Efficient Non-conjugate Gaussian Process Factor Models for Spike Count Data using Polynomial Approximations

Stephen L. Keeley [1] **David M. Zoltowski** [1] **Yiyi Yu** [2] **Jacob L. Yates** [3] **Spencer L. Smith** [2] **Jonathan W. Pillow** [1]

Abstract

Gaussian Process Factor Analysis (GPFA) has been broadly applied to the problem of identifying smooth, low-dimensional temporal structure underlying large-scale neural recordings. However, spike trains are non-Gaussian, which motivates combining GPFA with discrete observation models for binned spike count data. The drawback to this approach is that GPFA priors are not conjugate to count model likelihoods, which makes inference challenging. Here we address this obstacle by introducing a fast, approximate inference method for non-conjugate GPFA models. Our approach uses orthogonal second-order polynomials to approximate the nonlinear terms in the non-conjugate log-likelihood, resulting in a method we refer to as *polynomial approximate log-likelihood* (PAL) estimators. This approximation allows for accurate closed-form evaluation of marginal likelihoods and fast numerical optimization for parameters and hyperparameters. We derive PAL estimators for GPFA models with binomial, Poisson, and negative binomial observations and find the PAL estimation is highly accurate, and achieves faster convergence times compared to existing state-of-the-art inference methods. We also find that PAL hyperparameters can provide sensible initialization for black box variational inference (BBVI), which improves BBVI accuracy. We demonstrate that PAL estimators achieve fast and accurate extraction of latent structure from multi-neuron spike train data.[1]

1. Introduction

Recent advances in neural recording technologies have enabled the collection of increasingly high-dimensional neural data-sets. Making sense of such data requires new statistical methods for extracting shared latent structure underlying multi-neuron responses. Factor models provide one popular approach to this problem (Archer et al., 2015; Cunningham & Yu, 2014; Lakshmanan et al., 2015; Wu et al., 2017; Yu et al., 2009). These models seek to characterize the structure underlying neural data in terms of a small number of latent variables. These models have been widely successful in both uncovering interpretable structure from neural population data and providing insight into representations of stimulus input and behavior in population activity (Wu et al., 2017; Zhao & Park, 2017; Zhao et al., 2019). However, factor models can be cumbersome to learn when the prior distribution over the latent variables and the likelihood governing the observations are non-conjugate. This arises commonly for neural data, where binned spiking observations are best characterized by count models (e.g., binomial, Poisson, and negative-binomial).

Formally, latent factor models seek to explain shared structure underlying high-dimensional observations $(\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_T}) \in \mathbb{R}^{N \times T}$ in terms of low-dimensional latent variables $(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_T}) \in \mathbb{R}^{P \times T}$, where $N > P$ and the observations are ordered sequentially in time from $t = 1$ to $t = T$. A popular approach is to model the time series of latent variables with a Gaussian process (GP), which makes few assumptions about latent trajectories beyond the fact that they evolve smoothly in time. When combined with a Gaussian observations model, the resulting approach is known as Gaussian Process Factor Analysis (GPFA) (Yu et al., 2009). Recent work has extended GPFA to incorporate Poisson observations, which provides a more appropriate model for spike train data (Buesing et al., 2012; Macke et al., 2011; Wu et al., 2017; Zhao & Park, 2017; Zhao et al., 2019). However, closed-form inference under GPFA models is only possible when the model likelihood and prior are conjugate. Consequently, Poisson and other non-conjugate models require approximations to fit hyperparameters or obtain parametric expressions for the posterior distribution over latents.

[1] Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA [2] Dept. of Electrical Computer Engineering, University of California Santa Barbara Santa Barbara, CA, USA [3] Department of Biology, University of Maryland, MD, USA. Correspondence to: Stephen Keeley <skeeley@princeton.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

[1] Code available at https://github.com/skeeley/Count_GPFA

Here, we introduce a novel procedure for learning non-conjugate GPFA models with count observations, which we refer to as Polynomial Approximate Log-likelihood (PAL). This method exploits an idea for rapid inference in generalized linear models using so-called "approximate sufficient statistics" (Huggins et al., 2017; Zoltowski & Pillow, 2018), and extends it to the latent variable model setting. The basic idea involves approximating the nonlinear terms in the model log-likelihood using orthogonal polynomials. When the polynomial approximation is second-order, the likelihood term can be explicitly marginalized to obtain a closed-form expression for the marginal likelihood, and an approximately Gaussian posterior distribution over the latents. We explicitly derive PAL estimators for three GPFA models with different count statistics. This includes the previously implemented Poisson count-observation GPFA model (Zhao & Park, 2017; Zhao et al., 2019), as well as GPFA with binomial and negative-binomial observations. These three distributions (binomial, Poisson, and negative-binomial) have different dispersion characteristics which reflect various spiking properties in neurons in different areas of the brain (Charles et al., 2018; Goris et al., 2014; Linderman et al., 2016).

We compare our PAL approach to Black Box Variational Inference (BBVI), a state-of-the-art method for approximate inference in non-conjugate models that is renowned for its simplicity and adaptability (Archer et al., 2015; Gao et al., 2015; Ranganath et al., 2014) and the variational latent Gaussian Process (vLGP) (Zhao & Park, 2017), a previous algorithm used for Poisson noise GPFA. We find that PAL estimation exhibits comparable performance to these methods, but PAL compares favorably to both of them in that it provides a closed-form expression for marginal likelihood that can be optimized directly; it therefore requires no careful tuning of learning rates, number of Monte Carlo samples, or stopping criteria, and does not suffer from high-variance estimates due to sampling-based evaluation of marginal likelihood. We also find that PAL is faster than these existing algorithms and can accurately recover latent structure in simulated neural data.

We further demonstrate that PAL hyperparameters can be used to initialize BBVI to stabilize and improve inference. We use this combined BBVI + PAL on two different multi-neuron datasets, one from mouse visual cortex and one from primate parietal cortex, under three different choices of count model (binomial, Poisson, and negative binomial). We show that PAL initialized BBVI performs as good or better than BBVI alone. The PAL approach therefore offers a promising avenue for future work on non-conjugate models that arise frequently in the analysis of biological and other data.

2. Count-GPFA models

Consider a dataset consisting of count observations from N neurons over T time bins, $\mathbf{Y} \in \mathbb{N}^{N \times T}$. The count-GPFA model seeks to describe these data in terms of a nonlinearly transformed linear projection of lower-dimensional latent variable $\mathbf{X} \in \mathbb{R}^{P \times T}$, $P < N$, where each latent variable evolves according to an independent Gaussian process. Thus the timecourse of the j'th latent variable, which forms the j'th row of $\mathbf{X}$, has a multivariate normal distribution:

$$\mathbf{x}_j \sim \mathcal{N}(0, K_j), \tag{1}$$

where each K is a $T \times T$ covariance matrix whose (t, t')'th entry is given by the covariance function $k(t, t')$. In this paper, we use the common Gaussian or "squared exponential" covariance function: $k(t, t') = \exp(-(t - t')^2/(2\ell^2))$, which is governed by a single hyperparameter, the "length scale" ℓ, which controls smoothness of the latent process.

The count-GPFA observation model can then be written:

$$\mathbf{Y}|\mathbf{W}, \mathbf{X} \sim \mathcal{P}(f(\mathbf{W}\mathbf{X})) \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{N \times P}$ is a loading matrix, $f(\cdot)$ denotes a nonlinear function that transforms $\mathbf{WX}$ to the appropriate range for a count random variable (e.g., the non-negative reals), and $\mathcal{P}$ denotes a probability distribution for count data.

Fitting the count-GPFA model to data involves inferring the loading weights $\mathbf{W}$ and hyperparameters $\theta = \{\ell_1, \dots \ell_j\}$ via numerical optimization of the marginal likelihood:

$$P(\mathbf{Y}|\mathbf{W}, \theta) = \int P(\mathbf{Y}|\mathbf{W}, \mathbf{X}) P(\mathbf{X}|\theta) d\mathbf{X}. \tag{3}$$

However, non-conjugacy of the count model likelihood $P(\mathbf{Y}|\mathbf{W}, \mathbf{X})$ and Gaussian prior over latents $P(\mathbf{X}|\theta)$ means that this integral cannot be computed in closed form. Likewise, the posterior distribution over latents given the data, given by: $P(\mathbf{X}|\mathbf{Y}, \mathbf{W}, \theta) = P(\mathbf{Y}|\mathbf{X}, \mathbf{W}) P(\mathbf{X}|\theta) / P(\mathbf{Y}|\mathbf{W}, \theta)$, has no closed form expression, where the desired normalizing constant is the marginal likelihood. Fitting and inference therefore rely on approximate inference methods.

3. Polynomial Approximate Log-likelihood (PAL)

Here we propose Polynomial Approximate Log-likelihood (PAL), an approximation scheme for efficient inference in non-conjugate Gaussian latent variable models. The core idea is to approximate terms in the observation model log-likelihood that are nonlinear in $\mathbf{X}$ using orthogonal polynomials. Our approach is inspired by recent work on "polynomial approximate sufficient statistics" for generalized linear

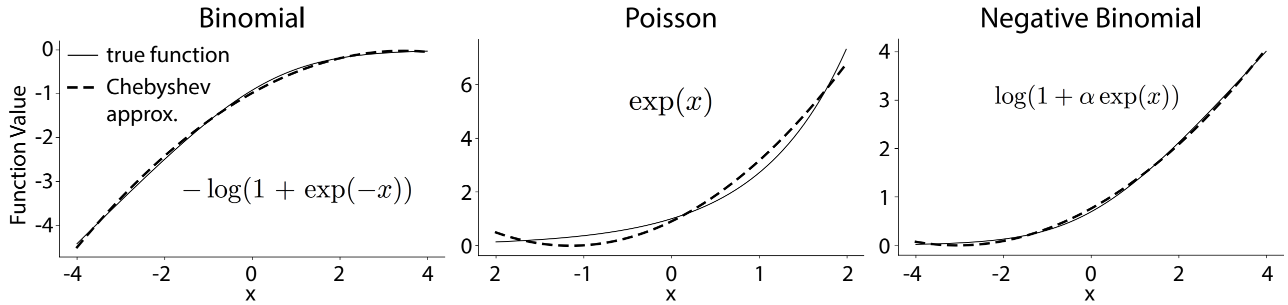

Figure 1: Comparison of nonlinear term found in the log-likelihood for binomial, Poisson, and negative-Binomial observation models (solid) with corresponding second-order Chebyshev approximation (dashed).

models (PASS-GLMs) (Huggins et al., 2017; Zoltowski & Pillow, 2018). In that work, the $\mathbf{X}$ were observed regressors, and the method provided so-called "approximate sufficient statistics" that could be computed with a single pass over the data.

Here, the $\mathbf{X}$ are (unobserved) latent variables instead of regressors, and the goal of the approximation is efficient marginalization rather than a set of sufficient statistics. We consider second-order polynomial approximations to the log-likelihood, which allow for analytic marginalization over latents. PAL therefore enables closed-form evaluation of the approximate marginal likelihood, allowing efficient optimization of parameters and hyperparameters.

We derive PAL estimators for GPFA under three different non-conjugate observation models: binomial, Poisson, and negative binomial (NB). These models range from underdispersed or "sub-Poisson" for binomial to overdispersed or "supra-Poisson" for NB, thus spanning the range of dispersion behaviors found in different brain areas (Charles et al., 2018; Gao et al., 2015; Goris et al., 2014; Kara et al., 2000; Maimon & Assad, 2009; Pillow & Scott, 2012).

All PAL count-GPFA models have the same general form for the approximate log marginal likelihood (log evidence):

$$\mathcal{E}(\mathbf{y}|\mathbf{W},\theta) \approx \frac{1}{2}\log|\mathbf{\Sigma}| + \frac{1}{2}\boldsymbol{\mu}^\top\mathbf{\Sigma}^{-1}\boldsymbol{\mu} - \frac{1}{2}\log|\mathbf{K}|, \quad (4)$$

where $\mathbf{\Sigma}$ denotes an approximate posterior covariance and $\boldsymbol{\mu}$ denotes an approximate posterior mean, and $\mathbf{K}$ is the prior covariance over all latents (a block-diagonal matrix, with one block for each latent). The form of the first two terms varies across models, which we derive for three specific models below. See Table 1 for a summary of the results for all count-GPFA models. For clarity, we define $\mathbf{H} = \mathbf{\Sigma}^{-1} - \mathbf{K}^{-1}$ in this table to succinctly present approximate posterior covariances.

3.1. PAL for Poisson-GPFA

We begin with the Poisson observation model, which is the most common model for spike counts and a popular choice for latent variable models of spike train data (Duncker & Sahani, 2018; Wu et al., 2017; Zhao & Park, 2017). For this model, spike count y given a spike rate parameter λ is distributed according to:

$$P(y|\lambda) = \tfrac{1}{y!}(\Delta\lambda)^y e^{-(\Delta\lambda)}, \quad (5)$$

where Δ is the time bin size (which we set here to 1, resulting in spike rates in units of spikes/bin). We use an exponential nonlinearity from latents to spike rates, so the vector of spike rates at time t is:

$$\boldsymbol{\lambda}_t = \exp(\mathbf{W}\mathbf{x}_t). \quad (6)$$

This choice of nonlinearity gives rise to a log-likelihood with a single nonlinear term, although other nonlinearities can be considered (Zoltowski & Pillow, 2018).

The Poisson log-likelihood for the entire dataset can be written conveniently in vector form as:

$$\mathcal{L}(\mathbf{y},\mathbf{x}|\tilde{\mathbf{W}}) = \mathbf{y}^\top\tilde{\mathbf{W}}\mathbf{x} - \mathbf{1}^\top\exp(\tilde{\mathbf{W}}\mathbf{x}) + const \quad (7)$$

where $\mathbf{y} = \text{vec}(\mathbf{Y})$ is a $NT \times 1$ vector of concatenated spike count observations from all N neurons and T time bins, $\mathbf{x} = \text{vec}(\mathbf{X})$ is a $PT \times 1$ vector of concatenated latent vectors across P latent time series, $\tilde{\mathbf{W}} = \mathbf{W} \otimes \mathbf{I}_T$ is a $NT \times PT$ Kronecker-structured matrix, and $\mathbf{1}$ is a length-NT vector of ones.

The only nonlinear term in the log-likelihood is the exponential term $\exp(\tilde{\mathbf{W}}\mathbf{x})$. We therefore approximate the exponential function with a second-order polynomial:

$$\exp(x) \approx ax^2 + bx + c, \quad (8)$$

with coefficients a, b, and c given by a Chebyshev polynomial approximation to $\exp(x)$ over an interval $\psi = [x_0, x_1]$, which we set independently for each neuron (Mason &

	binomial	Poisson	negative binomial
spike rate λ_{it}	$n\sigma(\mathbf{w}_i{}^\top\mathbf{x}_t)$	$\exp(\mathbf{w}_i{}^\top\mathbf{x}_t)$	$\exp(\mathbf{w}_i{}^\top\mathbf{x}_t)$
nonlinear term	$-\log(1+e^{-x})$	e^x	$\log(1+\alpha e^x)$
$\mathbf{H}$	$2n\tilde{\mathbf{W}}^\top \operatorname{diag}(\mathbf{a})\tilde{\mathbf{W}}$	$2\tilde{\mathbf{W}}^\top \operatorname{diag}(\mathbf{a})\tilde{\mathbf{W}}$	$2\tilde{\mathbf{W}}^\top \operatorname{diag}((\alpha^{-1}+\mathbf{y})\circ\mathbf{a})\tilde{\mathbf{W}}$
posterior mean $\boldsymbol{\mu}$	$\boldsymbol{\Sigma}\tilde{\mathbf{W}}^\top(\mathbf{y}-n-n\mathbf{b})$	$\boldsymbol{\Sigma}\tilde{\mathbf{W}}^\top(\mathbf{y}-\mathbf{b})$	$\boldsymbol{\Sigma}\tilde{\mathbf{W}}^\top(\mathbf{y}-\mathbf{y}\circ\mathbf{b}-\alpha^{-1}\mathbf{b})$

Table 1: Summary of PAL expressions for count-GPFA models. Top line gives the spike rate of neuron i at time t given the latent vector $\mathbf{x}_t$ and loading weights $\mathbf{w}_i$ for neuron i. Second line gives the nonlinear term of the log-likelihood that must be approximated under PAL. The third row, $\mathbf{H}$ is defined by $\mathbf{H} = \boldsymbol{\Sigma}^{-1} - \mathbf{K}^{-1}$, which succinctly presents posterior covariances, and the fourth line $\boldsymbol{\mu}$ shows approximate posterior means.

Handscomb, 2002; Zoltowski & Pillow, 2018). We use Chebyshev polynomials because they provide efficient near-minimax polynomial approximations (Huggins et al., 2017). Specifically, we computed the truncated Chebyshev expansion of the exponential $\exp(x) = \sum_{m=0}^{2} = \beta_m T_m$ where T_m is the degree-m Chebyshev polynomial of the first kind over $[x_0, x_1]$ and β_m are the expansion coefficients over that interval. The coefficients a, b, and c are given by collecting the terms to rewrite the expansion in the monomial basis.

We selected the interval $[x_0, x_1]$ independently for each neuron by computing the log of the mean firing rate of each neuron, $\log\lambda_i$. Since the nonlinearity is over the input $\mathbf{W}\mathbf{x}$, and the firing rate is $\lambda = \exp(\mathbf{W}\mathbf{x})$, we take the log of λ_i as we wish to center the nonlinear approximation at the center of the empirical neuronal rate to maximize accuracy. See Figure 1 as an example of a range centered at 0, corresponding to a simulated GP drawn with mean 0. We then chose the limits of the range to be $[\log\lambda_i - 2, \log\lambda_i + 2]$, resulting in an approximation range extending from e^{-2} to e^{2} times the mean firing rate. We found that this range balanced coverage in firing rate space with approximation accuracy. After selecting the range centers for each neuron, we computed the polynomial coefficients (a_i, b_i, c_i) for neuron i by gridding the interval of interest at a resolution of $dx = 0.01$ and solving for the coefficients that minimize the least squares between the true function and its polynomial approximation. For more detail, see (Zoltowski & Pillow, 2018).

Given coefficients for each neuron, the exponential term in the Poisson log-likelihood can be approximated:

$$\begin{aligned}&\mathbf{1}^\top \exp(\tilde{\mathbf{W}}\mathbf{x})\\ &\quad\approx \sum_{t=1}^{T}\sum_{i=1}^{N}\Big(a_i(\mathbf{W}\mathbf{x}_t)_i \circ (\mathbf{W}\mathbf{x}_t)_i + b_i(\mathbf{W}\mathbf{x}_t)_i + c_i\Big)\\ &\quad= \mathbf{x}^\top\tilde{\mathbf{W}}^\top \operatorname{diag}(\mathbf{a})\tilde{\mathbf{W}}\mathbf{x} + \mathbf{b}^\top\tilde{\mathbf{W}} + \mathit{const}, \qquad (9)\end{aligned}$$

where $\circ$ denotes Hadamard (element-wise) multiplication, and the second line involves the concatentation of the polynomial coefficients for each neuron and time bin: $\mathbf{a} = [a_1\mathbf{1}, \dots a_N\mathbf{1}]^\top$, $\mathbf{b} = [b_1\mathbf{1}, \dots b_N\mathbf{1}]^\top$, and we can ignore the constants c_i.

We now substitute the polynomial approximation into the log-likelihood and add the log prior, giving:

$$\begin{aligned}\mathcal{L}(\mathbf{y}, \mathbf{x}|\tilde{\mathbf{W}}, \theta) \approx&\\ &\mathbf{y}^\top\tilde{\mathbf{W}}\mathbf{x} - \mathbf{x}^\top\tilde{\mathbf{W}}^\top \operatorname{diag}(\mathbf{a})\tilde{\mathbf{W}}\mathbf{x}\\ &- \mathbf{b}^\top\tilde{\mathbf{W}}\mathbf{x} - \frac{1}{2}\mathbf{x}^\top\mathbf{K}^{-1}\mathbf{x} - \frac{1}{2}\log|\mathbf{K}|. \qquad (10)\end{aligned}$$

Since this approximation is quadratic in $\mathbf{x}$ we can exponentiate and then analytically marginalize $\mathbf{x}$ to obtain an approximation to the log-likelihood that follows equation (4) where:

$$\boldsymbol{\Sigma}^{-1} = 2\tilde{\mathbf{W}}^\top \operatorname{diag}(\mathbf{a})\tilde{\mathbf{W}} + \mathbf{K}^{-1} \qquad (11)$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\tilde{\mathbf{W}}^\top(\mathbf{y} - \mathbf{b}), \qquad (12)$$

and we have dropped terms that do not depend on $\tilde{\mathbf{W}}$ or θ.

3.2. PAL for Binomial-GPFA

Deriving the PAL estimator for a binomial observation model follows a similar logic to the Poisson case. Recall that for binomial model, spike count y is distributed according to :

$$P(y|p, n) = \binom{n}{y} p^y (1-p)^{(n-y)}. \qquad (13)$$

For this model, we map latents through a sigmoidal nonlinearity, $\sigma(x) = 1/(1 + \exp(-x))$, to obtain the binomial parameter p, and we set the number-of-trials parameter, n, to be the maximum number of observed spikes in a single time bin. The vector of spike rates at time t for this model is thus given by:

$$\boldsymbol{\lambda}_t = n\sigma(\mathbf{W}\mathbf{x}_t). \qquad (14)$$

We can write the log-likelihood in vectorized form as:

$$\begin{aligned}\mathcal{L}(\mathbf{y}|\mathbf{x}, \tilde{\mathbf{W}}) = (-n + \mathbf{y})\tilde{\mathbf{W}}\mathbf{x} -&\\ n\log(1 + \exp(-\tilde{\mathbf{W}}\mathbf{x})) + \mathit{const}&\end{aligned} \qquad (15)$$

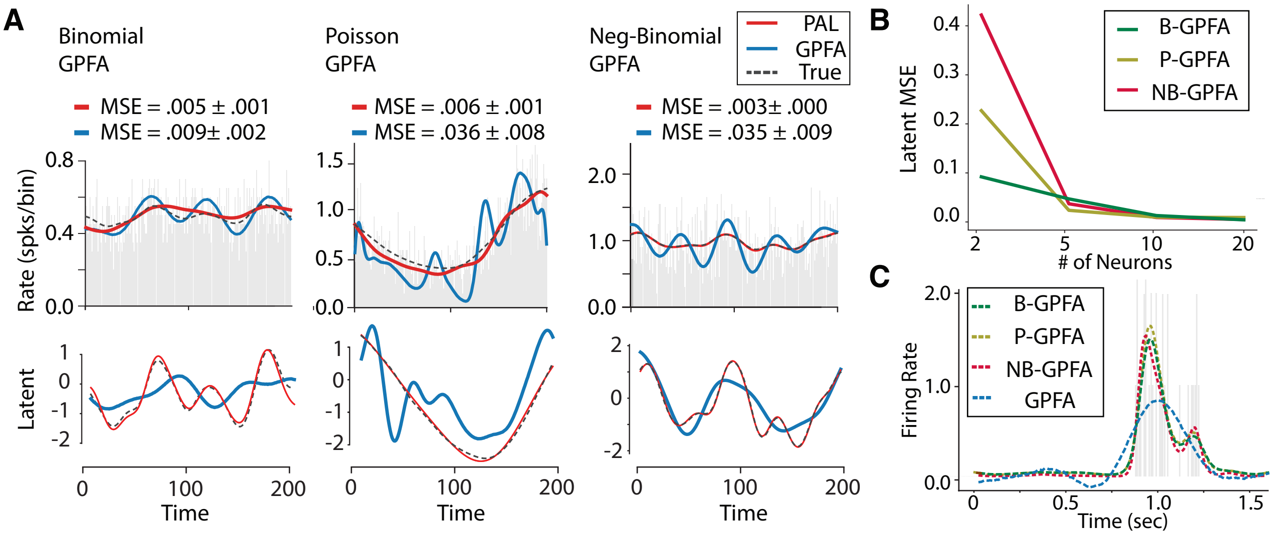

Figure 2: PAL inference for population data from binomial, Poisson, and negative binomial GPFA models. **A.** One example simulated neuron (out of 20) are shown for each model. Inferred rate for neurons for PAL inference compared to GPFA (Top). Latent trajectories recovered for each model (bottom). **B.** Error of recovered latent structure falls to zero with increasing numbers of observed neurons, as expected. **C.** Example PAL fits for all count-GPFA models compared to standard GPFA for spiking data from an example neuron. Light grey histogram denotes spike-count observations.

where we have ignored terms that do not depend on $\tilde{\mathbf{W}}\mathbf{x}$.

The problematic term here is the nonlinear second term, $\log(1+\exp(-x))$, which we approximate, as before, using a second-order Chebyshev polynomial approximation. In this case, we choose the center of the non-linearity to be the inverse sigmoid function of the empirical mean rate for each neuron $\sigma^{-1}(\lambda_i)$. We use a range of $[\sigma^{-1}(\lambda_i)-4, \sigma^{-1}(\lambda_i)+4]$ for the Chebyshev approximation. As in the Poisson case, we do this so the range for each neuron is centered at the mean empirical value of the input to the non-linearity, $\tilde{\mathbf{W}}\mathbf{x}$. The resulting approximation to the log-likelihood is:

$$\begin{aligned}\mathcal{L}(\mathbf{y}|\mathbf{x},\tilde{\mathbf{W}}) \approx& -n\mathbf{x}^\top\tilde{\mathbf{W}}^\top \operatorname{diag}(\mathbf{a})\tilde{\mathbf{W}}\mathbf{x}\\ &+ (\mathbf{y}-n-n\mathbf{b})^\top\tilde{\mathbf{W}}\mathbf{x} + const\end{aligned} \tag{16}$$

As in the Poisson case, we can add the log-prior to the above expression, exponentiate and marginalize over $\mathbf{x}$ to obtain an approximation to the log marginal likelihood in the same form as equation (4). In this case, we obtain matrix and vector terms:

$$\begin{aligned}\boldsymbol{\Sigma}^{-1} &= 2n\tilde{\mathbf{W}}^\top \operatorname{diag}(\mathbf{a})\tilde{\mathbf{W}} + \mathbf{K}^{-1}\\ \boldsymbol{\mu} &= \boldsymbol{\Sigma}\tilde{\mathbf{W}}^\top(\mathbf{y}-n-n\mathbf{b}).\end{aligned} \tag{17}$$

3.3. PAL for negative-binomial GPFA

Lastly, we consider a negative binomial observation model, which covers the over-dispersed spike responses (Goris et al., 2014; Linderman et al., 2016; Pillow & Scott, 2012). For negative-binomial GPFA, we parametrize the negative binomial distribution in terms of mean parameter m, and scale parameter $r = 1/\alpha$:

$$P(y|m,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y+1)}\Big(\frac{1}{1+\alpha m}\Big)^{\alpha^{-1}}\Big(\frac{\alpha m}{1+\alpha m}\Big)^{y} \tag{18}$$

This form of the distribution maps to the standard negative-binomial distribution, $p(y|p,r) = \binom{y+r-1}{y}(1-p)^r p^y$, via $p = \frac{r}{m+r}$. Parameterizing the negative binomial model this way makes for a simple expression of the expected spike count, which is equal to the model parameter m. Let us define this mean rate in the factor analytic framework as $m = \exp(\tilde{\mathbf{W}}\mathbf{x})$. This allows us to write the log-likelihood in vector form as:

$$\begin{aligned}\mathcal{L}(\mathbf{y}|\tilde{\mathbf{W}},&\mathbf{x},\alpha) = \mathbf{y}^\top\tilde{\mathbf{W}}\mathbf{x} -\\ &(\alpha^{-1}+\mathbf{y}^\top)\log(1+\alpha\exp(\tilde{\mathbf{W}}\mathbf{x})) + const.\end{aligned} \tag{19}$$

To derive a PAL estimator, we use a quadratic approximation to the nonlinear term $\log(1+\alpha\exp(x))$ on a per-neuron basis. We set $\alpha = 1$ for simulations, but this quantity may be learned in an outer loop. We choose the center of the nonlinear range to be the same as in the Poisson case, with the center value being the log of the mean firing rate of the neuron (see right panel of Figure 1 for example of centering with an average log-rate of 0). The range limits are $[\log\lambda_i - 4, \log\lambda_i + 4]$, where λ_i is the average value of m across time, per neuron. As in the previous cases, we obtain a quadratic approximate log-likelihood which has the

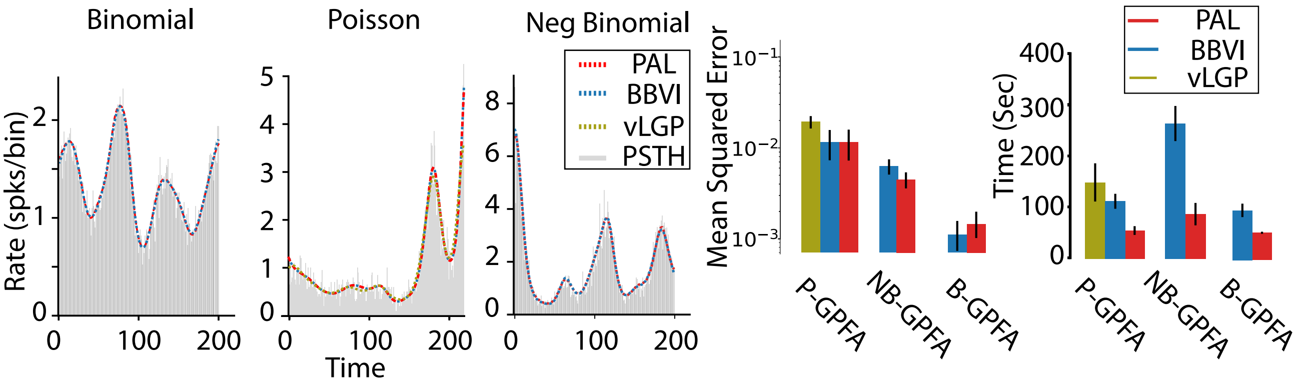

Figure 3: Comparison of vLGP, BBVI and PAL-based estimation in count-GPFA models. (**Left**) Reconstructed spike rates for each inference procedure, plotted alongside the spiking activity summed across trials (PSTH). Here, PAL-learned hyperparameters followed by a MAP estimation (red), BBVI (blue) and vLGP (yellow) all yield very similar and highly accurate spike rate reconstructions. Each of these closely match true spiking data (gray PSTH). (**Middle**) MSE of reconstructed rates shows good performance for all methods. (**Right**) PAL is faster than competing algorithms.

following form:

$$\mathcal{L}(\mathbf{y}|\mathbf{x}, \tilde{\mathbf{W}}, \alpha) \approx -\mathbf{x}\tilde{\mathbf{W}}^\top \operatorname{diag}((\alpha^{-1} + \mathbf{y}) \circ \mathbf{a})\tilde{\mathbf{W}}\mathbf{x} + (\mathbf{y} - \mathbf{y} \circ \mathbf{b} - \alpha^{-1}\mathbf{b})^\top \tilde{\mathbf{W}}\mathbf{x} + \mathit{const} \tag{20}$$

We then add the log prior and marginalize $\mathbf{x}$ to obtain an approximation to the log marginal likelihood for negative-binomial GPFA that follows the same form as equation (4) with

$$\mathbf{\Sigma}^{-1} = 2\tilde{\mathbf{W}}^\top \operatorname{diag}((\alpha^{-1} + \mathbf{y}) \circ \mathbf{a})\tilde{\mathbf{W}} + \mathbf{K}^{-1} \tag{21}$$

$$\mu = \mathbf{\Sigma}\tilde{\mathbf{W}}^\top (\mathbf{y} - \mathbf{y} \circ \mathbf{b} - \alpha^{-1}\mathbf{b}) \tag{22}$$

A summary of the features of all count GPFA models is given in Table 1. This table lists the nonlinear term for each model, the expected number of spikes for the ith neuron as a function of the latents, $\mathbf{X}$, loadings matrix $\mathbf{W}$, and the mean and covariance of the polynomial-approximated marginal distribution. We use n to refer to the maximal spike count observed in the data, and $\mathbf{w}_i$ to denote the ith column of $\mathbf{W}$.

3.4. Evaluating PAL performance

To assess the accuracy of the PAL estimator, we first analyzed its performance on simulated data. For 20 trials with 200 time points, we simulated count observations from 20 neurons with 2 latent processes with length scales $\ell_1 = 15$ and $\ell_2 = 60$ and each entry of $\mathbf{W}$ drawn uniformly in $[0, 2]$. We then fit each model by directly optimizing equation 4 to obtain parameter estimates $\hat{\mathbf{W}}$ and hyperparameter estimates $\hat{\ell}$. Conditioned on these estimates, we then maximized the conditional posterior to obtain $\hat{\mathbf{X}}_{MAP}$, the MAP estimate of the latent process. As a control, we compared PAL performance to standard Gaussian-noise GPFA. We found that the rates estimated using this procedure were similar to the true model rates and showed substantial improvement above Gaussian GPFA (Figure 2A, top). Additionally, PAL inference accurately captures latent structure (Figure 2A, bottom), whereas GPFA cannot. To identify latent structure in these simulated data, we regress learned latents onto the true latents as latent factors models are identifiable only up to a rotation matrix. Accurate identification of latent structure is a primary feature of this inference procedure, as latents have functional importance in neuroscience settings (Duncker & Sahani, 2018; Yu et al., 2009; Zhao & Park, 2017).

We additionally demonstrate PAL's accuracy by showing error of recovered latent structure as a function of the number of observed neurons. For each count-GPFA model, as we consider more and more data (from 2 to 20 neurons), PAL more and more accurately recovers latent structure, as expected (Figure 2B). Fits for real neural data are shown for an example neuron in Figure 2C. This is the first half of a trial for an example neuron from the mouse dataset (for more information, see section 5). PAL fits to count-GPFA better describe the neural spike-count data than standard GPFA. The background histogram in light grey in Figure 2C shows the true spike counts, and each of the dotted lines show the estimated neural firing rates under each GPFA model. Standard GPFA inference problematically yields negative rates and fails to capture quick changes in firing rate.

4. Comparison to other approaches

Variational inference (Blei et al., 2003) represents a common alternate approach to performing inference in non-conjugate factor models. This approach has been previously

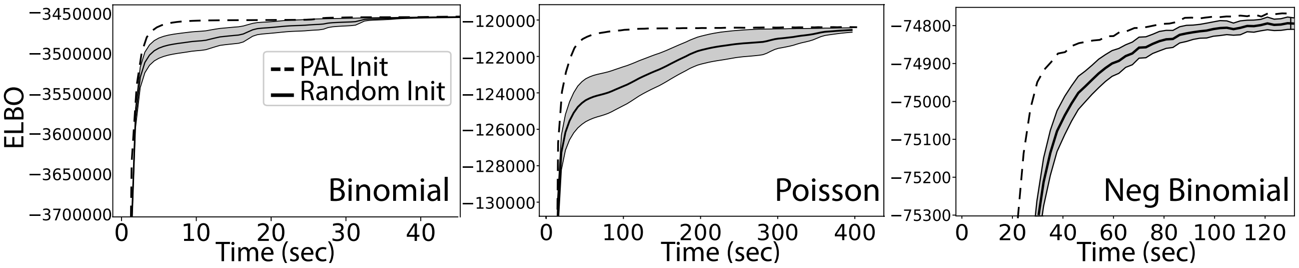

Figure 4: Optimization time for full BBVI from either a random initialization or PAL initialization for all count-GPFA models. Data is shown as mean and standard error for 10 trajectories.

used in the setting of Poisson-GPFA (Duncker & Sahani, 2018; Zhao & Park, 2017), and could in principal be used for the two other count-GPFA models we have introduced here. We therefore show comparisons of PAL to a variant of variational inference called black-box variational inference (BBVI), which uses Monte-Carlo samples to approximate the expectation term in the ELBO (Kingma et al., 2015). We additionally compare to an existing inference method available for Poisson GPFA called the variational Latent Gaussian Process (vLGP) (Zhao & Park, 2017).

On simulated data, BBVI, PAL and vLGP inference procedures achieve highly accurate reconstructions of true spike rates. The rate reconstructions of three example neurons for each model are shown on the left panel of Figure 3. Here, each model nearly perfectly predicts the simulated neural data. Average MSE across all neurons for these count GPFA models are shown in the middle panel of Figure 3. Though all inference methods achieve accurate results, times-to-convergence are faster and much more stable using the PAL approach (Figure 3, right panel). The time to converge is determined by the average times-to-convergence of ten runs of each optimization procedure. In the BBVI case, convergence was determined when the ELBO was within 99.8% of the maximal ELBO value identified. For occasional BBVI runs for each count model, this value was not achieved for the duration of the inference procedure, as the the ELBO was stuck at a local maxima. These convergence times were discarded when calculating the mean convergence time, and demonstrative of the irregularity of the BBVI inference procedure. For vLGP, we set the number of maximum iterations to 50, and the minimum to 10. The algorithm typically did not converge before the maximum number of iterations was up. We also note a slight accuracy improvement of PAL compared to vLGP. However, it is important to note that vLGP assigns a per-trial latent whereas our algorithm assumes the latent is same across all trials. This is possibly a confounding factor when comparing performance time of vLGP and PAL, both of which achieve high accuracy.

4.1. PAL initialization for BBVI

The PAL method can additionally be used in conjunction with BBVI to speed up and improve inference. Because the PAL method involves approximating a nonlinearity in a specified range with a quadratic, if the input ($\mathbf{Wx}$) is not within the range of the Chebyshev approximation the estimate will be inaccurate. Moreover, there is a significant limitation using BBVI. In particular, the use of sampling in the optimization poses considerable challenges in convergence. In fact, this is a well-known problem, and a variety of techniques have been introduced to reduce variance in the gradient estimates (Hoffman et al., 2013; Roeder et al., 2017; Salimbeni et al., 2018). We offer an alternative solution; we can overcome the limitations of both BBVI and PAL by combining them.

Initializing the BBVI algorithm with the hyperparameters provided by PAL optimization of equation 4 allows for a rapid and stable BBVI. That is, instead of following up our PAL hyper-parameter identification with a MAP estimation of the latents, we use it to seed an approximate accurate hyperparameters to BBVI. This procedure is more stable than full BBVI with random initial hyperparameters. We demonstrate this for all count-GPFA models. Figure 4 shows the evolution of the ELBO in time during optimization for all models on simulated data (though the same effect is observed in real data). In each case, BBVI is run 10 times, either initializing randomly or initializing at the PAL-optimal hyperparameters (ℓ and $\mathbf{W}$). Standard error is shown in grey for the random initialization, but not shown for PAL-initialized optimization, as this trajectory follows nearly identically for each run. An initial sharp increase in the ELBO is always observed in all models, as here latent structure is approximately identified, but hyperparameters are tuned at the end of the BBVI optimization procedure. Here, we have cut off the initial rise in ELBO for clarity. Figure 4 demonstrates the end of the optimization procedure, where randomly initialized BBVI attempts to find hyperparameters along varying trajectories, but PAL initialized BBVI quickly converges to high ELBO values.

Thus, initializing BBVI with PAL hyperparameter estimates avoids local optima. This PAL initialization procedure can therefore be considered alongside other methods for providing a way to stabilize and improve BBVI.

5. Applications to neural data

To examine the performance of these methods on real data, we applied the binomial, Poisson, and negative binomial count-GPFA models to neural data sets from two different species. We compared the three approaches outlined: PAL followed by a MAP estimate (section 3), BBVI, and PAL-initialized BBVI (section 4). We tested these models on one data set from primate parietal and high-level visual cortices, and the other from mouse V1. The first dataset consisted of 14 simultaneously recorded neurons from the middle temporal visual (MT) and lateral intraparietal (LIP) area. These data were 50 1.4-second trials of a visual perceptual decision-making task (Yates et al., 2017). In this task, random moving dots provided visual evidence towards left or right targets (choices). The trial contained a stimulus onset time, an evidence accumulation (decision making) period, and a decision. For the mouse data, spike times from 17 V1 neurons were recorded during passive viewing of 20 repeated 32-second trials of a gratings stimulus. The stimulus was a random flashing of gratings, with 8 orientations at fixed spatial and temporal frequencies. The gratings were presented for 4 seconds each. The spike times were determind by de-convolving calcium imaging traces. Additional details of the data can be found in (Yu et al., 2018).

For both the mouse and the primate data we asserted a latent dimensionality of three for all count-GPFA models. The PAL method demonstrated good empirical fits (see example mouse neuron in Figure 2C) and in general good cross validation performance compared to BBVI (Figure 5A, B). However, for the mouse data, PAL performance was notably weaker under a Poisson-GPFA model (Figure 5A). This was likely because these neurons were high-variance, exhibiting no activity for much of the trial, with some abrupt changes in firing throughout the trial. In this case, the exponential non-linearity is not accurately captured by the polynomial approximation. As seen in Figure 1, the exponential non-linearity for Poisson-GPFA is approximated least accurately compared to the approximations for the other two models. Thus, under a Poisson-GPFA model, using PAL in conjunction with BBVI performed notably better for the mouse data. For the other two count models (binomial and negative-binomial) we found that the PAL estimation procedure was approximately as accurate as BBVI. Interestingly, for the mouse data, the count-GPFA model that had highest cross-validated log-likelihood was Binomial-GPFA, which is a count-model not often considered in neuroscience settings.

For the primate data, all GPFA models and inference methods exhibited equal performance (Figure 5B), with a small bias favoring PAL-initialized negative binomial GPFA. The lack of major differences in performance here is likely because these data were high-spike rates neurons with many trials. This resulted in low-variance, stable rates that could be accurately captured by the PAL non-linearity and quickly and easily recovered using BBVI.

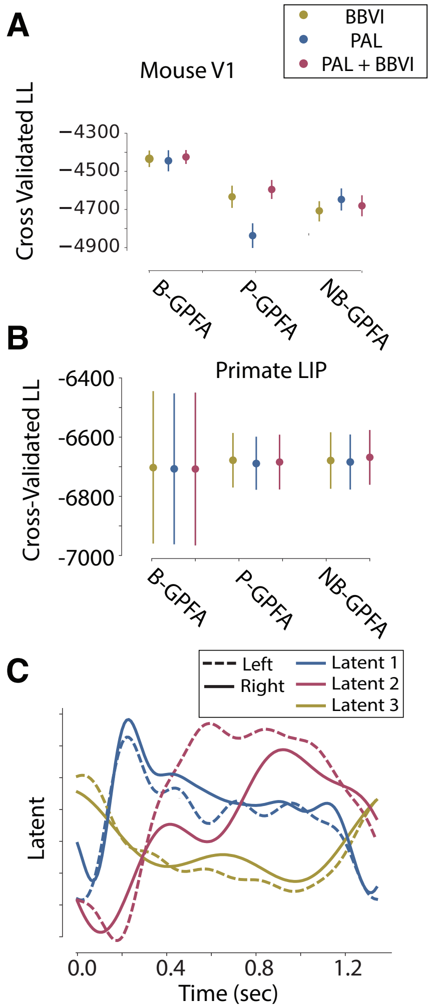

Figure 5: Average cross validated log-likelihood on hold-out trials for PAL, BBVI and PAL + BBVI inference methods for count-GPFA models for mouse (**A**) and primate (**B**) data. (**C**) Latent structure underlying monkey LIP responses during left- and right-choice trials, showing that one latent dimension captured meaningful differences between the encoding of left and right choices.

To give insight into the scientific uses of this model, we show results of the Binomial-GPFA model fit to monkey LIP data. We fit a Binomial-GPFA model with 3 latent dimensions to two different subsets of the data: one consisting of the leftward choice trials and another consisting of rightward choice trials. We then compared the latents inferred for each condition in order to examine how the latent variables encode the animal's choice. Figure 5C shows the latent structure of the neural data for each condition. Two of the latents are closely overlapping, which suggests the presence of shared structure across the two conditions. Interestingly, one latent (red) diverges approximately 400 ms after trial onset, which falls into the portion of the trial where the animal is putatively making its choice. This suggests that this latent may encode the choice variable in these neural data, and is a promising future direction of further exploration for count-GPFA models.

6. Conclusion

We have a developed novel technique for learning Gaussian process factor analytic models with count observations using polynomial approximate log-likelihood (PAL) which allows for rapid closed-form evaluation of marginal likelihoods. We develop our PAL approach for three count-observation models: binomial, Poisson, and negative-binomial. In each case, our approximation can accurately estimate model parameters and achieve good performance on both simulated and real neural data. PAL can additionally provide initial values for black box variational inference. Both PAL and BBVI have their own limitations – PAL provides an approximation to the model non-linearity that is only accurate within a particular range, and BBVI inference procedure is sampling-based and can get stuck in local optima. Combining the procedures by using the PAL method to identify approximate hyperparameters can thus stabilize BBVI and make it more reliable, overcoming well-known BBVI optimization limitations. Overall, our PAL inference method is a novel approach to learning non-conjugate models that is fast and achieves high accuracy.

Acknowledgements

SLK was supported by NIH grant F32MH115445-03, DMZ was supported by NIH grant T32MH065214 and JWP was supported by grants from the Simons Collaboration on the Global Brain (SCGB AWD543027), the NIH BRAIN initiative (NS104899 and R01EB026946), and a U19 NIH-NINDS BRAIN Initiative Award (5U19NS104648). YY and SLS were supported by the NIH (R01EY024294 and R01NS091335), the NSF (1450824 and 1707287) the Simons Foundation (SCGB 325407) and the McKnight Foundation. The authors thank Adam Charles, Mikio Aoi, Nicholas Roy, and the anonymous reviewers for providing helpful comments.

References

Archer, E., Park, I. M., Buesing, L., Cunningham, J., and Paninski, L. Black box variational inference for state space models. *stat*, 1050:23, 2015.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Buesing, L., Macke, J. H., and Sahani, M. Spectral learning of linear dynamics from generalised-linear observations with application to neural population data. In *Adv neur inf proc sys*, pp. 1682–1690, 2012.

Charles, A. S., Park, M., Weller, J. P., Horwitz, G. D., and Pillow, J. W. Dethroning the fano factor: A flexible, model-based approach to partitioning neural variability. *Neural computation*, 30(4):1012–1045, 2018.

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

Duncker, L. and Sahani, M. Temporal alignment and latent gaussian process factor inference in population spike trains. *bioRxiv*, pp. 331751, 2018.

Gao, Y., Busing, L., Shenoy, K. V., and Cunningham, J. P. High-dimensional neural spike train analysis with generalized count linear dynamical systems. In *Advances in neural information processing systems*, pp. 2044–2052, 2015.

Goris, R., Movshon, J., and Simoncelli, E. Partitioning neuronal variability. *Nature neuroscience*, 17(6):858–865, 2014.

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

Huggins, J., Adams, R. P., and Broderick, T. Pass-glm: polynomial approximate sufficient statistics for scalable bayesian glm inference. In *Advances in Neural Information Processing Systems*, pp. 3614–3624, 2017.

Kara, P., Reinagel, P., and Reid, R. C. Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. *Neuron*, 27:636–646, 2000.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pp. 2575–2583, 2015.

Lakshmanan, K., Sadtler, P., Tyler-Kabara, E., Batista, A., and Yu, B. Extracting low-dimensional latent structure from time series in the presence of delays. *Neural computation*, 2015.

Linderman, S., Adams, R. P., and Pillow, J. W. Bayesian latent structure discovery from multi-neuron recordings. In *Advances in neural information processing systems*, pp. 2002–2010, 2016.

Macke, J. H., Buesing, L., Cunningham, J. P., Byron, M. Y., Shenoy, K. V., and Sahani, M. Empirical models of spiking in neural populations. In *Advances in neural information processing systems*, pp. 1350–1358, 2011.

Maimon, G. and Assad, J. A. Beyond poisson: increased spike-time regularity across primate parietal cortex. *Neuron*, 62(3):426–440, May 2009. doi: 10.1016/j.neuron.2009.03.021. URL http://dx.doi.org/10.1016/j.neuron.2009.03.021.

Mason, J. C. and Handscomb, D. C. *Chebyshev polynomials*. CRC Press, 2002.

Pillow, J. and Scott, J. Fully bayesian inference for neural models with negative-binomial spiking. In Bartlett, P., Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1907–1915, 2012. URL http://books.nips.cc/papers/files/nips25/NIPS2012_0942.pdf.

Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *Artificial Intelligence and Statistics*, pp. 814–822, 2014.

Roeder, G., Wu, Y., and Duvenaud, D. Sticking the landing: An asymptotically zero-variance gradient estimator for variational inference. *Advances in Neural Information Processing Systems*, 2017.

Salimbeni, H., Eleftheriadis, S., and Hensman, J. Natural gradients in practice: Non-conjugate variational inference in gaussian process models. *arXiv preprint arXiv:1803.09151*, 2018.

Wu, A., Roy, N. G., Keeley, S., and Pillow, J. W. Gaussian process based nonlinear latent structure discovery in multivariate spike train data. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3499–3508. Curran Associates, Inc., 2017.

Yates, J. L., Park, I. M., Katz, L. N., Pillow, J. W., and Huk, A. C. Functional dissection of signal and noise in mt and lip during decision-making. *Nature neuroscience*, 20(9):1285, 2017.

Yu, B., Cunningham, J., Santhanam, G., Ryu, S., Shenoy, K., and Sahani, M. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Adv neur inf proc sys*, pp. 1881–1888, 2009.

Yu, Y., Stirman, J. N., Dorsett, C. R., and Smith, S. L. Mesoscale correlation structure with single cell resolution during visual coding. *bioRxiv*, pp. 469114, 2018.

Zhao, Y. and Park, I. M. Variational latent gaussian process for recovering single-trial dynamics from population spike trains. *Neural computation*, 29(5):1293–1316, 2017.

Zhao, Y., Yates, J. L., Levi, A. J., Huk, A. C., and Park, I. M. Stimulus-choice (mis) alignment in primate mt cortex. *bioRxiv*, 2019.

Zoltowski, D. M. and Pillow, J. W. Scaling the poisson glm to massive neural datasets through polynomial approximations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 3521–3531. Curran Associates, Inc., 2018.