ \Lambda_{t,t-1} &= A^{(z_t)^T}\Sigma^{-(z_t)}A^{(z_t)} - A^{(z_t)^T}\Sigma^{-(z_t)}(\Sigma^{-(z_t)} + C^T R^{-1} C + \Lambda_{t+1,t})^{-1}\Sigma^{-(z_t)}A^{(z_t)}. \end{aligned} \tag{75}$$

The backwards message passing recursion is initialized with $m_{T+1,T} \sim \mathcal{N}^{-1}(\boldsymbol{x}_T; 0, 0)$. Let,

$$\begin{aligned} \Lambda^b_{t|t} &= C^T R^{-1} C + \Lambda_{t+1,t} \\ \theta^b_{t|t} &= C^T R^{-1}\boldsymbol{y}_t + \theta_{t+1,t}. \end{aligned} \tag{76}$$

Then we can define the following recursion, which we note is equivalent to a backwards running Kalman filter in information form,

$$\begin{aligned} \Lambda^b_{t-1|t-1} &= C^T R^{-1} C + A^{(z_t)^T}\Sigma^{-(z_t)}A^{(z_t)} - A^{(z_t)^T}\Sigma^{-(z_t)}(\Sigma^{-(z_t)} + C^T R^{-1} C + \Lambda_{t+1,t})^{-1}\Sigma^{-(z_t)}A^{(z_t)} \\ &= C^T R^{-1} C + A^{(z_t)^T}\Sigma^{-(z_t)}A^{(z_t)} - A^{(z_t)^T}\Sigma^{-(z_t)}(\Sigma^{-(z_t)} + \Lambda^b_{t|t})^{-1}\Sigma^{-(z_t)}A^{(z_t)} \\ \theta^b_{t-1|t-1} &= C^T R^{-1}\boldsymbol{y}_{t-1} - A^{(z_t)^T}\Sigma^{-(z_t)}\boldsymbol{\mu}^{(z_t)} + A^{(z_t)^T}\Sigma^{-(z_t)}(\Sigma^{-(z_t)} + C^T R^{-1} C + \Lambda_{t+1,t})^{-1} \\ &\qquad \cdot (C^T R^{-1}\boldsymbol{y}_t + \Sigma^{-(z_t)}\boldsymbol{\mu}^{(z_t)} + \theta_{t+1,t}) \\ &= C^T R^{-1}\boldsymbol{y}_{t-1} - A^{(z_t)^T}\Sigma^{-(z_t)}\boldsymbol{\mu}^{(z_t)} + A^{(z_t)^T}\Sigma^{-(z_t)}(\Sigma^{-(z_t)} + \Lambda^b_{t|t})^{-1}(\theta^b_{t|t} + \Sigma^{-(z_t)}\boldsymbol{\mu}^{(z_t)}) \end{aligned}$$

We initialize at time T with

$$\begin{aligned} \Lambda^b_{T|T} &= C^T R^{-1} C \\ \theta^b_{T|T} &= C^T R^{-1}\boldsymbol{y}_T \end{aligned} \tag{77}$$

An equivalent, but more numerically stable recursion is summarized in Algorithm 5.

After computing the messages $m_{t+1,t}(\boldsymbol{x}_t)$ backwards in time, we sample the state sequence $\boldsymbol{x}_{1:T}$ working forwards in time. As with the discrete mode sequence, one can decompose the posterior distribution of the state sequence as

$$\begin{aligned} p(\boldsymbol{x}_{1:T} \mid \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) = p(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}, \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) p(\boldsymbol{x}_{T-1} \mid \boldsymbol{x}_{T-2}, \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) \\ \cdots p(\boldsymbol{x}_2 \mid \boldsymbol{x}_1, \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) p(\boldsymbol{x}_1 \mid \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}). \end{aligned} \tag{78}$$

1) Initialize filter with

$$\Lambda^b_{T|T} = C^T R^{-1} C$$
$$\theta^b_{T|T} = C^T R^{-1} \boldsymbol{y}_T$$

2) Working backwards in time, for each $t \in \{T-1, \ldots, 1\}$:

a) Compute

$$\tilde{J}_{t+1} = \Lambda^b_{t+1|t+1}(\Lambda^b_{t+1|t+1} + \Sigma^{-(z_{t+1})})^{-1}$$
$$\tilde{L}_{t+1} = I - \tilde{J}_{t+1}.$$

b) Predict

$$\Lambda_{t+1,t} = A^{(z_{t+1})^T}(\tilde{L}_{t+1}\Lambda^b_{t+1|t+1}\tilde{L}^T_{t+1} + \tilde{J}_{t+1}\Sigma^{-(z_{t+1})}\tilde{J}^T_{t+1})A^{(z_{t+1})}$$
$$\theta_{t+1,t} = A^{(z_{t+1})^T}\tilde{L}_{t+1}(\theta^b_{t+1|t+1} - \Lambda^b_{t+1|t+1}\boldsymbol{\mu}^{(z_{t+1})})$$

c) Update

$$\Lambda^b_{t|t} = \Lambda_{t+1,t} + C^T R^{-1} C$$
$$\theta^b_{t|t} = \theta_{t+1,t} + C^T R^{-1} \boldsymbol{y}_t$$

3) Set

$$\Lambda^b_{0|0} = \Lambda_{1,0}$$
$$\theta^b_{0|0} = \theta_{1,0}$$

Algorithm 5: Numerically stable form of the backwards Kalman information filter.

where

$$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) \propto p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, A^{(z_t)}, \Sigma^{(z_t)}, \boldsymbol{\mu}^{(z_t)}) p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, R) m_{t+1,t}(\boldsymbol{x}_t). \tag{79}$$

For the HDP-SLDS, the product of these distributions is equivalent to

$$\begin{aligned} p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) &\propto \mathcal{N}(\boldsymbol{x}_t; A^{(z_t)}\boldsymbol{x}_{t-1} + \boldsymbol{\mu}^{(z_t)}, \Sigma^{(z_t)}) \mathcal{N}(\boldsymbol{y}_t; C\boldsymbol{x}_t, R) m_{t+1,t}(\boldsymbol{x}_t) \\ &\propto \mathcal{N}(\boldsymbol{x}_t; A^{(z_t)}\boldsymbol{x}_{t-1} + \boldsymbol{\mu}^{(z_t)}, \Sigma^{(z_t)}) \mathcal{N}^{-1}(\boldsymbol{x}_t; \theta^b_{t|t}, \Lambda^b_{t|t}) \\ &\propto \mathcal{N}^{-1}(\boldsymbol{x}_t; \Sigma^{-(z_t)}(A^{(z_t)}\boldsymbol{x}_{t-1} + \boldsymbol{\mu}^{(z_t)}) + \theta^b_{t|t}, \Sigma^{-(z_t)} + \Lambda^b_{t|t}), \end{aligned} \tag{80}$$

which is a simple Gaussian distribution so that the normalization constant is easily computed. Specifically, for each $t \in \{1, \ldots, T\}$ we sample $\boldsymbol{x}_t$ from

$$\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t; (\Sigma^{-(z_t)} + \Lambda^b_{t|t})^{-1}(\Sigma^{-(z_t)} A^{(z_t)} \boldsymbol{x}_{t-1} + \Sigma^{-(z_t)} \boldsymbol{\mu}^{(z_t)} + \theta^b_{t|t}), (\Sigma^{-(z_t)} + \Lambda^b_{t|t})^{-1}). \tag{81}$$

C. Mode Sequence Message Passing for Sequential Sampling

A similar sampling scheme to Carter and Kohn [37] is used for generating samples of the mode sequence $z_{1:T}$ having marginalized over the state sequence $\boldsymbol{x}_{1:T}$. Specifically, we sample z_t from:

$$\begin{aligned} p(z_t = k \mid z_{\setminus t}, \boldsymbol{y}_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) &\propto p(z_t = k \mid z_{\setminus t}, \boldsymbol{\pi}) p(\boldsymbol{y}_{1:T} \mid z_t = k, z_{\setminus t}) \\ &\propto \pi_{z_{t-1}}(k) \pi_k(z_{t+1}) p(\boldsymbol{y}_{1:T} \mid z_t = k, z_{\setminus t}). \end{aligned} \tag{82}$$

We omit the dependency on $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ for compactness. To compute the likelihood for each z_t, we combine forward and backward messages along with the local dynamics and measurements as follows:

$$\begin{aligned} p(\boldsymbol{y}_{1:T} \mid z_t = k, z_{\setminus t}) &\propto \int_{\mathcal{X}_{t-1}} \int_{\mathcal{X}_t} m_{t-2,t-1}(\boldsymbol{x}_{t-1}) p(\boldsymbol{y}_{t-1} \mid \boldsymbol{x}_{t-1}) p(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, z_t = k) \\ &\qquad p(\boldsymbol{y}_t | \boldsymbol{x}_t) m_{t+1,t}(\boldsymbol{x}_t) d\boldsymbol{x}_t d\boldsymbol{x}_{t-1} \qquad (83) \\ &\propto \int_{\mathcal{X}_t} \int_{\mathcal{X}_{t-1}} m_{t-2,t-1}(\boldsymbol{x}_{t-1}) p(\boldsymbol{y}_{t-1} \mid \boldsymbol{x}_{t-1}) p(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, z_t = k) d\boldsymbol{x}_{t-1} \\ &\qquad p(\boldsymbol{y}_t | \boldsymbol{x}_t) m_{t+1,t}(\boldsymbol{x}_t) d\boldsymbol{x}_t, \qquad (84) \end{aligned}$$

where the backwards messages are defined as in Appendix B and the forward messages by:

$$m_{t-1,t}(\boldsymbol{x}_t) \propto \int_{\mathcal{X}_{t-1}} p(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, z_t) p(\boldsymbol{y}_{t-1} | \boldsymbol{x}_{t-1}) m_{t-2,t-1}(\boldsymbol{x}_{t-1}) d\boldsymbol{x}_{t-1}. \qquad (85)$$

To derive the forward message passing recursions, assume that

$$m_{t-2,t-1}(\boldsymbol{x}_{t-1}) \propto \mathcal{N}^{-1}(\boldsymbol{x}_{t-1}; \theta_{t-2,t-1}, \Lambda_{t-2,t-1}) \qquad (86)$$

and z_t is known. The terms of the integrand of Eq. (85) can be written as:

$$\begin{aligned} p(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}, z_t) &= \mathcal{N}(\boldsymbol{x}_t; A^{(z_t)} \boldsymbol{x_{t-1}} + \boldsymbol{\mu}^{(z_t)}, \Sigma^{(z_t)}) \qquad (87) \\ &\propto \exp\left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix}^T \begin{bmatrix} \Sigma^{-(z_t)} & -\Sigma^{-(z_t)} A^{(z_t)} \\ -A^{(z_t)^T} \Sigma^{-(z_t)} & A^{(z_t)^T} \Sigma^{-(z_t)} A^{(z_t)} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix} \right. \\ &\qquad \left. + \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix}^T \begin{bmatrix} \Sigma^{-(z_t)} \boldsymbol{\mu}^{(z_t)} \\ -A^{(z_t)^T} \Sigma^{-(z_t)} \boldsymbol{\mu}^{(z_t)} \end{bmatrix} \right\} \end{aligned}$$

$$\begin{aligned} m_{t-2,t-1}(\boldsymbol{x}_{t-1}) p(\boldsymbol{y}_{t-1} | \boldsymbol{x}_{t-1}) &\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \Lambda^{-f}_{t-1|t-1} \theta^f_{t-1|t-1}, \Lambda^f_{t-1|t-1}) \qquad (88) \\ \propto \exp\left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix}^T \begin{bmatrix} 0 & 0 \\ 0 & \Lambda^f_{t-1|t-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix}^T \begin{bmatrix} 0 \\ \theta^f_{t-1|t-1} \end{bmatrix} \right\}, \end{aligned}$$

where, similar to the backwards recursions, we have made the following definitions

$$\begin{aligned} \theta^f_{t|t} &= \theta_{t-1,t} + C^T R^{-1} \boldsymbol{y}_t \\ \Lambda^f_{t|t} &= \Lambda_{t-1,t} + C^T R^{-1} C. \end{aligned} \qquad (89)$$

Combining these distributions and integrating over $\boldsymbol{x}_{t-1}$, we have

$$m_{t-1,t}(\boldsymbol{x}_t) \propto \mathcal{N}^{-1}(\boldsymbol{x}_t; \theta_{t-1,t}, \Lambda_{t-1,t}) \qquad (90)$$

with

$$\begin{aligned} \theta_{t-1,t} &= \Sigma^{-(z_t)} \boldsymbol{\mu}^{(z_t)} + \Sigma^{-(z_t)} A^{(z_t)} (A^{(z_t)^T} \Sigma^{-(z_t)} A^{(z_t)} + \Lambda^f_{t-1|t-1})^{-1} (\theta^f_{t-1|t-1} - A^{(z_t)^T} \Sigma^{-(z_t)} \boldsymbol{\mu}^{(z_t)}) \\ \Lambda_{t-1,t} &= \Sigma^{-(z_t)} - \Sigma^{-(z_t)} A^{(z_t)} (A^{(z_t)^T} \Sigma^{-(z_t)} A^{(z_t)} + \Lambda^f_{t-1|t-1})^{-1} A^{(z_t)^T} \Sigma^{-(z_t)}, \end{aligned}$$

or equivalently,

$$\begin{aligned} \theta_{t-1,t} &= \Lambda_{t-1,t} (\boldsymbol{\mu}^{(z_t)} + A^{(z_t)} \Lambda^{-f}_{t-1|t-1} \theta^f_{t-1|t-1}) \\ \Lambda_{t-1,t} &= (\Sigma^{(z_t)} + A^{(z_t)} \Lambda^{-f}_{t-1|t-1} A^{(z_t)^T})^{-1}. \end{aligned} \qquad (91)$$

Assuming $\boldsymbol{x}_0 \sim \mathcal{N}(0, P_0)$, we initialize at time $t = 0$ to

$$\begin{aligned} \theta_{-1,0} &= 0 \\ \Lambda_{-1,0} &= P_0^{-1}. \end{aligned} \qquad (92)$$

An equivalent, but more numerically stable recursion is summarized in Algorithm 6. However, this algorithm relies on the dynamic matrix $A^{(k)}$ being invertible.

1) Initialize filter with

$$\Lambda^b_{0|0} = P_0$$
$$\theta^b_{0|0} = 0$$

2) Working forwards in time, for each $t \in \{1, \ldots, T\}$:

a) Compute

$$M_t = A^{-(z_{t+1})^T} \Lambda^{-f}_{t|t} A^{-(z_{t+1})}$$
$$J_t = M_t (M_t + \Sigma^{-(z_{t+1})})^{-1}$$
$$L_t = I - J_t.$$

b) Predict

$$\Lambda_{t-1,t} = L_{t-1} M_{t-1} L^T_{t-1} + J_{t-1} \Sigma^{-(z_t)} J^T_{t-1}$$
$$\theta_{t-1,t} = L_{t-1} A^{-(z_t)^T} (\theta^f_{t-1|t-1} + \theta^f_{t-1|t-1} A^{-(z_t)} \boldsymbol{\mu}^{(z_t)})$$

c) Update

$$\Lambda^f_{t|t} = \Lambda_{t-1,t} + C^T R^{-1} C$$
$$\theta^f_{t|t} = \theta_{t-1,t} + C^T R^{-1} \boldsymbol{y}_t$$

Algorithm 6: Numerically stable form of the forward Kalman information filter.

We now return to the computation of the likelihood of Eq. (84). We note that the integral over $\boldsymbol{x}_{t-1}$ is equivalent to computing the message $m_{t-1,t}(\boldsymbol{x}_t)$ using $z_t = k$. However, we have to be careful that any constants that were previously ignored in this message passing are not a function of z_t. For the meantime, let us assume that there exists such a constant and let us denote this special message by

$$m_{t-1,t}(\boldsymbol{x}_t; z_t) \propto c(z_t) \mathcal{N}^{-1}(\boldsymbol{x}_t; \theta_{t-1,t}(z_t), \Lambda_{t-1,t}(z_t)). \tag{93}$$

Then, the likelihood can be written as

$$p(\boldsymbol{y}_{1:T} \mid z_t = k, z_{\setminus t}) \propto \int_{\mathcal{X}_t} m_{t-1,t}(\boldsymbol{x}_t; z_t = k) p(\boldsymbol{y}_t | \boldsymbol{x}_t) m_{t+1,t}(\boldsymbol{x}_t) d\boldsymbol{x}_t \tag{94}$$
$$\propto \int_{\mathcal{X}_t} c(k) \mathcal{N}^{-1}(\boldsymbol{x}_t; \theta_{t-1,t}(k), \Lambda_{t-1,t}(k)) \mathcal{N}^{-1}(\boldsymbol{x}_t; \theta^b_{t|t}, \Lambda^b_{t|t}) d\boldsymbol{x}_t \tag{95}$$

Combining the information parameters, and maintaining the term in the normalizing constant that is a function of k, this is equivalent to

$$p(\boldsymbol{y}_{1:T} \mid z_t = k, z_{\setminus t}) \propto c(k) |\Lambda_{t-1,t}(k)|^{1/2} \exp\left(-\frac{1}{2} \theta_{t-1,t}(k)^T \Lambda_{t-1,t}(k)^{-1} \theta_{t-1,t}(k)\right)$$
$$\int_{\mathcal{X}_t} \exp\left(-\frac{1}{2} \boldsymbol{x}_t^T (\Lambda_{t-1,t}(k) + \Lambda^b_{t|t}) \boldsymbol{x}_t + \boldsymbol{x}_t^T (\theta_{t-1,t}(k) + \theta^b_{t|t})\right) d\boldsymbol{x}_t \tag{96}$$

To compute this integral, we write the integrand in terms of a Gaussian distribution times a constant. The integral

is then simply that constant term:

$$
\begin{aligned}
p(\boldsymbol{y}_{1:T} \mid z_t = k, z_{\setminus t}) &\propto c(k)|\Lambda_{t-1,t}(k)|^{1/2} \exp\left(-\frac{1}{2}\theta_{t-1,t}(k)^T \Lambda_{t-1,t}(k)^{-1}\theta_{t-1,t}(k)\right) \\
&\quad |\Lambda_{t-1,t}(k) + \Lambda^b_{t|t}|^{-1/2} \exp\left(\frac{1}{2}(\theta_{t-1,t}(k) + \theta^b_{t|t})^T (\Lambda_{t-1,t}(k) + \Lambda^b_{t|t})^{-1}(\theta_{t-1,t}(k) + \theta^b_{t|t})\right) \\
&\quad \int_{\mathcal{X}_t} \mathcal{N}^{-1}(\boldsymbol{x}_t; \theta_{t-1,t}(k) + \theta^b_{t|t}, \Lambda_{t-1,t}(k) + \Lambda^b_{t|t}) d\boldsymbol{x}_t \\
&\propto c(k) \frac{|\Lambda_{t-1,t}(k)|^{1/2}}{|\Lambda_{t-1,t}(k) + \Lambda^b_{t|t}|^{1/2}} \\
&\quad \exp\Bigg(-\frac{1}{2}\theta_{t-1,t}(k)^T \Lambda_{t-1,t}(k)^{-1}\theta_{t-1,t}(k) \\
&\qquad + \frac{1}{2}(\theta_{t-1,t}(k) + \theta^b_{t|t})^T (\Lambda_{t-1,t}(k) + \Lambda^b_{t|t})^{-1}(\theta_{t-1,t}(k) + \theta^b_{t|t})\Bigg)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
p(z_t = k \mid z_{\setminus t}, \boldsymbol{y}_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto \pi_{z_{t-1}}(k)\pi_k(z_{t+1})c(k)|\Lambda^{(k)}_t|^{1/2}|\Lambda^{(k)}_t + \Lambda^b_{t|t}|^{-1/2} \\
\exp\left(-\frac{1}{2}\theta^{(k)^T}_t \Lambda^{-(k)}_t \theta^{(k)}_t + \frac{1}{2}(\theta^{(k)}_t + \theta^b_{t|t})^T (\Lambda^{(k)}_t + \Lambda^b_{t|t})^{-1}(\theta^{(k)}_t + \theta^b_{t|t})\right) \quad (97)
\end{aligned}
$$

We now show that $c(z_t)$ is not a function z_t. The only place where the previously ignored dependency on z_t arises is from $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, z_t)$. Namely,

$$
\begin{aligned}
p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, z_t) &= \frac{\exp(-\frac{1}{2}\boldsymbol{\mu}^{(z_t)^T}\Sigma^{-(z_t)}\boldsymbol{\mu}^{(z_t)})}{|\Sigma^{(z_t)}|^{1/2}} \cdot \text{exponential}_1 \\
&= c_1(z_t) \cdot \text{exponential}_1 \quad (98)
\end{aligned}
$$

where exponential_1 is the exponentiated quadratic of Eq. (87). Then, when compute the message $m_{t-1,t}(\boldsymbol{x}_t; z_t)$ we update the previous message $m_{t-2,t-1}(\boldsymbol{x}_{t-1})$ by incorporating the local likelihood $p(\boldsymbol{y}_{t-1} \mid \boldsymbol{x}_{t-1})$ and then propagating the state estimate with $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, z_t)$ and integrating over $\boldsymbol{x}_{t-1}$. Namely, we combine the distribution of Eq. (98) with the exponentiated quadratic of Eq. (88) and integrate over $\boldsymbol{x}_{t-1}$:

$$
m_{t-1,t}(\boldsymbol{x}_t; z_t) \propto c_1(z_t) \int_{\mathcal{X}_{t-1}} \text{exponential}_1 \cdot \text{exponential}_2 d\boldsymbol{x}_{t-1}, \quad (99)
$$

where exponential_2 is the exponentiated quadratic of Eq. (88).

Since $m_{t-2,t-1}(\boldsymbol{x}_{t-1}) \propto p(\boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{1:t-2}, z_{1:t-1})$, and the Markov properties of the state space model dictate

$$
\begin{aligned}
p(\boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{1:t-1}, z_{1:t-1}) &= p(\boldsymbol{y}_{t-1}|\boldsymbol{x}_{t-1}) p(\boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{1:t-2}, z_{1:t-1}) \\
&\propto p(\boldsymbol{y}_{t-1}|\boldsymbol{x}_{t-1}) m_{t-2,t-1}(\boldsymbol{x}_{t-1}), \quad (100)
\end{aligned}
$$

then

$$
p(\boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{1:t-1}, z_{1:t-1}) = c_2 \cdot \text{exponential}_2.
$$

We note that the normalizing constant c_2 is not a function of z_t since we have only considered z_τ for $\tau < t$.

Once again exploiting the conditional independencies induced by the Markov structure of our state space model, and plugging in Eq. (98) and Eq. (101),

$$
\begin{aligned}
p(\boldsymbol{x}_t, \boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{1:t-1}, z_{1:t}) &= p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_{t-1}, z_t) p(\boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{1:t-1}, z_{1:t-1}) \\
&= (c_1(z_t) \cdot \text{exponential}_1)(c_2 \cdot \text{exponential}_2) \\
&= c_1(z_t)c_2 \cdot \text{exponential}_1 \cdot \text{exponential}_2. \quad (101)
\end{aligned}
$$

Plugging this results into Eq. (99), we have

$$\begin{aligned} m_{t-1,t}(\boldsymbol{x}_t; z_t) &\propto c_1(z_t) \int_{\mathcal{X}_{t-1}} \frac{1}{c_1(z_t)c_2} p(\boldsymbol{x}_t, \boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{1:t-1}, z_{1:t}) d\boldsymbol{x}_{t-1} \\ &\propto \frac{1}{c_2} p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1}, z_{1:t}). \end{aligned} \tag{102}$$

Comparing Eq. (102) to Eq. (93), and noting that

$$p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1}, z_{1:t}) = \mathcal{N}^{-1}(\boldsymbol{x}_t; \theta_{t-1,t}(z_t), \Lambda_{t-1,t}(z_t)),$$

we see that $c(z_t) = \frac{1}{c_2}$ and is thus not a function of z_t.

Algebraically, we could derive this result as follows.

$$\begin{aligned} m_{t-1,t}(\boldsymbol{x}_t; z_t) &\propto c_1(z_t) \int_{\mathcal{X}_{t-1}} \text{exponential}_1 \cdot \text{exponential}_2 d\boldsymbol{x}_{t-1} \\ &= c_1(z_t)c_3(z_t) \int_{\mathcal{X}_{t-1}} \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}; \boldsymbol{\Lambda}(z_t)^{-1}\boldsymbol{\theta}(z_t), \boldsymbol{\Lambda}(z_t) \right) d\boldsymbol{x}_{t-1}, \end{aligned} \tag{103}$$

where $\boldsymbol{\theta}(z_t)$ and $\boldsymbol{\Lambda}(z_t)$ are the information parameters determined by combining the functional forms of exponential_1 and exponential_2, and

$$c_1(z_t)c_3(z_t) = \frac{\exp\{-\frac{1}{2}\boldsymbol{\mu}^{(z_t)^T}\Sigma^{-(z_t)}\boldsymbol{\mu}^{(z_t)}\}}{|\Sigma^{(z_t)}|^{1/2}} \frac{\exp\{\frac{1}{2}\boldsymbol{\theta}(z_t)^T\boldsymbol{\Lambda}(z_t)^{-1}\boldsymbol{\theta}(z_t)\}}{|\boldsymbol{\Lambda}(z_t)|^{1/2}}. \tag{104}$$

Computing these terms in parts, and using standard linear algebra properties of block matrices,

$$\begin{aligned} |\boldsymbol{\Lambda}(z_t)| &= |\Sigma^{-(z_t)}||(A^{(z_t)^T}\Sigma^{-(z_t)}A^{(z_t)} + \Lambda^f_{t-1|t-1}) - A^{(z_t)^T}\Sigma^{-(z_t)}A^{(z_t)}| \\ &= |\Sigma^{-(z_t)}||\Lambda^f_{t-1|t-1}| \end{aligned} \tag{105}$$

$$\begin{aligned} \boldsymbol{\Lambda}(z_t)^{-1} &= \begin{bmatrix} (\Sigma^{-(z_t)} - \Sigma^{-(z_t)}A^{(z_t)}\tilde{\boldsymbol{\Lambda}}(z_t)^{-1}A^{(z_t)^T}\Sigma^{-(z_t)})^{-1} & A^{(z_t)}\Lambda^f_{t-1|t-1} \\ \Lambda^f_{t-1|t-1}A^{(z_t)^T} & (\tilde{\boldsymbol{\Lambda}}(z_t) - A^{(z_t)^T}\Sigma^{-(z_t)}A^{(z_t)})^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \Sigma^{(z_t)} + A^{(z_t)}\Lambda^{-f}_{t-1|t-1}A^{(z_t)^T} & A^{(z_t)}\Lambda^f_{t-1|t-1} \\ \Lambda^f_{t-1|t-1}A^{(z_t)^T} & \Lambda^{-f}_{t-1|t-1} \end{bmatrix}, \end{aligned} \tag{106}$$

where $\tilde{\boldsymbol{\Lambda}}(z_t) = (A^{(z_t)^T}\Sigma^{-(z_t)}A^{(z_t)} + \Lambda^f_{t-1|t-1})$ and we have used the matrix inversion lemma in obtaining the last equality. Using this form of $\boldsymbol{\Lambda}(z_t)^{-1}$, we readily obtain

$$\boldsymbol{\theta}(z_t)^T\boldsymbol{\Lambda}(z_t)^{-1}\boldsymbol{\theta}(z_t) = \boldsymbol{\mu}^{(z_t)}\Sigma^{-(z_t)}\boldsymbol{\mu}^{(z_t)} + \theta^{f^T}_{t-1|t-1}\Lambda^{-f}_{t-1|t-1}\theta^f_{t-1|t-1}. \tag{107}$$

Thus,

$$c_1(z_t)c_3(z_t) = \frac{\exp\{\frac{1}{2}\theta^{f^T}_{t-1|t-1}\Lambda^{-f}_{t-1|t-1}\theta^f_{t-1|t-1}\}}{|\Lambda^f_{t-1|t-1}|^{1/2}}, \tag{108}$$

which does not depend upon the value of z_t.

APPENDIX D
DERIVATION OF MANEUVERING TARGET TRACKING SAMPLER

In this Appendix we derive the maneuvering target tracking (MTT) sampler outlined in Sec. V-B. Recall the MTT model of Eq. (42). As described in Sec. V-B, we are interested in jointly sampling the control input and dynamical mode $(\boldsymbol{u}_t, z_t)$, marginalizing over the state sequence $\boldsymbol{x}_{1:T}$, the transition distributions $\boldsymbol{\pi}$, and the dynamic parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}^{(k)}, \Sigma^{(k)}\}$. One can factor the desired conditional distribution factorizes as,

$$p(\boldsymbol{u}_t, z_t | z_{\setminus t}, \boldsymbol{u}_{\setminus t}, \boldsymbol{y}_{1:T}, \beta, \alpha, \kappa, \lambda) = p(z_t | z_{\setminus t}, \boldsymbol{u}_{\setminus t}, \boldsymbol{y}_{1:T}, \beta, \alpha, \kappa, \lambda) p(\boldsymbol{u}_t | z_{1:T}, \boldsymbol{u}_{\setminus t}, \boldsymbol{y}_{1:T}, \lambda). \tag{109}$$

The distribution in Eq.(109) is a hybrid distribution: each discrete value of the dynamical mode indicator variable z_t corresponds to a different continuous distribution on the control input $\boldsymbol{u}_t$. We analyze each of the conditional distributions of Eq. (109) by considering the joint distribution on all of the model parameters, and then marginalizing $\boldsymbol{x}_{1:T}$, $\boldsymbol{\pi}$, and θ_k. (Note that marginalization over θ_j for $j \neq k$ simply results in a constant.)

$$p(z_t = k | z_{\setminus t}, \boldsymbol{u}_{\setminus t}, \boldsymbol{y}_{1:T}, \beta, \alpha, \kappa, \lambda) \propto \int_{\boldsymbol{\pi}} \prod_j p(\pi_j | \beta, \alpha, \kappa) \prod_\tau p(z_\tau | \pi_{z_{\tau-1}}) d\pi$$
$$\int_{\mathcal{U}_t} \int p(\theta_k \mid \lambda) \prod_{\tau | z_\tau = k} p(\boldsymbol{u}_\tau | \theta_k) d\theta_k \int_{\mathcal{X}} \prod_\tau p(\boldsymbol{x}_\tau | \boldsymbol{x}_{\tau-1}, \boldsymbol{u}_\tau) p(\boldsymbol{y}_\tau | \boldsymbol{x}_\tau) d\boldsymbol{x}_{1:T} d\boldsymbol{u}_t. \quad (110)$$

Similarly, we can write the conditional density of $\boldsymbol{u}_t$ for each candidate z_t as,

$$p(\boldsymbol{u}_t | z_t = k, z_{\setminus t}, \boldsymbol{u}_{\setminus t}, \boldsymbol{y}_{1:T}, \lambda) \propto \int p(\theta_k \mid \lambda) \prod_{\tau | z_\tau = k} p(\boldsymbol{u}_\tau | \theta_k) d\theta_k \int_{\mathcal{X}} \prod_\tau p(\boldsymbol{x}_\tau | \boldsymbol{x}_{\tau-1}, \boldsymbol{u}_\tau) p(\boldsymbol{y}_\tau | \boldsymbol{x}_\tau) d\boldsymbol{x}_{1:T}. \quad (111)$$

A key step in deriving these conditional distributions is the marginalization of the state sequence $\boldsymbol{x}_{1:T}$. In performing this marginalization, one thing we harness is the fact that conditioning on the control input sequence simplifies the SLDS to an LDS with a deterministic control input $\boldsymbol{u}_{1:T}$. Thus, conditioning on $\boldsymbol{u}_{1:t-1,t+1:T}$ allows us to marginalize the state sequence in the following manner. We run a forward Kalman filter to pass a message from $t-2$ to $t-1$, which is updated by the local likelihood at $t-1$. A backward filter is also run to pass a message from $t+1$ to t, which is updated by the local likelihood at t. These updated messages are combined with the local dynamic $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{u}_t, \boldsymbol{\theta})$ and then marginalized over $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$, resulting in the likelihood of the observation sequence $\boldsymbol{y}_{1:T}$ as a function of $\boldsymbol{u}_t$, the variable of interest. Because the sampler conditions on control inputs, the filter for this time-invariant system can be efficiently implemented by pre-computing the error covariances and then solely computing local Kalman updates at every time step. Of note is that the computational complexity is linear in the training sequence length, as well as the number of currently instantiated maneuver modes. In the following sections, we evaluate each of the integrals of Eq. (111) and Eq. (110) in turn.

A. Chinese Restaurant Franchise

The integration over $\boldsymbol{\pi}$ appearing in the first line of Eq. (110) results in exactly the same predictive distribution as the sticky HDP-HMM [14].

B. Normal-Inverse-Wishart Posterior Update

The marginalization of θ_k, appearing both in Eq. (110) and Eq. (111), can be rewritten as follows:

$$\begin{aligned}\int p(\theta_k|\lambda) \prod_{\tau|z_\tau=k} p(\boldsymbol{u}_\tau|\theta_k) d\theta_k &= \int p(\boldsymbol{u}_t|\theta_k) p(\theta_k|\lambda) \prod_{\tau|z_\tau=k, \tau\neq t} p(\boldsymbol{u}_\tau|\theta_k) d\theta_k \\ &\propto \int p(\boldsymbol{u}_t|\theta_k) p(\theta_k|\{\boldsymbol{u}_\tau|z_\tau = k, \tau \neq t\}, \lambda) d\theta_k \\ &= p(\boldsymbol{u}_t|\{\boldsymbol{u}_\tau|z_\tau = k, \tau \neq t\}, \lambda). \end{aligned} \quad (112)$$

Here, the set $\{\boldsymbol{u}_\tau | z_\tau = k, \tau \neq t\}$ denotes all the observations $\boldsymbol{u}_\tau$ other than $\boldsymbol{u}_t$ that were drawn from the Gaussian parameterized by θ_k. When θ_k has a normal-inverse-Wishart prior $\mathcal{NIW}(\kappa, \boldsymbol{\vartheta}, \nu, \Delta)$, standard conjugacy results imply that the posterior is:

$$p(\boldsymbol{u}_t|\{\boldsymbol{u}_\tau|z_\tau = k, \tau \neq t\}, \kappa, \boldsymbol{\vartheta}, \nu, \Delta) \simeq \mathcal{N}\left(\boldsymbol{u}_t; \bar{\boldsymbol{\vartheta}}, \frac{(\bar{\kappa}+1)\bar{\nu}}{\bar{\kappa}(\bar{\nu}-d-1)}\bar{\Delta}\right) \triangleq \mathcal{N}(\boldsymbol{u}_t; \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k), \quad (113)$$

where

$$\begin{aligned} \bar{\kappa} &= \kappa + |\{\boldsymbol{u}_s | z_s = k, s \neq t\}| \\ \bar{\nu} &= \nu + |\{\boldsymbol{u}_s | z_s = k, s \neq t\}| \\ \bar{\kappa}\bar{\boldsymbol{\vartheta}} &= \kappa\boldsymbol{\vartheta} + \sum_{\boldsymbol{u}_s \in \{\boldsymbol{u}_s | z_s = k, s \neq t\}} \boldsymbol{u}_s \\ \bar{\nu}\bar{\Delta} &= \nu\Delta + \sum_{\boldsymbol{u}_s \in \{\boldsymbol{u}_s | z_s = k, s \neq t\}} \boldsymbol{u}_s\boldsymbol{u}_s^T + \kappa\boldsymbol{\vartheta}\boldsymbol{\vartheta}^T - \bar{\kappa}\bar{\boldsymbol{\vartheta}}\bar{\boldsymbol{\vartheta}}^T \end{aligned} \quad (114)$$

Here, we are using the moment-matched Gaussian approximation to the Student-t predictive distribution for $\boldsymbol{u}_t$ induced by marginalizing θ_k.

C. Marginalization by Message Passing

When considering the control input $\boldsymbol{u}_t$ and conditioning on the values of all $\boldsymbol{u}_\tau$, $\tau \neq t$, the marginalization over all states $\boldsymbol{x}_{1:T}$ can be equated to a message passing scheme that relies on the conditionally linear dynamical system induced by fixing $\boldsymbol{u}_\tau$, $\tau \neq t$. Specifically,

$$\begin{aligned}
&\int_{\mathcal{X}} \prod_{\tau} p(\boldsymbol{x}_\tau|\boldsymbol{x}_{\tau-1},\boldsymbol{u}_\tau)p(\boldsymbol{y}_\tau|\boldsymbol{x}_\tau)dx \\
&\quad\propto \int_{\mathcal{X}_{t-1}}\int_{\mathcal{X}_t} m_{t-1,t-2}(\boldsymbol{x}_{t-1})p(\boldsymbol{y}_{t-1}|\boldsymbol{x}_{t-1})p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{u}_t)p(\boldsymbol{y}_t|\boldsymbol{x}_t)m_{t,t+1}(\boldsymbol{x}_t)dx_t dx_{t-1} \\
&\quad\propto p(\boldsymbol{y}_{1:T}|\boldsymbol{u}_t;\boldsymbol{u}_{\setminus t}),
\end{aligned} \tag{115}$$

where we recall the definitions of the forward messages $m_{t-1,t}(\boldsymbol{x}_t)$ and backward messages $m_{t+1,t}(\boldsymbol{x}_t)$ from Appendix B. For our MTT model of Eq. (42), however, instead of accounting for a process noise mean $\boldsymbol{\mu}^{(z_\tau)}$ at time τ in the filtering equations, we must account for the control input $\boldsymbol{u}_\tau$. Conditioning on $\boldsymbol{u}_\tau$, one can equate $B\boldsymbol{u}_\tau$ with a process noise mean, and thus we simply replace $\boldsymbol{\mu}^{(z_\tau)}$ with $B\boldsymbol{u}_\tau$ in the filtering equations of Appendix B. Similarly, we replace the process noise covariance term $\Sigma^{(z_\tau)}$ with our process noise covariance Q. (Note that although $\boldsymbol{u}_\tau(z_\tau) \sim \mathcal{N}(\boldsymbol{\mu}^{(z_\tau)},\Sigma^{(z_\tau)})$, we condition on the value $\boldsymbol{u}_\tau$ so that the MTT parameters $\{\boldsymbol{\mu}^{(z_\tau)},\Sigma^{(z_\tau)}\}$ do not factor into the message passing equations.)

D. Combining Messages

To compute the likelihood of Eq. (115), we take the filtered estimates of $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$, combine them with the local dynamics and local likelihood, and marginalize over $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$. To aid in this computation, we consider the exponentiated quadratic form of each term in the integrand of Eq. (115). We then join these terms and use standard Gaussian integration formulas to arrive at the desired likelihood. The derivation of this likelihood greatly parallels that for the sequential mode sequence sampler of Appendix C.

Recall the forward filter recursions of Appendix B in terms of information parameters

$$\{\theta_{t-1,t},\Lambda_{t-1,t},\theta^f_{t|t},\Lambda^f_{t|t}\},$$

and the backward filter recursions in terms of

$$\{\theta_{t+1,t},\Lambda_{t+1,t},\theta^b_{t|t},\Lambda^b_{t|t}\}.$$

Replace $\boldsymbol{\mu}^{(z_t)}$ with $B\boldsymbol{u}_t$ and $\Sigma^{(z_t)}$ with Q where appropriate. We many then write $m_{t,t+1}(\boldsymbol{x}_t)$ updated with the likelihood $p(\boldsymbol{y}_{t-1}|\boldsymbol{x}_{t-1})$ in exponentiated quadratic form as:

$$\begin{aligned}
&m_{t-1,t-2}(\boldsymbol{x}_{t-1})p(\boldsymbol{y}_{t-1}|\boldsymbol{x}_{t-1}) \\
&\qquad\propto \exp\left\{ -\frac{1}{2}\begin{bmatrix}\boldsymbol{x}_{t-1}\\ \boldsymbol{x}_t\end{bmatrix}^T \begin{bmatrix} C^TR^{-1}C+\Lambda_{t-1,t-2} & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_{t-1}\\ \boldsymbol{x}_t\end{bmatrix} \right. \\
&\qquad\qquad\left. +\begin{bmatrix}\boldsymbol{x}_{t-1}\\ \boldsymbol{x}_t\end{bmatrix}^T \begin{bmatrix} C^TR^{-1}\boldsymbol{y}_{t-1}+\theta_{t-1,t-2}\\ 0\end{bmatrix}\right\}.
\end{aligned}$$

The local dynamics can similarly be written as

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{u}_t) \propto \exp\left\{ -\frac{1}{2}\begin{bmatrix}\boldsymbol{u}_t\\ \boldsymbol{x}_{t-1}\\ \boldsymbol{x}_t\end{bmatrix}^T \begin{bmatrix} B^TQ^{-1}B & B^TQ^{-1}A & -B^TQ^{-1}\\ A^TQ^{-1}B & A^TQ^{-1}A & -A^TQ^{-1}\\ -Q^{-1}B & -Q^{-1}A & Q^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{u}_t\\ \boldsymbol{x}_{t-1}\\ \boldsymbol{x}_t\end{bmatrix} + \begin{bmatrix}\boldsymbol{u}_t\\ \boldsymbol{x}_{t-1}\\ \boldsymbol{x}_t\end{bmatrix}^T \begin{bmatrix}0\\0\\0\end{bmatrix}\right\}.$$

Finally, the backward message $m_{t,t+1}(\boldsymbol{x}_t)$ updated with the likelihood $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ can be written as

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t)m_{t,t+1}(\boldsymbol{x}_t) \propto \exp\left\{ -\frac{1}{2}\begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} 0 & 0 \\ 0 & C^TR^{-1}C + \Lambda_{t,t+1} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix} + \begin{bmatrix} \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} 0 \\ C^TR^{-1}\boldsymbol{y}_t + \theta_{t,t+1} \end{bmatrix} \right\}.$$

Using the definitions

$$\begin{aligned} \Lambda^b_{t|t} &= C^TR^{-1}C + \Lambda_{t+1,t} \\ \theta^b_{t|t} &= C^TR^{-1}\boldsymbol{y}_t + \theta_{t+1,t} \\ \Lambda^f_{t|t} &= C^TR^{-1}C + \Lambda_{t-1,t} \\ \theta^f_{t|t} &= C^TR^{-1}\boldsymbol{y}_t + \theta_{t-1,t}, \end{aligned}$$

we may express the entire integrand as

$$\begin{aligned} &m_{t-1,t-2}(\boldsymbol{x}_{t-1})p(\boldsymbol{y}_{t-1}|\boldsymbol{x}_{t-1})p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{u}_t)p(\boldsymbol{y}_t|\boldsymbol{x}_t)m_{t,t+1}(\boldsymbol{x}_t) \propto \\ &\exp\left\{ -\frac{1}{2}\begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} B^TQ^{-1}B & B^TQ^{-1}A & -B^TQ^{-1} \\ A^TQ^{-1}B & A^TQ^{-1}A + \Lambda^f_{t-1|t-1} & -A^TQ^{-1} \\ -Q^{-1}B & -Q^{-1}A & Q^{-1} + \Lambda^b_{t|t} \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix} \right. \\ &\qquad \left. + \begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{x}_{t-1} \\ \boldsymbol{x}_t \end{bmatrix}^T \begin{bmatrix} 0 \\ \theta^f_{t-1|t-1} \\ \theta^b_{t|t} \end{bmatrix} \right\} \end{aligned}$$

Integrating over $\boldsymbol{x}_t$, we obtain an expression proportional to

$$\mathcal{N}^{-1}\left(\begin{bmatrix} \boldsymbol{u}_t^T \\ \boldsymbol{x}_{t-1} \end{bmatrix}; \theta\left(\begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix}\right), \Lambda\left(\begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix}\right)\right),$$

with

$$\begin{aligned} \Lambda\left(\begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix}\right) &= \begin{bmatrix} B^TQ^{-1}B & B^TQ^{-1}A \\ A^TQ^{-1}B & A^TQ^{-1}A + \Lambda^f_{t-1|t-1} \end{bmatrix} - \begin{bmatrix} B^TQ^{-1} \\ A^TQ^{-1} \end{bmatrix} (Q^{-1} + \Lambda^b_{t|t})^{-1} \begin{bmatrix} Q^{-1}B & Q^{-1}A \end{bmatrix} \\ &= \begin{bmatrix} B^T\Sigma_t^{-1}B & B^T\Sigma_t^{-1}A \\ A^T\Sigma_t^{-1}B & A^T\Sigma_t^{-1}A \end{bmatrix} \\ \theta\left(\begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix}\right) &= \begin{bmatrix} 0 \\ \theta^f_{t-1|t-1} \end{bmatrix} + \begin{bmatrix} B^TQ^{-1} \\ A^TQ^{-1} \end{bmatrix} (Q^{-1} + \Lambda^b_{t|t})^{-1}\theta^b_{t|t} = \begin{bmatrix} B^TQ^{-1}K_t^{-1}\theta^b_{t|t} \\ \theta^f_{t-1|t-1} + A^TQ^{-1}K_t^{-1}\theta^b_{t|t} \end{bmatrix}. \end{aligned}$$

Here, we have defined

$$\Sigma_t = Q^{-1} + Q^{-1}(Q^{-1} + \Lambda^b_{t|t})^{-1}Q^{-1} = Q^{-1} + Q^{-1}K_t^{-1}Q^{-1}.$$

Finally, integrating over $\boldsymbol{x}_{t-1}$ yields an expression proportional to

$$\mathcal{N}^{-1}(\boldsymbol{u}_t^T; \theta(\boldsymbol{u}_t), \Lambda(\boldsymbol{u}_t)),$$

with

$$\begin{aligned} \Lambda(\boldsymbol{u}_t) &= B^T\Sigma_t^{-1}B - B^T\Sigma_t^{-1}A(A^T\Sigma_t^{-1}A + \Lambda^f_{t-1|t-1})^{-1}A^T\Sigma_t^{-1}B \\ \theta(\boldsymbol{u}_t) &= B^TQ^{-1}K_t^{-1}\theta^b_{t|t} - B^T\Sigma_t^{-1}A(A^T\Sigma_t^{-1}A + \Lambda^f_{t-1|t-1})^{-1}(\theta^f_{t-1|t-1} + A^TQ^{-1}K_t^{-1}\theta^b_{t|t}). \end{aligned}$$

E. Joining Distributions that Depend on $\boldsymbol{u}_t$

We have derived two terms which depend on $\boldsymbol{u}_t$: a prior and a likelihood. Normally, one would consider $p(\boldsymbol{u}_t|\theta_k)$ the prior on $\boldsymbol{u}_t$. However, through marginalization of this parameter, we induced dependencies between the control inputs $\boldsymbol{u}_\tau$ and all the $\boldsymbol{u}_\tau$ that were drawn from a distribution parameterized by θ_k inform us of the distribution over $\boldsymbol{u}_t$. Therefore, we treat $p(\boldsymbol{u}_t|\{\boldsymbol{u}_\tau|z_\tau = k, \tau \neq t\})$ as a prior distribution on $\boldsymbol{u}_t$. The likelihood function $p(\boldsymbol{y}_{1:T}|\boldsymbol{u}_t; \boldsymbol{u}_{\setminus t})$ describes the likelihood of an observation sequence $\boldsymbol{y}_{1:T}$ given the input sequence $\boldsymbol{u}_{1:T}$, containing the random variable is $\boldsymbol{u}_t$.

We multiply the prior distribution by the likelihood function to get the following quadratic expression:

$$\begin{aligned}
&p(\boldsymbol{u}_t|\{\boldsymbol{u}_\tau|z_\tau = k, \tau \neq t\})p(\boldsymbol{y}_{1:T}|\boldsymbol{u}_t; \boldsymbol{u}_{\setminus t}) \\
&\quad \propto \frac{1}{(2\pi)^{N/2}|\hat{\Sigma}_k|^{1/2}} \exp\Bigg\{ -\frac{1}{2}(\boldsymbol{u}_t - \hat{\boldsymbol{\mu}}_k)^T\hat{\Sigma}_k^{-1}(\boldsymbol{u}_t - \hat{\boldsymbol{\mu}}_k) \\
&\qquad\qquad - \frac{1}{2}(\boldsymbol{u}_t - \Lambda(\boldsymbol{u}_t)^{-1}\theta(\boldsymbol{u}_t))^T\Lambda(\boldsymbol{u}_t)(\boldsymbol{u}_t - \Lambda_t y^{-1}\theta(\boldsymbol{u}_t))\Bigg\} \\
&\quad = \frac{1}{(2\pi)^{N/2}|\hat{\Sigma}_k|^{1/2}} \exp\Bigg\{ -\frac{1}{2}\Bigg[\boldsymbol{u}_t^T(\hat{\Sigma}_k^{-1} + \Lambda(\boldsymbol{u}_t))\boldsymbol{u}_t - 2\boldsymbol{u}_t^T(\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\mu}}_k \\
&\qquad\qquad + \theta(\boldsymbol{u}_t)) + \hat{\boldsymbol{\mu}}_k^T\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\mu}}_k + \theta(\boldsymbol{u}_t)^T\Lambda(\boldsymbol{u}_t)^{-1}\theta(\boldsymbol{u}_t)\Bigg]\Bigg\} \\
&\quad = \frac{(2\pi)^{N/2}|(\hat{\Sigma}_k^{-1} + \Lambda(\boldsymbol{u}_t))^{-1}|^{1/2}}{(2\pi)^{N/2}|\hat{\Sigma}_k|^{1/2}} \exp\Bigg\{ -\frac{1}{2}\Bigg[\hat{\boldsymbol{\mu}}_k^T\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\mu}}_k + \theta(\boldsymbol{u}_t)^T\Lambda(\boldsymbol{u}_t)^{-1}\theta(\boldsymbol{u}_t) \\
&\qquad\qquad - (\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\mu}}_k + \theta(\boldsymbol{u}_t))^T(\hat{\Sigma}_k^{-1} + \Lambda(\boldsymbol{u}_t))^{-1}(\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\mu}}_k + \theta(\boldsymbol{u}_t))\Bigg]\Bigg\} \\
&\qquad \cdot \mathcal{N}(\boldsymbol{u}_t; (\hat{\Sigma}_k^{-1} + \Lambda(\boldsymbol{u}_t))^{-1}(\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\mu}}_k + \theta(\boldsymbol{u}_t)), (\hat{\Sigma}_k^{-1} + \Lambda(\boldsymbol{u}_t))^{-1}) \\
&\quad \triangleq C_k \cdot \mathcal{N}(\boldsymbol{u}_t; (\hat{\Sigma}_k^{-1} + \Lambda(\boldsymbol{u}_t))^{-1}(\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\mu}}_k + \theta(\boldsymbol{u}_t)), (\hat{\Sigma}_k^{-1} + \Lambda(\boldsymbol{u}_t))^{-1}),
\end{aligned} \tag{116}$$

where we note that the defined constant C_k is a function of $z_t = k$, but not of $\boldsymbol{u}_t$.

F. Resulting $(\boldsymbol{u}_t, z_t)$ Sampling Distributions

We write Eq. (110) and Eq. (111) in terms of the derived distributions:

$$\begin{aligned}
p(z_t = k|z_{\setminus t}, \boldsymbol{u}_{\setminus t}, \boldsymbol{y}_{1:T}, \beta, \alpha, \kappa, \lambda) &\propto p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) \\
&\qquad \int_{\mathcal{U}_t} p(\boldsymbol{u}_t|\{\boldsymbol{u}_\tau|z_\tau = k, \tau \neq t\})p(\boldsymbol{y}_{1:T}|\boldsymbol{u}_t; \boldsymbol{u}_{\setminus t})d\boldsymbol{u}_t, \quad (117) \\
p(\boldsymbol{u}_t|z_t = k, z_{\setminus t}, \boldsymbol{u}_{\setminus t}, \boldsymbol{y}_{1:T}, \lambda) &\propto p(\boldsymbol{u}_t|\{\boldsymbol{u}_\tau|z_\tau = k, \tau \neq t\})p(\boldsymbol{y}_{1:T}|\boldsymbol{u}_t; \boldsymbol{u}_{\setminus t}). \quad (118)
\end{aligned}$$

Thus, the distribution over z_t, marginalizing $\boldsymbol{u}_t$, is given by

$$\begin{aligned}
&p(z_t = k|z_{\setminus t}, \boldsymbol{u}_{\setminus t}, \boldsymbol{y}_{1:T}, \beta, \alpha, \kappa, \lambda) \\
&\quad \propto p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) \int_{\mathcal{U}_t} C_k \cdot \mathcal{N}(\boldsymbol{u}_t; (\hat{\Sigma}_k^{-1} + \Lambda(\boldsymbol{u}_t))^{-1}(\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\mu}}_k + \theta(\boldsymbol{u}_t)), (\hat{\Sigma}_k^{-1} + \Lambda(\boldsymbol{u}_t))^{-1})d\boldsymbol{u}_t \\
&\quad \propto C_k \cdot p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa).
\end{aligned} \tag{119}$$

and the distribution over $\boldsymbol{u}_t$ (for $z_t = k$ fixed) is

$$p(\boldsymbol{u}_t|z_t = k, z_{\setminus t}, \boldsymbol{u}_{\setminus t}, \boldsymbol{y}_{1:T}, \lambda) = \mathcal{N}(\boldsymbol{u}_t; (\hat{\Sigma}_k^{-1} + \Lambda(\boldsymbol{u}_t))^{-1}(\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\mu}}_k + \theta(\boldsymbol{u}_t)), (\hat{\Sigma}_k^{-1} + \Lambda(\boldsymbol{u}_t))^{-1}). \tag{120}$$