We similarly define a *switching* $\mathrm{VAR}(r)$ process by

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

B. Dirichlet Processes and the Sticky HDP-HMM

To examine a Bayesian nonparametric SLDS and thus relax the assumption that the number of dynamical modes is known and fixed, it is useful to first analyze such methods for the simpler HMM. One can equivalently represent the finite HMM of (4) via a set of *transition probability measures* $G_j = \sum_{k=1}^{K} \pi_{jk} \delta_{\theta_k}$, where δ_θ is a mass concentrated at θ. We then operate directly in the parameter space Θ and transition between emission parameters with probabilities given by $\{G_j\}$. That is

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

Here, $\theta'_t \in \{\theta_1, \ldots, \theta_K\}$ and is equivalent to θ_{z_t} of (4). A Bayesian nonparametric HMM takes G_j to be *random*[1] with an infinite collection of atoms corresponding to the infinite HMM mode space.

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

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

on a parameter space Θ that is endowed with a *base measure* H. The weights are sampled via a *stick-breaking construction* [38]:

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

In effect, we have divided a unit-length stick into lengths given by the weights β_k: the k^{th} weight is a random proportion ν_k of the remaining stick after the previous $(k-1)$ weights have been defined. Letting $\beta = [\beta_1 \quad \beta_2 \quad \ldots]$, we denote this distribution by $\beta \sim \mathrm{GEM}(\gamma)$.

The DP has proven useful in many applications due to its clustering properties, which are clearly seen by examining the *predictive distribution* of draws $\theta'_i \sim G_0$. Because probability measures drawn from a DP are discrete, there is a strictly positive probability of multiple observations θ'_i taking identical values within the set $\{\theta_k\}$, with θ_k defined as in (7). For each value θ'_i, let z_i be an indicator random variable that picks out the unique value θ_k such that $\theta'_i = \theta_{z_i}$. Blackwell and MacQueen [39] introduced a Pólya urn representation of the θ'_i

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

[1]Formally, a random measure on a measurable space Θ with sigma algebra $\mathcal{A}$ is defined as a stochastic process whose index set is $\mathcal{A}$. That is, $G(A)$ is a random variable for each $A \in \mathcal{A}$.

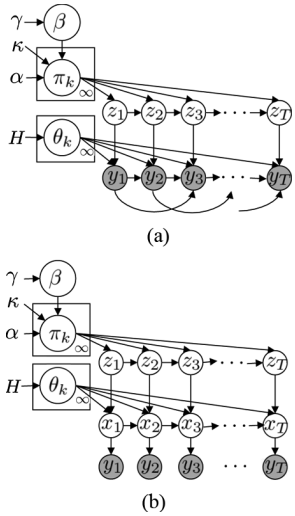

Fig. 1. Sticky HDP-HMM prior on (a) switching VAR(2) and (b) SLDS processes with the mode evolving as $z_{t+1} | \{\pi_k\}_{k=1}^{\infty}, z_t \sim \pi_{z_t}$ for $\pi_k | \alpha, \kappa, \beta \sim \mathrm{DP}(\alpha + \kappa, (\alpha\beta + \kappa\delta_k)/(\alpha + \kappa))$. Here, $\beta \mid \gamma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

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

A hierarchical extension of the DP, the hierarchical Dirichlet process (HDP) [16], has proven useful in defining a prior on the set of HMM transition probability measures G_j. The HDP defines a collection of probability measures $\{G_j\}$ on the same support points $\{\theta_1, \theta_2, \ldots\}$ by assuming that each discrete measure G_j is a variation on a global discrete measure G_0. Specifically, the Bayesian hierarchical specification takes $G_j \sim \mathrm{DP}(\alpha, G_0)$, with G_0 itself a draw from a $\mathrm{DP}(\gamma, H)$. Through this construction, one can show that the probability measures are described as

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

Here, we use the notation $\pi_j = [\pi_{j1} \quad \pi_{j2} \quad \ldots]$. Applying the HDP prior to the HMM, we obtain the *HDP-HMM* of Teh *et al.* [16]. This corresponds to the model in Fig. 1(a), but without the edges between the observations.

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

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

However, it does not differentiate self-transitions from moves between modes. When modeling dynamical processes with mode persistence, the flexible nature of the HDP-HMM prior allows for mode sequences with unrealistically fast dynamics to have large posterior probability. Recently, it has been shown