Variational bayesian inference for point process generalized linear models in neural spike trains analysis

Variational Bayesian Inference for Point Process Generalized Linear Models in Neural Spike Trains Analysis

Zhe Chen, Fabian Kloosterman, Matthew A. Wilson, and Emery N. Brown

***Abstract*— Point process generalized linear models (GLMs) have been widely used for neural spike trains analysis. Statistical inference for GLMs include maximum likelihood and Bayesian estimation. Variational Bayesian (VB) methods provide a computationally appealing means to infer the posterior density of unknown parameters, in which conjugate priors are designed for the regression coefficients in logistic and Poisson regression. In this paper, we develop and apply VB inference for point process GLMs in neural spike train analysis. The hierarchical Bayesian framework allows us to tackle the variable selection problem. We assess and validate our methods with ensemble neuronal recordings from rat's hippocampal place cells and entorhinal cortical cells during foraging in an open field environment.**

***Index Terms*— point process, generalized linear model, conjugate prior, logistic regression, Poisson regression, variational Bayes.**

I. Introduction

Point process generalized linear models (GLMs) have recently been widely used for neural spike train analysis [21], [13], [14]. Statistical inference procedures, either maximum likelihood or Bayesian approaches, have been developed for neural encoding and inferring functional connectivity [1], [13], [16], [19], [5]. In maximum likelihood estimation, parameters are treated as deterministic variables, and only point estimates are produced, with uncertainties represented by standard error or bootstrapped variance. Whereas in Bayesian estimation, parameters are viewed as random variables, which are associated with their posterior probability densities. The advantages of the Bayesian estimate over the maximum likelihood estimate (m.l.e.) are its ease of incorporating priors and its full characterization of the posterior [7], [9]. A hierarchical Bayesian model also enables us to model the uncertainty of the hyperparameters, such that the final performance is robust to the priors. In addition, maximum likelihood estimate has the tendency of overfitting the data using a large set of parameters (since the likelihood function is not bounded), and the model selection is typically achieved by Akaike's information criterion (AIC) or tedious cross-validation. Penalized maximum likelihood estimation attempts to improve this issue [5], but it still does not admit *automatic* variable selection.

Three schools of Bayesian inference algorithms are popular in the statistics/machine learning fields: i) Laplace approximation; ii) Monte Carlo Markov chain (MCMC); iii) variational Bayes (VB). The VB inference is particularly appealing because of its improved performance over the Laplace approximation and its smaller computational burden than the MCMC methods [2]. We extend the variational logistic regression model [10], [2] with hierarchical Bayesian modeling and develop a new VB Poisson regression model using a generalized conjugate prior. We use two VB-inference algorithms for point process GLM in neural spike train analysis for neural encoding and functional connectivity analysis. We demonstrate the effectiveness of our method with ensemble neuronal recordings from a foraging rat within a two-dimensonal open circular environment.

Support from NIH Grants DP1-OD003646 and R01-DA015644.

The authors are with the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Z. Chen and E. N. Brown are also with the Neuroscience Statistics Research Laboratory, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA. (Email: zhechen@mit.edu)

TABLE I
EXAMPLES OF EXPONENTIAL FAMILY IN A CANONICAL FORM.

prob. dist.	link func.	θ	$b(\theta)$	$\dot{b}(\theta)$	$\ddot{b}(\theta)$
Bernoulli$(1,\pi)$	logit	$\log\frac{\pi}{1-\pi}$	$\log(1+e^{\theta})$	π	$1-\pi$
Poisson(λ)	log	$\log\lambda$	$\exp(\theta)$	λ	λ

II. Exponential Family and Generalized Linear Models

In the framework of generalized linear model (GLM) [11], we assume that the observations $\{y_{1:T}\}$ follow an exponential family distribution with the form:

$$p(y_t|\theta_t) = \exp(y_t\theta_t - b(\theta_t) + c(y_t)), \tag{1}$$

where θ denotes the canonical parameter, and $c(y_t)$ is a normalizing constant. Suppose $\theta_t = g(\eta_t) = g(\boldsymbol{\beta}\boldsymbol{x}_t)$ ($\boldsymbol{\beta} \in \mathbb{R}^d$), where g is called the *link function*. Assume $b(\theta_t)$ is twice differentiable, then $\mathbb{E}_{y|\theta}[y_t] = \dot{b}(\theta_t) = \frac{\partial b(\theta_t)}{\partial \theta_t}$, $\mathrm{Var}[y_t] = \ddot{b}(\theta_t) = \frac{\partial^2 b(\theta_t)}{\partial \theta_t \partial \theta_t'}$ ($'$ denotes the transpose). Using a canonical link function, the natural parameter relates to the linear predictor by $\theta_t = \eta_t = \boldsymbol{\beta}\boldsymbol{x}_t$. Table I lists two probability distributions of exponential family (in a canonical form) for modeling discrete data.

Two popular GLMs used in modeling discrete data for neural spike trains are logistic regression and Poisson regression. In logistic regression, spike train observations 0 and 1 are treated as independent Bernoulli random variables, whereas in Poisson regression, spike counts follows an inhomogeneous Poisson distribution. The Poisson distribution is an approximation of the Binomial(n,π) distribution if π is sufficiently small, and $n\pi$ remains roughly a constant when $n \to \infty$. When the discretization bin Δ is sufficiently small, we can approximate $\pi = \lambda\Delta$.

A point process is a stochastic process with 0 and 1 observations. Let $\boldsymbol{x}_t$ denote the input covariate at time t, y_t denote the observed response variable, which equals to 1 if there is an event (spike) at time t and 0 otherwise. Denote $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$ and $\boldsymbol{y} = [y_1, \ldots, y_T]$. Within the point process GLM framework, we can write down the log-likelihood function for logistic (Bernoulli) and Poisson regression, respectively, as follows [21]:

$$\mathcal{L}_{Be} = \sum_{t=1}^{T} \Big[y_t \log \pi_t(\boldsymbol{\beta}) + (1-y_t)\log(1-\pi_t(\boldsymbol{\beta})) \Big] \tag{2}$$

$$\mathcal{L}_{Po} = \sum_{t=1}^{T} \Big[y_t \log(\lambda_t(\boldsymbol{\beta})) - \lambda_t(\boldsymbol{\beta}) \Big] \tag{3}$$

where $\pi_t(\boldsymbol{\beta}) = \sigma(\boldsymbol{\beta}\boldsymbol{x}_t)$ ($\sigma(\cdot)$ is a logistic sigmoid function) and $\lambda_t(\boldsymbol{\beta}) = \exp(\boldsymbol{\beta}\boldsymbol{x}_t)$. Maximizing (2) and (3) with respect to $\boldsymbol{\beta}$ yields the m.l.e. Statistical inference algorithms for m.l.e. include

the expectation-maximization, iteratively reweighted least squares, and conjugate gradient algorithms.

III. Bayesian Inference and Variational Bayes Method

The goal of Bayesian inference is to estimate the parameter posterior $p(\boldsymbol{\beta}|\boldsymbol{y})$ given a specific parameter prior $p(\boldsymbol{\beta})$. Normally, because the posterior is analytically non-trackable, we will need to resort to strategies for approximation. These methods include the Laplace approximation for log-posterior [3], [9], [2], expectation propagation (EP) for moment matching [17], and MCMC sampling [16], [9]. However, the Laplace and EP approximations are less accurate (esp. when the posterior has multiple modes or the mode is not near the majority of the probability mass); the MCMC methods are more general but require a high demand of computing power and experience difficulties of assessing the convergence of Markov chains. As an alternative Bayesian inference procedure, VB methods attempt to maximize the lower bound of the marginal likelihood (a.k.a. *evidence*) or the marginal log-likelihood [2] :

$$\begin{aligned}\log p(\boldsymbol{y}|\boldsymbol{x}) &= \log \int\int p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\beta})p(\boldsymbol{\beta}|\alpha)p(\alpha)d\boldsymbol{\beta}d\alpha \\ &\geq \int\int q(\boldsymbol{\beta},\alpha)\log\frac{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\beta})p(\boldsymbol{\beta}|\alpha)p(\alpha)}{q(\boldsymbol{\beta},\alpha)}d\boldsymbol{\beta}d\alpha \qquad (4)\end{aligned}$$

where $p(\boldsymbol{\beta}|\alpha)$ denotes the prior distribution of $\boldsymbol{\beta}$, specified by the hyperparameter α. The variational distribution $q(\boldsymbol{\beta},\alpha) = q(\boldsymbol{\beta})q(\alpha)$ has a factorial form, which attempts to approximate the posterior $p(\boldsymbol{\beta},\alpha|\boldsymbol{y})$. This approximation leads to an analytical posterior form if the distributions are conjugate-exponential.

An VB inference algorithm for logistic regression has been developed in the field of machine learning [10], [2]. The basic idea is to derive a variational lower bound (using a data-dependent variational parameter $\boldsymbol{\xi} = \{\xi_t\}$) for the marginal log-likelihood function (2). However, the hyperparameters therein are fixed a priori, so their model is empirical Bayesian. Here, we extend the model with hierarchical Bayesian modeling using *automatic relevance determination* (ARD) [12] for the purpose of variable selection. Such a fully Bayesian inference allows us to design a separate prior for each element β_i in vector $\boldsymbol{\beta}$ and to set a conjugate prior $p(\boldsymbol{\alpha})$ for the hyperparameters using a common gamma hyperprior. Our prior distributions are set up as follows:

$$p(\boldsymbol{\beta}|\boldsymbol{\alpha}) \sim \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\mu}_0, \boldsymbol{A}^{-1}) \propto \exp\Big(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)'\boldsymbol{A}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\Big),$$
$$p(\boldsymbol{\alpha}) = \prod_{i=1}^{d}\text{Gamma}(\alpha_i|a_0,b_0)$$

where $\boldsymbol{A} = \text{diag}\{\boldsymbol{\alpha}\} \equiv \text{diag}\{\alpha_1,\ldots,\alpha_d\}$ (a non-ARD formulation is equivalent to setting $\boldsymbol{A} = \alpha\mathbf{I}$ as a special case). Here, we assume that the mean hyperparameter is fixed (e.g., $\boldsymbol{\mu}_0 = \mathbf{0}$ or $\boldsymbol{\mu}_0 = \boldsymbol{\beta}_{\text{m.l.e.}}$). Applying the VB inference yields the variational posteriors

$$q(\boldsymbol{\beta}|\boldsymbol{y}) = \tilde{p}(\boldsymbol{\beta},\boldsymbol{\xi})\mathbb{E}_{q(\boldsymbol{\alpha})}[p(\boldsymbol{\beta}|\boldsymbol{\alpha})] = \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\mu}_T,\boldsymbol{\Sigma}_T) \qquad (5)$$

$$q(\boldsymbol{\alpha}|\boldsymbol{y}) = \mathbb{E}_{q(\boldsymbol{\beta})}[p(\boldsymbol{\beta}|\boldsymbol{\alpha})]p(\boldsymbol{\alpha}) = \prod_{i=1}^{d}\text{Gamma}(\alpha_i|a_T,b_{i,T}) \qquad (6)$$

where $\tilde{p}(\boldsymbol{\beta},\boldsymbol{\xi})$ denotes the variational likelihood bound for logistic regression, $\boldsymbol{\Sigma}_T^{-1} = \mathbb{E}_{q(\boldsymbol{\alpha})}[\boldsymbol{A}] + 2\sum_{t=1}^{T}\phi(\xi_t)\boldsymbol{x}_t\boldsymbol{x}_t'$, $\boldsymbol{\mu}_T = \boldsymbol{\Sigma}_T\big(\mathbb{E}_{q(\boldsymbol{\alpha})}[\boldsymbol{A}]\boldsymbol{\mu}_0 + \sum_{t=1}^{T}(y_t-0.5)\boldsymbol{x}_t\big)$, $\phi(\xi) = \tanh(\xi/2)/(4\xi)$, $\xi_t^2 = \boldsymbol{x}_t'(\boldsymbol{\Sigma}_T + \boldsymbol{\mu}_T\boldsymbol{\mu}_T')\boldsymbol{x}_t$, $\mathbb{E}_{q(\boldsymbol{\alpha})}[\boldsymbol{A}] = \text{diag}\{a_T/b_{i,T}\}$, $a_T = a_0 + 0.5$, $b_{i,T} = b_0 + 0.5[(\boldsymbol{\mu}_T)_i^2 + (\boldsymbol{\Sigma}_T)_{ii}]$. Finally, we can derive the variational lower bound of marginal $\mathcal{L}_{Be}$:

$$\begin{aligned}\tilde{\mathcal{L}}_{Be} &= \frac{1}{2}\Big\{\boldsymbol{\mu}_T'\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\mu}_T + \log|\boldsymbol{\Sigma}_T| + \sum_{t=1}^{T}\Big(2\log\sigma(\xi_t) - \xi_t \\ &\quad + 2\phi(\xi_t)\xi_t^2\Big)\Big\} + \sum_{i=1}^{d}\Big\{-\log\Gamma(a_0) + a_0\log b_0 \\ &\quad - b_0\frac{a_T}{b_{i,T}} - a_T\log b_{i,T} + \log\Gamma(a_T) + a_T\Big\} \qquad (7)\end{aligned}$$

The VB inference alternatingly updates (5) and (6) to monotonically increase $\tilde{\mathcal{L}}_{Be}$. The criterion of algorithmic convergence is set until the consecutive change of (7) is sufficiently small.

In Poisson regression, a Laplace approximation of (3) would yields a Gaussian likelihood with both mean and variance equal to $\lambda_t = \boldsymbol{\beta}\boldsymbol{x}_t$ [9], [11], but this approximation is poor for small values of λ_t (which is typically the case for neuronal data). Note that unlike logistic regression, it is difficult to derive a *tight* variational lower bound of marginal $\mathcal{L}_{Po}$ (because of its likelihood form). Here we adapt a generalized conjugate prior as formulated in [4], [8]

$$\begin{aligned}p(\boldsymbol{\beta}|\alpha_0,\boldsymbol{y}_0) &= h(\alpha_0,\boldsymbol{y}_0)\exp\{\alpha_0[\boldsymbol{y}_0'\boldsymbol{\theta} - \mathbf{1}'b(\boldsymbol{\theta})]\} \\ &\propto \exp\{\alpha_0[\boldsymbol{y}_0'\boldsymbol{\theta}(\boldsymbol{\eta}) - \mathbf{1}'b(\boldsymbol{\theta}(\boldsymbol{\eta}))]\} \\ &\equiv \exp\{\alpha_0[\boldsymbol{y}_0'\boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\beta}) - \mathbf{1}'b(\boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\beta}))]\} \qquad (8)\end{aligned}$$

where $\mathbf{1}$ is an all-ones vector, $\alpha_0 > 0$ is a scalar prior parameter, $\boldsymbol{y}_0 = (y_{0,1},\ldots,y_{0,T})$ is a vector of prior parameters as *pseudo-observations*, and $h(\alpha_0,\boldsymbol{y}_0)$ is a normalizing term such that $\int p(\boldsymbol{\beta}|\alpha_0,\boldsymbol{y}_0)d\boldsymbol{\beta} = 1$. In the case of Poisson GLM with a canonical log link function, $\boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$, $b(\boldsymbol{\theta}(\boldsymbol{X}\boldsymbol{\beta})) = \exp(\boldsymbol{X}\boldsymbol{\beta}) = [\lambda_1,\ldots,\lambda_T]$. The role of a_0 is to control the heaviness of the prior distribution: when $\alpha_0 = 0$, (8) reduces a uniform improper prior for $\boldsymbol{\beta}$; when a_0 gets large, (8) becomes more informative; when $\alpha_0 \to \infty$, the prior reduces to a point mass at its mode. The role of $\boldsymbol{y}_0$ is to represent a prior prediction for the marginal mean $\mathbb{E}[\boldsymbol{y}] = \mathbb{E}_{p(\boldsymbol{\beta})}[b(\boldsymbol{\theta})]$. The parameter α_0 can be viewed as a *precision parameter* that quantifies the strength of belief in $\boldsymbol{y}_0$. We denote the generalized conjugate prior in (8) by $\{\boldsymbol{\beta}|\alpha_0,\boldsymbol{y}_0\} \sim D(\alpha_0,\boldsymbol{y}_0)$. As shown in [4], as the number of samples $T \to \infty$, (8) approaches to a Normal distribution, where the mode coincides with the m.l.e. from $p(\boldsymbol{y}_0|\boldsymbol{\beta})$. The elicitation of $\boldsymbol{y}_0$ and α_0 is discussed in [4].

Combining the likelihood (1) and the generalized conjugate prior (8), we obtain the posterior of $\boldsymbol{\beta}$ in a similar form of prior [4]:

$$\{\boldsymbol{\beta}|\alpha_0,\boldsymbol{y}_0,\boldsymbol{y}\} \sim D\Big(\alpha_0+1, \frac{\alpha_0\boldsymbol{y}_0+\boldsymbol{y}}{\alpha_0+1}\Big). \qquad (9)$$

To conduct hierarchical Bayesian inference, we also set a gamma hyperprior for α_0: $p(\alpha_0) = \text{Gamma}(\alpha_0|a_0,b_0)$, where (a_0,b_0) are the pre-specified hyperprior parameters. Then the joint prior probability distribution is written as [4]:

$$\begin{aligned}p(\boldsymbol{\beta},\alpha_0|\boldsymbol{y}_0) &= p(\boldsymbol{\beta}|\alpha_0,\boldsymbol{y}_0)p(\alpha_0) \qquad (10)\\ &\propto \exp\{\alpha_0[\boldsymbol{y}_0\boldsymbol{\theta} - \mathbf{1}'b(\boldsymbol{\theta})\}\alpha_0^{a_0-1}\exp(-b_0\alpha_0)\end{aligned}$$

Note that $\alpha_0 = 0$ would yield a posterior of $\boldsymbol{\beta}$ close to the m.l.e. To perform exact Bayesian inference, we need to integrate over α_0 to obtain the posterior of $\boldsymbol{\beta}$:

$$\begin{aligned}p(\boldsymbol{\beta}|\boldsymbol{y}_0,\boldsymbol{y}) &= \int p(\boldsymbol{\beta}|\alpha_0,\boldsymbol{y}_0,\boldsymbol{y})p(\alpha_0)d\alpha_0 \\ &\sim \int D\Big(\alpha_0+1,\frac{\alpha_0\boldsymbol{y}_0+\boldsymbol{y}}{\alpha_0+1}\Big)\text{Gamma}(\alpha_0|a_0,b_0)d\alpha_0 \qquad (11)\end{aligned}$$

Since the integration in (11) has no an analytical form, we can use either VB approach or Gaussian approximation. Below we briefly describe these two solutions.

Using the variational approximation $q(\boldsymbol{\beta}, \alpha_0) = q(\boldsymbol{\beta})q(\alpha_0)$ (note that the data likelihood (1) is independent of α_0), we can derive the variational log-posteriors for both $\boldsymbol{\beta}$ and α_0:

$$\begin{aligned}\log q(\boldsymbol{\beta}|\boldsymbol{y}) &= \log p(\boldsymbol{y}|\boldsymbol{\beta}) + \mathbb{E}_{q(\alpha_0)}[\log p(\boldsymbol{\beta}|\alpha_0)] + \text{const.} \\ &= \log D\left(\mathbb{E}_{q(\alpha_0)}[\alpha_0] + 1, \frac{\mathbb{E}_{q(\alpha_0)}[\alpha_0]\boldsymbol{y}_0 + \boldsymbol{y}}{\mathbb{E}_{q(\alpha_0)}[\alpha_0] + 1}\right) \quad (12) \\ \log q(\alpha_0|\boldsymbol{y}) &= \mathbb{E}_{q(\boldsymbol{\beta})}[\log p(\boldsymbol{\beta}|\alpha_0)] + \log p(\alpha_0) + \text{const.} \\ &= \log \text{Gamma}(\alpha_0|a_T, b_T) \quad (13)\end{aligned}$$

where $\mathbb{E}_{q(\alpha_0)}[\alpha_0] = a_T/b_T$, $a_T = a_0 + d/2$, $b_T = \boldsymbol{y}_0'\boldsymbol{X}\langle\boldsymbol{\beta}\rangle - \mathbf{1}'\langle \exp(\boldsymbol{X}\boldsymbol{\beta})\rangle - b_0$ (the expectation $\langle\cdot\rangle$ is taken w.r.t. $q(\boldsymbol{\beta}|\boldsymbol{y})$). Note that the two variational log-posteriors in (12) and (13) are coupled and the algorithm requires iterative updates until convergence. Finally, we can derive the variational lower bound of marginal L_{Po}:

$$\begin{aligned}\tilde{\mathcal{L}}_{Po} &= \mathbb{E}_{q(\boldsymbol{\beta})}[\log p(\boldsymbol{y}|\boldsymbol{\beta})] + \mathbb{E}_{q(\boldsymbol{\beta},\alpha_0)}[\log p(\boldsymbol{\beta}|\alpha_0)] \\ &+\mathbb{E}_{q(\alpha_0)}[\log p(\alpha_0)] - \mathbb{E}_{q(\boldsymbol{\beta})}[\log q(\boldsymbol{\beta})] - \mathbb{E}_{q(\alpha_0)}[\log q(\alpha_0)]. \quad (14)\end{aligned}$$

The VB approach requires numerical approximation (e.g., by importance sampling) while evaluating (12) and (13) and its convergence might be slow. For this reason we resort to another approach for approximating (11), which will yield an analytic solution.

In light of the Theorem 2.3 in [4], when the sample size T is large, we approximate the posterior with a Normal distribution:

$$D\left(\overline{\alpha}_0 + 1, \frac{\overline{\alpha}_0\boldsymbol{y}_0 + \boldsymbol{y}}{\overline{\alpha}_0 + 1}\right) \overset{T\to\infty}{\longrightarrow} \mathcal{N}(\hat{\boldsymbol{\beta}}, \boldsymbol{\Sigma}) \quad (15)$$

where $\overline{\alpha}_0$ denotes the mean of gamma distribution, $\hat{\boldsymbol{\beta}}$ is the m.l.e. of $p(\frac{\overline{\alpha}_0\boldsymbol{y}_0+\boldsymbol{y}}{\overline{\alpha}_0+1}|\boldsymbol{\beta})$, $\boldsymbol{\Sigma}^{-1} = (\overline{\alpha}_0+1)\boldsymbol{X}'diag\{\hat{\boldsymbol{\lambda}}\}\boldsymbol{X}$ ($\hat{\boldsymbol{\lambda}} = \exp(\boldsymbol{X}\hat{\boldsymbol{\beta}})$). Let $\boldsymbol{\Sigma}_0 = (\boldsymbol{X}'diag\{\hat{\boldsymbol{\lambda}}\}\boldsymbol{X})^{-1}$, and substituting (15) into (11) yields a multivariate Student distribution for the posterior of $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \approx \frac{b_T^{a_T}\Gamma(a_T + d/2)}{(\pi)^{d/2}|\boldsymbol{\Sigma}_0|^{1/2}\Gamma(a_T)}\left[b_T + \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]^{-(a_T+\frac{d}{2})}$$

which has a heavy-tail shape that favors the sparsity of solution. For point process observations, when $\overline{\alpha}_0 = 0$ or $\boldsymbol{y}_0 = \boldsymbol{y}$, the mean $\hat{\boldsymbol{\beta}}$ will coincide with the m.l.e. of $p(\boldsymbol{y}|\boldsymbol{\beta})$. When $\overline{\alpha}_0$ increases, the role of prior gradually dominates the data likelihood.

IV. Experimental Data and Results

The experimental data studied here are multiple neural spike trains simultaneously recorded from an awake rat hippocampus (HPC) and entorhinal cortex (EC), during a freely foraging task in an two-dimensional open circular environment. A total of 17 neurons were used in the present study. Among these 17 cells, 8 of them (#1-8) were recorded from EC (including putative interneurons, head-directional cells, and grid cells), and 9 of them were recorded from the HPC-CA1 area. We binned the total 44-min spike train recordings with 2-ms temporal resolution.

First, we study the neural encoding problem by fitting the two-dimensional receptive fields of place cells and grid cell. Let (c_x, c_y) denote the Cartesian coordinates in the two-dimensional environment. Similar to [1], we assume that the covariates $\boldsymbol{x}$ consist of the a set of orthogonal two-dimensional Zernike polynomials:

$$\boldsymbol{\beta}\boldsymbol{x}_t = \sum_{\ell=0}^{L}\sum_{m=-\ell}^{\ell}\beta_{\ell,m}Z_\ell^m(\rho(t), \psi(t)) \quad (16)$$

where Z_ℓ^m denotes the mth component of the ℓth-order Zernike polynomial, $\beta_{\ell,m}$ denotes the associated coefficients, $\rho(t) = r^{-1}\sqrt{(c_x(t) - c_1)^2 + (c_y(t) - c_2(t))^2}$ denotes the normalized radial distance, $\psi(t) = \tan^{-1}[(c_y(t) - c_2)/(c_x(t) - c_1)]$ denotes the phase (in rad), r and (c_1, c_2) denotes the radius and the center of the circular environment, respectively. Here, we have $r = 85$ cm, $c_1 = 125$ cm, $c_2 = 105$ cm. In order to fit non-regular receptive fields of place cells and grid cells, we chose $L = 6$, which produced $d = 28$ nonzero Zernike polynomials. Now the task of Bayesian inference is to infer the posterior of vector $\boldsymbol{\beta} = \{\beta_{\ell,m}\}$, and to select the most relevant bases (i.e., variable selection) for neural encoding of individual cells. We used three cells to illustrate our result: two HPC place cells (#9 and 11) and one EC grid cell (#3). Both VB logistic and Poisson regression were examined. In VB-logistic regression, we fixed the hyperprior parameters as $a_0 = b_0 = 0.0001$ such that it is close to the non-informative Jeffrey's prior. In Poisson case, we set $\boldsymbol{y}_0 = \mathbf{1}$ (such that the prior mode of $\boldsymbol{\beta}$ is $\mathbf{0}$), and varied the hyperparameter values for a_0 and b_0.The fitting results are illustrated in Fig. 1. As a comparison, the m.l.e. solutions are also presented (only results from logistic regression, the Poisson regression results are similar). Compared to the m.l.e. solution, In VB-logistic regression, the ARD switches off the redundant parameters and implicitly performs a variable selection. In other words, we might use a relatively compact model to achieve a comparable goodness-of-fit (in terms of receptive field fitting and the *Kolmogorov-Smirnov (KS) test* statistic [1]). In the approximate VB-Poisson regression, the sparsity of the solution highly depends on the $\overline{\alpha}_0$. In general, the VB approach produces a sparser solution than the m.l.e. Variable selection can be done by discarding the $\{\beta_i\}$ whose value is close to 0, without compromising the quality of data fitting. In all but one fitting cases, the m.l.e. achieves the lowest KS statistic, which is not surprising since it attempts to fit all given variables (the exception is in the grid cell's fitting, the lowest KS statistic was achieved by VB-logistic regression). Whereas in VB methods, we can adapt the priors to adjust the sparsity of the solution to avoid overfitting (when a uniform improper prior is used, the Bayesian solution reduces to the m.l.e.).

Next, we study the functional connectivity among all 17 cells using a network likelihood model presented earlier [13], [5], where the covariates $\boldsymbol{x}$ consist of 28 Zernike polynomials plus the spike counts from all ensemble cells within a number of previous temporal windows. Here, we have limited the firing history up to past 100 ms in our analysis, with a constant history window length of 5 ms. In this case, the size of the parameter space is large: d=28+(20×17)=368. Variable selection is important in this case. To demonstrate this, we use VB-logistic regression and set the hyperprior parameters as $a_0 = b_0 = 0.01$. To identify the significant nonzero connection weights between neurons, the criterion was set in a way that the mean±SD of $\{\beta_i\}$ is not overlapping with zero (if the weight is positive/negative, the contribution from one trigger cell to the target cell is excitatory/inhibitory). As an example, we use one HPC cell (#11) as the target cell, and model the other 16 neurons as trigger cells. The results on inferred spiking dependence coefficients are illustrated in Fig. 2. In this case, the pairwise functional connectivity is not significant, except the self-excitatory effect. Overall, the mean estimates of $\{\beta_i\}$ are very sparse (close to 0) and their SDs are much smaller than those obtained from m.l.e. In Fig. 3, we show one snapshot of inferred "sparse" functional connectivity among the 17 neurons, and we also show the averaged normalized connectivity strength (absolute value). As seen, self-excitation is dominant among the cells. Notably, our initial investigations of synthetic spike train data have confirmed that compared to m.l.e., the VB-approach has a greater chance to discover the 'true' connectivity coefficients in that it will be less likely to misidentify non-significant connections (results not shown due to space limit).

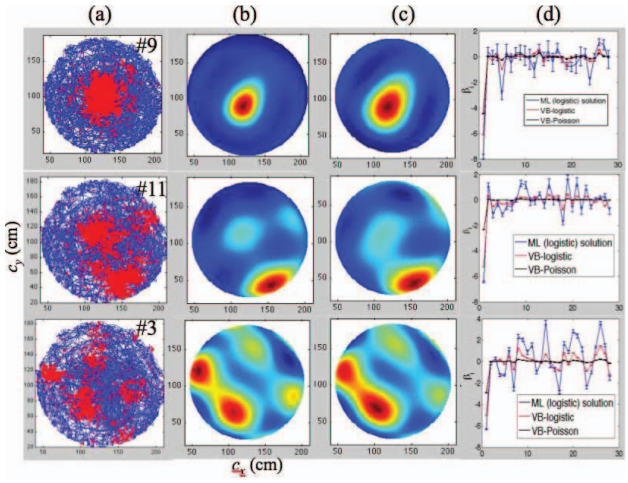

Fig. 1. (a) Spike maps from two place cells (#9, 11) and one grid cell (#3). The red dots indicate the spikes and blue curves indicate the rat's moving trajectory (unit: cm). (b) Fitted receptive fields with VB-logistic regression (the fitting result with m.l.e. is similar). (c) Fitting receptive fields with VB-Poisson regression ($\overline{\alpha}_0 = 0.01, 0.1, 0.05$ for the three cells, respectively). (d) Comparison of sparseness of 28 coefficient estimates of $\{\beta_i\}$ between the m.l.e. and VB-posterior mean. Error bars show SE or SD.

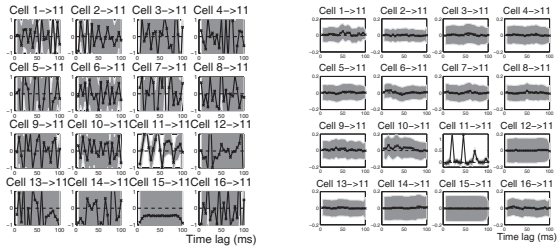

Fig. 2. The inferred coefficients of functional connectivity to a target cell #11 from m.l.e. (left) and VB-logistic regression (right). In the VB approach, a self-exciatory 40 Hz-rhythm is recovered, which is not clear from m.l.e.

V. Discussion

In this paper, we propose variational Bayesian inference methods for two point process GLMs in modeling neural spike trains. Unlike other Bayesian methods for GLM [7], Bayesian model averaging is conducted in a hierarchical modeling context (by adopting hyperpriors), and variable selection can be performed using either the ARD principle (in VB-logistic regression) or pseudo-observations (in VB-Poisson regression). Compared to the m.l.e., the Bayesian solution yields a full parameter posterior and it avoids overfitting via model averaging. In order to allow tractable hierarchical Bayesian inference, our methods use variational approximation to infer variational posteriors of parameters and hyperparameters. Our Bayesian inference procedures depend on the conjugate priors selected for the point process GLM. Hierarchical Bayesian modeling reduces the sensitivity of prior assumptions by adopting non-informative hyperpriors. In contrast to the non-hierarchical empirical Bayesian solutions, which impose *global* priors directly on the canonical or nuisance parameters [6], the priors of our VB solutions impose on a *hierarchical* structure on the regression parameters β.

In the future work, we plan to investigate the priors' sensitivity to the fitting performance using independent validation data sets. In addition, the point process GLMs can be generalized with inclusion

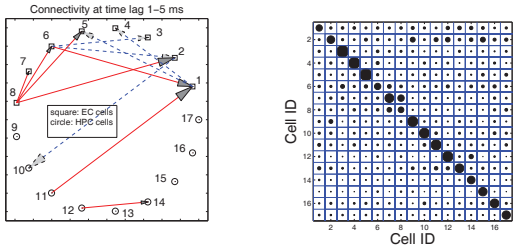

Fig. 3. Left: one snapshot of inferred functional interactions with VB-logistic regression (numbers indicate cell ID; arrows indicate the directional dependence; solid/dashed lines: excitatory/inhibitory connections). Right: averaged connectivity strength (the value is proportional to the circle area).

of latent state variables [18]. In that case, the inference of variational posteriors of state and parameters can be tackled by the VB-EM algorithm or its recent extensions [15], [20].

References

[1] R. Barbieri, L. M. Frank, D. P. Nguyen, M. C. Quirk, V. Solo, M. A. Wilson, and E. N. Brown, "Dynamic analyses of information encoding in neural ensembles," *Neural Comput.*, vol. 16, pp. 277–307, 2004.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York: Springer, 2006.

[3] E. N. Brown, L. M. Frank, D. Tang, M. C. Quirk, and M. A. Wilson, "A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells," *J. Neurosci.*, vol. 18, pp. 7411–7425, 1998.

[4] M-H. Chen and J. G. Ibrahim, "Conjugate priors for generalized linear models," *Statistica Sinica*, vol. 13, pp. 461–476, 2003.

[5] Z. Chen, D. F. Putrino, D. E. Ba, S. Ghosh, R. Barbieri, and E. N. Brown, "A regularized point process generalized linear model for assessing the functional connectivity in the cat motor cortex," in *Proc. IEEE EMBC'09* (pp. 5006–5009), Minneapolis, MN, 2009.

[6] S. Das and D. K. Dey, "On Bayesian analysis of generalized linear models using Jacobian technique," *The American Statistician*, vol. 60, pp. 264–268, 2006.

[7] D. K. Dey, S. K. Ghosh, and B. K. Mallick, editors, *Generalized Linear Models: A Bayesian Perspective*, New York: Marcel Dekker, 2000.

[8] P. Diaconis and D. Ylvisaker, "Conjugate priors for exponential families," *Ann. Statist.*, vol. 7, pp. 269–281, 1979.

[9] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis* (2nd ed.), Chapman & Hall/CRC, 2004.

[10] T. S. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statist. Comput.*, vol. 10, pp. 25–37, 2000.

[11] P. McCullagh and J. Nelder, *Generalized Linear Models* (2nd ed.), London: Chapman & Hall, 1989.

[12] R. Neal, *Bayesian Learning for Neural Networks*. Springer, 1996.

[13] M. Okatan, M. A. Wilson, and E. N. Brown, "Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity," *Neural Comput.*, vol. 17, pp. 1927–1961, 2005.

[14] L. Paninski, "Maximum likelihood estimation of cascade point-process neural encoding models," *Network*, vol. 15, pp. 243–262, 2004.

[15] Y. Qi and T. S. Jaakkola, "Parameter expanded variational Bayesian methods," *Adv. Neural Info. Proc. Syst. 19*, MIT Press, 2007.

[16] F. Rigat, M. de Gunst, and J. van Pelt, "Bayesian modelling and analysis of spatio-temporal neuronal networks," *Bayesian Analysis*, vol. 1, no. 4, pp. 733–764, 2006.

[17] M. Seeger, S. Gerwinn, and M. Bethge, "Bayesian inference for sparse generalized linear models," in *Proc. ECML'07*, pp. 298–309, 2007.

[18] A. Smith and E. N. Brown, "Estimating a state-space model from point process observations," *Neural Comput.*, vol. 15, pp. 965–991, 2003.

[19] I. H. Stevenson, et al., "Bayesian inference of functional connectivity and network structure from spikes," *IEEE Trans. Neural Syst. Rehab. Engr.*, vol. 17, pp. 203–213, 2009.

[20] J. Sung, Z. Ghahramani and S-Y. Bang, "Latent-space variational Bayes," *IEEE Trans. PAMI*, vol. 30, no. 2, pp. 2236–2242, 2008.

[21] W. Truccolo, et al., "A point process framework for relating neural spiking activity to spiking history, neural ensemble and covariate effects," *J. Neurophy.*, vol. 93, pp. 1074–1089, 2005.