Developing a tempered HDP-HMM for Systems with State Persistence

Emily B. Fox*, Erik B. Sudderth†, Michael I. Jordan† and Alan S. Willsky*
Department of Electrical Engineering and Computer Science
*Massachusetts Institute of Technology
†University of California, Berkeley
ebfox@mit.edu, sudderth@eecs.berkeley.edu, jordan@eecs.berkeley.edu, willsky@mit.edu

I. Introduction

Many real-world processes, as diverse as speech signals, the human genome, and financial time-series, can be modeled via a hidden Markov model (HMM). The HMM assumes that observations are generated by a hidden, discrete-valued Markov process representing the system's *state* evolution. An extension to the HMM is the switching linear dynamic system (SLDS), which allows for more complicated dynamics generating the observations, but still follows the Markov state-switching of the HMM. For both the HMM and the SLDS, the state sequence's Markov structure accounts for the temporal persistence of certain regimes of operation.

Recently, the hierarchical Dirichlet process (HDP) [1] has been applied to the problem of learning hidden Markov models (HMM) with unknown state space cardinality, and is referred to as a HDP-HMM. A Dirichlet process is a distribution over random probability measures on infinite parameter spaces. This process provides a practical, data-driven prior towards models whose complexity grows as more data is observed. A specific hierarchical layering of these Dirichlet processes results in the HDP. When applied as a prior on the parameters of an HMM, the Dirichlet process encourages simple models of state dynamics, but allows additional states to be created as new behaviors are observed. The hierarchical structure allows for consistent learning of temporal state dependencies. In addition, the HDP has a number of properties that allow for computationally efficient learning algorithms, even on large datasets.

The original HDP-HMM addresses the statistical issue of coping with an unknown and potentially infinite state space, but allows for learning models with unrealistically rapid dynamics. For many ap-

plications, the state sequence's Markov structure is an approximation to a system with more complex temporal behavior, perhaps better approximated as semi-Markov with some non-exponentially distributed state duration. Setting a high probability of self-transition is a common approach to modeling states that persist over lengthy periods of time. One of the main limitations of the original HDP-HMM formulation is that it cannot be biased towards learning transition densities that favor such self-transitions. This results in a large sensitivity to noise, since the HDP-HMM can explain the data by rapidly switching among redundant states. Although the Dirichlet process induces a weak bias towards simple explanations employing fewer model components, when state-switching probabilities are unconstrained there can be significant posterior uncertainty in the underlying model.

Existing learning algorithms for HDP-HMMs are based on Markov chain Monte Carlo (MCMC) methods, such as Gibbs sampling, with an implementation that sequentially samples the state for each time step [1]. This sequential sampler leads to a slow mixing rate since global assignment changes are constrained to occur coordinate by coordinate, making it difficult to transition between different modes of the posterior. Existing HMM algorithms, such as the forward-backward algorithm [2], provide efficient methods for jointly sampling the entire state sequence conditioned on the observations and model parameters. While the original MCMC algorithm marginalized out the infinite set of infinite dimensional transition densities, we explore the use of truncated approximations to the Dirichlet process to make joint sampling tractable.

In this paper we revisit the HDP-HMM, and develop methods which allow more efficient and effective learning from realistic time series. In Sec. II, we begin by presenting some of the theoretical background of Dirichlet processes. Then, in Sec. III, we briefly describe the hierarchical Dirichlet process and, in Sec. IV, how it relates to learning HMMs. The revised formulation is described in Sec. V while Section V-C outlines the procedure for the blocked resampling of the state sequence. In Sec. **??**, we offer a model and inference algorithm for an HDP-HMM with non-standard emission distributions. We present results from simulated datasets in Sec. VII.

II. Dirichlet Processes

A Dirichlet process defines a distribution over probability measures on a parameter space Θ, which might be countably or uncountably infinite. This stochastic process is uniquely defined by a concentration parameter, α, and base measure, H, on the parameter space Θ; we denote it by $\text{DP}(\alpha, H)$. Consider a random probability measure $G \sim \text{DP}(\alpha, H)$. The Dirichlet process is formally defined by the property

that for any finite partition[1] $\{A_1, \ldots, A_K\}$ of the parameter space Θ,

$$(G(A_1), \ldots, G(A_K)) \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_K)). \tag{1}$$

That is, the measure of a random probability distribution $G \sim \text{DP}(\alpha, H)$ on every finite partition of the parameter space Θ follows a specific Dirichlet *distribution*. The Dirichlet process was first introduced by Ferguson [3] using Kolmogorov's consistency conditions. A more practically insightful definition of the Dirichlet process was given by Sethuraman [4]. Consider a probability mass function (pmf) $\{\pi_k\}_{k=1}^{\infty}$ on a countably infinite set, where the discrete probabilities are constructively defined as follows:

$$\begin{aligned} \beta'_k &\sim \text{Beta}(1, \alpha) \qquad k = 1, 2, \ldots \\ \pi_k &= \beta'_k \prod_{\ell=1}^{k-1} (1 - \beta'_\ell) \quad k = 1, 2, \ldots . \end{aligned} \tag{2}$$

In effect, we have divided a unit-length stick by the weights π_k. The k^{th} weight is a random proportion β'_k of the remaining stick after the previous $(k-1)$ weights have been defined. This *stick-breaking construction* is typically denoted by $\pi \sim \text{GEM}(\alpha)$. Sethuraman showed that with probability one, a random draw $G \sim DP(\alpha, H)$ can be expressed as

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta - \theta_k) \qquad \theta_k \sim H, \quad k = 1, 2, \ldots, \tag{3}$$

where the notation $\delta(\theta - \theta_k)$ indicates a Dirac delta at $\theta = \theta_k$.

From this definition, we see that the Dirichlet process actually defines a distribution over discrete probability measures. The stick-breaking construction also gives us insight into how the concentration parameter α controls the relative proportion of the mixture weights π_k, and thus determines the model complexity in terms of the expected number of components with significant probability mass.[2]

The Dirichlet process has a number of properties which make inference using this nonparametric prior computationally tractable. Because random probability measures drawn from a Dirichlet process are discrete, there is a strictly positive probability of multiple observations $\bar{\theta}_i \sim G$ taking identical values. For each observation $\bar{\theta}_i \sim G$, let z_i be an indicator random variable for the unique values θ_k such that $\bar{\theta}_i = \theta_{z_i}$. Blackwell and MacQueen [6] introduced a Pólya urn representation of the Dirichlet process,

[1]A partition of a set A is a set of disjoint, non-empty subsets of A such that every element of A is contained in exactly one of these subsets. More formally, $\{A_k\}_{k=1}^{K}$ is a partition of A if $\cup_k A_k = A$ and for each $j \neq k$, $A_k \cap A_j = \emptyset$.

[2]If the value of α is unknown, the model may be augmented with a gamma prior distribution on α, so that the parameter is learned from the data [5]. See Section V-D.

which can be equivalently described by the following predictive distribution on these indicator random variables:

$$p(z_{N+1} = z \mid z_{1:N}, \alpha) = \frac{\alpha}{\alpha + N}\delta(z, \tilde{k}) + \frac{1}{\alpha + N}\sum_{k=1}^{K} N_k \delta(z, k). \tag{4}$$

Here, N_k is the number of indicator random variables taking the value k, and $\tilde{k}$ is a previously unseen value. We use the notation $\delta(z, k)$ to indicate the Kronecker delta. This representation can be used to sample observations from a Dirichlet process without explicitly constructing the countably infinite random probability measure $G \sim \text{DP}(\alpha, H)$.

The predictive distribution of Eq. (4) is commonly referred to as the *Chinese restaurant process*. The analogy is as follows. Take $\bar{\theta}_i$ 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 α, the customer starts a new, previously unoccupied table $\tilde{k}$. From the Chinese restaurant process, we see that the Dirichlet process has a reinforcement property that leads to favoring simpler models.

We have shown that if $z_i \sim \pi$ and $\pi \sim \text{GEM}(\alpha)$, then we can integrate out π to determine the predictive likelihood of z_i. Another important distribution is that over the number K of unique values of z_i drawn from π given the total number of N draws. When π is distributed according to a stick-breaking construction with concentration parameter α, this distribution is given by [7]:

$$p(K \mid N, \alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} s(N, K) \alpha^K, \tag{5}$$

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

The Dirichlet process is most commonly used as a prior distribution on the parameters of a mixture model when the number of mixture components is unknown *a priori*. Such a model is called a *Dirichlet process mixture model* and is depicted by the graphs of Fig.1(a)-(b). The parameter with which an observation is associated implicitly partitions or clusters the data. In addition, the Chinese restaurant process representation indicates that the Dirichlet process 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 learning finite, representative mixture models. It can be shown under mild conditions that if the data are generated by a finite mixture, then the Dirichlet process posterior is guaranteed to converge (in distribution) to that finite set of mixture parameters [8].

We now describe how the Dirichlet process mixture model can be derived as the limit of a sequence

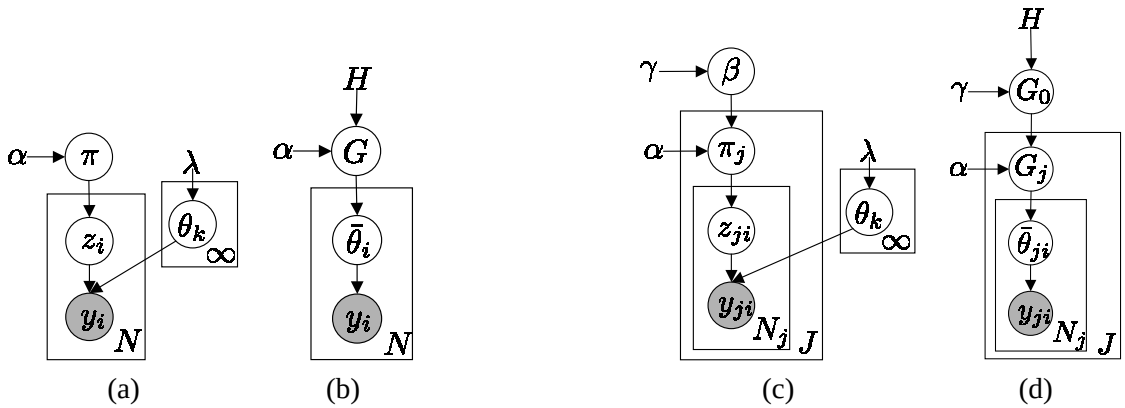


Fig. 1. Dirichlet process (left) and hierarchical Dirichlet process (right) mixture models represented by two graphs. (a) Indicator variable representation in which $\pi \sim \text{GEM}(\alpha)$, $\theta_k \sim H(\lambda)$, $z_i \sim \pi$, and $y_i \sim f(y \mid \theta_{z_i})$. (b) Alternative representation with $G \sim \text{DP}(\alpha, H)$, $\bar{\theta}_i \sim G$, and $y_i \sim f(y \mid \bar{\theta}_i)$. (c) Indicator variable representation in which $\beta \sim \text{GEM}(\gamma)$, $\pi_k \sim \text{DP}(\alpha, \beta)$, $\theta_k \sim H(\lambda)$, $z_{ji} \sim \pi_j$, and $y_{ji} \sim f(y \mid \theta_{z_{ji}})$. (d) Alternative representation with $G_0 \sim \text{DP}(\gamma, H)$, $G_j \sim \text{DP}(\alpha, G_0)$, $\bar{\theta}_{ji} \sim G_j$, and $y_{ji} \sim f(y \mid \bar{\theta}_{ji})$. Plate notation is used to compactly represent replicated variables of the graph [9].

of finite mixture models. Let us assume that there are L components to our finite mixture model and we place a Dirichlet *distribution* prior on these mixture weights. Our finite mixture model is described by:

$$\begin{aligned} \pi &\sim \text{Dir}(\alpha/L, \ldots, \alpha/L) & z_i &\sim \pi \\ \theta_k &\sim H(\lambda) & y_i &\sim F(\theta_{z_i}). \end{aligned} \tag{6}$$

Let $G^L(\theta) = \sum_{k=1}^{L} \pi_k \delta(\theta - \theta_k)$. Then, it can be shown that for every measurable function f integrable with respect to the measure H, this finite distribution G^L converges in distribution to a countably infinite distribution G distributed according to a Dirichlet *process* [10], [11]. That is,

$$\int_\theta f(\theta) dG^L(\theta) \xrightarrow{\mathcal{D}} \int_\theta f(\theta) dG(\theta), \tag{7}$$

as $L \to \infty$ for $G \sim DP(\alpha, H)$.

III. Hierarchical Dirichlet Processes

There are many scenarios in which groups of data are thought to be produced by related, yet unique, 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 disparate model from that of other data. In such scenarios we can take a hierarchical Bayesian approach and place a global Dirichlet process prior $DP(\alpha, G_0)$ on the parameter space Θ. We then draw group specific distributions $G_j \sim DP(\alpha, G_0)$, which will be discrete so that parameters are shared within the group. However, if the base measure G_0 is absolutely continuous with

respect to the Lebesgue measure, parameters will not be shared between groups. Only in the case where the base measure G_0 is discrete will there be a strictly positive probability of the group specific distributions having overlapping support (i.e. sharing parameters between groups.) To overcome this difficulty, the base measure G_0 should itself be a random measure distributed according to a Dirichlet process $DP(\gamma, H)$. This results in what is termed a *hierarchical Dirichlet process* (HDP) [1] and is depicted by the graphs of Fig. 1(c)-(d).

We now describe the HDP with a bit more formality. Let $(y_{j1}, \ldots, y_{jN_j})$ be the set of observations in group j. We assume there are J such groups of data. Then, the generative model can be written as:

$$\begin{aligned}
&G_0(\theta) = \textstyle\sum_{k=1}^{\infty} \beta_k \delta(\theta - \theta_k) && \beta \sim \text{GEM}(\gamma) \\
&&& \theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots \\
&G_j(\theta) = \textstyle\sum_{t=1}^{\infty} \tilde{\pi}_{jt} \delta(\theta - \tilde{\theta}_{jt}) && \tilde{\pi}_j \sim \text{GEM}(\alpha) \quad j = 1, \ldots, J \\
&&& \tilde{\theta}_{jt} \sim G_0 \qquad t = 1, 2, \ldots \\
&\bar{\theta}_{ji} \sim G_j && y_{ji} \sim F(\bar{\theta}_{ji}) \qquad j = 1, \ldots, J, \quad i = 1, \ldots, N_j.
\end{aligned} \tag{8}$$

See Fig. 1(d).

The Chinese restaurant process analogy of the Dirichlet process can be extended to a *Chinese restaurant franchise* for the HDP. The analogy is as follows. There are J restaurants, each with infinitely many tables. Each customer is pre-assigned to a given restaurant determined by its group j. Upon entering the j^{th} restaurant, a customer y_{ji} sits at a table $t_{ji} \sim \tilde{\pi}_j$. Each table then chooses a dish $\tilde{\theta}_{jt} \sim G_0$, or equivalently, an index for a dish $k_{jt} \sim \beta$. Therefore, customer y_{ji} eats dish $\bar{\theta}_{ji} = \tilde{\theta}_{jt_{ji}} = \theta_{k_{jt_{ji}}}$. The generative model is summarized below and is depicted in the graph of Fig. 2(a):

$$\begin{aligned}
&\beta \sim \text{GEM}(\gamma) && k_{jt} \sim \beta \\
&\tilde{\pi}_j \sim \text{GEM}(\alpha) && t_{ji} \sim \tilde{\pi}_j \\
&\theta_k \sim H(\lambda) && y_{ji} \sim F(\theta_{k_{jt_{ji}}}).
\end{aligned} \tag{9}$$

Let $\tilde{n}_{jt}$ be the number of *customers* in restaurant j sitting at table t and m_{jk} be the number of *tables* in restaurant j serving dish k. As with the Chinese restaurant process, the stick-breaking densities $\tilde{\pi}_j$

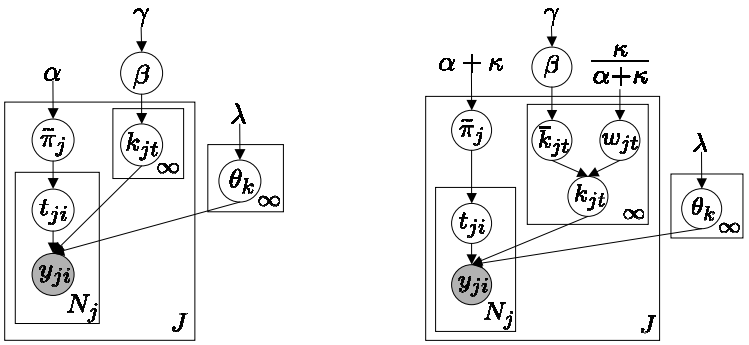


Fig. 2. Graph of the (a) Chinese restaurant franchise, and (b) tempered Chinese restaurant franchise. For the Chinese restaurant franchise, each customer (observation) y_{ji} is assigned to a table $t_{ji} \sim \tilde{\pi}_j$ in restaurant j, where $\tilde{\pi}_j \sim \text{GEM}(\alpha)$. Each table t then chooses a global dish index $k_{jt} \sim \beta$, where $\beta \sim \text{GEM}(\gamma)$. The likelihood of the observation is given by $y_{ji} \sim F(\theta_{k_{jt_{ji}}})$. For the tempered franchise, there is an *actual* restaurant serving dishes k_{jt}, which may have either arisen from the dish $\bar{k}_{jt}$ served in the *underlying* restaurant if $w_{jt} = 0$ or from having been overridden by dish j if $w_{jt} = 1$.

and β may be marginalized to yield the following predictive distributions:

$$p(t_{ji} \mid t_{j1}, \ldots, t_{ji-1}, \alpha) \propto \sum_{t=1}^{T_j} \tilde{n}_{jt}\delta(t_{ji}, t) + \alpha\delta(t_{ji}, \tilde{t}_j) \tag{10}$$

$$p(k_{jt} \mid \underline{k}_1, \underline{k}_2, \ldots, \underline{k}_{j-1}, k_{j1}, \ldots, k_{jt-1}, \gamma) \propto \sum_{k=1}^{K} m_{\cdot k}\delta(k_{jt}, k) + \gamma\delta(k_{jt}, \tilde{k}), \tag{11}$$

where $m_{\cdot k} = \sum_j m_{jk}$ and $\underline{k}_j = (k_{j1}, \ldots, k_{jT_j})$. Here, T_j is the number of currently occupied tables in restaurant j, and K is the total number of unique dishes being served in the franchise. The variables $\tilde{t}_j$ and $\tilde{k}$ represent choosing a currently uninstantiated table or dish, respectively.

Since each distribution G_j uses a discrete base measure G_0, multiple $\tilde{\theta}_{jt}$ may take an identical value θ_k for multiple unique values of t implying that multiple tables in the same restaurant may be serving the same dish, as depicted in Fig. 3. We can write G_j as a function of these unique dishes:

$$G_j(\theta) = \sum_{k=1}^{\infty} \pi_{jk}\delta(\theta - \theta_k), \quad \pi_j \sim \text{DP}(\alpha, \beta), \quad \theta_k \sim H, \tag{12}$$

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

$$\pi_{jk} = \sum_{t|k_{jt}=k} \tilde{\pi}_{jt}. \tag{13}$$

Let z_{ji} be the indicator random variable for the unique dish that observation y_{ji} eats. That is, $z_{ji} = k_{jt_{ji}}$.

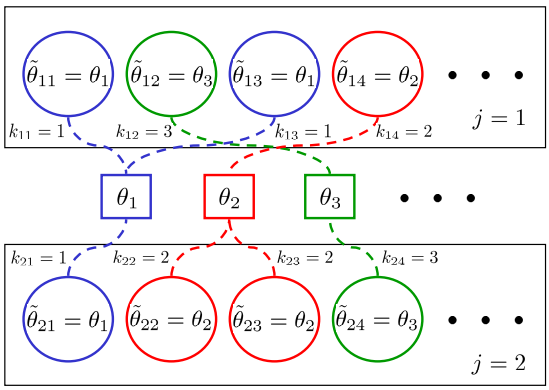


Fig. 3. Chinese restaurant franchise with $J = 2$ restaurants. The currently occupied tables each choose a dish $\tilde{\theta}_{jt} \sim G_j$, where $G_j \sim \text{DP}(\alpha, G_0)$ is a discrete probability measure so that multiple tables may serve the same dish. Since G_1 has overlapping support with G_2, parameters (i.e. dishes) are shared between restaurants.

A third equivalent representation of the generative model is in terms of these indicator random variables:

$$\begin{aligned} &\beta \sim \text{GEM}(\gamma) \\ &\pi_j \sim \text{DP}(\alpha, \beta) \quad z_{ji} \sim \pi_j \\ &\theta_k \sim H(\lambda) \qquad y_{ji} \sim F(\theta_{z_{ji}}), \end{aligned} \tag{14}$$

and is shown in Fig. 1(c).

As with the Dirichlet process, the HDP mixture model has an interpretation as the limit of a finite mixture model. In terms of the parameter indicator random variable representation, we write:

$$\begin{aligned} &\beta \sim \text{Dir}(\gamma/L, \ldots, \gamma/L) \\ &\pi_j \sim \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_L) \quad z_{ji} \sim \pi_j \\ &\theta_k \sim H \qquad\qquad\qquad\quad y_{ji} \sim F(\theta_{z_{ji}}). \end{aligned} \tag{15}$$

As $L \to \infty$, this model converges in distribution to that of the HDP mixture model [1].

IV. HDP-HMM

Hierarchical Dirichlet processes can be applied as a prior on the state values of a HMM with unknown state space cardinality, as described in [1]. Let us denote the state of the Markov chain at time t by z_t. Here, we have intentionally reused the notation z for this random variable for reasons that will become clear. Assume there are potentially countably infinitely many HMM state values. For each of these HMM states, there is a countably infinite transition density over the next HMM state. Let π_k be the transition density for HMM state k. Then, the Markov structure on the state sequence dictates that $z_t \sim \pi_{z_{t-1}}$. In terms of the previous HDP description, we see that z_{t-1} specifies the group with which y_t is associated.

Namely, all observations y_t with $z_{t-1} = j$ are assigned to group j since $z_t \sim \pi_j$. The current HMM state z_t then determines which of the global parameters θ_k are used to generate the observation y_t. The HDP-HMM is depicted by the graph in Fig. 4(a).

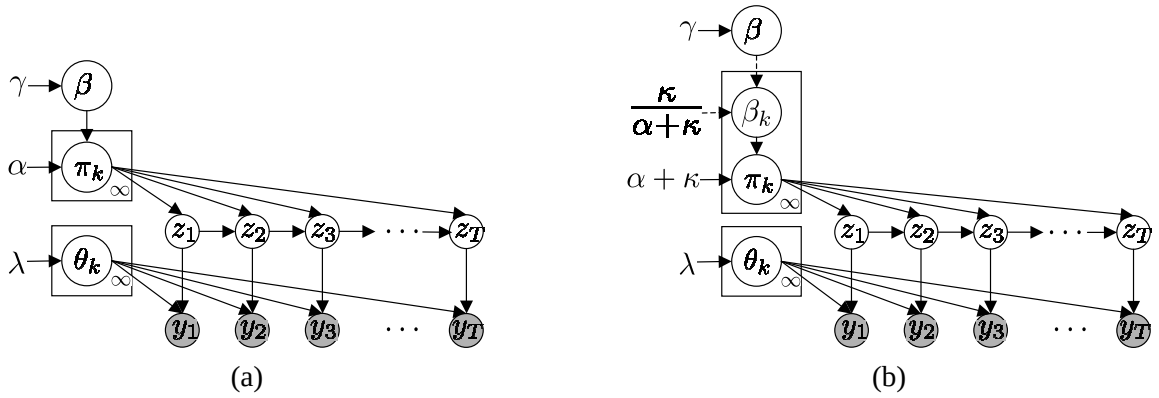


Fig. 4. Graph of the (a) HDP-HMM and (b) tempered HDP-HMM. The state z_t, taking values within a potentially countably infinite state space, indexes the transition density π_k from which the subsequent state z_{t+1} is drawn. That is, $z_{t+1} \sim \pi_{z_t}$. These transition densities have a hierarchical Dirichlet process prior. The HDP-HMM takes $\pi_k \sim \text{DP}(\alpha, \beta)$ with the global base measure defined as $\beta \sim \text{GEM}(\gamma)$. The tempered HDP-HMM instead employs $\pi_k \sim \text{DP}(\alpha + \kappa, \beta_k)$ with a state-specific base measure $\beta_k = (\alpha\beta + \kappa\delta_k)/(\alpha + \kappa)$, which is a deterministic function of the global base measure $\beta \sim \text{GEM}(\gamma)$ and the hyperparameters α and κ. The observation likelihood distributions are defined by the parameters θ_k so that $y_t \sim F(\theta_{z_t})$.

This model can also be described in terms of the Chinese restaurant franchise. 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 of indices $f : t \to ji$. The parent then chooses a table $t_{ji} \sim \tilde{\pi}_j$ and that table is served a dish indexed by $k_{jt} \sim \beta$. The index of the dish the parent is eating, $k_{jt_{ji}} = z_{ji} = z_t$, determines the parameter of the parent's likelihood distribution, θ_{z_t}, as well as the restaurant (or group) of the child z_{t+1}. This analogy is not very intuitive or useful for the basic HDP-HMM, but will be important in developing the tempered HDP-HMM.

A. Inference for HDP-HMM

In this section we describe one of the three Markov chain Monte Carlo (MCMC) HDP sampling algorithms presented in [1]. Specifically, we consider the direct assignment Rao-Blackwellized Gibbs sampler, which is cited as the best-suited to the HDP-HMM application. In the Chinese restaurant franchise, an observation y_{ji} is assigned to a table t_{ji}, and each table is then assigned a dish k_{jt} so that y_{ji} is indirectly associated with parameter $\theta_{k_{jt_{ji}}}$. The direct assignment sampler circumvents this complicated bookkeeping by directly associating an observation y_{ji} with a parameter via the indicator

random variable z_{ji}, where in the HDP-HMM we have $z_t = z_{ji}$. In this sampler, a set of auxiliary variables m_{jk} must be added, as will be discussed subsequently.

Throughout the remainder of the paper, we will use the following notational conventions. Given a random sequence $\{x_1, x_2, \ldots, x_T\}$, we use the shorthand $x_{1:t}$ to be the sequence $\{x_1, x_2, \ldots, x_t\}$ and $x_{\backslash t}$ to be 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, $\{x_{a_1 a_2}, \forall a_1, a_2\}$.

To derive the direct assignment sampler, we first assume that we sample: table assignments for each observation, t_{ji}; dish assignments for each of these tables, k_{jt}; and the global mixture weights, β. Because of the properties of the HDP, and more specifically the stick-breaking densities, we are able to marginalize the group-specific densities $\tilde{\pi}_j$ 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. (11).) The marginalization of these variables is referred to as *Rao-Blackwellization* [12]. 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 then proceed to show that directly sampling z_{ji} instead of t_{ji} and k_{jt} is sufficient when a set of auxiliary variables is additionally sampled.

1) Sampling β: We begin by examining the posterior distribution of β. Recall the HDP mixture model defined in Eq. (8). At any given iteration of the sampler, let us assume that there are K unique dishes being served and let us consider the finite partition $\{\theta_1, \theta_2, \ldots, \theta_K, \theta_{\tilde{k}}\}$ of the parameter space Θ, where $\theta_{\tilde{k}} = \Theta \backslash \bigcup_{k=1}^{K} \{\theta_k\}$ is the set of all currently unrepresented parameters. By definition of the Dirichlet process, G_0 has the following distribution on this finite partition:

$$\begin{aligned}(G_0(\theta_1), \ldots, G_0(\theta_K), G_0(\theta_{\tilde{k}})) &\sim \text{Dir}(\gamma H(\theta_1), \ldots, \gamma H(\theta_K), \gamma H(\theta_{\tilde{k}})) \\ &\sim \text{Dir}(0, \ldots, 0, \gamma), \end{aligned} \tag{16}$$

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

For every currently instantiated table t, k_{jt} associates the table-specific dish $\tilde{\theta}_{jt}$ with one among the unique set of dishes $\{\theta_1, \ldots, \theta_K\}$. We have used m_{jk} to denote how many of the table-specific dishes in restaurant j are dish θ_k. Therefore, we have $m_{\cdot k}$ observations $\tilde{\theta}_{jt} \sim G_0$ in the franchise that fall within the single-element partition $\{\theta_k\}$. By the properties of the Dirichlet distribution we have,

$$p((G_0(\theta_1), \ldots, G_0(\theta_K), G_0(\theta_{\tilde{k}})) | \{\tilde{\theta}_{jt}\}, \gamma) \propto \text{Dir}(m_{\cdot 1}, \ldots, m_{\cdot K}, \gamma). \tag{17}$$

Since $(G_0(\theta_1), \ldots, G_0(\theta_K), G_0(\theta_{\tilde{k}}))$ are by definition equal to $(\beta_1, \ldots, \beta_K, \beta_{\tilde{k}})$ and from the conditional

independencies illustrated in Fig. 2, the desired posterior distribution of β is

$$p((\beta_1, \ldots, \beta_K, \beta_{\tilde{k}}) \mid \boldsymbol{t}, \boldsymbol{k}, y_{1:T}, \gamma) \quad \propto \quad \text{Dir}(m_{\cdot 1}, \ldots, m_{\cdot K}, \gamma). \tag{18}$$

From the above, we see that $\{m_{\cdot k}\}_{k=1}^{K}$ is a set of sufficient statistics for re-sampling β defined on this partition.

2) Sampling z_t: We now determine the posterior distribution of z_t:

$$p(z_t = k \mid z_{\setminus t}, y_{1:T}, \beta, \alpha, \lambda) \propto p(z_t = k \mid z_{\setminus t}, \beta, \alpha) p(y_t \mid y_{\setminus t}, z_t = k, z_{\setminus t}, \lambda) \tag{19}$$

The properties of the Dirichlet process dictate that on the finite partition $\{1, \ldots, K, \tilde{k}\}$ we have the following form for the group-specific transition densities:

$$p(\pi_j \mid \alpha, \beta) \quad \propto \quad \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_K, \alpha\beta_{\tilde{k}}). \tag{20}$$

We use the above definition of π_j and the Dirichlet distribution's conjugacy to the multinomial observations z_t to marginalize π_j and derive the following conditional distribution over the states assignments:

$$p(z_t = k \mid z_{\setminus t}, \beta, \alpha) \propto \begin{cases} (\alpha\beta_k + n^{-t}_{z_{t-1}k})(\frac{\alpha\beta_{z_{t+1}} + n^{-t}_{kz_{t+1}} + \delta(z_{t-1},k)\delta(k,z_{t+1})}{\alpha + n^{-t}_{k\cdot} + \delta(z_{t-1},k)}) & k \in 1, \ldots, K \\ \alpha\beta_{\tilde{k}}\beta_{z_{t+1}} & k = \tilde{k}. \end{cases} \tag{21}$$

For a detailed derivation, see Appendix I-A. The notation n_{jk} represents the number of Markov chain transitions from state j to k, which can be computed from $z_{1:T}$. Furthermore, we use $n_{j\cdot}$ to indicate the number of transitions from j to any other state (i.e. $n_{j\cdot} = \sum_k n_{jk}$) and n^{-t}_{jk} 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}. Let $z_{t-1} = j$ and $z_{t+1} = \ell$. The intuition behind this distribution is that we choose a state k with prior probability as a function of how many times we have seen other j to k and k to ℓ transitions. Note that there is a minor dependency on whether either or both of these transitions correspond to a self-transition (i.e. $k = j$ or $k = \ell$.)

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:

$$p(y_t \mid y_{\setminus t}, z_t = k, z_{\setminus t}, \lambda) \quad \propto \quad \int_{\theta_k} p(y_t \mid \theta_k) p(\theta_k \mid \{y_\tau \mid z_\tau = k, \tau \neq t\}, \lambda) d\theta_k. \tag{22}$$

Note that the set $\{y_\tau \mid z_\tau = k, \tau \neq t\}$ denotes all the observations y_τ other than y_t that were drawn from the observation likelihood distribution parameterized by θ_k. By placing a conjugate prior on the

parameter space, there is a closed form distribution for this marginal likelihood. Further details may be found in Appendix I-B.

From the above conditional distributions for β and z_t, we see that the only effect of t_{ji} and k_{jt} is via m_{jk}, the number of tables serving dish k in restaurant j. Thus, it is sufficient to sample m_{jk} instead of t_{ji} and k_{jt}, when given the state index z_t.

3) Sampling m_{jk}: Having the state index assignments $z_{1:T}$ effectively partitions the data (customers) into both restaurants and dishes, though the table assignments are unknown. For example, $z_{t-1} = j$ and $z_t = k$ tells us that customer y_t is in restaurant j and eating dish k, though there may be multiple tables serving this dish so that the customer's table cannot be disambiguated. 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:

$$\begin{aligned} p(t_{ji} = t \mid k_{jt} = k, \boldsymbol{t}^{-ji}, \boldsymbol{k}^{-jt}, y_{1:T}, \beta, \alpha) &\propto p(t_{ji} \mid t_{j1}, \ldots, t_{ji-1}, t_{ji+1}, \ldots, t_{jT_j}, \alpha) p(k_{jt} = k \mid \beta) \\ &\propto \begin{cases} \tilde{n}_{jt}^{-ji}, & t \in \{1, \ldots, T_j\}; \\ \alpha\beta_k, & t = \tilde{t}_j. \end{cases} \end{aligned} \tag{23}$$

Here, $\tilde{n}_{jt}^{-ji}$ is the number of customers sitting at table t in restaurant j, not counting customer 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. The form of this distribution implies that a customer's table assignment conditioned on a dish assignment k follows a Dirichlet process with concentration parameter $\alpha\beta_k$. That is,

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

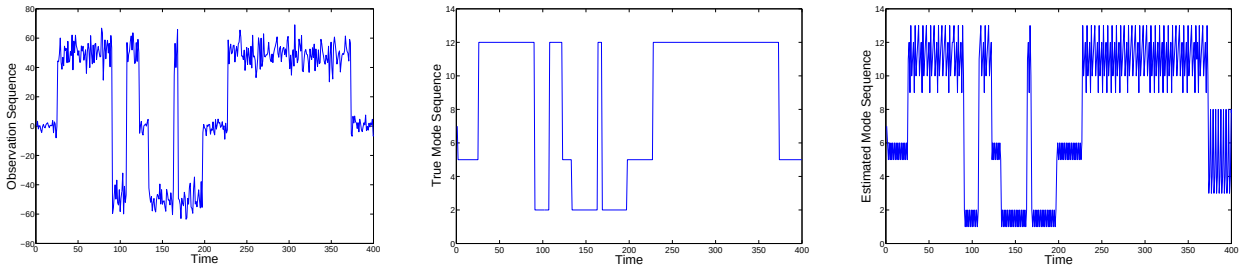
Then, Eq. (5) provides the form for the density over the number of unique components (i.e. tables) generated by sampling n_{jk} times from this conditional stick-breaking density:

$$p(m_{jk} = m \mid n_{jk}, \beta, \alpha) = \frac{\Gamma(\alpha\beta_k)}{\Gamma(\alpha\beta_k + n_{jk})} s(n_{jk}, m) (\alpha\beta_k)^m. \tag{24}$$

In terms of the Chinese restaurant franchise, the number of transitions from state j to k, n_{jk}, is the total number of customers in restaurant j eating dish k (i.e. $n_{jk} = \sum_{t|k_{jt}=k} \tilde{n}_{jt}$.) 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. (23), rather than having to compute a large array of Stirling numbers. See Algorithm 1 for an outline of the HDP-HMM direct assignment Gibbs sampler.

V. TEMPERED HDP-HMM

For the sake of argument, let us assume that we have continuous observations of our discrete state space, such as generated by a Gaussian likelihood distribution. Let us also assume that the state of the HMM typically persists over a period of time. In such scenarios, the unconstrained nature of the HDP-HMM transition probabilities obscures the learning procedure and results in a sensitivity to the within-state variations in the observations. This sensitivity is especially pronounced when the degree of the state-specific variation is an unknown parameter of the model. For example, with Gaussian likelihoods parameterized by unknown means and covariances, the sampler may divide the observations generated from a single state into two states with slightly different expected means, each with small expected covariances, and then quickly switch between these two states. The HDP-HMM reinforces this assignment since the predictive distribution of state transitions dictates that if the system is in one state at a given time, it is likely to be in the other state at the next time step based on having already seen many of these transitions. This scenario is depicted in Fig. 5.



(a) (b) (c)

Fig. 5. Qualitative plots showing the sensitivity of the original HDP-HMM direct assignment sampler to variations in the observations. (a) Observation sequence; (b) true HMM state sequence; and (c) estimated HMM state sequence after 100 iterations of the Gibbs sampler. In plot (c), we see that individual true states are divided into multiple estimated states, each with high probability of switching to one of the others.

Although the MCMC sampler is guaranteed to converge to the true posterior distribution, many fast state-switching sequences have large posterior probability in the standard HDP-HMM, thus slowing the rate at which the sampler explores the entire sequence space. The true state sequence might have only marginally larger posterior probability than these other explanations of the observations. Furthermore, when observations are high-dimensional, this fragmentation of data into redundant states may reduce predictive performance. For scenarios where the HMM is actually approximating a semi-Markov process, one would like to be able to incorporate the fact that slow state-switching is preferable to fast state-switching. That is, the probability of a self-transition should be biased towards larger values. To this end,

we modify the standard HDP-HMM as follows:

$$\begin{aligned} \beta &\sim \text{GEM}(\gamma) \\ \pi_j &\sim \text{DP}\left(\alpha+\kappa, \frac{\alpha\beta+\kappa\delta_j}{\alpha+\kappa}\right), \end{aligned} \tag{25}$$

where $(\alpha\beta + \kappa\delta_j)$ indicates that an amount κ is added to the j^{th} component of $\alpha\beta$. Now, each state-specific transition density has a unique base measure with an additional weight, determined by κ, on a transition to that given state. See Fig. 4(b).

The concept behind this κ parameter is reminiscent of the self-transition bias parameter in the infinite HMM [13]. 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. In [13], only a heuristic approximation to a Gibbs sampling algorithm was presented for inference in this model. The full connection between the infinite HMM and the HDP formulation, as well as developing a globally consistent inference algorithm, was made in [1]. However, in the HDP-HMM formulation of [1], there was no mention of the self-transition bias parameter.

To better understand the form of Eq. (25), it is useful to return to the formal definition of the Dirichlet process. Consider a finite partition $(Z_1, \ldots, Z_K)$ of the positive integers $\mathbb{Z}_+$. Then

$$(\pi_j(Z_1), \ldots, \pi_j(Z_K)) \sim \text{Dir}(\alpha\beta(Z_1) + \kappa\delta_j(Z_1), \ldots, \alpha\beta(Z_K) + \kappa\delta_j(Z_K)) \tag{26}$$

so that κ is only added to the Dirichlet parameter of the arbitrarily small partition containing j, which corresponds to a self-transition.

We will refer to this model as the *tempered* HDP-HMM.

A. Chinese Restaurant Franchises with Loyal Customers

We further abuse the analogy of the Chinese restaurants by extending it to the tempered HDP-HMM, where we now have a franchise of restaurants that each have a loyal set of customers. Each restaurant in the franchise has a specialty dish with the same index as that of the restaurant. Although this dish is

also served in other restaurants, it is more popular in the dish's namesake restaurant. We can see this increased popularity in the specialty dish from the fact that

$$k_{jt} \sim \frac{\alpha\beta + \kappa\delta_j}{\alpha + \kappa}. \tag{27}$$

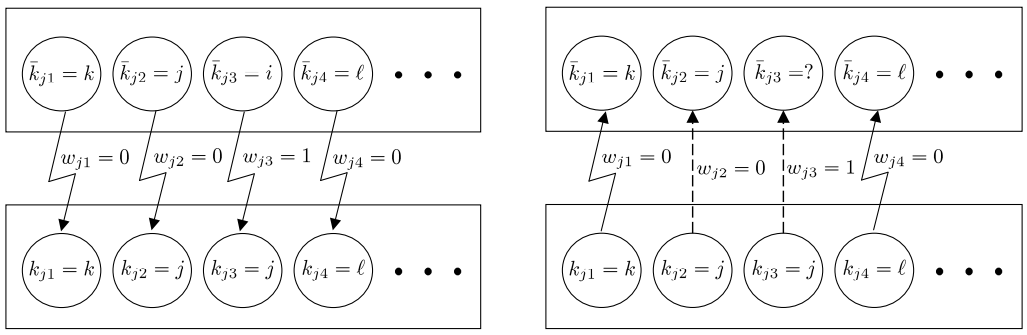
Noting that $z_t = z_{ji} = k_{jt_{ji}}$ and $z_{t+1} \sim \pi_{z_t}$, we see that children are more likely to eat in the same restaurant as their parent and, in turn, more likely to eat the restaurant's specialty dish. This develops family loyalty to a given restaurant in the franchise. However, if the parent chooses a dish that is not 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 tempered HDP-HMM, the 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, which is derived in Sec. V-B, is simplified if we introduce a set of auxiliary random variables $\bar{k}_{jt}$ and w_{jt} as follows:

$$\begin{aligned} \bar{k}_{jt} &\sim \beta \\ w_{jt} &\sim \text{Ber}\left(\frac{\kappa}{\alpha + \kappa}\right) \\ k_{jt} &= \begin{cases} \bar{k}_{jt}, & w_{jt} = 0; \\ j, & w_{jt} = 1, \end{cases} \end{aligned} \tag{28}$$

where $\text{Ber}(p)$ represents the Bernoulli distribution with p the probability of success. We will describe this formulation in terms of an *underlying* and an *actual* restaurant. The *underlying* restaurant corresponds to the process of choosing a dish without taking the restaurant's specialty into consideration (i.e. the original Chinese restaurant franchise.) With some probability, the considered decision to order a given dish is overridden (perhaps by a waiter's suggestion) and the table is served the specialty dish. The dishes the waiters actually serve the tables correspond to the *actual* restaurant. This generative process is depicted in Fig. 6(a). We refer to $\bar{k}_{jt}$ as the *considered* dish index and w_{jt} as the *override* variable. Our inference algorithm, described in Sec. V-B, will aim 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. 6(b).

This formulation is equivalent to the original formulation, which can be clearly seen if we rewrite the



(a) Generative (b) Inference

Fig. 6. (a) Generative model of the dish indices $\bar{k}_{jt}$ of underlying restaurant (top) being converted to indices k_{jt} in the actual restaurant (bottom) via override variables w_{jt}. (b) Inference perspective of trying to infer $\bar{k}_{jt}$ and w_{jt} given k_{jt}. If $k_{jt} \neq j$, then the override variable w_{jt} is automatically 0 and the underlying restaurant serves dish $\bar{k}_{jt} = k_{jt}$, as indicated by the jagged arrow. If the actual restaurant j serves dish $k_{jt} = j$ 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 the considered dish is also $\bar{k}_{jt} = k_{jt} = j$. However, if the considered dish was overridden, then that dish $\bar{k}_{jt}$ could have taken any value, as denoted by the question mark.

base measure as:

$$k_{jt} \sim \sum_{k=1}^{K} \frac{\alpha}{\alpha+\kappa}\beta_k \delta(k_{jt}, k) + \frac{\alpha}{\alpha+\kappa}\beta_{\tilde{k}}\delta(k_{jt}, \tilde{k}) + \frac{\kappa}{\alpha+\kappa}\delta(k_{jt}, j). \tag{29}$$

The graphical model of the Chinese restaurant franchise for the tempered HDP-HMM is shown in Fig. 2(b). Although not explicitly present in this graph, the tempered HDP-HMM still has a Markov structure on the indicator random variables z_t, which based on the value of their parent z_{t-1} are mapped to a group-specific index ji. As with the HDP-HMM, during the MCMC inference procedure the assignments of observations to groups is dynamically changing with the sampled value of the parent indicator random variable.

B. Direct Assignment Method for the Tempered HDP-HMM

In this section, we derive the tempered HDP-HMM direct assignment Gibbs sampler. Throughout this section, we will refer to the random variables in the graph of Fig. 2(b). As before, let us begin by assuming that we sample all the assignment random variables of the Chinese restaurant franchise. In the tempered HDP-HMM, this now includes: table assignments for each customer, t_{ji}; served dish assignments for each table, k_{jt}; considered dish assignments, $\bar{k}_{jt}$; and dish override variables, w_{jt}. We still sample the global weights, β, as well. We then show, as we did for the HDP-HMM, that we can rely solely on sampling the state variables z_t instead of assignment variables t_{ji}, k_{jt} and $\bar{k}_{jt}$ if we add auxiliary variables to our sampler.

1) Sampling β: Previously, we derived that the number of tables m_{jk} was a set of sufficient statistics for sampling β. Now, since $\bar{k}_{jt}$ is drawn from β, we have

$$\begin{aligned} p(\beta \mid \boldsymbol{t}, \boldsymbol{k}, \bar{\boldsymbol{k}}, \boldsymbol{w}, y_{1:T}, \gamma) &\propto p(\beta \mid \gamma) \prod p(\bar{k}_{jt} \mid \beta) \\ &\propto \text{Dir}(\bar{m}_{\cdot 1}, \bar{m}_{\cdot 2}, \ldots, \bar{m}_{\cdot K}, \gamma), \end{aligned} \tag{30}$$

where $\bar{m}_{jk}$ represents the number of tables that considered ordering a dish k.

2) Jointly Sampling m_{jk}, w_{jt}, and $\bar{m}_{jk}$: The auxiliary variables m_{jk}, w_{jt}, and $\bar{m}_{jk}$ can be jointly sampled given the state sequence $z_{1:T}$ and global density β. The joint conditional density can be decomposed as follows:

$$p(\boldsymbol{m}, \boldsymbol{w}, \bar{\boldsymbol{m}} \mid z_{1:T}, \beta, \alpha, \kappa) = p(\bar{\boldsymbol{m}} \mid \boldsymbol{m}, \boldsymbol{w}, z_{1:T}, \beta, \alpha, \kappa) p(\boldsymbol{w} \mid \boldsymbol{m}, z_{1:T}, \beta, \alpha, \kappa) p(\boldsymbol{m} \mid z_{1:T}, \beta, \alpha, \kappa) \tag{31}$$

We start by examining $p(\boldsymbol{m} \mid z_{1:T}, \beta, \alpha, \kappa)$, where m_{jk} is the number of tables with *served* dish k. This distribution is derived as in the original HDP-HMM by using Eq. (5). However, we now have concentration parameter $\alpha + \kappa$ and base measure $(\alpha\beta + \kappa\delta_j)/(\alpha + \kappa)$ so that

$$p(m_{jk} = m \mid n_{jk}, \beta, \alpha, \kappa) = \frac{\Gamma(\alpha\beta_k + \kappa\delta(j,k))}{\Gamma(\alpha\beta_k + \kappa\delta(j,k) + n_{jk})} s(n_{jk}, m) (\alpha\beta_k + \kappa\delta(j,k))^m. \tag{32}$$

Note that this distribution only differs from that of Eq. (24) when $j = k$.

We now derive the conditional distribution $p(\boldsymbol{w} \mid \boldsymbol{m}, z_{1:T}, \beta)$ over the override variables w_{jt}. The table counts m_{jk} inform us that for each table $t \in \mathcal{T}_{jk}$, where $|\mathcal{T}_{jk}| = m_{jk}$, the dish assignment is $k_{jt} = k$. Thus, we can equivalently examine the distribution $p(w_{jt} \mid k_{jt}, \beta)$ over each override variable independently since

$$p(\boldsymbol{w} \mid \boldsymbol{m}, z_{1:T}, \beta, \alpha, \kappa) = p(\boldsymbol{w} \mid \boldsymbol{k}, \beta, \alpha, \kappa) = \prod_{j,t} p(w_{jt} \mid k_{jt}, \beta, \alpha, \kappa). \tag{33}$$

Note that we only need to consider the tables t with served dish $k_{jt} = j$, corresponding to that restaurant's specialty, since these are the tables where the considered dish $\bar{k}_{jt}$ may have been overridden via $w_{jt} = 1$. For all other tables, we can automatically deduce that $w_{jt} = 0$.

For the tables with $k_{jt} = j$, 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.

If we define $\rho = \frac{\kappa}{\alpha+\kappa}$ to be the prior probability that $w_{jt} = 1$, then

$$\begin{aligned} p(w_{jt} \mid k_{jt} = j, \beta, \rho) &= \sum_{\bar{k}_{jt}=1}^{K} p(\bar{k}_{jt}, w_{jt} \mid k_{jt} = j, \beta) + p(\bar{k}_{jt} = \tilde{k}, w_{jt} \mid k_{jt} = j, \beta) \\ &\propto \sum_{\bar{k}_{jt}=1}^{K} p(k_{jt} = j \mid \bar{k}_{jt}, w_{jt}) p(\bar{k}_{jt} \mid \beta) p(w_{jt} \mid \rho) \\ &\quad + p(k_{jt} = j \mid \bar{k}_{jt} = \tilde{k}, w_{jt}) p(\bar{k}_{jt} = \tilde{k} \mid \beta) p(w_{jt} \mid \rho) \\ &\propto \begin{cases} \beta_j(1-\rho), & w_{jt} = 0; \\ \rho, & w_{jt} = 1. \end{cases} \end{aligned} \tag{34}$$

The above 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.

Let $\mathcal{T}_{jj} = \{t | k_{jt} = j\}$. For each table $t \in \mathcal{T}_{jj}$, that is, each table served dish j in the actual restaurant j, we independently draw a sample of w_{jt} from the above distribution. Thus, in total we draw m_{jj} i.i.d. samples of w_{jt}, 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 the *underlying* restaurant j. Any table that was overridden is an uninformative observation for the posterior of $\bar{m}_{jk}$ so that

$$\bar{m}_{jk} = \begin{cases} m_{jk}, & j \neq k; \\ m_{jj} - w_{j\cdot}, & j = k. \end{cases} \tag{35}$$

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.

3) Sampling z_t: Finally, we need the predictive distribution of z_t for the tempered HDP-HMM. We first note that the prior distribution of π_j is now

$$\pi_j \sim \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_j + \kappa, \ldots, \alpha\beta_{\tilde{k}}). \tag{36}$$

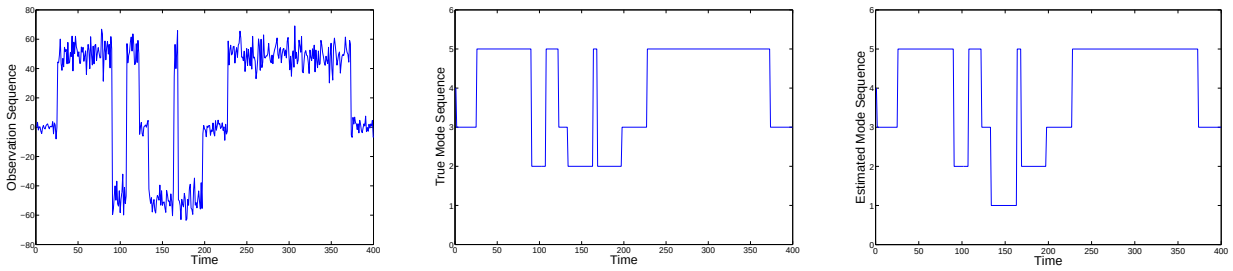
Using this tempered group-specific transition density, one can re-derive the predictive distribution of z_t to be:

$$p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) \propto \begin{cases} (\alpha\beta_k + n^{-t}_{z_{t-1}k} + \kappa\delta(z_{t-1}, k)) \left(\frac{\alpha\beta_{z_{t+1}} + n^{-t}_{kz_{t+1}} + \kappa\delta(k, z_{t+1}) + \delta(z_{t-1}, k)\delta(k, z_{t+1})}{\alpha + n^{-t}_{k\cdot} + \kappa + \delta(z_{t-1}, k)} \right) & k \in 1, \ldots, K \\ \frac{\alpha^2 \beta_{\tilde{k}} \beta_{z_{t+1}}}{\alpha + \kappa} & k = \tilde{k} \end{cases} \quad (37)$$

See Appendix I-A for a complete derivation. The resulting tempered HDP-HMM direct assignment Gibbs sampler is outlined in Algorithm 1.

C. Exploiting the HMM Structure

The tempered HDP-HMM reduces the posterior uncertainty caused by fast state-switching explanations of the data; however, the bias towards self-transitions introduces a mixing rate problem for the MCMC sampler. Specifically, two continuous periods of observations of a given state that are separated in time may be individually grouped into separate states (see Fig. 7.) If this occurs, the high probability of self-transition within each state makes it challenging for the sequential sampler to group those two examples into a single state. In this section, we consider a method of leveraging the Markov structure of the HDP-HMM to mitigate this problem.



(a) (b) (c)

Fig. 7. Qualitative plots showing the sequential Gibbs sampler splitting two temporally separated examples of the same true state into two states. (a) Observation sequence; (b) true HMM state sequence; (c) estimated HMM state sequence for a given iteration of the Gibbs sampler. In plot (c), we see that a single true state was divided into two estimated states, each with high probability of self-transition.

A variant of the HMM forward-backward procedure [2] allows us to jointly sample the state sequence $z_{1:T}$ given the observation sequence $y_{1:T}$, transitions densities π_j, and model parameters θ_k. With the sequential sampler, we were not exploiting the simple Markov structure of the graphical model. Note,

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

1) Set $z_{1:T} = z_{1:T}^{(n-1)}$ and $\beta = \beta^{(n-1)}$. For each $t \in \{1, \ldots, T\}$, 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 $z_t = k$:
   $$(\hat{\mu}_k, \hat{\Sigma}_k) \leftarrow (\hat{\mu}_k, \hat{\Sigma}_k) \ominus y_t$$

   b) For each of the K currently instantiated states, determine the predictive likelihood
   $$f_k(y_t) = (\alpha\beta_k + n_{z_{t-1}k}) \left(\frac{\alpha\beta_{z_{t+1}} + n_{kz_{t+1}} + \kappa\delta(k, z_{t+1})}{\alpha + n_{k\cdot} + \kappa} \right) \mathcal{N}(y_t; \hat{\mu}_k, \hat{\Sigma}_k)$$
   for $z_{t-1} \neq k$, otherwise see Eq. (37). Also determine likelihood $f_{\tilde{k}}(y_t)$ of a new state $\tilde{k}$.

   c) Sample the new state assignment z_t:
   $$z_t \sim \sum_{k=1}^{K} f_k(y_t)\delta(z_t, k) + f_{\tilde{k}}(y_t)\delta(z_t, \tilde{k})$$
   If $z_t = \tilde{k}$, 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 $\beta_{\tilde{k}} \leftarrow (1-b)\beta_{\tilde{k}}$.

   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 $z_t = k$:
   $$(\hat{\mu}_k, \hat{\Sigma}_k) \leftarrow (\hat{\mu}_k, \hat{\Sigma}_k) \oplus y_t$$

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 m, w, and $\bar{m}$ as follows:

   a) For each $(j, k) \in \{1, \ldots, K\}^2$, define $\mathcal{J}_{jk} = \{\tau \mid z_{\tau-1} = j, z_\tau = k\}$. Set $m_{jk} = 0$ and $n = 0$ and for each $\tau \in \mathcal{J}_{jk}$, 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 $j \in \{1, \ldots, K\}$, 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
   $$\beta^{(n)} \sim \text{Dir}(\bar{m}_{\cdot 1}, \ldots, \bar{m}_{\cdot K}, \gamma)$$

Algorithm 1: Direct assignment Rao–Blackwellized Gibbs sampler for the tempered 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 Appendix I-B). The $\oplus$ and $\ominus$ operators update cached mean and covariance statistics as assignments are added or removed from a given component. Hyperparameters may be resampled, according to the formulas in Appendix V-D, as a final step.

however, that in order to take advantage of this procedure, we must sample the transition densities and model parameters, which were previously integrated out.

To sample the transition densities in practice, we must somehow approximate these theoretically countably infinite 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 Dirichlet process,

$$\text{GEM}_L(\alpha) \triangleq \text{Dir}(\alpha/L, \ldots, \alpha/L), \tag{38}$$

where L is a number that exceeds the total number of expected HMM states. Note that both of these approximations, which are presented and compared in [10], [11], encourage learning models with fewer than L components while allowing the generation of new components, upper bounded by L, as new data is observed. We choose to use the second approximation because of its simplicity and computational efficiency.

The weak limit approximation to the Dirichlet process gives us the following form for the prior distribution on the global weights β:

$$\beta \sim \text{Dir}(\gamma/L, \ldots, \gamma/L). \tag{39}$$

On this partition, the prior distribution over the transition densities are Dirichlet with parametrization:

$$\pi_j \sim \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_j + \kappa, \ldots, \alpha\beta_L). \tag{40}$$

The posterior distributions are then given by:

$$\begin{aligned} \beta &\sim \text{Dir}(\gamma/L + \bar{m}_{\cdot 1}, \ldots, \gamma/L + \bar{m}_{\cdot L}) \\ \pi_j &\sim \text{Dir}(\alpha\beta_1 + n_{j1}, \ldots, \alpha\beta_j + \kappa + n_{jj}, \ldots, \alpha\beta_L + n_{jL}). \end{aligned} \tag{41}$$

Depending on the form of the observation likelihood distribution and prior distribution on the parameters θ_k of this likelihood, we sample our model parameters, one for each currently instantiated state, from the updated posterior distribution:

$$\theta_j \sim p(\theta \mid \{y_t \mid z_t = j\}, \lambda) \tag{42}$$

Now that we are sampling θ_j directly rather than marginalizing these parameters as in the direct assign-

ment sampler, we can use a non-conjugate base measure on the parameter space Θ (see Appendix II.)

To derive the forward-backward procedure for jointly sampling $z_{1:T}$ given $y_{1:T}$, we first note that the chain rule and Markov structure allows us to decompose the joint distribution as follows:

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

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:

$$\begin{aligned} p(z_1 \mid y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) &\propto p(z_1) p(y_1 \mid \theta_{z_1}) \sum_{z_{2:T}} \prod_t p(z_t \mid \pi_{z_{t-1}}) p(y_t \mid \theta_{z_t}) \\ &\propto p(z_1) p(y_1 \mid \theta_{z_1}) \sum_{z_2} p(z_2 \mid \pi_{z_1}) p(y_2 \mid \theta_{z_2}) m_{3,2}(z_2) \\ &\propto p(z_1) p(y_1 \mid \theta_{z_1}) m_{2,1}(z_1), \end{aligned} \tag{43}$$

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 given by:

$$\begin{aligned} m_{t,t-1}(z_{t-1}) &\propto \begin{cases} \sum_{z_t} p(z_t \mid \pi_{z_{t-1}}) p(y_t \mid \theta_{z_t}) m_{t+1,t}(z_t), & t \leq T; \\ 1, & t = T+1; \end{cases} \\ &\propto p(y_{t:T} \mid z_{t-1}, \boldsymbol{\pi}, \boldsymbol{\theta}). \end{aligned} \tag{44}$$

The general conditional distribution of z_t is:

$$p(z_t \mid z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto p(z_t \mid \pi_{z_{t-1}}) p(y_t \mid \theta_{z_t}) m_{t+1,t}(z_t). \tag{45}$$

If the k^{th} state-specific observation likelihood distribution is Gaussian with mean μ_k and covariance Σ_k, then the above distributions are given by:

$$p(z_t = i \mid z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\theta}) \propto \pi_{z_{t-1}}(i) \mathcal{N}(y_t; \mu_i, \Sigma_i) m_{t+1,t}(i) \tag{46}$$

$$m_{t+1,t}(i) = \sum_{j=1}^{L} \pi_i(j) \mathcal{N}(y_{t+1}; \mu_j, \Sigma_j) m_{t+2,t+1}(j) \tag{47}$$

$$m_{T+1,T}(i) = 1 \quad i = 1, \ldots, L. \tag{48}$$

The Gibbs sampler using blocked re-sampling of $z_{1:T}$ is outlined in Algorithm 2.

Given a previous set of state-specific transition densities $\boldsymbol{\pi}^{(n-1)}$, the global transition density $\beta^{(n-1)}$, and observation likelihood 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, for each $t \in \{T, \ldots, 1\}$ calculate messages $m_{t,t-1}(k)$:
   a) For each $k \in \{1, \ldots, L\}$, initialize messages to
   $$m_{T+1,T}(k) = 1$$
   b) For each $k \in \{1, \ldots, L\}$, 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 $(j,k) \in \{1, \ldots, L\}^2$:
   a) For each $k \in \{1, \ldots, L\}$, 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 m, w, and $\bar{m}$ as in step 3 of Algorithm 1.
4) Update the global transition density by sampling
   $$\beta \sim \text{Dir}(\gamma/L + \bar{m}_{\cdot 1}, \ldots, \gamma/L + \bar{m}_{\cdot K})$$
5) For each $k \in \{1, \ldots, L\}$, sample a new transition density and observation likelihood parameters based on the sampled state assignments
   $$\begin{aligned} \pi_k &\sim \text{Dir}(\alpha\beta_1 + n_{k1}, \ldots, \alpha\beta_k + \kappa + n_{kk}, \ldots, \alpha\beta_L + n_{kL}) \\ \theta_k &\sim p(\theta \mid \lambda, \mathcal{Y}_k) \end{aligned}$$
   See Appendix II for details on resampling θ_k.
6) Fix $\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}$, $\beta^{(n)} = \beta$, and $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}$.

Algorithm 2: Blocked-z Gibbs sampler for the tempered HDP-HMM. The algorithm for the HDP-HMM follows directly by setting $\kappa = 0$. Here, we assume Gaussian observations with an independent Gaussian prior on the mean and inverse-Wishart (IW) prior on the covariance (see Appendix I-B). The quantity $\mathcal{Y}_k$ is a set of statistics for 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. Hyperparameters may be resampled, according to the formulas in Appendix V-D, as a final step.

D. Hyperparameter Re-Sampling

In the discussion thus far, we have assumed that the hyperparameter values are known. However, one may place a prior over these parameters and sample them as well. The sampling equations for the HDP-HMM can be found in [1]. Our derivation of the tempered sampling equations roughly follows that that of the original HDP-HMM, the details of which can be found in Appendix III.

Since we have the deterministic relationships

$$\begin{aligned} \alpha &= (1-\rho)(\alpha+\kappa) \\ \kappa &= \rho(\alpha+\kappa), \end{aligned} \tag{49}$$

we can treat ρ and $\alpha+\kappa$ as our hyperparameters and sample these values instead of sampling α and κ directly. This greatly simplifies the inference procedure as we will see below.

If we place a Beta(c,d) prior on ρ, the posterior distribution given samples of w_{jt} is simply an updated beta distribution:

$$p(\rho \mid \boldsymbol{w}) \propto \rho^{\sum_{j=1}^{J} w_{j\cdot}+c-1}(1-\rho)^{m_{\cdot\cdot}-\sum_{j=1}^{J} w_{j\cdot}+d-1} \propto \text{Beta}(\sum_{j=1}^{J} w_{j\cdot}+c, m_{\cdot\cdot}-\sum_{j=1}^{J} w_{j\cdot}+d), \tag{50}$$

where $m_{\cdot\cdot}=\sum_k m_{\cdot k}$ is the total number of tables in the *actual* franchise. Here, $m_{\cdot\cdot}$ represents the number of draws of $w_{jt}\sim \text{Ber}(\rho)$ and $\sum_j w_{j\cdot}$ the number of Bernoulli successes.

When given the total number of tables in the *actual* franchise, $m_{\cdot\cdot}$, the posterior distribution of the tempered concentration parameter $\alpha+\kappa$ follows the same distribution as that of α in the original HDP-HMM with auxiliary variables r_j and s_j:

$$p(\alpha+\kappa \mid r,s,\boldsymbol{m}) \propto (\alpha+\kappa)^{a-1+m_{\cdot\cdot}-\sum_{j=1}^{J} s_j} e^{-(\alpha+\kappa)(b-\sum_{j=1}^{J}\log r_j)} \tag{51}$$

$$p(r_j \mid \alpha+\kappa, n_{j\cdot}) \propto r_j^{(\alpha+\kappa)}(1-r_j)^{n_{j\cdot}-1} \propto \text{Beta}(\alpha+\kappa+1, n_{j\cdot}) \tag{52}$$

$$p(s_j \mid \alpha+\kappa, n_{j\cdot}) \propto \left(\frac{n_{j\cdot}}{\alpha+\kappa}\right)^{s_j} \tag{53}$$

Similarly, the posterior distribution of γ is the same as in the original HDP-HMM formulation if we now use the total number of tables in the *underlying* franchise, $\bar{m}_{\cdot\cdot}$, and the number of unique dishes *considered*, $\bar{K}$, along with auxiliary variable η:

$$p(\gamma \mid \eta, \bar{K}, \bar{m}_{\cdot\cdot}) \propto \pi_{\bar{m}}\text{Gamma}(a+\bar{K}, b-\log\eta) + (1-\pi_{\bar{m}})\text{Gamma}(a+\bar{K}-1, b-\log\eta) \tag{54}$$

$$p(\eta \mid \gamma, \bar{K}) \propto \eta^{\gamma}(1-\eta)^{\bar{m}_{\cdot\cdot}-1} \propto \text{Beta}(\gamma+1, \bar{m}_{\cdot\cdot}), \tag{55}$$

where,

$$\pi_{\bar{m}} = \frac{a + \bar{K} - 1}{\bar{m}_{\cdot\cdot}(b - \log \eta)}. \tag{56}$$

VI. Tempered HDP-HMM with Non-Standard Emission Densities

With the tempered HDP-HMM, we may now examine more complicated emission densities. So far, we have assumed that the state-conditional emission distribution was a simple parametric distribution such as a Gaussian. Often, however, the underlying state process we aim to capture may be better described as generating observations from some multimodal or otherwise more general emission distribution. We approximate each of these state-specific non-standard emission distributions by an infinite Gaussian mixture model with a Dirichlet process prior. This formulation is related to the nested Dirichlet process of [14], which uses a Dirichlet process to partition data into groups, and then models each group via a Dirichlet process mixture. What allows us to distinguish between the underlying HDP-HMM states is the structure on this state sequence and the bias towards self-transitions. If the model was free to both rapidly switch between HDP-HMM states and associate multiple Gaussians with each state, there would be a considerable amount of posterior uncertainty. Thus, it is only with the tempered HDP-HMM that we can effectively learn such models.

The generative model is as follows. We augment the HDP-HMM state z_t with a term s_t, which indexes the mixture component of the z_t^{th} emission density. The state evolution of z_t is described by the same Markov process as before:

$$\begin{aligned} \beta &\sim \text{GEM}(\gamma) \\ \pi_k &\sim \text{DP}(\alpha + \kappa, \frac{\alpha\beta + \kappa\delta_k}{\alpha + \kappa}) \\ z_t &\sim \pi_{z_{t-1}}. \end{aligned} \tag{57}$$

For each HDP-HMM state value k, there is a unique stick-breaking density ψ_k defining the mixture weights of the k^{th} emission density. These state-specific mixture weights and associated parameters have a Dirichlet process prior $\text{DP}(\sigma, H)$. Conditioned on z_t, the mixture index s_t is generated as:

$$\begin{aligned} \psi_k &\sim \text{GEM}(\sigma) \\ s_t &\sim \psi_{z_t}. \end{aligned} \tag{58}$$

Given the augmented state (z_t, s_t), the observation y_t is then generated by the Gaussian mixture compo-

nent parameterized by θ_{z_t,s_t}:

$$\begin{aligned} \theta_{k,j} &\sim H(\lambda) \\ y_t &\sim F(\theta_{z_t,s_t}). \end{aligned} \tag{59}$$

Note that both the HDP-HMM state index and mixture component index are allowed to take values in a countably infinite set. See Fig. 8 for a graphical model of this process.

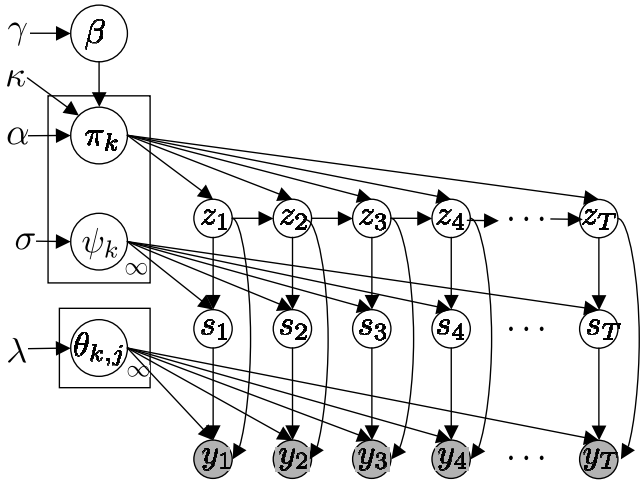


Fig. 8. Graphical model of a tempered HDP-HMM with infinite Gaussian mixture observation likelihoods. The model is as before, but with an added term s_t indexing the state-specific mixture component generating observation y_t. The mixture component index s_t is drawn from the z_t^{th} stick-breaking density ψ_{z_t}, where $\psi_k \sim \text{GEM}(\sigma)$. The parameters $\theta_{k,j}$ index the mean and covariance of the j^{th} Gaussian component of the k^{th} mixture density. Thus, $y_t \sim F(\theta_{z_t,s_t})$.

A. Direct Assignment Sampler

Much of the direct assignment sampler for the tempered HDP-HMM with infinite Gaussian mixture emissions remains the same as for the regular tempered HDP-HMM. Specifically, the sampling of global transition density β, number of tables in the actual restaurants m, override variables w, and number of tables in the underlying restaurants $\bar{m}$ is as presented in Eq. (30)-(35). The difference arises in how we sample our augmented state (z_t, s_t).

We can write the conditional distribution on the augmented state, having marginalized out the transition densities π_k and mixture component densities ψ_k, as:

$$\begin{aligned} p(z_t = k, s_t = j \mid z_{\backslash t}, s_{\backslash t}, y_{1:T}, \beta, \alpha, \sigma, \kappa, \lambda) &= p(s_t = j \mid z_t = k, z_{\backslash t}, s_{\backslash t}, y_{1:T}, \sigma, \lambda) \\ &\quad\; p(z_t = k \mid z_{\backslash t}, s_{\backslash t}, y_{1:T}, \beta, \alpha, \kappa, \lambda). \end{aligned} \tag{60}$$

The terms of this distribution, derived in Appendix I-C, are given by:

$$\begin{aligned} p(z_t = k \mid z_{\setminus t}, s_{\setminus t}, y_{1:T}, \beta, \alpha, \kappa, \lambda) &\propto p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) \\ &\quad \sum_{s_t} p(s_t \mid \{s_\tau \mid z_\tau = k, \tau \neq t\}, \sigma) p(y_t \mid \{y_\tau \mid z_\tau = k, s_t, \tau \neq t\}, \lambda) \quad (61) \\ p(s_t = j \mid z_t = k, z_{\setminus t}, s_{\setminus t}, y_{1:T}, \sigma, \lambda) &\propto p(s_t = j \mid \{s_\tau \mid z_\tau = k, \tau \neq t\}, \sigma) \\ &\quad p(y_t \mid \{y_\tau \mid z_\tau = k, s_t = j, \tau \neq t\}, \lambda). \quad (62) \end{aligned}$$

The component $p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa)$ of this pmf is as in Eq. (37) while $p(s_t \mid \{s_\tau \mid z_\tau = k, \tau \neq t\}, \sigma)$ is simply the Chinese restaurant process for the Dirichlet process associated with the state $z_t = k$. Let n'_{kj} be the number of observations with $(z_\tau = k, s_\tau = j)$. Then,

$$p(s_t = j \mid \{s_\tau \mid z_\tau = k, \tau \neq t\}, \sigma) \propto \begin{cases} \frac{n'^{-t}_{kj}}{\sigma + n'^{-t}_{k\cdot}}, & j \in \{1, \ldots, K'_k\}; \\ \frac{\sigma}{\sigma + n'^{-t}_{k\cdot}}, & j = \tilde{k}'_k, \end{cases} \quad (63)$$

where K'_k are the number of currently instantiated mixture components for the k^{th} emission density and $\tilde{k}'_k$ represents a new, previously unseen component. The component $p(y_t \mid \{y_\tau \mid z_\tau = k, s_t = j, \tau \neq t\}, \lambda)$ is the observation likelihood of y_t given an assignment $(z_t = k, s_t = j)$ conditioned on all other observations with this assignment, having marginalized out the parameter $\theta_{k,j}$. This distribution is further discussed in Appendix I-C.

The direct assignment sampler blocks the sampling of (z_t, s_t) and first draws z_t from the pmf defined by Eq. (61) and then s_t from the pmf of Eq. (62), conditioned on the sampled value of z_t. See Algorithm 3 for an outline of the direct assignment sampler for the tempered HDP-HMM with infinite Gaussian mixture emissions.

B. Blocked Sampler

In order to implement a blocked sampling of $(z_{1:T}, s_{1:T})$, we once again use the weak limit approximation to the Dirichlet process. Our derivations in this section are similar to those of Sec. V-C. For the tempered HDP-HMM with infinite Gaussian mixture emissions, the prior distributions on β, π_k, and ψ_k are defined as:

$$\begin{aligned} \beta &\sim \text{Dir}(\gamma/L, \ldots, \gamma/L) \\ \pi_k &\sim \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_k + \kappa, \ldots, \alpha\beta_L) \\ \psi_k &\sim \text{Dir}(\sigma/L', \ldots, \sigma/L'), \quad (64) \end{aligned}$$

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

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 $t \in \{1, \ldots, T\}$, sequentially

   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 $(z_t, s_t) = (k, j)$:
   $$(\hat{\mu}_{k,j}, \hat{\Sigma}_{k,j}) \leftarrow (\hat{\mu}_{k,j}, \hat{\Sigma}_{k,j}) \ominus y_t$$

   b) For each of the K currently instantiated HDP-HMM states compute

      i) The predictive conditional likelihood for each of the K'_k currently instantiated mixture components associated with this HDP-HMM state
      $$f'_{k,j}(y_t) \quad = \quad \left(\frac{n'_{kj}}{\sigma + n'_{k\cdot}}\right) \mathcal{N}(y_t; \hat{\mu}_{k,j}, \hat{\Sigma}_{k,j})$$
      and for a new mixture component $\tilde{k}'_k$
      $$f'_{k,\tilde{k}'_k}(y_t) \quad = \quad \frac{\sigma}{\sigma + n'_{k\cdot}} \mathcal{N}(y_t; \hat{\mu}_0, \hat{\Sigma}_0).$$

      ii) The predictive conditional likelihood of the HDP-HMM state without knowledge of the current mixture component
      $$f_k(y_t) \quad = \quad (\alpha\beta_k + n_{z_{t-1}k}) \left(\frac{\alpha\beta_{z_{t+1}} + n_{kz_{t+1}} + \kappa\delta(k, z_{t+1})}{\alpha + n_{k\cdot} + \kappa}\right) \left(\sum_{j=1}^{K'_k} f'_{k,j}(y_t) + f'_{k,\tilde{k}'_k}(y_t)\right)$$
      for $z_{t-1} \neq k$, otherwise see Appendix I-C. Repeat this procedure for a new HDP-HMM state $\tilde{k}$ with $K'_{\tilde{k}}$ initialized to 0, implying we only consider mixture component $\tilde{k}'_{\tilde{k}}$.

   c) Sample the new augmented state assignment (z_t, s_t) by first sampling z_t:
   $$z_t \quad \sim \quad \sum_{k=1}^{K} f_k(y_t)\delta(z_t, k) + f_{\tilde{k}}(y_t)\delta(z_t, \tilde{k}).$$
   Then, conditioned on a new assignment $z_t = k$, sample s_t:
   $$s_t \quad \sim \quad \sum_{j=1}^{K'_k} f'_{k,j}(y_t)\delta(s_t, j) + f'_{k,\tilde{k}'_k}(y_t)\delta(s_t, \tilde{k}'_k).$$
   If $z_t = \tilde{k}$, 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 $s_t = \tilde{k}'_{z_t}$, then increment K'_{z_t}.

   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 $(z_t, s_t) = (k, j)$:
   $$(\hat{\mu}_{k,j}, \hat{\Sigma}_{k,j}) \leftarrow (\hat{\mu}_{k,j}, \hat{\Sigma}_{k,j}) \oplus y_t$$

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 m, w, and $\bar{m}$ as in step 3 of Algorithm 1.

4) Sample the global transition density $\beta^{(n)}$ as in step 4 of Algorithm 1.

Algorithm 3: Direct assignment Rao–Blackwellized Gibbs sampler for the tempered HDP-HMM with infinite Gaussian mixture emissions. Hyperparameters may be resampled, according to the formulas in Appendix V-D, as a final step.

where L' is the approximation level for the Gaussian mixture emissions. The posterior distributions are given by:

$$\begin{aligned}
\beta &\sim \text{Dir}(\gamma/L + \bar{m}_{\cdot 1}, \ldots, \gamma/L + \bar{m}_{\cdot L}) \\
\pi_k &\sim \text{Dir}(\alpha\beta_1 + n_{k1}, \ldots, \alpha\beta_k + \kappa + n_{kk}, \ldots, \alpha\beta_L + n_{kL}) \\
\psi_k &\sim \text{Dir}(\sigma/L' + n'_{k1}, \ldots, \sigma/L' + n'_{kL'}).
\end{aligned} \tag{65}$$

For this model, the parameter $\theta_{k,j}$ defines the mean and covariance for the j^{th} Gaussian mixture component of the k^{th} emission distribution. The posterior of this parameter is determined by the observations assigned to this component, namely,

$$\theta_{k,j} \sim p(\theta \mid \{y_t \mid (z_t = k, s_t = j)\}, \lambda). \tag{66}$$

We now examine how to sample this augmented state (z_t, s_t). The conditional distribution of (z_t, s_t) for the forward-backward procedure is derived as:

$$p(z_t, s_t \mid z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{\theta}) \propto p(z_t \mid \pi_{z_{t-1}}) p(s_t \mid \psi_{z_t}) p(y_t \mid \theta_{z_t, s_t}) m_{t+1,t}(z_t). \tag{67}$$

Since the Markov 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:

$$m_{t,t-1}(z_{t-1}) \propto \begin{cases} \sum_{z_t} \sum_{s_t} p(z_t \mid \pi_{z_{t-1}}) p(s_t \mid \psi_{z_t}) p(y_t \mid \theta_{z_t, s_t}) m_{t+1,t}(z_t), & t \leq T; \\ 1, & t = T+1. \end{cases} \tag{68}$$

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

$$p(z_t = k, s_t = j \mid z_{t-1}, y_{1:T}, \boldsymbol{\pi}, \boldsymbol{\psi}, \boldsymbol{\theta}) \propto \pi_{z_{t-1}}(k) \psi_k(j) \mathcal{N}(y_t; \mu_{k,j}, \Sigma_{k,j}) m_{t+1,t}(k) \tag{69}$$

$$m_{t+1,t}(k) = \sum_{i=1}^{L} \sum_{\ell=1}^{L'} \pi_k(i) \psi_i(\ell) \mathcal{N}(y_{t+1}; \mu_{i,\ell}, \Sigma_{i,\ell}) m_{t+2,t+1}(i) \tag{70}$$

$$m_{T+1,T}(k) = 1 \quad k = 1, \ldots, L. \tag{71}$$

Algorithm 4 outlines the blocked-state sampler for the tempered HDP-HMM with infinite Gaussian mixture emissions.

Given a previous set of state-specific transition densities $\boldsymbol{\pi}^{(n-1)}$ and likelihood mixture weights $\boldsymbol{\psi}^{(n-1)}$, the global transition density $\beta^{(n-1)}$, and observation likelihood 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, for each $t \in \{T, \ldots, 1\}$ calculate messages $m_{t,t-1}(k)$:
   a) For each $k \in \{1, \ldots, L\}$, initialize messages to
   $$m_{T+1,T}(k) = 1$$
   b) For each $k \in \{1, \ldots, L\}$, compute
   $$m_{t,t-1}(k) = \sum_{i=1}^{L} \sum_{\ell=1}^{L'} \pi_k(i)\psi_i(\ell)\mathcal{N}(y_{t+1}; \mu_{i,\ell}, \Sigma_{i,\ell}) m_{t+2,t+1}(i)$$
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 $(i,k) \in \{1, \ldots, L\}^2$ and $(k,j) \in \{1, \ldots, L\} \times \{1, \ldots, L'\}$.
   a) For each $(k,j) \in \{1, \ldots, L\} \times \{1, \ldots, L'\}$, 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 m, w, and $\bar{m}$ as in step 3 of Algorithm 1.
4) Update the global transition density β by sampling as in step 4 of Algorithm 2.
5) For each $k \in \{1, \ldots, L\}$, sample a new transition density π_k and likelihood mixture weights ψ_k:
   $$\begin{aligned} \pi_k &\sim \text{Dir}(\alpha\beta_1 + n_{k1}, \ldots, \alpha\beta_k + \kappa + n_{kk}, \ldots, \alpha\beta_L + n_{kL}) \\ \psi_k &\sim \text{Dir}(\sigma/L' + n'_{k1}, \ldots, \sigma/L' + n'_{kL'}) \end{aligned}$$
   a) For each $j \in \{1, \ldots, L'\}$, 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 Appendix II for details on resampling $\theta_{k,j}$.
6) Fix $\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}$, $\boldsymbol{\psi}^{(n)} = \boldsymbol{\psi}$, $\beta^{(n)} = \beta$, and $\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta}$.

Algorithm 4: Blocked-state Gibbs sampler for the tempered HDP-HMM with infinite Gaussian mixture emissions. Here, we use an independent Gaussian prior on the mean and inverse-Wishart (IW) prior on the covariance (see Appendix I-B). The quantity $\mathcal{Y}_{k,j}$ is a set of statistics for 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. Hyperparameters may be resampled, according to the formulas in Appendix V-D, as a final step.

VII. RESULTS

To analyze the performance of the tempered HDP-HMM as compared to the original, we generated test data and applied the direct assignment and blocked-z sampler to both model variants (i.e. the HDP-HMM with and without the κ term.) The test data sequence is shown in Fig. 5(a) and was generated by a three-state HMM with 0.97 probability of self-transition and equally likely transitions to the other two states. The Gaussian observation likelihood densities had means 50, 0, and -50 and variances 50, 10, and 50, respectively. We ran 100 iterations of each of the Gibbs samplers with 200 different initializations. For the blocked-z sampler, we used a truncation level of $L = 15$, though the sampler learns to use a strict subset of the pool of available states.

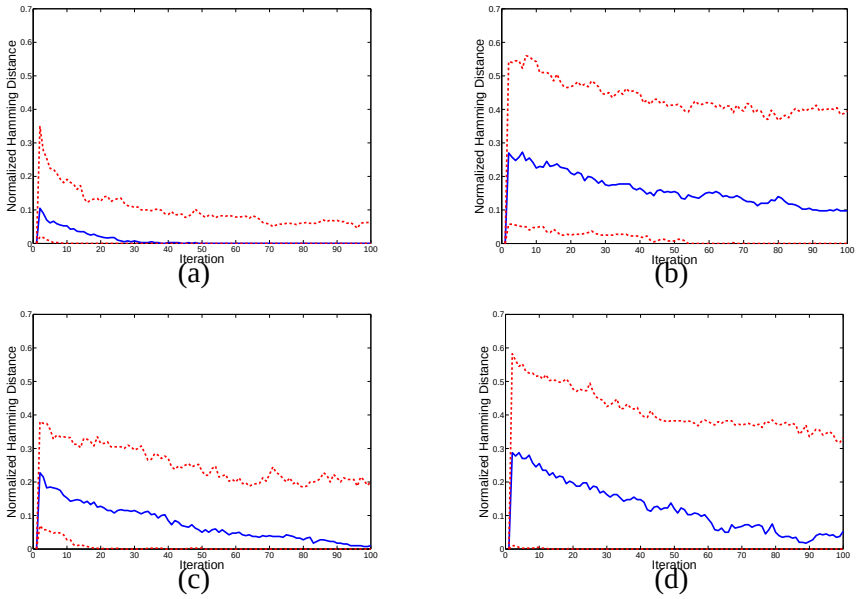


Fig. 9. Plots of Hamming distance between true and estimated state sequences over 100 iterations for the: (a) blocked-z sampler for the tempered HDP-HMM, (b) direct assignment sampler for the tempered HDP-HMM, (c) blocked-z sampler for the original HDP-HMM, and (d) direct assignment sampler for the original HDP-HMM. These plots show the median (solid blue) and 10^{th} and 90^{th} quantiles (dashed red) from 200 initializations of the sampler.

In Fig. 9, we plot the 10, 50, and 90-quantiles of the Hamming distance between the true and estimated state sequences over the 100 Gibbs iterations for each of the four samplers. To calculate the Hamming distance, we first map the randomly chosen indices of the estimated state sequence to the set of indices that maximize the overlap with the true sequence. We do this in a greedy fashion by starting with the most frequent state index of the true sequence and finding the corresponding state index of the estimated sequence with the most overlap. We use this corresponding state index to relabel the index of the true sequence and add it to the list of used indices. We then iterate with the next most frequent state index. If

the estimated state sequence has fewer states than the true state sequence, the extra true states are labeled with one of the remaining unused indices in $\{1, \ldots, L\}$.

In Fig. 10, we plot the 10, 50, and 90-quantiles of the log-likelihood of the observation sequence given the estimated set of parameters $\boldsymbol{\theta}$, $\boldsymbol{\pi}$, $z_{1:T}$, β, and the hyperparameters. For the direct assignment samplers, where $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ are integrated out, these parameters are sampled from the posterior distributions $p(\boldsymbol{\pi} \mid z_{1:T}, \beta)$ and $p(\boldsymbol{\theta} \mid z_{1:T}, y_{1:T})$, respectively.

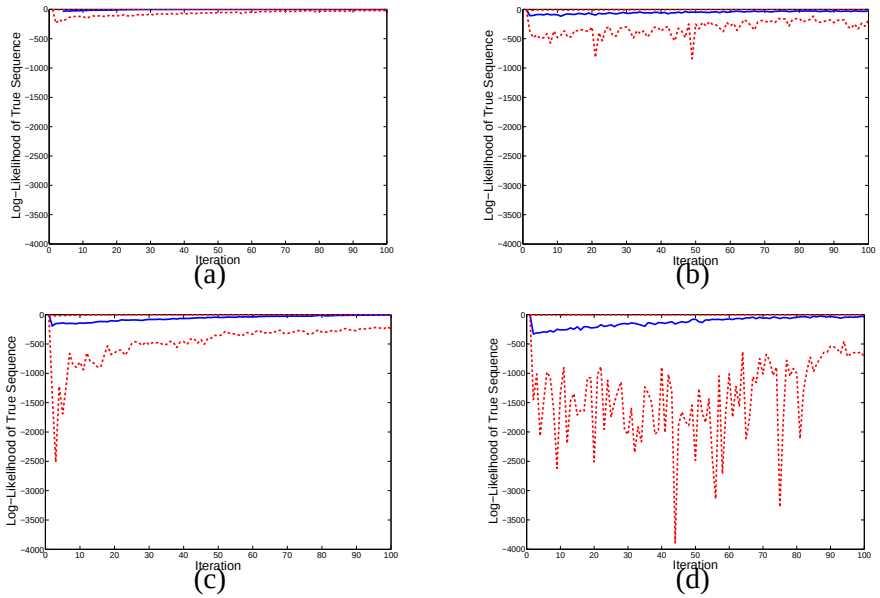


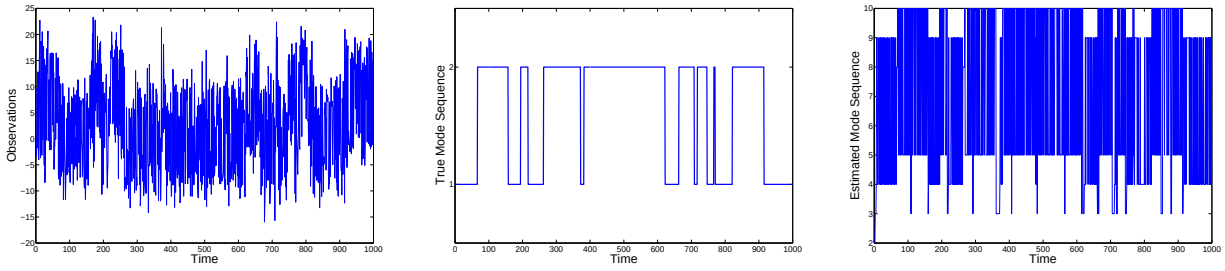
Fig. 10. Plots of log-likelihood of the observation sequence given the estimated set of parameters over 100 iterations for the: (a) blocked-z sampler for the tempered HDP-HMM, (b) direct assignment sampler for the tempered HDP-HMM, (c) blocked-z sampler for the original HDP-HMM, and (d) direct assignment sampler for the original HDP-HMM. These plots show the median (solid blue) and 10^{th} and 90^{th} quantiles (dashed red) from 200 initializations of the sampler.

From these plots, we see the performance gain of the blocked-z sampler for the tempered HDP-HMM as compared to the other samplers, both in terms of Hamming error and estimated model likelihood. As expected, the tempered HDP-HMM with the sequential, direct assignment sampler has the next largest likelihood of the estimated model (due to avoiding the fast state-switching sequences), but gets stuck in state sequence assignments that are hard to move away from, as conveyed by the flatness of the Hamming error versus iteration number plot in Fig. 9(b). For example, the estimated state sequence of Fig. 7(c) might have similar parameters associated with states four and five so that the model 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-z sampler with the original HDP-HMM improves the Hamming distance performance relative

to the sequential, direct assignment sampler for both the original and tempered HDP-HMM; however, the likelihood of the models estimated by both of the original HDP-HMM samplers are dramatically worse than those for the tempered HDP-HMM due to poor parameter estimates associated with the fast state-switching assignments (see Fig. 5(c).)

Now that we have established the benefit of the tempered HDP-HMM for modeling processes where the underlying state persists for lengthy periods of time, we may analyze extensions of this model to non-standard emission densities, as discussed in Sec. VI. To test the model of Sec. VI, we generated data from a two-state HMM, where each state had a two-Gaussian mixture emission distribution. For one state, the Gaussian mixture components were defined by means 0 and 10 while the other state's components had means -7 and 7. Each Gaussian mixture component had variance 10 and was equally weighted in the mixture. The choice of these parameter values enabled each emission distribution to be sufficiently multimodal while still maintaining significant overlap in the observation spaces of these two states. The probability of self-transition was set to 0.98. The large probability of self-transition is what disambiguates this process from one with four states, each with a single Gaussian emission distribution.

The resulting observation and true state sequences are shown in Fig. 11(a) and (b), respectively. In Fig. 11(c) we plot an estimated state sequence from the sampler using the tempered HDP-HMM when constrained to single Gaussian emissions. With such a model, a good explanation of the data is to create a state for each mixture component and then quickly switch between these states. Although not the desired effect in this scenario, this behavior demonstrates the flexibility of the tempered HDP-HMM: if the best explanation of the data according to the model is fast state-switching, the tempered HDP-HMM still allows for this by learning a small bias towards self-transitions.



(a) (b) (c)

Fig. 11. Qualitative plots showing the performance of the tempered HDP-HMM with single Gaussian emissions when the data was generated by an HMM with Gaussian mixture emissions. (a) Observation sequence; (b) true HMM state sequence; and (c) estimated HMM state sequence using the tempered HDP-HMM model. In plot (c), we see that since the model is constrained to single Gaussian emission distribution, the best explanation of the data is to separate into the components of the Gaussian mixture emissions and quickly switch between this set of states.

We tested the performance of the tempered HDP-HMM with infinite Gaussian mixture emissions against that of the tempered HDP-HMM with single Gaussian emissions. We then compared these results to those corresponding to the original HDP-HMM (i.e. no bias towards self-transitions.) We only present results from blocked-state sampling since we have seen the clear advantages of this method over the sequential, direct assignment sampler. For both the HDP-HMM portion of the model and the Dirichlet process mixture model emissions, we use a truncation level of $L = L' = 15$. The resulting performance, in terms of the Hamming distance metric, are summarized in Fig. 12.

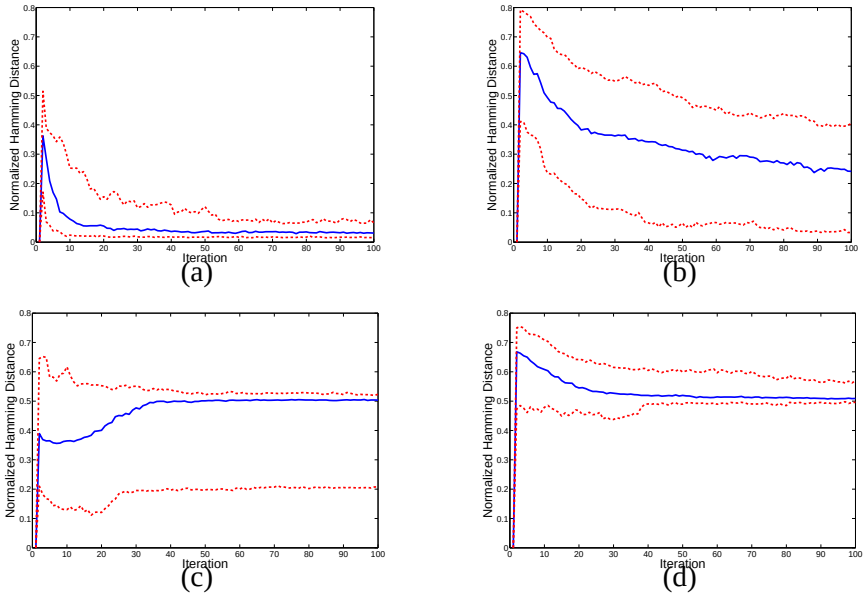


Fig. 12. Plots of Hamming distance between true and estimated state sequences over 100 iterations of the blocked-state sampler for the: (a) tempered HDP-HMM with infinite Gaussian mixture emissions, (b) original HDP-HMM with infinite Gaussian mixture emissions, (c) tempered HDP-HMM with single Gaussian emissions, and (d) original HDP-HMM with single Gaussian emissions. These plots show the median (solid blue) and 10^{th} and 90^{th} quantiles (dashed red) from 200 initializations of the sampler.

The results are rather intuitive and can be explained as follows. When the original HDP-HMM is constrained to single Gaussian emissions, the best explanation of the data is to associate each true Gaussian mixture component with a separate state and then quickly switch between these states, resulting in the large Hamming distances of Fig. 12(d). When using the tempered HDP-HMM with single Gaussian emissions, the bias towards self-transitions occasionally leads to more accurate state sequence estimates by grouping an individual true state's Gaussian mixture components into a single Gaussian with large variance. This behavior explains why the 10^{th} quantile of Fig. 12(c) is lower than that of the original HDP-HMM. The initial dip of the median and 10^{th} quantile is explained by the tempered HDP-HMM

initially grouping Gaussian mixture components and then preferring the split assignment. The original HDP-HMM with infinite Gaussian mixture emissions has improved performance over either of the models constrained to single Gaussian emissions. However, the tempered HDP-HMM with infinite Gaussian mixture emissions has by far the best performance due to the incorporated bias towards self-transitions so that fast state-switching is a less preferable explanation of the data.

VIII. Discussion

We have demonstrated that the original HDP-HMM is underconstrained in describing processes where the underlying state persists for lengthy periods of time. As an alternative, we have presented a tempered HDP-HMM, which allows for efficiently learning representative models of such processes. We have also extended the HDP-HMM to allow for non-standard emission densities approximated by infinite Gaussian mixtures. We are able to disambiguate such models because of the tempered HDP-HMM's bias towards self-transitions. We are currently investigating more challenging datasets such as using the tempered HDP-HMM for speaker diarization, the problem of partitioning an audio segment into homogeneous regions corresponding to an unknown number of distinct speakers.

Appendix I
Predictive Distribution of State Assignments

In this appendix we derive the predictive distribution for state assignments, $p(z_t = k \mid z_{\setminus t}, y_{1:T}, \beta, \alpha, \kappa)$, as used by the direct assignment sampler. For these derivations we will include the κ term of the tempered HDP-HMM, though the derivations for the original HDP-HMM follow directly by setting $\kappa = 0$. We derive the desired predictive distribution by considering the joint distribution over all random variables in the model and then marginalizing the transition densities π and parameters θ:

$$\begin{aligned} p(z_t = k \mid z_{\setminus t}, y_{1:T}, \beta, \alpha, \kappa) &\propto \int_{\boldsymbol{\pi}} \prod_i p(\pi_i \mid \alpha, \beta, \kappa) \prod_\tau p(z_\tau \mid \pi_{z_{\tau-1}}) d\boldsymbol{\pi} \\ &\quad \int_{\boldsymbol{\theta}} \prod_k p(\theta_k \mid \lambda) \prod_\tau p(y_\tau \mid \theta_{z_\tau}) d\boldsymbol{\theta} \\ &\propto p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) p(y_t \mid y_{\setminus t}, z_t = k, z_{\setminus t}, \lambda) \end{aligned} \tag{72}$$

The term $p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa)$, which arises from integration over π, is the Chinese restaurant franchise while $p(y_t \mid y_{\setminus t}, z_{t=k}, z_{\setminus t}, \lambda)$ is the observation likelihood of an assignment $z_t = k$ having marginalized the parameter θ_k. These distributions are further examined in the following two sections. We then examine the predictive distribution for the tempered HDP-HMM with infinite Gaussian mixture emission densities.

A. Chinese Restaurant Franchise

Let $\beta_{\tilde{k}} = \sum_{k=K+1}^{\infty} \beta_k$, where K is the number of currently instantiated states and $K+1$ indexes a potentially new state. Then, $\pi_j \sim \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_j + \kappa, \ldots, \alpha\beta_{\tilde{k}})$. Marginalizing over π induces a prior predictive distribution on z_t that is a variant of the Chinese restaurant franchise. Our result differs from that of the standard Chinese restaurant franchise because the indicator random variables z_t have a Markov structure. We analyze this distribution by continuing from the integration over π in Eq. (72):

$$\begin{aligned} p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) &\propto \int_{\boldsymbol{\pi}} \prod_i p(\pi_i \mid \alpha, \beta, \kappa) \prod_\tau p(z_\tau \mid \pi_{z_{\tau-1}}) d\boldsymbol{\pi} \qquad (73) \\ &\propto \int_{\boldsymbol{\pi}} p(z_{t+1} \mid \pi_k) p(z_t = k \mid \pi_{z_{t-1}}) \prod_i (p(\pi_i \mid \alpha, \beta, \kappa) \prod_{\tau \mid z_{\tau-1}=i, \tau \neq t, t+1} p(z_\tau \mid \pi_i)) d\boldsymbol{\pi} \\ &\propto \int_{\boldsymbol{\pi}} p(z_{t+1} \mid \pi_k) p(z_t = k \mid \pi_{z_{t-1}}) \prod_i p(\pi_i \mid \{z_\tau \mid z_{\tau-1} = i, \tau \neq t, t+1\}, \beta, \alpha, \kappa) d\boldsymbol{\pi}. \end{aligned}$$

Let $z_{t-1} = j$. By marginalization of the transition density π_j, the proposed state assignment $z_t = k$ is affected by all other states that were also drawn from π_j. In addition, this proposed assignment affects the likelihood of z_{t+1}, which is now considered to have been drawn from π_k, as dictated by z_t. We need to examine two scenarios: $k = j$, in which case z_t and z_{t+1} are both distributed according to the same transition density; and $k \neq j$, where these states are sampled from independent transition densities. We start by considering $k \neq j$, that is, an assignment of the state changing from j to k at time t:

$$\begin{aligned} p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) &\propto \int_{\pi_k} p(z_{t+1} \mid \pi_k) p(\pi_k \mid \{z_\tau \mid z_{\tau-1} = k, \tau \neq t, t+1\}, \beta, \alpha, \kappa) d\pi_k \\ &\quad \int_{\pi_j} p(z_t = k \mid \pi_j) p(\pi_j \mid \{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\}, \beta, \alpha, \kappa) d\pi_j \\ &\propto p(z_{t+1} \mid \{z_\tau \mid z_{\tau-1} = k, \tau \neq t, t+1\}, \beta, \alpha, \kappa) \qquad (74) \\ &\quad p(z_t = k \mid \{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\}, \beta, \alpha, \kappa). \end{aligned}$$

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

$$\begin{aligned} p(z_t = j \mid z_{\setminus t}, \beta, \alpha, \kappa) &\propto \int_{\pi_j} p(z_{t+1} \mid \pi_j) p(z_t = j \mid \pi_j) p(\pi_j \mid \{z_\tau \mid z_{\tau-1} = k, \tau \neq t, t+1\}, \beta, \alpha, \kappa) d\pi_j \\ &\propto p(z_t = j, z_{t+1} \mid \{z_\tau \mid z_{\tau-1} = k, \tau \neq t, t+1\}, \beta, \alpha, \kappa). \qquad (75) \end{aligned}$$

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. Consider the distribution of a generic

set of observations generated from a single transition density π_i given the hyperparameters α, β, and κ:

$$\begin{aligned}
p(\{z_\tau \mid z_{\tau-1} = i\} \mid \beta, \alpha, \kappa) &= \int_{\pi_i} p(\pi_i, \{z_\tau \mid z_{\tau-1} = i\} \mid \beta, \alpha, \kappa) d\pi_i \\
&= \int_{\pi_i} p(\pi_i \mid \beta, \alpha, \kappa) p(\{z_\tau \mid z_{\tau-1} = i\} \mid \pi_i) d\pi_i \\
&= \int_{\pi_i} \frac{\Gamma(\sum_k \alpha\beta_k + \kappa\delta(k,i))}{\prod_k \Gamma(\alpha\beta_k + \kappa\delta(k,i))} \prod_{k=1}^{K+1} \pi_{jk}^{\alpha\beta_k + \kappa\delta(k,i)-1} \prod_{k=1}^{K+1} \pi_{jk}^{n_{jk}} d\pi_i \\
&= \frac{\Gamma(\sum_k \alpha\beta_k + \kappa\delta(k,i))}{\prod_k \Gamma(\alpha\beta_k + \kappa\delta(k,i))} \int_{\pi_i} \prod_{k=1}^{K+1} \pi_{jk}^{\alpha\beta_k + \kappa\delta(k,i) + n_{jk} - 1} d\pi_i \\
&= \frac{\Gamma(\sum_k \alpha\beta_k + \kappa\delta(k,i))}{\prod_k \Gamma(\alpha\beta_k + \kappa\delta(k,i))} \frac{\prod_k \Gamma(\alpha\beta_k + \kappa\delta(k,i) + n_{jk})}{\Gamma(\sum_k \alpha\beta_k + \kappa\delta(k,i) + n_{jk})} \\
&= \frac{\Gamma(\alpha + \kappa)}{\Gamma(\alpha + \kappa + n_{i.})} \prod_k \frac{\Gamma(\alpha\beta_k + \kappa\delta(k,i) + n_{jk})}{\Gamma(\alpha\beta_k + \kappa\delta(k,i))}. \qquad (76)
\end{aligned}$$

We use Eq. (76) to determine that the first component of Eq. (74) is

$$\begin{aligned}
p(z_t = k \mid \{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\}, \beta, \alpha, \kappa) &= \frac{p(\{z_\tau \mid z_{\tau-1} = j, \tau \neq t+1, z_t = k\} \mid \beta, \alpha, \kappa)}{p(\{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\} \mid \beta, \alpha, \kappa)} \\
&= \frac{\Gamma(\alpha + \kappa + n_{j.}^{-t})}{\Gamma(\alpha + n_{j.}^{-t} + 1)} \frac{\Gamma(\alpha\beta_k + \kappa + n_{jk}^{-t} + 1)}{\Gamma(\alpha\beta_k + n_{jk}^{-t})} \\
&= \frac{\alpha\beta_k + n_{jk}^{-t}}{\alpha + n_{j.}^{-t}}. \qquad (77)
\end{aligned}$$

where n_{jk}^{-t} is the number of transitions from maneuver j to maneuver 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. (74) is derived to be

$$p(z_{t+1} \mid \{z_\tau \mid z_{\tau-1} = k, \tau \neq t, t+1\}, \beta, \alpha, \kappa) = \frac{\alpha\beta_\ell + \kappa\delta(\ell, k) + n_{kl}^{-t}}{\alpha + \kappa + n_{k.}^{-t}}, \qquad (78)$$

where $z_{t+1} = \ell$. For $k = j$, the distribution of Eq. (75) reduces to

$$\begin{aligned}
p(z_t = j, z_{t+1} \mid \{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\}, \beta, \alpha, \kappa) &= \frac{p(\{z_\tau \mid z_{\tau-1} = j\} \mid \beta, \alpha, \kappa)}{p(\{z_\tau \mid z_{\tau-1} = j, \tau \neq t, t+1\} \mid \beta, \alpha, \kappa)} \\
&= \begin{cases} \frac{\Gamma(\alpha+\kappa+n_{j\cdot}^{-t})}{\Gamma(\alpha+\kappa+n_{j\cdot}^{-t}+2)} \frac{\Gamma(\alpha\beta_j+\kappa+n_{jj}^{-t}+1)}{\Gamma(\alpha\beta_j+\kappa+n_{jj}^{-t})} \frac{\Gamma(\alpha\beta_\ell+n_{j\ell}^{-t}+1)}{\Gamma(\alpha\beta_\ell+n_{j\ell}^{-t})}, & z_{t+1} = \ell, \ell \neq j; \\ \frac{\Gamma(\alpha+\kappa+n_{j\cdot}^{-t})}{\Gamma(\alpha+\kappa+n_{j\cdot}^{-t}+2)} \frac{\Gamma(\alpha\beta_j+\kappa+n_{jj}^{-t}+2)}{\Gamma(\alpha\beta_j+\kappa+n_{jj}^{-t})}, & z_{t+1} = j. \end{cases} \\
&= \begin{cases} \frac{(\alpha\beta_j+\kappa+n_{jj}^{-t})(\alpha\beta_\ell+n_{j\ell}^{-t})}{(\alpha+\kappa+n_{j\cdot}^{-t}+1)(\alpha+\kappa+n_{j\cdot}^{-t})}, & z_{t+1} = \ell, \ell \neq j; \\ \frac{(\alpha\beta_j+\kappa+n_{jj}^{-t}+1)(\alpha\beta_j+\kappa+n_{jj}^{-t})}{(\alpha+\kappa+n_{j\cdot}^{-t}+1)(\alpha+\kappa+n_{j\cdot}^{-t})}, & z_{t+1} = j. \end{cases} \\
&= \frac{(\alpha\beta_j + \kappa + n_{jj}^{-t})(\alpha\beta_\ell + n_{j\ell}^{-t} + (\kappa+1)\delta(j,\ell))}{(\alpha + \kappa + n_{j\cdot}^{-t})(\alpha + \kappa + n_{j\cdot}^{-t} + 1)}
\end{aligned} \tag{79}$$

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

$$\begin{aligned}
&p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) \\
&\quad \propto \begin{cases} (\alpha\beta_k + n_{z_{t-1}k}^{-t} + \kappa\delta(z_{t-1}, k))(\frac{\alpha\beta_{z_{t+1}} + n_{kz_{t+1}}^{-t} + \kappa\delta(k,z_{t+1}) + \delta(z_{t-1},k)\delta(k,z_{t+1})}{\alpha + n_{k\cdot}^{-t} + \kappa + \delta(z_{t-1},k)}) & k \in 1, \ldots, K \\ \frac{\alpha^2\beta_{K+1}\beta_{z_{t+1}}}{\alpha+\kappa} & k = K+1 \end{cases}
\end{aligned} \tag{80}$$

B. Observation Likelihoods

We now further examine the observation likelihood term of Eq. (72). 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:

$$\begin{aligned}
p(y_t \mid y_{\setminus t}, z_t = k, z_{\setminus t}, \lambda) &\propto \int_{\theta_k} p(y_t \mid \theta_k) p(\theta_k \mid \lambda) \prod_{\tau \mid z_\tau = k, \tau \neq t} p(y_\tau \mid \theta_k) d\theta_k \\
&\propto \int_{\theta_k} p(y_t \mid \theta_k) p(\theta_k \mid \{y_\tau \mid z_\tau = k, \tau \neq t\}, \lambda) d\theta_k \\
&\propto p(y_t \mid \{y_\tau \mid z_\tau = k, \tau \neq t\}, \lambda).
\end{aligned} \tag{81}$$

Note that the set $\{y_\tau \mid z_\tau = k, \tau \neq t\}$ denotes all the observations y_τ other than y_t that were drawn from the observation likelihood distribution parameterized by θ_k.

If we consider Gaussian observation likelihoods, the conjugate distribution for the unknown mean and covariance parameters is the normal-inverse-Wishart, 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 $\mathcal{NIW}(\bar{\zeta}, \bar{\vartheta}, \bar{\nu}, \bar{\Delta})$,

where

$$\bar{\zeta} = \zeta + |\{y_s \mid z_s = k, s \neq t\}| \triangleq \zeta + |Y_k| \tag{82}$$

$$\bar{\nu} = \nu + |Y_k| \tag{83}$$

$$\bar{\zeta}\bar{\vartheta} = \zeta\vartheta + \sum_{y_s \in Y_k} y_s \tag{84}$$

$$\bar{\nu}\bar{\Delta} = \nu\Delta + \sum_{y_s \in Y_k} y_s y_s^T + \zeta\vartheta\vartheta^T - \bar{\zeta}\bar{\vartheta}\bar{\vartheta}^T. \tag{85}$$

Marginalizing θ_k induces a multivariate Student-t predictive distribution for y_t, which can be approximated by a moment-matched Gaussian,

$$p(y_t \mid \{y_\tau \mid z_\tau = k, \tau \neq t\}, \zeta, \vartheta, \nu, \Delta) \simeq \mathcal{N}(y_t; \bar{\vartheta}, \frac{(\bar{\zeta}+1)\bar{\nu}}{\bar{\zeta}(\bar{\nu}-d-1)}\bar{\Delta}) \triangleq \mathcal{N}(y_t; \hat{\mu}_k, \hat{\Sigma}_k). \tag{86}$$

C. Tempered HDP-HMM with Infinite Gaussian Mixture Emissions

In this section we derive the predictive distribution on the augmented state (z_t, s_t) of the tempered HDP-HMM with infinite Gaussian mixture emissions. We use the chain rule to write:

$$\begin{aligned} p(z_t = k, s_t = j \mid z_{\setminus t}, s_{\setminus t}, y_{1:T}, \beta, \alpha, \sigma, \kappa, \lambda) \quad &= \quad p(s_t = j \mid z_t = k, z_{\setminus t}, s_{\setminus t}, y_{1:T}, \sigma, \lambda) \\ & \quad\quad p(z_t = k \mid z_{\setminus t}, s_{\setminus t}, y_{1:T}, \beta, \alpha, \kappa, \lambda). \end{aligned} \tag{87}$$

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 necessary parameters. For the conditional distribution of $z_t = k$ when *not* given s_t, this amounts to:

$$\begin{aligned} p(z_t = k \mid z_{\setminus t}, s_{\setminus t}, y_{1:T}, \beta, \alpha, \kappa, \lambda) \quad &\propto \quad \int_{\boldsymbol{\pi}} \prod_j p(\pi_j \mid \alpha, \beta, \kappa) \prod_\tau p(z_\tau \mid \pi_{z_{\tau-1}}) d\boldsymbol{\pi} \\ & \quad\quad \sum_{s_t} \int_{\boldsymbol{\psi}} \prod_j p(\psi_j \mid \sigma) \prod_\tau p(s_\tau \mid \psi_{z_\tau}) d\boldsymbol{\psi} \\ & \quad\quad\quad \int_{\boldsymbol{\theta}} \prod_{i,\ell} p(\theta_{i,\ell} \mid \lambda) \prod_\tau p(y_\tau \mid \theta_{z_\tau, s_\tau}) d\boldsymbol{\theta} \end{aligned} \tag{88}$$

$$\propto p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa) \sum_{s_t} p(s_t \mid \{s_\tau \mid z_\tau = k, \tau \neq t\}, \sigma) p(y_t \mid \{y_\tau \mid z_\tau = k, s_t, \tau \neq t\}, \lambda).$$

The component $p(z_t = k \mid z_{\setminus t}, \beta, \alpha, \kappa)$ is as in Eq. (80) while $p(s_t \mid \{s_\tau \mid z_\tau = k, \tau \neq t\}, \sigma)$ is the Chinese restaurant process for the Dirichlet process associated with the state $z_t = k$. We similarly derive the

conditional distribution of an assignment $s_t = j$ given $z_t = k$ as:

$$\begin{aligned} p(s_t = j \mid z_t = k, z_{\setminus t}, s_{\setminus t}, y_{1:T}, \sigma, \lambda) \quad &\propto \quad p(s_t = j \mid \{s_\tau \mid z_\tau = k, \tau \neq t\}, \sigma) \\ & \quad\quad p(y_t \mid \{y_\tau \mid z_\tau = k, s_t = j, \tau \neq t\}, \lambda). \end{aligned} \tag{89}$$

The observation 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. (86) where now we only consider the observations y_τ that are assigned to HDP-HMM state $z_\tau = k$ *and* mixture component $s_\tau = k$.

APPENDIX II
NON-CONJUGATE BASE MEASURES AND THE BLOCKED-STATE SAMPLER

Since the blocked-state sampler instantiates the parameters θ, 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 parameter space is over 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 $\text{IW}(\nu, \Delta)$ 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:

$$\begin{aligned} \Sigma_k \mid \mu_k \quad &\sim \quad \text{IW}(\nu_k \Delta_k, \nu_k) \\ \mu_k \mid \Sigma_k \quad &\sim \quad \mathcal{N}(\bar{\mu}_k, \bar{\Sigma}_k), \end{aligned} \tag{90}$$

where

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

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 tempered HDP-HMM with infinite Gaussian mixture emissions follows 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 III
HYPERPARAMETERS

In this appendix we expound upon the derivations of the conditional hyperparameter distributions used for resampling these random variables. The hyperparameters of our model include α, κ, γ, σ, and λ, though λ is considered fixed. Many of these derivations follow directly from those presented in [5], [1].

We have shown that it is sufficient to parameterize our model by $\alpha + \kappa$ and $\rho = \kappa/\alpha + \kappa$ instead of by α and κ independently. This greatly simplifies the resampling of these hyperparameters. Let us assume that there are J restaurants in the franchise at a given iteration of the sampler. As depicted in Fig. 2(b), the generative model dictates that for each restaurant j we have $\tilde{\pi}_j \sim \text{GEM}(\alpha + \kappa)$, 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 construction over table assignments resulting in $m_{j\cdot}$ unique tables. By Eq. (5) and using the fact that each restaurant is mutually conditionally independent, we may write:

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

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

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

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 $r = \{r_1,\ldots,r_J\}$, where each $r_j \in [0,1]$. Now, the integration introduced by the beta function is over the domain of each r_j so that we can represent the joint distribution of $\alpha+\kappa$

and r as

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

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

$$p(\alpha+\kappa, r, s \mid m_{1.}, \ldots, m_{J.}, n_{1.}, \ldots, n_{J.}) \propto (\alpha+\kappa)^{a+m_{..}-1} e^{-(\alpha+\kappa)b} \prod_{j=1}^{J} (\frac{n_{j.}}{\alpha+\kappa})^{s_j} r_j^{\alpha+\kappa}(1-r_j)^{n_{j.}-1}.$$

Each conditional distribution is as follows:

$$\begin{aligned} p(\alpha+\kappa \mid r, s, m_{1.}, \ldots, m_{J.}, n_{1.}, \ldots, n_{J.}) &\propto (\alpha+\kappa)^{a+m_{..}-1-\sum_{j=1}^{J} s_j} e^{-(\alpha+\kappa)(b-\sum_{j=1}^{J} \log r_j)} \\ p(r_j \mid \alpha+\kappa, r_{\setminus j}, s, m_{1.}, \ldots, m_{J.}, n_{1.}, \ldots, n_{J.}) &\propto r_j^{\alpha+\kappa}(1-r_j)^{n_{j.}-1} \\ p(s_j \mid \alpha+\kappa, r, s_{\setminus j}, m_{1.}, \ldots, m_{J.}, n_{1.}, \ldots, n_{J.}) &\propto (\frac{n_{j.}}{\alpha+\kappa})^{s_j}. \end{aligned} \tag{91}$$

We may similarly derive the conditional distribution of γ. The generative model depicted in Fig. 2(b) dictates that $\beta \sim \text{GEM}(\gamma)$ and that each table t considers ordering a dish $\bar{k}_{jt} \sim \beta$. From Eq. (35), we see that the sampled value $\bar{m}_{j.}$ 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, $\bar{m}_{..}$ 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:

$$\bar{K} = \sum_{j=1}^{J} \mathbf{1}(\bar{m}_{.j} > 0) = K - \sum_{j=1}^{J} \mathbf{1}(\bar{m}_{.j} = 0 \text{ and } m_{jj} > 0). \tag{92}$$

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 j arose from an override in restaurant j and this dish was never considered in any other restaurant. That is, there were no informative considerations of dish j, implying $\bar{m}_{.j} = 0$, while dish j was served in restaurant

j, implying $m_{jj} > 0$ so that j is counted in K. This is equivalent to counting how many dishes j had an informative table consider ordering dish j, regardless of the restaurant. We may now use Eq. (5) to form the condition distribution on γ:

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

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

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

Here, we have used the fact that there is a Gamma(a, b) prior on γ. The resulting conditional distributions are:

$$\begin{aligned} p(\gamma \mid \eta, \bar{K}, \bar{m}_{..}) &\propto \gamma^{a+\bar{K}-2}(\gamma + \bar{m}_{..})e^{-\gamma(b-\log\eta)} \\ &\propto \pi_{\bar{m}}\text{Gamma}(a + \bar{K}, b - \log\eta) + (1 - \pi_{\bar{m}})\text{Gamma}(a + \bar{K} - 1, b - \log\eta) \\ p(\eta \mid \gamma, \bar{K}, \bar{m}_{..}) &\propto \eta^{\gamma}(1-\eta)^{\bar{m}_{..}-1} \propto \text{Beta}(\gamma + 1, \bar{m}_{..}), \end{aligned} \tag{93}$$

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

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 $n_{j.}$ 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:

$$\begin{aligned} p(\sigma \mid r', s', K'_{1.}, \ldots, K'_{J.}, n_{1.}, \ldots, n_{J.}) &\propto (\sigma)^{a+K'_{..}-1-\sum_{j=1}^{J} s'_j} e^{-(\sigma)(b-\sum_{j=1}^{J}\log r'_j)} \\ p(r'_j \mid \sigma, r'_{\setminus j}, s', K'_{1.}, \ldots, K'_{J.}, n_{1.}, \ldots, n_{J.}) &\propto r_j^{'\sigma}(1-r'_j)^{n_{j.}-1} \\ p(s'_j \mid \sigma, r', s'_{\setminus j}, K'_{1.}, \ldots, K'_{J.}, n_{1.}, \ldots, n_{J.}) &\propto (\frac{n_{j.}}{\sigma})^{s'_j}. \end{aligned} \tag{94}$$

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.

Finally, we derive the conditional distribution of ρ. We place a Beta(c, d) prior on ρ and have $m_{\cdot\cdot}$ total draws of $w_{jt} \sim \text{Ber}(\rho)$, with $w_{\cdot\cdot}$ successes from these draws. Here, each success represents a table's considered dish being overridden by the house specialty dish. Using these facts, we have

$$\begin{aligned} p(\rho \mid w) &\propto p(w \mid \rho)p(\rho) \\ &\propto \begin{pmatrix} m_{\cdot\cdot} \\ w_{\cdot\cdot} \end{pmatrix} \rho^{w}_{\cdot\cdot}(1-\rho)^{m_{\cdot\cdot}-w_{\cdot\cdot}} \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)} \rho^{c-1}(1-\rho)^{d-1} \\ &\propto \rho^{w_{\cdot\cdot}+c-1}(1-\rho)^{m_{\cdot\cdot}-w_{\cdot\cdot}+d-1} \propto \text{Beta}(w_{\cdot\cdot}+c, m_{\cdot\cdot}-w_{\cdot\cdot}+d). \end{aligned} \tag{95}$$

REFERENCES

[1] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.

[2] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[3] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[4] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[5] M. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *J Amer Stat Assoc*, vol. 90, no. 430, pp. 577–588, 1995.

[6] D. Blackwell and J. MacQueen, "Ferguson distributions via Polya urn schemes," *The Annals of Statistics*, vol. 1, no. 2, pp. 353–355, 1973.

[7] C. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *The Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

[8] H. Ishwaran and M. Zarepour, "Dirichlet prior sieves in finite normal mixtures," *Statistica Sinica*, vol. 12, pp. 941–963, 2002.

[9] E. Sudderth, "Graphical models for visual object recognition and tracking," PhD Thesis, MIT, Cambridge,MA, 2006.

[10] H. Ishwaran and M. Zarepour, "Exact and approximate sum–representations for the Dirichlet process," *Canadian Journal of Statistics*, vol. 30, pp. 269–283, 2002.

[11] ——, "Markov chain Monte Carlo in approximate Dirichlet and beta two–parameter process hierarchical models," *Biometrika*, vol. 87, no. 2, pp. 371–390, 2000.

[12] G. Casella and C. Robert, "Rao-Blackwellisation of sampling schemes," *Biometrika*, vol. 83, no. 1, pp. 81–94, 1996.

[13] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden Markov model," in *NIPS 14*. MIT Press, 2002, pp. 577–584.

[14] A. Rodriguez, D. Dunson, and A. Gelfand, "The nested Dirichlet process," *Institute of Statistics and Decision Sciences, Duke University, Technical Report #06-19.*