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.

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 (20), we decompose the prior over the lag blocks $A_i^{(k)}$

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

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 (20) and (22) on a vectorization of $\mathbf{A}^{(k)}$. Our ARD prior on $\mathbf{A}^{(k)}$ is equivalent to a $\mathcal{N}\left(0, \Sigma_0^{(k)}\right)$ prior on vec $\left(\mathbf{A}^{(k)}\right)$, where

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

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

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

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

$$\begin{aligned} p&\left(\mathrm{vec}\left(\mathbf{A}^{(k)}\right) \mid \mathbf{D}^{(k)}, \Sigma^{(k)}, \Sigma_0^{(k)}\right) \\ &= \mathcal{N}^{-1}\left(\sum_{t|z_t=k} \tilde{\Psi}_{t-1}^T \Sigma^{-(k)} \boldsymbol{\psi}_t, \right. \\ &\qquad \left. \Sigma_0^{-(k)} + \sum_{t|z_t=k} \tilde{\Psi}_{t-1}^T \Sigma^{-(k)} \tilde{\Psi}_{t-1}\right). \end{aligned} \quad (25)$$

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

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

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) remains the same. Thus, the gamma prior is an informative prior and the choice of a and b should depend upon the cardinality of $\mathcal{S}_\ell$ (see Section IV-B, for an example). For the HDP-SLDS, this cardinality is given by the maximal state dimension n and for the HDP-AR-HMM, by the square of the observation dimensionality d^2.

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

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

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

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

4) 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

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

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 [43]. We describe the sampler in terms of the pseudo-observations $\boldsymbol{\psi}_t$, as defined by (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 Section III-A. For the ARD prior, we then sample $\boldsymbol{\alpha}^{(k)}$ given $\mathbf{A}^{(k)}$. In practice we iterate multiple