Detection of Short-Term Temporal Dependencies in Hawkes Processes with Heterogeneous Background Dynamics

Yu Chen[1,*] **Fengpei Li**[1,*] **Anderson Schneider**[1] **Yuriy Nevmyvaka**[1] **Asohan Amarasingham**[2] **Henry Lam**[3]

[1]Machine Learning Research, Morgan Stanley, New York, NY
[2]Department of Mathematics and Biology, City College and The Graduate Center, City University of New York, New York, NY
[3]Department of Industrial Engineering & Operations Research, Columbia University, New York, NY
[*]Authors have equal contribution

Abstract

Many kinds of simultaneously-observed *event sequences* exhibit mutually exciting or inhibiting patterns. Reliable detection of such temporal dependencies is crucial for scientific investigation. A common model is the Multivariate Hawkes Process (MHP), whose impact function naturally encodes a causal structure in Granger causality. However, the vast majority of existing methods use a transformed *standard* MHP intensity with a constant baseline, which may be inconsistent with real-world data. On the other hand, modeling irregular and unknown background dynamics directly is a challenge, as one struggles to distinguish the effect of mutual interaction from that of fluctuations in background dynamics. In this paper, we address the short-term temporal dependency detection issue. We show that maximum likelihood estimation (MLE) for cross-impact from MHP has an error that can not be eliminated, but may be reduced by an order of magnitude using a heterogeneous intensity not for the target HP but for the interacting HP. Then we propose a robust and computationally-efficient modification of MLE that does not rely on the prior estimation of the heterogeneous intensity and is thus applicable in a data-limited regime (e.g., few-shot, unrepeated observations). Extensive experiments on various datasets show that our method outperforms existing ones by notable margins, with highlighted novel applications in neuroscience.

1 INTRODUCTION

A substantial amount of timestamp data manifest as a sequence of apparently irregular and asynchronous events. These are recorded in continuous time and observed in domains such as computational biology (e.g., neuronal spike trains [Kass and Ventura, 2001, Pillow et al., 2008], genomic events [Reynaud-Bouret and Schbath, 2010]), quantitative finance (e.g., limit order book modeling for high-frequency trading [Bacry et al., 2015, Bowsher, 2007]), credit risk modeling [Errais et al., 2010]), social media user activity [Farajtabar et al., 2015, Zhou et al., 2013a], e-healthcare ([Wang et al., 2018]) and seismology (e.g., earthquake aftershock [Ogata, 1988]). Besides asynchronicity, such sequence data often exhibit mutual interaction patterns in which the occurrence of one event can excite or inhibit the likelihood of another. For example, news-driven trading in behavioral finance studies the mutual excitation between investor-sentiment shocks and negative price jumps [Yang et al., 2018], while in cortical networks inhibitory connectivity in firing-rate between neurons and synapses may underlie memory maintenance [Mongillo et al., 2018]. Such an interaction patterns has been variously called a *temporal dependency* [Zuo et al., 2020], *cross-correlation* [Zhang et al., 2020], a *coupling effect* [Pillow et al., 2008] or *Granger causality* [Xu et al., 2016]. As [Eichler et al., 2017] note, although stand-alone notions of Granger causality can not establish cause-effect links, the detection of temporal dependencies remains useful for both prediction and scientific investigation.

Temporal point processes (TPP) [Cox and Isham, 1980] are a powerful tool for modeling event sequences. Multivariate Hawkes processes (MHP) [Hawkes, 1971], as a special type of TPP, have been widely used as the *de facto* tool for capturing temporal dependencies among event processes (see above, e.g.,[Bacry et al., 2015, Farajtabar et al., 2015, Wang and Zhang, 2022, Zuo et al., 2020]). An MHP models occurrence probability using a history-dependent conditional *intensity* and its *impact function* (also called *coupling filter*, *trigger kernel*, *influence function*, see [Pillow et al., 2008, Zhou et al., 2021b, 2013a]) is particularly well-suited to detect mutual excitatory effects. Inhibitory effects can also be incorporated, but some nonlinear link function is required to map the MHP intensity into $\mathbb{R}^+$ (e.g., notably a clip-

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

ping function $x^+ = \max(x, 0)$ in [Hansen et al., 2015] or sigmoid function in [Zhou et al., 2021b]).

Despite the expressiveness of impact functions, the background component in MHP intensity is assumed to be time-invariant. Possibly due to the extra modeling difficulty entailed, virtually all existing studies on MHP use, implicitly or explicitly, nonlinear transform of standard MHP intensity with constant baseline, including modern DL-based methods (e.g., Transformer HP [Zuo et al., 2020] HP in infinite relational model or Dirichlet mixture model [Blundell et al., 2012, Xu and Zha, 2017], sigmoid nonlinear MHP with Pólya-Gamma variable augmentation [Zhou et al., 2021b], self-attentive HP and recurrent neural network [Zhang et al., 2020]). Notable exceptions which incorporate temporal heterogeneity include [Mei and Eisner, 2017], a neurally self-modulating HP with LSTM and [Zhou et al., 2021a], where a state-switching latent process is proposed (yet still assuming constant background within each state) and [Hawkes, 2018] where the heterogeneous background is briefly discussed as a generalization of MHP to represent "exogenous economic activity".

However, real-world event dynamics are often decisively *temporally heterogeneous*. For example, Twitter has information bursts spurred by exogenous events (e.g., breaking news or sports games)[Wang and Zhang, 2022], the firing of neurons is commonly driven by varying visual stimuli [Siegle et al., 2021], and trading activity has a diurnal variation (e.g., more trades occur around market open/close than around noon [Bowsher, 2007]). Under *unknown* heterogeneous dynamics, temporal dependency detection is difficult as one struggles to distinguish the effect of mutual interaction from that of background intensity fluctuation (e.g., did the arrival of orders for stock A stimulate that for stock B, or did they both simply experience a nonlocalized spike in trading activity?).

In this paper, we show that the maximum likelihood estimation (MLE) of short-term temporal dependency detection for standard MHPs has non-negligible errors in the presence of heterogeneous background dynamics. However, this error decreases by an order of magnitude (in terms of impact window or kernel width) if the heterogeneous background between the *target* HP (recipients of the impact) and *source* HP (initiators of the impact) is *uncorrelated* (or *orthogonal* in the Hilbert space sense, $L_2[0, T]$ or $C[0, T]$, where T is observation horizon). Thus, loosely speaking, MHP can still estimate short-term cross-impact reasonably well, unless the heterogeneous intensity between the target HP and source HP shares common/correlated background dynamics. Building on this insight, we propose a robust and computationally-efficient modification of MLE, which utilizes a nonparametric estimate of heterogeneous intensity – not of the target HP, but of the source HP. By focusing on the background intensity of the source, we reduce the inference difficulty, and the error, due to the coupling between the target HP background and impact function, by regressing the commonly-varying background out of the target HP intensity.

The contribution of this paper can be summarized as:

- To the best of our knowledge, our work is the first to formally report and analyze the error of MLE of short-term temporal dependencies in MHPs due to heterogeneous background dynamics. We investigate the relation between estimation error and background-correlation among interacting HPs, which motivates a novel method to reduce the error.
- Through extensive numerical experiments, we show that our method exhibits superior performance and is robust, cost-efficient, applicable in a data-limited regime (e.g., when lacking repeated observations), and suitable for inference.
- Finally, we apply our method to mouse visual cortex data and discover distant interactions between neurons on a fine timescale in both top-down and bottom-up pathways, showcasing the method's direct applicability in neuroscience.

2 RELATED WORK

Hawkes process. Many efforts have been devoted to detecting temporal dependency among point processes, e.g. [Chwialkowski and Gretton, 2014, Gunawardana et al., 2011]. Among point processes, Hawkes processes stand out as the most commonly used tool for modeling complex temporal dependencies in event sequences. The paper [Eichler et al., 2017] established the link between Granger causality and impact functions in MHP and many methods are proposed to learn the temporal dependency in MHP, via group sparsity, [Xu et al., 2016], nonparametric learning using Euler-Lagrange equation [Zhou et al., 2013b], isotonic nonlinear link function [Wang et al., 2016], online learning [Yang et al., 2017] and modern DL-based methods (see intro, [Zuo et al., 2020, Blundell et al., 2012, Xu and Zha, 2017, Zhang et al., 2020]. However, these methods use direct or nonlinear transform of *standard* MHP time-invariant base intensity, overlooking the heterogeneity in event dynamics. Notably, [Mei and Eisner, 2017] implicitly allows for heterogeneity. Latent variable augmentation is proposed in [Zhou et al., 2021b,a, 2022, 2020] to incorporate the time-varying background, but the modeling of heterogeneity typically relies on piecewise constants. Moreover, most methods are data-intensive (e.g., as reported in [Yang et al., 2017], methods as [Zhou et al., 2013b] require more than $10^5 d$ arrival data to obtain good results on $d \leq 5$ event streams) and computationally-extensive (e.g., MCMC, EM algorithm or complex neural architecture) which is unsuitable for inference in data-limited regimes. Indeed, often in practice, only short/unrepeated sequences are available

[Salehi et al., 2019], which not only amplifies the risk of overfitting but also makes estimation of heterogeneous background infeasible.

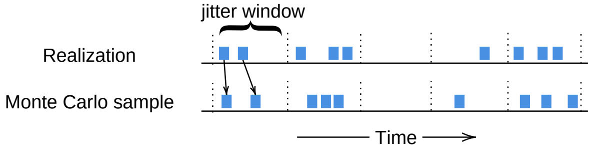

Figure 1: Construction of Monte Carlo samples from the conditional null hypothesis in the conditional-inference based CCG technique. Blue dots are timestamps.

Conditional inference based cross-correlogram (CCG). Heterogeneous dynamics are ubiquitous in neuroscience [Farajtabar et al., 2015]. Due to the limitations of TPP and MHP in this regime, a popular method in neuroscience for detecting temporal dependencies in cross-correlograms (CCG) is via conditional inference. Conditional hypothesis testing with a carefully designed null hypothesis can bypass the background heterogeneity issue. Particularly, given realizations of two point processes, CCG assesses temporal dependence between events by testing hypotheses about conditional distributions of CCG-statistics, conditioning on coarse timescale-statistics which reflect background dynamics. As shown in [Amarasingham et al., 2012], the method relies on conditional inference, where the samples from the null are generated by: shifting the timestamps within each *jitter window* (reflecting prior knowledge on the timescale of interactions) by a random amount, which is small enough to preserve coarse-timescale statistics, but large enough to break the finely-timed interaction pattern (see Figure 1). However, this method requires prior knowledge of timescales and assumes that the timescale of background activity (parameterized as the jitter window width) is larger than that of the interaction effect (See Figure 3) and discussion below. Additional details in Appendix A.3, C.6. Also, the outcome of the hypothesis test alone does not measure the strength of the coupling effect directly.

3 ANALYSIS AND METHODS

3.1 BASIC CONCEPTS

A temporal point process is a stochastic process whose realization consists of a list of discrete event timestamps $\{t_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}^+$, which can be equivalently represented by a counting process $\{N(t), 0 \le t \le T\}$ [Daley and Vere-Jones, 2008]. Formally, given a probability triple $(\Omega, \{\mathcal{H}_t\}_{0\le t\le T}, \mathbb{P})$, $N(t) := N((0,t], \omega)$ is a realization (i.e., $\omega \in \Omega$) of counting measure N for the number of points in $(0,t]$ and $\mathcal{H}_t$ is the σ-algebra generated from $N(B)$ for Borel subsets $B \subseteq (0,t]$ (or $(-\infty, t]$, we do not distinguish them here). The intensity of the point process is $\lambda(t) := \lim_{\delta\to 0} \frac{1}{\delta}\mathbb{P}(N(t+\delta) - N(t) > 0|\mathcal{H}_t)$. It can be shown (see [Ogata, 1978]) for $\mathcal{H}_t$-progressively measurable $\lambda(t), f(t)$ with left continuous (thus predictable) $f(t)$ that $\mathbb{E}[dN(t)|\mathcal{H}_t] = \lambda(t)dt$ and

$$\begin{aligned}\mathbb{E}\int_0^T f(t)dN(t) =& \mathbb{E}\int_0^T f(t)\mathbb{E}[dN(t)|\mathcal{H}_t] \\ =& \mathbb{E}\int_0^T f(t)\lambda(t)dt. \qquad (1)\end{aligned}$$

sssss For the multivariate Hawkes process, the density has the form

$$\lambda_j(t) = \alpha_j + \sum_{i=1}^{d} \int_0^t h_{i\to j}(t-s)dN_i(s) \qquad (2)$$

for $1 \le i, j \le d$, where d is the dimension (number of event streams), α_j is the *baseline* intensity for process N_j and $h_{i\to j}$ is the impact function from N_i to N_j. Standard MHP models mutual excitatory behavior and requires $h_{i\to j} \ge 0$ to avoid negative intensity which is meaningless. However, one can simply set $\lambda \leftarrow \max(\lambda, 0)$ [Hansen et al., 2015] to extend MHP for modeling mutual inhibitory behavior.

3.2 HETEROGENEOUS EVENT DYNAMICS

The standard MHP assumes the baseline intensity α to be a constant (2), which is incongruous with the heterogeneous event dynamics frequently observed in real-world scenarios. To accommodate heterogeneity, instead of using (2) as building blocks to construct a complex structure, we directly proposed a generalized MHP intensity for $1 \le i, j \le d$:

$$\lambda_j(t) = \alpha_j + f_j(t) + \sum_{i=1}^{d} \int_0^t h_{i\to j}(t-s)dN_i(s) \qquad (3)$$

where $f_j(t)$ is the fluctuation in the background intensity. For now, we do not restrict whether f_j is stochastic or deterministic, but simply assume it is $\mathcal{H}_t$-adapted. For identifiability between α and f, we assume $\int_0^T \mathbb{E}[f(t)]dt = 0$ (or more generally $\int_0^P f(t)dt = 0$ if it is deterministic and perodic with peroid P or $\mathbb{E}[f] = 0$ if $f(t)$ is stationary).

The main approaches for learning MHP falls under two directions: maximum likelihood-based (MLE) approaches [Ogata, 1978, Zhou et al., 2013a, Yang et al., 2017] and moment-matching flavored approaches based on higher-order statistics [Da Fonseca and Zaatour, 2014]. Due to the unknown statistical property of f, the moment-based methods are not applicable for (3). To investigate the applicability of the MLE approach for (3), we study a representative model for subsequent discussion. However, we emphasize that our proposed method applies generally to models from (3).

3.3 REPRESENTATIVE MODEL

Consider two point processes N_i, N_j as shown in Figure 2. The intensity functions are,

$$\begin{aligned}\lambda_j(t) &= \alpha_j + f_j(t) + \int_0^t h_{i\to j}(t-s)dN_i(s) \\ \lambda_i(t) &= \alpha_i + f_i(t)\end{aligned} \tag{4}$$

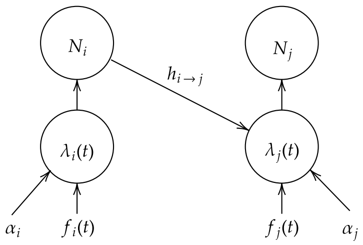

Figure 2: Illustrative MHP with a heterogeneous background. Two events stream N_i, N_j with intensities $\lambda_i(t), \lambda_j(t)$, baseline $\alpha_i + f_i(t), \alpha_j + f_j(t)$ and the one-way impact function $h_{i\to j}$.

where $h_{i\to j}$ is the impact function and $f_i(t), f_j(t)$ are unknown fluctuations. There are various methods of learning the form $h_{i\to j}$ with data-driven and nonparametric techniques ([Zhou et al., 2013b, Xu et al., 2016, Yang et al., 2017]. To facilitate the discussion of MLE, we assume the form of impact has been learned within a 1-D parametric family $h_{i\to j}(\cdot) \in \{\theta \cdot \mathbf{1}_{[0,\sigma_h]}(\cdot)\}_{\theta\in\Theta}$ which is widely applied in neuroscience (here, $\mathbf{1}_{[0,\sigma_h]}(t) = 1$ if $0 \le t \le \sigma_h$ and 0 otherwise). We set the ground truth impact to be $h_{i\to j} = c \cdot \mathbf{1}_{[0,\sigma_h]}$ for a given $c > 0$. Moreover, we focus on the recovery of impact function (i.e., estimation of c) and treat other parameters as *nuisance* parameters, as in *profile likelihood*[Murphy and Van der Vaart, 2000].

In MHP (2), not considering f_j, one parameterizes λ_j as

$$\lambda_{\boldsymbol{\theta}}(t) = \theta_1 + \theta_2 \int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s)dN_i(s), \tag{5}$$

which is misspecified and maximizes the log-likelihood:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \ell(\boldsymbol{\theta}; \mathcal{H}_T) \\ &= -\int_0^T \lambda_{\boldsymbol{\theta}}(t)dt + \int_0^T \log \lambda_{\boldsymbol{\theta}}(t) dN_j(t),\end{aligned}$$

see, e.g., [Ogata, 1978]. In the misspecified model, one would expect $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_{KL}$, the minimizer in KL-divergence information criterion [White, 1982]:

$$\boldsymbol{\theta}_{KL} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \Lambda(\boldsymbol{\theta}) := \mathbb{E}\ell(\boldsymbol{\theta}),$$

under suitable regularity conditions, including μ-strong convexity and L-Lipschitz gradient of Λ. We want to quantify the error between $[\boldsymbol{\theta}_{KL}]_2$ and c. We list technical conditions in Appendix B, along with proofs for the following results.

Proposition 1. *Under regularity conditions specified in Appendix B, for deterministic f_i and f_j in* (3)*, the error satisfies*

$$|[\boldsymbol{\theta}_{KL}]_2 - c| = \Theta\left(\left|\int_0^T \frac{f_i(t)f_j(t)}{\alpha_j + c}dt \cdot \sigma_h + o(\sigma_h)\right|\right).$$

Proposition 2. *Under the same condition as in Proposition 1, if f_i and f_j are stationary, the error satisfies*

$$|[\boldsymbol{\theta}_{KL}]_2 - c| = \Theta\left(\left|\frac{Cov(f_i, f_j)}{\alpha_j + c}\sigma_h + o(\sigma_h)\right|\right).$$

where the big-Θ notation stands for a growth function with the same rate in upper and lower bound, i.e. $f(x) = \Theta(g(x))$ if there exists $0 \le m \le M$ s.t. $mg(x) \le f(x) \le Mg(x), \forall x$.

Proposition 1 and 2 suggest that, under heterogeneous event dynamics, the error in estimating the impact function scales linearly with σ_h, with the coefficient determined by the "inner product" between f_i and f_j. In fact, if we define $\langle f_i, f_j\rangle = \mathbb{E}\int_0^T f_i(t)f_j(t)dt$, then we can unify (and generalize to a non-stationary case) the result in Proposition 1 and 2. We see that, for short-term temporal dependency detection $\sigma_h \to 0$, the ratio between estimation error and interaction timescale σ_h is non-vanishing and non-negligible unless the two HPs have *uncorrelated* background ($\langle f_i, f_j\rangle = 0$).

How could one reduce the order of this error term? The most natural way is to observe or estimate f_j directly. Indeed, given access to f_j, MLE is no longer misspecified. However, as discussed in [Zhou et al., 2020], the "exogenous component" (the baseline intensity) and the "endogenous" component (the impact function) are "coupled" in the likelihood, which hampers inference. In [Zhou et al., 2020], a *branching* structure is used to decouple these two components in HP, which does not apply to MLE because when the same, typically limited data are used to estimate both f_j and $h_{i\to j}$, the results are generally non-reliable (indeed, a naive use of MLE for fitting both would result in delta measures around the event timestamps for N_j). However, since the correlation between f_i and f_j results in a large error, one conjectures whether estimation of f_i, or entities highly correlated with f_i, could help regress out the common varying intensity out of f_j. Indeed, we have the following:

Proposition 3. *Under the same condition as in Proposition 2, if we let $r := \max\{\|f_i - \mathbb{E}f_i\|_\infty, \|f_j - \mathbb{E}f_j\|_\infty\}$, if we have access to $g = \frac{f_i - \mathbb{E}[f_i]}{\sqrt{Var(f_i)}}$ (i.e., normalized basis for f_i) in the likelihood* (5) *so that one parameterizes*

$$\lambda_{\boldsymbol{\theta}}(t) = \theta_0 + \theta_1 g + \theta_2 \int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s)dN_i(s), \tag{6}$$

then

$$\begin{aligned}[\boldsymbol{\theta}_{KL}]_1 =& \mathbb{E}[(f_j - \mathbb{E}[f_j])g] + o(r^2 + \sigma_h), \\ [\boldsymbol{\theta}_{KL}]_2 =& o(r^2 + \sigma_h).\end{aligned}$$

Although we can not directly observe f_i, Proposition 3 suggests that using f_i as a basis may reduce the error. Moreover, the form of $[\boldsymbol{\theta}_{KL}]_1 \approx \langle f_j, g\rangle$ also suggests using a "project f_j on f_i" as basis to modify the MLE.

3.4 PROPOSED METHOD

Inspired by the analysis above, we now propose our modification for estimating impact. In particular, we minimize the following expression modified from the likelihood function $\tilde{\ell}$:,

$$\min_{h_{i\to j},\beta_j,\beta_w,\sigma_w} \left\{ -\sum_{s\in N_j} \log\tilde{\lambda}_j(s) + \int_0^T \tilde{\lambda}_j(s)\mathrm{d}s \right\} \tag{7}$$

$$\tilde{\lambda}_j(t) := \left(\beta_j + \beta_w\, \overline{\mathbf{s}_i}(t) + \int_0^t h_{i\to j}(t-s)dN_i(s)\right)_+ \tag{8}$$

$$\overline{\mathbf{s}_i}(t) = \int_0^T W(t-s;\sigma_w)dN_i(s) \tag{9}$$

where $\overline{\mathbf{s}_i}$ can be regarded as the coarsened point process smoothed by a Gaussian kernel $W(\tau;\sigma_w) = \frac{1}{\sqrt{2\pi\sigma_w^2}}\exp(-\frac{\tau^2}{2\sigma_w^2})$ with scale σ_w, serving as a substitute basis for f_i. We also specify an algorithm that can be implemented in continuous time, which does not require one to discretize the time points [Eden and Brown, 2008, Foufoula-Georgiou and Lettenmaier, 1986], so that the memory requirement is proportional to the number of time points instead of the number of time bins. The optimization algorithm is detailed in Appendix A. Empirical and theoretical analysis of the estimator will be discussed in section 4.1.

3.5 OTHER USE CASES OF THE METHOD

Before experiments, we present some generality in the application of the method, with details left to Appendix D.

Hypothesis testing see Appendix D.2. We compare our model with conditional inference via CCG and standard MHP, in hypothesis testing. Both our model and CCG have proper uniform p-value distribution under the null of no interaction [Wasserman, 2004, Theorem 10.14], where the standard MHP fails. Moreover, our method is also more powerful/sensitive at detecting weak signals with small sample sizes, see Figure 3. Figure 3 shows a simulation example of fine timescale interaction between two point processes. Synthetic data is generated by HP with one process inhibiting the other and a common fluctuating background in Figure 3A. Figure 3B is the result of the conditional inference via cross-correlogram (CCG). The curve is mostly in the negative region indicating some inhibitory influence, yet the majority part of the curve stays within the acceptance band (i.e., not statistically significant). Figure 3CD show the result of the standard MHP vs our method, where the impact function is represented as lag period. As shown, our method accurately detects the inhibitory relation and the estimated error is close to the true function (red curve), with the improvement compared to CCG in the statistical power and standard MHP in terms of error. A similar observation in real data will be shown in Figure 7. A more detailed comparison between these models is in Appendix D.2.

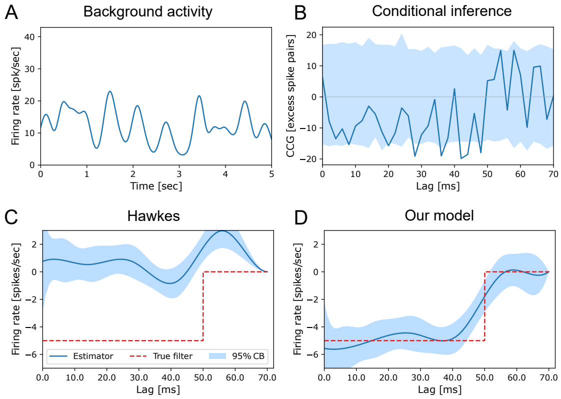

Figure 3: Impact function estimation with background fluctuation in simulation. **A**: Shared background intensity. **B**: CCG-based conditional inference. The 95% acceptance band is constructed using Monte Carlo samples from the null distribution. **C**: Standard MHP. The red curve is the ground truth. **D**: Our model. The band in C and D is also 95% pointwise CI.

Non-parametric fitting for the impact function see Appendix D.1. Our method does not have constraints on modeling the impact function, which can be easily extended to non-parametric fitting. One option is the general additive model using splines [Pillow et al., 2008, Hastie et al., 2009, ch. 5]. By leveraging the integral trick (Appendix A), time points do not need to be discretized and computational cost is small.

Bayesian inference see Appendix D.3. The method can be adopted for Bayesian inference where the uncertainty of the smoothing kernel width σ_w is evaluated using a sampling-based inference algorithm. The simulation shows that incorporating the uncertainty of σ_w does not affect the estimation of the temporal dependency significantly.

4 EXPERIMENTS

In this section, we empirically verify the method through multiple simulation studies, then apply the new tool to the neuroscience dataset where we discover a network of interacting neurons on a fine timescale. For simulations, continuous-time point processes are generated using Lewis' thinning algorithm [Lewis and Shedler, 1979, Ogata, 1981]. The gradient descent-based optimization algorithm is in Appendix A. Our code is available `https://github.com/AlbertYuChen/point_process_coupling_public`.

4.1 SIMULATION STUDY

4.1.1 Toy Example with background fluctuation

In this synthetic dataset, the dynamic baselines have known form so that their correlation or the "inner product" between the source and target processes, as discussed in Section 3.3, can be calculated in closed-form. The background activities are $f_i(t) = A\sin(2\pi(t - \phi_{\text{rnd}}))$, $f_j(t) = A\sin(2\pi(t - \phi_{\text{rnd}} - \phi_{\text{lag}}))$, where A is the amplitude, T is the length of the trial. We sample $\phi_{\text{rnd}} \sim \text{Unif}(0, 1)$ and set it to vary from trial to trial so the same background is never repeatedly observed. Here ϕ_{lag} controls the correlation between f_i, f_j, which we quantify using the *normalized* dot product $\langle f_i, f_j \rangle := \frac{1}{TA^2}\int_0^T f_i(s)f_j(s)\mathrm{d}s$. When $\phi_{\text{lag}} = 0$ and 0.5, the dot product achieves the largest positive and negative value respectively; when $\phi_{\text{lag}} = 0.25$, the dot product is zero.

For the problem we are considering, short-term temporal dependency detection with dynamic background, there really is no "state-of-the-art" model as we are not mainly interested in predicting future observations, but we aim at gaining insight into the relationship between features and responses for scientific discovery, which is a more challenging task [Fan et al., 2020]. Although many recent point process models, such as [Mei and Eisner, 2017, Zhang et al., 2020, Zuo et al., 2020], are designed for the prediction task, one popular representative deep learning-based model by Mei and Eisner [2017] using recurrent neural networks is included as the baseline model. The performance of three models are compared: standard MHP, our model, and Neural Hawkes [Mei and Eisner, 2017]. Some other deep learning models are not considered due to the convoluted black-box structure. For example in [Zhang et al., 2020], the intensity function is

$$\begin{aligned}\lambda_i(t) =& \text{softplus}(\mu_{u,i+1} + \\ &(\eta_{u,i+1} - \mu_{u,i+1})\exp(-\gamma_{u,i+1}(t - t_i))),\end{aligned}$$

where the variables μ, η, γ are all functions of latent variables obtained through attention network. Another example is [Zuo et al., 2020], where the intensity function is

$$\lambda_k(t) = f_k(\alpha_k \frac{t - t_j}{t_j} + \boldsymbol{w}_k^T \boldsymbol{h}(t_j) + b_k),$$

where t_j is the last event (not necessarily type k) and h is the latent variable that carries more history information extracted from transformers. Just by observing the intensity form above, one realizes that these models, designed for the event sequence prediction, are very difficult to draw inference on the coupling effect. The method in [Mei and Eisner, 2017] is the simplest framework we found where one can split out the coupling effect with minimum modification of the model.

The impact function is the square window impact function with a given width, so only the amplitude needs to be estimated. Neural Hawkes takes intervals of the superimposed point processes one by one in sequence. The impact function from source to target is modeled as

$$\boldsymbol{c}(t) = \bar{\boldsymbol{c}}_{i+1} + (\boldsymbol{c}_{i+1} - \bar{\boldsymbol{c}}_{i+1})\mathbb{I}_{[0,\sigma_h]}(t - t_i^{\text{source}}), \tag{10}$$

$$\boldsymbol{h}(t) = \boldsymbol{o}_i \odot \tanh(\boldsymbol{c}(t)), \tag{11}$$

$$\lambda_{\text{target}} = \left(\boldsymbol{W}_{\text{target}}^T \boldsymbol{h}\right)_+, \tag{12}$$

which is slightly modified for the context (original kernel in [Mei and Eisner, 2017] is exponential). The impact function is extracted from the model (the original model does not directly offer an estimated parameter) as $h_{\text{source}\to\text{target}}(t) = \boldsymbol{W}_{\text{target}}^T \left[\boldsymbol{o}_i \odot \tanh((\boldsymbol{c}_{i+1} - \bar{\boldsymbol{c}}_{i+1})\mathbb{I}_{[0,\sigma_h]}(t)\right]$ which could capture a time point's impact on the intensity. Instead of modeling multiple points in the history at once as in the standard MHP, Neural Hawkes considers non-linear mapping, which only receives one last interval, while the history effect is carried over $\boldsymbol{c}_{i+1}$, $\bar{\boldsymbol{c}}_{i+1}$, and $\boldsymbol{o}_i$ through a recurrent neural network. The result is shown in Figure 4 while details are left in Appendix C.1.

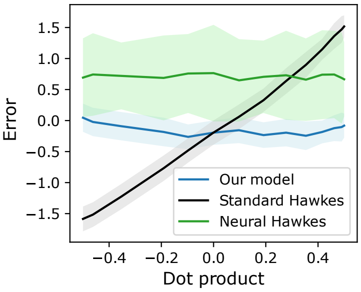

Figure 4: A comparison of impact function estimation between standard MHP, Neural Hawkes, and our model under dynamic background. The confidence band is created from 100 simulations.

As shown, the bias of standard MHP is nearly linearly correlated with the dot product, as suggested by theoretical analysis. The error of Neural Hawkes is less susceptible to this correlation, which corroborates the ability of a recurrent structure to capture the interaction effect despite dynamic background. However, the error and variance of the impact estimation from Neural Hawkes are visibly non-negligible. This is likely due to the fact that neural network models typically need large datasets for training. In contrast, our model performs satisfactorily in this example.

4.1.2 Background kernel smoothing

The kernel-smoothed basis in (9) plays a key role in our method. This section studies the relationship between the kernel width and the error of the estimator. In special cases, we are able to approximate the behavior of the estimator with an analytical formula. Following the model framework in (4), assuming the background activity is generated similar to the *linear Cox process* [Diggle, 1985] or the *cluster process* [Daley and Vere-Jones, 2003, Definition 6.3.I.]:

$$f_i = f_j := \sum_i \phi_{\sigma_I}(t - t_i^c) \tag{13}$$

where $\phi_{\sigma_I}(\cdot)$ is some positive and even function, i.e., $\phi_{\sigma_I}(\cdot) > 0$ and $\phi_{\sigma_I}(\tau) = \phi_{\sigma_I}(-\tau)$. Here t_i^c are the centers of the windows generated by a Poisson process with intensity ρ. f_i is second-order stationary with a *reduced covariance density* defined as follows (also see Appendix E).

$$\begin{aligned}\breve{c}_\Lambda(u) :=& \mathbb{E}[f_i(x) f_i(x+u)] - \mathbb{E}[f_i(x)]\mathbb{E}[f_i(x+u)]\\ =& \rho[\phi_{\sigma_I} * \phi_{\sigma_I}](u)\\ \breve{c}_N(u) :=& \mathbb{E}\left[\frac{dN_i(x) dN_i(x+u)}{(\mathrm{d}t)^2}\right] - \mathbb{E}\left[\frac{dN_i(x)}{\mathrm{d}t}\right]\mathbb{E}\left[\frac{dN_i(x+u)}{\mathrm{d}t}\right]\\ =& \rho \cdot [\phi_{\sigma_I} * \phi_{\sigma_I}](u) + (\rho + \alpha_i)\delta(u)\end{aligned} \tag{14}$$

which describes the smoothness of background activity, and α_i is the constant in (4). If adjacent points with lag u have larger covariance $\breve{c}_\Lambda(u)$, the background would be smoother. The impact functions are $h_{i\to j}(t) = \alpha_{i\to j} h(t)$, with amplitude to be fitted, for example $h(t) = \mathbb{I}_{[0,\sigma_h]}(t)$. Then the error in model (7) may be approximated as,

$$\text{error}(\hat{\alpha}_{i\to j}) \approx \frac{\langle W, W\rangle_{\breve{c}_N} \langle h, \mathbf{1}\rangle_{\breve{c}_\Lambda} - \langle h, W\rangle_{\breve{c}_N} \langle W, \mathbf{1}\rangle_{\breve{c}_\Lambda}}{\langle W, W\rangle_{\breve{c}_N} \langle h, h^-\rangle_{\breve{c}_N} - \langle W, \mathbf{1}\rangle_{\breve{c}_\Lambda}^2} \tag{15}$$

$\mathbf{1}$ is the constant and $h^-(\tau) = h(-\tau)$. The special inner product here are defined as $\langle g_1, g_2\rangle_{\breve{c}} := \int [g_1 * g_2](s)\breve{c}(s)\mathrm{d}s$ with $*$ denoting the convolution. The derivation of the analytical formula is in Appendix E. Simulation and analytical results are presented in Figure 5. The error and log-likelihood are plotted as functions of the smoothing kernel width σ_w in (9). The MLE, indicated by the vertical line in Figure 5, achieves a small error that agrees with the example in section 4.1.1. Interestingly, when the kernel width is too small or too large, including the theoretical limits by taking $\sigma_w \to 0$ or $\sigma_w \to \infty$, the model fails under heterogeneity. In this case, the error is close to that of standard MHP. Details are in Appendix C.2.

In Figure 5, when σ_w is between 20 ms and 120 ms, the error can be negative. The error as a function of the background smoothing kernel has two roots. The roots are related to the timescale of the coupling effect σ_h and the timescale of the background σ_I as in (13). In Figure 6, if σ_I increases, the root on the right, corresponding to the MLE, will move toward the right, as the background smoothing kernel W captures the fluctuation of the background. If σ_h increases, the root on the left will move toward the right. This can be intuitively interpreted by (15). Let W_h be the kernel with $\sigma_w \approx \sigma_h$, then $\langle W_h, W_h\rangle_{\breve{c}_N} \langle h, \mathbf{1}\rangle_{\breve{c}_\Lambda} \approx \langle h, W_h\rangle_{\breve{c}_N} \langle W_h, \mathbf{1}\rangle_{\breve{c}_\Lambda}$. So $\sigma_w = \sigma_h$ is close to the root of (15). Changing the amplitude of the impact function $\alpha_{i\to j}$ in a certain range does not influence the bias curve. More details are in Appendix C.3.

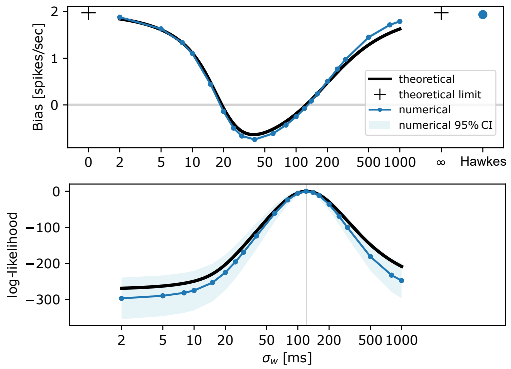

Figure 5: Error and likelihood of the estimator as functions of background smoothing kernel width σ_w in (9). Numerical and theoretical results as in (15) are shown in blue and dark respectively. The error of standard MHP is the blue dot on the right.

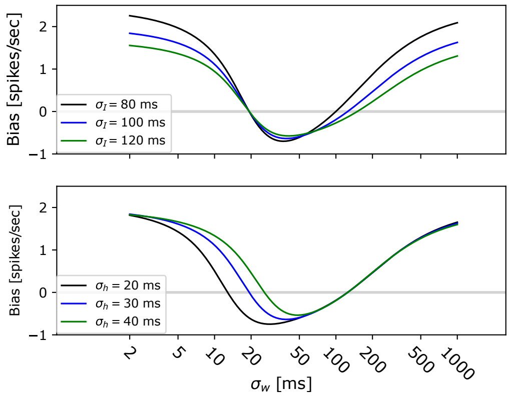

Figure 6: Error function with different background timescales σ_I (top) or coupling effect timescales σ_h (bottom). Numerical results match the theoretical results well, so only theoretical results are presented according to (15).

4.1.3 Two-way cross connections and self-connections

The model in Figure 2 only shows one cross-connection $i \to j$. This simulation scenario includes the most general two-way MHP cross/self connections between processes ($i \to j$ and $j \to i$), and self-connections ($i \to i$ and $j \to j$). The comparison between our model and standard MHP is in Table 1. Details of the experiment are in Appendix C.4.

Our model considerably outperforms the standard MHP in estimating cross-impact connections. However, both models perform poorly on self-connection estimation, as they are considered nuisance parameters in our method.

	Our model		Standard Hawkes	
	i	j	i	j
i	$1.70(0.18)$	$\mathbf{0.21}(0.14)$	$2.39(0.18)$	$2.39(0.19)$
j	$\mathbf{0.22}(0.15)$	$1.66(0.18)$	$2.40(0.19)$	$2.39(0.18)$

Table 1: Comparison between our model and standard MHP model in full connection task. Rows are source nodes, columns are target nodes. Each cell shows the mean absolute error with standard deviation. Unit in spikes/sec.

4.1.4 Multivariate Hawkes model

It is natural to extend our bivariate regression-type method to a multivariate regression-type model. The coupling effect in multivariate processes can be regarded as a form of graph structure recovery in graphical models, where each point process is considered a node. From this perspective, e.g., [Meinshausen and Bühlmann, 2006, Murphy, 2012, sec. 19.4.4], multivariate regression extends the bivariate case by studying *pairwise* conditional relations for all possible pairs. More specifically, given a pair of random variables X, Y, let Z represent the totality of all other random variables excluding X, Y. The multivariate regression infers if a *bivariate* relation $X \perp Y | Z$ holds, also known as the *global Markov property* [Koller and Friedman, 2009]. A similar concept in standard MHP can be found in [Eichler et al., 2017]. In our MHP setting, this is equivalent to estimating the impact functions between N_i and N_j given the observations of all other processes and so that their effect enters as the dynamic background. Notice that the standard MHP cannot model this extension because even if the baseline intensity of each point process is constant, the totality of random effect from all other nodes excluding two nodes will not necessarily give a constant baseline to the nodes under consideration. Consider the intensity function in the multivariate point process,

$$\begin{aligned}\lambda_j(t) = \alpha_j + \int_0^t h_{i\to j}(t-s)dN_i(s) \\ \underbrace{+f_j(t) + \sum_{r\neq i,j}\int_0^t h_{r\to j}(t-s)dN_r(s)}_{\tilde{f}_j(t)}\end{aligned} \quad (16)$$

where f_j together with input from other processes are treated as a new background $\tilde{f}_j(t)$. This perspective exactly reduces the MHP to model (8).

The performance of the model is evaluated using simulation dataset, which involves 6 processes and all processes are driven by fluctuating background. The coupling effects between nodes can be positive, negative or zero. Table 2 shows that our method outperforms the standard Hawkes model in multivariate processes scenario. Details of the experiment are in Appendix C.9.

	Bias (std)	RMSE (std)
Hawkes	1.52 (0.040)	1.54 (0.41)
Ours	**0.028** (0.040)	**0.25** (0.33)

Table 2: Comparison between the standard Hawkes model and our model. The unit is [spikes/sec].

4.1.5 Other simulation scenarios

Other properties of the model and empirical verifications are briefly summarized in this section due to the page limit.

Varying-timescale background. See Appendix C.5. We violate the settings in section 4.1.2 by relaxing the fixed background timescale σ_I in (13) to randomly changing timescale to test the robustness of the model.

Fast-changing background. See Appendix C.6. In extreme cases, the background activity f_i can have fast-changing activities. In this situation, the conditional inference-based method will be limited by its formalization of the null hypothesis, which implicitly assumes the timescale of the coupling effect is smaller than that of the background. Our model is still able to accurately estimate the cross-impact effect while the conditional inference-based method fails.

Asymptotic Normality. See Appendix C.7. Similar to profile likelihood, the approximate normality of the estimator is observed in simulations. The property may be convenient for model inference, details are also in Appendix D.2.

Selection of impact function length. See Appendix C.8. In practice, the timescale of the interaction effect is typically unknown. When users are not confident with the prior knowledge of the timescale of the coupling effect, our methods can be adapted to use a shorter impact function or nonparametric fitting first, as shown in Appendix D.1.

4.2 NEUROPIXELS DATA

Spiking neural activities likely come with non-stationary background signals due to external stimuli or inter-area interactions. With recent advances in high-density electrophysiological recording technologies, such as Neuropixels, hundreds of neurons from multiple brain regions can be recorded simultaneously. This offers opportunities to further investigate the interactions between brain areas [Siegle et al., 2021, Chen et al., 2022]. However, point-to-point coupling effects on fine timescales across regions is not yet well studied. Here, we apply our method to the hierarchical mouse visual system across 5 brain areas: V1, LM, RL, AL, and

AM in ascending order with V1 as the primary visual cortex, thought to process simple visual features, and AM as the higher-order cortex thought to handle sophisticated signals [Harris et al., 2019, Siegle et al., 2021] (Figure 8). We aim to fit the coupling effect across brain regions and discover the excitatory or inhibitory interactions on a fine timescale. Details are in Appendix F.

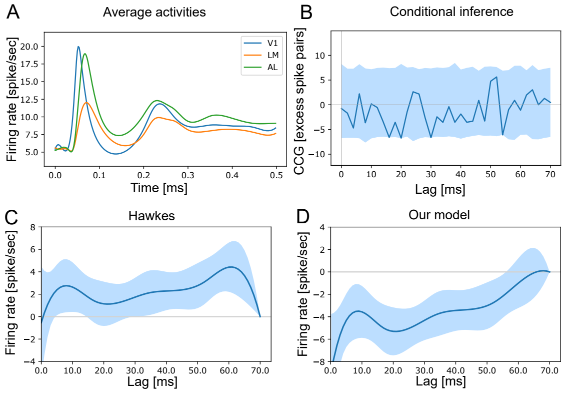

Figure 7: Neuropixels data. **A**: Activities of three brain areas showing correlated backgrounds. B,C,D are results of a pair of neurons. **B**: CCG. **C**: standard MHP. **D**: Our method.

Figure 7A demonstrates the averaged activities of 3 brain regions with large correlations providing a clue for the background artifact. Results are very similar to the simulation in Figure 3; CCG in Figure 7B shows some negative but not statistically significant effects. Our method is more sensitive in detecting the effect between 0 and 50 ms lag. In contrast, due to the background artifacts, the standard Hawkes model detects non-significant or slightly positive coupling effects.

Figure 8 shows the discovered neuronal network of 190 neurons. Multiple significant impact functions are selected with Bonferroni correction at level 0.01. 766 directed edges are split into bottom-up connections and top-down connections [Siegle et al., 2021, Harris et al., 2019]. The impact function is fitted using a 50 ms square window determined by exploring CCG and non-parametric fitting (see examples in Appendix F). Our main findings using MHP extension are: (a) Most edges concentrate at a few neurons, and (b) the active senders or receivers are consistent across top-down and bottom-up networks. The real data has no ground truth so we cannot directly evaluate the performance of this multivariate extension. However, the findings directly corroborate previous neuroscience studies [Harris et al., 2019, Glickfeld and Olsen, 2017] based on anatomical analysis, whereas our findings are entirely data-driven. The findings are also complementary to [Jia et al., 2020, Siegle et al., 2021] using the traditional CCG method (in section 2, our method outperforms CCG in both computation and performance).

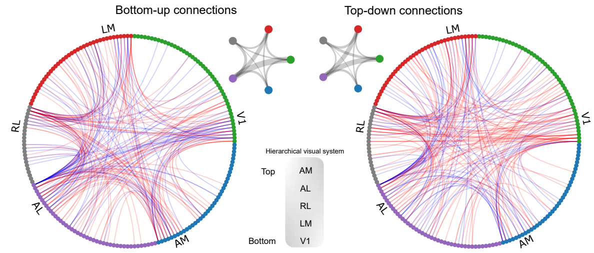

Figure 8: Network of coupling neurons in mouse visual system. Excitatory (positive) and inhibitory (negative) impact functions are shown in red and blue edges. 20% randomly selected edges are shown. The coupling filter connecting a lower-order region to a higher-order region, for example from V1 to RL, is categorized into the bottom-up graph on the left; the graph on the right shows the top-down connections [Siegle et al., 2021, Harris et al., 2019]. The small graphs at the corner count the total number of edges between areas.

Figure 9 compares the histograms of the impact function amplitudes between the standard Hawkes model and our model. It is suspected that the standard Hawkes model may falsely detect more positive relations. Goodness-of-fit analysis and more details of the experiment can be found in Appendix F. As shown, the above discoveries are greatly facilitated by our method.

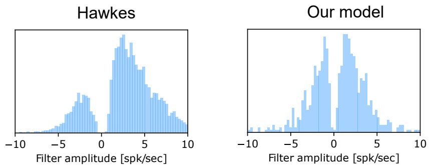

Figure 9: Histograms of estimated impact function amplitudes.

5 CONCLUSION

We report and analyze the error of MLE from MHP in short-term temporal dependency detection due to heterogeneous background, which we believe is common but largely overlooked. We developed a flexible, robust, and computationally efficient model to address this problem in an attempt to generalize the use case for MHP in practice. Finally, we applied the new tool to a neuroscience dataset and discovered the structure of a patterned neuronal network across visual cortices in the mouse visual system.

Acknowledgements

Yu Chen was supported by NIMH grant RO1 MH064537 when the work was done at Carnegie Mellon University.

References

Asohan Amarasingham, Matthew T Harrison, Nicholas G Hatsopoulos, and Stuart Geman. Conditional modeling and the jitter method of spike resampling. *Journal of Neurophysiology*, 107(2):517–531, 2012.

Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.

Charles Blundell, Jeff Beck, and Katherine A Heller. Modelling reciprocating relationships with hawkes processes. *Advances in neural information processing systems*, 25, 2012.

Clive G Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912, 2007.

Yu Chen, Hannah Douglas, Bryan J Medina, Motolani Olarinre, Joshua H Siegle, and Robert E Kass. Population burst propagation across interacting areas of the brain. *Journal of Neurophysiology*, 128(6):1578–1592, 2022.

Kacper Chwialkowski and Arthur Gretton. A kernel independence test for random processes. In *International Conference on Machine Learning*, pages 1422–1430. PMLR, 2014.

David Roxbee Cox and Valerie Isham. *Point processes*, volume 12. CRC Press, 1980.

José Da Fonseca and Riadh Zaatour. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579, 2014.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer, 2008.

Peter Diggle. A kernel method for smoothing point process data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):138–147, 1985.

Uri T Eden and Emery N Brown. Continuous-time filters for state estimation from point process models of neural data. *Statistica Sinica*, 18(4):1293, 2008.

Michael Eichler, Rainer Dahlhaus, and Johannes Dueck. Graphical modeling for multivariate hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

Eymen Errais, Kay Giesecke, and Lisa R Goldberg. Affine point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics*, 1(1):642–665, 2010.

Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou. *Statistical foundations of data science*. CRC press, 2020.

Mehrdad Farajtabar, Yichen Wang, Manuel Gomez Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. Coevolve: A joint point process model for information diffusion and network co-evolution. *Advances in Neural Information Processing Systems*, 28, 2015.

Efi Foufoula-Georgiou and Dennis P Lettenmaier. Continuous-time versus discrete-time point process models for rainfall occurrence series. *Water Resources Research*, 22(4):531–542, 1986.

Lindsey L Glickfeld and Shawn R Olsen. Higher-order areas of the mouse visual cortex. *Annual review of vision science*, 3:251–273, 2017.

Asela Gunawardana, Christopher Meek, and Puyang Xu. A model for temporal dependencies in event streams. *Advances in neural information processing systems*, 24, 2011.

Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83 – 143, 2015. doi: 10.3150/13-BEJ562. URL `https://doi.org/10.3150/13-BEJ562`.

Julie A Harris, Stefan Mihalas, Karla E Hirokawa, Jennifer D Whitesell, Hannah Choi, Amy Bernard, Phillip Bohn, Shiella Caldejon, Linzy Casal, Andrew Cho, et al. Hierarchical organization of cortical and thalamic connectivity. *Nature*, 575(7781):195–202, 2019.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Alan G Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443, 1971.

Alan G Hawkes. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198, 2018.

Xiaoxuan Jia, Joshua H Siegle, Séverine Durand, Greggory Heller, Tamina Ramirez, and Shawn R Olsen. Multi-area functional modules mediate feedforward and recurrent processing in visual cortical hierarchy. *bioRxiv*, 2020.

Robert E Kass and Valérie Ventura. A spike-train probability model. *Neural computation*, 13(8):1713–1720, 2001.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

PA W Lewis and Gerald S Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979.

Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems*, 30, 2017.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. 2006.

Gianluigi Mongillo, Simon Rumpel, and Yonatan Loewenstein. Inhibitory connectivity defines the realm of excitatory plasticity. *Nature neuroscience*, 2018.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Susan A Murphy and Aad W Van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

Yoshiko Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 1978.

Yosihiko Ogata. On lewis' simulation method for point processes. *IEEE transactions on information theory*, 1981.

Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 1988.

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):995–999, 2008.

Patricia Reynaud-Bouret and Sophie Schbath. Adaptive estimation for hawkes processes; application to genome analysis. 2010.

Farnood Salehi, William Trouleau, Matthias Grossglauser, and Patrick Thiran. Learning hawkes processes from a handful of events. *Advances in Neural Information Processing Systems*, 32, 2019.

Joshua H Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K Ramirez, Hannah Choi, Jennifer A Luviano, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021.

Liwen Wang and Lin Zhang. Hawkes processes for understanding heterogeneity in information propagation on twitter. *Frontiers in Physics*, page 970, 2022.

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2447–2456, 2018.

Yichen Wang, Bo Xie, Nan Du, and Le Song. Isotonic hawkes processes. In *International conference on machine learning*, pages 2226–2234. PMLR, 2016.

Larry Wasserman. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25, 1982.

Hongteng Xu and Hongyuan Zha. A dirichlet mixture model of hawkes processes for event sequence clustering. *Advances in neural information processing systems*, 30, 2017.

Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning granger causality for hawkes processes. In *International conference on machine learning*, pages 1717–1726. PMLR, 2016.

Steve Y Yang, Anqi Liu, Jing Chen, and Alan Hawkes. Applications of a multivariate hawkes process to joint modeling of sentiment and market return events. *Quantitative finance*, 18(2):295–310, 2018.

Yingxiang Yang, Jalal Etesami, Niao He, and Negar Kiyavash. Online learning for multivariate hawkes processes. *Advances in Neural Information Processing Systems*, 30, 2017.

Qiang Zhang, Aldo Lipani, Omer Kirnap, and Emine Yilmaz. Self-attentive hawkes process. In *International conference on machine learning*, pages 11183–11193. PMLR, 2020.

Feng Zhou, Zhidong Li, Xuhui Fan, Yang Wang, Arcot Sowmya, and Fang Chen. Efficient inference for nonparametric hawkes processes using auxiliary latent variables. *Journal of Machine Learning Research*, 2020.

Feng Zhou, Quyu Kong, Yixuan Zhang, Cheng Feng, and Jun Zhu. Nonlinear hawkes processes in time-varying system. *arXiv preprint arXiv:2106.04844*, 2021a.

Feng Zhou, Yixuan Zhang, and Jun Zhu. Efficient inference of flexible interaction in spiking-neuron networks. In *International Conference on Learning Representations*, 2021b. URL `https://openreview.net/forum?id=aGfU_xziEX8`.

Feng Zhou, Quyu Kong, Zhijie Deng, Jichao Kan, Yixuan Zhang, Cheng Feng, and Jun Zhu. Efficient inference for dynamic flexible interactions of neural populations. *Journal of Machine Learning Research*, 2022.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649. PMLR, 2013a.

Ke Zhou, Hongyuan Zha, and Le Song. Learning triggering kernels for multi-dimensional hawkes processes. In *International conference on machine learning*, pages 1301–1309. PMLR, 2013b.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. In *International conference on machine learning*, pages 11692–11702. PMLR, 2020.