
K²IE: Kernel Method-based Kernel Intensity Estimators for Inhomogeneous Poisson Processes

Hideaki Kim¹ Tomoharu Iwata¹ Akinori Fujino¹

Abstract

Kernel method-based intensity estimators, formulated within reproducing kernel Hilbert spaces (RKHSs), and classical kernel intensity estimators (KIEs) have been among the most easy-to-implement and feasible methods for estimating the intensity functions of inhomogeneous Poisson processes. While both approaches share the term “kernel”, they are founded on distinct theoretical principles, each with its own strengths and limitations. In this paper, we propose a novel regularized kernel method for Poisson processes based on the least squares loss and show that the resulting intensity estimator involves a specialized variant of the representer theorem: it has the dual coefficient of unity and coincides with classical KIEs. This result provides new theoretical insights into the connection between classical KIEs and kernel method-based intensity estimators, while enabling us to develop an efficient KIE by leveraging advanced techniques from RKHS theory. We refer to the proposed model as the *kernel method-based kernel intensity estimator* (K²IE). Through experiments on synthetic datasets, we show that K²IE achieves comparable predictive performance while significantly surpassing the state-of-the-art kernel method-based estimator in computational efficiency.

1. Introduction

Poisson processes have been the gold standard for modeling point patterns that occur randomly in multi-dimensional domains. They are characterized by an intensity function, that is, the instantaneous probability of events occurring at any point in the domain, which allows us to

assess the risk of experiencing events at specified domains and forecast the timings/locations of future events. Poisson processes have a variety of applications in reliability engineering (Lai & Xie, 2006), clinical research (Cox, 1972; Clark et al., 2003; Lánczky & Györfy, 2021), seismology (Ogata, 1988), epidemiology (Gatrell et al., 1996), ecology (Heikkinen & Arjas, 1999), and more.

Kernel intensity estimators (KIEs) are the simplest non-parametric approaches to estimating intensity functions (Ram-lau-Hansen, 1983; Diggle, 1985), with advantages that include superior computational efficiency and theoretical tractability. They represent the underlying intensity function as a sum of smoothing kernels¹ evaluated at data points, where rescaled versions of density functions are usually adopted as smoothing kernels to correct the edge effects, that is, the estimation biases around the edges of observation domains.

Recently, Flaxman et al. (2017) developed a feasible Reproducing Kernel Hilbert Space (RKHS) formulation for inhomogeneous Poisson processes. They showed that the representer theorem (Wahba, 1990; Schölkopf et al., 2001) holds for a penalized maximum likelihood estimation under the constraint that the square root of the intensity function lies in an RKHS: the obtained square root of the intensity estimator is given by a linear combination of transformed RKHS kernels¹ evaluated at data points. The transformed RKHS kernels, often referred to as the *equivalent RKHS kernels*, naturally account for edge effects through likelihood functions, and the intensity estimator has been shown to outperform KIEs in scenarios where edge effects are prominent, such as in high-dimensional domains. Although the kernel method-based intensity estimator has a form similar to KIEs, it requires fitting the dual coefficient using a gradient descent method, making it less favorable than KIEs in terms of computational efficiency.

In this paper, we propose a penalized least squares loss formulation for estimating intensity functions under the con-

¹NTT Corporation, Japan. Correspondence to: Hideaki Kim <hideaki.kin@ntt.com>.

¹Traditionally, the term “kernel” is used to refer both to the weights assigned to data points in kernel density estimation and to the positive-definite kernels that define an RKHS. To avoid ambiguity, this paper employs two distinct terms “smoothing” and “RKHS” kernels, following Flaxman et al. (2017).

straint that the intensity function resides in an RKHS. The least squares loss is motivated by the empirical risk minimization principle (van de Geer, 2000) and has demonstrated notable computational advantages in recent studies on Poisson processes (Hansen et al., 2015; Bacry et al., 2020; Cai et al., 2024). Utilizing advanced variational analysis via path integral representation (Kim, 2021), we show that a specialized variant of the representer theorem holds for the functional optimization problem: the resulting intensity estimator, which we call the *kernel method-based kernel intensity estimator* (K²IE), has the unit dual coefficient and requires no optimization of dual coefficients given a kernel hyper-parameter, which is consistent with KIEs; furthermore, it employs the equivalent RKHS kernels appeared in Flaxman’s model and can effectively address edge effects based on the RKHS theory. This result not only establishes a significant theoretical connection between classical KIEs and kernel method-based intensity estimators but also enables the development of a more scalable intensity estimator based on kernel methods.

In Section 2, we outline related works of intensity estimation. In Section 3, we derive K²IE via functional analysis with path integral representation of RKHS norm. In Section 4, we compare K²IE with conventional nonparametric intensity estimators on synthetic datasets, and confirm the effectiveness of the proposed method². Finally, Section 5 states our conclusions.

2. Background

Let a set of N point events, $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$, being observed in a d -dimensional compact space, $\mathcal{X} \subset \mathbb{R}^d$. We consider a learning problem of intensity function in the framework of inhomogeneous Poisson processes, where intensity function, $\lambda(\mathbf{x}) : \mathcal{X} \rightarrow \mathbb{R}_+$, represents an instantaneous probability of events occurring at any point in $\mathcal{X}$:

$$\lambda(\mathbf{x}) = \lim_{|d\mathbf{x}| \rightarrow 0} \mathbb{E}[\mathcal{N}(d\mathbf{x})] / |d\mathbf{x}|, \quad (1)$$

where $\mathcal{N}(d\mathbf{x})$ is the number of events occurring in $d\mathbf{x} \subset \mathcal{X}$, and $|\cdot|$ represents the measure of domain.

One of the most significant applications of Poisson processes lies in evaluating the risk of experiencing events (e.g., traffic accidents and disaster events) within specified regions. Given an intensity function $\lambda(\mathbf{x})$, the probability distribution of event counts (i.e., the Poisson distribution) over an arbitrary compact region $\mathcal{S} \subset \mathcal{X}$, denoted by $P_{\mathcal{S}}(\cdot)$, is calculated as follows:

$$P_{\mathcal{S}}(n) = \frac{\Lambda^n e^{-\Lambda}}{n!}, \quad \Lambda = \int_{\mathcal{S}} \lambda(\mathbf{x}) d\mathbf{x}, \quad (2)$$

where $n \in \{0, 1, 2, \dots\}$. Using Equation (2), we can per-

form binary classification to determine the occurrence of future events within $\mathcal{S}$.

2.1. Kernel Intensity Estimator

Kernel smoothing is a classical approach to nonparametric intensity estimation (Diggle, 1985), expressed as:

$$\hat{\lambda}(\mathbf{x}) = \sum_{n=1}^N g(\mathbf{x}, \mathbf{x}_n) / \nu(\mathbf{x}), \quad \nu(\mathbf{x}) = \int_{\mathcal{X}} g(\mathbf{x}, \mathbf{s}) d\mathbf{s}, \quad (3)$$

where $g : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ represents a non-negative smoothing kernel¹, and $\nu(\mathbf{x})$ is an edge-correction term³. This method, commonly referred to as the kernel intensity estimator (KIE), is closely related to the well-known kernel density estimator (Parzen, 1962; Davis et al., 2011), especially with bounded support (Jones, 1993). KIEs offer several advantages, including theoretical tractability, superior computational efficiency, and ease of implementation.

The smoothing kernel $g(\mathbf{x}, \mathbf{x}')$ involves a bandwidth hyperparameter, which can be optimized using standard techniques such as cross-validation (Cronie et al., 2024) or Silverman’s rule-of-thumb (Silverman, 2018). It is worth noting that widely-used cross-validation methods involving test log-likelihood functions require the integration of intensity functions over a test region $\mathcal{S} \subset \mathcal{X}$, where KIEs need to rely on time-consuming Monte Carlo integration because $g(\mathbf{x}, \cdot) / \nu(\mathbf{x})$ in (3) usually cannot be integrated in a closed form (e.g., Gaussian smoothing kernels).

2.2. Kernel Method-based Intensity Estimator

Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ denote a continuous positive semi-definite kernel. Then there exists a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ (Schölkopf & Smola, 2018; Shawe-Taylor & Cristianini, 2004) associated with kernel $k(\cdot, \cdot)$. Flaxman et al. (2017) modeled the intensity function as the square of a latent function in the RKHS,

$$\lambda(\mathbf{x}) = f^2(\mathbf{x}), \quad f(\cdot) \in \mathcal{H}_k, \quad (4)$$

and proposed a regularized minimization problem with log-likelihood loss functional as follows:

$$\min_{f \in \mathcal{H}_k} \left\{ - \sum_{n=1}^N \log(f^2(\mathbf{x}_n)) + \int_{\mathcal{X}} f^2(\mathbf{x}) d\mathbf{x} + \frac{1}{\gamma} \|f\|_{\mathcal{H}_k}^2 \right\}, \quad (5)$$

where $\|\cdot\|_{\mathcal{H}_k}^2$ represents the squared Hilbert space norm, and γ represents the regularization hyper-parameter. Through Mercer’s theorem (Mercer, 1909), Flaxman et al. (2017) showed that the representer theorem (Wahba, 1990; Schölkopf et al., 2001) does hold in an appropriately trans-

²Codes are available at: <https://github.com/HidKim/K2IE>

³Intensity estimates near the edge of the domain $\mathcal{X}$ are biased downwards since no points are observed outside $\mathcal{X}$.

formed RKHS, resulting in the following solution:

$$\hat{f}(\mathbf{x}) = \sum_{n=1}^N \alpha_n h(\mathbf{x}, \mathbf{x}_n), \quad (6)$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)^\top$ is the dual coefficient, and $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a transformed RKHS kernel defined in terms of the Mercer expansion of $k(\mathbf{x}, \mathbf{x}')$ as

$$h(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{\infty} \frac{\eta_m}{1/\gamma + \eta_m} e_m(\mathbf{x}) e_m(\mathbf{x}'), \quad (7)$$

$$\int_{\mathcal{X}} k(\mathbf{x}, \mathbf{s}) e_m(\mathbf{s}) d\mathbf{s} = \eta_m e_m(\mathbf{x}),$$

where $\{e_m(\cdot)\}_{m=1}^{\infty}$ is the eigenfunctions of the integral operator $\int_{\mathcal{X}} k(\mathbf{x}, \mathbf{s}) d\mathbf{s}$. Recently, Kim et al. (2022) rewrote the definition (7) in terms of a Fredholm integral equation of the second kind (Polyanin & Manzhirov, 1998),

$$\frac{1}{\gamma} h(\mathbf{x}, \mathbf{x}') + \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{s}) h(\mathbf{s}, \mathbf{x}') d\mathbf{s} = k(\mathbf{x}, \mathbf{x}'), \quad (8)$$

which enables us to utilize the established approximation techniques for solving Fredholm integral equations (Polyanin & Manzhirov, 1998; Atkinson, 2010). We will discuss how to solve Equation (8) in Section 3.2. The transformed RKHS kernel is referred to as the *equivalent RKHS kernel* (Flaxman et al., 2017; Walder & Bishop, 2017; Kim et al., 2022).

The dual coefficient $\boldsymbol{\alpha}$ in the intensity estimator (6) solves the following dual optimization problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \left\{ -\sum_{n=1}^N \log \sum_{n'=1}^N \alpha_{n'} h(\mathbf{x}_n, \mathbf{x}_{n'}) + \frac{1}{\gamma} \boldsymbol{\alpha}^\top \mathbf{H} \boldsymbol{\alpha} \right\}, \quad (9)$$

where $\mathbf{H} := [h(\mathbf{x}_n, \mathbf{x}_{n'})]_{nn'}$. The computational complexity of solving (9) is naively $\mathcal{O}(qN^2)$ for q iterations of gradient descent methods, but reduces to $\mathcal{O}(qMN)$ when the equivalent RKHS kernel is given in degenerate form with rank M ($< N$) such that $h(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^M \psi_m(\mathbf{x}) \psi_m(\mathbf{x}')$.

Flaxman’s intensity estimator (6) differs from the classical KIE (3) in that it does not require explicit edge correction. Instead, the equivalent RKHS kernel $h(\cdot, \cdot)$ naturally accounts for the effects of the finite observation domain $\mathcal{X} \subset \mathbb{R}^d$ through the second term on the left-hand side of Equation (8). Flaxman et al. (2017) demonstrated that the intensity estimator (6) outperforms the KIE in terms of predictive performance, especially in high-dimensional settings. However, it demands the model fitting (9) unlike KIEs, which makes KIEs more favorable regarding computational efficiency.

Other Related Works

Gaussian Cox Processes (GCPs) provide a Bayesian alternative to kernel intensity estimators and kernel method-based models, where Gaussian processes are used to model latent intensity functions via positive-valued link functions (Møller et al., 1998). This framework enables principled interval estimation of intensity functions, along with hyperparameter inference in a fully Bayesian manner (Rathbun & Cressie, 1994; Cunningham et al., 2007; Adams et al., 2009; Diggle et al., 2013; Gunter et al., 2014; Lloyd et al., 2015; Teng et al., 2017; Donner & Opper, 2018; John & Hensman, 2018; Aglietti et al., 2019). Although GCPs are typically more computationally expensive than their non-Bayesian counterparts, extending kernel-based models into the Gaussian process framework can yield efficient Bayesian alternatives. For instance, Walder & Bishop (2017) proposed a Bayesian variant of Flaxman’s model, known as the permanent process, and Sellier & Dellaportas (2023) further extended this approach within the generalized stationary kernel framework.

While neural network-based methods often trade computational efficiency for expressive power, Tsuchida et al. (2024) recently introduced the squared neural family, a model that simultaneously achieves expressiveness and analytical tractability. Like Flaxman’s model, it ensures non-negativity and closed-form expressiveness of the intensity function through the use of a squared link function—an elegant property that merits further attention.

3. Method

3.1. Kernel Method-based Kernel Intensity Estimator

In this paper, we introduce the least squares loss functional for Poisson processes (Hansen et al., 2015) given by

$$-2 \sum_{n=1}^N \lambda(\mathbf{x}_n) + \int_{\mathcal{X}} \lambda(\mathbf{x})^2 d\mathbf{x}. \quad (10)$$

This loss functional comes from the empirical risk minimization principle (van de Geer, 2000) and has demonstrated notable computational advantages in recent Poisson process literature (Hansen et al., 2015; Bacry et al., 2020; Cai et al., 2024). For readers unfamiliar with the loss defined in (10), we briefly explain the origin of the term *squares loss* in Appendix A. We model the intensity function as a latent function in RKHS $\mathcal{H}_k$, and consider the problem of minimization of the penalized least squares loss as follows:

$$\min_{\lambda \in \mathcal{H}_k} \left\{ -2 \sum_{n=1}^N \lambda(\mathbf{x}_n) + \int_{\mathcal{X}} \lambda(\mathbf{x})^2 d\mathbf{x} + \frac{1}{\gamma} \|\lambda\|_{\mathcal{H}_k}^2 \right\}, \quad (11)$$

where $\|\cdot\|_{\mathcal{H}_k}^2$ represents the squared Hilbert space norm, and γ represents the regularization hyper-parameter.

Through variational analysis, Theorem 1 below shows that the resulting kernel method-based intensity estimator is consistent with the classical KIE (3). While the main proof relies on the path integral representation (Kim, 2021), for completeness, we also provide an alternative derivation based on Mercer’s theorem in Appendix B. To the best of our knowledge, this paper is the first to prove that the representer theorem holds for the penalized minimization of the least squares loss for the intensity estimation in RKHS.

Theorem 1. *The solution of the functional optimization problem (11), denoted as $\hat{\lambda}(\cdot)$, involves the representer theorem under a transformed RKHS kernel $h(\cdot, \cdot)$ defined by Equation (8), and its dual coefficient is equal to unity:*

$$\hat{\lambda}(\mathbf{x}) = \sum_{n=1}^N h(\mathbf{x}, \mathbf{x}_n), \quad \mathbf{x} \in \mathcal{X}. \quad (12)$$

Proof. Let $\mathcal{K} \cdot (\mathbf{x}) = \int_{\mathcal{X}} \cdot k(\mathbf{x}, \mathbf{s}) d\mathbf{s}$ be the integral operator with RKHS kernel $k(\cdot, \cdot)$, and $\mathcal{K}^* \cdot (\mathbf{x}) = \int_{\mathcal{X}} \cdot k^*(\mathbf{x}, \mathbf{s}) d\mathbf{s}$ be its inverse operator. Then, through the path integral representation of Gaussian processes (Kim, 2021), the squared Hilbert space norm can be represented in a functional form,

$$\|\lambda\|_{\mathcal{H}_k}^2 = \iint_{\mathcal{X} \times \mathcal{X}} k^*(\mathbf{x}, \mathbf{s}) \lambda(\mathbf{x}) \lambda(\mathbf{s}) d\mathbf{x} d\mathbf{s}.$$

Using the representation, the objective functional in Equation (11) can be rewritten as follows:

$$S(\lambda) = -2 \sum_{n=1}^N \lambda(\mathbf{x}_n) + \iint_{\mathcal{X} \times \mathcal{X}} q^*(\mathbf{x}, \mathbf{s}) \lambda(\mathbf{x}) \lambda(\mathbf{s}) d\mathbf{x} d\mathbf{s},$$

where $q^*(\cdot, \cdot)$ is the weighted sum of $k^*(\cdot, \cdot)$ and the Dirac delta function $\delta(\cdot)$,

$$q^*(\mathbf{x}, \mathbf{s}) = \delta(\mathbf{x} - \mathbf{s}) + \frac{1}{\gamma} k^*(\mathbf{x}, \mathbf{s}). \quad (13)$$

The solution of Equation (11), $\hat{\lambda}(\mathbf{x})$, is obtained by solving the equation where the functional derivative of $S(\hat{\lambda})$ is equal to zero:

$$\frac{\delta S}{\delta \hat{\lambda}} = -2 \sum_{n=1}^N \delta(\mathbf{x} - \mathbf{x}_n) + 2 \int_{\mathcal{X}} q^*(\mathbf{x}, \mathbf{s}) \hat{\lambda}(\mathbf{s}) d\mathbf{s} = 0.$$

Let $\mathcal{Q}^* \cdot (\mathbf{x}) = \int_{\mathcal{X}} \cdot q^*(\mathbf{x}, \mathbf{s}) d\mathbf{s}$ be the integral operator associated with $q^*(\cdot, \cdot)$, and $\mathcal{Q} \cdot (\mathbf{x}) = \int_{\mathcal{X}} \cdot q(\mathbf{x}, \mathbf{s}) d\mathbf{s}$ be its inverse operator. Then applying operator $\mathcal{Q}$ to the equation, $\delta S / \delta \hat{\lambda} = 0$, leads to a representation of the form,

$$\hat{\lambda}(\mathbf{x}) = \sum_{n=1}^N q(\mathbf{x}, \mathbf{x}_n),$$

where the relation, $(\mathcal{Q}\mathcal{Q}^*) \cdot (\mathbf{x}) = \int_{\mathcal{X}} \cdot \delta(\mathbf{x} - \mathbf{s}) d\mathbf{s}$, was used. Furthermore, the following derivation shows that $q(\cdot, \cdot)$ is

equal to the equivalent RKHS kernel $h(\cdot, \cdot)$ defined by (8): applying operator $\mathcal{Q}$ to Equation (13) leads to the relation,

$$\delta(\mathbf{x} - \mathbf{x}') = q(\mathbf{x}, \mathbf{x}') + \frac{1}{\gamma} \int_{\mathcal{X}} k^*(\mathbf{x}, \mathbf{s}) q(\mathbf{s}, \mathbf{x}') d\mathbf{s},$$

and applying operator $\mathcal{K}$ to both sides of the relation yields:

$$k(\mathbf{x}, \mathbf{x}') = \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{s}) q(\mathbf{s}, \mathbf{x}') d\mathbf{s} + \frac{1}{\gamma} q(\mathbf{x}, \mathbf{x}'),$$

which is identical to Equation (8). ■

Theorem 1 demonstrates, under the least squares loss functional, a strong connection between classical KIEs and modern kernel methods. From the perspective of KIE theory, Theorem 1 implies that the equivalent RKHS kernels $h(\cdot, \cdot)$ are smoothing kernels constructed based on RKHS kernels. Hence, we call the proposed model (12) the *kernel method-based kernel intensity estimator* (K²IE). As Flaxman et al. (2017) discussed, the equivalent RKHS kernels implicitly incorporate edge effects in an effective manner. Therefore, our K²IE is expected to combine the computational efficiency of KIEs with the effectiveness of Flaxman’s kernel method-based estimator.

Similar to the conventional kernel method-based estimator (6), the support of K²IE in Theorem 1 lies within the observation domain $\mathcal{X}$, i.e., it concerns interpolation. However, by broadening the support of the RKHS kernel $k(\cdot, \cdot)$, the support in Theorem 1 can be naturally extended: In other words, K²IE defined by Equation (12) can be applied in its current form to extrapolation as well. A proof of this claim is provided in Appendix C.

Unlike conventional methods, K²IE has the limitation of not guaranteeing the non-negativity of intensity functions. The equivalent RKHS kernels may generally take negative values, and since K²IE is constructed as a linear combination of the equivalent RKHS kernels, it can yield negative values in certain regions, particularly in areas with no observed events. This issue is caused by the fact that K²IE models intensity function by an RKHS function $f(\cdot) \in \mathcal{H}_k$, while conventional methods by $\sigma(f(\cdot))$ for a non-negative link function $\sigma(\cdot)$. In practice, K²IE does not have large negative values because the second term of the objective function (11) penalizes them. Thus we can deal with the issue by applying $\max(u, 0)$ for intensity-related values u , such as $u = \lambda(\mathbf{x})$ and $u = \int_{\mathcal{S}} \lambda(\mathbf{x}) d\mathbf{x}$ over a domain $\mathcal{S}$.

3.2. Construction of Equivalent RKHS Kernel

The primary task in K²IE is to derive the equivalent RKHS kernel that satisfies the integral equation (8). The methodology varies depending on whether the observation domain $\mathcal{X}$ is infinite or finite, as elaborated in the subsequent

sections. Here, we assume that RKHS kernels are shift-invariant, i.e., $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}')$, which includes popular RKHS kernels such as Gaussian, Matérn, and Laplace kernels.

3.2.1. INFINITE OBSERVATION DOMAIN

If the observation domain is infinite, i.e., $\mathcal{X} = \mathbb{R}^d$, the integral equation (8) can be solved by using the Fourier transform as follows:

$$h(\mathbf{x} - \mathbf{x}') = \mathcal{F}^{-1} \left[\frac{\tilde{k}(\boldsymbol{\omega})}{\gamma^{-1} + \tilde{k}(\boldsymbol{\omega})} \right] (\mathbf{x} - \mathbf{x}'), \quad (14)$$

where $\mathcal{F}^{-1}[\cdot](\mathbf{x})$ denotes the inverse Fourier transform, and $\tilde{k}(\boldsymbol{\omega} \in \mathbb{R}^d)$ represents the Fourier transform of the shift-invariant RKHS kernel $k(\mathbf{x} - \mathbf{x}')$. Notably, the equivalent RKHS kernel $h(\cdot, \cdot)$ is also shift-invariant due to the symmetry of the integral equation (8). Approximation methods are required because the inverse Fourier transform in (14) generally cannot be expressed in closed form. One promising approach is the random feature map (Rahimi & Recht, 2007), where the equivalent RKHS kernel is approximated via Monte Carlo sampling from a probability distribution, $p(\cdot) \propto \tilde{k}(\cdot)/(\gamma^{-1} + \tilde{k}(\cdot))$, such that $h(\mathbf{x} - \mathbf{x}') = \mathbb{E}_{\boldsymbol{\omega} \sim p(\cdot)} [\exp(i\boldsymbol{\omega}^\top \mathbf{x}) \exp(i\boldsymbol{\omega}^\top \mathbf{x}')] = \mathbb{E}_{\boldsymbol{\omega} \sim p(\cdot)} [\exp(i\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}'))]$. Another feasible approach is to apply the fast Fourier transform to (14).

When $\mathcal{X} = \mathbb{R}^d$, the edge-correction term in KIE (3) vanishes, suggesting that the choice of the smoothing kernels $g(\mathbf{x}, \mathbf{x}')$ in KIE (3) is effectively equivalent to the selection of the equivalent RKHS kernels $h(\mathbf{x}, \mathbf{x}')$ in K²IE (12). Through $h(\mathbf{x}, \mathbf{x}')$, however, we could find smoothing kernels more robust to the squared error than popular ones such as Gaussian smoothing kernels. It is an interesting topic, but this paper focused on the case of a finite observation domain, where edge correction plays a crucial role.

3.2.2. FINITE OBSERVATION DOMAIN

Next, we consider a scenario where the observation domain is expressed as a union of a finite number of hyper-rectangular regions:

$$\mathcal{X} = \bigcup_{j=1}^J \mathcal{X}_j, \quad \mathcal{X}_j = \prod_{i=1}^d [X_{ij}^{\min}, X_{ij}^{\max}], \quad (15)$$

where J denotes the number of hyper-rectangular regions, and $\mathcal{X}_j \cap \mathcal{X}_{j'} \neq \emptyset$. While prior studies typically assume a single hyper-rectangular region, the assumption (15) enables us to deal with more complicated observation domains, such as disjoint or irregularly shaped regions, often encountered in practical applications.

The Fredholm integral equation (8) generally cannot be solved in closed form, and Flaxman et al. (2017) proposed

using Nyström approximation (Williams & Seeger, 2000), which approximates the integral term through numerical integration. While this approach has the advantage of being applicable to any RKHS kernel, it could potentially degrade the accuracy of the edge correction because the integral term is critical for the edge correction. To address the issue, we adopt the degenerate approach (Kim et al., 2022), which approximates RKHS kernels using $2M$ random Fourier features (Rahimi & Recht, 2007),

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &\simeq \sum_{m=1}^{2M} \phi_m(\mathbf{x}) \phi_m(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}'), \quad (16) \\ \phi_m(\mathbf{x}) &= M^{-1/2} \cos(\boldsymbol{\omega}_m^\top \mathbf{x} + \theta_m) \\ \boldsymbol{\omega}_{m \leq M} &\sim \tilde{k}(\boldsymbol{\omega}), \quad \boldsymbol{\omega}_{m > M} = \boldsymbol{\omega}_{m-M}, \\ \theta_{m \leq M} &= 0, \quad \theta_{m > M} = -\pi/2, \end{aligned}$$

and allows the integral term to be handled without any error as follows:

$$\begin{aligned} h(\mathbf{x}, \mathbf{x}') &= \boldsymbol{\phi}(\mathbf{x})^\top (\gamma^{-1} \mathbf{I}_{2M} + \mathbf{A})^{-1} \boldsymbol{\phi}(\mathbf{x}'), \\ \mathbf{A} &= \sum_{j=1}^J \mathbf{A}^j, \quad \mathbf{A}^j = \int_{\mathcal{X}_j} \boldsymbol{\phi}(\mathbf{x}) \boldsymbol{\phi}(\mathbf{x})^\top d\mathbf{x}, \end{aligned} \quad (17)$$

where $\mathbf{I}_{2M}$ represents the identity matrix of size $2M$. Notably, $2M \times 2M$ matrix $\mathbf{A}$, which involves the edge-correction, can be computed in a closed form:

$$\begin{aligned} (\mathbf{A}^j)_{mm'} &= \int_{\mathcal{X}_j} \phi_m(\mathbf{x}) \phi_{m'}(\mathbf{x}) d\mathbf{x} \\ &= \frac{1}{2M} \left[\zeta_j(\boldsymbol{\omega}_m + \boldsymbol{\omega}_{m'}, \theta_m + \theta_{m'}) \right. \\ &\quad \left. + \zeta_j(\boldsymbol{\omega}_m - \boldsymbol{\omega}_{m'}, \theta_m - \theta_{m'}) \right], \\ \zeta_j(\boldsymbol{\omega}, \theta) &= \cos \left[\frac{1}{2} \sum_{i=1}^d \omega^i (X_{ij}^{\max} + X_{ij}^{\min}) + \theta \right] \\ &\quad \cdot \prod_{i=1}^d (X_{ij}^{\max} - X_{ij}^{\min}) \operatorname{sinc} \left[\frac{1}{2} \omega^i (X_{ij}^{\max} - X_{ij}^{\min}) \right], \end{aligned} \quad (18)$$

where $\operatorname{sinc}(x) = \sin(x)/x$ is the unnormalized sinc function, and $\boldsymbol{\omega} = (\omega^1, \dots, \omega^d)^\top$. The relation (17) suggests that the equivalent kernel $h(\mathbf{x}, \mathbf{x}')$ has degenerate form of rank $2M$, which is obtained through Cholesky decomposition as $h(\mathbf{x}, \mathbf{x}') = (\mathbf{L}\boldsymbol{\phi}(\mathbf{x}))^\top (\mathbf{L}\boldsymbol{\phi}(\mathbf{x}'))$, where $\mathbf{L}^\top \mathbf{L} = (\gamma^{-1} \mathbf{I}_{2M} + \mathbf{A})^{-1}$. To enhance the approximation accuracy of the random Fourier features, we employed the quasi-Monte Carlo feature maps (Yang et al., 2014) in this paper.

The degenerate form of equivalent kernel (17) offers an additional advantage. For cross-validation with the least squares loss, K²IE needs to evaluate the integral of the

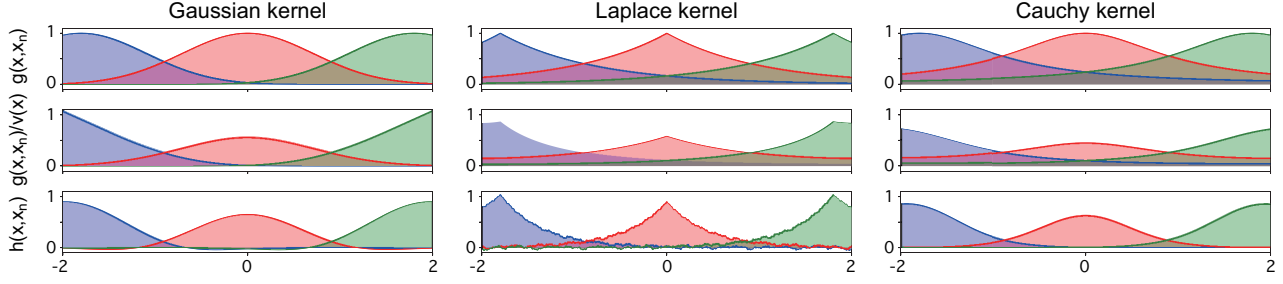


Figure 1: Examples of smoothing kernels $g(x, x_n)$, smoothing kernels with edge-correction $g(x, x_n)/\nu(x)$, and equivalent kernels $h(x, x_n)$ for data points $x_n \in \{-1.8, 0, 1.8\}$. Gaussian, Laplace, and Cauchy RKHS/smoothing kernels are $e^{-|x-x_n|^2}$, $e^{-|x-x_n|}$, and $\frac{1}{1+|x-x_n|^2}$, respectively. The regularization hyper-parameter, the number of random features, and the observation domain were set as $(\gamma, 2M, \mathcal{X}) = (2, 500, [-2, 2])$.

squared intensity function, $\int_{\mathcal{X}} (\sum_n h(x, x_n))^2 dx$, which requires $\mathcal{O}(N^2)$ computation naively or $\mathcal{O}(NZ)$ computation with $Z(\gg 1)$ points Monte Carlo integration. But the integral of the squared intensity function can be obtained analytically with $\mathcal{O}(M^2 + MN)$ computation under (17) as follows:

$$\begin{aligned} \int_{\mathcal{X}} dx \left[\sum_n h(x, x_n) \right]^2 &= \boldsymbol{\xi}^\top \mathbf{A} \boldsymbol{\xi}, \\ \boldsymbol{\xi} &= (\gamma^{-1} \mathbf{I}_{2M} + \mathbf{A})^{-1} \left(\sum_n \phi(x_n) \right). \end{aligned} \quad (19)$$

Therefore, regarding hyperparameter tuning, KIE and FIE require MC integration and solving a dual optimization problem for each cross-validation, respectively, whereas K^2IE requires neither, which is beneficial especially in multi-dimensional settings.

The comparison of K^2IE with KIE suggests that from the viewpoint of KIE, the primary distinction between them lies in how smoothing kernels with edge-correction are constructed: In KIE, the smoothing kernels with edge-correction are constructed by rescaling density functions with their integrals over the observation domain; In contrast, K^2IE constructs smoothing kernels as the solution to the integral equation (8). During model training, KIE benefits from more computational efficiency than K^2IE , which requires solving the integral equation. However, K^2IE offers computational advantages during inference as it can perform the intensity function integration needed in predictive tasks (e.g., see Equation (2)) analytically, while KIE relies on Monte Carlo integration. Furthermore, as demonstrated by (Flaxman et al., 2017), the smoothing kernel in K^2IE , i.e., the equivalent kernel is expected to achieve more effective edge correction, particularly in high-dimensional domain settings.

Figure 1 illustrates examples of smoothing kernels with and without edge correction in KIE, as well as the equiv-

alent RKHS kernels in K^2IE , showing that both the edge-corrected smoothing kernels and the equivalent RKHS kernels assign greater weight to data points near the boundary of the observation domain ($|x| \simeq 2$) compared to those at the center. Interestingly, K^2IE applies edge correction more conservatively through the equivalent RKHS kernels $h(x, x_n)$, that is, differentiates the weights between the center and the boundary less significantly compared to KIE with $g(x, x_n)/\nu(x)$.

4. Experiments

We evaluated the validity and the potential efficiency of our proposed K^2IE by comparing it with prior nonparametric approaches, including the kernel intensity estimator with edge correction (KIE) (Diggle, 1985) and Flaxman’s kernel method-based intensity estimator (FIE) (Flaxman et al., 2017), using synthetic datasets. For K^2IE and FIE, the number of random features $2M$ was fixed at 500 (see Appendix D for an ablation study on the feature number $2M$).

For both the smoothing and RKHS kernels, we employed a multiplicative Gaussian function, $z(\mathbf{x}, \mathbf{x}') = e^{-|\boldsymbol{\beta} \circ (\mathbf{x} - \mathbf{x}')|^2}$, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^\top$ is the inverse scale hyper-parameter, and $\circ$ denotes the Hadamard product. KIE optimized the hyper-parameter $\boldsymbol{\beta}$ through 5-fold cross-validation based on the negative log-likelihood function; FIE optimized the hyper-parameters, $(\boldsymbol{\beta}, \gamma)$, using the same cross-validation procedure as KIE; For K^2IE , the hyper-parameters, $(\boldsymbol{\beta}, \gamma)$, were optimized via 5-fold cross-validation with the least squares loss function (10). For all models, the Monte Carlo cross-validation with p -thinning (Cronie et al., 2024) was adopted, where p was fixed at 0.6. A 10×10 logarithmic grid search was conducted for $\gamma \in [0.1, 100]$ and $\boldsymbol{\beta} \in [0.1, 100] \cdot \bar{\boldsymbol{\beta}}$, where $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \dots, \bar{\beta}_d)^\top$ for $\bar{\beta}_i = 1/[\max_j(X_{ij}^{\max}) - \min_j(X_{ij}^{\min})]$. For FIE, the gradient descent algorithm Adam (Kingma & Ba, 2014) was employed to solve the dual optimization problem (9).

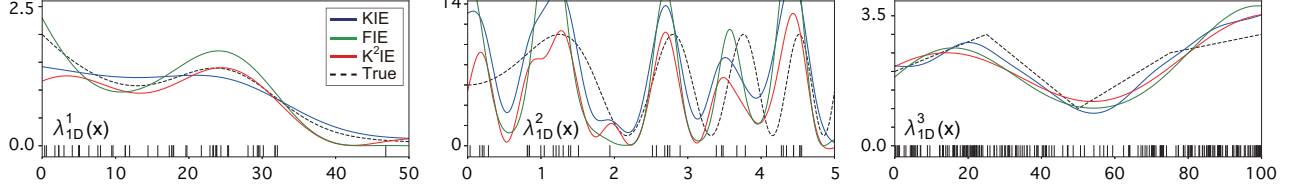


Figure 2: Examples of the estimated intensity functions on 1D synthetic data. The vertical lines represent the locations of observed events.

Table 1: Results on 1D synthetic data across 100 trials with standard errors in brackets. $\tilde{N}$ denotes the average data size per trial. The performances not significantly ($p > 10^{-2}$) different from the best one under the Mann-Whitney U test (Holm, 1979) are shown in bold for L^2 and $|L|$.

	$\lambda_{1D}^1(x) : \tilde{N}=46$				$\lambda_{1D}^2(x) : \tilde{N}=33$				$\lambda_{1D}^3(x) : \tilde{N}=226$			
	$L^2 \downarrow$	$ L \downarrow$	$\rho \uparrow$	$cpu \downarrow$	$L^2 \downarrow$	$ L \downarrow$	$\rho \uparrow$	$cpu \downarrow$	$L^2 \downarrow$	$ L \downarrow$	$\rho \uparrow$	$cpu \downarrow$
KIE	0.09 (0.07)	0.23 (0.08)	—	—	12.6 (2.82)	2.97 (0.28)	—	—	0.15 (0.07)	0.30 (0.08)	—	—
FIE	0.11 (0.11)	0.24 (0.09)	0.34	1.06 (0.17)	13.2 (3.41)	3.04 (0.28)	0.46	0.29 (0.30)	0.17 (0.09)	0.33 (0.09)	0.33	0.86 (0.39)
K²IE	0.12 (0.08)	0.26 (0.08)	0.26	0.01 (0.00)	13.9 (5.03)	3.09 (0.45)	0.48	0.01 (0.00)	0.18 (0.08)	0.34 (0.09)	0.31	0.01 (0.00)
	$10 \times \lambda_{1D}^1(x) : \tilde{N}=466$				$10 \times \lambda_{1D}^2(x) : \tilde{N}=328$				$10 \times \lambda_{1D}^3(x) : \tilde{N}=2250$			
	$L^2 \downarrow$	$ L \downarrow$	$\rho \uparrow$	$cpu \downarrow$	$L^2 \downarrow$	$ L \downarrow$	$\rho \uparrow$	$cpu \downarrow$	$L^2 \downarrow$	$ L \downarrow$	$\rho \uparrow$	$cpu \downarrow$
KIE	1.43 (1.03)	0.87 (0.29)	—	—	289 (71.3)	13.5 (1.92)	—	—	2.84 (1.68)	1.29 (0.34)	—	—
FIE	1.74 (1.53)	0.93 (0.39)	0.49	1.77 (0.13)	277 (80.6)	13.0 (2.09)	0.64	0.55 (0.33)	2.70 (1.79)	1.25 (0.37)	0.63	0.61 (0.13)
K²IE	1.67 (0.71)	0.92 (0.36)	0.49	0.01 (0.00)	266 (74.6)	12.7 (1.98)	0.77	0.01 (0.00)	3.24 (2.08)	1.34 (0.41)	0.47	0.01 (0.00)

Predictive performance was assessed using the integrated squared error (L^2) and the integrated absolute error ($|L|$) (Kowalczyk & Kozłowski, 1998), defined as follows:

$$L^2 = \frac{1}{|\mathcal{X}|} \int_{\mathcal{X}} (\lambda^*(x) - \hat{\lambda}(x))^2 dx, \quad (20)$$

$$|L| = \frac{1}{|\mathcal{X}|} \int_{\mathcal{X}} |\lambda^*(x) - \hat{\lambda}(x)| dx,$$

where $\lambda^*(x)$ and $\hat{\lambda}(x)$ denote the true and estimated intensity functions, respectively. Following (Flaxman et al., 2017), the fraction of times that L^2 is smaller than KIE across the trials, denoted by ρ , was also reported, where ρ was not defined for KIE. Efficiency was evaluated based on the CPU time (in seconds), cpu , required to execute the model fitting given the optimized hyper-parameters.

All models were implemented using TensorFlow-2.10² and executed on a MacBook Pro equipped with a 12-core CPU (Apple M2 Max), with the GPU disabled.

4.1. 1D Synthetic Data

In accordance with previous studies (Adams et al., 2009; John & Hensman, 2018; Aglietti et al., 2019; Kim, 2021),

we generated 1D datasets based on three types of intensity functions:

$$\begin{aligned} \lambda_{1D}^1(x) &= 2e^{-x/15} + e^{-(x-25)/10}^2, & \mathcal{X} &= [0, 50], \\ \lambda_{1D}^2(x) &= 5 \sin(x^2) + 6, & \mathcal{X} &= [0, 5], \\ \lambda_{1D}^3(x) &= \text{piecewise linear function}, & \mathcal{X} &= [0, 100], \end{aligned} \quad (21)$$

where $\lambda_{1D}^3(x)$ passes through the points: (0, 2), (25, 3), (50, 1), (75, 2.5), and (100, 3). Furthermore, to evaluate the scalability of K²IE and FIE with respect to data size, we generated 1D datasets using intensity functions scaled by a factor of ten, denoted by $10 \times \lambda_{1D}^q(x)$ for $q \in \{1, 2, 3\}$. For each intensity function, we simulated 100 trial sequences and performed intensity estimation 100 times using the compared methods.

Table 1 displays the predictive performance on the 1D synthetic datasets. It shows that our proposal K²IE matched the predictive performances of FIE across all three datasets, while achieving significantly faster model fitting in terms of CPU time. K²IE achieved comparable predictive performances with KIE on $\lambda_{1D}^2(x)$, but was outperformed by KIE on $\lambda_{1D}^1(x)$ and $\lambda_{1D}^3(x)$. This result is consistent with (Flax-

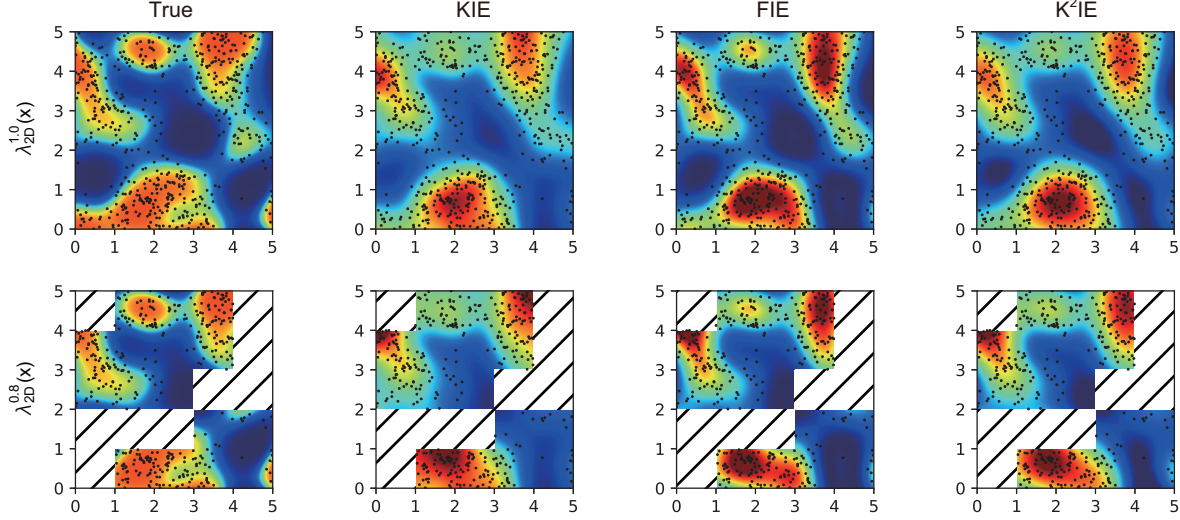


Figure 3: Examples of the estimated intensity functions on 2D synthetic data $\lambda_{2D}^{1.0}$ and $\lambda_{2D}^{0.8}$. The black dots represent the locations of observed events, and the unobserved regions are indicated by hatched lines.

Table 2: Results on 2D synthetic data across 100 trials with standard errors in brackets. $\tilde{N}$ denotes the average data size per trial. The notation follows Table 1.

	$\lambda_{2D}^{1.0}(x) : \tilde{N}=543$				$\lambda_{2D}^{0.9}(x) : \tilde{N}=483$				$\lambda_{2D}^{0.8}(x) : \tilde{N}=428$			
	$L^2 \downarrow$	$ L \downarrow$	$\rho \uparrow$	$cpu \downarrow$	$L^2 \downarrow$	$ L \downarrow$	$\rho \uparrow$	$cpu \downarrow$	$L^2 \downarrow$	$ L \downarrow$	$\rho \uparrow$	$cpu \downarrow$
KIE	63.3 (8.96)	6.36 (0.40)	— —	— —	63.5 (8.92)	6.35 (0.45)	— —	— —	64.5 (10.9)	6.34 (0.52)	— —	— —
FIE	56.47 (12.2)	5.38 (0.60)	0.80 —	1.54 (0.34)	59.8 (13.4)	5.53 (0.62)	0.71 —	1.45 (0.33)	62.3 (13.5)	5.64 (0.65)	0.64 —	1.50 (0.34)
K²IE	53.0 (10.2)	5.54 (0.49)	0.97 —	0.16 (0.03)	55.1 (11.1)	5.63 (0.51)	0.90 —	0.14 (0.02)	57.9 (12.2)	5.77 (0.55)	0.85 —	0.13 (0.03)

man et al., 2017), demonstrating that KIE performed very well in low-dimensional settings. It is worth noting that the discrepancy of predictive performances between KIE and K²IE became negligible on $10 \times \lambda_{1D}^{1,2,3}(x)$, where the dataset size increases. Also, Table 1 demonstrates that the CPU time of FIE increases with the data size, while that of K²IE remains nearly constant. Figure 2 displays some estimation results.

4.2. 2D Synthetic Data

Following the procedure in (Lloyd et al., 2015), we generated a 2D dataset from a sigmoidal Gaussian Cox process. Specifically, we first sampled a 2D function from a Gaussian process with an RKHS kernel, $k(x, x') = e^{-|x-x'|^2/2}$, over the domain $\mathcal{X} = [0, 5] \times [0, 5]$. The intensity function was then obtained by applying a sigmoid link function, $\sigma(z) = 50/(1 + e^{-20z})$, to the sampled function. Using the intensity function, we simulated 100 trials of event data and conducted intensity estimation 100 times

using the compared methods. The resulting dataset contained approximately 540 data points per trial.

In this study, we considered a scenario where the observation domain was divided into $5 \times 5 = 25$ sub-domains,

$$\mathcal{X} = \bigcup_{j=1}^{25} \mathcal{X}_j, \quad \mathcal{X}_j : \text{evenly partitioned 2D domain}, \quad (22)$$

with some of the sub-domains being missing. For each trial of the dataset, we randomly selected each sub-domain with a probability $p \in \{1.0, 0.9, 0.8\}$, thereby generating three datasets, denoted as $\lambda_{2D}^{1.0}(x)$, $\lambda_{2D}^{0.9}(x)$, and $\lambda_{2D}^{0.8}(x)$, respectively.

Table 2 displays the predictive performance on the 2D synthetic datasets, which shows that K²IE and FIE consistently outperformed KIE in all datasets. This result suggests that K²IE and FIE could more effectively handle edge effects than KIE in multi-dimensional settings. Notably, regarding the integrated squared error L^2 , K²IE achieved superior

predictive performance, on average, than FIE, despite both methods employing the same equivalent kernels (with hyperparameters optimized individually for each model). It might be due to the fact that KIE is based on the minimization of the least squares loss (see Section 3.1). Another possible explanation for this result is that the optimization of the dual coefficient required in FIE may become unstable. Indeed, [John & Hensman \(2018\)](#) reported that FIE can yield unreasonable solutions for highly modulating intensity functions. In contrast, K²IE is expected to work more robustly, as it does not require the optimization of dual coefficients. Figure 2 displays some estimation results on $\lambda_{2D}^{1,0}(x)$ and $\lambda_{2D}^{0,8}(x)$.

Additional experiments with a scalable Bayesian approach and on a real-world dataset are provided in Appendix E.

5. Discussions

We have proposed a novel penalized least squares loss formulation for estimating intensity functions that resides in an RKHS. Through the path integral representation of the squared Hilbert space norm, we showed that the optimization problem encompasses a representer theorem, and derived a feasible intensity 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-based estimator.

LIMITATIONS AND FUTURE WORK

As noted at the end of Section 3.1, a key limitation of our K²IE lies in its lack of a general guarantee for the non-negativity of the resulting intensity function. To investigate the effect, we conducted an analysis of how frequently K²IE produces negative values using the 2D synthetic dataset $\lambda_{2D}^{1,0}$. Specifically, we evaluated the estimated intensity values at 500×500 grid points within the observation domain and computed the ratio of negative values. The mean $\pm$ standard deviation of this ratio across 100 trials was 0.059 ± 0.016 , indicating that K²IE can indeed produce negative estimates in practice—particularly in regions with sparse data—highlighting the necessity of a post-hoc correction such as clipping via $\max(\hat{\lambda}(x), 0)$ in applications where negative intensity values are not permitted. As a direction for future work, we explore the technical challenges involved in incorporating non-negativity constraints directly into the functional optimization problem defined in Equation (11).

One natural approach is to model the intensity function as a non-negative transformation $\sigma(f(x))$ of a latent function $f(x)$ residing in an RKHS. In this setting, the functional analysis of the objective in Equation (11) yields the follow-

ing condition that the optimal function $\hat{f}(x)$ must satisfy:

$$\begin{aligned} \frac{1}{\gamma} f(x) + \int_{\mathcal{X}} k(x, s) \sigma(f(s)) \sigma'(f(s)) ds \\ = \sum_n k(x, x_n) \sigma'(f(x_n)), \end{aligned}$$

where $\sigma'(y) = \frac{d\sigma}{dy}(y)$. When $\sigma(y) = y$, the above equation reduces to a Fredholm integral equation, for which Theorem 1 provides a tractable solution. However, if σ is non-linear, even for simple cases like $\sigma(y) = y^2$, the resulting nonlinear integral equation becomes analytically and numerically challenging to solve.

An alternative approach is to impose non-negativity constraints at a finite set of virtual points, which leads to a dual optimization problem. Although this approach may reduce the risk of negative estimates at the virtual points, it neither guarantees global non-negativity nor preserves the computational simplicity of K²IE, due to the added complexity introduced by the dual optimization.

Does K²IE truly fail to guarantee non-negativity in its original form? Interestingly, a sufficient condition to ensure the non-negativity of the equivalent kernels arising in Flaxman’s estimator (and, of course, in K²IE) has been established by [Kim \(2024\)](#). Specifically, when RKHS kernels belong to the class of inverse M-kernels (IMKs), the associated equivalent kernels $h(x, x')$ are guaranteed to be non-negative. This suggests that K²IE, as defined by a sum of equivalent kernels in Equation (12), may be a non-negative intensity estimator whenever the RKHS kernel is an IMK. In one-dimensional cases, the Laplace kernel is known to be an IMK, but no general construction of IMKs is currently known in higher dimensions—posing an intriguing open problem.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

- Adams, R. P., Murray, I., and MacKay, D. J. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *International Conference on Machine Learning*, pp. 9–16, 2009.
- Aglietti, V., Bonilla, E. V., Damoulas, T., and Cripps, S. Structured variational inference in continuous Cox process models. In *Advances in Neural Information Processing Systems* 32, 2019.
- Atkinson, K. 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.
- Bacry, E., Bompaire, M., Gaïffas, S., and Muzy, J.-F. Sparse and low-rank multivariate Hawkes processes. *Journal of Machine Learning Research*, 21(50):1–32, 2020.
- Cai, B., Zhang, J., and Guan, Y. Latent network structure learning from high-dimensional multivariate point processes. *Journal of the American Statistical Association*, 119(545):95–108, 2024.
- Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. Survival analysis part i: basic concepts and first analyses. *British Journal of Cancer*, 89(2):232–238, 2003.
- Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
- Cronie, O., Moradi, M., and Biscio, C. A. A cross-validation-based statistical theory for point processes. *Biometrika*, 111(2):625–641, 2024.
- Cunningham, J. P., Byron, M. Y., Shenoy, K. V., and Sahani, M. Inferring neural firing rates from spike trains using Gaussian processes. In *Advances in Neural Information Processing Systems 20*, 2007.
- Daley, D. J. and Vere-Jones, D. *An Introduction to the Theory of Point Processes*. Springer-Verlag, New York, 1988.
- Davis, R. A., Lii, K.-S., and Politis, D. N. Remarks on some nonparametric estimates of a density function. *Selected Works of Murray Rosenblatt*, pp. 95–100, 2011.
- Diggle, P. A kernel method for smoothing point process data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):138–147, 1985.
- Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.
- Donner, C. and Opper, M. Efficient Bayesian inference of sigmoidal Gaussian Cox processes. *Journal of Machine Learning Research*, 19:1–34, 2018.
- Flaxman, S., Teh, Y. W., and Sejdinovic, D. Poisson intensity estimation with reproducing kernels. In *Artificial Intelligence and Statistics*, pp. 270–279. PMLR, 2017.
- Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers*, pp. 256–274, 1996.
- Gunter, T., Lloyd, C., Osborne, M. A., and Roberts, S. J. Efficient Bayesian nonparametric modelling of structured point processes. In *Uncertainty in Artificial Intelligence*, 2014.
- Hansen, N. R., Reynaud-Bouret, P., and Rivoirard, V. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.
- Heikkinen, J. and Arjas, E. Modeling a Poisson forest in variable elevations: A nonparametric Bayesian approach. *Biometrics*, 55(3):738–745, 1999.
- Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pp. 65–70, 1979.
- Hubbel, S. P. and Foster, R. B. Diversity of canopy trees in a neotropical forest and implications for conservation. *Tropical Rain Forest: Ecology and Management*, pp. 25–41, 1983.
- John, S. T. and Hensman, J. Large-scale Cox process inference using variational Fourier features. In *International Conference on Machine Learning*, volume 80, pp. 2362–2370. PMLR, 2018.
- Jones, M. C. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3:135–146, 1993.
- Kim, H. Fast Bayesian inference for Gaussian Cox processes via path integral formulation. In *Advances in Neural Information Processing Systems 34*, 2021.
- Kim, H. Inverse M-kernels for linear universal approximators of non-negative functions. In *Advances in Neural Information Processing Systems 37*, 2024.
- Kim, H., Asami, T., and Toda, H. Fast Bayesian estimation of point process intensity as function of covariates. In *Advances in Neural Information Processing Systems 35*, 2022.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Kowalczyk, Z. and Kozłowski, J. Integrated squared error and integrated absolute error in recursive identification of continuous-time plants. In *UKACC International Conference on Control’98 (Conf. Publ. No. 455)*, volume 1, pp. 693–698. IET, 1998.

- Lai, C. D. and Xie, M. *Stochastic Ageing and Dependence for Reliability*. Springer Science & Business Media, 2006.
- Lánczky, A. and Györfy, B. Web-based survival analysis tool tailored for medical research (kmplot): development and implementation. *Journal of Medical Internet Research*, 23(7):e27633, 2021.
- Lloyd, C., Gunter, T., Osborne, M., and Roberts, S. Variational inference for Gaussian process modulated Poisson processes. In *International Conference on Machine Learning*, volume 37, pp. 1814–1822. PMLR, 2015.
- Mercer, J. 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.
- Møller, J., Syversveen, A. R., and Waagepetersen, R. P. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.
- Ogata, Y. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.
- Parzen, E. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3): 1065–1076, 1962.
- Polyanin, A. D. and Manzhirov, A. V. *Handbook of Integral Equations*. CRC press, 1998.
- Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, 2007.
- Ramlau-Hansen, H. Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, pp. 453–466, 1983.
- Rathbun, S. L. and Cressie, N. Asymptotic properties of estimators for the parameters of spatial inhomogeneous Poisson point processes. *Advances in Applied Probability*, 26(1):122–154, 1994.
- Scholkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2018.
- Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pp. 416–426. Springer, 2001.
- Sellier, J. and Dellaportas, P. Sparse spectral Bayesian permanent process with generalized kernel. In *International Conference on Artificial Intelligence and Statistics*, pp. 2769–2791. PMLR, 2023.
- Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.
- Teng, M., Nathoo, F., and Johnson, T. D. Bayesian computation for log-Gaussian Cox processes: A comparative analysis of methods. *Journal of Statistical Computation and Simulation*, 87:2227–2252, 2017.
- Tsuchida, R., Ong, C. S., and Sejdinovic, D. Exact, fast and expressive Poisson point processes via squared neural families. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20559–20566, 2024.
- van de Geer, S. *Empirical Processes in M-estimation*, volume 6. Cambridge University Press, 2000.
- Wahba, G. *Spline Models for Observational Data*, volume 59. SIAM, 1990.
- Walder, C. J. and Bishop, A. N. Fast Bayesian intensity estimation for the permanent process. In *International Conference on Machine Learning*, volume 70, pp. 3579–3588. PMLR, 2017.
- Williams, C. and Seeger, M. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, 2000.
- Yang, J., Sindhwani, V., Avron, H., and Mahoney, M. Quasi-Monte Carlo feature maps for shift-invariant kernels. In *International Conference on Machine Learning*, pp. 485–493. PMLR, 2014.

A. Explanation of the Least Squares Loss

Let $\mathbb{E}$ denote the expectation with respect to data points generated from the true intensity function $\lambda^*(\mathbf{x})$. We consider the expected integrated squared loss between the estimator $\hat{\lambda}(\mathbf{x})$ and the true intensity function $\lambda^*(\mathbf{x})$, defined as:

$$\begin{aligned} \mathbb{E} \left[\int_{\mathcal{X}} |\hat{\lambda}(\mathbf{x}) - \lambda^*(\mathbf{x})|^2 d\mathbf{x} \right] &= \mathbb{E} \left[\int_{\mathcal{X}} \hat{\lambda}^2(\mathbf{x}) d\mathbf{x} \right] \\ &\quad - 2\mathbb{E} \left[\int_{\mathcal{X}} \hat{\lambda}(\mathbf{x}) \lambda^*(\mathbf{x}) d\mathbf{x} \right] + \mathbb{E} \left[\int_{\mathcal{X}} \lambda^{*2}(\mathbf{x}) d\mathbf{x} \right]. \end{aligned}$$

The third term on the right-hand side is independent of the estimator and can, therefore, be omitted. The second term

can be decomposed as follows:

$$\begin{aligned} 2\mathbb{E}\left[\int_{\mathcal{X}} \hat{\lambda}(\mathbf{x})\lambda^*(\mathbf{x})d\mathbf{x}\right] &= 2\mathbb{E}\left[\int_{\mathcal{X}} \hat{\lambda}(\mathbf{x})\sum_{n=1}^N\delta(\mathbf{x}-\mathbf{x}_n)d\mathbf{x}\right] \\ &+ 2\mathbb{E}\left[\int_{\mathcal{X}} \hat{\lambda}(\mathbf{x})\left(\lambda^*(\mathbf{x})-\sum_{n=1}^N\delta(\mathbf{x}-\mathbf{x}_n)\right)d\mathbf{x}\right], \end{aligned}$$

where the second term on the right-hand side vanishes due to Campbell's theorem (Daley & Vere-Jones, 1988):

$$\begin{aligned} \int_{\mathcal{X}} \mathbb{E}[\hat{\lambda}(\mathbf{x})]\lambda^*(\mathbf{x})d\mathbf{x} - \sum_{n=1}^N \mathbb{E}[\hat{\lambda}(\mathbf{x}_n)] \\ = \int_{\mathcal{X}} \mathbb{E}[\hat{\lambda}(\mathbf{x})]\lambda^*(\mathbf{x})d\mathbf{x} - \int_{\mathcal{X}} \mathbb{E}[\hat{\lambda}(\mathbf{x})]\lambda^*(\mathbf{x})d\mathbf{x} = 0. \end{aligned}$$

Combining the above expressions yields the following identity:

$$\begin{aligned} \mathbb{E}\left[\int_{\mathcal{X}} |\hat{\lambda}(\mathbf{x}) - \lambda^*(\mathbf{x})|^2 d\mathbf{x}\right] \\ = \mathbb{E}\left[\int_{\mathcal{X}} \hat{\lambda}^2(\mathbf{x})d\mathbf{x} - 2\sum_{n=1}^N \hat{\lambda}(\mathbf{x}_n)\right] + C, \end{aligned}$$

where C is a constant. This shows that the least squares loss defined in (10) corresponds to the empirical integrated squared loss.

B. Proof of Theorem 1 via Mercer's Theorem

We present a proof of Theorem 1 based on Mercer's Theorem, following an approach similar to that of Flaxman et al. (2017).

Proof. Using the Mercer expansion of the RKHS kernel given in Equation (7), any function $\lambda \in \mathcal{H}_k$ can be expressed as $\lambda(\cdot) = \sum_m b_m e_m(\cdot)$, where $\{b_m\}_m$ are the expansion coefficients and the RKHS norm is given by $\|\lambda\|_{\mathcal{H}_k}^2 = \sum_m b_m^2/\eta_m < \infty$. Substituting this into the objective in Equation (11), we obtain:

$$\begin{aligned} -2\sum_{n=1}^N \lambda(\mathbf{x}_n) + \frac{1}{\gamma}\|\lambda\|_{\mathcal{H}_k}^2 + \int_{\mathcal{X}} \lambda(\mathbf{x})^2 d\mathbf{x} \\ = -2\sum_{n=1}^N \lambda(\mathbf{x}_n) + \frac{1}{\gamma}\sum_m b_m^2/\eta_m \\ + \sum_m \sum_{m'} b_m b_{m'} \int_{\mathcal{X}} e_m(\mathbf{x})e_{m'}(\mathbf{x})d\mathbf{x} \\ = -2\sum_{n=1}^N \lambda(\mathbf{x}_n) + \frac{1}{\gamma}\sum_m b_m^2/\eta_m + \sum_m b_m^2 \\ = -2\sum_{n=1}^N \lambda(\mathbf{x}_n) + \sum_m \left(\frac{\eta_m}{\eta_m + 1/\gamma}\right)^{-1} b_m^2, \end{aligned}$$

where the orthogonality condition, $\int_{\mathcal{X}} e_m(\mathbf{x})e_{m'}(\mathbf{x})d\mathbf{x} = \delta_{mm'}$, is used. The above equation shows that if we define a new RKHS kernel $q(\cdot, \cdot)$ as

$$q(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{\infty} \frac{\eta_m}{\eta_m + 1/\gamma} e_m(\mathbf{x})e_m(\mathbf{x}'),$$

the optimization problem in Equation (11) reduces to:

$$\min_{\lambda \in \mathcal{H}_q} \left\{ -2\sum_{n=1}^N \lambda(\mathbf{x}_n) + \|\lambda\|_{\mathcal{H}_q}^2 \right\},$$

where $\|\cdot\|_{\mathcal{H}_q}^2$ represents the squared norm of an RKHS $\mathcal{H}_q$ associated with $q(\cdot, \cdot)$. By construction, $q(\cdot, \cdot)$ coincides with the equivalent kernel defined in Equation (7). According to the classical representer theorem (Schölkopf et al., 2001), the optimal solution to this problem lies in the span of kernel evaluations at the data points:

$$\hat{\lambda}(\mathbf{x}) = \sum_{n=1}^N \alpha_n q(\mathbf{x}, \mathbf{x}_n),$$

where the dual coefficients $\alpha = (\alpha_1, \dots, \alpha_N)^\top$ minimize the objective. Taking the gradient of the objective with respect to α yields:

$$\begin{aligned} \frac{\partial}{\partial \alpha_n} \left[-2\sum_{n=1}^N \lambda(\mathbf{x}_n) + \|\lambda\|_{\mathcal{H}_q}^2 \right] \\ = -2\sum_{n'=1}^N q(\mathbf{x}_{n'}, \mathbf{x}_n) + 2\alpha_n \sum_{n'=1}^N q(\mathbf{x}_{n'}, \mathbf{x}_n) = 0, \\ \therefore \alpha_n = 1. \end{aligned}$$

This completes the proof. ■

C. Extension of Theorem 1

Proposition 2. Let $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a continuous positive semi-definite kernel. Then the solution $\hat{\lambda}(\cdot)$ to the optimization problem (11) admits a representer theorem with respect to a transformed RKHS kernel $h(\cdot, \cdot)$, which is defined via the following Fredholm integral equation:

$$\begin{aligned} \frac{1}{\gamma}h(\mathbf{x}, \mathbf{x}') + \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{s})h(\mathbf{s}, \mathbf{x}')d\mathbf{s} &= k(\mathbf{x}, \mathbf{x}'), \\ (\mathbf{x}, \mathbf{x}') &\in \mathbb{R}^d \times \mathbb{R}^d. \end{aligned} \quad (\text{C1})$$

Moreover, its dual coefficient is equal to unity:

$$\hat{\lambda}(\mathbf{x}) = \sum_{n=1}^N h(\mathbf{x}, \mathbf{x}_n), \quad \mathbf{x} \in \mathbb{R}^d.$$

Proof. Let the integral operator associated with the RKHS kernel $k(\cdot, \cdot)$ be defined as $\mathcal{K} \cdot (\mathbf{x}) = \int_{\mathbb{R}^d} k(\mathbf{x}, \mathbf{s})d\mathbf{s}$, and its

inverse operator be denoted by $\mathcal{K}^* \cdot (\mathbf{x}) = \int_{\mathbb{R}^d} k^*(\mathbf{x}, \mathbf{s}) d\mathbf{s}$. Using the path integral formulation of Gaussian processes (Kim, 2021), the squared norm in the RKHS can be expressed in the functional form:

$$\|\lambda\|_{\mathcal{H}_k}^2 = \iint_{\mathbb{R}^d \times \mathbb{R}^d} k^*(\mathbf{x}, \mathbf{s}) \lambda(\mathbf{x}) \lambda(\mathbf{s}) d\mathbf{x} d\mathbf{s}.$$

Based on this representation, the objective functional in Equation (11) becomes:

$$S(\lambda) = -2 \sum_{n=1}^N \lambda(\mathbf{x}_n) + \iint_{\mathbb{R}^d \times \mathbb{R}^d} h^*(\mathbf{x}, \mathbf{s}) \lambda(\mathbf{x}) \lambda(\mathbf{s}) d\mathbf{x} d\mathbf{s},$$

where $h^*(\cdot, \cdot)$ is defined in terms of $k^*(\cdot, \cdot)$, the Dirac delta function $\delta(\cdot)$, and the indicator function $\mathbf{1}_{(\cdot)}$,

$$h^*(\mathbf{x}, \mathbf{s}) = \delta(\mathbf{x} - \mathbf{s}) \mathbf{1}_{\mathbf{s} \in \mathcal{X}} + \frac{1}{\gamma} k^*(\mathbf{x}, \mathbf{s}). \quad (\text{C2})$$

The minimizer $\hat{\lambda}(\mathbf{x})$ of $S(\lambda)$ satisfies the equation obtained by setting the functional derivative to zero:

$$\frac{\delta S}{\delta \hat{\lambda}} = -2 \sum_{n=1}^N \delta(\mathbf{x} - \mathbf{x}_n) + 2 \int_{\mathbb{R}^d} h^*(\mathbf{x}, \mathbf{s}) \hat{\lambda}(\mathbf{s}) d\mathbf{s} = 0.$$

Define the integral operator corresponding to $h^*(\cdot, \cdot)$ by $\mathcal{H}^* \cdot (\mathbf{x}) = \int_{\mathbb{R}^d} h^*(\mathbf{x}, \mathbf{s}) d\mathbf{s}$, and its inverse operator by $\mathcal{H} \cdot (\mathbf{x}) = \int_{\mathbb{R}^d} h(\mathbf{x}, \mathbf{s}) d\mathbf{s}$. Applying $\mathcal{H}$ to both sides of the functional equation yields:

$$\hat{\lambda}(\mathbf{x}) = \sum_{n=1}^N h(\mathbf{x}, \mathbf{x}_n), \quad \mathbf{x} \in \mathbb{R}^d,$$

where we have used the identity, $(\mathcal{H}\mathcal{H}^*) \cdot (\mathbf{x}) = \int_{\mathbb{R}^d} \delta(\mathbf{x} - \mathbf{s}) d\mathbf{s}$. Furthermore, applying the operator $\mathcal{H}$ to Equation (C2) leads to the relation,

$$\delta(\mathbf{x} - \mathbf{x}') = h(\mathbf{x}, \mathbf{x}') \mathbf{1}_{\mathbf{x} \in \mathcal{X}} + \frac{1}{\gamma} \int_{\mathbb{R}^d} k^*(\mathbf{x}, \mathbf{s}) h(\mathbf{s}, \mathbf{x}') d\mathbf{s},$$

and subsequent application of the operator $\mathcal{K}$ results in

$$k(\mathbf{x}, \mathbf{x}') = \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{s}) h(\mathbf{s}, \mathbf{x}') d\mathbf{s} + \frac{1}{\gamma} h(\mathbf{x}, \mathbf{x}'),$$

$$(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^d \times \mathbb{R}^d,$$

which is identical to Equation (C1). ■

D. Ablation Study on the Number of Random Features

We conducted an ablation study to investigate the effect of the number of random features ($2M$) on the predictive performance of K^2IE using the 2D synthetic dataset $\lambda_{2D}^{1,0}$. As

Table D1: Predictive performance of K^2IE on the 2D synthetic data $\lambda_{2D}^{1,0}$ as a function of the number of feature maps. Brackets represent standard errors over 100 trials.

	$2M$	20	100	300	500
L^2		147 (8.28)	75.4 (10.4)	53.2 (10.7)	53.0 (10.2)
$ L $		9.807 (0.26)	6.681 (0.45)	5.56 (0.52)	5.54 (0.49)

shown in Table D1, both the integrated squared error and the integrated absolute error consistently decrease as M increases. These results indicate that K^2IE benefits from more random features, and that the setting $2M = 500$, used in Section 4, provides sufficiently accurate and stable estimates.

E. Additional Experiments

E.1. Comparison with a Variational Bayesian model

We conducted an additional experiment on the 2D synthetic dataset $\lambda_{2D}^{1,0}$ to compare against a scalable Bayesian model. Here, we adopted a variational Bayesian approach based on a Gaussian Cox process with a quadratic link function (Lloyd et al., 2015), where a Gaussian RKHS kernel and 10×10 inducing points were employed. We employed a gradient descent algorithm, *Adam* (Kingma & Ba, 2014), to perform the model fitting, where the number of iterations and the learning parameter were set as 5000 and 0.01, respectively. L^2 , $|L|$, and *cpu* achieved by the Bayesian model were 63.9 (12.2), 5.55 (0.46), and 51.8 (32.2), respectively, where standard deviations are in brackets. The result highlights the high efficiency of K^2IE .

E.2. Comparison on a Real-world Dataset

We conducted an additional experiment using an open 2D real-world dataset, *bei*, in the R package *spatsta* (GPL-3). It consists of locations of 3605 trees of the species *Beilschmiedia pendula* in a tropical rain forest (Hubbel & Foster, 1983).

Following (Cronie et al., 2024), we randomly labeled the data points with independent and identically distributed marks $\{1, 2, 3\}$ from a multinomial distribution with parameters $(p_1, p_2, p_3) = (0.3, 0.3, 0.7)$, and assigned the points with label 1 and 2 to training data and test data, respectively; we repeated it 100 times for evaluation. A 10×10 logarithmic grid search was conducted for $\gamma \in [0.001, 1]$ and $\beta \in [0.1, 100] \cdot \bar{\beta}$, where $\bar{\beta} = (\bar{\beta}_1, \dots, \bar{\beta}_d)^\top$ for $\bar{\beta}_i = 1/[\max_j(X_{ij}^{\max}) - \min_j(X_{ij}^{\min})]$.

Let the observation domain $\mathcal{X}$ be regularly divided into 10×10 sub-domains as $\mathcal{X} = \bigcup_{j=1}^{100} \mathcal{X}_j$. We evaluated the pre-

Table E1: Results on the real-world data *bei* across 100 trials with standard errors in brackets. The notation follows Table 1.

	$L_s \downarrow$	$L_c \downarrow$	cpu
KIE	-5.80 (0.32)	267 (11.5)	– –
FIE	-5.16 (0.26)	287 (15.1)	5.15 (1.57)
K ² IE	-6.16 (0.44)	279 (13.2)	0.17 (0.04)

dictive performance of the estimator $\hat{\lambda}(x)$ based on the test least squares loss (L_s) and the test negative log-likelihood of counts (L_c):

$$L_s = \int_{\mathcal{X}} \hat{\lambda}^2(\mathbf{x}) d\mathbf{x} - 2 \sum_{n \in D_{\text{test}}} \hat{\lambda}(\mathbf{x}_n),$$

$$L_c = \sum_{j=1}^{100} \left[\hat{\Lambda}_j - N_j \log \hat{\Lambda}_j + \log(N_j!) \right], \quad \hat{\Lambda}_j = \int_{\mathcal{X}_j} \hat{\lambda}(\mathbf{x}) d\mathbf{x},$$

where D_{test} denotes the test data, and N_j represents the number of test data points observed within $\mathcal{X}_j$. Table E1 displays the results, showing that K²IE achieved the best performance on L_s but was outperformed by KIE on L_c , which could be because the hyperparameters were optimized based on the least squares loss and the log-likelihood for K²IE and KIE, respectively.