Journal of Neuroscience Methods 263 (2016) 36–47

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Computational Neuroscience

A Bayesian nonparametric approach for uncovering rat hippocampal population codes during spatial navigation

Scott W. Linderman[a], Matthew J. Johnson[a,b], Matthew A. Wilson[c], Zhe Chen[d,*]

[a] Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[b] Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA
[c] Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[d] Department of Psychiatry, Department of Neuroscience and Physiology, New York University School of Medicine, New York, NY 10016, USA

HIGHLIGHTS

- Bayesian nonparametric HDP-HMM can efficiently perform model selection and identify model complexity.
- MCMC inference outperforms other inference methods such as variational Bayes.
- The HDP-HMM and MCMC inference can efficiently uncover rat hippocampal population codes during spatial navigation.

ARTICLE INFO

Article history:
Received 6 September 2015
Received in revised form 25 January 2016
Accepted 25 January 2016
Available online 5 February 2016

Keywords:
Hidden Markov model
Stick-breaking process
Hierarchical Dirichlet process
Bayesian nonparametric inference
Variational Bayes
Markov chain Monte Carlo
Gibbs sampler
Hamiltonian Monte Carlo
Population codes

ABSTRACT

Background: Rodent hippocampal population codes represent important spatial information about the environment during navigation. Computational methods have been developed to uncover the neural representation of spatial topology embedded in rodent hippocampal ensemble spike activity.
New method: We extend our previous work and propose a novel Bayesian nonparametric approach to infer rat hippocampal population codes during spatial navigation. To tackle the model selection problem, we leverage a Bayesian nonparametric model. Specifically, we apply a hierarchical Dirichlet process-hidden Markov model (HDP-HMM) using two Bayesian inference methods, one based on Markov chain Monte Carlo (MCMC) and the other based on variational Bayes (VB).
Results: The effectiveness of our Bayesian approaches is demonstrated on recordings from a freely behaving rat navigating in an open field environment.
Comparison with existing methods: The HDP-HMM outperforms the finite-state HMM in both simulated and experimental data. For HPD-HMM, the MCMC-based inference with Hamiltonian Monte Carlo (HMC) hyperparameter sampling is flexible and efficient, and outperforms VB and MCMC approaches with hyperparameters set by empirical Bayes.
Conclusion: The Bayesian nonparametric HDP-HMM method can efficiently perform model selection and identify model parameters, which can used for modeling latent-state neuronal population dynamics.

1. Introduction

A fundamental goal in neuroscience is to understand how populations of neurons represent and transmit information about the external world. The hippocampus is known to encode information relevant to spatial navigation and episodic memory. Spatial representation of the environment is pivotal for navigation in rodents (O'Keefe and Nadel, 1978). One type of spatial representation is a topological map, which contains only relative ordering or connectivity information between spatial locations and is invariant to orientation or deformation. A relevant question of interest is: how can neurons downstream of the hippocampus infer representations of space from hippocampal spike activity without *a priori* place field information (namely, without the measurement of spatial correlates)? Several reports have been dedicated to the mathematical analysis of this problem (Curto and Itskov, 2008; Dabaghian et al., 2012, 2014); however, a data-driven approach for analyzing ensemble hippocampal spike data is needed. This

[*] Corresponding author. Tel.: +1 646 754 4765.
E-mail addresses: slinderman@seas.harvard.edu (S.W. Linderman), mattjj@csail.mit.edu (M.J. Johnson), mwilson@mit.edu (M.A. Wilson), Zhe.Chen3@nyumc.org (Z. Chen).

http://dx.doi.org/10.1016/j.jneumeth.2016.01.022
0165-0270/© 2016 Elsevier B.V. All rights reserved.

paper employs probabilistic modeling and inference methods to uncover the spatial representation (or topological map) based on the ensemble spike activity.

Bayesian statistical modeling is a consistent and principled framework for dealing with uncertainties about the observed data (Scott, 2002). The goal of Bayesian inference is to incorporate prior knowledge and constraints of the problem and to infer the posterior distribution of unobserved variables of interest (Gelman et al., 2013). In recent years, cutting-edge Bayesian methods have become increasingly popular for data analyses in neuroscience, medicine and biology (Mishchenko et al., 2011; Chen et al., 2011; Chen, 2013; Davidson et al., 2009; Kloosterman et al., 2014; Yau et al., 2011). Specifically, thanks to ever-growing computing power, Markov chain Monte Carlo (MCMC) methods have been widely used in Bayesian inference.

In our previous work (Chen et al., 2012a, 2014), we have developed a *parametric Bayesian* approach to uncover the neural representation of spatial topology embedded in rodent hippocampal population codes during spatial navigation. Here we extend the preceding work and consider a *nonparametric Bayesian* approach. The Bayesian nonparametric method brings additional flexibility to the probabilistic model, which allows us to model the complex structure of neural data (Teh and Jordan, 2010; Wood and Black, 2008; Shalchyan and Farina, 2014). Specifically, we leverage the hierarchical Dirichlet process-HMM (HDP-HMM) (Teh et al., 2006), which extends the finite-state hidden Markov model (HMM) with a nonparametric HDP prior, and derive corresponding Bayesian inference algorithms. We consider both deterministic and stochastic approaches for fully Bayesian inference. Based on deterministic approximation, we extend the work of (Chen et al., 2012a; Johnson and Willsky, 2014) and use a variational Bayes (VB) method for approximate Bayesian inference. For MCMC, we adapt the Gibbs sampling approach of (Teh et al., 2006), and integrate it with a Hamiltonian Monte Carlo (HMC) method (Neal, 2010) for hyperparameter inference. To the best of our knowledge, the application of the HDP-HMM to hippocampal ensemble neuronal spike trains and the HMC hyperparameter inference algorithm is novel.

We test the statistical model and inference methods with both simulation data and experimental data. The latter consists of a recording of rat dorsal hippocampal ensemble spike activity during open field navigation. Using a decoding analysis and predictive likelihood, we verify and compare the performance of the proposed Bayesian inference algorithms. We also discuss the results of model selection related to the sample size and the choice of concentration parameter or hyperparameters. Our methods provide an extended tool to analyze rodent hippocampal population codes, which may further empower us to explore important neuroscience questions about neural representation, learning and memory.

2. Methods: modeling and inference

2.1. Basic probabilistic model

In our previous work (Chen et al., 2012a), we used a finite m-state HMM to characterize the population spiking activity from a population of C hippocampal place cells. It was assumed that first, the animal's spatial location during locomotion, modeled as a latent state process, followed a first-order discrete-state Markov chain $\mathcal{S} = S_{1:T} \equiv \{S_t\} \in \{1, \ldots, m\}$, and second, the spike counts of individual place cells at time t, conditional on the hidden state S_t, followed a Poisson probability with their respective tuning curve functions $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}_c\} = \{\{\lambda_{c,i}\}\}$. Essentially, we employed a Markov-driven population Poisson firing model with the following probabilistic models

$$\begin{aligned}
p(\boldsymbol{y}_{1:T}, S_{1:T}|\boldsymbol{\pi}, \boldsymbol{P}, \boldsymbol{\Lambda}) &= p(\mathcal{S}_1|\boldsymbol{\pi})\prod_{t=2}^{T} p(S_t|S_{t-1}, \boldsymbol{P})\prod_{t=1}^{T} p(\boldsymbol{y}_t|S_t, \boldsymbol{\Lambda}),\\
p(S_1|\boldsymbol{\pi}) &= Multinomial(S_1|\boldsymbol{\pi}),\\
p(S_t|S_{t-1}, \boldsymbol{P}) &= Multinomial(S_t|\boldsymbol{P}_{S_{t-1},:}),\\
p(\boldsymbol{y}_t|S_t, \boldsymbol{\Lambda}) &= \prod_{c=1}^{C} Poisson(y_{c,t}|\lambda_{c,S_t}).
\end{aligned} \tag{1}$$

where $\boldsymbol{P} = \{P_{ij}\}$ denotes an m-by-m state transition probability matrix, with P_{ij} representing the transition probability from state i to j (since $\sum_{k=1}^{m} P_{ik} = 1$, each row of matrix $\boldsymbol{P}$ specifies a multinomial likelihood); $y_{c,t}$ denotes the number of spike counts from cell c within the t-th temporal bin (here we assume that the rate is defined in the unit bin size of 250 ms) and $\boldsymbol{y}_{1:T} = \{y_{c,t}\}_{C\times T}$ denotes time series of C-dimensional population response vector; and $Poisson(y_{c,t}|\lambda_{c,i})$ defines a Poisson distribution with the rate parameter $\lambda_{c,i}$ when $S_t = i$. Finally, $\log p(\boldsymbol{y}_{1:T}|\mathcal{S}, \boldsymbol{\theta})$ defines the observed data log likelihood given the hidden state sequence $\mathcal{S}$ and all parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{P}, \boldsymbol{\Lambda}\}$ (where $\boldsymbol{\pi} = \{\pi_i\}$ denotes a probability vector for the initial state S_1).

The hidden variables $\mathcal{S} = \{S_{1:T}\}$ are treated as the missing data, $\boldsymbol{y}_{1:T}$ as the observed (incomplete) data, and their combination $\{S_{1:T}, \boldsymbol{y}_{1:T}\}$ as the complete data.

A Bayesian version of this model introduces prior distributions over the parameters. We use the following prior distributions,

$$\begin{aligned}
\alpha_0 &\sim Gamma(a_{\alpha_0}, 1)\\
\boldsymbol{\pi} &\sim Dir(\alpha_0 \mathbf{1}),\\
\boldsymbol{P}_{i,:} &\sim Dir(\alpha_0 \mathbf{1}),\\
\lambda_{c,i} &\sim Gamma(a_c^0, b_c^0).
\end{aligned} \tag{2}$$

where *Dir* denotes the Dirichlet prior distribution, and $Gamma(a_c^0, b_c^0)$ denotes the gamma prior distribution with shape parameter a_c^0 and scale parameter b_c^0.

2.2. HDP-HMM

Model selection is an important issue for statistical modeling and data analysis. We have previously proposed a *Bayesian deviance information criterion* to select the model size m of HMM (Chen et al., 2012a, 2014). Here we extend the finite-state HMM to an HDP-HMM, a Bayesian nonparametric extension of the HMM that allows for a potentially infinite number of hidden states (Teh et al., 2006; Beal et al., 2002). Namely, the HDP-HMM treats the priors via a stochastic process. Instead of imposing a Dirichlet prior distribution on the rows of the finite state transition matrix $\boldsymbol{P}$, we use a HDP that allows for a countably infinite number of states.

Specifically, we sample a distribution over latent states, G_0, from a Dirichlet process (DP) (Ferguson, 1973) prior, $G_0 \sim \mathrm{DP}(\gamma, H)$, where γ is the concentration parameter and H is the base measure. Moreover, we place a prior distribution over the concentration parameter, $\gamma \sim Gamma(a_\gamma, 1)$. Given the concentration, one may sample from the DP via the "stick-breaking construction" (Sethuraman, 1994). First, sample the stick-breaking weights, $\boldsymbol{\beta}$,

$$\tilde{\beta}_i \sim Beta(1, \gamma), \qquad \beta_i = \tilde{\beta}_i \prod_{j=1}^{i-1} (1 - \beta_j) \tag{3}$$

where $\beta_1 = \tilde{\beta}_1$, $\sum_{i=1}^{\infty} \beta_i = 1$, and $Beta(a, b)$ defines a beta distribution with two positive parameters a and b.

The stick-breaking construction of (3) is sometimes denoted as $\boldsymbol{\beta} \sim \texttt{GEM}(\gamma)$, after Griffiths, Engen, and McCloskey (Ewens, 1990). The name "stick-breaking" comes from the interpretation of β_i as the length of the piece of a unit-length stick assigned to the i-th value. After the first $i-1$ values having their portions assigned, the length of the remainder of the stick is broken according to a sample $\tilde{\pi}_i$ from a beta distribution, and $\tilde{\beta}_i$ indicates the portion of the remainder to be assigned to the i-th value. Therefore, the stick-breaking process $\texttt{GEM}(\gamma)$ also defines a DP—the smaller γ, the less (in a statistical sense) of the stick will be left for subsequent values.

After sampling $\boldsymbol{\beta}$, we next sample the latent state variables, in this case $\boldsymbol{\lambda}_c$, from the base measure H. Our draw from the DP(γ, H) prior is then given by

$$G_0 = \sum_{j=1}^{\infty} \beta_j \delta_{\boldsymbol{\lambda}_c^{(j)}}. \tag{4}$$

Thus, the stick breaking construction makes clear that draws from a Dirichlet process distribution are discrete with probability one.

Given a countably infinite set of shared states, we may then sample the rows of the transition matrix, $\boldsymbol{P}_{i,:} \sim \mathrm{DP}(\alpha_0, \boldsymbol{\beta})$. We place the same prior over π. The base measure in this case is $\boldsymbol{\beta}$, a countably infinite vector of stick-breaking weights, that serves as the mean of the DP prior over the rows of $\boldsymbol{P}$. The concentration parameter, α_0, governs how concentrated the rows are about the mean. Since the base measure $\boldsymbol{\beta}$ is discrete, each row of $\boldsymbol{P}$ will be able to "see" the same set of states. By contrast, if we remove the HDP prior and treat each row of P as an independent draw from a DP with base measure H, each row would see a disjoint set of states with probability one. In other words, the hierarchical prior is required to provide a discrete (but countably infinite) set of latent states for the HMM.

2.3. Overdispersed Poisson model

An interesting consequence of this Bayesian model is that it naturally leads to a distribution of spike counts that is overdispersed relative to simple Poisson model, a feature that has been observed in neural recordings (Goris et al., 2014). Recent work has explored the negative binomial (NB) distribution as an alternative to Poisson model, since its two parameters allow for Fano factors greater than one. The NB distribution can also be seen as a continuous mixture of Poisson distributions (i.e., a compound probability distribution) where the mixing distribution of the Poisson rate is a gamma distribution (Gelman et al., 2013). In other words, the NB distribution is viewed as a gamma-Poisson (mixture distribution): a $Poisson(\lambda)$ distribution whose rate λ is itself a gamma random variable. In our case, the gamma prior over firing rates leads to a negative binomial marginal distribution over $y_{c,t}$.

Though the marginal spike count at a particular time t may be marginally distributed according to a negative binomial distribution, it is not necessarily true that a sequence of time bins, $\boldsymbol{y}_{1:T}$, will be i.i.d. negative binomials. This arises from the correlations induced by the state transition matrix. Instead, $\boldsymbol{y}_t$ will follow a finite mixture of Poisson distributions, with one component for each latent state. The mixture will be weighted by the marginal probability of the corresponding latent state. However, as the number of visited states grows, and the marginal probability of latent states becomes more uniform, the resulting marginal distribution over the spike count sequences inherits the over dispersed nature of the NB distribution. This is particularly true of an HDP-HMM with a high concentration.

2.4. Markov chain Monte Carlo (MCMC) inference

Several MCMC-based inference methods have been developed for the HDP-HMM (Beal et al., 2002; Teh et al., 2006; Van Gael et al., 2008). Some of these previous works use a collapsed Gibbs sampler in which the transition matrix P and the observation parameters $\boldsymbol{\Lambda}$ are integrated out (Teh et al., 2006; Van Gael et al., 2008). In this work, however, we use a "weak limit" approximation in which the DP prior is approximated with a symmetric Dirichlet prior. Specifically, we let

$$\begin{aligned}
\gamma &\sim Gamma(a_{\gamma}, 1)\\
\alpha_0 &\sim Gamma(a_{\alpha_0}, 1)\\
\boldsymbol{\beta}|\gamma &\sim Dir(\gamma/M, \ldots, \gamma/M),\\
\boldsymbol{\pi}|\alpha_0, \boldsymbol{\beta} &\sim Dir(\alpha_0\beta_1, \ldots, \alpha_0\beta_M),\\
\boldsymbol{P}_{i,:}|\alpha_0, \boldsymbol{\beta} &\sim Dir(\alpha_0\beta_1, \ldots, \alpha_0\beta_M).
\end{aligned} \tag{5}$$

where M denotes a truncation level for approximating the infinity (which is different from m in the finite-state setting). It can be shown that this prior will weakly converge to the DP prior as the dimensionality of the Dirichlet distribution approaches infinity (Johnson and Willsky, 2014; Ishwaran and Zarepour, 2002). With this approximation we can capitalize on forward-backward sampling algorithms to jointly update the latent states $\mathcal{S}$.

Previous work has typically been presented with Gaussian or multinomial likelihood models, with the acknowledgement that the same methods work with any exponential family likelihood when the base measure H is a conjugate prior. Here we present the Gibbs sampling algorithm of Teh et al. (2006) for the HDP-HMM applied to the special case of independent Poisson observations, and we derive Hamiltonian Monte Carlo (HMC) transitions to sample the cell-specific hyperparameters of the firing rate priors.

We begin by defining Gibbs updates for the neuronal firing rates $\boldsymbol{\Lambda}$. Since we are using gamma priors with independent Poisson observations, the model is fully conjugate and simple Gibbs updates suffice. Therefore, we have

$$\lambda_{c,i}|\boldsymbol{y}, \mathcal{S} \sim Gamma\left(\alpha_c^0 + \sum_{t=1}^{T} y_{c,t}\mathbb{I}[S_t = i], \quad \beta_c^0 + \sum_{t=1}^{T}\mathbb{I}[S_t = i]\right). \tag{6}$$

Under the weak limit approximation, the priors on $\boldsymbol{P}_{i,:}$ and $\boldsymbol{\pi}$ reduce to Dirichlet distributions, which are also conjugate with the finite HMM. Hence we can derive conjugate Gibbs updates for these parameters as well. They take the form:

$$\begin{aligned}
\boldsymbol{\pi}|\alpha_0, \boldsymbol{\beta} &\sim Dir\left(\alpha_0\boldsymbol{\beta} + 1_{S_1}\right),\\
\boldsymbol{P}_{i,:}|\alpha_0, \boldsymbol{\beta} &\sim Dir\left(\alpha_0\boldsymbol{\beta} + \boldsymbol{n}_i\right),\\
n_{i,j} &= \sum_{t=1}^{T-1}\mathbb{I}[S_t = i, S_{t+1} = j],
\end{aligned} \tag{7}$$

where 1_j is a unit vector with a one in the j-th entry.

Conditioned upon the firing rates, the initial state distribution, and the transition matrix, we can jointly update the latent states of the HDP-HMM using a *forward filtering, backward sampling* algorithm. Details of this well-known algorithm can be found in (Johnson, 2014); the intuition is that in the backward pass of the algorithm, we have a conditional distribution over S_t given $S_{t+1:T}$. We can iteratively sample from these distributions as we go backward in time to generate a full sample from $p(\mathcal{S}|\boldsymbol{P}, \boldsymbol{\pi}, \boldsymbol{\Lambda})$. Jointly sampling these latent states allows us to avoid issues with mixing when individually sampling states that are highly correlated with one another.

Finally, the Dirichlet parameters $\boldsymbol{\beta}$ and the concentration parameters α_0 and γ can be updated as in (Teh et al., 2006). A single iteration of the final algorithm consists of an update for each parameter of the model. The aforementioned updates are based upon previous work; one novel direction that we explore in this work is the sampling of the hyperparameters of the gamma firing rate priors.

2.4.1. Setting firing rate hyperparameters

We consider three approaches to setting the hyperparameters of the gamma priors for Poisson firing rates, namely, $\{\alpha_c^0, \beta_c^0\}$ for cell c. Note that these parameters are distinct from the parameters of the DP prior.

- In the first approach, we estimate these parameters using an empirical Bayesian (EB) procedure, that is, by maximizing the marginal likelihood of the spike counts. For each cell, this may be easily done using standard maximum likelihood estimation for the negative binomial model. In practice, we found that without regularization this approach leads to extreme values of the hyperparameters.
- Our second approach samples these hyperparameters using Hamiltonian Monte Carlo (HMC) (Neal, 2010). We note that for fixed values of the "shape" parameter α_c^0, the conditional distribution of the "scale" parameter, β_c^0 is conjugate with a gamma prior distribution. However, setting the shape parameter *a priori* is challenging because it can have a strong influence on the firing rate distribution. HMC allows us to jointly sample both the shape and the scale parameters simultaneously. To implement HMC, we must have access to both the log probability of the parameters as well as its gradient. Since both parameters are restricted to be positive, we instead reparameterize the problem in terms of their logs. For cell c, the conditional log probability equal to,

$$\begin{aligned}\mathcal{L} &= \log p(\log\alpha_c^0, \log\beta_c^0 | \boldsymbol{\Lambda}_{c,:}) \\ &= \sum_{i=1}^{m} \log p(\lambda_{c,i}|\alpha_c^0, \beta_c^0) + \text{const.} \\ &= \sum_{i=1}^{m} \alpha_c^0 \log\beta_c^0 - \log\Gamma(\alpha_c^0) + (\alpha_c^0 - 1)\log\lambda_{c,i} - \beta_c^0\lambda_{c,i}.\end{aligned} \tag{8}$$

  Taking gradients with respect to both parameters yields,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\log\alpha_c^0} &= \sum_{i=1}^{m}\left[\log\beta_c^0 - \Psi(\alpha_c^0) + \log\lambda_{c,i}\right] \times \alpha_c^0, \\ \frac{\partial\mathcal{L}}{\partial\log\beta_c^0} &= \sum_{i=1}^{m}\left[\frac{\alpha_c^0}{\beta_c^0} - \lambda_{c,i}\right] \times \beta_c^0.\end{aligned} \tag{9}$$

  The HMC algorithm uses these gradients to inform a stochastic walk over the posterior distribution. With knowledge of the gradients, HMC can sometimes make large updates to parameters, especially in cases where the parameters are highly correlated under the posterior.
- In the final approach, we fix the shape hyperparameter, α_c^0, and infer the scale hyperparameter, β_c^0. We place a gamma prior on the scale, $\beta_c^0 \sim Gamma(\mu, \nu)$. Given α_c^0, the conditional distribution of the scale hyperparameter is

$$\beta_c^0|\alpha_c^0, \{\lambda_{c,i}\}, \mathcal{S} \sim Gamma(\mu + \sum_{i=1}^{m}\mathbb{I}[n_i > 0]\cdot\alpha_c^0,\ \nu + \sum_{i=1}^{m}\mathbb{I}[n_i > 0]\cdot\lambda_{c,i}) \tag{10}$$

$$n_i = \sum_{t=1}^{T}\mathbb{I}[S_t = i].$$

In the following experiments, we set the shape parameter to $\alpha_c^0 = 1$, and we set the scale prior parameters to $\mu = 1$ and $\nu = 1$. This is equivalent to an exponential prior on rates, $\lambda_{c,i} \sim Exp(\beta_c^0)$, and an exponential prior on the scale $\beta_c^0 \sim Exp(1)$. One could perform cross validation over the shape parameter, but the exponential prior is a rather weak assumption that enables fully-Bayesian inference.

2.4.2. Predictive log likelihood

Upon completing the parameter and hyperparameter inference, we evaluate the performance of our algorithm in terms of its predictive log likelihood on held out test data. We approximate the predictive log likelihood with samples from the posterior distribution generated by our MCMC algorithm. That is,

$$\begin{aligned}\log p(\boldsymbol{y}_{test}|\boldsymbol{y}_{1:T}) &= \log\sum_{\mathcal{S}_{test}}\int_{\boldsymbol{\Theta}} p\left(\boldsymbol{y}_{test}, \mathcal{S}_{test}|\boldsymbol{\theta}\right)\ p\left(\boldsymbol{\theta}|\boldsymbol{y}_{train}\right)\mathrm{d}\boldsymbol{\theta}, \\ &\approx \log\frac{1}{N}\sum_{n=1}^{N}\sum_{\mathcal{S}_{test}} p\left(\boldsymbol{y}_{test}, \mathcal{S}_{test}|\boldsymbol{\theta}_n\right),\end{aligned} \tag{11}$$

where $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{P}, \boldsymbol{\pi})$ and $\{\boldsymbol{\theta}_n\}_{n=1}^{N} \sim p(\boldsymbol{\theta}|\boldsymbol{y}_{train})$. The summation over latent state sequences for the test data is performed with the message-passing algorithm for HMMs.

2.5. Variational Bayes (VB) inference

We build upon our previous work (Chen et al., 2012a, 2014; Johnson and Willsky, 2014) to develop a variational inference algorithm for fitting the HDP-HMM to hippocampal spike trains. Our objective is to approximate the posterior distribution of the HDP-HMM with a distribution from a more tractable family. As usual, we choose a factorized approximation that allows for tractable optimization of the parameters of the variational model. Specifically, we let,

$$p(\mathcal{S}, \boldsymbol{\Lambda}, \boldsymbol{P}, \boldsymbol{\pi}, \boldsymbol{\beta}|\boldsymbol{y}_{1:T}) \approx q(\mathcal{S})q(\boldsymbol{\Lambda})q(\boldsymbol{P})q(\boldsymbol{\pi})q(\boldsymbol{\beta}). \tag{12}$$

Since the independent Poisson observations are conjugate with the gamma firing rate prior distributions, choosing a set of independent gamma distributions for $q(\boldsymbol{\Lambda})$ allows for simple variational updates.

$$\begin{aligned}q(\boldsymbol{\Lambda}) &= \prod_{i=1}^{M}\prod_{c=1}^{C} Gamma\left(\tilde{\alpha}_{c,i},\ \tilde{\beta}_{c,i}\right), \\ \tilde{\alpha}_{c,i} &\leftarrow \alpha_c^0 + \sum_{t=1}^{T} y_{c,t}\mathbb{E}_q[\mathbb{I}[S_t = i]], \\ \tilde{\beta}_{c,i} &\leftarrow \beta_c^0 + \sum_{t=1}^{T}\mathbb{E}_q[\mathbb{I}[S_t = i]].\end{aligned} \tag{13}$$

Following (Johnson and Willsky, 2014), we use a "direct assignment" truncation for the HDP (Bryant and Sudderth, 2012; Liang et al., 2007). In this scheme, a truncation level M is chosen *a priori* and $q(\mathcal{S})$ is limited to support only states $S_t \in \{1, \ldots, M\}$. The advantage of this approximation is that conjugacy is retained with $\boldsymbol{\Lambda}$, $\boldsymbol{P}$, and $\boldsymbol{\pi}$, and the variational approximation $q(\mathcal{S})$ reduces to

$$\begin{aligned}q(\mathcal{S}) &= HMM(\tilde{\boldsymbol{P}}, \tilde{\boldsymbol{\pi}}, \tilde{\boldsymbol{\Lambda}}), \\ \tilde{\boldsymbol{P}} &= \exp\left\{\mathbb{E}_q[\ln\boldsymbol{P}]\right\}, \\ \tilde{\boldsymbol{\pi}} &= \exp\left\{\mathbb{E}_q[\ln\boldsymbol{\pi}]\right\}, \\ \tilde{\boldsymbol{\Lambda}} &= \exp\left\{\mathbb{E}_q[\ln p(\boldsymbol{y}|\boldsymbol{\Lambda})]\right\}.\end{aligned} \tag{14}$$

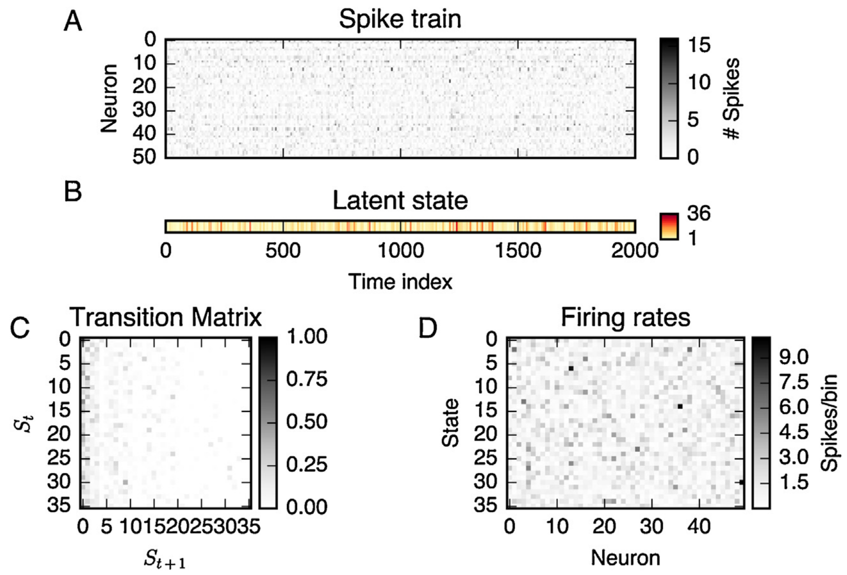

Fig. 1. An example of a synthetic dataset drawn from an HDP-HMM. (A) Simulated population spike trains or spike counts. (B) Inferred latent state sequence. (C) Inferred state transition matrix $\boldsymbol{P}$. (D) Inferred neuronal firing rate vectors $\boldsymbol{\lambda}_i$ specific to each state.

Expectations $\mathbb{E}_q[S_t = i]$ can then be computed using standard the message-passing algorithm for the HMM.

With the direct assignment truncation, the variational factors for $\boldsymbol{P}_{i,:}$ and $\boldsymbol{\pi}$ take on Dirichlet priors. Unlike in the finite-state HMM, however, these Dirichlet priors are now over $M+1$ dimensions since the final dimension accounts for all states $i > M$. Under the HDP prior we had $P_{i,:} \sim \texttt{DP}(\alpha_0 \cdot \boldsymbol{\beta})$, and under the truncation the DP parameter becomes $\alpha_0 \cdot \boldsymbol{\beta}_{1:M+1}$. Again, leveraging the conjugacy of the model, we arrive at the following variational updates:

$$
\begin{aligned}
q(\boldsymbol{P}) &= \prod_{i=1}^{M} \mathrm{Dir}(\tilde{\boldsymbol{\alpha}}_P^{(i)}), \\
(\tilde{\boldsymbol{\alpha}}_P^{(i)})_j &\leftarrow \alpha_0 \beta_j + \mathbb{E}_q[\mathbb{I}[S_t = i] \cdot \mathbb{I}[S_{t+1} = j]].
\end{aligned}
\tag{15}
$$

We use an analogous update for $\boldsymbol{\pi}$.

The principal drawback of the direct assignment truncation is that the prior for $\boldsymbol{\beta}$ is no longer conjugate. This could be avoided with the fully conjugate approach of Hoffman et al. (2013), however, this results in extra bookkeeping and the duplication of states. Instead, following (Johnson and Willsky, 2014; Bryant and Sudderth, 2012; Liang et al., 2007), we use a point estimate for this parameter by setting $q(\boldsymbol{\beta}) = \delta_{\boldsymbol{\beta}^*}$ and use gradient ascent with backtracking line search to update this parameter during inference.

There are a number of hyperparameters to set for the VB approach. The hyperparameters α_c^0 and β_c^0 of gamma prior on firing rates can be set with empirical Bayes, as above. We resort to cross validation to set the Dirichlet parameter α_0 and the GEM parameter γ.

Finally, in order to compute the predictive log likelihood on held out test data, we draw multiple samples $\{(\boldsymbol{\Lambda}, \mathcal{S}, \boldsymbol{P}, \boldsymbol{\pi}, \boldsymbol{\beta})_n\}_{n=1}^N$ from the variational posterior and approximate the predictive log likelihood as

$$
\begin{aligned}
\ln p(\boldsymbol{y}_{test} | \boldsymbol{y}_{1:T}) &\approx \ln \mathbb{E}_q \left[p(\boldsymbol{y}_{test} | \mathcal{S}, \boldsymbol{\Lambda}, \boldsymbol{P}, \boldsymbol{\pi}, \boldsymbol{\beta}) \right] \\
&\approx \ln \frac{1}{N} \sum_{n=1}^{N} p(\boldsymbol{y}_{test} | (\mathcal{S}, \boldsymbol{\Lambda}, \boldsymbol{P}, \boldsymbol{\pi}, \boldsymbol{\beta})_n).
\end{aligned}
\tag{16}
$$

3. Results

The inference algorithms were implemented based upon the PyHSMM framework of Johnson (2014). The codebase was written in Python with C offloads for the message passing algorithms. We have extended the codebase to perform hyperparameter inference using the methods described above, and expanded it to tailor to neural spike train analysis. Our source code is publicly available (https://github.com/slinderman/pyhsmm_spiketrains).

3.1. Simulation data

Setup. First, we simulate synthetic spike count data using an HDP-HMM with $C = 50$ neurons, $T = 2000$ time bins, and Dirichlet concentration parameters $\alpha_0 = 12.0$ and $\gamma = 12.0$. These configuration yield state sequences that tend to visit 30-45 states. All of neuronal firing rate parameters are drawn from a gamma distribution: $Gamma(\alpha_c^0 = 1, \beta_c^0 = 1)$ (with mean 1.0 and standard deviation 1.0). An example of one such synthetic dataset is shown in Fig. 1. The states have been ordered according to their occupancy (i.e., how many times they are visited during the simulation), such that the columns of the transition matrix exhibit a decrease in probability as the incoming state number, S_{t+1}, increases. This is a characteristic of the HDP-HMM, indicating the tendency of the model to reuse states with high occupancy.

We compare six combinations of model, inference algorithm, and hyperparameter selection approaches: (i) HMM with the correct number of states, fit by Gibbs sampling with fixed $\alpha_c^0 = 1$; (ii) HMM with the correct number of states, fit by VB with hyperparameters set by empirical Bayes; (iii) HDP-HMM fit by Gibbs sampling with fixed $\alpha_c^0 = 1$; (iv) HDP-HMM fit by Gibbs sampling and HMC for hyperparameter updates; (v) HDP-HMM fit by MCMC with hyperparameters set by empirical Bayes; and (vi) HDP-HMM fit by VB with hyperparameters set by empirical Bayes. For the MCMC methods, we set gamma priors over the concentration parameters (α_0 and γ); for the VB methods, we set α_0 and γ to their true values. Alternatively, they can be selected by cross validation. We set both the weak limit approximation for MCMC and the direct assignment truncation level for VB to $M = 100$.

We collect 5000 samples from the MCMC algorithms and use the last 2000 samples for computing predictive log likelihoods.

Table 1
Comparison of Hamming error (see Eq. (18)) computed from the same nine simulated data sets as above. The VB inference methods tend to overestimate the number of states and therefore have much larger Hamming error.

Dataset	1	2	3	4	5
HMM (Gibbs)	9	401	13	24	615
HMM (VB)	166	290	295	123	124
HDP-HMM (Gibbs)	2	**3**	5	**1**	6
HDP-HMM (HMC)	3	4	3	2	**4**
HDP-HMM (EB)	**1**	**3**	**2**	3	12
HDP-HMM (VB)	432	586	340	264	675

Note: The minimum Hamming error in each column is marked in bold font.

Table 2
Comparison of predictive log likelihood computed from 5 simulated data sets, measured in bits/spike improvement over a baseline of independent, homogeneous Poisson processes (the best result in each data set is marked in bold font).

Dataset	1	2	3	4	5
HMM (Gibbs)	0.315	0.300	0.312	0.310	0.250
HMM (VB)	0.298	0.290	0.313	0.306	0.252
HDP-HMM (Gibbs)	**0.323**	**0.307**	**0.321**	**0.318**	**0.259**
HDP-HMM (HMC)	**0.323**	0.306	0.320	**0.318**	**0.259**
HDP-HMM (EB)	0.322	0.306	**0.321**	**0.318**	**0.259**
HDP-HMM (VB)	0.312	0.291	0.309	0.305	0.244

For visualization, we use the final sample to extract the transition matrix and the firing rates. The number of samples and the amount of burn-in iterations were chosen by examining the log probability and parameter traces for convergence. It is found that the MCMC algorithm converges within hundreds of iterations. For further convergence diagnosis of a single Gibbs chain, one may use the autocorrelation tools suggested in (Raftery and Lewis, 1992; Cowles and Carlin, 1996).

We run the VB algorithm for 200 steps to guarantee convergence of the variational lower bound. Again, this is assessed by examining the variational lower bound and is found to converge to a local maxima within tens of iterations.

Assessment. We use two criteria for result assessment with simulation data. The first criterion is based on the Hamming error between the true and inferred state sequences. To compute this, we first relabel the inferred states in order to maximize overlap with the true states. Let $\mathcal{S}$ be the true state sequence and $\mathcal{S}'$ be the inferred state sequence. We define the overlap matrix $O \in \mathbb{N}^{M\times M}$ whose entries $O_{i,j}$ is the number of times the true state is i and the inferred state is j:

$$O_{i,j} = \sum_{t=1}^{T} \mathbb{I}[S_t = i]\,\mathbb{I}[S'_t = j]. \tag{17}$$

We use the Hungarian method (Kuhn, 1955) to find a relabeling of the inferred states that maximizes overlap, and then we measure the Hamming error between the true state sequence $\mathcal{S}$, and the relabeled sequence of inferred states, $\widetilde{\mathcal{S}}'$:

$$\mathrm{err}(\mathcal{S}, \widetilde{\mathcal{S}}') = \sum_{t=1}^{T} \mathbb{I}[S_t \neq \widetilde{S}'_t]. \tag{18}$$

Table 1 summarizes the Hamming error for all six models on five synthetic datasets. We see that the HDP-HMM fit via Gibbs sampling with firing rate hyperparameters set via empirical Bayes outperforms the other models and inference algorithms on three of five datasets, but the HDP-HMM with hyperparameter HMC sampling are very comparable. By contrast, when the models are fit with VB inference, the inferred state sequences tend to use more than the true number of states, which results in very poor Hamming error. Similarly, the HMM fit via Gibbs sampling does not factor in the penalty on additional states and instead tends to use all states equally, resulting in large Hamming error.

The second criterion is the model's predictive log likelihood (unit: bits/spike) on a held out sequence of $T_{test} = 1000$ time steps. We compare the predictive log likelihood to that of a set of independent Poisson processes. Their rates and the corresponding predictive log likelihood are given by,

$$\widehat{\lambda}_c = \frac{1}{T_{train}} \sum_{t=1}^{T_{train}} y_{c,t}, \tag{19}$$

$$\log p(\boldsymbol{y}_{test} | \boldsymbol{y}_{train}) = \sum_{c=1}^{C} \left[-T_{test}\widehat{\lambda}_c + \sum_{t=1}^{T_{test}} y_{c,t} \log \widehat{\lambda}_c \right]. \tag{20}$$

The improvement obtained by a model is measured in bits, and is normalized by the number of spikes in the test dataset in order to obtain comparable units for each of the test datasets.

Table 2 summarizes the predictive log likelihood comparison. For all five datasets, the HDP-HMM fit via Gibbs sampling with fixed α_c^0 performs best, though in general the increase over fitting the HDP-HMM when using HMC or EB for hyperparameter selection is small. By contrast, the improvement compared to fitting with VB inference or using a parametric HMM is quite significant.

Though computation cost is often a major factor with Bayesian inference, with the optimized PyHSMM package, the models can be fit to the synthetic data within less minutes on an Apple MacBook Air. The runtime necessarily grows the number of neurons and the truncation limit on the number of latent states. As the model complexity grows, we must also run our MCMC algorithm for more iterations, which often motivates the use of variational inference algorithms instead. Given our optimized implementation and the performance improvements yielded by MCMC, we opted for a fully-Bayesian approach using MCMC with HMC for hyperparameter sampling in our subsequent experiments.

Fig. 2 shows example traces from the MCMC combined with HMC algorithm for the HDP-HMM running on synthetic dataset 1. This is the same data from which Fig. 1 is generated. The first 5 Markov chain iterations have been omitted to highlight the variation in the latter samples (the first few iterations rapidly move away from the initial conditions). We see that the log likelihood of the data rapidly converges to nearly that of the true model (horizontal dotted line), and the number of states quickly converges to around $m = 35$. Note that the nuisance parameters α_0 and γ do not converge to the true values — this is due to the fact that the solution is insensitive to these parameters or the presence of local optima. However, even the concentration parameters are different from the true values, they are still consistent with the inferred state transition matrix.

Sensitivity of the number of latent states. To test the sensitivity of the number of inferred states to changes in the data, we vary a number of parameters and plotted the number of inferred states in Fig. 3. In all cases, we use synthetic dataset 1, shown in Fig. 1, and HDP-HMMs fit via Gibbs sampling with fixed α_c^0. First, we vary the number of neurons C, and find that the number of inferred states was relatively stable around the true number of states ($m = 35$). By contrast, as we increase the recording length T, the number of inferred states increases as well. This is because the true underlying data actually does visit more states as we simulate it for longer time. In general, we expect the number of inferred states to grow with the complexity of the data. Next, we vary the scale of the firing rate by multiplying the true model's firing rate by a factor of 0.1, 0.5, 1.0, 2.0, or 10.0, and sampling a new spike count. When the firing rates are very low, most bins do not contain any spikes, and hence it is not possible to resolve as many states. By contrast, when the rate is

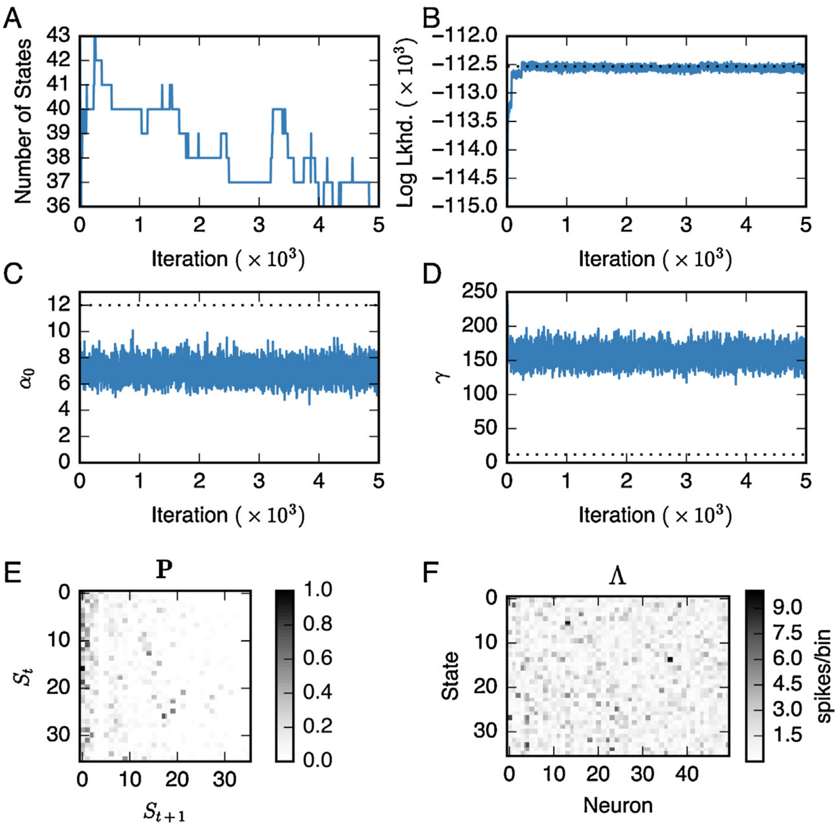

Fig. 2. MCMC state trajectories for an HDP-HMM fit to the synthetic dataset shown in Fig. 1. True values are shown by the dotted black lines. The first five iterations of the Markov chain are omitted since they differ greatly from the final states. The chain quickly converges to nearly the correct number of states (A) and achieves close to the true log likelihood (B). (C) and (D) The chain trajectories of hyperparameters α_0 and γ. (E) and (F) Inferred state transition matrix $\boldsymbol{P}$ and neuronal firing matrix $\boldsymbol{\Lambda} = \{\lambda_{c,i}\}$ drawn from the last iteration.

increased, the number of inferred states is slightly lower than the true number, which is likely the result of a slight mismatch with the prior on the firing rate scale (parameters μ and ν in Section 2.4.1). Finally, we considerate the effect of time bin size by scaling up the bin sizes by factors of 2 through 10. For example, when scaling by a factor of 2, we add the spike counts in each pair of adjacent bins. This has a similar effect to decreasing the recording length by a factor of 2, and hence we see the number of inferred states decrease with bin size.

3.2. Rat hippocampal neuronal ensemble data

Next, we apply the proposed methods to experimental data of the rat hippocampus. Experiments were conducted under the supervision of the Massachusetts Institute of Technology (MIT) Committee on Animal Care and followed the NIH guidelines. The micro-drive arrays containing multiple tetrodes were implanted above the right dorsal hippocampus of male Long-Evans rats. The tetrodes were slowly lowered into the brain reaching the cell layer

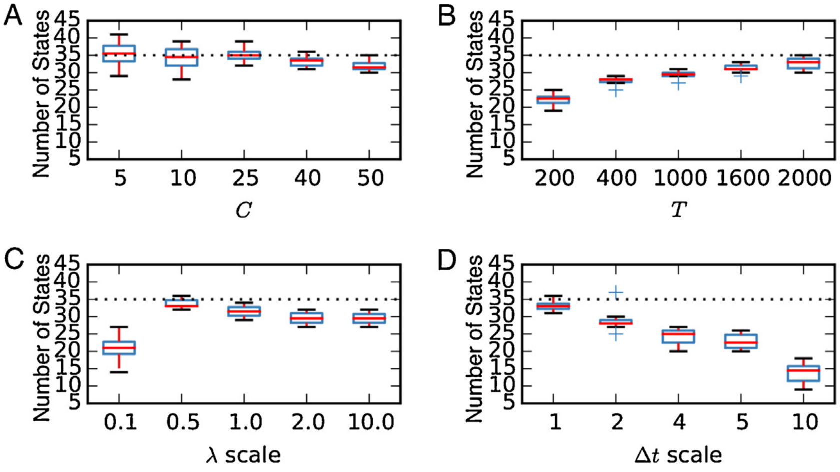

Fig. 3. In a synthetic data experiment, we generated a spike train for a population of $C_{\text{true}} = 50$ cells and $T_{\text{true}} = 2000$ time bins. Then we varied the number of neurons C, recording duration T, scale of the firing rate λ, temporal bin size Δt, and inferred the number of inferred latent states from the data. Horizontal dashed lines indicate the ground truth. Box plots are obtained from 10 independent Monte Carlo chains, and each chain was run for 1000 iterations; the number of states in the last iteration is used.

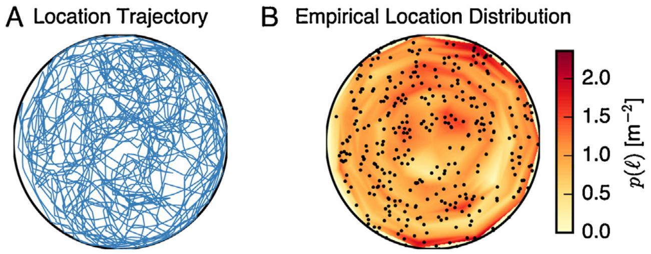

Fig. 4. One rat's behavioral trajectory (A) and spatial occupancy (B) in the open field environment.

Table 3
A comparison of HMMs, HDP-HMMs, and inference algorithms on the rat hippocampal data. Performance is measured in predictive log likelihood and mean decoding error on two minutes of held out test data (the best result is marked in bold font).

	Pred. log likelihood (bits/spike)	Decoding error (cm)
HMM (m = 25)	0.712	10.85 ± 6.43
HMM (m = 45)	0.706	10.71 ± 6.67
HMM (m = 65)	0.717	11.01 ± 6.93
HDP-HMM (Gibbs)	**0.722**	**9.56** ± **5.31**
HDP-HMM (HMC)	0.646	9.96 ± 6.05
HDP-HMM (EB)	0.579	10.81 ± 6.78
HDP-HMM (VB)	0.602	10.93 ± 6.24

of CA1 two to four weeks following the date of surgery. Recorded spikes were manually clustered and sorted to obtain single units using a custom software (XClust, M.A.W.).

For demonstration purpose, an ensemble spike train recording of C = 47 putative pyramidal neurons was collected from a single rat for a duration of 9.8 minutes. Once stable hippocampal units were obtained, the rat was allowed to freely forage in an approximately circular open field environment (radius: ~60 cm). We bin the ensemble spike activity with a bin size of 250 ms and obtain the population vector $\boldsymbol{y}_t$ in time. To identify the period of rodent locomotion during spatial navigation, we use a velocity threshold (>10 cm/s) to select the RUN epochs and merge them together. One animal's RUN trajectory and spatial occupancy are shown in Fig. 4A and B, respectively. The empirical probability of a location, $p(\ell)$, is determined by dividing the arena into 220 bins of equal area (11 angular bins and 20 radial bins) and counting the fraction of time points in which the rat is in the corresponding bin.

In experimental data analysis, we focus on nonparametric Bayesian inference for HDP-HMM. For all methods, we increase the truncation level to a large value of M = 100. To discover the model order of the variational solutions, we use the number of states visited by the most likely state sequence under the variational posterior. The MCMC algorithms yield samples of state sequences from which the model order can be directly counted.

We perform a quantitative comparison between HMMs, HDP-HMMs, inference algorithms, and hyperparameter setting algorithms, where performance is measured in terms of both predictive log likelihood and decoding error. For both metrics, we train the models on the first 7.8 minutes of data and test on the final two minutes of data for prediction. The results are summarized in Table 3. We find that the HDP-HMM fit by Gibbs sampling with fixed firing rate scale ($\alpha_c^0 = 1$) again outperforms the competing models in both measures.

For the purpose of result assessment, we plot the state-space or state-location map (Fig. 5A), which shows the mean value of the spatial position that each state represented. The size of the black dot is proportional to the occupancy of the state. To compute an "empirical" distribution over locations for a given state, we first compute the posterior distribution over latent states with our inference algorithms. This gives us a set of probabilities $\Pr(S_t = i)$ for all time bins t and states i. Then we compute the average location for each state i by weighting the animal's location (x_t, y_t) by the probability that the animal was in state i at time t. Summing over time yields a weighted set of locations, which we then bin into equal-area arcs and normalize to get an empirical distribution over locations for each state i.

The empirical location distributions for the top three states as measured by occupancy are shown in Fig. 5B). In Fig. 5C, we show the estimated animal's spatial trajectories in black, along with the reconstructed location in from the HDP-HMM with Gibbs sampling in blue. To reconstruct the position, we use the mean of each latent state's location distribution weighted

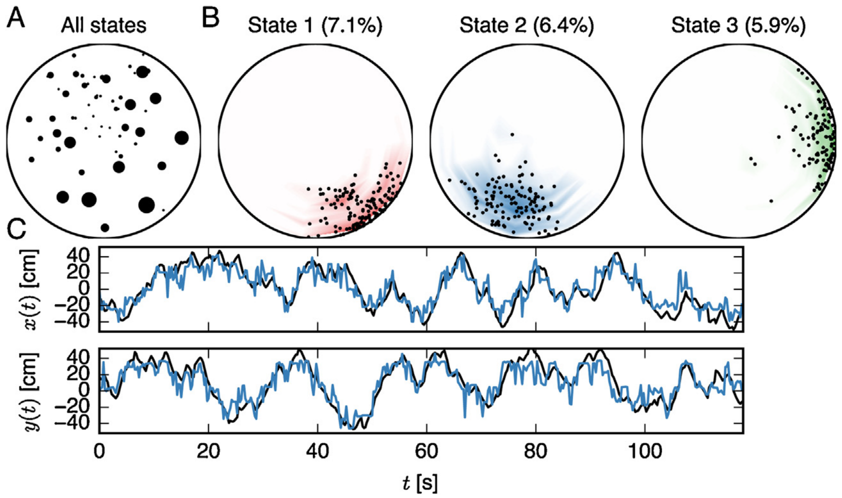

Fig. 5. Estimation result from HDP-HMM (Gibbs) for the rat hippocampal ensemble spike data. (A) Estimated state space map, where the mean value of the spatial position for each latent state is shown by a black dot. The size of the dot is proportional to the occupancy of the state. (B) Probability distributions over location corresponding to the top three latent states, measured by state occupancy. The small black dots indicate the location of the animal while in that state, and are used to compute the empirical distribution over location indicated by colored shading. (C) The true (black) and reconstructed (blue) trajectories shown in Cartesian coordinate. For each time bin, we use the mean location of the latent states to determine an estimate of the animal's location. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

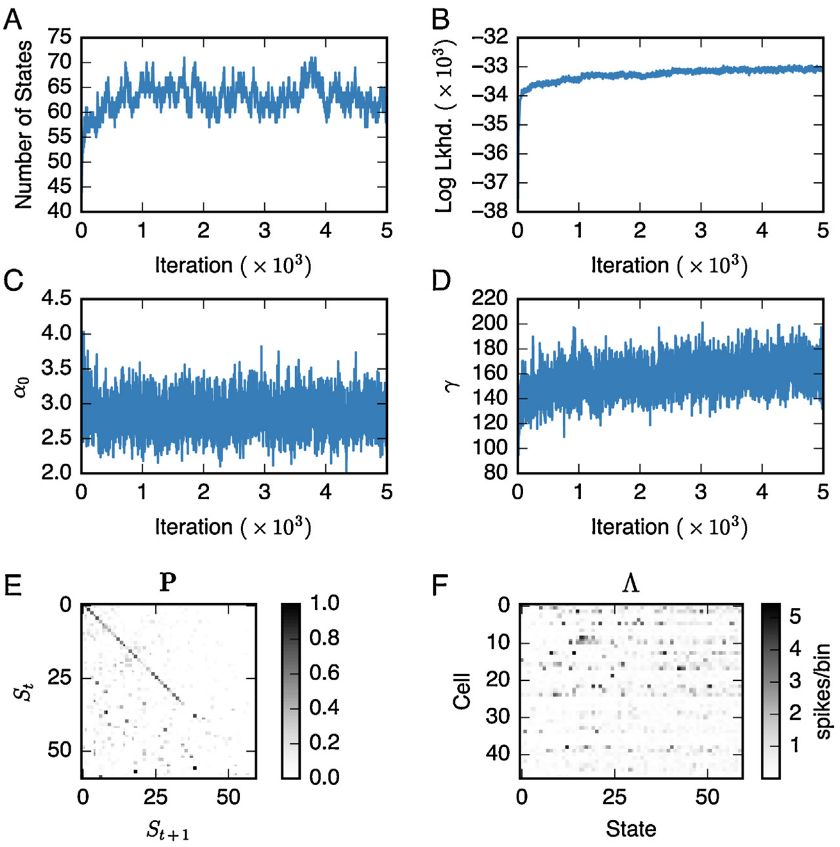

Fig. 6. Estimation result from HDP-HMM (Gibbs) for the rat hippocampal ensemble spike data. (A) The total number of states (solid blue) slowly increases as states are allocated for a small number of time bins. The number of states converges after 2500 iterations. (B) The log likelihood of the training data grows consistently as highly specific states are added. (C) and (D) The concentration parameters, α_0 and γ also converge after 2500 iterations. (E) and (F) The inferred state transition matrix $\boldsymbol{P}$ and neuronal firing matrix $\boldsymbol{\Lambda} = \{\lambda_{c,i}\}$ drawn from the last iteration. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

by the marginal probability of that state under the HDP-HMM. That is,

$$\hat{x}_t = \sum_{i=1}^{m} \bar{x}_i Pr(\mathcal{S}_t = i), \qquad \hat{y}_t = \sum_{i=1}^{m} \bar{y}_i Pr(\mathcal{S}_t = i), \tag{21}$$

where $\bar{x}_i$ and $\bar{y}_i$ denote the average location of the rat while in inferred state i (corresponding to the black dots in Fig. 5A). Note that the animal's position is not used in model inference, only during result assessment. In the illustrated example (HDP-HMM with MCMC+HMC), the mean reconstruction error in Euclidean distance is 9.07 cm.

As the parameter sample traces in Fig. 6 show, the Markov chain converges in around 2500 iterations. After this point, the total number of states stabilizes to around 65. The concentration parameters α_0 and γ converge within a similar number of iterations. Finally, we show the transition matrix $\boldsymbol{P}$ and firing rate matrix $\boldsymbol{\Lambda}$ obtained from the final Markov chain sample.

We again evaluate the sensitivity of these model fits to the choice of hyperparameters. For the HDP-HMM fit via Gibbs sampling with fixed α_c^0, the primary hyperparameters of interest are the concentration hyperparameters, a_{α_0} and a_γ in Eq. (5), where we have assumed $\alpha_0 \sim Gamma(a_{\alpha_0}, 1)$ and $\gamma \sim Gamma(a_\gamma, 1)$. Fig. 7 shows the inferred number of states as we vary these two hyperparameters over orders of magnitude. We find that the number of inferred states is stable around 65, indicating the performance robustness to the choice of these hyperparameters.

Looking into the inferred states, we can reconstruct the "place fields" or "state fields" of hippocampal neurons. To do so, we combine the state-location maps (Fig. 5B) with the firing rate of the individual neuron in those states (Fig. 6F) and weight by the marginal probability of the latent state. Together, these give rise to the inferred neuron's place field. Note that, again, the position data was only used in reconstruction but not in the inference procedure. Four pairs of inferred and true place fields are shown in

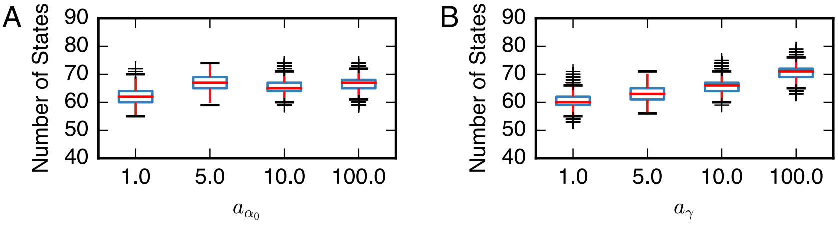

Fig. 7. Measuring the effect of concentration hyperparameters on the number of inferred latent states. We find that the concentration hyperparameters of the `Gamma` priors on the concentration parameters, α_0 and γ, have a minimal effect. Box plots are obtained from 10 independent Monte Carlo chains, and each chain was run for 1000 iterations; the number of states in the last iteration is used.

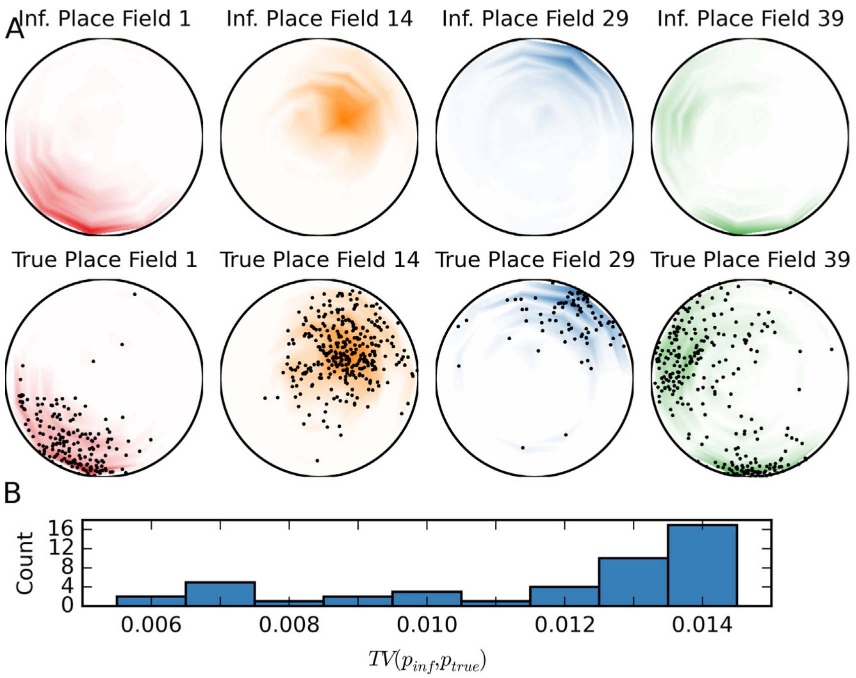

Fig. 8. (A) Comparison of inferred and true place fields for four randomly selected hippocampal neurons. The inferred place field (top row) for cell c is a combination of location distributions for each state i weighted by the inferred firing rates $\lambda_{c,i}$, whereas the true place field (bottom row) for cell c is a histogram of locations in which cell c fires. The black dots show the rat's locations used for each histogram. The inferred place fields closely match the true place fields. With adequate spike data recording, we expect a higher latent state dimensionality to yield higher spatial resolution in the inferred place fields. (B) Summary statistics of total variation (TV) distance between the inferred and true place fields for 47 neurons. The TV distance for the four examples in panel (A) are 0.0123, 0.0067, 0.0136, and 0.0091, respectively.

Fig. 8A. On the top row is the inferred place field; on the bottom is the true place field computed using the locations of the rat when cell c fired shown by black dots. We further assess the difference between the true and estimated place fields in population. Specifically, we compute the *total variation distance* between the inferred and true place fields for all 47 neurons, and the histogram statistic is shown in Fig. 8B.

In addition, we evaluate the model in terms of the information latent states convey about the rat's position in the circular environment. To do so, we divide the environment into 121 bins of equal area and treat the rat's position as a discrete random variable. Likewise, we treat the latent state as a discrete random variable, and we compute the discrete mutual information between these two variables. We investigate the information content of each individual state by constructing a binary random variable indicating whether or not the model is in state i and measuring its mutual information with the rat's position. The result is shown in Fig. 9, where the latent states are ordered in decreasing order of occupancy. As expected, states that are more frequently occupied carry more information about the rat's position.

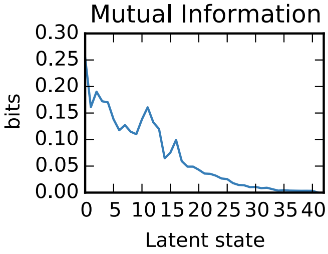

Fig. 9. Mutual information of the inferred states and the rat's position. Latent state are ordered by their occupancy, i.e. the number of times the rat was in that state.

4. Extensions and discussion

4.1. Hidden semi-Markovian models

In experimental data analysis, a striking feature of the inferred state transition matrix (Fig. 6E) is that the first 40 states exhibit strong self-transitions. This is a common feature of time series and has been addressed by a number of augmented Markovian models. In particular, hidden semi-Markovian models (HSMMs) explicitly model the duration of time spent in each state separately from the rest of the state transition matrix (Johnson and Willsky, 2013). Building this into the model allows the Dirichlet or HDP prior over state transition vectors to explain the rest of the transitions, which are often more similar. Alternatively, the “sticky” HMM and HDP-HMM accomplish a similar effect (Fox et al., 2008).

4.2. Statistical and computational considerations

We have seen a great advantage in Bayesian nonparametric formalism (i.e., HDP-HMM vs. HMM) regarding automatic model selection. This is especially important for sparse sample size or short recording in some neuroscience applications, where cross validation on data is infeasible.

For any statistical estimation, we need to consider the “bias vs. variance” problem. In VB inference, there is a potential estimate bias due to bound optimization (since we optimize the lower bound of the marginal likelihood). In addition, because of the mean-field approximation, the parameter's variance tends to be underestimated. In MCMC inference, the estimate is asymptotically unbiased, however, if the Markov chain mixes slowly, the estimate's variance can be inaccurate.

Computationally, the fully-Bayesian HDP-HMM inference is the most demanding. In practice, one can choose various inference tools with gradually increased computational resources (VB, empirical

Bayes, Gibbs sampling or HMC) depending on the data sample size and complexity. In addition, the convergence of these algorithm may vary according to the choice of hyperparameters.

4.3. Latent state dimensionality and continuous latent state

In experimental data analysis, the number of identified states from HDP-HMM depends on the data as well as the priors of hyperparameters. Given the same size of environment, different numbers of cells or different recording duration may yield different estimation results (Fig. 3), since the nonparametric prior allocates states in accordance with the complexity of the data. We found that the weak priors over the concentration parameters have a minimal effect on the number of inferred states (Fig. 7). Fixing the scale hyperparameter of the firing rate prior distribution and performing Gibbs sampling over the scale of the prior is a simple and robust method.

In our problem, we formulate the latent state is discrete (finite or infinite) and infer the state-transition matrix, from which we can derive the "topology graph" of the unknown environment (Chen et al., 2012a, 2014). In parallel to the discrete-state HMM or HDP-HMM, we can also formulate a continuous state-space model, where the state is Gaussian and the observation is Poisson (Brown et al., 1998; Smith and Brown, 2003; Yu et al., 2009; Buesing et al., 2012). Various inference algorithms (Gaussian or variational approximation) have been developed for such models in the literature. Different from discrete state, the continuous-state has a smoother representation (due to infinite spatial resolution). However, similar to the HMM, we will need to deal with the model selection (dimensionality of latent state) problem, which is often tackled by cross validation (Yu et al., 2009). In addition, continuous latent state is subject to sign/scale ambiguity. For the purpose of representing space and spatial topology, the discrete-state representation is more appropriate. Provided that the animal's behavior (spatial location) is available, the continuous representation of space will be more accurate. In general, nonparametric Bayesian inference can be applied to continuous or discrete states, as well as continuous or discrete observations (Teh et al., 2006; Van Gael et al., 2008; Fox et al., 2008, 2010; Chen, 2015).

4.4. Robustness of the population firing model

A key assumption in our probabilistic model is the Poisson likelihood. Although this assumption may not be true in experimental data, our results have showed excellent performance. To further assess the robustness of HDP-HMM-Poisson model in experimental data analysis, at every temporal bin we further add additional homogeneous non-Poissonian noise to the observed population spike counts by drawing from a NB distribution (with varying levels of mean 0.25–1.0 and variance 0.5–2.0), and repeat the decoding error analysis. We have found that, as a general trend, the median decoding error gradually grows as increasing noise mean or variance; yet the decoding performance remains quite satisfactory (results not shown).

4.5. Use of soft-labeled spikes

Thus far, we have assumed that all recorded ensemble spikes are sorted and clustered into single units. Nevertheless, it is known that spike sorting is complex, time-consuming and error-prone (Wood and Black, 2008; Shalchyan and Farina, 2014). On the one hand, sorting error is inevitable when there are overlapping features (such as spike energy, amplitudes or principal components). On the other hand, traditional spike-sorting procedures often throw away considerable non-clusterable "noisy" spikes, which might contain informative tuning information. How to use these noisy spikes and maximize the information efficiency remains an open question. In other words, can we conduct the ensemble spike analysis using unsorted spikes?

Motivated from a sorting-free ensemble decoding analysis (Chen et al., 2012; Kloosterman et al., 2014), we may use a soft-clustering method based on a Gaussian mixtures model (parameterized by an augmented vector $\boldsymbol{\xi} = \{\ell_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^{K}$ that characterizes the weights, mean, and covariance parameters of the Gaussian mixtures). By clustering the spike waveform feature space, we assign each spike with a "soft" class label (about the unit identity) according to the posterior probability within the K-mixtures. In the feature space, the points close to (far away from) the c-th cluster center are associated with a probability assignment value close to (smaller than) 1 in the c-th class. Because of the soft membership of individual spikes, the spike count $y_{c,t}$ ($c = 1, \ldots, K$) within a time interval can be a non-integer value. Consequently, we replace the variable C with K to indicate that the number of neurons is unknown, and rewrite the log likelihood as follows

$$\log p(\boldsymbol{y}_{1:T}|\mathcal{S}, \boldsymbol{\theta}, \boldsymbol{\xi}) = \sum_{t=1}^{T}\sum_{c=1}^{K} \log p(y_{c,t}|S_t, \boldsymbol{\theta}, \boldsymbol{\xi}) \qquad (22)$$

In this case, the inference procedure consists of two steps. At the first stage, d-dimensional spike waveform features are clustered using a "constrained" Gaussian mixture model (Zou and Adams, 2012), which can be either finite or infinite. In the case of infinite Gaussian mixtures, we can also resort to the nonparametric Bayesian approach (Rasmussen, 2000; Görür and Rasmussen, 2010; Wood and Black, 2008). Upon completing the inference, each spike will be given a posterior probability of being assigned to each cluster. At the second stage, we sum the soft-labeled spikes to obtain the probabilistic spike count $y_{c,t}$ for all K-clusters, and the remaining nonparametric Bayesian (MCMC or VB) inference procedure remains unchanged. A detailed investigation of this idea will be pursued in future work.

5. Conclusion

In this paper, we have explored the use of HDP-HMMs with Poisson likelihoods to analyze rat hippocampal ensemble spike data during spatial navigation. Compared to the parametric finite-state HMM, the HDP-HMM allows more flexibility to model the experimental data (without relying on time-consuming cross-validation in model selection). We evaluate two Bayesian nonparametric inference algorithms for HDP-HMM, one based on VB and the other based on MCMC. Furthermore, we consider two approaches for hyperparameter selection, an issue that is particularly important for the real-life application. It is found that the MCMC algorithm with HMC updates for the hyperparameters is robust and achieves the best performance in all simulated and experimental data. Our investigation shows a promising direction in applying nonparametric Bayesian methods for ensemble neuronal spike data analysis.

The unsupervised Bayesian inference approach allows us (or hippocampus downstream structures) to read out spatial information from hippocampal neuronal ensembles without *a priori* place receptive field information. One important future research direction is to apply this method to investigate sleep-associated hippocampal ensemble spike activity during either slow wave sleep (SWS) or rapid eye movement (REM) sleep (Louie and Wilson, 2001; Lee and Wilson, 2002). Traditionally, one would rely on place receptive fields estimated from pre-sleep run behavior to infer the content of population spike activity in sleep, but this approach is accompanied with many well-known statistical challenges, such as non-stationarity, firing rate remapping, and timescale warping. Our approach proposed here can provide an effective and complementary paradigm to investigate neural representation of hippocampal population codes without direct measurement of spatial correlate

(Chen et al., 2015). The same principle can also be applied to neocortical ensemble spike data (Ji and Wilson, 2007; Peyrache et al., 2009; Gulati et al., 2014).

Conflict of interest

The authors declare that there is no conflict of commercial interest.

Acknowledgments

Z.C. is supported by an NSF-CRCNS award (No. IIS-1307645) from the National Science Foundation. S.W.L. is supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship and by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. M.J.J. is supported by the Harvard/MIT Joint Research Grants Program. M.A.W. is supported by the NIH Grant R01-MH06197, TR01-GM10498, ONR-MURI N00014-10-1-0936 grant. This material is also based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. We also thank the members of the Wilson laboratory for sharing the experimental data and valuable discussions. S.W.L. and M.J.J. thank Ryan Adams for his support and helpful advice.

References

Beal MJ, Ghahramani Z, Rasmussen CE. The infinite hidden Markov model. In: Advances in neural information processing systems 14. Cambridge, MA: MIT Press; 2002. p. 577–85.

Brown EN, Frank LM, Tang Dc, Quirk M, Wilson MA. A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. J Neurosci 1998;18(18):7411–25.

Bryant M, Sudderth EB. Truly nonparametric online variational inference for hierarchical Dirichlet processes. In: Advances in neural information processing systems 25; 2012. p. 2699–707.

Buesing L, Macke JH, Sahani M. Learning stable, regularised latent models of neural population dynamics. Netw Comput Neural Syst 2012;23:24–47.

Chen Z. An overview of Bayesian methods for neural spike train analysis. Comput Intell Neurosci 2013;2013:251905.

Chen Z, editor. Advanced state space methods in neural and clinical data. Cambridge University Press; 2015.

Chen Z, Gomperts SN, Yamamoto J, Wilson MA. Neural representation of spatial topology in the rodent hippocampus. Neural Comput 2014;26(1):1–39.

Chen Z, Grosmark A, Penagos H, Buzsaki G, Wilson MA. Statistical analysis towards sleep-associated hippocampal ensemble spike activity; 2016, Manuscript under review.

Chen Z, Kloosterman F, Brown EN, Wilson MA. Uncovering spatial topology represented by rat hippocampal population neuronal codes. J Comput Neurosci 2012a;33(2):227–55.

Chen Z, Kloosterman F, Layton S, Wilson MA. Transductive neural decoding for unsorted neuronal spikes of rat hippocampus. In: Proceedings of IEEE engineering in medicine and biology conference; 2012. p. 1310–3.

Chen Z, Putrino DF, Ghosh S, Barbieri R, Brown EN. Statistical inference for assessing functional connectivity of neuronal ensembles with sparse spiking data. IEEE Trans Neural Syst Rehabil Eng 2011;19(2):121–35.

Cowles MK, Carlin BP. Markov chain Monte Carlo convergence diagnostics: a comparative review. J Am Stat Assoc 1996;91:883–904.

Curto C, Itskov V. Cell groups reveal structure of stimulus space. PLoS Comput Biol 2008;4(10):e1000205.

Dabaghian Y, Brandt VL, Frank L. Reconceiving the hippocampal map as a topological template. eLife 2014;3:e03476.

Dabaghian Y, Memoli F, Frank L, Carlsson G. A topological paradigm for hippocampal spatial map formation using persistent homology. PLoS Comput Biol 2012;8(8):e1002581.

Davidson TJ, Kloosterman F, Wilson MA. Hippocampal replay of extended experience. Neuron 2009;63(4):497–507.

Ewens WJ. Population genetics theory–the past and the future. In: Lessard S, editor. Mathematical and statistical developments of evolutionary theory. Springer; 1990. p. 177–227.

Ferguson TS. A Bayesian analysis of some nonparametric problems. Ann Stat 1973:209–30.

Fox EB, Sudderth EB, Jordan MI, Willsky AS. An HDP-HMM for systems with state persistence. In: Proceedings of the 25th international conference on machine learning; 2008. p. 312–9.

Fox EB, Sudderth EB, Jordan MI, Willsky AS. Bayesian nonparametric methods for learning Markov switching processes. IEEE Signal Process Mag 2010;27:43–54.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd ed. CRC Press; 2013.

Goris RL, Movshon JA, Simoncelli EP. Partitioning neuronal variability. Nat Neurosci 2014;17:858–65.

Görür D, Rasmussen CE. Dirichlet process Gaussian mixture models: choice of the base distribution. J Comput Sci Technol 2010;25(4):653–64.

Gulati T, Ramanathan DS, Wong CC, Ganguly K. Reactivation of emergent task-related ensembles during slow-wave sleep after neuroprosthetic learning. Nat Neurosci 2014;17(8):1107–13.

Hoffman MD, Blei DM, Wang C, Paisley J. Stochastic variational inference. J Mach Learn Res 2013;14(1):1303–47.

Ishwaran H, Zarepour M. Exact and approximate sum representations for the Dirichlet process. Can J Stat 2002;30(2):269–83.

Ji D, Wilson MA. Coordinated memory replay in the visual cortex and hippocampus during sleep. Nat Neurosci 2007;10(1):100–7.

Johnson MJ. Bayesian time series models and scalable inference. Ph.D. thesis. Massachusetts Institute of Technology; June 2014.

Johnson MJ, Willsky AS. Bayesian nonparametric hidden semi-Markov models. J Mach Learn Res 2013;14(1):673–701.

Johnson MJ, Willsky AS. Stochastic variational inference for Bayesian time series models. JMLR W&CP 2014;32:1854–62.

Kloosterman F, Layton SP, Chen Z, Wilson MA. Bayesian decoding using unsorted spikes in the rat hippocampus. J Neurophysiol 2014;111(1):217–27.

Kuhn HW. The Hungarian method for the assignment problem. Nav Res Logist Q 1955;2(1–2):83–97.

Lee AK, Wilson MA. Memory of sequential experience in the hippocampus during slow wave sleep. Neuron 2002;36(6):1183–94.

Liang P, Petrov S, Jordan MI, Klein D. The infinite PCFG using hierarchical Dirichlet processes. In: Proceedings of empirical methods in natural language processing (EMNLP); 2007. p. 688–97.

Louie K, Wilson MA. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. Neuron 2001;29:145–56.

Mishchenko Y, Vogelstein JT, Paninski L. A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. Ann Appl Stat 2011;5:1229–61.

Neal RM. MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo 2010:113–62.

O'Keefe J, Nadel L. The Hippocampus as a Cognitive Map, Vol. 3. Clarendon Press; 1978.

Peyrache A, Khamassi M, Benchenane K, Wiener S, Battaglia F. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nat Neurosci 2009;12(7):919–26.

Raftery AE, Lewis S. How many iterations in the Gibbs sampler? In: Bernardo JM, Berger J, Dawid AP, Smith AFM, editors. Bayesian Stat. Oxford University Press; 1992. p. 763–73.

Rasmussen CE. The infinite Gaussian mixture model. In: Advances in neural information processing systems 12. Cambridge, MA: MIT Press; 2000. p. 554–60.

Scott SL. Bayesian methods for hidden Markov models: Recursive computing in the 21st century. J Am Stat Assoc 2002;97(457):337–51.

Sethuraman J. A constructive definition of Dirichlet priors. Stat Sin 1994;4:639–50.

Shalchyan V, Farina D. A non-parametric bayesian approach for clustering and tracking non-stationarities of neural spikes. J Neurosci Methods 2014;223:85–91.

Smith AC, Brown EN. Estimating a state-space model from point process observations. Neural Comput 2003;15(5):965–91.

Teh YW, Jordan MI. Hierarchical Bayesian nonparametric models with applications. In: Hjort NL, Holmes C, Müller P, Walker SG, editors. Bayesian Nonparametr. Cambridge University Press; 2010. p. 158–207.

Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. J Am Stat Assoc 2006;101:1566–81.

Van Gael J, Saatci Y, Teh YW, Ghahramani Z. Beam sampling for the infinite hidden Markov model. In: Proceedings of the 25th international conference on machine learning; 2008. p. 1088–95.

Wood F, Black MJ. A nonparametric Bayesian alternative to spike sorting. J Neurosc Methods 2008;173(1):1–12.

Yau C, Papaspiliopoulos O, Roberts GO, Holmes C. Bayesian non-parametric hidden Markov models with applications in genomics. J R Stat Soc Ser B Stat Methodol 2011;73(1):37–57.

Yu BM, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. J Neurophysiol 2009;102:614–35.

Zou JY, Adams RP. Priors for diversity in generative latent variable models. In: Advances in neural information processing systems 24. Cambridge, MA: MIT Press; 2012.