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

	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	$\bar{\boldsymbol{\psi}}_t = [\boldsymbol{y}_{t-1}^T \dots \boldsymbol{y}_{t-r}^T]^T$	$\bar{\boldsymbol{\psi}}_t = \boldsymbol{x}_{t-1}$.

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 [40] is conjugate to the likelihood model defined in (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 (15) are independent conditioned on the mode sequence $z_{1:T}$. We note that while the MNIW prior does not enforce stability constraints on each mode, this prior is still a reasonable choice since each mode need not have stable dynamics for the SLDS to be stable [41] and conditioned on data from a stable mode, the posterior distribution will likely be sharply peaked around stable dynamic matrices.

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

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

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

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

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

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

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

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

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

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

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

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

Each precision parameter $\alpha_j^{(k)}$ is given a $\mathrm{Gamma}(a, b)$ prior. The zero-mean Gaussian prior penalizes nonzero 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

$$\begin{aligned} \mathcal{H} = [\, C^T \quad CA^T \quad \cdots \quad (CA^{d-1})^T \,]^T \\ \times [\, G \quad AG \quad \cdots \quad A^{d-1}G \,] \equiv \mathcal{OR} \end{aligned} \tag{21}$$

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.

3) 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 jth 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 $C = [\, I_d \quad 0 \,]$ does not interfere with learning a sparse realization. We could avoid restricting our attention to models satisfying Criterion 3.1 by considering a more general model where the measurement

[2]If $A \sim \mathcal{MN}(M, V, K)$, then $\mathrm{vec}(A) \sim \mathcal{N}(\mathrm{vec}(M), K^{-1} \otimes V)$, with $\otimes$ denoting the Kronecker product.