8

Switching Linear Dynamical Systems with Count Observations

The past two chapters have explored different notions of latent state: a dynamic network in Chapter 6 and a discrete latent state in Chapter 7. These states are a powerful addition to the autoregressive models of the earlier chapters. In this chapter, we consider one final extension — a continuous latent state that evolves over time. These continuous latent state space models are one of the most common methods in computational neuroscience (Smith and Brown, 2003; Paninski et al., 2010; Macke et al., 2011; Buesing et al., 2012a; Petreska et al., 2011; Cunningham and Yu, 2014).

The simplest form of continuous state space model assumes that the latent state obeys linear dynamics. Here, however, we will consider a more general case in which the dynamics are only *conditionally* linear given a dynamic discrete latent state (Petreska et al., 2011). This is known as a *switching* linear dynamical system (Murphy, 2012; Fox, 2009). By switching between different linear dynamical regimes, we obtain highly nonlinear patterns of dynamics. Moreover, this switching linear dynamical system will recover a number of common models as special cases.

The challenge, as should be expected by now, is in performing efficient inference in the face of discrete observations. The aforementioned existing methods have relied upon a

Laplace approximation, which approximates the conditional distribution with a Gaussian. Given the tools developed in previous chapters, we can now develop asymptotically exact Gibbs sampling algorithms. In particular, the Pólya-gamma augmentations introduced in Chapter 5 will make it easy to develop efficient algorithms that leverage many of the standard tools that exist for Gaussian observation models. Once we have augmented our observations with Pólya-gamma auxiliary variables, the observations are conditionally Gaussian distributed. Thus, all of our tools for efficient Bayesian inference in linear Gaussian models are at our disposal.

Finally, we will consider a problem that we have given little consideration thus far, namely, the problem of model comparison. We have tacitly assumed that predictive likelihoods provide a sufficient means of comparing two models. In practice, this has led to some difficulty, as we encountered with the network model comparison in Chapters 3 and 5. The root of the problem is that predictive likelihood comparisons only implicitly depend on model complexity. More complex models are more prone to overfitting, which should manifest itself in decreased predictive performance. However, there are more direct means of assessing the balance between model complexity and predictive capability. In theory, the marginal likelihood — the denominator in Bayes' rule — should provide a better estimate of the trade-off between how well a model fits the data and the size of the hypothesis class (MacKay, 1992; Kass and Raftery, 1995).

We will show how the conditionally conjugate models derived via Pólya-gamma augmentation enable principled marginal likelihood estimation with annealed importance sampling (AIS) (Neal, 2001). In order to make this practically feasible, however, we must dive into the inner workings of the Pólya-gamma distribution and develop a novel sampling algorithm capable of efficiently generating random variates in the "small shape" regime required by AIS.

A Hierarchy of Latent State Space Models

Consider a general class of models with a continuous latent state, $\boldsymbol{x}_t \in \mathbb{R}^D$, that obeys affine, but potentially nonstationary, dynamics at discrete time t,

$$\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{A}_t \boldsymbol{x}_{t-1} + \boldsymbol{b}_t,\ \boldsymbol{\Sigma}_t).$$

Let the initial state distribution have mean $\boldsymbol{\mu}_1$. Furthermore, assume a linear activation model $\boldsymbol{\psi}_t = \boldsymbol{C}\boldsymbol{x}_t$, where the mean spike count, $s_{t,n}$ is a nonlinear function of the activation, $\psi_{t,n}$, and neuron-specific parameters, ν_n. We refer to the collection of model parameters as,

$$\boldsymbol{\theta} = \left\{ \{\boldsymbol{A}_t, \boldsymbol{b}_t, \boldsymbol{\Sigma}_t\}_{t=1}^T, \boldsymbol{\mu}_1, \boldsymbol{C}, \{\nu_n\}_{n=1}^N \right\}$$

Given these parameters, we can summarize this probabilistic model. In keeping with standard texts (e.g. Murphy, 2012, Chapter 18), we use "Matlab" notation to refer to a sequence of spike count vectors, $\boldsymbol{s}_{1:T}$, and a sequence of latent state vectors, $\boldsymbol{x}_{1:T}$. We have,

$$p(\boldsymbol{s}_{1:T}, \boldsymbol{x}_{1:T} \mid \boldsymbol{\theta}) = p(\boldsymbol{\theta})\, p(\boldsymbol{x}_{1:T} \mid \boldsymbol{\theta})\, p(\boldsymbol{s}_{1:T} \mid \boldsymbol{x}_{1:T}, \boldsymbol{\theta})$$

where

$$\begin{aligned} p(\boldsymbol{x}_{1:T} \mid \boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \prod_{t=2}^{T} \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{A}_t \boldsymbol{x}_{t-1} + \boldsymbol{b}_t, \boldsymbol{\Sigma}_t) \\ p(\boldsymbol{s}_{1:T} \mid \boldsymbol{x}_{1:T}, \boldsymbol{\theta}) &= \prod_{t=1}^{T} p(\boldsymbol{s}_t \mid \boldsymbol{C}\boldsymbol{x}_t, \{\nu_n\}) \\ &= \prod_{t=1}^{T} \prod_{n=1}^{N} p(s_{t,n} \mid \psi_{t,n}, \nu_n). \end{aligned} \tag{8.1}$$

Now consider the special case where there are only $K < T$ unique dynamics and covariance matrices, $\{\boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, and that at any instant in time, the chosen dynamics are specified by the discrete latent variable $z_t \in \{1, \ldots, K\}$. Moreover, suppose this discrete

latent variable follows a Markov model with initial state distribution $\boldsymbol{\pi}_0$ and transition probabilities $\{\boldsymbol{\pi}_k\}_{k=1}^K$, as in the last chapter. Then the dynamics for $\boldsymbol{z}_{1:T}$ and $\boldsymbol{x}_{1:T}$ are,

$$p(\boldsymbol{z}_{1:T} \mid \boldsymbol{\theta}) = \text{Discrete}(z_1 \mid \boldsymbol{\pi}_0) \prod_{t=2}^{T} \text{Discrete}(z_t \mid \boldsymbol{\pi}_{z_{t-1}}).$$

$$p(\boldsymbol{x}_{1:T} \mid \boldsymbol{z}_{1:T}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}_1 \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{z_1}) \prod_{t=2}^{T} \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{A}_{z_t}\boldsymbol{x}_{t-1} + \boldsymbol{b}_{z_t}, \boldsymbol{\Sigma}_{z_t}),$$

This is a *switching linear dynamical system* (SLDS) model (Murphy, 2012; Fox, 2009). At any point in time, the latent state obeys linear dynamics. The particular choice of dynamics switches between K discrete values according to a Markov model.

The SLDS contains a number of other models as special cases:

- When there is only one discrete latent state ($K = 1$), this reduces to a standard linear dynamical system (LDS).
- When there is one discrete latent state and no continuous dynamics ($\boldsymbol{A}_k \equiv 0$), this reduces to factor analysis (FA).
- When (i) the state dimensionality is equal to the number of neurons ($D = N$); (ii) there are no continuous dynamics ($\boldsymbol{A}_k \equiv 0$); and (iii) the emission matrix is the identity ($\boldsymbol{C} \equiv \boldsymbol{I}$), the SLDS reduces to a hidden Markov model. At each point in time, the firing rate is determined solely by $\boldsymbol{b}_{z_t}$.
- When the the conditions of the HMM are met *and* the discrete transition matrix, $\boldsymbol{P}$, has identical rows ($\boldsymbol{\pi}_k \equiv \boldsymbol{\pi}_0$), the SLDS further reduces to a simple mixture model. At each point in time, the discrete latent state is drawn from $z_t \sim \text{Discrete}(\boldsymbol{\pi}_0)$.

The graphical models corresponding to these special cases are shown in Figure 8.1, with the omission of some model parameters to conserve space. This figure is adapted from Figure 2.2. The only model that is not captured here is the autoregressive model since, here, all interaction between spike counts arises through the latent state. Next we show how a single, unified algorithm can support efficient inference in the SLDS and all its special cases.

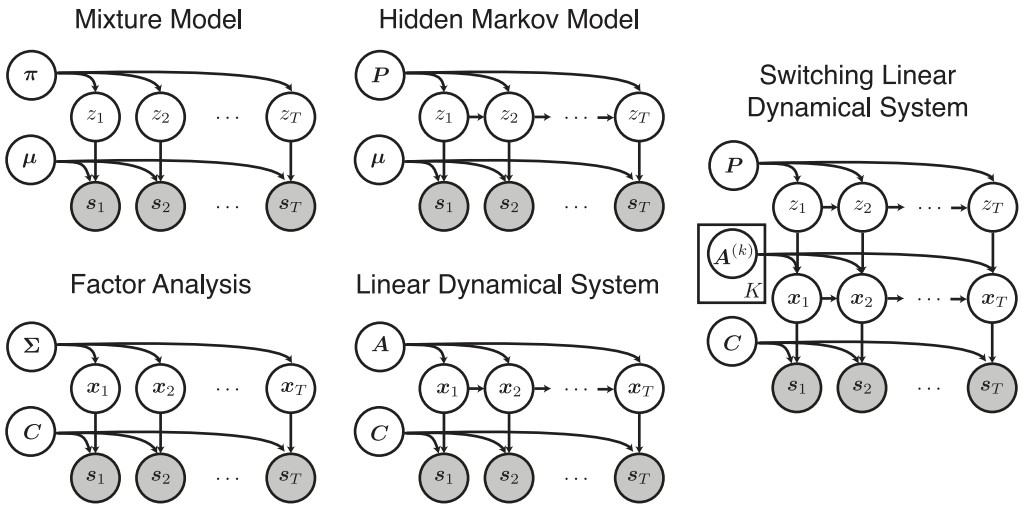

Figure 8.1: Special cases of the switching linear dynamical system. Adapted from Figure 2.2.

Markov Chain Monte Carlo Inference

First we show how the continuous latent states, $\boldsymbol{x}_{1:T}$, can be updated with a block Gibbs sampler when the observations are conditionally Gaussian distributed. The key elements of the inference algorithm will be conserved when we move to discrete count observations. Given the Gaussian inference algorithm, we will show how the Pólya-gamma augmentation explored in Chapter 5 enables efficient Bayesian inference in discrete models as well.

Block Gibbs Sampling Latent States with Gaussian Observations

Suppose the spike counts, $\boldsymbol{s}_t$ are conditionally distributed according to a Gaussian distribution. Moreover, assume the distribution has nonstationary precision, $\boldsymbol{\Omega}_t$, such that

$$p(\boldsymbol{s}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{s}_t \mid \boldsymbol{C}\boldsymbol{x}_t, \boldsymbol{\Omega}_t^{-1}). \tag{8.2}$$

In this case, the conditional density over continuous latent states, $p(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{\theta})$, is jointly Gaussian as well. We perform a block Gibbs update for the entire latent state sequence, $\boldsymbol{x}_{1:T}$, using a forward filtering-backward sampling algorithm, just as we did for the HMM in Section 7.2.1.

The marginal "filtered" distribution given observations up to time t is a Gaussian, which we will denote by,

$$p(\boldsymbol{x}_t \mid \boldsymbol{s}_{1:t}, \boldsymbol{z}_{1:t}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{m}_t, \boldsymbol{V}_t),$$

where $\boldsymbol{m}_t$ and $\boldsymbol{V}_t$ are the filtered mean and covariance, respectively. Kalman filtering is an iterative for computing the filtered means and variances of a Gaussian linear dynamical system, and it is analogous to the HMM filtering algorithms of the previous chapter. Here, we follow the presentation of Murphy (2012, Chapter 18). Kalman filtering consists of iterating forward in time from $t = 1$ to $t = T$. Assume that at iteration t we have already computed $\boldsymbol{m}_{t-1}$ and $\boldsymbol{V}_{t-1}$. As with the HMM, given the Markovian structure of the probabilistic model, the conditional distribution of $\boldsymbol{x}_t$ factors into,

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

We will show that both of these factors are Gaussian distributions, and hence their product is as well.

The first step is to *predict* $\boldsymbol{x}_t$ given observations $\boldsymbol{s}_{1:t-1}$. To do so, we marginalize over the previous latent state, $\boldsymbol{x}_{t-1}$,

$$\begin{aligned} p(\boldsymbol{x}_t \mid \boldsymbol{s}_{1:t-1}, \boldsymbol{z}_{1:t}, \boldsymbol{\theta}) &\propto \int p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, z_t, \boldsymbol{\theta})\, p(\boldsymbol{x}_{t-1} \mid \boldsymbol{s}_{1:t-1}, \boldsymbol{z}_{1:t-1}, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x}_{t-1} \\ &= \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{m}_{t|t-1}, \boldsymbol{V}_{t|t-1}), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{m}_{t|t-1} &\triangleq \boldsymbol{A}_t \boldsymbol{m}_{t-1} + \boldsymbol{b}_t, \\ \boldsymbol{V}_{t|t-1} &\triangleq \boldsymbol{A}_t \boldsymbol{V}_{t-1} \boldsymbol{A}_t^\mathsf{T} + \boldsymbol{\Sigma}_t. \end{aligned}$$

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

distribution,

$$\begin{aligned} \boldsymbol{m}_t &= \boldsymbol{m}_{t|t-1} + \boldsymbol{K}_t(\boldsymbol{s}_t - \boldsymbol{C}\boldsymbol{m}_{t|t-1}), \\ \boldsymbol{V}_t &= (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{C})\boldsymbol{V}_{t|t-1}, \end{aligned} \tag{8.3}$$

where $\boldsymbol{K}_t$ is the "Kalman gain" matrix,

$$\boldsymbol{K}_t \triangleq \boldsymbol{V}_{t|t-1}\boldsymbol{C}^\mathsf{T}\left[\boldsymbol{C}\boldsymbol{V}_{t|t-1}\boldsymbol{C}^\mathsf{T} + \boldsymbol{\Omega}_t^{-1}\right]^{-1}.$$

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

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

Thus, we can sample in reverse order, starting with $\boldsymbol{x}_T$ and ending with $\boldsymbol{x}_1$. The conditional distribution of $\boldsymbol{x}_t$

$$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1:T}, \boldsymbol{s}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{\theta}) \propto \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{m}_t, \boldsymbol{V}_t)\, \mathcal{N}(\boldsymbol{x}_{t+1} \mid \boldsymbol{A}_{t+1}\boldsymbol{x}_t + \boldsymbol{b}_{t+1}, \boldsymbol{\Sigma}_{t+1}), \tag{8.4}$$

which is yet another Gaussian distribution. Now we can write the complete algorithm for block Gibbs sampling the continuous latent states, $\boldsymbol{x}_{1:T}$.

Require: $\boldsymbol{s}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{\theta}$
  for $t = 1, \ldots, T$ **do**
    Compute $\boldsymbol{m}_t$ and $\boldsymbol{V}_t$ ▷ Eq. 8.3
  end for
  for $t = T, \ldots, 1$ **do**
    Sample $\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}, \boldsymbol{m}_t, \boldsymbol{V}_t, \boldsymbol{\theta}$ ▷ Eq. 8.4
  end for

Algorithm 8.1: Forward filtering-backward sampling (FFBS) algorithm for the Gaussian linear dynamical system. Note the similarity to the FFBS algorithm for HMMs in Alg. 7.1.

The conditional distribution of the latent states is only Gaussian if the observations are as well. Fortunately, the observations become conditionally Gaussian after augmenting the data with Pólya-gamma auxiliary variables. Recall from Chapter 5 that the Pólya-gamma augmentation is an auxiliary variable scheme that applies to models with logistic link functions (Polson et al., 2013). Specifically, this augmentation can be used to develop Gibbs for models with likelihoods of the form,

$$\begin{aligned} p(s \mid \psi, \nu) &= c(s, \nu)\, \sigma(\psi)^{a(s,\nu)}\, (1 - \sigma(\psi))^{d(s,\nu)} \\ &= c(s, \nu) \frac{(e^{\psi})^{a(s,\nu)}}{(1 + e^{\psi})^{b(s,\nu)}}. \end{aligned}$$

These are called *logistic likelihoods* because the latent variables are transformed by a logistic function, $\sigma(\psi) = e^{\psi}/(1 + e^{\psi})$. Bernoulli, binomial, negative binomial, and multinomial likelihoods can all be put in this form. For example, in the Bernoulli case,

$$\mathrm{Bern}(s \mid \psi) = \sigma(\psi)^s (1 - \sigma(\psi))^{1-s} = \frac{(e^{\psi})^s}{(1 + e^{\psi})}.$$

Thus, $a(s, \nu) = s$, $b(s, \nu) \equiv 1$, and $c(s, \nu) \equiv 1$. We refer the reader back to Table 5.1 for the formulation of other count distributions.

The augmentation is based on an integral identity derived from the Laplace transform of the Pólya-gamma density. If $p_{\mathrm{PG}}(\omega \mid b, 0)$ is the density of the Pólya-gamma distribution, $\mathrm{PG}(b, 0)$, then,

$$\frac{(e^{\psi})^a}{(1 + e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2}\, p_{\mathrm{PG}}(\omega \mid b, 0)\, \mathrm{d}\omega, \tag{8.5}$$

where $\kappa = a - b/2$. The integral on the right-hand side is the Laplace transform of the Pólya-gamma density evaluated at $\psi^2/2$, and the left-hand side is the same form found in discrete distributions with logistic link functions. Importantly, viewed as a function of ψ for fixed ω, the right-hand side is an unnormalized Gaussian density. Thus, the identity in (8.5) transforms a logistic likelihood to a Gaussian likelihood conditioned on an auxiliary

variable, ω.

Now, let us return to the likelihood of (8.1), where $\psi_{t,n} = [\boldsymbol{C}\boldsymbol{x}_t]_n = \boldsymbol{c}_n^\mathsf{T}\boldsymbol{x}_t$ is the activation of neuron n at time t. As a function of $\boldsymbol{x}_t$, the likelihood is proportional to,

$$\begin{aligned} p(\boldsymbol{s}_t \,|\, \boldsymbol{x}_t, \boldsymbol{\theta}) &\propto \prod_{n=1}^N \frac{(e^{\psi_{t,n}})^{a(s_{t,n},\nu_n)}}{(1+e^{\psi_{t,n}})^{b(s_{t,n},\nu_n)}} \\ &\propto \prod_{n=1}^N e^{\kappa(s_{t,n},\nu_n)\psi_{t,n}} \int_0^\infty e^{-\omega_{t,n}\psi_{t,n}^2/2}\, p_{\mathrm{PG}}(\omega_{t,n} \,|\, b(s_{t,n},\nu_n), 0)\, \mathrm{d}\omega_{t,n}. \end{aligned}$$

By introducing $\omega_{t,n}$ as auxiliary variables, the likelihood of $\boldsymbol{x}_t$ is proportional to a multivariate Gaussian distribution,

$$\begin{aligned} p(\boldsymbol{s}_t \,|\, \boldsymbol{x}_t, \boldsymbol{\omega}_t, \{\nu_n\}) &\propto \prod_{n=1}^N \mathcal{N}(\boldsymbol{c}_n^\mathsf{T}\boldsymbol{x}_t \,|\, \omega_{t,n}^{-1}\kappa(s_{t,n},\nu_n),\, \omega_{t,n}^{-1}) \\ &\propto \mathcal{N}(\widehat{\boldsymbol{s}}_t \,|\, \boldsymbol{C}\boldsymbol{x}_t,\, \boldsymbol{\Omega}_t^{-1}), \end{aligned} \tag{8.6}$$

where

$$\begin{aligned} \boldsymbol{\kappa}_t &= [\kappa(s_{t,1},\nu_1), \ldots, \kappa(s_{t,N},\nu_N)]^\mathsf{T} \\ \widehat{\boldsymbol{s}}_t &= \boldsymbol{\Omega}_t^{-1}\boldsymbol{\kappa}_t \\ \boldsymbol{\Omega}_t &= \mathrm{diag}\left([\omega_{t,1}, \ldots, \omega_{t,N}]\right). \end{aligned}$$

Note the similarity between the augmented likelihood of (8.6) and the Gaussian likelihood of (8.2). The only difference is that, here, the precision is given by the auxiliary variables, and the "effective" observations, $\widehat{s}_{t,n}$, are a function of $s_{t,n}$, $\omega_{t,n}$, and ν_n. Thus, given a set of Pólya-gamma auxiliary variables, the block Gibbs updates in Algorithm 8.1 will apply equally well to the setting with discrete count observations.

Moreover, by the exponential tilting property of the Pólya-gamma distribution, the

conditional distribution of $\omega_{t,n}$ is proportional to a Pólya-gamma distribution:

$$\begin{aligned} p(\omega_{t,n} \mid \psi_{t,n}, s_{t,n}, \nu_n) &\propto e^{-\omega_{t,n}\psi_{t,n}^2/2}\, p_{\mathrm{PG}}(\omega_{t,n} \mid b(s_{t,n}, \nu_n),\, 0) \\ &\propto p_{\mathrm{PG}}(\omega_{t,n} \mid b(s_{t,n}, \nu_n),\, \psi_{t,n}). \end{aligned} \tag{8.7}$$

These auxiliary variables are conditionally independent of each other, and hence amenable to block parallel Gibbs sampling. Efficient Pólya-gamma sampling algorithms have been developed for the regimes typically encountered in Bernoulli, binomial, and negative binomial models (Windle et al., 2014).

The proposed algorithm for sampling the latent variables and parameters of an SLDS is summarized in Algorithm 8.2.

Require: $\boldsymbol{s}_{1:T}$ and $\boldsymbol{z}_{1:T}$, $\boldsymbol{x}_{1:T}$, and $\boldsymbol{\theta}$ from previous iteration
  Sample $\boldsymbol{\theta} \mid \boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}, \boldsymbol{s}_{1:T}$
  Sample $\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T}, \boldsymbol{\theta}$ ▷ Algorithm 7.1
  for $t = 1, \ldots, T$ **do** ▷ In parallel
    for $n = 1, \ldots, N$ **do** ▷ In parallel
      Sample $\omega_{t,n} \mid s_{t,n}, \boldsymbol{x}_t, \boldsymbol{\theta}$ ▷ Eq. 8.7
    end for
  end for
  Compute $\boldsymbol{\Omega}_{1:T}$ and $\widehat{\boldsymbol{s}}_{1:T}$ ▷ Eq. 8.6
  Sample $\boldsymbol{x}_{1:T} \mid \widehat{\boldsymbol{s}}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{\Omega}_{1:T}, \boldsymbol{\theta}$ ▷ Algorithm 8.1

Algorithm 8.2: Single iteration of Gibbs sampler for an switching LDS with discrete count observations.

Missing Data

Sometimes we only have partial observations. For example, in some cases we have multiple recordings from the same circuit, but each recording only provides access to a subset of the population of neurons (Turaga et al., 2013). In other cases, we simply hold out some of the data for predictive likelihood comparisons. With Gaussian observations, we can implement this by replacing the missing data point, $s_{t,n}$, with a zero mean, zero precision observation. In the discrete count model, this can be implemented by setting the auxiliary variable, $\omega_{t,n}$, and the effective observation, $\widehat{s}_{t,n}$, to zero. Recall that the Pólya-gamma auxiliary variables

specify the precision of the effective observations. By setting this to zero, we effectively remove this data point.

Alternative Approaches

Most alternative approaches to performing Bayesian inference in latent state space models with discrete observations have relied on a Laplace approximation (Tierney and Kadane, 1986) to the conditional distribution, $p(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{\theta})$ (Smith and Brown, 2003; Paninski et al., 2010; Macke et al., 2011). * Given a Gaussian approximation, the model parameters, $\boldsymbol{\theta}$, can be optimized such that they maximize the *expected* joint log probability under the approximate Gaussian distribution on $\boldsymbol{x}_{1:T}$. This constitutes an approximate expectation-maximization (EM) algorithm (Dempster et al., 1977).

For completeness, we describe the fundamentals of this approach, largely following the presentation of Macke et al. (2011). Consider a generative model in which $s_{t,n} \sim \mathrm{Poisson}(\exp\{\boldsymbol{c}_n^\mathsf{T}\boldsymbol{x}_t\})$. The conditional log probability of $\boldsymbol{x}_{1:T}$ is given by,

$$\begin{aligned}\log p(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{\theta}) \simeq &\sum_{t=1}^{T}\sum_{n=1}^{N} s_{t,n}(\boldsymbol{c}_n^\mathsf{T}\boldsymbol{x}_t) - \exp\{\boldsymbol{c}_n^\mathsf{T}\boldsymbol{x}_t\} + \\ &- \frac{1}{2}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1)^\mathsf{T}\boldsymbol{\Sigma}_{z_1}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) + \\ &- \frac{1}{2}\sum_{t=2}^{T}(\boldsymbol{x}_t - \boldsymbol{A}_{z_t}\boldsymbol{x}_{t-1} - \boldsymbol{b}_{z_t})^\mathsf{T}\boldsymbol{\Sigma}_{z_t}^{-1}(\boldsymbol{x}_t - \boldsymbol{A}_{z_t}\boldsymbol{x}_{t-1} - \boldsymbol{b}_{z_t}),\end{aligned}$$

where $\simeq$ denotes equality up to an additive constant. This log probability is concave and can be efficiently maximized to obtain the mean of the Laplace approximation,

$$\boldsymbol{\mu}^* = \arg\max_{\boldsymbol{x}_{1:T}} \log p(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{\theta}).$$

Once the mean has been found, the optimal covariance is given by the inverse Hessian of

* While the Laplace approximation is most common, see Buesing et al. (2012b) and Pfau et al. (2013) for some interesting new directions.

the log posterior evaluated at $\boldsymbol{\mu}^*$,

$$\boldsymbol{\Sigma}^* = -\left[\nabla^2_{\boldsymbol{x}_{1:T}} \log p(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{\theta}) \Big|_{\boldsymbol{x}_{1:T}=\boldsymbol{\mu}^*}\right]^{-1}.$$

By exploiting the chain structure of the graphical model, this inverse Hessian can be computed in time linear in T using essentially the same forward-backward approaches used during sampling.

The mean and covariance parameterize a Gaussian approximation,

$$p(\boldsymbol{x}_{1:T} \mid \boldsymbol{s}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{\theta}) \approx q(\boldsymbol{x}_{1:T}) = \mathcal{N}(\boldsymbol{x}_{1:T} \mid \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*).$$

Given this approximation, the parameters are updated by maximizing the expected log probability,

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_q \left[\log p(\boldsymbol{s}_{1:T}, \boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{\theta})\right].$$

For example, consider this expectation as a function of the emission matrix, $\boldsymbol{C}$,

$$\begin{aligned}\mathbb{E}_q \left[\log p(\boldsymbol{s}_{1:T}, \boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{\theta})\right] &\simeq \sum_{t=1}^{T}\sum_{n=1}^{N} s_{t,n}(\boldsymbol{c}_n^\mathsf{T}\mathbb{E}_q[\boldsymbol{x}_t]) - \mathbb{E}_q\left[\exp\{\boldsymbol{c}_n^\mathsf{T}\boldsymbol{x}_t\}\right], \\ &= \sum_{t=1}^{T}\sum_{n=1}^{N} s_{t,n}(\boldsymbol{c}_n^\mathsf{T}\boldsymbol{\mu}_t^*) - \exp\left\{\boldsymbol{c}_n^\mathsf{T}\boldsymbol{\mu}_t^* + \frac{1}{2}\boldsymbol{c}_n^\mathsf{T}\boldsymbol{\Sigma}_{tt}^*\boldsymbol{c}_n\right\}.\end{aligned}$$

The last line follows from the moment generating function of the multivariate Gaussian distribution. This objective function is concave in $\boldsymbol{c}_n$. Note that closed form, concave expectations arise from the particular choice of exponential link function. Other models may require Monte Carlo estimates of the expectation inside the optimization. As more and more approximation is required, the performance of these methods tends to suffer.

Finally, we must handle the discrete latent states, $\boldsymbol{z}_{1:T}$. The simplest approach would be to alternate between updating the discrete and continuous latent states, as in the MCMC algorithm presented above. Alternatively, Petreska et al. (2011) have suggested a joint update

for both $\boldsymbol{x}_{1:T}$ and $\boldsymbol{z}_{1:T}$ based on an approximate filtering technique.

From our perspective, the principal advantages of the Pólya-gamma augmentation are: (i) it allows for simple block Gibbs updates that leverage off-the-shelf code for Gaussian models; (ii) it provides an asymptotically unbiased MCMC algorithm; (iii) the stochasticity of the MCMC transitions allows the sampling algorithm to escape local modes, to which expectation-maximization algorithms are prone (Bishop, 2006); and (iv) once we have an MCMC algorithm, a number of natural extensions are clear, like the marginal likelihood estimation methods we discuss next.

Model Comparison via Marginal Likelihood Estimation

The marginal likelihood is the probability of the data, $\boldsymbol{s}$, having integrated out the latent variables, $\boldsymbol{z}$ and $\boldsymbol{x}$, and the parameters, $\boldsymbol{\theta}$,

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

By integrating over the latent variables and parameters, the marginal likelihood captures a tradeoff between a model's complexity and its ability to explain the data. As such, it is a natural criterion for model comparison. In some cases, like linear Gaussian models with Gaussian observations, the marginal likelihood can be computed exactly. In these cases, marginal likelihood is often the gold-standard for model selection (Kass and Raftery, 1995; Grosse et al., 2015).

Unfortunately, seemingly small changes to the model can render the integration over parameters and latent variables intractable. For example, in linear Gaussian models with discrete observations, the marginal likelihood is no longer tractable. Instead, we must resort to approximate methods like annealed importance sampling (AIS) (Neal, 2001). AIS is based on sampling from a sequence of intermediate distributions that *anneal* between a tractable distribution and the intractable posterior. While AIS has proven highly effective for a variety of models (Grosse et al., 2015), the accuracy of the method hinges upon the efficiency of the Markov transition operators that target the intermediate distributions. Unfortunately, while the posterior distribution may admit efficient MCMC algorithms, the intermediate

distributions may not. We show how the Pólya-gamma augmentation strategies above can be extended to perform efficient annealed importance sampling in the class of switching linear dynamical systems models with count observations.

Annealed Importance Sampling

AIS starts with a sample from a tractable distribution with a computable normalization constant. The prior distribution often suffices. Given this initial sample, a sequence of Markov transition operators is applied. The stationary distributions of these transition operators interpolate between the tractable initial distribution and the intractable posterior. The posterior density is proportional to the joint density, and the normalizing constant is the marginal likelihood of interest. Formally, the annealing path is a sequence of distributions, $q_1(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})$ to $q_M(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x} \mid \boldsymbol{s})$, where

$$q_m(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x}) = \frac{f_m(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})}{\mathcal{Z}_m}, \qquad f_M(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x}) = p(\theta, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{s}), \qquad \mathcal{Z}_M = p(\boldsymbol{s}).$$

Let $q_1(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})$ be the normalized prior distribution such that $f_1(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})$ and $\mathcal{Z}_1 = 1$. Then, let $f_m(\boldsymbol{z}, \theta)$ be a geometric average of the prior and the joint:

$$\begin{aligned} f_m(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x}) &= \Big[p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})\Big]^{1-\beta_m} \Big[p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{s})\Big]^{\beta_m} \\ &= p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})\, p(\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})^{\beta_m}, \end{aligned}$$

with β_m monotonically increasing from $\beta_1 = 0$ to $\beta_M = 1$. As we anneal between $\beta = 0$ and $\beta = 1$, the intermediate distributions interpolate between the prior and the posterior. This geometric path is most common, but any path that starts with a tractable distribution and ends with the posterior will suffice (e.g. Grosse et al., 2013).

In addition to a annealing path, we also need a sequence of MCMC transition operators that leave the intermediate distributions q_m invariant,

$$\mathcal{T}_m(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x} \to \boldsymbol{\theta}', \boldsymbol{z}', \boldsymbol{x}').$$

Starting with a sample from the prior and applying these transition operators

for $m = 1, \ldots, M$ yields a sample that is closer in distribution to the posterior. AIS uses this procedure as a proposal distribution for importance sampling. The importance weights are given by a product of ratios between f_m and f_{m-1}. Since the target density is the unnormalized posterior density, the importance weights will be unbiased estimates of the normalization constant, namely the marginal likelihood, $\mathcal{Z}_M = p(\boldsymbol{s})$. The annealed importance sampling algorithm is summarized in Algorithm 8.3.

for $\ell = 1$ to L **do**
  $w^{(\ell)} \leftarrow \mathcal{Z}_1$
  Sample $\boldsymbol{\theta}^{(1)}, \boldsymbol{z}^{(1)}, \boldsymbol{x}^{(1)} \sim q_1(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})$
  for $m = 2$ to M **do**
    $w^{(\ell)} \leftarrow w^{(\ell)} \times \frac{f_m(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{z}^{(m-1)}, \boldsymbol{x}^{(m-1)})}{f_{m-1}(\boldsymbol{\theta}^{(m-1)}, \boldsymbol{z}^{(m-1)}, \boldsymbol{x}^{(m-1)})}$
    Sample $\boldsymbol{\theta}^{(m)}, \boldsymbol{z}^{(m)}, \boldsymbol{x}^{(m)} \sim \mathcal{T}_m(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x} \leftarrow \boldsymbol{\theta}^{(m-1)}, \boldsymbol{z}^{(m-1)}, \boldsymbol{x}^{(m-1)})$
  end for
end for
return $\widehat{\mathcal{Z}}_M = \frac{1}{L} \sum_{\ell=1}^{L} w^{(\ell)}$

Algorithm 8.3: Annealed Importance Sampling (AIS). Adapted from (Grosse et al., 2015).

How can we reduce the variance of this estimator? First, we can increase the number of intermediate distributions; second, we can design rapidly mixing transition operators, $\mathcal{T}_m$. In this section, we develop transition operators that are both computationally efficient, allowing us to run more transitions in a fixed amount of time, and more effective, in that they reach the equilibrium distribution more quickly.

With a geometric annealing path, the intermediate distributions of the switching LDS are given by,

$$f_m(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x}) \prod_{t=1}^{T} \prod_{n=1}^{N} c(s_{t,n}, \nu_n)^{\beta_m} \frac{(e^{\psi_{t,n}})^{a(s_{t,n},\nu_n)\cdot\beta_m}}{(1 + e^{\psi_{t,n}})^{b(s_{t,n},\nu_n)\cdot\beta_m}}. \tag{8.8}$$

where, again, ν_n is a parameter in $\boldsymbol{\theta}$, and $\psi_{t,n}$ is a function of $\boldsymbol{x}$ and $\boldsymbol{\theta}$. Raising the likelihood to the power β_m does change its functional form; it only changes the power in the exponent. Most importantly, it is still amenable to Pólya-gamma augmentation! Thus, the Gibbs sweep defined in Algorithm 8.2 can be used as a transition operator, $\mathcal{T}_m$. The only

differences in targeting f_m are that,

$$\kappa(s_{t,n}, \nu_n) = \left(a(s_{t,n}, \nu) - \frac{1}{2} b(s_{t,n}, \nu_n)\right) \cdot \beta_m,$$

and

$$p(\omega_{t,n} \mid \psi_{t,n}, s_{t,n}, \nu_n) \propto p_{\mathrm{PG}}(\omega_{t,n} \mid \underbrace{b(s_{t,n}, \nu_n) \cdot \beta_m}_{\text{often } <1}, \psi_{t,n}).$$

This provides some intuition into how AIS works. When β_m approaches zero, the intermediate distribution reduces to the prior. This is equivalent to setting κ and ω to zero. As $b \to 0$, the density Pólya-gamma density, $\mathrm{PG}(\omega \mid b, \psi)$, approaches a delta function at zero.

Note, however, that in order to implement $\mathcal{T}_m$ efficiently, we must be able to sample from the Pólya-gamma conditional distribution in the regime where $b(s_{t,n}, \nu_n) \cdot \beta_m < 1$. For Bernoulli observations, $b(s_{t,n}, \nu_n) \equiv 1$, so we will be in this regime for all $\beta_m \in [0, 1)$. While efficient samplers exist for Pólya-gamma distributed variables when $b(s_{t,n}, \nu_n) \cdot \beta_m \geq 1$ (Windle et al., 2014), the default method for this "small shape" regime is to approximate a Pólya-gamma sample with a truncated sum of gamma random variates (Polson et al., 2013). As the number of random variates in the sum approaches infinity, the approximate sample converges to a true draw from the Pólya-gamma distribution. To get a reasonably accurate draw, we typically need to sample around 200 gamma random variates per Pólya-gamma sample. With TN auxiliary variables, this quickly becomes prohibitively expensive. Next, we develop a novel sampling algorithm that makes these conditional updates very efficient, and renders AIS with Pólya-gamma augmented transitions highly effective.

A Novel Sampling Algorithm for the Pólya-gamma Distribution

The Pólya-gamma distribution, $\mathrm{PG}(b, \psi)$, is closely related to the Jacobi distribution, $J^*(b, \psi)$, surveyed by Biane et al. (2001) and elaborated upon in Windle et al. (2014).

Specifically,

$$Y \sim J^*(b, \tfrac{\psi}{2}) \implies \frac{1}{4}Y \sim \mathrm{PG}(b, \psi).$$

Thus, to develop a sampler for the Pólya-gamma distribution, it is sufficient to be able to sample the Jacobi distribution.

As derived by Windle et al. (2014), the density of $J^*(b, \psi)$ can be written as an infinite alternating sum,

$$p_{J^*}(\omega \mid b, \psi) = \cosh^b(\psi) e^{-\omega\psi^2/2} \frac{2^b}{\Gamma(b)} \sum_{n=0}^{\infty} (-1)^n \frac{\Gamma(n+b)}{\Gamma(n+1)} \frac{(2n+b)}{\sqrt{2\pi\omega^3}} \exp\left\{-\frac{(2n+b)^2}{2\omega}\right\}. \quad (8.9)$$

Windle et al. (2014) developed a number of methods for sampling this distribution. Most rely on finding tractable upper bounds on the density that can serve as a proposal distribution. Given a sample from the proposal, it is possible to accept or reject using the *alternating series method* (Devroye, 1986). We will go into more detail on this shortly.

We take the same basic approach, but we present a novel means of finding an upper bound on the Jacobi density. Massaging terms in (8.9), we can factor it into the product of three terms:

$$p_{J^*}(\omega \mid b, \psi) = \alpha^{-1}(b, \psi)\, p_{\mathrm{IG}}\left(\omega \,\middle|\, \frac{b}{|\psi|},\, b^2\right) \Phi(\omega \mid b). \quad (8.10)$$

The first term, $\alpha^{-1}(b, \psi)$, is a scaling constant greater than one,

$$\alpha^{-1}(b, \psi) = 2^b \cosh^b(\psi)\, e^{-b|\psi|} = \left(1 + e^{-2|\psi|}\right)^b \geq 1.$$

The second is an inverse Gaussian density,

$$p_{\mathrm{IG}}\left(\omega \,\middle|\, \frac{b}{|\psi|},\, b^2\right) = \frac{b}{\sqrt{2\pi\omega^3}} \exp\left\{-\frac{\psi^2}{2\omega}\left(\omega - \frac{b}{|\psi|}\right)^2\right\}.$$

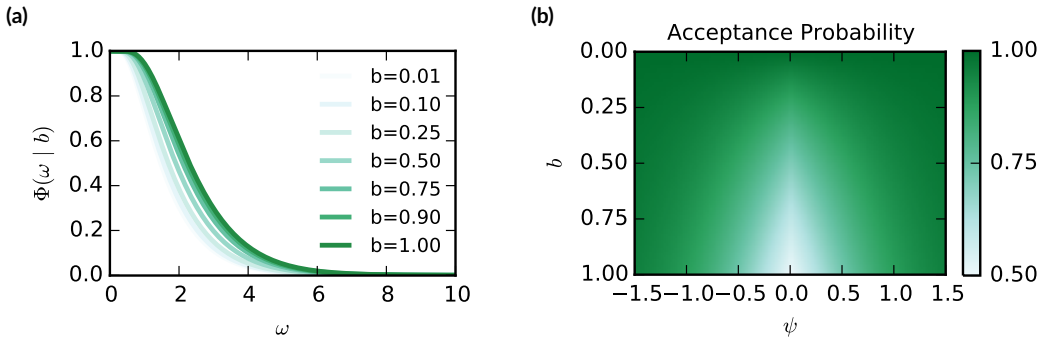

Figure 8.2: **(a)** Plot of $\Phi(\omega \mid b)$, the conditional acceptance probability for a proposed value of ω, for a range of $b \in (0, 1]$. In all cases, this function is monotonically decreasing from 1 to 0 as a function of ω, and thus defines a cumulative distribution function. **(b)** Acceptance probability, $\alpha(b, \psi)$, as a function of b and ψ.

When $\psi = 0$, the inverse Gaussian density reduces to an inverse gamma density,

$$p_{\mathrm{IGa}}(\omega \mid \tfrac{1}{2}, \tfrac{b^2}{2}) = \frac{b}{\sqrt{2\pi\omega^3}} \exp\left\{-\frac{b^2}{2\omega}\right\}.$$

Finally, the third term we have called $\Phi(\omega \mid b)$,

$$\begin{aligned}
\Phi(\omega \mid b) &= \sum_{n=0}^{\infty} (-1)^n \phi_n(\omega \mid b) \\
\phi_n(\omega \mid b) &= \frac{\Gamma(n+b)}{\Gamma(n+1)} \frac{2n+b}{\Gamma(b+1)} \exp\left\{-\frac{2n(n+b)}{\omega}\right\},
\end{aligned}$$

where each term, $\phi_n(\omega \mid b)$, is nonnegative, and $\phi_0(\omega \mid b) = 1$.

Figure 8.2a plots $\Phi(\omega \mid b)$ for various values of b. In all cases, it appears that $\Phi(\omega \mid b)$ is monotonically decreasing and its range is $[0, 1]$. We have not proven this, but our numerical experiments suggest that it is true. We formalize this as a conjecture:

Conjecture 1. *For all $b > 0$, $\Phi(\omega \mid b)$ is a monotonically decreasing function of ω with,*

$$\lim_{0 \leftarrow \omega} \Phi(\omega \mid b) = 1,$$
$$\textit{and,} \quad \lim_{\omega \to \infty} \Phi(\omega \mid b) = 0.$$

Assuming this conjecture is true, as our plots suggest, all three terms in (8.10) are

nonnegative. With $\Phi(\omega \mid b) \leq 1$, the product $\alpha^{-1}(b, \psi)\, p_{\text{IG}}(\omega \mid \frac{b}{|\psi|}, b^2)$ must dominate $p_{J^*}(\omega \mid b, \psi)$. Thus, the inverse Gaussian is a natural proposal distribution for a rejection sampling algorithm. To determine whether a proposed value of ω is accepted, we must sample $u \sim \text{Unif}(0, 1)$, and check whether $u < \Phi(\omega \mid b)$.

The acceptance probability is $\alpha(b, \psi)$, the inverse of the scaling constant. It is bounded between $[\frac{1}{2}, 1]$ when $b \leq 1$. The lower bound (worst case) is achieved when $\psi = 0$ and $b = 1$. The upper bound (best case) is approached as b goes to zero or $|\psi|$ goes to infinity. This is illustrated in Figure 8.2b for a range of b and ψ. In fact, this rejection sampling algorithm works for $b \geq 1$ as well, but as b increases, the acceptance probability goes to zero. For this regime, the existing approaches of Windle et al. (2014) are a better choice.

Determining acceptance In order to determine whether to accept or reject a proposed value of ω, we need to compare against $\Phi(\omega \mid b)$. This function is not analytically tractable; however, it is still possible to determine whether or not to accept with finite computation. To do so, we use a slight modification of the alternating series method (Devroye, 1986). We exploit the fact that $\Phi(\omega \mid b)$ is an alternating sum, and the terms, $\phi_n(\omega \mid b)$, are eventually monotonically decreasing as a function of the index n for all fixed values of ω and b. We formalize this with the following lemma,

Lemma 1. *For all fixed values of ω and b,*

$$\exists m : \forall n \geq m : \phi_{n+1}(\omega \mid b) < \phi_n(\omega \mid b).$$

Proof. We show that the ratio of ϕ_{n+1} to ϕ_n is a decreasing function whose limit is zero.

$$\begin{aligned}
\frac{\phi_{n+1}(\omega \mid b)}{\phi_n(\omega \mid b)} &= \frac{\Gamma(n+1)\Gamma(n+1+b)(2n+2+b)\exp\left\{-\frac{2(n+1)(n+1+b)}{\omega}\right\}}{\Gamma(n+2)\Gamma(n+b)(2n+b)\exp\left\{-\frac{2n(n+b)}{\omega}\right\}} \\
&= \frac{(n+b)(2n+b+2)}{(n+1)(2n+b)}\exp\left\{-\frac{4n+2b+2}{\omega}\right\} \\
&= \ell(n)r(n),
\end{aligned}$$

where

$$\begin{aligned}\ell(n) &= \frac{2n^2 + nb + 2n + b + 2(n+1)b + b^2}{2n^2 + nb + 2n + b} \\ &= 1 + \mathcal{O}\left(\tfrac{1}{n}\right), \\ r(n) &= \exp\left\{-\frac{4n + 2b + 2}{\omega}\right\}.\end{aligned}$$

Observe that $\ell(n)$ is monotonically decreasing toward one as n approaches infinity. The rate of convergence is inverse polynomial in n. In contrast, $r(n)$ decreases to zero exponentially quickly as n approaches infinity. Thus, there exists a threshold m such that this ratio is less than one for all $n \geq m$. Equivalently,

$$\forall\, n \geq m :\ \phi_{n+1}(\omega \mid b) \leq \phi_n(\omega \mid b).$$

□

Lemma 1 guarantees that once we have computed the increasing terms, all subsequent partial sums for even n are upper bounds, and all subsequent partial sums for odd n are lower bounds on $\Phi(\omega \mid b)$. To determine acceptance of u, we evaluate until we find an upper bound less than u, at which point we reject, or a lower bound greater than u, at which point we accept. In practice, determining acceptance takes only a small number of iterations.

Algorithm 8.4 provides pseudocode for the final rejection sampling algorithm.

Require: $b > 0, \psi \in \mathbb{R}$
  accept $\leftarrow$ False
  while not accept **do**
    $\omega \sim \mathrm{IG}\left(\frac{b}{|\psi/2|}, b^2\right)$ ▷ Inv. Gaussian proposal
    $u \sim \mathrm{Unif}(0, 1)$ ▷ Sample acceptance variable
    $\Phi = 1$ ▷ Initialize partial sum with first term
    for $n = 1$ to ∞ **do**
      $\Phi \leftarrow \Phi + (-1)^n \phi_n(\omega \mid b)$ ▷ Update partial sum
      if $\phi_n(\omega \mid b) < \phi_{n-1}(\omega \mid b)$ **then** ▷ Check if terms are decreasing
        if n odd **and** $u \leq \Phi$ **then** ▷ Compare to lower bound
          accept $\leftarrow$ True ▷ Accept and return
          break
        end if
        if n even **and** $u > \Phi$ **then** ▷ Compare to upper bound
          break ▷ Reject and make new proposal
        end if
      end if
    end for
  end while
  return $\frac{1}{4}\omega$

Algorithm 8.4: A rejection sampling algorithm for the Pólya-gamma distribution that is most efficient in the "small-shape" ($b < 1$) regime.

Conclusion

This chapter has explored various facets of modeling neural spike trains with switching linear dynamical systems models. This powerful model for nonlinear dynamical systems contains a number of simpler models as special cases. We have shown how a simple MCMC inference algorithm based on the Pólya-gamma augmentation provides a unified means of performing inference for the SLDS and its special cases.

As we consider hierarchical models like these — models constructed out of layers of latent structure — we must turn our attention to the important question of model selection. How should we justify our modeling choices? Marginal likelihood estimates provide one answer to this question. We have shown how the same Pólya-gamma augmentations can be

applied inside annealed importance sampling algorithms, one of the most successful means of approximating marginal likelihoods. In order to make these methods work in practice, however, we needed to improve the efficiency of sampling the Pólya-gamma distribution in the "small shape" regime. By leveraging a particular decomposition of the related Jacobi density, we derived a novel rejection sampling algorithm with acceptance probability of at least one half.

Next, we turn our attention to another important question. For all their structure, what can these models teach us about neural computation? The next chapter provides some initial attempts to connect the methods we have developed thus far to more abstract theoretical models of neural computation.

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.