7

Bayesian Nonparametric Hidden Markov Models

The dynamic network models of the Chapter 6 introduced an important idea — the notion of a time-varying latent state. While we did not explicitly frame it in these terms, the dynamic weight matrices are an example of a latent population state. There, the state had a concrete biophysical interpretation and was imbued with dynamics derived from synaptic plasticity, but in general, latent state space models need not be tied to specific biophysical phenomena. In this chapter, we explore a very common state space model, the hidden Markov model (HMM), which can model neural spike trains as a progression of discrete latent states. By relating these states to externally measured covariates, we can gain insight into the computations performed by neural circuits.

Hidden Markov models have found numerous applications in neural modeling. Recently, Latimer et al. (2015) have used HMMs to argue that decision-making related activity in macaque lateral intraparietal (LIP) area is better characterized by a discrete step between "undecided" and "decided" states. Similarly, Miller and Katz (2010) have used HMMs to model the dynamics of neural activity during taste processing and decision making. In modeling the activity in hippocampal place cells, Chen et al. (2012; 2014) have used HMMs to interrogate discrete states of hippocampal activity and reconstruct topological maps of the environment from neural activity. This chapter provides further analysis of the same

hippocampal dataset, which was also used in Chapters 3 and 5.

A major challenge with using HMMs in practice is determining the number of latent states. It is common to fit models of varying dimensionality and compare them on the basis of a cross-validation metric, such as the predictive log-likelihood assigned to held-out data. However, this approach has a number of drawbacks. First, it can be computationally expensive to fit and compare a sequence of models. Second, it makes inefficient use of the data since we can only use a fraction of the data to train the model. While this is primarily a concern in the "small data" regime, it is still relevant to neural data analysis. Third, as with basic mixture models, the standard model does not explicitly penalize duplicate states. This penalty is only implicitly enforced through cross-validation. In terms of interpretability, it is desirable to incorporate an explicit penalty on excess states.

Here we extend the preceding work and consider a *Bayesian nonparametric* approach (Orbanz and Teh, 2011). The Bayesian nonparametric approach brings additional flexibility to the probabilistic model (Teh and Jordan, 2010; Wood and Black, 2008; Shalchyan and Farina, 2014). Specifically, we use a hierarchical Dirichlet process hidden Markov model (HDP-HMM) (Teh et al., 2006), which extends the finite-state HMM with an HDP prior that theoretically allows a countably infinite number of states. The Dirichlet process provides a nonparametric prior over atomic probability measures with support for a countably infinite number of outcomes. This is a natural way to model distributions over infinitely many states, e.g. to model the rows of a transition matrix. The *hierarchical* part of the HDP prior allows sharing between the rows. Even though these priors support infinitely many states, we can still perform efficient inference by leveraging truncation methods and constructive definitions of the HDP that only instantiate variables for states that are visited in the data.

Nevertheless, inference in Bayesian nonparametric models can be finicky. Hyperparameters of the HDP prior can have a strong effect, as can the particular choice of inference algorithm. Here we provide an assessment of various inference approaches. We consider both MCMC and variational inference algorithms. For MCMC, we adapt the Gibbs sampling approach of (Teh et al., 2006), and consider various methods of inferring hyperparameters.

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.

Probabilistic Model

First we present a standard Bayesian hidden Markov model with a fixed, finite number of states. We introduce notation and prior distributions for the various model parameters. Then we extend this to the nonparametric case with a countably infinite number of latent states.

Parametric Hidden Markov Models

Consider a finite K-state HMM applied to population spiking activity from a population of N neurons. We assume that the discrete latent state follows a first-order Markov chain $\boldsymbol{z} = [z_1, \ldots, z_T]$ with $z_t \in \{1, \ldots, K\}$, and that the spike counts of individual neurons at time t follow a Poisson distribution whose rate depends on the latent state, z_t. This is summarized in the following probabilistic model:

$$p(\boldsymbol{S}, \boldsymbol{z} \mid \boldsymbol{\pi}^{(0)}, \boldsymbol{P}, \boldsymbol{\Lambda}) = p(z_1 \mid \boldsymbol{\pi}) \prod_{t=2}^{T} p(z_t \mid z_{t-1}, \boldsymbol{P}) \prod_{t=1}^{T} p(\boldsymbol{s}_t \mid z_t, \boldsymbol{\Lambda}), \tag{7.1}$$

where

$$\begin{aligned} p(z_1 \mid \boldsymbol{\pi}^{(0)}) &= \mathrm{Discrete}(z_1 \mid \boldsymbol{\pi}^{(0)}), \\ p(z_t \mid z_{t-1}, \boldsymbol{P}) &= \mathrm{Discrete}(z_t \mid \boldsymbol{\pi}^{(z_{t-1})}), \\ p(\boldsymbol{s}_t \mid z_t, \boldsymbol{\Lambda}) &= \prod_{n=1}^{N} \mathrm{Poisson}(s_{t,n} \mid \lambda_{z_t,n}). \end{aligned}$$

Here, $\boldsymbol{\pi}^{(0)} \in [0,1]^K$ is a discrete probability distribution over initial states, and

$$\boldsymbol{P} = \begin{bmatrix} — & \boldsymbol{\pi}^{(1)} & — \\ & \vdots & \\ — & \boldsymbol{\pi}^{(K)} & — \end{bmatrix},$$

is a $K \times K$ transition matrix where the row, $\boldsymbol{\pi}^{(k)} \in [0,1]^K$, specifies a discrete conditional distribution over z_t given that $z_{t-1} = k$. The state-conditional firing rates are collected in the matrix,

$$\boldsymbol{\Lambda} = \begin{bmatrix} — & \boldsymbol{\lambda}^{(1)} & — \\ & \vdots & \\ — & \boldsymbol{\lambda}^{(K)} & — \end{bmatrix} = \begin{bmatrix} \lambda_{1,1} & \dots & \lambda_{1,N} \\ \vdots & & \vdots \\ \lambda_{K,1} & \dots & \lambda_{K,N} \end{bmatrix},$$

In a Bayesian HMM, we introduce prior distributions over the parameters. We use the following prior distributions,

$$\begin{aligned} \alpha_0 &\sim \text{Gamma}(a_{\alpha_0}, 1.0) \\ \boldsymbol{\pi}^{(0)} &\sim \text{Dir}(\alpha_0 \mathbf{1}), \\ \boldsymbol{\pi}^{(k)} &\sim \text{Dir}(\alpha_0 \mathbf{1}), \\ \lambda_{k,n} &\sim \text{Gamma}(\kappa_n, \nu_n). \end{aligned}$$

where gamma prior on firing rates has neuron-specific shape parameters, κ_n, and scale parameters, ν_n.

Nonparametric Hidden Markov Models

Model selection is an important issue for statistical modeling and data analysis. 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). 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, $G_0 \sim \mathcal{DP}(\gamma, H)$, where γ is the concentration parameter and H is the base measure. Moreover, we place a prior distribution over the concentration parameter, $\gamma \sim \text{Gamma}(a_\gamma, 1.0)$. Given the concentration, one may sample from the DP via the "stick-breaking construction" (Sethuraman, 1994). First, sample the stick-breaking weights, $\boldsymbol{\beta}$,

$$\widetilde{\beta}_k \sim \text{Beta}(1, \gamma), \quad \beta_k = \widetilde{\beta}_k \prod_{j=1}^{k-1} \Big(1 - \widetilde{\beta}_j\Big), \tag{7.2}$$

where $\beta_1 = \widetilde{\beta}_1$, $\sum_{k=1}^{\infty} \beta_k = 1$.

The stick-breaking construction of (7.2) is sometimes denoted as $\boldsymbol{\beta} \sim \text{GEM}(\gamma)$, after Griffiths, Engen, and McCloskey (Ewens, 1990). The name "stick-breaking" comes from the interpretation of β_k as the length of the piece of a unit-length stick assigned to the k-th value. After the first $k-1$ values having their portions assigned, the length of the remainder of the stick is broken according to a sample $\widetilde{\beta}_k$ from a beta distribution, and β_k indicates the portion of the remainder to be assigned to the k-th value. Therefore, the stick-breaking process $\text{GEM}(\gamma)$ also defines a DP— smaller values of γ will lead to larger values of $\widetilde{\beta}_k$, which means most of the probability mass will be allocated to the first "sticks," i.e. the small values of k.

After sampling $\boldsymbol{\beta}$, we next sample the latent state variables, in this case $\boldsymbol{\lambda}^{(k)}$, from the base measure H. For us, H is simply a set of independent gamma distributions for each neuron. Our draw from the $\mathcal{DP}(\gamma, H)$ prior is then given by

$$G_0(\boldsymbol{\lambda}) = \sum_{k=1}^{\infty} \beta_k \, \delta_{\boldsymbol{\lambda}^{(k)}}(\boldsymbol{\lambda}).$$

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, $\boldsymbol{\pi}^{(k)} \sim \mathcal{DP}(\alpha_0, \boldsymbol{\beta})$. We place the same prior over $\boldsymbol{\pi}^{(0)}$. 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 $\boldsymbol{P}$ 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.

Markov Chain Monte Carlo Inference

Several MCMC-based inference methods have been developed for the HDP-HMM (Teh et al., 2006; Van Gael et al., 2008). Some of these previous works use a collapsed Gibbs sampler in which the transition matrix $\boldsymbol{P}$ 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

$$\begin{aligned} \gamma &\sim \text{Gamma}(a_\gamma, 1), \\ \alpha_0 &\sim \text{Gamma}(a_{\alpha_0}, 1), \\ \boldsymbol{\beta} \,|\, \gamma &\sim \text{Dir}(\gamma/K_{\text{max}}, \ldots, \gamma/K_{\text{max}}), \\ \boldsymbol{\pi}^{(0)} \,|\, \alpha_0, \boldsymbol{\beta} &\sim \text{Dir}(\alpha_0\beta_1, \ldots, \alpha_0\beta_{K_{\text{max}}}), \\ \boldsymbol{\pi}^{(k)} \,|\, \alpha_0, \boldsymbol{\beta} &\sim \text{Dir}(\alpha_0\beta_1, \ldots, \alpha_0\beta_{K_{\text{max}}}). \end{aligned} \tag{7.3}$$

where K_{max} denotes a truncation level. 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 $\boldsymbol{z}$.

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) (Neal, 2010) transitions to sample the neuron-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

$$\lambda_{k,n} \mid \boldsymbol{S}, \boldsymbol{z} \sim \text{Gamma}\left(\kappa_n + \sum_{t=1}^{T} s_{t,n}\, \mathbb{I}[z_t = k],\ \nu_n + \sum_{t=1}^{T} \mathbb{I}[z_t = k]\right).$$

Under the weak limit approximation the priors on $\boldsymbol{\pi}^{(k)}$ and $\boldsymbol{\pi}^{(0)}$ 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:

$$\begin{aligned}
\boldsymbol{\pi}^{(0)} \mid \alpha_0, \boldsymbol{\beta} &\sim \text{Dir}\left(\alpha_0 \boldsymbol{\beta} + \mathbf{1}_{z_1}\right), \\
\boldsymbol{\pi}^{(k)} \mid \alpha_0, \boldsymbol{\beta} &\sim \text{Dir}\left(\alpha_0 \boldsymbol{\beta} + \boldsymbol{n}_k\right), \\
n_{i,j} &= \sum_{t=1}^{T-1} \mathbb{I}[z_t = i] \cdot \mathbb{I}[z_{t+1} = j],
\end{aligned}$$

where $\mathbf{1}_k$ is a unit vector with a one in the k-th entry.

The Dirichlet parameters $\boldsymbol{\beta}$ and the concentration parameters α_0 and γ can be updated as in (Teh et al., 2006).

Block Gibbs updates for the latent states

Conditioned upon the firing rates, the initial state distribution, and the transition matrix, which we collectively refer to as $\boldsymbol{\theta}$, we can jointly update the latent states of the HDP-HMM using a *forward filtering, backward sampling* algorithm. Jointly sampling these latent states allows us to avoid issues with mixing when individually sampling states that are highly correlated with one another. We provide a brief overview of this algorithm here. Complete details of this algorithm can be found in, for example, Johnson (2014).

First, we "filter" the data to get the marginal distribution over z_t given the observations

up to time t. We use "Matlab" notation to refer to a set of variables, $\boldsymbol{s}_{1:t} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_t\}$. Since z_t is discrete, its filtered distribution is parameterized by a probability vector, which we call $\boldsymbol{m}_t$.

We compute these filtered probability distributions iteratively. Assume that at iteration t we have already computed $\boldsymbol{m}_{t-1}$. Given the Markovian structure of the probabilistic model, the conditional distribution of z_t factors into,

$$p(z_t \mid \boldsymbol{s}_{1:t}, \boldsymbol{\theta}) \propto \underbrace{p(\boldsymbol{s}_t \mid z_t, \boldsymbol{\theta})}_{\text{condition}} \underbrace{p(z_t \mid \boldsymbol{s}_{1:t-1}, \boldsymbol{\theta})}_{\text{predict}}.$$

The *prediction* step involves a marginalization over the previous latent state, z_{t-1},

$$\begin{aligned} p(z_t \mid \boldsymbol{s}_{1:t-1}\boldsymbol{\theta}) &\propto \sum_{k=1}^{K} p(z_t \mid z_{t-1} = k, \boldsymbol{\theta})\, p(z_{t-1} = k \mid \boldsymbol{s}_{1:t-1}, \boldsymbol{\theta}) \\ &= \text{Discrete}(z_t \mid \boldsymbol{m}_{t|t-1}), \end{aligned}$$

where

$$m_{t|t-1,k} \propto \sum_{j=1}^{K} p(z_t = k \mid z_{t-1} = j, \boldsymbol{\theta}) \cdot m_{t-1,j}.$$

Then, we *condition* on the current observations, $\boldsymbol{s}_t$, to get the parameters of the filtered distribution,

$$m_{t,k} \propto p(\boldsymbol{s}_t \mid z_t = k, \boldsymbol{\theta}) \cdot m_{t|t-1,k}. \tag{7.4}$$

Once we have computed the filtered distributions for all time bins, we can sample from the joint distribution over $\boldsymbol{z}_{1:T}$ by applying the chain rule,

$$\begin{aligned} p(\boldsymbol{z}_{1:T} \mid \boldsymbol{s}_{1:T}, \boldsymbol{\theta}) &= p(z_T \mid \boldsymbol{s}_{1:T}, \boldsymbol{\theta}) \prod_t p(z_t \mid \boldsymbol{z}_{t+1:T}, \boldsymbol{s}_{1:T}, \boldsymbol{\theta}) \\ &\propto p(z_T \mid \boldsymbol{s}_{1:T}, \boldsymbol{\theta}) \prod_t p(z_t \mid \boldsymbol{s}_{1:t}, \boldsymbol{\theta})\, p(z_{t+1} \mid z_t, \boldsymbol{\theta}). \end{aligned}$$

Thus, we can sample in reverse order, starting with z_T and ending with z_1. The conditional distribution of z_t is,

$$p(z_t \mid \boldsymbol{z}_{t+1:T}, \boldsymbol{s}_{1:T}, \boldsymbol{\theta}) \propto p(z_t \mid \boldsymbol{m}_t)\, p(z_{t+1} \mid z_t, \boldsymbol{\theta}), \tag{7.5}$$

which is another discrete distribution. The final algorithm for block Gibbs sampling $\boldsymbol{z}_{1:T}$ is:

Require: $\boldsymbol{s}_{1:T}, \boldsymbol{\theta}$
  for $t = 1, \ldots, T$ **do**
    Compute $\boldsymbol{m}_t$ ▷ Eq. 7.4
  end for
  for $t = T, \ldots, 1$ **do**
    Sample $z_t \mid z_{t+1}, \boldsymbol{m}_t, \boldsymbol{\theta}$ ▷ Eq. 7.5
  end for

Algorithm 7.1: Forward filtering, backward sampling algorithm for the hidden Markov model.

A single iteration of the comple Gibbs sampling 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 chapter is the sampling of the hyperparameters of the gamma firing rate priors.

Setting firing rate hyperparameters

We consider three approaches to setting the hyperparameters of the gamma priors for Poisson firing rates, namely, $\{\kappa_n, \nu_n\}$ for the n-th neuron.

- 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 neuron, 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 κ_n, the conditional distribution of the "scale" parameter, ν_n 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 re-parameterize the problem in terms of their logs. For neuron n, the conditional log probability equal to,

$$\begin{aligned}\mathcal{L} &= \log p(\log \kappa_n, \log \nu_n \mid \mathbf{\Lambda}) \\ &= \sum_{k=1}^{K} \log p(\lambda_{k,n} \mid \kappa_n, \nu_n) + \text{const.} \\ &= \sum_{k=1}^{K} \kappa_n \log \nu_n - \log \Gamma(\kappa_n) + (\kappa_n - 1) \log \lambda_{k,n} - \nu_n \lambda_{k,n}.\end{aligned}$$

Taking gradients with respect to both parameters yields,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \log \kappa_n} &= \sum_{k=1}^{K} \left[\log \nu_n - \Psi(\kappa_n) + \log \lambda_{k,n}\right] \times \kappa_n, \\ \frac{\partial \mathcal{L}}{\partial \log \nu_n} &= \sum_{k=1}^{K} \left[\frac{\kappa_n}{\nu_n} - \lambda_{k,n}\right] \times \nu_n.\end{aligned}$$

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, κ_n, and infer the scale, ν_n. We place a gamma prior on the scale, $\nu_n \sim \text{Gamma}(\mu, \nu_0)$. Given κ_n, the condi-

tional distribution of the scale is

$$\nu_n \mid \kappa_n, \{\lambda_{k,n}\}, \boldsymbol{z} \sim \text{Gamma}\left(\mu + \sum_{k=1}^{K} \mathbb{I}[n_k > 0] \cdot \kappa_n,\ \nu_0 + \sum_{k=1}^{K} \mathbb{I}[n_k > 0] \cdot \lambda_{k,n}\right)$$

$$n_k = \sum_{t=1}^{T} \mathbb{I}[z_t = k].$$

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

Predictive log likelihood

With the parameter and hyperparameter inference complete, 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,

$$\begin{aligned}\log p(\boldsymbol{S}_{\text{test}} \mid \boldsymbol{S}_{1:T}) &= \log \sum_{\boldsymbol{z}_{\text{test}}} \int_{\boldsymbol{\Theta}} p\left(\boldsymbol{S}_{\text{test}}, \boldsymbol{z}_{\text{test}} \mid \boldsymbol{\theta}\right)\ p\left(\boldsymbol{\theta} \mid \boldsymbol{S}_{\text{train}}\right) \mathrm{d}\boldsymbol{\theta}, \\ &\approx \log \frac{1}{L} \sum_{\ell=1}^{L} \sum_{\boldsymbol{z}_{\text{test}}} p\left(\boldsymbol{S}_{\text{test}}, \boldsymbol{z}_{\text{test}} \mid \boldsymbol{\theta}^{(\ell)}\right),\end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{P}, \boldsymbol{\pi}^{(0)})$ and $\{\boldsymbol{\theta}^{(\ell)}\}_{\ell=1}^{L} \sim p(\boldsymbol{\theta} \mid \boldsymbol{S}_{\text{train}})$. The summation over latent state sequences for the test data is performed with the message-passing algorithm for HMMs.

Variational Inference

We build upon our previous work (Chen et al., 2012; 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,

$$p(\boldsymbol{z}, \boldsymbol{\Lambda}, \boldsymbol{P}, \boldsymbol{\pi}^{(0)}, \boldsymbol{\beta} \mid \boldsymbol{S}_{1:T}) \approx q(\boldsymbol{z})\, q(\boldsymbol{\Lambda})\, q(\boldsymbol{P})\, q(\boldsymbol{\pi}^{(0)})\, q(\boldsymbol{\beta}).$$

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.

$$\begin{aligned}
q(\boldsymbol{\Lambda}) &= \prod_{k=1}^{K} \prod_{n=1}^{N} \text{Gamma}\left(\widetilde{\kappa}_{k,n},\, \widetilde{\nu}_{k,n}\right), \\
\widetilde{\kappa}_{k,n} &\leftarrow \kappa_n + \sum_{t=1}^{T} s_{t,n} \mathbb{E}_q\big[\mathbb{I}[z_t = k]\big], \\
\widetilde{\nu}_{k,n} &\leftarrow \nu_n + \sum_{t=1}^{T} \mathbb{E}_q\big[\mathbb{I}[z_t = k]\big].
\end{aligned}$$

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 K_{max} is chosen *a priori* and $q(\boldsymbol{z})$ is limited to support only states $z_t \in \{1, \ldots, K_{\mathsf{max}}\}$. The advantage of this approximation is that conjugacy is retained with $\boldsymbol{\Lambda}$, $\boldsymbol{P}$, and $\boldsymbol{\pi}^{(0)}$, and the variational approximation $q(\boldsymbol{z})$ reduces to*

$$\begin{aligned}
q(\boldsymbol{z}) &= \text{HMM}(\widetilde{\boldsymbol{P}}, \widetilde{\boldsymbol{\pi}}^{(0)}, \widetilde{\boldsymbol{\Lambda}}), \\
\widetilde{\boldsymbol{P}} &= \exp\left\{\mathbb{E}_q[\ln \boldsymbol{P}]\right\}, \\
\widetilde{\boldsymbol{\pi}}^{(0)} &= \exp\left\{\mathbb{E}_q[\ln \boldsymbol{\pi}^{(0)}]\right\}, \\
\widetilde{\boldsymbol{\Lambda}} &= \exp\left\{\mathbb{E}_q[\ln p(\boldsymbol{S} \mid \boldsymbol{\Lambda})]\right\}.
\end{aligned}$$

Expectations $\mathbb{E}_q[z_t = k]$ can then be computed using standard message-passing algorithms

*In a slight abuse of notation, $\widetilde{\boldsymbol{\Lambda}}$ refers to the expected observation likelihood for each latent state. That is, $\widetilde{\boldsymbol{\Lambda}}$ is a matrix where $\widetilde{\Lambda}_{t,k} = \exp\{\mathbb{E}_{q(\boldsymbol{\Lambda})}[\ln p(\boldsymbol{s}_t \mid z_t = k, \boldsymbol{\Lambda}]\}$.

for HMMs.

With the direct assignment truncation, the variational factors for the rows $\boldsymbol{\pi}^{(k)}$ and the initial distribution $\boldsymbol{\pi}^{(0)}$ are Dirichlet distributions. Unlike in the finite-state HMM, however, these Dirichlet factors are now over $K_{\mathsf{max}} + 1$ dimensions since the final dimension accounts for all states $k > K_{\mathsf{max}}$. Under the HDP prior we had $\boldsymbol{\pi}^{(k)} \sim \mathcal{DP}(\alpha_0 \cdot \boldsymbol{\beta})$, and under the truncation the DP parameter becomes $\alpha_0 \cdot \boldsymbol{\beta}_{1:K_{\mathsf{max}}+1}$. Again, leveraging the conjugacy of the model, we arrive at the following variational updates:

$$q(\boldsymbol{P}) = \prod_{k=1}^{K_{\mathsf{max}}} \mathrm{Dir}(\widetilde{\boldsymbol{n}}_k),$$

$$\widetilde{n}_{i,j} \leftarrow \alpha_0 \beta_j + \mathbb{E}_q[\mathbb{I}[z_t = i] \cdot \mathbb{I}[z_{t+1} = j]].$$

We use an analogous update for $\boldsymbol{\pi}^{(0)}$.

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 $q(\boldsymbol{\beta}) = \delta_{\boldsymbol{\beta}^*}(\boldsymbol{\beta})$ and use gradient ascent to update this parameter during inference.

There are a number of hyperparameters to set for the variational approach as well. The hyperparameters κ_n and ν_n 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.

Predictive Log Likelihood

Finally, in order to compute predictive log likelihoods on held-out test data, we draw multiple samples, $\{\boldsymbol{\theta}^{(\ell)}\}_{\ell=1}^{L}$ for $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{z}, \boldsymbol{P}, \boldsymbol{\pi}^{(0)}, \boldsymbol{\beta})$, from the variational posterior, q, and

approximate the predictive log likelihood as

$$\begin{aligned}\ln p(\boldsymbol{S}_{\text{test}} \mid \boldsymbol{S}) &\approx \ln \mathbb{E}_q \left[p(\boldsymbol{S}_{\text{test}} \mid \boldsymbol{\theta}) \right] \\ &\approx \ln \frac{1}{L} \sum_{\ell=1}^{L} p(\boldsymbol{S}_{\text{test}} \mid \boldsymbol{\theta}^{(\ell)}).\end{aligned}$$

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

Synthetic Data Experiments

Setup First, we simulate synthetic spike count data using an HDP-HMM with $N = 50$ 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: $\text{Gamma}(\kappa_n = 1, \nu_n = 1)$ (with mean 1.0 and standard deviation 1.0).

An example of one such synthetic dataset is shown in Fig. 7.1. 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, z_{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 $\kappa_n = 1$; (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 $\kappa_n = 1$; (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

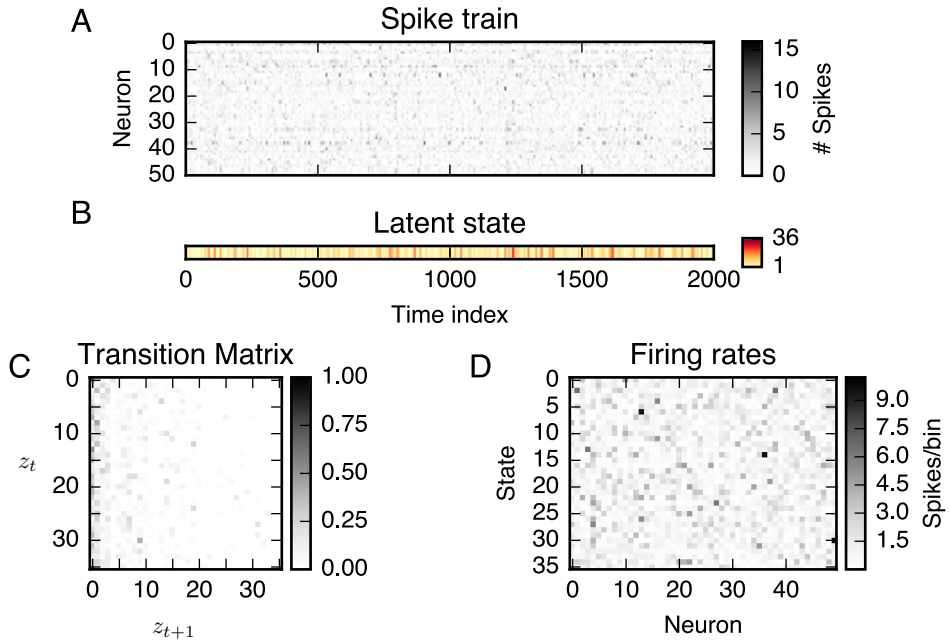

Figure 7.1: An example of a synthetic dataset drawn from an HDP-HMM. (A) Simulated population spike trains or spike counts. (B) Inferred latent state sequence. (C) Inferred state transition matrix $\boldsymbol{P}$. (D) Inferred neuronal firing rate matrix, $\boldsymbol{\Lambda}$.

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 $K_{\mathsf{max}} = 100$.

We collect 5000 samples from the MCMC algorithms and use the last 2000 for computing predictive log likelihoods. 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 $\boldsymbol{z}$ be the true state sequence and $\boldsymbol{z}'$ be the inferred state sequence. We define the overlap matrix $O \in \mathbb{N}^{K_{\max} \times K_{\max}}$ whose entries $O_{i,j}$ is the number of times the true state is i and the inferred state is j:

$$O_{i,j} = \sum_{t=1}^{T} \mathbb{I}[z_t = i]\, \mathbb{I}[z'_t = j].$$

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 $\boldsymbol{z}$, and the relabeled sequence of inferred states, $\widetilde{\boldsymbol{z}}'$:

$$\text{err}(\boldsymbol{z}, \widetilde{\boldsymbol{z}}') = \sum_{t=1}^{T} \mathbb{I}[z_t \neq \widetilde{z}'_t]. \tag{7.6}$$

Table 7.1 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 high Hamming error.

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

Table 7.1: Comparison of Hamming error (see Eq. 7.6) 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 higher 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

likelihood are given by,

$$\widehat{\lambda}_n = \frac{1}{T_{\mathsf{train}}} \sum_{t=1}^{T_{\mathsf{train}}} s_{t,n},$$

$$\log p(\boldsymbol{S}_{\mathsf{test}} \mid \boldsymbol{S}_{\mathsf{train}}) = \sum_{n=1}^{N} \left[-T_{\mathsf{test}} \widehat{\lambda}_n + \sum_{t=1}^{T_{\mathsf{test}}} s_{t,n} \log \widehat{\lambda}_n \right].$$

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 7.2 summarizes the predictive log likelihood comparison. For all five datasets, the HDP-HMM fit via Gibbs sampling with fixed κ_n 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 in under 10 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

Table 7.2: Comparison of predictive log likelihood (bits/spike) computed from 9 simulated data sets, measured in bits per 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

MCMC with HMC for hyperparameter sampling in our subsequent experiments.

Figure 7.2 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. 7.1 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 $K = 35$. 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. 7.3. In all cases, we use synthetic dataset 1, shown in Fig. 7.1, and HDP-HMMs fit via Gibbs sampling with fixed κ_n. First, we vary the number of observed neurons, N, and find that the number of inferred states was relatively stable around the true number of states ($K = 35$). By contrast, as we increase the observed 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

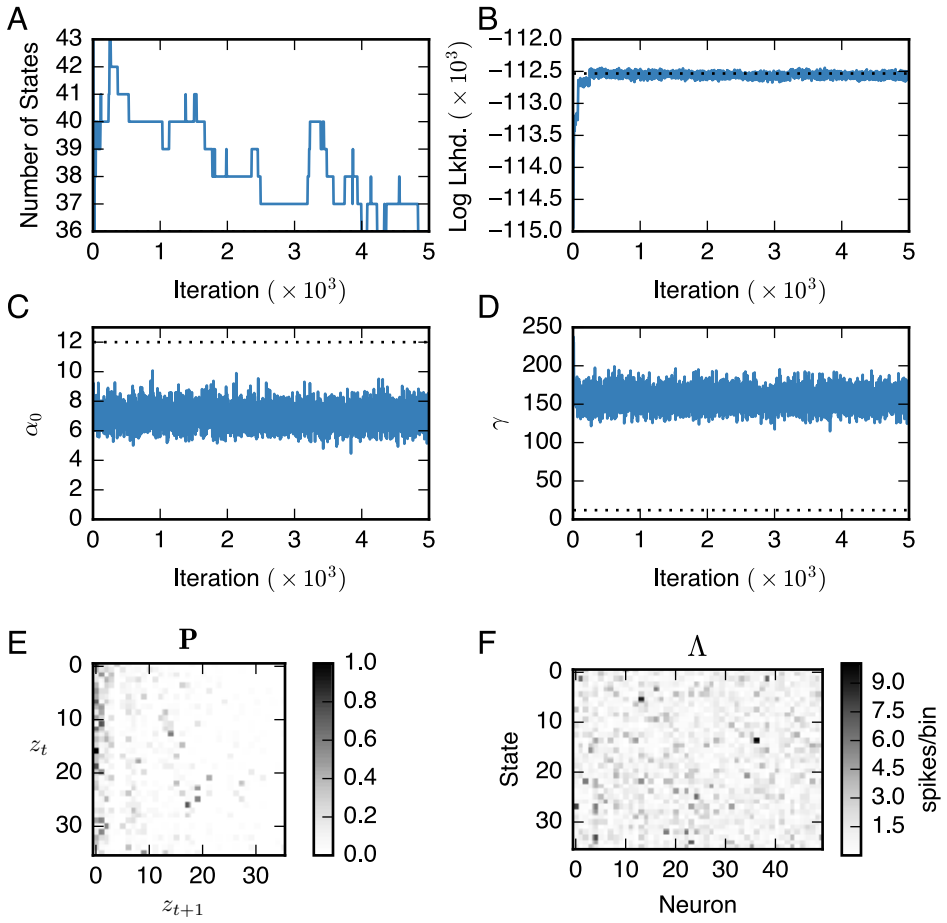

Figure 7.2: MCMC state trajectories for an HDP-HMM fit to the synthetic dataset shown in Fig. 7.1. True values are shown by the dotted black lines. The first five iterations of the Markov chain are omitted since they differ greatly from the final states. The chain quickly converges to nearly the correct number of states (A) and achieves close to the true log likelihood (B). (C, D) The chain trajectories of hyperparameters α_0 and γ. (E, F) Inferred state transition matrix and neuronal firing map drawn from the last iteration.

of 0.1, 0.5, 1.0, 2.0, or 10.0, and sampling a new spike count. When the 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 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_0 in Section 7.2.2). 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

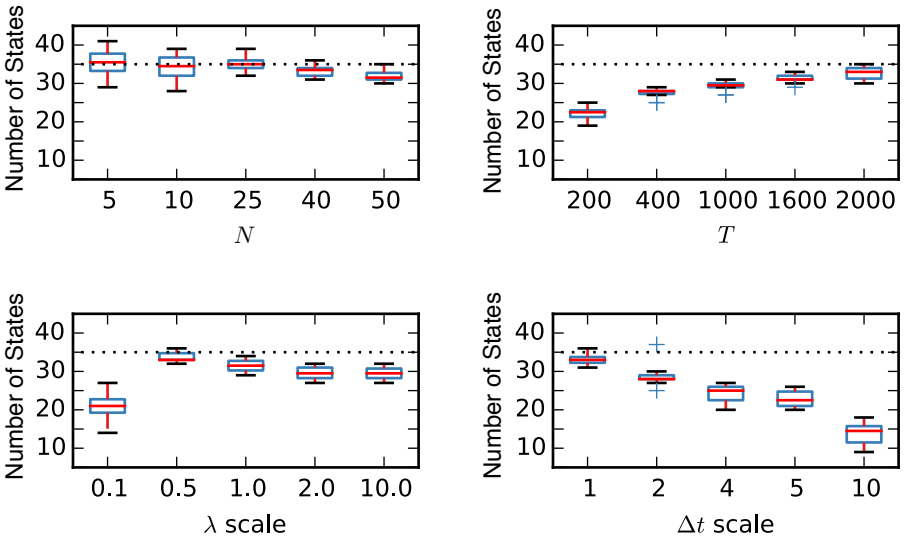

Figure 7.3: In a synthetic data experiment, we generated a spike train for a population of $N = 50$ neurons and $T_{\text{true}} = 2000$ time bins. Then we varied the number of observed neurons N, recording duration T, scale of the firing rate λ, temporal bin size Δt, and measured the number of inferred latent states. Horizontal dashed lines indicate the ground truth.

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.

Hippocampal Place Cells

Next, we apply the proposed methods to experimental data of the rat hippocampus. This is the same dataset studied in previous chapters, but here we applied additional preprocessing. We bin the ensemble spike activity with a bin size of 250 ms and obtain the population vector $\boldsymbol{z}$ 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. The result is a recording that is 9.8 minutes in duration. One animal's RUN trajectory and spatial occupancy are shown in Fig. 7.4 (left and right panels, 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

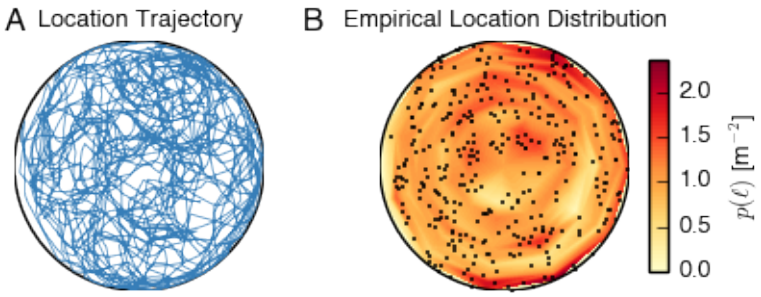

Figure 7.4: One rat's behavioral trajectory (left) and spatial occupancy (right) in the open field environment.

the rat is in the corresponding bin.

In experimental data analysis, we focus on Bayesian nonparametric inference for HDP-HMM. For all methods, we increase the truncation level to a large value of $K_{\text{max}} = 100$. 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 decoding error and predictive log likelihood. 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 7.3. We find that the HDP-HMM fit by Gibbs sampling with fixed firing rate scale ($\kappa_n = 1$) again outperforms the competing models in both

Table 7.3: 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 ($K = 25$)	0.712	10.85 ± 6.43
HMM ($K = 45$)	0.706	10.71 ± 6.67
HMM ($K = 65$)	0.717	11.01 ± 6.93
HDP-HMM (Gibbs)	**0.722**	$\mathbf{9.56 \pm 5.31}$
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

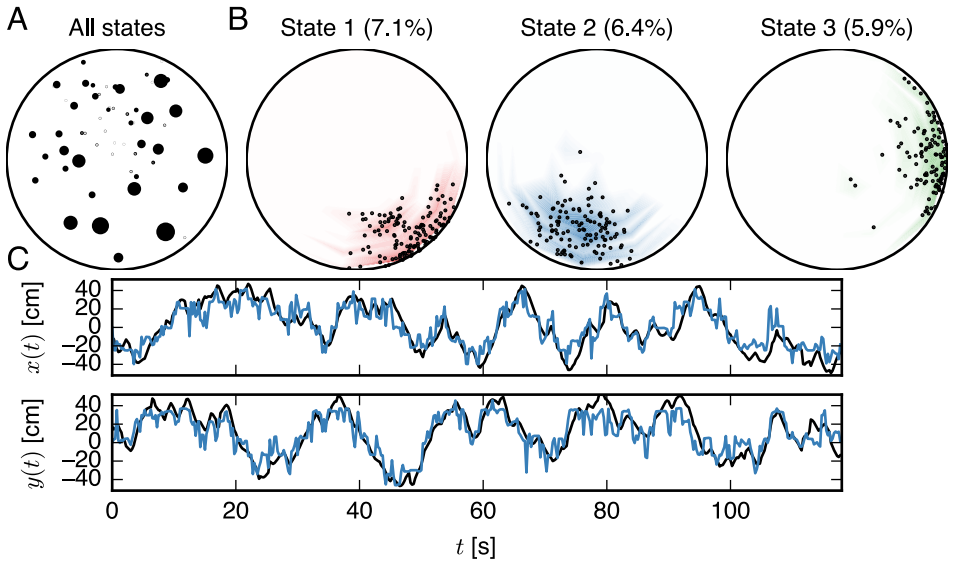

Figure 7.5: Estimation result from HDP-HMM (Gibbs) for the rat hippocampal ensemble spike data. (A) Estimated state space map, where the mean value of the spatial position for each latent state is shown by a black dot. The size of the dot is proportional to the occupancy of the state. (B) Probability 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 and reconstructed trajectories in Cartesian coordinate. The true trajectory is shown in black and the reconstructed trajectory is shown in blue. For each time bin, we use the mean location of the latent states to determine an estimate of the animal's location.

measures.

For the purpose of result assessment, we plot the state-space or state-location map (Fig. 7.5A), 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 $\Pr(z_t = k)$ for all time bins t and states k. Then we compute the average location for each state k by weighting the animal's location, (x_t, y_t) by the probability that the animal was in state k 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 k.

The empirical location distribution for the top three states as measured by occupancy

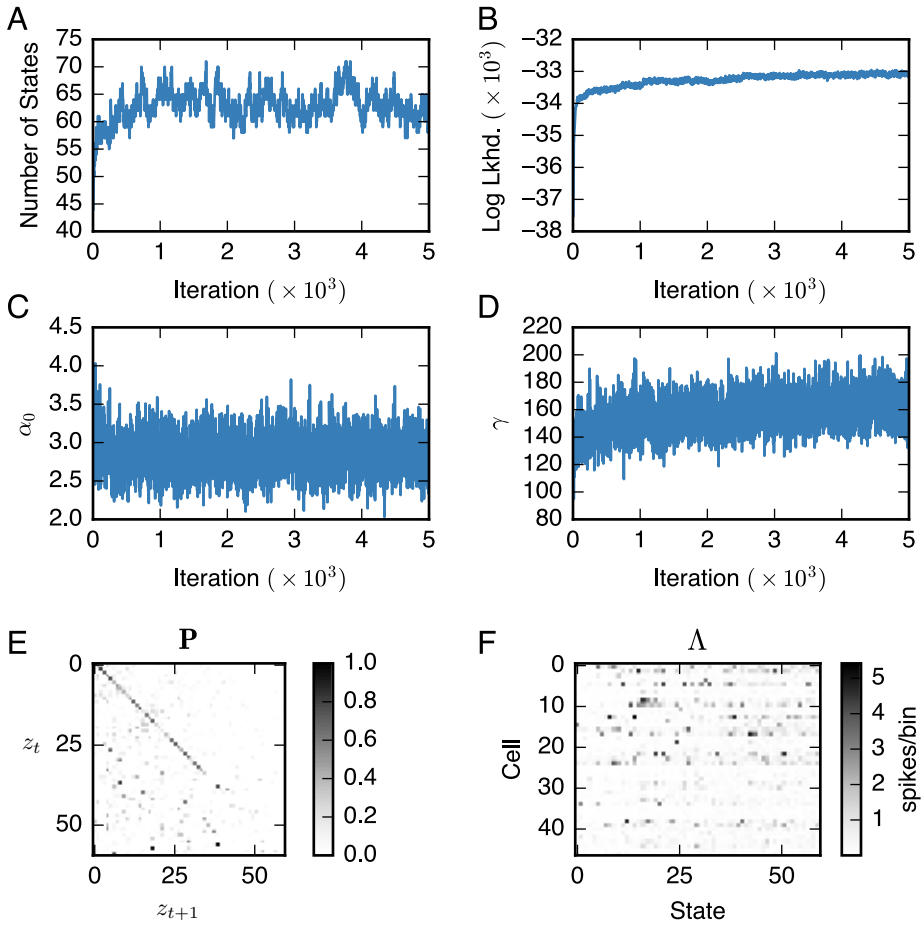

Figure 7.6: Estimation result from HDP-HMM (Gibbs) for the rat hippocampal ensemble spike data. (A) The total number of states (solid blue) slowly increases as states are allocated for a small number of time bins. The number of states converges after 2500 iterations. (B) The log likelihood of the training data grows consistently as highly specific states are added. (C, D) The concentration parameters, α_0 and γ also converge after 2500 iterations. (E, F) The inferred state transition matrix and firing rate samples drawn from the last iteration.

are shown in Fig. 7.5B). In Fig. 7.5C, 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 distribu-

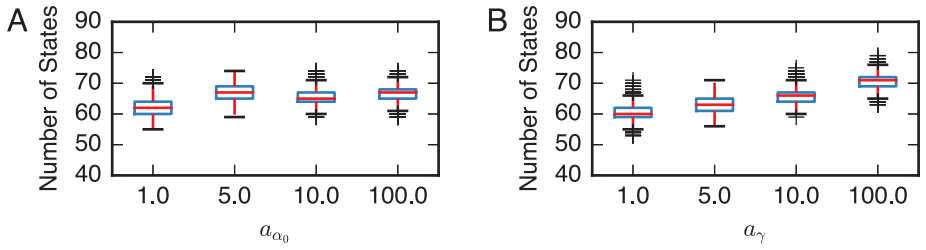

Figure 7.7: Measuring the effect of concentration hyperparameters on the number of inferred latent states. We find that the concentration hyperparameters of the gamma priors on the concentration parameters, α_0 and γ, have a minimal effect.

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

$$\hat{x}_t = \sum_{k=1}^{K} \bar{x}_k \Pr(\boldsymbol{z}_t = k), \qquad \hat{y}_t = \sum_{k=1}^{K} \bar{y}_k \Pr(\boldsymbol{z}_t = k),$$

where $\bar{x}_k$ and $\bar{y}_k$ denote the average location of the rat while in inferred state k (corresponding to the black dots in Fig. 7.5A). 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. 7.6 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 evaluated the sensitivity of these model fits to the choice of hyperparameters. For the HDP-HMM fit via Gibbs sampling with fixed κ_n, the primary hyperparameters of interest are the concentration hyperparameters, a_{α_0} and a_γ in Eq. 7.3, where we have assumed $\alpha_0 \sim \mathrm{Gamma}(a_{\alpha_0}, 1)$ and $\gamma \sim \mathrm{Gamma}(a_\gamma, 1)$. Figure 7.7 shows the inferred number of states as we vary these two hyperparameters over orders of magnitude. We found 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

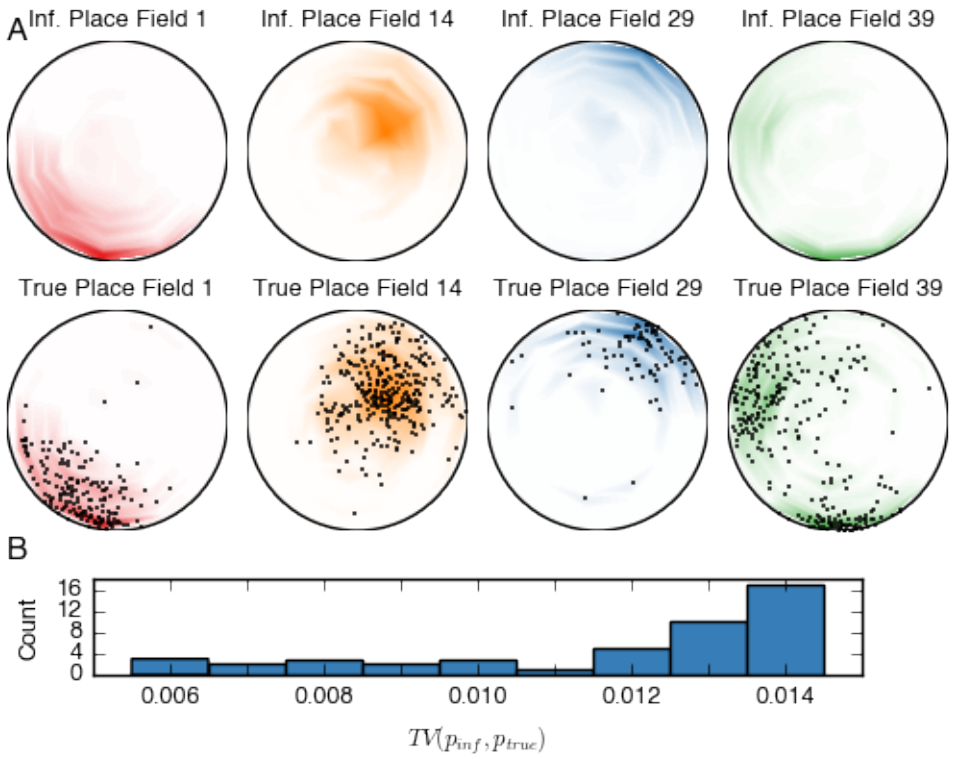

Figure 7.8: Comparison of inferred and true place fields for four randomly selected hippocampal neurons. The inferred place field (top row) for neuron n is a combination of location distributions for each state k weighted by the inferred firing rates $\lambda_{k,n}$, whereas the true place field (bottom row) for neuron n is a histogram of locations in which neuron n 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.

hippocampal neurons. To do so, we combine the state-location maps (Fig. 7.5B) with the firing rate of the individual neuron in those states (Fig. 7.6F) 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. 7.8. 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 neuron n fired shown by black dots.

Extensions

Hidden Semi-Markovian Models A striking feature of the inferred state transition matrix in Fig. 7.6E 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, "sticky" HMMs and HDP-HMMs accomplish a similar effect (Fox et al., 2008).

Dependent Observation Models The HMMs in this chapter used conditionally independent Poisson observations. Given the latent state, each neuron fires independently of the others, and also independently of its previous spike counts. One way to extend these models is by introducing dependencies in the observation models. For example, we can combine the autoregressive models of previous chapters with the discrete latent states of an HMM with a model of the form,

$$p(\boldsymbol{s}_t \mid z_t) = \prod_{n=1}^{N} \text{Poisson}(s_{t,n} \mid \lambda_{t,n}),$$
$$\lambda_{t,n} = g\left(\psi_n^{(z_t)} + \boldsymbol{w}_n^{(z_t)} \boldsymbol{s}_{t-1}\right).$$

As in previous chapters, this can easily be extended to higher-order autoregressive models. Expectation-maximization algorithms for this type of model were developed by Escola et al. (2011). Alternatively, we can use the Pólya-gamma augmentation schemes of Chapter 5, and we have presented a preliminary versions of this approach in Johnson et al. (2015).

Input-Output HMMs Hidden Markov models are "open loop" systems: the next state depends only on the previous state. In practice, it is natural to expect that transitions are not only state-dependent but also a function of some external variables. For example, in the hippocampus where states correspond to actual locations, whether or not the rat transitions

into a state may depend on instantaneous properties of that location. If more complex experimental setups there may be food or obstacles in the environment that affect where the rat goes next.

These types of external variables can be modeled with an input-output HMM (IOHMM) (Bengio and Frasconi, 1995). Suppose we have an external input, $\boldsymbol{u}_t \in \mathbb{R}^D$. We can model the transition probability as,

$$\boldsymbol{\pi}^{(k)} \propto \exp\{\boldsymbol{\psi}^{(k)} + \boldsymbol{W}\boldsymbol{u}_t\},$$

that is, as a "soft-max" function of a baseline probability plus a weighted combination of input covariates. Performing inference in this type of model is not much more challenging than in the standard HMM. When sampling the latent states, we simply compute the instantaneous transition probabilities for each time step. In order to update the transition weights, $\boldsymbol{W}$, and the baseline probabilities, $\boldsymbol{\psi}^{(k)}$, we can either use HMC or our recently developed Pólya-gamma augmentation scheme for multinomial models (Linderman and Johnson, 2015).

Conclusion

This chapter explored the idea of dynamic latent states underlying neural activity. Specifically, we developed Bayesian nonparametric hidden Markov models (HDP-HMMs) and corresponding MCMC and variational inference algorithms. Since these models can be quite sensitive to hyperparameter settings, we performed a thorough assessment of inference results on both synthetic data and real recordings from rat hippocampal place cells. In the next chapter, we will build on these ideas, developing more sophisticated latent state space models with a mix of discrete and continuous latent states. As we will see, HMMs are only one in a hierarchy of state space models.

References

Yashar Ahmadian, Jonathan W Pillow, and Liam Paninski. Efficient Markov chain Monte Carlo methods for decoding neural spike trains. *Neural Computation*, 23(1):46–96, 2011.

Misha B Ahrens, Michael B Orger, Drew N Robson, Jennifer M Li, and Philipp J Keller. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. *Nature Methods*, 10(5):413–420, 2013.

Laurence Aitchison and Peter E Latham. Synaptic sampling: A connection between PSP variability and uncertainty explains neurophysiological observations. *arXiv preprint arXiv:1505.04544*, 2015.

Laurence Aitchison and Máté Lengyel. The Hamiltonian brain. *arXiv preprint arXiv:1407.0973*, 2014.

David J Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

Charles H Anderson and David C Van Essen. Neurobiological computational systems. *Computational Intelligence Imitating Life*, pages 1–11, 1994.

Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

Michael J Barber, John W Clark, and Charles H Anderson. Neural representation of probabilistic information. *Neural Computation*, 15(8):1843–64, August 2003.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden Markov model. *Advances in Neural Information Processing Systems 14*, pages 577–585, 2002.

Jeffrey M Beck and Alexandre Pouget. Exact inferences in a neural implementation of a hidden Markov model. *Neural Computation*, 19(5):1344–1361, 2007.

Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Marginalization in neural circuits with divisive normalization. *The Journal of Neuroscience*, 31(43):15310–15319, 2011.

Jeffrey M Beck, Katherine A Heller, and Alexandre Pouget. Complex inference in neural circuits with probabilistic population codes and topic models. *Advances in Neural Information Processing Systems*, pages 3059–3067, 2012.

Yoshua Bengio and Paolo Frasconi. An input output HMM architecture. *Advances in Neural Information Processing Systems*, pages 427–434, 1995.

Pietro Berkes, Gergo Orbán, Máté Lengyel, and József Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013): 83–7, January 2011.

Gordon J Berman, Daniel M Choi, William Bialek, and Joshua W Shaevitz. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of The Royal Society Interface*, 11(99):20140672, 2014.

Philippe Biane, Jim Pitman, and Marc Yor. Probability laws related to the Jacobi theta and Riemann zeta functions, and Brownian excursions. *Bulletin of the American Mathematical Society*, 38(4):435–465, 2001.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232, 2014.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Carolyn R Block and Richard Block. *Street gang crime in Chicago*. US Department of Justice, Office of Justice Programs, National Institute of Justice, 1993.

Carolyn R Block, Richard Block, and Illinois Criminal Justice Information Authority. Homicides in Chicago, 1965-1995. ICPSR06399-v5. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], July 2005.

Charles Blundell, Katherine A Heller, and Jeffrey M Beck. Modelling reciprocating relationships with Hawkes processes. *Advances in Neural Information Processing Systems*, pages 2600–2608, 2012.

George EP Box. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, pages 383–430, 1980.

David H Brainard and William T Freeman. Bayesian color constancy. *Journal of the Optical Society of America A*, 14(7):1393–1411, 1997.

Kevin L Briggman, Henry DI Abarbanel, and William B Kristan. Optical imaging of neuronal populations during decision-making. *Science*, 307(5711):896–901, 2005.

David R. Brillinger. Maximum likelihood analysis of spike trains of interacting nerve cells. *Biological Cybernetics*, 59(3):189–200, August 1988.

David R Brillinger, Hugh L Bryant Jr, and Jose P Segundo. Identification of synaptic interactions. *Biological Cybernetics*, 22(4):213–228, 1976.

Michael Bryant and Erik B Sudderth. Truly nonparametric online variational inference for hierarchical Dirichlet processes. *Advances in Neural Information Processing Systems 25*, pages 2699–2707, 2012.

Lars Buesing, Johannes Bill, Bernhard Nessler, and Wolfgang Maass. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, 7(11):e1002211, November 2011.

Lars Buesing, Jakob H. Macke, and Maneesh Sahani. Learning stable, regularised latent models of neural population dynamics. *Network: Computation in Neural Systems*, 23: 24–47, 2012a.

Lars Buesing, Jakob H Macke, and Maneesh Sahani. Spectral learning of linear dynamics from generalised-linear observations with application to neural population data. *Advances in Neural Information Processing Systems*, pages 1682–1690, 2012b.

Lars Buesing, Timothy A Machado, John P Cunningham, and Liam Paninski. Clustered factor analysis of multineuronal spike data. *Advances in Neural Information Processing Systems*, pages 3500–3508, 2014.

Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

Santiago Ramón Cajal. *Textura del Sistema Nervioso del Hombre y los Vertebrados*, volume 1. Imprenta y Librería de Nicolás Moya, Madrid, Spain, 1899.

Natalia Caporale and Yang Dan. Spike timing-dependent plasticity: a Hebbian learning rule. *Annual Review of Neuroscience*, 31:25–46, 2008.

Nick Chater and Christopher D Manning. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7):335–344, 2006.

Zhe Chen, Fabian Kloosterman, Emery N Brown, and Matthew A Wilson. Uncovering spatial topology represented by rat hippocampal population neuronal codes. *Journal of Computational Neuroscience*, 33(2):227–255, 2012.

Zhe Chen, Stephen N Gomperts, Jun Yamamoto, and Matthew A Wilson. Neural representation of spatial topology in the rodent hippocampus. *Neural Computation*, 26(1): 1–39, 2014.

Sharat Chikkerur, Thomas Serre, Cheston Tan, and Tomaso Poggio. What and where: A Bayesian inference theory of attention. *Vision Research*, 50(22):2233–2247, 2010.

Yoon Sik Cho, Aram Galstyan, Jeff Brantingham, and George Tita. Latent point process models for spatial-temporal networks. *arXiv:1302.2671*, 2013.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

Aaron C Courville, Nathaniel D Daw, and David S Touretzky. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10(7):294–300, 2006.

Ronald L Cowan and Charles J Wilson. Spontaneous firing patterns and axonal projections of single corticostriatal neurons in the rat medial agranular cortex. *Journal of Neurophysiology*, 71(1):17–32, 1994.

W Maxwell Cowan, Thomas C Südhof, and Charles F Stevens. *Synapses*. Johns Hopkins University Press, 2003.

Mary Kathryn Cowles and Bradley P Carlin. Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91: 883–904, 1996.

John P Cunningham and Byron M Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014.

Paul Dagum and Michael Luby. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60(1):141–153, 1993.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: Volume I: Elementary Theory and Methods*. Springer Science & Business Media, 2 edition, 2003.

Peter Dayan and Larry F Abbott. *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. MIT Press, 2001.

Peter Dayan and Joshua A Solomon. Selective Bayes: Attentional load and crowding. *Vision Research*, 50(22):2248–2260, 2010.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

Sophie Deneve. Bayesian spiking neurons I: inference. *Neural Computation*, 20(1):91–117, January 2008.

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, USA, 1986.

Christopher DuBois, Carter Butts, and Padhraic Smyth. Stochastic block modeling of relational event dynamics. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 238–246, 2013.

Seif Eldawlatly, Yang Zhou, Rong Jin, and Karim G Oweiss. On the use of dynamic Bayesian networks in reconstructing functional neuronal networks from spike train ensembles. *Neural Computation*, 22(1):158–189, 2010.

Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

Sean Escola, Alfredo Fontanini, Don Katz, and Liam Paninski. Hidden Markov models for the stimulus-response relationships of multistate neural systems. *Neural Computation*, 23(5):1071–1132, 2011.

Warren John Ewens. Population genetics theory—the past and the future. In S. Lessard, editor, *Mathematical and Statistical Developments of Evolutionary Theory*, pages 177–227. Springer, 1990.

Daniel E Feldman. The spike-timing dependence of plasticity. *Neuron*, 75(4):556–71, August 2012.

Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991.

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

Christopher R Fetsch, Amanda H Turner, Gregory C DeAngelis, and Dora E Angelaki. Dynamic reweighting of visual and vestibular cues during self-motion perception. *The Journal of Neuroscience*, 29(49):15601–15612, 2009.

Christopher R Fetsch, Alexandre Pouget, Gregory C DeAngelis, and Dora E Angelaki. Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, 15(1):146–154, 2012.

József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14(3):119–130, 2010.

Alyson K Fletcher, Sundeep Rangan, Lav R Varshney, and Aniruddha Bhargava. Neural reconstruction with approximate message passing (neuramp). *Advances in Neural Information Processing Systems*, pages 2555–2563, 2011.

Emily B Fox. *Bayesian nonparametric learning of complex dynamical phenomena*. PhD thesis, Massachusetts Institute of Technology, 2009.

Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. An HDP-HMM for systems with state persistence. *Proceedings of the International Conference on Machine Learning*, pages 312–319, 2008.

Jeremy Freeman, Greg D Field, Peter H Li, Martin Greschner, Deborah E Gunning, Keith Mathieson, Alexander Sher, Alan M Litke, Liam Paninski, Eero P Simoncelli, et al. Mapping nonlinear receptive field structure in primate retina at single cone resolution. *eLife*, 4:e05241, 2015.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, 11(2):127–38, February 2010.

Karl J Friston. Functional and effective connectivity in neuroimaging: a synthesis. *Human Brain Mapping*, 2(1-2):56–78, 1994.

Deep Ganguli and Eero P Simoncelli. Implicit encoding of prior probabilities in optimal neural populations. *Advances in Neural Information Processing Systems*, pages 6–9, 2010.

Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology*, 32:148–155, 2015.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC press, 3rd edition, 2013.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

Felipe Gerhard, Tilman Kispersky, Gabrielle J Gutierrez, Eve Marder, Mark Kramer, and Uri Eden. Successful reconstruction of a physiological circuit with known connectivity from spiking activity alone. *PLoS Computational Biology*, 9(7):e1003138, 2013.

Samuel J Gershman, Matthew D Hoffman, and David M Blei. Nonparametric variational inference. *Proceedings of the International Conference on Machine Learning*, pages 663–670, 2012a.

Samuel J Gershman, Edward Vul, and Joshua B Tenenbaum. Multistability and perceptual inference. *Neural Computation*, 24(1):1–24, 2012b.

Sebastian Gerwinn, Jakob Macke, Matthias Seeger, and Matthias Bethge. Bayesian inference for spiking neuron models with a sparsity prior. *Advances in Neural Information Processing Systems*, pages 529–536, 2008.

Charles J Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, pages 473–483, 1992.

Walter R Gilks. *Markov Chain Monte Carlo*. Wiley Online Library, 2005.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1019–1028, 2010.

Noah Goodman, Vikash Mansinghka, Daniel M Roy, Keith Bonawitz, and Joshua B Tenenbaum. Church: a language for generative models. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 220––229, 2008.

Noah D Goodman, Joshua B Tenenbaum, and Tobias Gerstenberg. Concepts in a probabilistic language of thought. Technical report, Center for Brains, Minds and Machines (CBMM), 2014.

Agnieszka Grabska-Barwinska, Jeff Beck, Alexandre Pouget, and Peter Latham. Demixing odors-fast inference in olfaction. *Advances in Neural Information Processing Systems*, pages 1968–1976, 2013.

SG Gregory, KF Barlow, KE McLay, R Kaul, D Swarbreck, A Dunham, CE Scott, KL Howe, K Woodfine, CCA Spencer, et al. The DNA sequence and biological annotation of human chromosome 1. *Nature*, 441(7091):315–321, 2006.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. In Ron Sun, editor, *The Cambridge Handbook of Computational Psychology*. Cambridge University Press, 2008.

Roger B Grosse, Chris J Maddison, and Ruslan R Salakhutdinov. Annealing between distributions by averaging moments. *Advances in Neural Information Processing Systems*, pages 2769–2777, 2013.

Roger B Grosse, Zoubin Ghahramani, and Ryan P Adams. Sandwiching the marginal likelihood using bidirectional Monte Carlo. *arXiv preprint arXiv:1511.02543*, 2015.

Yong Gu, Dora E Angelaki, and Gregory C DeAngelis. Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience*, 11(10):1201–1210, 2008.

Fangjian Guo, Charles Blundell, Hanna Wallach, and Katherine A Heller. The Bayesian echo chamber: Modeling influence in conversations. *arXiv preprint arXiv:1411.2674*, 2014.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83, 1971.

Moritz Helmstaedter, Kevin L Briggman, Srinivas C Turaga, Viren Jain, H Sebastian Seung, and Winfried Denk. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500(7461):168–174, 2013.

Geoffrey E Hinton. How neural networks learn from experience. *Scientific American*, 1992.

Geoffrey E Hinton and Terrence J Sejnowski. Optimal perceptual inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1983.

Daniel R Hochbaum, Yongxin Zhao, Samouil L Farhi, Nathan Klapoetke, Christopher A Werley, Vikrant Kapoor, Peng Zou, Joel M Kralj, Dougal Maclaurin, Niklas Smedemark-Margulies, et al. All-optical electrophysiology in mammalian neurons using engineered microbial rhodopsins. *Nature Methods*, 2014.

Peter D Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. *Advances in Neural Information Processing Systems*, 20:1–8, 2008.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Douglas N. Hoover. Relations on probability spaces and arrays of random variables. Technical report, Institute for Advanced Study, Princeton, 1979.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

Patrik O Hoyer and Aapo Hyvarinen. Interpreting neural response variability as Monte Carlo sampling of the posterior. *Advances in neural information processing systems*, pages 293–300, 2003.

Yanping Huang and Rajesh P. N. Rao. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):580–593, September 2011.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962.

Hemant Ishwaran and Mahmoud Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

Tomoharu Iwata, Amar Shah, and Zoubin Ghahramani. Discovering latent influence in online social activities via shared cascade Poisson processes. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 266–274, 2013.

Mehrdad Jazayeri and Michael N Shadlen. Temporal context calibrates interval timing. *Nature Neuroscience*, 13(8):1020–1026, 2010.

Mehrdad Jazayeri and Michael N Shadlen. A neural mechanism for sensing and reproducing a time interval. *Current Biology*, 25(20):2599–2609, 2015.

Matthew J Johnson. *Bayesian time series models and scalable inference*. PhD thesis, Massachusetts Institute of Technology, June 2014.

Matthew J Johnson and Alan S Willsky. Bayesian nonparametric hidden semi-Markov models. *Journal of Machine Learning Research*, 14(1):673–701, 2013.

Matthew J Johnson and Alan S Willsky. Stochastic variational inference for Bayesian time series models. *Proceedings of the International Conference on Machine Learning*, 32:1854–1862, 2014.

Matthew J Johnson, Scott W Linderman, Sandeep R Datta, and Ryan P Adams. Discovering switching autoregressive dynamics in neural spike train recordings. *Computational and Systems Neuroscience (Cosyne) Abstracts*, 2015.

Lauren M Jones, Alfredo Fontanini, Brian F Sadacca, Paul Miller, and Donald B Katz. Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. *Proceedings of the National Academy of Sciences*, 104(47):18772–18777, 2007.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

Eric R Kandel, James H Schwartz, Thomas M Jessell, et al. *Principles of neural science*, volume 4. McGraw-Hill New York, 2000.

David Kappel, Stefan Habenschuss, Robert Legenstein, and Wolfgang Maass. Network plasticity as Bayesian inference. *PLoS Computational Biology*, 11(11):e1004485, 2015a.

David Kappel, Stefan Habenschuss, Robert Legenstein, and Wolfgang Maass. Synaptic sampling: A Bayesian approach to neural network plasticity and rewiring. *Advances in Neural Information Processing Systems*, pages 370–378, 2015b.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

Jason ND Kerr and Winfried Denk. Imaging in vivo: watching the brain in action. *Nature Reviews Neuroscience*, 9(3):195–205, 2008.

Roozbeh Kiani and Michael N Shadlen. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, 324(5928):759–64, May 2009.

John F. C. Kingman. *Poisson Processes (Oxford Studies in Probability)*. Oxford University Press, January 1993. ISBN 0198536933.

David C Knill and Whitman Richards. *Perception as Bayesian inference*. Cambridge University Press, 1996.

Konrad P Körding and Daniel M Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–7, January 2004.

Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, and David Blei. Automatic variational inference in Stan. *Advances in Neural Information Processing Systems*, pages 568–576, 2015.

Stephen W Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, 16(1):37–68, 1953.

Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Kenneth W Latimer, Jacob L Yates, Miriam LR Meister, Alexander C Huk, and Jonathan W Pillow. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science*, 349(6244):184–187, 2015.

Tai Sing Lee and David Mumford. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7):1434–1448, 2003.

Robert Legenstein and Wolfgang Maass. Ensembles of spiking neurons with noise support optimal probabilistic inference in a dynamically changing environment. *PLoS Computational Biology*, 10(10):e1003859, 2014.

William C Lemon, Stefan R Pulver, Burkhard Höckendorf, Katie McDole, Kristin Branson, Jeremy Freeman, and Philipp J Keller. Whole-central nervous system functional imaging in larval Drosophila. *Nature Communications*, 6, 2015.

Michael S Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation in Neural Systems*, 9(4):R53–R78, 1998.

Percy Liang, Slav Petrov, Michael I Jordan, and Dan Klein. The infinite PCFG using hierarchical Dirichlet processes. *Proceedings of Empirical Methods in Natural Language Processing*, pages 688–697, 2007.

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

Jeff W Lichtman, Jean Livet, and Joshua R Sanes. A technicolour approach to the connectome. *Nature Reviews Neuroscience*, 9(6):417–422, 2008.

Scott W Linderman and Ryan P. Adams. Discovering latent network structure in point process data. *Proceedings of the International Conference on Machine Learning*, pages 1413–1421, 2014.

Scott W Linderman and Ryan P Adams. Scalable Bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*, 2015.

Scott W Linderman and Ryan P Johnson, Matthew Jand Adams. Dependent multinomial models made easy: Stick-breaking with the Pólya-gamma augmentation. *Advances in Neural Information Processing Systems*, pages 3438–3446, 2015.

Scott W Linderman, Christopher H Stock, and Ryan P Adams. A framework for studying synaptic plasticity with neural spike train data. *Advances in Neural Information Processing Systems*, pages 2330–2338, 2014.

Scott W Linderman, Ryan P Adams, and Jonathan W Pillow. Inferring structured connectivity from spike trains under negative-binomial generalized linear models. *Computational and Systems Neuroscience (Cosyne) Abstracts*, 2015.

Scott W Linderman, Matthew J Johnson, Matthew W Wilson, and Zhe Chen. A nonparametric Bayesian approach to uncovering rat hippocampal population codes during spatial navigation. *Journal of Neuroscience Methods*, 263:36–47, 2016a.

Scott W Linderman, Aaron Tucker, and Matthew J Johnson. Bayesian latent state space models of neural activity. *Computational and Systems Neuroscience (Cosyne) Abstracts*, 2016b.

Fredrik Lindsten, Michael I Jordan, and Thomas B Schön. Ancestor sampling for particle Gibbs. *Advances in Neural Information Processing Systems*, pages 2600–2608, 2012.

Shai Litvak and Shimon Ullman. Cortical circuitry implementing graphical models. *Neural Computation*, 21(11):3010–3056, 2009.

James Robert Lloyd, Peter Orbanz, Zoubin Ghahramani, and Daniel M Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. *Advances in Neural Information Processing Systems*, 2012.

Wei Ji Ma and Mehrdad Jazayeri. Neural coding of uncertainty and probability. *Annual Review of Neuroscience*, 37:205–220, 2014.

Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–8, November 2006.

David JC MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

Jakob H Macke, Lars Buesing, John P Cunningham, M Yu Byron, Krishna V Shenoy, and Maneesh Sahani. Empirical models of spiking in neural populations. *Advances in neural information processing systems*, pages 1350–1358, 2011.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Vikash Mansinghka, Daniel Selsam, and Yura Perov. Venture: a higher-order probabilistic programming platform with programmable inference. *arXiv preprint arXiv:1404.0099*, 2014.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press, 1982.

Paul Miller and Donald B Katz. Stochastic transitions between neural states in taste processing and decision-making. *The Journal of Neuroscience*, 30(7):2559–2570, 2010.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023—-1032, 1988.

Shakir Mohamed, Zoubin Ghahramani, and Katherine A Heller. Bayesian and L1 approaches for sparse unsupervised learning. *Proceedings of the International Conference on Machine Learning*, pages 751–758, 2012.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

Michael L Morgan, Gregory C DeAngelis, and Dora E Angelaki. Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron*, 59(4):662–673, 2008.

Abigail Morrison, Markus Diesmann, and Wulfram Gerstner. Phenomenological models of synaptic plasticity based on spike timing. *Biological Cybernetics*, 98(6):459–478, 2008.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Radford M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, 2010.

John A Nelder and R Jacob Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

Bernhard Nessler, Michael Pfeiffer, Lars Buesing, and Wolfgang Maass. Bayesian computation emerges in generic cortical microcircuits through spike-timing-dependent plasticity. *PLoS Computational Biology*, 9(4):e1003037, 2013.

Mark EJ Newman. The structure and function of complex networks. *Society for Industrial and Applied Mathematics (SIAM) Review*, 45(2):167–256, 2003.

Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.

Seung Wook Oh, Julie A Harris, Lydia Ng, Brent Winslow, Nicholas Cain, Stefan Mihalas, Quanxin Wang, Chris Lau, Leonard Kuan, Alex M Henry, et al. A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–214, 2014.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

John O'Keefe and Lynn Nadel. *The Hippocampus as a Cognitive Map*, volume 3. Clarendon Press, 1978.

Peter Orbanz and Daniel M Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):437–461, 2015.

Peter Orbanz and Yee Whye Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pages 81–89. Springer, 2011.

Adam M Packer, Darcy S Peterka, Jan J Hirtz, Rohit Prakash, Karl Deisseroth, and Rafael Yuste. Two-photon optogenetics of dendritic spines and neural circuits. *Nature Methods*, 9(12):1202–1205, 2012.

Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, January 2004.

Liam Paninski, Yashar Ahmadian, Daniel Gil Ferreira, Shinsuke Koyama, Kamiar Rahnama Rad, Michael Vidne, Joshua Vogelstein, and Wei Wu. A new look at state-space models for neural data. *Journal of Computational Neuroscience*, 29(1-2):107–126, 2010.

Andrew V Papachristos. Murder by structure: Dominance relations and the social structure of gang homicide. *American Journal of Sociology*, 115(1):74–128, 2009.

Il Memming Park and Jonathan W Pillow. Bayesian spike-triggered covariance analysis. *Advances in Neural Information Processing Systems*, pages 1692–1700, 2011.

Patrick O Perry and Patrick J Wolfe. Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2013.

Biljana Petreska, Byron Yu, John P Cunningham, Gopal Santhanam, Stephen I Ryu, Krishna V Shenoy, and Maneesh Sahani. Dynamical segmentation of single trials from population neural data. *Advances in Neural Information Processing Systems*, pages 756–764, 2011.

David Pfau, Eftychios A Pnevmatikakis, and Liam Paninski. Robust learning of low-dimensional dynamics from large neural ensembles. *Advances in Neural Information Processing Systems*, pages 2391–2399, 2013.

Jonathan W. Pillow and James Scott. Fully Bayesian inference for neural models with negative-binomial spiking. *Advances in Neural Information Processing Systems*, pages 1898–1906, 2012.

Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

Eftychios A Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron*, 2016.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya-gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

Ruben Portugues, Claudia E Feierstein, Florian Engert, and Michael B Orger. Whole-brain activity maps reveal stereotyped, distributed networks for visuomotor behavior. *Neuron*, 81(6):1328–1343, 2014.

Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9):1170–1178, 2013.

Robert Prevedel, Young-Gyu Yoon, Maximilian Hoffmann, Nikita Pak, Gordon Wetzstein, Saul Kato, Tina Schrödel, Ramesh Raskar, Manuel Zimmer, Edward S Boyden, et al. Simultaneous whole-animal 3d imaging of neuronal activity using light-field microscopy. *Nature Methods*, 11(7):727–730, 2014.

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

Adrian E Raftery and Steven Lewis. How many iterations in the Gibbs sampler? *Bayesian Statistics*, pages 763–773, 1992.

Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 33:275––283, 2014.

Rajesh P. N. Rao. Bayesian computation in recurrent neural circuits. *Neural Computation*, 16(1):1–38, January 2004.

Rajesh P. N. Rao. Neural models of Bayesian belief propagation. In *Bayesian brain: Probabilistic approaches to neural computation*, pages 236–264. MIT Press Cambridge, MA, 2007.

Rajesh P. N. Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1): 79–87, January 1999.

Danilo J Rezende, Daan Wierstra, and Wulfram Gerstner. Variational learning for recurrent spiking networks. *Advances in Neural Information Processing Systems*, pages 136–144, 2011.

Fred Rieke, David Warland, Rob de Ruyter van Steveninck, and William Bialek. *Spikes: exploring the neural code*. MIT press, 1999.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Dan Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1):273–302, 1996.

Maneesh Sahani. *Latent variable models for neural data analysis*. PhD thesis, California Institute of Technology, 1999.

Maneesh Sahani and Peter Dayan. Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. *Neural Computation*, 2279:2255–2279, 2003.

Joshua R Sanes and Richard H Masland. The types of retinal ganglion cells: current status and implications for neuronal classification. *Annual Review of Neuroscience*, 38:221–246, 2015.

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

Ben Shababo, Brooks Paige, Ari Pakman, and Liam Paninski. Bayesian inference and online experimental design for mapping neural microcircuits. *Advances in Neural Information Processing Systems*, pages 1304–1312, 2013.

Vahid Shalchyan and Dario Farina. A non-parametric Bayesian approach for clustering and tracking non-stationarities of neural spikes. *Journal of Neuroscience Methods*, 223: 85–91, 2014.

Lei Shi and Thomas L Griffiths. Neural implementation of hierarchical Bayesian inference by importance sampling. *Advances in Neural Information Processing Systems*, 2009.

Yousheng Shu, Andrea Hasenstaub, and David A McCormick. Turning on and off recurrent balanced cortical activity. *Nature*, 423(6937):288–293, 2003.

Jack W Silverstein. The spectral radii and norms of large dimensional non-central random matrices. *Stochastic Models*, 10(3):525–532, 1994.

Aleksandr Simma and Michael I Jordan. Modeling events with cascades of Poisson processes. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010.

Eero P Simoncelli. Optimal estimation in sensory systems. *The Cognitive Neurosciences, IV*, 2009.

Anne C Smith and Emery N Brown. Estimating a state-space model from point process observations. *Neural Computation*, 15(5):965–91, May 2003.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

Sen Song, Kenneth D Miller, and Lawerence F Abbott. Competitive Hebbian learning through spike-timing-dependent synaptic plasticitye. *Nature Neuroscience*, 3(9):919–26, September 2000. ISSN 1097-6256.

Daniel Soudry, Suraj Keshri, Patrick Stinson, Min-hwan Oh, Garud Iyengar, and Liam Paninski. Efficient "shotgun" inference of neural connectivity from highly sub-sampled

activity data. *PLoS Computational Biology*, 11(10):1–30, 10 2015. doi: 10.1371/journal.pcbi.1004464.

Olaf Sporns, Giulio Tononi, and Rolf Kötter. The human connectome: a structural description of the human brain. *PLoS Computational Biology*, 1(4):e42, 2005.

Olav Stetter, Demian Battaglia, Jordi Soriano, and Theo Geisel. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS Computational Biology*, 8(8):e1002653, 2012.

Ian Stevenson and Konrad Koerding. Inferring spike-timing-dependent plasticity from spike train data. *Advances in Neural Information Processing Systems*, pages 2582–2590, 2011.

Ian H Stevenson, James M Rebesco, Nicholas G Hatsopoulos, Zach Haga, Lee E Miller, and Konrad P Körding. Bayesian inference of functional connectivity and network structure from spikes. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 17(3):203–213, 2009.

Alan A Stocker and Eero P Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–85, April 2006.

Yee Whye Teh and Michael I Jordan. Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics*, pages 158–207, 2010.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

Joshua B Tenenbaum, Thomas L Griffiths, and Charles Kemp. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318, 2006.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

Wilson Truccolo, Uri T. Eden, Matthew R. Fellows, John P. Donoghue, and Emery N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, 93(2):1074–1089, 2005. doi: 10.1152/jn.00697.2004.

Philip Tully, Matthias Hennig, and Anders Lansner. Synaptic and nonsynaptic plasticity approximating probabilistic inference. *Frontiers in Synaptic Neuroscience*, 6(8), 2014.

Srini Turaga, Lars Buesing, Adam M Packer, Henry Dalgleish, Noah Pettit, Michael Hausser, and Jakob Macke. Inferring neural population dynamics from multiple partial recordings of the same neural circuit. *Advances in Neural Information Processing Systems*, pages 539–547, 2013.

Leslie G Valiant. *Circuits of the Mind*. Oxford University Press, Inc., 1994.

Leslie G Valiant. Memorization and association on a realistic neural model. *Neural Computation*, 17(3):527–555, 2005.

Leslie G Valiant. A quantitative theory of neural computation. *Biological Cybernetics*, 95 (3):205–211, 2006.

Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. Beam sampling for the infinite hidden Markov model. *Proceedings of the International Conference on Machine Learning*, pages 1088–1095, 2008.

Michael Vidne, Yashar Ahmadian, Jonathon Shlens, Jonathan W Pillow, Jayant Kulkarni, Alan M Litke, EJ Chichilnisky, Eero Simoncelli, and Liam Paninski. Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. *Journal of Computational Neuroscience*, 33(1):97–121, 2012.

Joshua T Vogelstein, Brendon O Watson, Adam M Packer, Rafael Yuste, Bruno Jedynak, and Liam Paninski. Spike inference from calcium imaging using sequential Monte Carlo methods. *Biophysical Journal*, 97(2):636–655, 2009.

Joshua T Vogelstein, Adam M Packer, Timothy A Machado, Tanya Sippy, Baktash Babadi, Rafael Yuste, and Liam Paninski. Fast nonnegative deconvolution for spike train

inference from population calcium imaging. *Journal of Neurophysiology*, 104(6):3691–3704, 2010.

Hermann von Helmholtz and James Powell Cocke Southall. *Treatise on Physiological Optics: Translated from the 3rd German Ed.* Optical Society of America, 1925.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

Yair Weiss, Eero P Simoncelli, and Edward H Adelson. Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6):598–604, 2002.

Mike West, P Jeff Harrison, and Helio S Migon. Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83, 1985.

John G White, Eileen Southgate, J Nichol Thomson, and Sydney Brenner. The structure of the nervous system of the nematode Caenorhabditis elegans: the mind of a worm. *Philosophical Transactions of the Royal Society of London: Series B (Biological Sciences)*, 314:1–340, 1986.

Louise Whiteley and Maneesh Sahani. Attention in a Bayesian framework. *Frontiers in Human Neuroscience*, 6, 2012.

Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.

Jesse Windle, Nicholas G Polson, and James G Scott. Sampling Pólya-gamma random variates: alternate and approximate techniques. *arXiv preprint arXiv:1405.0506*, 2014.

Frank Wood and Michael J Black. A nonparametric Bayesian alternative to spike sorting. *Journal of Neuroscience Methods*, 173(1):1–12, 2008.

Frank Wood, Jan Willem van de Meent, and Vikash Mansinghka. A new approach to probabilistic programming inference. *arXiv preprint arXiv:1507.00996*, 2015.

Tianming Yang and Michael N Shadlen. Probabilistic reasoning by neurons. *Nature*, 447 (7148):1075–80, June 2007.

Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen I. Ryu, Krishna V. Shenoy, and Maneesh Sahani. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology*, 102:614–635, 2009.

Alan Yuille and Daniel Kersten. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, 2006.

Richard S Zemel, Peter Dayan, and Alexandre Pouget. Probabilistic interpretation of population codes. *Neural Computation*, 10(2):403–30, February 1998.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 16, 2013.

Mingyuan Zhou, Lingbo Li, Lawrence Carin, and David B Dunson. Lognormal and gamma mixed negative binomial regression. *Proceedings of the International Conference on Machine Learning*, pages 1343–1350, 2012.