A REPRESENTER THEOREM FOR HAWKES PROCESSES VIA PENALIZED LEAST SQUARES MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The representer theorem is a cornerstone of kernel methods, which aim to estimate latent functions in reproducing kernel Hilbert spaces (RKHSs) in a nonparametric manner. Its significance lies in converting inherently infinite-dimensional optimization problems into finite-dimensional ones over dual coefficients, thereby enabling practical and computationally tractable algorithms. In this paper, we address the problem of estimating the latent triggering kernels–functions that encode the interaction structure between events–for linear multivariate Hawkes processes based on observed event sequences within an RKHS framework. We show that, under the principle of penalized least squares minimization, a novel form of representer theorem emerges: a family of transformed kernels can be defined via a system of simultaneous integral equations, and the optimal estimator of each triggering kernel is expressed as a linear combination of these transformed kernels evaluated at the data points. Remarkably, the dual coefficients are all analytically fixed to unity, obviating the need to solve a costly optimization problem to obtain the dual coefficients. This leads to a highly efficient estimator capable of handling large-scale data more effectively than conventional nonparametric approaches. Empirical evaluations on synthetic datasets reveal that the proposed method attains competitive predictive accuracy while substantially improving computational efficiency over existing state-of-the-art kernel method-based estimators.

1 INTRODUCTION

Nonparametric estimation of latent functions remains a central topic in both theoretical and applied research, spanning domains such as signal and image processing (Liu et al., 2011; Takeda et al., 2007), system control (Liu et al., 2018), geostatistics (Chiles & Delfiner, 2012), bioinformatics (Schölkopf et al., 2004), and clinical studies (Collett, 2023). Among various nonparametric approaches, kernel methods stand out as one of the most powerful and mature frameworks. These methods enable flexible function approximation by embedding data into high-dimensional reproducing kernel Hilbert spaces (RKHSs) (Schölkopf & Smola, 2018; Shawe-Taylor & Cristianini, 2004). In classical supervised settings with i.i.d. data, the representer theorem plays a pivotal role in kernel methods. It states that the solution to a broad class of infinite-dimensional optimization problems in RKHSs admits a finite-dimensional representation: the optimal function estimator can be expressed as a linear combination of kernel functions evaluated at the training points (Schölkopf et al., 2001; Wahba, 1990). This linear form not only provides theoretical insight but also brings practical advantages in optimization and inference.

Recently, the kernel method literature has begun to address the nonparametric estimation of intensity functions in point process models. The problems are fundamentally more challenging than i.i.d. cases, primarily because the loss functions to minimize (e.g., negative log-likelihood functions) involve integrals of latent intensity functions over observation domains and violate independence assumptions, which renders classical representer theorems inapplicable. A seminal contribution by Flaxman et al. (2017) demonstrated that a representer theorem can still hold for a point process: specifically, they showed that if the square root of the intensity function lies in an RKHS, then the solution to the penalized maximum likelihood estimation problem admits a finite-dimensional representation. Interestingly, the optimal estimator is expressed not via standard RKHS kernels but via *equivalent kernels*–RKHS kernels transformed through a Fredholm integral equation. The result

has since been extended to settings with covariate-dependent intensity functions (Kim et al., 2022) and survival point processes (Kim, 2023), providing a broader foundation for kernel-based learning in point process models.

More recently, Bonnet & Sangnier (2025) addressed a more intricate setting of multivariate Hawkes processes (Brémaud & Massoulié, 1996; Hawkes, 1971), which offer a powerful framework for modeling self- and mutually-interacting event dynamics in real-world applications such as finance (Bacry et al., 2015), neuroscience (Gerhard et al., 2017), social networks (Zhou et al., 2013), and seismology (Ogata, 1988). By leveraging the approximations of both the log-likelihood and least-squares loss functions, they obtained a representer theorem for the estimation of triggering kernels in an RKHS. To ensure the non-negativity of the intensity functions, the model employs non-linear link functions, allowing it to capture both excitatory and inhibitory interactions. Although the method demonstrates strong empirical performance, it requires solving a non-linear optimization problem over dual coefficients whose dimensionality scales with the data size, posing serious scalability issues for a large scale of datasets often seen in multivariate Hawkes processes.

In this paper, we consider a kernel method-based least squares loss formulation for estimating latent triggering kernels in linear multivariate Hawkes processes, where the identity link function is assumed. By leveraging variational analysis, we establish a novel representer theorem tailored to the functional optimization problem: the obtained estimator of each triggering kernel admits a linear expansion in terms of *equivalent kernels* defined through a system of Fredholm integral equations. Notably, all dual coefficients are analytically fixed to unity, eliminating the need to solve a costly coefficient optimization problem. To the best of our knowledge, this paper is the first to establish a representer theorem for the non-approximated penalized least squares formulation of linear Hawkes processes. Furthermore, we propose an efficient algorithm to solve the integral equations using the random feature map approximation of RKHS kernels (Rahimi & Recht, 2007), where all required integrals are obtained in a closed form, in contrast to the Bonnet & Sangnier (2025) model that relies on Riemann approximation. Consequently, the proposed estimator[1] consists solely of additive matrix operations and an inversion of a matrix whose size is independent of the data size. This yields a highly lightweight and scalable estimator that remains effective even on large-scale event data, offering a practical and theoretically grounded solution for learning in multivariate Hawkes processes.

2 Proposed Method

2.1 Preliminary: Linear Hawkes Processes

We consider a multivariate linear Hawkes process (Brémaud & Massoulié, 1996; Hawkes, 1971) on a time domain $\mathbb{R}_+$, i.e., a U-dimensional counting process $(N_1(t), \dots, N_U(t))$ characterized by the following conditional intensity functions:

$$\lambda_i(t) = \mu_i + \sum_{j \in \mathcal{U}} \int_0^t g_{ij}(t-s) dN_j(s), \qquad t \in \mathbb{R}_+,\ i \in \mathcal{U} \coloneqq [\![1, U]\!], \tag{1}$$

where $\mu_i \in \mathbb{R}_+$ denotes the baseline intensity for dimension i, and $g_{ij}(t-s) : \mathbb{R}_+ \to \mathbb{R}$ is the triggering kernel quantifying the change in the dimension i's intensity at time t caused by the event of dimension j occurring at time s.

Let $\{(t_n, u_n) \in \mathbb{R}_+ \times \mathcal{U}\}_{n=1}^{N(T)}$ denote a sequence of $N(T) = \sum_{i \in \mathcal{U}} N_i(T)$ observed events over an interval $[0, T]$, where each pair (t_n, u_n) indicates that the n-th event occurred at time t_n on dimension u_n. In the literature on point processes, two common approaches have been used to estimate the intensity functions: one based on the negative log-likelihood function (Daley & Vere-Jones, 2006), and the other on the least squares contrast (Hansen et al., 2015), defined respectively as

$$L_{\mathrm{LL}} = \sum_{i \in \mathcal{U}} \Big[\int_0^T \lambda_i(t) dt - \sum_{n \in \mathcal{N}_i} \log\big(\lambda_i(t_n)\big) \Big], \quad L_{\mathrm{LS}} = \sum_{i \in \mathcal{U}} \Big[\int_0^T \lambda_i^2(t) dt - 2 \sum_{n \in \mathcal{N}_i} \lambda_i(t_n) \Big], \tag{2}$$

[1] Code and data to reproduce the results will be public at `https://github.com/****`.

where $\mathcal{N}_i = \{t_n : u_n = i\}_{n=1}^{N(T)}$ denotes a subset of event times associated with dimension i. Notably, the least squares contrast, L_{LS}, arises from the principle of empirical risk minimization (van de Geer, 2000), and has recently attracted attention due to its favorable computational properties in Hawkes process modeling (Bacry et al., 2020; Cai et al., 2024).

2.2 A Representer Theorem for Linear Hawkes Processes

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 RKHS kernel $k(\cdot, \cdot)$.

Given an observed sequence of events $\{(t_n, u_n)\}_{n=1}^{N(T)}$ over an interval $[0, T]$, we consider the following regularized optimization problem of triggering kernels, $g = \{g_{ij}(\cdot)\}_{(i,j)\in\mathcal{U}^2}$, and baseline intensities, $\mu = \{\mu_i\}_{i\in\mathcal{U}}$, in the linear Hawkes process (1):

$$\hat{g},\ \hat{\mu} \ = \ \underset{g\in\mathcal{H}_k^{U^2},\ \mu\in\mathbb{R}^U}{\arg\min} \left[L(g, \mu) + \frac{1}{\gamma} \sum_{(i,j)\in\mathcal{U}^2} ||g_{ij}||^2_{\mathcal{H}_k} \right], \tag{3}$$

where L represents the loss functional, $||\cdot||^2_{\mathcal{H}_k}$ represents the squared Hilbert space norm, and $\gamma \in \mathbb{R}_+$ represents the regularization hyperparameter. In this paper, we adopt the least squares contrast for point processes, denoted by $L_{\text{LS}}(g, \mu)$, as a loss functional, which takes a quadratic form in terms of the triggering kernels and baseline intensities,

$$\begin{aligned} L_{\text{LS}}(g, \mu) = \sum_{i\in\mathcal{U}} \Bigg[\int_0^T \Big(\mu_i + \sum_{n\in\mathcal{N}} g_{iu_n}(t - t_n) \mathbf{1}_{0<t-t_n\le A} \Big)^2 dt \\ - 2 \sum_{n'\in\mathcal{N}_i} \Big(\mu_i + \sum_{n\in\mathcal{N}} g_{iu_n}(t_{n'} - t_n) \mathbf{1}_{0<t_{n'}-t_n\le A} \Big) \Bigg], \end{aligned} \tag{4}$$

where $\mathbf{1}_{(\cdot)}$ denotes the indicator, and $\mathcal{N} = \{t_n\}_{n=1}^{N(T)}$ represents the whole event times observed. Here, a finite support window $A \in \mathbb{R}_+$ for the triggering kernels is introduced, as is commonly done in Hawkes process modeling to reduce computational cost (Bonnet et al., 2023; Halpin, 2013). Theorem 1 establishes a novel representer theorem for the functional optimization problem defined in (3-4). Notably, all dual coefficients are analytically fixed to unity, eliminating the need for their optimization. While the proof relies on the path integral representation (Kim, 2021), for completeness, we also provide an alternative derivation through Mercer's theorem (Mercer, 1909) in Appendix E.

Theorem 1. *Given the estimation of the baseline intensity $\{\hat{\mu}_i\}_{i\in\mathcal{U}}$, the solutions of the functional optimization problem (3-4), denoted as $\{\hat{g}_{ij}(\cdot)\}_{(i,j)\in\mathcal{U}^2}$, involve the representer theorem under a set of equivalent kernels[2], $\{h_j(\cdot, \cdot)\}_{j\in\mathcal{U}}$, and their dual coefficients are equal to unity:*

$$\hat{g}_{ij}(s) = \sum_{n\in\mathcal{N}_i} \alpha_n^{ij} h_j(s, t_n) - \hat{\mu}_i \int_0^T h_j(s, t) dt, \quad \alpha_n^{ij} = 1,\ \ s \in \mathcal{T},\ (i, j) \in \mathcal{U}^2, \tag{5}$$

where $\{\alpha_n^{ij}\}$ denote the dual coefficients, and the equivalent kernels $\{h_j(\cdot, \cdot)\}_{j\in\mathcal{U}}$ are defined through a system of Fredholm integral equations,

$$\begin{aligned} \frac{1}{\gamma} h_j(s, s') + \sum_{l\in\mathcal{U}} \int_0^T V_{jl}(s, t) h_l(t, s') dt = \sum_{n\in\mathcal{N}_j} k(s, s' - t_n) \mathbf{1}_{0<s'-t_n\le A}, \\ V_{jl}(s, t) = \sum_{n\in\mathcal{N}_j} \sum_{n'\in\mathcal{N}_l} k(s, t + t_{n'} - t_n) \mathbf{1}_{\max(t_n, t_{n'})<t+t_{n'}\le \min(T, A+t_n, A+t_{n'})}. \end{aligned} \tag{6}$$

Proof. Let $\mathcal{K} \cdot (s) = \int_{\mathcal{T}} \cdot\, k(s, t) dt$ be the integral operator with RKHS kernel $k(\cdot, \cdot)$, and $\mathcal{K}^* \cdot (s) = \int_{\mathcal{T}} \cdot\, k^*(s, t) dt$ be its inverse operator. Then, through the path integral representation (Kim, 2021), the squared RKHS norm term can be represented in a functional form,

$$\sum_{(i,j)\in\mathcal{U}^2} ||g_{ij}||^2_{\mathcal{H}_k} = \sum_{(i,j)\in\mathcal{U}^2} \iint_{\mathcal{T}\times\mathcal{T}} k^*(s, t) g_{ij}(s) g_{ij}(t) ds dt.$$

[2]Following Flaxman et al. (2017), we call the transformed kernel functions where a representer theorem holds the *equivalent kernels*.

Using the representation, the functional derivatives of the least squares term and the penalization term in (3), with respect to $g_{ij}(\cdot)$, can be written as follows:

$$\frac{\delta L_{\mathrm{LS}}}{\delta g_{ij}(s)} = 2\int_0^T \Big(\hat{\mu}_i + \sum_{n'\in\mathcal{N}} g_{iu_{n'}}(t-t_{n'})\mathbf{1}_{0<t-t_{n'}\le A}\Big)\sum_{n\in\mathcal{N}_j}\delta(s-(t-t_n))\mathbf{1}_{0<t-t_n\le A}dt$$
$$- 2\sum_{n'\in\mathcal{N}_i}\sum_{n\in\mathcal{N}_j}\delta(s-(t_{n'}-t_n))\mathbf{1}_{0<t_{n'}-t_n\le A},$$

$$\frac{\delta}{\delta g_{ij}(s)}\sum_{i,j}\|g_{ij}\|^2_{\mathcal{H}_k} = \frac{\delta}{\delta g_{ij}(s)}\iint_{\mathcal{T}\times\mathcal{T}} k^*(t,t')g_{ij}(t)g_{ij}(t')dtdt' = 2\int_{\mathcal{T}} k^*(s,t)g_{ij}(t)dt,$$

where $\delta(\cdot)$ denotes the Dirac delta function. The optimal estimator $\hat{g}_{ij}(\cdot)$ should solve the equation where the functional derivative of the penalized least squares contrast is equal to zero:

$$\frac{\delta}{\delta g_{ij}(s)}\left[L_{\mathrm{LS}}(g,\hat{\mu}) + \frac{1}{\gamma}\sum_{i,j}\|g_{ij}\|^2_{\mathcal{H}_k}\right]\Big|_{g=\hat{g}} = 0, \qquad s\in\mathcal{T},\ (i,j)\in\mathcal{U}^2.$$

Then applying operator $\mathcal{K}$ to the equation leads to the following simultaneous Fredholm integral equations of the second kind:

$$\begin{aligned}&\frac{1}{\gamma}\hat{g}_{ij}(s) + \sum_{l\in\mathcal{U}}\int_0^T V_{jl}(s,t)\hat{g}_{il}(t)dt \\ &\quad= \sum_{n'\in\mathcal{N}_i}\sum_{n\in\mathcal{N}_j} k(s,t_{n'}-t_n)\mathbf{1}_{0<t_{n'}-t_n\le A} - \hat{\mu}_i\sum_{n\in\mathcal{N}_j}\int_0^T k(s,t-t_n)\mathbf{1}_{0<t-t_n\le A}dt,\end{aligned} \tag{7}$$

where $V_{jl}(s,t)$ is defined in (6), the second term on the left-hand side of Equation (7) is derived using the following relation,

$$\begin{aligned}&\sum_{n'\in\mathcal{N}}\sum_{n\in\mathcal{N}_j}\int_0^T \hat{g}_{iu_{n'}}(t-t_{n'})k(s,t-t_n)\mathbf{1}_{0<t-t_n\le A}\mathbf{1}_{0<t-t_{n'}\le A}dt \\ &= \sum_{l\in\mathcal{U}}\sum_{n\in\mathcal{N}_j}\sum_{n'\in\mathcal{N}_l}\int_{-t_{n'}}^{T-t_{n'}} \hat{g}_{il}(t)k(s,t+t_{n'}-t_n)\mathbf{1}_{0<t-t_n+t_{n'}\le A}\mathbf{1}_{0<t\le A}dt \qquad (t-t_{n'}\to t) \\ &= \sum_{l\in\mathcal{U}}\int_0^T \hat{g}_{il}(t)\sum_{n\in\mathcal{N}_j}\sum_{n'\in\mathcal{N}_l} k(s,t+t_{n'}-t_n)\mathbf{1}_{\max(t_n,t_{n'})<t+t_{n'}\le\min(T,A+t_n,A+t_{n'})}\,dt,\end{aligned}$$

and the relation, $(\mathcal{K}\mathcal{K}^*)\cdot(s) = \int_{\mathcal{T}}\cdot\,\delta(s-t)dt$, was used. Equation (7) indicates that the optimal estimator $\hat{g}_{ij}(\cdot)$ admits a linear representation in terms of a set of transformed kernel functions, $\{h_{ij}(\cdot,\cdot)\}_{(i,j)\in\mathcal{U}^2}$, as

$$\hat{g}_{ij}(s) = \sum_{n\in\mathcal{N}_i} h_{ij}(s,t_n) - \hat{\mu}_i\int_0^T h_{ij}(s,t)dt, \qquad s\in\mathcal{T},\ (i,j)\in\mathcal{U}^2,$$

where the transformed kernel functions are defined by a system of simultaneous Fredholm integral equations,

$$\frac{1}{\gamma}h_{ij}(s,s') + \sum_{l\in\mathcal{U}}\int_0^T V_{jl}(s,t)h_{il}(t,s')dt = \sum_{n\in\mathcal{N}_j} k(s,s'-t_n)\mathbf{1}_{0<s'-t_n\le A}. \tag{8}$$

Since the coefficients in Equation (8) are independent of the index i, its solution $h_{ij}(\cdot,\cdot)$ is also independent of i, allowing us to write as $h_{ij}(\cdot,\cdot) = h_j(\cdot,\cdot)$. This completes the proof. ■

While Bonnet & Sangnier (2025) has also explored representer theorems under the least squares contrast for linear Hawkes processes, their formulation relies on a discretized approximation (see Proposition 1 in (Bonnet & Sangnier, 2025)), which introduces additional optimization over dual coefficients and obscures the elegant mathematical properties of the least squares contrast. To the

best of our knowledge, this work is the first to establish a representer theorem for the non-discretized penalized least squares formulation of linear Hawkes processes.

In Theorem 1, the optimal estimators of baseline intensities, $\{\hat{\mu}_i\}_{i\in\mathcal{U}}$ are treated as given constants. Proposition 2 demonstrates that by substituting Equation (5) into Equation (3), $\{\hat{\mu}_i\}_{i\in\mathcal{U}}$ can be obtained in closed form in terms of the equivalent kernels. The proof is provided in Appendix B.

Proposition 2. *The solutions, $\{\hat{\mu}_i\}_{i\in\mathcal{U}}$, of the functional optimization problem (3-4) have closed forms in terms of the equivalent kernels defined by Equation (6) as follows:*

$$\hat{\mu}_i = \frac{|\mathcal{N}_i| - \sum_{n\in\mathcal{N}}\sum_{n'\in\mathcal{N}_i}\int_0^T h_{u_n}(t-t_n,t_{n'})\mathbf{1}_{0<t-t_n\le A}dt}{T - \sum_{n\in\mathcal{N}}\int_0^T\int_0^T h_{u_n}(t-t_n,s)\mathbf{1}_{0<t-t_n\le A}dtds}, \qquad i\in\mathcal{U}, \tag{9}$$

where $|\mathcal{N}_i|$ denotes the numbe of observed events associated with dimension i.

2.3 Construction of Equivalent Kernels

In Section 2.2, we showed that the optimal estimators of g and μ can be expressed in closed form using the equivalent kernels $\{h_j(\cdot,\cdot)\}_{j\in\mathcal{U}}$. However, obtaining $\{h_j(\cdot,\cdot)\}_{j\in\mathcal{U}}$ in practice requires solving the coupled integral equations (6), which is generally a non-trivial task. In Proposition 3, we propose a solution based on the degenerate kernel approximation methods (Atkinson, 2010; Polyanin & Manzhirov, 1998). The proof is provided in Appendix C.

Proposition 3. *Let an RKHS kernel $k(\cdot,\cdot)$ have a degenerate form with M feature maps $\{\phi_m(s)\}$,*

$$k(s,s') = \sum_{m=1}^{M}\phi_m(s)\phi_m(s') = \boldsymbol{\phi}(s)^\top\boldsymbol{\phi}(s'), \tag{10}$$

where $\boldsymbol{\phi}(s) = (\phi_1(s),\ldots,\phi_M(s))^\top$. Then the solution of the simultaneous Fredholm integral equations (6) can be obtained in closed form as follows:

$$h_j(s,s') = \boldsymbol{\phi}(s)^\top\left[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)^{-1}\tilde{\boldsymbol{\phi}}(s')\right]_{1+(j-1)M:jM}, \quad j\in\mathcal{U}, \tag{11}$$

where $[\cdot]_{a:b}$ denotes the slice of matrix between the a-th row and the b-th one, $\boldsymbol{I}_{MU}\in\mathbb{R}^{MU\times MU}$ denotes the identity matrix, $\boldsymbol{\Xi} = [\boldsymbol{\Xi}_{ij}]\in\mathbb{R}^{MU\times MU}$ is defined as a symmetric block matrix whose (i,j)-th block is given by an M-by-M submatrix,

$$\boldsymbol{\Xi}_{ij} = \sum_{n\in\mathcal{N}_i}\sum_{n'\in\mathcal{N}_j}\mathbf{1}_{\max(t_n,t_{n'})<\min(T,A+t_n,A+t_{n'})}\int_{\max(t_n,t_{n'})}^{\min(T,A+t_n,A+t_{n'})}\boldsymbol{\phi}(t-t_n)\boldsymbol{\phi}(t-t_{n'})^\top dt, \tag{12}$$

and $\tilde{\boldsymbol{\phi}}(s):\mathcal{T}\to\mathbb{R}^{MU}$ denotes a concatenated vector function,

$$\tilde{\boldsymbol{\phi}}(s) = \Big[\tilde{\boldsymbol{\phi}}_1(s)\mid\tilde{\boldsymbol{\phi}}_2(s)\mid\ldots\mid\tilde{\boldsymbol{\phi}}_U(s)\Big], \qquad \tilde{\boldsymbol{\phi}}_i(s) = \sum_{n\in\mathcal{N}_i}\boldsymbol{\phi}(s-t_n)\mathbf{1}_{0<s-t_n\le A}. \tag{13}$$

Proposition 4 shows that substituting the equivalent kernels (11) into Equations (5) and (9), we can obtain the optimal estimators in terms of the feature maps. The proof is provided in Appendix D.

Proposition 4. *For a degenerate form of RKHS kernel in (10), the optimal estimators, $\hat{g}$ and $\hat{\mu}$, are obtained in closed form in terms of the feature maps:*

$$\begin{aligned}
\hat{g}_{ij}(s) &= \boldsymbol{\phi}(s)^\top\left[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)^{-1}\Big(\sum_{n\in\mathcal{N}_i}\tilde{\boldsymbol{\phi}}(t_n) - \hat{\mu}_i\int_0^T\tilde{\boldsymbol{\phi}}(t)dt\Big)\right]_{1+(j-1)M:jM},\\
\hat{\mu}_i &= \frac{|\mathcal{N}_i| - \big(\int_0^T\tilde{\boldsymbol{\phi}}(t)dt\big)^\top\big(\frac{1}{\gamma}\boldsymbol{I}_{MU}+\boldsymbol{\Xi}\big)^{-1}\big(\sum_{n\in\mathcal{N}_i}\tilde{\boldsymbol{\phi}}(t_n)\big)}{T - \big(\int_0^T\tilde{\boldsymbol{\phi}}(t)dt\big)^\top\big(\frac{1}{\gamma}\boldsymbol{I}_{MU}+\boldsymbol{\Xi}\big)^{-1}\big(\int_0^T\tilde{\boldsymbol{\phi}}(t)dt\big)}.
\end{aligned} \tag{14}$$

In this paper, we assume that RKHS kernels are shift-invariant, i.e., $k(s,s') = k(|s-s'|)$, which includes popular kernels such as Gaussian, Matérn, and Laplace kernels. We employ the random

Fourier feature method (Rahimi & Recht, 2007), approximating the shift-invariant RKHS kernel as a sum of Fourier features sampled from the Fourier transform of the kernel, denoted by $\tilde{k}(\omega)$, as

$$\phi_m(s) = \sqrt{\frac{2}{M}}\cos(\omega_m s + \theta_m), \quad \omega_m = \begin{cases} \sim \tilde{k}(\omega) & m \leq \frac{M}{2} \\ \omega_{m-\frac{M}{2}} & m > \frac{M}{2} \end{cases}, \; \theta_m = \begin{cases} 0 & m \leq \frac{M}{2} \\ -\frac{\pi}{2} & m > \frac{M}{2} \end{cases}. \quad (15)$$

To enhance the approximation accuracy of the random Fourier features, we employed the quasi-Monte Carlo feature maps (Yang et al., 2014), and used $M = 100$ in Section 4. Then the integral operations appeared in (12) and (14) can be performed analytically as follows:

$$\begin{aligned} &\int_0^T \tilde{\boldsymbol{\phi}}_i(s)ds = \sqrt{\frac{2}{M}}\frac{1}{\boldsymbol{\omega}} \circ \Big[\sin\big(\boldsymbol{\omega}\cdot\min(T, A - t_n) + \boldsymbol{\theta}\big) - \sin\big(\boldsymbol{\theta}\big)\Big] \in \mathbb{R}^M, \\ &\int_a^b \boldsymbol{\phi}(t - t_n)\boldsymbol{\phi}(t - t_{n'})^\top dt = \frac{b-a}{M}\Big[\zeta(\boldsymbol{\omega}, \boldsymbol{\omega}^\top, \boldsymbol{\theta}, \boldsymbol{\theta}^\top) + \zeta(\boldsymbol{\omega}, -\boldsymbol{\omega}^\top, \boldsymbol{\theta}, -\boldsymbol{\theta}^\top)\Big] \in \mathbb{R}^{M\times M}, \end{aligned} \quad (16)$$

where $\circ$ denotes the Hadamard product, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_M)^\top$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)^\top$, and

$$\zeta(\omega, \omega', \theta, \theta') = \cos\big[(b+a)(\omega+\omega')/2 + \theta + \theta' - \omega t_n - \omega' t_{n'}\big]\operatorname{sinc}\big[(b-a)(\omega+\omega')/2\big]. \quad (17)$$

Here, $\operatorname{sinc}(x) = \sin(x)/x$ is the unnormalized sinc function. As a result, the optimal estimators are obtained in closed form without requiring any discretization approximation of the integral operators. It is worth noting that the number of feature maps, M, required to approximate a one-dimensional RKHS kernel remains modest regardless of the data size, whereas the number of discretization nodes needed for accurate integral evaluations grows with the data size. See Section 3 for details.

2.4 Complexity Analysis

The computational complexity of obtaining our estimators in Equation (14) is $\mathcal{O}(\underline{N}^2M^2U^2 + M^3U^3)$, where $\underline{N} = \max(|\mathcal{N}_1|, \ldots, |\mathcal{N}_U|)$: the first term arises from the computation of $\boldsymbol{\Xi}$, and the second from the inversion of $\big(\gamma^{-1}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\big)$. Its memory complexity is $\mathcal{O}(M^2U^2)$, which stems of $\boldsymbol{\Xi}$. In contrast, the prior kernel-based method (Bonnet & Sangnier, 2025) requires the computation of $\mathcal{O}(\underline{N}^4U^2P)$, where P denotes the number of iterations needed for convergence in an iterative optimization algorithm. Its memory complexity is $\mathcal{O}(\underline{N}^2U^2)$. Therefore, our approach achieves significantly better scalability with respect to the data size compared to the previous method, making it well-suited for large-scale data scenarios. Moreover, our method requires only a single matrix inversion and avoids the need to carefully tune convergence criteria and learning rates, offering a more stable and practical solution, which is in contrast to the prior kernel method-based method that relies on iterative optimization.

3 Related Work

Hawkes processes (Hawkes, 1971), particularly in the multivariate setting, have been extensively studied due to their expressive power in modeling self- and mutually-exciting temporal dynamics on networks. One of the simplest approaches to learning the triggering kernels in Hawkes processes is parametric modeling, where exponential kernels are particularly popular owing to their ability to compactly encode interaction strength and temporal decay. In the case of linear Hawkes processes (1), maximum likelihood estimation has been the gold standard (Bacry et al., 2015; Ozaki, 1979; Zhou et al., 2013). However, several alternative estimation strategies have been proposed, including least squares-based approaches that exploit analytic tractability (Bacry et al., 2020), spectral methods (Adamopoulos, 1976), and moment-matching methods (Da Fonseca & Zaatour, 2014).

Most of the above models assume mutual excitation, i.e., non-negative triggering kernels. Recently, however, Bonnet et al. (2023) introduced a flexible non-linear Hawkes model (see Equation (18)) with exponential triggering kernels, which enables us to estimate both excitatory and inhibitory interactions efficiently within exponential forms. While the complexity of model fitting scales linearly with the number of events, the model requires processing event times successively to evaluate the likelihood function and is therefore difficult to parallelize across multiple cores, limiting scalability.

In the nonparametric regime, a wide variety of methods have been explored for linear Hawkes processes, which include piece-wise constant (Reynaud-Bouret et al., 2014) and Gaussian mixture (Xu

et al., 2016) representations of triggering kernels, and the estimation method via the solution of Wiener–Hopf equations (Bacry & Muzy, 2016). For non-linear Hawkes processes, nonparametric formulations such as those using Bernstein-type polynomials (Lemonnier & Vayatis, 2014) and B-spline expansions (Cai et al., 2024) have been proposed. Additionally, many neural network-based models have been developed to learn event dynamics directly from data, ranging from RNN-based approaches (Mei & Eisner, 2017) to Transformers (Zuo et al., 2020). For a more comprehensive survey, we refer readers to (Bonnet & Sangnier, 2025; Bonnet et al., 2023).

While prior works using reproducing kernel Hilbert spaces (RKHSs) remain relatively underexplored in the context of Hawkes processes, two notable exceptions exist. Yang et al. (2017) propose an online estimation method in RKHS under a regret minimization framework, which fundamentally differs from the batch learning setting considered in this paper. Bonnet & Sangnier (2025), on the other hand, considered a non-linear multivariate Hawkes process,

$$\lambda_i(t) = \varphi\Big(\mu_i + \sum_{j\in\mathcal{U}} \int_0^t g_{ij}(t-s)dN_j(s)\Big), \qquad t \in \mathbb{R}_+,\ i \in \mathcal{U} \coloneqq [\![1, U]\!], \tag{18}$$

where $\varphi(x) = \log(1 + e^{wx})/w$ is a non-negative soft-plus function ($w = 100$), and demonstrated a representer theorem for the problem by adopting an approximate likelihood function (Lemonnier & Vayatis, 2014) or an upper bound on the least squares loss (Lemonnier & Vayatis, 2014) as the objective. Specifically, they showed that in both cases, the optimal estimator of each triggering kernel admits a linear representation in terms of RKHS kernels as follows:

$$\hat{g}_{ij}(\cdot) = \alpha_0^{ij} \sum_{n\in\mathcal{N}_j} \int_0^T k(\cdot, t-t_n)\mathbf{1}_{0<t-t_n\le A}dt + \sum_{n\in\mathcal{N}_i} \alpha_1^{nj} \sum_{n'\in\mathcal{N}_j} k(\cdot, t_n - t_{n'})\mathbf{1}_{0<t_n-t_{n'}\le A}, \tag{19}$$

where $\{\alpha_0^{ij}, \alpha_1^{nj}\}$ denote dual coefficients of dimension $U(N(T) + U)$. Rather than solving the associated dual optimization problem directly, they adopted the linear representation (19) as a semi-parametric model and estimated the dual coefficients by maximizing the objective in (3), where the integrals in the loss functionals (2) were evaluated via discretization approximation. For a detailed formulation of the resulting objective function, see Section 3.3 of (Bonnet & Sangnier, 2025).

While the approach (18-19) exhibits strong empirical performance, it involves solving a non-linear optimization problem over dual coefficients, which demands the computation of $\mathcal{O}(\underline{N}^4U^2)$ for each evaluation of the objective function, resulting in significant scalability challenges for large-scale datasets, as is often the case in multivariate Hawkes processes. What makes the situation worse is that the intensity functions in Hawkes processes is usually discontinuous at the observed event times, and to accurately evaluate the integrals of the intensity functions in the loss functionals (2), a dense set of discretization nodes is required along with the number of events increasing. As a result, the computational cost can grow significantly as the dataset becomes larger.

4 Experiments

We evaluated the validity of our proposed method (`Ours`) by comparing it with prior parametric and non-parametric approaches. In accordance with Bonnet & Sangnier (2025), we adopted the following four approaches as baselines: `Exp` is the state-of-the-art parametric approach based on a non-linear Hawkes process with exponential triggering kernels (Bonnet et al., 2023); `Gau` is a non-parametric approach based on a linear Hawkes process (Xu et al., 2016), where the triggering kernels are represented as Gaussian mixtures; `Ber` is a non-parametric approach based on a non-linear Hawkes process (Lemonnier & Vayatis, 2014), where the triggering kernels are represented as Bernstein-type polynomials; `Bonnet` is the state-of-the-art kernel method-based approach (Bonnet & Sangnier, 2025), which assumes the triggering kernels lie in an RKHS. For `Bonnet`, the node size of discretization approximation was set at $\max(1000, 2|\mathcal{N}_1|, \ldots, 2|\mathcal{N}_U|)$ folowing (Bonnet & Sangnier, 2025). For `Ours` and `Bonnet`, a Gaussian RKHS kernel was employed: $k(s, s') = e^{-(\beta|s-s'|)^2}$, where β is the inverse scale hyperparameter. For `Gau` and `Ber`, the number of basis functions was set at 50. For the models except `Exp`, the support window A was set at 5.

Except for `Exp`, the models have hyperparameters to optimize. `Gau` and `Ber` have the regularization hyperparameter γ for a quadratic penalty on the coefficients of mixture models, and `Ours` and

Table 1: Results of `Exp` (Bonnet et al., 2023), `Gau` (Xu et al., 2016), `Ber` (Lemonnier & Vayatis, 2014), `Bonnet` (Bonnet & Sangnier, 2025), and `Ours` on mutually-exciting scenario across 10 trials with standard errors in brackets. $\tilde{N}$ is the average data size per trial. cpu is the CPU time in seconds. The performances not significantly ($p \geq 0.01$) different from the best one under the Mann-Whitney U test (Holm, 1979) are shown in bold.

		Exp		Gau		Ber		Bonnet		Ours	
T	$\tilde{N}$	Δ^2	cpu	Δ^2	cpu	Δ^2	cpu	Δ^2	cpu	Δ^2	cpu
2000	1318	0.29 (0.01)	46.1 (23.3)	**0.20** (0.02)	**3.42** (1.33)	0.51 (0.16)	**4.06** (2.03)	**0.21** (0.05)	135 (144)	0.38 (0.15)	**3.16** (2.42)
3000	2055	0.29 (0.00)	74.7 (41.3)	**0.19** (0.02)	**4.94** (3.00)	0.69 (0.50)	**6.61** (4.71)	**0.19** (0.04)	272 (298)	0.27 (0.06)	**4.76** (3.97)
5000	4081	0.28 (0.00)	134 (64.9)	**0.18** (0.01)	**12.2** (8.91)	0.30 (0.06)	**19.0** (15.4)	**0.15** (0.02)	1358 (1270)	**0.20** (0.06)	**10.7** (7.71)
7000	5380	0.28 (0.00)	180.4 (56.6)	**0.18** (0.02)	**17.3** (8.00)	0.27 (0.04)	29.0 (16.1)	**0.14** (0.04)	2070 (1210)	**0.16** (0.04)	**13.1** (5.55)

Table 2: Results on refractory scenario data across 10 trials. Notations follow Table 1.

		Exp		Gau		Ber		Bonnet		Ours	
T	$\tilde{N}$	Δ^2	cpu	Δ^2	cpu	Δ^2	cpu	Δ^2	cpu	Δ^2	cpu
2000	2050	2.19 (0.03)	124 (61.3)	1.51 (0.02)	**7.03** (3.58)	1.23 (0.31)	**10.8** (6.19)	**0.63** (0.19)	413 (291)	0.95 (0.24)	**5.04** (3.92)
3000	2956	2.19 (0.02)	183 (47.9)	1.50 (0.02)	**10.8** (4.11)	0.96 (0.14)	18.5 (7.58)	**0.56** (0.22)	927 (460)	**0.85** (0.19)	**7.66** (4.74)
5000	5222	2.20 (0.02)	355 (74.0)	1.47 (0.01)	**24.7** (8.41)	0.82 (0.16)	44.0 (15.8)	**0.44** (0.18)	3197 (1323)	**0.59** (0.13)	**14.9** (5.48)
7000	6887	2.20 (0.02)	503 (108)	1.47 (0.01)	37.3 (14.5)	0.71 (0.09)	74.7 (29.0)	**0.41** (0.19)	5884 (2972)	**0.50** (0.07)	**16.1** (5.26)

`Bonnet` have the inverse scale hyperparameter β in addition to γ. We optimized the hyperparameters on the grids of $\gamma \in \{0.1, 0.5, 1.0\}$ and $\beta \in \{0.5, 1.0, 1.5\}$, based on the negative log-likelihood (`Gau`, `Ber`, `Bonnet`) and the least squares loss (`Ours`) minimization. Specifically, for a sequence of events observed in an interval $[0, T]$, each model was fitted with the events in [0, 0.8T], evaluated the negative log-likelihood/least squares loss for the rest of the data in $[0.8T, T]$, and the hyperparameters to minimize the criteria were chosen.

Predictive performance was assessed using the integrated squared error (Δ^2) defined as follows:

$$\Delta^2 = \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} \int_0^A \left| g_{ij}(s) - \hat{g}_{ij}(s) \right|^2 ds, \tag{20}$$

where $A = 5$, and $g_{ij}(s)$ and $\hat{g}_{ij}(s)$ denote the true and estimated triggering kernels, respectively. Efficiency was evaluated based on the CPU time, denoted by cpu, required to execute the model fitting given the optimized hyperparameters.

The four baselines were implemented using the Python code in Bonnet & Sangnier (2025) (MIT License), and our model using TensorFlow-2.10 (Abadi et al., 2015). All the experiments were executed on a MacBook Pro equipped with a 12-core CPU (Apple M2 Max), with the GPU disabled.

4.1 Mutually-exciting Scenario

We consider synthetic data generated from a 3-variate linear Hawkes process (1) with baseline intensities, $\mu_i = 0.01$ for $i \in \mathcal{U}$, and mutually-exciting triggering kernels ($g_{ij} > 0$) defined as follows:

$$\begin{array}{lll} g_{11}(s) = 0.5e^{-s} & g_{12}(s) = 0.5e^{-10(s-1)^2} & g_{13}(s) = 0.5e^{-20(s-3)^2} \\ g_{21}(s) = 2^{-5s-1} & g_{22}(s) = 0.3e^{-0.5s} & g_{23}(s) = 0.5e^{-20(s-2)^2} \\ g_{31}(s) = 0.2e^{-3(s-2)^2} & g_{32}(s) = 0.25(1 + \cos(\pi s))e^{-s} & g_{33}(s) = 0.5e^{-s} \end{array}, \tag{21}$$

of which setting is a modification of the one that appeared in Bonnet & Sangnier (2025). We simulated 10 trial sequences of events over the interval $[0, T]$ and performed the estimation of triggering

kernels 10 times using the compared methods. To clarify the model efficiency regarding data size, we set the horizon at $T \in [2000, 3000, 5000, 7000]$.

Table 1 displays the predictive error and computational efficiency on the mutually-exciting scenario dataset. Some estimation results are displayed in Appendix A. The results demonstrate that, given a sufficiently large amount of data, our proposal achieved comparable predictive accuracy to `Bonnet`, the SOTA kernel-based approach, while requiring several orders of magnitude less computation time for model fitting. In small data regimes, `Bonnet` tended to achieve higher accuracy than `Ours`, which may be attributed to the theoretical advantage of negative log-likelihood over least squares loss in reducing estimation biases (Bacry et al., 2016). `Gau` consistently achieved high predictive accuracy with small computation time, because it is the only baseline specifically designed for mutually-exciting interactions, aligned with the underlying process. `Exp` performed well only for the exponential triggering kernels $\{g_{ii}(\cdot)\}_{i\in\mathcal{U}}$ (see Appendix A).

4.2 Refractory Scenario

Refractory phenomena arise in point processes exhibiting short-term self-inhibition, where the occurrence of an event temporarily suppresses the likelihood of subsequent events. Such behavior is observed in neuronal spike trains (Berry & Meister, 1997) and in sequences of mainshock events (Rotondi & Varini, 2019). Here, we consider synthetic data generated from a 3-variate non-linear Hawkes process (18) with short-term self-inhibition adopted in (Bonnet & Sangnier, 2025),

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

and various non-inhibitory inter-interactions,

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

The remaining experimental conditions follow those of the mutually-exciting scenario. We adopted the soft-plus function as the link function, which is consistent with the assumption in `Bonnet`.

Table 2 presents the predictive error and computational efficiency on the refractory scenario dataset. Some estimation results are displayed in Appendix A. As shown, `Ours` was outperformed by `Bonnet` in terms of accuracy on small datasets, but the gap became less significant as the dataset size increased. In contrast to the mutually-exciting scenario, `Gau` performed poorly here due to its inability to model inhibitory interaction. While `Ber` succeeded in reconstructing the inhibitory interactions, it still fell short of the kernel method-based approaches in terms of accuracy. Increasing the component number may improve `Ber`'s performance, but at the cost of higher computational time. Our proposed method was the fastest, achieving a speed-up of several hundred times compared to the SOTA `Bonnet`. Note that `Ber` exhibited unstable behavior in some trials, where Δ^2 exceeded 1000. Such outlier samples were excluded from the results in Table 2.

5 Conclusion

We have proposed a novel penalized least squares loss formulation for estimating triggering kernels in multivariate Hawkes processes that reside in an RKHS. We demonstrated that a novel representer theorem holds for the optimization problem and derived a feasible estimator based on kernel methods. We evaluated the proposed estimator on synthetic data, confirming that it achieved comparable predictive accuracy while being substantially faster than the state-of-the-art kernel method estimator.

Limitations: Our proposed method is based on a linear Hawkes process, which does not guarantee the non-negativity of the intensity function. As a result, when using the estimated triggering kernel to predict future intensity values, post-hoc clippings such as applying $\max(\lambda(t), 0)$ are required. If it is known a priori that the underlying triggering kernels are excitatory, it is more appropriate to enforce non-negativity directly on the estimated triggering kernels. Moreover, the computational complexity of our method scales cubically with the dimensionality U of the Hawkes process, making it less suitable for high-dimensional processes, while it empirically works robustly for moderate U (see Appendix F.2). This issue, stemming from the matrix inversion, may be mitigated using iterative solvers such as the conjugate gradient method.

Reproducibility Statement

We provide detailed implementation instructions and reproducibility guidelines in Section 4, and the full implementation (python code) is submitted as supplementary materials.

The Use of Large Language Models (LLMs)

We used LLMs to polish writing and correct typos.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

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

Kendall Atkinson. A personal perspective on the history of the numerical analysis of Fredholm integral equations of the second kind. In *The Birth of Numerical Analysis*, pp. 53–72. World Scientific, 2010.

Emmanuel Bacry and Jean-François Muzy. First-and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62 (4):2184–2202, 2016.

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

Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Mean-field inference of Hawkes point processes. *Journal of Physics A: Mathematical and Theoretical*, 49 (17):174006, 2016.

Emmanuel Bacry, Martin Bompaire, Stéphane Gaïffas, and Jean-François Muzy. Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21(50):1–32, 2020.

Michael Berry and Markus Meister. Refractoriness and neural precision. In *Advances in Neural Information Processing Systems 10*, 1997.

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.

Biao Cai, Jingfei Zhang, and Yongtao Guan. Latent network structure learning from high-dimensional multivariate point processes. *Journal of the American Statistical Association*, 119 (545):95–108, 2024.

Jean-Paul Chiles and Pierre Delfiner. *Geostatistics: Modeling Spatial Uncertainty*, volume 713. John Wiley & Sons, 2012.

David Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, 2023.

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

Daryl J Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer Science & Business Media, 2006.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1555–1564, 2016.

Seth Flaxman, Yee Whye Teh, and Dino Sejdinovic. Poisson intensity estimation with reproducing kernels. In *Artificial Intelligence and Statistics*, pp. 270–279. PMLR, 2017.

Felipe Gerhard, Moritz Deger, and Wilson Truccolo. On the stability and dynamics of stochastic spiking neuron models: Nonlinear Hawkes process and point process GLMs. *PLoS Computational Biology*, 13(2):e1005390, 2017.

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.

Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

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

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

Hideaki Kim. Fast Bayesian inference for Gaussian Cox processes via path integral formulation. In *Advances in Neural Information Processing Systems 34*, 2021.

Hideaki Kim. Survival permanental processes for survival analysis with time-varying covariates. *Advances in Neural Information Processing Systems 36*, 2023.

Hideaki Kim, Taichi Asami, and Hiroyuki Toda. Fast Bayesian estimation of point process intensity as function of covariates. In *Advances in Neural Information Processing Systems 35*, 2022.

Remi Lemonnier and Nicolas Vayatis. Nonparametric Markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate Hawkes processes. In *Machine Learning and Knowledge Discovery in Databases*, pp. 161–176. Springer, 2014.

Miao Liu, Girish Chowdhary, Bruno Castra Da Silva, Shih-Yuan Liu, and Jonathan P. How. Gaussian processes for learning and control: A tutorial with examples. *IEEE Control Systems Magazine*, 38(5):53–86, 2018.

Weifeng Liu, Jose C. Principe, and Simon Haykin. *Kernel Adaptive Filtering: A Comprehensive Introduction*. John Wiley & Sons, 2011.

Hongyuan Mei and Jason M. Eisner. The neural Hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems 30*, 2017.

James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

Yosihiko Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30:243–261, 1978.

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

Tohru Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31:145–155, 1979.

Andrei D. Polyanin and Alexander V. Manzhirov. *Handbook of Integral Equations*. CRC press, 1998.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, 2007.

Patricia Reynaud-Bouret, Vincent Rivoirard, Franck Grammont, and Christine Tuleau-Malot. Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience*, 4:1–41, 2014.

Renata. Rotondi and Elisa Varini. Failure models driven by a self-correcting point process in earthquake occurrence modeling. *Stochastic Environmental Research and Risk Assessment*, 33(3): 709–724, 2019.

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

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pp. 416–426. Springer, 2001.

Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert. *Kernel Methods in Computational Biology*. MIT press, 2004.

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

Hiroyuki Takeda, Sina Farsiu, and Peyman Milanfar. Kernel regression for image processing and reconstruction. *IEEE Transactions on Image Processing*, 16(2):349–366, 2007.

Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge University Press, 2000.

Grace Wahba. *Spline Models for Observational Data*, volume 59. SIAM, 1990.

Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning Granger causality for Hawkes processes. In *International Conference on Machine Learning*, pp. 1717–1726. PMLR, 2016.

Jiyan Yang, Vikas Sindhwani, Haim Avron, and Michael Mahoney. Quasi-Monte Carlo feature maps for shift-invariant kernels. In *International Conference on Machine Learning*, pp. 485–493. PMLR, 2014.

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

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.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer Hawkes process. In *International Conference on Machine Learning*, pp. 11692–11702. PMLR, 2020.

A EXAMPLES OF THE ESTIMATED TRIGGERING KERNELS ON SYNTHETIC DATA

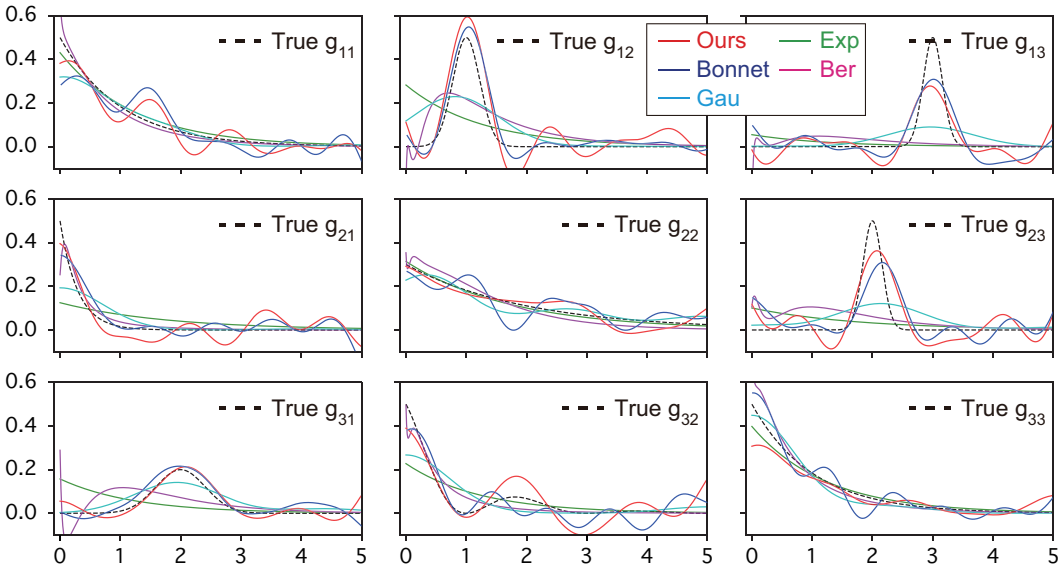

Figure A1: Examples of the estimated triggering kernels in the mutually-exciting scenario. Dashed lines represent the true triggering kernels.

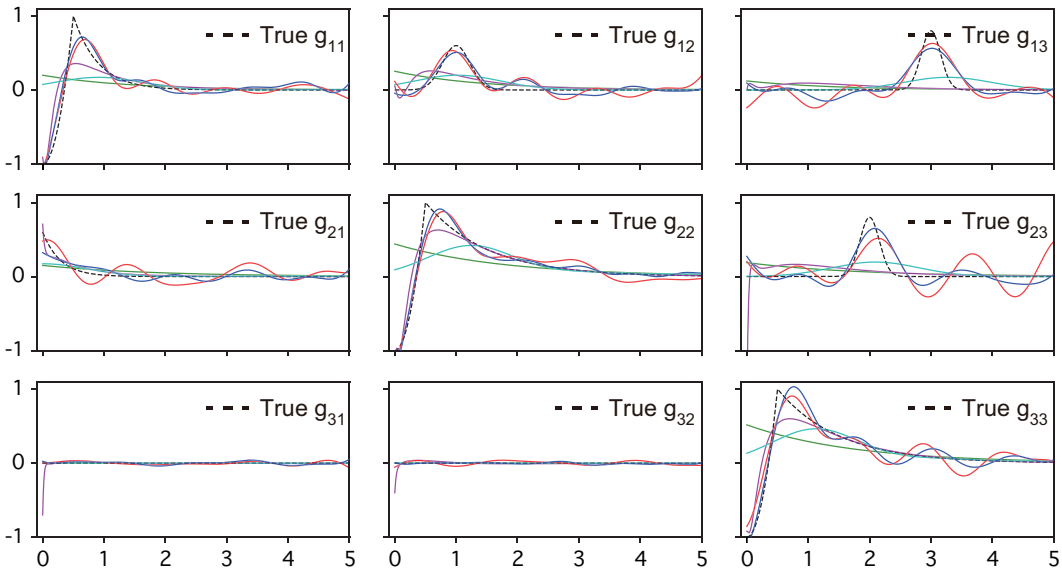

Figure A2: Examples of the estimated triggering kernels in the refractory scenario. Dashed lines represent the true triggering kernels.

B PROOF OF PROPOSITION 2

Proof. Given the estimated baseline intensities $\{\hat{\mu}_i\}_{i\in\mathcal{U}}$, the optimal estimators of the triggering kernels, $\{\hat{g}_{ij}(\cdot)\}_{(i,j)\in\mathcal{U}^2}$, are determined by the equations,

$$\frac{\delta}{\delta g_{ij}(s)}\left[L_{\text{LS}}(g,\hat{\mu})+\frac{1}{\gamma}\sum_{(i,j)\in\mathcal{U}^2}\|g_{ij}\|^2_{\mathcal{H}_k}\right]\Big|_{g=\hat{g}}=0,\qquad s\in\mathcal{T},\ (i,j)\in\mathcal{U}^2. \tag{B1}$$

These equations are equivalently expressed as:

$$\begin{aligned}\int_0^T\Big(\hat{\mu}_i+\sum_{n'\in\mathcal{N}}\hat{g}_{iu_{n'}}(t-t_{n'})\mathbf{1}_{0<t-t_{n'}\le A}\Big)\sum_{n\in\mathcal{N}_j}\delta(s-(t-t_n))\mathbf{1}_{0<t-t_n\le A}dt\\ -\sum_{n'\in\mathcal{N}_i}\sum_{n\in\mathcal{N}_j}\delta(s-(t_{n'}-t_n))\mathbf{1}_{0<t_{n'}-t_n\le A}+\frac{1}{\gamma}\int_{\mathcal{T}}k^*(s,t)\hat{g}_{ij}(t)dt=0.\end{aligned} \tag{B2}$$

Applying the operator $\int_{\mathcal{T}}\cdot\,\hat{g}_{ij}(s)ds$ to both sides of Equations (B2) yields the following representation of the RKHS regularization term under the estimated triggering kernels:

$$\begin{aligned}\frac{1}{\gamma}\|\hat{g}_{ij}\|^2_{\mathcal{H}_k}&=\frac{1}{\gamma}\iint_{\mathcal{T}\times\mathcal{T}}k^*(s,t)\hat{g}_{ij}(s)\hat{g}_{ij}(t)dsdt\\&=\sum_{n'\in\mathcal{N}_i}\sum_{n\in\mathcal{N}_j}\hat{g}_{ij}(t_{n'}-t_n)\mathbf{1}_{0<t_{n'}-t_n\le A}\\&\quad-\int_0^T\Big(\hat{\mu}_i+\sum_{n'\in\mathcal{N}}\hat{g}_{iu_{n'}}(t-t_{n'})\mathbf{1}_{0<t-t_{n'}\le A}\Big)\sum_{n\in\mathcal{N}_j}\hat{g}_{ij}(t-t_n)\mathbf{1}_{0<t-t_n\le A}dt.\end{aligned} \tag{B3}$$

Substituting this representation into the objective function in (3) leads to:

$$\begin{aligned}&L_{\text{LS}}(\hat{g},\hat{\mu})+\frac{1}{\gamma}\sum_{(i,j)\in\mathcal{U}^2}\|\hat{g}_{ij}\|^2_{\mathcal{H}_k}\\&=\sum_{i\in\mathcal{U}}\Big[\int_0^T\Big(\hat{\mu}_i+\sum_{n\in\mathcal{N}}\hat{g}_{iu_n}(t-t_n)\mathbf{1}_{0<t-t_n\le A}\Big)^2dt+\sum_{n'\in\mathcal{N}_i}\sum_{n\in\mathcal{N}}\hat{g}_{iu_n}(t_{n'}-t_n)\mathbf{1}_{0<t_{n'}-t_n\le A}\\&\qquad-2\sum_{n'\in\mathcal{N}_i}\Big(\hat{\mu}_i+\sum_{n\in\mathcal{N}}\hat{g}_{iu_n}(t_{n'}-t_n)\mathbf{1}_{0<t_{n'}-t_n\le A}\Big)\\&\qquad-\int_0^T\Big(\hat{\mu}_i+\sum_{n'\in\mathcal{N}}\hat{g}_{iu_{n'}}(t-t_{n'})\mathbf{1}_{0<t-t_{n'}\le A}\Big)\sum_{n\in\mathcal{N}}\hat{g}_{iu_n}(t-t_n)\mathbf{1}_{0<t-t_n\le A}dt\Big]\\&=\sum_{i\in\mathcal{U}}\Big[T\hat{\mu}_i^2+\hat{\mu}_i\int_0^T\sum_{n\in\mathcal{N}}\hat{g}_{iu_n}(t-t_n)\mathbf{1}_{0<t-t_n\le A}dt-2|\mathcal{N}_i|\hat{\mu}_i\\&\qquad\qquad-\sum_{n'\in\mathcal{N}_i}\sum_{n\in\mathcal{N}}\hat{g}_{iu_n}(t_{n'}-t_n)\mathbf{1}_{0<t_{n'}-t_n\le A}\Big].\end{aligned} \tag{B4}$$

By invoking the representer theorem (5), the objective can be rewritten as a function of the estimated baseline intensities:

$$\begin{aligned}Z(\hat{\mu})&=L_{\text{LS}}(\hat{g}(\hat{\mu}),\hat{\mu})+\frac{1}{\gamma}\sum_{(i,j)\in\mathcal{U}^2}\|\hat{g}_{ij}(\hat{\mu})\|^2_{\mathcal{H}_k}\\&=\sum_{i\in\mathcal{U}}\Big[\hat{\mu}_i^2\Big(T-\sum_{n\in\mathcal{N}}\int_0^T\int_0^T h_{u_n}(t-t_n,s)\mathbf{1}_{0<t-t_n\le A}dtds\Big)\\&\qquad-\hat{\mu}_i\Big(2|\mathcal{N}_i|-\sum_{n\in\mathcal{N}}\sum_{n'\in\mathcal{N}_i}\int_0^T h_{u_n}(t-t_n,t_{n'})\mathbf{1}_{0<t-t_n\le A}dt\\&\qquad\qquad-\sum_{n\in\mathcal{N}}\sum_{n'\in\mathcal{N}_i}\int_0^T h_{u_n}(t_{n'}-t_n,t)\mathbf{1}_{0<t_{n'}-t_n\le A}dt\Big)\Big]+C,\end{aligned} \tag{B5}$$

where C is the constant term. $Z(\hat{\mu})$ can be more simplified as

$$Z(\hat{\mu}) = \sum_{i\in\mathcal{U}}\Bigg[\hat{\mu}_i^2\Bigg(T - \sum_{n\in\mathcal{N}}\int_0^T\int_0^T h_{u_n}(t-t_n,s)\mathbf{1}_{0<t-t_n\leq A}dtds\Bigg) - 2\hat{\mu}_i\Bigg(|\mathcal{N}_i| - \sum_{n\in\mathcal{N}}\sum_{n'\in\mathcal{N}_i}\int_0^T h_{u_n}(t-t_n,t_{n'})\mathbf{1}_{0<t-t_n\leq A}dt\Bigg)\Bigg] + C, \tag{B6}$$

where the identity

$$\sum_{n\in\mathcal{N}}\sum_{n'\in\mathcal{N}_i}\int_0^T h_{u_n}(t-t_n,t_{n'})\mathbf{1}_{0<t-t_n\leq A}dt = \sum_{n\in\mathcal{N}}\sum_{n'\in\mathcal{N}_i}\int_0^T h_{u_n}(t_{n'}-t_n,t)\mathbf{1}_{0<t_{n'}-t_n\leq A}dt, \tag{B7}$$

is used (for proof, see Appendix D). Finally, the optimal estimators $\hat{\mu}_i$ should solve the equation where the functional derivative of $Z(\hat{\mu})$ regarding μ_i is equal to zero ($dZ/d\mu_i = 0$), which leads to the following representation in terms of the equivalent kernels:

$$\hat{\mu}_i = \frac{|\mathcal{N}_i| - \sum_{n\in\mathcal{N}}\sum_{n'\in\mathcal{N}_i}\int_0^T h_{u_n}(t-t_n,t_{n'})\mathbf{1}_{0<t-t_n\leq A}dt}{T - \sum_{n\in\mathcal{N}}\int_0^T\int_0^T h_{u_n}(t-t_n,s)\mathbf{1}_{0<t-t_n\leq A}dtds}, \qquad i\in\mathcal{U}. \tag{B8}$$

This completes the proof. ■

C Proof of Proposition 3

Proof. For an RKHS kernel $k(\cdot,\cdot)$ with the degenerate form given in Equation (10), the coefficient functions, $\{V_{jl}(s,t)\}_{(j,l)\in\mathcal{U}^2}$, appearing in the system of integral equations (6) can also be expressed in degenerate forms as follows:

$$\begin{aligned} V_{jl}(s,t) &= \sum_{m=1}^{M}\phi_m(s)\psi_m^{jl}(t), \\ \psi_m^{jl}(t) &= \sum_{n\in\mathcal{N}_j}\sum_{n'\in\mathcal{N}_l}\phi_m(t+t_{n'}-t_n)\mathbf{1}_{\max(t_n,t_{n'})<t+t_{n'}\leq\min(T,A+t_n,A+t_{n'})}. \end{aligned} \tag{C1}$$

Substituting Equation (C1) into Equation (6), we find that the solutions, $\{h_j(s,s')\}_{j\in\mathcal{U}}$, admit degenerate forms as

$$\begin{aligned} h_j(s,s') &= \gamma\sum_{n\in\mathcal{N}_j}k(s,s'-t_n)\mathbf{1}_{0<s'-t_n\leq A} - \gamma\sum_{l\in\mathcal{U}}\int_0^T V_{jl}(s,t)h_l(t,s')dt, \\ &= \gamma\sum_{m=1}^{M}\phi_m(s)\Bigg[\sum_{n\in\mathcal{N}_j}\phi_m(s'-t_n)\mathbf{1}_{0<s'-t_n\leq A} - \sum_{l\in\mathcal{U}}\int_0^T\psi_m^{jl}(t)h_l(t,s')dt\Bigg] \\ &= \sum_{m=1}^{M}\phi_m(s)c_m^j(s'), \end{aligned} \tag{C2}$$

where $\{c_m^j(s')\}_{(m,j)\in[\![1,M]\!]\times\mathcal{U}}$ are unknown coefficient functions. By substituting Equations (C2) and (C1) into Equation (6), we obtain the following linear system that the coefficient functions must satisfy:

$$\sum_m\phi_m(s)\Bigg[\frac{1}{\gamma}c_m^j(s') + \sum_{l\in\mathcal{U}}\int_0^T\psi_m^{jl}(t)\sum_{m'}\phi_{m'}(t)c_{m'}^l(s')dt - \sum_{n\in\mathcal{N}_j}\phi_m(s'-t_n)\mathbf{1}_{0<s'-t_n\leq A}\Bigg] = 0,$$

$$\therefore\ \frac{1}{\gamma}c_m^j(s) + \sum_{l\in\mathcal{U}}\sum_{m'}c_{m'}^l(s)\int_0^T\psi_m^{jl}(t)\phi_{m'}(t)dt - \sum_{n\in\mathcal{N}_j}\phi_m(s-t_n)\mathbf{1}_{0<s-t_n\leq A} = 0, \tag{C3}$$

for $(m, j) \in [\![1, M]\!] \times \mathcal{U}$. Let us define the MU-dimensional stacked vector of coefficient functions as

$$\tilde{\boldsymbol{c}}(s) = (c_1^1(s), c_2^1(s), \ldots, c_M^1(s), c_1^2(s), c_2^2(s), \ldots, c_M^2(s), c_1^3(s), \ldots, c_M^U(s))^\top. \tag{C4}$$

Then, the linear system can be written compactly as

$$\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)\tilde{\boldsymbol{c}}(s) = \tilde{\boldsymbol{\phi}}(s), \tag{C5}$$

where $\tilde{\boldsymbol{\phi}}(s)$ and $\boldsymbol{\Xi}$ are defined in Equations (12) and (13), respectively. Substituting Equation C5) into Equation (C2) yields the solution to the system of integral equations (6) as,

$$\begin{aligned} h_j(s, s') &= \sum_{m=1}^{M} \phi_m(s) c_m^j(s') \\ &= \boldsymbol{\phi}(s)^\top \big[\boldsymbol{c}(s')\big]_{1+(j-1)M:jM} \\ &= \boldsymbol{\phi}(s)^\top \Big[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)^{-1} \tilde{\boldsymbol{\phi}}(s')\Big]_{1+(j-1)M:jM}. \end{aligned} \tag{C6}$$

This completes the proof. ■

D Proof of Proposition 4

Proof. Substituting Equation (11) into Equation (5) yields the expression for the estimated triggering kernels in terms of the feature maps:

$$\hat{g}_{ij}(s) = \boldsymbol{\phi}(s)^\top \Big[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)^{-1} \Big(\sum_{n \in \mathcal{N}_i} \tilde{\boldsymbol{\phi}}(t_n) - \hat{\mu}_i \int_0^T \tilde{\boldsymbol{\phi}}(t) dt\Big)\Big]_{1+(j-1)M:jM}. \tag{D1}$$

By using Equation (11), the double integral in the denominator of Equation (9) can be rewritten using the feature maps as follows:

$$\begin{aligned} &\sum_{n \in \mathcal{N}} \int_0^T \int_0^T h_{u_n}(t - t_n, s) \mathbf{1}_{0 < t - t_n \le A} dt ds \\ &= \sum_{n \in \mathcal{N}} \Big[\int_0^T \boldsymbol{\phi}(t - t_n) \mathbf{1}_{0 < t - t_n \le A} dt\Big]^\top \Big[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)^{-1} \int_0^T \tilde{\boldsymbol{\phi}}(s) ds\Big]_{1+(u_n-1)M:u_nM} \\ &= \sum_{l \in \mathcal{U}} \Big[\sum_{n \in \mathcal{N}_l} \int_0^T \boldsymbol{\phi}(t - t_n) \mathbf{1}_{0 < t - t_n \le A} dt\Big]^\top \Big[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)^{-1} \int_0^T \tilde{\boldsymbol{\phi}}(s) ds\Big]_{1+(l-1)M:lM} \\ &= \sum_{l \in \mathcal{U}} \Big[\int_0^T \tilde{\boldsymbol{\phi}}_l(t) dt\Big]^\top \Big[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)^{-1} \int_0^T \tilde{\boldsymbol{\phi}}(s) ds\Big]_{1+(l-1)M:lM} \\ &= \Big(\int_0^T \tilde{\boldsymbol{\phi}}(t) dt\Big)^\top \Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)^{-1} \Big(\int_0^T \tilde{\boldsymbol{\phi}}(t) dt\Big). \end{aligned} \tag{D2}$$

Similarly, the integral term in the numerator of Equation (9) becomes,

$$\begin{aligned} &\sum_{n \in \mathcal{N}} \sum_{n' \in \mathcal{N}_i} \int_0^T h_{u_n}(t - t_n, t_{n'}) \mathbf{1}_{0 < t - t_n \le A} dt \\ &= \sum_{n \in \mathcal{N}} \Big[\int_0^T \boldsymbol{\phi}(t - t_n) \mathbf{1}_{0 < t - t_n \le A} dt\Big]^\top \Big[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)^{-1} \Big(\sum_{n' \in \mathcal{N}_i} \tilde{\boldsymbol{\phi}}(t_{n'})\Big)\Big]_{1+(u_n-1)M:u_nM} \\ &= \sum_{l \in \mathcal{U}} \Big(\int_0^T \tilde{\boldsymbol{\phi}}_l(t) dt\Big)^\top \Big[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)^{-1} \Big(\sum_{n' \in \mathcal{N}_i} \tilde{\boldsymbol{\phi}}(t_{n'})\Big)\Big]_{1+(l-1)M:lM} \\ &= \Big(\int_0^T \tilde{\boldsymbol{\phi}}(t) dt\Big)^\top \Big(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\Big)^{-1} \Big(\sum_{n' \in \mathcal{N}_i} \tilde{\boldsymbol{\phi}}(t_{n'}) dt\Big). \end{aligned} \tag{D3}$$

This completes the proof. ■

Furthermore, from Equation (D3) and the following identity:

$$
\begin{aligned}
&\sum_{n\in\mathcal{N}}\sum_{n'\in\mathcal{N}_i}\int_0^T h_{u_n}(t_{n'}-t_n,t)\mathbf{1}_{0<t_{n'}-t_n\le A}dt \\
&=\sum_{n\in\mathcal{N}}\Big[\sum_{n'\in\mathcal{N}_i}\boldsymbol{\phi}(t_{n'}-t_n)\mathbf{1}_{0<t_{n'}-t_n\le A}\Big]^\top\Big[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU}+\boldsymbol{\Xi}\Big)^{-1}\Big(\int_0^T\tilde{\boldsymbol{\phi}}(t)dt\Big)\Big]_{1+(u_n-1)M:u_nM} \\
&=\sum_{l\in\mathcal{U}}\Big[\sum_{n'\in\mathcal{N}_i}\sum_{n\in\mathcal{N}_l}\boldsymbol{\phi}(t_{n'}-t_n)\mathbf{1}_{0<t_{n'}-t_n\le A}\Big]^\top\Big[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU}+\boldsymbol{\Xi}\Big)^{-1}\Big(\int_0^T\tilde{\boldsymbol{\phi}}(t)dt\Big)\Big]_{1+(l-1)M:lM} \\
&=\sum_{l\in\mathcal{U}}\Big[\sum_{n'\in\mathcal{N}_i}\tilde{\boldsymbol{\phi}}_l(t_{n'})\Big]^\top\Big[\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU}+\boldsymbol{\Xi}\Big)^{-1}\Big(\int_0^T\tilde{\boldsymbol{\phi}}(t)dt\Big)\Big]_{1+(l-1)M:lM} \\
&=\Big(\sum_{n'\in\mathcal{N}_i}\tilde{\boldsymbol{\phi}}(t_{n'})dt\Big)^\top\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU}+\boldsymbol{\Xi}\Big)^{-1}\Big(\int_0^T\tilde{\boldsymbol{\phi}}(t)dt\Big) \\
&=\Big(\int_0^T\tilde{\boldsymbol{\phi}}(t)dt\Big)^\top\Big(\frac{1}{\gamma}\boldsymbol{I}_{MU}+\boldsymbol{\Xi}\Big)^{-1}\Big(\sum_{n'\in\mathcal{N}_i}\tilde{\boldsymbol{\phi}}(t_{n'})dt\Big),
\end{aligned}
$$

where the final equality holds because $\big(\gamma^{-1}\boldsymbol{I}_{MU}+\boldsymbol{\Xi}\big)$ is symmetric, we obtain the relation in Equation (B7), which holds for any $M\le\infty$ and feature map $\boldsymbol{\phi}(s)$.

E Proof of Theorem 1 via Mercer's Theorem

Proof. Through Mercer's theorem, the RKHS kernel $k(\cdot,\cdot)$ can be expressed through its Mercer's expansion:

$$
k(t,s)=\sum_{m=1}^{\infty}e_m(t)e_m(s),\quad \int_{\mathcal{T}}e_m(t)e_{m'}(t)dt=\eta_m\delta_{mm'},\tag{E1}
$$

where $\{e_m(\cdot)\}_{m=1}^{\infty}$ and $\{\eta_m\}_{m=1}^{\infty}$ denote the eigenfunctions and the eigenvalues, respectively, of the integral operator $\int_{\mathcal{T}}\cdot\,k(t,s)ds$. Accordingly, the triggering kernels in the RKHS, $\{g_{ij}(\cdot)\in\mathcal{H}_k\}_{(i,j)\in\mathcal{U}^2}$, and their squared RKHS norms, $||g_{ij}||^2_{\mathcal{H}_k}$, admit the representation

$$
g_{ij}(s)=\sum_{m=1}^{\infty}b_{ij}^m\,e_m(s),\qquad ||g_{ij}||^2_{\mathcal{H}_k}=\sum_{m=1}^{\infty}(b_{ij}^m)^2,\qquad (i,j)\in\mathcal{U}^2,\tag{E2}
$$

where $b=\{b_{ij}^m\in\mathbb{R}\}$ is the expansion coefficient. Using this representation, the optimization problem (3-4) can be reformulated as follows:

$$
\hat{b},\ \hat{\mu}\ =\underset{b,\ \mu}{\arg\min}\Big[L_{\text{LS}}(b,\mu)+\frac{1}{\gamma}\sum_{(i,j)\in\mathcal{U}^2}\sum_{m=1}^{\infty}(b_{ij}^m)^2\Big],\tag{E3}
$$

where

$$
\begin{aligned}
L_{\text{LS}}(b,\mu)=\sum_{i\in\mathcal{U}}\Big[\int_0^T\Big(\mu_i+\sum_{n\in\mathcal{N}}\sum_{m=1}^{\infty}b_{iu_n}^m e_m(t-t_n)\mathbf{1}_{0<t-t_n\le A}\Big)^2dt \\
-2\sum_{n'\in\mathcal{N}_i}\Big(\mu_i+\sum_{n\in\mathcal{N}}\sum_{m=1}^{\infty}b_{iu_n}^m e_m(t_{n'}-t_n)\mathbf{1}_{0<t_{n'}-t_n\le A}\Big)\Big].
\end{aligned}\tag{E4}
$$

Given the estimate of the baseline intensity $\hat{\mu}$, the optimal coefficient vector $\hat{b}$ must satisfy the equation obtained by setting the gradient of the objective with respect to b equal to zero:

$$
\frac{\partial}{\partial b_{ij}^m}\Big[L_{\text{LS}}(b,\hat{\mu})+\frac{1}{\gamma}\sum_{(i,j)\in\mathcal{U}^2}\sum_{m'=1}^{\infty}(b_{ij}^{m'})^2\Big]\Big|_{b=\hat{b}}=0,\quad (i,j)\in\mathcal{U}^2,\ \ m\in\{1,2,\dots\}.\tag{E5}
$$

Table F3: Average CPU time in seconds across 10 trials. $\tilde{N}$ denotes the average data size per trial.

T	$\tilde{N}$	Bonnet *cpu*	Ours *cpu*
10000	8248	9250	19.1
15000	12748	26406	29.5

Table F4: Average CPU time in seconds across 5 trials.

U	Exp *cpu*	Gau *cpu*	Ber *cpu*	Bonnet *cpu*	Ours *cpu*
3	124	7.03	10.8	413	5.04
15	1369	300	254	8513	29.9

Equation (E5) can be written explicitly as

$$\begin{aligned}
&\frac{1}{2}\frac{\partial}{\partial b_{ij}^m}\left[L_{\mathrm{LS}}(b,\hat{\mu}) + \frac{1}{\gamma}\sum_{(i,j)\in\mathcal{U}^2}\sum_{m'=1}^{\infty}(b_{ij}^{m'})^2\right]\Big|_{b=\hat{b}} \\
&= \int_0^T \Big(\hat{\mu}_i + \sum_{n'\in\mathcal{N}}\sum_{m'=1}^{\infty}\hat{b}_{iu_{n'}}^{m'} e_{m'}(t-t_{n'})\mathbf{1}_{0<t-t_{n'}\le A}\Big)\sum_{n\in\mathcal{N}_j} e_m(t-t_n)\mathbf{1}_{0<t-t_n\le A}dt \\
&\qquad - \sum_{n'\in\mathcal{N}_i}\sum_{n\in\mathcal{N}_j} e_m(t_{n'}-t_n)\mathbf{1}_{0<t_{n'}-t_n\le A} + \frac{1}{\gamma}\hat{b}_{ij}^m \\
&= 0.
\end{aligned} \tag{E6}$$

Operating $\sum_{m=1}^{\infty}\cdot\, e_m(s)$ on both sides of Equation (E6) yields the following system of Fredholm integral equations of the second kind:

$$\begin{aligned}
&\hat{\mu}_i\sum_{n\in\mathcal{N}_j}\int_0^T k(s,t-t_n)\mathbf{1}_{0<t-t_n\le A}dt + \sum_{l\in\mathcal{U}}\int_0^T V_{jl}(s,t)\hat{g}_{il}(t)dt \\
&\qquad - \sum_{n'\in\mathcal{N}_i}\sum_{n\in\mathcal{N}_j} k(s,t_{n'}-t_n)\mathbf{1}_{0<t_{n'}-t_n\le A} + \frac{1}{\gamma}\hat{g}_{ij}(s) = 0,
\end{aligned} \tag{E7}$$

where we used the relation, $\hat{g}_{ij}(s) = \sum_{m=1}^{\infty}\hat{b}_{ij}^m\ e_m(s)$, and the kernel trick, $k(t,s) = \sum_{m=1}^{\infty} e_m(t)e_m(s)$; here, $V_{jl}(s,t)$ is defined in Equation (6). The resulting system of integral equations coincides with Equation (7) in the proof shown in the main text. Therefore, the remainder of the proof proceeds as in Equations (7-8), which completes the proof. ■

F Additional Experiments

F.1 Scalability on Larger Data Size

In Section 2.4, we discussed the scalability of the proposed method on the data size. To confirm it, we conducted an additional experiment in the refractory scenario with $T \in \{10000, 15000\}$, and evaluated the CPU times of `Ours` and `Bonnet` on these larger datasets. The results in Table F3 demonstrate that `Ours` remains scalable for the larger data sizes.

F.2 Scalability on Larger Dimensionality

The computational cost of our method (`Ours`) scales cubically with the dimensionality U, which is a disadvantage compared to the quadratic scaling of the prior kernel method (`Bonnet`). We conducted an additional experiment under a refractory scenario with $T = 2000$ and $U = 15$ to examine this

Table F5: Results of `Exp` (Bonnet et al., 2023), `Gau` (Xu et al., 2016), `Ber` (Lemonnier & Vayatis, 2014), `Bonnet` (Bonnet & Sangnier, 2025), and `Ours` on financial dataset across 10 trials with standard errors in brackets. nll is the negative log-likelihood on test data. N is the data size per trial. cpu is the CPU time in seconds. The performances not significantly ($p \geq 0.01$) different from the best one under the Mann-Whitney U test (Holm, 1979) are shown in bold.

	`Exp`		`Gau`		`Ber`		`Bonnet`		`Ours`	
N	nll	cpu	nll	cpu	nll	cpu	nll	cpu	nll	cpu
3319	8086 (720)	38.2 (15.8)	2899 (1015)	7.17 (0.69)	**1487** (1291)	13.3 (1.40)	**1829** (1147)	550 (71.3)	**1721** (896)	**3.4** (1.43)

issue. The triggering kernel matrix was constructed as a $U \times U$ block-diagonal matrix obtained by placing copies of the 3×3 triggering kernel matrix, $g(s) = [g_{ij}(s)]_{ij}$, used in Section 4.2, along the diagonal. We fixed the hyperparameters to $(\gamma, \beta) = (1, 1)$ and evaluated only the computation time.

The results in Table F4 (note that the $U = 3$ case is identical to that reported in Table 2 for $T = 2000$) show that all methods exhibited an increase in computation time as U grows. However, the increase for `Ours` is more moderate compared to the conventional methods (`Exp`, `Gau`, `Ber`, and `Bonnet`). Although this trend may contradict the complexity analysis presented in Section 2.4, it can be attributed to the fact that our method relies solely on matrix additions and matrix inversions (performed via Cholesky decomposition), which are highly amenable to parallelization across multiple CPU cores.

F.3 Effects of Hyperparameter Grid on Performance

For the proposed model `Ours`, we conducted an additional experiment under the refractory scenario with $T = 5000$, where the grid was refined from 3×3 to 10×10. The resulting squared error Δ^2 was 0.58 ± 0.12, which represents only a marginal improvement over the 3×3 grid (0.59 ± 0.13). This result suggests that the performance in Tables 1-2, especially the gap between `Bonnet` and `Ours`, could not be solely attributed to the hyperparameter tuning strategy. Since `Bonnet` is based on the likelihood function, it is expected to achieve higher accuracy than `Ours`, which relies on the least squares loss. Note that maximum likelihood estimation is known to be statistically efficient asymptotically for Hawkes processes (see (Ogata, 1978)).

F.4 Experiments on real-world data

We conducted an experiment on a financial dataset that is widely used for evaluating Hawkes process models (Du et al., 2016). This dataset contains transaction records of a single stock over one day, with two event types ($U = 2$): "buy" and "sell". The event sequence is further partitioned by timestamps. From the 100 sequences available on the GitHub repository of Zuo et al. (2020)[3], each of which size is 3319, we randomly constructed 10 pairs of sequences; for a pair, one was used for model training, and the other was used to evaluate the negative log-likelihood as the predictive error (the lower, the better). For our proposed method, we applied a post-hoc clipping, $\max(\hat{\lambda}(t), 10^{-2})$ to the estimated intensity function $\hat{\lambda}(\cdot)$. For the models except `Exp`, the support window A was set at 3, and the hyperparameters were optimized on the grids of $\gamma \in \{0.01, 0.1, 1.0\}$ and $\beta \in \{0.5, 1.0, 10\}$, based on the negative log-likelihood minimization. All other procedures followed those described in the main experiments.

Table F5 summarizes the results: compared to the baseline methods, our approach achieves robust performance while remaining computationally efficient.

F.5 Effects of support window on performance

The support window A can be regarded as a hyperparameter that controls the shape of the triggering kernel and can therefore be selected from data, for example, via cross-validation. In general, if A

[3] https://github.com/SimiaoZuo/Transformer-Hawkes-Process

Table F6: Performance of `Ours` regarding the support window A on mutually-exciting scenario (T = 5000) across 10 trials with standard errors in brackets. Δ^2 and cpu denote the integrated squared error and the CPU time in seconds, respectively.

$A = 1$		$A = 2$		$A = 5$		$A = 10$		$A = 20$	
Δ^2	cpu	Δ^2	cpu	Δ^2	cpu	Δ^2	cpu	Δ^2	cpu
0.51	1.75	0.24	3.89	0.20	10.7	0.29	22.5	0.33	39.1
(0.04)	(1.24)	(0.03)	(2.74)	(0.06)	(7.71)	(0.11)	(16.9)	(0.07)	32.4

Table F7: Training time (in seconds) of `Ours` regarding the dimensionality U across 5 trials with standard errors in brackets.

$U = 10$	$U = 50$	$U = 100$	$U = 200$	$U = 300$	$U = 500$
cpu	cpu	cpu	cpu	cpu	cpu
0.68	5.05	26.9	160	478	2010
(0.26)	(0.82)	(0.67)	(1.91)	(6.70)	(15.5)

is too small, estimation methods cannot capture the true shape of the underlying triggering kernel. Conversely, if A is too large, non-negligible values remain in regions where the true triggering kernel is essentially zero, thereby increasing the estimation error. It is worth noting that choosing A too small is substantially more detrimental than choosing A too large; therefore, it is preferable to set A on the larger side.

To verify this behavior, we evaluated the predictive performance of the proposed method `Ours` for the support window $A \in \{1, 2, 5, 10, 20\}$, using data from the mutually-exciting scenario (T = 5000). Here, predictive performance was assessed using the integrated squared error (Δ^2) defined as follows:

$$\Delta^2 = \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} \int_0^{20} \left| g_{ij}(s) - \hat{g}_{ij}(s) \right|^2 ds, \tag{F1}$$

where $g_{ij}(s)$ and $\hat{g}_{ij}(s)$ denote the true and estimated triggering kernels, respectively. Table F6 summarizes the results, suggesting that an optimal value of A exists.

F.6 Scaling Curve vs. U

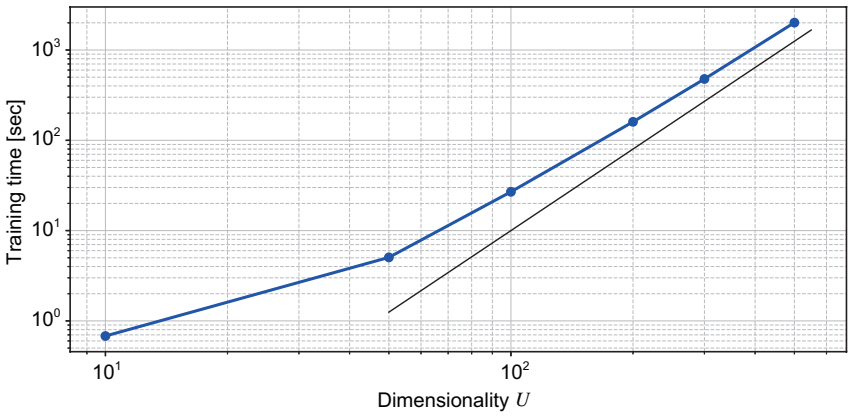

Figure F1: Scaling curve of the training time regarding the dimensionality. The blue line represents the training time of `Ours` with respect to the dimensionality U. The black line represents a reference curve of $\mathcal{O}(U^3)$.

We conducted an additional experiment to empirically characterize how the computational cost of our method scales with respect to U. The computational burden of solving the linear system is most evident in settings where the number of events per dimension is small while U itself is large.

Table F8: Estimation accuracy of triggering kernels and predictive performance by `Ours`. Results are on the mutually exciting scenario dataset, averaged over 10 trials, with standard errors in brackets. nll is the negative log-likelihood on the test data. Δ^2 denotes the integrated squared error for triggering kernel estimation, of which results are a reproduction of those presented in Table 1.

$T = 2000$		$T = 3000$		$T = 5000$	
nll	Δ^2	nll	Δ^2	nll	Δ^2
3294	0.38	3013	0.27	2842	0.20
(481)	(0.15)	(226)	(0.06)	(194)	(0.16)

Accordingly, we set $\mu_i = 0.1$, $g_{ij}(\cdot) = 0$, and $T = 100$, and generated synthetic datasets for $U \in \{10, 50, 100, 200, 300, 500\}$. We then measured the training time of the proposed method. Table F7 and Figure F1 summarize the results: empirically, the training time grows sub-cubically in U up to around $U \simeq 100$, and the scaling curve gradually approaches $\mathcal{O}(U^3)$. At least up to $U = 500$, the observed scaling is therefore better than $\mathcal{O}(U^3)$. We attribute this behavior largely to the highly optimized implementation of Cholesky factorization in TensorFlow, which efficiently exploits the 12-core CPU architecture.

Additionally, we implemented a naïve conjugate gradient (CG) solver to compare its runtime with the Cholesky factorization (CF) approach, using a very simple preconditioner given by the inverse of the diagonal of $(\frac{1}{\gamma} I_{MU} + \Xi)$. Unfortunately, this configuration led to worse runtime than CF, suggesting that a substantially more suitable preconditioner is required to fully benefit from CG. Designing such a preconditioner is nontrivial and beyond what can reasonably be accomplished within the rebuttal period. Based on these findings, our current model should be regarded as being practically targeted at event data with up to a few hundred dimensions.

F.7 Effects of Estimation Accuracy of Triggering Kernels on Predictive Performance

We conducted an additional experiment to examine the relationship between the estimation accuracy of the latent triggering kernel and the predictive performance of the point process model. Based on the mutually-exciting scenario dataset in Section 4.1, we estimated the triggering kernel for each trial of $T \in \{2000, 3000, 5000\}$ by using the proposed model (`Ours`), and evaluated its predictive performance on $T = 7000$ data. Here, predictive performance was assessed using the negative log-likelihood (the lower, the better). Since the $T = 7000$ data is composed of 10 trials, we report the average predictive performance over these trials. In addition, because each training dataset also contains 10 trials, we repeated the above prediction experiment 10 times. We applied a post-hoc clipping, $\max(\hat{\lambda}(t), 10^{-4})$, to the estimated intensity function $\hat{\lambda}(\cdot)$. The results, summarized in Table F8, show that the predictive performance of the point process model indeed improves as the estimation accuracy of the latent triggering kernel increases.

G Existence of Equivalent Kernels

We demonstrate that the equivalent kernels, $\{h_j(\cdot, \cdot)\}_{j \in \mathcal{U}}$, defined by the system of Fredholm integral equations (6) exist and are uniquely determined for any positive semi-definite kernel $k(\cdot, \cdot)$.

A positive semi-definite kernel can be represented as the inner product of an M-dimensional feature vector, $\boldsymbol{\phi}(\cdot) = (\phi_1(\cdot), \ldots, \phi_M(\cdot))^\top$, for some $M \leq \infty$. Under this representation, the equivalent kernels, $\{h_j(\cdot, \cdot)\}_{j \in \mathcal{U}}$, can be expressed by Equation (11). The block matrix $\boldsymbol{\Xi} = [\boldsymbol{\Xi}_{ij}] \in \mathbb{R}^{MU \times MU}$ in Equation (11) has a set of submatrices defined in Equation (12), which can be rewritten as follows:

$$\begin{aligned} \boldsymbol{\Xi}_{ij} &= \sum_{n \in \mathcal{N}_i} \sum_{n' \in \mathcal{N}_j} \mathbf{1}_{\max(t_n, t_{n'}) < \min(T, A+t_n, A+t_{n'})} \int_{\max(t_n, t_{n'})}^{\min(T, A+t_n, A+t_{n'})} \boldsymbol{\phi}(t - t_n) \boldsymbol{\phi}(t - t_{n'})^\top dt \\ &= \sum_{n \in \mathcal{N}_i} \sum_{n' \in \mathcal{N}_j} \int_0^T \xi_n(t) \xi_{n'}(t) \boldsymbol{\phi}(t - t_n) \boldsymbol{\phi}(t - t_{n'})^\top dt, \end{aligned} \tag{G1}$$

where $\xi_n(t) = \mathbf{1}_{t_n<t<\min(T,A+t_n)}$. Therefore, for all $\boldsymbol{c} = (c_{11}, \ldots, c_{1M}, c_{21}, \ldots, c_{UM})^\top \in \mathbb{R}^{MU}$, the following inequality holds:

$$
\begin{aligned}
&\boldsymbol{c}^\top \boldsymbol{\Xi}\, \boldsymbol{c} \\
&= \sum_{i\in\mathcal{U}} \sum_{j\in\mathcal{U}} \sum_{m=1}^{M} \sum_{m'=1}^{M} c_{im} c_{jm'} \sum_{n\in\mathcal{N}_i} \sum_{n'\in\mathcal{N}_j} \int_0^T \xi_n(t)\xi_{n'}(t)\phi_m(t-t_n)\phi_{m'}(t-t_{n'})dt, \\
&= \int_0^T \left[\sum_{i\in\mathcal{U}} \sum_{m=1}^{M} c_{im} \sum_{n\in\mathcal{N}_i} \xi_n(t)\phi_m(t-t_n)\right] \left[\sum_{j\in\mathcal{U}} \sum_{m'=1}^{M} c_{jm'} \sum_{n'\in\mathcal{N}_j} \xi_{n'}(t)\phi_{m'}(t-t_{n'})\right] dt \quad \text{(G2)} \\
&= \int_0^T \left[\sum_{i\in\mathcal{U}} \sum_{m=1}^{M} c_{im} \sum_{n\in\mathcal{N}_i} \xi_n(t)\phi_m(t-t_n)\right]^2 dt \\
&\geq 0.
\end{aligned}
$$

This relation shows that $\left(\frac{1}{\gamma}\boldsymbol{I}_{MU} + \boldsymbol{\Xi}\right)$ in Equation (11) is positive definite and invertible for $\gamma > 0$. Therefore, the equivalent kernels defined by Equations (6) and (11) exist and are uniquely determined for any positive semi-definite kernels $k(\cdot,\cdot)$.