3

Hawkes Processes with Latent Network Structure

Networks are fundamental models for complex systems, enabling us to reason about systems by studying the relationships between their parts. As we discussed in Chapter 1, networks are employed in a myriad of ways throughout neuroscience. Whether they represent synapses in the human connectome, population-level interactions in brain circuits, pairwise correlations in fMRI recordings, or coupling in theoretical models, networks serve as an abstraction for the messy details of systems. A network consists of a set of nodes, which may represent neurons, populations, voxels, etc., depending on the application. Connecting these nodes is a set of edges, which represent interactions between pairs of nodes, like the effect of one neuron's spikes on the subsequent activity of its downstream neighbors. By reducing a system to a network of nodes and edges, we create a simplified object for analysis.

A great deal can be learned by considering simple network properties like the average number of connections per node, or higher order statistics like "betweenness" and the number of connected triangles (Bullmore and Sporns, 2009). Some network analyses involve probabilistic modeling. For example, we may look for clusters or features of nodes that have similar patterns of connectivity. Other analyses are more supervised in nature; in a "link prediction" task we seek to predict whether or not a pair of nodes is connected, given partial observations of the network (Liben-Nowell and Kleinberg, 2007). Tradition-

ally, network analysis has focused on *explicit network* problems in which the network itself is considered to be the observed data. That is, the nodes are and edges are considered known. A rich literature has arisen in recent years for applying statistical machine learning models to this type of problem, e.g., Liben-Nowell and Kleinberg (2007); Hoff (2008); Goldenberg et al. (2010).

In practice, however, we are often confronted with *implicit networks* that cannot be observed directly, but about which we wish to perform analysis. In an implicit network, the nodes or edges of the network may not be directly observed, but the graph structure may be inferred from noisy emissions. These noisy observations are assumed to have been generated according to an underlying model that respects the latent network structure.

For example, in connectomics problems the network must be inferred from noisy electron microscopy images; in spike train modeling the network must be inferred from weak anatomical evidence and noisy measurements of population activity; and in fMRI experiments, the network is derived from noisy blood-oxygen-level dependent (BOLD) responses. In all cases, if we knew the underlying network (i.e. if we knew how the nodes were connected), we could assign a likelihood to the observed electron microscopy images or to the measured activity. By combining this likelihood with a prior distribution that reflects our intuitions about the network structure, we construct a probabilistic model to tackle these implicit network problems.

In this chapter, we consider the case where our observations come in the form of neural spike trains, and our intuition is that a spike on one neuron will influence the activity of downstream neurons. We formalize this with a probabilistic model based on mutually interacting point processes. Specifically, we combine the Hawkes process (Hawkes, 1971) with hierarchical prior distributions on the network. This combination allows us to reason about properties of neurons that govern network structure, which in turn governs the dynamics of activity via a Hawkes process.

Figure 3.1 illustrates the components of the probabilistic model. At the highest level, we have a network of neurons. The pattern of connectivity in the network may be governed by latent variables like cell types and locations. In this case there are four types of cells (different colors in Fig. 3.1a), and each cell has a location in the two-dimensional plane. Nearby cells of the same type are more likely to connect. The network governs the dynamics of

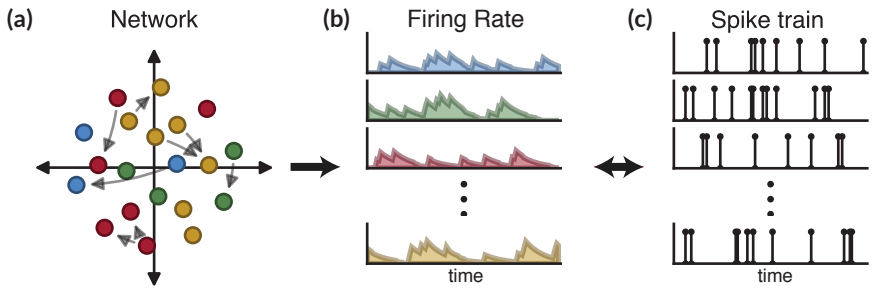

Figure 3.1: Components of the generative model. **(a)** Each neuron is endowed with latent variables, like locations in space and discrete types (illustrated with different colors). These variables determine the probability of connections and the strength of those connections. In this example, nearby neurons of the same type are most likely to connect. **(b)** The network parameterizes an autoregressive model with a time-varying firing rate, which specifies the instantaneous probability of an action potential. **(c)** Spikes are randomly generated according to the firing rate. Each spike induces an impulse response on the firing rate of downstream neurons.

the firing rate (Fig. 3.1b), which in turn gives rise to the observed spikes (Fig. 3.1c). In this case, spikes on one neuron induce impulse responses that feed back into the firing rate of downstream neurons. These firing rate–spike train dynamics are modeled with Hawkes processes.

The rest of the chapter is organized as follows. In Section 3.1 we introduce a compositional probabilistic model for networks, and in Section 3.2 we introduce Hawkes processes. Section 3.3 stitches these two components together into a joint model for implicit networks with spike train observations. Before diving into inference applications, we first consider the theoretical consequences of a particular network model on the stability of the system, and provide some intuition on how the network properties affect the asymptotic behavior of the system. Then, in Section 3.4, we derive a Gibbs sampling algorithm with an elegant auxiliary variable formulation that allows efficient parallelism. Since this is the first technical chapter, we go through these derivations in substantial detail. In later chapters we will move more quickly. Finally, the remaining sections consider applications, first to synthetic data, and then to biological recordings. While the primary emphasis is on modeling neural data, we also explore some applications in areas of finance and criminology.

Name	$\boldsymbol{\vartheta}$	$\mathrm{dom}(\boldsymbol{z}_n)$	$\rho_{n\to n'}$
Empty Model	—	—	0
Dense Model	—	—	1
Bernoulli Model	ρ	—	ρ
Stochastic Block Model	$\{\{\rho_{k\to k'}\}\}$	$\{1,\ldots,K\}$	$\rho_{z_n\to z_{n'}}$
Latent Distance Model	γ_0	$\mathbb{R}^K$	$\sigma(-\|\|\boldsymbol{z}_n-\boldsymbol{z}_{n'}\|\|_2^2+\gamma_0)$

Table 3.1: Binary adjacency matrix models.

3.1 Probabilistic Network Models

Networks of N nodes can be represented by $N \times N$ matrices. Unweighted networks correspond to binary adjacency matrices $\boldsymbol{A}$ where $a_{m,n} = a_{m\to n} = 1$ indicates a directed edge from node m to node n. We use the arrow notation ($\to$) to remind the reader of the directionality of the connection. When these edges have scalar weights associated with them, we can encode the weights in a second matrix, $\boldsymbol{W} \in \mathbb{R}^{N\times N}$. The complete network is then defined by the elementwise product, $\boldsymbol{A} \odot \boldsymbol{W}$. The binary adjacency matrix captures the sparsity pattern, and the real-valued weight matrix captures the strength of the connections. From a modeling perspective, separating these two matrices allows us to separate our prior intuitions about sparsity and strength. This is known as a spike-and-slab model (Mitchell and Beauchamp, 1988).

Hierarchical models can be constructed by incorporating latent variables into the prior distributions over $\boldsymbol{A}$ and $\boldsymbol{W}$. Unsurprisingly, the same types of motifs that recur throughout probabilistic modeling — discrete latent types and continuous latent features— also form the building blocks of standard network models. We briefly outline a few simple models that are used in this and following chapters.

Table 3.1 summarizes a few models for binary adjacency matrices. In all cases, the distri-

bution over $\boldsymbol{A}$ factorizes into a product over edges,

$$\begin{aligned} p(\boldsymbol{A} \mid \boldsymbol{z}, \boldsymbol{\vartheta}) &= \prod_{n=1}^{N} \prod_{n'=1}^{N} p(a_{n \to n'} \mid \boldsymbol{z}_n, \boldsymbol{z}_{n'}, \boldsymbol{\vartheta}) \\ &= \prod_{n=1}^{N} \prod_{n'=1}^{N} \mathrm{Bern}(a_{n \to n'} \mid \rho_{n \to n'}). \end{aligned}$$

The difference is in how the local latent variables, $\boldsymbol{z}_{n'}$ and $\boldsymbol{z}_n$, and the global network parameters, $\boldsymbol{\vartheta}$, combine to determine the probability, $\rho_{n \to n'}$. We describe these models below:

- *Empty Model:* The empty model is essentially a null model. According to this model, there are no connections between neurons. Nevertheless, it is useful to list it here because the empty model provides a baseline for more sophisticated models, capturing the null hypothesis that neurons are independent.

- *Dense Model:* At the other extreme, the dense model corresponds to the hypothesis that all pairs of neurons are connected. In the models of neural activity that follow, the dense model will reduce to the standard models in use today, which do not incorporate structured prior distributions over the network.

- *Bernoulli Model:* The Bernoulli model is a spike-and-slab model in which each connection is an independent and identically distributed Bernoulli random variable. This is also known as an Erdős-Rényi model.

- *Stochastic Block Model (SBM):* In the stochastic block model (SBM) (Nowicki and Snijders, 2001), each neuron has an associated class, z_n. The probability of connection depends on the class of the two neurons. This is the network equivalent of a mixture model. In a Bayesian framework, we assume the class assignments are drawn from a discrete prior, $z_n \sim \mathrm{Discrete}(\boldsymbol{\pi})$, and the class weights are given a conjugate, symmetric Dirichlet prior, $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha \mathbf{1}_K)$. The connection probabilities are given a conjugate beta prior, $\beta_{k \to k'} \sim \mathrm{Beta}(\alpha, \beta)$.

Name	$\boldsymbol{\vartheta}$	$\mathrm{dom}(\boldsymbol{z})$	$\mu_{n\to n'}$
Independent Model	μ	—	μ
Stochastic Block Model	$\{\{\mu_{k\to k'}\}\}$	$\{1,\ldots,K\}$	$\mu_{z_n\to z_{n'}}$
Latent Distance Model	μ_0	$\mathbb{R}^K$	$-\lvert\lvert\boldsymbol{z}_n-\boldsymbol{z}_{n'}\rvert\rvert_2^2+\mu_0$

Table 3.2: General weight models.

- *Latent Distance Model:* The latent distance model (Hoff, 2008) encodes the belief that connection probability should decrease with distance between latent locations. The locations are given spherical Gaussian priors, $\boldsymbol{z}_n \sim \mathcal{N}(0, \tau\boldsymbol{I})$, and the scale is drawn from an inverse gamma prior, $\tau \sim \mathrm{IGa}(1,1)$. The offset is given a standard normal prior, $\gamma_0 \sim \mathcal{N}(0,1)$.

The same ideas can be applied to models for the scalar weight matrix, $\boldsymbol{W}$, but rather than modeling the connection probability, we now model the mean weight, $\mu_{n\to n'}$. The resulting distribution is of the form,

$$\begin{aligned} p(\boldsymbol{W} \mid \boldsymbol{z}, \boldsymbol{\vartheta}) &= \prod_{m=1}^{N}\prod_{n=1}^{N} p(w_{m\to n} \mid z_n, z_{n'}, \boldsymbol{\vartheta}) \\ &= \prod_{m=1}^{N}\prod_{n=1}^{N} p(w_{m\to n} \mid \mu_{m\to n}, \boldsymbol{\vartheta}). \end{aligned}$$

We do not specify the exact functional form of the distribution since this will depend on the model for neural activity. The linear models with nonnegative weights in this chapter will use a gamma prior, whereas the nonlinear autoregressive models in Chapter 5 will use a Gaussian distribution. Table 3.2 lists some examples of weight models analogous to the adjacency matrix models above. While we have only shown models for the mean weight, the same latent variables may also parameterize the variance of the weight distribution. For example in a Gaussian SBM, each directed pair of classes may have an associated variance, $\sigma^2_{k\to k'}$.

Probabilistic network models like these are unified under an elegant theoretical framework due to Aldous and Hoover (Aldous, 1981; Hoover, 1979). Conceptually, the Aldous-Hoover representation characterizes the class of *exchangeable* random graphs, that is, graph

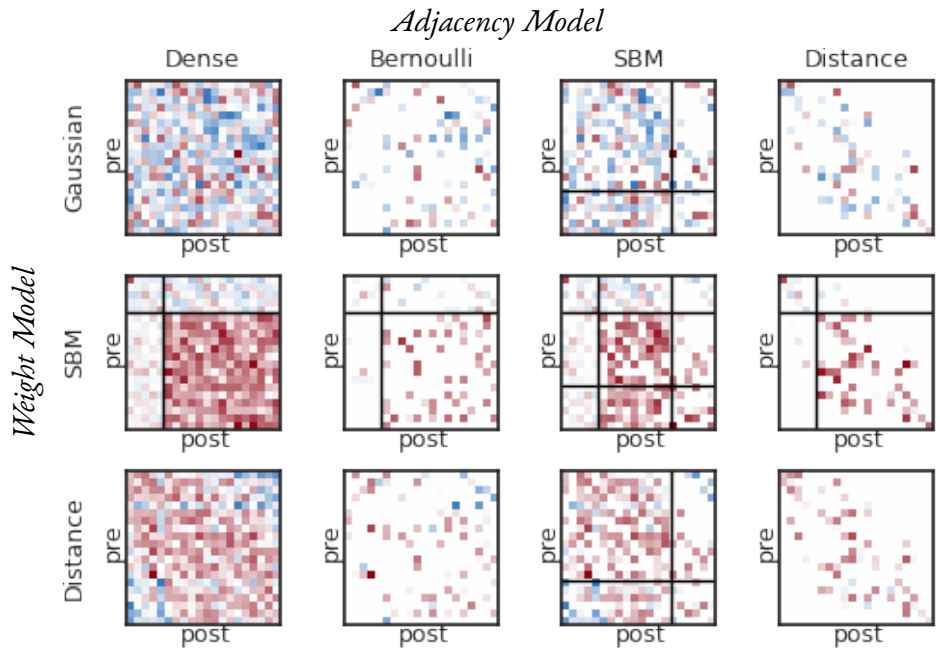

Figure 3.2: Example network models. Each row corresponds to a fixed weight matrix, $\boldsymbol{W}$, for three different weight models, and each column corresponds to a fixed adjacency matrix, $\boldsymbol{A}$, for four different adjacency models. The panels show the elementwise product of the two. Color denotes the weight (blue is negative, red is positive). In the SBM, the rows and columns are sorted by type, and in the distance model, they are sorted by location.

models for which the joint probability is invariant under permutations of the node labels. Just as de Finetti's theorem equates exchangeable sequences to independent draws from a random probability measure, Aldous-Hoover renders the entries of $\boldsymbol{A}$ and $\boldsymbol{W}$ conditionally independent given latent variables $\boldsymbol{z}$ and global parameters $\boldsymbol{\vartheta}$. Lloyd et al. (2012) and Orbanz and Roy (2015) review this theoretical framework and its applications in probabilistic machine learning.

Note that we have associated each neuron with a single latent variable, $\boldsymbol{z}_n$. This suggests that both the adjacency matrix and the weight matrix are governed by the same latent variable, but in general they can have separate variables. Whether or not they are shared is a modeling decision. We will collectively refer to all latent variables of the network as $\boldsymbol{z}_n$ and whether they govern the adjacency model or the weight model will be clear from context.

Figure 3.2 shows how a variety of networks can be constructed by combining different priors on the weights (rows) with priors on the pattern of connectivity (columns).

Each row corresponds to a fixed weight matrix drawn from either an independent model, a stochastic block model (SBM), or a latent distance model. In these cases, the weights are Gaussian distributed with unit variance and model-specific mean. Each column corresponds to a fixed adjacency matrix drawn from either a dense model, an independent Bernoulli model, an SBM, or a latent distance model. The matrices show the element-wise product, which encodes a weighted, directed network. Next, we introduce a model for spike trains that leverages an underlying network.

3.2 Hawkes Processes

Hawkes processes (Hawkes, 1971) are a special type of point process that allow spikes to influence the future firing rate. This is achieved via a linear superposition of Poisson processes. Before jumping into the details, a brief primer on Poisson processes is in order.

3.2.1 Poisson Processes

Point processes are fundamental statistical objects that yield random finite sets of spikes $\{s_m\}_{m=1}^M \subset \mathcal{V}$, where $\mathcal{V}$ is a compact subset of $\mathbb{R}^D$ (Daley and Vere-Jones, 2003). When modeling neural spike trains, we typically let $\mathcal{V}$ be the interval $[0, T]$. The Poisson process is the canonical example. It is governed by a nonnegative firing rate or intensity function, $\lambda(t) : \mathcal{V} \to \mathbb{R}_+$. The number of spikes in a subset $\mathcal{V}' \subset \mathcal{V}$ follows a Poisson distribution with mean $\int_{\mathcal{V}'} \lambda(t)\, \mathrm{d}t$. Moreover, the number of spikes in disjoint subsets are independent (Kingman, 1993).

We use the notation $\{s_m\}_{m=1}^M \sim \mathcal{PP}(\lambda(t))$ to indicate that a set of spikes $\{s_m\}_{m=1}^M$ is drawn from a Poisson process with rate $\lambda(t)$. There are many ways to sample a Poisson process; one way is to sample a Poisson number of spikes with mean $\int_{\mathcal{V}} \lambda(t) \mathrm{d}t$ and then sample the individual spike times, s_m, independently from the density, $p(s) = \frac{\lambda(s)}{\int_{\mathcal{V}} \lambda(t)\, \mathrm{d}t}$.

Thus, after accounting for the $M!$ permutations, the likelihood of a set of spikes is given by,

$$\begin{aligned} p(\{s_m\}_{m=1}^M \mid \lambda(t)) &= \text{Poisson}\left(M \,\middle|\, \int_{\mathcal{V}} \lambda(t)\,\mathrm{d}t\right) \left(\prod_{m=1}^M \frac{\lambda(s_m)}{\int_{\mathcal{V}} \lambda(t)\,\mathrm{d}t}\right) M! \\ &= \frac{M!}{M!} \left(\int_{\mathcal{V}} \lambda(t)\,\mathrm{d}t\right)^M \exp\left\{-\int_{\mathcal{V}} \lambda(t)\,\mathrm{d}t\right\} \left(\prod_{m=1}^M \frac{\lambda(s_m)}{\int_{\mathcal{V}} \lambda(t)\,\mathrm{d}t}\right) \\ &= \exp\left\{-\int_{\mathcal{V}} \lambda(t)\mathrm{d}t\right\} \prod_{m=1}^M \lambda(s_m). \end{aligned} \tag{3.1}$$

Poisson Superposition Principle We will make use of a special property of Poisson processes called the *Poisson superposition principle*, which states that if we have sets of spikes from independent Poisson processes, then the union of spikes is distributed according to a Poisson process as well (Kingman, 1993). Moreover, the rate of this process equals the sum of rates from the individual processes. Formally, suppose we are given sets of spikes drawn independently from K Poisson processes with rates $\lambda_1(t), \ldots, \lambda_K(t)$. Call the union of the spikes $\{s_m, \omega_m\}$, where $s_m \in \mathcal{V}$ is the location of the spike and $\omega_m \in \{1, \ldots, K\}$ denotes which process it came from. Let $\lambda_{\text{tot}}(t) = \sum_{k=1}^K \lambda_k(t)$. The likelihood of the full set of spikes is,

$$\begin{aligned} p(\{s_m, \omega_m\}_{m=1}^M \mid \{\lambda_k(t)\}_{k=1}^K) &= \prod_{k=1}^K \mathcal{PP}(\{s_m : \omega_m = k\} \mid \lambda_k(t)) \\ &= \prod_{k=1}^K \left[\exp\left\{-\int_{\mathcal{V}} \lambda_k(t)\,\mathrm{d}t\right\} \prod_{m=1}^M \lambda_k(s_m)^{\mathbb{I}[\omega_m = k]}\right] \\ &= \exp\left\{-\int_{\mathcal{V}} \lambda_{\text{tot}}(t)\,\mathrm{d}t\right\} \prod_{m=1}^M \prod_{k=1}^K \lambda_k(s_m)^{\mathbb{I}[\omega_m = k]}. \end{aligned}$$

The Poisson superposition principle states that the marginal distribution, summing over

all possible process assignments, $\{\omega_m\}$, is a Poisson process with rate $\lambda_{\text{tot}}(t)$. That is,

$$\begin{aligned} p(\{s_m\}_{m=1}^M \mid \{\lambda_k(t)\}_{k=1}^K) &= \sum_{\omega_1=1}^K \cdots \sum_{\omega_M=1}^K p(\{s_m, \omega_m\}_{m=1}^M \mid \{\lambda_k(t)\}_{k=1}^K) \\ &= \exp\left\{-\int_{\mathcal{V}} \lambda_{\text{tot}}(t)\,\mathrm{d}t\right\} \prod_{m=1}^M \sum_{\omega_m=1}^K \prod_{k=1}^K \lambda_k(s_m)^{\mathbb{I}[\omega_m=k]} \\ &= \exp\left\{-\int_{\mathcal{V}} \lambda_{\text{tot}}(t)\,\mathrm{d}t\right\} \prod_{m=1}^M \lambda_{\text{tot}}(s_m) \\ &= \mathcal{PP}(\{s_m\} \mid \lambda_{\text{tot}}(t)). \end{aligned}$$

Furthermore, the conditional distribution of ω_m is,

$$p(\omega_m = k \mid s_m, \{\lambda_k(t)\}_{k=1}^K) = \frac{\lambda_k(s_m)}{\sum_{k'=1}^K \lambda_{k'}(s_m)}.$$

In other words, given a set of spikes drawn from rate $\lambda_{\text{tot}}(t)$, we can attribute each spike to one of the K additive contributions to the rate function by sampling a discrete distribution with probabilities given by the relative rate at the time of the spike. This is known as *Poisson thinning* (Kingman, 1993).

3.2.2 Including Spike History with Hawkes Processes

Though the Poisson process has many nice properties, it cannot capture interactions between spikes. For this we turn to a more general model known as the Hawkes process (Hawkes, 1971). First, consider the spike train of a single neuron, $\{s_m\}_{m=1}^M \subset [0, T]$. In a Hawkes process, the firing rate, $\lambda(t \mid \mathcal{H}_t)$, is a function of the spike history, $\mathcal{H}_t = \{s_m : s_m < t\}$. The neuron has a baseline firing rate, $\lambda^{(0)}$. On top of this baseline, each spike adds a nonnegative impulse response, $h(\Delta t)$, to the subsequent firing rate. This allows for spike-driven dynamics that are not possible in Poisson processes. Causality and locality of influence are enforced by limiting the support of $h(\Delta t)$

to $\Delta t \in [0, \Delta t_{\mathsf{max}}]$. The rate is thus given by,

$$\lambda(t \mid \mathcal{H}_t) = \lambda^{(0)} + \sum_{m=1}^{M} h(t - s_m).$$

When the impulse response is equal to zero, the Hawkes process reduces to a standard Poisson process with rate $\lambda^{(0)}$.

By the Poisson superposition principle, these additive components can be considered independent processes, each giving rise to their own spikes. This suggests a convenient latent variable representation in which each spike is attributed to either the background rate or the impulse response of a preceding spike. We augment our data with an auxiliary variable $\omega_m \in \{0, \ldots, m-1\}$ to indicate the *origin* of the m-th spike (0 if the spike is due to the background rate and $1 \ldots m-1$ if it was spawned by a preceding spike).

This is easily extended to a population of N neurons by considering a Hawkes process that gives rise to sets of *marked* spikes $\{s_m, c_m\}_{m=1}^{M}$, where $c_m \in \{1, \ldots, N\}$ specifies the neuron on which the m-th spike occurred. As in the single neuron case, the rate of the n-th neuron, $\lambda_n(t \mid \mathcal{H}_t)$, depends on the spike history, but here the spike history contains the spikes of all neurons through time t. The multi-neuronal generalization also allows for different background rates for each neuron, $\lambda_n^{(0)}$, and different impulse responses for each pair of neurons. For example, the impulse response from neuron n to neuron n', which we now call $h_{n \to n'}(\Delta t)$, may differ from that of the reverse connection. As before, we do require that the impulse responses be causal and have bounded support. Putting it all together, the rate of the n-th neuron is,

$$\lambda_n(t \mid \mathcal{H}_t) = \lambda_n^{(0)} + \sum_{m=1}^{M} h_{c_m \to n}(t - s_m). \tag{3.2}$$

After augmenting the data with auxiliary variables denoting the origin of each spike, the multi-neuronal Hawkes process likelihood reduces to a product of Poisson process likeli-

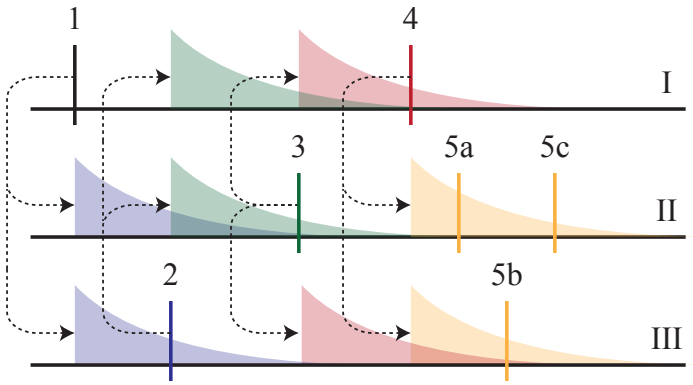

Figure 3.3: Illustration of a Hawkes process. Spikes induce impulse responses on connected processes and spawn "child" spikes. See the main text for a complete description.

hoods for each background rate and each impulse response:

$$p(\{(s_m, c_m, \omega_m)\}_{m=1}^M \mid \{\lambda_n^{(0)}\}, \{\{h_{n\to n'}(\Delta t)\}\}) = \prod_{n=1}^N \mathcal{PP}\big(\{s_m : c_m = n \wedge \omega_m = 0\} \mid \lambda_n^{(0)}\big) \times \prod_{m=1}^M \prod_{n'=1}^N \mathcal{PP}\big(\{s_{m'} : c_{m'} = n' \wedge \omega_{m'} = m\} \mid h_{c_m\to n'}(t - s_m)\big).$$

Combining this with Eq. 3.1, we can write the augmented likelihood as,

$$p(\{s_m, c_m, \omega_m\}_{m=1}^M \mid \{\lambda_n^{(0)}\}, \{\{h_{n\to n'}(\Delta t)\}\}) = \prod_{n=1}^N \left[\exp\left\{-\int_0^T \lambda_n^{(0)} \mathrm{d}t\right\} \prod_{m=1}^M (\lambda_n^{(0)})^{\mathbb{I}[c_m=n]\mathbb{I}[\omega_m=0]}\right] \times \prod_{m=1}^M \prod_{n'=1}^N \left[\exp\left\{-\int_{s_m}^T h_{c_m\to n'}(t - s_m)\mathrm{d}t\right\} \prod_{m'=1}^M h_{c_m\to c_{m'}}(s_{m'} - s_m)^{\mathbb{I}[c_{m'}=n']\mathbb{I}[\omega_{m'}=m]}\right]. \quad (3.3)$$

The second line corresponds to the likelihood of the background processes; the third and fourth correspond to the likelihood of the induced processes triggered by each spike.

Figure 3.3 illustrates a causal cascades of spikes for a simple network of three processes (I-III). The first spike is caused by the background rate ($\omega_1 = 0$), and it induces impulse responses on processes II and III. Spike 2 is spawned by the impulse on the third process ($\omega_2 = 1$), and feeds back onto processes I and II. In some cases a single parent spike induces multiple children, e.g., spike 4 spawns spikes 5a-c. In this simple example, processes excite one another, but do not excite themselves.

3.3 The Network Hawkes Model

In order to combine Hawkes processes and random network models, we decompose the Hawkes impulse response $h_{n\to n'}(\Delta t)$ as follows:

$$h_{n\to n'}(\Delta t) = a_{n\to n'} \cdot w_{n\to n'} \cdot \hbar(\Delta t; \theta_{n\to n'}). \tag{3.4}$$

Here, $a_{n\to n'}$ is an entry in the binary adjacency matrix, $\boldsymbol{A} \in \{0,1\}^{N\times N}$, and $w_{n\to n'}$ is the corresponding entry in the nonnegative weight matrix, $\boldsymbol{W} \in \mathbb{R}_+^{N\times N}$. Together these specify the *sparsity structure* and *strength* of the interaction network, respectively. The nonnegative function $\hbar(\Delta t; \theta_{n\to n'})$ captures the temporal aspect of the interaction. It is parameterized by $\theta_{n\to n'}$ and satisfies two properties: a) it has bounded support for $\Delta t \in [0, \Delta t_{\mathsf{max}}]$, and b) it integrates to one. In other words, $\hbar$ is a probability density with compact support.

Decomposing the impulse response as in Equation 3.4 has many advantages. It allows us to express our separate beliefs about the sparsity structure of the interaction network and the strength of the interactions by using probabilistic network models as priors on $\boldsymbol{A}$ and $\boldsymbol{W}$. The empty graph model recovers independent background processes, and the complete graph recovers the standard Hawkes process introduced by Hawkes (1971). Making $\hbar$ a probability density endows $\boldsymbol{W}$ with units of "expected number of spikes" and allows us to compare the relative strength of interactions. The form suggests an intuitive generative model: for each impulse response draw $k \sim \text{Poisson}(w_{n\to n'})$ number of induced spikes and draw the k child spike times i.i.d. from $\hbar$. As we will see, this enables computationally tractable conjugate priors.

We can now write down the joint probability of the probabilistic model,

$$p(\{s_m, c_m, \omega_m\}, \boldsymbol{A}, \boldsymbol{W}, \{\{\theta_{n\to n'}\}\}, \{\lambda_n^{(0)}\}, \{\boldsymbol{z}_n\}, \boldsymbol{\vartheta}) =$$
$$p(\boldsymbol{\vartheta}) \times \overbrace{p(\{\boldsymbol{z}_n\} \mid \boldsymbol{\vartheta})}^{\text{latent variables}} \times \overbrace{p(\boldsymbol{A}, \boldsymbol{W} \mid \{\boldsymbol{z}_n\}, \boldsymbol{\vartheta})}^{\text{network}} \times \overbrace{p(\{\lambda_n^{(0)}\})}^{\text{background}} \times \overbrace{p(\{\theta_{n\to n'}\})}^{\text{impulses}}$$
$$\times \overbrace{p\left(\{s_m, c_m, \omega_m\} \mid \boldsymbol{A}, \boldsymbol{W}, \{\theta_{n\to n'}\}, \{\lambda_n^{(0)}\}\right)}^{\text{augmented likelihood}}. \quad (3.5)$$

Before deriving inference algorithms for this model, however, we pause to consider some of its theoretical properties.

3.3.1 Stability of Network Hawkes Processes

Due to their recurrent, mutually-excitatory nature, Hawkes processes can easily be unstable and give rise to an infinite number of spikes. A stable system must satisfy [*]

$$\lambda_{\mathsf{max}} = \max \; |\operatorname{eig}(\boldsymbol{A} \odot \boldsymbol{W})| < 1$$

(c.f. Daley and Vere-Jones (2003)). For the generative model, we would like to set our hyperparameters such that the prior distribution places little mass on unstable networks. In order to do so, we use tools from random matrix theory.

The circular law describes the asymptotic eigenvalue distribution for $N \times N$ random matrices with entries that are i.i.d. with zero mean and variance σ^2. As N grows, the eigenvalues are uniformly distributed over a disk in the complex plane centered at the origin and with radius $\sigma\sqrt{N}$. In our case, however, the mean of the entries, $\mu = \mathbb{E}[a_{n\to n'} \cdot w_{n\to n'}]$, is not zero. Silverstein (1994) has analyzed such "non-central" random matrices and shown that the largest eigenvalue is asymptotically distributed as $\lambda_{\mathsf{max}} \sim \mathcal{N}(\mu N, \sigma^2)$.

In the simple case of $w_{n\to n'} \sim \operatorname{Gamma}(\kappa, \nu)$ and $a_{n\to n'} \sim \operatorname{Bern}(\rho)$, we have $\mu = \rho\kappa/\nu$ and $\sigma = \sqrt{\rho((1-\rho)\kappa^2 + \kappa)}/\nu$. We are using the rate parameterization of

[*] In this context, λ_{max} refers to an eigenvalue rather than a rate.

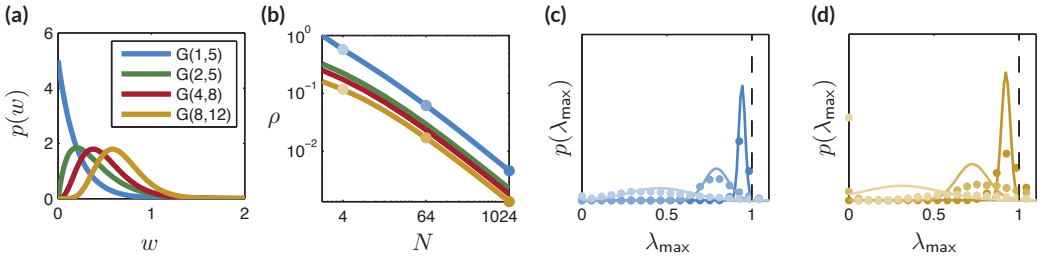

Figure 3.4: Empirical and theoretical distribution of the maximum eigenvalue for independent Bernoulli graphs with gamma weights. (a) Four gamma weight distributions. The colors correspond to the curves in the remaining panels. (b) Sparsity that theoretically yields 99% probability of stability as a function of $p(w)$ and N. (c) and (d) Theoretical (solid) and empirical (dots) distribution of the maximum eigenvalue. Color corresponds to the weight distribution in (a) and intensity indicates N and ρ shown in (b).

the gamma density,

$$\text{Gamma}(w \mid \kappa, \nu) = \frac{\nu^\kappa}{\Gamma(\kappa)} w^{\kappa-1} e^{-\nu w}.$$

For a given N, κ and ν, we can tune the sparsity parameter ρ to achieve stability with high probability. We simply set ρ such that the minimum of $\sigma\sqrt{N}$ and, say, $\mu N + 3\sigma$, equals one. Figures 3.4a and 3.4b show a variety of weight distributions and the maximum stable ρ. Increasing the network size, the mean, or the variance will require a concomitant increase in sparsity.

This approach relies on asymptotic eigenvalue distributions, and it is unclear how quickly the spectra of random matrices will converge to this distribution. To test this, we computed the empirical eigenvalue distribution for random matrices of various size, mean, and variance. We generated 10^4 random matrices for each weight distribution in Figure 3.4a with sizes $N = 4, 64$, and 1024, and ρ set to the theoretical maximum indicated by dots in Figure 3.4b. The theoretical and empirical distributions of the maximum eigenvalue are shown in Figures 3.4c and 3.4d. We find that for small mean and variance weights, for example $\text{Gamma}(1, 5)$ in the Figure 3.4c, the empirical results closely match the theory. As the weights grow larger, as in $\text{Gamma}(8, 12)$ in 3.4d, the empirical eigenvalue distributions have increased variance and lead to a greater than expected probability of unstable matrices for the range of network sizes tested here. We conclude that networks with strong weights should be counterbalanced by strong sparsity limits, or additional structure in the

adjacency matrix that prohibits excitatory feedback loops.

3.4 Bayesian Inference with Gibbs Sampling

We present a Gibbs sampling procedure for inferring the model parameters, $\boldsymbol{A}$, $\boldsymbol{W}$, $\{\lambda_n^{(0)}\}$, $\{\theta_{n\to n'}\}$, and the parameters of the network, $\{\boldsymbol{z}_n\}$ and $\boldsymbol{\vartheta}$. In order to simplify our Gibbs updates, we will also sample a set of parent assignments for each spike $\{\omega_m\}$. Incorporating these parent variables enables conjugate prior distributions and a simple and efficient Gibbs sampling algorithm.

Sampling weights $\boldsymbol{W}$. To derive the updates for weights, recall from (3.4) that $w_{n\to n'}$ only appears in the impulse responses for which $c_m = n$ and $c_{m'} = n'$, so the likelihood is proportional to,

$$\begin{aligned} p(\{s_m, c_m, \omega_m\}_{m=1}^M \mid a_{n\to n'}, w_{n\to n'}, \theta_{n\to n'}) \\ \propto \prod_{m=1}^{M} \left[\exp\left\{-\int_{s_m}^{T} a_{n\to n'} \cdot w_{n\to n'} \cdot \hbar(t - s_m; \theta_{n\to n'})\,\mathrm{d}t\right\}\right]^{\mathbb{I}[c_m=n]} \\ \times \prod_{m=1}^{M}\prod_{m'=1}^{M} \left[w_{n\to n'}\right]^{\mathbb{I}[c_m=n]\mathbb{I}[c_{m'}=n']\mathbb{I}[\omega_{m'}=m]}. \end{aligned}$$

If $a_{n\to n'} = 0$, the impulse response is deterministically zero and, as a result, none of the spikes on neuron n' will be attributed to spikes on neuron n. Thus, the likelihood does not depend on $w_{n\to n'}$. If $a_{n\to n'} = 1$, the likelihood is more complicated. Note, however, that if $s_m < T - \Delta t_{\mathsf{max}}$,

$$\begin{aligned} -\int_{s_m}^{T} a_{n\to n'} \cdot w_{n\to n'} \cdot \hbar(t - s_m; \theta_{n\to n'})\,\mathrm{d}t &= -w_{n\to n'} \int_{s_m}^{T} \hbar(t - s_m; \theta_{n\to n'})\,\mathrm{d}t \\ &= -w_{n\to n'}, \end{aligned}$$

since $\hbar$ is a density defined on $[0, \Delta t_{\mathsf{max}}]$. In general, it is safe to ignore the impulse responses from spikes that occur in the time after $T - \Delta t_{\mathsf{max}}$ since this will be quite small compared to the total recording duration. With this approximation, the conditional distri-

bution of $w_{n\to n'}$ reduces to,

$$p(\{s_m, c_m, \omega_m\}_{m=1}^M \mid a_{n\to n'} = 1, w_{n\to n'}) \propto e^{-M_n \cdot w_{n\to n'}} \, (w_{n\to n'})^{M_{n\to n'}}.$$

where

$$M_n = \sum_{m=1}^{M} \mathbb{I}[c_m = n],$$
$$M_{n\to n'} = \sum_{m=1}^{M} \mathbb{I}[c_m = n] \, \mathbb{I}[c_{m'} = n'] \, \mathbb{I}[\omega_{m'} = m].$$

These sufficient statistics count the number of spikes caused by an connection $n \to n'$ and the total unweighted rate induced by spikes on neuron n.

Now that we have simplified the augmented log likelihood, we see that it is conjugate with a gamma prior on the weights, $w_{n\to n'} \sim \mathrm{Gamma}(\kappa_{n\to n'}, \nu_{n\to n'})$. In Section 3.1 the weight models specified the mean, $\mu_{n\to n'}$. For a gamma distribution, $\mu_{n\to n'} = \frac{\kappa_{n\to n'}}{\nu_{n\to n'}}$. The simplest way to reconcile these is to fix the shape parameter $\kappa_{n\to n'} \equiv \kappa$, then we can compute the mean for any rate parameter, $\nu_{n\to n'}$.

Assuming κ and $\nu_{n\to n'}$ are given, the conditional distribution of the weights is,

$$p(w_{n\to n'} \mid \{s_m, c_m, \omega_m\}_{m=1}^M, a_{n\to n'} = 1, \kappa, \nu_{n\to n'}) = \mathrm{Gamma}(w_{n\to n'} \mid \widetilde{\kappa}_{n\to n'}, \widetilde{\nu}_{n\to n'}),$$

where

$$\widetilde{\kappa}_{n\to n'} = \kappa + M_{n\to n'},$$
$$\widetilde{\nu}_{n\to n'} = \nu_{n\to n'} + M_n.$$

Sampling constant background rates. Similarly, the likelihood of a constant background rate, $\lambda_n^{(0)}$, is conjugate with a gamma prior $\lambda_n^{(0)} \sim \mathrm{Gamma}(\alpha_0, \beta_0)$. The con-

ditional distribution is,

$$p(\lambda_n^{(0)} \mid \{s_m, c_m, \omega_m\}_{m=1}^M, \alpha_0, \beta_0) = \mathrm{Gamma}(\lambda_n^{(0)} \mid \widetilde{\alpha}_{0,n}, \widetilde{\beta}_{0,n}),$$
$$\widetilde{\alpha}_{0,n} = \alpha_0 + \sum_m \mathbb{I}[c_m = n]\, \mathbb{I}[\omega_m = 0]$$
$$\widetilde{\beta}_{0,n} = \beta_0 + T$$

Sampling impulse response parameters $\theta_{n\to n'}$. The logistic-normal density with parameters $\theta_{n\to n'} = \{\mu_{n\to n'}, \tau_{n\to n'}\}$ provides a flexible model for the impulse response:

$$\hbar(\Delta t;\, \mu_{n\to n'}, \tau_{n\to n'}) = \frac{1}{Z} \exp\left\{ \frac{-\tau_{n\to n'}}{2} \left(\sigma^{-1}\left(\frac{\Delta t}{\Delta t_{\mathsf{max}}} \right) - \mu_{n\to n'} \right)^2 \right\}$$
$$\sigma^{-1}(x) = \ln(x/(1-x))$$
$$Z = \frac{\Delta t(\Delta t_{\mathsf{max}} - \Delta t)}{\Delta t_{\mathsf{max}}} \left(\frac{\tau_{n\to n'}}{2\pi} \right)^{-\frac{1}{2}}.$$

Given the auxiliary parent variables, the likelihood is conjugate with a normal-gamma prior $\mu_{n\to n'}, \tau_{n\to n'} \sim \mathcal{NG}(\mu_\mu, \kappa_\mu, \alpha_\tau, \beta_\tau)$. The sufficient statistics are,

$$x_{m\to m'} \triangleq \ln(s_{m'} - s_m) - \ln(t_{\mathsf{max}} - (s_{m'} - s_m)),$$
$$\bar{x}_{n\to n'} = \frac{1}{M_{n\to n'}} \sum_{m=1}^M \sum_{m'=1}^M \mathbb{I}[c_m = n]\, \mathbb{I}[c_{m'} = n']\, \mathbb{I}[\omega_{m'} = m]\, x_{m\to m'},$$
$$v_{n\to n'} = \sum_{m=1}^M \sum_{m'=1}^M \mathbb{I}[c_m = n]\, \mathbb{I}[c_{m'} = n']\, \mathbb{I}[\omega_{m'} = m]\, (x_{m\to m'} - \bar{x})^2.$$

Intuitively, these correspond to the number of spikes attributed to a connection and the mean and variance of their (transformed) delays. The parameters of the normal-gamma

conditional distribution are,

$$\widetilde{\mu}_{n\to n'} = \frac{\kappa_\mu \mu_\mu + M_{n\to n'}\bar{x}_{n\to n'}}{\kappa_\mu + M_{n\to n'}}, \quad \widetilde{\kappa}_{n\to n'} = \kappa_\mu + M_{n\to n'},$$
$$\widetilde{\alpha}_{n\to n'} = \alpha_\tau + \frac{M_{n\to n'}}{2}, \qquad \widetilde{\beta}_{n\to n'} = \frac{v_{n\to n'}}{2} + \frac{M_{n\to n'}\kappa_\mu}{M_{n\to n'}+\kappa_\mu}\frac{(\bar{x}_{n\to n'}-\mu_\mu)^2}{2}.$$

Collapsed Gibbs sampling $\boldsymbol{A}$ and $\boldsymbol{\omega}$. With Aldous-Hoover graph priors, the entries in the binary adjacency matrix $\boldsymbol{A}$ are conditionally independent given the parameters of the prior. The likelihood introduces dependencies between the rows of $\boldsymbol{A}$, but each column can be sampled in parallel. This allows us to parallelize over columns and achieve an $\mathcal{O}(N)$ speedup.

Gibbs updates are complicated, however, by the strong dependencies between the graph and the parent variables. Specifically, if $\omega_{m'} = m$, then we must have $a_{c_m, c_{m'}} = 1$. To improve the performance of our sampling algorithm, first we update $\boldsymbol{A} \mid \{s_m, c_m\}, \boldsymbol{W}, \theta_{n\to n'}$ by marginalizing the parent variables. By the Poisson superposition principle, the marginal distribution is still a Poisson process:

$$\begin{aligned} p(a_{n\to n'} \mid \{s_m, c_m\}, &\boldsymbol{A}_{\neg n\to n'}, \boldsymbol{W}, \boldsymbol{\theta}, \{\boldsymbol{z}_n\}, \boldsymbol{\vartheta}) \\ &\propto \mathcal{PP}\big(\{s_m : c_m = n'\} \,\big|\, \lambda_{n'}(t \mid \mathcal{H}_t)\big) \times p(a_{n\to n'} \mid \boldsymbol{z}_n, \boldsymbol{z}_{n'}, \boldsymbol{\vartheta}) \\ &= \exp\left\{-\int_0^T \lambda_{n'}(t \mid \mathcal{H}_t)\,\mathrm{d}t\right\} \prod_{m=1}^{M} \Big[\lambda_{n'}(s_m \mid \mathcal{H}_t)^{\mathbb{I}[c_m = n']}\Big]\, p(a_{n\to n'} \mid \boldsymbol{z}_n, \boldsymbol{z}_{n'}, \boldsymbol{\vartheta}), \end{aligned}$$

where $\lambda_{n'}(t \mid \mathcal{H}_t)$ depends on $\boldsymbol{A}$, $\boldsymbol{W}$, and $\boldsymbol{\theta}$ through (3.2). Importantly, the integral of the

rate function appearing in the likelihood can be computed without numerical quadrature,

$$\begin{aligned}\int_0^T \lambda_{n'}(t \mid \mathcal{H}_t)\,\mathrm{d}t &= \lambda_{n'}^{(0)} T + \sum_{m=1}^{M} a_{c_m \to n'} \cdot w_{c_m \to n'} \int_0^T \hbar(t - s_m; \theta_{c_m \to n'})\,\mathrm{d}t \\ &\approx \lambda_{n'}^{(0)} T + \sum_{m=1}^{M} a_{c_m \to n'} \cdot w_{c_m \to n'} \\ &= \lambda_{n'}^{(0)} T + \sum_{n=1}^{N} a_{n \to n'} \cdot w_{n \to n'} \cdot M_n.\end{aligned}$$

Again, the approximation stems from ignoring spikes that occur in the final interval of the recording, $(T - \Delta t_{\mathsf{max}}, T]$. For each column, we iterate over incoming edges, $a_{n \to n'}$ and sample from its collapsed distribution, holding all other parameters fixed.

Once the adjacency matrix has been updated, the parent variables are updated by Poisson thinning — that is, by sampling from their discrete conditional distribution. Again, these are all conditionally independent, so the M auxiliary variables can be sampled in parallel.

Sampling Network Variables and Parameters Given the network and the spike train, the conditional distributions for the latent variables, $\{\boldsymbol{z}_n\}$, and the parameters, $\boldsymbol{\vartheta}$ are easy by design.

- *Latent class updates*: If a stochastic block model is used for either the adjacency matrix or the weights, then it is necessary to sample the class assignments from their conditional distribution. We iterate over each neuron and update its assignment given the rest by sampling from the conditional distribution. For example, if z_n governs a stochastic block model for the adjacency matrix, the conditional distribution of the label for neuron n is given by,

$$p(z_n = k \mid \boldsymbol{z}_{\neg n}, \boldsymbol{A}, \boldsymbol{\vartheta}) \propto \pi_k \prod_{n'=1}^{N} p(a_{n' \to n} \mid \rho_{z_{n'} \to k})\, p(a_{n \to n'} \mid \rho_{k \to z_{n'}}), \tag{3.6}$$

  where $\boldsymbol{\vartheta} = \{\boldsymbol{\pi}, \{\rho_{k \to k'}\}\}$. For stochastic block models of the weight matrix, $\boldsymbol{W}$, the conditional distribution depends on $w_{n' \to n}$ and $w_{n \to n'}$ instead.

Given the class assignments and the network, the parameters $\rho_{k\to k'}$, $\mu_{k\to k'}$, and $\boldsymbol{\pi}$ are easily updating according to their conditional distributions, since the model is conjugate.

- *Latent location updates*: We resample the locations using hybrid Monte Carlo (HMC) (Neal, 2010). Since the latent variables are continuous and unconstrained, this method is quite effective.

  In addition to the locations, the latent distance model is parameterized by a location scale, η. Given the locations and an inverse gamma prior, the inverse gamma conditional distribution can be computed in closed form.

  The remaining parameters include the log-odds, γ_0, if the distance model applies to the adjacency matrix, and the baseline mean, μ_0, if it applies to the weight matrix. These can be sampled alongside the locations with HMC.

Computational concerns. The complexity of inference is primarily driven by the number of spikes, M. We must update the auxiliary variable of each spike, and in the worst case there are m potential parents for the m-th spike. Hence, this operation can be at worst $\mathcal{O}(M^2)$ complexity. In practice, compact impulse responses limit the number of potential spike parents and significantly reduce the memory requirements and running time of our algorithm. If we could bound the maximum firing rate at λ_{max}, the complexity of resampling parent variables would be $\sim M\, N\, \lambda_{\mathsf{max}}\, \Delta t_{\mathsf{max}}$. However, these auxiliary variables are conditionally independent so we can save a factor of M by parallelizing their updates, as we do in our parallel implementation.

The second most computationally expensive operation is updating the adjacency matrix and the weights. Note, however, that the columns of the weighted adjacency matrix are conditionally independent. Thus, we can save a factor of N by using block parallel Gibbs sampling. In order to perform these updates, we must compute sufficient statistics that involve sums over M spikes. We have implemented a multithreaded inference algorithm in Python and C that capitalizes on these opportunities for parallelism.[†]

[†] https://github.com/slinderman/pyhawkes

First, we test our inference algorithm on synthetic data generated from the network Hawkes model. We perform two tests: (i) a link prediction task where the process identities are given and the goal is to simply infer whether or not an interaction exists, and (ii) a spike prediction task where we measure the probability of held-out spike sequences.

The network Hawkes model can be used for link prediction by considering the posterior probability of interactions $p(a_{n\to n'} \mid \{s_m, c_m\})$. By thresholding at varying probabilities we compute a ROC curve. A standard Hawkes process assumes a complete set of interactions ($a_{n\to n'} \equiv 1$), but we can similarly threshold its inferred weight matrix to perform link prediction.

Cross correlation provides a simple alternative measure of interaction. By binning the data and summing the cross-correlation over offsets $\Delta t \in [0, \Delta t_{\mathsf{max}})$, we obtain a measure of directed interaction. A probabilistic alternative is offered by the generalized linear model for point processes (GLM), a popular model for spiking dynamics in computational neuroscience (Paninski, 2004). The GLM allows for constant background rates and both excitatory and inhibitory interactions. Impulse responses are modeled with linear basis functions. Area under the impulse response provides a measure of directed excitatory interaction that we use to compute a ROC curve. In Chapter 5, we will discuss generalized linear models for spike trains in great detail.

We sampled ten network Hawkes processes of 30 nodes each with independent Bernoulli graph models, constant background rates, and the conjugate priors described above. The Hawkes processes were simulated for $T = 1000$ seconds. We used the models above to predict the presence or absence of interactions. The results of this experiment are shown in the ROC curves of Figure 3.5a. The network Hawkes model accurately identifies the sparse interactions, outperforming all other models. With the Hawkes process and the GLM we can evaluate the log likelihood of held-out test data. On this task, the network Hawkes outperforms the competitors for all networks. On average, the network Hawkes model achieves $2.2 \pm .1$ bits/spike improvement in predictive log likelihood over a homogeneous Poisson process. Figure 3.5b shows that on average the standard Hawkes and the GLM provide only 60% and 72%, respectively, of this predictive power.

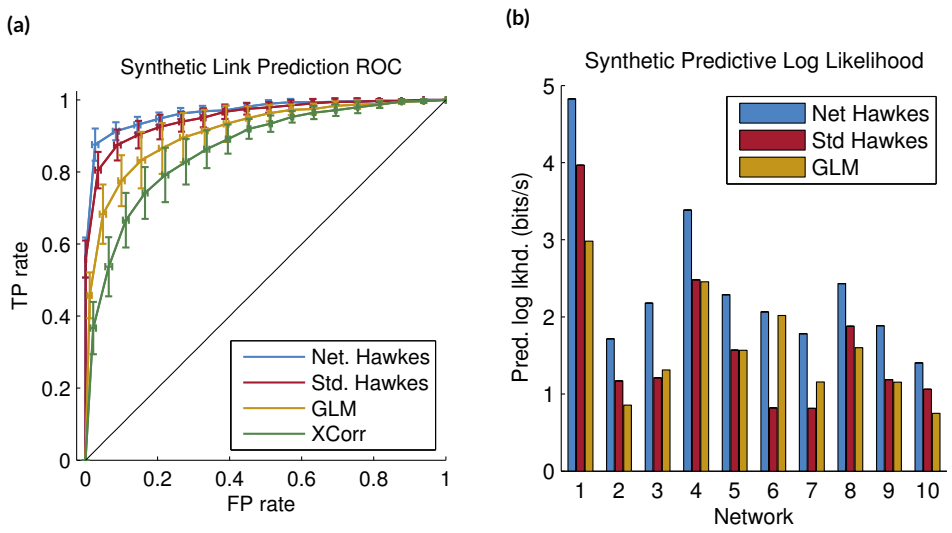

Figure 3.5: (a) Comparison of models on a link prediction test averaged across ten randomly sampled synthetic networks of 30 nodes each. The network Hawkes model with the correct independent Bernoulli graph prior outperforms a standard Hawkes model, GLM, and simple thresholding of the cross-correlation matrix. **(b)** Comparison of predictive log likelihoods, compared to a baseline of a Poisson process with constant rate. Improvement in predictive likelihood over baseline is normalized by the number of spikes in the test data to obtain units of "bits per spike." The network Hawkes model outperforms the competitors in all sample networks.

3.6 Modeling Hippocampal Place Cells

Our first real dataset consists of a simultaneously recorded population of 49 hippocampal place cells from a rat freely foraging in a circular arena roughly 120cm in diameter. The data is courtesy of the lab of Prof. Matthew Wilson at MIT.[‡] The recording duration was roughly 25 minutes. The first 20 minutes were used for model fitting and the last 5 were reserved for predictive tests. Over the entire 25 minute recording, each neuron fired on average 1979 ± 4117 spikes (min: 48, max: 27572) for a total of 97000 spikes. This corresponds to a firing rate of 1.35 ± 2.82Hz (min: 0.32Hz, max: 18.88Hz). The rat's location, $\boldsymbol{x}(t)$, was recorded along with the corresponding spike times. From this, the place field of the n-

[‡] The experiments were conducted under the supervision of the Massachusetts Institute of Technology (MIT) Committee on Animal Care and followed the NIH guidelines. The micro-drive arrays containing multiple tetrodes were implanted above the right dorsal hippocampus of male Long-Evans rats. The tetrodes were slowly lowered into the brain reaching the cell layer of CA1 two to four weeks following the date of surgery. Recorded spikes were manually clustered and sorted to obtain single units using custom software.

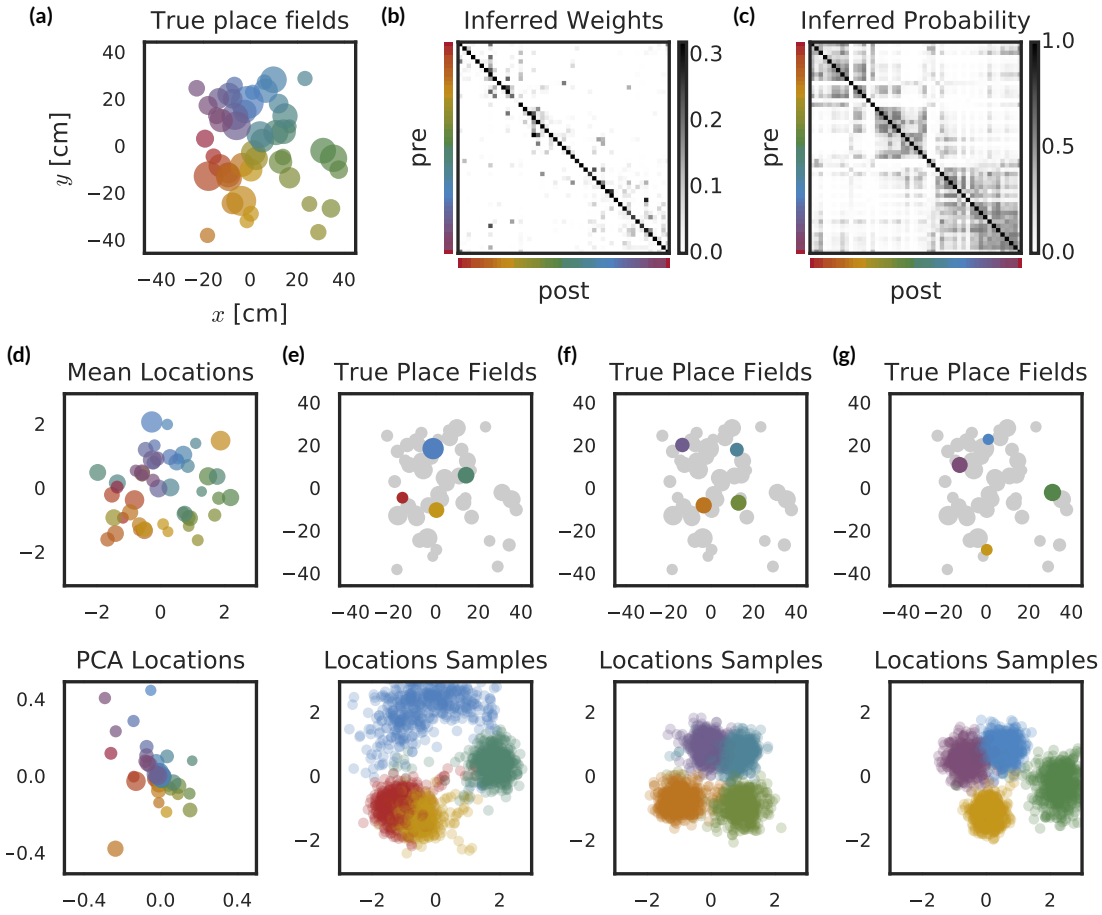

Figure 3.6: Inferred weights and locations of hippocampal place cells using a latent distance model as a prior distribution over the adjacency matrix. **(a):** True place field centers. Marker size is proportional to the size of the place field. Neurons are false colored for identification. **(b):** Expected weights under posterior. Neurons are sorted by location, and colorbars on x and y axes map to colors in **(a)**. **(c):** Expected connection probability under posterior according to latent distance model. **(d):** Mean posterior locations under the latent distance model. For comparison, the embedding found by PCA is plotted below. **(e-g):** Posterior distribution over locations shown four cells at a time. **top:** True locations of the cells. **bottom:** 250 samples from the posterior distribution over neuron locations for the four cells colored above.

th neuron is computed as,

$$\bar{\boldsymbol{x}}_n = \frac{1}{M_n} \sum_{m=1}^{M} \boldsymbol{x}(s_m) \cdot \mathbb{I}[c_m = n],$$

where, again, M_n is the number of spikes fired by neuron n. Likewise, the covariance of the place field is given by,

$$\boldsymbol{V}_n = \left[\frac{1}{M_n}\sum_{m=1}^{M} \boldsymbol{x}(s_m)\boldsymbol{x}(s_m)^\mathsf{T} \cdot \mathbb{I}[c_m = n]\right] - \bar{\boldsymbol{x}}_n \bar{\boldsymbol{x}}_n^\mathsf{T}.$$

This gives us an estimate of the size of the place field. Let $\boldsymbol{X} = \{\boldsymbol{x}_n\}$ denote the 49×2 matrix of place fields.

We compare a few different models for this data. Our baseline is a set of independent Poisson processes with constant firing rates set by maximum likelihood. Next, we consider a standard, densely connected Hawkes process. Third, we fit a network Hawkes process with an independent Bernoulli prior on the adjacency matrix and an independent gamma prior on the weight matrix. This induces sparsity in the connectivity. Finally, we fit a network Hawkes model with a latent distance prior on the adjacency matrix and an independent gamma prior on the weights. In fitting this last model, we infer a distribution over sets of locations, $\boldsymbol{Z} = \{\boldsymbol{z}_n\}$, for the population. Intuitively, we expect these locations to mirror the true place fields since nearby cells are likely to have correlated firing rates, which should be captured by excitatory impulse responses between nearby cells.

Figure 3.6 shows the posterior distribution from the network Hawkes model with a latent distance model prior. Figure 3.6a shows the true place fields of the 49 neurons. The marker size is proportional to the size of the place field, as measured by the largest eigenvalue of $\boldsymbol{V}_n$. The neurons are false colored for identification. Figure 3.6b and 3.6c show the expected weights, $\mathbb{E}[\boldsymbol{W}]$, and the matrix of expected connection probabilities, $\mathbb{E}[\rho_{n\to n'}]$, respectively. The colorbars on the axes map to colors in Figure 3.6a. Nearby neurons have higher probability of connection, as expected. This is reflected in the inferred locations.

Since the latent distance model is invariant to rotation, for each sample $\boldsymbol{Z}^{(\ell)}$, we find the orthogonal matrix, $\boldsymbol{R}^{(\ell)}$, that minimizes $||\boldsymbol{X} - \boldsymbol{R}^{(\ell)}\boldsymbol{Z}^{(\ell)}||_F$ and apply it to obtain a rotated set of locations, $\widetilde{\boldsymbol{Z}}^{(\ell)} = \boldsymbol{R}^{(\ell)}\boldsymbol{Z}^{(\ell)}$. Doing this for each sample yields a set of locations $\{\widetilde{\boldsymbol{Z}}^{(\ell)}\}_{\ell=1}^{L}$. Figure 3.6d (top) shows the mean posterior locations, $\mathbb{E}[\widetilde{\boldsymbol{Z}}]$, and we see that it is qualitatively very similar to the true locations. In contrast, the two dimensional

Hippocampal Model	Pred. log lkhd. (bits/spike)
Standard Hawkes	$0.750 \pm 9.7 \times 10^{-5}$
Net. Hawkes ($\boldsymbol{A} \sim$ Bernoulli Model)	$0.768 \pm 9.2 \times 10^{-5}$
Net. Hawkes ($\boldsymbol{A} \sim$ Latent Distance Model)	$0.766 \pm 9.4 \times 10^{-5}$

Table 3.3: Comparison of hippocampal models on a spike prediction task relative to a homogeneous Poisson process baseline.

PCA embedding is highly skewed.[§]

Finally, panels (e-g) show the samples from the posterior distribution of $\widetilde{\boldsymbol{Z}}$. Since this is difficult to visualize, we show the marginal distribution of four neurons at a time. The true location of the four place fields is identified in the upper panel, and the sampled locations are scattered in the lower panel. Importantly, the relative arrangement of locations is well preserved in the inferred locations. Moreover, cells with larger place fields tend to have larger posterior variance in their locations. This is to be expected since large place fields imply imprecise coding of space and higher correlation with other cells. This illustrates how Hawkes processes combined with latent variable models can provide interpretable portraits of complex datasets and find low-dimensional embeddings that recover intuitive structure.

Table 3.3 lists the predictive likelihoods of the various models relative to a homogeneous Poisson process baseline in units of bits per spike. We see that the sparsity of the network Hawkes model leads to improved predictive performance. In this case, however, the simple independent Bernoulli model and the latent distance model both have similar predictive likelihoods. When the network is largely determined by the training data, the prior has little effect. Thus, the two sparse priors may yield similar predictive performance, even though the latent distance model identifies meaningful latent structure. We will consider additional ways of disambiguating different network models in Chapter 5.

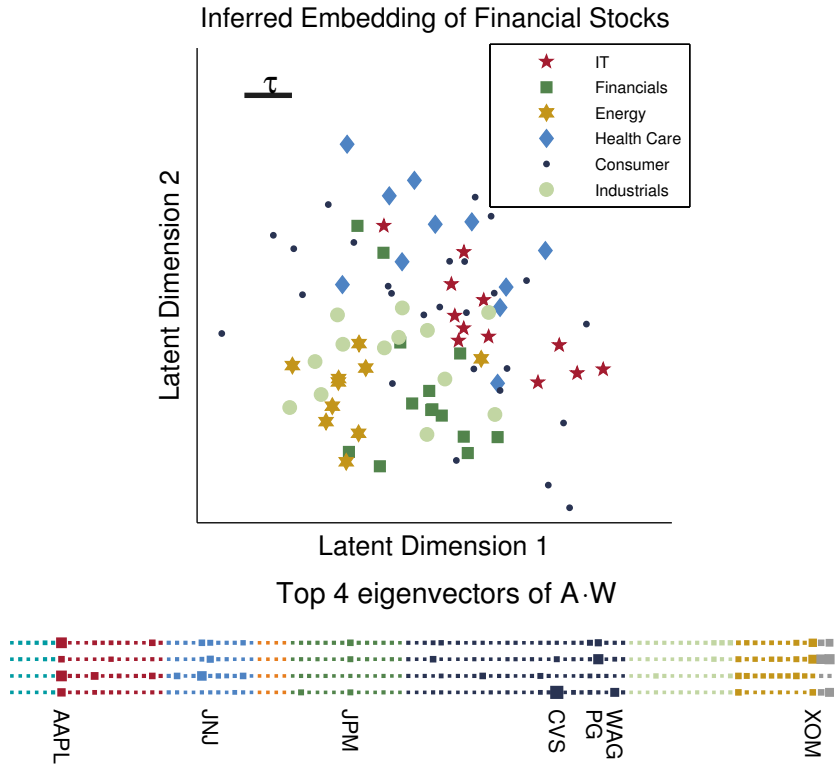

Figure 3.7: Top: A sample from the posterior distribution over embeddings of stocks from the six largest sectors of the S&P100 under a latent distance graph model with two latent dimensions. Scale bar: the characteristic length scale of the latent distance model. The latent embedding tends to embed stocks such that they are nearby to, and hence more likely to interact with, others in their sector. **Bottom**: Hinton diagram of the top 4 eigenvectors. Size indicates magnitude of each stock's component in the eigenvector and colors denote sectors as in the top panel, with the addition of Materials (aqua), Utilities (orange), and Telecomm (gray). We show the eigenvectors corresponding to the four largest eigenvalues $\lambda_{\text{max}} = 0.74$ (top row) to $\lambda_4 = 0.34$ (bottom row).

3.7 Trades on the S&P 100

While the focus of this thesis is on modeling *neural* spike trains, these models have broad applicability outside neuroscience as well. Here we present one example in which we study the trades on the S&P 100 index collected at 1s intervals during the week of Sep. 28 through Oct. 2, 2009. Every time a stock price changes by $\pm 0.1\%$ of its current price a spike is logged on the stock's process, yielding a total of $N = 100$ processes and $M = 182,037$

[§]First, we binned the spikes into 250ms bins, then we smoothed the spike counts with a Gaussian kernel of width 1s to estimate the firing rate. Finally, we applied PCA to the firing rate matrix and used the top two principal components as the embedding.

Financial Model	Pred. log lkhd. (bits/spike)
Independent LGCP	0.594
Standard Hawkes	0.912
Net. Hawkes ($\boldsymbol{A} \sim$ Bernoulli Model)	0.903
Net. Hawkes ($\boldsymbol{A} \sim$ Latent Distance Model)	0.888

Table 3.4: Comparison of financial models on a spike prediction task, relative to a homogeneous Poisson process baseline.

spikes.

Trading volume varies substantially over the course of the day, with peaks at the opening and closing of the market. Rather than attempting to model this background fluctuation with a constant background rate, here we use a log Gaussian Cox process (LGCP) (Møller et al., 1998) with a periodic kernel instead. Complete details of inference are given in Linderman and Adams (2014). We look for short-term interactions on top of this background rate with time scales of $\Delta t_{\text{max}} = 60\text{s}$.

In Figure 3.4 we compare the predictive performance of independent LGCPs, a standard Hawkes process with LGCP background rates, and the network Hawkes model with LGCP background rates under two graph priors. The models are trained on four days of data and tested on the fifth. Though the network Hawkes is slightly outperformed by the standard Hawkes, the difference is small relative to the performance improvement from considering interactions, and the inferred network parameters provide interpretable insight into the market structure.

In the latent distance model for $\boldsymbol{A}$, each stock has a latent embedding $\boldsymbol{z}_n \in \mathbb{R}^2$ such that nearby stocks are more likely to interact, as described in Section 3.1. Figure 3.7 shows a sample from the posterior distribution over embeddings in $\mathbb{R}^2$. We have plotted stocks in the six largest sectors, as listed on Bloomberg.com. Some sectors, notably energy and financials, tend to cluster together, indicating an increased probability of interaction between stocks in the same sector. Other sectors, such as consumer goods, are broadly distributed, suggesting that these stocks are less influenced by others in their sector. For the consumer industry, which is driven by slowly varying factors like inventory, this may not be surprising.

The Hinton diagram in the bottom panel of Figure 3.7 shows the top 4 eigenvectors of

the interaction network. All eigenvalues are less than 1, indicating that the system is stable. The top row corresponds to first eigenvector ($\lambda_{\max} = 0.74$). Apple (AAPL), J.P. Morgan (JPM), and Exxon Mobil (XOM) have notably large entries in the eigenvector, suggesting that their activity will spawn cascades of self-excitation.

3.8 Gangs of Chicago

As a second example of applications outside neuroscience, we study spatiotemporal patterns of gang-related homicide in Chicago. Sociologists have suggested that gang-related homicide is mediated by underlying social networks and occurs in mutually-exciting, retaliatory patterns (Papachristos, 2009). This is consistent with a spatiotemporal Hawkes process in which processes correspond to gang territories and homicides incite further homicides in rival territories.

We study gang-related homicides between 1980 and 1995 (Block et al., 2005). Homicides are labeled by the community in which they occurred. Over this time-frame there were $M = 1637$ gang-related homicides in the 77 communities of Chicago.

We evaluate our model with a spike-prediction task, training on 1980-1993 and testing on 1994-1995. We use a LGCP temporal background rate in all model variations. Our baseline is a single process with a uniform spatial rate for the city. Here, however, the analogous "neurons" are not so clear. We consider two models: (i) the "community" model, which considers each community a separate "neuron," or process, and (ii) the "cluster" model, which groups communities into processes. The number of clusters is chosen by cross-validation (again, see Linderman and Adams (2014)). For each process identity model, we compare four graph models: (i) independent LGCPs (*empty*), (ii) a standard Hawkes process with all possible interactions (*complete*), (iii) a network Hawkes model with a sparsity-inducing independent Bernoulli graph prior, and (iv) a network Hawkes model with a latent distance model that prefers short-range interactions.

The community process identity model improves predictive performance by accounting for higher rates in South and West Chicago where gangs are deeply entrenched. Allowing for interactions between community areas, however, results in a decrease in predictive power due to overfitting (there is insufficient data to fit all 77^2 potential interactions).

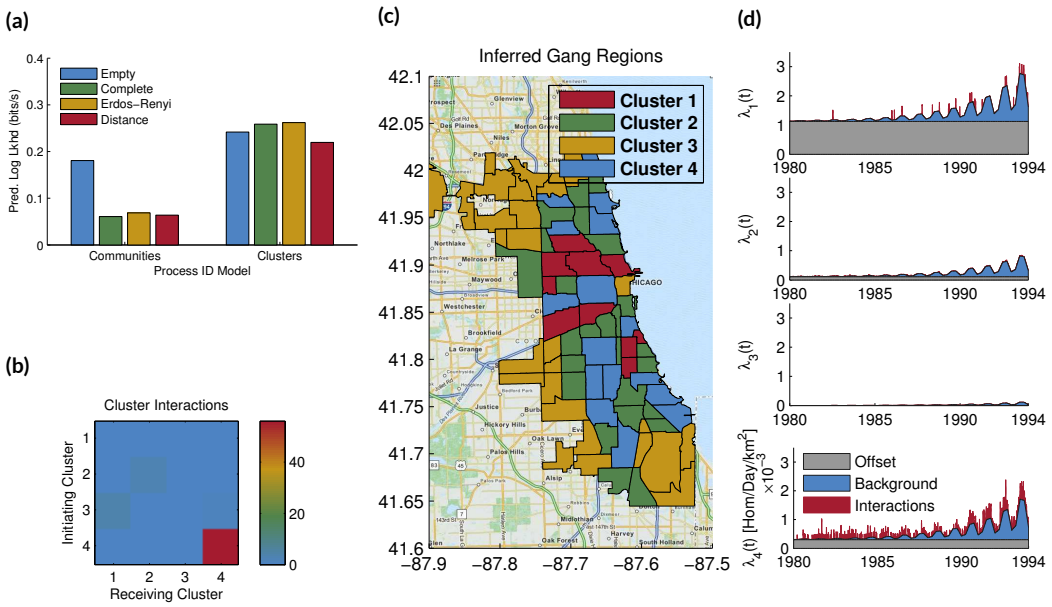

Figure 3.8: Inferred interactions among clusters of community areas in the city of Chicago. **(a)** Predictive log likelihood for "communities" and "clusters" process identity models and four graph models. Panels **(b-d)** present results for the model with the highest predictive log likelihood: an independent Bernoulli graph with $N = 4$ clusters. **(b)** The weighted interaction network in units of induced homicides over the training period (1980-1993). **(c)** Inferred clustering of the 77 community areas. **(d)** The intensity for each cluster, broken down into the offset, the shared background rate, and the interactions (units of 10^{-3} homicides per day per square kilometer).

Interestingly, sparse graph priors do not help. They bias the model toward sparser but stronger interactions which are not supported by the test data. These results are shown in the "communities" group of Figure 3.8a. Clustering the communities improves predictive performance for all graph models, as seen in the "clusters" group. Moreover, the clustered models benefit from the inclusion of excitatory interactions, with the highest predictive log likelihoods coming from a four-cluster independent Bernoulli graph model with interactions shown in Figure 3.8b. Distance-dependent graph priors do not improve predictive performance on this dataset, suggesting that either interactions do not occur over short distances, or that local rivalries are not substantial enough to be discovered in our dataset. More data is necessary to conclusively say which.

Looking into the inferred clusters in Figure 3.8c and their rates in 3.8d, we can interpret the clusters as "safe suburbs" in gold, "buffer neighborhoods" in green, and "gang terri-

tories" in red and blue. Self-excitation in the blue cluster (Figure 3.8b) suggests that these regions are prone to bursts of activity, as one might expect during a turf-war. This interpretation is supported by reports of "a burst of street-gang violence in 1990 and 1991" in West Englewood (41.77°N, −87.67°W) (Block and Block, 1993).

Figure 3.8d also shows a significant increase in the homicide rate between 1989 and 1995, consistent with reports of escalating gang warfare (Block and Block, 1993). In addition to this long-term trend, homicide rates show a pronounced seasonal effect, peaking in the summer and tapering in the winter. An LGCP with a quadratic kernel point-wise added to a periodic kernel captures both effects.

3.9 Related Work

Hawkes processes and latent network discovery have been a subject of recent interest in the machine learning community. Much of this interest stems from the growth of social networking applications which produce massive amounts of spiking data. Gomez-Rodriguez et al. (2010) introduced one of the earliest algorithms for discovering latent networks from cascades of spikes in social network data. They developed a highly scalable approximate inference algorithm, but they did not explore the potential of random network models or emphasize the point process nature of the data. Simma and Jordan (2010) studied this problem from the context of Hawkes processes and developed an expectation-maximization inference algorithm that could scale to massive datasets, like the interactions between authors on Wikipedia. We have adapted their latent variable formulation in our fully-Bayesian inference algorithm and introduced a framework for prior distributions over the latent network.

Others have considered special cases of the model we have proposed. Blundell et al. (2012) combine Hawkes processes and the Infinite Relational Model (a specific exchangeable graph model with an Aldous-Hoover representation) to cluster processes and discover interactions in email networks. Cho et al. (2013) applied Hawkes processes to gang incidents in Los Angeles. They developed a spatial Gaussian mixture model (GMM) for process identities, but did not explore structured network priors. We experimented with this process identity model but found that it suffers in predictive log likelihood tests.

Iwata et al. (2013) developed a stochastic EM algorithm for Hawkes processes, leveraging

similar conjugacy properties, but without network priors. Zhou et al. (2013) have developed a promising optimization-based approach to discovering low-rank networks in Hawkes processes, similar to some of the network models we explored. Guo et al. (2014) have developed a similar model to ours. They focus on applying Hawkes processes to language modeling and incorporating features of the discrete events. DuBois et al. (2013) also explored the use of infinite relational models as a prior in conjunction with a point process observation model build on a Gibbs sampling algorithm.

Perry and Wolfe (2013) derived a partial likelihood inference algorithm for Hawkes processes with a similar emphasis on structural patterns in the network of interactions. They provide an estimator capable of discovering homophily and other network effects. Our fully-Bayesian approach generalizes this method to capitalize on recent developments in random network models (Lloyd et al., 2012).

Finally, generalized linear models (GLMs) are widely used in computational neuroscience (Paninski, 2004). GLMs allow for both excitatory and inhibitory interactions, but, as we have shown, when the data consists of purely excitatory interactions, Hawkes processes outperform GLMs in link- and spike-prediction tests. We will discuss these models in Chapter 5

3.10 Conclusion

This chapter developed a framework for discovering latent network structure from spiking data with mutually excitatory interactions. Our auxiliary variable formulation of the multivariate Hawkes process supports a broad class of prior distributions on latent network structure. This allows us to connect interpretable latent variables, like neuron types and features, to a dynamic model for spike trains. Our parallel MCMC algorithm allowed us to reason about uncertainty in the latent network in a fully-Bayesian manner. We leveraged results from random matrix theory to analyze the conditions under which random network models will be stable, and our applications uncovered interpretable latent networks in a variety of synthetic and real-world problems.

Hawkes processes are the point process analogue of linear autoregressive models. The firing rate is a sum of nonnegative impulse responses induced by preceding spikes. As we gen-

eralize these models in the following chapters, we will exploit this relationship and consider natural extensions like discrete time, nonlinear, and nonstationary versions of the model.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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