

A Computationally Efficient Method for Modeling Neural Spiking Activity with Point Processes Nonparametrically

Todd P. Coleman and Sridevi Sarma

March 5, 2007

Abstract

Point process models have been shown to be useful in characterizing neural spiking activity (NSA) as a function of extrinsic and intrinsic factors. Most point process models of NSA are parametric as they are often efficiently computable. However, if the actual point process does not lie in the assumed parametric class of functions, misleading inferences can arise. Nonparametric methods are attractive due to fewer assumptions, but computation grows with the size of the data. We propose a computationally efficient method for nonparametric maximum likelihood estimation when the conditional intensity function, which characterizes the point process in its entirety, is assumed to be a Lipschitz continuous function but otherwise arbitrary. We show that by exploiting much structure, the problem becomes efficiently solvable and we compare our nonparametric estimation method to the most commonly used parametric approaches on goldfish retinal ganglion neural data. In this example, our nonparametric method gives a superior absolute goodness-of-fit measure than all parametric approaches analyzed.

1 Introduction

Characterizing neural spiking activity as a function of environmental stimuli, and intrinsic effects such as a neuron's own spiking history and concurrent ensemble activity is important in neuroscience. Such a characterization is complex and there is increasing need for a broad class of models to capture such details. Point process models have been shown to be useful in characterizing neural spiking activity. For example, they have successfully characterized stimulus-response relationships computed from data recorded in hippocampal cells of a rat [1], retinal ganglion cells in salamander, rabbit, and cat, as well as from lateral geniculate nucleus neurons in cat [15]. These models also captured the dynamics of supplementary eye field cells of a macaque monkey [14], the influence of one neuron's spiking activity on another's in sea hare (*Aplysia californica*) [4], and ensemble activity in hippocampal cells of rats [11].

Most point process models developed to date are parametric. That is, the conditional intensity function is assumed to belong to a class of parametric functions. Parametric models have several advantages. First, they are often efficiently computable. Second, the parameters may be related back to physiological and environmental factors. Finally, they have nice asymptotic properties when the actual distribution lies in the assumed parametric class. However, if the true conditional intensity function does not lie in the assumed class, large errors may occur resulting in misleading inferences. In contrast, there are very few computable nonparametric methods for estimating the conditional intensity function of a point process model. Furthermore, even if a method is computable, it is very difficult to show nice asymptotic properties of the estimate. Yet, nonparametric estimation methods have the advantage of not assuming that the conditional intensity function lies in some known class of functions.

We propose a computationally efficient method to build nonparametric point process models, which only involves computing a convex optimization problem subject to linear constraints. We show that point process estimation is a simple application of a nonparametric nonlinear regression problem [5]. The only restriction on the conditional intensity function is that its logarithm must be Lipschitz continuous.

Although this method is nonparametric, it does require an assumption of the Lipschitz continuity constant. Enforcing the constant to be too small results in over-constraining the estimate of the function, while making the constant too large results in overfitting the data. Thus, model selection [2] is fundamentally required to develop an estimate of the constant based upon the data. We use the time-rescaling theorem [10, 8] and the Kolmogorov-Smirnov statistic [13] as a principled method to perform model selection in this scenario. We compare our nonparametric estimation method to the most commonly used parametric approaches on goldfish retinal ganglion neural data. In this example, our nonparametric method gives a superior absolute goodness-of-fit measure than all parametric approaches analyzed.

2 Methods

2.1 Point Process Models

A point process is a stochastic process that describes a collection of binary events in a given continuum (eg. number of bus arrivals, number of neural spikes in some time interval), and is characterized entirely by the conditional intensity function (CIF) defined below.

Consider the time interval $(0, T]$ as the continuum, and events as neuronal spike times. Let t_1, t_2, \dots, t_S denote the times of each neural spike such that $0 < T_1 < t_2 \dots < t_S \leq T$. Then, if $N(t)$ is the sample path of the associated counting process (i.e. $N(t) \in \mathbb{Z}_+ \forall t, N(0) = 0$ and $N(t) \geq N(\tau)$ for $t \geq \tau$), the conditional intensity function is defined as follows:

$$\lambda(t|X_t) \triangleq \lim_{\Delta \rightarrow 0} \frac{\text{Prob}(N(t, t+\Delta)=1|X_t)}{\Delta}$$

$$\Leftrightarrow \text{Prob}(N(t, t+\Delta) = 1|X_t) \simeq \lambda(t|X_t) \Delta,$$

where X_t is a vector of real-values in general whose components carry information about intrinsic and extrinsic factors that impact the probability that a neuron fires at time t . Also note that for $(a, b]$, $N(a, b) \triangleq N(b) - N(a)$ is the number of counts in the interval $(a, b]$. See [10] for more on point process theory.

One can also compute the inter-spike interval (ISI) probability density at time t , $p(t)$, in terms of the conditional intensity function and the previous spike time t_{prev} as follows

$$p(t) = \lambda(t|X_t) \exp\left\{-\int_{t_{prev}}^t \lambda(u|X_u) du\right\}.$$

In turn, the CIF can also be written in terms of $p(t)$ as follows

$$\lambda(t|X_t) = \frac{p(t)}{1 - \int_0^t p(u) du}. \quad (1)$$

When modeling neural spike train data as a point process, one is ultimately interested in estimating either the conditional intensity function or the ISI distribution. As mentioned in the introduction, most point process models of NSA have been parametric. We highlight commonly used models below.

The *exponential*, *gamma*, and *inverse Gaussian* distributions have been used to model ISI densities for neural spiking activity [9].

- *Exponential*: the ISI densities are i.i.d. with distribution $p(t) = \lambda e^{-\lambda t}$ $t \geq 0$, which comes from a simple stochastic model that the neuron spike times constitute a *Poisson* arrival process (a special point process) with constant CIF or rate function λ .
- *Gamma*: the ISI densities are i.i.d. gamma random variables with parameters $\beta > 0$ and $\alpha > 0$. That is,

$$p(t|\beta, \alpha) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} \exp -\beta t.$$

This is the ISI probability model obtained from a simple stochastic integrate-and-fire model in which the inputs to the neuron are Poisson with constant rate β , and α depends on the firing threshold of the neuron. One can show, via transformation (1) that the corresponding conditional intensity is a function of time and parameters (β, α) . For details refer to [9].

- *Inverse Gaussian*: Finally, if one models neural spiking activity as a stochastic integrate-and-fire model in which the membrane voltage of the neuron is represented as a random walk with drift [19], then the resulting ISI densities are i.i.d. inverse Gaussian. That is

$$p(t|\mu, \eta) = \left(\frac{\eta}{2\pi t^3}\right)^{\frac{1}{2}} \exp\left\{\frac{-\eta(t-\mu)^2}{2\mu^2 t}\right\},$$

where the positive real parameters (μ, η) depend on the neuron's properties. The associated CIF can also be computed using (1) and is a function of (μ, η) and time.

Since the gamma and inverse Gaussian ISI probability densities can be derived from elementary stochastic integrate-and-fire models, these distributions seem more plausible than the exponential in modeling neural spiking activity [9]. Also note that fitting the above models with spike train data requires construction of the likelihoods as a function of the parameters and then estimating the parameters via some optimization algorithm. See [8, 9] for construction of conditional intensity functions and likelihoods for the above ISI distributions.

The *generalized linear model* (GLM), has recently been used to model conditional intensity functions of point processes. In its simplest form, the GLM represents the conditional intensity function parametrically as follows:

$$\log(\lambda(t|X_t)) = \sum_i \alpha_i g_i^{ext}(t) + \beta_i g_i^{int}(t),$$

where $g_i^{ext}(t)$ and $g_i^{int}(t)$ are known basis functions that capture the extrinsic and intrinsic effects (carried in X_t) on the probability of spiking at time t , respectively. The extrinsic and intrinsic associated parameters to be estimated are $\alpha_i, \beta_i \in \mathbb{R}$, respectively.

2.1.1 The Likelihood Function

As mentioned above, most estimates of the CIF and ISI density are parametric, and the parameters are estimated via algorithms such as maximum likelihood (ML) or method of moments [8]. In this paper, we compute ML estimates for all models generated.

For any point process with conditional intensity function $\lambda(t|X_t)$ and arrival times $\{t_1, t_2, \dots, t_S\}$ in $[0, T]$, it can be shown [8] that the likelihood $\phi(\{\lambda(t|X_t)\})$ can be expressed as

$$\begin{aligned} \phi(\{\lambda(t|X_t)\}) &= f_{T_1, \dots, T_S}(t_1, \dots, t_S) \\ &= \left[\prod_{j=1}^S \lambda(t_j|X_{t_j}) \right] \exp \left\{ - \int_0^T \lambda(t|X_t) dt \right\} \\ &= \exp \left\{ \int_0^T \log \lambda(t|X_t) dN(t) - \lambda(t|X_t) dt \right\}. \end{aligned}$$

By discretizing $[0, T]$ into n small discrete length Δ intervals so that

$$dN_i \triangleq N((i+1)\Delta) - N(i\Delta) \in \{0, 1\},$$

we get that the likelihood function of spike observations and discrete samples of the CIF is

$$\sum_{i=1}^n -dN_i \lambda_i(X_i) + \sum_{i=1}^n \lambda_i(X_i) \Delta. \quad (2)$$

where $\lambda_i(X_i) = \lambda(t|X_t)$ for $t = i\Delta$, $i = 1, 2, \dots, n$.

2.1.2 Goodness of Fit Test

Before making inferences from the data, one must first check the validity of the point process model. The **time-rescaling theorem** can be used to develop goodness-of-fit tests for point process models and is stated below.

Theorem 2.1. *By defining $T_0 = 0$ and for $i = 1, \dots, S$:*

$$\begin{aligned}\Lambda(t) &= \int_0^t \lambda(\tau|X_\tau) d\tau \\ \Lambda_k &= \Lambda(T_k) = \int_0^{T_k} \lambda(t|X_t) dt \\ Z_k &= \Lambda_k - \Lambda_{k-1} = \int_{T_{k-1}}^{T_k} \lambda(t|X_t) dt,\end{aligned}\tag{3}$$

if we observe the S arrival times $\{T_k\}$ of occurrence of a point process with conditional intensity function $\lambda(t|X_t)$, then $\{\Lambda_k\}$ are arrival times of a unit-rate Poisson process. Equivalently, the time-rescaled inter-arrival times, or inter-spike-intervals (ISIs), $\{Z_k\}$, given by (3), are i.i.d. unit-rate exponentials.

Various proofs of the time rescaling theorem are given in [16, 18, 10, 7]

To compute how well a model $\hat{\lambda}(t|X_t)$ fits spike train sample $t = (t_1, \dots, t_S)$, we first generate time-rescaled ISIs $Z_k = \Lambda_k - \Lambda_{k-1}$, which, if $\hat{\lambda} = \lambda$, are i.i.d unit-rate exponential random variables. Then, the Z_k 's are transformed to $u_k = 1 - \exp(-Z_k)$, which, if $\hat{\lambda} = \lambda$, are i.i.d. uniform $[0, 1]$ random variables. To assess goodness-of-fit of $\hat{\lambda}(t|X_t)$, we compute the Kolmogorov-Smirnov (KS) statistic to assess agreement between the u_k 's and a uniform distribution [13]. The KS statistic measures the difference between the empirical distribution generated from the u_k 's ($F_{empirical}(t)$) and the cumulative distribution of a uniform random variable ($F_{uniform}(t) = t$). Specifically, the KS statistic is

$$\max_{t \in [0,1]} |F_{empirical}(t) - F_{uniform}(t)|.$$

KS plots, which plot $F_{uniform}(t)$ versus $F_{empirical}(t)$, are used to visualize the goodness-of-fit for each model. The model is better if its corresponding KS plot lies on the 45 degree line. In addition, one can compute 95% confidence bounds for the degree of agreement using the distribution of the KS statistic [13].

We use KS statistic to test absolute goodness of fit for all point process models computed in our examples.

2.1.3 Computing Uncertainty

Once a point process model is estimated and an acceptable goodness-of-fit is established, one would like to make reliable inferences from the data about the underlying neural process being studied. Basic questions of inference require characterizing uncertainty in either the estimate itself or a statistic computed from the estimate (eg. its mean or standard error).

Uncertainty can be computed exactly by repeating the experiment and building a model an infinite number of times. However, in practice, only a finite data set is available and uncertainty must be approximated. There are methods to compute approximations analytically, which include Propagation of Errors Method (Lindberg, 2000), and the Delta Method (Green, 2003). The general method of getting formulas for propagating errors involves computing the total differential of a function, which may be difficult to compute. The delta method derives an approximate probability distribution for a given function of an estimate, but requires knowledge of the limiting distribution of that estimate. If uncertainty cannot be characterized analytically, then bootstrap algorithms can be used to numerically approximate uncertainty (Davison and Hinkley, 1997). However, bootstrapping typically relies on the strong assumption that the data samples (spike times) are i.i.d. If the conditional intensity function of a point process model does not depend on the history of the process, then the arrival times are independent events. However, if one is modeling neural spike trains, then in general each spike time is dependent on the times of the previous spikes, which makes bootstrapping more challenging.

In [17], a new bootstrap technique applicable to any point process model that does not require any distribution approximations or additional assumptions on the data was introduced. The method applies the time-rescaling theorem to represent the original spike-time samples as a set of i.i.d. samples, and then implements one of two bootstrap techniques to the transformed data. When bootstrap samples are generated, they are transformed back into spike arrival time samples, and then bootstrap replications of the estimate of conditional intensity function and/or statistics are computed. We use the parametric bootstrap method described in [17] to compute uncertainties in our point process model estimates.

2.2 The Dynamic Maximum Likelihood Estimation Problem

We consider a nonlinear nonparametric regression problem, where we would like to infer the structure of an unknown nonlinear function z , only assuming that it is Lipschitz continuous and not assuming any a priori parametric structure. In particular, we would like to do this within the context of point process observations where $z = \log \lambda(X)$.

Recall that a function $z : \mathbb{R}^m \rightarrow \mathbb{R}$ is *Lipschitz continuous* with parameter $K > 0$ if

$$|z(X_1) - z(X_2)| \leq K \|X_1 - X_2\|_\infty$$

holds for all X_1, X_2 .

We consider the likelihood function of a point process and follow the methodology of [6] (section 3) to get the following Dynamic Maximum Likelihood Estimation (DMLE) problem:

$$\begin{aligned}
& \min -\log \phi \left(\left\{ \hat{\lambda}(t|X_t) \right\} \right) \\
& \text{s.t. } \left| \log \hat{\lambda}_{j_1} - \log \hat{\lambda}_{j_2} \right| \leq K \|X_{j_1} - X_{j_2}\|_\infty, \\
& \quad j_1, j_2 = 1, \dots, n
\end{aligned}$$

which by way of (2) corresponds to

$$\begin{aligned}
& \min \quad \left[\sum_{i=1}^n -dN_i \log \hat{\lambda}_i \right] + \sum_{i=1}^n \hat{\lambda}_i \Delta \\
& \text{s.t. } \quad \left| \log \hat{\lambda}_{j_1} - \log \hat{\lambda}_{j_2} \right| \leq K \|X_{j_1} - X_{j_2}\|_\infty, \\
& \quad \quad \quad j_1, j_2 = 1, \dots, n.
\end{aligned} \tag{4}$$

Note that we are minimizing a convex function over a convex set.

2.2.1 Computational Reduction of the DMLE

Now we uncover a computationally efficient method to solve the DMLE problem (4) by exploiting duality, separable structure, and network flow problems.

By introducing $z_i = \log \lambda_i(X_i)$, *lipschitzconstraint* = $K \|X_{j_1} - X_{j_2}\|_\infty$ and the set $J = \{(j_1, j_2) : j_1 < j_2, j_2 = 1, 2, \dots, n\}$ we have:

$$\min \quad \sum_{i=1}^n -N_i z_i + \Delta \exp(z_i) \tag{5a}$$

$$\text{s.t. } \quad z_{j_1} - z_{j_2} \leq C_{j_1, j_2} \quad j \in J. \tag{5b}$$

We now exploit the special structure of this problem. Note from Section 5.1.6 in [3] that this problem is in what is called as ‘separable’ form:

$$\begin{aligned}
& \min \quad \sum_{i=1}^n f_i(z_i) \\
& \text{s.t. } \quad \sum_{i=1}^n g_{i,j}(z_i) \leq 0, \quad (j_1, j_2) \in J
\end{aligned}$$

where the equivalence is made by the following

- From (5a) $f_i(z_i) = -N_i z_i + \Delta \exp(z_i)$.
- From (5b), for $j = (j_1, j_2) \in J$

$$g_{i,j}(z_i) = \begin{cases} z_i - \frac{1}{2}C_{i,j_2} & \text{if } i = j_1 \\ -z_i - \frac{1}{2}C_{j_1,i} & \text{if } i = j_2 \\ 0 & \text{otherwise.} \end{cases}$$

This class of problems has a special structure that allows for simplified analysis and separation. This can be seen by taking a look at the dual. We can define the Lagrangian by

$$\begin{aligned} L(z, \mu) &= \sum_{i=1}^n \left[f_i(z_i) + \sum_{j \in J} \mu_j g_{i,j}(z_i) \right] \\ &= \sum_{i=1}^n L_i(z_i, \mu), \end{aligned}$$

where

$$\begin{aligned} L_i(z_i, \mu) &\triangleq f_i(z_i) + \sum_{j \in J} \mu_j g_{i,j}(z_i) \\ &= -dN_i z_i + \Delta \exp(z_i) + \left[\sum_{j=(i,j_2)} \mu_j \left(z_i - \frac{1}{2} C_{i,j_2} \right) \right] \\ &\quad + \left[\sum_{j=(j_1,i)} \mu_j \left(-z_i - \frac{1}{2} C_{i,j_2} \right) \right] \\ &= -dN_i z_i + \Delta \exp(z_i) + \left[\sum_{j_2=1}^n \mu_{i,j_2} \left(z_i - \frac{1}{2} C_{i,j_2} \right) \right] \\ &\quad + \left[\sum_{j_1=1}^n \mu_{j_1,i} \left(-z_i - \frac{1}{2} C_{i,j_2} \right) \right] \end{aligned} \tag{6}$$

(7)

$$\begin{aligned} q(\mu) &= \inf_{z_i, i=1, \dots, n} \{L(z, \mu)\} = \sum_{i=1}^n \inf_{z_i} \{L_i(z_i, \mu)\} \\ &= \sum_{i=1}^n q_i(\mu). \end{aligned}$$

Let $q_i(\mu) \triangleq \inf_{z_i} \{L_i(z_i, \mu)\}$, and our dual problem of $\max_{\mu} q(\mu)$ such that $\mu \geq 0$ can be analyzed in more depth by looking at the structure of each $q_i(\mu)$.

Now we exploit the separability and calculate the closed-form expression for each $q_i(\mu)$. Note that the Lagrangian $L_i(z_i, \mu)$ given by (6) satisfies $\frac{\partial L_i(z_i, \mu)}{\partial z_i} = -dN_i + \Delta \exp(z_i) + \sum_{j_2} \mu_{i,j_2} - \sum_{j_1} \mu_{j_1,i}$.

So by defining $W_i(\mu) \triangleq dN_i + \sum_{j_1} \mu_{j_1,i} - \sum_{j_2} \mu_{i,j_2}$ and taking $\frac{\partial L_i(z_i, \mu)}{\partial z_i} = 0$, we have $\Delta \exp(z_i^*) = W_i(\mu)$, which implies that $z_i^* = -\log \Delta + \log W_i(\mu)$. Note that $W_i(\mu) \geq 0$ for all i, μ . We can now express $q_i(\mu)$ as the Lagrangian evaluated at z_i^* :

$$\begin{aligned}
q_i(\mu) &= -dN_i z_i^* + \Delta \exp(z_i^*) + \left[\sum_{j_2} \mu_{i,j_2} \left(z_i^* - \frac{1}{2} C_{i,j_2} \right) \right] + \\
&\quad + \left[\sum_{j_1} \mu_{j_1,i} \left(-z_i^* - \frac{1}{2} C_{i,j_2} \right) \right] \\
&= \Delta \exp(z_i^*) - z_i^* \left(dN_i + \left[\sum_{j_1} \mu_{j_1,i} - \sum_{j_2} \mu_{i,j_2} \right] \right) - \\
&\quad \sum_{j_2} \frac{1}{2} \mu_{i,j_2} C_{i,j_2} - \sum_{j_1} \frac{1}{2} \mu_{j_1,i} C_{j_1,i} \\
&= W_i(\mu) - [-\log \Delta + \log W_i(\mu)] W_i(\mu) - \\
&\quad \sum_{j_2} \frac{1}{2} \mu_{i,j_2} C_{i,j_2} - \sum_{j_1} \frac{1}{2} \mu_{j_1,i} C_{j_1,i} \\
&= (1 + \log \Delta) W_i(\mu) - W_i(\mu) \log W_i(\mu) \\
&\quad - \sum_{j_2} \frac{1}{2} \mu_{i,j_2} C_{i,j_2} - \sum_{j_1} \frac{1}{2} \mu_{j_1,i} C_{j_1,i}.
\end{aligned}$$

We simplify the expression for $q(\mu) = \sum_{i=1}^n q_i(\mu)$. It follows from the definition of $W_i(\mu)$ that (1) $\sum_{i=1}^n W_i(\mu) = \sum_{i=1}^n dN_i$ and (2) the appearance of the term $-\frac{1}{2} C_{j_1,j_2} \mu_{j_1,j_2}$ in $q(\mu)$ appears twice, once for $q_{j_1}(\mu)$ and once for $q_{j_2}(\mu)$.

Thus, $q(\mu) = [\sum_{i=1}^n (1 + \log \Delta) dN_i - W_i(\mu) \log W_i(\mu)] + [\sum_{j_1} \sum_{j_2} -C_{j_1,j_2} \mu_{j_1,j_2}]$, and the bottom line is that performing our nonparametric regression is equivalent to solving the following dual optimization problem.

DUAL PROBLEM:

$$\begin{aligned} \max q(\mu) \\ \text{s.t. } \mu \geq 0 \end{aligned}$$

where $\mu = (\mu_{j_1, j_2})$ has n^2 components and the function q is specified as

$$q(\mu) = \left[\sum_{i=1}^n (1 + \log \Delta) dN_i - W_i(\mu) \log W_i(\mu) \right] + \left[\sum_{j_1} \sum_{j_2} -C_{j_1, j_2} \mu_{j_1, j_2} \right]$$

where

$$\begin{aligned} W_i(\mu) &\triangleq dN_i + \sum_{j_1} \mu_{j_1, i} - \sum_{j_2} \mu_{i, j_2} \\ C_{j_1, j_2} &\triangleq K \|X_{j_1} - X_{j_2}\|_{\infty}. \end{aligned}$$

The optimal solution to the primal, given the dual solution, is given in closed form simply as

$$\begin{aligned} z_i^* &= -\log \Delta + \log W_i(\mu^*) \\ \Rightarrow \hat{\lambda}_i^* &= \frac{1}{\Delta} W_i(\mu^*) \end{aligned}$$

It turns out that we can cut computational costs by looking more carefully at the dual problem. Consider, for a fixed W , the set $S(W) = \{\mu : \mu \geq 0, W_i - dN_i = \sum_{j_1} \mu_{j_1, i} - \sum_{j_2} \mu_{i, j_2}, i = 1, \dots, n\}$, where $W_i \geq 0, i = 1, \dots, n$ and $\sum_i W_i = \sum_i dN_i$. This can be described succinctly by saying that

$$\begin{aligned} A\mu &= b, \\ \mu &\geq 0 \end{aligned}$$

where A is a matrix consisting of 0, 1 and -1 and entries, with n rows and n^2 columns; and $b = W - dN$ is the column constraint matrix. Note that

$$\sum_{i=1}^n b_i = \sum_{i=1}^n (W_i - dN_i) = \left(\sum_{i=1}^n W_i \right) - \left(\sum_{i=1}^n dN_i \right) = 0. \quad (8)$$

Therefore, we are fixing the values of W and varying the remaining free parameters μ , consistent with this fixed W , to optimize $q(\mu)$. Note that maximizing $q(\mu)$ subject to this fixed value of W is equivalent to performing the following linear program (LP):

$$\begin{aligned} \min \quad & \sum_{j_1} \sum_{j_2} C_{j_1, j_2} \mu_{j_1, j_2} \\ \text{s.t.} \quad & A\mu = b \\ & \mu \geq 0 \end{aligned}$$

COMPUTATIONALLY EFFICIENT DUAL PROBLEM:

Perform

$$\begin{aligned} \min \quad & \tilde{q}(W) \\ \text{s.t.} \quad & W \geq 0, \\ & \sum_{i=1}^n W_i = \sum_{i=1}^n dN_i \end{aligned}$$

where the function \tilde{q} is specified as $\tilde{q}(W) \triangleq \sum_{i=1}^n W_i \log W_i + R(W)$,

$$\begin{aligned} R(W) &= \min C_{j_1, j_2} \mu_{j_1, j_2} \\ \text{s.t.} \quad & A\mu = b \\ & \mu \geq 0 \\ & = \max \sum_i^n p_i(W_i - dN_i) \\ \text{s.t.} \quad & p_{j_1} - p_{j_2} \leq C_{j_1, j_2}, \quad j_1, j_2 = 1, \dots, n \end{aligned}$$

This can efficiently be solved by performing a simple line search method (i.e. *fmincon*) in Matlab as we have n unknown variables, $n+1$ constraints, and each call to the dual function involves the computation of $R(W)$, which can be done with special-purpose network flow algorithms that are computationally efficient.

The optimal solution to the primal, given the dual solution, is given in closed form simply as

$$\begin{aligned} z_i^* &= -\log \Delta + \log W_i^* \\ \Rightarrow \hat{\lambda}_i^* &= \frac{1}{\Delta} W_i^*. \end{aligned}$$

This LP is in standard form and thus there exists a basic feasible solution that is optimal [5]. Also note that any column of A has exactly two nonzero values, with one being 1 and the other being -1 . Combining this observation with (8), it follows that this is a network flow problem [5] (chapter 7). We have a graph $G = (V, E)$ with a set V of vertices, ranging from 1 to n where each vertex corresponds to a millisecond observation. The edge set E corresponds to directed edges (j_1, j_2) in the graph, and each edge (j_1, j_2) has an associated cost C_{j_1, j_2} . Note that the graph is fully connected, because we want the Lipschitz constraints to be satisfied for all possible pairs. To enforce absolute value constraints, what this means is that the directed edges (j_1, j_2) and (j_2, j_1) are both in the edge set E and they both have the same cost C_{j_1, j_2} . There is a flow $b_i = W_i - dN_i$ that is supplied to the network at node i .

By virtue of the dual, this can also be expressed as

$$\begin{aligned} \max \quad & \sum_i^n p_i(W_i - dN_i) \\ \text{s.t.} \quad & p_{j_1} - p_{j_2} \leq C_{j_1, j_2}, \quad j_1, j_2 = 1, \dots, n. \end{aligned}$$

In general, network flow problems are efficiently solvable and the network simplex method, dual ascent methods, and likewise allow for extremely efficient solutions. They are often orders of magnitudes faster than implementations of the standard LP simplex method. Thus we now have the equivalent problem (shown in box above) that can be solved efficiently.

3 Results

We study a spike train data series from a goldfish retinal ganglion cell neuron recorded in vitro. The retinae were removed from the goldfish and maintained in a flow of moist oxygen and recordings were made with an extracellular microelectrode under constant illumination (constant stimulus). The data are 975 spikes recorded over 30 seconds from neuron 78 in [12]. Figure 1 illustrates the histogram of ISIs in this data.

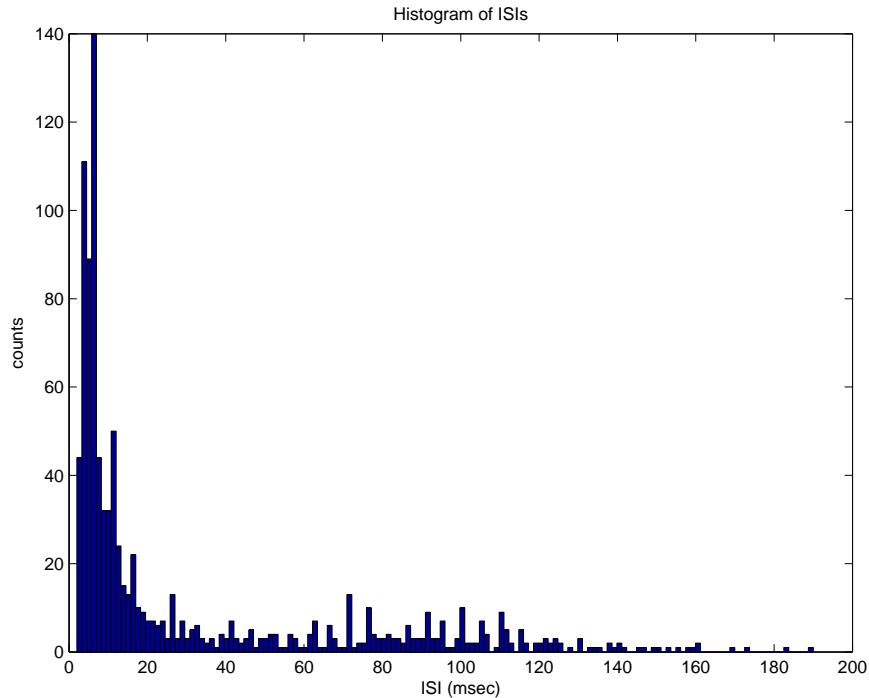


Figure 1: Histogram of ISIs

We built various nonparametric (using our DMLE) and parametric models for the first 3 seconds of this data. We use a resolution of $\Delta = 1\text{ms}$, where each X_i is equal to the time (in ms) since the last spike (purely an intrinsic factor). For our nonparametric model, we do not know a priori the Lipschitz constant K and so we estimate it from the data. We use the time-rescaling theorem and the KS statistic as a principled methodology for estimating K . Note that

estimating K has a flavor of model selection [2]: by virtue of (5), the feasible sets are nested increasingly with K . This is analogous to the traditional model selection problem whereby one tries to develop the correct parameter dimension of a model. As in the case of traditional model selection, allowing for K to be too large results in over-fitting and constraining K to be too small overconstrains the estimate of the function we are after. The parametric models built are enumerated below.

1. ISI Density is Exponential: $\hat{\lambda}^{exp}$
2. ISI Density is Gamma: $\hat{\lambda}^{gam}(t, \hat{\beta}_{ML}, \hat{\alpha}_{ML})$
3. ISI Density is Inverse Gaussian: $\hat{\lambda}^{invg}(t, \hat{\mu}_{ML}, \hat{\eta}_{ML})$
4. CIF is a GLM: $\hat{\lambda}^{glm}(t) = \exp\{\sum_{i=1}^{10} \hat{\alpha}_i^{ML} N(t - l(i, 1), t - l(i, 2))\}$ ¹

Figure 2 shows the KS plots for the 4 parametric models and 3 nonparametric models (with 3 different Lipschitz constants). The optimal nonparametric model ($K = 14.1$) has the best absolute goodness of fit.

We note that as K tends to ∞ , the Lipschitz constraints disappear and the optimal DMLE solution is just

$$\hat{\lambda}_i = \begin{cases} 0 & dN_i = 0 \\ \frac{1}{\Delta} & dN_i = 1. \end{cases}$$

Loosely speaking, the above is simply a model that 'equals' the data. As a consequence, note that each time-rescaled ISI will become

$$Z_k = \int_{T_{k-1}}^{T_k} \lambda(t|X_t) dt \simeq \sum_{i=i_{k-1}+1}^{i_k} \hat{\lambda}_i \Delta = \frac{1}{\Delta} \Delta = 1$$

and so each $\tau_k = 1 - \exp(-Z_k) = 1 - \exp(-1)$. See the vertical line in the KS plot (Figure2).

Note that we can see this phenomena manifesting itself already at $K = 100$. Thus over-fitting, as it pertains to selection of K using our methodology, is captured with the KS plot.

3.1 Computing Uncertainty

In this section, we compute 95% confidence intervals for each point process estimate. In particular, we do the following for each case:

- Exponential
- Gamma

¹Note that the history dependence (X_i) is more flexible for the GLM estimate. In particular, $l = [1\ 5; 6\ 10; 11\ 20; 21\ 30; 31\ 35; 36\ 40; 41\ 45; 46\ 50; 51\ 60; 61\ 100]$ msec.

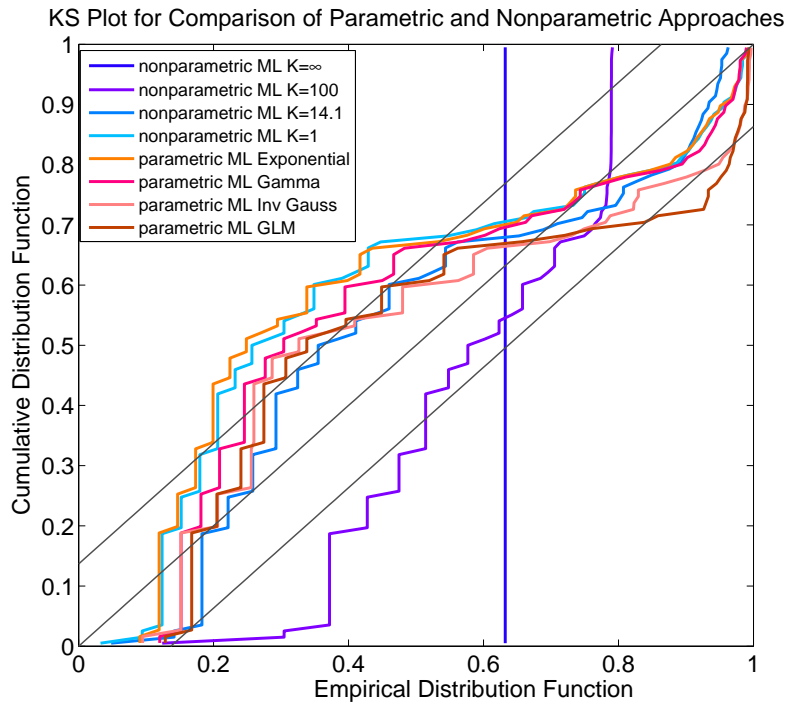


Figure 2: Using the time-rescaling theorem and Kolmogorov-Smirnov plots for model selection of the Lipschitz parameter K . The optimization algorithm was applied to 3000ms of goldfish retinal ganglionic neural data and plotted are the 95% confidence intervals as well as empirical CDFs of the time-rescaled ISIs for $K = 1.0$, $K = 14.1$, and $K = 100$.

- Inverse Gaussian
- GLM
- Nonparametric ($K = 14.1$)

4 Future Work

Future work includes showing consistency in our nonparametric estimate of the conditional intensity function, and understanding the KS statistic from an information-theoretic model selection perspective.

References

- [1] Barbieri R, Quirk MC, Frank LM, Wilson M, Brown EN, “Construction and Analysis of non-Poisson stimulus response models of neural spike train activity,” *Journal of Neurosci Methods*, 105: 25-37, 2001.
- [2] Barron AR, Rissanen J, Yu B. “The Minimum Description Length Principle in Coding and Modeling.” *IEEE Transactions on Information Theory*, vol 44, no. 6, 1998.
- [3] Bertsekas, D. “Nonlinear Programming, 2nd Edition.” *Athena Scientific, Belmont, MA*, 1999.
- [4] Brillinger, DR, ”Maximum likelihood analysis of spike trains of interacting nerve cells,” *Biological Cybernetics*, Vol. 59 (1988), pp. 189-200.
- [5] Bertsimas, D. and Tsitsiklis, JN. “Introduction to Linear Optimization, 2nd Edition.” *Athena Scientific, Belmont, MA*, 1997.
- [6] Bertsimas, D., Gamarnik, D., Tsitsiklis, JN. “Estimation of Time-Varying Parameters in Statistical Models: an Optimization Approach.” *Machine Learning*, vol 35., pp. 225-245, 1999.
- [7] Brown EN. etal, “The Time-Rescaling Theorem and Its Application to Neural Spike Train Data Analysis,” *Neural Computation*, 14: 325-346, 2001.
- [8] Brown EN etal., “Likelihood Methods for Neural Spike Train Data Analysis,” *Computational Neuroscience: A Comprehensive Approach*, chapter 9: 253-286, 2004.
- [9] Brown EN. “Theory of Point Processes for Neural Systems. In: Chow CC, Gutkin B, Hansel D, Meunier C, Dalibard J”, eds. *Methods and Models in Neurophysics. Paris, Elsevier* 2005, Chapter 14, pp. 691-726.
- [10] Daley D.J., Vere-Jones D., “An Introduction to the Theory of Point Processes,” *New York : Springer*, c2003.
- [11] Harris, Kenneth D, Csicsvari, Jozsef, Hirase, Hajime, Dragoi, George, Buzsaki, Gyorgy, “Organizations of cell assemblies in the hippocampus.,” *Nature*, Vol 424, July, 2003.
- [12] Iyengar S., Liao Q. “Modeling neural activity using the generalized inverse Gaussian distribution.” *Biol. Cyber.*, vol 77, pp. 289-295, 1997.
- [13] Johnson A, Kotz S.,” Distributions in statistics: Continuous univariate distributions,” *New York: Wiley*, 1970.
- [14] Kass RE, Ventura V, “A Spike Train Probability Model,”*Neural Computation*, 13:1713-1720, 2001.
- [15] Keat J, Reinagel P, Reid RC, Meister M, “Predicting every spike: a model for the responses of visual neurons,” *Neuron*, 30 (3): 803-17, June 2001.
- [16] Ogata Y. “Statistical Models for earthquake occurrences and residual analysis for point processes,” *J Am Stat Assoc*’ 83: 9-27, 1988.

- [17] Sarma S., Czanner G., Nguyen D., Wirth S., Wilson M., Suzuki W., Brown E., “Bootstrap Methods for Point Process Models of Neural Spiking Activity.” *preprint*.
- [18] Truccolo W., Eden U., Fellows M.R., Donoghue J., Brown EN, “A Point Process Framework for Relating Neural Spiking Activity to Spiking History, Neural Ensemble, and Extrinsic Covariate Effects,” *Journal of Neurophysiology*’ 93: 1074-1089, 2005.
- [19] Tuckwell, HC. “Introduction to Theoretical Neurobiology: Nonlinear and Stochastic Theories, Volume 2.” *Cambridge: Cambridge University Press*, 1988.