A REPRESENTER THEOREM FOR COARSE-GRAINED HAWKES PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Hawkes processes are a standard framework for modeling event sequences and estimating triggering kernels that characterize inter-event influences from event data. In many practical settings, however, observations are limited to event counts within fixed time bins or windows rather than exact event timestamps. This challenging observation regime has motivated growing interest in estimating triggering kernels on aggregated event data. Among the approaches developed for this setting, the recently proposed *coarse-grained Hawkes process* (CGHP) offers a parametric framework that combines computational efficiency with robustness to temporal aggregation. In this paper, we propose a nonparametric extension of CGHP within a reproducing kernel Hilbert space (RKHS) framework, enabling flexible estimation of triggering kernels from aggregated observations. By exploiting the mathematical structure of CGHP, we establish a representer theorem, a cornerstone of kernel methods, that reduces the infinite-dimensional estimation problem to a finite-dimensional one, providing the foundation for an efficient estimation algorithm. To the best of our knowledge, this is the first representer theorem for a Hawkes process model designed for aggregated event data. Experiments on synthetic and real-world datasets demonstrate that the proposed method achieves higher estimation accuracy than the state-of-the-art nonparametric approach, with particularly pronounced improvements under coarse temporal aggregation.

1 INTRODUCTION

Hawkes processes (Hawkes, 1971a;b) are a fundamental class of point processes for modeling how past events influence future occurrences, in which the inter-event interactions are characterized by triggering kernels that describe their strength and temporal evolution through the conditional intensity function. Since their introduction in 1971, Hawkes processes have become a gold standard for analyzing interacting event sequences across diverse domains, including seismology (Adamopoulos, 1976; Ogata, 1988), finance (Bacry et al., 2015; Hawkes, 2018), criminology (Mohler et al., 2011), social media analytics (Zhou et al., 2013; Zhao et al., 2015), and neuroscience (Chornoboy et al., 1988). More recently, increasing interest has focused on nonparametric methods that recover flexible triggering kernel structures without imposing restrictive functional forms (Bonnet & Sangnier, 2025; Kim & Iwata, 2026).

Most existing Hawkes process models assume access to exact event timestamps. In many applications, however, only event counts within fixed time bins are observed due to reporting constraints, limited temporal resolution of sensing, or privacy requirements. For example, infectious-disease surveillance data are often reported as daily or weekly case counts (WHO, 2023), while crime records may be temporally coarsened before public release (UK Police, 2026). Although such aggregation removes within-bin timing information, the resulting count sequences can still retain substantial information about the underlying interaction structure, thereby motivating growing interest in developing Hawkes process models to recover triggering kernels from aggregated event observations. Conventional approaches range from treating unobserved event times as latent variables and relying on Monte Carlo inference (Shlomovich et al., 2022a;b; Chen et al., 2025) to constructing tractable discrete-time approximations of the underlying continuous-time processes (Rizoiu et al., 2022; Koyama, 2025). Most approaches, however, remain parametric, assuming standard forms of triggering kernels such as exponential or power-law functions, which may lead to model misspecification when the true interaction patterns are more complex.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

A notable nonparametric alternative is the integer-valued autoregressive (INAR) model (Kirchner (2016; 2017)), which approximates Hawkes processes in a discretized way: its parameters provide pointwise estimates of the triggering kernels, which are subsequently smoothed with cubic splines to obtain a continuous estimate. This method is computationally efficient and scales well to large datasets. However, the INAR approximation does not account for intra-bin excitation/inhibition among events, which can introduce substantial estimation bias when the aggregation scale or the bin width is not sufficiently small relative to the timescale of the triggering dynamics.

In this paper, we propose a nonparametric approach to triggering-kernel estimation that combines computational efficiency with robustness to temporal aggregation. We build on the recently proposed *coarse-grained Hawkes process* (CGHP), a tractable parametric approximation to multivariate Hawkes processes that explicitly accounts for intra-bin event interactions (Koyama, 2025). We extend CGHP within a reproducing kernel Hilbert space (RKHS) framework (Schölkopf & Smola, 2018; Shawe-Taylor & Cristianini, 2004), enabling flexible estimation of triggering kernels from aggregated event observations. By exploiting the mathematical structure of the CGHP objective, we establish a new representer theorem–a cornerstone of kernel methods–that reduces the infinite-dimensional functional estimation problem to a finite-dimensional one, thereby yielding an efficient estimation algorithm. To the best of our knowledge, this is the first representer theorem for a Hawkes process model specifically designed for aggregated event data.

We evaluate the proposed method on synthetic multivariate Hawkes processes with excitatory and partially inhibitory interactions, as well as on large-scale real-world social media data. The results demonstrate that our method achieves higher estimation accuracy than the state-of-the-art nonparametric approach, with particularly pronounced gains under coarse temporal aggregation, while maintaining high computational efficiency.

2 Proposed Method

2.1 Preliminary: Coarse-Grained Hawkes Processes

We consider a multivariate linear Hawkes process (Hawkes, 1971a;b) on the time domain $\mathbb{R}_+$, i.e., a d-dimensional counting process $(N_1(t), \dots, N_d(t))$ characterized by the following conditional intensity functions:

$$\lambda_i(t) = \mu_i + \sum_{j \in \mathcal{U}} \int_0^t \phi_{ij}(t-s) dN_j(s), \qquad t \in \mathbb{R}_+, \;\; i \in \mathcal{U} = \{1, \dots, d\}, \tag{1}$$

where $\mu_i \in \mathbb{R}_+$ denotes the baseline intensity of dimension i, and $\phi_{ij}(t-s) : \mathbb{R}_+ \to \mathbb{R}$ is the triggering kernel that quantifies the effect of an event of dimension j occurring at time s on the intensity of dimension i at time t. We consider the setting in which exact event timestamps are unavailable and only the event counts occurring within fixed time intervals, hereafter referred to as *bins*, are observed. Our goal is to estimate the Hawkes process parameters, $\{\mu_i\}$ and $\{\phi_{ij}(\cdot)\}$, from such aggregated event data.

Let $(0, T]$ denote the observation period, n the number of bins, and $\Delta = T/n$ the bin width. The aggregated event data, $\{\boldsymbol{X}_k\}_{k=1}^n$, are then defined as

$$\boldsymbol{X}_k = (X_k^1, \dots, X_k^d)^\top \in \mathbb{R}^d, \quad X_k^i = N_i(k\Delta) - N_i((k-1)\Delta), \quad k = 1, \dots, n, \tag{2}$$

where X_k^i denotes the number of events of dimension i observed in the k-th bin, $((k-1)\Delta, k\Delta]$. *Coarse-grained Hawkes processes* (CGHPs), recently proposed by Koyama (2025), offer a tractable approximation to the discrete-time count process $\{\boldsymbol{X}_k\}_{k=1}^n$ through an autoregressive representation. Specifically, the conditional expectation and the covariance matrix of $\boldsymbol{X}_k$, given the history $\mathcal{F}_{k-1} = \{\boldsymbol{X}_l : l \leq k-1\}$, are given by

$$\mathbb{E}\big[\boldsymbol{X}_k | \mathcal{F}_{k-1}\big] = (\boldsymbol{I}_d - \boldsymbol{\Phi}_0)^{-1} \left(\boldsymbol{\mu}\Delta + \sum_{l \leq k-1} \boldsymbol{\Phi}_{k-l} \boldsymbol{X}_l \right) \coloneqq \boldsymbol{\lambda}_k^*, \tag{3}$$

$$\mathrm{Cov}\big[\boldsymbol{X}_k | \mathcal{F}_{k-1}\big] = (\boldsymbol{I}_d - \boldsymbol{\Phi}_0)^{-1} \mathrm{diag}(\boldsymbol{\lambda}_k^*) (\boldsymbol{I}_d - \boldsymbol{\Phi}_0)^{-\top}, \tag{4}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$ is the baseline intensity vector, $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix, and $\{\boldsymbol{\Phi}_l \in \mathbb{R}^{d\times d}\}_{l\geq 0}$ is the *coarse-grained kernel*, defined as

$$\begin{aligned}[\boldsymbol{\Phi}_{k-l}]_{ij} &= \frac{1}{\Delta}\int_{(k-1)\Delta}^{k\Delta} dt \int_{(l-1)\Delta}^{l\Delta} du\ \phi_{ij}(t-u) \\ &= \int_{((k-l)-1)\Delta}^{((k-l)+1)\Delta} \phi_{ij}(v)\big(1-|(k-l)-v/\Delta|\big)\, dv,\end{aligned} \tag{5}$$

for $1 \leq i, j \leq d$. The key idea underlying CGHPs is to obtain a tractable characterization of inter-bin interactions by assuming that the latent event times are uniformly distributed within each bin. Under this approximation, the coarse-grained kernel $\boldsymbol{\Phi}_{k-l}$ in Eq. (5) captures the influence of the l-th bin's count on the k-th one. Notably, CGHPs explicitly account for intra-bin interactions through $\boldsymbol{\Phi}_0$, leading to improved approximation accuracy for the finite bin width Δ compared with the conventional INAR-based approach (Kirchner, 2017). We refer the reader to Koyama (2025) for further details. Based on the second-order statistical structure defined in Eqs. (3-4), the CGHP loss function is defined by the negative log-likelihood of a multivariate normal distribution:

$$L_{\mathrm{CG}} = \sum_{k=1}^{n}\left(\boldsymbol{\Upsilon}_k - \sum_{l=0}^{k-1}\boldsymbol{\Phi}_l\boldsymbol{\Upsilon}_{k-l}\right)^\top \boldsymbol{\Lambda}^{-1}\left(\boldsymbol{\Upsilon}_k - \sum_{l=0}^{k-1}\boldsymbol{\Phi}_l\boldsymbol{\Upsilon}_{k-l}\right) - 2n\log|\boldsymbol{I}_d - \boldsymbol{\Phi}_0|, \tag{6}$$

where $\hat{\boldsymbol{\lambda}} = n^{-1}\sum_{k=1}^{n}\boldsymbol{X}_k$ is the empirical mean of the event counts, $\boldsymbol{\Upsilon}_k = \boldsymbol{X}_k - \hat{\boldsymbol{\lambda}}$ is the centered count data, and $\boldsymbol{\Lambda} = \mathrm{diag}(\hat{\boldsymbol{\lambda}})$. Accordingly, under the assumption that each triggering kernel $\phi_{ij}(\cdot)$ is represented by a prescribed parametric family characterized by a parameter vector $\boldsymbol{\theta}$, the parameters of the underlying Hawkes process are estimated as follows:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L_{\mathrm{CG}}(\boldsymbol{\theta}), \quad \hat{\boldsymbol{\mu}} = (\boldsymbol{I}_d - \hat{\boldsymbol{\Omega}})\hat{\boldsymbol{\lambda}}/\Delta, \quad \hat{\boldsymbol{\Omega}} = \sum_{k=0}^{\infty}\boldsymbol{\Phi}_k(\hat{\boldsymbol{\theta}}), \tag{7}$$

where $\hat{\boldsymbol{\Omega}}$ denotes the estimated branching-ratio matrix.

2.2 A Representer Theorem for Coarse-Grained Hawkes Processes

We impose causality and finite-support assumptions on the triggering kernels, $\boldsymbol{\phi}(\cdot) = \{\phi_{ij}(\cdot)\}$, such that

$$\boldsymbol{\phi}(v) = \boldsymbol{\psi}(v)\mathbf{1}_{0<v\leq R}, \qquad v \in \mathbb{R}, \tag{8}$$

where $R \in \mathbb{R}_+$, $\mathbf{1}_{(\cdot)}$, and $\boldsymbol{\psi}(\cdot) = \{\psi_{ij}(\cdot)\}_{1\leq i,j\leq d}$ denote the support window, the indicator function, and the untruncated triggering kernels, respectively. Let R be an integer multiple of Δ. Then, the coarse-grained kernels, $\{\boldsymbol{\Phi}_l\}_{l=0}^{n-1}$, defined in Eq. (5) are equal to zero when the integration interval falls outside the support window, i.e., $(l-1)\Delta \geq R$:

$$\boldsymbol{\Phi}_l = 0, \qquad \overline{n} \leq l \leq n-1, \qquad \overline{n} = \min(n, R/\Delta + 1). \tag{9}$$

which can substantially reduce the computational cost when R/Δ is sufficiently small relative to the data length n. Finite support windows are commonly adopted in the literature to reduce computational burden for nonparametric estimation (Bonnet et al., 2023; Halpin, 2013).

Let $k : \mathcal{T}\times\mathcal{T} \to \mathbb{R}$ be a positive semi-definite kernel on a one-dimensional compact space $\mathcal{T} \subset \mathbb{R}$. Then there exists a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ (Schölkopf & Smola, 2018; Shawe-Taylor & Cristianini, 2004) associated with the RKHS kernel $k(\cdot,\cdot)$. Given aggregated event data $\{\boldsymbol{X}_k\}_{k=1}^{n}$ with bin width Δ, we consider the following regularized optimization problem over the untruncated triggering kernels, $\boldsymbol{\psi}(\cdot)$, within the CGHP framework:

$$\hat{\boldsymbol{\psi}} = \arg\min_{\boldsymbol{\psi}\in\mathcal{H}_k^{d^2}}\left[L_{\mathrm{CG}}\big(\{\boldsymbol{\Phi}_l(\boldsymbol{\psi})\}\big) + \frac{1}{\gamma}\sum_{1\leq i,j\leq d}||\psi_{ij}||_{\mathcal{H}_k}^2\right], \tag{10}$$

where $||\cdot||_{\mathcal{H}_k}^2$ denotes the squared Hilbert space norm, and $\gamma \in \mathbb{R}_+$ is a regularization hyperparameter. Since the loss function $L_{\mathrm{CG}}\big(\{\boldsymbol{\Phi}_l(\boldsymbol{\psi})\}\big)$ involves integrals of the functional variable $\boldsymbol{\psi}(\cdot)$ through $\{\boldsymbol{\Phi}_l\}$ (see Eq. 5), the functional optimization problem may appear intractable. Nevertheless, as in classical kernel methods, its solution admits a finite-dimensional representation in the RKHS $\mathcal{H}_k$ as a linear combination of functions associated with the RKHS kernel. Theorem 1 formalizes this result by establishing a novel representer theorem for the functional optimization problem in Eq. (10). The proof is provided in Appendix A.

Theorem 1. *Given aggregated event data $\{\boldsymbol{X}_k\}_{k=1}^n$ with bin width Δ, the solutions of the functional optimization problem (10), denoted as $\hat{\boldsymbol{\psi}}(\cdot) = \{\hat{\psi}_{ij}(\cdot)\}_{1\le i,j\le d}$, admit the following finite-dimensional representation in terms of the RKHS kernel $k(\cdot,\cdot)$:*

$$\hat{\psi}_{ij}(v) = \sum_{l=0}^{\overline{n}-1} \alpha_l^{ij} h_l(v), \quad h_l(v) = \int_{(l-1)\Delta}^{(l+1)\Delta} k(v,s)\big(1-|l-s/\Delta|\big)\mathbf{1}_{0<s\le R}\, ds, \tag{11}$$

where $\{\alpha_l^{ij} \in \mathbb{R}\}$ denotes the corresponding dual coefficients, and $\overline{n} = \min(n, R/\Delta+1)$.

Corollary 2 provides a finite-dimensional formulation to obtain the coefficients $\{\alpha_l^{ij}\}$ in Eq. (11). The proof is provided in Appendix B. Its extension to the multi-trial setting is shown in Appendix C.

Corollary 2. *Let $\mathbf{A} \in \mathbb{R}^{d\times d\times \overline{n}}$ and $\mathbf{B} \in \mathbb{R}^{d\times d\times \overline{n}}$ be 3-D tensors defined by $[\mathbf{A}]_{ijl} \coloneqq \alpha_{l-1}^{ij}$ and $[\mathbf{B}]_{ijl} \coloneqq [\boldsymbol{\Phi}_{l-1}]_{ij}$ for $1 \le l \le \overline{n}$, respectively. Then the dual coefficient tensor $\mathbf{A}$ is related to the coarse-grained kernel tensor $\mathbf{B}$ through a positive semidefinite matrix $\boldsymbol{H} \in \mathbb{R}^{\overline{n}\times\overline{n}}$ as*

$$\mathbf{A}_{ij:} = \boldsymbol{H}^{-1}\mathbf{B}_{ij:}, \quad [\boldsymbol{H}]_{ll'} = \int_{((l-1)-1)\Delta}^{((l-1)+1)\Delta} h_{l'-1}(v)\big(1-|l-1-v/\Delta|\big)\mathbf{1}_{0<v\le R}\, dv, \tag{12}$$

and $\mathbf{B}$ is the minimizer of the following objective:

$$\sum_{k=1}^{n}\Big(\boldsymbol{\Upsilon}_k - \sum_{l=1}^{\overline{k}} \mathbf{B}_{::l}\boldsymbol{\Upsilon}_{k-l+1}\Big)^{\top}\boldsymbol{\Lambda}^{-1}\Big(\boldsymbol{\Upsilon}_k - \sum_{l=1}^{\overline{k}} \mathbf{B}_{::l}\boldsymbol{\Upsilon}_{k-l+1}\Big) - 2n\log|\boldsymbol{I}_d - \mathbf{B}_{::1}| + \frac{1}{\gamma}\sum_{i,j}\mathbf{B}_{ij:}^{\top}\boldsymbol{H}^{-1}\mathbf{B}_{ij:}, \tag{13}$$

where $\mathbf{B}_{ij:}$ denotes a 1-D slice of the 3-D tensor, $\mathbf{B}_{::l}$ denotes a 2-D slice, and $\overline{k} = \min(k, R/\Delta+1)$.

2.3 Recipes for the Dual Coefficient Optimization

Corollary 2 indicates that to obtain nonparametric estimates of the triggering kernels, we need to solve the optimization problem (13) regarding $(d^2\overline{n})$-dimensional variable $\mathbf{B}$, using an iterative optimization method. A key issue is to ensure that the log-determinant, $\log|\boldsymbol{I}_d - \mathbf{B}_{::1}|$, in the second term of the objective remains well-defined, or equivalently, $|\boldsymbol{I}_d - \mathbf{B}_{::1}| > 0$, throughout the optimization. Although this constraint could be handled by projected gradient descent methods, we instead adopt a parameterization of $\mathbf{B}_{::1}$ via LDU decomposition: specifically, we introduce a unit lower-triangular matrix $\boldsymbol{L}$, a unit upper-triangular matrix $\boldsymbol{U}$, and a diagonal matrix $\boldsymbol{D}$ with strictly positive diagonal entries, as parameterization variables, so that $\mathbf{B}_{::1} = \boldsymbol{I}_d - \boldsymbol{LDU}$; the positivity of the diagonal entries of $\boldsymbol{D}$ is enforced using the softplus function. Then, the determinant term keeps positive at every iteration: $|\boldsymbol{I}_d - \mathbf{B}_{::1}| = |\boldsymbol{L}||\boldsymbol{D}||\boldsymbol{U}| = \prod_{i=1}^d [\boldsymbol{D}]_{ii} > 0$, where we have used $|\boldsymbol{L}| = |\boldsymbol{U}| = 1$. No additional constraints are imposed on the other slices of $\mathbf{B}$.

Evaluating $\boldsymbol{H}$ in the objective function and $\{h_l(\cdot)\}_{l=0}^{\overline{n}-1}$ in the triggering-kernel estimator requires up to n^2 weighted (double) integrals involving the RKHS kernel $k(\cdot,\cdot)$. Proposition 3 shows that, for shift-invariant RKHS kernels, including the widely used Gaussian and Laplace RKHS kernels, all of these integrals can be expressed in closed form in terms of their moment functions up to the third order (the proof is provided in Appendix D). This result enables efficient evaluation of both $\boldsymbol{H}$ and $\{h_l(\cdot)\}_{l=0}^{\overline{n}-1}$ without resorting to numerical integration during the optimization.

Proposition 3. *Let an RKHS kernel be shift-invariant, $k(v,s) = k(\theta|v-s|)$, where $\theta \in \mathbb{R}$ denotes the inverse of the scale parameter. Then $\boldsymbol{H}$ in Eq. (12) can be expressed in terms of its moment functions, $F_\eta^i(v) = \int_0^1 s^i k(\eta|v-s|)ds$ for $i \in \{0,1,2,3\}$, as follows:*

$$\begin{aligned}
[\boldsymbol{H}]_{ll'} &= \Delta^2\Big[\big(\mathbf{1}_{1<l\le n^*+1}\mathbf{1}_{1<l'\le n^*+1} + \mathbf{1}_{1\le l\le n^*}\mathbf{1}_{1\le l'\le n^*}\big)\big(Z_1(l-l') + Z_1(l'-l)\big) \\
&\qquad + \mathbf{1}_{1<l\le n^*+1}\mathbf{1}_{1\le l'\le n^*}Z_2(l'-l) + \mathbf{1}_{1<l'\le n^*+1}\mathbf{1}_{1\le l\le n^*}Z_2(l-l')\Big], \\
Z_1(l-l') &= \frac{1}{3}F_{\Delta\theta}^0(l-l') - \frac{1}{2}F_{\Delta\theta}^1(l-l') + \frac{1}{6}F_{\Delta\theta}^3(l-l'), \\
Z_2(l-l') &= \frac{1}{6}F_{\Delta\theta}^0(l-l'+1) + \frac{1}{2}F_{\Delta\theta}^1(l-l'+1) - \frac{1}{2}F_{\Delta\theta}^2(l-l'+1) \\
&\qquad + \frac{1}{6}\Big[F_{\Delta\theta}^3(l-l'+2) - F_{\Delta\theta}^3(l-l'+1)\Big], \quad 1 \le l, l' \le \overline{n},
\end{aligned} \tag{14}$$

where the support window R is assumed to be an integer multiple of Δ, $n^* = R/\Delta \in \{1, 2, \dots\}$, and $\overline{n} = \min(n, R/\Delta + 1)$. Similarly, $\{h_l(\cdot)\}_{l=0}^{\overline{n}-1}$ in Eq. (11) can be expressed as follows:

$$h_l(v) = \Delta\Big[F^1_{\Delta\theta}(1 - l + v/\Delta)\mathbf{1}_{0<l\leq n^*} + F^1_{\Delta\theta}(1 + l - v/\Delta)\mathbf{1}_{0\leq l<n^*}\Big], \quad 0 \leq l \leq \overline{n} - 1. \quad (15)$$

For commonly used shift-invariant RKHS kernels, such as the Gaussian and Laplace RKHS kernels, the corresponding moment functions can be obtained analytically (see Table D1). For more complex shift-invariant kernels, such as the periodic squared-exponential kernels, however, closed-form expressions may not be available. Importantly, each of $\{F^i_\eta(s)\}_{i=0}^3$ is a function of only two variables and is independent of both the data and the hyperparameters. Therefore, a highly accurate lookup table can be precomputed once on a sufficiently fine grid, after which evaluations of $\{F^i_\eta(s)\}_{i=0}^3$ can be performed efficiently by linear interpolation. Consequently, Proposition 3 ensures that the dual optimization problem can still be solved efficiently even when the moment functions do not admit closed-form expressions.

2.4 Complexity Analysis

The computational complexity of solving the dual optimization problem in Eq. (13) using a gradient-based method is naively $\mathcal{O}(d^2 n\overline{n}P + \overline{n}^3)$, where P denotes the number of iterations required for convergence. The first term arises from evaluating $\{\sum_{l=1}^k \mathbf{B}_{::l}\boldsymbol{\Upsilon}_{k-l+1}\}_{k=1}^n$, at each optimization step, while the second term corresponds to the inversion of $\boldsymbol{H}$. Importantly, the sum $\sum_{l=1}^k \mathbf{B}_{::l}\boldsymbol{\Upsilon}_{k-l+1}$ is a discrete convolution along the lag index l. Hence, the entire sequence $\{\sum_{l=1}^k \mathbf{B}_{::l}\boldsymbol{\Upsilon}_{k-l+1}\}_{k=1}^n$ can be evaluated simultaneously using the Fast Fourier Transform (FFT) (Cooley & Tukey, 1965), reducing its computational cost to $\mathcal{O}(d^2 n \log n)$. Consequently, with the FFT-based implementation, the overall computational complexity of solving Eq. (13) becomes $\mathcal{O}((d^3 + d^2\overline{n}^2 + d^2 n \log n)P + \overline{n}^3)$, where the d^3 term arises from the matrix multiplications required by the LDU-based parameterization $\mathbf{B}_{::1} = \boldsymbol{I}_d - \boldsymbol{LDU}$, as detailed in Section 2.3. The FFT-based implementation is particularly advantageous when d is moderate and $1 \ll \overline{n} \ll n$. Regardless of whether FFT is used, when $\overline{n} \ll n$, the computational cost of our nonparametric method scales favorably with the number of observations n, making it well-suited to large-scale datasets.

3 Related Work

While triggering-kernel estimation from event sequences has a long history in Hawkes process research, estimation from temporally aggregated data has attracted systematic attention more recently.

A pioneering contribution is the nonparametric estimation framework of Kirchner (2016; 2017), which approximates the bin-count sequence of a multivariate Hawkes process using an integer-valued autoregressive (INAR) model. Conditional least-squares estimates of the autoregressive coefficients yield pointwise estimates of the triggering kernels after rescaling the coefficients by the bin width, which can subsequently be smoothed using cubic splines to obtain continuous function estimates. The availability of a closed-form least-squares estimator makes this approach computationally attractive. However, the INAR approximation neglects interactions among latent events within the same bin, potentially introducing systematic bias in the estimates when the bin width is not sufficiently small relative to the characteristic timescale of the triggering kernels.

A direct way to address this limitation is to use Monte Carlo (MC) methods to integrate over the latent within-bin event times in likelihood-based estimation. Shlomovich et al. (2022b;a) introduced an MC expectation–maximization framework, whereas Chen et al. (2025) developed a pseudo-marginal Metropolis–Hastings approach to maximum likelihood estimation. Zhou & Papadogeorgou (2026) recently proposed a Bayesian estimation method based on Markov Chain Monte Carlo. Although these methods can provide accurate estimates, repeated MC computation poses challenges for large-scale applications. These approaches assume parametric triggering kernels, and extending them to more flexible nonparametric representations may further increase the computational burden.

Computationally tractable alternatives include spectral estimation based on the Whittle likelihood (Cheysson & Lang, 2022), the mean behavior Poisson process (MBPP) (Rizoiu et al., 2022), and the coarse-grained Hawkes process (CGHP) (Koyama, 2025), all of which adopt a parametric formulation. Spectral estimation exploits the second-order dependence retained in aggregated counts

via their spectral density; MBPP approximates a Hawkes process with an inhomogeneous Poisson process whose intensity equals the expected Hawkes intensity; and CGHP provides a tractable approximation for Hawkes processes under the assumption that latent event times are uniformly distributed within a bin. Our method extends CGHP to nonparametric triggering-kernel estimation within a reproducing kernel Hilbert space (RKHS). It retains the underlying treatment of temporal aggregation while replacing a prescribed kernel family with RKHS-regularized function estimation. Our contribution, therefore, lies in this nonparametric formulation and its associated representer theorem, rather than in a new approximation to the bin-count process.

RKHS-based nonparametric estimation of triggering kernels has also been explored using exact event timestamps (Yang et al., 2017; Bonnet & Sangnier, 2025; Kim & Iwata, 2026). Particularly relevant is the work of Kim & Iwata (2026), establishing an exact representer theorem for multivariate linear Hawkes processes under a least-squares contrast. The CGHP loss in Eq. (6) similarly builds on a weighted quadratic criterion, augmented by a log-determinant term reflecting the covariance structure of aggregated observations. In this sense, our representer theorem can be viewed as an aggregation-aware counterpart of their exact-timestamp result, connecting RKHS-based Hawkes process estimation across temporal resolutions.

4 Experiments

We evaluated the performance of our proposed method, $\texttt{CGH}_{\texttt{rkhs}}$, against parametric and nonparametric baselines. The parametric baseline, $\texttt{CGH}_{\texttt{exp}}$, is based on CGHP with exponential triggering kernels (Koyama, 2025), whereas `INAR` is a state-of-the-art nonparametric estimator based on the INAR model (Kirchner, 2016; 2017). For $\texttt{CGH}_{\texttt{rkhs}}$, we used a Gaussian RKHS kernel, $k(v, s) = \exp[-(\theta|v - s|)^2]$, where θ is the inverse scale hyperparameter. Both θ and the regularization hyperparameter γ were selected by cross-validation. For `INAR`, we applied cubic smoothing splines to the pointwise triggering-kernel estimates using `make_smoothing_spline` from `scipy.interpolate`. The smoothing parameter was selected automatically by minimizing the generalized cross-validation criterion (Woltring, 1986). All models were fitted using Adam with a learning rate of 0.02.

The models were implemented in TensorFlow-2.10, with SciPy used for spline smoothing. All experiments were conducted on a MacBook Pro equipped with a 16-core CPU (Apple M4 Max), with the GPU disabled[1].

4.1 Synthetic Data

We evaluated the estimation accuracy of the competing methods using synthetic data from three-dimensional Hawkes processes under three scenarios. In the two scenarios with exclusively excitatory triggering kernels, event sequences were simulated using the Python library `tick` (Bacry et al., 2017). The third scenario involved triggering kernels with negative values; in this setting, we simulated a nonlinear Hawkes process using the Python code from (Bonnet & Sangnier, 2025). In all scenarios, the simulated event sequences were aggregated into bin counts. For each scenario, we prepared 100 trials of sequences for every combination of observation horizon $T \in \{1000, 2,000, 10,000\}$ and bin width $\Delta \in \{1.0, 0.5, 0.2, 0.1\}$.

For $\texttt{CGH}_{\texttt{rkhs}}$, hyperparameters were selected separately for each trial as follows: the model was fitted to the first 70% of bins, and the CGHP loss was evaluated on the remaining 30%. We selected the hyperparameters that minimized this validation loss over the candidate grids of $\theta \in \{0.1, 0.2, \ldots, 1.0\}$ and $\gamma \in \{0.2, 0.4, \ldots, 1.0\}$.

Estimation accuracy, denoted by δ^2, was assessed using the integrated squared error between the true and estimated triggering kernels, denoted by $\phi^*_{ij}(\cdot)$ and $\hat{\phi}_{ij}(\cdot)$, respectively:

$$\delta^2 = \frac{1}{Rd^2} \sum_{i=1}^{d} \sum_{j=1}^{d} \int_0^R \left| \phi^*_{ij}(v) - \hat{\phi}_{ij}(v) \right|^2 dv, \tag{16}$$

where R is the support window. Computational efficiency, denoted by τ, was measured by the time required for model fitting with the selected hyperparameters.

[1]Code is submitted as supplementary material.

Table 1: Results on the exponential scenario across 100 trials with standard deviations in brackets. Integrated squared errors δ^2 not significantly ($p \geq 0.01$) different from the best one under the Mann-Whitney U test (Holm, 1979) are shown in bold.

$\delta^2 \times 10^4$	$T = 1000$				$T = 2000$				$T = 10000$			
	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1
$\texttt{CGH}_{\texttt{exp}}$	**6.30** (4.33)	**4.34** (2.85)	**2.83** (1.71)	**2.38** (1.08)	**4.86** (4.21)	**2.87** (2.30)	**1.54** (1.03)	**1.19** (0.51)	**4.17** (4.43)	**2.06** (2.24)	**0.70** (0.64)	**0.39** (0.30)
INAR	11.5 (5.43)	8.42 (3.70)	6.44 (2.22)	6.04 (2.11)	8.73 (4.43)	5.45 (2.10)	3.63 (1.13)	3.36 (0.99)	5.56 (3.22)	2.47 (1.16)	1.23 (0.45)	0.96 (0.30)
$\texttt{CGH}_{\texttt{rkhs}}$	8.52 (3.69)	7.16 (2.74)	6.50 (2.30)	6.58 (2.10)	5.92 (2.58)	4.89 (1.78)	4.31 (1.60)	4.31 (1.46)	**2.35** (1.43)	**1.42** (0.55)	1.15 (0.33)	1.14 (0.38)
τ [sec]	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1
$\texttt{CGH}_{\texttt{exp}}$	4.06	6.54	16.0	20.9	5.43	7.06	14.4	27.7	12.7	18.6	38.5	53.3
INAR	0.82	1.20	2.27	3.39	1.45	1.62	3.18	5.49	3.16	4.94	10.8	15.9
$\texttt{CGH}_{\texttt{rkhs}}$	0.64	0.68	0.87	1.25	0.92	0.75	1.06	2.17	1.21	1.44	2.64	4.06

Exponential Scenario

We generated event sequences from 3-variate Hawkes processes (1) with unit baseline intensity, $\{\mu_i = 1\}_{i=1}^3$, and exponential triggering kernels,

$$\phi_{ij}(v) = \eta_{ij}\xi_{ij}e^{-\xi_{ij}v}, \qquad 1 \leq i, j \leq 3, \tag{17}$$

where the decay parameters, $\{\xi_{ij}\}$, and the adjacency matrix, $[\eta_{ij}]_{ij}$, were randomly generated. Further details are provided in Appendix E.1.

Results: Table 1 reports the triggering-kernel estimation error δ^2 and computation time τ. Representative examples of the estimated triggering kernels are provided in Figure F1. Across all observation horizons T considered, the parametric CGHP with exponential triggering kernels, $\texttt{CGH}_{\texttt{exp}}$, achieves competitive performance, which is reasonable because the assumed triggering kernels match the latent data-generating process. Among the nonparametric methods, our $\texttt{CGH}_{\texttt{rkhs}}$ outperforms INAR at the coarser bin widths $\Delta \in \{1.0, 0.5\}$, demonstrating its advantage under coarse temporal aggregation.

In terms of computation time, the nonparametric $\texttt{CGH}_{\texttt{rkhs}}$ is generally faster than the parametric $\texttt{CGH}_{\texttt{exp}}$. One possible explanation is the structure of their optimization objectives: the objective of $\texttt{CGH}_{\texttt{rkhs}}$ in Corollary 2 is quadratic in $\mathbf{B}$ except for $\mathbf{B}_{::1}$. This landscape may facilitate gradient-based optimization compared with the highly nonlinear objective of $\texttt{CGH}_{\texttt{exp}}$. Although $\texttt{CGH}_{\texttt{rkhs}}$ is faster than INAR, this comparison may be implementation-dependent and should not be interpreted as an inherent computational advantage over INAR: INAR admits a closed-form least-squares solution (Kirchner, 2017), which may yield shorter runtimes than iterative optimization.

Gamma-mixture Scenario

To clarify the benefits of nonparametric estimation, we generated event sequences from 3-variate Hawkes processes (1) with unit baseline intensity, $\{\mu_i = 1\}_{i=1}^3$, and 2-component gamma-mixture triggering kernels,

$$\phi_{ij}(v) = \eta_{ij}\sum_{q=1}^{2} r_{ij}^q g(v|\xi_{ij}^q, \kappa_{ij}^q), \quad g(v|\xi,\kappa) = \xi^\kappa \frac{v^{\kappa-1}e^{-\xi v}}{\Gamma(\kappa)}, \quad \sum_{q=1}^{2} r_{ij}^q = 1, \quad 1 \leq i, j \leq 3, \tag{18}$$

which was a different scenario from the parametric assumption of $\texttt{CGH}_{\texttt{exp}}$. For each trial, the gamma distribution parameters $\{(\xi_{ij}^q, \kappa_{ij}^q)\}$, the mixture ratio (r_{ij}^1, r_{ij}^2), and the adjacency matrix $[\eta_{ij}]_{ij}$, were randomly sampled. Further details are provided in Appendix E.2.

Results: Table 2 summarizes the results. Except for the shortest observation period ($T = 1000$), the nonparametric methods $\texttt{CGH}_{\texttt{rkhs}}$ and INAR achieve significantly higher estimation accuracy than the parametric method $\texttt{CGH}_{\texttt{exp}}$. This advantage reflects the inability of the assumed exponential kernels to capture the underlying data-generating process, highlighting the importance of

Table 2: Results on the gamma-mixture scenario across 100 trials with standard deviations in brackets. The notation follows Table 1.

$\delta^2 \times 10^3$	$T = 1000$				$T = 2000$				$T = 10000$			
	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1
$\texttt{CGH}_{\texttt{exp}}$	**1.86** (1.07)	**1.73** (0.98)	**1.69** (0.96)	1.71 (0.95)	**1.72** (1.05)	1.62 (0.97)	1.61 (0.96)	1.60 (0.95)	1.66 (1.04)	1.57 (0.97)	1.55 (0.96)	1.54 (0.95)
INAR	**2.11** (1.12)	**1.76** (0.88)	**1.45** (0.61)	**1.32** (0.50)	**1.91** (1.06)	1.49 (0.84)	**1.06** (0.49)	**0.82** (0.28)	1.74 (1.06)	1.29 (0.89)	0.45 (0.27)	**0.25** (0.11)
$\texttt{CGH}_{\texttt{rkhs}}$	**1.83** (0.95)	**1.57** (0.69)	**1.43** (0.60)	**1.40** (0.55)	**1.55** (0.80)	**1.18** (0.57)	**0.98** (0.40)	**0.94** (0.36)	**1.05** (0.69)	**0.64** (0.49)	**0.29** (0.10)	0.29 (0.12)
τ [sec]	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1
$\texttt{CGH}_{\texttt{exp}}$	5.54	8.65	24.0	34.0	7.45	10.0	21.8	40.1	14.9	24.4	53.9	84.9
INAR	0.84	1.25	6.11	3.14	1.53	1.72	3.34	5.58	3.42	5.32	10.6	16.4
$\texttt{CGH}_{\texttt{rkhs}}$	0.77	0.69	0.96	1.16	0.91	0.80	1.08	1.65	1.17	1.42	2.56	4.07

Table 3: Results on the refractory scenario across 100 trials with standard deviations in brackets. The notation follows Table 1.

$\delta^2 \times 10^2$	$T = 1000$				$T = 2000$				$T = 10000$			
	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1
$\texttt{CGH}_{\texttt{exp}}$	**4.89** (0.39)	5.02 (1.05)	4.76 (0.25)	4.75 (0.20)	4.77 (0.15)	4.75 (0.24)	4.70 (0.11)	4.70 (0.16)	4.71 (0.03)	4.69 (0.03)	4.66 (0.03)	4.65 (0.03)
INAR	5.32 (0.53)	4.95 (0.36)	4.04 (0.76)	**3.33** (0.93)	5.15 (0.33)	4.74 (0.28)	3.63 (0.76)	**2.53** (0.64)	5.04 (0.14)	4.45 (0.21)	2.79 (0.65)	**1.03** (0.33)
$\texttt{CGH}_{\texttt{rkhs}}$	4.93 (1.29)	**4.06** (1.09)	**3.22** (1.04)	**3.26** (1.03)	**4.05** (0.99)	**3.27** (1.04)	**2.68** (1.15)	**2.74** (1.19)	**3.04** (0.66)	**2.05** (1.21)	**1.81** (1.32)	**1.91** (1.38)
τ [sec]	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1	Δ=1	0.5	0.2	0.1
$\texttt{CGH}_{\texttt{exp}}$	18.9	45.3	86.1	107	32.7	54.6	123	170	133	142	351	517
INAR	3.81	6.90	9.00	9.34	6.79	8.32	14.0	16.0	22.4	23.8	38.3	46.4
$\texttt{CGH}_{\texttt{rkhs}}$	0.93	0.80	0.89	0.90	1.71	0.89	1.08	1.33	6.36	2.42	2.55	3.60

nonparametric flexibility. For $T = 1000$, all methods perform poorly, suggesting that the available data are insufficient to reliably recover the complex shapes of the underlying triggering kernels.

Among the nonparametric methods, our proposed $\texttt{CGH}_{\texttt{rkhs}}$ outperforms INAR at relatively large bin widths ($\Delta \in \{1.0, 0.5, 0.2\}$), with increasingly pronounced gains as the observation period grows. However, when a long observation period is combined with fine temporal resolution, as in $(T, \Delta) = (10000, 0.1)$, INAR tends to achieve higher accuracy. One possible explanation is the stationary-mean approximation used by CGHP: in Eq. (7), CGHP estimates the baseline intensity $\boldsymbol{\mu}$ approximately with the empirical stationary mean of the bin counts, whereas INAR estimates $\boldsymbol{\mu}$ directly within its optimization problem. Extending our method to estimate $\boldsymbol{\mu}$ without relying on this approximation is an important direction for future work. Representative examples of the estimated triggering kernels are provided in Appendix F.

Refractory Scenario

All methods evaluated here are based on linear Hawkes processes. Although linear Hawkes processes primarily target purely excitatory triggering kernels, the corresponding estimators may also perform well when inhibitory effects are present but not dominant. To assess their performance in this regime, we generated data using the *refractory scenario* in Kim & Iwata (2026), which exhibits short-term self-inhibition. Further details are provided in Appendix E.3.

Results: Table 3 shows that our method, $\texttt{CGH}_{\texttt{rkhs}}$, achieves the highest estimation accuracy at relatively large bin widths ($\Delta \in \{1.0, 0.5, 0.2\}$) in this scenario. The only exception occurs when both the amount of data and the temporal resolution are severely limited ($T = 1000$, $\Delta = 1$), in which case the parametric method $\texttt{CGH}_{\texttt{exp}}$ performs best. Figure F3 further demonstrates that $\texttt{CGH}_{\texttt{rkhs}}$ successfully captures the short-term self-inhibitory effect.

Table 4: Results on the SNS data across 45 train–test splits with standard deviations in brackets. Test log-likelihood scores ll not significantly ($p \geq 0.01$) different from the best one under the Mann-Whitney U test (Holm, 1979) are shown in bold.

	$\Delta = 0.2$ h			$\Delta = 0.16$ h			$\Delta = 0.1$ h			$\Delta = 0.05$ h		
	ll	ρ	τ[sec]	ll	ρ	τ[sec]	ll	ρ	τ[sec]	ll	ρ	τ[sec]
$\texttt{CGH}_{\texttt{exp}}$	**30.8** (7.45)	0.42 -	96.7 (131)	28.7 (7.68)	0.02 -	26.3 (45.1)	25.2 (7.92)	0.00 -	138 (294)	20.3 (7.89)	0.00 -	121 (250)
INAR	25.9 (3.83)	0.00 -	0.95 (0.64)	30.7 (4.09)	0.00 -	0.83 (0.07)	**36.7** (4.63)	0.33 -	2.24 (1.15)	**39.5** (5.66)	0.02 -	5.60 (1.17)
$\texttt{CGH}_{\texttt{rkhs}}$	**31.1** (3.69)	0.58 -	1.06 (0.83)	**34.0** (4.84)	0.98 -	0.63 (0.03)	**36.9** (5.43)	0.67 -	1.26 (0.77)	**41.3** (5.35)	0.98 -	2.36 (1.22)

4.2 Real-world Data

We evaluated the predictive performance of the competing methods on a large-scale SNS dataset (Zhao et al., 2015) comprising original tweets, their associated retweet sequences, and the follower counts of each retweeting user. From it, we extracted a dataset comprising 104,833 events across 1,000 retweet sequences, in which users were grouped into three categories ($d = 3$) based on follower counts: *small*, *middle*, and *large*. Further details are provided in Appendix E.4.

We partitioned the 1,000 sequences into ten equally sized groups and constructed all ${}_{10}\mathrm{C}_2 = 45$ train–test splits by holding out two groups for testing and using the remaining for training. For each split, we aggregated the training events into bins of width $\Delta \in \{0.2, 0.16, 0.1, 0.05\}$ hours and fitted each competing method to the aggregated data. Predictive performance was then evaluated based on the log-likelihood of the test event sequences, denoted by ll, where a common post-hoc clipping, $\max(\hat{\lambda}(\cdot), 10^{-8})$, was applied to the estimated intensity function $\hat{\lambda}(\cdot)$ for each competing method. For reference, we also computed each method's win rate across the 45 splits, denoted by ρ.

For $\texttt{CGH}_{\texttt{rkhs}}$, the hyperparameters were selected separately for each train–test split. Specifically, the first 70% of the training sequences were used for model fitting and the remaining 30% for validation. The validation criterion was the Poisson negative log-likelihood of the bin counts, with each Poisson mean given by Eq. (3). We selected the hyperparameters that minimized this criterion over the candidate grids, $\theta \in \{1, 2, \ldots, 10\}$ and $\gamma \in \{1, 2, \ldots, 10, 20, \ldots, 100\} \times 10^{-5}$.

Results: Table 4 summarizes the results. At coarse-to-intermediate bin widths ($\Delta \in \{0.2, 0.16\}$), $\texttt{CGH}_{\texttt{rkhs}}$ achieves significantly higher predictive performance ll than the nonparametric INAR, whereas the difference is no longer statistically significant at finer resolutions ($\Delta \in \{0.1, 0.05\}$). Figure F4 shows that $\texttt{CGH}_{\texttt{rkhs}}$ yields higher triggering-kernel estimates at short lags than INAR. These findings support the benefits of our CGHP-based formulation, particularly its robustness to coarse temporal aggregation. Notably, its win rate ρ remains high even at finer resolutions, suggesting a potential practical advantage across a broad range of aggregation scales.

Interestingly, the parametric baseline $\texttt{CGH}_{\texttt{exp}}$ performs better with larger bins. One possible explanation is that coarse aggregation obscures fine-scale kernel structure, potentially leading to an underestimation of the exponential decay rate. The triggering patterns in Figure F4 suggest longer-tailed dynamics than those captured by the exponential model, and a slower fitted decay may therefore partially compensate for the misspecified exponential form over the observed lag range, resulting in better predictive performance at coarser resolutions.

5 Conclusion

We presented an RKHS-based nonparametric extension of coarse-grained Hawkes processes for estimating triggering kernels from aggregated event data. We established a novel representer theorem to reduce the infinite-dimensional optimization problem to a finite-dimensional one, enabling efficient estimation while retaining intra-bin interactions. Experiments on both synthetic and real-world data demonstrated that the proposed method achieved improved triggering-kernel recovery and predictive performance compared to INAR under coarse temporal aggregation. These results highlight the value of combining nonparametric flexibility with aggregation-aware modeling.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods and assist with translation. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. Additionally, we used generative AI tools to edit a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed implementation instructions and reproducibility guidelines in Sections 4 and E, and the Python code to reproduce the results is available[1].

REFERENCES

L. Adamopoulos. Cluster models for earthquakes: Regional comparisons. *Journal of the International Association for Mathematical Geology*, 8:463–475, 1976.

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

Emmanuel Bacry, Martin Bompaire, Stéphane Gaïffas, and Soren Poulsen. Tick: A python library for statistical learning, with a particular emphasis on time-dependent modelling. *arXiv preprint arXiv:1707.03003*, 2017.

Anna Bonnet and Maxime Sangnier. Nonparametric estimation of Hawkes processes with RKHSs. In *Artificial Intelligence and Statistics*, pp. 3574–3582. PMLR, 2025.

Anna Bonnet, Miguel Martinez Herrera, and Maxime Sangnier. Inference of multivariate exponential Hawkes processes with inhibition and application to neuronal activity. *Statistics and Computing*, 33(4):91, 2023.

Pierre Brémaud and Laurent Massoulié. Stability of nonlinear Hawkes processes. *The Annals of Probability*, pp. 1563–1588, 1996.

Feng Chen, Tsz-Kit Jeffrey Kwan, and Tom Stindl. Estimating the Hawkes process from a discretely observed sample path. *Journal of Computational and Graphical Statistics*, 34(4):1419–1431, 2025.

Felix Cheysson and Gabriel Lang. Spectral estimation of Hawkes processes from count data. *The Annals of Statistics*, 50(3):1722–1746, 2022.

E.S. Chornoboy, L.P. Schramm, and A.F. Karr. Maximum likelihood identification of neural point process systems. *Biological Cybernetics*, 59(4):265–275, 1988.

James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

Peter F. Halpin. A scalable EM algorithm for Hawkes processes. In *New Developments in Quantitative Psychology: Presentations from the 77th Annual Psychometric Society Meeting*, pp. 403–414. Springer, 2013.

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

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

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

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pp. 65–70, 1979.

Hideaki Kim and Tomoharu Iwata. A representer theorem for Hawkes processes via penalized least squares minimization. In *International Conference on Learning Representations*, volume 2026, pp. 2780–2803, 2026.

Matthias Kirchner. Hawkes and INAR (∞) processes. *Stochastic Processes and their Applications*, 126(8):2494–2525, 2016.

Matthias Kirchner. An estimation procedure for the Hawkes process. *Quantitative Finance*, 17(4): 571–595, 2017.

Shinsuke Koyama. Coarse-grained Hawkes processes. *Entropy*, 27(6):555, 2025.

George O. Mohler, Martin B. Short, P. Jeffrey Brantingham, Frederic Paik Schoenberg, and George E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

Marian-Andrei Rizoiu, Alexander Soen, Shidi Li, Pio Calderon, Leanne J. Dong, Aditya Krishna Menon, and Lexing Xie. Interval-censored Hawkes processes. *Journal of Machine Learning Research*, 23(338):1–84, 2022.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2018.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Leigh Shlomovich, Edward A.K. Cohen, and Niall Adams. A parameter estimation method for multivariate binned Hawkes processes. *Statistics and Computing*, 32(6):98, 2022a.

Leigh Shlomovich, Edward A.K. Cohen, Niall Adams, and Lekha Patel. Parameter estimation of binned Hawkes processes. *Journal of Computational and Graphical Statistics*, 31(4):990–1000, 2022b.

UK Police. ASB incidents, crime and outcomes. data.police.uk, 2026. URL `https://data.police.uk/about/`. Accessed: September 8, 2026.

WHO. Weekly COVID-19 cases and deaths by date reported to WHO. WHO COVID-19 Dashboard, 2023. URL `https://data.who.int/dashboards/covid19/data`. Accessed: September 8, 2026.

Herman J. Woltring. A fortran package for generalized, cross-validatory spline smoothing and differentiation. *Advances in Engineering Software (1978)*, 8(2):104–113, 1986.

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

Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1513–1522, 2015.

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

Lingxiao Zhou and Georgia Papadogeorgou. Bayesian inference for aggregated Hawkes processes. *Bayesian Analysis*, 21(3):1247–1275, 2026.

A Proof of Theorem 1

Proof. By Mercer's theorem, the RKHS kernel $k(\cdot,\cdot)$ admits the following Mercer expansion:

$$k(v,s)=\sum_{m=1}^{\infty} e_m(v)e_m(s), \quad \int_{\mathcal{T}} e_m(s)e_{m'}(s)ds=\eta_m\delta_{mm'}, \tag{19}$$

where $\{e_m(\cdot)\}_{m=1}^{\infty}$ and $\{\eta_m\}_{m=1}^{\infty}$ denote the eigenfunctions and the eigenvalues, respectively, of the integral operator $(T_k f)(v)=\int_{\mathcal{T}} k(v,s)f(s)\,ds$. Accordingly, the untruncated triggering kernels and their squared norms in $\mathcal{H}_k$ can be represented as

$$\psi_{ij}(v)=\sum_{m=1}^{\infty} b_m^{ij}\, e_m(v), \quad ||\psi_{ij}||^2_{\mathcal{H}_k}=\sum_{m=1}^{\infty}(b_m^{ij})^2, \quad 1\le i,j\le d, \tag{20}$$

where $\boldsymbol{b}=\{b_m^{ij}\in\mathbb{R}\}$ denotes the set of expansion coefficients. Using this representation, the optimization problem (10) can be reformulated as

$$\hat{\boldsymbol{b}}=\arg\min_{\boldsymbol{b}}\left[L_{\mathrm{CG}}\big(\{\boldsymbol{\Phi}_l(\boldsymbol{b})\}\big)+\frac{1}{\gamma}\sum_{i,j,m}(b_m^{ij})^2\right], \tag{21}$$

where $[\boldsymbol{\Phi}_l(\boldsymbol{b})]_{ij}=\sum_m b_m^{ij}\int_{(l-1)\Delta}^{(l+1)\Delta} e_m(s)\big(1-|l-s/\Delta|\big)\mathbf{1}_{0<s\le R}\,ds$. The optimal coefficient tensor $\hat{\boldsymbol{b}}$ satisfies the following stationarity condition:

$$\frac{\partial}{\partial b_m^{ij}}\left[L_{\mathrm{CG}}\big(\{\boldsymbol{\Phi}_l\}\big)+\frac{1}{\gamma}\sum_{i,j,m}(b_m^{ij})^2\right]\Bigg|_{\boldsymbol{b}=\hat{\boldsymbol{b}}}=\sum_{l=0}^{\overline{n}-1} c_l^{ij}\frac{\partial[\boldsymbol{\Phi}_l]_{ij}}{\partial b_m^{ij}}+\frac{2}{\gamma}b_m^{ij}\Bigg|_{\boldsymbol{b}=\hat{\boldsymbol{b}}}=0, \quad c_l^{ij}=\frac{\partial L_{\mathrm{CG}}}{\partial[\boldsymbol{\Phi}_l]_{ij}}, \tag{22}$$

where c_l^{ij} is a scalar constant and we use the relation (9). Applying $\sum_{m=1}^{\infty}\cdot\, e_m(v)$ to both sides of Eq. (22) yields the representation in terms of RKHS kernels in Eq. (11), where we use the relation (20) and the kernel trick, $k(v,s)=\sum_m e_m(v)e_m(s)$. This completes the proof. ■

B Proof of Corollary 2

Proof. Substituting the representation (11) into Eq. (5) yields the relation (12). Through Mercer expansion (19), the representation (11) can be expressed in terms of the eigenfunctions. Comparing this expansion with (20) yields the following relation between $\{b_m^{ij}\}$ and $\{\alpha_l^{ij}\}$: $b_m^{ij}=\sum_l \alpha_l^{ij}\int_{(l-1)\Delta}^{(l+1)\Delta} e_m(s)\big(1-|l-s/\Delta|\big)\mathbf{1}_{0<s\le R}\,ds$. Using this relation, the regularization term in Eq. (21) can be rewritten in terms of $\mathbf{B}$ as follows:

$$\begin{aligned}\sum_{i,j,m}(b_m^{ij})^2&=\sum_{i,j,l,l'}\alpha_l^{ij}\alpha_{l'}^{ij}\int_{(l-1)\Delta}^{(l+1)\Delta}\int_{(l'-1)\Delta}^{(l'+1)\Delta}k(v,s)\big(1-|l-v/\Delta|\big)\big(1-|l'-s/\Delta|\big)\mathbf{1}_{0<v\le R}\mathbf{1}_{0<s\le R}\,dvds\\&=\sum_{i,j,l,l'}\alpha_l^{ij}\alpha_{l'}^{ij}[\boldsymbol{H}]_{ll'}=\sum_{i,j}\mathbf{A}_{ij:}^{\top}\boldsymbol{H}\mathbf{A}_{ij:}=\sum_{i,j}\mathbf{B}_{ij:}^{\top}\boldsymbol{H}^{-1}\mathbf{B}_{ij:},\end{aligned}$$

where we use the kernel trick, $k(v,s)=\sum_m e_m(v)e_m(s)$, and the relation (12). The loss function term in Eq. (21) can likewise be expressed in terms of $\mathbf{B}$ through the relation $\boldsymbol{\Phi}_{l-1}=\mathbf{B}_{::l}$. Consequently, the objective in Eq.(21) can be rewritten as Eq.(13), which completes the proof. ■

C Extension of Corollary 2 to the Multi-trial Setting

The CGHP framework of Koyama (2025) and the corresponding formulation in Corollary 2 assume data from a single trial. In practice, however, observations from multiple trials are often available. Here, we extend both CGHP and our proposed model to the simplest multi-trial setting, in which all trials share the same observation period and bin boundaries

Let $(0,T]$ denote the observation period, n the number of bins for each trial, $\Delta=T/n$ the bin width, and M the number of trials. The aggregated event data, $\{\{\boldsymbol{X}_{k,m}\}_{k=1}^{n}\}_{m=1}^{M}$, are then defined as

$$\boldsymbol{X}_{k,m}=(X_{k,m}^1,\ldots,X_{k,m}^d)^{\top}\in\mathbb{R}^d, \quad X_{k,m}^i=N_i^m(k\Delta)-N_i^m((k-1)\Delta),$$

where $(N_1^m(t), \ldots, N_d^m(t))$ denotes a d-dimensional counting process for the m-th trial, and $X_{k,m}^i$ denotes the number of events of dimension i for the trial observed in the k-th bin, $((k-1)\Delta, k\Delta]$. The CGHP loss function in Eq. (6) can be extended as

$$L_{\text{CG}} = \sum_{m=1}^{M}\sum_{k=1}^{n}\Big(\boldsymbol{\Upsilon}_{k,m} - \sum_{l=0}^{k-1}\boldsymbol{\Phi}_l\boldsymbol{\Upsilon}_{k-l,m}\Big)^{\top}\boldsymbol{\Lambda}^{-1}\Big(\boldsymbol{\Upsilon}_{k,m} - \sum_{l=0}^{k-1}\boldsymbol{\Phi}_l\boldsymbol{\Upsilon}_{k-l,m}\Big) - 2nM\log|\boldsymbol{I}_d - \boldsymbol{\Phi}_0|,$$

where $\hat{\boldsymbol{\lambda}} = (nM)^{-1}\sum_{m=1}^{M}\sum_{k=1}^{n}\boldsymbol{X}_{k,m}$ is the empirical mean of the event counts, $\boldsymbol{\Upsilon}_{k,m} = \boldsymbol{X}_{k,m} - \hat{\boldsymbol{\lambda}}$ is the centered count data, and $\boldsymbol{\Lambda} = \text{diag}(\hat{\boldsymbol{\lambda}})$. Accordingly, Theorem 1 remains valid without modification, while the objective in Corollary 2 for determining the corresponding dual coefficients is replaced by the following:

$$\begin{aligned}\sum_{m=1}^{M}\sum_{k=1}^{n}\Big(\boldsymbol{\Upsilon}_{k,m} - \sum_{l=1}^{\overline{k}}\mathbf{B}_{::l}\boldsymbol{\Upsilon}_{k-l+1,m}\Big)^{\top}\boldsymbol{\Lambda}^{-1}\Big(\boldsymbol{\Upsilon}_{k,m} - \sum_{l=1}^{\overline{k}}\mathbf{B}_{::l}\boldsymbol{\Upsilon}_{k-l+1,m}\Big)\\ - 2nM\log|\boldsymbol{I}_d - \mathbf{B}_{::1}| + \frac{1}{\gamma}\sum_{i,j}\mathbf{B}_{ij:}^{\top}\boldsymbol{H}^{-1}\mathbf{B}_{ij:}\,.\end{aligned} \tag{23}$$

D Proof of Proposition 3

Proof. For a shift-invariant RKHS kernel of the form $k(v,s) = k(\theta|v-s|)$, the integrated kernel functions $\{h_l(\cdot)\}_{l=0}^{\overline{n}-1}$ in Eq. (11) admit a representation in terms of the first-order moment function $F_\eta^1(v) = \int_0^1 s\,k(\eta|v-s|)ds$:

$$\begin{aligned}
h_l(v) &= \int_{(l-1)\Delta}^{(l+1)\Delta} k\big(\theta\big|v-s\big|\big)\big(1-\big|l-s/\Delta\big|\big)\mathbf{1}_{0<s\le R}\,ds\\
&= \Delta\int_{-1}^{1} k\big(\Delta\theta\big|v/\Delta - l - u\big|\big)\big(1-|u|\big)\mathbf{1}_{0<u+l\le n^*}du \qquad (u = s/\Delta - l,\ n^* = R/\Delta)\\
&= \Delta\bigg[\int_{-1}^{0}(1+u) + \int_{0}^{1}(1-u)\bigg]k\big(\Delta\theta\big|v/\Delta - l - u\big|\big)\mathbf{1}_{0<u+l\le n^*}du\\
&= \Delta\bigg[\int_{-1}^{0}(1+u) - \int_{1}^{0}(1-u)\bigg]k\big(\Delta\theta\big|v/\Delta - l - u\big|\big)\mathbf{1}_{0<u+l\le n^*}du\\
&= \Delta\int_0^1 u\,\Big[k\big(\Delta\theta\big|v/\Delta + 1 - l - u\big|\big)\mathbf{1}_{0<u+l-1\le n^*}\\
&\qquad\qquad + k\big(\Delta\theta\big|v/\Delta - 1 - l + u\big|\big)\mathbf{1}_{-n^*\le u-l-1<0}\Big]du\\
&= \Delta\Big[F_{\Delta\theta}^1\big(1-l+v/\Delta\big)\mathbf{1}_{0<l\le n^*} + F_{\Delta\theta}^1\big(1+l-v/\Delta\big)\mathbf{1}_{0\le l<n^*}\Big], \quad 0\le l\le \overline{n}-1,
\end{aligned}$$

where the final equality follows from the assumption that $n^* = R/\Delta$ is a positive integer and the symmetry identity $k\big(\Delta\theta\big|v/\Delta - 1 - l + u\big|\big) = k\big(\Delta\theta\big|-v/\Delta + 1 + l - u\big|\big)$.

We next evaluate the matrix $\boldsymbol{H}$, whose entries are weighted double integrals of the RKHS kernel. Rescaling the integration variables and splitting each interval at zero yields a decomposition into four double integrals over unit squares:

$$\begin{aligned}
[\boldsymbol{H}]_{(l+1)(l'+1)} &= \int_{(l-1)\Delta}^{(l+1)\Delta}\int_{(l'-1)\Delta}^{(l'+1)\Delta} k\big(\theta\big|v-s\big|\big)\big(1-\big|l-v/\Delta\big|\big)\big(1-\big|l'-s/\Delta\big|\big)\mathbf{1}_{0<v\le R}\mathbf{1}_{0<s\le R}\,dv\,ds\\
&= \Delta^2\int_{-1}^{1} dv\,\big(1-|v|\big)\int_{-1}^{1} ds\,\big(1-|s|\big)\,k\big(\Delta\theta\big|v-s+l-l'\big|\big)\mathbf{1}_{0<v+l\le n^*}\mathbf{1}_{0<s+l'\le n^*}\\
&\qquad\qquad (s/\Delta - l' \to s,\ v/\Delta - l \to v)\\
&= \Delta^2\bigg\{\int_{-1}^{0} dv\,(1+v) + \int_{0}^{1} dv\,(1-v)\bigg\}\bigg\{\int_{-1}^{0} ds\,(1+s) + \int_{0}^{1} ds\,(1-s)\bigg\}\\
&\qquad\qquad \times k\big(\Delta\theta\big|v-s+l-l'\big|\big)\mathbf{1}_{0<v+l\le n^*}\mathbf{1}_{0<s+l'\le n^*}.
\end{aligned}$$

Further changes of variables reduce these four integrals to two canonical forms over the unit square:

$$
\begin{aligned}
&\Delta^{-2}[\boldsymbol{H}]_{(l+1)(l'+1)} \\
&= \int_{-1}^{0} dv \int_{-1}^{0} ds\ (1+v)(1+s) k\big(\Delta\theta\big|v-s+l-l'\big|\big)\mathbf{1}_{0<v+l\le n^*}\mathbf{1}_{0<s+l'\le n^*} \\
&+ \int_{0}^{1} dv \int_{0}^{1} ds\ (1-v)(1-s) k\big(\Delta\theta\big|v-s+l-l'\big|\big)\mathbf{1}_{0<v+l\le n^*}\mathbf{1}_{0<s+l'\le n^*} \\
&+ \int_{-1}^{0} dv \int_{0}^{1} ds\ (1+v)(1-s) k\big(\Delta\theta\big|v-s+l-l'\big|\big)\mathbf{1}_{0<v+l\le n^*}\mathbf{1}_{0<s+l'\le n^*} \\
&+ \int_{0}^{1} dv \int_{-1}^{0} ds\ (1-v)(1+s) k\big(\Delta\theta\big|v-s+l-l'\big|\big)\mathbf{1}_{0<v+l\le n^*}\mathbf{1}_{0<s+l'\le n^*} \\
&= \int_{0}^{1} dv \int_{0}^{1} ds\ vs\ k\big(\Delta\theta\big|v-s+l-l'\big|\big)\mathbf{1}_{0<v+l-1\le n^*}\mathbf{1}_{0<s+l'-1\le n^*} \qquad (1+v\to v,\ 1+s\to s) \\
&+ \int_{0}^{1} dv \int_{0}^{1} ds\ vs\ k\big(\Delta\theta\big|v-s+l'-l\big|\big)\mathbf{1}_{-n^*\le v-l-1<0}\mathbf{1}_{-n^*\le s-l'-1<0} \qquad (1-v\to v,\ 1-s\to s) \\
&+ \int_{0}^{1} dv \int_{0}^{1} ds\ vs\ k\big(\Delta\theta\big|v+s+l-l'-2\big|\big)\mathbf{1}_{0<v+l-1\le n^*}\mathbf{1}_{-n^*\le s-l'-1<0} \quad (1+v\to v,\ 1-s\to s) \\
&+ \int_{0}^{1} dv \int_{0}^{1} ds\ vs\ k\big(\Delta\theta\big|v+s+l'-l-2\big|\big)\mathbf{1}_{-n^*\le v-l-1<0}\mathbf{1}_{0<s+l'-1\le n^*} \quad (1-v\to v,\ 1+s\to s) \\
&= \mathbf{1}_{0<l\le n^*}\mathbf{1}_{0<l'\le n^*} Z_0(l-l') + \mathbf{1}_{0\le l<n^*}\mathbf{1}_{0\le l'<n^*} Z_0(l'-l) \\
&\qquad + \mathbf{1}_{0<l\le n^*}\mathbf{1}_{0\le l'<n^*} Z_2(l'-l) + \mathbf{1}_{0\le l<n^*}\mathbf{1}_{0<l'\le n^*} Z_2(l-l'),
\end{aligned}
\tag{24}
$$

where

$$
\begin{aligned}
Z_0(l-l') &= \int_0^1 dv \int_0^1 ds\ vs\ k\big(\Delta\theta\big|v-s+l-l'\big|\big), \\
Z_2(l-l') &= \int_0^1 dv \int_0^1 ds\ vs\ k\big(\Delta\theta\big|v+s+l'-l-2\big|\big).
\end{aligned}
\tag{25}
$$

The first integral in Eq. (25) satisfies the symmetry relation

$$
\begin{aligned}
Z_0(l-l') &= \int_0^1 dv \int_0^1 ds\ vs\ k\big(\Delta\theta\big|-(v-s+l-l')\big|\big) \\
&= \int_0^1 ds \int_0^1 dv\ sv\ k\big(\Delta\theta\big|s-v+l'-l\big|\big) = Z_0(l'-l),
\end{aligned}
\tag{26}
$$

which allows Eq. (24) to be rewritten as

$$
\begin{aligned}
[\boldsymbol{H}]_{(l+1)(l'+1)} = \Delta^2\Big[&\big(\mathbf{1}_{0<l\le n^*}\mathbf{1}_{0<l'\le n^*} + \mathbf{1}_{0\le l<n^*}\mathbf{1}_{0\le l'<n^*}\big) Z_0(l-l') \\
&+ \mathbf{1}_{0<l\le n^*}\mathbf{1}_{0\le l'<n^*} Z_2(l'-l) + \mathbf{1}_{0\le l<n^*}\mathbf{1}_{0<l'\le n^*} Z_2(l-l')\Big]
\end{aligned}
\tag{27}
$$

Reducing the double integral $Z_0(l-l')$ to a one-dimensional integral yields an expression in terms of the moment functions $\{F^i_\eta(v) = \int_0^1 s^i k(\eta|v-s|)ds\}_{i=0}^3$:

$$
\begin{aligned}
Z_0(l-l') &= \int_{-1}^{1} dy \left[\frac{y}{2}s^2 + \frac{s^3}{3}\right]_{\max(0,-y)}^{\min(1,1-y)} k\big(\Delta\theta\big|y+l-l'\big|\big) \\
&= \left\{\int_{-1}^{0}\left(\frac{1}{3}+\frac{1}{2}y-\frac{1}{6}y^3\right) + \int_0^1\left(\frac{1}{3}-\frac{1}{2}y+\frac{1}{6}y^3\right)\right\} k\big(\Delta\theta\big|y+l-l'\big|\big)dy \\
&= \int_0^1 \left(\frac{1}{3}-\frac{1}{2}y+\frac{1}{6}y^3\right)\Big(k\big(\Delta\theta|y+l'-l|\big)+k\big(\Delta\theta|y+l-l'|\big)\Big)dy \\
&= Z_1(l-l') + Z_1(l'-l), \\
Z_1(l-l') &= \frac{1}{3}F^0_{\Delta\theta}(l-l') - \frac{1}{2}F^1_{\Delta\theta}(l-l') + \frac{1}{6}F^3_{\Delta\theta}(l-l').
\end{aligned}
\tag{28}
$$

Table D1: Examples of moment functions for RKHS kernels: $\{F_\eta^i(z) = \int_0^1 s^i k(\eta|z-s|)ds\}_{i=0}^3$. $\mathrm{erf}(z)$ and $\mathrm{sgn}(z)$ denote the error function and the sign function, respectively.

$k(\eta\|z\|)$	Gaussian RKHS kernel: $e^{-(\eta\|z\|)^2}$	Laplace RKHS kernel: $e^{-\eta\|z\|}$
$F_\eta^0(z)$	$\frac{\sqrt{\pi}}{2\eta}\Big(\mathrm{erf}(z_0)+\mathrm{erf}(z_1)\Big)$	$\frac{1}{\eta}\Big[\mathrm{sgn}(z_0)\Big(1-e^{-\|z_0\|}\Big)+\mathrm{sgn}(z_1)\Big(1-e^{-\|z_1\|}\Big)\Big]$
$F_\eta^1(z)$	$\frac{\sqrt{\pi}z_0}{2\eta^2}\Big(\mathrm{erf}(z_0)+\mathrm{erf}(z_1)\Big) + \frac{1}{2\eta^2}\Big(e^{-\|z_0\|^2}-e^{-\|z_1\|^2}\Big)$	$\frac{z_0}{\eta^2}\Big(\mathrm{sgn}(z_0)+\mathrm{sgn}(z_1)\Big)+\frac{1}{\eta^2}e^{-\|z_0\|} - \frac{1}{\eta^2}\Big(1+\eta\,\mathrm{sgn}(z_1)\Big)e^{-\|z_1\|}$
$F_\eta^2(z)$	$\frac{\sqrt{\pi}(1+2\|z_0\|^2)}{4\eta^3}\Big(\mathrm{erf}(z_0)+\mathrm{erf}(z_1)\Big) + \frac{1}{2\eta^3}\Big[z_0e^{-\|z_0\|^2}-\big(2z_0+z_1\big)e^{-\|z_1\|^2}\Big]$	$\frac{2+\|z_0\|^2}{\eta^3}\Big(\mathrm{sgn}(z_0)+\mathrm{sgn}(z_1)\Big)-\frac{2}{\eta^3}\,\mathrm{sgn}(z_0)e^{-\|z_0\|} - \frac{1}{\eta^3}\Big[2\eta+\big(2+\eta^2\big)\,\mathrm{sgn}(z_1)\Big]e^{-\|z_1\|}$
$F_\eta^3(z)$	$\frac{\sqrt{\pi}z_0(2\|z_0\|^2+3)}{4\eta^4}\Big(\mathrm{erf}(z_0)+\mathrm{erf}(z_1)\Big) + \frac{1}{2\eta^4}\Big[\big(\|z_0\|^2+1\big)e^{-\|z_0\|^2} - \big(3\|z_0\|^2+3z_0z_1+\|z_1\|^2+1\big)e^{-\|z_1\|^2}\Big]$	$\frac{6z_0+\|z_0\|^3}{\eta^4}\Big(\mathrm{sgn}(z_0)+\mathrm{sgn}(z_1)\Big)+\frac{6}{\eta^4}e^{-\|z_0\|} - \frac{1}{\eta^4}\Big[6+3\eta^2+\eta\big(6+\eta^2\big)\,\mathrm{sgn}(z_1)\Big]e^{-\|z_1\|}$

$z_0 = \eta z, \quad z_1 = \eta(1-z)$

Similarly, the double integral $Z_2(l-l')$ can be reduced to a one-dimensional integral and expressed in terms of the moment functions:

$$\begin{aligned} Z_2(l-l') &= \int_0^2 dy \left[\frac{y}{2}s^2-\frac{s^3}{3}\right]_{\max(0,y-1)}^{\min(1,y)} k\big(\Delta\theta\big|y+l'-l-2\big|\big) \\ &= \left\{\int_0^1 \frac{1}{6}y^3+\int_1^2\left(-\frac{2}{3}+y-\frac{1}{6}y^3\right)\right\}k\big(\Delta\theta\big|y+l'-l-2\big|\big)dy \\ &= \frac{1}{6}F_{\Delta\theta}^0(l-l'+1)+\frac{1}{2}F_{\Delta\theta}^1(l-l'+1)-\frac{1}{2}F_{\Delta\theta}^2(l-l'+1) \\ &\qquad + \frac{1}{6}\Big[F_{\Delta\theta}^3(l-l'+2)-F_{\Delta\theta}^3(l-l'+1)\Big]. \end{aligned} \tag{29}$$

Substituting Eqs. (28) and (29) into Eq. (27) yields Eq. (14), completing the proof. ■

E Settings of Experiments

E.1 Exponential Scenario

We generated event sequences from 3-variate Hawkes processes (1) with unit baseline intensity, $\{\mu_i = 1\}_{i=1}^3$, and exponential triggering kernels,

$$\phi_{ij}(v) = \eta_{ij}\xi_{ij}e^{-\xi_{ij}v}, \qquad 1 \le i,j \le 3.$$

For each trial, the decay parameters, $\{\xi_{ij}\}$, were uniformly sampled from $\{0.2, 1.0, 2.0\}$, and the adjacency matrix, $[\eta_{ij}]_{ij}$, was randomly generated so that the resulting matrix was entrywise positive and had the spectral radius less than one. The support window was fixed at $R = 10$ for both $\texttt{CGH}_{\texttt{rkhs}}$ and `INAR`.

E.2 Gamma-mixture Scenario

We generated event sequences from 3-variate Hawkes processes (1) with unit baseline intensity, $\{\mu_i = 1\}_{i=1}^3$, and 2-component gamma-mixture triggering kernels,

$$\phi_{ij}(v) = \eta_{ij}\sum_{q=1}^2 r_{ij}^q g(v|\xi_{ij}^q,\kappa_{ij}^q), \quad g(v|\xi,\kappa) = \xi^\kappa\frac{v^{\kappa-1}e^{-\xi v}}{\Gamma(\kappa)}, \quad \sum_{q=1}^2 r_{ij}^q = 1, \quad 1 \le i,j \le 3.$$

For each trial, the decay and shape parameters, $\{\xi_{ij}^q\}$ and $\{\kappa_{ij}^q\}$, were uniformly sampled from $\{5, 10\}$, the mixture ratio $(r_{ij}^1 \geq 0, r_{ij}^2 \geq 0)$ was sampled from the uniform Dirichlet distribution, and the adjacency matrix $[\eta_{ij}]_{ij}$ was randomly generated so that the resulting matrix was entrywise positive and had the spectral radius less than one. The support window was fixed at $R = 10$ for $\texttt{CGH}_{\texttt{rkhs}}$ and `INAR`.

E.3 Refractory Scenario

Kim & Iwata (2026) considered synthetic data generated from a 3-variate nonlinear Hawkes process (Brémaud & Massoulié, 1996) with short-term self-inhibition,

$$\begin{aligned} \phi_{11}(s) &= (8s^2 - 1)\mathbf{1}_{s \leq 0.5} + e^{-2.5(s-0.5)}\mathbf{1}_{s>0.5}, \\ \phi_{22}(s) &= \phi_{33}(s) = (8s^2 - 1)\mathbf{1}_{s \leq 0.5} + e^{-(s-0.5)}\mathbf{1}_{s>0.5}, \end{aligned} \tag{30}$$

and various non-inhibitory cross-component interactions,

$$\begin{array}{lll} \phi_{12}(s) = 0.6e^{-10(s-1)^2} & \phi_{13}(s) = 0.8e^{-20(s-3)^2} & \phi_{21}(s) = 0.6 \cdot 2^{-5s} \\ \phi_{23}(s) = 0.8e^{-20(s-2)^2} & \phi_{31}(s) = 0 & \phi_{32}(s) = 0 \end{array} . \tag{31}$$

The baseline intensity was set as $\{\mu_i = 0.01\}_{i=1}^3$, and the support window as $R = 5$. The softplus function was adopted as the link function for the nonlinear Hawkes process.

E.4 SNS Data

A large-scale SNS dataset (Zhao et al., 2015) consists of timestamps of original tweets and their associated retweet sequences, together with the follower counts of each retweeting user. For each original tweet, we extracted the retweets that occurred between 2 and 22 hours after its publication. We retained retweet sequences containing between 100 and 110 retweets, with at least one additional retweet occurring after the observation period. Of the 2,490 sequences that satisfied these criteria, we randomly selected 1,000 for our experiments, yielding a dataset comprising 104,833 events across 1,000 sequences. We assigned users to three categories ($d = 3$) based on the empirical distribution of follower counts: the bottom 30% were labeled *small*, the middle 40% *middle*, and the top 30% *large*. We excluded the first two hours because the average retweet activity exhibited pronounced temporal variation during this initial period (see Figure 1 in (Zhao et al., 2015)). Removing this early transient makes the retained observation window more consistent with the constant-baseline assumption underlying the three models being compared. For both $\texttt{CGH}_{\texttt{rkhs}}$ and `INAR`, the support window was set to $R = 1.92$ hours when $\Delta = 0.16$ hours and to $R = 2$ hours otherwise.

F EXAMPLES OF ESTIMATED TRIGGERING KERNELS

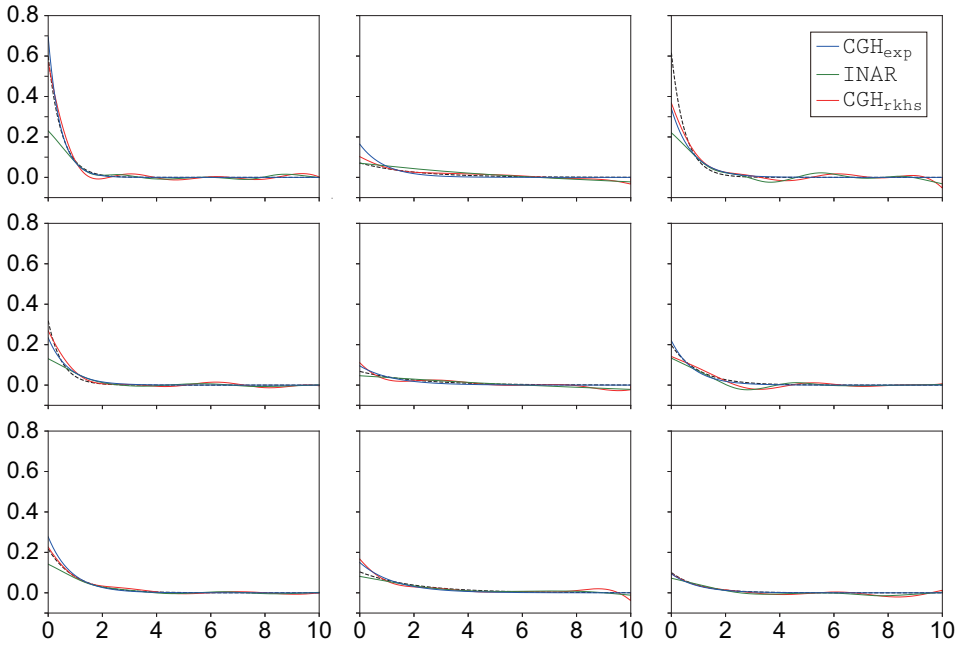

Figure F1: Examples of the estimated triggering kernels in the exponential scenario ($T = 10^4$, $\Delta = 1.0$). Dashed lines represent the true triggering kernels.

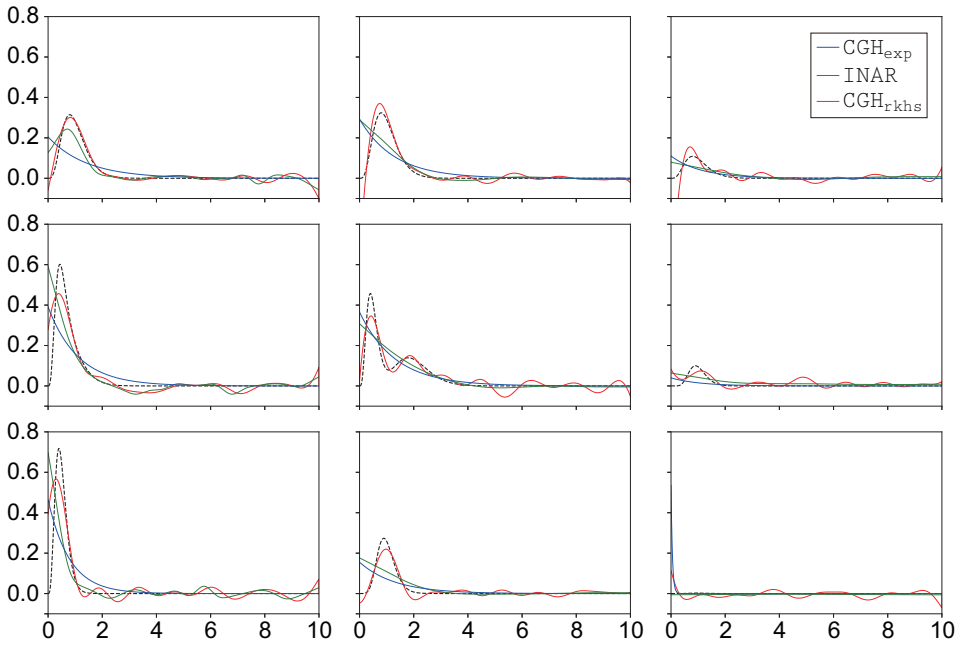

Figure F2: Examples of the estimated triggering kernels in the gamma-mixture scenario ($T = 10^4$, $\Delta = 0.5$). Dashed lines represent the true triggering kernels.

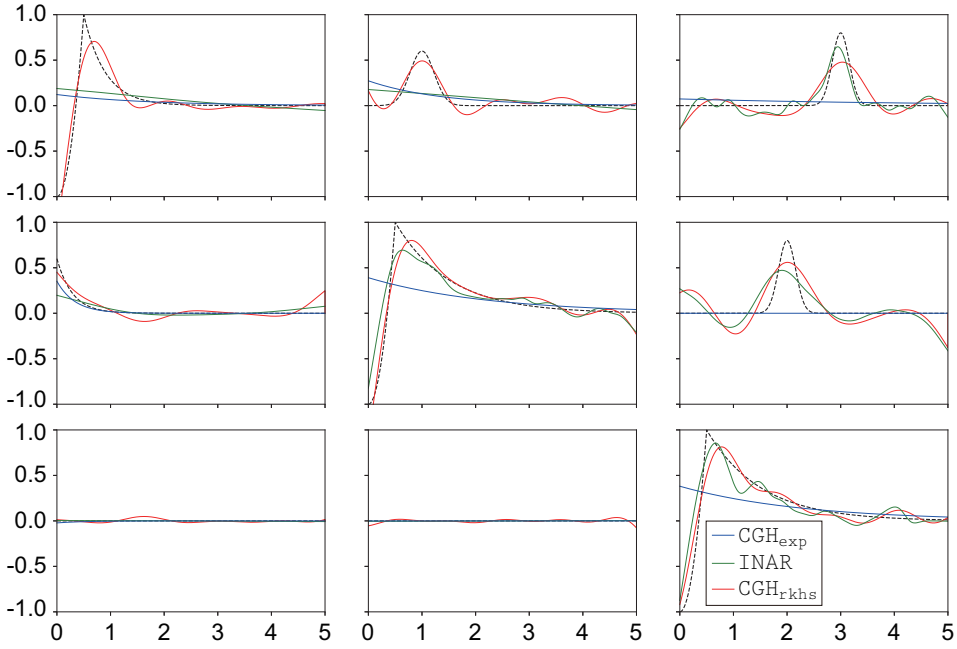

Figure F3: Examples of the estimated triggering kernels in the refractory scenario ($T = 10^4$, $\Delta = 0.2$). Dashed lines represent the true triggering kernels.

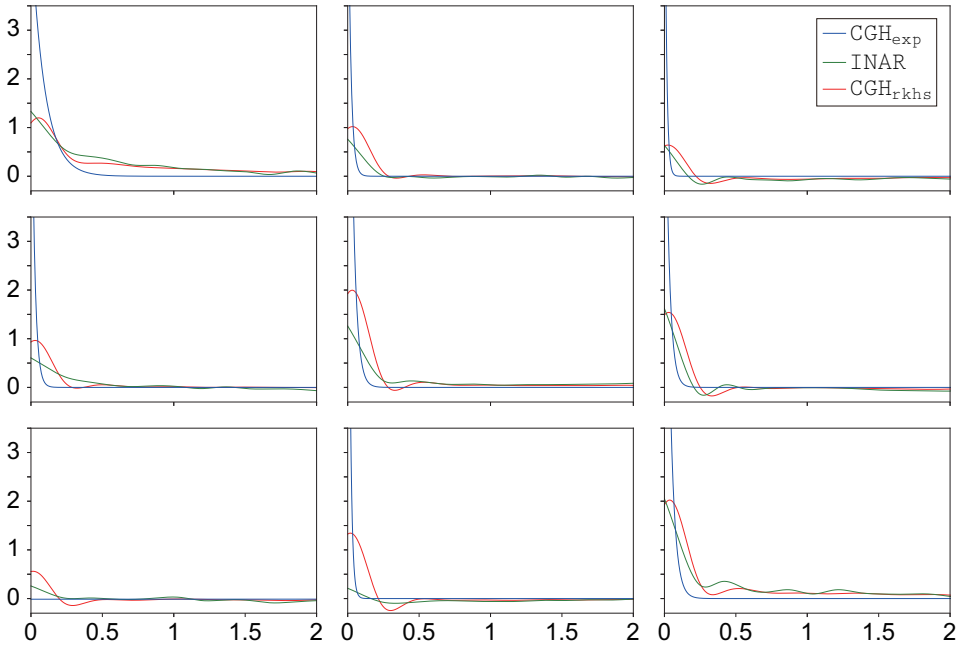

Figure F4: Examples of the estimated triggering kernels in the SNS data ($\Delta = 0.16$ h).