Network Modeling for Short Over-Dispersed Spike-Counts: A Hierarchical Parametric Empirical Bayes Framework

Qi She, *Member, IEEE,* Beth Jelfs, *Member, IEEE,* Adam S. Charles, *Member, IEEE,* and Rosa H.M. Chan, *Senior Member, IEEE,*

Abstract—Accurate statistical models of neural spike responses can characterize the information carried by neural populations. Yet, challenges in recording at the level of individual neurons commonly results in relatively limited samples of spike counts, which can lead to model overfitting. Moreover, current models assume spike counts to be Poisson-distributed, which ignores the fact that many neurons demonstrate over-dispersed spiking behavior. The Negative Binomial Generalized Linear Model (NB-GLM) provides a powerful tool for modeling over-dispersed spike counts. However, maximum likelihood based standard NB-GLM leads to unstable and inaccurate parameter estimations. Thus, we propose a hierarchical parametric empirical Bayes method for estimating the parameters of the NB-GLM. Our method integrates Generalized Linear Models (GLMs) and empirical Bayes theory to: (1) effectively capture over-dispersion nature of spike counts from retinal ganglion neural responses; (2) significantly reduce mean square error of parameter estimations when compared to maximum likelihood based method for NB-GLMs; (3) provide an efficient alternative to fully Bayesian inference with low computational cost for hierarchical models; and (4) give insightful findings on both neural interactions and spiking behaviors of real retina cells. We apply our approach to study both simulated data and experimental neural data from the retina. The simulation results indicate the new framework can efficiently and accurately retrieve the weights of functional connections among neural populations and predict mean spike counts. The results from the retinal datasets demonstrate the proposed method outperforms both standard Poisson and Negative Binomial GLMs in terms of the predictive log-likelihood of held-out data.

I. INTRODUCTION

UNDERSTANDING functional connectivity among neurons is vital to deducing how populations of neurons process information. Functional connectivity focuses on statistical dependencies between neural time series (*e.g.*, spike counts, membrane potential, local field potential, EEG and fMRI) [1–4]. With the recent increase in accessibility of datasets containing spiking activities from large-scale neural populations, it is now possible to test the effectiveness of different methods for extracting functional dependences at the neuronal level. *Here, we consider the problem of recovering the connectivity between neurons in a network merely by observing their simultaneous spiking activity (*e.g.*, spike counts).*

Qi She and Rosa H.M. Chan are with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, e-mail: (qishe2-c@my.cityu.edu.hk, rosachan@cityu.edu.hk).

Beth Jelfs is with the School of Engineering, RMIT University, Australia, e-mail: (beth.jelfs@rmit.edu.au).

Adam S. Charles is with the Princeton Neuroscience Institute, Princeton University, New Jersey, USA, e-mail: (adamsc@princeton.edu).

Two of the most commonly used models for simultaneously recorded spiking activity are the Generalized Linear Models (GLMs) [5–9] and Latent Variable Models (LVMs) [10–14]. In the supervised setting, GLMs have used stimuli and spiking histories as covariates driving the spiking of a neural population [15]. Also GLMs are closely related to the well-known *Hawkes process* model [16], which has similarly been used extensively for network inference [17–20]. The GLM essentially introduces a nonlinearity to the Hawkes process that ensures positive rates and allows for super- or sub-linear influences between nodes. In the unsupervised setting, LVMs focus on extracting a low-dimensional, smooth, and time-evolving latent structure that can capture the variability of the recorded data, both temporally and spatially. However, in both these settings, the spike counts in each time bin are often assumed to be conditionally Poisson, given the shared signal [21]. While the Poisson assumption gives algorithmic conveniences, it implies the conditional mean and variance of spike counts are equal. This ignores the fact that in some cases the variance of spike counts could be much larger than its mean [22, 23], that is, the data is over-dispersed. The Negative Binomial (NB) model has been proposed as a solution to handling over-dispersed spike counts [24, 25]. *Here we intend to extract functional dependences among neurons and give insights over neural interactions. Thus,* NB-GLM *is a natural extension to achieve this goal, while simultaneously capturing the over-dispersion of each neuron.*

Despite the ease of implementation of maximum likelihood estimation for the NB-GLM, when the recorded length of spike-train data is short, and a large number of neurons are recorded simultaneously, the accuracy of the estimated coefficients using GLMs with NB responses is low [2, 3, 26]. Unfortunately, in typical experimental settings, we cannot obtain long sequences of high-quality neural data due to (i) the short lifetime of some neurons, (ii) the limited viable time of recording materials and (iii) the micro-movement of recording electrodes during an activity of the animal [27]. Hence, the size of dataset is often small, be that either due to the length of the experiment or even the need for real-time inference [28–30]. In this case, the maximum likelihood estimator of the parameters in the NB distribution leads to a large mean square error (MSE) under a standard GLM. To alleviate this problem, one can employ regularization priors in the form of a hierarchical model, as a means to trade off between bias and variance. *The key challenges of hierarchical modeling are how to flexibly*

design prior structures and efficiently solve the non-trivial inference problem, which are main focuses of our work.

In this paper, we propose a hierarchical empirical Bayes estimator for the probability parameters of NB-GLM, which helps to model Short Over-Dispersed Spike-Counts (we call "$\mathcal{SODS}$"). Finally, it can capture accurate spiking behavior of neurons and meanwhile recover functional connectivity under the GLM framework. Our hierarchical framework places a prior distribution on the parameters of the NB distribution, which can be estimated using empirical Bayes. The hyperparameters of the prior distribution are estimated using maximum marginal likelihood methods. The estimated value can then be used to obtain the mean spike counts. In summary, our main contributions are four-fold:

1) *Provide a hierarchical extension of the NB-GLM for modeling the statistical dependences among neural responses including a flexible link function;*
2) *Develop an efficient empirical Bayes method for inference of the hierarchical NB-GLM parameters;*
3) *Present more accurate prediction performance on retinal ganglion cells compared with state-of-the-art methods;*
4) *Give insightful findings on both neural interactions and spiking behaviors of real retina cells.*

Generally, this paper is organized as follows. In Section II, we review the properties of the Negative Binomial Distribution and the differences between full and empirical Bayes approaches. In Section III, we introduce the $\mathcal{SODS}$. Section IV discusses parameter estimations in $\mathcal{SODS}$, via numerical optimization of the maximum marginal likelihood, and the roles of these parameters. Results for both simulated and experimental data are presented in Section V. Discussion of our contributions and findings are concluded in Section VI.

II. REVIEW

A. Negative Binomial Distribution

A discrete random variable Y follows the Negative Binomial distribution $\mathrm{NB}(\boldsymbol{r}, \theta)$, with shape parameter $\boldsymbol{r}$ and probability parameter θ, if

$$P(Y = y) = \binom{\boldsymbol{r} + y - 1}{y} \theta^{\boldsymbol{r}} (1 - \theta)^y. \quad (1)$$

This can be seen as a extension of the Poisson distribution $\mathrm{Poisson}(\lambda)$, in which the rate parameter λ is generated from the Gamma distribution:

$$Y \mid \lambda \quad \sim \quad \mathrm{Poisson}(\lambda), \quad (2)$$

$$\lambda \mid \boldsymbol{r}, \theta \quad \sim \quad \mathrm{Gamma}\left(\boldsymbol{r}, \frac{\theta}{1 - \theta}\right). \quad (3)$$

Recall that $Y \sim \mathrm{Poisson}(\lambda)$ if $P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}$, and $\lambda \sim \mathrm{Gamma}(\boldsymbol{r}, \phi)$ if $p(\lambda) = \lambda^{\boldsymbol{r}-1} \phi^{\boldsymbol{r}} \frac{\exp(-\lambda\phi)}{\Gamma(\boldsymbol{r})}$. The mean and variance of the NB distribution are $\mathbb{E}[Y] = \frac{(1-\theta)\boldsymbol{r}}{\theta}$ and $\mathrm{Var}[Y] = \frac{(1-\theta)\boldsymbol{r}}{\theta^2}$, which has $\mathrm{Var}[Y] > \mathbb{E}[Y]$ since $0 < \theta < 1$.

Figure 1 (a) shows the relationship between variance and mean of Negative Binomial and Poisson distributions. The variance of the NB distribution is larger than the mean, which shows super-Poisson variability [23, 31]. Figure 1 (b) shows

the probability mass function of NB distribution with different combinations of parameters $\boldsymbol{r}$ and θ.
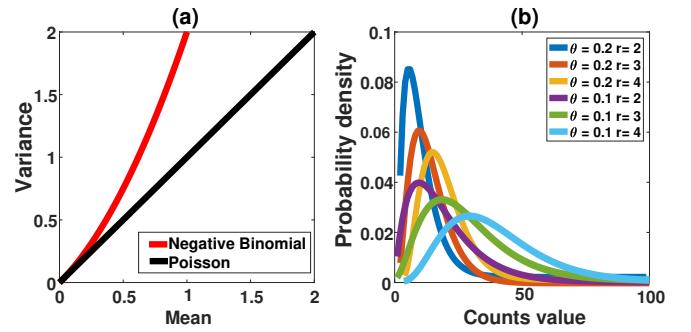

Fig. 1. (a) The relationship between variance and mean of Poisson and Negative Binomial distributions. Negative Binomial shows super-Poisson variability (variance larger than mean). (b) The probability density function of NB distribution with different parameters ($\theta = \{0.1, 0.2\}$, $\boldsymbol{r} = \{2, 3, 4\}$).

B. Fully Bayesian and Empirical Bayes Inference

Functional connectivity is modeled as an input-output system, which links the Negative Binomial output and spiking activities of input neurons via a hierarchical model. In the hierarchical setting, we use either fully Bayesian inference or empirical Bayes to estimate the model parameters. Fully Bayesian inference assumes specific hyperprior over the hyperparameters, which need to be integrated out. As we often cannot obtain the closed form of this marginalization, fully Bayesian inference requires a sampling strategy to approximate this distribution. Correspondingly, this comes at a high computational cost, especially for high-dimensional data [32].

On the other hand, the empirical Bayes inference sets the parameters in the highest level of the hierarchical model with their most likely value. Setting the hyperparameters by maximizing the marginal likelihood function incurs a much lower computational cost. Hence, by combining empirical Bayes with the Negative Binomial GLM we can produce an estimator for the parameters of the Negative Binomial distribution which should efficiently handle both over-dispersion and smaller data-sets. *The key is to establish a network model in this framework and still capture super-Poisson spiking behavior.*

III. PROPOSED METHOD

A. Hierarchical Negative Binomial Model

Figure. 2 (a) illustrates a demo of simple network considered in this work. We represent functional dependences in this graph with the connection strengths (weights) between neurons. Note that we can use input neurons' spiking activities (*e.g.*, neurons #2: $x_1(t)$, #3: $x_2(t)$, #4: $x_3(t)$, #5: $x_4(t)$) as regressors to predict an output neuron's spike counts (*e.g.*, neuron #1: $y(t)$). Figure. 2 (b) presents neuron #1 and #5 have excitatory and inhibitory effects on neuron #3 via a flexible link function and a NB distribution model, respectively. *The accurate modeling of both link functions and the NB model can help to effectively retrieve intrinsic coupling strengths.*

Let Y_{ij} be the spike counts recorded from the jth experimental trial at time i. We assume that $\{Y_i\}_{i=1}^{K}$ are generated
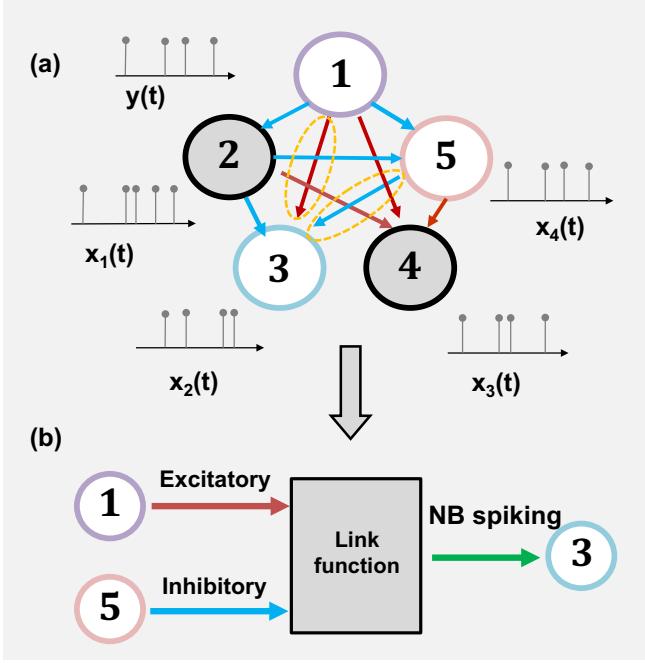
Fig. 2. (a) A simple network model considered in our work, with excitatory and inhibitory effects and generates NB spiking behavior. (b) A illustration of neuron #1 and #5 have effects on #3 through a flexible link function and then a NB distribution. The observed data are multiple spike-train data recorded simultaneously, which are presented as $x_{1:4}(t)$ and $y(t)$. Each gray line in x and y signals indicates one spike obtained.

from the Negative Binomial distribution (with shape parameter r and probability parameter θ_i):

$$Y_{ij} \mid r, \theta_i \sim \text{NB}(r, \theta_i). \tag{4}$$

We use the beta distribution, the conjugate prior of the Negative Binomial distribution, as the prior for θ_i:

$$\theta_i \sim \text{Beta}(\alpha_i, \beta_i), \tag{5}$$

i.e.,

$$p(\theta_i) = \frac{\theta_i^{\alpha_i-1}(1-\theta_i)^{\beta_i-1}}{\text{B}(\alpha_i, \beta_i)},$$

where α_i, β_i are the hyperparameters, and

$$\text{B}(\alpha_i, \beta_i) = \int_0^1 x^{\alpha_i-1}(1-x)^{\beta_i-1}\text{d}x = \frac{\Gamma(\alpha_i)\Gamma(\beta_i)}{\Gamma(\alpha_i+\beta_i)}, \tag{6}$$

is the beta function, and $\Gamma(t)$ is the Gamma function.

We introduce the hyperparameter $\sigma \equiv \alpha_i + \beta_i$, which can be interpreted as a precision parameter that reflects the degree of prior belief in the GLM, and is fixed across different time bins. The prior mean is $\mu_i \equiv \mathbb{E}(\theta_i|\alpha_i, \beta_i) = \frac{\alpha_i}{\sigma}$, and $\alpha_i = \sigma\mu_i, \beta_i = \sigma(1-\mu_i)$. We can thus determine the beta distribution by learning μ_i and σ. In particular, we learn μ_i by using a Generalized Linear Model (GLM) with the mean counts of input neurons at the previous time step (x_{i-1}) (see graphical model in Figure. 3). A vector of functional weights, ω, capture the directed effects of input neurons on the output neuron. μ_i is modeled as:

$$g(\mu_i) = x_{i-1}^\top \omega. \tag{7}$$

Normally, the link function $g(\cdot)$ is predefined using specific form such as log, $logit$, $probit$, $identity$, and $log-log$ [33]. However, we do not want to constrain the link function to be a fixed form. Hence, we propose a family of link functions governed by a hyperparameter, γ, such that,

$$g(\mu_i, \gamma) = \log\left(\frac{(1-\mu_i)^{-\gamma}-1}{\gamma}\right). \tag{8}$$

We design this link family with three considerations: (1) it can represent many widely used link functions. For instance, the *logit* function when $\gamma = 1$, the complementary $log-log$ link function when $\gamma \approx 0$, and the log function if $\gamma = -1$; (2) It should constrain the prior mean, modeled as the mean value of the probability parameter, to $\mu_i > 0$ and (3) it can be inversed and provide gradients for hyperparameters γ and ω (discussed in Section IV-A) easily. Note that the hyperparameter γ, is a flexible parameter which determines the specific form of the link function, $g(\cdot)$, therefore ensuring the flexibility of the nonlinear transformation from the regressors to the output.

Denoting the inverse link function by $g^{-1}\left(x_{i-1}^\top\omega, \gamma\right)$. Thus, the prior mean becomes

$$\mu_i = g^{-1}\left(x_{i-1}^\top\omega, \gamma\right) = 1 - \left(\gamma e^{x_{i-1}^\top\omega} + 1\right)^{-\frac{1}{\gamma}}. \tag{9}$$

In the subsequent paragraphs, we let $\zeta \equiv \{r, \omega, \sigma, \gamma\}$ denote all the model parameters.

Table I provides a complete summary of all the variables used in the "$\mathcal{SODS}$" estimator.

TABLE I
SUMMARY OF VARIABLE DEFINITIONS.

Variable	Definition
y_{ij}	Spike counts of j-th trial at i-th time bin
x_{i-1}	Vector of regressors at (i-1)-th time step
λ_i	Mean of Poisson distribution (firing rate of neurons)
θ_i	Probability parameter of Negative Binomial distribution
r	# failures in Negative Binomial distribution
α_i, β_i	Parameters of beta distribution
σ	Degree of freedom of prior distribution ($\sigma \equiv \alpha_i + \beta_i$)
ω	Vector of weights
$g(\cdot)$	Family of link functions
γ	Parameter determining specific form of link function
μ_i	Mean of prior distribution
n_i	Number of trials at i-th time bin
$\bar{y}_i$	Mean spike counts across all trials
π_i	Weight of the observation component in our estimator
K	Data length (the total number of bins)
A_i, B_{ij}, C_{ij}	Components of the gradients
p	Element number of ω
N	Total number of neurons
ζ	$(r, \omega, \sigma, \gamma)$
k	Vector of Lagrangian coefficients
m	Variable number from 1 to $N+3$
d	Vector of search direction
Δ_t	Step length of iteration

Figure. 3 shows the graphical model of the proposed hierarchical structure. The observation data are Y_i and x_{i-1}; μ_i and θ_i are latent variables; $\zeta \equiv \{r, \omega, \sigma, \gamma\}$ are global parameters, which are consistent across all time steps.

B. Empirical Bayes Estimator: $\mathcal{SODS}$

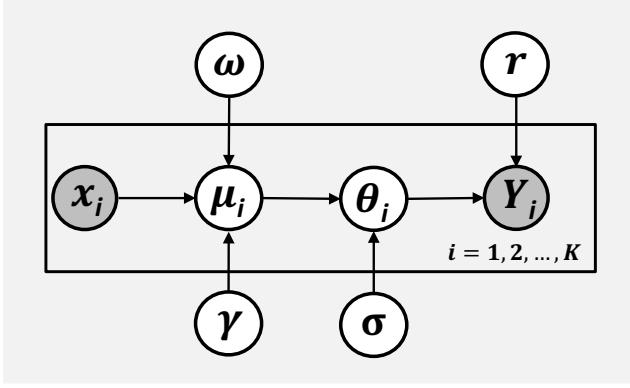First, we study the posterior distribution of θ_i. As the Beta distribution is the conjugate prior of the Negative Binomial

Fig. 3. Graphical representation of the proposed model. The prior mean μ_i is formed from the GLM of the input regressors x_{i-1}, the weight vector ω, and link function parameterized by γ. μ_i is the mean of the beta prior of NB probability parameter θ_i. σ is the degree of freedom of the prior beta distribution. Finally, θ_i, together with the shape parameter for the NB distribution r, generate the observed spike counts Y_i. Shaded nodes x_{i-1} and Y_i denote observed random variables; μ_i and θ_i are latent random variables. r, ω, σ, and γ are hyperparameters. The rectangular box is "plate notation", which denotes replication.

likelihood function, the posterior distribution of θ_i given $Y_{ij} = y_{ij}$ follows the beta distribution [24]:

$$\theta_i \mid y_{ij} \sim \text{Beta}\Big(\sigma\mu_i + n_i r, \sigma(1-\mu_i) + n_i\overline{y}_i\Big), \quad (10)$$

where n_i is the number of trials in the ith time bin, and $\overline{y}_i$ is the mean count across all training trials at bin i. Substituting (9) into (10), we get

$$\theta_i \mid y_{ij}, \zeta \sim \text{Beta}(\sigma g^{-1}(x_{i-1}^\top \omega, \gamma) + n_i r,$$
$$\sigma(1 - g^{-1}(x_{i-1}^\top \omega, \gamma)) + n_i \overline{y}_i). \quad (11)$$

We take the mean of this posterior distribution as the estimator for θ_i, we call this estimator derived from our model as "$\mathcal{SODS}$" estimator, and denoted as θ^{SODS}:

$$\theta_i^{\text{SODS}} = \mathbb{E}(\theta_i \mid y_{ij}, \zeta) = \frac{n_i r + \sigma g^{-1}(x_{i-1}^\top \omega, \gamma)}{n_i r + n_i \overline{y}_i + \sigma}, \quad (12)$$

which can be rewritten as

$$\theta_i^{\text{SODS}} = \pi_i \left(\frac{r}{r + \overline{y}_i}\right) + (1 - \pi_i) g^{-1}(x_{i-1}^\top \omega, \gamma), \quad (13)$$

where $\pi_i = \frac{r + \overline{y}_i}{r + \overline{y}_i + \sigma/n_i} \in (0, 1)$. Hence, θ_i^{SODS} is a convex combination of the data-driven estimate of θ_i and the prior mean of the GLM. We can consider π_i as the parameter to achieve a trade-off between bias and variance. σ can be viewed as a precision parameter. When $\sigma \to 0$, thus $\pi_i \to 1$, it results in θ_i^{SODS} only reflecting the observed data. When $\sigma \to \infty$, thus $\pi_i \to 0$, the estimator reduces to be standard Negative Binomial GLM, which links the probability parameter with the input regressors via a link function:

$$\mathbb{E}(\theta_i \mid y_{ij}, \zeta) = g^{-1}(x_{i-1}^\top \omega, \gamma). \quad (14)$$

With the estimator θ_i^{SODS}, the mean spike counts can then be obtained from Eq. (1):

$$\mathbb{E}[Y_i \mid \theta_i^{\text{SODS}}] = r\Big(\frac{1}{\theta_i^{\text{SODS}}} - 1\Big)$$
$$= r\frac{n_i \overline{y}_i + \sigma - \sigma g^{-1}(x_{i-1}^\top \omega, \gamma)}{n_i r + \sigma g^{-1}(x_{i-1}^\top \omega, \gamma)}. \quad (15)$$

IV. Maximum Marginal Likelihood

θ^{SODS} depends on $\zeta \equiv \{r, \omega, \sigma, \gamma\}$. To estimate ζ, we use the empirical Bayes approach. We first derive the marginal likelihood function, where the marginal distribution is the spike counts conditioned only on the hyperparameters. We then optimize the marginal likelihood using gradient-based *Broyden-Fletcher-Goldfarb-Shanno* (*BFGS*) algorithm. Finally, we discuss the roles of model parameters in our model and how to use prior knowledge to set the initial value to give more stable and accurate optimization results.

Since using the maximum marginal likelihood approach does not include any assumptions on the hyperparameters, we have the benefit of relatively low computational cost for estimating high-dimensional parameters. To derive the marginal likelihood, we need to integrated out the probability parameter θ_i, as $p(y_{ij}) = \int p(\theta_i)p(y_{ij} \mid \theta_i)d\theta_i$. Reformulating the Negative Binomial likelihood as,

$$p(y_{ij} \mid \theta_i) = \frac{\Gamma(r + y_{ij})}{\Gamma(y_{ij} + 1)\Gamma(r)}\theta_i^r(1 - \theta_i)^{y_{ij}}$$
$$= \frac{\Gamma(r + y_{ij})}{\Gamma(y_{ij})\Gamma(r)}\frac{\Gamma(y_{ij})}{\Gamma(y_{ij} + 1)}\theta_i^r(1 - \theta_i)^{y_{ij}}$$
$$= \frac{\theta_i^r(1 - \theta_i)^{y_{ij}}}{B(r, y_{ij})y_{ij}}, \quad (16)$$

then, the marginal likelihood is

$$p(y_{ij}) = \int_0^1 p(\theta_i)\frac{\theta_i^r(1 - \theta_i)^{y_{ij}}}{B(r, y_{ij})y_{ij}}d\theta_i$$
$$= \frac{1}{B(r, y_{ij})B(\alpha_i, \beta_i)y_{ij}}\int_0^1 \theta_i^{r+\alpha_i-1}(1 - \theta_i)^{y_{ij}+\beta_i-1}d\theta_i$$
$$= \frac{B(r + \alpha_i, y_{ij} + \beta_i)}{B(r, y_{ij})B(\alpha_i, \beta_i)y_{ij}}. \quad (17)$$

Substituting α_i and β_i into Eq. 17 with $\alpha_i = \sigma g^{-1}(x_{i-1}^\top \omega, \gamma)$, $\beta_i = \sigma - \sigma g^{-1}(x_{i-1}^\top \omega, \gamma)$, the marginal density of the spike counts conditioned on ζ is

$$p_i(y_{ij}|\zeta) = \frac{B\Big(r + \sigma g^{-1}(x_{i-1}^\top \omega, \gamma), y_{ij} + \sigma - \sigma g^{-1}(x_{i-1}^\top \omega, \gamma)\Big)}{B(r, y_{ij})B\Big(\sigma g^{-1}(x_{i-1}^\top \omega, \gamma), \sigma - \sigma g^{-1}(x_{i-1}^\top \omega, \gamma)\Big)y_{ij}},$$

and conditioning on y_{ij}, the log marginal likelihood $\ell(\zeta) = \sum_{i=1}^K \sum_{j=1}^{n_i} \log p_i(y_{ij})$ of the conditional posterior is

$$\ell(\zeta) \propto \sum_{i=1}^K \sum_{j=1}^{n_i}\Bigg[\log\Gamma\Big(r + \sigma g^{-1}(x_{i-1}^\top \omega, \gamma)\Big) + \log\Gamma(r + y_{ij})$$
$$+ \log\Gamma(\sigma) - \log\Gamma(r + y_{ij} + \sigma) - \log\Gamma(r)$$
$$+ \log\Gamma\Big(y_{ij} + \sigma - \sigma g^{-1}(x_{i-1}^\top \omega, \gamma)\Big)$$
$$- \log\Gamma\Big(\sigma - \sigma g^{-1}(x_{i-1}^\top \omega, \gamma)\Big)$$
$$- \log\Gamma\Big(\sigma g^{-1}(x_{i-1}^\top \omega, \gamma)\Big)\Bigg]. \quad (18)$$

r, ω, σ and γ can then be obtained by maximizing Eq. (18).

A. Optimization of Hyperparameters

In hierarchical modeling, it is hard to obtain closed-forms for the estimators. Thus, we use numerical maximization. Here, we need to approximate the Hessian matrix at each iteration, which is achieved by applying use the Quasi-Newton method to the approximation of the Hessian matrix. Specifically, $BFGS$ [34] was used to approximate the Hessian matrix at each iteration. We derive the gradients w.r.t. $r, \omega, \sigma, \gamma$ as

$$\frac{\partial \ell(\zeta)}{\partial r} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \{ \Psi \left[r + \sigma g^{-1}(x_{i-1}^\top \omega, \gamma) \right] + \Psi(r + y_{ij}) $$
$$- \Psi(r + y_{ij} + \sigma) - \Psi(r) \}$$

$$\frac{\partial \ell(\zeta)}{\partial \omega_p} = \sigma \sum_{i=1}^{K} \sum_{j=1}^{n_i} \frac{\partial g^{-1}(x_{i-1}^\top \omega, \gamma)}{\partial \omega_p} (A_i - B_{ij})$$

$$\frac{\partial \ell(\zeta)}{\partial \sigma} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \{ A_i g^{-1}(x_{i-1}^\top \omega, \gamma) $$
$$+ B_{ij} \left[1 - g^{-1}(x_{i-1}^\top \omega, \gamma) \right] + C_{ij} \}$$

$$\frac{\partial \ell(\zeta)}{\partial \gamma} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \frac{\partial g^{-1}(x_{i-1}^\top \omega, \gamma)}{\partial \gamma} (A_i - B_{ij}), \qquad (19)$$

where

$$A_i = \Psi \left(r + \sigma g^{-1}(x_{i-1}^\top \omega, \gamma) \right) - \Psi \left(\sigma g^{-1}(x_{i-1}^\top \omega, \gamma) \right)$$
$$B_{ij} = \Psi \left(y_{ij} + \sigma - \sigma g^{-1}(x_{i-1}^\top \omega, \gamma) \right)$$
$$- \Psi \left(\sigma - \sigma g^{-1}(x_{i-1}^\top \omega, \gamma) \right)$$
$$C_{ij} = \Psi(\sigma) - \Psi(r + y_{ij} + \sigma), \qquad (20)$$

$\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$ is the digamma function, ω_p is the individual element in the vector ω with $p = (1, 2, \ldots, N)$ and N is the number of involved neurons. Moreover, in (19), the ease of implementation of gradients calculations give

$$\frac{\partial g^{-1}(x_{i-1}^\top \omega, \gamma)}{\partial \omega_p} = x_p e^{x_{i-1}^\top \omega} \left(\gamma e^{x_{i-1}^\top \omega} + 1 \right)^{-\frac{1}{\gamma} - 1}$$

$$\frac{\partial g^{-1}(x_{i-1}^\top \omega, \gamma)}{\partial \gamma} = - \left(\gamma e^{x_{i-1}^\top \omega} + 1 \right)^{-\frac{1}{\gamma}}$$
$$\left[\frac{\log \left(\gamma e^{x_{i-1}^\top \omega} + 1 \right)}{\gamma^2} - \frac{e^{x_{i-1}^\top \omega}}{\gamma(\gamma e^{x_{i-1}^\top \omega} + 1)} \right] .$$

However, during optimizations, we found sometimes γ will have negative value (<0), while a probability parameter should be within $[0, 1]$. To solve this problem, we further enforce the constraint $\gamma > 0$ using Sequential Quadratic Programming (SQP) [35], applying both SQP and the Quasi-Newton method at each updating step. In SQP, to form a quadratic program and find a line search direction by minimizing the quadratic subproblem We form a Lagrangian function with all the hyperparameters $\zeta = \{r, \omega, \sigma, \gamma\}$, as

$$L(\zeta, \mathbf{k}) = \ell(\zeta) + \sum_{m=1}^{N+3} k_m \zeta_m, \qquad (21)$$

where, $\mathbf{k} = \{k_1, k_2, \ldots, k_{N+3}\}$ are the Lagrangian coefficients, and ζ_m is the element in ζ. The total number of parameters to be estimated is $N + 3$, where N (the number of

neurons involved in the network) is equivalent to the length of vector ω and (r, σ, γ) are the other three parameters. We then form the quadratic programming subproblem

$$\min_{\mathbf{d} \in \mathbb{R}^d} \quad \nabla \ell^\top(\zeta)\mathbf{d} + \frac{1}{2}\mathbf{d}^\top \nabla^2 L(\zeta, \mathbf{k})\mathbf{d},$$
$$s.t. \quad \nabla \zeta_m^\top d + \zeta_m \leq 0, \quad m = 1, \ldots, N + 3. \qquad (22)$$

These quadratic problems are solved using an active-set algorithm [36]. At each iteration step t, we find the linear search direction $\mathbf{d}$, and then use a line search procedure to find the step length parameter, Δ, which achieves a sufficient decrease in the merit function $f_m(\zeta(t) + \Delta_t \cdot \mathbf{d}(t)) < f_m(\zeta(t))$[37]. We update the group parameters ζ until converged as below

$$\zeta(t + 1) = \zeta(t) + \Delta_t \cdot \mathbf{d}(t). \qquad (23)$$

B. The Role of the Hyperparameters

The parameters in the "$\mathcal{SODS}$" estimator play different roles in explaining the neural spike train dataset. In this section we discuss each of them in turn, and present rules to tune the initial values used in the optimization procedure, in order to improve the estimation efficiently.

- r is the shape parameter for Negative Binomial response. Physically, it controls underlying firing rates of neurons. In real situations, the actual firing rate of the underlying neural population may not be very high, such as in hippocampal areas. In this case, to get reasonable mean spike counts, we should make sure the initial value of r is small, as this helps the spike count observations match the low firing rates. Accordingly, if we believe a brain area has a high firing rate, such as in the motor cortex, we can initialize r to a higher value. In Figure. 4, we keep the same probability parameter, and show the influence of different values of r on the spike counts. The results indicate that for the Negative Binomial distribution, larger values of r give larger spike counts.
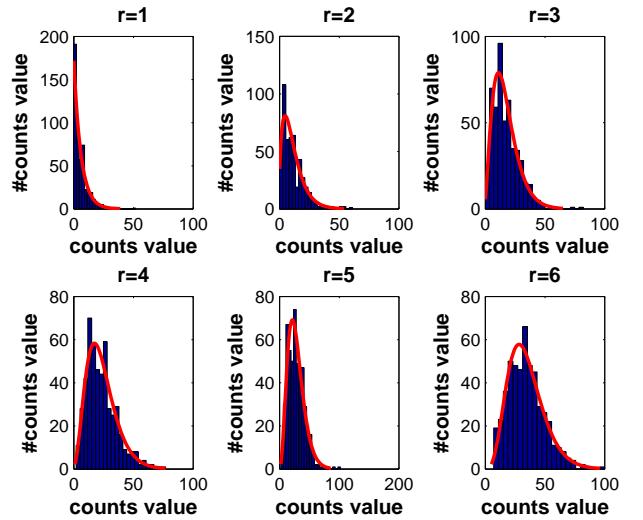

Fig. 4. Negative Binomial distribution with different r and the same θ. Larger r leads to higher probability to generate large counts value.

- $\star$ ω is a vector of coupling weights, which help to capture the directed effects of input neurons on the output neuron.

It is the core part to introduce functional connectivity into our hierarchical model. This vector can also include other factors such as the spiking history of the output neuron or external stimuli *e.g.*, if we have prior knowledge, for instance the pixels of an image shown to excite the retinal neurons. These weights can be positive or negative, which can be explained as neurons having either an excitatory or inhibitory effect on the output neuron. In section V-A, using simulated data, we test the ability of proposed estimator to capture these excitatory and inhibitory effects of functional connectivity. The initial values of elements in $\boldsymbol{\omega}$ are randomly chosen from (-1,1).

- $\boldsymbol{\sigma}$ is the degrees of freedom of the beta distribution. It controls the balance between our limited data sample and prior knowledge. From Eq. (13), we can see our proposed estimator is the weighted combination of the observed data $\theta_i^{obs} = \frac{\boldsymbol{r}}{\boldsymbol{r}+\overline{y}_i}$ and standard GLM estimation $\theta_i^{\text{GLM}} = g^{-1}\left(\boldsymbol{x}_{i-1}^{\top}\boldsymbol{\omega}, \boldsymbol{\gamma}\right)$. The weights of each component are $\pi_i = \frac{n_i \boldsymbol{r} + n_i \overline{y}_i}{n_i \boldsymbol{r} + n_i \overline{y}_i + \boldsymbol{\sigma}}$ and $1 - \pi_i = \frac{\boldsymbol{\sigma}}{n_i \boldsymbol{r} + n_i \overline{y}_i + \boldsymbol{\sigma}}$. Thus, if $\boldsymbol{\sigma}$ is large, the proposed method is close to the GLM of the prior mean; when it is small, the estimator is approaching the observed data. The initial value of $\boldsymbol{\sigma}$ should be determined by the number of trials n_i, such that, the more trials we have, smaller $\boldsymbol{\sigma}$ should be, which means more confidence we can have in the observed data.

- $\boldsymbol{\gamma}$ conveys the nonlinear effects of input neurons on the output neurons, which selects the best fit link function for the dataset. When $\gamma = 1$, it is the commonly used *logit* function; when $\gamma \approx 0$, it becomes the complementary $log - log$ link function; and when $\gamma = -1$, it is log link function. Regularly, GLMs choose the link function by specifying a parametric link, our work, however, determines the unknown parameter automatically. Learning from the dataset itself allows our approach to automatically select a suitable link function. The initial value of $\boldsymbol{\gamma}$ is determined so as to result in relatively low firing rate, which has empirically been shown to give good performance for spike count prediction.

We summarize the derivation of the "$\mathcal{SODS}$" estimator and the mean spike counts in Algorithm 1, which shows the steps for empirical Bayes inference on our hierarchical model.

V. RESULTS

In this section, we test our framework on both simulated and experimental recordings. The simulated data is generated via the process outlined in Figure. 3, and the general setting up of simulations is listed in Table. II. The experimental datasets are 4 retinal datasets taken from visual experiments, each with different numbers of neurons [38]. We use them to validate the stability and fitness of our model for neural recordings.

A. Simulated Data

The hyperparameters $\boldsymbol{r}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \boldsymbol{\gamma}$ were given different combinations for comparison. The range of configurations of the parameters tested for the simulated data are shown in the Table II, where N_s is the number of simulated trials, N is the number of neurons in our simulation, and K represents

---

Algorithm 1 The Hierarchical Parametric Empirical Bayes Framework for Short Over-Dispersed Spike-Counts

---
Input: $\boldsymbol{x}_{i-1} = [x_{i-1,1}, x_{i-1,2}, \ldots, x_{i-1,N}]^{\top}$ and y_{ij} ($i = 1, ..., K$ and $j = 1, ..., n_i$).

Output: $\mathbb{E}(\theta_i \mid \boldsymbol{r}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \boldsymbol{\gamma})$ and $\mathbb{E}[Y_{ij}|\theta_i]$.

1: Initialize $\boldsymbol{r}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \boldsymbol{\gamma}$ based on Section IV-B.
2: Form the linear regressors $\eta_i = \boldsymbol{x}_{i-1}^{\top}\boldsymbol{\omega}$.
3: Form the link function $\eta_i = g(u_i, \boldsymbol{\gamma})$ based on Eq. (8).
4: Derive the prior information as $u_i = g^{-1}(\boldsymbol{x}_{i-1}^{\top}\boldsymbol{\omega}, \boldsymbol{\gamma})$.
5: Construct the beta Negative Binomial log marginal likelihood function as Eq. (18).
6: **repeat**
7:     Compute gradients of hyperparameters using Eq. (19)
$$\nabla\ell(\boldsymbol{r}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \boldsymbol{\gamma}) = \left(\frac{\partial\ell}{\partial\boldsymbol{r}}, \frac{\partial\ell}{\partial\omega_p}, \frac{\partial\ell}{\partial\boldsymbol{\sigma}}, \frac{\partial\ell}{\partial\boldsymbol{\gamma}}\right).$$
8:     Apply $BFGS$ to update Hessian matrix of Lagrangian function in Eq. (21). We denote it as $\nabla^2 L(\boldsymbol{\zeta}, \mathbf{k})$.
9:     Solve SQP sub-problem with active-set algorithm
$$\min_{\mathbf{d}\in\mathbb{R}^d} \quad \nabla\ell^{\top}(\boldsymbol{\zeta})\mathbf{d} + \frac{1}{2}\mathbf{d}^{\top}\nabla^2 L(\boldsymbol{\zeta}, \mathbf{k})\mathbf{d},$$
$$s.t. \quad \nabla\zeta_m^{\top}d + \zeta_m \leq 0, \quad m = 1, \ldots, N+3.$$
10:     Choose Δ_t so that merit function achieves
$$f_m(\boldsymbol{\zeta}(t) + \Delta_t \cdot \mathbf{d}(t)) < f_m(\boldsymbol{\zeta}(t))$$
11:     Update
$$\boldsymbol{\zeta}(t+1) = \boldsymbol{\zeta}(t) + \Delta_t \cdot \mathbf{d}(t).$$
12: **until** convergence $|\nabla\ell^{\top}(\boldsymbol{\zeta}(t+1))\boldsymbol{\zeta}(t+1)| < 1 \times 10^{-8}$
13: Calculate the empirical Bayes estimation of probability parameter $\mathbb{E}(\theta_i \mid \boldsymbol{r}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \boldsymbol{\gamma})$ from Eq. (12) as
$$\theta_i^{\text{SODS}} = \mathbb{E}(\theta_i|\boldsymbol{r}, \boldsymbol{\omega}, \boldsymbol{\sigma}, \boldsymbol{\gamma}) = \frac{n_i\boldsymbol{r} + \boldsymbol{\sigma}g^{-1}\left(\boldsymbol{x}_{i-1}\top\boldsymbol{\omega}, \boldsymbol{\gamma}\right)}{n_i\boldsymbol{r} + n_i\overline{y}_i + \boldsymbol{\sigma}}.$$
14: Obtain the mean spike counts based on Eq. (1):
$$\mathbb{E}[Y_{ij}|\theta_i] = \boldsymbol{r}\left(\frac{n_i\overline{y}_i + \boldsymbol{\sigma} - \boldsymbol{\sigma}g^{-1}\left(\boldsymbol{x}_{i-1}\top\boldsymbol{\omega}, \boldsymbol{\gamma}\right)}{n_i\boldsymbol{r} + \boldsymbol{\sigma}g^{-1}\left(\boldsymbol{x}_{i-1}\top\boldsymbol{\omega}, \boldsymbol{\gamma}\right)}\right).$$

---

the data length (bins) of each time series in the training dataset. The spike counts of output neuron is sampled from Negative Binomial model. Depending on the combination of these parameters we can provide different types of observed spike counts. Basically, we conducted two simulation tests:

- Interaction estimation. We randomly assigned excitatory ($\boldsymbol{\omega} > 0$) or inhibitory weights ($\boldsymbol{\omega} < 0$) to the neural population. Our goal was to identify whether we could recover the weights of the interactions accurately.

- Performance of the "$\mathcal{SODS}$" estimator. In the simulation process, we have the underlying truth regarding the mean spike counts based on Eq. (1), $\mathbb{E}[Y_i|\theta_i] = \boldsymbol{r}(\frac{1}{\theta_i} - 1)$. We tested the performance of the "$\mathcal{SODS}$" estimator by calculating MSE with this true value.

Table III, IV and V show comparisons of the MSE of the standard Negative Binomial GLM and the "$\mathcal{SODS}$" model,

TABLE II
SETS OF THE PARAMETERS USED FOR THE SIMULATION.

N_s (# of simulated trials)	N (# of neurons)	K (data length)	ω (weights)	r (shape parameter of NB)	σ (degrees of freedom)	γ (link function)
1, 10, 20, 50, 100	10, 20, 50	100, 500, 1000, 1500, 2000	(-1,+1)	1, 3, 5	50, 100, 200	1, 3, 5, 7, 9

with different parameter settings. For example, in Table III, each model is trained using a training set of length K and then tested by computing the MSE between the mean spike counts and the true value for a testing trial of same data length. The results are averaged across 50 randomly selected initializations of unknown parameters. The link function of the NB-GLM was selected as the $probit$ function, we have compared the performance of different link functions (results not shown) such as log, $logit$, $probit$, $identity$, and $log-log$, and found in most cases the $probit$ gives the best performance. For simplicity, we demonstrate results of standard NB-GLM with the $probit$ link function.

TABLE III
MEAN SQUARE ERROR COMPARISON OF THE NEGATIVE BINOMIAL GLM AND SODS MODEL FOR $N = 20$, $r = 5$, $\sigma = 50$, $\gamma = 7$.

NB-GLM / $\mathcal{SODS}$ Model	$K = 100$	$K = 500$	$K = 1000$	$K = 2000$
$N_s = 10$	1.432 **1.230**	0.497 **0.383**	0.450 **0.255**	0.380 **0.190**
$N_s = 50$	1.450 **0.731**	0.292 **0.204**	0.062 **0.052**	0.050 **0.047**
$N_s = 100$	0.640 **0.428**	0.056 **0.048**	0.049 **0.047**	0.032 **0.032**

TABLE IV
MEAN SQUARE ERROR COMPARISON OF THE NEGATIVE BINOMIAL GLM AND SODS MODEL FOR $N_s = 50$, $K = 500$, $r = 5$, $\sigma = 50$.

NB-GLM / $\mathcal{SODS}$ Model	$\gamma = 1$	$\gamma = 3$	$\gamma = 5$	$\gamma = 7$	$\gamma = 9$
$N = 10$	0.129 **0.090**	0.132 **0.090**	0.137 **0.092**	0.145 **0.103**	0.187 **0.173**
$N = 20$	0.234 **0.165**	0.265 **0.175**	0.274 **0.184**	0.292 **0.204**	0.354 **0.320**
$N = 50$	0.562 **0.407**	0.605 **0.421**	0.680 **0.430**	0.756 **0.450**	1.252 **0.950**

TABLE V
MEAN SQUARE ERROR COMPARISON OF THE NEGATIVE BINOMIAL GLM AND SODS MODEL FOR $N = 20$, $N_s = 50$, $K = 500$, $\gamma = 7$.

NB-GLM / $\mathcal{SODS}$ Model	$\sigma = 50$	$\sigma = 100$	$\sigma = 200$
$r = 1$	0.049 **0.021**	0.102 **0.080**	0.637 **0.492**
$r = 3$	0.224 **0.125**	0.265 **0.213**	0.774 **0.684**
$r = 5$	0.292 **0.204**	0.455 **0.321**	1.580 **0.912**

Comparing the configurations in Table III we find, as would be expected, increasing N_s and K gives a better estimation. When the number of training samples is large the two models have similarly good performances on the simulated data,

however, when the number of samples decreases the $\mathcal{SODS}$ outperforms standard Negative Binomial GLMs. Table IV shows that more neurons and larger γ decrease estimation accuracy, but $\mathcal{SODS}$ model still consistently outperforms NB-GLM. In Table V, it is expected to see the increased MSE with larger r and σ since larger r leads to more possible values of spike counts and larger σ leads to more degrees of freedom.

The left panel of Figure. 5 shows the MSE with error bars of $\mathcal{SODS}$ with different number of training trials (N_s) and data lengths (K). 10 training trials and 1000 data length already give relatively small MSE. The scatter plot in the right panel of Figure. 5 provides us with a clear view of the comparison between true and estimated mean spike counts.
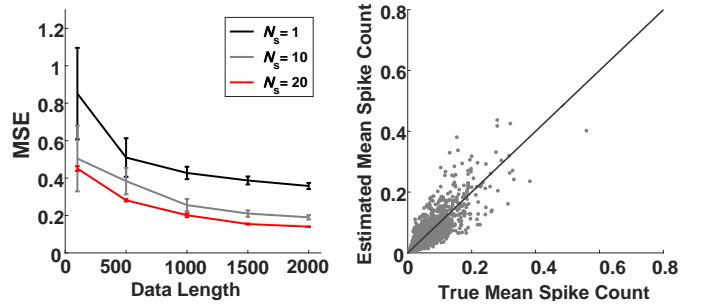

Fig. 5. Goodness-of-fit for simulated data - *left panel*: the mean square error across data lengths (K) for different numbers of trials ($N_s = \{1, 10, 20\}$); *right panel*: scatter plot of estimated mean spike counts and true mean spike counts with 1000 time bins. Each gray dot is one spike count value.

Next, we tested the accuracy of the estimation of the weights ω describing the directed effect of the input neurons on the output neuron. We explored the effect of different data lengths on the $BFGS$ method which was applied to maximize the log marginal likelihood. The results shown in Figure. 6 are taken for several different combinations of the parameter configurations in Table II, as we can see: (1) the relative standard error is large when the actual weights are close to 0; and (2) 1,000 bins is sufficient to provide accurate and efficient weight estimations, which is consistent with results in shown in the Figure. 5.

B. Experimental Data

The experimental data used here is taken from multi-unit recordings of retinal ganglion cells from the $\mathrm{ret}-1$ database [38], curated at $\mathrm{CRCNS.org}$. This database has single-unit neural responses recorded from isolated retina from mice using a 61-electrode array in response to various visual stimuli. It aims to learn how different visual stimuli influence the spiking activity of retina cells. For population activity, network models using GLM framework are quite
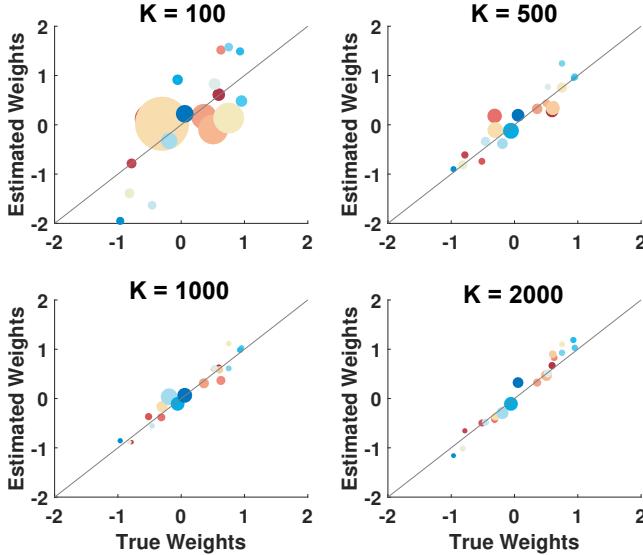
Fig. 6. Scatter plots of estimated weights of a 20-neuron network with different number of data points in each trial (total 10 trials, data length K = 100, 500, 1000, 2000). Different colors indicate different neurons. The size of the ball shows the relative standard error (standard error/mean value). The initial values of the weights to be optimized was selected randomly from -1 to 1 and simulated 20 times. 1000 bins is sufficient to achieve accurate and stable estimations of coupling weights inside neural network.

popular [2, 4, 7]. We test our framework with state-of-the-art methods on 4 datasets containing 37, 26, 15, and 14 neurons, respectively. The experimental data (spike counts) were binned into 16 ms. This bin size is a trade-off between how finely time is discretized and the computational costs. From each experimental dataset we created 5 random training and testing splits with all experimental trials, 4 splits were used for training and 1 split was used for testing with 5-fold cross validation performed. The training dataset was used to estimate the unknown parameters $r, \omega, \sigma, \gamma$, and θ_i^{SODS}. We then used the $\mathcal{SODS}$ model to compute the log-likelihoods of the held-out test data versus both the standard Poisson GLM and Negative Binomial GLM.

Figures 7 and 8 show the percentage increase in the log-likelihood of the $\mathcal{SODS}$ model over the Poisson GLM and Negative Binomial GLM for each of the datasets. We denote $\ell_{\mathrm{SODS}}, \ell_{\mathrm{NB}}, \ell_{\mathrm{Poisson}}$ as the predictive log-likelihoods of each model. The percentage log-likelihood increase is calculated by $\frac{\ell_{\mathrm{SODS}} - \ell_{\mathrm{NB}}}{|\ell_{\mathrm{NB}}|} \times 100\%$ and $\frac{\ell_{\mathrm{SODS}} - \ell_{\mathrm{Poisson}}}{|\ell_{\mathrm{Poisson}}|} \times 100\%$. In Figure 7 and 8, a positive percentage value indicates an improvement in the predictive log-likelihood for the held-out data of the $\mathcal{SODS}$ compared to the comparative method and conversely a negative value indicates a decrease in predictive log-likelihood. The majority of results (between 62% and 100% of neurons in each dataset) present higher prediction log-likelihoods for the test data when using the $\mathcal{SODS}$ model. Notably, the improvement offered by the $\mathcal{SODS}$ model is larger when compared with the Poisson GLM than when compared with the Negative Binomial GLM, indicating that the datasets analyzed do indeed exhibit over-dispersion property. However, when compared with the Negative Binomial GLM, the $\mathcal{SODS}$ still offered an increase in performance for a similar number of recorded neurons as for the Poisson GLM. This is despite
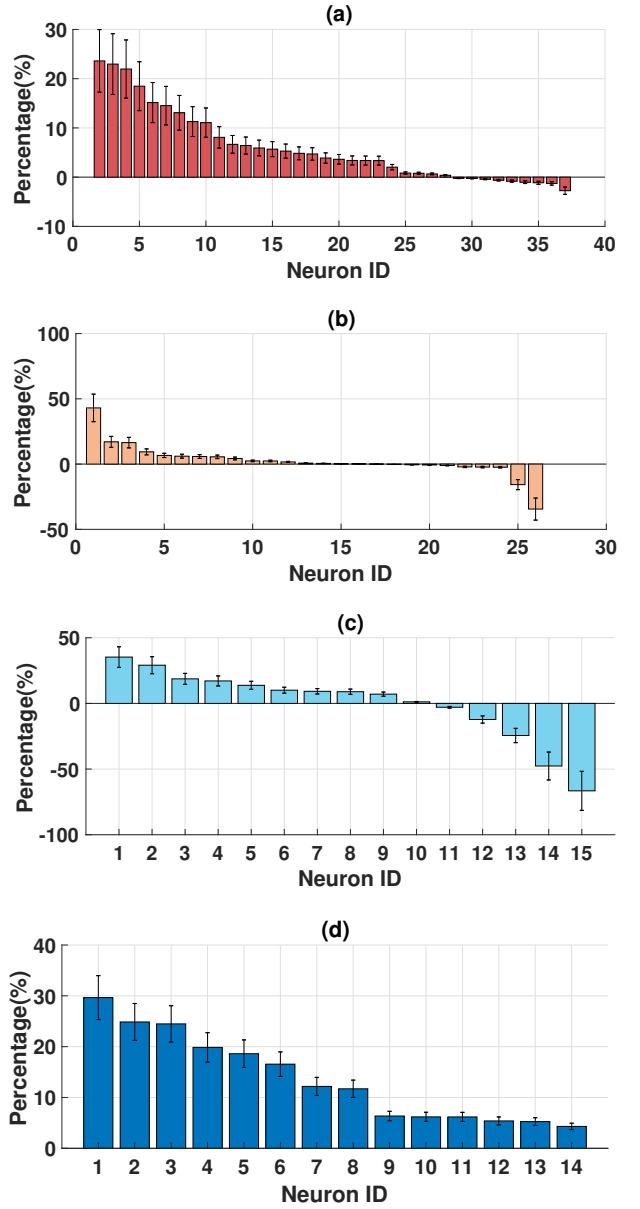

Fig. 7. $\mathcal{SODS}$ compared with Negative Binomial GLM: error bar of the percentage increase in predictive log likelihood (testing set) of all neurons in dataset (a) #62413 (37 neurons); (b) #62423 (26 neurons); (c) #62814 (15 neurons); (d) #62871 (14 neurons). The neuron ID in each dataset is sorted according to the improvements of prediction performance.

the Negative Binomial GLM having the same assumption on the observed spike counts as our $\mathcal{SODS}$ model.

Figure 9 shows the network weights estimated using $\mathcal{SODS}$ for two experimental datasets. Around 60% of total weights strength are positive ($\frac{|\omega_+|}{|\omega_+| + |\omega_-|}$) in #62814 dataset, while 63% of total weights strength are negative ($\frac{|\omega_-|}{|\omega_+| + |\omega_-|}$) in #62871 dataset. The results give valuable insights into the neural interactions among neurons: the kind of coupling weights dominates the whole neural connections (excitatory or inhibitory). The recovered directed weighted neural network provides a quantitative way to intuitively view the information flow under neural circuits. #62814 dataset show highly mutual excitations within the functional neural network, while #62871 is shown to have a inhibitory-dominated underlying
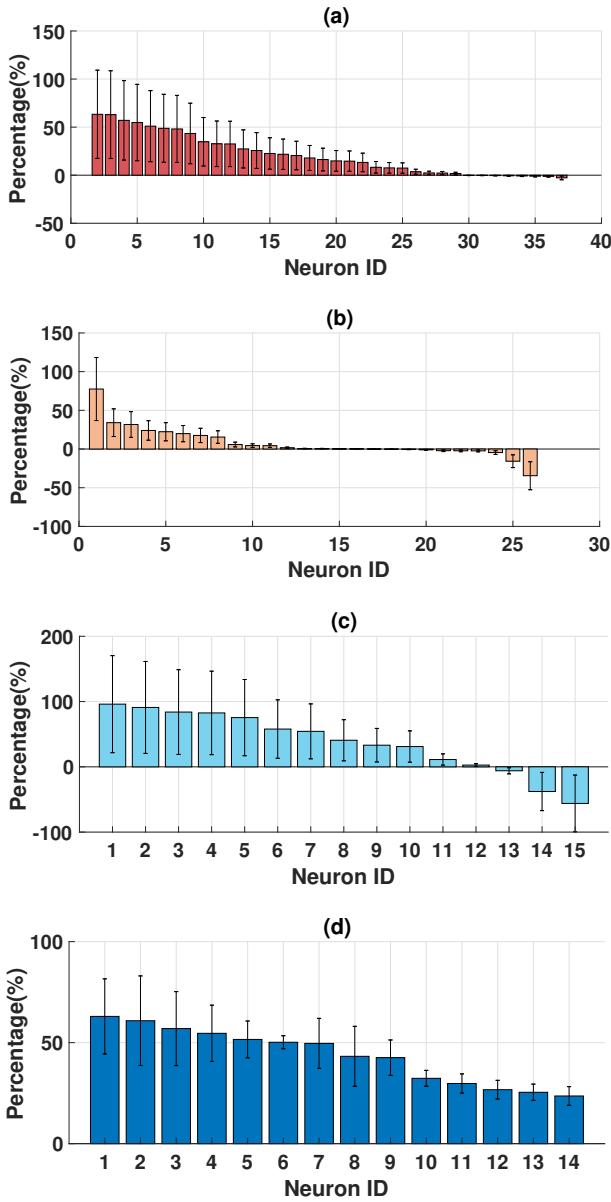
(a)

(b)

(c)

(d)

Fig. 8. $\mathcal{SODS}$ compared with the Poisson GLM: error bar of the percentage increase in predictive log likelihood (testing set) of all neurons in dataset **(a)** #62413 (37 neurons); **(b)** #62423 (26 neurons); **(c)** #62814 (15 neurons); **(d)** #62871 (14 neurons). The neuron ID in each dataset is sorted according to the improvements of prediction performance.

neural network.

VI. DISCUSSION AND CONCLUSION

The bio-signal processing community has shown great interest in multivariate regression methods [39–43]. These methods can provide a clear view of the nature of neuronal interactions. Linderman *et al.* [44] developed a fully Bayesian inference method for Negative Binomial responses that yields regularized estimations for all of the hyperparameters. Although it can have uncertainties (probability distributions) on all the parameters, applying fully Bayesian approaches to hierarchical models is computationally intensive. As an alternative, empirical Bayes can provide a bias-variance trade-off which can achieve a small mean square error at a lower computational
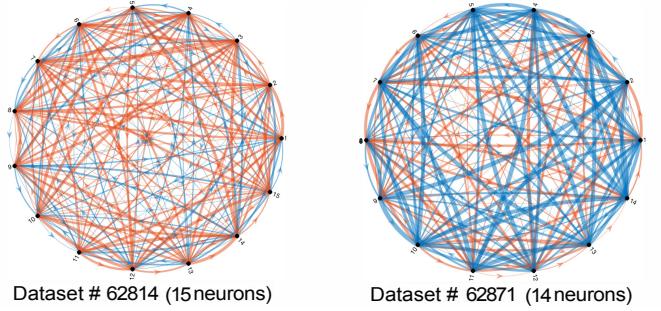

Fig. 9. Estimated network weights of two experimental datasets using our model. The neurons are laid in circle with each black dot indicating one neuron. Red/blue lines highlight the positive/negative weights, and the thickness indicates the strength of weights. Neural interactions recovered in dataset # 62814 (15 neurons) show an excitatory-dominated property; while the results from dataset # 62871 (14 neurons) indicate the underlying neural circuit is more inhibitory-dominated.

cost. To estimate the unknown parameters of the model, Paninski *et al.* [5] used maximum likelihood estimation, but when the dataset is small, the estimation becomes biased. *The "SODS" estimator developed here, to model over-dispersed spiking behavior and extract latent interactions among neural populations, combines both of the above methods.* It has the benefit of providing a bias-variance trade-off estimator for Negative Binomial responses, while not needing the intensive computation as fully Bayesian inference.

We take advantage of the beneficial properties of both GLMs and empirical Bayes inference to propose the "$\mathcal{SODS}$" estimator. We used the Negative Binomial distribution to model the spike counts of each neuron. The Negative Binomial distribution was selected as it allows for over-dispersed spike counts using a dispersion parameter superior to standard Poisson model. The beta distribution is employed as the prior information for the probability parameter in the Negative Binomial distribution, which allows for a closed-form posterior distribution. We propose a flexible link function family in order to model the prior mean using regressors. By using the recorded data from other neurons as the covariates, we can then infer the functional weights among the neural population. Unlike fully Bayesian inference, which utilizes hyperpriors, we instead estimate the hyperparameters by maximizing the marginal likelihood. The proposed "$\mathcal{SODS}$" estimator is a shrinkage estimator and the weights we estimate can be viewed as the hidden functional dependences. By taking the neurons as nodes in our functional neural network, and their spike-train data as the observations, our empirical Bayes inference method can be used to identify the neural interactions, including excitatory and inhibitory behaviors.

We have validated our method using both simulated data and experimental retinal neuron data. By using intensive simulations we have shown, that on our simulated system, the $\mathcal{SODS}$ outperforms the standard Negative Binomial GLM. Furthermore, the proposed approach implements a flexible link function, which is unlike the standard Negative Binomial GLM, allowing for selecting the best link for each dataset. From the simulation data we found that by efficiently maximizing the marginal likelihood, we can accurately estimate the model parameters. For the experimental data, when compared

with two of the most widely used regression methods: Poisson and Negative Binomial regressions, there was substantial improvement in the predictive log likelihood of the held-out data. While the results presented here are promising, going forward, we are interested in extending our model. For instance, the incorporation of Hebbian learning rules could account for time-varying weights. Applying prior knowledge regarding network structure, such as random, small world or scale-free networks, could also be a promising avenue for future research. Finally, the ability of our model to operate in data-limited cases would open possibilities for future applications to real-time settings, such as for closed loop experiments or improved brain-machine interface (BMI) devices. We implemented "$\mathcal{SODS}$" in Matlab (R2015b) and plan to release the code on Github shortly.

ACKNOWLEDGMENTS

REFERENCES

[1] I. M. Park, S. Seth, A. R. Paiva, L. Li, and J. C. Principe, "Kernel methods on spike train space for neuroscience: a tutorial," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 149–160, 2013.

[2] Z. Chen, D. F. Putrino, S. Ghosh, R. Barbieri, and E. N. Brown, "Statistical inference for assessing functional connectivity of neuronal ensembles with sparse spiking data," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 19, no. 2, pp. 121–135, 2011.

[3] M. Okatan, M. A. Wilson, and E. N. Brown, "Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity," *Neural Computation*, vol. 17, no. 9, pp. 1927–1961, 2005.

[4] D. Song, R. H. Chan, B. S. Robinson, V. Z. Marmarelis, I. Opris, R. E. Hampson, S. A. Deadwyler, and T. W. Berger, "Identification of functional synaptic plasticity from spiking activities using nonlinear dynamical modeling," *Journal of Neuroscience Methods*, vol. 244, pp. 123–135, 2015.

[5] L. Paninski, "Maximum likelihood estimation of cascade point-process neural encoding models," *Network: Computation in Neural Systems*, vol. 15, no. 4, pp. 243–262, 2004.

[6] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, "A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects," *Journal of Neurophysiology*, vol. 93, no. 2, pp. 1074–1089, 2005.

[7] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, "Spatio-temporal correlations and visual signalling in a complete neuronal population," *Nature*, vol. 454, no. 7207, pp. 995–999, 2008.

[8] A.-K. Seghouane and A. Shah, "Sparse estimation of the hemodynamic response functionin functional near infrared spectroscopy," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2074–2078.

[9] A. Kazemipour, M. Wu, and B. Babadi, "Robust estimation of self-exciting generalized linear models with application to neuronal modeling," *IEEE Transactions on Signal Processing*, vol. 65, no. 14, pp. 3733–3748, 2017.

[10] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 886–902, 1998.

[11] N. D. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," in *Advances in Neural Information Processing Systems*, no. 3, 2004, pp. 329–336.

[12] N. Lawrence, "Probabilistic non-linear principal component analysis with gaussian process latent variable models," *The Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.

[13] K. C. Lakshmanan, P. T. Sadtler, E. C. Tyler-Kabara, A. P. Batista, and M. Y. Byron, "Extracting low-dimensional latent structure from time series in the presence of delays," *Neural Computation*, vol. 27, no. 9, pp. 1825–1856, 2015.

[14] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan, "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 145–163, 2015.

[15] J. W. Pillow, L. Paninski, and E. P. Simoncelli, "Maximum likelihood estimation of a stochastic integrate-and-fire neural model." in *Advances in Neural Information Processing Systems*, 2003, pp. 1311–1318.

[16] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[17] C. Blundell, J. Beck, and K. A. Heller, "Modelling reciprocating relationships with hawkes processes," in *Advances in Neural Information Processing Systems*, 2012, pp. 2600–2608.

[18] M. G. Moore and M. A. Davenport, "A hawkes' eye view of network information flow," in *Statistical Signal Processing Workshop (SSP), 2016 IEEE*. IEEE, 2016, pp. 1–5.

[19] S. Chen, A. Shojaie, E. Shea-Brown, and D. Witten, "The multivariate hawkes process in high dimensions: Beyond mutual excitation," *arXiv preprint arXiv:1707.04928*, 2017.

[20] B. Mark, G. Raskutti, and R. Willett, "Network estimation from point process data," *arXiv preprint arXiv:1802.04838*, 2018.

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

[22] M. M. Churchland, M. Y. Byron, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, Corrado *et al.*, "Stimulus onset quenches neural variability: a widespread cortical phenomenon," *Nature Neuroscience*, vol. 13, no. 3, pp. 369–378, 2010.

[23] R. L. Goris, J. A. Movshon, and E. P. Simoncelli, "Partitioning neuronal variability," *Nature Neuroscience*, vol. 17, no. 6, pp. 858–865, 2014.

[24] J. F. Lawless, "Negative binomial and mixed poisson regression," *Canadian Journal of Statistics*, vol. 15, no. 3, pp. 209–225, 1987.

[25] B. Chen *et al.*, "Mean-value deviance detection of transient signals modeled as overdispersed dft data," *IEEE Transaction on Signal Processing*, 1998.

[26] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple neural spike train data analysis: state-of-the-art and future challenges," *Nature Neuroscience*, vol. 7, no. 5, p. 456, 2004.

[27] M. E. Spira and A. Hai, "Multi-electrode array technologies for neuroscience and cardiology," *Nature Nanotechnology*, vol. 8, no. 2, pp. 83–94, 2013.

[28] M. D. Mauk and D. V. Buonomano, "The neural basis of temporal processing," *Annual Review of Neuroscience*, vol. 27, pp. 307–340, 2004.

[29] N. Bertschinger and T. Natschläger, "Real-time computation

at the edge of chaos in recurrent neural networks," *Neural Computation*, vol. 16, no. 7, pp. 1413–1436, 2004.

[30] Q. She, G. Chen, and R. H. Chan, "Evaluating the small-world-ness of a sampled network: Functional connectivity of entorhinal-hippocampal circuitry," *Scientific Reports*, vol. 6, 2016.

[31] A. S. Charles, M. Park, J. P. Weller, G. D. Horwitz, and J. W. Pillow, "Dethroning the fano factor: a flexible, model-based approach to partitioning neural variability," *bioRxiv*, 2017. [Online]. Available: http://www.biorxiv.org/content/early/2017/07/19/165670

[32] J. Scott and J. W. Pillow, "Fully bayesian inference for neural models with negative-binomial spiking," in *Advances in Neural Information Processing Systems*, 2012, pp. 1898–1906.

[33] J. A. Nelder and R. J. Baker, *Generalized linear models*. Wiley Online Library, 1972.

[34] D. F. Shanno, "On broyden-fletcher-goldfarb-shanno method," *Journal of Optimization Theory and Applications*, vol. 46, no. 1, pp. 87–94, 1985.

[35] Y. Taniai and J. Nishii, "Optimality of upper-arm reaching trajectories based on the expected value of the metabolic energy cost," *Neural Computation*, vol. 27, no. 8, pp. 1721–1737, 2015.

[36] N. Gillis and F. Glineur, "Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization," *Neural Computation*, vol. 24, no. 4, pp. 1085–1105, 2012.

[37] P. E. Gill, W. Murray, and M. A. Saunders, "Snopt: An sqp algorithm for large-scale constrained optimization," *SIAM review*, vol. 47, no. 1, pp. 99–131, 2005.

[38] J. L. Lefebvre, Y. Zhang, M. Meister, X. Wang, and J. R. Sanes, "γ-protocadherins regulate neuronal survival but are dispensable for circuit formation in retina," *Development*, vol. 135, no. 24, pp. 4141–4151, 2008.

[39] D. F. Schmidt and E. Makalic, "Estimating the order of an autoregressive model using normalized maximum likelihood," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 479–487, 2011.

[40] C. D. Giurcăneanu and F. A. A. Saip, "New insights on ar order selection with information theoretic criteria based on localized estimators," *Digital Signal Processing*, vol. 32, pp. 37–47, 2014.

[41] W. Dai, H. Xiong, J. Wang, S. Cheng, and Y. F. Zheng, "Generalized context modeling with multi-directional structuring and mdl-based model selection for heterogeneous data compression," *IEEE Transactions on Signal Processing*, vol. 63, no. 21, pp. 5650–5664, 2015.

[42] R. M. Rangayyan, *Biomedical signal analysis*. John Wiley & Sons, 2015, vol. 33.

[43] A. Sheikhattar, J. B. Fritz, S. A. Shamma, and B. Babadi, "Recursive sparse point process regression with application to spectrotemporal receptive field plasticity analysis," *IEEE Transactions on Signal Processing*, vol. 64, no. 8, pp. 2026–2039, 2016.

[44] S. W. Linderman and R. P. Adams, "Discovering latent network structure in point process data," in *International Conference on Machine Learning*, 2014, pp. 1413–1421.

Qi She (M'17) received the B.E. degree in 2014 from Nanjing University of Posts and Telecommunications within Top 1% Ranking (2/230). Currently, he is a Ph.D. candidate at Department of Electronic Engineering, College of Science and Engineering, City University of Hong Kong. During 2017-2018, he is a Visiting Student Research Collaborator (VSRC) in Pillow lab of Princeton Neuroscience Institute, Princeton University. His research focuses on statistical signal processing methods to extract hidden structure from high-dimensional spike-train data and infer brain connectivity using Bayesian framework. He is interested in studying how information is encoded, decoded in the brain, especially using latent variable models for modeling neural dynamics.

Beth Jelfs is currently the recipient of a Vice-Chancellors Research Fellowship at RMIT University, Australia. She graduated from the University of Leicester, UK with a MEng (1st Hons.) in Electronic & Software Engineering, receiving the British Computer Society prize for top graduate 2005, and from Imperial College London, UK with a Ph.D. in Electrical and Electronic Engineering in 2010. She has previously held research positions in the Department of Electronic Engineering at City University of Hong Kong, Hong Kong, the Department of Medical Physics and Bioengineering at University College London, UK and the Departments of Chemistry and Physics at University of Oxford, UK. Her current research interests include statistical and adaptive signal processing, machine learning and signal modality characterization and in particular their application to biomedical data.

Adam S. Charles (M'15) received both a B.E. and M.E in Electrical and Computer Engineering in 2009 from The Cooper Union in New York City, New York. He received his Ph.D. in Electrical and Computer Engineering from The Georgia Institute of Technology in 2015, where his research was awarded a Sigma Xi Best Doctoral Thesis award as well as an Electrical and Computer Engineering Research Excellence award. Currently, Dr. Charles is a post-doctoral fellow at the Princeton Neuroscience Institute in Princeton New Jersey, where he studies theoretical and computational neuroscience. His research interests currently include neural imaging technologies, inference of sparse and structured signals, and statistical models of neural systems.

Rosa H.M. Chan (M'01-SM'17) is currently an Associate Professor in the Department of Electronic Engineering at City University of Hong Kong. She received the B. Eng. (1st Hon.) degree in Automation and Computer-Aided Engineering from The Chinese University of Hong Kong in 2003. She was later awarded the Croucher Scholarship and Sir Edward Youde Memorial Fellowship for Overseas Studies in 2004. She received her Ph. D. degree in Biomedical Engineering in 2011 from the University of Southern California (USC), where she also received her M. S. degrees in Biomedical Engineering, Electrical Engineering, and Aerospace Engineering. Her research interests include computational neuroscience, neural prosthesis, and brain-computer interface applications. She was the co-recipient of the Outstanding Paper Award of IEEE Transactions on Neural Systems and Rehabilitation Engineering in 2013, for their research breakthroughs in mathematical modelling for cognitive prosthesis. Dr. Chan was the Chair of the Hong Kong-Macau Joint Chapter of IEEE Engineering in Medicine and Biology Society (EMBS) in 2014 and is elected to the IEEE EMBS AdCom as Asia Pacific Representative (2018-2020).