1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–15

Multivariate Mutually Regressive Point Processes

Ifigeneia Apostolopoulou IAPOSTOL@ANDREW.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA 15213, USA

Scott W. Linderman SCOTT.LINDERMAN@COLUMBIA.EDU
Columbia University, New York, NY 10027, USA

Kyle Miller MILLE856@ANDREW.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA 15213, USA

Artur Dubrawski AWD@CS.CMU.EDU
Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

Many real-world datasets involve sequences of interdependent events unfolding over time, which are naturally modeled as realizations of a point process. Despite many potential applications, existing point process models are limited in their ability to capture complex patterns of interaction. Hawkes processes (Hawkes, 1971) admit many efficient inference algorithms, but are limited to mutually excitatory interactions. Nonlinear Hawkes processes allow for more complex influence patterns, but we typically must resort to discrete-time approximations to estimate their parameters. In this paper, we introduce a new general class of point processes models extended with a nonlinear component that accounts for inhibitory interactions. We derive a fully Bayesian, continuous time inference algorithm for these processes using Pólya-Gamma augmentation and Poisson thinning. We illustrate the proposed model with an application to analyze multi-neuronal spike train recordings.

1. Introduction

Many natural phenomena and practical applications involve asynchronous and irregular events. Modeling correlations between events of various types may reveal informative patterns, such as when past occurrences may help predict next occurrences, or guide interventions to trigger or prevent future events. Probabilistic modeling of intensities of such events can be useful in many applied domains and help understanding of complex mechanisms which govern social media dynamics, neuron activity, medical events, consumer behavior, high frequency financial markets, or dynamics of crime.

Hawkes processes are the most widely used point process model which takes into consideration the influence of past events on the occurrence of future events (Hawkes, 1971; Hawkes and Oakes, 1974). A Hawkes process is actually a class of models in which impulse responses induced by past events linearly combine to increase the probability of future events. Therefore, it can be viewed as a mutually exciting point process.

A simple nonlinear generalization of the Hawkes process allows for both excitatory and inhibitory interactions, but evaluating the probability density of these models requires computing the integrated intensity, which is generally intractable. Instead, we are forced to use discrete time approximations, which reduce to a Poisson Generalized Linear Model (Poisson-

GLM). This discrete-time approximation of the continuous-time Poisson process represents the logarithm of the intensity function as a linear combination of general functions. The observed time window is divided into short discrete intervals, and the total number of events over each sub-interval is aggregated. As a result, parameter estimation amounts to fitting a generalized linear model, making learning of these models from data very efficient. Unlike the Hawkes model, the Poisson-GLM allows for both excitatory and inhibitory interactions, but still in a discrete-time formulation. One of its notable applications is in neuroscience, where it is widely used for modeling multi-neuronal spiking dynamics (Pillow et al., 2008). However, the estimated regression coefficients may vary widely depending on the boundaries chosen for aggregation (Fotheringham and Wong, 1991). Moreover, they may be suitable for one-step predictions but they could turn out to be problematic generative models due to stochastic instability (Gerhard et al., 2017).

There is currently a significant gap between real-world applications and statistical theory for point process models that support general temporal interactions (either excitatory or inhibitory) in a continuous-time regime. To this end, we develop a new class of point process models—*Multivariate Mutually Regressive Point Processes*—that allow for nonlinear temporal interactions while still admitting an efficient, fully-Bayesian inference algorithm in continuous time.

2. Proposed Model

We are interested in learning distributions over event sequences $(\dot{n}_1, \dot{t}_1), (\dot{n}_2, \dot{t}_2), \ldots,$ where each $\dot{n}_i \in \{1, 2, \ldots, N\}$ is a label that represents the type of the i-th observed event and $\dot{t}_i$ is the time of its occurrence (assuming that no two or more events can arrive at the same time). Let $\dot{t}^i_m$ denote the arrival time of the i-th observed event of type m, assuming a temporal ordering. The N-dimensional point process which generates the stream of events is mutually regressive in the sense that the history of events of any type m can affect the probability of an arrival of an event of any type n in the future. Specifically, the intensity function $\lambda_n(t)$ of the point process which models the occurrence of events of type n is defined as follows:

$$\lambda_n(t) = \lambda^*_n \, p_n(t), \qquad p_n(t) = \sigma(\psi_n(t)), \qquad \psi_n(t) = b_n + \sum_{m=1}^{N} h_{m\to n}(t),$$
$$h_{m\to n}(t) = w_{m\to n} \, h_m(t), \qquad h_m(t) = c \sum_{\dot{t}^i_m < t} e^{-\delta(t-\dot{t}^i_m)}, \tag{1}$$

where $\lambda^*_n > 0$, $c > 0$ and $\delta > 0$. In the above equations, $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. The weight $w_{m\to n}$ models the influence of type m on type n and h_m is the aggregated temporal influence of type m up to time t. We can simulate from this model via Poisson thinning (Lewis and Shedler, 1979; Adams et al., 2009). First, we sample a set of events from a homogeneous Poisson process with intensity λ^*_n, then we sequentially proceed through the simulated events and accept them with probability $\lambda_n(t)/\lambda^*_n = p_n(t)$, the relative intensity at that point in time. Importantly, the relative intensity only depends on the preceding events that were *accepted*; rejected events have no influence on the future intensity. The correctness of this procedure is proven in Appendix A.2.3. Note that a positive weight $w_{m\rightarrow n}$ means that events of type m excite future events of type n since $h_{m\rightarrow n}$ increases $p_n(t)$. Similarly, a negative weight shows that events of type m are inhibitory for events of type n. In a matrix form, the model can be written as follows:

$$\psi_n(t) = \boldsymbol{w}_n^T \boldsymbol{h}(t), \quad \boldsymbol{w}_n = [b_n, w_{1\rightarrow n}, w_{2\rightarrow n}, \ldots, w_{n\rightarrow n}]^T, \quad \boldsymbol{h}(t) = [1, h_1(t), h_2(t), \ldots, h_N(t)]^T. \tag{2}$$

Note that instead of a constant intensity λ_n^*, a cluster-based Hawkes process intensity function (Hawkes and Oakes, 1974; Hawkes, 1971) can be used such that:

$$\lambda_n^*(t) = \lambda_n^* + \sum_{i:\dot{t}_i<t} \alpha_{\dot{n}_i,n}\, e^{-\delta_{\dot{n}_i,n}(t-\dot{t}_i)}, \tag{3}$$

where $\alpha_{\dot{n}_i,n} \geq 0$, and $\delta_{\dot{n}_i,n} > 0$. By the superposition theorem for Poisson processes, the additive terms in Equation 3 can be viewed as independent nonhomogeneous processes (with intensity function that varies in time) characterized by the intensity function $\alpha_{\dot{n}_i,n}\, e^{-\delta_{\dot{n}_i,n}(t-\dot{t}_i)}$, triggered by the observed event i.

3. Bayesian Inference via Augmentation

The likelihood of the sequence $\dot{\mathcal{S}}_n \triangleq \{\dot{t}_n^i\}_{i=1}^{K_n}$ of K_n events generated by a point process with intensity function $\lambda_n(t)$ in the time window $[0,T]$ is (Rubin, 1972):

$$p(\dot{S}_n \mid \lambda_n(t)) = \exp\left\{-\int_0^T \lambda_n(\tau)\,\mathrm{d}\tau\right\} \prod_{i=1}^{K_n} \lambda_n(\dot{t}_n^i). \tag{4}$$

However, due to the sigmoid term in the intensity function (1), the integral is intractable (Adams et al., 2009). Let $\tilde{\mathcal{S}}_n \triangleq \{\tilde{t}_n^i\}_{i=1}^{M_n}$ be a sequence of M_n latent (thinned) events generated by the Poisson process characterized by the intensity function $\lambda_n^*\,(1-p_n(t))$. Its superposition with the intensity function of Equation 1 is a Poisson process with intensity λ_n^*. Define the merged event sequence to be the set

$$S_n \triangleq \dot{S}_n \cup \tilde{S}_n = \{t_n^i, s_n^i\}_{i=1}^{K_n+M_n}, \tag{5}$$

where $s_n^i \triangleq \mathbb{I}(t_n^i \in \dot{S}_n)$ is the label indicating whether event i is realized or thinned. The likelihood of the merged set of events is

$$p(S_n \mid \lambda_n^*) = \exp(-\lambda_n^*\,T)\,{\lambda_n^*}^{K_n+M_n} \prod_{i=1}^{M_n+K_n} p_n(t_n^i)^{s_n^i}\,(1-p_n(t_n^i))^{1-s_n^i}. \tag{6}$$

Note that in this augmented space of observed and thinned events, the likelihood is tractable; the hard integral in (4) is effectively replaced by Markov Chain Monte Carlo (MCMC) over an extended state space. Each MCMC iteration of our algorithm samples thinned events from the Poisson process $\lambda_n^*\,(1-p_n(t))$. Given S_n, conjugate updates can be derived for λ_n^* (or $\lambda_n^*(t)$). Finally, the inference of the weights $w_{m\rightarrow n}$ of the thinning procedure amounts to solving a logistic regression problem that can be tackled effectively via Pólya-Gamma augmentation as in (e.g. Polson et al., 2013; Linderman et al., 2015, 2016). The complete algorithm and its derivation are provided in Appendix A.

4. Experiment: Multi-Neuron Spike Train Analysis

In this section, we apply the proposed model to a data set consisting of spike train recordings from neurons in the cat primary visual cortex (area 17). The data were recorded in the laboratory of Tim Blanche (Blanche, 2009), and are available from the NSF-funded CRCNS data repository http://www.dickimaw-books.com/software/makejmlrbookgui/videos/. The datasets were acquired with multi-channel silicon electrode arrays that enable simultaneous recording from more than a hundred single units at once. This is of utmost importance because recordings from multiple neurons at a time are necessary if conclusions about cortical circuit function or network dynamics are to be derived. Specifically, we used the spike times of ten simultaneously recorded cells from anaesthetized cat area 17 in response to oriented drifting bars.

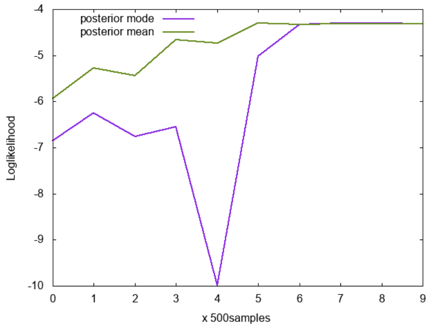

Figure 1: Learning Curves for the Multivariate Mutually Regressive Point Process on Neuron Spike Data.

We trained our model on a stream consisting of 2000 spikes (events) observed in a window of $T = 30966$ microseconds. In Fig. 3 and Fig. 4 in Appendix B, we visualize the temporal patterns of the training and the testing spike sequence. We ran the Markov Chain for 5000 iterations and we used the last 2000 to obtain the posterior modes of the model parameters. We manually set the parameters in Equation 1: $\delta = 0.001$ and $c = 100$. We used normal prior (of zero mean and standard deviation of 10) for the inhibition weights. For the bias term in the sigmoid function, we set a high mean value (100) and a standard deviation of 10 for its normal prior. In this way, excitatory effects are mostly demonstrated by a large $\alpha_{m,n}$, while a negative $w_{m \to n}$ that cancels out the bias term in the sigmoid indicates an inhibitory relationship. We utilized the mutually exciting intensity defined in Equation 3. We assumed gamma priors (with inverse-scale parameterization) for λ_n^* (with $\alpha = 0.08$ and $\beta = 2$) and $\alpha_{m,n}$ (with $\alpha = 1$ and $\beta = 2$), and an exponential prior for $\delta_{m,n}$ with $\lambda = 100$. Figure 1 shows the log likelihood (normalized per number of events) realized by the inference for an increasing number of samples. The normalized log likelihood realized for a held-out sequence (with the same observation window) was -5.36 (close to the log likelihood for the sequence used for the training once the Markov Chain has converged).

We computed the full Hawkes log likelihood, which accounts for both the cluster structure and the generated thinned events. We assigned each event to a posterior parent (the nonhomogeneous Poisson process that has most likely triggered the event). Given the posterior mode point parameters, we augmented the spike sequence with latent events. We sampled 100 latent sequences, and we report the mean of the full log likelihoods realized. We did the same both for the training and the testing spike streams. Strong inhibitory effects are discovered from neuron 6 on neuron 10, from neuron 1 on neuron 9, from neurons 6 and 4 on neuron 5, from neurons 6 and 1 on neuron 3 and from neuron 2 on neuron 1 (refer to Figure 3 in Appendix B for the ordering of the neurons). In Figure 2 in Appendix B, the computational demands of the inference procedure are demonstrated.

References

Ryan Prescott Adams, Iain Murray, and David JC MacKay. Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16. ACM, 2009.

Tim Blanche. Large-scale neuronal recordings in primary visual cortex. 2009.

A Stewart Fotheringham and David WS Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A*, 23(7):1025–1044, 1991.

Felipe Gerhard, Moritz Deger, and Wilson Truccolo. On the stability and dynamics of stochastic spiking neuron models: Nonlinear hawkes process and point process glms. *PLoS computational biology*, 13(2):e1005390, 2017.

Heikki Haario, Eero Saksman, Johanna Tamminen, et al. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

Alan G Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.

Peter A Lewis and Gerald S Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics (NRL)*, 26(3):403–413, 1979.

Scott Linderman, Matthew Johnson, and Ryan P Adams. Dependent multinomial models made easy: Stick-breaking with the pólya-gamma augmentation. In *Advances in Neural Information Processing Systems*, pages 3456–3464, 2015.

Scott Linderman, Ryan P Adams, and Jonathan W Pillow. Bayesian latent structure discovery from multi-neuron recordings. In *Advances in neural information processing systems*, pages 2002–2010, 2016.

Scott W Linderman and Ryan P Adams. Scalable bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*, 2015.

Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995, 2008.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108 (504):1339–1349, 2013.

Vinayak Rao, Ryan P Adams, and David D Dunson. Bayesian inference for matérn repulsive processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79 (3):877–897, 2017.

Jakob Gulddahl Rasmussen. Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642, 2013.

Izhak Rubin. Regular point processes and their detection. *IEEE Transactions on Information Theory*, 18(5):547–557, 1972.

Acknowledgments

We would like to thank Alex Reinhart for helpful discussions. The first author would like to acknowledge Alexander S. Onassis Public Benefit Foundation and A.G. Leventis Foundation, for graduate fellowships supporting this work. This work has been partially supported by DARPA (FA8750-17-2-0130) and NIH (NR013912).

Appendix A. Bayesian Inference

A.1. Derivation of the Likelihood

The full likelihood including both the observed events and the thinned events is,

$$\mathcal{L} \triangleq \mathcal{L}\big(\{\lambda_n^*, \boldsymbol{w}_n, \dot{\mathcal{S}}_n, \tilde{\mathcal{S}}_n\}_{n=1}^N; \boldsymbol{\theta}\big), \tag{7}$$

where N is the number of types, $\dot{\mathcal{S}}_n \triangleq \{\dot{t}_n^i\}_{i=1}^{K_n}$ is the sequence of the K_n realized (observed) events of type n, and $\tilde{\mathcal{S}}_n \triangleq \{\tilde{t}_n^i,\}_{i=1}^{M_n}$ is the sequence of the thinned events of type n. We also assume a temporal order of the events such that $\dot{t}_1^i < \dot{t}_2^i < \dot{t}_3^i < \cdots < \dot{t}_{K_n}^i$. Similarly, for the thinned events, $\tilde{t}_1^i < \tilde{t}_2^i < \tilde{t}_3^i < \cdots < \tilde{t}_{K_n}^i$.

We assume that $w_{m \to n} \sim \mathcal{N}(\mu_n, \sigma_n^2)$. In matrix notation $\boldsymbol{w}_n \sim \mathcal{N}(\boldsymbol{\mu}_n, \Sigma_n)$, where $\boldsymbol{\mu}_n = [\mu_n, \mu_n, \dots, \mu_n]$, and $\Sigma_n = diag(\sigma_n^2)$. $\boldsymbol{\theta}$ is the set of hyperparameters. With a gamma prior for λ_n^* such that $\lambda_n^* \sim \text{Gamma}(\alpha_*, \beta_*)$, a Normal prior for $\mu_n \sim \mathcal{N}(\mu_0, \sigma_0^2)$, and a gamma prior for the precision $\tau_n = 1/\sigma_n^2$ such that $\tau_n \sim \text{Gamma}(\alpha_0, \beta_0)$, the parameters are $\boldsymbol{\theta} = \{\alpha_*, \beta_*, \mu_0, \sigma_0^2, \alpha_0, \beta_0\}$. In future work, we will consider a Normal-Wishart prior for the prior mean and the covariance matrix of the weights $\boldsymbol{w}_n$ in the probit model.

The thinned events are not observed; hence they constitute latent variables of the model. The merged sequence of the realized and thinned events is denoted by:

$$\mathcal{S}_n \triangleq \{t_n^i, s_n^i\}_{i=1}^{K_n+M_n}, \tag{8}$$

such that $t_1^i < t_2^i < t_3^i < \cdots < t_{K_n+M_n}^i$. The variables s_n^i correspond to the label that indicates whether the event will be realized ($s_n^i = 1$) or thinned ($s_n^i = 0$). We assume that:

$$\{t_n^1, t_n^2, \dots, t_n^{K_n+M_n}\} \sim \mathcal{PP}(\lambda_n^*). \tag{9}$$

The likelihood function of a sequence of events which is generated by an inhomogeneous Poisson process: $\{t_n\}_{n=1}^N \sim \mathcal{PP}(\lambda(t))$ in the observation window $[0, T]$ is given by (Rubin, 1972),

$$p(\{t_n\}_{n=1}^N \mid \lambda(t)) = \exp\left\{-\int_0^T \lambda(\tau)\mathrm{d}\tau\right\} \prod_{n=1}^N \lambda(t_n). \tag{10}$$

By combining Equations 1, 9, and 10, the likelihood of a sequence of both the observed and the thinned events will be:

$$\mathcal{L} = p(\boldsymbol{\theta}) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{w}_n; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \prod_{n=1}^{N} \left(\lambda_n^{*\,M_n+K_n} \exp(-\lambda_n^* T)\right) \prod_{n=1}^{N} \prod_{i=1}^{K_n+M_n} \frac{e^{\psi_n(t_n^i) * s_n^i}}{e^{\psi_n(t_n^i)} + 1}. \tag{11}$$

The first term in the product corresponds to the priors for the mean and variance of the inhibition weights and the intensity parameters. The second term corresponds to the normal prior for the inhibition weights. The third term corresponds to the likelihood of the homogeneous Poisson processes which generates the sequence S_n of the realized and thinned events of type n. Finally, the last term corresponds to the realization of the Bernoulli process of the thinning procedure.

The likelihood contribution of the thinning acceptance/ rejection of an event at time t_n^i:

$$\ell_n^i = \frac{e^{\psi_n(t_n^i) * s_n^i}}{e^{\psi_n(t_n^i)} + 1}, \tag{12}$$

can be rewritten according to Theorem 1 in (Polson et al., 2013) as:

$$\ell_n^i \propto \exp(\kappa_n^i {\boldsymbol{w}_n}^T \boldsymbol{h}(t_n^i)) \int_0^{\infty} \exp\left\{-\frac{1}{2}\omega_n^i (\boldsymbol{w}_n^T \boldsymbol{h}(t_n^i))^2\right\} \mathcal{PG}_m(\omega_n^i \mid 1, 0)\, d\omega_n^i, \tag{13}$$

where $\kappa_n^i = s_n^i - 1/2$, and $\mathcal{PG}_m(\omega_n^i \mid 1, 0)$ is the density of a Pólya-Gamma distribution with parameters $(1, 0)$. Combined with a prior on $\boldsymbol{w}_n$, the integrand in Equation 13 defines a joint density on $(s_n^i, \omega_n^i, \boldsymbol{w}_n)$, where ω_n^i is a latent Pólya-Gamma random variable. The augmented (with the latent variables ω_n^i) likelihood will be $\mathcal{LP} \triangleq \mathcal{L}\big(\{\lambda_n^*, \boldsymbol{w}_n, \mathcal{SP}_n\}_{n=1}^N; \boldsymbol{\theta}\big)$, where

$$\dot{\mathcal{SP}}_n \triangleq \{\dot{t}_n^i, \dot{\omega}_n^i\}_{i=1}^{K_n}, \qquad \tilde{\mathcal{SP}}_n \triangleq \{\tilde{t}_n^i, \tilde{\omega}_n^i\}_{i=1}^{M_n}, \qquad \mathcal{SP}_n \triangleq \{t_n^i, s_n^i, \omega_n^i\}_{i=1}^{K_n+M_n}. \tag{14}$$

From Equations 11, 12, 13, we obtain:

$$\begin{aligned}\mathcal{LP} = p(\boldsymbol{\theta}) \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{w}_n; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n, \boldsymbol{\theta}) \prod_{n=1}^{N} \Big[\lambda_n^{*\,M_n+K_n} \exp\big(-\lambda_n^* T)\Big] \times \\ \prod_{n=1}^{N} \prod_{i=1}^{K_n+M_n} \exp\left\{\kappa_n^i {\boldsymbol{w}_n}^T \boldsymbol{h}(t_n^i) - \tfrac{1}{2}\omega_n^i (\boldsymbol{w}_n^T \boldsymbol{h}(t_n^i))^2\right\} \mathcal{PG}_m(\omega_n^i \mid 1, 0)\big).\end{aligned} \tag{15}$$

A.2. Gibbs Sampling Updates

Algorithm 1 summarizes the steps of our Markov Chain Monte Carlo inference algorithm. In the following subsections, we provide the mathematical derivations.

A.2.1. Gibbs Sampling for the interaction weights

From Equation 15 and by keeping only the terms which contain $\boldsymbol{w}_n$,

$$p(\boldsymbol{w}_n \mid \dots) = p(\boldsymbol{w}_n \mid \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n, \mathcal{SP}_n, \{\dot{\mathcal{S}}_{n'}\}_{n'=1, n' \neq n}^{N}). \tag{16}$$

Algorithm 1 Bayesian Inference for Mutually Regressive Point Processes

```
1. Input: Sequence of events $\{\dot{\mathcal{S}}_n\}_{n=1}^{N}$.
2. Output: Samples from $\{p(\boldsymbol{w}_n \mid \dot{\mathcal{S}}_n), p(\lambda_n^* \mid \dot{\mathcal{S}}_n)\}_{n=1}^{N}$.
3. Initialize randomly the model parameters from the priors:
   $\{\boldsymbol{w}_n \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n), \lambda_n^* \sim \mathcal{G}_m(\alpha_*, \beta_*)\}_{n=1}^{N}$.
4. Repeat
   (a) Sample the thinned events: $\{\tilde{\mathcal{S}}_n \sim \mathcal{PP}(\lambda_n^*\,(1 - p_n(t)))\}_{n=1}^{N}$ via Poisson thinning.
   (b) Sample the latent Pólya-Gamma variables: $\{\omega_n^i \sim \mathcal{PG}_m(1, \psi_n(t_n^i))\}_{n=1}^{N}$.
   (c) Sample the weights (Eq 22 & 23): $\boldsymbol{w}_n \sim \mathcal{N}(\tilde{\boldsymbol{\Sigma}}_n, \tilde{\boldsymbol{\mu}}_n)$ for $n = 1, \ldots, N$.
   (d) Sample the parameters of the exogenous intensities (Eq 33, 34 & 35):
       $\lambda_n^* \sim \mathrm{Gamma}(\tilde{\alpha}_n, \tilde{\beta}_n)$ for $n = 1, \ldots, N$.
```

In words, the posterior of the weight $\boldsymbol{w}_n$ depends on the realized and thinned events of type n and their corresponding Pólya-Gamma latent variables, and the arrival times of the realized events of the rest of the types. Note that the dependence on the event history of the other types stems from the definition of $\psi_n(t)$ in Equations (1) and (2).

$$\begin{aligned}
p(\boldsymbol{w}_n \mid \dots) &\propto \mathcal{N}(\boldsymbol{w}_n; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n, \boldsymbol{\theta}) \prod_{k=1}^{K_n+M_n} \exp(\kappa_n^k \boldsymbol{w}_n{}^T \boldsymbol{h}(t_n^k) - \tfrac{1}{2}\omega_n^k (\boldsymbol{w}_n^T \boldsymbol{h}(t_n^k))^2) \\
&\propto \mathcal{N}(\boldsymbol{w}_n; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n, \boldsymbol{\theta}) \prod_{k=1}^{K_n+M_n} \exp\big(-\frac{\omega_n^k}{2}(\boldsymbol{w}_n^\intercal \boldsymbol{h}(t_n^k) - \kappa_n^k/\omega_n^k)^2\big) \\
&\propto \exp\big\{-\frac{1}{2}(\boldsymbol{w}_n - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{w}_n - \boldsymbol{\mu}_n)\big\} \exp\big\{-\frac{1}{2}(\boldsymbol{z}_n - \boldsymbol{H}_n \boldsymbol{w}_n)^T \boldsymbol{\Omega}_n (\boldsymbol{z}_n - \boldsymbol{H}_n \boldsymbol{w}_n)\big\},
\end{aligned} \tag{17}$$

where:

$$\boldsymbol{\Omega}_n = \mathrm{diag}(\omega_n^1, \omega_n^2, \dots, \omega_n^{K_n+M_n}), \tag{18}$$

$$\boldsymbol{z}_n = [\kappa_n^1/\omega_n^1, \dots, \kappa_n^{K_n+M_n}/\omega_n^{K_n+M_n}]^T, \tag{19}$$

$$\boldsymbol{H}_n = [\boldsymbol{h}(t_n^1), \dots, \boldsymbol{h}(t_n^{K_n+M_n})]^T. \tag{20}$$

Finally,

$$p(\boldsymbol{w}_n \mid \dots) = \mathcal{N}(\boldsymbol{w}_n; \tilde{\boldsymbol{\Sigma}}_n, \tilde{\boldsymbol{\mu}}_n). \tag{21}$$

By equating the quadratic and linear terms of $\boldsymbol{w}_n$, we get $\tilde{\boldsymbol{\Sigma}}_n$ and $\tilde{\boldsymbol{\mu}}_n$ respectively:

$$\tilde{\boldsymbol{\Sigma}}_n = \big(\boldsymbol{\Sigma}_n^{-1} + \boldsymbol{H}_n^T \boldsymbol{\Omega}_n \boldsymbol{H}_n\big)^{-1}, \tag{22}$$

$$\tilde{\boldsymbol{\mu}}_n = \tilde{\boldsymbol{\Sigma}}_n \big(\boldsymbol{\Sigma}_n^{-1} \boldsymbol{\mu}_n + \boldsymbol{H}_n^T \boldsymbol{\Omega}_n \boldsymbol{z}_n\big). \tag{23}$$

Note that the labels of the events contribute to the new sample of the weight through the terms κ_n^k. The sign of these terms (positive for a realized event and negative for a thinned event) steers the new sample of the weight to either a positive or a negative value according to Equation 23. The contribution of these terms is weighted by the history of the corresponding event.

A.2.2. Gibbs Sampling for the Pólya-Gamma latent variables

From Theorem 1 in (Polson et al., 2013), for $\alpha = 1$ and $\beta = 1$, we obtain

$$p(\omega_n^k \mid \dots) = p(\omega_n^k \mid t_k^n, \boldsymbol{w}_n, \{\dot{\mathcal{S}}_{n'}\}_{n'=1}^N) = \mathcal{PG}_m(\omega_n^k; 1, \psi_n(t_n^k)). \tag{24}$$

A.2.3. Gibbs Sampling for the thinned events

We will prove that:

$$p(\tilde{\mathcal{S}}_n \mid \{\dot{\mathcal{S}}_n\}_{n=1}^N, \boldsymbol{w}_n, \lambda_n^*) = \exp\left\{-\int_0^T \lambda_n^* \left(1 - p_n(\tau)\right) \mathrm{d}\tau\right\} \prod_{i=1}^{M_n} \left(\lambda_n^* \left(1 - p_n(\tilde{t}_n^i)\right)\right), \tag{25}$$

for $n = 1, 2, \dots, N$. The proof is similar in spirit to the proof in (Rao et al., 2017). Intuitively, according to Equation 10, this implies that the thinned events are generated from:

$$\tilde{\mathcal{S}}_n \sim \mathcal{PP}(\lambda_n^* \left(1 - p_n(t)\right), \tag{26}$$

which is the symmetric case of Equation 1 for the sequence of the observed events. Note that there are no interactions between the thinned events; only the observed events can affect the thinning probability of a latent event.

Equivalently, we will prove that:

$$p(\dot{\mathcal{S}}_n \mid \{\dot{\mathcal{S}}_{n'}\}_{n'=1, n'\neq n}^N, \boldsymbol{w}_n, \lambda_n^*) = \exp\left\{-\int_0^T \lambda_n^* \, p_n(\tau) \, \mathrm{d}\tau\right\} \prod_{i=1}^{K_n} \left(\lambda_n^* \, p_n(\dot{t}_n^i)\right). \tag{27}$$

Note that there is no circularity since the probability $p_n(t)$ depends on the events which occurred before time t. Therefore, the joint likelihoods in Equation 27 are well defined. From Equation 11, it holds that:

$$p(\mathcal{S}_n \mid \{\dot{\mathcal{S}}_{n'}\}_{n'=1, n'\neq n}^N, \boldsymbol{w}_n, \lambda_n^*) = exp(-\lambda_n^* \, T) \prod_{i=1}^{K_n} \left(\lambda_n^* \, p_n(\dot{t}_n^i)\right) \prod_{i=1}^{M_n} \left(\lambda_n^* \left(1 - p_n(\tilde{t}_n^i)\right)\right). \tag{28}$$

Note that this is different from the joint probability $p(t_n^1, t_n^2, \dots, t_n^{K_n+M_n})$ since it also involves the probability of the outcome of the thinning for each event. Then, from Bayes rule and Equations 27 and 28, Equation 25 follows:

$$p(\tilde{\mathcal{S}}_n \mid \{\dot{\mathcal{S}}_n\}_{n=1}^N, \boldsymbol{w}_n, \lambda_n^*) = \frac{p(S_n \mid \{\dot{\mathcal{S}}_{n'}\}_{n'=1, n'\neq n}^N, \boldsymbol{w}_n, \lambda_n^*)}{p(\dot{\mathcal{S}}_n \mid \{\dot{\mathcal{S}}_{n'}\}_{n'=1, n'\neq n}^N, \boldsymbol{w}_n, \lambda_n^*)}. \tag{29}$$

In Equation 28, we first marginalize over the arrival times of the thinned events:

$$p(\dot{\mathcal{S}}_n, M_n \mid \{\dot{\mathcal{S}}_{n'}\}_{n'=1,n'\neq n}^{N}, \boldsymbol{w}_n, \lambda_n^*) = \exp(-\lambda_n^* T) \prod_{i=1}^{K_n} \left(\lambda_n^* \, p_n(\dot{t}_n^i)\right) \frac{(\int_0^T \lambda_n^* \, (1 - p_n(\tau)) d\tau)^{M_n}}{M_n!}. \tag{30}$$

Note that the ordering of $\tilde{t}_n^i$ reduces the integration interval for the marginalization. Subsequently, we marginalize over the number of the thinned events M_n:

$$\begin{aligned} p(\dot{\mathcal{S}}_n \mid \{\dot{\mathcal{S}}_{n'}\}_{n'=1,n'\neq n}^{N}, \boldsymbol{w}_n, \lambda_n^*) &= \exp(-\lambda_n^* T) \prod_{i=1}^{K_n} \left(\lambda_n^* \, p_n(\dot{t}_n^i)\right) \sum_{M_n=0}^{\infty} \frac{(\int_0^T \lambda_n^* \, (1 - p_n(\tau)) d\tau)^{M_n}}{M_n!} \\ &= \exp(-\lambda_n^* T) \prod_{i=1}^{K_n} \left(\lambda_n^* \, p_n(\dot{t}_n^i)\right) \exp\left\{\int_0^T \lambda_n^* \, (1 - p_n(\tau)) d\tau\right\}, \end{aligned} \tag{31}$$

and Equation 27 follows immediately. We note that the derivations are similar if a Hawkes intensity as that in Equation 3 is used.

A.2.4. Gibbs Sampling for the Intensity Parameters

Given $\mathcal{S}_n$, conjugate prior updates are possible for the intensity parameters (Rasmussen, 2013). This is due to the independence between the parameters in the logistic related portion $p_n(t)$ and the intensity-related portion λ_n^* (or $\lambda_n^*(t)$) of the model.

From Equation 15, and by keeping the terms in which λ_n^* appears, we obtain:

$$p(\lambda_n^* \mid \mathcal{S}_n) \propto {\lambda_n^*}^{K_n + M_n} \exp(-\lambda_n^* T) \, \mathrm{Gamma}(\lambda_n^* \mid \alpha_*, \beta_*), \tag{32}$$

where $\mathrm{Gamma}(\lambda_n^* \mid \alpha_*, \beta_*)$ is the pdf of a Gamma distribution. Therefore,

$$p(\lambda_n^* \mid \mathcal{S}_n) = \mathrm{Gamma}(\lambda_n^* \mid \tilde{\alpha}_n, \tilde{\beta}_n), \tag{33}$$

$$\tilde{\alpha}_n = K_n + M_n + \alpha_*, \tag{34}$$

$$\tilde{\beta}_n = T + \beta_*. \tag{35}$$

Similarly, in case of a Hawkes intensity (Equation 3) and by adopting its clustering interpretation, the update for the excitatory coefficients $\alpha_{n,m}$ is given by:

$$\begin{aligned} &p(\alpha_{n,m} \mid \mathcal{S}_n) \propto \\ &\quad exp\left\{-\alpha_{n,m} \sum_{i:\dot{n}_i = n} \int_{\dot{t}_i}^{T} exp\left(-\delta_{n,m}(\tau - \dot{t}_i)\right) d\tau\right\} \alpha_{n,m}^{N_{n,m}} \, \mathrm{Gamma}(\alpha_{n,m} \mid \alpha_*, \beta_*) \Rightarrow \\ &p(\alpha_{n,m} \mid \mathcal{S}_n) \propto \\ &\quad exp\left\{-\alpha_{n,m} \sum_{i:\dot{n}_i = n} \frac{1}{\delta_{n,m}} \left(1 - exp(-\delta_{n,m}(T - \dot{t}_i))\right)\right\} \alpha_{n,m}^{N_{n,m}} \, \mathrm{Gamma}(\alpha_{n,m} \mid \alpha_*, \beta_*), \end{aligned} \tag{36}$$

where $N_{n,m}$ is the number of events of type m triggered by an event of type n. Finally,

$$p(\alpha_{n,m} \mid \mathcal{S}_n) \sim \mathrm{Gamma}(\tilde{\alpha}_{n,m}, \tilde{\beta}_{n,m}), \tag{37}$$

$$\tilde{\alpha}_{n,m} = \alpha_* + N_{n,m}, \tag{38}$$

$$\tilde{\beta}_{n,m} = \beta_* + \sum_{i:\dot{n}_i=n} \frac{1}{\delta_{n,m}} \big(1 - exp\big(- \delta_{n,m}(T - \dot{t}_i)\big)\big)\,. \tag{39}$$

The parent of each event follows a multinomial distribution (the updates are provided in Linderman and Adams (2015)). An Adaptive Metropolis Hastings (Haario et al., 2001) update can be used for the coefficients $\delta_{n,n'}$, while the Hastings ratio can be found in (Rasmussen, 2013).

Appendix B. Experiment: Multi-Neuron Spike Train Analysis

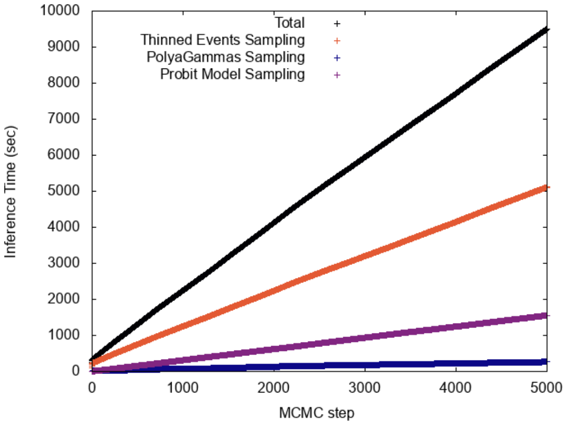

Figure 2: Inference Time of Multi-Neuron Spike Train as a Multivariate Mutually Regressive Point Process.

Figure 2 illustrates the computational demands of the main steps of the inference provided in Algorithm 1. The time required for the computation of the kernel history of the thinned events (note that for the realized events the history is computed only once) according to Equations 1 and 2 and for the sampling the parameters of the Hawkes intensity is counted towards the total inference time in Figure 2. This time reflects execution time among 4 threads for the steps of the inference that can be parallelized. Specifically, the sampling of the Pólya-Gamma variable and the simulation of the nonhomogeneous Poisson process $\alpha_{\dot{n}_i,n}\, e^{-\delta_{\dot{n}_i,n}(t-\dot{t}_i)}\,(1-p_n(t))$ for sampling of the thinned events is inherently parallel across the events. The most computationally demanding step of the inference is the sampling of the thinned events. However, the time required for the simulation of the triggered nonhomogeneous Poisson processes is reduced as the Markov Chain starts to converge since $(1-p_n(t))$ remains large only for the real inhibitory relationships.

Figures 3 and 4 illustrate the arrivals of spikes emitted by 10 neurons in the cat visual area cortex.

Figure 5 shows the intensity functions learned from the spike train of Figure 3.

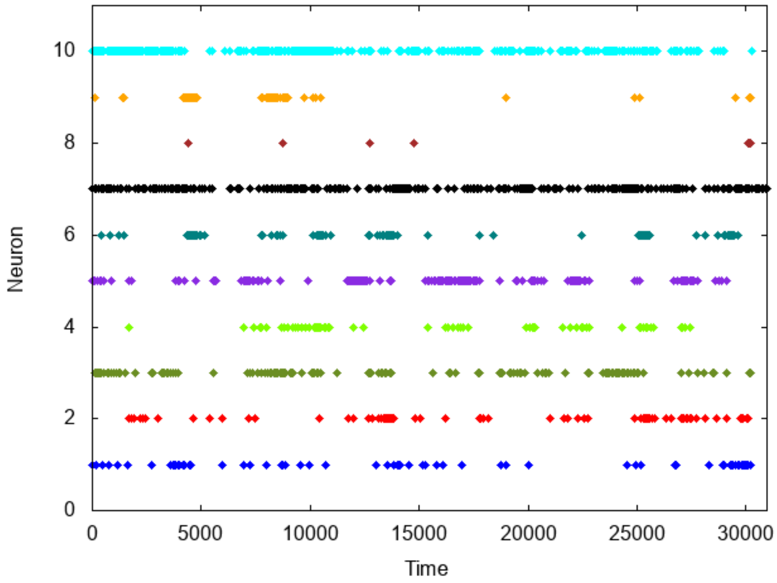

Figure 3: The Multi-Neuron Spike Train used for training.

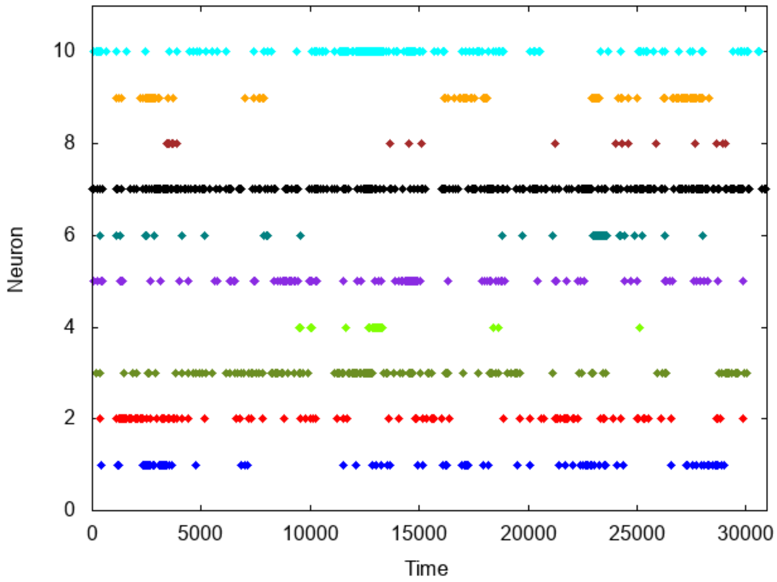

Figure 4: The Multi-Neuron Spike Train used for testing.

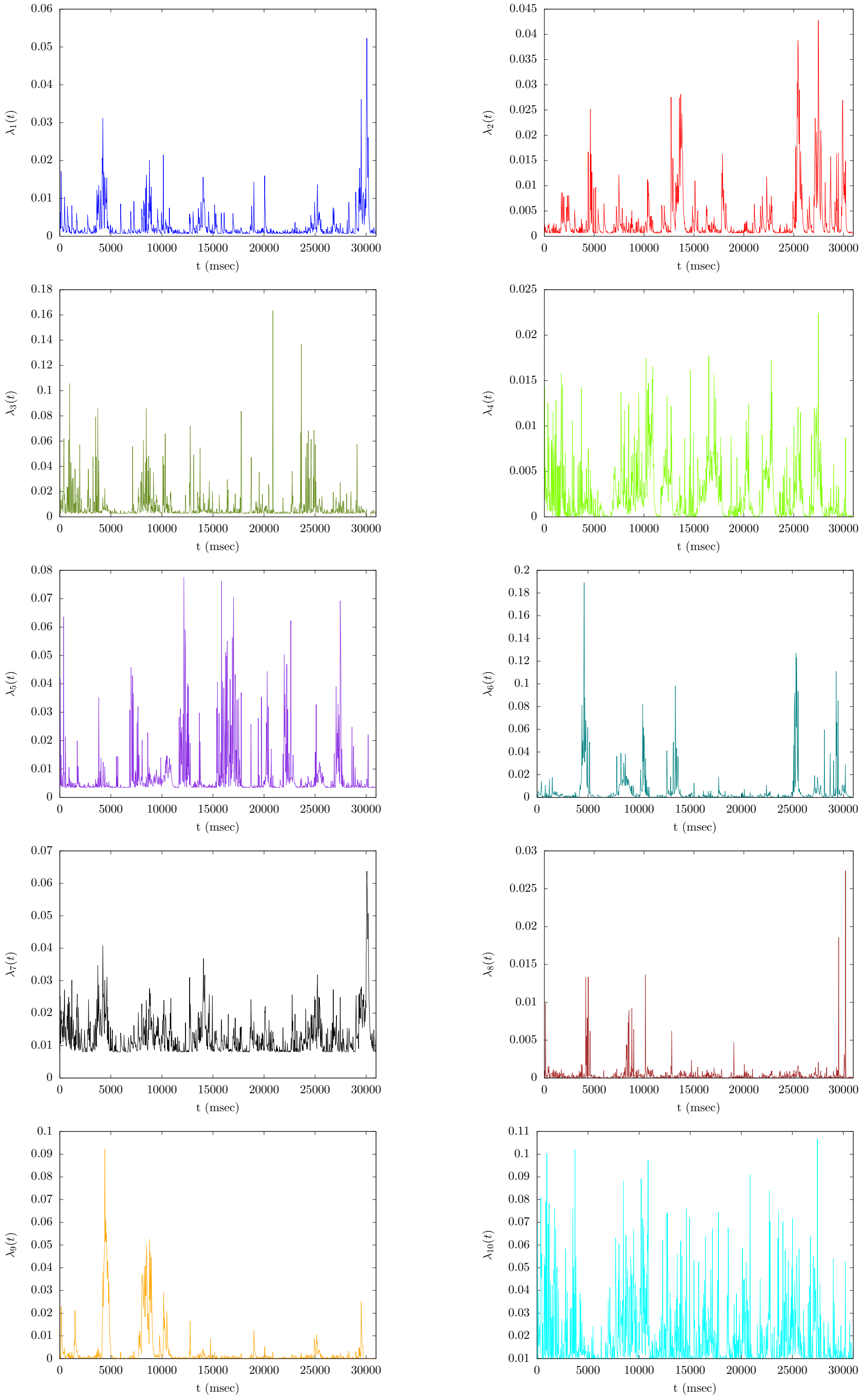

Figure 5: Learned intensity functions for neurons in the cat primary visual cortex.

Appendix C. Reviews

C.1. Review 1

2: (accept) The authors designed a point process model called multivariate mutually regressive point processes. It captures both excitatory and inhibitory interactions between temporal events in a continuous-time regime. The paper also derived a full Bayesian MCMC inference algorithm for this model using Polya-Gamma augmentation and Poisson thinning. Experiments on multi-neuronal spike train recordings showed the method is able to discover inhibitory patterns between neurons.

The paper is very well-written and contains an enough level of technical details. Although I'm not an expert in point processes, I think its contribution is big enough for acceptance, in the sense that it designed a new model to explicitly learn the inhibitory interactions of events in continuous time domain. Meanwhile, I have some minor questions and suggestions:

1) In the experiment section, the authors manually assigned some value to c and delta. Is there any reason why they are not learned, in a similar way like w and lambda*?

2) Since the proposed point process isn't a Poisson process, it could be misleading to say so at the beginning of section 3. For general point processes, a probably better reference to cite might be Rubin, I. (1972). Regular point processes and their detection. IEEE Transactions on Information Theory, 18(5), 547-557.

3) Some typos: In Eq.(27), the curly brace should end before the product symbol. In Eq.(31), there shouldn't be a product over n at the beginning of the second line.

4) It's probably due to the limit of space, but more comparisons with previous models, both theoretically and experimentally, would be very helpful to better show the value of this work.

C.2. Review 2

2: (accept) Solid, promising work. I look forward to reading the full-paper version somewhere in the near future.

A few detailed comments:

– The accent on the 'o' in 'Polya' is in the wrong direction throughout the paper.

– Nice introduction. Clear, concise review of the background. It raises an obvious question that is not addressed: where, how, when, is the discrete-time approximation problematic? A convincing argument that the Poisson GLM falls short in some way would give much stronger motivation than "There is currently a significant gap between real-world applications and statistical theory..."

– Fix this on p. 3: (Polson et al., 2013; Linderman et al., 2015, 2016, e.g.).

– Figure 1: legend clashes with the plot.

– In the experiments section, why are only 10 cells used? Is the full set intractable? How were the 10 cells chosen? The way in which they were chosen (e.g., randomly sampled) may bias what you infer, i.e., by sampling uniformly at random, you turn a network with spatial and/or functional structure into a vertex-exchangeable one.

– Section 4 is hard to read, presumably because it was written to fit the space constraint.

– ”Strong inhibitory effects are discovered from neuron 6 on neuron 10, from neuron 1 on neuron 9, from neurons 6 and 4 on neuron 5, from neurons 6 and 1 on neuron 3 and from neuron 2 on neuron 1...” — turn this into a picture!

I didn’t go through the appendix in any detail, but it ”looks right.”

C.3. Review 3

1: (weak accept) The paper is well written. But, the following works (by Donner and Opper)

https://journals.aps.org/pre/abstract/10.1103/PhysRevE.96.062104

https://arxiv.org/pdf/1808.00831.pdf

–which are not in the reference list of the manuscript–, limit considerably the novelty of work. There are of course some difference between the current and previous works such as the definition of aggregated temporal infuence and optimizing the rate parameter \lambda_n*. This could (or not) lead to better results than the results by [Donner and Opper]. But the essential idea is strongly similar to the previous works.

C.4. Review 4

1: (weak accept) This paper introduced a point process model allowing for excitatory and inhibitory interactions in the continuous time setting. A Bayesian inference algorithm is derived, and an analysis is run on neural spike train recordings.

This is a well-written paper with a nice idea and algorithm. The detail in the main paper and appendices were particularly nice.

My main concern is regarding the experiment. How important are the values of the parameters of various priors? Why is the bias term in the sigmoid function given a high prior mean value (why is an asymmetry between learnt parameters for excitatory vs inhibitory effects desired)? Also, looking at figures in the Appendix, I was not sure about some of the inhibitory effects inferred (as one example, to my eye, it does not look like there is a strong inhibitory effect from neuron 1 on neuron 9).