PROCEEDINGS OF NEURAL INFORMATION PROCESSING SYSTEMS 2008

Nonparametric Bayesian Learning of Switching Linear Dynamical Systems

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

Erik B. Sudderth[†], **Michael I. Jordan**[†‡]
[†]Electrical Engineering & Computer Science and [‡]Statistics, University of California, Berkeley
{sudderth, jordan}@eecs.berkeley.edu

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

Abstract

Many nonlinear 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 nonparametric Bayesian approach utilizes a hierarchical Dirichlet process prior to learn an unknown number of persistent, smooth dynamical modes. 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, and the IBOVESPA stock index.

1 Introduction

Linear dynamical systems (LDSs) are useful in describing dynamical phenomena as diverse as human motion [9], financial time-series [4], maneuvering targets [6, 10], and the dance of honey bees [8]. However, such phenomena often exhibit structural changes over time and the LDS models which describe them must also change. For example, a coasting 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 nonparametric Bayesian approach for learning *switching* LDS (SLDS) models. We also consider a special case of the SLDS—the switching vector autoregressive (VAR) process—in which direct observations of the underlying dynamical process are assumed available. Although a special case of the general linear systems framework, autoregressive models have simplifying properties that often make them a practical choice in applications.

One can view switching dynamical processes as an extension of hidden Markov models (HMMs) in which each HMM state, or *mode*, is associated with a dynamical process. Existing methods for learning SLDSs and switching VAR processes rely on either fixing the number of HMM modes, such as in [8], or considering a change-point detection formulation where each inferred change is to a new, previously unseen dynamical mode, such as in [14]. 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 [2, 12]. In this paper we make use of a variant of the HDP-HMM—the *sticky HDP-HMM* of [5]—that provides improved control over the number of modes inferred by the HDP-HMM; such control is crucial for the problems we examine. Although the HDP-HMM and its sticky extension are very flexible time series models, they do make a strong Markovian assumption that observations are conditionally independent given the HMM mode. This assumption is often insufficient for capturing the temporal dependencies of the observations in real data. Our nonparametric Bayesian approach for learning switching dynamical processes extends the sticky HDP-HMM formulation to learn an unknown number of persistent, smooth dynamical modes and thereby capture a wider range of temporal dependencies.

2 Background: Switching Linear Dynamic Systems

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

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

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

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

Here, the observations depend linearly on the previous r observation vectors. Every VAR(r) process can be described in SS form by, for example, the following transformation:

$$\boldsymbol{x}_t = \begin{bmatrix} A_1 & A_2 & \dots & A_r \\ I & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & I & 0 \end{bmatrix} \boldsymbol{x}_{t-1} + \begin{bmatrix} I \\ 0 \\ \vdots \\ 0 \end{bmatrix} \boldsymbol{e}_t \qquad \boldsymbol{y}_t = [I \quad 0 \quad \dots \quad 0]\, \boldsymbol{x}_t. \tag{3}$$

Note that there are many such equivalent *minimal* SS representations that result in the same input-output relationship, where minimality implies that there does not exist a realization with lower state dimension. On the other hand, not every SS model may be expressed as a VAR(r) process for finite r [1]. We can thus conclude that considering a class of SS models with state dimension $r \cdot d$ and arbitrary dynamic matrix A subsumes the class of VAR(r) processes.

The dynamical phenomena we examine in this paper exhibit behaviors better modeled as switches between a set of linear dynamical models. Due to uncertainty in the mode of the process, the overall model is nonlinear. We define a *switching linear dynamical system* (SLDS) by

$$\boldsymbol{x}_t = A^{(z_t)}\boldsymbol{x}_{t-1} + \boldsymbol{e}_t(z_t) \qquad \boldsymbol{y}_t = C\boldsymbol{x}_t + \boldsymbol{w}_t. \tag{4}$$

The first-order Markov process z_t indexes the mode-specific LDS at time t, which is driven by Gaussian noise $\boldsymbol{e}_t(z_t) \sim \mathcal{N}(0, \Sigma^{(z_t)})$. We similarly define a *switching* VAR(r) process by

$$\boldsymbol{y}_t = \sum_{i=1}^{r} A_i^{(z_t)} \boldsymbol{y}_{t-i} + \boldsymbol{e}_t(z_t) \qquad \boldsymbol{e}_t(z_t) \sim \mathcal{N}(0, \Sigma^{(z_t)}). \tag{5}$$

Note that the underlying state dynamics of the SLDS are equivalent to a switching VAR(1) process.

3 Background: Dirichlet Processes and the Sticky HDP-HMM

A Dirichlet process (DP), denoted by $\mathrm{DP}(\gamma, H)$, is a distribution on discrete measures

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad \theta_k \sim H \tag{6}$$

on a parameter space Θ. The weights are generated via a *stick-breaking construction* [11]:

$$\beta_k = \beta'_k \prod_{\ell=1}^{k-1} (1 - \beta'_\ell) \qquad \beta'_k \sim \text{Beta}(1, \gamma). \tag{7}$$

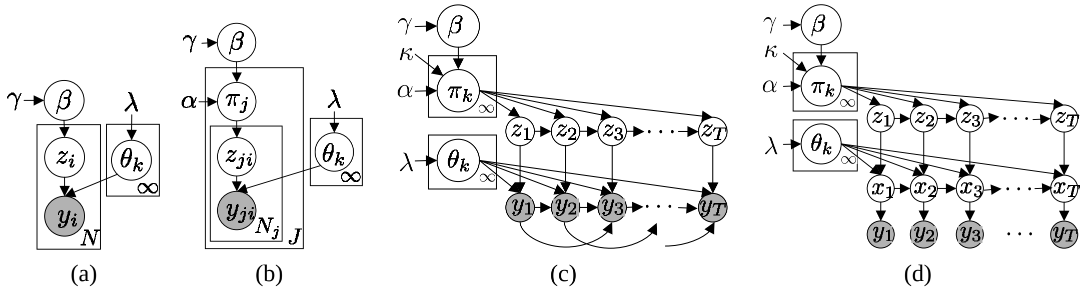


Figure 1: For all graphs, $\beta \sim \text{GEM}(\gamma)$ and $\theta_k \sim H(\lambda)$. (a) DP mixture model in which $z_i \sim \beta$ and $y_i \sim f(y \mid \theta_{z_i})$. (b) HDP mixture model with $\pi_j \sim \text{DP}(\alpha, \beta)$, $z_{ji} \sim \pi_j$, and $y_{ji} \sim f(y \mid \theta_{z_{ji}})$. (c)-(d) Sticky HDP-HMM prior on switching VAR(2) and SLDS processes with the mode evolving as $z_{t+1} \sim \pi_{z_t}$ for $\pi_k \sim \text{DP}(\alpha+\kappa, (\alpha\beta + \kappa\delta_k)/(\alpha+\kappa))$. The dynamical processes are as in Eq. (13).

We denote this distribution by $\beta \sim \text{GEM}(\gamma)$. The DP is commonly used as a prior on the parameters of a mixture model, resulting in a *DP mixture model* (see Fig.1(a)). To generate observations, we choose $\bar{\theta}_i \sim G_0$ and $y_i \sim F(\bar{\theta}_i)$. This sampling process is often described via a discrete variable $z_i \sim \beta$ indicating which component generates $y_i \sim F(\theta_{z_i})$.

The *hierarchical Dirichlet process* (HDP) [12] extends the DP to cases in which groups of data are produced by related, yet distinct, generative processes. Taking a hierarchical Bayesian approach, the HDP draws G_0 from a Dirichlet process prior $\text{DP}(\gamma, H)$, and then draws group specific distributions $G_j \sim \text{DP}(\alpha, G_0)$. Here, the base measure G_0 acts as an "average" distribution ($E[G_j \mid G_0] = G_0$) encoding the frequency of each shared, global parameter:

$$G_j = \sum_{t=1}^{\infty} \tilde{\pi}_{jt} \delta_{\tilde{\theta}_{jt}} \qquad \tilde{\pi}_j \sim \text{GEM}(\alpha) \tag{8}$$

$$= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\theta_k} \qquad \pi_j \sim \text{DP}(\alpha, \beta)\,. \tag{9}$$

Because G_0 is discrete, multiple $\tilde{\theta}_{jt} \sim G_0$ may take identical values θ_k. Eq. (9) aggregates these probabilities, allowing an observation y_{ji} to be directly associated with the unique global parameters via an indicator random variable $z_{ji} \sim \pi_j$. See Fig. 1(b).

An alternative, non–constructive characterization of samples $G_0 \sim \text{DP}(\gamma, H)$ from a Dirichlet process states that for every finite partition $\{A_1, \ldots, A_K\}$ of Θ,

$$(G_0(A_1), \ldots, G_0(A_K)) \sim \text{Dir}(\gamma H(A_1), \ldots, \gamma H(A_K)). \tag{10}$$

Using this expression, it can be shown that the following finite, hierarchical mixture model converges in distribution to the HDP as $L \to \infty$ [7, 12]:

$$\beta \sim \text{Dir}(\gamma/L, \ldots, \gamma/L) \qquad \pi_j \sim \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_L). \tag{11}$$

This *weak limit* approximation is used by the sampler of Sec. 4.2.

The HDP can be used to develop an HMM with a potentially infinite mode space [2, 12]. For this HDP-HMM, each HDP group-specific distribution, π_j, is a mode-specific transition distribution and, due to the infinite mode space, there are infinitely many groups. Let z_t denote the mode of the Markov chain at time t. For discrete Markov processes $z_t \sim \pi_{z_{t-1}}$, so that z_{t-1} indexes the group to which y_t is assigned. The current HMM mode z_t then indexes the parameter θ_{z_t} used to generate observation y_t. See Fig. 1(c), ignoring the direct correlation in the observations.

By sampling $\pi_j \sim \text{DP}(\alpha, \beta)$, the HDP prior encourages modes to have similar transition distributions ($E[\pi_{jk} \mid \beta] = \beta_k$). 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 [5] that one may mitigate this problem by instead considering a *sticky* HDP-HMM where π_j is distributed as follows:

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

Here, $(\alpha\beta + \kappa\delta_j)$ indicates that an amount $\kappa > 0$ is added to the j^{th} component of $\alpha\beta$. The measure of π_j over a finite partition $(Z_1, \dots, Z_K)$ of the positive integers $\mathbb{Z}_+$, as described by Eq. (10), adds an amount κ only to the arbitrarily small partition containing j, corresponding to a self-transition. When $\kappa = 0$ the original HDP-HMM is recovered. We place a vague prior on κ and learn the self-transition bias from the data.

4 The HDP-SLDS and HDP-AR-HMM Models

For greater modeling flexibility, we take a nonparametric approach in defining the mode space of our switching dynamical processes. Specifically, we develop extensions of the sticky HDP-HMM for both the SLDS and switching VAR models. For the SLDS, we consider conditionally-dependent emissions of which only noisy observations are available (see Fig. 1(d)). For this model, which we refer to as the *HDP-SLDS*, we place a prior on the parameters of the SLDS and infer their posterior from the data. We do, however, fix the measurement matrix, C, for reasons of identifiability. Let $\tilde{C} \in \mathbb{R}^{d \times n}$, $n \geq d$, be the measurement matrix associated with a dynamical system defined by $\tilde{A}$, and assume $\tilde{C}$ has full row rank. Then, without loss of generality, we may consider $C = [I \;\; 0]$ since there exists an invertible transformation T such that the pair $C = \tilde{C}T = [I \;\; 0]$ and $A = T^{-1}\tilde{A}T$ defines an equivalent input-output system. The dimensionality of I is determined by that of the data. Our choice of the number of columns of zeros is, in essence, a choice of model order.

The previous work of Fox et al. [6] considered a related, yet simpler formulation for modeling a maneuvering target as a fixed LDS driven by a switching exogenous input. Since the number of maneuver modes was assumed unknown, the exogenous input was taken to be the emissions of a HDP-HMM. This work can be viewed as an extension of the work by Caron et. al. [3] in which the exogenous input was an independent noise process generated from a DP mixture model. The HDP-SLDS is a major departure from these works since the dynamic parameters themselves change with the mode and are learned from the data, providing a much more expressive model.

The switching VAR(r) process can similarly be posed as an HDP-HMM in which the observations are modeled as conditionally VAR(r). This model is referred to as the *HDP-AR-HMM* and is depicted in Fig. 1(c). The generative processes for these two models are summarized as follows:

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

(13)

Here, π_j is as defined in Sec. 3 and the additive noise processes as in Sec. 2.

4.1 Posterior Inference of Dynamic Parameters

In this section we focus on developing a prior to regularize the learning of different dynamical modes conditioned on a fixed mode assignment $z_{1:T}$. For the SLDS, we analyze the posterior distribution of the dynamic parameters given a fixed, known state sequence $\boldsymbol{x}_{1:T}$. Methods for learning the number of modes and resampling the sequences $\boldsymbol{x}_{1:T}$ and $z_{1:T}$ are discussed in Sec. 4.2.

Conditioned on the mode sequence, one may partition the observations into K different linear regression problems, where $K = |\{z_1, \dots, z_T\}|$. That is, for each mode k, we may form a matrix $\mathbf{Y}^{(k)}$ with N_k columns consisting of the observations $\boldsymbol{y}_t$ with $z_t = k$. Then,

$$\mathbf{Y}^{(k)} = \mathbf{A}^{(k)}\bar{\mathbf{Y}}^{(k)} + \mathbf{E}^{(k)}, \tag{14}$$

where $\mathbf{A}^{(k)} = [A_1^{(k)} \dots A_r^{(k)}]$, $\bar{\mathbf{Y}}^{(k)}$ is a matrix of lagged observations, and $\mathbf{E}^{(k)}$ the associated noise vectors. Let $\mathbf{D}^{(k)} = \{\mathbf{Y}^{(k)}, \bar{\mathbf{Y}}^{(k)}\}$. The posterior distribution over the VAR(r) parameters associated with the k^{th} mode decomposes as follows:

$$p(\mathbf{A}^{(k)}, \Sigma^{(k)} \mid \mathbf{D}^{(k)}) = p(\mathbf{A}^{(k)} \mid \Sigma^{(k)}, \mathbf{D}^{(k)}) p(\Sigma^{(k)} \mid \mathbf{D}^{(k)}). \tag{15}$$

We place a conjugate *matrix-normal inverse-Wishart* prior on the parameters $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$ [13], providing a reasonable combination of flexibility and analytical convenience. A matrix $\boldsymbol{A} \in \mathbb{R}^{d \times m}$ has a matrix-normal distribution $\mathcal{MN}(\boldsymbol{A}; \boldsymbol{M}, \boldsymbol{V}, \boldsymbol{K})$ if

$$p(\boldsymbol{A}) = \frac{|\boldsymbol{K}|^{\frac{d}{2}}}{|2\pi\boldsymbol{V}|^{\frac{m}{2}}} e^{-\frac{1}{2}tr\left((\boldsymbol{A}-\boldsymbol{M})^T \boldsymbol{V}^{-1} (\boldsymbol{A}-\boldsymbol{M})\boldsymbol{K}\right)}, \tag{16}$$

where $\boldsymbol{M}$ is the mean matrix and $\boldsymbol{V}$ and $\boldsymbol{K}^{-1}$ are the covariances along the rows and columns, respectively. A vectorization of the matrix $\boldsymbol{A}$ results in

$$p(\mathrm{vec}(\boldsymbol{A})) = \mathcal{N}(\mathrm{vec}(\boldsymbol{M}), \boldsymbol{K}^{-1} \otimes \boldsymbol{V}), \tag{17}$$

where $\otimes$ denotes the Kronecker product. The resulting posterior is derived as

$$p(\mathbf{A}^{(k)} \mid \Sigma^{(k)}, \mathbf{D}^{(k)}) = \mathcal{MN}(\mathbf{A}^{(k)}; \mathbf{S}_{y\bar{y}}^{(k)} \mathbf{S}_{\bar{y}\bar{y}}^{-(k)}, \Sigma^{-(k)}, \mathbf{S}_{\bar{y}\bar{y}}^{(k)}), \tag{18}$$

with $\boldsymbol{B}^{-(k)}$ denoting $(\boldsymbol{B}^{(k)})^{-1}$ for a given matrix $\boldsymbol{B}$, and

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

We place an inverse-Wishart prior $\mathrm{IW}(S_0, n_0)$ on $\Sigma^{(k)}$. Then,

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

where $\mathbf{S}_{y|\bar{y}}^{(k)} = \mathbf{S}_{yy}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)} \mathbf{S}_{\bar{y}\bar{y}}^{-(k)} \mathbf{S}_{y\bar{y}}^{(k)^T}$. When $\boldsymbol{A}$ is simply a vector, the matrix-normal inverse-Wishart prior reduces to the normal inverse-Wishart prior with scale parameter $\boldsymbol{K}$.

For the HDP-SLDS, we additionally place an $\mathrm{IW}(R_0, r_0)$ prior on the measurement noise covariance R, which is shared between modes. The posterior distribution is given by

$$p(R \mid \boldsymbol{y}_{1:T}, \boldsymbol{x}_{1:T}) = \mathrm{IW}(S_R + R_0, T + r_0), \tag{20}$$

with $S_R = \sum_{t=1}^{T} (\boldsymbol{y}_t - C\boldsymbol{x}_t)(\boldsymbol{y}_t - C\boldsymbol{x}_t)^T$. Further details are provided in supplemental Appendix I.

4.2 Gibbs Sampler

For the switching VAR(r) process, our sampler iterates between sampling the mode sequence, $z_{1:T}$, and both the dynamic and sticky HDP-HMM parameters. The sampler for the SLDS is identical to that of a switching VAR(1) process with the additional step of sampling the state sequence, $\boldsymbol{x}_{1:T}$, and conditioning on the state sequence when resampling dynamic parameters. The resulting Gibbs sampler is described below and further elaborated upon in supplemental Appendix II.

Sampling Dynamic Parameters Conditioned on a sample of the mode sequence, $z_{1:T}$, and the observations, $\boldsymbol{y}_{1:T}$, or state sequence, $\boldsymbol{x}_{1:T}$, we can sample the dynamic parameters $\boldsymbol{\theta} = \{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$ from the posterior density described in Sec. 4.1. For the HDP-SLDS, we additionally sample R.

Sampling $z_{1:T}$ As shown in [5], the mixing rate of the Gibbs sampler for the HDP-HMM can be dramatically improved by using a truncated approximation to the HDP, such as the weak limit approximation, and jointly sampling the mode sequence using a variant of the forward-backward algorithm. Specifically, we compute backward messages $m_{t+1,t}(z_t) \propto p(\boldsymbol{y}_{t+1:T} | z_t, \boldsymbol{y}_{t-r+1:t}, \boldsymbol{\pi}, \boldsymbol{\theta})$ and then recursively sample each z_t conditioned on z_{t-1} from

$$p(z_t \mid z_{t-1}, \boldsymbol{y}_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto p(z_t \mid \pi_{z_{t-1}}) p(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-r:t-1}, \mathbf{A}^{(z_t)}, \Sigma^{(z_t)}) m_{t+1,t}(z_t), \tag{21}$$

where $p(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-r:t-1}, \mathbf{A}^{(z_t)}, \Sigma^{(z_t)}) = \mathcal{N}(\sum_{i=1}^{r} A_i^{(z_t)} \boldsymbol{y}_{t-i}, \Sigma^{(z_t)})$. Joint sampling of the mode sequence is especially important when the observations are directly correlated via a dynamical process since this correlation further slows the mixing rate of the direct assignment sampler of [12]. Note that the approximation of Eq. (11) retains the HDP's nonparametric nature by encouraging the use of fewer than L components while allowing the generation of new components, upper bounded by L, as new data are observed.

Sampling $\boldsymbol{x}_{1:T}$ (HDP-SLDS only) Conditioned on the mode sequence $z_{1:T}$ and the set of dynamic parameters $\boldsymbol{\theta}$, 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 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

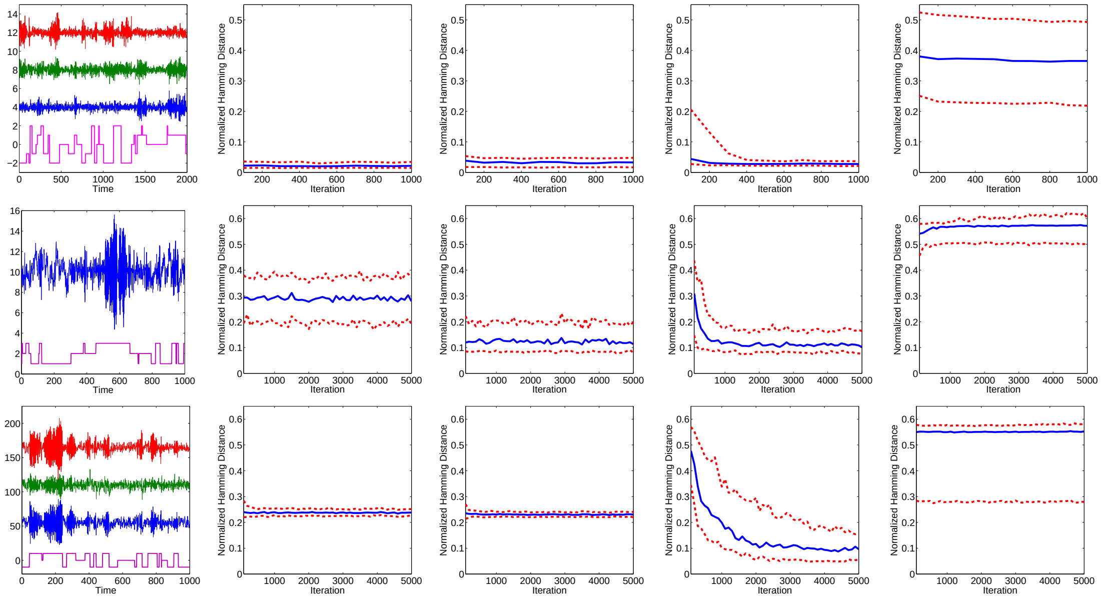
$$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) \propto p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, A^{(z_t)}, \Sigma^{(z_t)}) p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, R) m_{t+1,t}(\boldsymbol{x}_t). \tag{22}$$

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

$$\theta_{t,t-1} = A^{(z_t)^T} \Sigma^{-(z_t)} (\Sigma^{-(z_t)} + C^T R^{-1} C + \Lambda_{t+1,t})^{-1} (C^T R^{-1} \boldsymbol{y}_t + \theta_{t+1,t}) \tag{23}$$

$$\Lambda_{t,t-1} = A^{(z_t)^T} \Sigma^{-(z_t)} A^{(z_t)} - A^{(z_t)^T} \Sigma^{-(z_t)} (\Sigma^{-(z_t)} + C^T R^{-1} C + \Lambda_{t+1,t})^{-1} \Sigma^{-(z_t)} A^{(z_t)}.$$

See supplemental Appendix II for a more numerically stable version of this recursion.



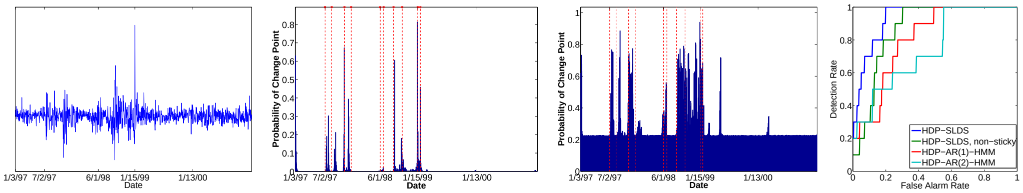
(a) (b) (c) (d) (e)

Figure 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.

5 Results

Synthetic Data In Fig. 2, we compare the performance of the HDP-VAR(1)-HMM, HDP-VAR(2)-HMM, HDP-SLDS, and a baseline sticky HDP-HMM on three sets of test data (see Fig. 2(a)). The Hamming distance error is calculated by first choosing the optimal mapping of indices maximizing overlap between the true and estimated mode sequences. For the first scenario, the data were generated from a 5-mode switching VAR(1) process. The three switching linear dynamical models provide comparable performance since both the HDP-VAR(2)-HMM and HDP-SLDS with $C = I$ contain the class of HDP-VAR(1)-HMMs. Note that the HDP-SLDS sampler is slower to mix since the hidden, three-dimensional continuous state is also sampled. In the second scenario, 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 $C = [1 \;\; 0]$ is comparable after burn-in. As shown in Sec. 2, this HDP-SLDS model encompasses the class of HDP-AR(2)-HMMs. The data in the third scenario were generated from a 3-mode SLDS model with $C = I$. Here, we clearly see that neither the HDP-VAR(1)-HMM nor HDP-VAR(2)-HMM is equivalent to the HDP-SLDS. 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. Finally, note that all of the switching models yielded significant improvements relative to the baseline sticky HDP-HMM, even when the latter was given first differences of the observations. This input representation, which is equivalent to an HDP-VAR(1)-HMM with random walk dynamics ($A^{(k)} = I$ for all k), is more effective than using raw observations for HDP-HMM learning, but still much less effective than richer models which switch among learned LDS.

IBOVESPA Stock Index We test the HDP-SLDS model on the IBOVESPA stock index (Sao Paulo Stock Exchange) over the period of 01/03/1997 to 01/16/2001. There are ten key world events shown in Fig. 3 and cited in [4] as affecting the emerging Brazilian market during this time period. In [4], a 2-mode Markov switching stochastic volatility (MSSV) model is used to identify periods of higher volatility in the daily returns. 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. The HDP-SLDS is able to infer very similar change points to those presented in [4]. Interestingly, the HDP-SLDS consistently identifies three regimes of volatility versus the assumed 2-mode model. In Fig. 3, the overall performance of the



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

Figure 3: (a) IBOVESPA stock index daily returns from 01/03/1997 to 01/16/2001. (b) Plot of the estimated probability of a change point on each day using 3000 Gibbs samples for the HDP-SLDS. The 10 key events are indicated with red lines. (c) Similar plot for the *non-sticky* HDP-SLDS with no bias towards self-transitions. (d) ROC curves for the HDP-SLDS, non-sticky HDP-SLDS, HDP-AR(1)-HMM, and HDP-AR(2)-HMM.

HDP-SLDS is compared to that of the HDP-AR(1)-HMM, HDP-AR(2)-HMM, and HDP-SLDS with no bias for self-transitions (i.e., $\kappa = 0$.) The ROC curves shown in Fig. 3(d) 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. We clearly see the advantage of using a SLDS model combined with the sticky HDP-HMM prior on the mode sequence. Without the sticky extension, the HDP-SLDS over-segments the data and rapidly switches between redundant states which leads to a dramatically larger number of inferred change points.

Dancing Honey Bees We test the HDP-VAR(1)-HMM on a set of six dancing honey bee sequences, aiming to segment the sequences into the three dances displayed in Fig. 4. (Note that we did not see performance gains by considering the HDP-SLDS, so we omit showing results for that architecture.) The data consist of measurements $\boldsymbol{y}_t = [\cos(\theta_t) \;\; \sin(\theta_t) \;\; x_t \;\; y_t]^T$, where (x_t, y_t) denotes the 2D coordinates of the bee's body and θ_t its head angle. We compare our results to those of Xuan and Murphy [14], who used a change-point detection technique for inference on this dataset. As shown in Fig. 4(d)-(e), our model achieves a superior segmentation compared to the change-point formulation in almost all cases, while also identifying modes which reoccur over time.

Oh et al. [8] 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 [8] 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 1 we report the performance of these methods as well as the median performance (over 100 trials) of the unsupervised HDP-VAR(1)-HMM to provide a sense of the level of performance achievable without detailed, manual supervision. As seen in Table 1, 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. 4(a)-(c)); 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. This 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. As we see in Table 1, this considerably improved performance for these three sequences.

Not depicted in the plots in Fig. 4 is the extreme variation in head angle during the waggle dances of sequences 1 to 3. 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.

6 Discussion

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 non-

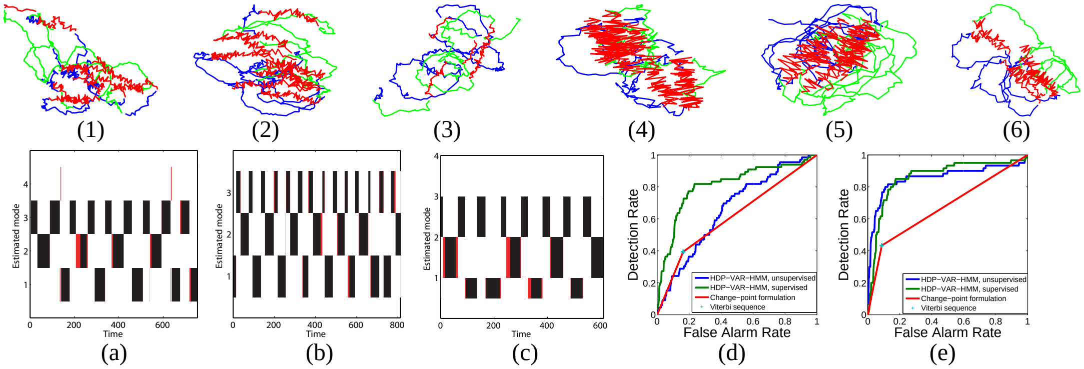


Figure 4: (top) Trajectories of the dancing honey bees for sequences 1 to 6, colored by *waggle* (red), *turn right* (blue), and *turn left* (green) dances. (a)-(c) Estimated mode sequences representing the median error for sequences 4, 5, and 6 at the 200th Gibbs iteration, with errors indicated in red. (d)-(e) ROC curves for the unsupervised HDP-VAR-HMM, partially supervised HDP-VAR-HMM, and change-point formulation of [14] using the Viterbi sequence for segmenting datasets 1-3 and 4-6, respectively.

Sequence	1	2	3	4	5	6
HDP-VAR(1)-HMM unsupervised	46.5	44.1	45.6	83.2	93.2	88.7
HDP-VAR(1)-HMM partially supervised	65.9	88.5	79.2	86.9	92.3	89.1
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 1: 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) [8].

parametric Bayesian 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, 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. An interesting direction for future research is learning models of varying order for each mode.

References

[1] M. Aoki and A. Havenner. State space modeling of multiple time series. *Econ. Rev.*, 10(1):1–59, 1991.

[2] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden Markov model. In *NIPS*, 2002.

[3] F. Caron, M. Davy, A. Doucet, E. Duflos, and P. Vanheeghe. Bayesian inference for dynamic models with Dirichlet process mixtures. In *Int. Conf. Inf. Fusion*, July 2006.

[4] C. Carvalho and H. Lopes. Simulation-based sequential analysis of Markov switching stochastic volatility models. *Comp. Stat. & Data Anal.*, 2006.

[5] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. An HDP-HMM for systems with state persistence. In *ICML*, 2008.

[6] E. B. Fox, E. B. Sudderth, and A. S. Willsky. Hierarchical Dirichlet processes for tracking maneuvering targets. In *Int. Conf. Inf. Fusion*, July 2007.

[7] H. Ishwaran and M. Zarepour. Exact and approximate sum–representations for the Dirichlet process. *Can. J. Stat.*, 30:269–283, 2002.

[8] S. Oh, J. Rehg, T. Balch, and F. Dellaert. Learning and inferring motion patterns using parametric segmental switching linear dynamic systems. *IJCV*, 77(1–3):103–124, 2008.

[9] J. M. Pavlović, V. Rehg and J. MacCormick. Learning switching linear models of human motion. In *NIPS*, 2000.

[10] X. Rong Li and V. Jilkov. Survey of maneuvering target tracking. Part V: Multiple-model methods. *IEEE Trans. Aerosp. Electron. Syst.*, 41(4):1255–1321, 2005.

[11] J. Sethuraman. A constructive definition of Dirichlet priors. *Stat. Sinica*, 4:639–650, 1994.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Amer. Stat. Assoc.*, 101(476):1566–1581, 2006.

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

[14] X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. In *ICML*, 2007.

APPENDIX I
DYNAMIC PARAMETER POSTERIOR

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

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

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

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

We define the following notation for Eq. 2:

$$\mathbf{Y}^{(k)} = \mathbf{A}^{(k)} \bar{\mathbf{Y}}^{(k)} + \mathbf{E}^{(k)}. \tag{3}$$

Let $\mathbf{D}^{(k)} = \{\mathbf{Y}^{(k)}, \bar{\mathbf{Y}}^{(k)}\}$. We place a matrix-normal inverse-Wishart prior on the dynamic parameters $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\}$ and show that the posterior remains matrix-normal inverse Wishart. The matrix-normal inverse-Wishart prior is given by placing a matrix-normal prior $\mathcal{MN}(\mathbf{A}^{(k)}; \boldsymbol{M}, \Sigma^{(k)}, \boldsymbol{K})$ on $\mathbf{A}^{(k)}$ given $\Sigma^{(k)}$:

$$p(\mathbf{A}^{(k)} \mid \Sigma^{(k)}) = \frac{|\boldsymbol{K}|^{d/2}}{|2\pi\Sigma^{(k)}|^{m/2}} \exp\left(-\frac{1}{2} tr((\boldsymbol{A} - \boldsymbol{M})^T \Sigma^{-(k)} (\boldsymbol{A} - \boldsymbol{M})\boldsymbol{K})\right) \tag{4}$$

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

$$p(\Sigma^{(k)}) = \frac{|S_0|^{n/2} |\Sigma^{(k)}|^{-(d+n+1)/2}}{2^{nd/2}\Gamma_d(n/2)} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)} S_0)\right) \tag{5}$$

where $\Gamma_d(n/2)$ 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, $\bar{\boldsymbol{y}}_t = [\boldsymbol{y}_{t-1}^T \cdots \boldsymbol{y}_{t-r}^T]^T$, we have

$$\begin{aligned} p(\mathbf{D}^{(k)} | \mathbf{A}^{(k)}, \Sigma^{(k)}) &= \frac{1}{|2\pi\Sigma^{(k)}|^{N_k/2}} \exp\left(-\frac{1}{2} \sum_i (\boldsymbol{y}_{t_i} - \mathbf{A}^{(k)} \bar{\boldsymbol{y}}_{t_i})^T \Sigma^{-(k)} (\boldsymbol{y}_{t_i} - \mathbf{A}^{(k)} \bar{\boldsymbol{y}}_{t_i})\right) \\ &= \frac{1}{|2\pi\Sigma^{(k)}|^{N_k/2}} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)} (\mathbf{Y}^{(k)} - \mathbf{A}^{(k)} \bar{\mathbf{Y}}^{(k)})(\mathbf{Y}^{(k)} - \mathbf{A}^{(k)} \bar{\mathbf{Y}}^{(k)})^T)\right) \\ &= \frac{1}{|2\pi\Sigma^{(k)}|^{N_k/2}} \exp\left(-\frac{1}{2} tr((\mathbf{Y}^{(k)} - \mathbf{A}^{(k)} \bar{\mathbf{Y}}^{(k)})^T \Sigma^{-(k)} (\mathbf{Y}^{(k)} - \mathbf{A}^{(k)} \bar{\mathbf{Y}}^{(k)})\boldsymbol{I})\right) \\ &= \mathcal{MN}(\mathbf{Y}^{(k)}; \mathbf{A}^{(k)} \bar{\mathbf{Y}}^{(k)}, \Sigma^{(k)}, \boldsymbol{I}). \end{aligned} \tag{6}$$

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

$$p(\mathbf{D}^{(k)}, \mathbf{A}^{(k)} \mid \Sigma^{(k)}) \;\; = \;\; p(\mathbf{D}^{(k)} \mid \mathbf{A}^{(k)}, \Sigma^{(k)}) p(\mathbf{A}^{(k)} \mid \Sigma^{(k)}). \tag{7}$$

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

$$\begin{aligned}
&\log p(\mathbf{D}^{(k)}, \mathbf{A}^{(k)} \mid \Sigma^{(k)}) + C \\
&\qquad = -\frac{1}{2} tr((\mathbf{Y}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{Y}}^{(k)})^T \Sigma^{-(k)} (\mathbf{Y}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{Y}}^{(k)}) + (\mathbf{A}^{(k)} - \boldsymbol{M})^T \Sigma^{-(k)} (\mathbf{A}^{(k)} - \boldsymbol{M})\boldsymbol{K}) \\
&\qquad = -\frac{1}{2} tr(\Sigma^{-(k)} \{(\mathbf{Y}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{Y}}^{(k)})(\mathbf{Y}^{(k)} - \mathbf{A}^{(k)}\bar{\mathbf{Y}}^{(k)})^T + \ (\mathbf{A}^{(k)} - \boldsymbol{M})\boldsymbol{K}(\mathbf{A}^{(k)} - \boldsymbol{M})^T\}) \\
&\qquad = -\frac{1}{2} tr(\Sigma^{-(k)} \{\mathbf{A}^{(k)}\mathbf{S}_{\bar{y}\bar{y}}^{(k)}\mathbf{A}^{(k)^T} - 2\mathbf{S}_{y\bar{y}}^{(k)}\mathbf{A}^{(k)^T} + \mathbf{S}_{yy}^{(k)}\}) \\
&\qquad = -\frac{1}{2} tr(\Sigma^{-(k)} \{(\mathbf{A}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)}\mathbf{S}_{\bar{y}\bar{y}}^{-(k)})\mathbf{S}_{\bar{y}\bar{y}}^{(k)}(\mathbf{A}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)}\mathbf{S}_{\bar{y}\bar{y}}^{-(k)})^T + \mathbf{S}_{y|\bar{y}}^{(k)}\}),
\end{aligned} \tag{8}$$

for $C = -\log \frac{1}{|2\pi\boldsymbol{\Sigma}^{(k)}|^{N_k/2}} \frac{|\boldsymbol{K}|^{d/2}}{|2\pi\boldsymbol{\Sigma}^{(k)}|^{rN_k/2}}$ and $\mathbf{S}_{y|\bar{y}}^{(k)} = \mathbf{S}_{yy}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)}\mathbf{S}_{\bar{y}\bar{y}}^{-(k)}\mathbf{S}_{y\bar{y}}^{(k)^T}$ using the definitions

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

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

$$\begin{aligned}
p(\mathbf{A}^{(k)} \mid \mathbf{D}^{(k)}, \Sigma^{(k)}) &\propto \exp\left(-\frac{1}{2} tr((\mathbf{A}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)}\mathbf{S}_{\bar{y}\bar{y}}^{-(k)})^T \Sigma^{-(k)} (\mathbf{A}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)}\mathbf{S}_{\bar{y}\bar{y}}^{-(k)})\mathbf{S}_{\bar{y}\bar{y}}^{(k)})\right) \\
&= \mathcal{MN}(\mathbf{A}^{(k)}; \mathbf{S}_{y\bar{y}}^{(k)}\mathbf{S}_{\bar{y}\bar{y}}^{-(k)}, \Sigma^{-(k)}, \mathbf{S}_{\bar{y}\bar{y}}^{(k)}).
\end{aligned} \tag{9}$$

Marginalizing Eq. 8 over the dynamic matrix $\mathbf{A}^{(k)}$, we derive

$$\begin{aligned}
p(\mathbf{D}^{(k)} \mid \Sigma^{(k)}) &= \int_{\mathbf{A}^{(k)}} p(\mathbf{D}^{(k)}, \mathbf{A}^{(k)} \mid \Sigma^{(k)}) d\mathbf{A}^{(k)} \\
&= \int_{\mathbf{A}^{(k)}} \frac{|\boldsymbol{K}^{d/2}|}{|2\pi\Sigma^{(k)}|^{N_k/2}|2\pi\Sigma^{(k)}|^{rN_k/2}} \\
&\quad \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)}(\mathbf{A}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)}\mathbf{S}_{\bar{y}\bar{y}}^{-(k)})\mathbf{S}_{\bar{y}\bar{y}}^{(k)}(\mathbf{A}^{(k)} - \mathbf{S}_{y\bar{y}}^{(k)}\mathbf{S}_{\bar{y}\bar{y}}^{-(k)})^T)\right) \\
&\quad \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)}\mathbf{S}_{y|\bar{y}}^{(k)})\right) d\mathbf{A}^{(k)} \\
&= \frac{|\boldsymbol{K}|^{d/2}}{|2\pi\Sigma^{(k)}|^{N_k/2}} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)}\mathbf{S}_{y|\bar{y}}^{(k)})\right) \\
&\quad \int_{\mathbf{A}^{(k)}} \frac{1}{|\mathbf{S}_{\bar{y}\bar{y}}^{(k)}|^{d/2}} \mathcal{MN}(\mathbf{A}^{(k)}; \mathbf{S}_{y\bar{y}}^{(k)}\mathbf{S}_{\bar{y}\bar{y}}^{-(k)}, \Sigma^{-(k)}, \mathbf{S}_{\bar{y}\bar{y}}^{(k)}) d\mathbf{A}^{(k)} \\
&= \frac{|\boldsymbol{K}|^{d/2}}{|2\pi\Sigma^{(k)}|^{N_k/2}|\mathbf{S}_{\bar{y}\bar{y}}^{(k)}|^{d/2}} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)}\mathbf{S}_{y|\bar{y}}^{(k)})\right)
\end{aligned} \tag{10}$$

Using the above, the posterior of the covariance parameter is computed as

$$\begin{aligned}
&p(\Sigma^{(k)} \mid \mathbf{D}^{(k)}) \propto p(\mathbf{D}^{(k)} \mid \Sigma^{(k)}) p(\Sigma^{(k)}) \\
&\qquad \propto \frac{|\boldsymbol{K}|^{d/2}}{|2\pi\Sigma^{(k)}|^{N_k/2}|\mathbf{S}_{\bar{y}\bar{y}}^{(k)}|^{d/2}} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)}\mathbf{S}_{y|\bar{y}}^{(k)})\right) |\Sigma^{(k)}|^{-(d+n+1)/2} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)} S_0)\right) \\
&\qquad \propto |\Sigma^{(k)}|^{-(d+N_k+n+1)/2} \exp\left(-\frac{1}{2} tr(\Sigma^{-(k)}(\mathbf{S}_{y|\bar{y}}^{(k)} + S_0))\right) \\
&\qquad = \text{IW}(\mathbf{S}_{y|\bar{y}}^{(k)} + S_0, N_k + n).
\end{aligned} \tag{11}$$

APPENDIX II
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

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:

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

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:

$$\begin{aligned} p(z_1 \mid \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) &\propto p(z_1) p(\boldsymbol{y}_1 \mid \theta_{z_1}, \boldsymbol{y}_{1-r:0}) \sum_{z_{2:T}} \prod_t p(z_t \mid \pi_{z_{t-1}}) p(\boldsymbol{y}_t \mid \theta_{z_t}, \boldsymbol{y}_{t-r:t-1}) \\ &\propto p(z_1) p(\boldsymbol{y}_1 \mid \theta_{z_1}, \boldsymbol{y}_{1-r:0}) \sum_{z_2} p(z_2 \mid \pi_{z_1}) p(\boldsymbol{y}_2 \mid \theta_{z_2}, \boldsymbol{y}_{2-r:1}) m_{3,2}(z_2) \\ &\propto p(z_1) p(\boldsymbol{y}_1 \mid \theta_{z_1}, \boldsymbol{y}_{1-r:0}) m_{2,1}(z_1), \end{aligned} \tag{12}$$

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:

$$m_{t,t-1}(z_{t-1}) \propto \begin{cases} \sum_{z_t} p(z_t \mid \pi_{z_{t-1}}) p(\boldsymbol{y}_t \mid \theta_{z_t}, \boldsymbol{y}_{t-r:t-1}) m_{t+1,t}(z_t), & t \leq T; \\ 1, & t = T+1. \end{cases} \tag{13}$$

The general conditional distribution of z_t is:

$$p(z_t \mid z_{t-1}, \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto p(z_t \mid \pi_{z_{t-1}}) p(\boldsymbol{y}_t \mid \theta_{z_t}, \boldsymbol{y}_{t-r:t-1}) m_{t+1,t}(z_t). \tag{14}$$

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

$$p(z_t = k \mid z_{t-1}, \boldsymbol{y}_{1-r:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto \pi_{z_{t-1}}(k) \mathcal{N}(\boldsymbol{y}_t; \sum_{i=1}^{r} A_i^{(k)} \boldsymbol{y}_{t-i}, \Sigma^{(k)}) m_{t+1,t}(k) \tag{15}$$

$$m_{t+1,t}(k) = \sum_{j=1}^{L} \pi_k(j) \mathcal{N}(\boldsymbol{y}_{t+1}; \sum_{i=1}^{r} A_i^{(j)} \boldsymbol{y}_{t-i}, \Sigma^{(j)}) m_{t+2,t+1}(j) \tag{16}$$

$$m_{T+1,T}(k) = 1 \quad k = 1, \ldots, L. \tag{17}$$

B. State Sequence Message Passing

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 $\Sigma = \Lambda^{-1}$. The backwards messages for the HDP-SLDS can be recursively defined by

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

For this model, the densities of Eq. 18 can be expressed as

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

The product of these quadratics is given by:

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

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

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

where,

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

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

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

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

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

We initialize at time T with

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

An equivalent, but more numerically stable recursion is summarized in Algorithm 1.

```
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 $t \in \{T-1, \dots, 1\}$:
   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
       $\Lambda_{t+1,t} = A^{(z_{t+1})^T}(\tilde{L}_{t+1}\Lambda^b_{t+1|t+1}\tilde{L}^T_{t+1} + \tilde{J}_{t+1}\Sigma^{-(z_{t+1})}\tilde{J}^T_{t+1})A^{(z_{t+1})}$
       $\theta_{t+1,t} = A^{(z_{t+1})^T}\tilde{L}_{t+1}\theta^b_{t+1|t+1}$
   c) Update
       $\Lambda^b_{t|t} = \Lambda_{t+1,t} + C^T R^{-1} C$
       $\theta^b_{t|t} = \theta_{t+1,t} + C^T R^{-1} \boldsymbol{y}_t$
```

Algorithm 1: Numerically stable form of the backwards Kalman information filter.

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

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

where

$$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) \propto p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, A^{(z_t)}, \Sigma^{(z_t)}) p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, R) m_{t+1,t}(\boldsymbol{x}_t). \tag{18}$$

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

$$\begin{aligned} p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) & \propto \mathcal{N}(\boldsymbol{x}_t; A^{(z_t)}\boldsymbol{x}_{t-1}, \Sigma^{(z_t)}) \mathcal{N}(\boldsymbol{y}_t; C\boldsymbol{x}_t, R) m_{t+1,t}(\boldsymbol{x}_t) \\ & \propto \mathcal{N}(\boldsymbol{x}_t; A^{(z_t)}\boldsymbol{x}_{t-1}, \Sigma^{(z_t)}) \mathcal{N}^{-1}(\boldsymbol{x}_t; \theta^b_{t|t}, \Lambda^b_{t|t}) \\ & \propto \mathcal{N}^{-1}(\boldsymbol{x}_t; \Sigma^{-(z_t)} A^{(z_t)} \boldsymbol{x}_{t-1} + \theta^b_{t|t}, \Sigma^{-(z_t)} + \Lambda^b_{t|t}), \end{aligned} \tag{19}$$

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

$$\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t; (\Sigma^{-(z_t)} + \Lambda^b_{t|t})^{-1}(\Sigma^{-(z_t)} A^{(z_t)} \boldsymbol{x}_{t-1} + \theta^b_{t|t}), (\Sigma^{-(z_t)} + \Lambda^b_{t|t})^{-1}). \tag{20}$$

```
Given a previous set of mode-specific transition probabilities $\boldsymbol{\pi}^{(n-1)}$, the global transition distribution $\beta^{(n-1)}$, the dynamic parameters $\boldsymbol{\theta}^{(n-1)}$, and pseudo-observations $\tilde{\boldsymbol{y}}_{1:T}^{(n-1)}$:
1) Set $\boldsymbol{\pi} = \boldsymbol{\pi}^{(n-1)}$, $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\} = \{\mathbf{A}^{(k)}, \Sigma^{(k)}\}^{(n-1)}$, and $\tilde{\boldsymbol{y}}_{1:T} = \tilde{\boldsymbol{y}}_{1:T}^{(n-1)}$.
2) Calculate messages $m_{t,t-1}(k)$ and the sample mode sequence $z_{1:T}$:
   a) For each $k \in \{1, \dots, L\}$, initialize messages to $m_{T+1,T}(k) = 1$.
   b) For each $t \in \{T, \dots, 1\}$ and $k \in \{1, \dots, L\}$, compute
      $$m_{t,t-1}(k) = \sum_{j=1}^{L} \pi_k(j) \mathcal{N}\left(\tilde{\boldsymbol{y}}_t; \sum_{i=1}^{r} A_i^{(j)} \tilde{\boldsymbol{y}}_{t-i}, \Sigma^{(j)}\right) m_{t+1,t}(j)$$
   c) Working sequentially forward in time, starting with transitions counts $n_{jk} = 0$ for each $(j, k)$:
      i) For each $k \in \{1, \dots, L\}$, compute the probability
         $$f_k(\tilde{\boldsymbol{y}}_t) = \pi_{z_{t-1}}(k) \mathcal{N}\left(\boldsymbol{y}_t; \sum_{i=1}^{r} A_i^{(k)} \tilde{\boldsymbol{y}}_{t-i}, \Sigma^{(k)}\right) m_{t+1,t}(k)$$
      ii) Sample a mode assignment $z_t$ as follows and increment $n_{z_{t-1} z_t}$:
         $$z_t \sim \sum_{k=1}^{L} f_k(\tilde{\boldsymbol{y}}_t) \delta(z_t, k)$$
3) If HDP-AR-HMM, set pseudo-observations $\tilde{\boldsymbol{y}}_{1:T} = \boldsymbol{y}_{1:T}$.
4) If HDP-SLDS, calculate messages $m_{t,t-1}(\boldsymbol{x}_{t-1})$ and the sample state sequence $\boldsymbol{x}_{1:T}$:
   a) Initialize messages to $m_{T+1,T}(\boldsymbol{x}_T) = \mathcal{N}^{-1}(\boldsymbol{x}_T; 0, 0)$.
   b) For each $t \in \{T, \dots, 1\}$, recursively compute $\{\theta_{t|t}^b, \Lambda_{t|t}^b\}$ as in Algorithm 1.
   c) Working sequentially forward in time sample
      $$\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{x}_t; (\Sigma^{-(z_t)} + \Lambda_{t|t}^b)^{-1}(\Sigma^{-(z_t)} A^{(z_t)} \boldsymbol{x}_{t-1} + \theta_{t|t}^b), (\Sigma^{-(z_t)} + \Lambda_{t|t}^b)^{-1}).$$
   d) Set pseudo-observations $\tilde{\boldsymbol{y}}_{1:T} = \boldsymbol{x}_{1:T}$.
5) For each $k \in \{1, \dots, L\}$, compute sufficient statistics using pseudo-observations $\tilde{\boldsymbol{y}}_{1:T}$:
   $$\mathbf{S}_{\bar{y}\bar{y}}^{(k)} = \bar{\mathbf{Y}}^{(k)} \bar{\mathbf{Y}}^{(k)^T} + \boldsymbol{K} \qquad \mathbf{S}_{y\bar{y}}^{(k)} = \mathbf{Y}^{(k)} \bar{\mathbf{Y}}^{(k)^T} + \boldsymbol{MK} \qquad \mathbf{S}_{yy}^{(k)} = \mathbf{Y}^{(k)} \mathbf{Y}^{(k)^T} + \boldsymbol{MKM}^T.$$
6) Sample auxiliary variables $\boldsymbol{m}$, $\boldsymbol{w}$, and $\bar{\boldsymbol{m}}$ and then hyperparameters $\alpha$, $\gamma$, and $\kappa$ as in [5], [12].
7) Update the global transition distribution by sampling
   $$\beta \sim \text{Dir}(\gamma/L + \bar{m}_{.1}, \dots, \gamma/L + \bar{m}_{.L})$$
8) For each $k \in \{1, \dots, L\}$, sample a new transition distribution and dynamic parameters based on the sampled mode assignments and sufficient statistics of the pseudo-observations:
   $$\pi_k \sim \text{Dir}(\alpha\beta_1 + n_{k1}, \dots, \alpha\beta_k + \kappa + n_{kk}, \dots, \alpha\beta_L + n_{kL})$$
   $$\Sigma^{(k)} \sim \text{IW}(\mathbf{S}_{y|\bar{y}}^{(k)} + S_0, \sum_{\ell=1}^{L} n_{k\ell} + n_0)$$
   $$\mathbf{A}^{(k)} \mid \Sigma^{(k)} \sim \mathcal{MN}(\mathbf{A}^{(k)}; \mathbf{S}_{yy}^{(k)} \mathbf{S}_{\bar{y}\bar{y}}^{-(k)}, \Sigma^{-(k)}, \mathbf{S}_{\bar{y}\bar{y}}^{(k)}).$$
   If HDP-SLDS, also sample the measurement noise covariance
   $$R \sim \text{IW}(\sum_{t=1}^{T} (\boldsymbol{y}_t - C\boldsymbol{x}_t)(\boldsymbol{y}_t - C\boldsymbol{x}_t)^T + R_0, T + r_0).$$
9) Fix $\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}$, $\beta^{(n)} = \beta$, $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}$, and $\tilde{\boldsymbol{y}}_{1:T}^{(n)} = \tilde{\boldsymbol{y}}_{1:T}$.
```

Algorithm 2: HDP-SLDS and HDP-AR-HMM Gibbs sampler.