

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:

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

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

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

We define the following notation for Eq. 2:

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

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 $\mathcal{MN}(\mathbf{A}^{(k)}; \mathbf{M}, \Sigma^{(k)}, \mathbf{K})$ on $\mathbf{A}^{(k)}$ given $\Sigma^{(k)}$:

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

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

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

where $\Gamma_d(n/2)$ is the multivariate gamma function and $\mathbf{B}^{-(k)}$ denotes $(\mathbf{B}^{(k)})^{-1}$ for some matrix \mathbf{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, \mathbf{y}_t , is conditionally Gaussian given the lag observations, $\bar{\mathbf{y}}_t = [\mathbf{y}_{t-1}^T \dots \mathbf{y}_{t-r}^T]^T$, we have

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

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

$$p(\mathbf{D}^{(k)}, \mathbf{A}^{(k)} | \Sigma^{(k)}) = p(\mathbf{D}^{(k)} | \mathbf{A}^{(k)}, \Sigma^{(k)}) p(\mathbf{A}^{(k)} | \Sigma^{(k)}). \quad (7)$$

Using the fact that both the likelihood term $p(\mathbf{D}^{(k)} | \mathbf{A}^{(k)}, \Sigma^{(k)})$ and the prior $p(\mathbf{A}^{(k)} | \Sigma^{(k)})$ are matrix-normally

distributed sharing a common parameter $\Sigma^{(k)}$, we have

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

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

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

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

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

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

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

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

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

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 $\mathbf{x}_{1:T}$.

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 $\mathbf{y}_{1:T}$, plus r initial observations $\mathbf{y}_{1-r:0}$, we first note that the chain rule and Markov structure allows us to decompose the joint distribution as follows:

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

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

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

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:

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

The general conditional distribution of z_t is:

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

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

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

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

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

B. State Sequence Message Passing

A similar sampling scheme is used for generating samples of the state sequence $\mathbf{x}_{1:T}$. Although we now have a continuous state space, the computation of the backwards messages $m_{t+1,t}(\mathbf{x}_t)$ is still analytically feasible since we are working with Gaussian densities. Assume, $m_{t+1,t}(\mathbf{x}_t) \propto \mathcal{N}^{-1}(\mathbf{x}_t; \theta_{t+1,t}, \Lambda_{t+1,t})$, 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

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

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

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

The product of these quadratics is given by:

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

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

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

where,

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

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

$$\begin{aligned}
\Lambda_{t|t}^b &= C^T R^{-1} C + \Lambda_{t+1,t} \\
\theta_{t|t}^b &= C^T R^{-1} \mathbf{y}_t + \theta_{t+1,t}
\end{aligned}$$

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

$$\begin{aligned}
\Lambda_{t-1|t-1}^b &= C^T R^{-1} C + A^{(z_t)T} \Sigma^{-(z_t)} A^{(z_t)} - A^{(z_t)T} \Sigma^{-(z_t)} (\Sigma^{-(z_t)} + C^T R^{-1} C + \Lambda_{t+1,t})^{-1} \Sigma^{-(z_t)} A^{(z_t)} \\
&= C^T R^{-1} C + A^{(z_t)T} \Sigma^{-(z_t)} A^{(z_t)} - A^{(z_t)T} \Sigma^{-(z_t)} (\Sigma^{-(z_t)} + \Lambda_{t|t}^b)^{-1} \Sigma^{-(z_t)} A^{(z_t)} \\
\theta_{t-1|t-1}^b &= C^T R^{-1} \mathbf{y}_{t-1} + A^{(z_t)T} \Sigma^{-(z_t)} (\Sigma^{-(z_t)} + C^T R^{-1} C + \Lambda_{t+1,t})^{-1} (C^T R^{-1} \mathbf{y}_t + \theta_{t+1,t}) \\
&= C^T R^{-1} \mathbf{y}_{t-1} + A^{(z_t)T} \Sigma^{-(z_t)} (\Sigma^{-(z_t)} + \Lambda_{t|t}^b)^{-1} \theta_{t|t}^b
\end{aligned}$$

We initialize at time T with

$$\begin{aligned}
\Lambda_{T|T}^b &= C^T R^{-1} C \\
\theta_{T|T}^b &= C^T R^{-1} \mathbf{y}_T
\end{aligned}$$

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

1) Initialize filter with

$$\Lambda_{T|T}^b = C^T R^{-1} C$$

$$\theta_{T|T}^b = C^T R^{-1} \mathbf{y}_T$$

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

a) Compute

$$\tilde{J}_{t+1} = \Lambda_{t+1|t+1}^b (\Lambda_{t+1|t+1}^b + \Sigma^{-(z_{t+1})})^{-1}$$

$$\tilde{L}_{t+1} = I - \tilde{J}_{t+1}.$$

b) Predict

$$\Lambda_{t+1,t} = A^{(z_{t+1})^T} (\tilde{L}_{t+1} \Lambda_{t+1|t+1}^b \tilde{L}_{t+1}^T + \tilde{J}_{t+1} \Sigma^{-(z_{t+1})} \tilde{J}_{t+1}^T) A^{(z_{t+1})}$$

$$\theta_{t+1,t} = A^{(z_{t+1})^T} \tilde{L}_{t+1} \theta_{t+1|t+1}^b$$

c) Update

$$\Lambda_{t|t}^b = \Lambda_{t+1,t} + C^T R^{-1} C$$

$$\theta_{t|t}^b = \theta_{t+1,t} + C^T R^{-1} \mathbf{y}_t$$

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

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

$$p(\mathbf{x}_{1:T} | \mathbf{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) = p(\mathbf{x}_T | \mathbf{x}_{T-1}, \mathbf{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) p(\mathbf{x}_{T-1} | \mathbf{x}_{T-2}, \mathbf{y}_{1:T}, z_{1:T}, \boldsymbol{\theta})$$

$$\cdots p(\mathbf{x}_2 | \mathbf{x}_1, \mathbf{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) p(\mathbf{x}_1 | \mathbf{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}).$$

where

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

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

$$p(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{y}_{1:T}, z_{1:T}, \boldsymbol{\theta}) \propto \mathcal{N}(\mathbf{x}_t; A^{(z_t)} \mathbf{x}_{t-1}, \Sigma^{(z_t)}) \mathcal{N}(\mathbf{y}_t; C \mathbf{x}_t, R) m_{t+1,t}(\mathbf{x}_t)$$

$$\propto \mathcal{N}(\mathbf{x}_t; A^{(z_t)} \mathbf{x}_{t-1}, \Sigma^{(z_t)}) \mathcal{N}^{-1}(\mathbf{x}_t; \theta_{t|t}^b, \Lambda_{t|t}^b)$$

$$\propto \mathcal{N}^{-1}(\mathbf{x}_t; \Sigma^{-(z_t)} A^{(z_t)} \mathbf{x}_{t-1} + \theta_{t|t}^b, \Sigma^{-(z_t)} + \Lambda_{t|t}^b), \quad (19)$$

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

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t; (\Sigma^{-(z_t)} + \Lambda_{t|t}^b)^{-1} (\Sigma^{-(z_t)} A^{(z_t)} \mathbf{x}_{t-1} + \theta_{t|t}^b), (\Sigma^{-(z_t)} + \Lambda_{t|t}^b)^{-1}). \quad (20)$$

Given a previous set of mode-specific transition probabilities $\pi^{(n-1)}$, the global transition distribution $\beta^{(n-1)}$, the dynamic parameters $\theta^{(n-1)}$, and pseudo-observations $\tilde{\mathbf{y}}_{1:T}^{(n-1)}$:

- 1) Set $\pi = \pi^{(n-1)}$, $\{\mathbf{A}^{(k)}, \Sigma^{(k)}\} = \{\mathbf{A}^{(k)}, \Sigma^{(k)}\}^{(n-1)}$, and $\tilde{\mathbf{y}}_{1:T} = \tilde{\mathbf{y}}_{1:T}^{(n-1)}$.
- 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

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

- 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

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

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

$$z_t \sim \sum_{k=1}^L f_k(\tilde{\mathbf{y}}_t) \delta(z_t, k)$$

- 3) If HDP-AR-HMM, set pseudo-observations $\tilde{\mathbf{y}}_{1:T} = \mathbf{y}_{1:T}$.
- 4) If HDP-SLDS, calculate messages $m_{t,t-1}(\mathbf{x}_{t-1})$ and the sample state sequence $\mathbf{x}_{1:T}$:
 - a) Initialize messages to $m_{T+1,T}(\mathbf{x}_T) = \mathcal{N}^{-1}(\mathbf{x}_T; 0, 0)$.
 - 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

$$\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t; (\Sigma^{-(z_t)} + \Lambda_{t|t}^b)^{-1} (\Sigma^{-(z_t)} A^{(z_t)} \mathbf{x}_{t-1} + \theta_{t|t}^b), (\Sigma^{-(z_t)} + \Lambda_{t|t}^b)^{-1}).$$

- d) Set pseudo-observations $\tilde{\mathbf{y}}_{1:T} = \mathbf{x}_{1:T}$.

- 5) For each $k \in \{1, \dots, L\}$, compute sufficient statistics using pseudo-observations $\tilde{\mathbf{y}}_{1:T}$:

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

- 6) Sample auxiliary variables \mathbf{m} , \mathbf{w} , and $\tilde{\mathbf{m}}$ and then hyperparameters α , γ , and κ as in [5], [12].
- 7) Update the global transition distribution by sampling

$$\beta \sim \text{Dir}(\gamma/L + \tilde{\mathbf{m}}_{\cdot 1}, \dots, \gamma/L + \tilde{\mathbf{m}}_{\cdot L})$$

- 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:

$$\pi_k \sim \text{Dir}(\alpha\beta_1 + n_{k1}, \dots, \alpha\beta_k + \kappa + n_{kk}, \dots, \alpha\beta_L + n_{kL})$$

$$\Sigma^{(k)} \sim \text{IW}(\mathbf{S}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}^{(k)} + S_0, \sum_{\ell=1}^L n_{k\ell} + n_0)$$

$$\mathbf{A}^{(k)} | \Sigma^{(k)} \sim \mathcal{MN}(\mathbf{A}^{(k)}; \mathbf{S}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}^{(k)} \mathbf{S}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}^{-(k)}, \Sigma^{-(k)}, \mathbf{S}_{\tilde{\mathbf{y}}\tilde{\mathbf{y}}}^{(k)}).$$

If HDP-SLDS, also sample the measurement noise covariance

$$\mathbf{R} \sim \text{IW}\left(\sum_{t=1}^T (\mathbf{y}_t - C\mathbf{x}_t)(\mathbf{y}_t - C\mathbf{x}_t)^T + R_0, T + r_0\right).$$

- 9) Fix $\pi^{(n)} = \pi$, $\beta^{(n)} = \beta$, $\theta^{(n)} = \theta$, and $\tilde{\mathbf{y}}_{1:T}^{(n)} = \tilde{\mathbf{y}}_{1:T}$.

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