the L possible states. See Section 6.3 for a comparison with standard parametric modeling.

There are multiple methods of approximating the countably infinite transition distributions via truncations. One approach is to terminate the stick-breaking construction after some portion of the stick has already been broken and assign the remaining weight to a single component. This approximation is referred to as the *truncated Dirichlet process*. Another method is to consider the *degree L weak limit approximation* to the DP [Ishwaran and Zarepour (2002c)],

$$\mathrm{GEM}_L(\alpha) \triangleq \mathrm{Dir}(\alpha/L, \dots, \alpha/L), \tag{5.4}$$

where L is a number that exceeds the total number of expected HMM states. Both of these approximations, which are presented in Ishwaran and Zarepour (2000a, 2002c), encourage the learning of models with fewer than L components while allowing the generation of new components, upper bounded by L, as new data are observed. We choose to use the second approximation because of its simplicity and computational efficiency. The two choices of approximations are compared in Kurihara, Welling and Teh (2007), and little to no practical differences are found. Using a weak limit approximation to the Dirichlet process prior on β (i.e., employing a finite Dirichlet prior) induces a finite Dirichlet prior on π_j:

$$\beta|\gamma \sim \mathrm{Dir}(\gamma/L, \dots, \gamma/L), \tag{5.5}$$

$$\pi_j|\alpha, \beta \sim \mathrm{Dir}(\alpha\beta_1, \dots, \alpha\beta_L). \tag{5.6}$$

As $L \to \infty$, this model converges in distribution to the HDP mixture model [Teh et al. (2006)].

The Gibbs sampler using blocked resampling of $z_{1:T}$ is derived in the Supplementary Material [Fox et al. (2010)]; an outline of the resulting algorithm is also presented (see Algorithm 3). A similar sampler has been used for inference in HDP hidden Markov trees [Kivinen, Sudderth and Jordan (2007)]. However, this work did not consider the complications introduced by multimodal emissions, which we explore in Section 7.

The blocked sampler is initialized by drawing L parameters θ_k from the base measure, β from its L-dimensional symmetric Dirichlet prior, and the L transition distributions π_k from the induced L-dimensional Dirichlet prior specified in equation (5.5). The hyperparameters are also drawn from the prior. Based on the sampled parameters and transition distributions, one can block sample $z_{1:T}$ and proceed as in Algorithm 3 of the Supplementary Material [Fox et al. (2010)].

5.3. *Hyperparameters*. We treat the hyperparameters in the sticky HDP-HMM as unknown quantities and perform full Bayesian inference over these quantities. This emphasizes the role of the data in determining the number of occupied states and the degree of self-transition bias. Our derivation of sampling updates for the hyperparameters of the sticky HDP-HMM is presented in the Supplementary