2

Background

This chapter lays the foundation for probabilistic modeling of neural spike trains. We start by introducing the language of generative models, which allow us to formalize, in probabilistic terms, our hypotheses about dynamics and low-dimensional structure. The key ingredients are latent variables that reflect the underlying state of the system and conditional distributions that relate these variables to the observed data. Once we understand the basics of this language, we can begin to articulate hypotheses about dynamical data in the form of generative time series models. Section 2.2 enumerates a few common motifs of time series modeling that will be used throughout this thesis. Finally, given a model and an observed spike train, we can invert the model and reason about the posterior distribution over latent variables using Bayesian inference algorithms such as Markov Chain Monte Carlo and mean field variational inference, which are introduced in Section 2.3. At the end of this chapter, we will have the basic foundation necessary to start looking for structure in neural data. The rest of the thesis will build upon this foundation by developing more sophisticated models and increasingly efficient inference algorithms, and by putting them to use on real neural recordings.

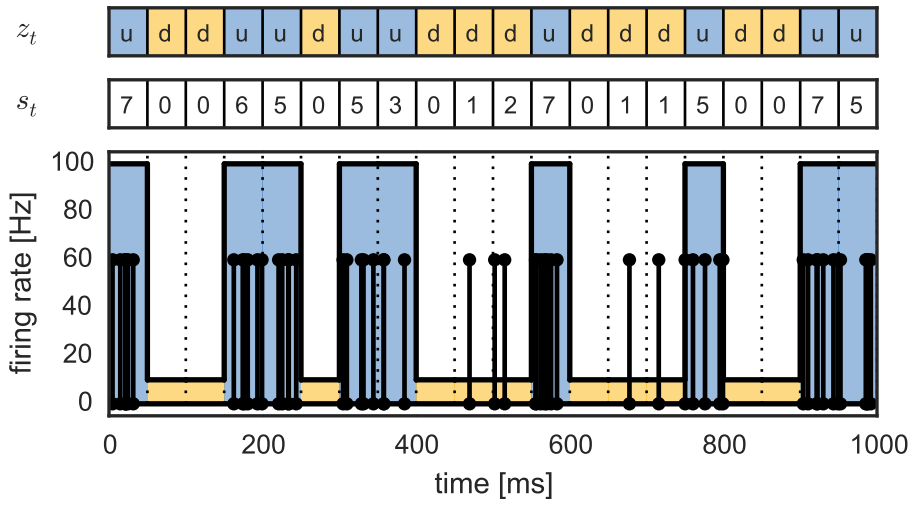

Figure 2.1: A simple neuron that randomly switches between an *up* and a *down* state every 50ms. Here, time bins are colored blue and yellow depending on the latent state, z_t. Each state has an associated firing rate from which a Poisson number of spikes, s_t, is drawn. Precise spike times are uniformly distributed over the 50ms interval.

2.1 Generative Probabilistic Models

Generative probabilistic models tell a story of how data comes to be. While this story never captures every physical detail, it serves as an idealized version, capturing the essence of the system. For example, when modeling a neural spike train, we will ignore the states of individual ion channels and the nonlinear dynamics of membrane potential and instead characterize the instantaneous *firing rate* of a neuron — the probability that a neuron spikes at any moment in time.

As a simple illustration, consider the following generative process. Suppose a neuron has two states, an *up* state and a *down* state. In the *up* state, it spikes at a high rate, say 100Hz, and in the *down* state it fires less frequently, say at 10Hz. Assume that every 50ms the neuron flips a coin to decide its new state and then fires a random number of spikes according to the firing rate associated with that state. For the sake of simplicity, assume the precise spike times are uniformly distributed over the 50ms interval. Once the interval has elapsed, the neuron flips another coin and its rate immediately changes to reflect its new state. Our goal is to infer the latent state of the neuron given the observed spikes.

Clearly, this generative story contains many simplifying assumptions and omits a great

amount of detail. In addition to assuming that spiking is adequately captured by firing rates, the notion that a neuron has only two firing rates and that it randomly switches between them is a gross simplification. Nevertheless, this very simple model captures patterns of spiking that have been observed in actual experiments (Cowan and Wilson, 1994; Shu et al., 2003).

We can formalize this generative story with a probabilistic model that specifies a distribution over latent states and observed spike counts. Let $s_t \in \mathbb{N}$ denote the number of spikes counted in the t-th time bin, and $z_t \in \{\textit{up}, \textit{down}\}$ denote the corresponding state of the neuron. The assumption that states are drawn from a coin flip corresponds to the prior distribution, $z_t \sim \text{Discrete}(\boldsymbol{\pi})$, where $\boldsymbol{\pi} = [\pi_{\textit{up}}, \pi_{\textit{down}}]$ is a nonnegative vector that sums to one and specifies the probability of *up* and *down* states.* Implicitly, we have assumed that $\boldsymbol{\pi} = \left[\frac{1}{2}, \frac{1}{2}\right]$, though this need not be the case. We previously said that the neurons fire a random number of spikes according to their state-dependent firing rate; now we will formalize this by assuming, $s_t \sim \text{Poisson}(\lambda_{z_t} \cdot \Delta t)$, where $\Delta t = 0.05$s, $\lambda_{\textit{up}} = 100$ spikes/s, and $\lambda_{\textit{down}} = 10$ spike/s.

Figure 2.1 shows a neural spike train sampled from this generative model. The time bins are colored blue or yellow depending on whether the neuron is in the *up* or *down* state, respectively. The precise spike times are denoted by black vertical lines with circular endpoints. Above, the vector of observed spike counts, $\boldsymbol{s} = \left[s_1, \ldots, s_T\right]$, and the vector of latent states, $\boldsymbol{z} = \left[z_1, \ldots, z_T\right]$, are shown. We will use this notation throughout the thesis: bold symbols like $\boldsymbol{s}$ will denote arrays of random variables; lowercase bold symbols will typically denote vectors.

The generative procedure defines the *likelihood* of any given set of observed spike counts and corresponding latent states. This can be written as a conditional distribution where the

*The notation $z \sim \mathrm{P}(\theta)$ means that the random variable z is sampled from (or distributed according to) the distribution P, which is parameterized by θ. When we write $\mathrm{P}(z \mid \theta)$ we refer to the density (assuming it exists) of P evaluated at z. A list of commonly used distributions and their densities is given in Appendix A.

state probabilities and firing rates are given. We have,

$$p(\boldsymbol{s}, \boldsymbol{z} \mid \boldsymbol{\pi}, \boldsymbol{\lambda}) = p(\boldsymbol{z} \mid \boldsymbol{\pi})\, p(\boldsymbol{s} \mid \boldsymbol{z}, \boldsymbol{\lambda}) \tag{2.1}$$

$$= \prod_{t=1}^{T} p(z_t \mid \boldsymbol{\pi})\, p(s_t \mid \lambda_{z_t}) \tag{2.2}$$

$$= \prod_{t=1}^{T} \mathrm{Discrete}(z_t \mid \boldsymbol{\pi})\, \mathrm{Poisson}(s_t \mid \lambda_{z_t} \cdot \Delta t). \tag{2.3}$$

Since Δt is a constant, we do not include it as a random variable in the joint distribution or explicitly condition on it.

The probabilistic model specifies the particular factorization of the likelihood implied by the generative story. Eq. 2.1 applies the product rule of probability, and reflects the assumptions that $\boldsymbol{z}$ depends only on $\boldsymbol{\pi}$ and $\boldsymbol{s}$ depends only on $\boldsymbol{z}$ and $\boldsymbol{\lambda}$. In going from (2.1) to (2.2), we have asserted that the latent states z_t and $z_{t'}$ are conditionally independent given $\boldsymbol{\pi}$, and that the spike counts s_t and $s_{t'}$ are conditionally independent given their corresponding latent states and firing rates. This conditional independence assumption, which was implicit in the generative story, becomes explicit when we factor the likelihood into a product over time bins. Eq. 2.3 specifies the functional form of the conditional distributions. When we hypothesize relationships between different variables, we are making assertions about the factorization and the form of the likelihood. In Section 2.2, we explore different patterns of conditional dependence that provide the building blocks of models for dynamic data.

So far, we have assumed that the firing rates and state probabilities are known, but in practice this is a bit unreasonable. To complete the probabilistic model, we need to combine the likelihood function with a *prior distribution* that captures our uncertainty about these parameters. For example, a more reasonable hypothesis is that neurons have two firing rates, and while we do not know their exact values, we can specify a distribution over them, $p(\boldsymbol{\lambda})$. Similarly, we may not know the exact probability of each state, $\boldsymbol{\pi}$, but perhaps we can specify a prior, $p(\boldsymbol{\pi})$, that captures our intuition that the states should be equally likely *a priori*. Putting this all together, we can now write down the *joint distribution* of our

probabilistic model — the product of the likelihood and the prior distributions:

$$p(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\lambda} \mid \alpha, \beta, \gamma) = p(\boldsymbol{s}, \boldsymbol{z} \mid \boldsymbol{\pi}, \boldsymbol{\lambda})\, p(\boldsymbol{\pi} \mid \gamma)\, p(\boldsymbol{\lambda} \mid \alpha, \beta). \tag{2.4}$$

When constructing a probabilistic model, we express these prior intuitions and simultaneously make inference easier by using *conjugate* prior distributions.

Conjugate Prior Distributions

A conjugate prior ensures that the conditional distribution of a parameter, given the data, will have a tractable form. Specifically, the conditional distribution will have the same form as the prior. For example, take the parameter, λ_{up}. If we look at the likelihood as a function of λ_{up} and ignore terms that do not depend on this parameter, we have,

$$\begin{aligned} p(\boldsymbol{s}, \boldsymbol{z} \mid \boldsymbol{\pi}, \boldsymbol{\lambda}) &\propto \prod_{t=1}^{T} \left[\text{Poisson}(s_t \mid \lambda_{up} \cdot \Delta t)\right]^{\mathbb{I}[z_t = up]} \\ &\propto \prod_{t=1}^{T} \left[\lambda_{up}^{s_t}\, e^{-\lambda_{up} \cdot \Delta t}\right]^{\mathbb{I}[z_t = up]} \\ &= \lambda_{up}^{s_{up}}\, e^{-\lambda_{up} \cdot t_{up}}, \end{aligned}$$

where

$$\begin{aligned} s_{up} &= \sum_{t=1}^{T} s_t \cdot \mathbb{I}[z_t = up], \\ t_{up} &= \sum_{t=1}^{T} \Delta t \cdot \mathbb{I}[z_t = up], \end{aligned}$$

and $\mathbb{I}[x]$ is an indicator function that equals one if x evaluates to true and equals zero otherwise.

Now consider a gamma prior distribution,

$$\begin{aligned} p(\lambda_{up} \mid \alpha, \beta) &= \mathrm{Gamma}(\lambda_{up} \mid \alpha, \beta) \\ &\propto \lambda_{up}^{\alpha-1}\, e^{-\lambda_{up} \cdot \beta}. \end{aligned}$$

The conditional distribution over λ_{up} given the observed spike counts, the latent states, and the prior is proportional to the likelihood times the prior. This simplifies to,

$$\begin{aligned} p(\lambda_{up} \mid \boldsymbol{s}, \boldsymbol{z}, \alpha, \beta) &\propto p(\boldsymbol{s}, \boldsymbol{z} \mid \boldsymbol{\pi}, \boldsymbol{\lambda})\, p(\lambda_{up} \mid \alpha, \beta) \\ &\propto \lambda_{up}^{s_{up}+\alpha-1}\, e^{-\lambda_{up}(t_{up}+\beta)} \\ &\propto \mathrm{Gamma}(\lambda_{up} \mid s_{up} + \alpha,\ t_{up} + \beta). \end{aligned}$$

Since both the prior and the the conditional distribution over λ_{up} are in the gamma family, we say gamma prior is conjugate with this product-of-Poissons likelihood. Moreover, the parameters of conditional distribution only depend on $\boldsymbol{s}$ and $\boldsymbol{z}$ through simple *sufficient statistics*, s_{up} and t_{up}. A Dirichlet prior distribution on the state probability, $\mathrm{Dir}(\boldsymbol{\pi} \mid \gamma)$, is similarly conjugate with the product of discrete densities in the likelihood that links $\boldsymbol{\pi}$ and $\boldsymbol{z}$. In fact, conjugate priors exist for all likelihoods in the *exponential family*. These ideas are thoroughly discussed in standard Bayesian statistics and machine learning textbooks like Gelman et al. (2013); Murphy (2012).

Latent Variables, Parameters, and Hyperparameters As our models become increasingly complicated, we will often distinguish between the different types of random variables. The states, $\boldsymbol{z}$, are called *local latent variables* because there is one for each data point. The unknown latent state probability and the firing rates, $\{\boldsymbol{\pi}, \boldsymbol{\lambda}\}$, are either called *parameters* or *global latent variables* because their dimension is fixed. The remaining values, $\{\alpha, \beta, \gamma\}$, are called *hyperparameters*. These are constants that we set prior to performing inference. Typically, these can be tuned by cross-validation, or simply set based on intuition and physical constraints. For conciseness, we will refer to the set of all parameters as $\boldsymbol{\theta}$ and the set of hyperparameters as $\boldsymbol{\eta}$.

2.1.1 Representations of Spike Trains

One of the first decisions we must make is how to represent our data. In this thesis we will focus solely on modeling spike trains, which are sequences of discrete events in time. These spike trains typically come from spike sorting algorithms applied to extracellular recordings from multi-electrode arrays (Lewicki, 1998) or from deconvolution algorithms applied to optically recorded calcium fluorescence traces (Pnevmatikakis et al., 2016; Vogelstein et al., 2010). Reducing the data to a set of spike times often results in enormous compression. Rather than considering electrode potentials, which may be sampled at upwards of 10kHz, or calcium fluorescence traces, which are highly autocorrelated due to the relatively slow dynamics of calcium concentration in cells, we only consider the times of action potentials.

The most general representation of a spike train is a set of real-valued times for each neuron. In Figure 2.1, this corresponds to the temporal locations of each black spike. When there is more than one neuron, we have a set of *marked* spike times, which we call,

$$\mathcal{S} = \{(s_m, c_m)\}_{m=1}^{M} \subset [0, T] \times \{1, \ldots, N\}.$$

Each member of this set consists of a real-valued spike time s_m in the interval $[0, T]$, and an integer, $c_m \in \{1, \ldots, N\}$, that specifies the index of the cell that generated this spike. M is the total number of spikes on all neurons.

This continuous-time representation is warranted when the temporal resolution of the data is considerably higher than the timescale of typical action potentials. For example, multi-elecrode arrays typically have sampling intervals of 0.1ms or smaller, whereas the width of action potentials is on the order of 1ms. This allows us to specify the spike time as an effectively real-valued number.

Sets of discrete events like these are typically modeled as realizations of a *marked point process* (Daley and Vere-Jones, 2003). Such a process is defined by its nonnegative firing rates[†], $\{\lambda_n(t \mid \mathcal{H}_t)\}_{n=1}^{N}$, where $\mathcal{H}_t$ captures the history of the process through time t. For example, the history may include the previous spikes, $\mathcal{H}_t = \{(s_m, c_m) : s_m < t\}$, as well as some external covariates. If we consider a small time window, $[t, t + \Delta t)$, and take the limit as Δt approaches zero, $\lambda_n(t \mid \mathcal{H}_t) \cdot \Delta t$ is the expected number of spikes fired by neuron n

[†]In the point process literature, these firing rates are called *conditional intensity functions*.

in the window $[t, t + \Delta t)$.

The limiting perspective on the conditional intensity functions suggests an alternative, discrete-time representation. Rather than modeling a set of continuous spike times and conditional firing rates, we may instead represent a spike count matrix, $\boldsymbol{S}$, and the corresponding rate matrix, $\boldsymbol{\Lambda}$, where,

$$\boldsymbol{S} = \begin{bmatrix} s_{1,1} & \cdots & s_{1,N} \\ \vdots & & \vdots \\ s_{T,1} & \cdots & s_{T,N} \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{1,1} & \cdots & \lambda_{1,N} \\ \vdots & & \vdots \\ \lambda_{T,1} & \cdots & \lambda_{T,N} \end{bmatrix}.$$

Here, $s_{t,n} \in \mathbb{N}$ denotes the number of spikes fired in the t-th time bin by the n-th neuron, and $\lambda_{t,n} \in \mathbb{R}_+$ denotes the corresponding firing rate. Sometimes, the effects we are interested in studying occur at relatively slow time scales, so discretizing may provide valuable compression while retaining most of the relevant information. For example, if we are studying neural dynamics on the order of minutes, then simply knowing how many spikes occurred each second may provide most of the relevant information, while precise, millisecond-resolution spike timing may be superfluous.

However, the primary reason to discretize spike times into a matrix of counts is that the statistics and machine learning community has developed a much broader set of models for matrices than for sets of continuous time events. In the next section, we will explore a number of common modeling motifs that can be applied to time series data represented as matrices, and many of the chapters of this thesis will focus on extending these motifs in novel ways.

2.2 Motifs of Time Series Models

The art of probabilistic modeling lies in balancing two conflicting concerns: our model should capture as much of the relevant structure in the data as possible, drawing on our intuition and our existing knowledge of the system, yet at the same time we wish to limit the complexity of the model so that we may perform inference efficiently. One way to balance

these goals is to compose our model out of common, well-studied motifs.

Motifs correspond to factorizations of probabilistic models. To visualize these motifs, we represent the probabilistic model in the form of a directed acylcic graph. Each node in the graph corresponds to a random variable, and shaded nodes indicate which variables are observed. The edges represent conditional dependencies. For example, in the mixture model shown in Figure 2.2, the spike count s_2 has incoming edges from the corresponding latent state z_2 and the firing rates, $\boldsymbol{\lambda}$. Thus, the joint probability distribution contains the factor, $p(s_2 \mid z_2, \boldsymbol{\lambda})$. Since the graph is directed and acyclic, we read off the factors starting with the root nodes, $p(\boldsymbol{\pi})$ and $p(\boldsymbol{\lambda})$, and ending with the leaf nodes, $p(s_t \mid z_t, \boldsymbol{\lambda})$. In this way, the graph captures the factorization of the joint probability distribution and specifies a particular subset of all possible joint distributions over this set of variables.

The edges of the graph do not, however, specify the type of the random variable or functional form of the factors. For example, a node may indicate either a discrete or a continuous random variable, and an edge may indicate an arbitrary form of dependence, like a linear relationship. In this way, two models may share the same graph but have fundamentally different interpretations. This is true of mixture models and factor analysis models shown in Figure 2.2. Some patterns of factorization, types, and dependencies are used over and over again and form the building blocks for more complex models. Next, we discuss a few of the common motifs shown in Figure 2.2.

Mixture Models Our working example from Section 2.1 is an instance of a simple mixture model. The firing rate assumes only two values, and the observed spike counts are a mixture of counts drawn from the *up* state and counts from the *down* state. We can easily extend this to populations of neurons and mixtures of more than two states. Suppose there are now K states, such that $z_t \in \{1, \ldots, K\}$. Furthermore, we generalize the rates $\lambda_{\textit{up}}$ and $\lambda_{\textit{down}}$, to vectors of rates, one for each neuron and state. In a slight abuse of notation, let $\boldsymbol{\lambda}_k = [\lambda_{k,1}, \ldots, \lambda_{k,N}]$ denote a vector of rates in which $\lambda_{k,n}$ is the firing rate of the n-th neuron in state k.

In a mixture model, the latent states are discrete, the time bins are conditionally independent, and the dependence of s_t on z_t and $\boldsymbol{\lambda}$ is linear. To see the latter claim, note that that the instantaneous firing rate of neuron n can be written, $\sum_{k=1}^{K} \mathbb{I}[z_t = k] \cdot \lambda_{k,n}$. These

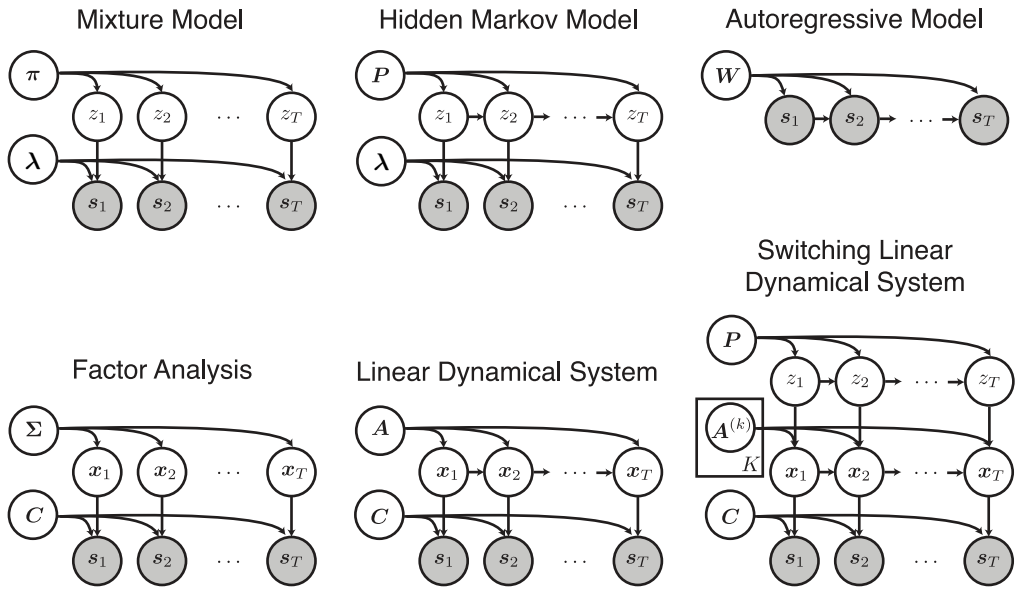

Figure 2.2: Motifs of time series models. By introducing conditional dependencies and layers of random variables, we construct models that reflect sophisticated hypotheses about the structure underlying the data. See Section 2.2 for detailed description.

three properties suggest multiple dimensions along which the mixture model may be generalized.

HIDDEN MARKOV MODELS First, let's address the conditional independence of time bins in the mixture model. According to this model, the distribution over latent states factors into a product, $p(\boldsymbol{z} \mid \boldsymbol{\pi}) = \prod_t p(z_t \mid \boldsymbol{\pi})$. This clearly ignores the temporal dynamics of neural data. Instead, we may hypothesis that latent states obey Markovian dynamics,

$$\begin{aligned} p(\boldsymbol{z} \mid \boldsymbol{\pi}^{(0)}, \boldsymbol{P}) &= p(z_1 \mid \boldsymbol{\pi}^{(0)}) \prod_{t=2}^{T} p(z_t \mid z_{t-1}, \boldsymbol{P}) \\ &= \text{Discrete}(z_1 \mid \boldsymbol{\pi}^{(0)}) \prod_{t=2}^{T} \text{Discrete}(z_t \mid \boldsymbol{\pi}^{(z_{t-1})}), \end{aligned}$$

where $\boldsymbol{\pi}^{(0)} \in [0,1]^K$ 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 $z_{t-1} = k$. This is known as a hidden Markov model (HMM) (Baum and Petrie, 1966; Rabiner, 1989), and the corresponding graphical model is shown in Figure 2.2. Chapter 7 studies some of the challenges involved in selecting the number of states, K, in a nonparametric way.

Autoregressive Models In an HMM, correlations in spike counts from one bin to the next arise from correlations in the underlying latent states. Alternatively, we may directly model the rate as a function of previous spike counts. For example, consider an autoregressive model with linear dynamics,

$$\lambda_{t,n} = \sum_{m=1}^{N} \sum_{d=1}^{D} w_{m \to n}^{(d)} \cdot s_{t-d,m}. \tag{2.5}$$

The weight, $w_{m \to n}^{(d)}$, specifies the influence that spikes on neuron m have on the rate of neuron n at an offset of d time bins in the future. Unlike the HMM, which has an autoregressive model for latent states, here the autoregression governs the rates directly. Moreover, this autoregressive model sums over the spike counts of all neurons over the past D time bins, allowing delayed interactions. Figure 2.2 shows the graph structure of an autoregressive model in the special case that $D = 1$.

In continuous time, autoregressive interactions like these are the basis of the Hawkes process (Hawkes, 1971), a mutually-excitatory point process. Chapter 3 will study these models in great detail, and Chapter 4 will extend the Hawkes process inference algorithms to their discrete time counterparts.

Since the firing rates must be nonnegative, the weights must be as well. That is, $w_{m \to n}^{(d)} \in \mathbb{R}_+$. This implicitly instantiates the hypothesis that interactions between spikes on one

neuron and the rate of another is always excitatory — a spike can never decrease the future firing rate. While this is not the most biologically realistic model given our knowledge of excitatory and inhibitory synapses, it is important to remember that this is simply a descriptive model of firing rate dynamics, and it does not necessarily map onto physical synaptic connections. As we will show, the weights inferred by this type of excitatory autoregressive model can still provide useful insight into the structure of neural activity.

Nonlinear Autoregressive Models In order to capture both excitatory and inhibitory autoregressive weights, we need to introduce a nonlinear function that ensures a nonnegative firing rate. Specifically, assume that,

$$\begin{aligned}\psi_{t,n} &= \sum_{m=1}^{N}\sum_{d=1}^{D} w_{m\to n}^{(d)} \cdot s_{t-d,m}, \\ \lambda_{t,n} &= g\left(\psi_{t,n}\right).\end{aligned}$$

The nonlinear function $g(\cdot) : \mathbb{R} \to \mathbb{R}_+$ maps a real valued "activation," $\psi_{t,n}$, into a nonnegative firing rate, $\lambda_{t,n}$. In this formulation, the weights may be either positive or negative to reflect either excitatory or inhibitory interactions, respectively. In computational neuroscience, this is often called a generalized linear model (GLM) (Paninski, 2004; Truccolo et al., 2005; Pillow et al., 2008), since the linear-nonlinear cascade that links spike history to firing rate is an instance of the GLM commonly used in statistics (Nelder and Baker, 1972). Chapter 5 combines these nonlinear autoregressive models with prior distributions on the underlying network and derives efficient Bayesian inference algorithms to fit them to data.

Factor Models HMM's introduced dynamics to the mixture model and nonlinear autoregressive models generalized the linear functional dependence. Factor models generalize the discrete nature of the random variables with a continuous analogue. For example, consider a model in which the discrete variable $z_t \in \{1, \ldots, K\}$ is replaced by a discrete

probability distribution $\boldsymbol{\pi}_t \in [0,1]^K$. The rate is then a nonnegative combination,

$$\lambda_{t,n} = \sum_{k=1}^{K} \pi_{t,k} \cdot \lambda_{k,n}$$

This is naturally interpreted as a *mixed membership model* in which the rates at each time bin derive from a mixture of discrete latent states with mixing weights $\boldsymbol{\pi}_t$. In text modeling, this motif is the basis of the latent Dirichlet allocation (LDA) model (Blei et al., 2003).

Alternatively, we may replace the discrete latent state with a continuous one, $\boldsymbol{x}_t \in \mathbb{R}^K$. As in the nonlinear autoregressive model, we can retain the linear form and introduce an elementwise nonlinearity to ensure nonnegative firing rates:

$$\begin{aligned} p(\boldsymbol{x}) &= \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{x}_t \mid \mathbf{0}, \boldsymbol{\Sigma}), \\ \psi_{t,n} &= \sum_{k=1}^{K} x_{t,k} \cdot c_{k,n}, \\ \lambda_{t,n} &= g(\psi_{t,n}). \end{aligned}$$

Here, $c_{k,n}$ is an entry in the real valued matrix $\boldsymbol{C} \in \mathbb{R}^{K \times N}$, and $\boldsymbol{\Sigma} = \operatorname{diag}([\sigma_1^2, \ldots, \sigma_K^2])$. This corresponds to a factor analysis model. Unlike standard factor analysis, however, here the observations are discrete spike counts rather than Gaussian observations.

Linear Dynamical Systems In the same way that HMM's extend mixture models with temporal dynamics, linear dynamical systems (LDSs) extend factor models with linear autoregressive dynamics in the latent state. We simply replace the prior on $\boldsymbol{x}$ with a model of the form,

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}_1 \mid \mathbf{0}, \boldsymbol{\Sigma}) \prod_{t=2}^{T} \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{A}\boldsymbol{x}_{t-1}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{A} \in \mathbb{R}^{K \times K}$ specifies the linear dynamics of the latent state. The elementwise nonlinear mapping from latent states to firing rates is the same as in the factor model, but now

the linear autoregressive nature of the dynamics induces correlations in spike counts from one time bin to the next.

Hierarchical Extensions These motifs — continuous and discrete latent states, linear autoregressive dynamics, and nonlinear link functions — provide a foundation for constructing probabilistic models for spike trains. Atop this foundation, we may layer additional random variables reflecting hypotheses about shared structure. For example, a switching linear dynamical system, shown in Figure 2.2 and studied in Chapter 8, combines discrete *and* continuous latent states (Murphy, 2012; Fox, 2009). Likewise, Chapters 3, 4, and 5 consider structured prior distributions on the weights of autoregressive models, and Chapter 7 considers nonparametric Bayesian priors on the number of states in an HMM. Once the dynamics model has been specified, it is easy to test a variety of hypotheses about hierarchical structure. In order to fit these models, however, we need efficient inference algorithms that capitalize on the compositional structure of the model.

2.3 Bayesian Inference

Given an observed spike train, our goal is to compute the posterior distribution over latent variables, $\boldsymbol{z}$, and parameters, $\boldsymbol{\theta}$, of the model. For example, in an HMM the latent variables are the dynamic latent states and the parameters are $\boldsymbol{\theta} = \{\boldsymbol{P}, \boldsymbol{\lambda}\}$, the transition matrix and the firing rates for each latent state. Bayes' rule relates the posterior distribution to the joint distribution of our probabilistic model,

$$p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{s}) = \frac{p(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\theta})}{p(\boldsymbol{s})} = \frac{p(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\theta})}{\int p(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{z} \, \mathrm{d}\boldsymbol{\theta}}. \tag{2.6}$$

Unfortunately, the denominator in Eq. 2.6 involves an integral that is intractable for all but the simplest models. Instead, we must resort to approximate algorithms like Markov Chain Monte Carlo (MCMC) and mean field variational inference. We will briefly describe each of these in turn.

2.3.1 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) algorithms are a workhorse of modern machine learning, and many texts are devoted to the subject (e.g. Geyer, 1992; Gilks, 2005; Robert and Casella, 2013). The fundamental idea is to generate a collection of samples from the posterior distribution and use them to estimate expectations. Specifically, given a set of samples,

$$\left\{ \left(\boldsymbol{z}^{(1)}, \boldsymbol{\theta}^{(1)}\right), \ldots, \left(\boldsymbol{z}^{(L)}, \boldsymbol{\theta}^{(L)}\right) \right\},$$

where

$$\boldsymbol{z}^{(\ell)}, \boldsymbol{\theta}^{(\ell)} \sim p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{s}),$$

we can form a Monte Carlo estimate of the expectation of a function $f(\boldsymbol{z}, \boldsymbol{\theta})$ with respect to the posterior,

$$\mathbb{E}_{p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{s})}\left[f(\boldsymbol{z}, \boldsymbol{\theta})\right] \approx \frac{1}{L} \sum_{\ell=1}^{L} f\left(\boldsymbol{z}^{(\ell)}, \boldsymbol{\theta}^{(\ell)}\right).$$

When the samples are independently drawn from the posterior, the strong law of large numbers states that the Monte Carlo estimate converges to the true expectation almost surely, implying that these Monte Carlo estimates are unbiased. Moreover, if the function f is real-valued, the variance of the Monte Carlo estimator scales as $\mathcal{O}(L^{-1})$ regardless of the dimension of $\boldsymbol{z}$ and $\boldsymbol{\theta}$.

To collect these samples, we design a Markov chain to stochastically explore the space of latent variables and parameters. The chain iteratively samples a new state according to its transition operator, $\mathcal{T}\left((\boldsymbol{z}, \boldsymbol{\theta}) \to (\boldsymbol{z}', \boldsymbol{\theta}')\right)$, which specifies the probability of transitioning from state $(\boldsymbol{z}, \boldsymbol{\theta})$ to state $(\boldsymbol{z}', \boldsymbol{\theta}')$. Each state the Markov chain visits is taken as a sample. If we design the Markov chain appropriately, we guarantee that the transition operator will asymptotically visit states according to their posterior probability.

When states are sampled with a Markov chain, it is no longer true that the samples are independent. In fact, the transition operator often leads to relatively local updates, which

in turn lead to autocorrelation in the sequence of samples. This does not affect the bias of the Monte Carlo estimate, but it does affect the constant in the asymptotic $\mathcal{O}(L^{-1})$ convergence rate. However, in addition to the increased variance, MCMC algorithms also suffer from a transient bias due to the fact that the initial state is not drawn from the posterior distribution (we would not be using MCMC if we could sample from the posterior directly). Fortunately, the transient bias of the Monte Carlo estimator also decays as $\mathcal{O}(L^{-1})$. Since the mean squared error of an estimator is equal to its variance plus its bias squared, and since both variance and bias scale inversely with L, the asymptotic effect of the transient bias is insignificant compared to that of the variance.

The critical property of our Markov chain is that, asymptotically, it visits states with probability equal to the true posterior probability. For this asymptotic guarantee to hold, the posterior distribution must be invariant with respect to the transition operator, which is defined by the following equivalence,

$$p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{s}) = \int p\big(\boldsymbol{z}', \boldsymbol{\theta}' \mid \boldsymbol{s}\big)\, \mathcal{T}\big((\boldsymbol{z}', \boldsymbol{\theta}') \to (\boldsymbol{z}, \boldsymbol{\theta})\big)\, \mathrm{d}\boldsymbol{z}'\, \mathrm{d}\boldsymbol{\theta}'. \tag{2.7}$$

Intuitively, an invariant, or "stationary" distribution with respect to $\mathcal{T}$ has the property that if we randomly sample a state from the stationary distribution and apply the transition operator, the resulting state will be drawn from the stationary distribution as well. When $\boldsymbol{z}$ and $\boldsymbol{\theta}$ are discrete, the stationary distribution is an eigenvector of a transition matrix with an eigenvalue of one.

In addition to leaving the posterior distribution invariant, the Markov chain must also converge to this stationary distribution regardless of where it starts. If this property holds, the Markov chain is *ergodic*, and the posterior distribution is the unique stationary distribution of the chain. One simple sufficient condition that ensures ergodicity is that the transition probability be strictly positive for all states.

Designing an MCMC algorithm thus boils down to designing a valid transition operator. This is typically done by composing a sequence of operators, $\mathcal{T} = \mathcal{T}_1 \circ \ldots \circ \mathcal{T}_K$, each of which leaves the stationary distribution intact. While there are many ways of developing these transition operators, one of the most common is to sample from the conditional distribution of one variable while holding the rest fixed. This leads to an algorithm called

Gibbs sampling (Geman and Geman, 1984).

Gibbs Sampling

Consider a transition operator, $\mathcal{T}_{\boldsymbol{z}}$, that only updates $\boldsymbol{z}$, holding $\boldsymbol{\theta}$ constant. In order to update $\boldsymbol{z}$, it samples from the conditional distribution, $p(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{s})$. To see that this transition operator leaves the posterior distribution invariant, we plug it into (2.7):

$$\begin{aligned}
&\int \mathcal{T}_{\boldsymbol{z}}\big((\boldsymbol{z}', \boldsymbol{\theta}') \to (\boldsymbol{z}, \boldsymbol{\theta})\big)\, p\big(\boldsymbol{z}', \boldsymbol{\theta}' \mid \boldsymbol{s}\big)\, \mathrm{d}\boldsymbol{z}'\, \mathrm{d}\boldsymbol{\theta}' \\
&\qquad = \int p(\boldsymbol{z} \mid \boldsymbol{\theta}', \boldsymbol{s})\, \delta_{\boldsymbol{\theta}'}(\boldsymbol{\theta})\, p\big(\boldsymbol{z}', \boldsymbol{\theta}' \mid \boldsymbol{s}\big)\, \mathrm{d}\boldsymbol{z}'\, \mathrm{d}\boldsymbol{\theta}' \\
&\qquad = \int p(\boldsymbol{z} \mid \boldsymbol{\theta}', \boldsymbol{s})\, \delta_{\boldsymbol{\theta}'}(\boldsymbol{\theta})\, p\big(\boldsymbol{\theta}' \mid \boldsymbol{s}\big) \underbrace{\int p\big(\boldsymbol{z}' \mid \boldsymbol{\theta}', \boldsymbol{s}\big)\, \mathrm{d}\boldsymbol{z}'}_{=1}\, \mathrm{d}\boldsymbol{\theta}' \\
&\qquad = p(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{s})\, p(\boldsymbol{\theta} \mid \boldsymbol{s}) \\
&\qquad = p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{s}).
\end{aligned}$$

The same holds for a transition operator that samples $p(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{s})$, or even a single element of these sets, $p(\theta_j \mid \boldsymbol{\theta}_{\neg j}, \boldsymbol{z}, \boldsymbol{s})$. Here, θ_j is one parameter, like the transition matrix in an HMM, and $\boldsymbol{\theta}_{\neg j}$ is the set of all other parameters except for the j-th.

Many compositional models are designed such that these conditional distributions are easy to sample from. For example, if the model is defined with conjugate prior distributions, as described above, the conditional distributions have closed forms that can often be sampled exactly. Moreover, some model motifs enable more efficient types of Gibbs updates outlined below:

- Block Gibbs Sampling: In some cases, entire subsets or "blocks" of random variables can be updated by a single transition operator. Consider the conditional distribution over a single latent state in an HMM, z_t, given all other variables,

  $$p(z_t \mid \boldsymbol{s}, \boldsymbol{z}_{\neg t}, \boldsymbol{\theta}) \propto p(z_t \mid z_{t-1}, \boldsymbol{\theta})\, p(z_{t+1} \mid z_t, \boldsymbol{\theta})\, p(\boldsymbol{s}_t \mid z_t, \boldsymbol{\theta}).$$

  A naïve Gibbs sampling algorithm would enumerate the K possible values of z_t,

compute their posterior probability, and sample accordingly. However, this would be horribly inefficient when the states are highly correlated. Given z_{t-1} and z_{t+1}, the state z_t may essentially be deterministic. Thus, even if there is genuine uncertainty over the state sequence as a whole, this simple transition operator may get stuck in a single state sequence assignment. This is an example of "poor mixing."

Instead, we could try to update the entire state sequence at once. The conditional distribution of $\boldsymbol{z}$ is proportional to the joint distribution,

$$p(\boldsymbol{z} \mid \boldsymbol{s}, \boldsymbol{\theta}) \propto p(z_1 \mid \boldsymbol{\theta})\, p(\boldsymbol{s}_1 \mid z_1, \boldsymbol{\theta}) \prod_{t=2}^{T} p(z_t \mid z_{t-1}, \boldsymbol{\theta})\, p(\boldsymbol{s}_t \mid z_t, \boldsymbol{\theta}).$$

While there are K^T possible assignments of $\boldsymbol{z}$, since the conditional distribution is chain structured (each state depends only on the previous state and the current spike counts), we can actually sample this distribution using dynamic programming without enumerating all possible assignments (e.g. Bishop, 2006).

- Block Parallel Gibbs Sampling: A special case of block Gibbs sampling occurs when an entire block of variables is conditionally independent given the rest. For example, consider the conditional distribution of $\boldsymbol{z}$ in a mixture model,

$$p(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{s}) \propto \prod_{t=1}^{T} p(z_t \mid \boldsymbol{\theta})\, p(\boldsymbol{s}_t \mid z_t, \boldsymbol{\theta}).$$

  Since the conditional distribution factors into a product, the individual latent variables are conditionally independent of one another. That is, the update to z_t does not depend on the updated value of $z_{t'}$. This allows us to sample new latent states in parallel using as many processors or threads as we have at our disposal.

- Collapsed Gibbs Sampling: Another special case of block Gibbs sampling occurs when the conditional distribution can be factored using the product rule. For example, consider a model with two highly correlated latent variables, z_1 and z_2. Naïvely alternating between sampling $p(z_1 \mid z_2, \boldsymbol{\theta}, \boldsymbol{s})$ and $p(z_2 \mid z_1, \boldsymbol{\theta}, \boldsymbol{s})$ will lead to poor mixing, so we would like to update them jointly. Suppose, however, that it is chal-

lenging to directly sample the full conditional distribution $p(z_1, z_2 \mid \boldsymbol{\theta}, \boldsymbol{s})$. By the sum and product rules of probability,

$$\begin{aligned} p(z_1, z_2 \mid \boldsymbol{\theta}, \boldsymbol{s}) &= p(z_2 \mid z_1, \boldsymbol{\theta}, \boldsymbol{s})\, p(z_1 \mid \boldsymbol{\theta}, \boldsymbol{s}) \\ &= p(z_2 \mid z_1, \boldsymbol{\theta}, \boldsymbol{s}) \int p(z_1, z_2 \mid \boldsymbol{\theta}, \boldsymbol{s})\, \mathrm{d}z_2. \end{aligned}$$

If it is possible "collapse" the second variable and obtain a tractable closed form solution for $p(z_1 \mid \boldsymbol{\theta}, \boldsymbol{s})$, then we can sample the pair of variables jointly in a two step procedure. First, sample z_1 from its marginal conditional distribution, $p(z_1 \mid \boldsymbol{\theta}, \boldsymbol{s})$, and then sample $p(z_2 \mid z_1, \boldsymbol{\theta}, \boldsymbol{s})$. We use this technique in the spike-and-slab models of Chapter 5.

- Augmented Gibbs Sampling: Just as it is possible to collapse some variables during block updates, in other cases it is possible to introduce *auxiliary* variables that make the model conditionally conjugate and thus easier to work with. For example, in some cases $p(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{s})$ is challenging to sample from, but by introducing an auxiliary variable, $\boldsymbol{\omega}$, it becomes easier to sample from the conditional distributions of the full model, $p(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\omega}, \boldsymbol{\theta})$. It is as if we "un-collapse" $\boldsymbol{\omega}$ and then perform augmented Gibbs sampling in two steps,

$$\begin{aligned} \boldsymbol{z}' &\sim p(\boldsymbol{z} \mid \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{s}), \\ \boldsymbol{\omega}' &\sim p(\boldsymbol{\omega} \mid \boldsymbol{z}', \boldsymbol{\theta}, \boldsymbol{s}). \end{aligned}$$

  As long as the original joint distribution, $p(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\theta})$, is equal to the marginal distribution, $\int p(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\omega}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\omega}$, the samples $\{\boldsymbol{z}^{(\ell)}, \boldsymbol{\theta}^{(\ell)}\}$ will be distributed according to $p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{s})$. Thus, simply discarding the samples of $\boldsymbol{\omega}$ leaves a set of samples drawn from the desired posterior. This technique of *data augmentation* is a powerful tool that we use throughout this thesis.

2.3.2 Mean Field Variational Inference

Variational inference methods (Jordan et al., 1999; Wainwright and Jordan, 2008) take a fundamentally different approach to approximating the posterior distribution. Rather than collecting a set of samples, variational methods attempt to find the distribution within a tractable family of distributions that most closely matches the true posterior. Thus, inference becomes an optimization problem.

Let's assume the variational posterior is parameterized by $\boldsymbol{\vartheta}$, and call the variational distribution, $q(\boldsymbol{z}, \boldsymbol{\theta}; \boldsymbol{\vartheta})$. [‡] To find the optimal $q(\cdot)$, we optimize a functional, $\mathcal{L}[q]$, called the *variational lower bound*, which provides a lower bound on the log marginal likelihood, $\log p(\boldsymbol{s})$. Specifically, we can write the log marginal likelihood as an expectation with respect to q,

$$\begin{aligned}
\log p(\boldsymbol{s}) &= \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{\vartheta})}\left[\log \frac{p(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\theta})}{p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{s})}\right] \\
&= \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{\vartheta})}\left[\log \frac{p(\boldsymbol{s}, \boldsymbol{z}, \boldsymbol{\theta})}{q(\boldsymbol{z}, \boldsymbol{\theta}; \boldsymbol{\vartheta})}\right] + \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{\vartheta})}\left[\log \frac{q(\boldsymbol{z}, \boldsymbol{\theta}; \boldsymbol{\vartheta})}{p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{s})}\right] \\
&= \mathcal{L}[q(\boldsymbol{z}, \boldsymbol{\theta}; \boldsymbol{\vartheta})] + \mathrm{KL}(q(\boldsymbol{z}, \boldsymbol{\theta}; \boldsymbol{\vartheta}) \,||\, p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{s}) \\
&\geq \mathcal{L}[q(\boldsymbol{z}, \boldsymbol{\theta}; \boldsymbol{\vartheta})].
\end{aligned}$$

where $\mathrm{KL}(q \,||\, p)$ is the KL-divergence between distributions q and p. The last line follows from the fact that the KL-divergence is nonnegative and equal to zero if and only if the distributions are identical. Thus, optimizing this functional is equivalent to minimizing the KL-divergence between the variational distribution and the true posterior.

Our goal is to maximize the variational lower bound over a parameterized family of tractable distributions, $\mathcal{Q}$. In *mean field variational inference*, we take $\mathcal{Q}$ to be the family of fully factorized distributions,

$$\mathcal{Q} = \left\{ q : q(\boldsymbol{z}, \boldsymbol{\theta}; \boldsymbol{\vartheta}) \propto \prod_{t=1}^{T} q_t(z_t; \vartheta_t) \prod_{j=1}^{J} q_j(\theta_j; \vartheta_j) \right\}.$$

[‡] We use a semicolon to indicate that $q(\boldsymbol{z}, \boldsymbol{\theta})$ is parameterized by $\boldsymbol{\vartheta}$. Since $\boldsymbol{\vartheta}$ is not a random variable, q is not strictly a conditional distribution and the vertical bar notation is inappropriate.

We will set these parameters, $\boldsymbol{\vartheta}$, in order to maximize the variational lower bound over the set of distributions in $\mathcal{Q}$.

In general, this objective function is not concave, so we should not expect to find a global optimum. However, we can still use local optimization and multiple random restarts with the hope of finding a global optimum. For mean field variational inference, a simple approach is to perform coordinate ascent on the parameters of one variational factor at a time, holding the rest fixed. Given the equivalence between maximizing the variational lower bound and minimizing the KL-divergence, we can derive the general form of a mean field update. Consider updating the variational factor for θ_j. We have,

$$\begin{aligned}
\mathrm{KL}(q \,||\, p) &= \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{\vartheta})}\left[\log q(\boldsymbol{z},\boldsymbol{\theta})\right] - \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{\vartheta})}\left[\log p(\boldsymbol{z},\boldsymbol{\theta} \mid \boldsymbol{s})\right] \\
&\simeq \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{\vartheta})}\left[\log q(\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{\vartheta})\right] - \mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{\vartheta}_)}\left[\log p(\boldsymbol{s},\boldsymbol{z},\boldsymbol{\theta})\right] \\
&\simeq \mathbb{E}_{q_j(\theta_j;\vartheta_j)}\left[\log q_j(\theta_j;\vartheta_j)\right] - \mathbb{E}_{q(\theta_j;\vartheta_j)}\left[\mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\theta}_{\neg j};\boldsymbol{\vartheta}_{\neg j})}\big[\log p(\boldsymbol{s},\boldsymbol{z},\boldsymbol{\theta})\big]\right] \\
&\simeq \mathrm{KL}\big(q_j(\theta_j;\vartheta_j) \,||\, \widetilde{p}_j(\theta_j)\big), \qquad (2.8)
\end{aligned}$$

where $\simeq$ denotes equality up to an additive term that is constant with respect to θ_j, and

$$\widetilde{p}_j(\theta_j) \propto \exp\left\{\mathbb{E}_{q(\boldsymbol{z},\boldsymbol{\theta}_{\neg j};\boldsymbol{\vartheta}_{\neg j})}\big[\log p(\boldsymbol{s},\boldsymbol{z},\boldsymbol{\theta})\big]\right\}. \qquad (2.9)$$

We are able to separate the expectations in the third line of (2.8) due to the factorized form of $q(\boldsymbol{z},\boldsymbol{\theta};\boldsymbol{\vartheta})$. Since KL-divergence is minimized when the two distributions are equal, the optimal $q_j(\theta_j;\vartheta_j)$, given the variational factors for the remaining variables, is equal to $\widetilde{p}_j(\theta_j)$. As in the Gibbs sampling algorithms developed before, the expectation in (2.9) is often greatly simplified by the factorization of the joint distribution in our probabilistic model. Moreover, when the model is constructed out of conjugate distributions, these mean field updates can be computed in closed form.

Structured Mean Field Just as block Gibbs sampling enables more efficient updates for sets of correlated random variables, structured mean field algorithms allow groups of random variables to share a variational factor. For example, in an HMM, we can group the latent states $\boldsymbol{z}$ together in a shared factor, $q(\boldsymbol{z})$, that does not necessarily factor into a

product over time bins. If the optimal shared factor given by (2.9) has a tractable form, we can perform coordinate ascent on the variational lower bound by updating the parameters of the shared factor, rather than sequentially updating individual factors for each time bin. As a result, our coordinate ascent algorithm will converge much more rapidly. The only other requirement is that it must be possible to compute the expectations with respect to the shared factor. In the case of HMMs, the same type of dynamic programming algorithm that enables efficient block sampling also enables efficient calculation of expectations.

2.3.3 Model Comparison

Now that we have developed the tools to formulate models and perform Bayesian inference, we need a way to compare and criticize our models. The easiest way, and the primary way used throughout this thesis, is to compare the models based on how well they predict held-out data. Suppose that at the beginning of the experiment, we reserve a set of spike counts, $\boldsymbol{s}_{\text{test}}$, to be used for model comparison. Once we have used Bayesian inference to compute a posterior distribution over the model's parameters and latent variables, we can then compute the predictive likelihood,

$$p(\boldsymbol{s}_{\text{test}} \mid \boldsymbol{s}_{\text{train}}) = \int p(\boldsymbol{s}_{\text{test}} \mid \boldsymbol{z}_{\text{test}}, \boldsymbol{\theta})\, p(\boldsymbol{z}_{\text{test}} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{s}_{\text{train}})\, \mathrm{d}\boldsymbol{z}_{\text{test}}\, \mathrm{d}\boldsymbol{\theta}. \tag{2.10}$$

Notice that this is an expectation with respect to the *posterior* distribution of $\boldsymbol{\theta}$ given the training data, and a *marginal* distribution in that it involves an integration over the latent variables associated with the test data. As usual, this integral is typically intractable, but we can construct a Monte Carlo estimate given samples from the approximate posterior,

$$p(\boldsymbol{s}_{\text{test}} \mid \boldsymbol{s}_{\text{train}}) \approx \frac{1}{L} \sum_{\ell=1}^{L} p(\boldsymbol{s}_{\text{test}} \mid \boldsymbol{z}_{\text{test}}^{(\ell)}, \boldsymbol{\theta}^{(\ell)})$$

where

$$\begin{aligned} \boldsymbol{\theta}^{(\ell)} &\sim p(\boldsymbol{\theta} \mid \boldsymbol{s}_{\text{train}}), \\ \boldsymbol{z}_{\text{test}}^{(\ell)} &\sim p(\boldsymbol{z}_{\text{test}} \mid \boldsymbol{\theta}^{(\ell)}). \end{aligned}$$

When Bayesian inference is performed with MCMC, the samples $\{\boldsymbol{\theta}^{(\ell)}\}$ are simply the states visited by the Markov chain. When variational methods are used, we assume they are drawn from the variational posterior, $q(\boldsymbol{\theta})$.

This is by no means the only method of comparing models. In "fully Bayesian" analyses, it is common to compare models on the basis of their *marginal likelihood*, $p(\boldsymbol{s})$ (Kass and Raftery, 1995). Recall that this is the quantity that variational methods attempt to lower bound. Unfortunately, we cannot evaluate the tightness of variational lower bounds because they depend on the KL-divergence, which is intractable.

Instead, we may resort to other methods of approximating the marginal likelihood. Notice that,

$$p(\boldsymbol{s}) = \int p(\boldsymbol{s} \mid \boldsymbol{z}, \boldsymbol{\theta})\, p(\boldsymbol{z} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{z}\, \mathrm{d}\boldsymbol{\theta} \tag{2.11}$$

is equal to the predictive likelihood in the absence of training data. Unfortunately, training data plays the crucial role of winnowing the posterior distribution over parameters. Without this constraint, simple Monte Carlo estimates like those used to approximate the predictive likelihood will suffer from extremely high variance. Instead, more sophisticated methods, like annealed importance sampling (Neal, 2001) are typically employed.

Finally, another means of evaluating and criticizing models is via *posterior predictive checks* (PPCs) (Box, 1980; Gelman et al., 2013; Blei, 2014). Though we do not make use of them in this thesis, we note that they provide a slightly different view on model performance. Rather than assessing how well the model predicts held-out data, they assess how well statistics of data simulated from the posterior distribution match statistics calculated from samples of the real data. Rather than evaluating how well one model performs relative to another, PPCs assess how well the model explains relevant aspects of the data.

2.4 Conclusion

With this background, we have the basic tools necessary to formulate models, perform Bayesian inference, and evaluate model performance. However, as we incorporate more structure into our model and scale up to larger datasets, inference quickly becomes computationally intractable. This thesis is about extending the frontier of models and motifs at

our disposal by leveraging model structure to develop efficient inference algorithms. One of the major techniques we use is the introduction of auxiliary variables that render the model conjugate and enable block parallel Gibbs samplers or structured mean field algorithms. Essentially, these methods provide nice "axes" for inference. While this increases the dimensionality of the posterior, it is sometimes easier to make two simple updates rather than one hard update. These insights enable us to push the frontier of modeling and inference for complex discrete datasets like neural spike trains, and extend the set of motifs in our modeling toolkit.

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. In *Advances in Neural Information Processing Systems 14*, pages 577–585, Cambridge, MA, 2002. MIT Press.

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 1–9, 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*, 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. *JOSA 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. In *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. In *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. In *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 et al. 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. In *Proceedings of the Sixteenth 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). In *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. In *Proceedings of the 25th 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. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (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 29th 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. In *Proceedings of the 21st Annual Conference on 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. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1019–1028. ACM, 2010.

Noah Goodman, Vikash Mansinghka, Daniel M Roy, Keith Bonawitz, and Joshua B Tenenbaum. Church: a language for generative models. *Proceedings of the 24th 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. In *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. In *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. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Washington DC*, 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*, 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. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 266–274. ACM, 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 31st 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. In *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. In *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. *JOSA 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 Comput Biol*, 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. In *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. In *Proceedings of the 31st 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. In *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. In *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. In *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. In *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. In *Proceedings of the 29th 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. *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. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 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. In *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. 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. In *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. In *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? In J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics*, pages 763–773. Oxford University Press, 1992.

Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. *17th 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. In *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. In *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 26th 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. In *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 Comput Biol*, 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. In *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. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 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. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*, pages 158–207. Cambridge University Press, 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. In *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. In *Proceedings of the 25th 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. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 16, 2013.

Mingyuan Zhou, Lingbo Li, Lawrence Carin, and David B Dunson. Lognormal and gamma mixed negative binomial regression. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1343–1350, 2012.