Detection of Short-Term Temporal Dependencies in Hawkes Processes with Heterogeneous Background Dynamics (Supplementary Material)

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 & Biology, City College & 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

In section A, we provide details of optimization algorithms, and condition inference based cross-correlogram. Proofs of the propositions in the main text can be found in section B. Section C lists all simulation scenarios for the model verification. Section D presents some use cases and versatility of our new tool. Section E derives the approximated analytical formulae of the model's properties in a special situation. Finally, more details about the neuroscience experiments are in section F.

A ALGORITHMS

A.1 UPDATING RULES

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

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

$$\overline{\mathbf{s}_i}(t)=\int_0^T W(t-s)\mathrm{d}N_i(s)=\sum_{t_m\in N_i}W(t-t_m)$$

Let ϕ_w,ϕ_h be the bases defined as,

$$\phi_w(t):=\int W(t-s)\mathrm{d}N_i(s),\qquad \phi_h(t):=\int h_{i\to j}(t-s)\mathrm{d}N_i(s) \tag{1}$$

The intensity can be rewritten in a linear form,

$$\tilde{\lambda}_j(t)=\beta_j\cdot 1+\beta_w\phi_w(t)+\beta_h\phi_h(t)=\Psi(t)\boldsymbol{\beta} \tag{2}$$

$\Psi(t)$ represents all bases, $\boldsymbol{\beta}$ is a vector of the coefficients. If the impact function is fitted using non-parametric method, such as general additive model or splines

$$h_{i\to j}(s)=\beta_{h,1}B_1(s)+...+\beta_{h,k}B_k(s)$$

where $B_1,...,B_k$ are spline bases for the impact function. Define

$$\phi_{h,1}(t):=\int B_1(t-s)\mathrm{d}N_i(s),...,\ \phi_{h,k}(t):=\int B_k(t-s)\mathrm{d}N_i(s)$$

The intensity still maintains the linear form:

$$\tilde{\lambda}_j(t)=\beta_j\cdot 1+\beta_w\phi_w(t)+\beta_{h,1}\phi_{h,1}(t)+...+\beta_{h,k}\phi_{h,k}(t)$$

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

The target equation of the model can be optimized using gradient descent. For a fixed σ_w, the target is convex and the optimization is efficient using Newton's method. The first-order and second-order derivatives of the target equations are,

$$\begin{aligned}\frac{\partial \tilde{\ell}}{\partial \boldsymbol{\beta}} &= -\int_0^T \frac{\Psi(s)}{\tilde{\lambda}_j(s)} \mathrm{d}N_j(s) + \int_0^T \Psi(s)\mathrm{d}s \\ \frac{\partial^2 \tilde{\ell}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \int_0^T \frac{\Psi(s)\Psi(s)^T}{\tilde{\lambda}_j(s)^2} \mathrm{d}N_j(s).\end{aligned}$$

The update of σ_w can be done as a separate step using a gradient as follows. If W is a Gaussian kernel function, then

$$\begin{aligned}\frac{\partial \tilde{\ell}}{\partial \sigma_w} &= -\sum_{t_n \in N_j} \frac{\beta_w}{\tilde{\lambda}(t_n)} \frac{\partial}{\partial \sigma_w} \overline{\mathbf{s}_i}(t_n) + \beta_w \frac{\partial}{\partial \sigma_w} \int_0^T \overline{\mathbf{s}_i}(u)\mathrm{d}u \\ &= -\sum_{t_n \in N_j} \frac{\beta_w}{\tilde{\lambda}(t_n)} \sum_{t_m \in N_i} \frac{\partial}{\partial \sigma_w} W(t_n - t_m) + \beta_w \sum_{t_m \in N_i} \frac{\partial}{\partial \sigma_w} \int_0^T W(u - t_m)\mathrm{d}u \\ &= -\sum_{t_n \in N_j} \frac{\beta_w}{\tilde{\lambda}(t_n)} \sum_{t_m \in N_i} \frac{\partial}{\partial \sigma_w} W(t_n - t_m) + \beta_w \sum_{t_m \in N_i} \Big(W(T - t_m) \frac{\partial}{\partial \sigma_w} W(T - t_m) - W(-t_m) \frac{\partial}{\partial \sigma_w} W(-t_m) \Big)\end{aligned} \tag{3}$$

where

$$\frac{\partial}{\partial \sigma_w} W(x) = \left(-\frac{1}{\sqrt{2\pi\sigma_w^2}} + \frac{x^2}{\sqrt{2\pi\sigma_w^4}} \right) \exp\left\{ -\frac{x^2}{2\sigma_w^2} \right\} \tag{4}$$

σ_w can be optimized using grid-search during warmup. In Appendix D.3, we discuss the sampling-based method, which shows incorporating the uncertainty of σ_w or fixing it at the optimal does not make a significant difference.

A.2 INTEGRAL TRICK

One computational advantage of the proposed model in main (2) is that the integral $\int \Psi(s)\mathrm{d}s$ can be calculated in the closed form if the bases are designed carefully. In contrast, models with intensity in logarithmic scale $\log \lambda(t) = \Psi(t)\boldsymbol{\beta}$ do not enjoy this computation convenience. For example, the derivative of the modified negative log-likelihood function becomes,

$$\frac{\partial \tilde{\ell}}{\partial \boldsymbol{\beta}} = -\int_0^T \Psi(s)\mathrm{d}N_j(s) + \int_0^T \Psi(s)e^{\Psi(s)\boldsymbol{\beta}}\mathrm{d}s.$$

Usually, it is not tractable to calculate the integral in the second term, so it is approximated by discretizing the continuous functions. This is the reason our model does not involve discretization or require specifying the time resolution. Another benefit of using a continuous-time model is that the number of data points is small, which is proportional to the number of spikes instead of the number of time bins. For example, if the bin width is 1 ms, then for one 1-second long trial, it needs to store 1000 data points. If the trial has 20 spikes, the continuous-time model only needs to keep 20 data points. The memory space is 50 times smaller.

If the regression bases have form Eq (1) with kernel, then

$$\int_0^T \Psi(t)\mathrm{d}t = \int_0^T \int_0^T K(t - s)N_i(\mathrm{d}s) = N_i(T) \int_{\mathrm{R}} K(s)\mathrm{d}s.$$

If K is a Normal window function or a square window function, the above integral is simple. The boundary effect can be removed in the integral by only considering a few time points close to 0 or T. Next, we show how to calculate such an integral if K is B-spline, which is widely used in non-parametric curve fitting. An example can be found in Appendix D.1.

The B-splines are defined using Cox-de Boor recursion equations. t_i are knots (with repeated padding). p is the degree of the spline polynomial. When $p = 3$, these are the cubic splines.

$$\begin{aligned}B_{i,0}(x) =& \mathbb{I}_{[t_i, t_{i+1})}(x) \\ B_{i,p}(x) =& \frac{x - t_i}{t_{i+p} - t_i} B_{i,p-1}(x) + \frac{t_{i+p+1} - x}{t_{i+p+1} - t_{i+1}} B_{i+1,p-1}.\end{aligned}$$

Knot padding is important to create proper splines. If $p = 3$ and the distinct knot locations are $(0, 1, 2)$, the input knots should be $(0, 0, 0, 0, 1, 2, 2, 2, 2)$. The knots need extra p repeated knots of the two ends. If there are K distinct knots, then there are $K + 2p$ input knots. The total number of basis is $K + p - 1$.

Lemma A.1. *For the B-spline curve defined above, the integral of the curve has closed-form as follows,*

$$\int_{-\infty}^{\infty} B_{i,p}(s)\mathrm{d}s = \frac{t_{i+p+1} - t_i}{p+1}. \tag{5}$$

Proof. The support of each basis spans over $p + 1$ knot-intervals (including the padded knots on the ends),

$$\text{supp}(B_{i,p}) =[t_i, t_{i+p+1})$$

$$\frac{\mathrm{d}}{\mathrm{d}x} B_{i,p}(x) = \frac{p}{t_{i+p} - t_i} B_{i,p-1}(x) - \frac{p}{t_{i+p+1} - t_{i+1}} B_{i+1,p-1}(x).$$

The support of the derivative is almost the same as the basis except for a few 0 derivative points.

$$\text{supp}(\frac{\mathrm{d}}{\mathrm{d}x} B_{i,p}) \subseteq[t_i, t_{i+p+1})$$

We reform the derivative properties to get the integral [Bhatti and Bracken, 2006].

$$\frac{\mathrm{d}}{\mathrm{d}x} \sum_{i=0}^{\infty} c_i B_{i,p+1}(x) = \sum_{i=0}^{\infty} (p+1) \frac{c_i - c_{i-1}}{t_{i+p+1} - t_i} B_{i,p}(x)$$

c_i are some arbitrary coefficients. Next we set $c_0, ..., c_{i-1} = 0$, $c_i, c_{i+1}, ... = 1$.

$$\frac{\mathrm{d}}{\mathrm{d}x} \sum_{j=i}^{\infty} c_j B_{j,p+1}(x) = \frac{\mathrm{d}}{\mathrm{d}x} \sum_{j=i}^{i+p} B_{j,p+1}(x) = \frac{p+1}{t_{i+p+1} - t_i} B_{i,p}(x)$$

The first equation simplifies the sum due to the supports of bases. Then take the integral on both side,

$$\int_{-\infty}^{x} B_{i,p}(s)\mathrm{d}s = \int_{t_i}^{x} B_{i,p}(s)\mathrm{d}s = \frac{t_{i+p+1} - t_i}{p+1} \sum_{j=i}^{\infty} B_{j,p+1}(s) = \frac{t_{i+p+1} - t_i}{p+1} \sum_{j=i}^{i+p} B_{j,p+1}(x)$$

The area under the curve of a basis is,

$$\int_{-\infty}^{\infty} B_{i,p}(s)\mathrm{d}s = \int_{t_i}^{t_{i+p+1}} B_{i,p}(s)\mathrm{d}s = \frac{t_{i+p+1} - t_i}{p+1} \sum_{j=i}^{i+p} B_{j,p+1}(t_{i+p+1})$$

Consider the summation term,

$$\begin{aligned}
\sum_{j=i}^{i+p} B_{j,p+1}(t_{i+p+1}) =& B_{i,p+1}(t_{i+p+1}) + B_{i+1,p+1}(t_{i+p+1}) + ... + B_{i+p,p+1}(t_{i+p+1}) \\
=& \Big(\frac{t_{i+p+1} - t_i}{t_{i+p+1} - t_i} B_{i,p}(t_{i+p+1}) + \frac{t_{i+p+2} - t_{i+p+1}}{t_{i+p+2} - t_{i+1}} B_{i+1,p}(t_{i+p+1}) \Big) \\
&+ \Big(\frac{t_{i+p+1} - t_{i+1}}{t_{i+p+2} - t_{i+1}} B_{i+1,p}(t_{i+p+1}) + \frac{t_{i+p+3} - t_{i+p+1}}{t_{i+p+3} - t_{i+2}} B_{i+2,p}(t_{i+p+1}) \Big) + ... \\
&+ \Big(\frac{t_{i+p+1} - t_{i+p}}{t_{i+2p+1} - t_{i+p}} B_{i+p,p}(t_{i+p+1}) + \frac{t_{i+2p+2} - t_{i+p+1}}{t_{i+2p+2} - t_{i+p+1}} B_{i+p+1,p}(t_{i+p+1}) \Big) \\
=& B_{i,p}(t_{i+p+1}) + B_{i+1,p}(t_{i+p+1}) + ... + B_{i+p+1,p}(t_{i+p+1}) \\
=& B_{i,p-1}(t_{i+p+1}) + B_{i+1,p}(t_{i+p+1}) + ... + B_{i+p+2,p-1}(t_{i+p+1}) \\
=& B_{i,0}(t_{i+p+1}) + B_{i+1,0}(t_{i+p+1}) + ... + B_{i+2p+1,0}(t_{i+p+1}) = 1
\end{aligned}$$

So the conclusion holds. □

Another popular example with closed-form integral is the exponential coupling function, which also enjoys the integral trick.

$$h_{i\to j}(\tau) = e^{-\gamma_{i\to j}(\tau)}\mathbb{I}(\tau \geq 0) \tag{6}$$

The modified negative log-likelihood is,

$$\begin{aligned}\tilde{\ell} = &-\sum_{t_n \in N_j} \log\left(\beta_j + \overline{\mathbf{s}_i}(t_n) + \alpha_{i\to j} \sum_{t_m \in N_i, t_m < t_n} e^{-\gamma_{i\to j}(t_n - t_m)}\right) \\ &+ \beta_j T + \int_0^T \overline{\mathbf{s}_i}(s)\mathrm{d}s + \frac{\alpha_{i\to j}}{\gamma_{i\to j}} \sum_{t_m \in N_i}\left(1 - e^{-\gamma_{i\to j}(T - t_m)}\right)\end{aligned} \tag{7}$$

where $\alpha_{i\to j}$ is the coefficient of the exponential basis. The derivatives of the target equation over $\beta_j, \beta_w, \alpha_{i\to j}$ are similar to other linear models. The derivative over the timescale parameter $\gamma_{i\to j}$ is

$$\begin{aligned}\frac{\partial \tilde{\ell}}{\partial \gamma_{i\to j}} = &\sum_{t_n \in N_j} \frac{\alpha_{i\to j}}{\tilde{\lambda}(t_n)} \sum_{t_m \in N_i, t_m < t_n} (t_n - t_m) e^{-\gamma_{i\to j}(t_n - t_m)} \\ &- \frac{\alpha_{i\to j}}{\gamma_{i\to j}^2} \sum_{t_m \in N_i} (1 - e^{-\gamma_{i\to j}(T - t_m)}) + \frac{\alpha_{i\to j}}{\gamma_{i\to j}} \sum_{t_m \in N_i} (T - t_m) e^{-\gamma_{i\to j}(T - t_m)}\end{aligned} \tag{8}$$

A.3 DETAILS OF CONDITIONAL INFERENCE BASED CROSS-CORRELOGRAM (CCG)

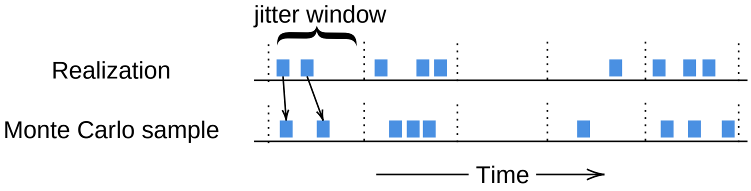

Figure 1: Construction of Monte Carlo samples from the null in conditional-inference based CCG. Blue dots are timestamps.

The calculation of the CCG hypothesis tests requires first discretizing the point processes. The time bin is usually set with a small size so that each bin contains at most one point. Consider two time-binned processes X_i and X_j with respect to the counting processes N_i and N_j. The correlation (or cross-correlation) of timing at a certain lag $\tau \in \mathbb{N}$ (X_i leads X_j) with fluctuating background activity is assessed in the following procedure:

1. Calculate the test statistic CCG at τ,
$$\mathrm{CCG}(\tau) = \sum_n X_i(n - \tau) X_j(n) \tag{9}$$
2. Divide the timeline into equal-width jitter windows Δ as shown in the figure above.
3. One Monte Carlo sample from the null distribution is generated by uniformly allocating the time points within each jitter window, as shown in the figure. All samples are drawn independently (conditioned on the observed data). Such a process can be applied to either N_i or N_j, or both. Then time-bin the "jittered" samples $\tilde{X}_i$, $\tilde{X}_j$.
4. Calculate the CCG of the Monte Carlo sample,
$$\mathrm{CCG}^{\mathrm{MC}}(\tau) = \sum_n \tilde{X}_i(n - \tau)\tilde{X}_j(n) \tag{10}$$
5. Repeat step 3 and 4 multiple times N_{MC} to acquire the null distribution of CCG. The p-value of the test, which is exact, is
$$\mathrm{p-value} = \frac{N_\tau + 1}{N_{\mathrm{MC}} + 1} \tag{11}$$
where N_τ is the number of Monte Carlo samples $\mathrm{CCG}(\tau) < \mathrm{CCG}^{\mathrm{MC}}(\tau)$ (or the other way if $\mathrm{CCG}(\tau)$ is on the left tail). Similarly, the acceptance band can be constructed from the null distribution.

Δ is chosen with prior knowledge of the data, which is roughly the timescale of the background activity, so that jittering the samples within window Δ maintains the intensity of N_i, N_j, but it can "break" the fine time structures without breaking the background timing association. As the time points are effectively 'jittered' by a small amount, the null samples are, heuristically, presumed to have the same intensity as the original process. In this way, the method bypasses the problem of estimating a dynamic background. For better visualization, the mean of the null is usually subtracted from the test statistic so the CCG curve and the acceptance band are centered around zero, as in the main Figures 3 and 7. In the main Figure 3 and 7, the jitter window is $\Delta = 120$ ms, the discretization time bin width is 2 ms, and the null distribution is computed from 1000 Monte Carlo samples. A rigorous explanation and variations on the theme are provided in Amarasingham et al. [2012].

B PROOFS

We list the regularity conditions needed for our statements here. These technical conditions first include follows from Assumptions A, B, C in Ogata [1978], which makes sure the consistency of $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_{KL} = \operatorname{argmin}_{\boldsymbol{\theta}} \Lambda(\boldsymbol{\theta}) := \mathbb{E}\ell(\boldsymbol{\theta})$ for the misspecified model. Then, we make the additional assumption:

Assumption 1. $\Lambda(\cdot)$ *is* μ*-strongly convex and has* L*-Lipschitz gradient.*

With this assumption, we can guarantee that for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, we have

$$\mu\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \leq \|\nabla\Lambda(\boldsymbol{\theta}_1) - \nabla\Lambda(\boldsymbol{\theta}_2)\| \leq L\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$

which gives $\|\nabla\Lambda(\boldsymbol{\theta}_1) - \nabla\Lambda(\boldsymbol{\theta}_2)\| = \Theta(\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|)$.

We state a Lemma. Before we state the Lemma, let as simplify the notation using 2,1 vs i, j and absorb the baseline and heterogeneity into one function f so that we have

$$\begin{aligned} \lambda_1(t) &= f_1(t) + c\int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s) \\ \lambda_2(t) &= f_2(t) \end{aligned} \tag{12}$$

This will simplify the notation in the Lemma below. We always explicitly make clear which version of notation we are using before stating results.

Lemma B.1. *Fixed a and b, assume f_2 is continuous everywhere with bounded gradient, then*

$$\begin{aligned} \mathbb{E}\left[\frac{1}{a + b\int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s)}\right] =& \frac{1 - f_2(t)\sigma_h}{a} + \frac{f_2(t)}{a+b}\sigma_h + o(\sigma_h) \\ =& \frac{1}{a} - \frac{b}{a(a+b)} f_2(t)\sigma_h + o(\sigma_h) \end{aligned} \tag{13}$$

Proof. Simply note that $\int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s)$ is a Poisson distribution with parameter (also mean) $\int_{t-\sigma_h}^t f_2(s)ds = f_2(t)\sigma_h + o(\sigma_h)$. The rest follows direct calculation using Poisson p.m.f. □

We also state another Lemma which will be useful later, based on taylor expansion of ratio function:

Lemma B.2. *Given finitely supported R.V. X and Y with mean μ_x and μ_y, if we define $r_x = \|X - \mu_x\|_\infty$ and $r_y = \|Y - \mu_y\|_\infty$ and $r = \max(r_x, r_y)$. Suppose $\frac{X}{Y}$ is always strictly bounded away from 0 by some fixed positive constant, then*

$$\frac{X}{Y} = \frac{\mu_x}{\mu_y} - \frac{\mu_x}{\mu_y^2}(Y - \mu_y) + \frac{1}{\mu_y}(X - \mu_x) + o(r)$$

or higher order approximation:

$$\frac{X}{Y} = \frac{\mu_x}{\mu_y} - \frac{\mu_x}{\mu_y^2}(Y - \mu_y) + \frac{1}{\mu_y}(X - \mu_x) + \frac{\mu_x}{\mu_y^3}(Y - \mu_y)^2 - \frac{1}{\mu_y^2}(X - \mu_x)(Y - \mu_y) + o(r^2).$$

Thus,

$$\mathbb{E}\frac{X}{Y} = \frac{\mu_x}{\mu_y} + \frac{\mathit{Var}(Y)\mu_x}{\mu_y^3} - \frac{\mathit{Cov}(X,Y)}{\mu_y^2} + o(r^2).$$

Proof. Omitted. □

Thus, if we use the above high order approximation:

$$R(X,Y) := \frac{\mu_x}{\mu_y} + \frac{\mathrm{Var}(Y)\mu_x}{\mu_y^3} - \frac{\mathrm{Cov}(X,Y)}{\mu_y^2},$$

we have $\mathbb{E}[\frac{X}{Y}] = R(X,Y) + o(r^2)$.

B.1 PROOF OF PROPOSITION 1.

We first prove Proposition 1. In this proof we use original intensity notation (4).

Proof of Proposition 1. Using (1) (or see Ogata [1978]), we can show that

$$\begin{aligned}\Lambda(\theta) =&\mathbb{E}\Big[-\int_0^P \lambda_\theta(t)dt + \int_0^P \log\lambda_\theta(t)dN_1(t)\Big]\\ =&\mathbb{E}\Big[-\int_0^P \lambda_1(t)\Big(\frac{\lambda_\theta(t)}{\lambda_1(t)} - \log\lambda_\theta(t)\Big)dt\Big]\end{aligned}$$

for

$$\lambda_1(t) =\alpha_1 + f_1(t) + c\int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s). \tag{14}$$

where we set $\int_0^T f_1(t)dt = 0$ to avoid identifiability issue with α_1 and parametrize

$$\lambda_\theta(t) = \theta_1 + \theta_2\int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s),$$

for MLE. Using Lemma B.1 and $\int_0^T f_1(t)dt = 0$, one can show

$$\begin{aligned}\frac{\partial\Lambda}{\partial\theta_2}_{|\boldsymbol{\theta}=(\alpha_1,c)} =&\mathbb{E}\Big[-\int_0^T \frac{\partial\lambda_\theta}{\partial\theta_2}\Big(1-\frac{\lambda_1^*(t)}{\lambda_\theta(t)}\Big)dt\Big]\\ =&-\int_0^T \mathbb{E}\Big[\frac{f_1(t)\int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s)}{\alpha_1 + c\int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s)}\Big]dt\\ =&-\int_0^T \frac{f_1(t)}{c} - \frac{f_1(t)\mu}{c}\mathbb{E}\Big[\frac{1}{\alpha_1 + c\int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s)}\Big]dt\\ =&-\int_0^T \frac{f_1(t)f_2(t)\sigma_h}{\alpha_1+c}dt + o(\sigma_h)dt\end{aligned}$$

Now, note $\boldsymbol{\theta}_{KL}$ corresponds to the point where $\frac{\partial\Lambda}{\partial\theta_2} = 0$. The rest follows from Assumption 1.

□

B.2 PRELIMINARY WORK FOR PROOFS OF PROPOSITION 2 AND 3.

Hence forth we use the alternative intensity notation (12). We lay some ground work before proving Proposition 2 and 3. Again we restate the notation (12), but also parametrize the density as θ (the impact function parameter) and $\boldsymbol{\eta}$ (all else is nuisance parameter):

$$\lambda_1(t) =f_1(t) + c\int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s)$$

$$\begin{aligned}\lambda_2(t) =& f_2(t)\\ \lambda_{\theta,\boldsymbol{\eta}}(t) =& \sum_{i=1}^{M} \eta_i g_i(t) + \theta \int_0^t \mathbf{1}_{[0,\sigma_h]}(t-s) dN_2(s)\end{aligned} \tag{15}$$

and re-define

$$\ell(\theta, \boldsymbol{\eta}; \mathcal{H}_T) = -\int_0^T \lambda_{\theta,\boldsymbol{\eta}}(t)dt + \int_0^T \log \lambda_{\theta,\boldsymbol{\eta}}(t) dN_1(t)$$

and consequently $\Lambda(\theta, \boldsymbol{\eta}) = \mathbb{E}[\ell(\theta, \boldsymbol{\eta})]$. However, we characterize a concept, on population level, similar to profile likelihood

$$\Lambda_p(\theta) = \sup_{\boldsymbol{\eta}} \mathbb{E}\Big[-\int_0^T \lambda_{\theta,\boldsymbol{\eta}}(t)dt + \int_0^T \log \lambda_{\theta,\boldsymbol{\eta}}(t) dN_1(t)\Big],$$

So, we can define

$$\boldsymbol{\eta}(\theta) = \underset{\boldsymbol{\eta}}{\operatorname{argmax}}\, \mathbb{E}\Big[-\int_0^T \lambda_{\theta,\boldsymbol{\eta}}(t)dt + \int_0^T \log \lambda_{\theta,\boldsymbol{\eta}}(t) dN_1(t)\Big]$$

so that $\Lambda(\theta, \boldsymbol{\eta}(\theta)) = \Lambda_p(\theta)$. As a result of Assumption 1, Λ_p is also μ-strongly convex with L-Lipschitz gradient. Now, we first characterize $\boldsymbol{\eta}(c)$ which corresponds to the equatiosn $\frac{\partial \Lambda(c,\boldsymbol{\eta})}{\partial \eta_i} = 0$ for all i. To analyze this term, we write:

$$\begin{aligned}&\frac{\partial \Lambda(c, \boldsymbol{\eta})}{\partial \eta_i}\\ =& -\mathbb{E}\Big[\int_0^T \frac{\partial \lambda_{\theta,\boldsymbol{\eta}}}{\partial \eta_i}(t)\Big(1 - \frac{\lambda_1(t)}{\lambda_{\theta,\boldsymbol{\eta}}(t)}\Big)dt\Big]\\ =& -\mathbb{E}\Big[\int_0^T g_i(t)\Big(\frac{\sum_{i=1}^M [\boldsymbol{\eta}(c)]_i g_i(t) - f_1(t)}{\sum_{i=1}^M [\boldsymbol{\eta}(c)]_i g_i(t) + c\int_0^T \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s)}\Big)dt\Big]\\ =& -\mathbb{E}\Big[\int_0^T \frac{g_i(t)\Big(\sum_{i=1}^M [\boldsymbol{\eta}(c)]_i g_i(t) - f_1(t)\Big)}{\sum_{i=1}^M [\boldsymbol{\eta}(c)]_i g_i(t)} - \frac{c\Big(\sum_{i=1}^M [\boldsymbol{\eta}(c)]_i g_i(t) - f_1(t)\Big) f_2(t)\sigma_h}{\sum_{i=1}^M [\boldsymbol{\eta}(c)]_i g_i(t)\Big(c + \sum_{i=1}^M [\boldsymbol{\eta}(c)]_i g_i(t)\Big)} + o(\sigma_h)dt\Big]\end{aligned}$$

Set $\frac{\partial \Lambda(c,\boldsymbol{\eta})}{\partial \eta_i} = 0$ for $1 \leq i \leq M$, we can solve for $\boldsymbol{\eta}(c)$, we can show, using Assumption 1, if we had a $\tilde{\boldsymbol{\eta}}_c$ that satisfies

$$0 = \mathbb{E}\Big[\int_0^T g_i(t)(1 - \frac{f_1(t)}{\sum_{i=1}^M [\tilde{\boldsymbol{\eta}}_c]_i g_i(t)})dt\Big]$$

for all i or, if all process are stationary:

$$0 = \mathbb{E}\Big[g_i\Big(1 - \frac{f_1}{\sum_{i=1}^M [\tilde{\boldsymbol{\eta}}_c]_i g_i}\Big)\Big], \tag{16}$$

for all i, then

$$\|\tilde{\boldsymbol{\eta}}_c - \boldsymbol{\eta}(c)\| = O(\sigma_h). \tag{17}$$

Now, we can investigate estimation for θ:

$$\begin{aligned}&\frac{\partial \Lambda_p(\theta)}{\partial \theta}_{|\theta=c}\\ =&\frac{\partial \Lambda(\theta, \boldsymbol{\eta})}{\partial \theta}_{|\theta=c,\boldsymbol{\eta}=\boldsymbol{\eta}(c)}\\ =& -\mathbb{E}\Big[\int_0^T \frac{\partial \lambda_{\theta,\boldsymbol{\eta}}}{\partial \theta}(t)\Big(1 - \frac{\lambda_1(t)}{\lambda_{\theta,\boldsymbol{\eta}}(t)}\Big)dt\Big]\\ =& -\mathbb{E}\Big[\int_0^T \Big(\int_0^T \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s)\Big)\Big(\frac{\sum_{i=1}^M [\boldsymbol{\eta}(c)]_i g_i(t) - f_1(t)}{\sum_{i=1}^M [\boldsymbol{\eta}(c)]_i g_i(t) + c\int_0^T \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s)}\Big)dt\Big]\end{aligned}$$

$$
\begin{aligned}
=&-\mathbb{E}\Big[\int_0^T \frac{\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)-f_1(t)}{c} - \frac{\Big(\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)-f_1(t)\Big)\Big(\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)\Big)}{c\Big(\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)+c\int_0^T \mathbf{1}_{[0,\sigma_h]}(t-s)dN_2(s)\Big)}dt\Big]\\
=&-\mathbb{E}\Big[\int_0^T \frac{\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)-f_1(t)}{c} - \frac{\Big(\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)-f_1(t)\Big)\Big(\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)\Big)}{c\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)}\\
&+\frac{c^2\Big(\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)-f_1(t)\Big)\Big(\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)\Big)f_2(t)\sigma_h}{\Big(c\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)\Big)\Big(c\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)+c^2\Big)}+o(\sigma_h)dt\Big]\\
=&-\mathbb{E}\Big[\int_0^T \frac{\Big(\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)-f_1(t)\Big)f_2(t)}{\sum_{i=1}^M[\boldsymbol{\eta}(c)]_i g_i(t)+c}\sigma_h+o(\sigma_h)dt\Big].
\end{aligned}
$$

As we can see, this gradient is off-zero by

$$
-\mathbb{E}\Big[\int_0^T \frac{\Big(\sum_{i=1}^M[\tilde{\boldsymbol{\eta}}_c]_i g_i(t)-f_1(t)\Big)f_2(t)}{\sum_{i=1}^M[\tilde{\boldsymbol{\eta}}_c]_i g_i(t)+c}dt\Big]\sigma_h+o(\sigma_h) \tag{18}
$$

so we expect the same order of error between $\hat{\theta}$ and c. Now we are ready to prove Proposition 2 and 3.

B.3 PROOF OF PROPOSITION 2.

We use the alternative intensity notation (15).

Proof of Proposition 2. When using the naive Hawkes, we only have $g(t) = 1$ the constant function, then we can solve for $\tilde{\eta}_c$, which is simply solving for the $\tilde{\eta}_c$

$$
0=\mathbb{E}\Big[\frac{\tilde{\eta}_c-f_1}{\tilde{\eta}_c}\Big]
$$

which is simply the baseline intensity mean $\tilde{\eta}_c = \mathbb{E}[f_1]$.

The gradient is thus off by zero by

$$
\mathbb{E}\Big[\frac{(f_1-\mathbb{E}f_1)f_2}{\mathbb{E}f_1+c}\Big]\sigma_h+o(\sigma_h)=\frac{\text{Cov}(f_1,f_2)}{\mathbb{E}f_1+c}\sigma_h+o(\sigma_h)
$$

The rest follows from Assumption 1 as in proof of Proposition 1. □

B.4 PROOF OF PROPOSITION 3.

We use the alternative intensity notation (15). In general case, we first assume all $f_1(t)$, $f_2(t)$ and $g_i(t)$ for $i > 1$ with $g_1 = 1$ are stationary and square integrable (so we drop the dependence on t and make this a subspace Hilbert space). For basis normalization, let us fix $g_1 = 1$ and all other $\|g_i\|_2 := \mathbb{E}[g_i^2] = 1$. More precisely, let $G \subset L^2(\Omega, \mathcal{F}, \mathbb{P})$ be a centered Hilbert Space, we assume all $\{g_i\}_{i>1} \subseteq G$, as well as $f_1 - \mathbb{E}f_1 \in G$ and $f_2 - \mathbb{E}f_2 \in G$. Furthermore, we assume the basis g_i are uncorrelated (orthogonal basis): $\langle g_i, g_j\rangle := \mathbb{E}[g_i g_j]$.

Proof of Proposition 3. Recall Lemma B.2, we first verify

$$
\begin{aligned}
[\tilde{\boldsymbol{\eta}}_c]_1 =&\mathbb{E}[f_1]\\
[\tilde{\boldsymbol{\eta}}_c]_i =&\mathbb{E}[(f_1-\mathbb{E}[f_1])g_i] \text{ for } i \geq 1
\end{aligned}
$$

is a solution for the high order approximation:

$$R(g_i(\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i - f_1), \sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i) = 0$$

To verify, first notice that

$$\mathbb{E}[f_1|\mathcal{G}] = \sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i$$

where $\mathcal{G}$ is the sigma-algebra generated by $g_1, g_2, ..., g_M$. This can be easily checked by notice that $\langle \sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i - f_1, g_i\rangle = 0$ (projection) for all i. Thus,

$$\mathbb{E}[g_i(\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i - f_1)] = \mathbb{E}[g_i\mathbb{E}[\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i - f_1|\mathcal{G}]] = 0$$

Then,

$$\begin{aligned}\text{Cov}(g_i(\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i - f_1), \sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i) =& \mathbb{E}[g_i(\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i - f_1)(\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i)] \\ =& \mathbb{E}[g_i(\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i)\mathbb{E}[\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i - f_1|\mathcal{G}]] \\ =& 0\end{aligned}$$

Then, we use Lemma B.2 and (16), (17),(18) to check

$$\mathbb{E}\left[g_i\left(1 - \frac{f_1}{\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i}\right)\right] = R(g_i(\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i - f_1), \sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i) + o(r^2) = o(r^2)$$

$$-\mathbb{E}\left[\int_0^T \frac{\left(\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i(t) - f_1(t)\right)f_2(t)}{\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i(t) + c} dt\right]\sigma_h = R((\sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i - f_1)f_2, \sum_{i=1}^{M}[\tilde{\boldsymbol{\eta}}_c]_i g_i + c) + o(r^2) = o(r^2)$$

The rest follows from $o(r^2) + o(\sigma_h) = o(r^2 + \sigma_h)$, Assumption 1 and the setting have $M = 2$ and $g_1 = 1$ and $g_2 = \frac{f_2 - \mathbb{E}[f_2]}{\sqrt{\text{Var}(f_2)}}$.

□

C SIMULATION STUDY AND EMPIRICAL VERIFICATION

This section presents a detailed simulation study and empirical verification of our model.

C.1 SINUOID BACKGROUND

Experiment settings The sinusoid function provides a simple way to control the correlation between the background activity of f_i and f_j. Consider the dynamic background as follows,

$$f_i(t) = A\sin(2\pi(t - \phi_{\text{rnd}})) \tag{19}$$

$$f_j(t) = A\sin(2\pi(t - \phi_{\text{rnd}} - \phi_{\text{lag}})) \tag{20}$$

where A is the amplitude, the length of a trial is $T = 5$ second, $\phi_{\text{rnd}} \sim \text{Uniform}(0, 1)$ varies from trial to trial so the background is not repeatedly observed, ϕ_{lag} controls the correlation between background signals, which is measured as the normalized dot product

$$\langle f_i, f_j\rangle := \frac{1}{TA^2}\int_0^T f_i(s)f_j(s)\mathrm{d}s \tag{21}$$

When $\phi_{\text{lag}} = 0, 0.5$, it achieves the largest positive or negative correlation respectively; when $\phi_{\text{lag}} = 0.25$, the correlation is zero. $A = 5$ spikes/second, $\alpha_i = \alpha_j = 30$ spikes/second. Each simulation includes 200 trials. $h_{i\to j}(t) = 2\mathbb{I}_{[0,\sigma_h]}(t)$, $\sigma_h = 30$ ms. $h_{j\to i}(t) = 0$. The error band in the main Figure 4 is obtained by repeating the simulation 100 times.

Neural Hawkes baseline model The Neural Hawkes is a representative deep learning framework for Hawkes processes [Mei and Eisner, 2017]. The key components are

$$\boldsymbol{c}(t) = \bar{\boldsymbol{c}}_{i+1} + (\boldsymbol{c}_{i+1} - \bar{\boldsymbol{c}}_{i+1}) \exp(-\boldsymbol{\delta}_{i+1}(t - t_i^{\text{source}})) \tag{22}$$

$$\boldsymbol{h}(t) = \boldsymbol{o}_i \odot \tanh(\boldsymbol{c}(t)) \tag{23}$$

$$\lambda_{\text{target}} = \left(\boldsymbol{W}_{\text{target}}^T \boldsymbol{h}\right)_+ \tag{24}$$

To make a fair comparison with other models, we replace the decay function with

$$\boldsymbol{c}(t) = \bar{\boldsymbol{c}}_{i+1} + (\boldsymbol{c}_{i+1} - \bar{\boldsymbol{c}}_{i+1}) \mathbb{I}_{[0,\sigma_h]}(t - t_i^{\text{source}}) \tag{25}$$

The model takes the superimposed time points as the input. At every step, the interval $t^{\text{target}} - t^{\text{source}}$ is passed to the model with labels of the source node and the target node. In the standard Hawkes model, the impact of the history points to the intensity of the target is $\sum_{t_m \in N_{\text{source}}, t_m < t} h_{\text{source}\to\text{target}}(t - t_m)$. In contrast, Neural Hawkes considers

$$\boldsymbol{W}_{\text{target}}^T \left[\boldsymbol{o}_i \odot \tanh\Big(\bar{\boldsymbol{c}}_{i+1} + (\boldsymbol{c}_{i+1} - \bar{\boldsymbol{c}}_{i+1}) h_{\text{source}\to\text{target}}(t - t_m)\Big)\right] \tag{26}$$

Although the model only takes one last time point t_m at each step, it can still incorporate the history (in a non-linear way) using intermediate variables $\boldsymbol{o}_i, \boldsymbol{c}_{i+1}, \bar{\boldsymbol{c}}_{i+1}$ carried by recurrent neural network (RNN). The process label of t_m is passed to the RNN through embedding, so the model can distinguish the source process. $\boldsymbol{W}_{\text{target}}$ varies from target to target, so the same source process can have different influences on different target processes.

After replacing the coupling window function from the default exponential function with the square window function, the amplitude of the coupling effect between a source node and a target node is assessed as $\boldsymbol{W}_{\text{target}}^T \left[\boldsymbol{o}_i \odot \tanh(\boldsymbol{c}_{i+1} - \bar{\boldsymbol{c}}_{i+1})\right]$ which contributes to the *increment* of the intensity. Since the output function is ReLU or softmax, which is close to the identity function on $\mathbb{R}_+$, the amplitude does not need to be rescaled. Our implemented baseline model is in the public repository.

C.2 SECOND-ORDER STATIONARY BACKGROUND

In this section, we further study the behavior of the model in (7) through simulations. The simulation setup in this section is the foundation of the following sections. In this special setup, we can approximate the bias, variance, risk, and likelihood of the estimator and they match the numerical results very well. Details and all derivations will be shown in Appendix E. Assume the fluctuating background activity f_i and f_j are second-order stationary stochastic processes, meaning $\mathbb{E}[f_i(t) f_i(t+u)]$ only depends on u but not t. A special case of the second-order stationary process is the *cluster point process* or *linear Cox process*, which is widely used in point process study Diggle [1985], Bartlett [1964] and [Daley and Vere-Jones, 2003, sec. 6.3]. We add the second-order stationary condition only to make theoretical derivations easier. More variant simulation scenarios will be shown later.

We first generate random background f_i, f_j, then generate point processes. Let $\phi_{\sigma_I}(\tau) = \frac{1}{\sqrt{2\pi\sigma_I^2}} \exp(-\frac{\tau^2}{2\sigma_I^2})$ be a Gaussian window function with scale σ_I. t_i^c are the time points of the center process determining the positions of Gaussian windows, which is generated by a homogeneous Poisson process with intensity ρ.

$$f_i(t) = f_j(t) = \sum_i \phi_{\sigma_I}(t - t_i^c) \tag{27}$$

For simplicity, we first consider the impact function in the form

$$h_{i\to j}(t) = \alpha_{i\to j} \cdot \mathbb{I}_{[0,\sigma_h]}(t) \tag{28}$$

where $\alpha_{i\to j}$ is the amplitude, and the filter length is σ_h. σ_I controls the timescale of the background activity. If σ_I is smaller, then f_i changes faster. σ_h controls the timescale of the point-to-point coupling effect. If σ_h is smaller, then neuron i influences neuron j in a shorter time range. The impact function estimator has form $\hat{h}_{i\to j} = \hat{\alpha}_{i\to j} \cdot \mathbb{I}_{[0,\sigma_h]}(t)$ with just one parameter $\hat{\alpha}_{i\to j}$ and the timescale σ_h is known. We use the thinning method to generate continuous-time point processes Ogata [1988].

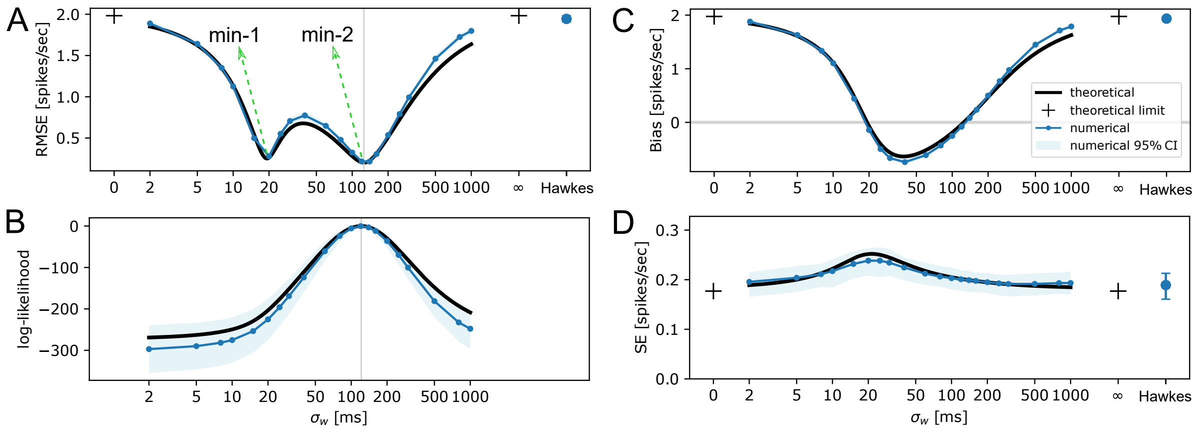

Figure 2: Numerical and analytical properties of the estimator $\hat{\alpha}_{i\to j}$. We show the properties of $\hat{\alpha}_{i\to j}$ as a function of smoothing kernel scale σ_w of W. For numerical cases, we evaluate the properties at different σ_w indicated by the blue dots. The x-axis is in logarithmic scale. The numerical (blue curves) and theoretical results (dark curves) are very close. The pointwise confidence interval for RMSE and SE is calculated using bootstrap (bootstrap the replicated estimators, not the point processes). The pointwise confidence interval for bias is calculated based on standard deviation. The blue band for the likelihood is 1.96×standard deviation. **A** The estimated risk root mean square error (RMSE) of the estimator $\hat{\alpha}_{i\to j}$. The two local minimums are labeled by "min-1" and "min-2". Our method prefers to select "min-2" indicated by the vertical line. The RMSE can be decomposed into bias (shown in **C**) and standard error (shown in **D**). **B** The maximum log-likelihood as function of σ_w. Since the likelihood functions may have different offsets, we align them by the peak (the maximum value across σ_w) to zero, then calculate the mean and pointwise standard deviation. The vertical line indicates the peak (numerical and theoretical peaks overlap), which matches the position of "min-2" in A. The theoretical extreme cases "0" and "∞" mean the scale σ_w of smoothing window W goes to limit 0 or ∞. The numerical case "no nuisance" represents the model without including the nuisance regressor $\overline{\mathbf{s}_i}$, which becomes a typical Hawkes process model ignoring the fluctuating background activity.

Figure 2 shows the numerical and analytical approximation of the properties of the estimator. The analytical formula is derived based on the second-order stationarity condition of the background activity, through which we would hope to provide insights into how the timescale of the activity is linked to the behaviors of the estimator. The activity f_i in the true model is set as a cluster process in (27) with $\sigma_I = 100$ ms. The square window filter width is $\sigma_h = 30$ ms and $\alpha_{i\to j} = 2$ spikes/sec. The firing rate of the center process $\rho = 30$ spikes/sec. The baselines are $\alpha_j = \alpha_i = 10$ spikes/sec. One simulation case has 200 trials and the length of the trial is 5 sec. Each trial is assigned with an independently generated f_i. Results in Figure 2 are obtained through 100 repetitions.

If the estimation only considers a constant baseline, as known as the standard Hawkes model, without considering the fluctuating background signal, the estimated impact function will be positively biased (Figure 2C), as the model struggles to distinguish the effect of mutual interaction from that of the correlated input between neurons. This explains why the estimated filter in the main Figure 3 is larger than the true filter.

Our model uses a smoothing kernel to eliminate the background artifacts as in main (9). When the background smoothing kernel width is too wide $\sigma_w \to \infty$ or too narrow $\sigma_w \to 0$, the nuisance variable is not able to capture any background activity, and the performance is as bad as the standard Hawkes with large positive bias (Figure 2 with labels "∞" points and "0" points). The bias becomes negative between $\sigma_w = 20$ ms and $\sigma_w = 125$ ms. The standard error in Figure 2D does not change too much as σ_w changes.

The estimated risk of the estimator has two local minimums, labeled "min-1" and "min-2" in the figure. The MLE points at "min-2", which is indicated by the vertical line in Figure 2A, B. The slope of the risk curve around "min-2" is smaller than the slope near "min-1" (the x-axis of the figure is in logarithmic scale), so the model is relatively less sensitive to the estimation or selection of σ_w near "min-2". As will be shown shortly, the position of "min-2" is related to the timescale σ_I in f_i and it is almost invariant of the impact function scale σ_h or amplitude $\alpha_{i\to j}$. The nuisance variable, the coarsened spike train $\overline{\mathbf{s}_i}$ (as in main (9)), can be interpreted as an approximation of the background activity, and σ_w reflects the timescale of the background.

C.3 INFLUENCES OF THE TIMESCALES OF COUPLING EFFECT AND BACKGROUND EFFECT

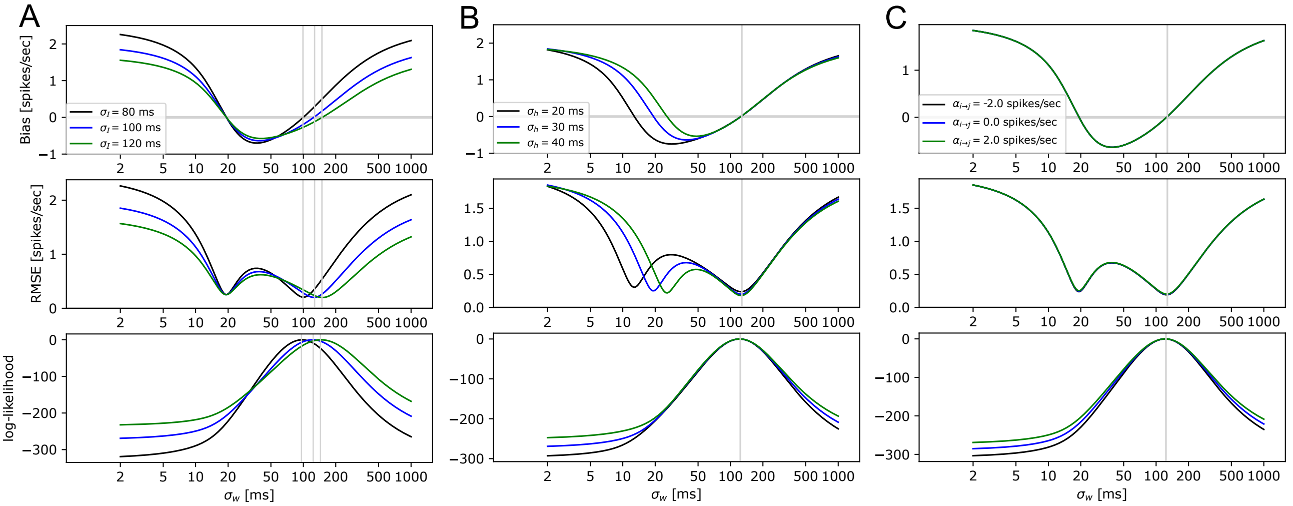

Figure 3: Relations between the estimator's properties and background activity timescale, impact function timescale, and impact function amplitude. We show the RMSE and log-likelihood curves as in Figure 2. The settings are the same as Figure 2. σ_I is tuned in A, σ_h is tuned in B, and $\alpha_{i\to j}$ is tuned in C. This figure only shows the analytical results. Numerical results match the analytical formula very well, so data is not shown. The log-likelihood functions may have different offsets, we align them by the peak to zero (maximum value across σ_w). The vertical line indicates the MLE. **A** $\sigma_I = 80, 100, 120$ ms. $\sigma_h = 30$ ms and $\alpha_{i\to j} = 2$ spikes/sec are fixed. **B** $\sigma_h = 20, 30, 40$ ms. $\sigma_I = 100$ ms and $\alpha_{i\to j} = 2$ spikes/sec are fixed. **C** $\alpha_{i\to j} = -2, 0, 2$ spikes/sec. $\sigma_I = 100$ ms and $\sigma_h = 30$ ms are fixed.

As pointed out in main (15) and derived in Appendix E, the properties of the estimator, such as bias and standard error, are related to the timescale of the background and the timescale of the coupling effect, but not the amplitude of the coupling effect. These connections are presented in Figure 3. The figure shows the relations between the estimator's properties and the timescale of the background activity (σ_I of f_i, f_j in (27)), the timescale of the point-to-point coupling activity (σ_h of the impact function in (28)), and the amplitude of the impact function ($\alpha_{i\to j}$ in (28)). In Figure 4A, the scale σ_I of the background activity f_i is related to the estimated smoothing kernel width σ_w. If σ_I is larger, the optimal σ_w also becomes larger. In Figure 4B, the scale σ_h of the impact function does not affect MLE too much, but it is related to the left root of the bias or the left local minimum of the risk. In Figure 4C, the amplitude of the impact function, whether it is positive or negative, does not change the bias or the RMSE, or the estimated σ_w. These properties suggest a simpler and heuristic way of estimating σ_w. Unlike the jitter-based conditional inference method [Amarasingham et al., 2012], our method does not rely on the assumption restricting the timescale of the background being larger than the timescale of the coupling effect. In Appendix C.6, we will show a scenario with $\sigma_I < \sigma_h$, which violates the assumption of the condition inference, but our method still works well.

The optimal smoothing kernel width σ_w is insensitive to the impact function's amplitude or timescale, which suggests a heuristic approximation for σ_w, meaning the range of the optimal kernel width σ_w can be determined before estimating the impact function. The variant of the model below without the impact function can be used for this purpose.

$$\min_{\beta_j,\beta_w,\sigma_w}\left\{-\sum_{s\in N_j}\log\tilde{\lambda}_j(s)+\int_0^T\tilde{\lambda}_j(s)\mathrm{d}s\right\} \tag{29}$$

$$\tilde{\lambda}_j(t) := \beta_j + \beta_w\,\overline{\mathbf{s}_i}(t) \tag{30}$$

$$\overline{\mathbf{s}_i}(t) = \int_0^T W(t-s)\mathrm{d}N_i(s) \tag{31}$$

C.4 CROSS-CONNECTIONS AND SELF-CONNECTIONS

As a test of a more general scenario, this simulation considers full connections cross processes and self-connection within processes. Simulation data is generated according to,

$$\begin{aligned}\lambda_j(t) &= \alpha_j + f_j(t) + \int_0^t h_{i\to j}(t-s)\mathrm{d}N_i(s) + \int_0^t h_{j\to j}(t-s)\mathrm{d}N_j(s)\\ \lambda_i(t) &= \alpha_i + f_i(t) + \int_0^t h_{j\to i}(t-s)\mathrm{d}N_j(s) + \int_0^t h_{i\to i}(t-s)\mathrm{d}N_i(s)\end{aligned} \tag{32}$$

The fluctuating background follows the linear Cox process with the same settings as in Appendix C.2. The only difference is that this scenario includes 4 impact functions $h_{i\to j} = -2, h_{j\to i} = -2, h_{i\to i} = 1, h_{j\to j} = 1$ spikes/sec. The number of samples in each simulation is 200 5-second trials, and the number of repetitions is 100, the same as Appendix C.2.

C.5 VARYING-TIMESCALE BACKGROUND

This section considers a variant of the scenario in Appendix C.2 with not only fluctuating background but also with time-varying timescale σ_I. The background activity f_i in (27) is composed of a sequence of Gaussian windows with fixed scale σ_I. The locations of the windows are randomly determined through a homogeneous Poisson process with intensity ρ. σ_I controls how fast the background changes; if σ_I is smaller, the activity will change faster. f_i is second-order stationary and some properties can be derived in closed-form formula.

Consider a similar process but the scale of the window σ_I is no longer fixed,

$$f_i = \sum_i \phi_{\sigma_{I,i}}(t - t_i^c) \tag{33}$$

where $\sigma_{I,i}$ changes randomly; every time point of the center process t_i^c is assigned with a different scale $\sigma_{I,i} \sim \mathrm{Uniform}(80, 140)$ ms. The process f_i changes faster at smaller $\sigma_{I,i}$, and changes slower at larger $\sigma_{I,i}$. The rest of the experiment settings is the same as Appendix C.2. The true impact function is a square window $h_{i\to j}(t) = \alpha_{i\to j} \cdot \mathbb{I}_{[0,\sigma_h]}(t)$, where the timescale is $\sigma_h = 30$ ms, the amplitude is $\alpha_{i\to j} = 2$ spikes/sec.

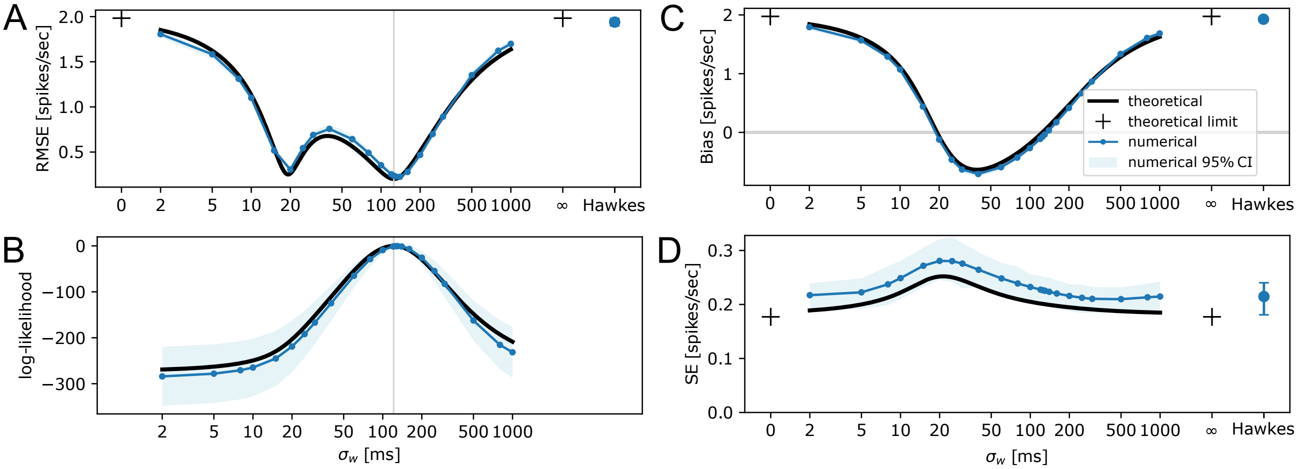

Figure 4: Simulation results of the impact function estimator with varying background activity timescale. The figure is presented in the same ways as Figure 2. The simulation details are in the text. The background activity f_i in (27) is replaced with (33) with varying timescale. The results are similar to Figure 2. The dark curves show the equivalent theoretical approximation using the model in main 7 with fixed timescale $\sigma_I = 100$ ms, which is manually tuned to match the numerical results.

As shown in Figure 4, the selected kernel width σ_w will balance the varying timescale, and it can still the select estimator with small risk and low bias, indicated by the vertical lines in Figure 4 A and B. Similar to Figure 2D, the SE does not change too much as the smoothing kernel width σ_w changes. The model can balance the bias, which can be explained by its properties in Figure 2C. If the timescale of the background σ_I is fixed and consider the bias of the estimator near the right root. If σ_w is larger than the right root, the bias will be positive, if σ_w is a little smaller than the right root, the bias will become negative. In this scenario, the timescale of the background σ_I varies. The optimal σ_w is relatively large for the activity with small σ_I, so the bias is positive for the fast-changing part. The optimal σ_w is relatively small for the sessions with large σ_I, so the bias is negative for the slow-changing part of the activity. With proper selection of σ_w, the estimator will balance the overall bias between fast- and slow-changing activities, and it can still achieve zero bias. Together with the SE, the risk properties remain similar in Figure 4 A.

To verify the reasoning, we compare the numerical results with the equivalent theoretical approximation shown in the dark curves in Figure 4. The theoretical method is for the model in main 7 with fixed timescale for the background by manually tune the timescale as $\sigma_I = 100$ ms to match the numerical curves. The behavior of the estimator for the varying-timescale background activity is almost equivalent to the case with fixed-timescale background activity. The SE of the numerical results is slightly larger though.

C.6 FAST-CHANGING BACKGROUND

In extreme cases, the background activity f_i can have fast-changing activities. In this situation, conditional inference-based method can be limited by its formalization of the null hypothesis:

> *samples from the null distribution are generated by jittering the time points by a random amount, small enough to maintain the fluctuating background intensity, but big enough to break the time association pattern.*

which implicitly assumes the timescale of the coupling effect is much smaller than the timescale of the background. This simulation scenario is similar to the setup in section 4.1, except that σ_I is set as a small value, which is comparable to or much smaller than the point-to-point interaction timescale σ_h. We will show in this section, this is not a necessary assumption or constraint for our method. Even the background changes faster than the coupling effect, our model can still have small error.

This simulation scenario is the same as Appendix C.2 except that the timescale in 27 is set to $\sigma_I = 20$ ms (Figure 5 A,B,C,D), and $\sigma_I = 8$ ms (Figure 5 E,F,G,H). The rest settings of the simulation scenarios are the same as the basic scenario in section C.2, where the true impact function is a square window $h_{i\to j}(t) = \alpha_{i\to j} \cdot \mathbb{I}_{[0,\sigma_h]}(t)$, the timescale is $\sigma_h = 30$ ms, the amplitude is $\alpha_{i\to j} = 2$ spikes/sec.

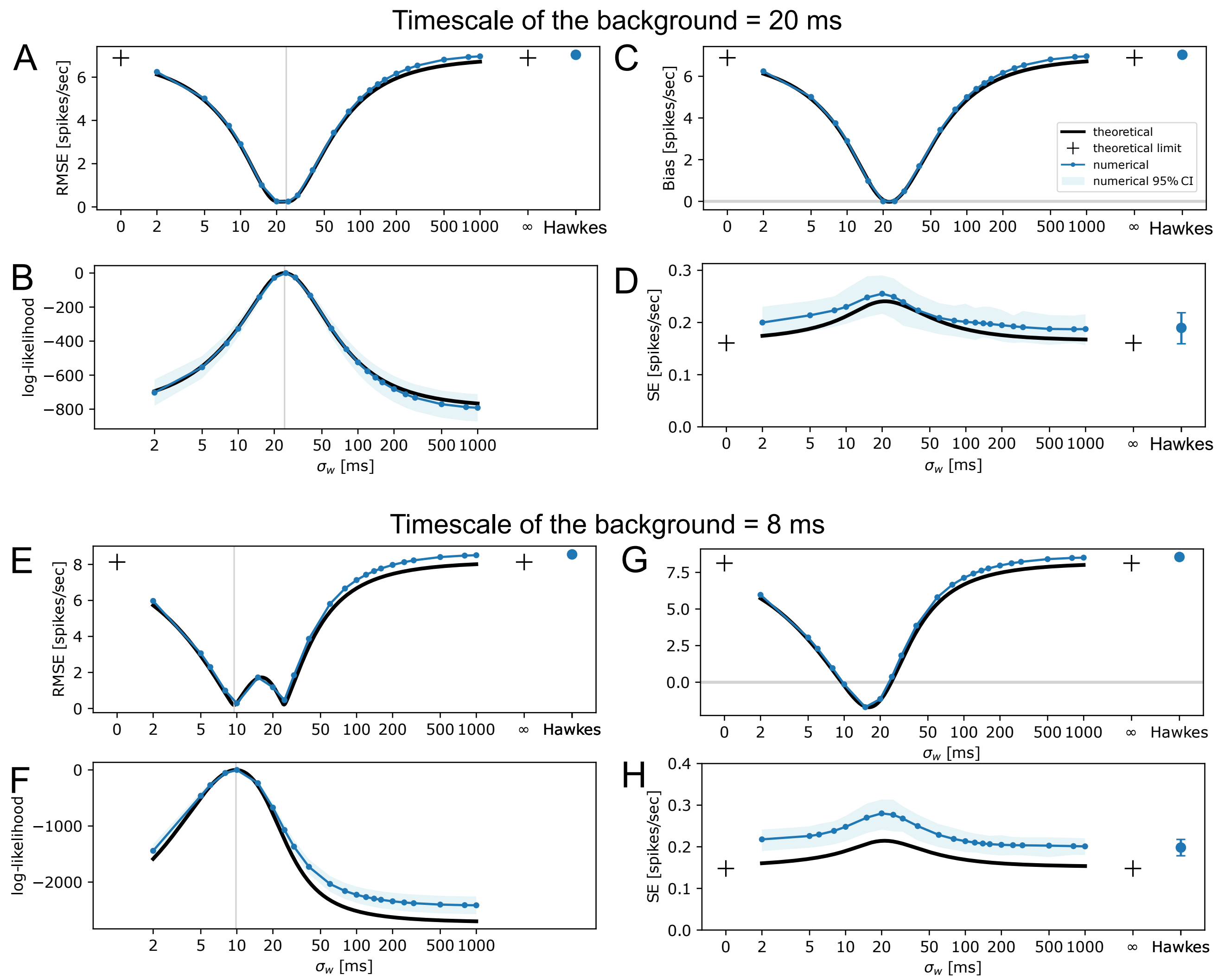

Figure 5: Fast-changing background with small σ_I. The experiment is similar to Figure 2 except that in **A, B, C, D** $\sigma_I = 20$ ms, in **E, F, G, H** $\sigma_I = 8$ ms.

As discussed in main section 4.1.2, Appendix C.3, Figure 2, and Figure 3, the right root of the bias curve (or the right local minimum of the risk curve) is associated with background timescale σ_I when $\sigma_I > \sigma_h$: if σ_I decreases, the right root will shift toward the left. Figure 5A shows a special case if σ_I keeps decreasing, two local minimums of the risk will overlap. In Figure 5C, two roots of the bias will merge to one. The property of SE does not change too much, see Figure 5D and 2D. If σ_I keeps decreasing when $\sigma_I < 20$ ms, the local minimum of the risk or the root of the bias corresponding to the MLE will move on the left side, see Figure 5E and F. We also notice that our analytical approximation of the standard error in Figure 5H begins to have a large error when σ_I is very small.

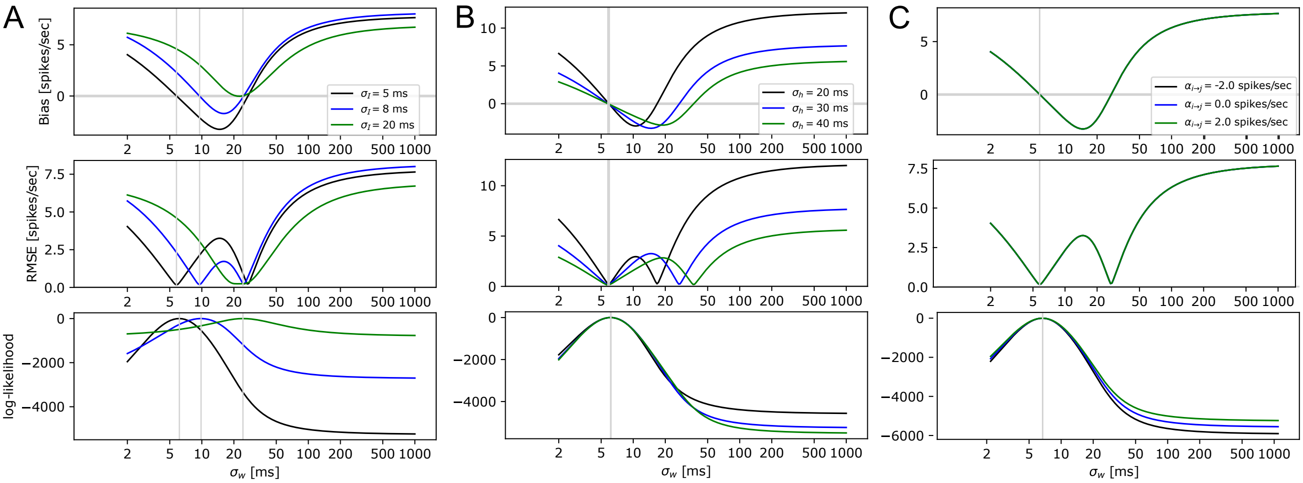

Figure 6: Properties of the estimator with fast-changing background. This figure is analog to Figure 3 but the time scale σ_I of the background activity f_i is very small. The settings are the same as Figure 2, 3, and 6 except for different tuning parameters. Only the analytical RMSE and log-likelihood curves are shown, which match the numerical results very well. Some numerical results have already been shown in Figure 5. The log-likelihood functions may have different offsets, we align them by the peak to zero (maximum value across σ_w). The vertical lines indicate the MLE. **A** Background timescale σ_I as the tuning variable. **B** Coupling effect timescale σ_h as the tuning variable. **C** Coupling effect amplitude $\alpha_{i\to j}$ as the tuning variable.

Similar to Figure 3, next, we explore how the timescale of the background activity σ_I, the timescale of coupling effect σ_h, and the amplitude of impact function $\alpha_{i\to j}$ are related to the above properties when σ_I is very small. When σ_I is tuned, it will change the optimal σ_w. If σ_I is around 20 ms, two local minimum values of the risk curve may merge to one, which agrees with the numerical result in Figure 5A. If $\sigma_I < 20$ ms, the root of the bias or the local minimum of the risk corresponding to the MLE will be on the left side. When σ_h is tuned, it will be the right local minimum of the risk or the right root of the bias that will be associated with the impact function timescale, that is opposite to the conclusion in Figure 3B. Similar to Appendix C.3, the timescale σ_h or the amplitude $\alpha_{i\to j}$ of the impact function do not affect σ_w of the MLE.

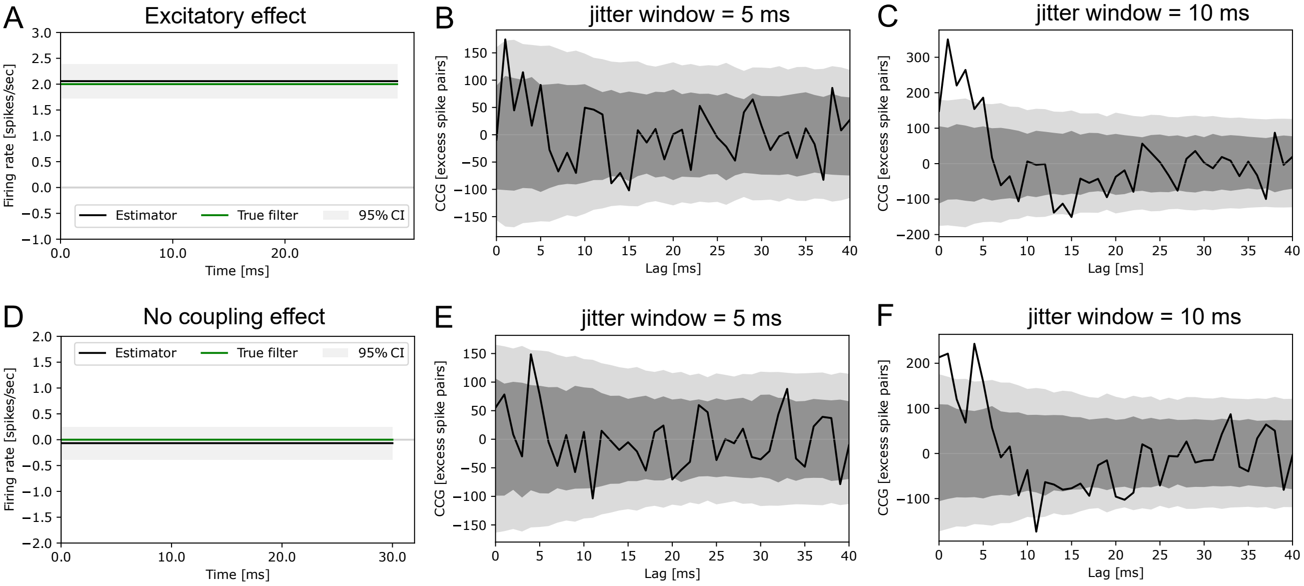

Figure 7: A comparison between the estimations of the coupling effect with fast-changing background. The figure compares the performance of the estimators of the coupling effect. The simulation settings are the same as Figure 2 except that the timescale of the background $\sigma_I = 5$ ms is very small. The timescale of the impact function as in Eq (28) is $\sigma_h = 30$ ms. In A, B, and C, the amplitude of the true impact function is $\alpha_{i\to j} = 2$ spikes/sec. In D, E, and F, the amplitude of the true impact function is $\alpha_{i\to j} = 0$ spikes/sec. **A, D** The estimator of the point process regression. It can accurately estimate the true filter, which is supported by the analysis in Figure 5 and 6. **B, E** Jitter-based CCG. The time bin for the spike train is 1 ms. The jitter window width is set as 5 ms, close to the timescale of the background activity. The dark grey band is pointwise 95% acceptance band, and the light grey band is simultaneous 95% acceptance band. The result is acquired from 1000 surrogate jitter samples. The CCG method detects a small excitatory effect before lag = 5 ms no matter whether the neurons have true coupling effect. **C, F** Similar to B except that the jitter window width is 10 ms. In both B and C, the jitter-based CCG method can only detect a small effect before 5 ms lag or 7 ms lag. A large part of the coupling effect between 0 to 30 ms is buried under the CI band. However, such an effect is due to the fast-changing background, but not the neuron-to-neuron coupling effect.

A significant advantage of our model over the jitter-based model is that the proposed model does not assume the background activity changes slower than the coupling effect, and the model can automatically find the optimal timescale. The jitter-based method can not avoid such an assumption due to its nature of conditional inference. The null hypothesis states that the coupling effects do not change faster than the jitter window width. Thus the samples under the null distribution are obtained by randomly jittering the points within the jitter window. If the background changes as fast as the coupling effect, such bootstrapping method can not maintain the temporal structure of the background activity, so it can not split the background artifacts and the coupling effects. In other words, if the jitter window is set a little larger than the coupling effects, it can not tell whether the detected effect belongs to the background or the point-to-point interaction. Some other bootstrapping methods have the same issue for exactly the same reason Cowling et al. [1996]. Figure 7 compares the point process regression method and jitter-based CCG method. The simulation scenario is the same as the basic model in Figure 2 except that the timescale of the background activity is very small $\sigma_I = 5$ ms. The numerical properties of the estimator have been shown in Figure 5E-H. The true impact function is a square window $h_{i\to j}(t) = \alpha_{i\to j} \cdot \mathbb{I}_{[0,\sigma_h]}(t)$. The timescale of the coupling effect is $\sigma_h = 30$ ms. The true amplitude of the impact function is $\alpha_{i\to j} = 2$ spikes/sec in Figure 7A,B,C, and $\alpha_{i\to j} = 0$ spikes/sec in Figure 7E,E,F. In both cases, the regression method can accurately estimate the true estimator, which agrees with the numerical and theoretical results in Figure 5. In Figure 7 B and C, the CCG method with jitter window width = 5 or 10 ms can detect some excitatory effect in a lag range smaller than 5 ms or 10 ms. Nevertheless, it misses the excitatory effect between lag=10 to 30 ms. It is unreasonable to use a larger jitter window, as it will not match the background timescale. The CCG results are similar to another example in Figure 7E, F, where there is no coupling effect. The detected significant data points are totally due to the fast-changing background. So for the results in Figure 7B and C, we can not conclude that the jitter method has removed the background artifacts and the significant effect is caused by the coupling effect.

C.7 ASYMPTOTIC NORMALITY OF THE ESTIMATOR

In this section, we perform simulations to verify the asymptotic normality property of the estimator empirically. The dataset is the same as Figure 2. The true impact function is a square window $h_{i\to j}(t) = \alpha_{i\to j} \cdot \mathbb{I}_{[0,\sigma_h]}(t)$, where the amplitude is $\alpha_{i\to j} = 2$ spikes/sec, and the timescale is $\sigma_h = 30$ ms. The estimator for the coupling effect $\hat{\alpha}_{i\to j}$ is 7. We compare the empirical distribution of the estimators with the theoretical distribution. As shown by Figure 8, the estimator has normal distribution at the optimal selection of $\sigma_w = 120$ ms. If σ_w is too small or too large, the empirical distributions will have visible deviations.

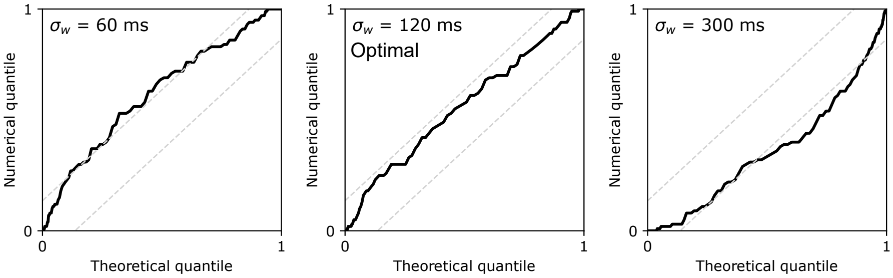

Figure 8: Normality of the estimator's distribution. The dataset is the same as Figure 2 including 100 repetitions. The figure shows the Q-Q plots of the empirical distribution of the estimator against the theoretical normal distribution, shown in the dark curves. The straight dashed grey lines are 95% uniform CI. In the first row, the empirical distribution matches the theoretical distribution very well at the optimal model ($\sigma_w = 120$ ms). We also evaluate the model at other different smoothing kernel widths. If σ_w is too small or too large $\sigma_w = 60, 300$ ms), the empirical distribution will have a visible deviation from the theoretical distribution.

C.8 SELECTION OF IMPACT FUNCTION LENGTH.

The simulation scenario in the main text in Figure 2 and many scenarios in the supplementary sections simplify the impact function estimation using a square window and assume the timescale of the coupling effect σ_h in Eq (28) is known. In this section, we show the consequences of unmatched impact function timescale. Because in practice, the timescale of the coupling effect is usually unknown.

We used the same dataset in Figure 2, where the true impact function is a square window. The amplitude $\alpha_{i\to j} = 2$ spikes/sec and the window width is $\sigma_h = 30$ ms. We applied two versions of the regression model to the dataset. Both versions used a square window as the impact function estimator, but one with shorter timescale $\sigma_{h,1} = 20$ ms, the other with longer timescale $\sigma_{h,2} = 40$ ms. We present the results in Figure 9 in the same way as Figure 2.

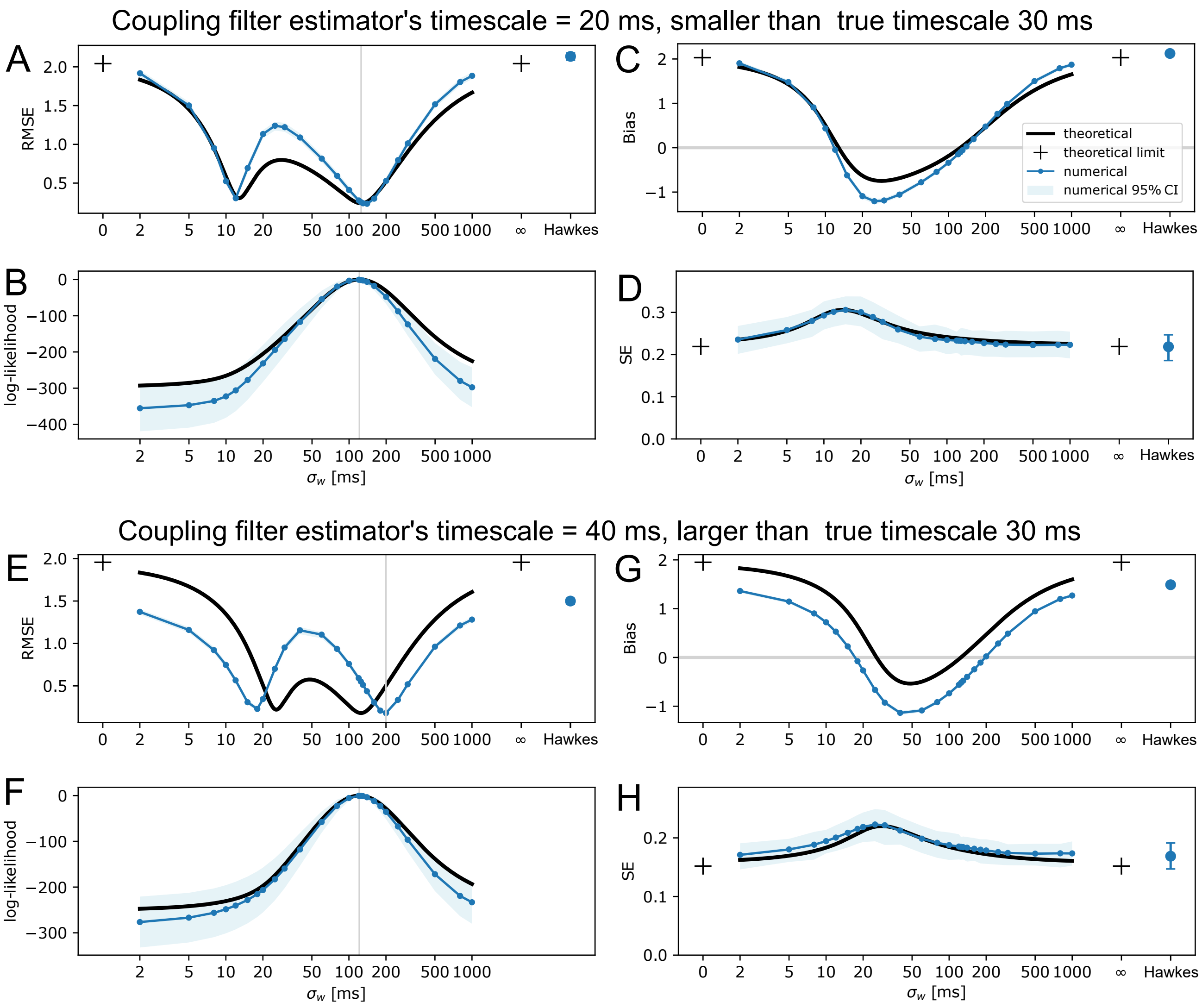

Figure 9: Consequences of unmatched impact function timescale. The dataset is the same as Figure 2, where the true impact function is a square window. The amplitude is $\alpha_{i\to j} = 2$ spikes/sec and the window width is $\sigma_h = 30$ ms. We tested the regression model with unmatched impact function width. The model in **A, B, C, D** estimates the impact function using a shorter timescale $\sigma_{h,1} = 20$ ms. The model in **E, F, G, H** estimates the impact function using a longer timescale $\sigma_{h,1} = 40$ ms. As a reference, the dark curves show the theoretical approximation using the basic regression model by setting the impact function timescale as 20 ms in A-D, and 40 ms in E-H. The rest settings are the same as the simulation.

If the impact function is estimated using a shorter timescale ($\sigma_{h,1} = 20$ ms) as shown in Figure 9A,B,C,D, the selected model still has the minimum risk indicated by the vertical line. The dark curves in Figure 9A,B,C,D, show the theoretical approximation of the properties using the basic regression model in (7) by setting the impact function timescale as 20 ms instead, which can be seen as the expected properties of the model. The absolute values of the bias are larger than expected if σ_w is between 10 ms and 120 ms or larger than 200 ms. But the roots of the bias still match the expected position. The SE is not affected by the unmatched timescale. So the optimal selection of σ_w does not change. As a contrast, if the impact function timescale of the estimator ($\sigma_{h,2} = 40$ ms) is longer than the truth (30 ms), the consequence is more severe. As shown in Figure 9 G, the actual bias is uniformly lower than the expected bias. The SE is not affected. So the consequence is that the selected smoothing kernel width σ_w (Figure 9 F vertical line) does not match the actual risk minimum (Figure 9 E vertical line).

By combining the results of the two cases, we recommend users select shorter impact function timescale if they are not confident about the impact function timescale, or using non-parametric fitting as in Supplementary D.1.

C.9 MULTIVARIATE REGRESSION AND PARTIAL RELATION

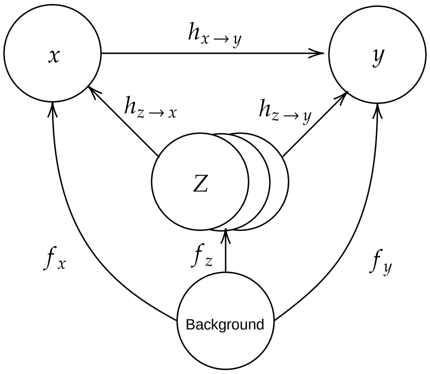

Figure 10: Diagram of multivariate point process network driven by background activity.

Multivariate regression is a natural extension of the basic regression model introduced in the main text. The diagram is shown in Figure 10.

$$\lambda_y(t|\mathcal{H}_t) = \alpha_y + \int_0^t h_{x\to y}(t-\tau)N_x(\mathrm{d}\tau) + \underbrace{f_y(t) + \sum_{z\in Z}\int_0^t h_{z\to y}(t-\tau)N_z(\mathrm{d}\tau)}_{\tilde{f}_y(t)} \tag{34}$$

where besides the interaction between node x and y, they are both connected to other nodes denoted by Z. The whole network is also driven by unobserved background activity. This scenario is motivated by the challenge in practice: only part of the network can be observed with a limited number of nodes; Besides interaction across nodes, the network is also driven by other factors, usually not directly observed, see Figure 10). The input of node y includes the coupling effect $h_{x\to y}$ from x, or $h_{z\to y}$ from other nodes $z \in Z$, and background influence f_y. The sum input of $h_{z\to y}$ and f_y can be seen as $\tilde{f}_y$. Similarly, the total input for x is denoted by $\tilde{f}_x$. Our goal is to estimate $h_{x\to y}$ as the target relation conditioning on both background and Z, by properly handling the correlation between $\tilde{f}_x$ and $\tilde{f}_y$. The multivariate regression problem is reduced to the pairwise bivariate regression problem as the main text. This case is also inspired by [Chen et al., 2017], where authors proposed that the coupling effect in multivariate point process regression problems can be approximated by pairwise cross-correlation very well. But their method assumes constant baselines and only positive impact function functions. Next, we demonstrate using simulations to show our model is promising to overcome these limitations.

The simulation scenarios follow the diagrams in Figure 10. The process in Z and x, y are all driven by fluctuating background $f_x = f_y = f_z$ set as the linear Cox process as 27 in section C.2, where the intensity of the center process is $\rho = 20$ spikes/sec, and the window function is Gaussian with scale $\sigma_I = 100$ ms. The constant baseline of all processes is 10 spikes/sec. The network includes 6 nodes, coupled with square window function $\alpha_{i\to j}\mathbb{I}_{[0,\sigma_h]}(t)$, $\sigma_h = 30$ ms as known. The amplitude $\alpha_{i\to j}$ are positive, negative, or zero. Each simulation has 200 trials with a 5-second duration. The performance in the main Table 2 is obtained from 100 repetitions.

D SOME USE CASES OF THE MODEL

D.1 NON-PARAMETRIC FITTING FOR THE IMPACT FUNCTION

For simplicity, the models presented in the main text and many sections in the appendix use a square window for the impact function. This section considers non-parametric fitting for the impact function through splines. The linear form of the intensity function in main (7) can be easily extended for this purpose, also see Appendix A. The impact function now is estimated as a linear combination of spline bases as follows,

$$h_{i\to j}(s) = \beta_{h,1}B_1(s) + ... + \beta_{h,k}B_k(s)$$

where $B_1, ..., B_k$ are spline bases. Define the covariates in the regression,

$$\phi_{h,1}(t) := \int B_1(t-s)N_i(\mathrm{d}s), ..., \ \phi_{h,k}(t) := \int B_k(t-s)N_i(\mathrm{d}s)$$

The intensity function becomes,

$$\tilde{\lambda}_j(t) = \beta_j + \beta_w \overline{\mathbf{s}_i}(t) + \beta_{h,1}\phi_{h,1}(t) + ... + \beta_{h,k}\phi_{h,k}(t)$$

$\overline{\mathbf{s}_i}(t)$ is same as the coarsened spike train in main (9). The coefficients of the impact function $\beta_{h,1}, ..., \beta_{h,k}$ can still be estimated using the model in 7. The optimization algorithm is in Appendix A. We applied the non-parametric fitting to the dataset in Figure 2. The true impact function is a square window $h_{i\to j}(t) = \alpha_{i\to j} \cdot \mathbb{I}_{[0,\sigma_h]}(t)$ where the amplitude of the impact function is $\alpha_{i\to j} = 2$ spikes/sec, the timescale of the square window is $\sigma_h = 30$ ms. The impact function is estimated in a lag window between 0 and 50 ms using B-splines with 9 equal-distance knots. We evaluate the risk using root-mean-integral-square error (RMISE). The RMISE between the true impact function $h(t)$ and the estimator $\hat{h}(t)$ is defined as follows. $L_h = 50$ ms is the length of the impact function.

$$\mathrm{RMISE}(h, \hat{h}) := \sqrt{\frac{1}{L_h}\int_0^{L_h} \left(\hat{h}(t) - h(t)\right)^2 \mathrm{d}t}$$

We evaluate the bias and the standard error of the filter at lag 5, 15, 25 ms. The result is shown in Figure 11 below.

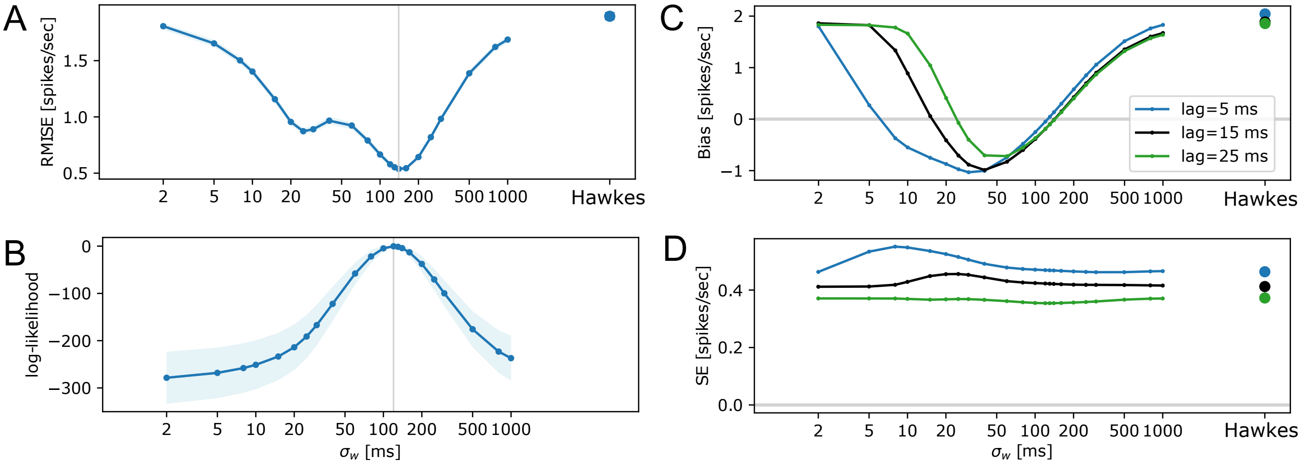

Figure 11: Non-parametric fitting for the impact function. The dataset and the non-parametric estimator are described in the test. The results are presented in the same way as Figure 2. **A** RMISE of the estimated impact function as a function of smoothing kernel width σ_w of W in main (9). The vertical line indicates the minimum risk. **B** The maximum log-likelihood as function of σ_w. Since the likelihood functions may have different offsets, we align them by the peak (the maximum value across σ_w) to zero, then calculate the mean and pointwise standard deviation. The vertical line indicates the peak of the mean log-likelihood. **C** The bias of the estimator is evaluated at lag = 5, 15, 25 ms. **D** The standard error of the estimator is evaluated at $h(\mathrm{lag})$, lag = 5, 15, 25 ms.

The risk curve and the log-likelihood curve are similar to the result in Figure 2. The optimal model with minimum risk can be selected by maximizing the likelihood, which is the same as the basic regression scenario in Figure 2. The difference is that, in the non-parametric fitting, the left local minimum risk has a higher value than the right local minimum. While in the basic fitting case, two local minimum values of the risk curve are close in Figure 2). This can be explained by decomposing the risk into bias and SE shown in Figure 11C and D. If the smoothing kernel width σ_w is around 130 ms, the bias values at different lags of the impact function are nearly the same. But if σ_w is around 20 ms, the bias values at different lags have large divergence: the beginning part of the estimator at lag=5 ms has a negative bias, the middle part at lag=15 ms has around zero bias, and the end part of the estimator at lag=25 ms has a positive bias. The SE of the estimator at different lags does not change a lot as σ_w varies. So overall, the RMISE has a much larger value near lag=20 ms than at lag=130 ms. These properties are further demonstrated in Figure 12.

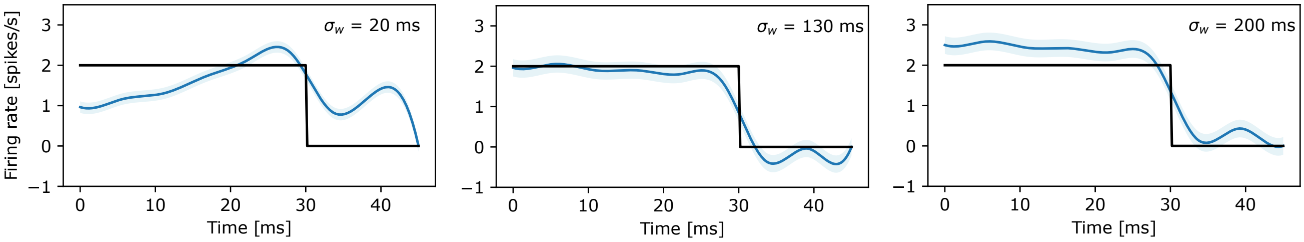

Figure 12: Non-parametric fitting for the impact function. The figure compares the true impact function (dark) and the estimator (blue). The light blue band is pointwise 95% CI. The impact functions were fitted in the same way as described in Figure 11. This figure picks out some fitted estimators with different smoothing kernel widths $\sigma_w = 20, 130, 200$ ms. If the smoothing kernel width is too small ($\sigma_w = 20$ ms), the bias values of the estimator at different lags have large differences. This matches the bias curves shown in Figure 11C. At the beginning part of the estimated impact function around lag=5 ms, the bias is negative, and at the end part around lag=25 ms, the bias is positive. If the smoothing kernel width is selected optimally ($\sigma_w = 130$ ms), the fitted impact function matches the true filter very well. If the smoothing kernel width is too wide ($\sigma_w = 200$ ms), the whole estimated impact function has uniform positive bias at different lags. This agrees with Figure 11C that multiple bias curves with different lags beyond $\sigma_w = 130$ ms are very close.

D.2 HYPOTHESIS TESTING EXAMPLE

The regression model can be adopted for hypothesis testing problems. The simulation scenario in this section is similar to the case in Figure 2. The background activity is a cluster point process in (27), and the impact function is a square window $h_{i\to j}(t) = \alpha_{i\to j} \cdot \mathbb{I}_{[0,\sigma_h]}(t)$ in (28). Each simulation dataset only has 10 trials. We reduce the sample size to make the tasks more difficult, and the performances of different estimators will be more distinguishable. The length of the trial is 5 seconds, the time scale of the background activity is $\sigma_I = 100$ ms. The intensity of the center process is $\rho = 30$ spikes/sec. The baselines of two neurons are $\alpha_i = \alpha_j = 10$ spikes/sec. The impact function is estimated using a square window with known timescale $\sigma_h = 30$ ms. The amplitude of the impact function is $\alpha_{i\to j} = 0$ spikes/sec in the null cases without coupling effects. We include three true positive scenarios with impact function amplitudes $\alpha_{i\to j} = 2, -2, 1$ spikes/sec respectively. The dataset generating and the model fitting procedure was repeated for 100 times.

Consider the null hypothesis

$$H_0 : \ \hat{\alpha}_{i\to j} = 0$$

$\hat{\alpha}_{i\to j}$ is the estimator for $\alpha_{i\to j}$. The inference method is a direct application of the properties of the estimator. The smooth kernel is $\sigma_w = 125$ ms, which is chosen by maximizing the likelihood. $\hat{\alpha}_{i\to j}$ has asymptotic normal distribution (see details in Appendix C.7 and Appendix E), so the p-value can be easily calculated accordingly. The alternative method is jitter-based CCG, where the time bin width is 2 ms, the jitter window width is 100 ms. The CCG with shorter or longer jitter window width, for example 60 ms or 140 ms, gives similar results, so the figures are not shown. The p-value of the method is obtained by considering the multiple testing across all time lags between 0 and 30 ms, which is the same as the true impact function length. The calculation detail is in Amarasingham et al. [2012] supplementary document.

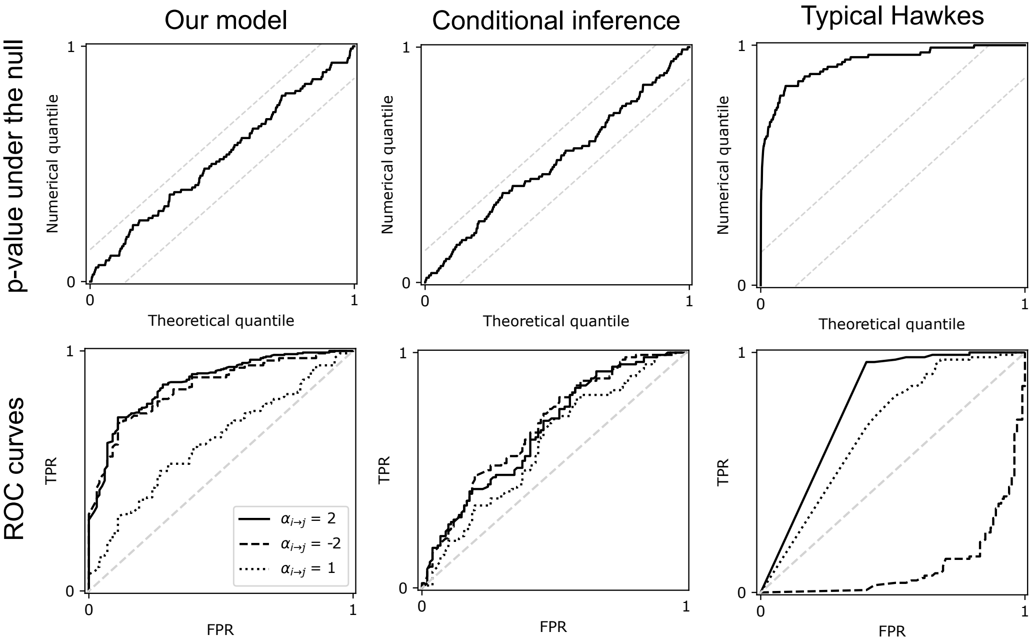

Figure 13: Hypothesis testing examples. 3 models are compared: our point process regression model (first column), jitter-based cross-correlation (CCG) method (second column), and the typical Hawkes model (third column). The simulation details are in the text. The first row shows the Q-Q plot between p-value distribution under the null (numerical quantile along the y-axis) and the uniform distribution (theoretical quantile along the x-axis). The dashed line is the 95% CI. The second row shows the results of ROC analysis with the false positive rate (FPR) along the x-axis, and the true positive rate (TPR) along the y-axis. The score of an outcome is the p-value of the hypothesis test.

Figure 13 first row verifies if the p-value distribution under the null is uniform. Both our method and conditional inference yield proper p-value distributions. As typical Hawkes model suffers large error due to background artifacts, the p-value under the null is ill. Figure 13 second row presents the ROC analysis. The score of a test outcome is the p-value. The our method has better performance when the amplitude of the coupling effect is $\alpha_{i\to j} = 2, -2$ spikes/sec. Neither methods has satisfactory performance if the coupling effect is as weak as $\alpha_{i\to j} = 1$ spikes/sec. Usually, the jitter-based method focuses on pointwise statistic at a specific time lag, which can ignore the connection between adjacent time lags. We think the power of the CCG method can be improved by considering the time lag dependency and designing the multiple hypothesis test more carefully, but it is not the main interest of this paper. Typical Hawkes model has positive bias in this setting. When the coupling effect is inhibitory $\alpha_{i\to j} = -2$, some of the outcome will be detected as excitatory instead of inhibitory, so the ROC curve is under the diagonal. When the coupling effect is excitatory $\alpha_{i\to j} = 1, 2$, the model will be over confident, so large p-values in the outcome are missing, for example, when $\alpha_{i\to j} = 2$ the smallest FPR observed is around 0.4.

D.3 SIMPLE BAYESIAN MODEL

The regression model in main (7) is probabilistic, so it can be easily adapted for Bayesian inference. In this section, we present some simple Bayesian models where the scale σ_w of the smoothing kernel W in (9) can be treated as a random variable. We want to investigate how incorporating the uncertainty of the smoothing kernel width affects the estimation of the impact function, and how the variance of the background timescale affects the uncertainty of the smoothing kernel width. The posterior of the impact function coefficients obtained using the sampling-based method can also verify the Normality property in the regression method when the sample size dominates the prior.

We consider two Bayesian models below and the basic point process regression model. The likelihood of the model is the same as the main (7). The impact function is estimated using a square window, same as (28). We choose non-informative flat priors for all the variables. As the sample size is large, the posterior does not heavily rely on the prior. Model 2 is similar to the regression model, where the kernel width σ_w is selected using the same way as the regression model and held as fixed. Model 2 and the regression model are expected to have similar results. In Model 1, σ_w is a random variable. We performed the estimation on two datasets: The first dataset is the same as the example in Figure 2 (details are in the main text); The second one is the same as the scenario in Appendix C.5, where the timescale of the background activity σ_I randomly changes in a continuous range between 80 ms and 140 ms. The true impact function is a square window $h_{i\to j}(t) = \alpha_{i\to j} \cdot \mathbb{I}_{[0,\sigma_h]}(t)$ where the amplitude of the impact function is $\alpha_{i\to j} = 2$ spikes/sec. The timescale of the square window is $\sigma_h = 30$ ms. We used the Hastings-Metropolis method for the model inference, which was a Monte Carlo Markov Chain (MCMC) sampler. The posterior was acquired by drawing 1000 samples. The model was initialized using the basic point process regression method main (7). The basic regression model approximates the estimator's distribution using Normal distribution; the mean is the MLE $\hat{\alpha}_{i\to j}$, and the standard error is from the Fishier information.

Model 1:

$$\beta_j, \beta_w, \alpha_{i\to j}, \sigma_w \propto 1$$
$$p(\beta_j, \beta_w, \alpha_{i\to j}, \sigma_w | \mathbf{s}_j, \mathbf{s}_i) \propto p(\mathbf{s}_j | \mathbf{s}_i, \beta_j, \beta_w, \alpha_{i\to j}, \sigma_w)$$

where $p(\mathbf{s}_j | \mathbf{s}_i, \beta_j, \beta_w, \alpha_{i\to j}, \sigma_w)$ is the likelihood function of the point process similar to main (7). σ_w is a variable of the model.

Model 2: σ_w is fixed and the parameter selection follows the regression method.

$$\beta_j, \beta_w, \alpha_{i\to j} \propto 1$$
$$p(\beta_j, \beta_w, \alpha_{i\to j} | \mathbf{s}_j, \mathbf{s}_i) \propto p(\mathbf{s}_j | \mathbf{s}_i, \beta_j, \beta_w, \alpha_{i\to j}, \sigma_w)$$

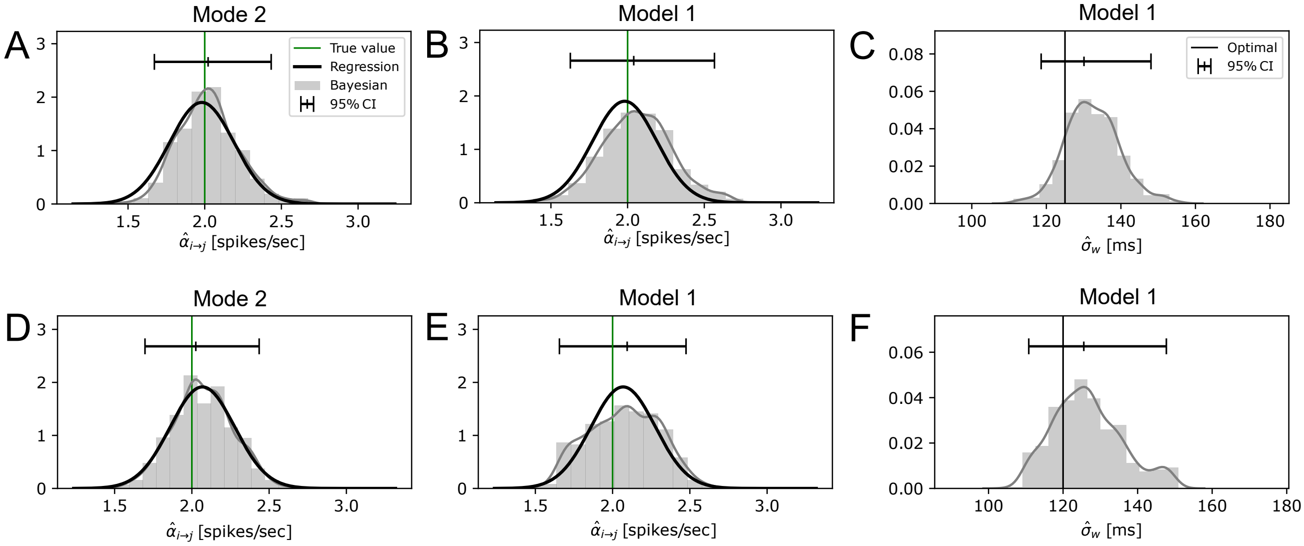

Figure 14: Applications of Bayesian model 1 and model 2 to two datasets. The figure shows the posterior of the estimated impact function amplitude $\hat{\alpha}_{i\to j}$ of Bayesian models 1 and 2 (grey histograms in A,B,D,E) and the posterior of the smoothing kernel scale $\hat{\sigma}_w$ of model 1 (grey histograms in C, F). The solid dark curves are the Normal distributions of $\hat{\alpha}_{i\to j}$ obtained using the point process regression in main (7). **A, B, C** Applications of Bayesian models 1, 2, and the basic regression model to the dataset in Figure (2). The timescale of the background activity f_i, f_j is fixed at $\sigma_I = 100$ ms. Details of the dataset description is in the main text. In C, the mode of the kernel scale is 130 ms, and the 95% CI is [119, 148] ms. The optimal kernel scale σ_w selected by the regression model is 125 ms. **D, E, F** Applications of Bayesian models 1, 2, and the basic regression model to the dataset in section C.5, where the timescale of the shared activity σ_I varies from 80 ms to 140 ms. In F, the mode of the kernel scale is 126 ms, and the 95% CI is [111, 148] ms. The optimal kernel scale σ_w selected by the regression model is 120 ms.

Figure 14 presents the estimated impact function coefficient of Bayesian models 1, 2, and the basic regression model using two simulation datasets. One dataset has fixed shared activity timescale $\sigma_I = 100$ ms, shown in plot A,B,C; the other dataset has a time-varying timescale in a continuous range between 80 ms and 140 ms, shown in plots D,E,F. In A and D, the posterior distributions of $\hat{\alpha}_{i\to j}$ (grey histogram) and the estimated distribution of the regression model (solid curves) are very close. This can be a side proof of the result in Appendix C.7, that $\hat{\alpha}_{i\to j}$ has asymptotic Normal distribution. In both datasets (first row and second row), by comparing the results between model 1 and model 2, incorporating the uncertainty of the smoothing kernel scale σ_w does not change the posterior of $\hat{\alpha}_{i\to j}$ too much. As shown in Figure 3, the selected smoothing kernel scale σ_w is related to the timescale of the shared activity σ_I. If σ_I increases, the corresponding selected σ_w will increase by around the same amount. By comparing Figure 14C and F, the CI width does not change a lot (from 29 ms in C to 37 ms in F) when the timescale of the background switched from a fixed value $\sigma_I = 100$ ms to a randomly varying value in $[80, 140]$ ms. So the uncertainty of the σ_w does not directly reflect the variance of the shared activity timescale.

E DERIVATIONS RELATED TO MAIN EQUATION 15

In this section, we provide derivations of the estimator properties, including bias, standard error, risk. All of these are based on the second-order stationary background similar to (27) in main 4.1.

Definition E.1. *Let ξ be a second-order stationary random measure on $\mathcal{X}$. It satisfies two properties Daley and Vere-Jones [2003]:*

1. *The first-moment measure is $M_{\xi,1}(A) := \mathbb{E}\xi(A)$, where A is a set in the Borel σ-field of $\mathcal{X}$, satisfies,*

$$M_{\xi,1}(\mathrm{d}x) = \bar{\lambda}\mathrm{d}x \tag{35}$$

   where $\bar{\lambda}$ is a constant, which is called the mean density.

2. *The second-moment measure is $M_{\xi,2}(A \times B) := \mathbb{E}\xi(A)\xi(B)$. A, B are sets in the Borel σ-field of $\mathcal{X}$. The second-moment can be expressed as the product of a Lebesgue component $\mathrm{d}x$ and a reduced measure, say $\breve{M}_{\xi,2}$. $\breve{m}_{\xi,2}$ is the density of the reduced measure $\breve{M}_{\xi,2}(\mathrm{d}u) = \breve{m}_{\xi,2}(u)\mathrm{d}u$. The following equation holds,*

$$\int_{\mathcal{X}}\int_{\mathcal{X}} f(s,t) M_{\xi,2}(\mathrm{d}s \times \mathrm{d}t) = \int_{\mathcal{X}}\int_{\mathcal{X}} f(x, x+u)\mathrm{d}x \cdot \breve{m}_{\xi,2}(u)\mathrm{d}u \tag{36}$$

The reduced second-moment measure $\breve{M}_{\xi,2}$ is symmetric, positive, positive-definite and translation-bounded. Details can be found in [Daley and Vere-Jones, 2003, proposition 8.1.I, 8.1.II]. The mean corrected process is $\tilde{\xi}(A) := \xi(A) - \bar{\lambda}\ell(A)$. Similarly, the reduced covariance measure and its density can be defined as,

$$\breve{C}_{\xi}(\mathrm{d}u) := \breve{M}_{\tilde{\xi},2}(\mathrm{d}u) = \breve{M}_{\xi,2}(\mathrm{d}u) - \bar{\lambda}^2\mathrm{d}u \tag{37}$$

$$\breve{c}_{\xi}(u) = \breve{m}_{\xi,2}(u) - \bar{\lambda}^2 \tag{38}$$

Lemma E.2. *Assuming $f_i = f_j$ is second-order stationary. The intensities of two coupling processes are $\lambda_j(t) = \alpha_j + f_j(t) + \int_0^t h_{i\to j}(t-\tau)N_i(\mathrm{d}\tau)$ and $\lambda_i(t) = \alpha_i + f_i(t)$. The impact function has format $h_{i\to j}(\tau) = \alpha_{i\to j}h(\tau)$ where only the amplitude needs to be fitted, the bias of the estimator $\hat{\alpha}_{i\to j}$ using model* (7) *is approximated as,*

$$\mathrm{bias}(\hat{\alpha}_{i\to j}) \approx \frac{\langle W, W\rangle_{\breve{c}_N}\langle h, \mathbf{1}\rangle_{\breve{c}_\Lambda} - \langle h, W\rangle_{\breve{c}_N}\langle W, \mathbf{1}\rangle_{\breve{c}_\Lambda}}{\langle W, W\rangle_{\breve{c}_N}\langle h, h^-\rangle_{\breve{c}_N} - \langle W, \mathbf{1}\rangle_{\breve{c}_\Lambda}^2} \tag{39}$$

where $\breve{c}_N$ is the reduced second-order moment measure intensity of spike count measure $N_i(\cdot)$; $\breve{c}_\Lambda$ is the reduced second-order moment measure intensity of the intensity measure $\Lambda_i(\cdot)$ as described in E.1. $*$ denotes the convolution, $\mathbf{1}$ is a constant function, $h^-(\tau) = h(-\tau)$. The operator between two functions g_1, g_2 is defined as

$$\langle g_1, g_2\rangle_{\breve{c}} := \int [g_1 * g_2](s)\breve{c}(\mathrm{d}s)\mathrm{d}s \tag{40}$$

Additionally, if the background activity f_i follows the cluster process in (27) *with parameters σ_I, ρ, and the impact function* *has form in* (28) *with parameters* σ_h*, then we have* $\mathrm{bias}(\hat{\alpha}_{i\to j}) \approx \mathrm{Numerator}/\mathrm{Denominator}$ *as follows,*

$$\begin{aligned}
\mathrm{Numerator} =& \left(\frac{\rho}{2\sqrt{\pi}\sqrt{\sigma_w^2+\sigma_I^2}} + \frac{\bar{\lambda}_i}{2\sqrt{\pi}\sigma_w} \right) \cdot \left(\frac{\rho}{2}\mathrm{erf}\left(\frac{\sigma_h}{2\sigma_I}\right) \right) \\
&- \left(\frac{\rho}{2}\mathrm{erf}\left(\frac{\sigma_h}{2\sqrt{\sigma_w^2/2+\sigma_I^2}} \right) + \frac{\bar{\lambda}_i}{2}\mathrm{erf}\left(\frac{\sigma_h}{\sqrt{2}\sigma_w} \right) \right) \cdot \left(\frac{\rho}{2\sqrt{\pi}\sqrt{\sigma_w^2/2+\sigma_I^2}} \right) \\
\mathrm{Denominator} =& \left(\frac{\rho}{2\sqrt{\pi}\sqrt{\sigma_w^2+\sigma_I^2}} + \frac{\bar{\lambda}_i}{2\sqrt{\pi}\sigma_w} \right) \cdot \\
&\left(\rho \left[\sigma_h \mathrm{erf}\left(\frac{\sigma_h}{2\sigma_I}\right) - \frac{2\sigma_I}{\sqrt{\pi}} \left(1 - e^{-\frac{\sigma_h^2}{4\sigma_I^2}} \right) \right] + \bar{\lambda}_i \sigma_h \right) \\
&- \left(\frac{\rho}{2}\mathrm{erf}\left(\frac{\sigma_h}{2\sqrt{\sigma_w^2/2+\sigma_I^2}} \right) + \frac{\bar{\lambda}_i}{2}\mathrm{erf}\left(\frac{\sigma_h}{\sqrt{2}\sigma_w} \right) \right)^2
\end{aligned} \tag{41}$$

$\bar{\lambda}_i = \mathbb{E}[N_i(\mathrm{d}t)/\mathrm{d}t] = \alpha_i + \rho$. $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{d}t$.

Proof. The bases of the regression model (8) include a constant, the nuisance variable $\overline{\mathbf{s}_i} = W * \mathbf{s}_i$, and the impact function term $h_{i\to j} * \mathbf{s}_i$. The target is,

$$\tilde{\ell} := \underbrace{-\int_0^T \log\tilde{\lambda}_j(s)\mathrm{d}N_j + \int_0^T \tilde{\lambda}_j(s)\mathrm{d}s}_{\tilde{\ell}_j} \underbrace{-\int_0^T \log\tilde{\lambda}_i(s)\mathrm{d}N_i + \int_0^T \tilde{\lambda}_i(s)\mathrm{d}s}_{\tilde{\ell}_i} \tag{42}$$

where $\tilde{\lambda}_j$ is rewritten as,

$$\tilde{\lambda}_j(s) = \beta_j + \beta_w \varphi_w(s) + \alpha_{i\to j}\varphi_h(s)$$

φ_w, φ_h are mean-subtracted bases,

$$\begin{aligned}
\varphi_w(s) :=& \int W(s-t)(N_i(\mathrm{d}t) - N_i(T)/T\mathrm{d}t) \\
\varphi_h(s) :=& \int h_{i\to j}(s-t)(N_i(\mathrm{d}t) - N_i(T)/T\mathrm{d}t)
\end{aligned}$$

where $N_i(T)/T \to \mathbb{E}[N_i(\mathrm{d}t)/\mathrm{d}t]$ as $T\to\infty$. When estimating the filter $h_{i\to j}$, minimizing the total negative log-likelihood $\tilde{\ell}$ is equivalent to minimizing the negative log-likelihood $\tilde{\ell}_j$, which can be approximated using the Laplace method with coefficients H and $\boldsymbol{b}$,

$$\begin{aligned}
\tilde{\ell}_j(\boldsymbol{\beta}) - \tilde{\ell}_j(\boldsymbol{\beta}_{\mathrm{MLE}}) =& \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_{\mathrm{MLE}})^T \frac{\partial^2\tilde{\ell}_j}{\partial\beta_{\mathrm{MLE}}\partial\beta_{\mathrm{MLE}}^T}(\boldsymbol{\beta}-\boldsymbol{\beta}_{\mathrm{MLE}}) + \left(\frac{\partial\tilde{\ell}_j}{\partial\beta_{\mathrm{MLE}}}\right)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_{\mathrm{MLE}}) + o(\|\boldsymbol{\beta}-\boldsymbol{\beta}_{\mathrm{MLE}}\|^2) \\
=& \boldsymbol{\beta}^T H\boldsymbol{\beta} + \boldsymbol{b}^T\boldsymbol{\beta} + \mathrm{const}
\end{aligned} \tag{43}$$

H is the Hessian matrix can be obtained from second-order derivative, analytical form of $\boldsymbol{b}$ needs further approximation.

$$\frac{\partial\tilde{\ell}_j}{\partial\boldsymbol{\beta}} \approx H\boldsymbol{\beta} + \boldsymbol{b}, \quad H = \frac{\partial^2\tilde{\ell}_j}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \tag{44}$$

The MLE thus can be expressed as $\boldsymbol{\beta}_{\mathrm{MLE}} \approx -H^{-1}\boldsymbol{b}$.

Define the following shorthands,

$$S_{ww} := \langle\varphi_w,\varphi_w\rangle, \quad S_{hh} := \langle\varphi_h,\varphi_h\rangle, \quad S_{hw} = S_{wh} := \langle\varphi_w,\varphi_h\rangle S_{w\lambda} := \langle\varphi_w,\lambda_i\rangle, \quad S_{h\lambda} := \langle\varphi_h,\lambda_i\rangle, \tag{45}$$

$\langle\cdot,\cdot\rangle$ denotes the inner product between two functions on interval $[0,T]$. Lemma E.10 will show the analytical forms of these inner products in the special case with f_i being the linear Cox process as in (27).

$$
\begin{aligned}
H =& \frac{\partial^2 \tilde{\ell}_j}{\partial \beta \partial \beta^T} = \int_0^T \frac{\Psi(s)\Psi(s)^T}{\tilde{\lambda}_j(s)^2} N_j(\mathrm{d}s) \approx \mathbb{E}_{N_j}\left[\int_0^T \frac{\Psi(s)\Psi(s)^T}{\tilde{\lambda}_j(s)^2} N_j(\mathrm{d}s) \middle| N_i \right] \\
=& \int_0^T \frac{\Psi(s)\Psi(s)^T}{\tilde{\lambda}_j(s)^2} \lambda_j(s)\mathrm{d}s \approx \frac{1}{\bar{\lambda}_j} \int_0^T \Psi(s)\Psi(s)^T \mathrm{d}s
\end{aligned}
$$

$\Psi(s) = (\mathbf{1}, \varphi_w, \varphi_h)^T$ is the vector of two bases. $1/T\langle \varphi_h, \mathbf{1} \rangle \to 0$ as $T \to \infty$. The parameter **b** in (43) can be solved using two special (suboptimal) solutions with $\hat{\boldsymbol{\beta}}^B$ at the conditions

$$
\hat{\alpha}_{i\to j}^B = 0, \quad \frac{\partial \tilde{\ell}_j}{\partial \beta_w} = 0, \quad \frac{\partial \tilde{\ell}_j}{\partial \beta_j^B} = 0
$$

and $\hat{\boldsymbol{\beta}}^C$ at conditions

$$
\hat{\beta}_w^C = 0, \quad \frac{\partial \tilde{\ell}_j}{\partial \alpha_{i\to j}^C} = 0, \quad \frac{\partial \tilde{\ell}_j}{\partial \beta_j^C} = 0
$$

Solution $\hat{\beta}_w^B$ corresponds to the model $\tilde{\lambda}_j = \beta_j + \beta_w \varphi_w$ without the impact function term at the condition $\hat{\alpha}_{i\to j}^B = 0$.

$$
\begin{aligned}
0 =& \frac{\partial \tilde{\ell}_j}{\partial \beta_w} = -\int_0^T \frac{\varphi_w(s)}{\tilde{\lambda}_j(s)} \mathrm{d}N_j(s) + \int_0^T \varphi_w(s)\mathrm{d}s \\
=& -\int_0^T \varphi_w(s) \frac{1}{\bar{\lambda}_j + \hat{\beta}_w^B \varphi_w(s)} \mathrm{d}N_j(s) = -\frac{1}{\bar{\lambda}_j} \int_0^T \varphi_w(s) \frac{1}{1 + \frac{\hat{\beta}_w^B}{\bar{\lambda}_j}\varphi_w(s)} \mathrm{d}N_j(s) \\
=& -\frac{1}{\bar{\lambda}_j} \int_0^T \varphi_w(s) \left(1 - \frac{\hat{\beta}_w^B}{\bar{\lambda}_j} \varphi_w(s) \right) \mathrm{d}N_j(s) + o\left(\frac{\hat{\beta}_w^B}{\bar{\lambda}_j^2} \int_0^T \varphi_w(s)\varphi_w(s)\mathrm{d}N_j(s) \right) \\
\approx& \mathbb{E}\left[-\frac{1}{\bar{\lambda}_j} \int_0^T \varphi_w(s) \left(1 - \frac{\hat{\beta}_w^B}{\bar{\lambda}_j}\varphi_w(s) \right) \mathrm{d}N_j(s) \middle| N_i \right] \\
=& -\frac{1}{\bar{\lambda}_j} \int_0^T \varphi_w(s) \left(1 - \frac{\hat{\beta}_w^B}{\bar{\lambda}_j}\varphi_w(s) \right) \lambda_j(s)\mathrm{d}s
\end{aligned}
$$

Then we can derive the $\hat{\beta}_w^B$,

$$
\hat{\beta}_w^B \approx \bar{\lambda}_j \frac{\langle \varphi_w, \lambda_j \rangle}{\langle \varphi_w^2, \lambda_j \rangle} \approx \bar{\lambda}_j \frac{\langle \varphi_w, \lambda_j \rangle}{\langle \varphi_w^2, \bar{\lambda}_j \rangle} = \frac{\langle \varphi_w, \lambda_j \rangle}{\langle \varphi_w, \varphi_w \rangle}
$$

Similarly, we have

$$
\hat{\alpha}_{i\to j}^C \approx \bar{\lambda}_j \frac{\langle \varphi_h, \lambda_j \rangle}{\langle \varphi_h^2, \lambda_j \rangle} \approx \frac{\langle \varphi_h, \lambda_j \rangle}{\langle \varphi_h, \varphi_h \rangle}
$$

The MLE then is,

$$
\hat{\beta} \approx -H^{-1}\boldsymbol{b} \tag{46}
$$

$$
H \approx \frac{1}{\bar{\lambda}_j} \begin{pmatrix} S_{ww} & S_{wh} & 0 \\ S_{hw} & S_{hh} & 0 \\ 0 & 0 & \hat{\beta}_j^2 T \end{pmatrix}, \quad \boldsymbol{b} \approx -\frac{1}{\bar{\lambda}_j} \begin{pmatrix} \langle \varphi_w, \lambda_j \rangle \\ \langle \varphi_h, \lambda_j \rangle \\ T\bar{\lambda}_j^3 \end{pmatrix}
$$

where $\bar{\lambda}_j = \mathbb{E}\lambda_j = \bar{\lambda}_i + \alpha_{i\to j}\sigma_h$. So we have the estimator $\hat{\alpha}_{i\to j}$,

$$
\begin{aligned}
\hat{\alpha}_{i\to j} \approx& \frac{S_{ww} \cdot \langle \varphi_h, \lambda_j \rangle - S_{hw} \cdot \langle \varphi_w, \lambda_j \rangle}{S_{ww}S_{hh} - S_{wh}^2} \\
=& \frac{S_{ww} \cdot \langle \varphi_h, \alpha_j + f_i + \alpha_{i\to j}\varphi_h \rangle - S_{hw} \cdot \langle \varphi_w, \alpha_j + f_i + \alpha_{i\to j}\varphi_h \rangle}{S_{ww}S_{hh} - S_{wh}^2}
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{S_{ww}\cdot\langle\varphi_h,f_i\rangle-S_{hw}\cdot\langle\varphi_w,f_i\rangle}{S_{ww}S_{hh}-S_{wh}^2}+\alpha_{i\to j}\cdot\frac{S_{ww}\cdot\langle\varphi_h,\varphi_h\rangle-S_{hw}\cdot\langle\varphi_w,\varphi_h\rangle}{S_{ww}S_{hh}-S_{wh}^2}\\
=&\frac{S_{ww}\cdot\langle\varphi_h,\alpha_i+f_i\rangle-S_{hw}\cdot\langle\varphi_w,\alpha_i+f_i\rangle}{S_{ww}S_{hh}-S_{wh}^2}+\alpha_{i\to j}\\
\approx&\alpha_{i\to j}+\frac{S_{ww}\langle\varphi_h,\lambda_i\rangle-S_{hw}\langle\varphi_w,\lambda_i\rangle}{S_{ww}S_{hh}-S_{hw}^2}
\end{aligned}
$$

So the bias of the estimator is approximately,

$$
\operatorname{bias}(\hat{\alpha}_{i\to j})\approx\frac{S_{ww}S_{h\lambda}-S_{hw}S_{w\lambda}}{S_{ww}S_{hh}-S_{hw}^2}\tag{47}
$$

Lemma E.10 shows the derivation of the inner products $S_{ww}, S_{hh}, S_{hw}, S_{w\lambda}, S_{h\lambda}$, which lead to the equations for the linear Cox background in (41). □

Corollary E.3. *If the regression model* (8) *does not include the nuisance variable, which becomes a typical Hawkes process, then the bias of the estimator is the following with similar derivation.*

$$
\operatorname{bias}(\hat{\alpha}_{i\to j})\approx\frac{S_{h\lambda}}{S_{hh}}\approx\frac{\frac{\rho}{2}\operatorname{erf}\left(\frac{\sigma_h}{2\sigma_I}\right)}{\rho\left[\sigma_h\operatorname{erf}\left(\frac{\sigma_h}{2\sigma_I}\right)-\frac{2\sigma_I}{\sqrt{\pi}}\left(1-e^{-\frac{\sigma_h^2}{4\sigma_I^2}}\right)\right]+\bar{\lambda}_i\sigma_h}\tag{48}
$$

Corollary E.4. *When the smoothing kernel becomes infinitely narrow, the bias in* (41) *satisfies*

$$
\lim_{\sigma_w\to 0}\operatorname{bias}(\hat{\alpha}_{i\to j})\to\frac{\frac{\rho}{2}\operatorname{erf}\left(\frac{\sigma_h}{2\sigma_I}\right)}{\rho\left[\sigma_h\operatorname{erf}\left(\frac{\sigma_h}{2\sigma_I}\right)-\frac{2\sigma_I}{\sqrt{\pi}}\left(1-e^{-\frac{\sigma_h^2}{4\sigma_I^2}}\right)\right]+\bar{\lambda}_i\sigma_h}\tag{49}
$$

Corollary E.5. *When the smoothing kernel becomes infinitely wide, the bias in* (41) *satisfies*

$$
\lim_{\sigma_w\to \infty}\operatorname{bias}(\hat{\alpha}_{i\to j})\to\frac{\frac{\rho}{2}\operatorname{erf}\left(\frac{\sigma_h}{2\sigma_I}\right)}{\rho\left[\sigma_h\operatorname{erf}\left(\frac{\sigma_h}{2\sigma_I}\right)-\frac{2\sigma_I}{\sqrt{\pi}}\left(1-e^{-\frac{\sigma_h^2}{4\sigma_I^2}}\right)\right]+\bar{\lambda}_i\sigma_h}\tag{50}
$$

The approximation is similar to Lemma E.2. Note that the three corollaries have the same results.

Lemma E.6. *Same as the settings in Lemma E.2, the variance of the estimator is approximated as,*

$$
\operatorname{Var}(\hat{\alpha}_{i\to j})\approx\frac{\bar{\lambda}_j}{T}\frac{\langle W,W\rangle_{\check{c}_N}}{\langle W,W\rangle_{\check{c}_N}\langle h,h^-\rangle_{\check{c}_N}-\langle W,\mathbf{1}\rangle_{\check{c}_\Lambda}^2}\tag{51}
$$

If f_i follows the cluster process in (27)*, then* $\operatorname{Var}(\hat{\alpha}_{i\to j})\approx\text{Numerator}/\text{Denominator}$

$$
\begin{aligned}
\text{Numerator}=&\frac{\bar{\lambda}_j}{T}\left(\frac{\rho}{2\sqrt{\pi}\sqrt{\sigma_w^2+\sigma_I^2}}+\frac{\bar{\lambda}_i}{2\sqrt{\pi}\sigma_w}\right)\\
\text{Denominator}=&\left(\frac{\rho}{2\sqrt{\pi}\sqrt{\sigma_w^2+\sigma_I^2}}+\frac{\bar{\lambda}_i}{2\sqrt{\pi}\sigma_w}\right)\cdot\\
&\left(\rho\left[\sigma_h\operatorname{erf}\left(\frac{\sigma_h}{2\sigma_I}\right)-\frac{2\sigma_I}{\sqrt{\pi}}\left(1-e^{-\frac{\sigma_h^2}{4\sigma_I^2}}\right)\right]+\bar{\lambda}_i\sigma_h\right)\\
&-\left(\frac{\rho}{2}\operatorname{erf}\left(\frac{\sigma_h}{2\sqrt{\sigma_w^2/2+\sigma_I^2}}\right)+\frac{\bar{\lambda}_i}{2}\operatorname{erf}\left(\frac{\sigma_h}{\sqrt{2}\sigma_w}\right)\right)^2
\end{aligned}\tag{52}
$$

The proof is similar to Lemma E.2 using the Fisher information.

Corollary E.7. *If the regression model* (8) *does not include the nuisance variable, which becomes a typical Hawkes process, then the variance of the estimator is*

$$\mathrm{Var}(\hat{\alpha}_{i\to j}) \approx \frac{\bar{\lambda}_j}{S_{hh}} \approx \frac{\bar{\lambda}_j}{T}\left(\rho\left[\sigma_h \mathrm{erf}\left(\frac{\sigma_h}{2\sigma_I}\right) - \frac{2\sigma_I}{\sqrt{\pi}}\left(1 - e^{-\frac{\sigma_h^2}{4\sigma_I^2}}\right)\right] + \bar{\lambda}_i\sigma_h\right)^{-1}$$

The proof is similar to Lemma E.2. The three corollaries above have the same results.

Corollary E.8. *If* $\sigma_w \to 0$ *of the variance in* (52) *will converge*

$$\lim_{\sigma_w\to 0}\mathrm{Var}(\hat{\alpha}_{i\to j}) \to \frac{\bar{\lambda}_j}{T}\left(\rho\left[\sigma_h \mathrm{erf}\left(\frac{\sigma_h}{2\sigma_I}\right) - \frac{2\sigma_I}{\sqrt{\pi}}\left(1 - e^{-\frac{\sigma_h^2}{4\sigma_I^2}}\right)\right] + \bar{\lambda}_i\sigma_h\right)^{-1} \tag{53}$$

Corollary E.9. *If* $\sigma_w \to \infty$ *of the variance in* (52) *will converge*

$$\lim_{\sigma_w\to \infty}\mathrm{Var}(\hat{\alpha}_{i\to j}) \to \frac{\bar{\lambda}_j}{T}\left(\rho\left[\sigma_h \mathrm{erf}\left(\frac{\sigma_h}{2\sigma_I}\right) - \frac{2\sigma_I}{\sqrt{\pi}}\left(1 - e^{-\frac{\sigma_h^2}{4\sigma_I^2}}\right)\right] + \bar{\lambda}_i\sigma_h\right)^{-1} \tag{54}$$

Lemma E.10. *If the linear Cox model in* (27) *and* (28)*, the inner products defined in* (45) *can be derived in analytical forms as follows.*

Proof. Apply E.1, E.11, and E.12,

$$\begin{aligned}
\frac{1}{T}S_{ww} \approx & \int_{\mathbb{R}} [W * W](s)\check{c}_N(\mathrm{d}s)\mathrm{d}s \\
= & \int_{\mathbb{R}} \frac{1}{2\sigma_w\sqrt{\pi}}\exp\left\{-\frac{s^2}{4\sigma_w^2}\right\}\left(\frac{\rho}{2\sigma_I\sqrt{\pi}}\exp\left\{-\frac{s^2}{4\sigma_I^2}\right\} + \bar{\lambda}_i\delta(s)\right)\mathrm{d}s \\
= & \frac{\rho}{2\sqrt{\pi}\sqrt{\sigma_w^2+\sigma_I^2}} + \frac{\bar{\lambda}_i}{2\sqrt{\pi}\sigma_w} \\
\frac{1}{T}S_{hh} \approx & \int_{\mathbb{R}} [h * h^-](s)\check{c}_N(\mathrm{d}s)\mathrm{d}s \\
= & \int_{\mathbb{R}} \left[\mathrm{rect}\left(\frac{u}{\sigma_h} - \frac{1}{2}\right) * \mathrm{rect}\left(-\frac{u}{\sigma_h} - \frac{1}{2}\right)\right](s)\left(\frac{\rho}{2\sigma_I\sqrt{\pi}}\exp\left\{-\frac{s^2}{4\sigma_I^2}\right\} + \bar{\lambda}_i\delta(s)\right)\mathrm{d}s \\
= & \rho\left[\sigma_h\mathrm{erf}\left(\frac{\sigma_h}{2\sigma_I}\right) - \frac{2\sigma_I}{\sqrt{\pi}}\left(1 - \exp\left\{\frac{\sigma_h^2}{4\sigma_I^2}\right\}\right)\right] + \bar{\lambda}_i\sigma_h \\
\frac{1}{T}S_{hw} \approx & \int_{\mathbb{R}} [h * W](s)\check{c}_N(\mathrm{d}s)\mathrm{d}s \\
= & \int_{\mathbb{R}} \left[\mathrm{rect}\left(\frac{u}{\sigma_h} - \frac{1}{2}\right) * \phi_{\sigma_W}(u)\right](s)\left(\frac{\rho}{2\sigma_I\sqrt{\pi}}\exp\left\{-\frac{s^2}{4\sigma_I^2}\right\} + \bar{\lambda}_i\delta(s)\right)\mathrm{d}s \\
= & \frac{\rho}{2}\mathrm{erf}\left(\frac{\sigma_h}{\sqrt{2\sigma_w^2+4\sigma_I^2}}\right) + \frac{\bar{\lambda}_i}{2}\mathrm{erf}\left(\frac{\sigma_h}{\sqrt{2}\sigma_w}\right) \\
\frac{1}{T}S_{w\lambda} \approx & \int_{\mathbb{R}} W(s)\check{c}_\Lambda(s)\mathrm{s} \\
= & \int_{\mathbb{R}} \frac{1}{\sigma_w\sqrt{2\pi}}\exp\left\{-\frac{s^2}{2\sigma_w^2}\right\}\left(\frac{\rho}{2\sigma_I\sqrt{\pi}}\exp\left\{-\frac{s^2}{4\sigma_I^2}\right\}\right)\mathrm{d}s \\
= & \frac{\rho}{\sqrt{2\sigma_w^2+4\sigma_I^2}\cdot\sqrt{\pi}} \\
\frac{1}{T}S_{h\lambda} \approx & \int_{\mathbb{R}} h(s)\check{c}_\Lambda(s)\mathrm{s}
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathbb{R}} \mathbb{I}_{[0,\sigma_h]}(s) \left(\frac{\rho}{2\sigma_I\sqrt{\pi}} \exp\left\{ -\frac{s^2}{4\sigma_I^2} \right\} \right) \mathrm{d}s \\
&= \frac{\rho}{2} \mathrm{erf}\left(\frac{\sigma_h}{2\sigma_I} \right)
\end{aligned}
$$

□

Lemma E.11. *Assume the point process $N_i(\cdot)$ is second-order stationary (definition E.1) with reduced second-order moment measure intensity $\breve{c}_N$, intensity function λ_i, and mean intensity $\bar{\lambda}_i$. Define two mean subtracted processes,*

$$
\begin{aligned}
\varphi_1(s) &:= \int W_1(s-t)(N_i(\mathrm{d}t) - \bar{\lambda}_i \mathrm{d}t) \\
\varphi_2(s) &:= \int W_2(s-t)(N_i(\mathrm{d}t) - \bar{\lambda}_i \mathrm{d}t)
\end{aligned}
$$

Then the inner product on interval $[0,T]$ between the processes is

$$
\langle \varphi_1, \varphi_2 \rangle \approx T \int_{\mathbb{R}} [W_1 * W_2^-](r) \breve{c}_N(r) \mathrm{d}r \tag{55}
$$

$W_2^-(x) := W_2(-x)$.

$$
\langle \varphi_w, \lambda_i \rangle \approx \int_{\mathbb{R}} W(r) \breve{c}_\Lambda(r) \mathrm{d}s\mathrm{d}t \tag{56}
$$

$\breve{c}_\Lambda$ and $\breve{c}_N$ are reduced second-order moment measure intensity corresponding to $\Lambda_i(\cdot)$ and $N_i(\cdot)$, see Lemma E.12. $\Lambda_i(A) := \int_A \lambda_i(t)\mathrm{d}t$ is the intensity measure.

Proof.

$$
\begin{aligned}
\langle \varphi_1, \varphi_2 \rangle &= \int_0^T \int_0^T \int_0^T \underbrace{W_1(t-u)}_{s:=t-u} \underbrace{W_2(t-v)}_{W_2^-(x):=W_2(-x)} \left(N_i(\mathrm{d}u) - \bar{\lambda}_i \mathrm{d}u\right) \left(N_i(\mathrm{d}v) - \bar{\lambda}_i \mathrm{d}v\right) \mathrm{d}t \\
&= \int_0^T \int_0^T \int_{-u}^{T-u} W_1(s) \underbrace{W_2^-((v-u)-s)}_{u:=u, r:v-u} \mathrm{d}s \left(N_i(\mathrm{d}u) - \bar{\lambda}_i \mathrm{d}u\right) \left(N_i(\mathrm{d}v) - \bar{\lambda}_i \mathrm{d}v\right) \\
&= \int_0^T \int_{-u}^{T-u} \int_{-u}^{T-u} W_1(s) W_2^-(r-s) \mathrm{d}s \cdot \breve{c}_N(r) \mathrm{d}r \cdot \mathrm{d}u \\
&\approx \int_0^T \int_{-u}^{T-u} [W_1 * W_2^-](r) \breve{c}_N(r) \mathrm{d}r \cdot \mathrm{d}u \approx T \int_{\mathbb{R}} [W_1 * W_2^-](r) \breve{c}_N(r) \mathrm{d}r
\end{aligned}
$$

The approximation error comes from the boundary effect. If the kernels W_1, W_2 decays fast, the error can be ignored.

$$
\begin{aligned}
\langle \varphi_w, \lambda_i \rangle &= \int_0^T \int_0^T W(t-u) \left(N_i(\mathrm{d}u)) - \bar{\lambda}_i \mathrm{d}u\right) \lambda_i(t) \mathrm{d}t \\
&= \int_0^T \int_0^T W(t-u) \left(N_i(\mathrm{d}u)) - \bar{\lambda}_i \mathrm{d}u\right) \left(\lambda_i(t)\mathrm{d}t - \bar{\lambda}_i \mathrm{d}t\right) \\
&\approx T \int_{\mathbb{R}} W(r) \breve{c}_\Lambda(r) \mathrm{d}s\mathrm{d}t
\end{aligned}
$$

□

Remark Many works that study the second-order stationary point process is in the frequency-domain Bartlett [1963], Brémaud et al. [2005], Brillinger [1972, 1974], Hawkes [1971], Lewis [1970], Mugglestone and Renshaw [1996] and [Daley and Vere-Jones, 2003, ch. 8]. All of our analysis is in time-domain. If we apply the Parseval's theorem to (55), it equivalently shifts almost all results into frequency-domain.

$$
\int_{\mathbb{R}} [W_1 * W_2^-](r) \breve{c}_N(r) \mathrm{d}r = \int_{\mathbb{R}} \widehat{W_1}(f) \cdot \widehat{W_2^-}(f) \cdot \Gamma_N(\mathrm{d}f)
$$

where $\widehat{W}_1, \widehat{W}_2^-$ are the spectrum of kernels, and Γ_N is called *Bartlett spectrum* for point process or *Bochner spectrum* for wide-sense process (see [Daley and Vere-Jones, 2003, ch. 8] and [Brémaud et al., 2005]). This can shift the time-domain analysis into the frequency-domain. This work does not include any frequency properties of the estimator, but it is promising to interpret some steps using the Bartlett spectrum measure in the future work.

Lemma E.12. *Consider the cluster process in* (27). *Let $\phi(\cdot)_{\sigma_I}$ be a window function with scale σ_I, t_i^c be the points of the center process which is generated by homogeneous Poisson process with intensity ρ. α_i is the baseline. The intensity function has form,*

$$\lambda_i(t) = \alpha_i + \sum_i \phi_{\sigma_I}\left(t - t_i^c\right) \tag{57}$$

$N_i(\cdot)$ is the corresponding count measure. Assume ϕ_{σ_I} is a Normal window with mean zero and standard deviation σ_I. The reduced covariance measure intensity of $\Lambda_i(t)$ is,

$$\breve{c}_\Lambda(u) = \rho \cdot [\phi_{\sigma_I} * \phi_{\sigma_I}](u) = \frac{\rho}{\sqrt{4\pi\sigma_I^2}} \exp\left\{-\frac{u^2}{4\sigma_I^2}\right\} \tag{58}$$

Similarly, the reduced covariance measure intensity the point process $N_i(t)$ is,

$$\breve{c}_N(u) = \rho \cdot [\phi_{\sigma_I} * \phi_{\sigma_I}](u) + \bar{\lambda}_i\delta(u) = \frac{\rho}{\sqrt{4\pi\sigma_I^2}} \exp\left\{-\frac{u^2}{4\sigma_I^2}\right\} + \bar{\lambda}_i\delta(u) \tag{59}$$

Proof. The first-moment property of the intensity is the following.

$$\bar{\lambda}_i = \mathbb{E}[\lambda(t)] = \mathbb{E}\left[\int_0^\infty \phi_{\sigma_I}\left(t-s\right) N(\mathrm{d}s)\right] = \int \phi_{\sigma_I}\left(t-s\right)(\alpha_i + \rho)\mathrm{d}s = \alpha_i + \rho$$

The *reduced covariance* for the second-moment stationary process is defined as,

$$\breve{c}_\Lambda(u) = \mathbb{E}[\lambda_i(x)\lambda_i(x+u)] - \mathbb{E}[\lambda_i(x)]\mathbb{E}[\lambda_i(x+u)] = \mathbb{E}[\lambda_i(x)\lambda_i(x+u)] - \bar{\lambda}_i^2$$

The second-moment measure of homogeneous Poisson process is [Hawkes, 1971],

$$\breve{M}_{N,2}^c(\mathrm{d}v) = \bar{\lambda}_i\delta(v)\mathrm{d}v + \bar{\lambda}_i^2\mathrm{d}v$$

The second equation holds due to the Campbell lemma [Kutoyants, 1998, Lemma 1.1]. $N^c(\cdot)$ is the count measure of the center process. $\Lambda_i(\cdot) := \int_A \lambda_i(t)\mathrm{d}t$ is the intensity measure with respect to the intensity λ_i.

$$\begin{aligned}
&\breve{m}_{\Lambda,2}(u) = \mathbb{E}\left[\frac{\Lambda_i(\mathrm{d}x)\Lambda_i(x+\mathrm{d}u)}{\mathrm{d}x\mathrm{d}u}\right] = \mathbb{E}[\lambda_i(x)\lambda_i(x+u)] \\
=&\mathbb{E}[(\alpha_i + f_i(x))(\alpha_i + f_i(x+u))] = \mathbb{E}[f_i(x)f_i(x+u)] + 2\rho\alpha_i + \alpha_i^2 \\
=&\mathbb{E}\left[\left(\int \phi_{\sigma_I}\left(x-s\right)\mathrm{d}N^c(s)\right)\left(\int \phi_{\sigma_I}\left(x+u-r\right)\mathrm{d}N^c(r)\right)\right] + 2\rho\alpha_i + \alpha_i^2 \\
=&\mathbb{E}\left[\iint \phi_{\sigma_I}\left(x-s\right)\phi_{\sigma_I}\left(x+u-r\right)\mathrm{d}N^c(s)\mathrm{d}N^c(r)\right] + 2\rho\alpha_i + \alpha_i^2 \\
=&\int \mathrm{d}s \int \phi_{\sigma_I}\left(x-s\right)\phi_{\sigma_I}\left(x+u-(s+v)\right)\breve{M}_{N,2}^c(\mathrm{d}v) + 2\rho\alpha_i + \alpha_i^2 \\
=&\bar{\lambda}_i^2 + \rho\int \phi_{\sigma_I}\left(s\right)\phi_{\sigma_I}\left(u-s\right)\mathrm{d}s = \bar{\lambda}_i^2 + \rho[\phi_{\sigma_I} * \phi_{\sigma_I}](u)
\end{aligned}$$

The reduced covariance measure intensity of the count measure can be derived as follows.

$$\begin{aligned}
&M_{N,2}(\mathrm{d}t \times (t+\mathrm{d}u)) = \mathrm{d}t \cdot \breve{M}_N(\mathrm{d}u) \\
=&\mathbb{E}\left[N_i(\mathrm{d}t)N_i(t+\mathrm{d}u)\right] = \mathbb{E}_\Lambda\left[\mathbb{E}_N\left[N(\mathrm{d}t)N(t+\mathrm{d}u)|\Lambda_i\right]\right] \\
=&\bar{\lambda}_i\delta(u)\mathrm{d}u + \mathbb{E}_\lambda\left[\Lambda_i(\mathrm{d}t)\Lambda_i(t+\mathrm{d}u)\right] = \bar{\lambda}_i\delta(u)\mathrm{d}u\mathrm{d}t + \breve{m}_\Lambda(u)\mathrm{d}u\mathrm{d}t
\end{aligned}$$

So we have,

$$\begin{aligned}\breve{m}_N(u) &= \bar{\lambda}_i \delta(u) + \breve{m}_\Lambda(u) \\ \breve{c}_N(u) &= \bar{\lambda}_i \delta(u) + \breve{c}_\Lambda(u)\end{aligned}$$

Similarly, the reduced second-order covariance intensity is,

$$\begin{aligned}&\mathbb{E}\left[N(\mathrm{d}t)\Lambda(t+\mathrm{d}u)\right] = \mathbb{E}_\Lambda\left[\mathbb{E}_N\left[N(\mathrm{d}t)\Lambda(t+\mathrm{d}u)|\lambda\right]\right] \\ =&\mathbb{E}_\Lambda\left[\Lambda(\mathrm{d}t)\Lambda(t+\mathrm{d}u)\right] = \breve{m}_\Lambda(u)\mathrm{d}u\mathrm{d}t\end{aligned}$$

□

F APPLICATION TO NEUROSCIENCE DATASET

F.1 MATERIALS

We applied our method to the Allen Brain Observatory Visual Coding Neuropixels Siegle et al. [2021]. It used multiple high-density extracellular electrophysiology probes to simultaneously record spiking activity from many areas in the mouse brain, especially the visual cortex. The animals were passively presented with visual stimuli while the head was fixed. The details of the experimental setup can be found in Siegle et al. [2021]. Our work used drifting gratings as the trials are long, and visual stimuli strongly elicit neural responses. The drifting gratings have 8 different orientations (0°, 45°, 90°, 135°,180°, 225°, 270°, 315°, clockwise from 0° = right-to-left). The temporal frequency is 8Hz. The spatial frequency is 0.04 cycles/deg and the contrast is 80% for all trials. The dataset assigns unique identities for all properties, such as conditions, trials, neurons, etc. In this paper, we refer to those identities directly. We analyzed mouse session `798911424`. The stimulus condition identities are: `246`, `254`, `256`, `263`, `267`, `276`, `281`, `284`. Each condition has 15 repeated trials, 120 trials in total. A trial lasts for 3 sec with a 2-second stimulus and a 1-second blank screen. 5 brain areas were recorded by separate Neuropixels probes simultaneously. The number of recorded neurons in each visual cortical area is roughly 100. We selected the top 50% most active neurons (thresholded by the mean firing rate) including 47 V1 neuron, 39 LM neurons, 23 RL neurons, 44 AL neurons, 39 AM neurons.

F.2 GOODNESS-OF-FIT

The goodness-of-fit test was assessed with the Kolmogorov-Smirnov (KS) test based on the time-rescaling theorem Bowsher [2007], Brown et al. [2002], Haslinger et al. [2010]. The theorem states that the transformed inter-spike intervals follow the unit exponential distribution. The KS test is used to compare the empirical distribution and the target distribution. A good fit should have a straight curve along the diagonal in the Q-Q plot. Figure 15 shows the results between all pairs of brain areas. The model does a good job with all curves staying along the diagonal without large deviations.

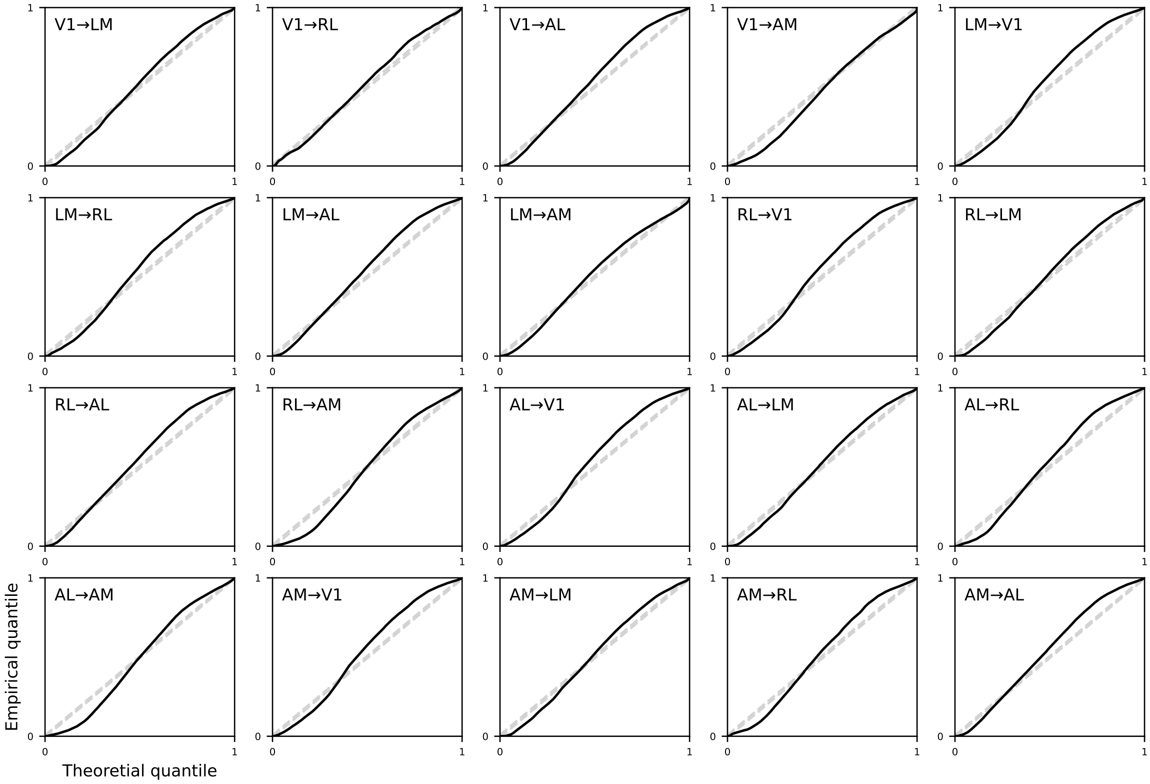

Figure 15: Goodness-of-fit tests between all pairs of brain areas. Each plot shows the results of an impact function of a pair of neurons. The connection direction is labeled at the corner. The grey dashed lines are 99% CI. A good fit should have a straight curve along the diagonal.

Next, we show examples by comparing the fitted impact function and conditional inference-based CCG as another way of verification. The CCG together with non-parametric fitting can be used to explore the timescale of the coupling effect. Figure 16 shows an example of excitatory, inhibitory, or neural coupling effects. Similar to Figure 3 and 7, the jitter-based CCG method may not be sensitive enough to detect the weak signals although it shows some clue of the coupling effect. When the coupling effects are fitted using square windows in Figure 16 second row, it will show a more significant excitatory or inhibitory coupling effect. We also estimate those effects using non-parametric fitting as shown in Figure 16 last row. Our method allows us to aggregate all the information in a lag window to estimate the impact function with one parameter using a square window, or a few parameters using B-splines. The results will be more effective and significant than the method using the pointwise statistic. We admit modeling using square windows loses some details of the coupling effect, but it is effective in capturing the general properties of the coupling effect with a limited dataset and large noise.

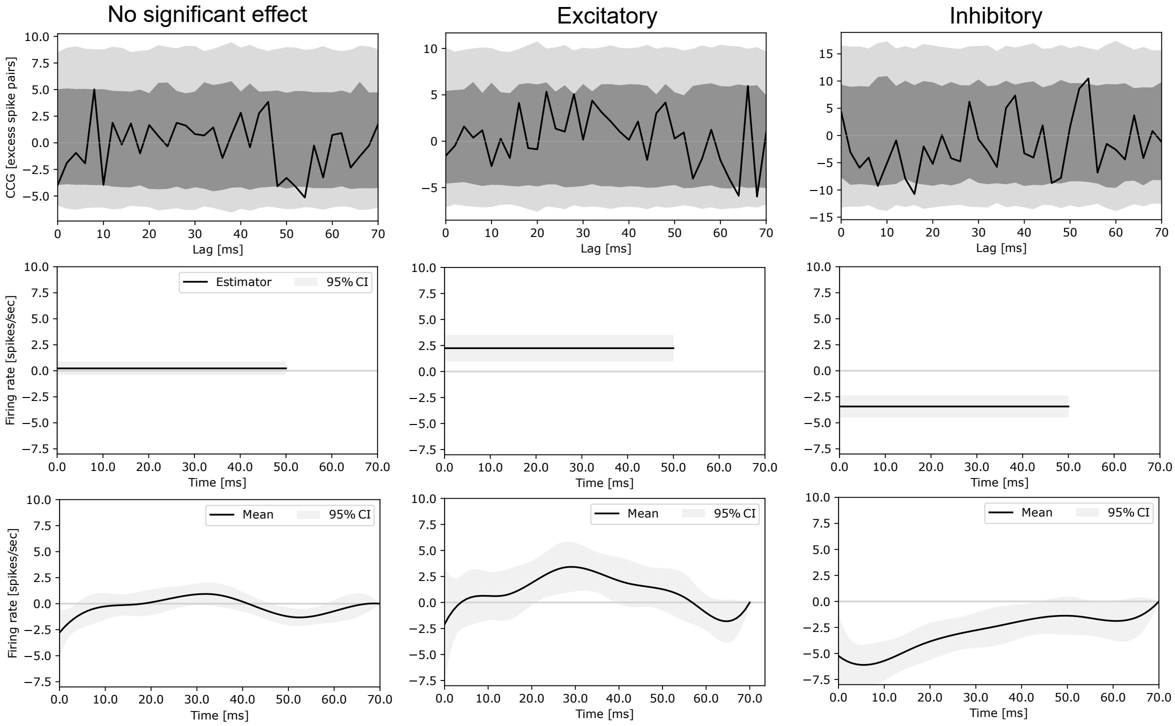

Figure 16: Comparison between impact functions and jitter-based CCG. The calculation of the jitter-based CCG in the first row is the same as Figure 3. The second row shows the fitted impact functions using square windows. The third row shows the fitted impact functions using the non-parametric method.

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.

Maurice S Bartlett. Statistical estimation of density functions. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 245–254, 1963.

Maurice S Bartlett. The spectral analysis of two-dimensional point processes. *Biometrika*, 51(3/4):299–311, 1964.

M Bhatti and P Bracken. The calculation of integrals involving b-splines by means of recursion relations. *Applied mathematics and computation*, 172(1):91–100, 2006.

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

Pierre Brémaud, Laurent Massoulié, and Andrea Ridolfi. Power spectra of random spike fields and related processes. *Advances in applied probability*, 37(4):1116–1146, 2005.

David R Brillinger. The spectral analysis of stationary interval functions. In *Vol. 1 Theory of Statistics*, pages 483–514. University of California Press, 1972.

David R Brillinger. Cross-spectral analysis of processes with stationary increments including the stationary $G/G/\infty$ queue. *The Annals of Probability*, pages 815–827, 1974.

Emery N Brown, Riccardo Barbieri, Valérie Ventura, Robert E Kass, and Loren M Frank. The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation*, 14(2):325–346, 2002.

Shizhe Chen, Daniela Witten, and Ali Shojaie. Nearly assumptionless screening for the mutually-exciting multivariate hawkes process. *Electronic journal of statistics*, 11(1):1207, 2017.

Ann Cowling, Peter Hall, and Michael J Phillips. Bootstrap confidence regions for the intensity of a poisson point process. *Journal of the American Statistical Association*, 91(436):1516–1524, 1996.

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

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.

Robert Haslinger, Gordon Pipa, and Emery Brown. Discrete time rescaling theorem: determining goodness of fit for discrete time statistical models of neural spiking. *Neural computation*, 22(10):2477–2506, 2010.

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

Yu A Kutoyants. *Statistical inference for spatial Poisson processes*, volume 134. Springer Science & Business Media, 1998.

PAW Lewis. Remarks on the theory, computation and application of the spectral analysis of series of events. *Journal of Sound and Vibration*, 12(3):353–375, 1970.

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.

Moira A Mugglestone and Eric Renshaw. A practical guide to the spectral analysis of spatial point processes. *Computational Statistics & Data Analysis*, 21(1):43–65, 1996.

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

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

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.