MIT LABORATORY FOR INFORMATION & DECISION SYSTEMS TECHNICAL REPORT P-2777
NOVEMBER 5, 2007, REVISED MAY 13, 2009

THE STICKY HDP-HMM: BAYESIAN NONPARAMETRIC HIDDEN MARKOV MODELS WITH PERSISTENT STATES

BY EMILY B. FOX[†], ERIK B. SUDDERTH[‡] MICHAEL I. JORDAN[‡] AND ALAN S. WILLSKY[†]

Massachusetts Institute of Technology[†] *and University of California, Berkeley*[‡]

We consider the problem of *speaker diarization*, the problem of segmenting an audio recording of a meeting into temporal segments corresponding to individual speakers. The problem is rendered particularly difficult by the fact that we are not allowed to assume knowledge of the number of people participating in the meeting. To address this problem, we take a Bayesian nonparametric approach to speaker diarization that builds on the hierarchical Dirichlet process hidden Markov model (HDP-HMM) of Teh et al. (2006). Although the basic HDP-HMM tends to over-segment the audio data—creating redundant states and rapidly switching among them—we describe an augmented HDP-HMM that provides effective control over the switching rate. We also show that this augmentation makes it possible to treat emission distributions nonparametrically. To scale the resulting architecture to realistic diarization problems, we develop a sampling algorithm that employs a truncated approximation of the Dirichlet process to jointly resample the full state sequence, greatly improving mixing rates. Working with a benchmark NIST data set, we show that our Bayesian nonparametric architecture yields state-of-the-art speaker diarization results.

1. Introduction. A recurring problem in many areas of information technology is that of segmenting a waveform into a set of time intervals that have a useful interpretation in some underlying domain. In this article we focus on a particular instance of this problem, namely the problem of *speaker diarization*. In speaker diarization, an audio recording is made of a meeting involving multiple human participants and the problem is to segment the recording into time intervals associated with individual speakers (Wooters and Huijbregts, 2007). This segmentation is to be carried out without a priori knowledge of the number of speakers involved in the meeting; moreover, we do not assume that we have a priori knowledge of the speech patterns of particular individuals.

Our approach to the speaker diarization problem is built on the framework of hidden Markov models (HMMs), which have been a major success story not only in speech technology but also in many other fields involving complex sequential data,

[*] This work was supported in part by MURIs funded through AFOSR Grant FA9550-06-1-0324 and ARO Grant W911NF-06-1-0076, by AFOSR under Grant FA9559-08-1-0180, and by DARPA IPTO Contract FA8750-05-2-0249.

Keywords and phrases: Bayesian nonparametrics, hierarchical Dirichlet processes, hidden Markov models, speaker diarization

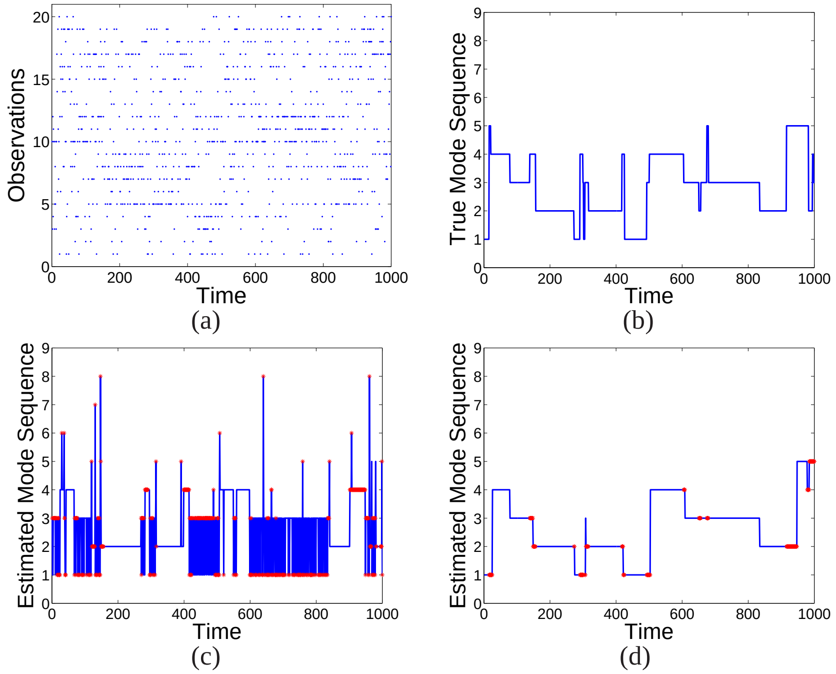
including genomics, structural biology, machine translation, cryptanalysis and finance. An alternative to HMMs in the speaker diarization setting would be to treat the problem as a changepoint detection problem, but a key aspect of speaker diarization is that speech data from a single individual generally recurs in multiple disjoint intervals. This suggests a Markovian framework in which the model transitions among states that are associated with the different speakers.

An apparent disadvantage of the HMM framework, however, is that classical treatments of the HMM generally require the number of states to be fixed a priori. While standard parametric model selection methods can be adapted to the HMM, there is little understanding of the strengths and weaknesses of such methods in this setting, and practical applications of HMMs generally fix the number of states using ad hoc approaches. It is not clear how to adapt HMMs to the diarization problem where the number of speakers is unknown.

In recent work, Teh et al. (2006) presented a Bayesian nonparametric version of the HMM in which a stochastic process—the *hierarchical Dirichlet process* (HDP)—defines a prior distribution on transition matrices over countably infinite state spaces. The resulting *HDP-HMM* is amenable to full Bayesian posterior inference over the number of states in the model. Moreover, this posterior distribution can be integrated over when making predictions, effectively averaging over models of varying complexity. The HDP-HMM has shown promise in a variety of applied problems, including visual scene recognition (Kivinen et al., 2007), music synthesis (Hoffman et al., 2008), and the modeling of genetic recombination (Xing and Sohn, 2007) and gene expression (Beal and Krishnamurthy, 2006).

While the HDP-HMM seems like a natural fit to the speaker diarization problem given its structural flexibility, as we show in Sec. 7, the HDP-HMM does not yield state-of-the-art performance in the speaker diarization setting. The problem is that the HDP-HMM inadequately models the temporal persistence of states. This problem arises in classical finite HMMs as well, where semi-Markovian models are often proposed as solutions. However, the problem is exacerbated in the nonparametric setting, in which the Bayesian bias towards simpler models is insufficient to prevent the HDP-HMM from giving high posterior probability to models with unrealistically rapid switching. This is demonstrated in Fig. 1, where we see that the HDP-HMM sampling algorithm creates redundant states and rapidly switches among them. (The figure also displays results from the augmented HDP-HMM—the "sticky HDP-HMM" that we describe in this paper.) The tendency to create redundant states is not necessarily a problem in settings in which model averaging is the goal. For speaker diarization, however, it is critical to infer the number of speakers as well as the transitions among speakers.

Thus, one of our major goals in this paper is to provide a general solution to the problem of state persistence in HDP-HMMs. Our approach is easily stated—we



(a)

(b)

(c)

(d)

FIG 1. *(a) Multinomial observation sequence; (b) true state sequence; (c)-(d) estimated state sequence after 30,000 Gibbs iterations for the original and sticky HDP-HMM, respectively, with errors indicated in red. Without an extra self-transition bias, the HDP-HMM rapidly transitions among redundant states.*

simply augment the HDP-HMM to include a parameter for self-transition bias, and place a separate prior on this parameter. The challenge is to execute this idea coherently in a Bayesian nonparametric framework. Earlier papers have also proposed self-transition parameters for HMMs with infinite state spaces (Beal et al., 2002; Xing and Sohn, 2007), but did not formulate general solutions that integrate fully with Bayesian nonparametric inference.

Another goal of the current paper is to develop a more fully nonparametric version of the HDP-HMM in which not only the transition distribution but also the emission distribution (the conditional distribution of observations given states) is treated nonparametrically. This is again motivated by the speaker diarization problem—in classical applications of HMMs to speech recognition problems it is often the case that emission distributions are found to be multimodal, and high-performance HMMs generally use finite Gaussian mixtures as emission distributions (Gales and Young, 2008). In the nonparametric setting it is natural to replace these finite mixtures with Dirichlet process mixtures. Unfortunately, this idea is not viable in practice, because of the tendency of the HDP-HMM to rapidly switch between redundant states. As we show, however, by incorporating an additional self-transition bias it is possible to make use of Dirichlet process mixtures for the emission distributions.

An important reason for the popularity of the classical HMM is its computational tractability. In particular, marginal probabilities and samples can be obtained from the HMM via an efficient dynamic programming algorithm known as the forward-backward algorithm (Rabiner, 1989). We show that this algorithm also plays an important role in computationally efficient inference for our generalized HDP-HMM. In particular, we develop a blocked Gibbs sampler which leverages forward–backward recursions to jointly resample the state and emission assignments for all observations.

The paper is organized as follows. In Sec. 2, we begin by summarizing some of the basic background on Dirichlet processes. Then, in Sec. 3, we briefly describe the hierarchical Dirichlet process and, in Sec. 4, discuss how it applies to HMMs and can be extended to account for state persistence. An efficient Gibbs sampler is also described in this section. In Sec. 6, we treat the case of nonparametric emission distributions. We discuss our application to speaker diarization in Sec. 7. A list of notational conventions can be found in the Supplementary Material.

2. Dirichlet Processes. A Dirichlet process (DP) is a distribution on probability measures on a measurable space Θ. This stochastic process is uniquely defined by a base measure H on Θ and a concentration parameter γ; we denote it by $\text{DP}(\gamma, H)$. Consider a random probability measure $G_0 \sim \text{DP}(\gamma, H)$. The DP is formally defined by the property that for any finite partition$\{A_1, \dots, A_K\}$ of Θ,

$$(G_0(A_1), \dots, G_0(A_K)) \mid \gamma, H \sim \text{Dir}(\gamma H(A_1), \dots, \gamma H(A_K)). \tag{2.1}$$

That is, the measure of a random probability distribution $G_0 \sim \text{DP}(\gamma, H)$ on every finite partition of Θ follows a finite-dimensional Dirichlet *distribution*. This definition of the DP is due to Ferguson (1973), who invoked Kolmogorov's consistency conditions to establish the existence of the DP as a stochastic process with Dirichlet marginals. A more constructive definition of the DP was given by Sethuraman (1994). Consider a probability mass function (pmf) $\{\beta_k\}_{k=1}^{\infty}$ on a countably infinite set, where the discrete probabilities are defined as follows:

$$\begin{aligned} v_k \mid \gamma &\sim \text{Beta}(1, \gamma) \qquad k = 1, 2, \dots \\ \beta_k &= v_k \prod_{\ell=1}^{k-1} (1 - v_\ell) \quad k = 1, 2, \dots . \end{aligned} \tag{2.2}$$

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 v_k of the remaining stick after the previous

$(k-1)$ weights have been defined. This *stick-breaking construction* is generally denoted by $\beta \sim \text{GEM}(\gamma)$. With probability one, a random draw $G_0 \sim DP(\gamma, H)$ can be expressed as

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \qquad \theta_k \mid H \sim H, \quad k = 1, 2, \dots, \tag{2.3}$$

where δ_θ denotes a unit-mass measure concentrated at θ. From this definition, we see that the DP actually defines a distribution over discrete probability measures. The stick-breaking construction also gives us insight into how the concentration parameter γ controls the relative proportion of the mixture weights β_k, and thus determines the model complexity in terms of the expected number of components with significant probability mass.

The DP has a number of properties which make inference based on this nonparametric prior computationally tractable. Consider a set of observations $\{\theta_i'\}$ with $\theta_i' \sim G_0$. 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 Eq. (2.3). 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 (1973) introduced a Pólya urn representation of the θ_i':

$$\begin{aligned} \theta_i' \mid \theta_1', \dots, \theta_{i-1}' &\sim \frac{\gamma}{\gamma + i - 1} H + \sum_{j=1}^{i-1} \frac{1}{\gamma + i - 1} \delta_{\theta_j'} \\ &\sim \frac{\gamma}{\gamma + i - 1} H + \sum_{k=1}^{K} \frac{N_k}{\gamma + i - 1} \delta_{\theta_k}, \end{aligned} \tag{2.4}$$

implying the following predictive distribution on the indicator random variables:

$$p(z_{N+1} = z \mid z_1, \dots, z_N, \gamma) = \frac{\gamma}{N + \gamma} \delta(z, K+1) + \frac{1}{N + \gamma} \sum_{k=1}^{K} N_k \delta(z, k). \tag{2.5}$$

Here, $N_k = \sum_{i=1}^{N} \delta(z_i, k)$ is the number of indicator random variables taking the value k, and $K + 1$ is a previously unseen value. We use the notation $\delta(z, k)$ to indicate the discrete Kronecker delta. This representation can be used to sample observations from a DP without explicitly constructing the countably infinite random probability measure $G_0 \sim \text{DP}(\gamma, H)$.

The distribution on partitions induced by the sequence of conditional distributions in Eq. (2.5) is commonly referred to as the *Chinese restaurant process*. The analogy, which is useful in developing various generalizations of the Dirichlet process we consider in this paper, is as follows. Take θ_i' to be a customer entering a restaurant with infinitely many tables, each serving a unique dish θ_k. Each arriving customer chooses a table, indicated by z_i, in proportion to how many customers are currently sitting at that table. With some positive probability proportional to γ, the customer starts a new, previously unoccupied table $K+1$. From the Chinese restaurant process, we see that the DP has a reinforcement property that leads to a clustering at the values θ_k.

From Eq. (2.5) we see that when $z_i \sim \beta$ and $\beta \sim \text{GEM}(\gamma)$, we can integrate out β to determine a closed-form predictive distribution for z_i. We can also find the distribution of the number of unique values of z_i resulting from N draws from the measure β. Letting K be the number of unique values of $\{z_1, \dots, z_N\}$, this distribution is given by (Antoniak, 1974):

$$p(K \mid N, \gamma) = \frac{\Gamma(\gamma)}{\Gamma(\gamma+N)} s(N, K)\gamma^K, \tag{2.6}$$

where $s(n, m)$ are unsigned Stirling numbers of the first kind.

The DP is commonly used as a prior on the parameters of a mixture model with a random number of components. Such a model is called a *Dirichlet process mixture model* and is depicted as a graphical model in Fig. 2(a)-(b). To generate observations, we choose $\theta_i' \sim G_0$ and $y_i \sim F(\theta_i')$ for an indexed family of distributions $F(\cdot)$. This sampling process is also often described in terms of the indicator random variables z_i; in particular, we have $z_i \sim \beta$ and $y_i \sim F(\theta_{z_i})$. The parameter with which an observation is associated implicitly partitions or clusters the data. In addition, the Chinese restaurant process representation indicates that the DP provides a prior that makes it more likely to associate an observation with a parameter to which other observations have already been associated. This reinforcement property is essential for inferring finite, compact mixture models. It can be shown under mild conditions that if the data were generated by a finite mixture, then the DP posterior is guaranteed to converge (in distribution) to that finite set of mixture parameters (Ishwaran and Zarepour, 2002a).

Finally, we can also obtain the DP mixture model as the limit of a sequence of finite mixture models. Let us assume that there are L components in a finite mixture model and we place a finite-dimensional Dirichlet prior on these mixture weights:

$$\beta \mid \gamma \sim \text{Dir}(\gamma/L, \dots, \gamma/L). \tag{2.7}$$

Let $G_0^L = \sum_{k=1}^L \beta_k \delta_{\theta_k}$. Then, it can be shown (Ishwaran and Zarepour, 2002b, 2000) that for every measurable function f integrable with respect to the measure H, this finite distribution G_0^L converges weakly to a countably infinite distribution G_0 distributed according to a Dirichlet *process*. That is, as $L \to \infty$,

$$\int_\theta f(\theta) dG_0^L(\theta) \xrightarrow{\mathcal{D}} \int_\theta f(\theta) dG_0(\theta), \quad G_0 \sim DP(\gamma, H). \tag{2.8}$$

3. Hierarchical Dirichlet Processes. There are many scenarios in which groups of data are thought to be produced by related, yet distinct, generative processes. For example, take a sensor network monitoring an environment where time-varying conditions may influence the quality of the data. Data collected under certain conditions should be grouped and described by a similar, but different model from that of other data. The hierarchical Dirichlet process (HDP) (Teh et al., 2006) extends the DP to such scenarios by taking a hierarchical Bayesian approach: a global Dirichlet process prior $\mathrm{DP}(\alpha, G_0)$ is placed on Θ and group-specific distributions are drawn from a global prior, $G_j \sim \mathrm{DP}(\alpha, G_0)$, where the base measure G_0 acts as an "average" distribution across all groups; indeed, we have $E[G_j \mid G_0] = G_0$. The base measure G_0 is itself distributed according to a Dirichlet process $\mathrm{DP}(\gamma, H)$, implying that atoms are shared not only within groups, but also between groups. If the base measure G_0 were instead fixed and absolutely continuous with respect to Lebesgue measure, there would be zero probability of the group-specific distributions having overlapping support.

We now describe the HDP more formally. Let $\{y_{j1}, \dots, y_{jN_j}\}$ be the set of observations in group j. We assume there are J such groups of data. Then, the generative model, depicted in Fig. 2(d), can be written as:

$$\begin{array}{lll} G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} & \beta \mid \gamma \sim \mathrm{GEM}(\gamma) & \\ & \theta_k \mid H, \lambda \sim H(\lambda) & k = 1, 2, \dots \\ & & \\ G_j = \sum_{t=1}^{\infty} \tilde{\pi}_{jt} \delta_{\theta^*_{jt}} & \tilde{\pi}_j \mid \alpha \sim \mathrm{GEM}(\alpha) & j = 1, \dots, J \\ & \theta^*_{jt} \mid G_0 \sim G_0 & t = 1, 2, \dots \\ & & \\ \theta'_{ji} \mid G_j \sim G_j & y_{ji} \mid \theta'_{ji} \sim F(\theta'_{ji}) & j = 1, \dots, J \\ & & i = 1, \dots, N_j. \end{array} \tag{3.1}$$

Teh et al. (2006) have also described the marginal probabilities obtained from integrating over the random measures G_0 and G_j. They show that these marginals can be described in terms of a *Chinese restaurant franchise* (CRF) that is an analog of the Chinese restaurant process. The CRF is comprised of J restaurants, each corresponding to an HDP group, and an infinite buffet line of dishes common to all restaurants. The process of seating customers at tables, however, is restaurant specific. We introduce indicator variables t_{ji} and k_{jt} to represent table and dish assignments. There are J restaurants (groups), each with infinitely many tables (clusters) at which customers (observations) sit. Each customer is pre-assigned to a given restaurant determined by that customer's group j. The table assignment for the i^{th} customer in the j restaurant is chosen as $t_{ji} \sim \tilde{\pi}_j$, and each table is assigned a dish (parameter) via $k_{jt} \sim \beta$. One can think of β as a set of ratings for the dishes served in the buffet line. Observation y_{ji} is then generated by global parameter

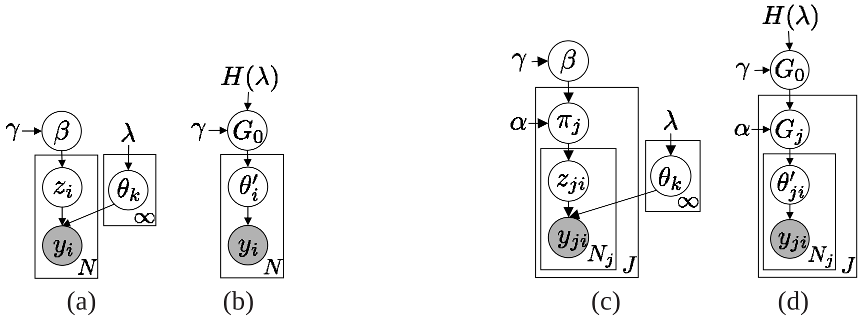


FIG 2. *Dirichlet process (left) and hierarchical Dirichlet process (right) mixture models represented in two different ways as graphical models. (a) Indicator variable representation in which $\beta|\gamma \sim$ GEM(γ), $\theta_k|H,\lambda \sim H(\lambda)$, $z_i|\beta \sim \beta$, and $y_i|\{\theta_k\}_{k=1}^{\infty}, z_i \sim F(\theta_{z_i})$. (b) Alternative representation with $G_0|\alpha, H \sim \mathrm{DP}(\alpha, H)$, $\theta'_i|G_0 \sim G_0$, and $y_i|\theta'_i \sim F(\theta'_i)$. (c) Indicator variable representation in which $\beta|\gamma \sim$ GEM(γ), $\pi_k|\alpha,\beta \sim \mathrm{DP}(\alpha,\beta)$, $\theta_k|H,\lambda \sim H(\lambda)$, $z_{ji}|\pi_j \sim \pi_j$, and $y_{ji}|\{\theta_k\}_{k=1}^{\infty}, z_{ji} \sim F(\theta_{z_{ji}})$. (d) Alternative representation with $G_0|\gamma, H \sim \mathrm{DP}(\gamma, H)$, $G_j|G_0 \sim \mathrm{DP}(\alpha, G_0)$, $\theta'_{ji}|G_j \sim G_j$, and $y_{ji}|\theta'_{ji} \sim F(\theta'_{ji})$. The "plate" notation is used to compactly represent replication (Teh et al., 2006).*

$\theta'_{ji} = \theta^*_{jt_{ji}} = \theta_{k_{jt_{ji}}}$. The generative model is summarized below and is depicted as a graphical model in Fig. 3(a):

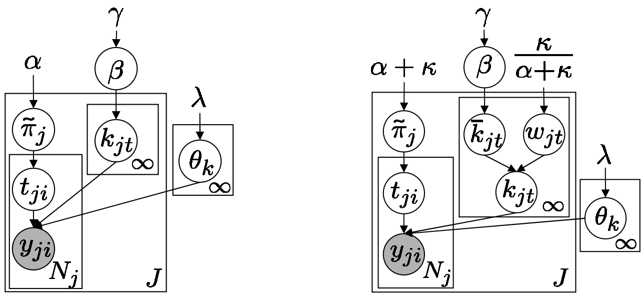
$$k_{jt} \mid \beta \sim \beta \qquad t_{ji} \mid \tilde{\pi}_j \sim \tilde{\pi}_j \qquad y_{ji} \mid \{\theta_k\}_{k=1}^{\infty}, \{k_{jt}\}_{t=1}^{\infty}, t_{ji} \sim F(\theta_{k_{jt_{ji}}}). \tag{3.2}$$

Marginalizing over the stick-breaking measures $\tilde{\pi}_j$ and β yields the following predictive distributions that describe the CRF:

$$p(t_{ji} \mid t_{j1}, \dots, t_{ji-1}, \alpha) \propto \sum_{t=1}^{T_j} \tilde{n}_{jt}\delta(t_{ji}, t) + \alpha\delta(t_{ji}, T_j + 1) \tag{3.3}$$

$$p(k_{jt} \mid \underline{k}_1, \underline{k}_2, \dots, \underline{k}_{j-1}, k_{j1}, \dots, k_{jt-1}, \gamma) \propto \sum_{k=1}^{K} m_{\cdot k}\delta(k_{jt}, k) + \gamma\delta(k_{jt}, K + 1),$$

where $m_{\cdot k} = \sum_j m_{jk}$ and $\underline{k}_j = \{k_{j1}, \dots, k_{jT_j}\}$. Here, $\tilde{n}_{jt}$ denotes the number of customers in restaurant j sitting at table t, m_{jk} the number of tables in restaurant j serving dish k, T_j the number of currently occupied tables in restaurant j, and K the total number of unique dishes being served in the franchise. Eq. (3.3) implies that upon entering the j^{th} restaurant in the CRF, customer y_{ji} sits at currently occupied tables t_{ji} with probability proportional to the number of currently seated customers, or starts a new table $T_j + 1$ with probability proportional to α. The first customer to sit at a table goes to the buffet line and picks a dish k_{jt} for their table, choosing the dish with probability proportional to the number of times that dish has been picked previously, or ordering a new dish θ_{K+1} with probability proportional



(a) (b)

FIG 3. *Graph of (a) CRF, and (b) CRF with loyal customers. Customers* y_{ji} *sit at table* $t_{ji}|\tilde{\pi}_j \sim \tilde{\pi}_j$. *In the CRF, each table chooses a dish* $k_{jt}|\beta \sim \beta$ *while in the CRF with loyal customers tables* consider *a dish* $\bar{k}_{jt}|\beta \sim \beta$*, but override variables* $w_{jt}|\alpha,\kappa \sim$ *Ber*$(\kappa/(\alpha+\kappa))$ *can force the* served *dish* k_{jt} *to be* j*. See Sec. 4.1.*

to γ. The intuition behind this predictive distribution is that integrating over the dish ratings β results in customers making decisions based on the observed popularity of the dishes.

Since each distribution G_j is drawn from a DP with a discrete base measure G_0, multiple θ^*_{jt} may take an identical value θ_k for multiple unique values of t, implying that multiple tables in the same restaurant may be served the same dish.We can write G_j as a function of these unique dishes:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk}\delta_{\theta_k}, \quad \pi_j \mid \alpha,\beta \sim \mathrm{DP}(\alpha,\beta)\,, \quad \theta_k \mid H \sim H, \tag{3.4}$$

where π_j now defines a restaurant-specific distribution over dishes served rather than over tables, with

$$\pi_{jk} = \sum_{t|k_{jt}=k} \tilde{\pi}_{jt}. \tag{3.5}$$

Let z_{ji} be the indicator random variable for the unique dish eaten by observation y_{ji}, so that $z_{ji} = k_{jt_{ji}}$. A third equivalent representation of the generative model is in terms of these indicator random variables:

$$\pi_j \mid \alpha,\beta \sim \mathrm{DP}(\alpha,\beta) \qquad z_{ji} \mid \pi_j \sim \pi_j \qquad y_{ji} \mid \{\theta_k\}, z_{ji} \sim F(\theta_{z_{ji}}), \tag{3.6}$$

and is shown in Fig. 2(c).

As with the DP, the HDP mixture model has an interpretation as the limit of a finite mixture model. Placing a finite Dirichlet prior on β induces a finite Dirichlet prior on π_j:

$$\begin{aligned} \beta \mid \gamma &\sim \mathrm{Dir}(\gamma/L,\ldots,\gamma/L) \\ \pi_j \mid \alpha,\beta &\sim \mathrm{Dir}(\alpha\beta_1,\ldots,\alpha\beta_L). \end{aligned} \tag{3.7}$$

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

4. The Sticky HDP-HMM. Recall that the hidden Markov model, or *HMM*, is a class of doubly stochastic processes based on an underlying, discrete-valued state sequence, which is modeled as Markovian (Rabiner, 1989). Let z_t denote the state of the Markov chain at time t and π_j the state-specific transition distribution for state j. Then, the Markovian structure on the state sequence dictates that $z_t \sim \pi_{z_{t-1}}$. The observations, y_t, are conditionally independent given this state sequence, with $y_t \sim F(\theta_{z_t})$ for some fixed distribution $F(\cdot)$.

The HDP can be used to develop an HMM with an infinite state space—the HDP-HMM (Teh et al., 2006). Conceptually, we envision a doubly-infinite transition matrix, with each row corresponding to a Chinese restaurant. That is, the groups in the HDP formalism here correspond to states, and each Chinese restaurant defines a distribution on next states. The CRF links these next-state distributions. Thus, in this application of the HDP, the group-specific distribution, π_j, is a state-specific transition distribution and, due to the infinite state space, there are infinitely many such groups. Since $z_t \sim \pi_{z_{t-1}}$, we see that z_{t-1} indexes the group to which y_t is assigned (i.e., all observations with $z_{t-1} = j$ are assigned to group j). Just as with the HMM, the current state z_t then indexes the parameter θ_{z_t} used to generate observation y_t (see Fig. 4(a)).

By defining $\pi_j \sim \mathrm{DP}(\alpha, \beta)$, the HDP prior encourages states to have similar transition distributions ($E[\pi_{jk} \mid \beta] = \beta_k$). However, it does not differentiate self-transitions from moves between different states. When modeling data with state persistence, the flexible nature of the HDP-HMM prior allows for state sequences with unrealistically fast dynamics to have large posterior probability. For example, with multinomial emissions, a good explanation of the data is to divide different observation values into unique states and then rapidly switch between them (see Fig. 1). In such cases, many models with redundant states may have large posterior probability, thus impeding our ability to identify a compact dynamical model which best explains the observations. The problem is compounded by the fact that once this alternating pattern has been instantiated by the sampler, its persistence is then reinforced by the properties of the Chinese restaurant franchise, thus slowing mixing rates. Furthermore, this fragmentation of data into redundant states can reduce predictive performance, as is discussed in Sec. 5. In many applications, one would like to be able to incorporate prior knowledge that slow, smoothly varying dynamics are more likely.

To address these issues, we propose to instead sample transition distributions π_j

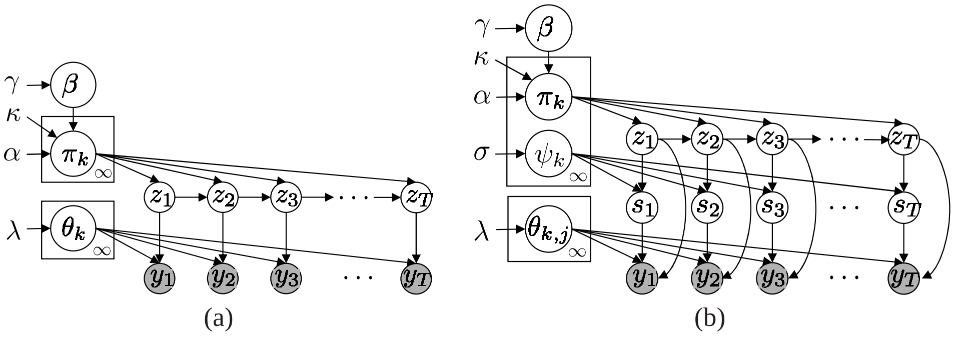


FIG 4. *(a) Graphical representation of the sticky HDP-HMM. The state evolves as $z_{t+1}|\{\pi_k\}_{k=1}^{\infty}, z_t \sim \pi_{z_t}$, where $\pi_k|\alpha,\kappa,\beta \sim$ DP$(\alpha+\kappa, (\alpha\beta+\kappa\delta_k)/(\alpha+\kappa))$ and $\beta|\gamma \sim$ GEM(γ), and observations are generated as $y_t|\{\theta_k\}_{k=1}^{\infty}, z_t \sim F(\theta_{z_t})$. The original HDP-HMM has $\kappa = 0$. (b) Sticky HDP-HMM with DP emissions, where s_t indexes the state-specific mixture component generating observation y_t. The DP prior dictates that $s_t|\{\psi_k\}_{k=1}^{\infty}, z_t \sim \psi_{z_t}$ for $\psi_k|\sigma \sim$ GEM(σ). The j^{th} Gaussian component of the k^{th} mixture density is parameterized by $\theta_{k,j}$ so $y_t|\{\theta_{k,j}\}_{k,j=1}^{\infty}, z_t, s_t \sim F(\theta_{z_t,s_t})$.*

as follows:

$$\begin{aligned}
&\beta \mid \gamma \sim \mathrm{GEM}(\gamma)\\
&\pi_j \mid \alpha,\kappa,\beta \sim \mathrm{DP}\left(\alpha+\kappa, \frac{\alpha\beta+\kappa\delta_j}{\alpha+\kappa}\right).
\end{aligned} \tag{4.1}$$

Here, $(\alpha\beta + \kappa\delta_j)$ indicates that an amount $\kappa > 0$ is added to the j^{th} component of $\alpha\beta$. Informally, what we are doing is increasing the expected probability of self-transition by an amount proportional to κ. More formally, over a finite partition $(Z_1, \ldots, Z_K)$ of the positive integers $\mathbb{Z}_+$, the prior on the measure π_j adds an amount κ only to the arbitrarily small partition containing j, corresponding to a self-transition. That is,

$$(\pi_j(Z_1), \ldots, \pi_j(Z_K)) \mid \alpha,\beta \sim \mathrm{Dir}(\alpha\beta(Z_1)+\kappa\delta_j(Z_1), \ldots, \alpha\beta(Z_K)+\kappa\delta_j(Z_K)) \tag{4.2}$$

When $\kappa = 0$ the original HDP-HMM of Teh et al. (2006) is recovered. Because positive κ values increase the prior probability $E[\pi_{jj} \mid \beta]$ of self-transitions, we refer to this extension as the *sticky* HDP-HMM. See Fig. 4(a).

The κ parameter is reminiscent of the self-transition bias parameter of the infinite HMM, a precursor of the HDP-HMM (Beal et al., 2002). The infinite HMM employs a two-level urn model. The top-level process places a probability on transitions to existing states in proportion to how many times these transitions have been seen, with an added bias towards a self-transition even if this has not previously occurred. With some remaining probability an oracle is called, representing the second-level urn. This oracle chooses an existing state in proportion to how many times the oracle previously chose that state, regardless of the state transition involved, or chooses a previously unvisited state. The oracle is included so that newly instantiated states may be visited from all currently instantiated states. While this urn model is an appealing description of probabilities on transitions, the lack of an underlying random measure makes it difficult to specify a coherent Bayesian inference procedure, and indeed the infinite HMM of Beal et al. (2002) relied on a heuristic approximation to a Gibbs sampler. The full connection between HMMs on an infinite state space and an underlying nonparametric Bayesian prior, as well as the development of a coherent inference algorithm, was made in Teh et al. (2006), but without the inclusion of a self-transition parameter (and hence with the potential pitfalls mentioned previously.)

4.1. *Chinese Restaurant Franchises with Loyal Customers.* We extend the Chinese restaurant metaphor to the sticky HDP-HMM, where our franchise now has restaurants with loyal customers. In addition to providing intuition for the predictive distribution on assignment variables, developing this metaphor aids in constructing the Gibbs samplers of Sec. 4.2 and Sec. 4.3. In the CRF with loyal customers, each restaurant in the franchise has a specialty dish with the same index as that of the restaurant. Although this dish is served elsewhere, it is more popular in the dish's namesake restaurant. We see this increased popularity in the specialty dish from the fact that a table's dish is now drawn from the *modified* dish ratings:

$$k_{jt} \mid \alpha, \kappa, \beta \sim \frac{\alpha\beta + \kappa\delta_j}{\alpha + \kappa}. \tag{4.3}$$

Specifically, we note that each restaurant has a set of restaurant-specific ratings of the buffet line that redistributes the shared ratings β so that there is more weight on the house-specialty dish.

Recall that while customers in the CRF of the HDP are pre-partitioned into restaurants based on the fixed group assignments, in the HDP-HMM the value of the state z_t determines the group assignment (and thus restaurant) of customer y_{t+1}. Therefore, we will describe a generative process that first assigns customers to restaurants and then assigns customers to tables and dishes. We will refer to z_t as the parent and z_{t+1} as the child. The parent enters a restaurant j determined by its parent (the grandparent), $z_{t-1} = j$. We assume there is a bijective mapping $f : t \to ji$ of time indices t to restaurant/customer indices ji. The parent then chooses a table $t_{ji} \sim \tilde{\pi}_j$ and that table is served a dish indexed by k_{jt}. Noting that $z_t = z_{ji} = k_{jt_{ji}}$ (i.e., the value of the state is the dish index), the increased popularity of the house specialty dish implies that children are more likely to eat in the same restaurant as their parent and, in turn, more likely to eat the restaurant's specialty dish. This develops family loyalty to a given restaurant in the franchise.

However, if the parent chooses a dish other than the house specialty, the child will then go to the restaurant where this dish is the specialty and will in turn be more likely to eat this dish, too. One might say that for the sticky HDP-HMM, children have similar tastebuds to their parents and will always go the restaurant that prepares their parent's dish best. Often, this keeps many generations eating in the same restaurant.

The inference algorithm for the sticky HDP-HMM, which is derived in the Supplementary Material,is simplified if we introduce a set of auxiliary random variables $\bar{k}_{jt}$ and w_{jt} as follows:

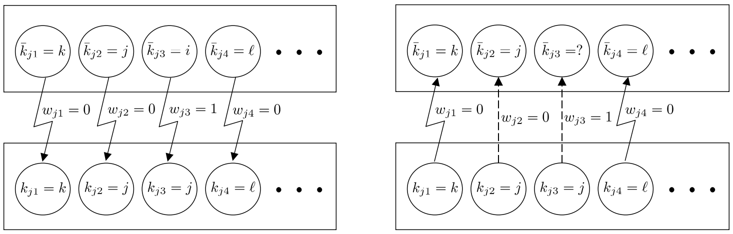
$$\begin{aligned} \bar{k}_{jt} \mid \beta &\sim \beta \\ w_{jt} \mid \alpha, \kappa &\sim \text{Ber}\left(\frac{\kappa}{\alpha+\kappa}\right) \triangleq \text{Ber}(\rho) \\ k_{jt} \mid \bar{k}_{jt}, w_{jt} &= \begin{cases} \bar{k}_{jt}, & w_{jt} = 0; \\ j, & w_{jt} = 1, \end{cases} \end{aligned} \tag{4.4}$$

where Ber(p) represents the Bernoulli distribution with parameter p. Here, we have defined a self-transition parameter $\rho = \kappa/(\alpha+\kappa)$. The table first chooses a dish $\bar{k}_{jt}$ without taking the restaurant's specialty into consideration (i.e., the original CRF). With some probability, this *considered* dish is overridden (perhaps by a waiter's suggestion) and the table is served the specialty dish j. Thus, k_{jt} represents the *served* dish. We refer to w_{jt} as the *override* variable. For the original HDP-HMM, when $\kappa = 0$, the considered dish is always the served dish since $w_{jt} = 0$ for all tables. This generative process is depicted in Fig. 5(a). Our inference algorithm, described in Sec. 4.2, aims to infer these variables conditioned on knowledge of the *served* dishes k_{jt}. For example, if the served dish of table t in restaurant j is indexed by j, the house specialty, the origin of this dish may either have been from considering $\bar{k}_{jt} = j$ or having been overridden by $w_{jt} = 1$. See Fig. 5(b).

A graphical model representation of the sticky HDP-HMM is shown in Fig. 3(b). Although not explicitly represented in this graph, the sticky HDP-HMM still induces a Markov structure on the indicator random variables z_t, which, based on the value of the parent state z_{t-1}, are mapped to a group-specific index ji. One can derive a distribution on partitions by marginalizing over the stick-breaking distributed measures $\tilde{\pi}_j$ and β, just as in the HDP. The CRF with loyal customers is then described by:

$$p(t_{ji} \mid t_{j1}, \dots, t_{ji-1}, \alpha, \kappa) \propto \sum_{t=1}^{T_j} \tilde{n}_{jt} \delta(t_{ji}, t) + (\alpha+\kappa)\delta(t_{ji}, T_j + 1) \tag{4.5}$$

$$p(\bar{k}_{jt} \mid \underline{\bar{k}}_1, \dots, \underline{\bar{k}}_{j-1}, \bar{k}_{j1}, \dots, \bar{k}_{jt-1}, \gamma) \propto \sum_{k=1}^{\bar{K}} \bar{m}_{\cdot k} \delta(\bar{k}_{jt}, k) + \gamma \delta(\bar{k}_{jt}, \bar{K} + 1),$$



(a) Generative (b) Inference

FIG 5. *(a) Generative model of considered dish indices $\bar{k}_{jt}$ (top) being converted to served dish indices k_{jt} (bottom) via override variables w_{jt}. (b) Perspective from the point of view of an inference algorithm that must infer $\bar{k}_{jt}$ and w_{jt} given k_{jt}. If $k_{jt} \neq j$, then the override variable w_{jt} is automatically 0 implying that $\bar{k}_{jt} = k_{jt}$, as indicated by the jagged arrow. If instead $k_{jt} = j$, then this could have arisen from the considered dish $\bar{k}_{jt}$ being overridden ($w_{jt} = 1$) or not ($w_{jt} = 0$). These scenarios are indicated by the dashed arrow. If the considered dish was not overridden, then $\bar{k}_{jt} = k_{jt} = j$. Otherwise, $\bar{k}_{jt}$ could have taken any value, as denoted by the question mark.*

where $\bar{m}_{jk}$ is the number of tables in restaurant j that *considered* dish k, and $\bar{K}$ the number of unique considered dishes in the franchise. The distributions on w_{jt} and k_{jt} remain as before, so that considered dishes are sometimes overridden by the house specialty.

Throughout the remainder of the paper, we use the following notational conventions. Given a random sequence $\{x_1, x_2, \ldots, x_T\}$, we use the shorthand $x_{1:t}$ to denote the sequence $\{x_1, x_2, \ldots, x_t\}$ and $x_{\setminus t}$ to denote the set $\{x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_T\}$. Also, for random variables with double subindices, such as $x_{a_1 a_2}$, we will use $\boldsymbol{x}$ to denote the entire set of such random variables, $\{x_{a_1 a_2}, \forall a_1, \forall a_2\}$, and the shorthand notation $x_{a_1 \cdot} = \sum_{a_2} x_{a_1 a_2}$, $x_{\cdot a_2} = \sum_{a_1} x_{a_1 a_2}$, and $x_{\cdot\cdot} = \sum_{a_1} \sum_{a_2} x_{a_1 a_2}$.

4.2. *Sampling via Direct Assignments.* In this section we present an inference algorithm for the sticky HDP-HMM of Sec. 4 and Fig. 4(a) that is a modified version of the direct assignment Rao-Blackwellized Gibbs sampler of Teh et al. (2006). This sampler circumvents the complicated bookkeeping of the CRF by sampling indicator random variables directly. The resulting sticky HDP-HMM direct assignment Gibbs sampler is outlined in Algorithm 1 of the Supplementary Material, which also contains the full derivations of this sampler.

The basic idea is that we marginalize over the infinite set of state-specific transition distributions π_k and parameters θ_k, and sequentially sample the state z_t given all other state assignments $z_{\setminus t}$, the observations $y_{1:T}$, and the global transition distribution β. A variant of the Chinese restaurant process gives us the prior probability of an assignment of z_t to a value k based on how many times we have seen other transitions from the previous state value z_{t-1} to k and k to the next state value z_{t+1}.As derived in the Supplementary Material, this conditional distribution is dependent upon whether either or both of the transitions z_{t-1} to k and k to z_{t+1} correspond to a self-transition, most strongly when $\kappa > 0$. The prior probability of an assignment of z_t to state k is then weighted by the likelihood of the observation y_t given all other observations assigned to state k.

Given a sample of the state sequence $z_{1:T}$, we can represent the posterior distribution of the global transition distribution β via a set of auxiliary random variables $\bar{m}_{jk}$, m_{jk}, and w_{jt}, which correspond to the j^{th} restaurant-specific set of table counts for each considered dish and served dish, and override variables of the CRF with loyal customers, respectively. The Gibbs sampler iterates between sequential sampling of the state z_t for each individual value of t given β and $z_{\setminus t}$; sampling of the auxiliary variables $\bar{m}_{jk}$, m_{jk}, and w_{jt} given $z_{1:T}$ and β; and sampling of β given these auxiliary variables.

4.3. *Blocked Sampling of State Sequences.* The HDP-HMM sequential, direct assignment sampler of Sec. 4.2 can exhibit slow mixing rates since global state sequence changes are forced to occur coordinate by coordinate. This phenomenon is explored in Scott (2002) for the finite HMM. Although the sticky HDP-HMM reduces the posterior uncertainty caused by fast state-switching explanations of the data, the self-transition bias can cause two continuous and temporally separated sets of observations of a given state to be grouped into two states. See Fig. 6(b) for an example. If this occurs, the high probability of self-transition makes it challenging for the sequential sampler to group those two examples into a single state.

Alternatively, we propose using a variant of the HMM forward-backward procedure (Rabiner, 1989) to harness the Markovian structure and jointly sample the state sequence $z_{1:T}$ given the observations $y_{1:T}$, transition probabilities π_k, and parameters θ_k. To take advantage of this procedure, we now must sample the previously marginalized transition distributions and model parameters. In practice, this requires approximating the countably infinite transition distributions. 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, 2002b),

$$\text{GEM}_L(\alpha) \triangleq \text{Dir}(\alpha/L, \dots, \alpha/L), \tag{4.6}$$

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 (2002b, 2000), 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 et al. (2007), and little to no practical differences are found.

The Gibbs sampler using blocked resampling of $z_{1:T}$ is derived in the Supplementary Material; 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 et al., 2007). However, this work did not consider the complications introduced by multimodal emissions, which we explore in Sec. 6. Recently, a slice sampler, referred to as *beam sampling* (Van Gael et al., 2008), has been developed for the HDP-HMM. This sampler harnesses the efficiencies of the forward-backward algorithm without having to fix a truncation level for the HDP. However, as we elaborate upon in Sec. 5.1, this sampler suffers from slower mixing rates than our blocked sampler, which utilizes a fixed-order, weak limit truncation of the HDP-HMM.

4.4. *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 Material; it roughly follows that of the original HDP-HMM (Teh et al., 2006). A key step which simplifies our inference procedure is to note that since we have the deterministic relationships

$$\begin{aligned} \alpha &= (1-\rho)(\alpha+\kappa) \\ \kappa &= \rho(\alpha+\kappa), \end{aligned} \tag{4.7}$$

we can treat ρ and $\alpha + \kappa$ as our hyperparameters and sample these values instead of sampling α and κ directly.

5. Experiments with Synthetic Data.

In this section, we explore the performance of the sticky HDP-HMM relative to the original model (i.e., the model with $\kappa = 0$) in a series of experiments with synthetic data. We judge performance according to two metrics: our ability to accurately segment the data according to the underlying state sequence, and the predictive likelihood of held-out data under the inferred model. We additionally assess the improvements in mixing rate achieved by using the blocked sampler of Sec. 4.3.

5.1. *Gaussian Emissions.* We begin our analysis of the sticky HDP-HMM performance by examining a set of simulated data generated from an HMM with Gaussian emissions. The first dataset is generated from an HMM with a high probability of self-transition. Here, we aim to show that the original HDP-HMM inadequately captures state persistence. The second dataset is from an HMM with a high probability of leaving the current state. In this scenario, our goal is to demonstrate that the sticky HDP-HMM is still able to capture rapid dynamics by inferring a small probability of self-transition.

For all of the experiments with simulated data, we used weakly informative hyperpriors. We placed a Gamma$(1, 0.01)$ prior on the concentration parameters γ and $(\alpha + \kappa)$. The self-transition proportion parameter ρ was given a Beta$(10, 1)$ prior. The parameters of the base measure were set from the data, as will be described for each scenario.

State Persistence. The data for the high persistence case were generated from a three-state HMM with a 0.98 probability of self-transition and equal probability of transitions to the other two states. The observation and true state sequences for the state persistence scenario are shown in Fig. 6(a). We placed a normal inverse-Wishart prior on the space of mean and variance parameters and set the hyperparameters as: 0.01 pseudocounts, mean equal to the empirical mean, three degrees of freedom, and scale matrix equal to 0.75 times the empirical variance. We used this conjugate base measure so that we may directly compare the performance of the blocked and direct assignment samplers.For the blocked sampler, we used a truncation level of $L = 20$.

In Fig. 6(d)-(h), we plot the 10^{th}, 50^{th}, and 90^{th} quantiles of the Hamming distance between the true and estimated state sequences over the 1000 Gibbs iterations using the direct assignment and blocked samplers on the sticky and original HDP-HMM models. To calculate the Hamming distance, we used the Munkres algorithm (Munkres, 1957) to map the randomly chosen indices of the estimated state sequence to the set of indices that maximize the overlap with the true sequence.

From these plots, we see that the burn-in rate of the blocked sampler using the sticky HDP-HMM is significantly faster than that of any other sampler-model combination. As expected, the sticky HDP-HMM with the sequential, direct assignment sampler gets stuck in state sequence assignments from which it is hard to move away, as conveyed by the flatness of the Hamming error versus iteration number plot in Fig. 6(g). For example, the estimated state sequence of Fig. 6(b) might have similar parameters associated with states 3, 7, 10 and 11 so that the likelihood is in essence the same as if these states were grouped, but this sequence has a large error in terms of Hamming distance and it would take many iterations to move away from this assignment. Incorporating the blocked sampler with the original HDP-HMM improves the Hamming distance performance relative to the sequential, direct assignment sampler for both the original and sticky HDP-HMM; however, the burn-in rate is still substantially slower than that of the blocked sampler on the sticky model.

Recently, a *beam sampling* algorithm (Van Gael et al., 2008) has been proposed

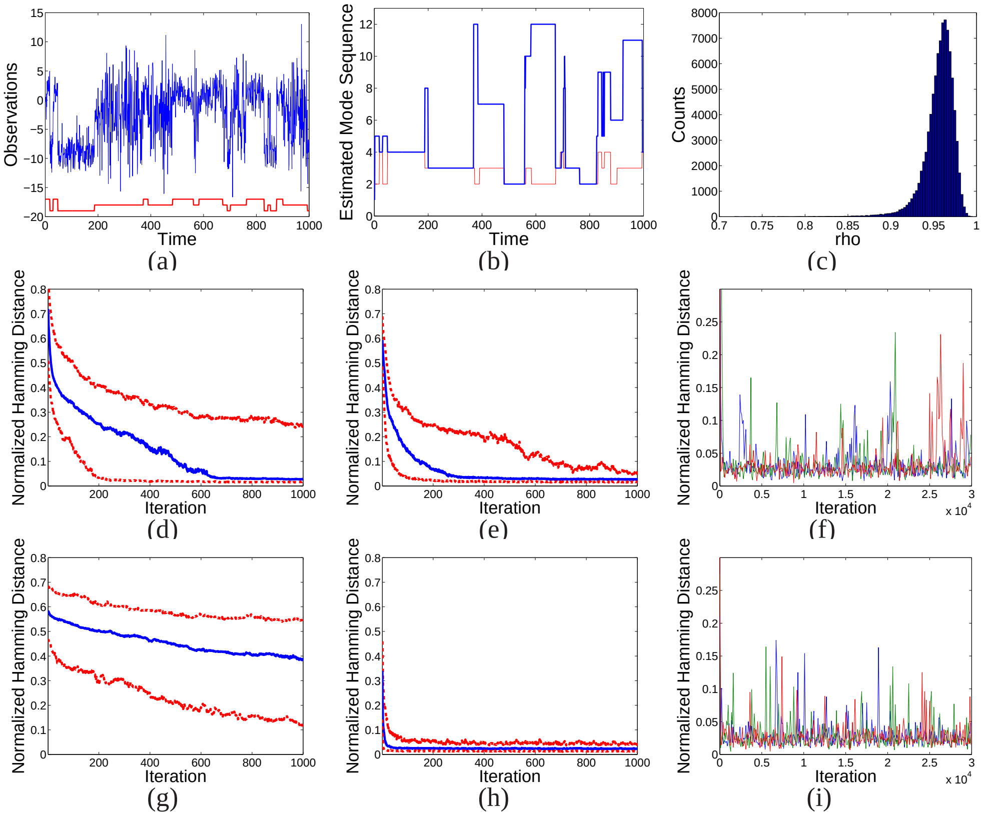
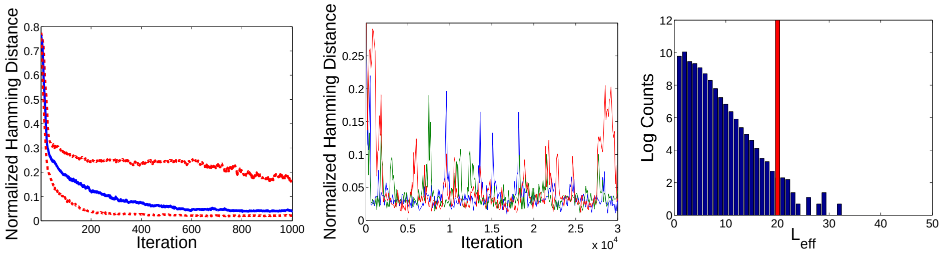


FIG 6. *(a) Observation sequence (blue) and true state sequence (red) for a three-state HMM with state persistence. (b) Example of the sticky HDP-HMM direct assignment Gibbs sampler splitting temporally separated examples of the same true state (red) into multiple estimate states (blue) at Gibbs iteration 1,000. (c) Histogram of the inferred self-transition proportion parameter, ρ, for the sticky HDP-HMM blocked sampler. For the original HDP-HMM, the median (solid blue) and 10^{th} and 90^{th} quantiles (dashed red) of Hamming distance between the true and estimated state sequences over the first 1,000 Gibbs samples from 200 chains are shown for the (d) direct assignment sampler, and (e) blocked sampler. (f) Hamming distance over 30,000 Gibbs samples from three chains of the original HDP-HMM blocked sampler. (g)-(i) Analogous plots to (d)-(f) for the sticky HDP-HMM.*

which adapts slice sampling methods (Robert and Casella, 2005) to the HDP-HMM. This approach uses a set of auxiliary slice variables, one for each observation, to effectively truncate the number of state transitions that must be considered at every Gibbs sampling iteration. Dynamic programming methods can then be used to jointly resample state assignments. The beam sampler was inspired by a related approach for DP mixture models (Walker, 2007), which is conceptually similar to retrospective sampling methods (Papaspiliopoulos and Roberts, 2008). In comparison to our fixed-order, weak limit truncation of the HDP-HMM, the beam sampler



(a) (b) (c)

FIG 7. *For the beam sampler, we plot: (a) the median (solid blue) and 10^{th} and 90^{th} quantiles (dashed red) of the Hamming distance between the true and estimated state sequences over the first 1,000 Gibbs samples from 200 chains, and (b) the Hamming distance over 30,000 Gibbs samples from three chains. (c) Histogram of the effective beam sampler truncation level, L_{eff}, over the 30,000 Gibbs iterations from the three chains (blue) compared to the fixed truncation level, $L = 20$, used above (red).*

provides an asymptotically exact algorithm. However, the beam sampler can be slow to mix relative to our blocked sampler on the fixed, truncated model (see Fig. 7 for an example comparison.) The issue is that in order to consider a transition which has low prior probability, one needs a correspondingly rare slice variable sample at that time. Thus, even if the likelihood cues are strong, to be able to consider state sequences with several low-prior-probability transitions, one needs to wait for several *rare events* to occur when drawing slice variables. By considering the full, exponentially large set of paths in the truncated state space, we avoid this problem. Of course, the trade-off between the computational cost of the blocked sampler on the fixed, truncated model ($O(TL^2)$) and the slower mixing rate of the beam sampler yields an application-dependent sampler choice.

The Hamming distance plots of Fig. 7(a) and (b), when compared to those of Fig. 6, depict the substantially slower mixing rate of the beam sampler than the blocked sampler. However, the theoretical computational benefit of the beam sampler can be seen in Fig. 7(c). In this plot, we present a histogram of the effective truncation level, L_{eff}, used over the 30,000 Gibbs iterations on three chains. We computed this effective truncation level by summing over the number of state transitions considered during a full sweep of sampling $z_{1:T}$ and then dividing this number by the length of the dataset, T, and taking the square root. On a more technical note, our fixed, truncated model allows for more vectorization of the code than the beam sampler. Thus, in practice, the difference in computation time between the samplers is significantly less than the $O(L^2/L_{eff}^2)$ factor obtained by counting state transitions.

From this point onwards, we present results only from blocked sampling since we have seen the clear advantages of this method over the sequential, direct assignment sampler.

Fast State-Switching. In order to warrant the general use of the sticky model, one would like to know that the incorporated sticky parameter does not preclude learning models with fast dynamics. To this end, we explored the performance of the sticky HDP-HMM on data generated from a model with a high probability of switching between states. Specifically, we generated observations from a four-state HMM with the following transition probability matrix:

$$\begin{bmatrix} 0.4 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.4 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.4 & 0.4 \\ 0.1 & 0.1 & 0.4 & 0.4 \end{bmatrix}. \tag{5.1}$$

We once again used a truncation level $L = 20$. Since we are restricting ourselves to the blocked Gibbs sampler, it is no longer necessary to use a conjugate base measure. Instead we placed an independent Gaussian prior on the mean parameter and an inverse-Wishart prior on the variance parameter. For the Gaussian prior, we set the mean and variance hyperparameters to be equal to the empirical mean and variance of the entire dataset. The inverse-Wishart hyperparameters were set such that the expected variance is equal to 0.75 times that of the entire dataset, with three degreee of freedom.

The results depicted in Fig. 8 confirm that by inferring a small probability of self-transition, the sticky HDP-HMM is indeed able to capture fast HMM dynamics, and just as quickly as the original HDP-HMM (although with higher variability.) Specifically, we see that the histogram of the self-transition proportion parameter ρ for this dataset (see Fig. 8(d)) is centered around a value close to the true probability of self-transition, which is substantially lower than the mean value of this parameter on the data with high persistence (Fig. 6(c).)

5.2. *Multinomial Emissions.* The difference in modeling power, rather than simply burn-in rate, between the sticky and original HDP-HMM is more pronounced when we consider multinomial emissions. This is because the multinomial observations are embedded in a discrete topological space in which there is no concept of similarity between non-identical observation values. In contrast, Gaussian emissions have a continuous range of values in $\mathbb{R}^n$ with a clear notion of *closeness* between observations under the Lebesgue measure, aiding in grouping observations under a single HMM state's Gaussian emission distribution, even in the absence of a self-transition bias.

To demonstrate the increased posterior uncertainty with discrete observations, we generated data from a five-state HMM with multinomial emissions with a 0.98 probability of self-transition and equal probability of transitions to the other four

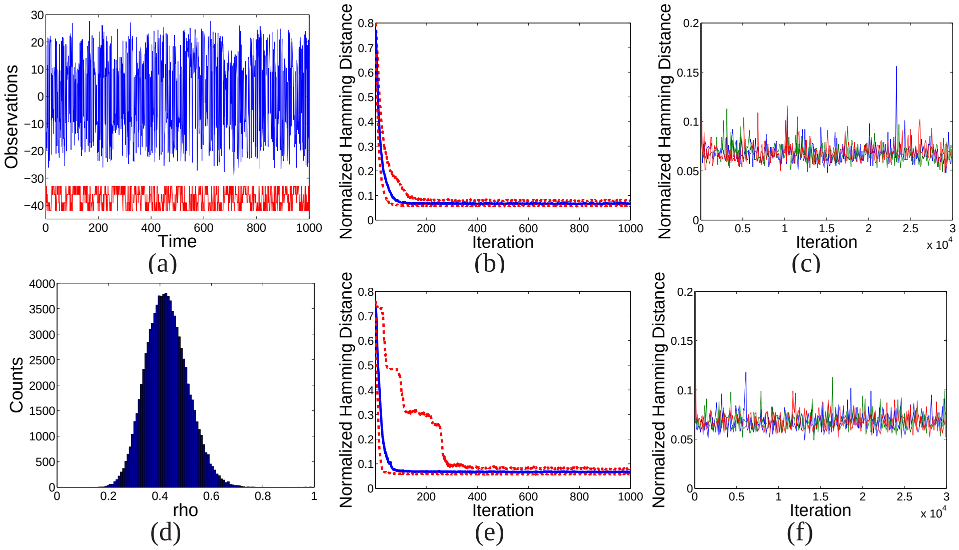
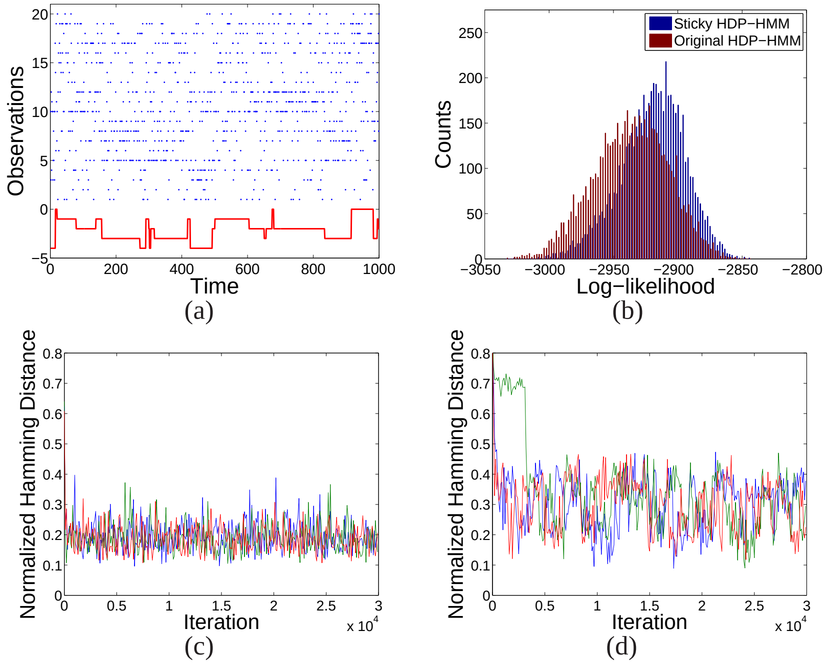


FIG 8. *(a) Observation sequence (blue) and true state sequence (red) for a four-state HMM with fast state switching. For the original HDP-HMM using a blocked Gibbs sampler: (b) the median (solid blue) and 10^{th} and 90^{th} quantiles (dashed red) of Hamming distance between the true and estimated state sequences over the first 1,000 Gibbs samples from 200 chains, and (c) Hamming distance over 30,000 Gibbs samples from three chains. (d) Histogram of the inferred self-transition parameter, ρ, for the sticky HDP-HMM blocked sampler. (e)-(f) Analogous plots to (b)-(c) for the sticky HDP-HMM.*

states. The vocabulary, or range of possible observation values, was set to 20. The observation and true state sequences are shown in Fig. 9(a). We placed a symmetric Dirichlet prior on the parameters of the multinomial distribution, with the Dirichlet hyperparameters equal to 2 (i.e., Dir$(2, \dots, 2)$.)

From Fig. 9, we see that even after burn-in, many fast-switching state sequences have significant posterior probability under the non-sticky model leading to sweeps through regions of larger Hamming distance error. A qualitative plot of one such inferred sequence after 30,000 Gibbs iterations is shown in Fig. 1(c). Such sequences have negligible posterior probability under the sticky HDP-HMM formulation.

In some applications, such as the speaker diarization problem that is explored in Sec. 7, one cares about the inferred segmentation of the data into a set of state labels. In this case, the advantage of incorporating the sticky parameter is clear. However, it is often the case that the metric of interest is the predictive power of the fitted model, not the accuracy of the inferred state sequence. To study performance under this metric, we simulated 10 test sequences using the same parameters that generated the training sequence. We then computed the likelihood of each of the test sequences under the set of parameters inferred at every 100^{th} Gibbs iter-



(a) (b) (c) (d)

FIG 9. *(a) Observation sequence (blue) and true state sequence (red) for a five-state HMM with multinomial observations. (b) Histogram of the predictive probability of test sequences using the inferred parameters sampled every* 100^{th} *iteration from Gibbs iterations 10,000 to 30,000 for the sticky and original HDP-HMM. The Hamming distances over 30,000 Gibbs samples from three chains are shown for the (b) sticky HDP-HMM and (c) original HDP-HMM.*

ation from iterations 10,000 to 30,000. This likelihood was computed by running the forward-backward algorithm of Rabiner (1989). We plot these results as a histogram in Fig. 9(b). From this plot, we see that the fragmentation of data into redundant HMM states can also degrade the predictive performance of the inferred model. Thus, the sticky parameter plays an important role in the Bayesian nonparametric learning of HMMs even in terms of model averaging.

5.3. *Comparison to Independent Sparse Dirichlet Prior.* We have alluded to the fact that the *shared* sparsity of the HDP-HMM induced by β is essential for inferring sparse representations of the data. Although this is clear from the perspective of the prior model, or equivalently the generative process, it is not immediately obvious how much this hierarchical Bayesian constraint helps us in posterior inference. Once we are in the realm of considering a fixed, truncated approximation to the HDP-HMM, one might propose an alternate model in which we simply place a sparse Dirichlet prior, $\mathrm{Dir}(\alpha/L, \dots, \alpha/L)$ with $\alpha/L < 1$, independently on each row of the transition matrix. This is equivalent to setting $\beta = [1/L, \dots, 1/L]$ in

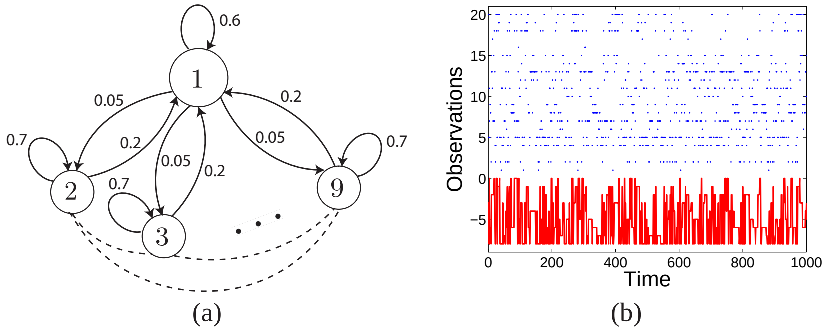
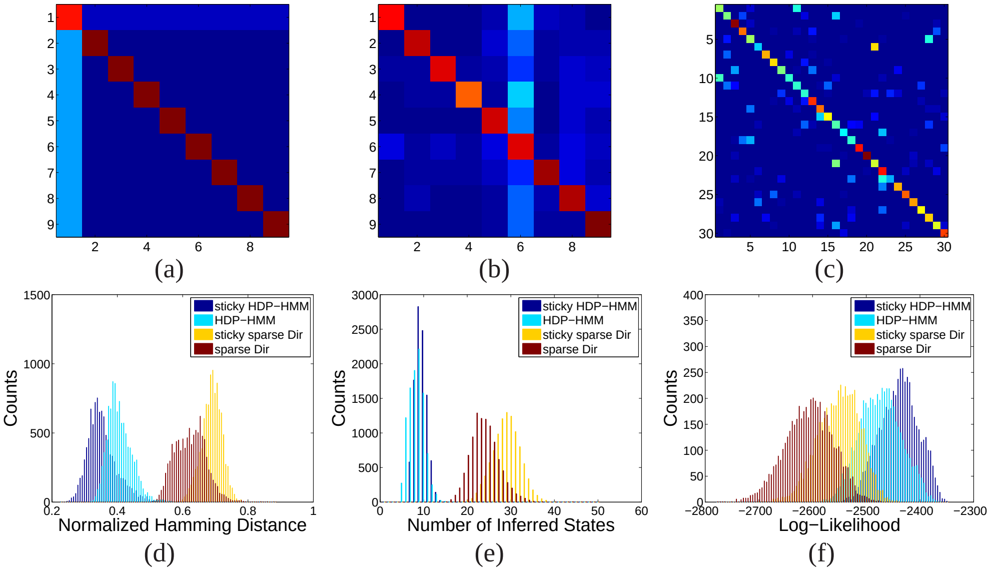


FIG 10. *(a)State transition diagram for a nine-state HMM with one main state (labeled 1) and eight sub-states (labeled 2 to 9.) All states have a significant probability of self-transition. From the main state, all other states are equally likely. From a sub-state, the most likely non-self-transition is a transition is back to the main state. However, all sub-states have a small probability of transitioning to another sub-state, as indicated by the dashed arcs. (b) Observation sequence (blue) and true state sequence (red) generated by the nine-state HMM with multinomial observations.*

the truncated HDP-HMM, which can also be achieved by letting the hyperparameter γ tend to infinity. Indeed, when the data do not exhibit shared sparsity or when the likelihood cues are sufficiently strong, the independent sparse Dirichlet prior model can perform as well as the truncated HDP-HMM. However, in scenarios such as the one depicted in Fig. 10, we see substantial differences in performance by considering the HDP-HMM, as well as the inclusion of the sticky parameter. We explored the relative performance of the HDP-HMM and sparse Dirichlet prior model, with and without the sticky parameter, on such a Markov model with multinomial emissions on a vocabulary of size 20. We placed a $\mathrm{Dir}(0.1, \dots, 0.1)$ prior on the parameters of the multinomial distribution. For the sparse Dirichlet prior model, we assumed a state space of size 50, which is the same as the truncation level we chose for the HDP-HMM (i.e., $L = 50$). The results are presented in Fig. 11. From these plots, we see that the hierarchical Bayesian approach of the HDP-HMM does, in fact, improve the fitting of a model with shared sparsity. The HDP-HMM consistently infers fewer HMM states and more representative model parameters. As a result, the HDP-HMM has higher predictive likelihood on test data, with an additional benefit gained from using the sticky parameter.

Note that the results of Fig. 11(f) also motivate the use of the sticky parameter in the more classical setting of a finite HMM with a standard Dirichlet sparsity prior. A motivating example of the use of sparse Dirichlet priors for finite HMMs is presented in Johnson (2007).

6. Multimodal Emission Densities. In many application domains, the data associated with each hidden state may have a complex, multimodal distribution. We propose to model such emission distributions nonparametrically, using a DP mix-



(a) (b) (c)

(d) (e) (f)

FIG 11. *(a) The true transition probability matrix (TPM) associated with the state transition diagram of Fig. 10. (b)-(c) The inferred TPM at the 30,000th Gibbs iteration for the sticky HDP-HMM and sticky sparse Dirichlet model, respectively, only examining those states with more than 1% of the assignments. For the HDP-HMM and sparse Dirichlet model, with and without the sticky parameter, we plot: (d) the Hamming distance error over 10,000 Gibbs iterations, (e) the inferred number of states with more than 1% of the assignments, and (f) the predictive probability of test sequences using the inferred parameters sampled every 100^{th} iteration from Gibbs iterations 5,000 to 10,000.*

ture of Gaussians. This formulation is related to the nested DP (Rodriguez et al., 2008), which uses a Dirichlet process to partition data into groups, and then models each group via a Dirichlet process mixture. The bias towards self-transitions allows us to distinguish between the underlying HDP-HMM states. If the model were free to both rapidly switch between HDP-HMM states and associate multiple Gaussians per state, there would be considerable posterior uncertainty. Thus, it is only with the sticky HDP-HMM that we can effectively fit such models.

We augment the HDP-HMM state z_t with a term s_t indexing the mixture component of the z_t^{th} emission density. For each HDP-HMM state, there is a unique stick-breaking measure $\psi_k \sim \mathrm{GEM}(\sigma)$ defining the mixture weights of the k^{th} emission density so that $s_t \sim \psi_{z_t}$. Given the augmented state (z_t, s_t), the observation y_t is generated by the Gaussian component with parameter θ_{z_t,s_t}. Note that both the HDP-HMM state index and mixture component index are allowed to take values in a countably infinite set. See Fig. 4(b).

6.1. *Direct Assignment Sampler.* Many of the steps of the direct assignment sampler for the sticky HDP-HMM with DP emissions remain the same as for the regular sticky HDP-HMM. Specifically, the sampling of the global transition distribution β, the table counts m_{jk} and $\bar{m}_{jk}$, and the override variables w_{jt} are unchanged. The difference arises in how we sample the augmented state (z_t, s_t).

The joint distribution on the augmented state, having marginalized the transition distributions π_k and emission mixture weights ψ_k, is given by

$$\begin{aligned} p(z_t = k, s_t = j \mid z_{\backslash t}, s_{\backslash t}, y_{1:T}, \beta, \alpha, \sigma, \kappa, \lambda) &= p(s_t = j \mid z_t = k, z_{\backslash t}, s_{\backslash t}, y_{1:T}, \sigma, \lambda) \\ &\quad p(z_t = k \mid z_{\backslash t}, s_{\backslash t}, y_{1:T}, \beta, \alpha, \kappa, \lambda). \end{aligned} \tag{6.1}$$

We then block-sample (z_t, s_t) by first sampling z_t, followed by s_t conditioned on the sampled value of z_t. The term $p(s_t = j \mid z_t = k, z_{\backslash t}, s_{\backslash t}, y_{1:T}, \sigma, \lambda)$ relies on how many observations are currently assigned to the j^{th} mixture component of state k. These conditional distributions are derived in the Supplementary Material, with the resulting Gibbs sampler outlined in Algorithm 2.

6.2. *Blocked Sampler.* To implement blocked resampling of $(z_{1:T}, s_{1:T})$, we use weak limit approximations to both the HDP-HMM and DP emissions, approximated to levels L and L', respectively. The posterior distributions for β and π_k remain unchanged from the sticky HDP-HMM; that of ψ_k is given by

$$\psi_k \mid z_{1:T}, s_{1:T}, \sigma \sim \text{Dir}(\sigma/L' + n'_{k1}, \dots, \sigma/L' + n'_{kL'}). \tag{6.2}$$

The procedure for sampling the augmented state $(z_{1:T}, s_{1:T})$ is derived in the Supplementary Material (see Algorithm 4).

6.3. *Assessing the Multimodal Emissions Model.* In this section, we evaluate the ability of the sticky HDP-HMM to infer multimodal emission distributions relative to the model without the sticky parameter. We generated data from a five-state HMM with mixture of Gaussian emissions, where the number of mixture components for each emission distribution was chosen randomly from a uniform distribution on $\{1, 2, \dots, 10\}$. Each component of the mixture was equally weighted and the probability of self-transition was set to 0.98, with equal probabilities of transitions to the other states. The large probability of self-transition is what disambiguates this process from one with many more HMM states, each with a single Gaussian emission distribution. The resulting observation and true state sequences are shown in Fig. 12(a) and (b).

We once again used a non-conjugate base measure and placed a Gaussian prior on the mean parameter and an independent inverse-Wishart prior on the variance parameter of each Gaussian mixture component. The hyperparameters for these distributions were set from the data in the same manner as in the fast-switching scenario. Consistent with the sticky HDP-HMM concentration parameters γ and

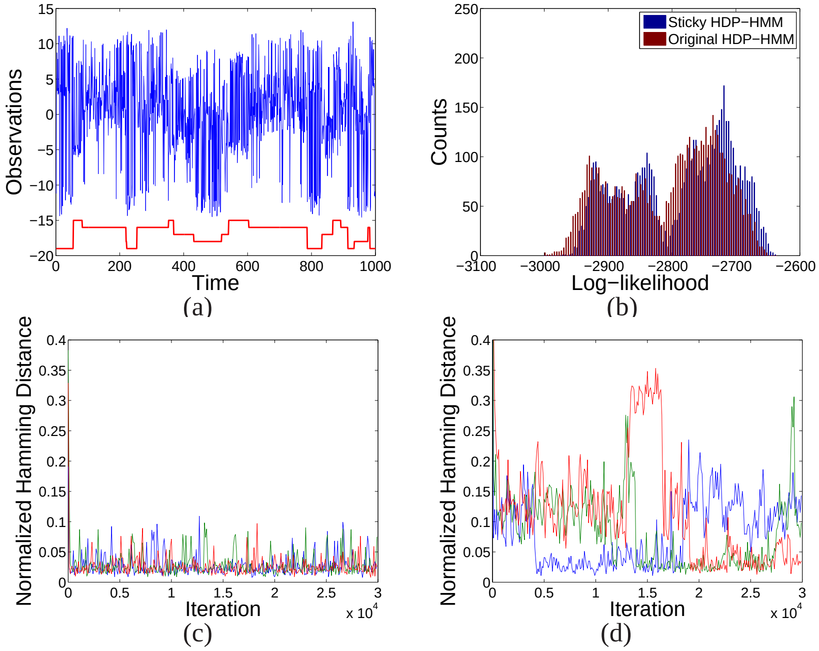


FIG 12. *(a) Observation sequence (blue) and true state sequence (red) for a five-state HMM with mixture of Gaussian observations. The Hamming distance over 30,000 Gibbs samples from three chains are shown for the (b) sticky HDP-HMM and (c) original HDP-HMM, both with DP emissions. (d) Histogram of the predictive probability of test sequences using the inferred parameters sampled every* 100^{th} *iteration from Gibbs iterations 10,000 to 30,000 for the sticky and original HDP-HMM.*

$(\alpha + \kappa)$, we placed a weakly informative Gamma$(1, 0.01)$ prior on the concentration parameter σ of the DP emissions. All results are for the blocked sampler with truncation levels $L = L' = 20$.

In Fig. 12, we compare the performance of the sticky HDP-HMM with DP emissions to that of the original HDP-HMM with DP emissions (i.e., DP emissions, but no bias towards self-transitions.) As with the multinomial observations, when the distance between observations does not directly factor into the grouping of observations into HMM states, there is a considerable amount of posterior uncertainty in the underlying HMM state. Even after 30,000 Gibbs samples, there are still state sequence sample paths with very rapid dynamics. The result of this fragmentation into redundant states is a slight reduction in predictive performance on test sequences, as in the multinomial emission case. See Fig. 12(b).

7. Speaker Diarization. Recall that the *speaker diarization* task involves segmenting an audio recording into speaker-homogeneous regions, while simultaneously identifying the number of speakers. In this section we present our results on applying the sticky HDP-HMM with DP emissions to the speaker diarization task.

The data used for our experiments are a standard benchmark data set distributed by NIST as part of the Rich Transcription 2004-2007 meeting recognition evaluations (NIST, 2007). We used the first 19 Mel Frequency Cepstral Coefficients (MFCCs), computed over a 30ms window every 10ms, as a feature vector. When working with this dataset, we discovered that: (1) the high frequency content of these features contained little discriminative information, and (2) without a minimum speaker duration, the sticky HDP-HMM inferred within-speaker dynamics in addition to global speaker changes. To address both of these issues, we defined the observations as averages over 250ms, non-overlapping blocks. A minimum speaker duration of 500ms was set by associating two of these observations with each hidden state. We also tied the covariances of within-state mixture components (i.e., each speaker-specific mixture component was forced to have identical covariance structure), and used a non-conjugate prior on the mean and covariance parameters. We placed a normal prior on the mean parameter with mean equal to the empirical mean and covariance equal to 0.75 times the empirical covariance, and an inverse-Wishart prior on the covariance parameter with 1000 degrees of freedom and expected covariance equal to the empirical covariance. For the concentration parameters, we placed a Gamma$(12, 2)$ prior on γ, a Gamma$(6, 1)$ prior on $\alpha + \kappa$, and a Gamma$(1, 0.5)$ prior on σ. The self-transition parameter ρ was given a Beta$(500, 5)$ prior. For each of the 21 meetings, we ran 10 chains of the blocked Gibbs sampler for 10,000 iterations for both the original and sticky HDP-HMM with DP emissions.

For the NIST speaker diarization evaluations, the goal is to produce a single segmentation for each meeting. Due to the label-switching issue (i.e., under our exchangeable prior, labels are arbitrary entities that do not necessarily remain consistent over Gibbs iterations), we cannot simply integrate over multiple Gibbs-sampled state sequences. We propose two solutions to this problem. The first is to simply choose from a fixed set of Gibbs samples the one that produces the largest likelihood given the estimated parameters (marginalizing over state sequences), and then produce the corresponding Viterbi state sequence. This heuristic, however, is sensitive to overfitting and will, in general, be biased towards solutions with more states. An alternative, more robust, metric we propose is what we refer to as the *minimum expected Hamming distance*. We first choose a large reference set of state sequences produced by the Gibbs sampler and a possibly smaller set of test sequences. Then, for each sequence in the test set, we compute the mean Hamming distance between the test sequence and those in the reference set. We then choose the test sequence that minimizes this expected Hamming distance. To compute the Hamming distance, we first find the optimal permutation of test labels to reference labels. This heuristic for choosing state sequence samples aims

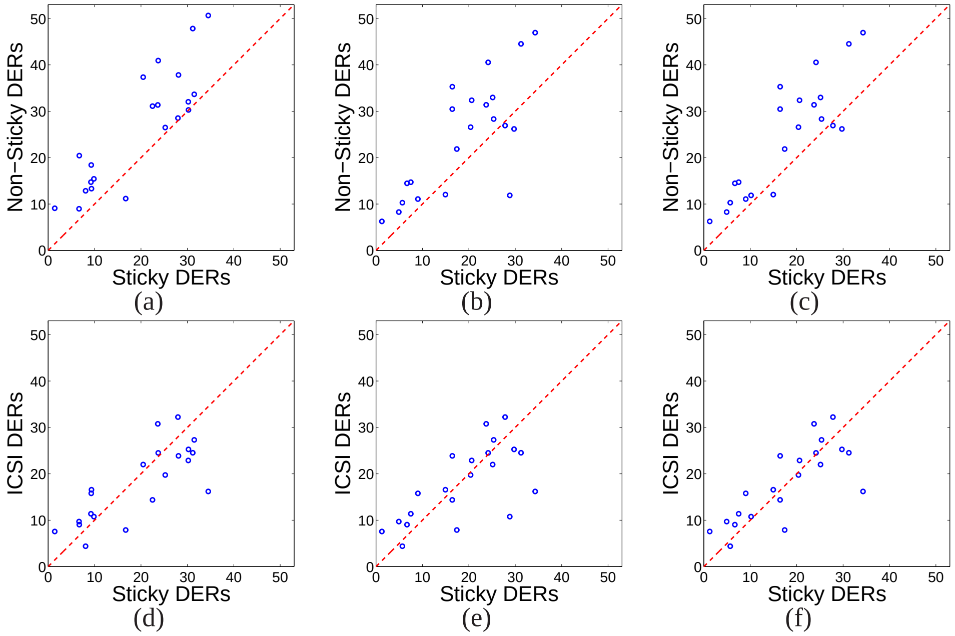


FIG 13. *(a)-(c) For each of the 21 meetings, comparison of diarizations using sticky vs. original HDP-HMM with DP emissions. In (a) we plot the DERs corresponding to the Viterbi state sequence using the parameters inferred at Gibbs iteration 10,000 that maximize the likelihood, and in (b) the DERs using the state sequences that minimize the expected Hamming distance. Plot (c) is the same as (b), except for running the 10 chains for meeting 16 out to 50,000 iterations. (d)-(f) Comparison of the sticky HDP-HMM with DP emissions to the ICSI errors under the same conditions.*

to produce segmentations of the data that are *typical* samples from the posterior. Jasra et al. (2005) provides an overview of some related techniques to address the label-switching issue. Although we could have chosen any loss function to minimize, we chose the Hamming distance metric because it is closely related to the official NIST *diarization error rate* (DER) that is calculated during the evaluations. The final metric by which the speaker diarization algorithms are judged is the *overall* DER, a weighted average based on the length of each meeting.

In Fig. 13(a), we report the DER of the chain with the largest likelihood given the parameters estimated at the $10,000^{th}$ Gibbs iteration for each of the 21 meetings, comparing the sticky and original HDP-HMM with DP emissions. We see that the sticky model's temporal smoothing provides substantial performance gains. Although not depicted in this paper, the likelihoods based on the parameter estimates under the original HDP-HMM are almost always higher than those under the sticky model. This phenomenon is due to the fact that without the sticky parameter, the HDP-HMM over-segments the data and thus produces parameter estimates more finely tuned to the data resulting in higher likelihoods. Since the original HDP-

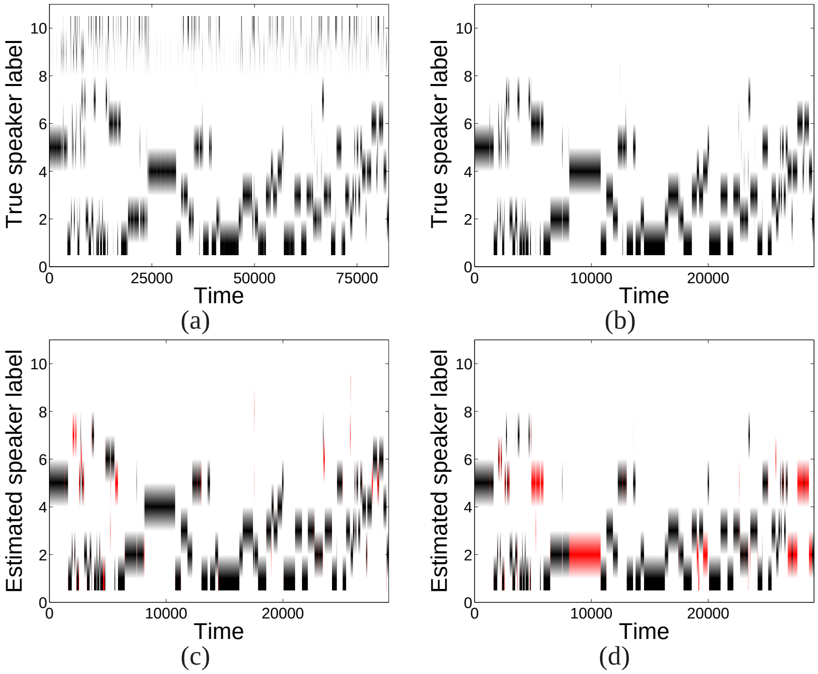
Overall DERs (%)	Min Hamming	Max Likelihood	2-Best	5-Best
Sticky HDP-HMM	19.01 (17.84)	19.37	16.97	14.61
Non-Sticky HDP-HMM	23.91	25.91	23.67	21.06

TABLE 1

Overall DERs for the sticky and original HDP-HMM with DP emissions using the minimum expected Hamming distance and maximum likelihood metrics for choosing state sequences at Gibbs iteration 10,000. For the maximum likelihood criterion, we show the best overall DER if we consider the top two or top five most-likely candidates. The number in the parentheses is the performance when running meeting 16 for 50,000 Gibbs iterations. The overall ICSI DER is 18.37%.

HMM is contained within the class of sticky models (i.e., when $\kappa = 0$), there is some probability that state sequences similar to those under the original model will eventually arise using the sticky model. Thus, the likelihood metric is not very robust as one would expect the performance under the sticky model to degrade given enough Gibbs chains and/or iterations. In Fig. 13(b), we instead report the DER of the chain whose state sequence estimate at Gibbs iteration 10,000 minimizes the expected Hamming distance to the sequences estimated every 100 Gibbs iteration, discarding the first 5,000 iterations. Due to the slow mixing rate of the chains in this application, we additionally discard samples whose normalized log-likelihood is below 0.1 units of the maximum at Gibbs iteration 10,000. From this figure, we see that the sticky model still significantly outperforms the original HDP-HMM, implying that most state sequences produced by the original model are worse, not just the one corresponding to the most-likely sample. One noticeable exception to this trend is the NIST_20051102-1323 meeting (meeting 16). For the sticky model, the state sequence using the maximum likelihood metric had very low DER (see Fig. 14(c)); however, there were many chains that merged speakers and produced segmentations similar to the one in Fig. 14(d), resulting in such a sequence minimizing the expected Hamming distance. See Sec. 8 for a discussion on the issue of merged speakers. Running meeting 16 for 50,000 Gibbs iterations improved the performance, as depicted by the revised results in Fig. 13(c). We summarize our overall performance in Table 1, and note that (when using the 50,000 Gibbs iterations for meeting 16) we obtain an overall DER of 17.84% using the sticky HDP-HMM versus the 23.91% of the original HDP-HMM model.

As a further comparison, the algorithm that was by far the best performer at the 2007 NIST competition—the algorithm developed by a team at the International Computer Science Institute (ICSI) (Wooters and Huijbregts, 2007), has an overall DER of 18.37%. The ICSI team's algorithm uses agglomerative clustering, and requires significant tuning of parameters on representative training data. In contrast, our hyperparameters are automatically set meeting-by-meeting, as outlined at the beginning of this section. An additional benefit of the sticky HDP-HMM over the



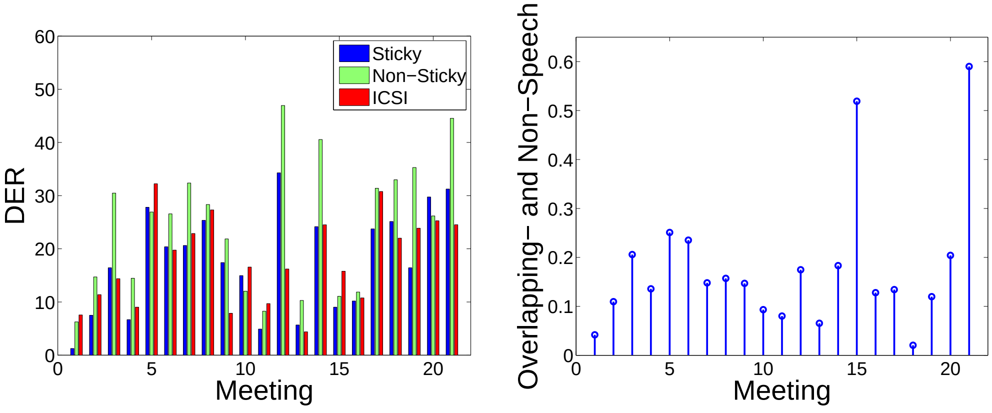
(a) (b)

(c) (d)

FIG 14. *(a) True state sequence for the NIST_20051102-1323 meeting (meeting 16), with labels 9 and 10 indicating times of overlapping- and non- speech, respectively, missed by the speech/non-speech preprocessor. (b) True state sequence with the overlapping- and non- speech time steps removed. (c)-(d) Plotted only over the time-steps as in (b), the state sequences inferred by the sticky HDP-HMM with DP emissions at Gibbs iteration 10,000 chosen using the most likely and minimum expected Hamming distance metrics, respectively. Incorrect labels are shown in red.*

ICSI approach is the fact that there is inherent posterior uncertainty in this task, and by taking a Bayesian approach we are able to provide several interpretations. Indeed, when considering the best per-meeting DER for the five most likely samples, our overall DER drops to 14.61% (see Table 1). Although not helpful in the NIST evaluations, providing multiple segmentations could be useful in practice.

To ensure a fair comparison, we use the same speech/non-speech pre-processing as ICSI, so that the differences in our performance are due to changes in the identified speakers. (Non-speech refers to time intervals in which nobody is speaking.) The pre-processing step of removing non-speech observations is important in ensuring that the fitted acoustic models are not corrupted by non-speech information. As depicted in Fig. 15, both our performance and that of ICSI depend significantly on the quality of this pre-processing step. In Fig. 15(a), we compare the meeting-by-meeting DERs of the sticky HDP-HMM, the original HDP-HMM, and the ICSI algorithm, and in Fig. 15(b) we plot the fraction of post-processed data that still contains overlapping- and non-speech.[1] It is clear from Fig. 15(a) that the sticky HDP-HMM with DP emissions provides performance comparable to that of the ICSI algorithm while the original HDP-HMM with DP emissions performs significantly worse. Overall, the results presented in this section demonstrate that the sticky HDP-HMM with DP emissions provides an elegant and empirically effective speaker diarization method.



(a) (b)

FIG 15. *(a) Chart comparing the DERs of the sticky and original HDP-HMM with DP emissions to those of ICSI for each of the 21 meetings. Here, we chose the state sequence at the* $10,000^{th}$ *Gibbs iteration that minimizes the expected Hamming distance. For meeting 16 using the sticky HDP-HMM with DP emissions, we chose between state sequences at Gibbs iteration 50,000. (b) Plot of the fraction of overlapping- or non- speech in the post-processed data for each of the 21 meetings.*

8. Discussion. We have developed a Bayesian nonparametric approach to the problem of speaker diarization, building on the HDP-HMM presented in Teh et al. (2006). Although the original HDP-HMM does not yield competitive speaker diarization performance due to its inadequate modeling of the temporal persistence of states, the sticky HDP-HMM that we have presented here resolves this problem and yields a state-of-the-art solution to the speaker diarization problem.

We have also shown that this sticky HDP-HMM allows a fully Bayesian nonparametric treatment of multimodal emissions, disambiguated by its bias towards self-transitions. Accommodating multimodal emissions is essential for the speaker diarization problem and is likely to be an important ingredient in other applications of the HDP-HMM to problems in speech technology.

We also presented efficient sampling techniques with mixing rates that improve on the state-of-the-art by harnessing the Markovian structure of the HDP-HMM.

[1] Not shown in this plot is the amount of actual speech removed by the speech/non-speech preprocessor.

Specifically, we proposed employing a truncated approximation to the HDP and block-sampling the state sequence using a variant of the forward-backward algorithm. Although the blocked samplers yield substantially improved mixing rates over the sequential, direct assignment samplers, there are still some pitfalls to these sampling methods. One issue is that for each new considered state, the parameter sampled from the prior distribution must better explain the data than the parameters associated with other states that have already been informed by the data. In high-dimensional applications, and in cases where state-specific emission distributions are not clearly distinguishable, this method for adding new states poses a significant challenge. The data in the speaker diarization task is both high-dimensional and often has only marginally distinguishable speakers, leading to extremely slow mixing rates, as indicated by trace plots of various indicators such as Hamming distance and log-likelihood for 100,000 Gibbs iterations of meeting 16. Many of our errors in this application can be attributed to merged speakers, as depicted in Fig. 14(d). On such large datasets, the computation cost of running hundreds of thousands of Gibbs iterations proves an insurmountable barrier. A direction for future work is to develop split-merge algorithms for the HDP and HDP-HMM similar to those developed in Jain and Neal (2004) for the DP mixture model.

A limitation of the HMM in general is that the observations are assumed conditionally i.i.d. given the state sequence. This assumption is often insufficient in capturing the complex temporal dependencies exhibited in real-world data. Another area of future work is to consider Bayesian nonparametric versions of models better suited to such applications, like the switching linear dynamical system (SLDS) and switching VAR process. A first attempt at developing such models is presented in Fox et al. (2009). An inspiration for the sticky HDP-HMM actually came from considering the original HDP-HMM as a prior for an SLDS. In such scenarios where one does not have direct observations of the underlying state sequence, the issues arising from not properly capturing state persistence are exacerbated. The sticky HDP-HMM presented in this paper provides a more robust building block for developing more complex Bayesian nonparametric dynamical models.

Acknowledgements. We thank O. Vinyals, G. Friedland, and N. Morgan for helpful discussions about the NIST dataset.

References.

C.E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

M.J. Beal and P. Krishnamurthy. Gene expression time course clustering with countably infinite hidden Markov models. In *Proc. Conference on Uncertainty in Artificial Intelligence*, 2006.

M.J. Beal, Z. Ghahramani, and C.E. Rasmussen. The infinite hidden Markov model. In *NIPS 14*, pages 577–584. MIT Press, 2002.

D. Blackwell and J.B. MacQueen. Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

G. Casella and C. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2): 209–230, 1973.

E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Nonparametric Bayesian learning of switching dynamical systems. In *Advances in Neural Information Processing Systems*, volume 21, pages 457–464, 2009.

M. Gales and S. Young. *The Application of Hidden Markov Models in Speech Recognition*. Now Publishers Inc, 2008.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 2004.

M. Hoffman, P. Cook, and D. Blei. Data-driven recomposition using the hierarchical Dirichlet process hidden Markov model. In *Proc. International Computer Music Conference*, 2008.

H. Ishwaran and M. Zarepour. Markov chain Monte Carlo in approximate Dirichlet and beta two–parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

H. Ishwaran and M. Zarepour. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12: 941–963, 2002a.

H. Ishwaran and M. Zarepour. Exact and approximate sum–representations for the Dirichlet process. *Canadian Journal of Statistics*, 30:269–283, 2002b.

S. Jain and R.M. Neal. A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13:158–182, 2004.

A. Jasra, C.C. Holmes, and D.A. Stephens. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005.

M. Johnson. Why doesn't EM find good HMM POS-taggers. In *Proc. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.

J.J. Kivinen, E.B. Sudderth, and M.I. Jordan. Learning multiscale representations of natural scenes using Dirichlet processes. In *Proc. International Conference on Computer Vision*, pages 1–8, 2007.

K. Kurihara, M. Welling, and Y.W. Teh. Collapsed variational Dirichlet process mixture models. In *Proc. International Joint Conferences on Artificial Intelligence*, 2007.

J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, 5(1):32–38, 1957.

NIST. Rich transcriptions database. *http://www.nist.gov/speech/tests/rt/*, 2007.

O. Papaspiliopoulos and G.O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95:169–186, 2008.

L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2005.

A. Rodriguez, D.B. Dunson, and A.E. Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association.*, 103(483):1131–1154, 2008.

S.L. Scott. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97(457):337–351, 2002.

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

J. Van Gael, Y. Saatci, Y.W. Teh, and Z. Ghahramani. Beam sampling for the infinite hidden Markov model. In *Proc. International Conference on Machine Learning*, July 2008.

S.G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics–Simulation and Computation*, 36:45–54, 2007.

C. Wooters and M. Huijbregts. The ICSI RT07s speaker diarization system. *To appear in LNCS*, 2007.

E.P. Xing and K-A Sohn. Hidden Markov Dirichlet process: Modeling genetic inference in open ancestral space. *Bayesian Analysis*, 2(3):501–528, 2007.

APPENDIX A: NOTATIONAL CONVENTIONS

General Notation

$\mathbb{Z}_+$	the set of positive integers
$\mathbb{R}$	the set of reals
$x_{1:t}$	the sequence $\{x_1, \dots, x_t\}$
$x_{\backslash t}$	the sequence $\{x_1, \dots, x_{t-1}, x_{t+1}, \dots, x_T\}$, where T is largest possible index
$x_{\cdot b}$	$\sum_a x_{ab}$
$x_{a \cdot}$	$\sum_b x_{ab}$
$x_{\cdot\cdot}$	$\sum_b \sum_a x_{ab}$
$\lvert \cdot \rvert$	cardinality of a set
$\delta(k, j)$	the discrete Kronecker delta
δ_θ	measure concentrated at θ
$E[\cdot]$	expectation of a random variable
$\text{DP}(\alpha, H)$	Dirichlet process distribution with concentration parameter α and base measure H
$\text{Dir}(\underline{\alpha})$	K-dimensional finite Dirichlet distribution with parameter $\underline{\alpha}$
$\text{Ber}(p)$	Bernoulli distribution with parameter p
$\text{GEM}(\gamma)$	stick-breaking distribution with parameter γ

Hierarchical Dirichlet Process and CRF with Loyal Customers

y_{ji}	i^{th} observation within j^{th} group
z_{ji}	index of mixture component that generated observation y_{ji}
θ'_{ji}	(non-unique) parameter associated with observation y_{ji}
θ^*_{jt}	(non-unique) parameter, or *dish*, served at table t in restaurant j
θ_k	k^{th} unique global parameter of the mixture model
t_{ji}	table assignment for observation, or *customer*, y_{ji}
$\bar{k}_{jt}$	considered dish assignment for table t in restaurant j
k_{jt}	served dish assignment for table t in restaurant j
$\underline{\bar{k}}_j$	the set of all considered dish assignments in restaurant j
$\underline{k}_j$	the set of all served dish assignments in restaurant j
w_{jt}	override variable for table t in restaurant j
$\tilde{n}_{jt}$	number of customers at table t in restaurant j
$\bar{m}_{jk}$	number of tables in restaurant j that considered dish k
m_{jk}	number of tables in restaurant j that were served dish k
T_j	number of currently occupied tables in restaurant j
$\bar{K}$	number of unique dishes considered in the franchise
K	number of unique dishes served in the franchise

Sticky HDP-HMM

y_t observation from the hidden Markov model at time t

z_t state of the Markov chain at time t

n_{jk} number of transitions from state j to state k in $z_{1:T}$

n_{jk}^{-t} number of transitions from state j to state k in $z_{1:T}$, not counting the transitions $z_{t-1} \rightarrow z_t$ or $z_t \rightarrow z_{t+1}$

κ self-transition parameter

ρ self-transition proportion parameter $\kappa/(\alpha + \kappa)$

with DP emissions

s_t index of mixture component that generated observation y_t

n'_{kj} number of observations assigned to the k^{th} state's j^{th} mixture component

n'^{-t}_{kj} number of observations assigned to the k^{th} state's j^{th} mixture component, not counting observation y_t

K'_k number of currently instantiated mixture components for the k^{th} state's emission distribution

APPENDIX B: DIRECT ASSIGNMENT SAMPLER

This supplementary material provides the derivations for the sequential, direct assignment Gibbs samplers for the sticky HDP-HMM and sticky HDP-HMM with DP emissions. Throughout this section, we will refer to the random variables in the graph of Fig. 3(b). For these derivations we include the κ term of the sticky HDP-HMM; the derivations for the original HDP-HMM follow directly by setting $\kappa = 0$. The resulting Gibbs samplers are outlined in Algorithms 1 and 2.

B.1. Sticky HDP-HMM.

To derive the direct assignment sampler for the sticky HDP-HMM, we first assume that we sample: table assignments for each customer, t_{ji}; served dish assignments for each table, k_{jt}; considered dish assignments, $\bar{k}_{jt}$; dish override variables, w_{jt}; and the global mixture weights, β. Because of the properties of the HDP, and more specifically the stick-breaking distribution, we are able to marginalize the group-specific distributions $\tilde{\pi}_j$ and parameters θ_k and still have closed-form distributions from which to sample (since exchangeability implies that we may treat every table and dish as if it were the last, as in Eq. (4.5).) The marginalization of these variables is referred to as *Rao-Blackwellization* (Casella and Robert, 1996). The assumption of having t_{ji} and k_{jt} is a stronger assumption than that of having z_{ji} since z_{ji} can be uniquely determined from t_{ji} and k_{jt}, though not vice versa. We proceed to show that directly sampling z_{ji} instead of t_{ji} and k_{jt} is sufficient when the auxiliary variables m_{jk}, $\bar{m}_{jk}$, and w_{jt} are additionally sampled.

B.1.1. *Sampling z_t.*

The posterior distribution of z_t factors as:

$$\begin{aligned} p(z_t = k \mid z_{\backslash t}, y_{1:T}, \beta, \alpha, \kappa) &\propto \int_{\boldsymbol{\pi}} \prod_i p(\pi_i \mid \alpha, \beta, \kappa) \prod_\tau p(z_\tau \mid \pi_{z_{\tau-1}}) d\boldsymbol{\pi} \\ &\quad \int \prod_k h(\theta_k \mid \lambda) \prod_\tau f(y_\tau \mid \theta_{z_\tau}) d\boldsymbol{\theta} \\ &\propto p(z_t = k \mid z_{\backslash t}, \beta, \alpha, \kappa) p(y_t \mid y_{\backslash t}, z_t = k, z_{\backslash t}, \lambda). \end{aligned} \tag{B.1}$$

Here, $f(\cdot \mid \theta)$ is the conditional density associated with the likelihood distribution $F(\theta)$ and $h(\cdot \mid \lambda)$ with the base measure $H(\lambda)$.

The term $p(z_t = k \mid z_{\backslash t}, \beta, \alpha, \kappa)$, which arises from integration over $\boldsymbol{\pi}$, is a variant of the Chinese restaurant franchise prior, while $p(y_t \mid y_{\backslash t}, z_{t=k}, z_{\backslash t}, \lambda)$ is the likelihood of an assignment $z_t = k$ having marginalized the parameter θ_k.

The conditional distribution $p(z_t = k \mid z_{\backslash t}, \beta, \alpha, \kappa)$ of Eq. (B.1) can be written as:

$$
\begin{aligned}
p(z_t = k \mid z_{\backslash t}, \beta, \alpha, \kappa) &\propto \int_{\boldsymbol{\pi}} p(z_{t+1} \mid \pi_k) p(z_t = k \mid \pi_{z_{t-1}}) \\
&\qquad \prod_i (p(\pi_i \mid \alpha, \beta, \kappa) \prod_{\tau \mid z_{\tau-1} = i, \tau \neq t, t+1} p(z_\tau \mid \pi_i)) d\boldsymbol{\pi} \\
&\propto \int_{\boldsymbol{\pi}} p(z_{t+1} \mid \pi_k) p(z_t = k \mid \pi_{z_{t-1}}) \\
&\qquad \prod_i p(\pi_i \mid \{z_\tau \mid z_{\tau-1} = i, \tau \neq t, t+1\}, \beta, \alpha, \kappa) d\boldsymbol{\pi}. \qquad \text{(B.2)}
\end{aligned}
$$

Let $z_{t-1} = j$. If $k \neq j$, that is, assuming a change in state value at time t, then

$$
\begin{aligned}
&p(z_t = k \mid z_{\backslash t}, \beta, \alpha, \kappa) \\
&\quad \propto \int_{\pi_k} p(z_{t+1} \mid \pi_k) p(\pi_k \mid \{z_\tau \mid z_{\tau-1} = k, \tau \neq t, t+1\}, \beta, \alpha, \kappa) d\pi_k \\
&\qquad \int_{\pi_j} p(z_t = k \mid \pi_j) p(\pi_j \mid \{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\}, \beta, \alpha, \kappa) d\pi_j \\
&\quad \propto p(z_{t+1} \mid \{z_\tau \mid z_{\tau-1} = k, \tau \neq t, t+1\}, \beta, \alpha, \kappa) \\
&\qquad p(z_t = k \mid \{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\}, \beta, \alpha, \kappa). \qquad \text{(B.3)}
\end{aligned}
$$

When considering the probability of a self-transition (i.e., $k = j$), we have

$$
\begin{aligned}
p(z_t = j \mid z_{\backslash t}, \beta, \alpha, \kappa) &\propto \int_{\pi_j} p(z_{t+1} \mid \pi_j) p(z_t = j \mid \pi_j) \\
&\qquad p(\pi_j \mid \{z_\tau \mid z_{\tau-1} = k, \tau \neq t, t+1\}, \beta, \alpha, \kappa) d\pi_j \\
&\propto p(z_t = j, z_{t+1} \mid \{z_\tau \mid z_{\tau-1} = k, \tau \neq t, t+1\}, \beta, \alpha, \kappa). \qquad \text{(B.4)}
\end{aligned}
$$

These predictive distributions can be derived by standard results arising from having placed a Dirichlet prior on the parameters defining these multinomial observations z_τ. The finite Dirichlet prior is induced by considering the finite partition $\{1, \dots, K, A_{\tilde{k}}\}$ of $\mathbb{Z}_+$, where $A_{\tilde{k}} = \{K+1, K+2, \dots\}$ is the set of unrepresented state values in $z_{\backslash t}$. The properties of the DP dictate that on this finite partition, we have the following form for the group-specific transition distributions:

$$
\pi_j \mid \alpha, \beta \sim \text{Dir}(\alpha\beta_1, \dots, \alpha\beta_j + \kappa, \dots, \alpha\beta_K, \alpha\beta_{\tilde{k}}), \qquad \text{(B.5)}
$$

where $\beta_{\tilde{k}} = \sum_{i=K+1}^{\infty} \beta_i$. Using this prior, we derive the distribution of a generic set of observations generated from a single transition distribution π_i given the hyperparameters α, β, and κ:

$$
\begin{aligned}
p(\{z_\tau \mid z_{\tau-1} = i\} \mid \beta, \alpha, \kappa) &= \int_{\pi_i} p(\pi_i \mid \beta, \alpha, \kappa) p(\{z_\tau \mid z_{\tau-1} = i\} \mid \pi_i) d\pi_i \\
&= \int_{\pi_i} \frac{\Gamma(\sum_k \alpha\beta_k + \kappa\delta(k,i))}{\prod_k \Gamma(\alpha\beta_k + \kappa\delta(k,i))} \prod_{k=1}^{K+1} \pi_{jk}^{\alpha\beta_k + \kappa\delta(k,i) - 1} \prod_{k=1}^{K+1} \pi_{jk}^{n_{jk}} d\pi_i \\
&= \frac{\Gamma(\sum_k \alpha\beta_k + \kappa\delta(k,i))}{\prod_k \Gamma(\alpha\beta_k + \kappa\delta(k,i))} \frac{\prod_k \Gamma(\alpha\beta_k + \kappa\delta(k,i) + n_{jk})}{\Gamma(\sum_k \alpha\beta_k + \kappa\delta(k,i) + n_{jk})} \\
&= \frac{\Gamma(\alpha + \kappa)}{\Gamma(\alpha + \kappa + n_{i\cdot})} \prod_k \frac{\Gamma(\alpha\beta_k + \kappa\delta(k,i) + n_{jk})}{\Gamma(\alpha\beta_k + \kappa\delta(k,i))}, \qquad \text{(B.6)}
\end{aligned}
$$

where we make a slight abuse of notation in taking $\beta_{K+1} = \beta_{\tilde{k}}$. We use Eq. (B.6) to determine that the first component of Eq. (B.3) is

$$
\begin{aligned}
&p(z_t = k \mid \{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\}, \beta, \alpha, \kappa) \\
&\qquad = \frac{p(\{z_\tau \mid z_{\tau-1} = j, \tau \neq t+1, z_t = k\} \mid \beta, \alpha, \kappa)}{p(\{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\} \mid \beta, \alpha, \kappa)} \\
&\qquad = \frac{\Gamma(\alpha + \kappa + n_{j\cdot}^{-t})}{\Gamma(\alpha + n_{j\cdot}^{-t} + 1)} \frac{\Gamma(\alpha\beta_k + \kappa + n_{jk}^{-t} + 1)}{\Gamma(\alpha\beta_k + n_{jk}^{-t})} \\
&\qquad = \frac{\alpha\beta_k + n_{jk}^{-t}}{\alpha + n_{j\cdot}^{-t}}. \qquad \text{(B.7)}
\end{aligned}
$$

Here, n_{jk}^{-t} denotes the number of transitions from state j to k not counting the transition from z_{t-1} to z_t or from z_t to z_{t+1}. Similarly, the second component of Eq. (B.3) is derived to be

(B.8)

$$
p(z_{t+1} = \ell \mid \{z_\tau \mid z_{\tau-1} = k, \tau \neq t, t+1\}, \beta, \alpha, \kappa) = \frac{\alpha\beta_\ell + \kappa\delta(\ell, k) + n_{k\ell}^{-t}}{\alpha + \kappa + n_{k\cdot}^{-t}},
$$

For $k = j$, the distribution of Eq. (B.4) reduces to

$$
\begin{aligned}
&p(z_t = j, z_{t+1} \mid \{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\}, \beta, \alpha, \kappa) \\
&\quad = \frac{p(\{z_\tau \mid z_{\tau-1} = j\} \mid \beta, \alpha, \kappa)}{p(\{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\} \mid \beta, \alpha, \kappa)} \\
&\quad = \begin{cases} \frac{\Gamma(\alpha+\kappa+n_{j\cdot}^{-t})}{\Gamma(\alpha+\kappa+n_{j\cdot}^{-t}+2)} \frac{\Gamma(\alpha\beta_j+\kappa+n_{jj}^{-t}+1)}{\Gamma(\alpha\beta_j+\kappa+n_{jj}^{-t})} \frac{\Gamma(\alpha\beta_\ell+n_{j\ell}^{-t}+1)}{\Gamma(\alpha\beta_\ell+n_{j\ell}^{-t})}, & z_{t+1} = \ell, \ell \neq j; \\ \frac{\Gamma(\alpha+\kappa+n_{j\cdot}^{-t})}{\Gamma(\alpha+\kappa+n_{j\cdot}^{-t}+2)} \frac{\Gamma(\alpha\beta_j+\kappa+n_{jj}^{-t}+2)}{\Gamma(\alpha\beta_j+\kappa+n_{jj}^{-t})}, & z_{t+1} = j; \end{cases} \\
&\quad = \begin{cases} \frac{(\alpha\beta_j+\kappa+n_{jj}^{-t})(\alpha\beta_\ell+n_{j\ell}^{-t})}{(\alpha+\kappa+n_{j\cdot}^{-t}+1)(\alpha+\kappa+n_{j\cdot}^{-t})}, & z_{t+1} = \ell, \ell \neq j; \\ \frac{(\alpha\beta_j+\kappa+n_{jj}^{-t}+1)(\alpha\beta_j+\kappa+n_{jj}^{-t})}{(\alpha+\kappa+n_{j\cdot}^{-t}+1)(\alpha+\kappa+n_{j\cdot}^{-t})}, & z_{t+1} = j; \end{cases} \\
&\quad = \frac{(\alpha\beta_j + \kappa + n_{jj}^{-t})(\alpha\beta_\ell + n_{j\ell}^{-t} + (\kappa+1)\delta(j,\ell))}{(\alpha + \kappa + n_{j\cdot}^{-t})(\alpha + \kappa + n_{j\cdot}^{-t} + 1)}. \qquad \text{(B.9)}
\end{aligned}
$$

Combining these cases, the prior predictive distribution of z_t is:

$$
\begin{aligned}
&p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) \qquad \text{(B.10)} \\
&\propto \begin{cases} (\alpha\beta_k + n_{z_{t-1}k}^{-t} + \kappa\delta(z_{t-1}, k)) \\ \qquad \left(\frac{\alpha\beta_{z_{t+1}} + n_{kz_{t+1}}^{-t} + \kappa\delta(k, z_{t+1}) + \delta(z_{t-1}, k)\delta(k, z_{t+1})}{\alpha + n_{k\cdot}^{-t} + \kappa + \delta(z_{t-1}, k)} \right) & k \in \{1, \dots, K\} \\ \frac{\alpha^2 \beta_{\tilde{k}} \beta_{z_{t+1}}}{\alpha+\kappa} & k = K+1. \end{cases}
\end{aligned}
$$

The conditional distribution of the observation y_t given an assignment $z_t = k$ and given all other observations y_τ, having marginalized out θ_k, can be written as follows:

$$
\begin{aligned}
p(y_t \mid y_{\setminus t}, z_t = k, z_{\setminus t}, \lambda) &\propto \int f(y_t \mid \theta_k) h(\theta_k \mid \lambda) \prod_{\tau \mid z_\tau = k, \tau \neq t} f(y_\tau \mid \theta_k) d\theta_k \\
&\propto \int f(y_t \mid \theta_k) p(\theta_k \mid \{y_\tau \mid z_\tau = k, \tau \neq t\}, \lambda) d\theta_k \\
&\propto p(y_t \mid \{y_\tau \mid z_\tau = k, \tau \neq t\}, \lambda). \qquad \text{(B.11)}
\end{aligned}
$$

There exists a closed-form distribution for this likelihood if we consider a conjugate distribution on the parameter space Θ.

Assuming our emission distributions are Gaussian with unknown mean and covariance parameters, the conjugate prior is normal-inverse-Wishart distribution, which we denote by $\mathcal{NIW}(\zeta, \vartheta, \nu, \Delta)$. Here, $\lambda = \{\zeta, \vartheta, \nu, \Delta\}$. Via conjugacy, the posterior distribution of $\theta_k = \{\mu_k, \Sigma_k\}$ given a set of Gaussian observations

$y_t \sim \mathcal{N}(\mu_k, \Sigma_k)$ is distributed as an updated normal-inverse-Wishart $\mathcal{NIW}(\bar{\zeta}_k, \bar{\vartheta}_k, \bar{\nu}_k, \bar{\Delta}_k)$, where

$$\begin{aligned}
\bar{\zeta}_k &= \zeta + |\{y_s \mid z_s = k, s \neq t\}| \triangleq \zeta + |Y_k| \\
\bar{\nu}_k &= \nu + |Y_k| \\
\bar{\zeta}_k \bar{\vartheta}_k &= \zeta\vartheta + \sum_{y_s \in Y_k} y_s \\
\bar{\nu}_k \bar{\Delta}_k &= \nu\Delta + \sum_{y_s \in Y_k} y_s y_s^T + \zeta\vartheta\vartheta^T - \bar{\zeta}_k \bar{\vartheta}_k \bar{\vartheta}_k^T.
\end{aligned}$$

Marginalizing θ_k induces a multivariate Student-t predictive distribution for y_t (Gelman et al., 2004):

$$\begin{aligned}
p(y_t \mid \{y_\tau \mid z_\tau = k, \tau \neq t\}, \zeta, \vartheta, \nu, \Delta) &= t_{\bar{\nu}_k - d - 1}\left(y_t; \bar{\vartheta}_k, \frac{(\bar{\zeta}_k + 1)\bar{\nu}_k}{\bar{\zeta}_k(\bar{\nu}_k - d - 1)}\bar{\Delta}_k\right) \\
&\triangleq t_{\hat{\nu}_k}(y_t; \hat{\mu}_k, \hat{\Sigma}_k). \qquad \text{(B.12)}
\end{aligned}$$

B.1.2. *Sampling* β. Let $\bar{K}$ be the number of unique dishes *considered*. We note that for the sticky HDP-HMM, every served dish had to be considered in some restaurant. The only scenario in which this would not be the case is if for some dish j, every table served dish j arose from an override decision. However, overrides resulting in dish j being served can only occur in restaurant j, and this restaurant would not exist if dish j was not considered (and thus served) in some other restaurant. Therefore, each served dish had to be considered by at least one table in the franchise. On the other hand, there may be some dishes considered that were never served. From this, we conclude that $\bar{K} \geq K$. We will assume that the K served dishes are indexed in $\{1, \dots, K\}$ and any considered, but not served, dish is indexed in $\{K+1, K+2, \dots\}$. For the sake of inference, we will see in the following section that $\tilde{K}$ never exceeds K, the number of unique considered dishes, implying that $\tilde{K} = K$.

Take a finite partition $\{\theta_1, \theta_2, \dots, \theta_{\bar{K}}, \Theta_{\tilde{k}}\}$ of the parameter space Θ, where $\Theta_{\tilde{k}} = \Theta \backslash \bigcup_{k=1}^{\bar{K}} \{\theta_k\}$ is the set of all currently unrepresented parameters. By definition of the Dirichlet process, G_0 has the following distribution on this finite partition:

$$\begin{aligned}
(G_0(\theta_1), \dots, G_0(\theta_{\bar{K}}), G_0(\Theta_{\tilde{k}})) \mid \gamma, H &\sim \mathrm{Dir}(\gamma H(\theta_1), \dots, \gamma H(\theta_{\bar{K}}), \gamma H(\Theta_{\tilde{k}})) \\
&\sim \mathrm{Dir}(0, \dots, 0, \gamma), \qquad \text{(B.13)}
\end{aligned}$$

where we have used the fact that H is absolutely continuous with respect to Lebesgue measure.

For every currently instantiated table t, the considered dish assignment variable $\bar{k}_{jt}$ associates the table-specific considered dish θ^*_{jt} with one among the unique set of dishes $\{\theta_1, \dots, \theta_{\bar{K}}\}$. Recalling that $\bar{m}_{jk}$ denotes how many of the tables in restaurant j considered dish θ_k, we see that we have $\bar{m}_{\cdot k}$ observations $\theta^*_{jt} \sim G_0$ in the franchise that fall within the single-element cell $\{\theta_k\}$. By the properties of the Dirichlet distribution, the posterior of G_0 is

$$(G_0(\theta_1), \dots, G_0(\theta_{\bar{K}}), G_0(\Theta_{\tilde{k}}))|\boldsymbol{\theta^*}, \gamma \sim \mathrm{Dir}(\bar{m}_{\cdot 1}, \dots, \bar{m}_{\cdot \bar{K}}, \gamma). \tag{B.14}$$

Since $(G_0(\theta_1), \dots, G_0(\theta_{\bar{K}}), G_0(\Theta_{\tilde{k}}))$ is by definition equal to $(\beta_1, \dots, \beta_{\bar{K}}, \beta_{\tilde{k}})$, and from the conditional independencies illustrated in Fig. 3, the desired posterior of β is

$$(\beta_1, \dots, \beta_{\bar{K}}, \beta_{\tilde{k}}) \mid \boldsymbol{t}, \boldsymbol{k}, \boldsymbol{\bar{k}}, \boldsymbol{w}, y_{1:T}, \gamma \sim \mathrm{Dir}(\bar{m}_{\cdot 1}, \dots, \bar{m}_{\cdot \bar{K}}, \gamma), \tag{B.15}$$

where here we define $\beta_{\tilde{k}} = \sum_{k=\bar{K}+1}^{\infty} \beta_k$. From the above, we see that $\{\bar{m}_{\cdot k}\}_{k=1}^{\bar{K}}$ is a set of sufficient statistics for resampling β defined on this partition. Thus, it is sufficient to sample $\bar{m}_{jk}$ instead of t_{ji} and k_{jt}, when given the state index z_t. The sampling of $\bar{m}_{jk}$, as well as the resampling of hyperparameters (see Supplementary Material D), is greatly simplified by additionally sampling auxiliary variables m_{jk} and w_{jt}, corresponding to the number of tables in restaurant j that were *served* dish k and the corresponding override variables.

B.1.3. *Jointly Sampling* m_{jk}, w_{jt}, *and* $\bar{m}_{jk}$. We jointly sample the auxiliary variables m_{jk}, w_{jt}, and $\bar{m}_{jk}$ from

$$\begin{aligned} p(\boldsymbol{m}, \boldsymbol{w}, \boldsymbol{\bar{m}} \mid z_{1:T}, \beta, \alpha, \kappa) &= p(\boldsymbol{\bar{m}} \mid \boldsymbol{m}, \boldsymbol{w}, z_{1:T}, \beta, \alpha, \kappa) \\ &\quad p(\boldsymbol{w} \mid \boldsymbol{m}, z_{1:T}, \beta, \alpha, \kappa) p(\boldsymbol{m} \mid z_{1:T}, \beta, \alpha, \kappa). \end{aligned} \tag{B.16}$$

We start by examining $p(\boldsymbol{m} \mid z_{1:T}, \beta, \alpha, \kappa)$. Having the state index assignments $z_{1:T}$ effectively partitions the data (customers) into both restaurants and dishes, though the table assignments are unknown since multiple tables can be served the same dish. Thus, sampling m_{jk} is in effect equivalent to sampling table assignments for each customer *after* knowing the dish assignment. This conditional distribution given by:

$$\begin{aligned} &p(t_{ji} = t \mid k_{jt} = k, \boldsymbol{t}^{-ji}, \boldsymbol{k}^{-jt}, y_{1:T}, \beta, \alpha, \kappa) \\ &\qquad \propto p(t_{ji} \mid t_{j1}, \dots, t_{ji-1}, t_{ji+1}, \dots, t_{jT_j}, \alpha, \kappa) p(k_{jt} = k \mid \beta, \alpha, \kappa) \\ &\qquad \propto \begin{cases} \tilde{n}_{jt}^{-ji}, & t \in \{1, \dots, T_j\}; \\ \alpha\beta_k + \kappa\delta(k, j), & t = T_j + 1, \end{cases} \end{aligned} \tag{B.17}$$

where $\tilde{n}_{jt}^{-ji}$ is the number of customers sitting at table t in restaurant j, not counting y_{ji}. Similarly, $\boldsymbol{t}^{-ji}$ are the table assignments for all customers except y_{ji} and $\boldsymbol{k}^{-jt}$ are the dish assignments for all tables except table t in restaurant j. We recall that T_j is the number of currently occupied tables in restaurant j. The form of Eq. (B.17) implies that a customer's table assignment conditioned on a dish assignment k follows a DP with concentration parameter $\alpha\beta_k + \kappa\delta(k, j)$. That is,

$$t_{ji} \mid k_{jt_{ji}} = k, \boldsymbol{t}^{-ji}, \boldsymbol{k}^{-jt_{ji}}, y_{1:T}, \beta, \alpha, \kappa \sim \tilde{\pi}', \qquad \tilde{\pi}' \sim \text{GEM}(\alpha\beta_k + \kappa\delta(k, j)).$$

Then, Eq. (2.6) provides the form for the distribution over the number of unique components (i.e., tables) generated by sampling n_{jk} times from this stick-breaking distributed measure, where we note that for the HDP-HMM n_{jk} is the number of customers in restaurant j eating dish k:

$$\text{(B.18)} \quad p(m_{jk} = m \mid n_{jk}, \beta, \alpha, \kappa) = \frac{\Gamma(\alpha\beta_k + \kappa\delta(k, j))}{\Gamma(\alpha\beta_k + \kappa\delta(k, j) + n_{jk})} s(n_{jk}, m)(\alpha\beta_k + \kappa\delta(k, j))^m.$$

For large n_{jk}, it is often more efficient to sample m_{jk} by simulating the table assignments of the Chinese restaurant, as described by Eq. (B.17), rather than having to compute a large array of Stirling numbers.

We now derive the conditional distribution for the override variables w_{jt}. The table counts provide that m_{jk} tables are serving dish k in restaurant j. If $k \neq j$, we automatically have m_{jk} tables with $w_{jt} = 0$ since the served dish is not the house specialty. Otherwise, for each of the m_{jj} tables t serving dish $k_{jt} = j$, we start by assuming we know the considered dish index $\bar{k}_{jt}$, from which inference of the override parameter is trivial. We then marginalize over all possible values of this index:

$$\begin{aligned}
&p(w_{jt} \mid k_{jt} = j, \beta, \rho) \\
&\quad = \sum_{\bar{k}_{jt}=1}^{\bar{K}} p(\bar{k}_{jt}, w_{jt} \mid k_{jt} = j, \beta) + p(\bar{k}_{jt} = \bar{K} + 1, w_{jt} \mid k_{jt} = j, \beta) \\
&\quad \propto \sum_{\bar{k}_{jt}=1}^{\bar{K}} p(k_{jt} = j \mid \bar{k}_{jt}, w_{jt}) p(\bar{k}_{jt} \mid \beta) p(w_{jt} \mid \rho) \\
&\qquad + p(k_{jt} = j \mid \bar{k}_{jt} = \bar{K} + 1, w_{jt}) p(\bar{k}_{jt} = \bar{K} + 1 \mid \beta) p(w_{jt} \mid \rho) \\
\text{(B.19)} &\quad \propto \begin{cases} \beta_j(1 - \rho), & w_{jt} = 0; \\ \rho, & w_{jt} = 1, \end{cases}
\end{aligned}$$

where $\rho = \frac{\kappa}{\alpha + \kappa}$ is the prior probability that $w_{jt} = 1$. This distribution implies that having observed a served dish $k_{jt} = j$ makes it more likely that the considered dish $\bar{k}_{jt}$ was overridden via choosing $w_{jt} = 1$ than the prior suggests. This is justified by the fact that if $w_{jt} = 1$, the considered dish $\bar{k}_{jt}$ could have taken any value and the served dish would still be $k_{jt} = j$. The only other explanation of the observation $k_{jt} = j$ is that the dish was not overridden, namely $w_{jt} = 0$ occurring with prior probability $(1 - \rho)$, *and* the table considered a dish $\bar{k}_{jt} = j$, occurring with probability β_j. These events are independent, resulting in the above distribution. We draw m_{jj} i.i.d. samples of w_{jt} from Eq. (B.19), with the total number of dish overrides in restaurant j given by $w_{j\cdot} = \sum_t w_{jt}$. The sum of these Bernoulli random variables results in a binomial random variable.

Given m_{jk} for all j and k and w_{jt} for each of these instantiated tables, we can now deterministically compute $\bar{m}_{jk}$, the number of tables that *considered* ordering dish k in restaurant j. Any table that was overridden is an uninformative observation for the posterior of $\bar{m}_{jk}$ so that

$$\bar{m}_{jk} = \begin{cases} m_{jk}, & j \neq k; \\ m_{jj} - w_{j\cdot}, & j = k. \end{cases} \tag{B.20}$$

Note that we are able to subtract off the sum of the override variables within a restaurant, $w_{j\cdot}$, since the only time $w_{jt} = 1$ is if table t is served dish j. From Eq. (B.20), we see that $\bar{K} = K$.

The resulting direct assignment Gibbs sampler is outlined in Algorithm 1.

B.2. Sticky HDP-HMM with DP emissions. In this section we derive the predictive distribution of the augmented state (z_t, s_t) of the sticky HDP-HMM with DP emissions. We use the chain rule to write:

$$\begin{aligned} &p(z_t = k, s_t = j \mid z_{\backslash t}, s_{\backslash t}, y_{1:T}, \beta, \alpha, \sigma, \kappa, \lambda) \\ &\qquad = p(s_t = j \mid z_t = k, z_{\backslash t}, s_{\backslash t}, y_{1:T}, \sigma, \lambda) \\ &\qquad\qquad p(z_t = k \mid z_{\backslash t}, s_{\backslash t}, y_{1:T}, \beta, \alpha, \kappa, \lambda). \end{aligned} \tag{B.21}$$

We can examine each term of this distribution by once again considering the joint distribution over all random variables in the model and then integrating over the appropriate parameters. For the conditional distribution of $z_t = k$ when *not* given

```
Given the previous state assignments $z_{1:T}^{(n-1)}$ and global transition distribution $\beta^{(n-1)}$:

  1. Set $z_{1:T} = z_{1:T}^{(n-1)}$ and $\beta = \beta^{(n-1)}$. For each $t \in \{1, \dots, T\}$, sequentially
      (a) Decrement $n_{z_{t-1}z_t}$ and $n_{z_t z_{t+1}}$ and remove $y_t$ from the cached statistics for the
          current assignment $z_t = k$:
              $(\hat{\mu}_k, \hat{\Sigma}_k) \leftarrow (\hat{\mu}_k, \hat{\Sigma}_k) \ominus y_t$
              $\hat{\nu}_k \leftarrow \hat{\nu}_k - 1$
      (b) For each of the $K$ currently instantiated states, determine
              $f_k(y_t) = (\alpha\beta_k + n_{z_{t-1}k}) \left( \frac{\alpha\beta_{z_{t+1}} + n_{kz_{t+1}} + \kappa\delta(k, z_{t+1})}{\alpha + n_{k\cdot} + \kappa} \right) t_{\hat{\nu}_k}(y_t; \hat{\mu}_k, \hat{\Sigma}_k)$
          for $z_{t-1} \neq k$, otherwise see Eq. (B.10). Also determine probability $f_{K+1}(y_t)$ of a
          new state $K+1$.
      (c) Sample the new state assignment $z_t$:
              $z_t \sim \sum_{k=1}^{K} f_k(y_t)\delta(z_t, k) + f_{K+1}(y_t)\delta(z_t, K+1)$
          If $z_t = K+1$, then increment $K$ and transform $\beta$ as follows. Sample $b \sim \text{Beta}(1, \gamma)$
          and assign $\beta_K \leftarrow b\beta_{\tilde{k}}$ and $\beta_{\tilde{k}} \leftarrow (1-b)\beta_{\tilde{k}}$, where $\beta_{\tilde{k}} = \sum_{k=K+1}^{\infty} \beta_k$.
      (d) Increment $n_{z_{t-1}z_t}$ and $n_{z_t z_{t+1}}$ and add $y_t$ to the cached statistics for the new
          assignment $z_t = k$:
              $(\hat{\mu}_k, \hat{\Sigma}_k) \leftarrow (\hat{\mu}_k, \hat{\Sigma}_k) \oplus y_t$
              $\hat{\nu}_k \leftarrow \hat{\nu}_k + 1$

  2. Fix $z_{1:T}^{(n)} = z_{1:T}$. If there exists a $j$ such that $n_{j\cdot} = 0$ and $n_{\cdot j} = 0$, remove $j$ and decrement
     $K$.

  3. Sample auxiliary variables $\boldsymbol{m}$, $\boldsymbol{w}$, and $\bar{\boldsymbol{m}}$ as follows:
      (a) For each $(j, k) \in \{1, \dots, K\}^2$, set $m_{jk} = 0$ and $n = 0$. For each customer in
          restaurant $j$ eating dish $k$, that is for $i = 1, \dots, n_{jk}$, sample
              $x \sim \text{Ber}\left( \frac{\alpha\beta_k + \kappa\delta(j,k)}{n + \alpha\beta_k + \kappa\delta(j,k)} \right)$
          Increment $n$, and if $x = 1$ increment $m_{jk}$.
      (b) For each $j \in \{1, \dots, K\}$, sample the number of override variables in restaurant $j$:
              $w_{j\cdot} \sim \text{Binomial}\left( m_{jj}, \frac{\rho}{\rho + \beta_j(1-\rho)} \right),$
          Set the number of informative tables in restaurant $j$ considering dish $k$ to:
              $\bar{m}_{jk} = \begin{cases} m_{jk}, & j \neq k; \\ m_{jj} - w_{j\cdot}, & j = k. \end{cases}$

  4. Sample the global transition distribution from
              $\beta^{(n)} \sim \text{Dir}(\bar{m}_{\cdot 1}, \dots, \bar{m}_{\cdot K}, \gamma)$

  5. Optionally, resample the hyperparameters $\gamma$, $\alpha$, and $\kappa$ as described in Supplementary
     Material D.
```

Algorithm 1: Direct assignment Rao–Blackwellized Gibbs sampler for the sticky HDP-HMM. The algorithm for the HDP-HMM follows directly by setting $\kappa = 0$. Here, we assume Gaussian observations with a normal-inverse-Wishart prior on the parameters of these distributions (see Supplementary Material B). The $\oplus$ and $\ominus$ operators update cached mean and covariance statistics as assignments are added or removed from a given component.

s_t, this amounts to:

$$\begin{aligned} p(z_t = k \mid z_{\backslash t}, s_{\backslash t}, y_{1:T}, \beta, \alpha, \kappa, \lambda) &\propto \int_{\boldsymbol{\pi}} \prod_j p(\pi_j \mid \alpha, \beta, \kappa) \prod_\tau p(z_\tau \mid \pi_{z_{\tau-1}}) d\boldsymbol{\pi} \\ &\quad \sum_{s_t} \int_{\boldsymbol{\psi}} \prod_j p(\psi_j \mid \sigma) \prod_\tau p(s_\tau \mid \psi_{z_\tau}) d\boldsymbol{\psi} \\ &\quad\quad \int \prod_{i,\ell} h(\theta_{i,\ell} \mid \lambda) \prod_\tau f(y_\tau \mid \theta_{z_\tau, s_\tau}) d\boldsymbol{\theta} \\ &\propto p(z_t = k \mid z_{\backslash t}, \beta, \alpha, \kappa) \\ &\quad \sum_{s_t} p(s_t \mid \{s_\tau \mid z_\tau = k, \tau \neq t\}, \sigma) \\ &\quad\quad p(y_t \mid \{y_\tau \mid z_\tau = k, s_t, \tau \neq t\}, \lambda). \end{aligned} \tag{B.22}$$

The term $p(z_t = k \mid z_{\backslash t}, \beta, \alpha, \kappa)$ is as in Eq. (B.10), while

$$p(s_t = j \mid \{s_\tau \mid z_\tau = k, \tau \neq t\}, \sigma) = \begin{cases} \frac{n'^{-t}_{kj}}{\sigma + n'^{-t}_{k\cdot}}, & j \in \{1, \ldots, K'_k\}; \\ \frac{\sigma}{\sigma + n'^{-t}_{k\cdot}}, & j = K'_k + 1, \end{cases} \tag{B.23}$$

which is the predictive distribution of the indicator random variables of the DP mixture model associated with $z_t = k$. Here, n'^{-t}_{kj} is the number of observations y_τ with $(z_\tau = k, s_\tau = j)$ for $\tau \neq t$, and K'_k is the number of currently instantiated mixture components for the k^{th} emission density.

We similarly derive the conditional distribution of an assignment $s_t = j$ given $z_t = k$ as:

$$\begin{aligned} p(s_t = j \mid z_t = k, z_{\backslash t}, s_{\backslash t}, y_{1:T}, \sigma, \lambda) &\propto p(s_t = j \mid \{s_\tau \mid z_\tau = k, \tau \neq t\}, \sigma) \\ &\quad p(y_t \mid \{y_\tau \mid z_\tau = k, s_t = j, \tau \neq t\}, \lambda). \end{aligned} \tag{B.24}$$

The likelihood component of these distributions,

$$p(y_t \mid \{y_\tau \mid z_\tau = k, s_t = j, \tau \neq t\}, \lambda),$$

is derived in the same fashion as Eq. (B.12) where now we only consider the observations y_τ that are assigned to HDP-HMM state $z_\tau = k$ *and* mixture component $s_\tau = k$.

The direct assignment Gibbs sampler for the sticky HDP-HMM with DP emissions is outlined in Algorithm 2.

Given a previous set of augmented state assignments $(z_{1:T}^{(n-1)}, s_{1:T}^{(n-1)})$ and the global transition distribution $\beta^{(n-1)}$:

1. Set $(z_{1:T}, s_{1:T}) = (z_{1:T}^{(n-1)}, s_{1:T}^{(n-1)})$ and $\beta = \beta^{(n-1)}$. For each $t \in \{1, \dots, T\}$,
   (a) Decrement $n_{z_{t-1}z_t}$, $n_{z_t z_{t+1}}$, and $n'_{z_t s_t}$ and remove y_t from the cached statistics for the current assignment $(z_t, s_t) = (k, j)$:
   $$(\hat{\mu}_{k,j}, \hat{\Sigma}_{k,j}) \leftarrow (\hat{\mu}_{k,j}, \hat{\Sigma}_{k,j}) \ominus y_t$$
   $$\hat{\nu}_{k,j} \leftarrow \hat{\nu}_{k,j} - 1$$
   (b) For each of the K currently instantiated HDP-HMM states, compute
      i. The predictive conditional distribution for each of the K'_k currently instantiated mixture components associated with this HDP-HMM state
      $$f'_{k,j}(y_t) \quad = \quad \left(\frac{n'_{kj}}{\sigma + n'_{k\cdot}} \right) t_{\hat{\nu}_{k,j}}(y_t; \hat{\mu}_{k,j}, \hat{\Sigma}_{k,j})$$
      and for a new mixture component $K'_k + 1$
      $$f'_{k,K'_k+1}(y_t) \quad = \quad \frac{\sigma}{\sigma + n'_{k\cdot}} t_{\hat{\nu}_0}(y_t; \hat{\mu}_0, \hat{\Sigma}_0).$$
      ii. The predictive conditional distribution of the HDP-HMM state without knowledge of the current mixture component
      $$f_k(y_t) = (\alpha\beta_k + n_{z_{t-1}k}) \left(\frac{\alpha\beta_{z_{t+1}} + n_{kz_{t+1}} + \kappa\delta(k, z_{t+1})}{\alpha + n_{k\cdot} + \kappa} \right) \left(\sum_{j=1}^{K'_k} f'_{k,j}(y_t) + f'_{k,K'_k+1}(y_t) \right)$$
      for $z_{t-1} \neq k$, otherwise see Supplementary Material B.2. Repeat this procedure for a new HDP-HMM state $K+1$ with K'_{K+1} initialized to 0.
   (c) Sample the new augmented state assignment (z_t, s_t) by first sampling z_t:
   $$z_t \quad \sim \quad \sum_{k=1}^{K} f_k(y_t)\delta(z_t, k) + f_{K+1}(y_t)\delta(z_t, K+1).$$
   Then, conditioned on a new assignment $z_t = k$, sample s_t:
   $$s_t \quad \sim \quad \sum_{j=1}^{K'_k} f'_{k,j}(y_t)\delta(s_t, j) + f'_{k,K'_k+1}(y_t)\delta(s_t, K'_k + 1).$$
   If $k = K + 1$, then increment K and transform β as follows. Sample $b \sim \text{Beta}(1, \gamma)$ and assign $\beta_K \leftarrow b\beta_{\tilde{k}}$ and $\beta_{\tilde{k}} \leftarrow (1-b)\beta_{\tilde{k}}$. If $s_t = K'_k + 1$, then increment K'_k.
   (d) Increment $n_{z_{t-1}z_t}$, $n_{z_t z_{t+1}}$, and $n'_{z_t s_t}$ and add y_t to the cached statistics for the new assignment $(z_t, s_t) = (k, j)$:
   $$(\hat{\mu}_{k,j}, \hat{\Sigma}_{k,j}) \leftarrow (\hat{\mu}_{k,j}, \hat{\Sigma}_{k,j}) \oplus y_t$$
   $$\hat{\nu}_{k,j} \leftarrow \hat{\nu}_{k,j} + 1$$
2. Fix $(z_{1:T}^{(n)}, s_{1:T}^{(n)}) = (z_{1:T}, s_{1:T})$. If there exists a k such that $n_{k\cdot} = 0$ and $n_{\cdot k} = 0$, remove k and decrement K. Similarly, if there is a (k, j) such that $n'_{kj} = 0$ then remove j and decrement K'_k.
3. Sample auxiliary variables $\boldsymbol{m}$, $\boldsymbol{w}$, and $\bar{\boldsymbol{m}}$ as in step 3 of Algorithm 1.
4. Sample the global transition distribution $\beta^{(n)}$ as in step 4 of Algorithm 1.
5. Optionally, resample the hyperparameters σ, γ, α, and κ as described in Supplementary Material D.

Algorithm 2: Direct assignment Rao–Blackwellized Gibbs sampler for the sticky HDP-HMM with DP emissions.

APPENDIX C: BLOCKED SAMPLER

In this section, we present the derivation of the blocked Gibbs samplers for the sticky HDP-HMM and sticky HDP-HMM with DP emissions. The resulting Gibbs samplers are outlined in Algorithms 3 and 4.

C.1. Sampling $\boldsymbol{\beta}$, $\boldsymbol{\pi}$, and $\boldsymbol{\psi}$. The order L weak limit approximation to the DP gives us the following form for the prior distribution on the global weights β:

$$\beta \mid \gamma \sim \text{Dir}(\gamma/L, \dots, \gamma/L). \tag{C.1}$$

On this partition, the prior distribution over the transition probabilities is Dirichlet with parametrization:

$$\pi_j \mid \alpha, \kappa, \beta \sim \text{Dir}(\alpha\beta_1, \dots, \alpha\beta_j + \kappa, \dots, \alpha\beta_L). \tag{C.2}$$

The posterior distributions are then given by:

$$\begin{aligned} \beta \mid \bar{\boldsymbol{m}}, \gamma &\sim \text{Dir}(\gamma/L + \bar{m}_{\cdot 1}, \dots, \gamma/L + \bar{m}_{\cdot L}) \\ \pi_j \mid z_{1:T}, \alpha, \beta &\sim \text{Dir}(\alpha\beta_1 + n_{j1}, \dots, \alpha\beta_j + \kappa + n_{jj}, \dots, \alpha\beta_L + n_{jL}), \end{aligned} \tag{C.3}$$

where we recall that n_{jk} is the number of j to k transitions in the state sequence $z_{1:T}$ and $\bar{m}_{jk}$ is the number of tables in restaurant j that considered dish k. The sampling of the auxiliary variables $\bar{m}_{jk}$ is as in Supplementary Material B.

For the sticky HDP-HMM with DP emissions, an order L' weak limit approximation to the DP prior on the emission parameters yields the following posterior distribution on the mixture weights ψ_k:

$$\psi_k \mid z_{1:T}, s_{1:T}, \sigma \sim \text{Dir}(\sigma/L' + n'_{k1}, \dots, \sigma/L' + n'_{kL'}), \tag{C.4}$$

where $n'_{k\ell}$ is the number of observations assigned to the ℓ^{th} mixture component of the k^{th} HMM state.

C.2. Sampling $\boldsymbol{z_{1:T}}$ for the Sticky HDP-HMM. To derive the forward-backward procedure for jointly sampling $z_{1:T}$ given $y_{1:T}$ for the sticky HDP-HMM, we first note that

$$\begin{aligned} p(z_{1:T} \mid y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) = p(z_T \mid z_{T-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(z_{T-1} \mid z_{T-2}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \\ \cdots p(z_2 \mid z_1, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) p(z_1 \mid y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}). \end{aligned}$$

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

$$\begin{aligned} p(z_1 \mid y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) &\propto p(z_1) f(y_1 \mid \theta_{z_1}) \sum_{z_{2:T}} \prod_t p(z_t \mid \pi_{z_{t-1}}) f(y_t \mid \theta_{z_t}) \\ &\propto p(z_1) f(y_1 \mid \theta_{z_1}) \sum_{z_2} p(z_2 \mid \pi_{z_1}) f(y_2 \mid \theta_{z_2}) m_{3,2}(z_2) \\ &\propto p(z_1) f(y_1 \mid \theta_{z_1}) m_{2,1}(z_1), \end{aligned} \tag{C.5}$$

where $m_{t,t-1}(z_{t-1})$ is the backward message passed from z_t to z_{t-1} and for an HMM is recursively defined by:

$$\begin{aligned} m_{t,t-1}(z_{t-1}) &\propto \begin{cases} \sum_{z_t} p(z_t \mid \pi_{z_{t-1}}) f(y_t \mid \theta_{z_t}) m_{t+1,t}(z_t), & t \leq T; \\ 1, & t = T+1; \end{cases} \\ &\propto p(y_{t:T} \mid z_{t-1}, \boldsymbol{\pi}, \boldsymbol{\theta}). \end{aligned} \tag{C.6}$$

The general conditional distribution of z_t is:

$$p(z_t \mid z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto p(z_t \mid \pi_{z_{t-1}}) f(y_t \mid \theta_{z_t}) m_{t+1,t}(z_t). \tag{C.7}$$

The resulting blocked Gibbs sampler is outlined in Algorithm 3.

C.3. Sampling $(z_{1:T}, s_{1:T})$ for the Sticky HDP-HMM with DP emissions. We now examine how to sample the augmented state (z_t, s_t) of the sticky HDP-HMM with DP emissions. The conditional distribution of (z_t, s_t) for the forward-backward procedure is derived as:

$$p(z_t, s_t \mid z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{\theta}) \propto p(z_t \mid \pi_{z_{t-1}}) p(s_t \mid \psi_{z_t}) f(y_t \mid \theta_{z_t, s_t}) m_{t+1,t}(z_t). \tag{C.8}$$

Since the Markovian structure is only on the z_t component of the augmented state, the backward message $m_{t,t-1}(z_{t-1})$ from (z_t, s_t) to (z_{t-1}, s_{t-1}) is solely a function of z_{t-1}. These messages are given by:

$$\begin{aligned} &m_{t,t-1}(z_{t-1}) \\ &\quad \propto \begin{cases} \sum_{z_t} \sum_{s_t} p(z_t \mid \pi_{z_{t-1}}) p(s_t \mid \psi_{z_t}) f(y_t \mid \theta_{z_t, s_t}) m_{t+1,t}(z_t), & t \leq T; \\ 1, & t = T+1. \end{cases} \end{aligned} \tag{C.9}$$

More specifically, since each component j of the k^{th} state-specific emission distribution is a Gaussian with parameters $\theta_{j,k} = \{\mu_{k,j}, \Sigma_{k,j}\}$, we have:

$$\begin{aligned} &p(z_t = k, s_t = j \mid z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{\theta}) \\ &\qquad \propto \pi_{z_{t-1}}(k) \psi_k(j) \mathcal{N}(y_t; \mu_{k,j}, \Sigma_{k,j}) m_{t+1,t}(k) \end{aligned} \tag{C.10}$$

Given a previous set of state-specific transition probabilities $\boldsymbol{\pi}^{(n-1)}$, the global transition distribution $\beta^{(n-1)}$, and emission parameters $\boldsymbol{\theta}^{(n-1)}$:

1. Set $\boldsymbol{\pi} = \boldsymbol{\pi}^{(n-1)}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(n-1)}$. Working sequentially backwards in time, calculate messages $m_{t,t-1}(k)$:
   (a) For each $k \in \{1, \dots, L\}$, initialize messages to
   $$m_{T+1,T}(k) = 1$$
   (b) For each $t \in \{T-1, \dots, 1\}$ and for each $k \in \{1, \dots, L\}$, compute
   $$m_{t,t-1}(k) = \sum_{j=1}^{L} \pi_k(j) \mathcal{N}(y_t; \mu_j, \Sigma_j) m_{t+1,t}(j)$$
2. Sample state assignments $z_{1:T}$ working sequentially forward in time, starting with $n_{jk} = 0$ and $\mathcal{Y}_k = \emptyset$ for each $(j,k) \in \{1, \dots, L\}^2$:
   (a) For each $k \in \{1, \dots, L\}$, compute the probability
   $$f_k(y_t) = \pi_{z_{t-1}}(k) \mathcal{N}(y_t; \mu_k, \Sigma_k) m_{t+1,t}(k)$$
   (b) Sample a state assignment z_t:
   $$z_t \sim \sum_{k=1}^{L} f_k(y_t) \delta(z_t, k)$$
   (c) Increment $n_{z_{t-1}z_t}$ and add y_t to the cached statistics for the new assignment $z_t = k$:
   $$\mathcal{Y}_k \leftarrow \mathcal{Y}_k \oplus y_t$$
3. Sample the auxiliary variables $\boldsymbol{m}$, $\boldsymbol{w}$, and $\bar{\boldsymbol{m}}$ as in step 3 of Algorithm 1.
4. Update the global transition distribution by sampling
   $$\beta \sim \text{Dir}(\gamma/L + \bar{m}_{\cdot 1}, \dots, \gamma/L + \bar{m}_{\cdot L})$$
5. For each $k \in \{1, \dots, L\}$, sample a new transition distribution and emission parameter based on the sampled state assignments
   $$\begin{aligned} \pi_k &\sim \text{Dir}(\alpha\beta_1 + n_{k1}, \dots, \alpha\beta_k + \kappa + n_{kk}, \dots, \alpha\beta_L + n_{kL}) \\ \theta_k &\sim p(\theta \mid \lambda, \mathcal{Y}_k) \end{aligned}$$
   See Supplementary Material C.4.1 for details on resampling θ_k.
6. Fix $\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}$, $\beta^{(n)} = \beta$, and $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}$.
7. Optionally, resample the hyperparameters γ, α, and κ as described in Supplementary Material D.

Algorithm 3: Blocked Gibbs sampler for the sticky HDP-HMM. The algorithm for the original HDP-HMM follows directly by setting $\kappa = 0$. Here, we assume Gaussian observations with an independent Gaussian prior on the mean and inverse-Wishart prior on the covariance (see Supplementary Material C.4.1). The set $\mathcal{Y}_k$ is comprised of the statistics obtained from the observations assigned to state k that are necessary for updating the parameter $\theta_k = \{\mu_k, \Sigma_k\}$. The $\oplus$ operator updates these cached statistics as a new assignment is made.

$$m_{t+1,t}(k) = \sum_{i=1}^{L}\sum_{\ell=1}^{L'} \pi_k(i)\psi_i(\ell)\mathcal{N}(y_{t+1};\mu_{i,\ell},\Sigma_{i,\ell})m_{t+2,t+1}(i) \tag{C.11}$$

$$m_{T+1,T}(k) = 1 \quad k = 1,\dots,L. \tag{C.12}$$

C.4. Sampling $\boldsymbol{\theta}$. Depending on the form of the emission distribution and base measure on the parameter space Θ, we sample parameters for each of the currently instantiated states from the updated posterior distribution. For the sticky HDP-HMM, this distribution is:

$$\theta_j \mid z_{1:T}, y_{1:T}, \lambda \sim p(\theta \mid \{y_t \mid z_t = j\}, \lambda). \tag{C.13}$$

For the sticky HDP-HMM with DP emissions, the posterior distribution for each Gaussian's mean and covariance, $\theta_{k,j}$, is determined by the observations assigned to this component, namely,

$$\theta_{k,j} \mid z_{1:T}, s_{1:T}, y_{1:T}, \lambda \sim p(\theta \mid \{y_t \mid (z_t = k, s_t = j)\}, \lambda). \tag{C.14}$$

The resulting blocked Gibbs sampler for sticky HDP-HMM with DP emissions is outlined in Algorithm 4.

C.4.1. *Non-Conjugate Base Measures.* Since the blocked sampler instantiates the parameters θ_k, rather than marginalizing them as in the direct assignment sampler, we can place a non-conjugate base measure on the parameter space Θ. Take, for example, the case of single Gaussian emission distributions where the parameters are the means and covariances of these distributions. Here, $\theta_k = \{\mu_k, \Sigma_k\}$. In this situation, one may place a Gaussian prior $\mathcal{N}(\mu_0, \Sigma_0)$ on the mean μ_k and an inverse-Wishart IW(ν, Δ) prior on the covariance Σ_k.

At any given iteration of the sampler, there is a set of observations $Y_k = \{y_t \mid z_t = k\}$ with cardinality $|Y_k|$. The posterior distributions over the mean and covariance parameters are:

$$\begin{aligned} \Sigma_k \mid \mu_k &\sim \text{IW}(\bar{\nu}_k\bar{\Delta}_k, \bar{\nu}_k) \\ \mu_k \mid \Sigma_k &\sim \mathcal{N}(\bar{\mu}_k, \bar{\Sigma}_k), \end{aligned} \tag{C.15}$$

where

$$\begin{aligned} \bar{\nu}_k &= \nu + |Y_k| \\ \bar{\nu}_k\bar{\Delta}_k &= \nu\Delta + \sum_{t\in Y_k}(y_t - \mu_k)(y_t - \mu_k)' \\ \bar{\Sigma}_k &= (\Sigma_0^{-1} + |Y_k|\Sigma_k^{-1})^{-1} \\ \bar{\mu}_k &= \bar{\Sigma}_k\left(\Sigma_0^{-1}\mu_0 + \Sigma_k\sum_{t\in Y_k} y_t\right). \end{aligned}$$

```
Given a previous set of state-specific transition probabilities $\boldsymbol{\pi}^{(n-1)}$, emission mixture weights $\boldsymbol{\psi}^{(n-1)}$, global transition distribution $\beta^{(n-1)}$, and emission parameters $\boldsymbol{\theta}^{(n-1)}$:

  1. Set $\boldsymbol{\pi} = \boldsymbol{\pi}^{(n-1)}$, $\boldsymbol{\psi} = \boldsymbol{\psi}^{(n-1)}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(n-1)}$. Working sequentially backwards in time,
     calculate messages $m_{t,t-1}(k)$ :
       (a) For each $k \in \{1, \dots, L\}$, initialize messages to
               $m_{T+1,T}(k) = 1$
       (b) For each $t \in \{T-1, \dots, 1\}$ and for each $k \in \{1, \dots, L\}$, compute
               $m_{t,t-1}(k) = \sum_{i=1}^{L} \sum_{\ell=1}^{L'} \pi_k(i)\psi_i(\ell)\mathcal{N}(y_{t+1}; \mu_{i,\ell}, \Sigma_{i,\ell})m_{t+1,t}(i)$
  2. Sample augmented state assignments $(z_{1:T}, s_{1:T})$ working sequentially forward in time.
     Start with $n_{ik} = 0$, $n'_{kj} = 0$, and $\mathcal{Y}_{k,j} = \emptyset$ for $(i,k) \in \{1, \dots, L\}^2$ and
     $(k,j) \in \{1, \dots, L\} \times \{1, \dots, L'\}$.
       (a) For each $(k,j) \in \{1, \dots, L\} \times \{1, \dots, L'\}$, compute the probability
               $f_{k,j}(y_t) = \pi_{z_{t-1}}(k)\psi_k(j)\mathcal{N}(y_t; \mu_{k,j}, \Sigma_{k,j})m_{t+1,t}(k)$
       (b) Sample an augmented state assignment $(z_t, s_t)$:
               $(z_t, s_t) \sim \sum_{k=1}^{L} \sum_{j=1}^{L'} f_{k,j}(y_t)\delta(z_t, k)\delta(s_t, j)$
       (c) Increment $n_{z_{t-1}z_t}$ and $n'_{z_t s_t}$ and add $y_t$ to the cached statistics for the new
           assignment $(z_t, s_t) = (k,j)$:
               $\mathcal{Y}_{k,j} \leftarrow \mathcal{Y}_{k,j} \oplus y_t$
  3. Sample the auxiliary variables $\boldsymbol{m}$, $\boldsymbol{w}$, and $\bar{\boldsymbol{m}}$ as in step 3 of Algorithm 1.
  4. Update the global transition distribution $\beta$ as in step 4 of Algorithm 3.
  5. For each $k \in \{1, \dots, L\}$,
       (a) Sample a new transition distribution $\pi_k$ and emission mixture weights $\psi_k$:
               $\pi_k \sim \mathrm{Dir}(\alpha\beta_1 + n_{k1}, \dots, \alpha\beta_k + \kappa + n_{kk}, \dots, \alpha\beta_L + n_{kL})$
               $\psi_k \sim \mathrm{Dir}(\sigma/L' + n'_{k1}, \dots, \sigma/L' + n'_{kL'})$
       (b) For each $j \in \{1, \dots, L'\}$, sample the parameters associated with the $j^{th}$ mixture
           component of the $k^{th}$ emission distribution:
               $\theta_{k,j} \sim p(\theta \mid \lambda, \mathcal{Y}_{k,j})$
           See Supplementary Material C.4.1 for details on resampling $\theta_{k,j}$.
  6. Fix $\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}$, $\boldsymbol{\psi}^{(n)} = \boldsymbol{\psi}$, $\beta^{(n)} = \beta$, and $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}$.
  7. Optionally, resample the hyperparameters $\sigma$, $\gamma$, $\alpha$, and $\kappa$ as described in Supplementary
     Material D.
```

Algorithm 4: Blocked Gibbs sampler for the sticky HDP-HMM with DP emissions. Here, we use an independent Gaussian prior on the mean and inverse-Wishart prior on the covariance (see Supplementary Material C.4.1). The set $\mathcal{Y}_{k,j}$ is comprised of the statistics obtained from the observations assigned to augmented state (k,j) that are necessary for updating the parameter $\theta_{k,j} = \{\mu_{k,j}, \Sigma_{k,j}\}$. The $\oplus$ operator updates these cached statistics as a new assignment is made.

The sampler alternates between sampling μ_k given Σ_k and Σ_k given μ_k several times before moving on to the next stage in the sampling algorithm. The equations for the sticky HDP-HMM with DP emissions follow directly by considering $Y_{k,j} = \{y_t \mid z_t = k, s_t = j\}$ when resampling parameter $\theta_{k,j} = \{\mu_{k,j}, \Sigma_{k,j}\}$.

APPENDIX D: HYPERPARAMETERS

In this section we present the derivations of the conditional distributions for the hyperparameters of the sticky HDP-HMM. These hyperparameters include α, κ, γ, σ, and λ, where λ is considered fixed. Many of these derivations follow directly from those presented in Escobar and West (1995); Teh et al. (2006).

We parameterize our model by $(\alpha+\kappa)$ and $\rho = \kappa/(\alpha+\kappa)$; this simplifies the resulting sampler. We place Gamma(a,b) priors on each of the concentration parameters $(\alpha+\kappa)$, γ, and σ, and a Beta(c,d) prior on ρ. The a and b parameters of the gamma hyperprior may differ for each of the concentration parameters. In the following sections, we derive the resulting posterior distribution of these hyperparameters.

D.1. Posterior of $(\alpha+\kappa)$. Let us assume that there are J restaurants in the franchise at a given iteration of the sampler. Note that for the HDP-HMM, $J = K$. As depicted in Fig. 3(b), the generative model dictates that for each restaurant j we have $\tilde{\pi}_j \sim$ GEM$(\alpha+\kappa)$, and a table assignment is determined for each customer by $t_{ji} \sim \tilde{\pi}_j$. In total there are $n_{j\cdot}$ draws from this stick-breaking measure over table assignments resulting in $m_{j\cdot}$ unique tables. By Eq. (2.6) and using the fact that the restaurants are mutually conditionally independent, we may write:

$$\begin{aligned} p(\alpha+\kappa \mid m_{1\cdot},\dots,m_{J\cdot},n_{1\cdot},\dots,n_{J\cdot}) & \\ &\propto p(\alpha+\kappa)p(m_{1\cdot},\dots,m_{J\cdot} \mid \alpha+\kappa, n_{1\cdot},\dots,n_{J\cdot}) \\ &\propto p(\alpha+\kappa)\prod_{j=1}^{J} p(m_{j\cdot} \mid \alpha+\kappa, n_{j\cdot}) \\ &\propto p(\alpha+\kappa)\prod_{j=1}^{J} s(n_{j\cdot},m_{j\cdot})(\alpha+\kappa)^{m_{j\cdot}}\frac{\Gamma(\alpha+\kappa)}{\Gamma(\alpha+\kappa+n_{j\cdot})} \\ &\propto p(\alpha+\kappa)(\alpha+\kappa)^{m_{\cdot\cdot}}\prod_{j=1}^{J}\frac{\Gamma(\alpha+\kappa)}{\Gamma(\alpha+\kappa+n_{j\cdot})}. \end{aligned} \tag{D.1}$$

Using the fact that the gamma function has the property $\Gamma(z+1) = z\Gamma(z)$ and is related to the beta function via $\beta(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$, we rewrite this distribution as

$$\begin{aligned} p(\alpha+\kappa \mid m_{1\cdot},\dots,m_{J\cdot},n_{1\cdot},\dots,n_{J\cdot}) & \\ &\propto p(\alpha+\kappa)(\alpha+\kappa)^{m_{\cdot\cdot}}\prod_{j=1}^{J}\frac{(\alpha+\kappa+n_{j\cdot})\beta(\alpha+\kappa+1,n_{j\cdot})}{(\alpha+\kappa)\Gamma(n_{j\cdot})} \\ &= p(\alpha+\kappa)(\alpha+\kappa)^{m_{\cdot\cdot}}\prod_{j=1}^{J}\left(1+\frac{n_{j\cdot}}{\alpha+\kappa}\right)\int_0^1 r_j^{\alpha+\kappa}(1-r_j)^{n_{j\cdot}-1}dr_j, \end{aligned} \tag{D.2}$$

where the second equality arises from the fact that $\beta(x,y)=\int_0^1 t^{x-1}(1-t)^{y-1}dt$. We introduce a set of auxiliary random variables $r=\{r_1,\dots,r_J\}$, where each $r_j\in[0,1]$. Now, we augment the posterior with these auxiliary variables as follows:

$$\begin{aligned}
&p(\alpha+\kappa,r\mid m_{1\cdot},\dots,m_{J\cdot},n_{1\cdot},\dots,n_{J\cdot})\\
&\qquad\propto p(\alpha+\kappa)(\alpha+\kappa)^{m_{\cdot\cdot}}\prod_{j=1}^J\left(1+\frac{n_{j\cdot}}{\alpha+\kappa}\right)r_j^{\alpha+\kappa}(1-r_j)^{n_{j\cdot}-1}\\
&\qquad\propto(\alpha+\kappa)^{a+m_{\cdot\cdot}-1}e^{-(\alpha+\kappa)b}\prod_{j=1}^J\left(1+\frac{n_{j\cdot}}{\alpha+\kappa}\right)r_j^{\alpha+\kappa}(1-r_j)^{n_{j\cdot}-1}\\
&\qquad=(\alpha+\kappa)^{a+m_{\cdot\cdot}-1}e^{-(\alpha+\kappa)b}\prod_{j=1}^J\sum_{s_j\in\{0,1\}}\left(\frac{n_{j\cdot}}{\alpha+\kappa}\right)^{s_j}r_j^{\alpha+\kappa}(1-r_j)^{n_{j\cdot}-1}.
\end{aligned}\tag{D.3}$$

Here, we have used the fact that we placed a Gamma(a,b) prior on $(\alpha+\kappa)$. We add another set of auxiliary variables $s=\{s_1,\dots,s_J\}$, with each $s_j\in\{0,1\}$, to further simplify this distribution. The joint distribution over $(\alpha+\kappa)$, r, and s is given by

$$\begin{aligned}
&p(\alpha+\kappa,r,s\mid m_{1\cdot},\dots,m_{J\cdot},n_{1\cdot},\dots,n_{J\cdot})\\
&\qquad\propto(\alpha+\kappa)^{a+m_{\cdot\cdot}-1}e^{-(\alpha+\kappa)b}\prod_{j=1}^J\left(\frac{n_{j\cdot}}{\alpha+\kappa}\right)^{s_j}r_j^{\alpha+\kappa}(1-r_j)^{n_{j\cdot}-1}.
\end{aligned}\tag{D.4}$$

Each conditional distribution is as follows:

$$\begin{aligned}
&p(\alpha+\kappa\mid r,s,m_{1\cdot},\dots,m_{J\cdot},n_{1\cdot},\dots,n_{J\cdot})\\
&\qquad\propto(\alpha+\kappa)^{a+m_{\cdot\cdot}-1-\sum_{j=1}^J s_j}e^{-(\alpha+\kappa)(b-\sum_{j=1}^J\log r_j)}\\
&\qquad=\text{Gamma}(a+m_{\cdot\cdot}-\sum_{j=1}^J s_j,b-\sum_{j=1}^J\log r_j)
\end{aligned}$$

$$\begin{aligned}
p(r_j\mid\alpha+\kappa,r_{\backslash j},s,m_{1\cdot},\dots,m_{J\cdot},n_{1\cdot},\dots,n_{J\cdot})&\propto r_j^{\alpha+\kappa}(1-r_j)^{n_{j\cdot}-1}\\
&=\text{Beta}(\alpha+\kappa+1,n_{j\cdot})\\
p(s_j\mid\alpha+\kappa,r,s_{\backslash j},m_{1\cdot},\dots,m_{J\cdot},n_{1\cdot},\dots,n_{J\cdot})&\propto\left(\frac{n_{j\cdot}}{\alpha+\kappa}\right)^{s_j}\\
&=\text{Ber}\left(\frac{n_{j\cdot}}{n_{j\cdot}+\alpha+\kappa}\right).
\end{aligned}$$

D.2. Posterior of γ. We may similarly derive the conditional distribution of γ. The generative model depicted in Fig. 3(b) dictates that $\beta\sim\text{GEM}(\gamma)$ and that each table t considers ordering a dish $\bar{k}_{jt} \sim \beta$. From Eq. (B.20), we see that the sampled value $\bar{m}_{j\cdot}$ represents the total number of tables in restaurant j where the considered dish $\bar{k}_{jt}$ was the served dish k_{jt} (i.e., the number of tables with considered dishes that were not overridden.) Thus, $\bar{m}_{\cdot\cdot}$ is the total number of *informative* draws from β. If K is the number of unique *served* dishes, which can be inferred from $z_{1:T}$, then the number of unique *considered* dishes at the informative tables is:

$$\bar{K} = \sum_{k=1}^{K} \mathbf{1}(\bar{m}_{\cdot k} > 0) = K - \sum_{k=1}^{K} \mathbf{1}(\bar{m}_{\cdot k} = 0 \text{ and } m_{kk} > 0). \tag{D.5}$$

We use the notation $\mathbf{1}(A)$ to represent an indicator random variable that is 1 if the event A occurs and 0 otherwise. The only case where $\bar{K}$ is not equivalent to K is if every instance of a served dish k arose from an override in restaurant k and this dish was never considered in any other restaurant. That is, there were no informative considerations of dish k, implying $\bar{m}_{\cdot k} = 0$, while dish k was served in restaurant k, implying $m_{kk} > 0$ so that k is counted in K. This is equivalent to counting how many dishes k had an informative table consider ordering dish k, regardless of the restaurant. We may now use Eq. (2.6) to form the conditional distribution on γ:

$$\begin{aligned} p(\gamma \mid \bar{K}, \bar{m}_{\cdot\cdot}) &\propto p(\gamma) p(\bar{K} \mid \gamma, \bar{m}_{\cdot\cdot}) \\ &\propto p(\gamma) s(\bar{m}_{\cdot\cdot}, \bar{K}) \gamma^{\bar{K}} \frac{\Gamma(\gamma)}{\Gamma(\gamma + \bar{m}_{\cdot\cdot})} \\ &\propto p(\gamma) \gamma^{\bar{K}} \frac{(\gamma + \bar{m}_{\cdot\cdot}) \beta(\gamma + 1, \bar{m}_{\cdot\cdot})}{\gamma \Gamma(\bar{m}_{\cdot\cdot})} \\ &\propto p(\gamma) \gamma^{\bar{K}-1} (\gamma + \bar{m}_{\cdot\cdot}) \int_0^1 \eta^{\gamma} (1-\eta)^{\bar{m}_{\cdot\cdot}-1} d\eta. \end{aligned} \tag{D.6}$$

As before, we introduce an auxiliary random variable $\eta \in [0, 1]$ so that the joint distribution over γ and η can be written as

$$\begin{aligned} p(\gamma, \eta \mid \bar{K}, \bar{m}_{\cdot\cdot}) &\propto p(\gamma) \gamma^{\bar{K}-1} (\gamma + \bar{m}_{\cdot\cdot}) \eta^{\gamma} (1-\eta)^{\bar{m}_{\cdot\cdot}-1} \\ &\propto \gamma^{a+\bar{K}-2} (\gamma + \bar{m}_{\cdot\cdot}) e^{-\gamma(b - \log \eta)} (1-\eta)^{\bar{m}_{\cdot\cdot}-1}. \end{aligned} \tag{D.7}$$

Here, we have used the fact that there is a Gamma(a, b) prior on γ. We may add an indicator random variable $\zeta \in \{0, 1\}$ as we did in Eq. (D.4), such that

$$p(\gamma, \eta, \zeta \mid \bar{K}, \bar{m}_{\cdot\cdot}) \propto \gamma^{a+\bar{K}-1} \left(\frac{\bar{m}_{\cdot\cdot}}{\gamma} \right)^{\zeta} e^{-\gamma(b - \log \eta)} (1-\eta)^{\bar{m}_{\cdot\cdot}-1}.$$

The resulting conditional distributions are given by:

$$
\begin{aligned}
p(\gamma \mid \eta, \zeta, \bar{K}, \bar{m}_{\cdot\cdot}) &\propto \gamma^{a+\bar{K}-1-\zeta} e^{-\gamma(b-\log\eta)} \\
&= \text{Gamma}(a+\bar{K}-\zeta, b-\log\eta) \\
p(\eta \mid \gamma, \zeta, \bar{K}, \bar{m}_{\cdot\cdot}) &\propto \eta^{\gamma}(1-\eta)^{\bar{m}_{\cdot\cdot}-1} = \text{Beta}(\gamma+1, \bar{m}_{\cdot\cdot}) \\
p(\zeta \mid \gamma, \eta, \bar{K}, \bar{m}_{\cdot\cdot}) &\propto \left(\frac{\bar{m}_{\cdot\cdot}}{\gamma}\right)^{\zeta} = \text{Ber}\left(\frac{\bar{m}_{\cdot\cdot}}{\bar{m}_{\cdot\cdot}+\gamma}\right). \qquad \text{(D.8)}
\end{aligned}
$$

Alternatively, we can directly identify Eq (D.7) as leading to a conditional distribution on γ that is a simple mixture of two Gamma distributions:

$$
\begin{aligned}
p(\gamma \mid \eta, \bar{K}, \bar{m}_{\cdot\cdot}) &\propto \gamma^{a+\bar{K}-2}(\gamma+\bar{m}_{\cdot\cdot}) e^{-\gamma(b-\log\eta)} \\
&\propto \pi_{\bar{m}}\text{Gamma}(a+\bar{K}, b-\log\eta) \qquad \text{(D.9)} \\
&\quad +(1-\pi_{\bar{m}})\text{Gamma}(a+\bar{K}-1, b-\log\eta) \\
p(\eta \mid \gamma, \bar{K}, \bar{m}_{\cdot\cdot}) &\propto \eta^{\gamma}(1-\eta)^{\bar{m}_{\cdot\cdot}-1} = \text{Beta}(\gamma+1, \bar{m}_{\cdot\cdot}), \qquad \text{(D.10)}
\end{aligned}
$$

where

$$\pi_{\bar{m}} = \frac{a+\bar{K}-1}{\bar{m}_{\cdot\cdot}(b-\log\eta)}.$$

The distribution in Eq. (D.3) would lead to a much more complicated mixture of Gamma distributions. The addition of auxiliary variables s_j greatly simplifies the interpretation of the distribution.

D.3. Posterior of $\boldsymbol{\sigma}$. The derivation of the conditional distribution on σ is similar to that of $(\alpha+\kappa)$ in that we have J distributions $\psi_j \sim \text{GEM}(\sigma)$. The state-specific mixture component index is generated as $s_t \sim \psi_{z_t}$ implying that we have $n_{j\cdot}$ total draws from ψ_j, one for each occurrence of $z_t = j$. Let K'_j be the number of unique mixture components associated with these draws from ψ_j. Then, after adding auxiliary variables r' and s', the conditional distributions of σ and these auxiliary variables are:

$$
\begin{aligned}
& p(\sigma \mid r', s', K'_{1\cdot}, \ldots, K'_{J\cdot}, n_{1\cdot}, \ldots, n_{J\cdot}) \\
&\qquad\qquad \propto (\sigma)^{a+K'_{\cdot\cdot}-1-\sum_{j=1}^{J} s'_j} e^{-(\sigma)(b-\sum_{j=1}^{J}\log r'_j)} \\
& p(r'_j \mid \sigma, r'_{\setminus j}, s', K'_{1\cdot}, \ldots, K'_{J\cdot}, n_{1\cdot}, \ldots, n_{J\cdot}) \propto r'^{\sigma}_j (1-r'_j)^{n_{j\cdot}-1} \\
& p(s'_j \mid \sigma, r', s'_{\setminus j}, K'_{1\cdot}, \ldots, K'_{J\cdot}, n_{1\cdot}, \ldots, n_{J\cdot}) \propto \left(\frac{n_{j\cdot}}{\sigma}\right)^{s'_j}.
\end{aligned}
$$

In practice, it is useful to alternate between sampling the auxiliary variables and concentration parameters α, γ, and σ for several iterations before moving to sampling the other variables of this model.

D.4. Posterior of $\boldsymbol{\rho}$. Finally, we derive the conditional distribution of ρ. We have $m_{\cdot\cdot} = \sum_k m_{\cdot k}$ total draws of $w_{jt} \sim \mathrm{Ber}(\rho)$, with $\sum_j w_{j\cdot}$ the number of Bernoulli successes. Here, each success represents a table's considered dish being overridden by the house specialty dish. Using these facts, and the Beta(c,d) prior on ρ, we have

$$\begin{aligned}
p(\rho \mid \boldsymbol{w}) &\propto p(\boldsymbol{w} \mid \rho)p(\rho) \\
&\propto \binom{m_{\cdot\cdot}}{\sum_j w_{j\cdot}} \rho^{\sum_j w_{j\cdot}}(1-\rho)^{m_{\cdot\cdot}-\sum_j w_{j\cdot}} \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)}\rho^{c-1}(1-\rho)^{d-1} \\
&\propto \rho^{\sum_j w_{j\cdot}+c-1}(1-\rho)^{m_{\cdot\cdot}-\sum_j w_{j\cdot}+d-1} \\
&\propto \mathrm{Beta}\left(\sum_j w_{j\cdot}+c, m_{\cdot\cdot}-\sum_j w_{j\cdot}+d\right). \qquad \text{(D.11)}
\end{aligned}$$

DEPARTMENT OF EECS
77 MASSACHUSETTS AVE.
CAMBRIDGE, MA 02139
E-MAIL: ebfox@mit.edu
willsky@mit.edu

DEPARTMENT OF EECS
527 SODA HALL
BERKELEY, CA 94720
E-MAIL: sudderth@eecs.berkeley.edu

DEPARTMENT OF STATISTICS AND DEPARTMENT OF EECS
427 EVANS HALL
BERKELEY, CA 94720
E-MAIL: jordan@stat.berkeley.edu