BAYESIAN LATENT STATE SPACE MODELS OF NEURAL ACTIVITY

By Scott W. Linderman, Aaron Tucker, and Matthew J. Johnson

Harvard University

Abstract. Latent state space models such as linear dynamical systems and hidden Markov models are extraordinarily powerful tools for gaining insight into the latent structure underlying neural activity. By beginning with simple hypotheses about the latent states of neural populations and incorporating additional beliefs about the nature of this state and its dynamics, we can compose a nested sequence of increasingly sophisticated models and evaluate them in a statistically rigorous manner. Unfortunately, inferring the latent states and parameters of these models is particularly challenging when presented with discrete spike counts, since the observations are not conjugate with latent Gaussian structure. Thus, we often resort to model-specific approximate inference algorithms which preclude rapid model iteration and typically provide only point estimates of the model parameters. As a result, it is difficult compare models in a way that is robust to the approximation and the particular estimates of the model parameters. Here, we develop a unified framework for composing latent state space models and performing efficient Bayesian inference by leveraging a data augmentation strategy to handle the discrete spike count observations. This framework is easily extensible, as we demonstrate by developing an array of latent state space models with a variety of discrete spike count distributions and fitting them to a simultaneously recorded population of hippocampal place cells. Our Bayesian approach yields a posterior distribution over latent states and parameters, which enables robust prediction and principled model comparison. Moreover, we show that our method is at least as fast as alternative approaches in real-world settings.

1. Latent State Space Models. Neural recordings often take the form of a matrix of observed spike counts, $\boldsymbol{S}$, whose entries, $s_{n,t}$, denote the number of spikes emitted by neuron n at time t. Depending on the recording, "neurons" may be replaced by voxels or features, and "spikes" may be replaced by fluorescence or another measurable signal. Latent state space models begin with the assumption that these observations are a noisy function of some unobserved instantaneous "activation," $\psi_{n,t}$, and then hypothesize various models for how the activation is coupled across neurons and time. For example, principal components analysis (PCA) corresponds to the hypothesis that the activation is well modeled as, $\boldsymbol{\psi}_t = \boldsymbol{C}\boldsymbol{x}_t$, where $\boldsymbol{x}_t$ represents a low dimensional state at time t, and $\boldsymbol{C}$ is a linear mapping from latent states to activations. This corresponds to the hypothesis that spike counts are often correlated across neurons. Linear dynamical systems (LDS's) extend PCA by modeling how the latent states evolve over time. Specifically, an LDS assumes, $\boldsymbol{x}_t \sim \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t$, where the matrix $\boldsymbol{A}$ specifies the linear dynamics and $\boldsymbol{\epsilon}_t$ represents Gaussian noise. This captures the temporal correlations in spike counts. Hidden Markov models (HMM's) characterize population activity in terms of a *discrete* latent state $z_t \in \{1, \ldots, K\}$, each of which corresponds to a different activation vector, $\boldsymbol{\psi}^{(k)}$. Then, the instantaneous activation is $\boldsymbol{\psi}_t = \boldsymbol{\psi}^{(z_t)}$. This may be an appropriate model for populations that switch between "on" and "off" states, for example. Combining HMM's and LDS's, we arrive at a *switching* linear dynamical system (SLDS) model with both continuous and discrete latent states. When in discrete state, z_t, the continuous state, $\boldsymbol{x}_t$, is governed by dynamics matrix, $\boldsymbol{A}^{(z_t)}$. Though the instantaneous dynamics are always linear, switching between these discrete latent states allows us to model complex, nonlinear dynamics of the neural population. From this perspective, it is clear that these seemingly disparate models are really special cases of one another, and they correspond to a sequence of models that encode increasingly sophisticated hypotheses about the latent structure underlying neural activity.

When the observed signal is well modeled by a Gaussian distribution, the conditional distribution over continuous latent states, $\boldsymbol{x}_t$, is Gaussian and can be computed in closed form, and we can leverage a host of off-the-shelf inference algorithms. However, when modeling discrete spike counts, a Bernoulli, Poisson model is more appropriate. In cases where the spike counts are overdispersed, a negative binomial model may provide an even better fit. Unfortunately, these discrete models are not conjugate with the Gaussian latent states and inference is considerably more complicated. Substantial work has gone into developing approximate inference algorithms for such models (Macke et al., 2011), but these methods rely on approximations to the model. Though these approximations are fast and effective in practice, they yield asymptotically biased inferences. Moreover, they often provide only a point estimate of the latent states and parameters, which does not permit robust model comparison. Here, we present a simpler, fully-conjugate alternative.

2. Bayesian Inference. Consider a Bernoulli observation model, $p(s_{n,t} \,|\, \psi_{n,t}) = \mathrm{Bern}(s_{n,t} \,|\, \sigma(\psi_{n,t}))$. Dropping the dependence on n and t for notational convenience, this likelihood is equal to, $p(s \,|\, \psi) = c(s)\,(e^{\psi})^{a(s)}\,(1+e^{\psi})^{-b(s)}$, where $a(s) = s$, $b(s) = 1$, and $c(s) = 1$. Polson, Scott and Windle (2013) have developed an elegant strategy for performing inference in models with likelihoods of this form. Their strategy leverages the following integral identity,

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

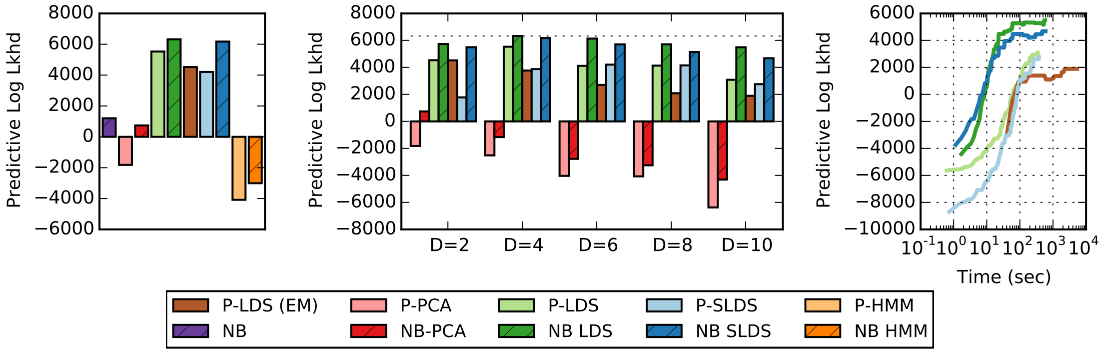

Fig 1: A comparison of latent state space models (PCA, LDS, SLDS, and HMM) with either Poisson (P) or negative binomial (NB) observations fit by our Pólya-gamma augmented Gibbs sampler to a population recording of hippocampal place cells. We measure predictive log likelihood on a heldout subset of spike counts and find that negative binomial dynamical systems provide the best account (left). The latent state space dimensionality, D, is fit by cross validation (center), and as the dimension increases the models over-fit the training data. Finally, we compare to a Poisson LDS (PLDS) model fit via EM and find that our algorithms are comparably fast, especially when D is large (right, $D = 10$).

where $\kappa(s) = a(s) - b(s)/2$ and PG denotes the Pólya-gamma density function. The integrand of (1) defines a joint density on (ψ, s, ω) which admits $p(\psi, s)$ as a marginal density. Moreover, inference in this augmented model is fast and efficient because, given ω, the conditional distribution of ψ is Gaussian:

$$(2) \qquad p(\psi \,|\, s, \omega) \propto p(\psi) e^{\kappa(s)\psi} e^{-\omega\psi^2/2} \propto p(\psi)\, \mathcal{N}(\omega^{-1}\kappa(s) \,|\, \psi, \omega^{-1}).$$

Thus, we can once again use off-the-shelf algorithms to infer ψ after augmentation. Furthermore, by the exponential tilting property of the Pólya-gamma distribution, we have $\omega \,|\, \psi, x \sim \mathrm{PG}(b(x), \psi)$, and efficient samplers exist for this distribution. Together, these facts imply that that we can perform inference over the joint distribution. With MCMC, we can compute the marginal posterior over ψ by simply discarding samples of ω. Moreover, this augmentation strategy applies to any model of the form in Eq. 1. For example, the negative binomial distribution, $s \sim \mathrm{NB}(r, \sigma(\psi))$, corresponds to $a(s) = s$, $b(s) = s + r$, and $c(s) = \binom{s+r-1}{r}$. The Poisson model, $s \sim \mathrm{Poiss}(e^{\psi})$, does not immediately match this form, but it can be approximated by $\mathrm{NB}(r, \sigma(\psi - \log r))$ for large r.

3. Results. First, we studied a population of 47 hippocampal place cells recorded from a freely moving rat in a circular arena.[1] Spikes were counted in 250ms bins for approximately ten minutes. We held out half of the spike counts (randomly sampled) to compute the predictive likelihood of each model relative to a constant-rate Poisson model baseline. Figure 1 shows the results of our comparison for a negative binomial model with constant activation and PCA, HMM's, LDS's, and switching LDS's with both Poisson and negative binomial observations, all fit using our augmented MCMC inference algorithm. We found that the simple PCA models over-fit the training data and performed poorly on generalization tasks. HMMs, which have no low dimensional continuous latent state to tether neurons together, performed even worse on predictive tasks. By contrast, the LDS and SLDS models and SLDS models exploited temporal dynamics in order to inform latent state estimates. We explored the effect of the latent state space dimensionality (center panel) and found that the LDS and SLDS models generalized well with 4 to 6 dimensional latent states. In all cases, the negative binomial observation models provided a better fit to the data, suggesting that these spike counts were indeed overdispersed (i.e. had variance larger than the mean). With our framework, fitting a negative binomial model requires simply changing the coefficients of the observation model.

We also compared to a Poisson LDS fit via EM, following (Macke et al., 2011), and found that predictive log likelihood estimates with state sequences drawn from the variational posterior yield poor predictive estimates. For this dataset, samples from the true, non-Gaussian posterior yield more accurate results. Contrary to common beliefs about MCMC, our inference algorithms are able to explore the posterior parameter space with block Gibbs updates and achieve performance that is comparable to EM algorithms, which must solve a large convex optimization problem at each iteration (right panel).

Our Bayesian treatment of latent state space models provides a unified framework for composing and comparing models of neural activity, and identifying latent structure underlying spike trains. As shown with hippocampal recordings, adopting such an approach allows us to find a parsimonious description of the neural activity quickly and efficiently in models that traditionally pose significant inferential challenges.

Macke, J. H., Buesing, L., Cunningham, J. P., Byron, M. Y., Shenoy, K. V. and Sahani, M. (2011). Empirical models of spiking in neural populations. In *Advances in Neural Information Processing Systems* 1350–1358.

Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American Statistical Association* **108** 1339–1349.

---

[1] Data courtesy of the Wilson lab at MIT.