A Mixing of two Hawkes processes

We first present a useful lemma, which provides the proof for full model case (i.e. under H_1). Another equivalent definition of conditional intensity $\lambda(t|\mathcal{H}_t)$ for a counting process $\{N(t) : t \geq 0\}$ with history $\mathcal{H}_t$ $(t \geq 0)$ is

$$\mathbb{P}(N(t+h) - N(t) = m|\mathcal{H}_t) = \begin{cases} \lambda(t|\mathcal{H}_t)h + o(h), & m = 1 \\ o(h), & m > 1 \\ 1 - \lambda(t|\mathcal{H}_t)h + o(h), & m = 0 \end{cases}$$

We will make use of this definition to prove the following lemma.

Lemma A.1. *Suppose we have n Hawkes processes $\{N_z(t) : t \geq 0\}$ $(z = 1, 2, \ldots, n)$ with conditional intensity specified by (1). Define the mixing to be $N(t) = \sum_{z=1}^{n} N_z(t)$. The conditional intensity of mixing of n Hawkes processes is sum of those n conditional intensities. That is,*

$$\lambda(t|\mathcal{H}_t) = \sum_{z=1}^{n} \lambda_z(t|\mathcal{H}_{z,t}),$$

where $\mathcal{H}_t = \cup_{z=1}^{n} \mathcal{H}_{z,t}$.

Proof of Lemma A.1. We prove by the definition of conditional intensity. For any non-negative integer $m \in \mathbb{Z}_+$, denote $\underbrace{\mathbf{m}} = (m_1, \ldots, m_n)$ and $M = \{\underbrace{\mathbf{m}} \mid m_1 + \cdots + m_n = m,\ m_i \in \mathbb{Z}_+\}$,

$$\mathbb{P}(N(t+h) - N(t) = m|\mathcal{H}_t) = \sum_{\underbrace{\mathbf{m}} \in M} \prod_{i=1}^{n} \mathbb{P}(N_i(t+h) - N_i(t) = m_i|\mathcal{H}_{i,t}).$$

Case 1: When $m > 1$, it is easy to see $\mathbb{P}(N(t+h) - N(t) = m|\mathcal{H}_t) = o(h)$, since either there are at least two m_i's ≥ 1 or at least one $m_i \geq 2$.

Case 2: When $m = 1$, there will be one and only one of all m_i's taking value 1 and the rest will be all zeros. Thus, we have

$$\begin{aligned} &\mathbb{P}(N(t+h) - N(t) = 1|\mathcal{H}_t) \\ =& \sum_{j=1}^{n} \mathbb{P}(N_j(t+h) - N_j(t) = 1|\mathcal{H}_{j,t}) \prod_{i \neq j}^{n} \mathbb{P}(N_i(t+h) - N_i(t) = 0|\mathcal{H}_{i,t}) \\ =& \sum_{j=1}^{n} (\lambda_j(t|\mathcal{H}_{j,t})h + \mathrm{o}(h)) \prod_{i \neq j}^{n} (1 - \lambda_i(t|\mathcal{H}_{i,t})h + o(h)) = \sum_{j=1}^{n} \lambda_j(t|\mathcal{H}_{j,t})h + \mathrm{o}(h). \end{aligned}$$

Case 3: When $m = 0$, all m_i's will be zeros and we will have

$$\begin{aligned} \mathbb{P}(N(t+h) - N(t) = 0|\mathcal{H}_t) &= \prod_{i=1}^{n} \mathbb{P}(N_i(t+h) - N_i(t) = 0|\mathcal{H}_{i,t}) \\ &= \prod_{j=1}^{n} (1 - \lambda_i(t|\mathcal{H}_{i,t})h + o(h)) = 1 - \sum_{i=1}^{n} \lambda_j(t|\mathcal{H}_{j,t})h + \mathrm{o}(h). \end{aligned}$$

Let $\lambda(t|\mathcal{H}_t) = \sum_{i=1}^{n} \lambda_i(t|\mathcal{H}_{i,t})$, and we will find out this is the conditional intensity for $N(t)$. □

Proof of Proposition 1. We can see under the alternative hypothesis, the result directly follows Lemma A.1. Under null hypothesis, by Lemma A.1, it is easy to show $N(t)$ defined in Proposition 1 has intensity

$$\lambda(t|\mathcal{H}_t) = \mu^{(1)} + \mu^{(2)} + \int_0^t \phi(t-u)\mathrm{d}(N_1(u) + N_2(u)) = \mu + \int_0^t \phi(t-u)\mathrm{d}N(u),$$

where $\mathcal{H}_t = \mathcal{H}_{1,t} \cup \mathcal{H}_{2,t}$.

By the definition of Hawkes Process in Section 2, we can see the mixing of two Hawkes processes under H_0 is still a Hawkes process. We complete the proof. □

B A non-parametric estimation of the Quasi-conditional intensity

B.1 Probability Weighted Histogram Estimation under null hypothesis

Here, we redefine the Quasi-parameter as $\theta = (\mu, \alpha^{(1)}, g_1^{(1)}, \ldots, g_{n_0}^{(1)}, \alpha^{(2)}, g_1^{(2)}, \ldots, g_{n_0}^{(2)})$, where $\mu \triangleq \mu^{(1)} + \mu^{(2)}$. This is because we will estimate the triggering magnitude and the temporal triggering function separately.

The full model Quasi-parameter space is given by

$$\Theta = \Big\{\theta \mid \mu > 0 \text{ and } \int_0^\infty \phi^{(z)}(u)du = \int_0^\infty \alpha^{(z)} g^{(z)}(u)du = \alpha^{(z)} < 1 \;\; (z = 1, 2)\Big\}.$$

Under H_0, we have

$$\theta_0 \in \Theta_0 = \{\theta \in \Theta \mid \alpha^{(1)} = \alpha^{(2)} = \alpha \quad \text{and} \quad g_k^{(1)} = g_k^{(2)} = g_k \;\; (k = 1, \ldots, n_0)\}.$$

Denote $\mathcal{H}_t = \mathcal{H}_{1,t} \cup \mathcal{H}_{2,t} = \{t_1, \ldots, t_N\}$. Define the branching structure as follows:

$$p_{ij} = \begin{cases} \text{probability event } i \text{ is triggered by event } j, & i > j \\ \text{probability event } i \text{ comes from background}, & i = j \\ 0, & i < j \end{cases}$$

Apparently, we want to estimate the Quasi-background intensity from background process only and Quasi-triggering function from the triggered events only. Instead of using a hard-threshold indicator, Zhuang et al. (2002) used a stochastic declustering procedure to separate the background events from triggered ones by assigning each event a weight, or rather the probability that this event comes from background or is direct offspring from an individual ancestor. Then, we can use a probability weighted estimator to estimate Quasi-background intensity and Quasi-triggering function. The algorithm is as follows:

Assume we have estimated branching structure $p_{ii}^{(v)}$ at iteration v, then we can estimate the Quasi-background intensity as follows:

$$\mu^{(v)} = \frac{1}{T} \sum_{i=1}^{N} p_{ii}^{(v)}. \tag{4}$$

For the Quasi-triggering component, as we assume g to be a p.d.f., we can estimate the magnitude of triggering effect from triggered events only:

$$\alpha^{(v)} = 1 - \sum_{i=1}^{N} p_{ii}^{(v)}/N. \tag{5}$$

For the temporal component in the Quasi-triggering function, for each bin (as we discretize in (2)), we estimate its parameter from those triggered events which falls into that bin, i.e.

$$g_k^{(v)} = \frac{\sum_{B_k} p_{ij}^{(v)}}{\Delta t_k \sum_{i=1}^{N} \sum_{j=1}^{i-1} p_{ij}^{(v)}}, \;\; (k = 1, \ldots, n_0). \tag{6}$$

After estimating the Quasi-conditional intensity function, we update the branching structure. More specifically, for $i > j$:

$$p_{ij}^{(v+1)} = \mathbb{P}(\text{i-th event is triggered by j-th event}|\mathcal{H}_{t_i}) = \frac{\alpha^{(v)} g^{(v)}(t_i - t_j)}{\mu^{(v)} + \sum_{j=1}^{i-1} \alpha^{(v)} g^{(v)}(t_i - t_j)}, \tag{7}$$

$$p_{ii}^{(v+1)} = \mathbb{P}(\text{i-th event comes from background}|\mathcal{H}_{t_i}) = \frac{\mu^{(v)}}{\mu^{(v)} + \sum_{j=1}^{i-1} \alpha^{(v)} g^{(v)}(t_i - t_j)}. \tag{8}$$

We summarize the algorithm as follows:

Algorithm 2 Probability Weighted Histogram Estimation of Quasi-log-likelihood under H_0

Initialize: choose stopping critical value ϵ (e.g. 10^{-3}), initialize $p_{ij}^{(0)}$ and set $p_{ij}^{(-1)} = \epsilon + p_{ij}^{(0)}$ and iteration index $v = 0$.
while $\max_{i>j} \left| p_{ij}^{(v)} - p_{ij}^{(v-1)} \right| < \varepsilon$ **do**

1. Estimate Quasi-background rate μ as in (4).
2. Estimate Quasi-triggering components magnitude α and temporal $g(t)$ as in (5) and (6).
3. Update probabilities $p_{ij}^{(v+1)}$'s as in (7) and (8).
4. $v \leftarrow v + 1$

end while

B.2 EM-type algorithm derivation

In Fox et al. (2016), they assumed the ground-truth takes piecewise constant form (2) and demonstrated that algorithm 2 is an EM-type algorithm under (2) by using complete data log-likelihood. However, they did not explicitly show the E-step also maximizes the complete data log-likelihood (or rather complete data Quasi-log-likelihood in our setting). We will first lower bound the Quasi-log-likelihood and then show that the algorithm iterates between maximizing this lower bound w.r.t. branching structure (p_{ij}'s) and w.r.t. the Quasi-conditional intensity (Quasi-background rate μ, Quasi-triggering magnitude α and temporal Quasi-triggering function g).

First recall the Quasi-log-likelihood function under H_0:

$$\ell_0(\theta) = -\mu T + \sum_{i=1}^{N} \log \left(\mu + \sum_{i>j} \phi\left(t_i - t_j\right) \right) - \sum_{j=1}^{N} \int_{t_j}^{T} \phi\left(t - t_j\right) dt$$

We can simplify the last term above by using integral approximation of Schoenberg (2013):

$$\sum_{j=1}^{N} \int_{t_j}^{T} \phi\left(t_i - t_j\right) dt = \sum_{j=1}^{N} \int_{t_j}^{T} \alpha g\left(t - t_j\right) dt \approx \sum_{j=1}^{N} \alpha \int_{t_j}^{\infty} g\left(t - t_j\right) dt = \alpha N$$

Thus we can ignore the last term when maximizing the log-likelihood function. Next, we lower bound the first term in the Quasi-log-likelihood function by Jensen's inequality:

$$\begin{aligned} \sum_{i=1}^{N} \log \left(\mu + \sum_{i>j} \phi\left(t_i - t_j\right) \right) &= \sum_{i=1}^{N} \log \left(p_{ii} \frac{\mu}{p_{ii}} + \sum_{i>j} p_{ij} \frac{\phi\left(t_i - t_j\right)}{p_{ij}} \right) \\ &\geq \sum_{i=1}^{N} p_{ii} \log \left(\frac{\mu}{p_{ii}} \right) + \sum_{i>j} p_{ij} \log \left(\frac{\phi\left(t_i - t_j\right)}{p_{ij}} \right), \end{aligned}$$

where p_{ij}'s satisfy $\sum_{i \geq j} p_{ij} = 1$. Then we can get a lower bound on the approximation of Quasi-log-likelihood under the piecewise constant parameterization:

$$-\alpha N - T\mu + \sum_{i=1}^{N} \left[p_{ii} \log\left(\mu\right) + \sum_{i>j} p_{ij} \left(\log \alpha + \log \left(\sum_{k=1}^{n_0} g_k \mathbf{1}_{B_k}\left(t_i - t_j\right) \right) \right) - \sum_{i \geq j} p_{ij} \log(p_{ij}) \right]$$

Denote this lower bound by $\widetilde{\ell}(\theta)$. We maximize this lower bound under the following constraints:

$$\begin{aligned} \sum_{k=1}^{n_0} g_k \Delta t_k = 1, \quad & (g(t) \text{ is a p.d.f.}) \\ \sum_{i \geq j} p_{ij} = 1, \quad & (p_{ij}\text{'s are probability weights}) \end{aligned}$$

By adding Lagrange multipliers, this is equivalent to maximizing the following objective:

$$\widetilde{L} = \sum_{i=1}^{N}\left[p_{ii}\log(\mu) + \sum_{i>j} p_{ij}\left(\log\alpha + \log\left(\sum_{k=1}^{n_0} g_k \mathbf{1}_{B_k}(t_i - t_j)\right)\right) - \sum_{i\geq j} p_{ij}\log(p_{ij})\right]$$
$$- \alpha N - T\mu - c_1\left(\sum_{k=1}^{n_0} g_k \Delta t_k - 1\right) - \sum_{i=1}^{N} c_2^{(i)}\left(\sum_{i\geq j} p_{ij} - 1\right).$$

M-step: By taking first order derivative w.r.t. μ and setting it to zero, we will have:

$$\frac{\partial \widetilde{L}}{\partial \mu} = \sum_{i=1}^{N}\left(\frac{p_{ii}}{\mu}\right) - T = 0.$$

Solving for μ and we will get

$$\mu = \frac{\sum_{i=1}^{N} p_{ii}}{T},$$

which is the same as the update in step 1 in Algorithm 2. This means when we have $p_{ij}^{(v)}$'s at iteration v, the update in step 1 in Algorithm 2 leads to a larger Quasi-log-likelihood value. Similarly taking derivative w.r.t. α and setting it to zero leads to the update in step 2: $\alpha^{(v)} = 1 - \sum_{i=1}^{N} p_{ii}^{(v)}/N$.

$$\frac{\partial \widetilde{L}}{\partial g_k} = \sum_{i=1}^{N}\sum_{i>j}\left(\frac{p_{ij}\mathbf{1}_{B_k}(t_i - t_j)}{g_k}\right) - c_1\Delta t_k = 0$$
$$\frac{\partial \widetilde{L}}{\partial c_1} = 1 - \sum_{k=1}^{n_0} g_k \Delta t_k = 0$$

We can solve for g_k and c_1 by some simple algebra and then get the update for g_k at iteration v (given $p_{ij}^{(v)}$'s) :

$$g_k^{(v)} = \frac{\sum_{i=1}^{N}\sum_{i>j} p_{ij}^{(v)}\mathbf{1}_{B_k}(t_i - t_j)}{\Delta t_k \sum_{j=1}^{N}\sum_{i>j} p_{ij}^{(v)}}.$$

E-step: As for p_{ij}'s, denote

$$\log\phi_{ij} = \log\alpha + \log\left(\sum_{k=1}^{n_0} g_k \mathbf{1}_{B_k}(t_i - t_j)\right).$$

Repeat the similar procedure, we will get:

$$\frac{\partial \widetilde{L}}{\partial p_{ij}} = -\log(p_{ij}) - 1 - c_2^{(i)} + \log\phi_{ij} = 0$$
$$\frac{\partial \widetilde{L}}{\partial p_{ii}} = -\log(p_{ii}) - 1 - c_2^{(i)} + \log\mu = 0$$
$$\frac{\partial \widetilde{L}}{\partial c_2^{(i)}} = \sum_{i\geq j} p_{ij} - 1 = 0$$

By the first two equations we have

$$\frac{p_{ii}}{p_{ij}} = \frac{\mu}{\phi_{ij}}.$$

Plug this back into the last equation and we will get the update in step 3 in Algorithm 2. Thus, we validate Algorithm 2 as an EM-type algorithm.

C Explicit form of GS statistic

Note that $\phi^{(z)}(u) = \sum_{k=1}^{n_0} \phi_k^{(z)} \mathbf{1}_{B_k}(u)$. To simplify the explicit expressions, we first define the following notations:

$$G(i, z'; z) = \sum_{j=1}^{N_{z'}} \phi^{(z)}(t_i^{(z')} - t_j^{(z)}),$$
$$G'_k(i, z'; z) = \sum_{j=1}^{N_{z'}} \mathbf{1}_{B_k}(t_i^{(z')} - t_j^{(z)})$$
$$\Delta_{z,i} = \mu + G(i, z; z) + G(i, z; z')$$

Here, $G(i, z'; z)$ represents the triggering effect of events in process z to $i-$th event in process z'. $G'_k(i, z'; z)$ is the partial derivative of $G(i, z'; z)$ w.r.t. $\phi_k^{(z)}$.

Note that $\phi^{(z)}(\cdot)$ and $\mathbf{1}_{B_k}(\cdot)$ $(k = 1, 2, \ldots, n_0)$ take value zero on $(-\infty, 0]$. Thus we have

$$\sum_{j<i} \phi^{(z)}(t_i^{(z)} - t_j^{(z)}) = \sum_{j=1}^{N_z} \phi^{(z)}(t_i^{(z)} - t_j^{(z)}),$$

which can be denoted by $G(i, z; z)$ we just defined. By our notations, the Quasi-log-likelihood takes the following form:

$$\ell_1(\mu, \phi^{(1)}, \phi^{(2)}|\mathcal{H}_t) = -\mu\, T + \sum_{z=1}^{2} \sum_{i=1}^{N_z} \log \Delta_{z,i} - \int_0^{T - t_i^{(z)}} \phi^{(z)}(u) du,$$

where $(\mu, \phi^{(1)}, \phi^{(2)}) = (\mu, \phi_1^{(1)}, \ldots, \phi_{n_0}^{(1)}, \phi_1^{(2)}, \ldots, \phi_{n_0}^{(2)})$. Those parameters are denoted by θ to simplify the notations. To get the explicit form of GS statistic, we only need to calculate the first two order partial derivative of $\ell_1(\mu, \phi^{(1)}, \phi^{(2)}|\mathcal{H}_t)$ w.r.t. θ.

First order partial derivatives:

$$\frac{\partial \ell_1(\mu, \phi^{(1)}, \phi^{(2)}|\mathcal{H}_t)}{\partial \mu} = \sum_{z=1}^{2} \sum_{i=1}^{N_z} \frac{1}{\Delta_{z,i}} - T,$$
$$\frac{\partial \ell_1(\mu, \phi^{(1)}, \phi^{(2)}|\mathcal{H}_t)}{\partial \phi_k^{(z)}} = \sum_{i=1}^{N_z} \frac{G'_k(i, z; z)}{\Delta_{z,i}} + \sum_{i=1}^{N_{z'}} \frac{G'_k(i, z'; z)}{\Delta_{z',i}} - \sum_{i=1}^{N_z} \int_0^{T - t_i^{(z)}} \mathbf{1}_{B_k}(u) du.$$

Here, we get the explicit expression for $S_T(\theta)$ and $A_T(\theta)$.

Second order partial derivatives:

$$\frac{\partial^2 \ell_1(\mu, \phi^{(1)}, \phi^{(2)}|\mathcal{H}_t)}{\partial \mu^2} = -\sum_{z=1}^{2} \sum_{i=1}^{N_z} \frac{1}{\Delta_{z,i}^2}$$
$$\frac{\partial^2 \ell_1(\mu, \phi^{(1)}, \phi^{(2)}|\mathcal{H}_t)}{\partial (\phi_k^{(z)})^2} = -\sum_{i=1}^{N_z} \left(\frac{G'_k(i, z; z)}{\Delta_{z,i}} \right)^2 - \sum_{i=1}^{N_{z'}} \left(\frac{G'_k(i, z'; z)}{\Delta_{z',i}} \right)^2$$

$$\frac{\partial^2 \ell_1(\mu, \phi^{(1)}, \phi^{(2)}|\mathcal{H}_t)}{\partial \mu \partial \phi_k^{(z)}} = -\sum_{i=1}^{N_z} \frac{G'_k(i, z; z)}{\Delta_{z,i}^2} + \sum_{i=1}^{N_{z'}} \frac{G'_k(i, z'; z)}{\Delta_{z',i}^2}$$
$$\frac{\partial^2 \ell_1(\mu, \phi^{(1)}, \phi^{(2)}|\mathcal{H}_t)}{\partial \phi_k^{(z)} \partial \phi_l^{(z')}} = -\sum_{i=1}^{N_z} \frac{G'_l(i, z; z') G'_k(i, z; z)}{\Delta_{z,i}^2} + \sum_{i=1}^{N_{z'}} \frac{G'_l(i, z'; z') G'_k(i, z'; z)}{\Delta_{z',i}^2}$$
$$\frac{\partial^2 \ell_1(\mu, \phi^{(1)}, \phi^{(2)}|\mathcal{H}_t)}{\partial \phi_k^{(z)} \partial \phi_l^{(z)}} = -\sum_{i=1}^{N_z} \frac{G'_k(i, z; z) G'_l(i, z; z)}{\Delta_{z,i}^2} + \sum_{i=1}^{N_{z'}} \frac{G'_k(i, z'; z) G'_l(i, z'; z)}{\Delta_{z',i}^2}$$

D Asymptotic properties of QMLE and GS test

D.1 Identifiability of the estimand and justification of our testing framework

Proof of Identifiability. One can verify that for each specific sample trajectory $\mathcal{H}_T$, $\ell_1(\theta|\mathcal{H}_T)$ is composed of two parts: a linear function of θ plus several logarithm of a linear function of θ. This means that $\ell_1(\theta|\mathcal{H}_T)$ is concave in θ. We will use a very simple example to elaborate on this.

Suppose we only have 3 events $\mathcal{H}_T = \{t_1^{(1)}, t_1^{(2)}, t_2^{(1)}\}$, where $T = t_2^{(1)}$. Then

$$\begin{aligned}\ell_1(\theta|\mathcal{H}_T) = &-\mu T - \sum_{k=1}^{n_0} \phi_k^{(1)} \int_0^{T-t_1^{(1)}} \mathbf{1}_{B_k}(u)du - \sum_{k=1}^{n_0} \phi_k^{(2)} \int_0^{T-t_1^{(2)}} \mathbf{1}_{B_k}(u)du \\ &+ \log\mu + \log(\mu + \phi_{k_1}^{(1)}) + \log(\mu + \phi_{k_2}^{(1)} + \phi_{k_3}^{(2)}),\end{aligned}$$

where k_1, k_2, k_3 are the indices of the bins which $t_1^{(2)} - t_1^{(1)}, t_2^{(1)} - t_1^{(1)}, t_2^{(1)} - t_1^{(2)}$ fall into, respectively. Note that $\int_0^t \mathbf{1}_{B_k}(u)du$ is the length of the intersection of $[0, t]$ and B_k, which is a constant. Thus, it is easy to see from the example that $\ell_1(\theta|\mathcal{H}_T)$ is concave in θ for any fixed trajectory.

Next, we can show that $\ell_1(\theta|\mathcal{H}_T)$ will remain the same for sample trajectories that are "close" to each other. In the simple example above, as long as $t_1^{(2)} - t_1^{(1)}, t_2^{(1)} - t_1^{(1)}, t_2^{(1)} - t_1^{(2)}$ remain in bins $B_{k_1}, B_{k_2}, B_{k_3}$, the value of the corresponding $\ell_1(\theta|\mathcal{H}_T)$ will not change. For fixed number of events N, we call all trajectories with N events that corresponds to the same Quasi-log-likelihood value a case. It is easy to see the number of all cases for fixed number of events N is countable. Then the expectation taken w.r.t. all possible trajectories will reduce to a countablely infinite summation. That is

$$\mathbb{E}[\ell_1(\theta|\mathcal{H}_T)] = \sum_i \ell_{1,i}(\theta) p_i,$$

where p_i is the probability of all sample trajectories such that $\ell_1(\theta|\mathcal{H}_T) \equiv \ell_{1,i}(\theta)$. Note that we just show $\ell_{1,i}(\theta)$ is concave in θ. Thus the objective is a linear combination of concave functions. This means θ_0 actually solves a concave program. It is a unique maximizer of the expected Quasi-log-likelihood, i.e. globally identifiable. We have a well-defined estimand here. □

Justification of our testing framework. By adopting the view in Akaike (1998), in (3) we are actually trying to find a $\theta_0 \in \Theta$ whose corresponding Quasi-likelihood has a minimum K-L divergence with the unknown ground-truth λ^*.

As is suggested in Akaike (1998), we can view this as a statistical decision problem where the loss function is $\log \lambda^*/\lambda_\theta$. For the simple example above, the loss function can be expressed by

$$\begin{aligned}&(\mu - \mu^*)T + \sum_{k=1}^{n_0} \int_0^{T-t_1^{(1)}} (\phi_k^{(1)} - \phi^{*(1)}(u))\mathbf{1}_{B_k}(u)du + \sum_{k=1}^{n_0} \int_0^{T-t_1^{(2)}} (\phi_k^{(2)} - \phi^{*(2)}(u))\mathbf{1}_{B_k}(u)du \\ &+ \log(\mu^*/\mu) + \log\frac{\mu^* + \phi^{*(1)}(t_1^{(2)} - t_1^{(1)})}{\mu + \phi_{k_1}^{(1)}} + \log\frac{\mu^* + \phi^{*(1)}(t_2^{(1)} - t_1^{(1)}) + \phi^{*(2)}(t_2^{(1)} - t_1^{(2)})}{\mu + \phi_{k_2}^{(1)} + \phi_{k_3}^{(2)}},\end{aligned}$$

Taking all possible sample trajectories into account, when $\phi^{*(1)} = \phi^{*(2)}$, apparently we will achieve minimum risk when $\theta_0 \in \Theta_0$. □

D.2 Proof of Lemma 1: consistency and asymptotic normality of QMLE

Proof. We will provide a generalization of the asymptotic properties MLEs under correct model specification for temporal Hawkes process in Ogata (1978) to model misspecification (or model mismatch) case.

We first show that the assumptions in Ogata (1978) hold for our Quasi-conditional intensity function.

(A) Since under our parameterization (2), we have $\int_0^\infty \alpha g(t)dt = \alpha < 1$, our point process model is stationary and ergodic. It is easy to check assumptions (A1) $\sim$ (A3).

(B) The Quasi-conditional intensity function we consider here is actually linear w.r.t. the parameters, then it is arbitrarily order continuous differentiable (i.e. smooth) and bounded within any compact set in the Quasi-parameter space. Assumptions (B1) $\sim$ (B7) hold trivially.

(C) By (2), the Quasi-temporal triggering function is truncated on $[0, T_0]$, which means and complete data conditional intensity function $\lambda(t|\mathcal{H}_{-\infty,t})$ will be exactly the same as $\lambda(t|\mathcal{H}_{0,t})$ as long as $t > T_0$. Since Assumptions (C1) $\sim$ (C4) only require stochastic approximations of $\lambda(t|\mathcal{H}_{0,t})$ to $\lambda(t|\mathcal{H}_{-\infty,t})$ when t goes to infinity, it is easy to see those assumptions are satisfied.

Next, since our parametric form (2) is only approximation to the true one, we need to slightly modify the theoretical results in Ogata (1978) for our QMLE. Here we will not mention theorems or lemmas that we do not need to modify under model mismatch (except that we should keep in mind that the "true" parameter in Ogata (1978) is understood as the maximizer of Quasi-likelihood) and it is easy to verify those theoretical results (from the beginning to Theorem 5) by just following the proof therein.

Before we proceed to the proof, we should note that the QMLE is $\widehat{\theta}_{QMLE}$ under H_0 and $\widetilde{\theta}_{QMLE}$ under H_1. Under $H_1 : \theta_0 \notin \Theta_0$, the estimator $\widetilde{\theta}_{QMLE}$ is obtained using the full model conditional intensity ℓ_1 instead of ℓ_0. The estimation is given in Algorithm 3 in Appendix E.

For simplicity, we denote $\bar{\theta}_{QMLE}$ to be $\widehat{\theta}_{QMLE}$ and $\widetilde{\theta}_{QMLE}$ under H_0 and H_1, respectively. That is,

$$\bar{\theta}_{QMLE} = \begin{cases} \widehat{\theta}_{QMLE}, & H_0 \text{ is true} \\ \widetilde{\theta}_{QMLE}, & H_1 \text{ is true} \end{cases}$$

Modifications on Theorem 1. Here θ_0 is not the true parameter of the true conditional intensity function. Instead, it is the maximizer of Quasi-log-likelihood, i.e. our approximation to the true log-likelihood function. By the definition of θ_0 and stationarity of the process, the first result still in this theorem still holds:

$$\frac{\partial \mathbb{E}\big[\ell_1(\theta|\mathcal{H}_t)\big]}{\partial \theta}\bigg|_{\theta=\theta_0} = 0.$$

However, the second result does not hold unless our approximation is indeed a correct specification of the model. More specifically, in general,

$$dN(t) = \lambda^*(t|\mathcal{H}_t)dt \neq \lambda(t|\mathcal{H}_t)dt,$$

where λ^* is the correct parametric form and typically unknown in practice.

Thus, we have

$$\mathbb{E}\bigg[\frac{\partial \ell_1(\theta|\mathcal{H}_t)}{\partial \theta_i}\frac{\partial \ell_1(\theta|\mathcal{H}_t)}{\partial \theta_j}\bigg]\bigg|_{\theta=\theta_0} \neq -\mathbb{E}\bigg[\frac{\partial^2 \ell_1(\theta|\mathcal{H}_t)}{\partial \theta_i \partial \theta_j}\bigg]\bigg|_{\theta=\theta_0}.$$

Using our notation, this can be re-expressed as $A(\theta_0) \neq B(\theta_0)$.

Modifications on Theorem 2. The convergence in our case is much stronger. By following the proof in Fox et al. (2016), the convergence in probability comes from Assumptions (C), where the convergence in the stochastic approximation is only in probability sense. However, we just show that the stochastic approximation holds for every sample path as long as $t > T_0$ based on our parameterization (2) that the Quasi-temporal triggering function is truncated, i.e. our convergence is in almost surely sense. Thus, we have:

$$\bar{\theta}_{QMLE} \overset{a.s.}{\to} \theta_0 \quad \text{as} \quad T \to \infty.$$

Modifications on Theorem 4. Since $A(\theta_0) \neq B(\theta_0)$, the convergence result should be

$$\frac{1}{\sqrt{T}}\frac{\partial \ell_1(\theta|\mathcal{H}_T)}{\partial \theta}\bigg|_{\theta=\theta_0} \overset{d}{\to} N(0, A(\theta_0)) \quad \text{as} \quad T \to \infty.$$

This is because

$$\mathbb{E}\bigg[\frac{\partial \ell_1(\theta|\mathcal{H}_1)}{\partial \theta}\frac{\partial \ell_1(\theta|\mathcal{H}_1)}{\partial \theta^\intercal}\bigg] = A(\theta_0) \neq B(\theta_0),$$

where the first equality comes from definition and stationarity of the process.

Modifications on Theorem 5. By the proof of this theorem one can reach this result:

$$\sqrt{T}(\bar{\theta}_{QMLE} - \theta_0) \xrightarrow{d} N\Big(0, B^{-1}(\theta_0)A(\theta_0)B^{-1}(\theta_0)\Big) \quad \text{as} \quad T \to \infty.$$

Again, since $A(\theta_0) \neq B(\theta_0)$, the asymptotic covariance matrix is not $B^{-1}(\theta_0)$ and that's the modification here. Besides, the asymptotic χ^2 distribution of log-likelihood ratio does not hold because of the model mismatch.

Here, we complete the proof. □

D.3 Proof of Theorem 1: asymptotic distribution under null hypothesis

This proof is highly involved. To help better understand this proof, we first provide a high level sketch on why our GS statistic follows a χ^2 distribution.

Proof Sketch. $\widehat{\theta}_{QMLE}$ solves the following problem

$$\max_{\theta \in \Theta} \ell_1(\theta|\mathcal{H}_T) \quad \text{s.t.} \quad h(\theta) = 0.$$

By adding Lagrange Multiplier ζ_T, we can derive that $\widehat{\theta}_{QMLE}$ satisfies:

$$\nabla \ell_1(\widehat{\theta}_{QMLE}|\mathcal{H}_T) + \zeta_T^\intercal \nabla h(\widehat{\theta}_{QMLE}) = S_T(\widehat{\theta}_{QMLE}) + \zeta_T^\intercal \nabla h(\widehat{\theta}_{QMLE}) = 0. \tag{9}$$

Following idea in Boos (1992), we can use Taylor expansion to expand $S_T(\theta_0)$ about $\widehat{\theta}_{QMLE}$ and $h(\widehat{\theta}_{QMLE})$ about θ_0 (note that we have $h(\theta_0) = 0$ under H_0):

$$\begin{aligned} S_T(\widehat{\theta}_{QMLE}) &= S_T(\theta_0) - B_T(\widehat{\theta}_{QMLE})(\widehat{\theta}_{QMLE} - \theta_0) + o(1), \\ 0 = h(\widehat{\theta}_{QMLE}) &= h(\theta_0) + \nabla h(\theta_0)(\widehat{\theta}_{QMLE} - \theta_0) + o(1). \end{aligned}$$

Note that by our notation $\nabla h(\theta) = H(\theta)$. Since $h(\theta)$ is linear in θ, its gradient is a constant matrix and we can denote $H = \nabla h(\theta)$.

Pre-multiply the first equation above by $H^\intercal \Big(HB_T^{-1}(\theta)H^\intercal\Big)^{-1} HB_T^{-1}(\theta)\Big|_{\theta=\widehat{\theta}_{QMLE}}$

$$\begin{aligned} &H^\intercal \Big(HB_T^{-1}(\theta)H^\intercal\Big)^{-1} HB_T^{-1}(\theta)S_T(\theta_0)\Big|_{\theta=\widehat{\theta}_{QMLE}} \\ &\quad = B_T^{\frac{1}{2}}(\theta)\left(B_T^{-\frac{1}{2}}(\theta)H^\intercal \Big(HB_T^{-1}(\theta)H^\intercal\Big)^{-1} HB_T^{-\frac{1}{2}}(\theta)\right) B_T^{-\frac{1}{2}}(\theta)S_T(\theta)\Big|_{\theta=\widehat{\theta}_{QMLE}} + o(1). \end{aligned}$$

The matrix in the middle of RHS is a projection matrix for the column space of $B_T^{-\frac{1}{2}}(\widehat{\theta}_{QMLE})H^\intercal$, and from (9) we know $B_T^{-\frac{1}{2}}(\widehat{\theta}_{QMLE})S_T(\widehat{\theta}_{QMLE})$ is already in this space. This means the RHS is exactly $S_T(\widehat{\theta}_{QMLE})$ and we will get:

$$S_T(\widehat{\theta}_{QMLE}) = H^\intercal \Big(HB_T^{-1}(\theta_0)H^\intercal\Big)^{-1} HB_T^{-1}(\theta_0)S_T(\theta_0) + o(1).$$

Rewrite GS statistic as

$$\widehat{GS}_T = \frac{1}{\sqrt{T}} S_T^\intercal(\widehat{\theta}_{QMLE}) \left(T\widehat{\Sigma}^{-1}\right) \frac{1}{\sqrt{T}} S_T(\widehat{\theta}_{QMLE}).$$

By Lemma 1, one can verify $S_T(\widehat{\theta}_{QMLE})/\sqrt{T}$ has a asymptotic normal distribution with $T\widehat{\Sigma}^{-1}$ being a consistent estimator of generalized inverse of its asymptotic covariance matrix.

Since H is of rank r, we verify that $\widehat{GS}_T \sim \chi_r^2$. □

Next, we present a more rigorous proof following the method in White (1982).

Proof. We first state some useful results:

By the almost surely convergence of QMLE (modifications of Theorems 2 and 5 in Ogata (1978)), we have that

$$\frac{1}{T}A_T(\widehat{\theta}_{QMLE}) \xrightarrow{a.s.} A(\theta_0) \quad \text{as} \quad T \to \infty$$
$$\frac{1}{T}B_T(\widehat{\theta}_{QMLE}) \xrightarrow{a.s.} B(\theta_0) \quad \text{as} \quad T \to \infty.$$

The modification of Theorem 1 in Ogata (1978) can be re-expressed as $S(\theta_0) = 0$.

The modification of Theorem 4 in Ogata (1978) can be re-expressed as follows

$$\frac{1}{\sqrt{T}}S_T(\theta_0) \xrightarrow{d} N\Big(0, A(\theta_0)\Big) \quad \text{as} \quad T \to \infty,$$

where S_T is the Quasi-score function (i.e. first order gradient of Quasi-log-likelihood function).

Under null hypothesis, the asymptotic χ^2 distribution of GS statistic under model mismatch (e.g. Theorem 3.5. in White (1982) and Section 4.2. in Boos (1992)) can be extended to temporal Hawkes process.

The QMLE actually solves the following optimization problem:

$$\max_{\theta \in \Theta_0} \ell_0(\theta|\mathcal{H}_T).$$

Since $\ell_1(\theta) = \ell_0(\theta)$ $(\forall \theta \in \Theta_0)$, equivalently it can be re-expressed as

$$\max_{\theta \in \Theta_0} \ell_1(\theta|\mathcal{H}_T),$$

or

$$\max_{\theta \in \Theta} \ell_1(\theta|\mathcal{H}_T) \quad \text{s.t.} \quad h(\theta) = 0.$$

We can reformulate this by adding Lagrange Multiplier ζ_T:

$$\max_{\theta \in \Theta} \frac{1}{T}\ell_1(\theta|\mathcal{H}_T) + \zeta_T^\intercal h(\theta).$$

Since h as well as ∇h both has full row rank r, by Lagrange Multiplier Theorem (e.g. Theorem 42.9 in Bartle (1976)), we can guarantee the existence of ζ_T, which satisfies:

$$\begin{aligned} &\frac{1}{T}\nabla \ell_1(\widehat{\theta}_{QMLE}|\mathcal{H}_T) + \Big(\nabla h(\widehat{\theta}_{QMLE})\Big)^\intercal \zeta_T = 0, \\ &h(\widehat{\theta}_{QMLE}) = 0. \end{aligned} \tag{10}$$

We denote $S_T(\theta) = \nabla \ell_1(\theta|\mathcal{H}_T)$. By the mean-value theorem for random functions (Lemma 3 in Jennrich (1969)), we have:

$$S_T(\widehat{\theta}_{QMLE}) = S_T(\theta_0) + B_T(\bar{\theta})(\widehat{\theta}_{QMLE} - \theta_0), \tag{11}$$
$$0 = h(\widehat{\theta}_{QMLE}) = h(\theta_0) + \nabla h(\widetilde{\theta})(\widehat{\theta}_{QMLE} - \theta_0), \tag{12}$$

where $\widetilde{\theta}$ and $\bar{\theta}$ lies on the segment joining $\widehat{\theta}_{QMLE}$ and θ_0. Since $\widehat{\theta}_{QMLE}$ converges to θ_0 almost surely, $\widetilde{\theta}$ and $\bar{\theta}$ both converge to θ_0 almost surely.

Under H_0: $\theta_0 \in \Theta_0$, we have $h(\theta_0) = 0$. Plug this back into the mean-value expansion (12) we will get:

$$\nabla h(\widetilde{\theta})\sqrt{T}(\widehat{\theta}_{QMLE} - \theta_0) = 0. \tag{13}$$

Multiply (10) by $\sqrt{T}$ and plug the mean-value expansion (11) into it, we will get:

$$\frac{1}{\sqrt{T}}S_T(\theta_0) + \frac{1}{T}B_T(\bar{\theta})\sqrt{T}(\widehat{\theta}_{QMLE} - \theta_0) + \sqrt{T}\Big(\nabla h(\widehat{\theta}_{QMLE})\Big)^\intercal \zeta_T = 0, \tag{14}$$

Since $B_T(\bar{\theta})/T \xrightarrow{a.s.} B(\theta_0)$, the non-singularity of $B_T(\bar{\theta})$ directly follows Assumption (B6) in Ogata (1978) for sufficiently large T. Pre-multiplying (14) by $\nabla h(\widetilde{\theta})B_T^{-1}(\bar{\theta})$ and plug (13) into it, we will get:

$$\begin{aligned}0 =&\nabla h(\widetilde{\theta})B_T^{-1}(\bar{\theta})\left(\frac{1}{\sqrt{T}}S_T(\theta_0)+\frac{1}{T}B_T(\bar{\theta})\sqrt{T}(\widehat{\theta}_{QMLE}-\theta_0)+\sqrt{T}\Big(\nabla h(\widehat{\theta}_{QMLE})\Big)^\intercal\zeta_T\right)\\ =&\nabla h(\widetilde{\theta})B_T^{-1}(\bar{\theta})\frac{1}{\sqrt{T}}S_T(\theta_0)+\nabla h(\widetilde{\theta})B_T^{-1}(\bar{\theta})\Big(\nabla h(\widehat{\theta}_{QMLE})\Big)^\intercal\sqrt{T}\zeta_T.\end{aligned}$$

Note that for our testing problem, since $h(\theta)$ is linear in θ, $\nabla h(\theta)$ does not depend on θ and has full row rank r. We denote this by H. It is easy to verify that $HB_T^{-1}(\bar{\theta})H^\intercal$ is non-singular for sufficiently large T. Thus, pre-multiply $(HB_T^{-1}(\bar{\theta})H^\intercal)^{-1}$ and rearrange the terms, we will get:

$$\sqrt{T}\zeta_T=-\Big(HB_T^{-1}(\bar{\theta})H^\intercal\Big)^{-1}HB_T^{-1}(\bar{\theta})\frac{1}{\sqrt{T}}S_T(\theta_0).$$

Note that we have shown that $S_T(\theta_0)/\sqrt{T}$ is asymptotically normally distributed with covariance matrix $A(\theta_0)$, thus we will have

$$\sqrt{T}\zeta_T\xrightarrow{d}N\Big(0,\Big(HB^{-1}(\theta_0)H^\intercal\Big)^{-1}HB^{-1}(\theta_0)A(\theta_0)B^{-1}(\theta_0)H^\intercal\Big(HB^{-1}(\theta_0)H^\intercal\Big)^{-1}\Big).\tag{15}$$

We denote this covariance matrix by $Q(\theta_0)$.

Denote

$$\sqrt{T}\widetilde{\zeta}_T(\theta)=-\Big(HB_T^{-1}(\theta)H^\intercal\Big)^{-1}HB_T^{-1}(\theta)\frac{1}{\sqrt{T}}S_T(\theta).\tag{16}$$

By 2c.4(x.a) in Rao et al. (1973), we will have

$$\sqrt{T}\zeta_T-\sqrt{T}\widetilde{\zeta}_T(\theta_0)\xrightarrow{p}0.$$

Meanwhile, by pre-multiplying (10) by $\Big(HB_T^{-1}(\widehat{\theta}_{QMLE})H^\intercal\Big)^{-1}HB_T^{-1}(\widehat{\theta}_{QMLE})$ (again the non-singularity holds for sufficiently large T), we will have

$$\sqrt{T}\zeta_T=\sqrt{T}\widetilde{\zeta}_T(\widehat{\theta}_{QMLE}).$$

Thus, by (15), we have when $T\to\infty$,

$$\sqrt{T}\widetilde{\zeta}_T(\widehat{\theta}_{QMLE})\xrightarrow{d}N\Big(0,Q(\theta_0)\Big).$$

We can easily re-write GS statistic $\widehat{GS}_T$ as a quadratic form of score function $S_T(\widehat{\theta}_{QMLE})$. By the notation we just defined in (16) we will have:

$$\widehat{GS}_T=\sqrt{T}\widetilde{\zeta}_T^\intercal(\theta)HB^{-1}(\theta)H^\intercal\left(HB^{-1}(\theta)\frac{A_T(\theta)}{T}B^{-1}(\theta_0)H^\intercal\right)^{-1}HB^{-1}(\theta)H^\intercal\sqrt{T}\widetilde{\zeta}_T(\theta)\Bigg|_{\theta=\widehat{\theta}_{QMLE}},$$

where the matrix in the middle

$$HB^{-1}(\theta)H^\intercal\left(HB^{-1}(\theta)\frac{A_T(\theta)}{T}B^{-1}(\theta_0)H^\intercal\right)^{-1}HB^{-1}(\theta)H^\intercal\Bigg|_{\theta=\widehat{\theta}_{QMLE}}$$

is a consistent estimator of $Q(\theta_0)$, since $\widehat{\theta}_{QMLE}$ converges to θ_0 almost surely.

By Lemma 3.3 in White (1980), we can verify the asymptotic χ^2 distribution of our GS statistic. □

D.4 Proof of Theorem 2: asymptotic power under alternative hypothesis

Proof. We make use of the Generalized Wald (GW) test statistic here, which is asymptotically equivalent to GS statistic under both H_0 and H_1. More specifically, by 2c.4(xiv) in Rao et al. (1973) (or Theorem 1 in 13.6 in Engle (1984)),

$$\widehat{GS}_T - \widehat{GW}_T \xrightarrow{p} 0,$$

where $\widehat{GW}_T$ is the GW test statistic. We define it as follows:

$$\widehat{GW}_T = \left. h(\theta)^\intercal \Big(H(\theta) B_T^{-1}(\theta) A_T(\theta) B_T^{-1}(\theta) H(\theta)^\intercal \Big)^{-1} h(\theta) \right|_{\theta=\widetilde{\theta}_{QMLE}}, \tag{17}$$

where $\widetilde{\theta}_{QMLE}$ is QMLE under H_1.

As we have mentioned above, $h(\theta)$ is linear in θ, thus its first order gradient is a constant matrix, i.e. $H(\theta) = H$. More specifically, $h(\theta) = H\theta$. Then it is not hard to verify the asymptotic normal distribution of $h(\widetilde{\theta}_{QMLE})$ based on asymptotically normality of $\widetilde{\theta}_{QMLE}$. That is

$$\sqrt{T}\Big(h(\widetilde{\theta}_{QMLE}) - h(\theta_0) \Big) \xrightarrow{d} N\Big(0, H B^{-1}(\theta_0) A(\theta_0) B^{-1}(\theta_0) H^\intercal \Big) \quad \text{as} \quad T \to \infty.$$

Then the noncentral χ^2 distribution of $\widehat{GW}_T$ as well as $\widehat{GS}_T$ directly follow.

Since

$$\theta_0^\intercal H^\intercal H \theta_0 = \Big(\alpha^{(1)} - \alpha^{(2)} \Big)^2 + \sum_{k=1}^{n_0} \Big(g_k^{(1)} - g_k^{(2)} \Big)^2 = \| \phi^{(1)} - \phi^{(2)} \|_2^2,$$

and H is of rank r, the noncentrality parameter is $T\| \phi^{(1)} - \phi^{(2)} \|_2^2$ and the degree of freedom is r. Thus, we get that the asymptotic power function is Marcum-Q-function. □

Another proof of consistency of GS test. We can re-express GW test statistic as:

$$\widehat{GW}_T = T \; h(\theta)^\intercal \left(H(\theta) \Big(\frac{B_T(\theta)}{T} \Big)^{-1} \frac{A_T(\theta)}{T} \Big(\frac{B_T(\theta)}{T} \Big)^{-1} H(\theta)^\intercal \right)^{-1} h(\theta) \Bigg|_{\theta=\widetilde{\theta}_{QMLE}}.$$

From Lemma 1 which we just prove, we have (i) $h(\widetilde{\theta}_{QMLE}) \to h(\theta_0) \neq 0$ almost surely, where the last inequality comes form $H_1 : \theta_0 \notin \Theta_0$; and (ii) $A_T(\widetilde{\theta}_{QMLE})/T$, $B_T(\widetilde{\theta}_{QMLE})/T$ converges to $A(\theta_0)$, $B(\theta_0)$ almost surely. Thus, we can verify

$$\widehat{GW}_T \to \infty \quad \text{as} \quad T \to \infty.$$

Thus, we have

$$\widehat{GS}_T \to \infty \quad \text{as} \quad T \to \infty,$$

which indicates the unit asymptotic power of the proposed GS test, i.e. this test is consistent. □

E Numerical experiments

E.1 Testing details

The testing procedures are detailed in Algorithm 1. We specify the data sequence sets we use, the initialization and other experiment configurations in Algorithm 1 here for all experiments above.

Validation of asymptotic properties in Section 5.1: (a) For each $\alpha \in \{1.5, 2, 2.5, 3, 3.5\}$, generate L data sequences as D_1 and another L data sequences as D_2; (b) Generate L data sequences from $\alpha = 1$ as D_1 and another L data sequences as D_2 from $\alpha = 4$; (c) Use the first pair of data sequence set in (a) (corresponding to $\alpha = 1.5$) as positive sample and data sequence set in (b) as the negative sample. For experiments in Sections 5.2 and 5.3, data generation mechanisms for D_1 and D_2 are the same.

The experiment configurations (initialization) are as follows: $L = 1,000$, $n_0 = 14$ and endpoints for those bins are $(0, .04, .08, .12, .16, .2, .26, .32, .38, .45, .55, .65, .75, 1, 2)$ for all experiments. (a) $N = 200$, $K = 20$; (b) $N \in \{50, 150, \ldots, 850\}$, $K = 5$; (c) $N \in \{25, 50, 100\}$, $K = 150$.

For the experiment on how n_0 influences our proposed test, the endpoints for bins with $n_0 = 2, 3, 4, 7, 14, 28$ are $(0, .6, 2)$, $(0, .2, .6, 2)$, $(0, .1, .2, .6, 2)$, $(0, .08, .16, .2, .32, .45, .65, 2)$, $(0, .04, .08, .12, .16, .2, .26, .32, .38, .45, .55, .65, .75, 1, 2)$ and $(0, .02, .04, .06, .08, .1, .12, .14, .16, .18, .2, .23, .26, .29, .32, .35, .38, .41, .45, .5, .55, .6, .65, .7, .75, .8, 1, 1.5, 2)$, respectively. We use $L = 1,000$ sequences in computing $\widehat{GS}_T$.

Goodness-of-fit in Section 5.4: D_1 is chosen to be the testing data and D_2 is generated from the model fitted on the training data. The endpoints of bins are

(i) $(0, .02, .04, .06, .08, .1, .12, .14, .16, .18, .2, .25, .3, .35, .4, .5)$ for `Exp` and `Matern` data;

(ii) $(0, .02, .04, .06, .08, .1, .12, .14, .16, .18, .2, .5, .6, .8, 1)$ for `MIMIC` data;

(iii) $(0, .05, .1, .15, .2, .25, .3, .35, .4, .45, .5, .6, .8, 1)$ for `MEME` data.

For 911 call data, $L = 364$ and we use the first 200 sequences to as the training data to fit the model and the rest 164 sequences as D_1. Then we generate 164 data sequences as D_2 to perform the testing procedure. We choose $N = 20$, $K = 1$ and use $(0, .02, .04, .06, .08, .1, .12, .14, .16, .18, .2, .5, 1)$ as endpoints for bins.

E.2 Additional experiments

Validation of our proposed method as an model free approach. We use different synthetic data to validate our theoretical results. Here, the triggering function used to generate synthetic data is power function (which is commonly used in seismology) : $\phi(t) = \alpha(P-1)c^{P-1}(t+c)^{-P}\mathbf{1}_{\{t>0\}}$ with parameters $\mu = 20, \alpha = 0.2, C = 2, P = 13, 14, \ldots, 17$. The experiment configurations are as follows: $L = 1,000$, $n_0 = 12$ and endpoints for those bins are $(0, .04, .08, .12, .16, .2, .24, .28, .32, .36, .4, .7, 2)$ for all experiments. (a) $N = 200$, $K = 20$; (b) $N \in \{50, 150, \ldots, 850\}$, $K = 5$; (c) $N \in \{50, 100, 200\}$, $K = 150$. See the results in Figure 8.

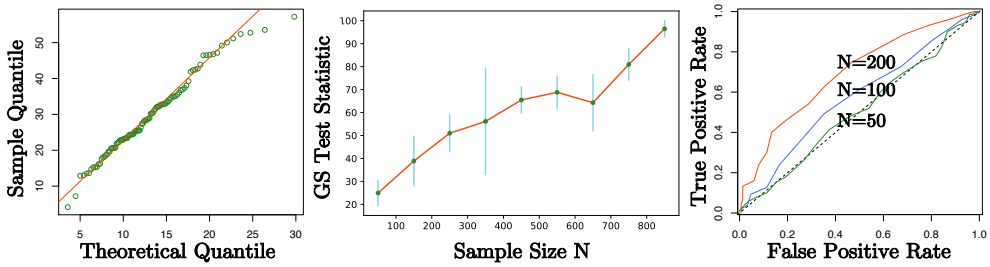

Figure 8: Simulation results: (a) Quantiles of calculated scores against theoretical quantiles of $\chi^2_{n_0+1}$ distribution under H_0; (b) mean and variance of scores with increasing N under H_1; (c) ROC curve for different N.

Comparison with Ripley's K function. See Figure 9.

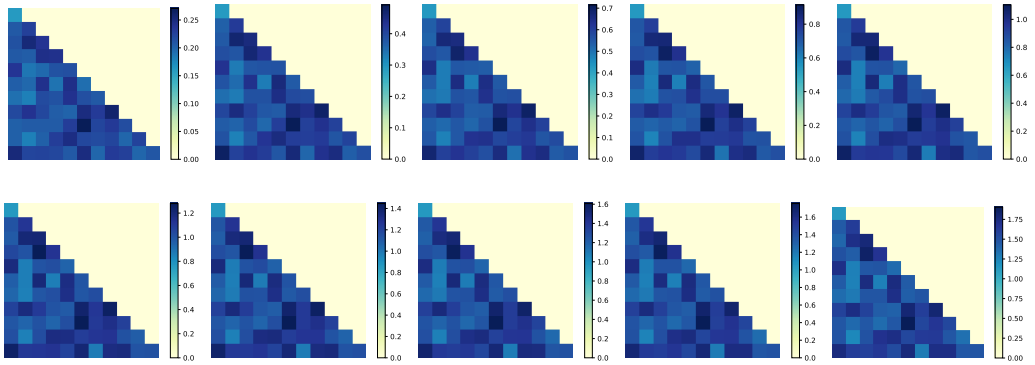

Figure 9: Heatmap of estimated Ripley's K function value $\widehat{K}(t)$ for $t = 1, \ldots, 5$ (top), $t = 6, \ldots, 10$ (bottom). For each pixel, the data sequence D_1 and D_2 are generated from the same distribution as in Figure 6 (a).

Algorithmic behavior of `Exp GD` method on `Exp` data. In our experiment, we saw a very interesting phenomenon — no matter where we initialize $\widehat{\alpha}$, using GD to maximize log-likelihood under correct model specification would yield very biased estimate.

As illustrated in Figure 10, we observe that when $\widehat{\alpha}$ is around the ground-truth 1, the log-likelihood is very large. But it keeps growing larger when $\widehat{\alpha}$ keeps decreasing. The same is also true for $\widehat{\beta}$. We can see that even though we got very large log-likelihood, the estimate is very biased. Clearly, overfitting occurs here — we only gain very little log-likelihood increment but the estimates are getting futther away from the ground-truth. Therefore, using log-likelihood as GOF would be questionable.

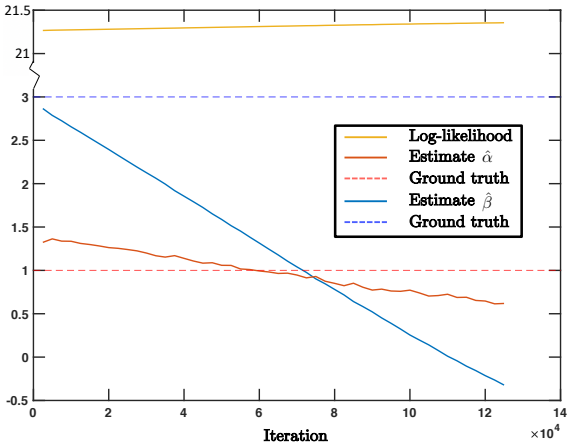

Figure 10: Algorithmic behavior illustration of `Exp GD` method on `Exp` data.

Probability weighted histogram estimation under H_1. As one may see from our proof in Appendix D, GS and GW tests are asymptotically equivalent. One would ask why we choose GS test over GW test. The reason is two-fold. Firstly, it is not computationally efficient, since using GW test involves estimating r more parameters. Secondly and most importantly, its power is far less than GS test. That's because, in empirical study, the QMLE $\widetilde{\theta}_{QMLE}$ does maximize the full model Quasi-likelihood but fails to differentiate two different triggering components, which makes $\|h(\widetilde{\theta}_{QMLE})\|_2$ much smaller than $\|h(\theta_0)\|_2$. We will further illustrate this by performing the estimation of the full model (Algorithm 3 in next section) and visualizing the estimation of triggering function as follows:

We can see a very interesting pattern: when the true triggering functions for D_1 and D_2 are different (off-diagonal panels), histogram estimation tends to yields two piecewise constant triggering function lie between those two different true ones. This means the estimated difference between two triggering functions are much smaller than the truth, or rather $\|h(\widetilde{\theta}_{QMLE})\|_2$ will be much smaller than it should be. By the form of GW test statistic (17), we can see the power of GW test statistic is highly dependent on $\|h(\widetilde{\theta}_{QMLE})\|_2$ and histogram estimation will make the resulting GW test less powerful.

F Some useful functions

F.1 Marcum-Q-function

In statistics, the Marcum-Q-function Q_M is defined as

$$Q_M(a,b) = \int_b^\infty x \left(\frac{x}{a}\right)^{M-1} \exp\left(-\frac{x^2+a^2}{2}\right) I_{M-1}(ax)dx,$$

or

$$Q_M(a,b) = \exp\left(-\frac{a^2+b^2}{2}\right) \sum_{k=1-M}^{\infty} \left(\frac{a}{b}\right)^k I_k(ab),$$

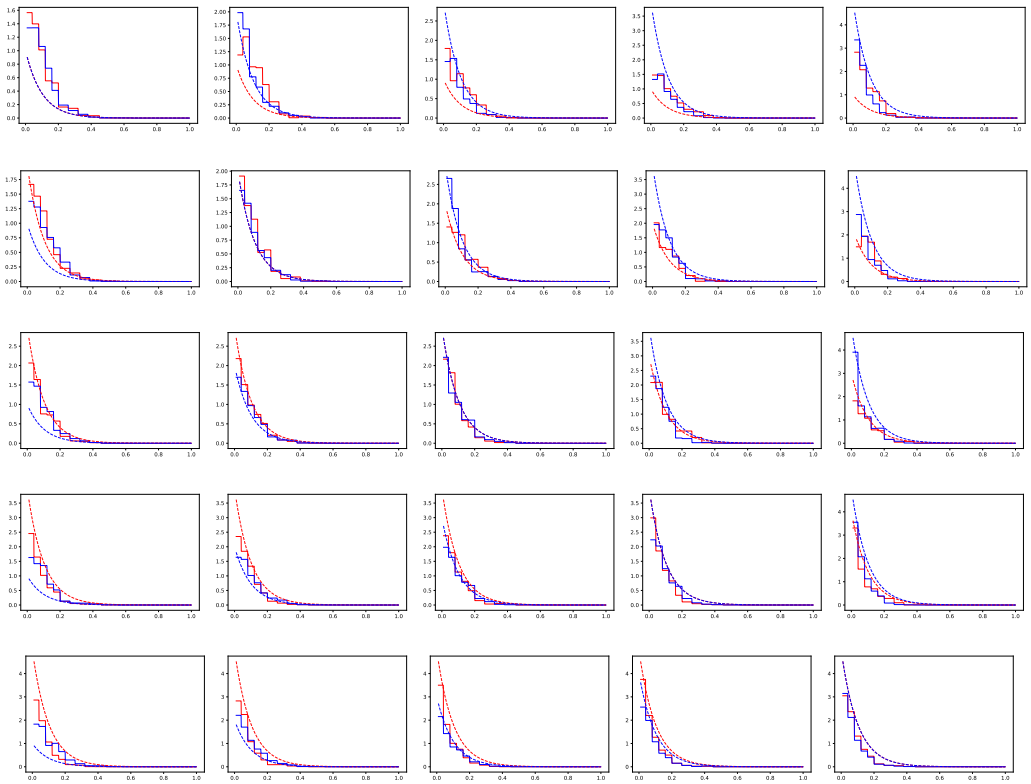

Figure 11: Histogram estimation under H_1: $\mu = 20, \alpha = 2, \beta = 1, 2, 3, 4, 5$. The solid line is the true triggering function whereas the dashed line is the estimated one (blue for D_1 and red for D_2).

with modified Bessel function $I_{M-1}(\cdot)$ of order $M-1$. Abdel-Aty (1954) proved the following approximation formula:

$$Q_{k/2}(\sqrt{\lambda},\sqrt{x}) \approx 1-\Phi\left\{\frac{\left(\frac{x}{k+\lambda}\right)^{1/3}-\left(1-\frac{2}{9f}\right)}{\sqrt{\frac{2}{9f}}}\right\},$$

where $f=\frac{(k+\lambda)^2}{k+2\lambda}=k+\frac{\lambda^2}{k+2\lambda}$ and $\Phi(\cdot)$ is CDF of standard Gaussian random variable. We can easily verify that $Q_{k/2}(\sqrt{\lambda},\sqrt{x})\to 1$ as $\lambda\to\infty$. Also, this is illustrated in Figure 3. What's more, by the Theorem 1 in Sun et al. (2010), Marcum-Q-function $Q_M(a,b)$ is monotonically increasing w.r.t. a.

F.2 Matérn covariance function

The Matérn covariance between two points separated by d distance units is defined as

$$C_{\rho,\nu}(d)=\sigma^2\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{d}{\rho}\right)^{\nu}K_\nu\left(\sqrt{2\nu}\frac{d}{\rho}\right).$$

where $\Gamma(\cdot)$ is the gamma function, $K_\nu(\cdot)$ is the modified Bessel function of the second kind, and ρ and ν are non-negative parameters of the covariance.

G Probability weighted histogram estimation under alternative hypothesis

Under H_1, the triggering mechanism is more complex compared to univariate Hawkes Process, since each event can be either from the background, direct offspring from an individual ancestor in Hawkes Process 1 or Hawkes Process 2 and the triggering effects of events in two different processes are different.

We denote the branching structure matrix $P_{(z)(z')}\in\mathbb{R}^{(N_1+N_2)\times(N_1+N_2)}$ $(z,z'\in\{1,2\})$. The element in $i-$th row and $j-$th column is defined to be the probability that event i in process z is triggered by event j in process z' if either $i\neq j$ or $z\neq z'$ (case 1) or the probability that event i in process z is a background event if $i=j$ and $z=z'$ (case 2). That is,

$$\left(P_{(z)(z')}\right)_{ij}=\begin{cases}\text{probability that event } i \text{ in process } z \text{ is triggered by event } j \text{ in process } z', & \text{case 1}\\ \text{probability that event } i \text{ in process } z \text{ is a background event,} & \text{case 2}\end{cases}$$

Note that the probability is zero when $t_i^{(z)}\le t_j^{(z')}$, which means the event that happens earlier in the process cannot be triggered by those which happen later.

As discussed above, we focus on differentiating difference in triggering effect. Thus we estimate the sum of two background intensities from all background events:

$$\mu^{(v)}=\frac{1}{T}\sum_{z=1}^{2}\sum_{i=1}^{N_z}\left(P^{(v)}_{(z)(z)}\right)_{ii}. \tag{18}$$

For the triggering components, we estimate the magnitude for process z $(z=1,2)$ using events from aggregated data triggered by process z and estimate the temporal triggering density function from those events which fall into the corresponding bin. Note that we have

$$\left(P^{(v)}_{(z')(z)}\right)_{ij}=\left(P^{(v)}_{(z')(z)}\right)_{ij}\mathbf{1}_{\{t_i^{(z')}>t_j^{(z)}\}}.$$

This is because the probability will be zero if $t_i^{(z')}\le t_j^{(z)}$ as discussed above. Thus, for $z=1,2$ and $k=1,\dots,n_0$, the estimators can be expressed as

$$\alpha_z^{(v)}=\frac{\sum_{i=1}^{N_z}\sum_{j=1}^{i-1}\left(P^{(v)}_{(z)(z)}\right)_{ij}+\sum_{i=1}^{N_{z'}}\sum_{j=1}^{N_z}\left(P^{(v)}_{(z')(z)}\right)_{ij}}{N_z}, \tag{19}$$

$$g_{z,k}^{(v)}=\frac{\sum_{i=1}^{N_z}\sum_{j=1}^{i-1}\left(P^{(v)}_{(z)(z)}\right)_{ij}\mathbf{1}_{B_k}\left(t_i^{(z)}-t_j^{(z)}\right)+\sum_{i=1}^{N_{z'}}\sum_{j=1}^{N_z}\left(P^{(v)}_{(z')(z)}\right)_{ij}\mathbf{1}_{B_k}\left(t_i^{(z')}-t_j^{(z)}\right)}{\Delta t_k\left(\sum_{i=1}^{N_z}\sum_{j=1}^{i-1}\left(P^{(v)}_{(z)(z)}\right)_{ij}+\sum_{i=1}^{N_{z'}}\sum_{j=1}^{N_z}\left(P^{(v)}_{(z')(z)}\right)_{ij}\right)}. \tag{20}$$

And the updates for the branching probabilities are similar, for $z = 1, 2$, $z \neq z'$ and $i = 1, 2, \ldots, N_z$:

$$\lambda_{z,i}^{(v)} = \mu^{(v)} + \sum_{j=1}^{i-1} \alpha_z^{(v)} g_z^{(v)} \left(t_i^{(z)} - t_j^{(z)}\right) + \sum_{j=1}^{N_{z'}} \alpha_{z'}^{(v)} g_{z'}^{(v)} \left(t_i^{(z)} - t_j^{(z')}\right) \mathbf{1}_{\{t_i^{(z)} > t_j^{(z')}\}}$$

$$\left(P_{(z)(z)}^{(v+1)}\right)_{ii} = \frac{\mu^{(v)}}{\lambda_{z,i}^{(v)}} \tag{21}$$

$$\left(P_{(z)(z)}^{(v+1)}\right)_{ij} = \frac{\alpha_z^{(v)} g_z^{(v)} \left(t_i^{(z)} - t_j^{(z)}\right)}{\lambda_{z,i}^{(v)}} \quad (\text{for } i > j) \tag{22}$$

$$\left(P_{(z)(z')}^{(v+1)}\right)_{ij} = \frac{\alpha_{z'}^{(v)} g_{z'}^{(v)} \left(t_i^{(z)} - t_j^{(z')}\right) \mathbf{1}_{\{t_i^{(z)} > t_j^{(z')}\}}}{\lambda_{z,i}^{(v)}} \tag{23}$$

Here we summarize the algorithm as follows:

Algorithm 3 Probability Weighted Histogram Estimation of Quasi-log-likelihood under H_1

Initialize: choose stopping critical value ϵ (e.g. 10^{-3}), initialize $P_{(z)(z')}^{(0)}$ and set $\left(P_{(z)(z')}^{(-1)}\right)_{ij} = \epsilon + \left(P_{(z)(z')}^{(0)}\right)_{ij}$ and iteration index $v = 0$.

while $\max_{t_i^{(z)} > t_j^{(z')}} \left| \left(P_{(z)(z')}^{(v)}\right)_{ij} - \left(P_{(z)(z')}^{(v-1)}\right)_{ij} \right| < \varepsilon$ **do**

1. Estimate background rate μ as in (18).
2. Estimate triggering components α_z, $g_z(t)$ as in (19) and (20).
3. Update probabilities $\left(P_{(z)(z')}^{(v+1)}\right)_{ij}$'s as in (21), (22) and (23).
4. $v = v + 1$

end while

We follow the derivation of EM-type algorithm in Appendix. B.2 and derive that Probability Weighted Histogram Estimation under the full model is again an EM-type algorithm. Similar to the proof framework above, we first use integral approximation of Schoenberg (2013) to approximate the Quasi-log-likelihood function and then lower bound it using Jensen's inequality:

$$\begin{aligned}
\widetilde{\ell}(\theta) \approx & \sum_{z=1}^{2} \sum_{i=1}^{N_z} \left[\left(P_{(z)(z)}\right)_{ii} \log \mu + \sum_{j<i} \left(P_{(z)(z)}\right)_{ij} \left(\log \alpha_z + \log \left(\sum_{k=1}^{n_0} g_{z,k} \mathbf{1}_{B_k} \left(t_i^{(z)} - t_j^{(z)}\right) \right) \right) \right. \\
& \left. + \sum_{j=1}^{N_{z'}} \left(P_{(z)(z')}\right)_{ij} \left(\log \alpha_{z'} + \log \left(\sum_{k=1}^{n_0} g_{z',k} \mathbf{1}_{B_k} \left(t_i^{(z)} - t_j^{(z')}\right) \right) \right) \right] - T\mu - N_1\alpha_1 - N_2\alpha_2 \\
& - \sum_{z=1}^{2} \sum_{i=1}^{N_z} \left(\sum_{i \geq j} \left(P_{(z)(z)}\right)_{ij} \log(\left(P_{(z)(z)}\right)_{ij}) + \sum_{j=1}^{N_{z'}} \left(P_{(z)(z')}\right)_{ij} \log(\left(P_{(z)(z')}\right)_{ij}) \mathbf{1}_{\{t_i^{(z)} > t_j^{(z')}\}} \right),
\end{aligned}$$

where $z' \neq z$. Note that the term $-N_1\alpha_1 - N_2\alpha_2$ comes from integral approximation. Add Lagrange multipliers and we will get the following objective function:

$$\begin{aligned}\widetilde{L}(\theta) = &\sum_{z=1}^{2}\sum_{i=1}^{N_z}\left[\left(P_{(z)(z)}\right)_{ii}\log\mu + \sum_{j<i}\left(P_{(z)(z)}\right)_{ij}\left(\log\alpha_z + \log\left(\sum_{k=1}^{n_0} g_{z,k}\mathbf{1}_{B_k}\left(t_i^{(z)} - t_j^{(z)}\right)\right)\right)\right.\\ &\left.+\sum_{j=1}^{N_{z'}}\left(P_{(z)(z')}\right)_{ij}\left(\log\alpha_{z'} + \log\left(\sum_{k=1}^{n_0} g_{z',k}\mathbf{1}_{B_k}\left(t_i^{(z)} - t_j^{(z')}\right)\right)\right)\right] - T\mu - N_1\alpha_1 - N_2\alpha_2\\ &-\sum_{z=1}^{2}\sum_{i=1}^{N_z}\left(\sum_{i\geq j}\left(P_{(z)(z)}\right)_{ij}\log(\left(P_{(z)(z)}\right)_{ij}) + \sum_{j=1}^{N_{z'}}\left(P_{(z)(z')}\right)_{ij}\log(\left(P_{(z)(z')}\right)_{ij})\mathbf{1}_{\{t_i^{(z)} > t_j^{(z')}\}}\right)\\ &-\sum_{z=1}^{2}\left[c_{z,1}\left(\sum_{k=1}^{n_0} g_{z,k}\Delta t_k - 1\right) - \sum_{i=1}^{N_z} c_{z,3}^{(i)}\left(\sum_{i\geq j}\left(P_{(z)(z)}\right)_{ij} + \sum_{j=1}^{N_{z'}}\left(P_{(z)(z')}\right)_{ij}\mathbf{1}_{\{t_i^{(z)} > t_j^{(z')}\}} - 1\right)\right].\end{aligned}$$

Then, by taking first order derivatives and setting them to zero we can validate Algorithm 3 as an EM-type algorithm.