Poisson intensity estimation with reproducing kernels

Seth Flaxman **Yee Whye Teh** **Dino Sejdinovic**
Department of Statistics, Oxford

Abstract

Despite the fundamental nature of the inhomogeneous Poisson process in the theory and application of stochastic processes, and its attractive generalizations (e.g. Cox process), few tractable nonparametric modeling approaches of intensity functions exist, especially in high dimensional settings. In this paper we develop a new, computationally tractable Reproducing Kernel Hilbert Space (RKHS) formulation for the inhomogeneous Poisson process. We model the square root of the intensity as an RKHS function. The modeling challenge is that the usual representer theorem arguments no longer apply due to the form of the inhomogeneous Poisson process likelihood. However, we prove that the representer theorem does hold in an appropriately transformed RKHS, guaranteeing that the optimization of the penalized likelihood can be cast as a tractable finite-dimensional problem. The resulting approach is simple to implement, and readily scales to high dimensions and large-scale datasets.

1 INTRODUCTION

Poisson processes are ubiquitous in statistical science, with a long history spanning both theory (e.g. Kingman (1993)) and applications (e.g. Diggle et al. (2013)), especially in the spatial statistics and time series literature. Despite their ubiquity, fundamental questions in their application to real datasets remain open. Namely, scalable nonparametric models for intensity functions of inhomogeneous Poisson processes are not well understood, especially in multiple dimensions since the standard approaches, based on kernel smoothing, are akin to density estimation. In this contribution, we propose a step towards such scalable nonparametric modeling and introduce a new Reproducing Kernel Hilbert Space (RKHS) formulation for inhomogeneous Poisson process modeling, which is based on the Empirical Risk Minimization (ERM) framework. We model the square root of the intensity as an RKHS function and consider a risk functional given by a penalized version of the inhomogeneous Poisson process likelihood. However, standard representer theorem arguments do not apply directly due to the form of the likelihood. Namely, the fundamental difference arises since the observation that *no points* occur in some region is just as important as the locations of the points that do occur. Thus, the likelihood depends not only on the evaluations of the intensity at the observed points, but also on its integral across the domain of interest. As we will see, this difficulty can be overcome by appropriately adjusting the RKHS under consideration. We prove a version of the representer theorem in this adjusted RKHS, which coincides with the original RKHS as a space of functions but has a different inner product structure. This allows us to cast the estimation problem as an optimization over a finite-dimensional subspace of the adjusted RKHS. The derived method is demonstrated to give better performance than a naïve unadjusted RKHS method which resorts to an optimization over a subspace without representer theorem guarantees. We describe cases where adjusted RKHS can be described with explicit Mercer expansions and propose numerical approximations where Mercer expansions are not available. We observe strong performance of the proposed method on a variety of synthetic, environmental, crime and bioinformatics data.

2 BACKGROUND AND RELATED WORK

2.1 Poisson process

We briefly state relevant definitions for point processes over domains $S \subset \mathbb{R}^D$, following Cressie and Wikle (2011). For Lebesgue measurable subsets $T \subset S$, $N(T)$ denotes the number of events in $T \subset S$. $N(\cdot)$ is a

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

stochastic process characterizing the point process. Our focus is on providing a nonparametric estimator for the first-order intensity of a point process, which is defined as:

$$\lambda(s) = \lim_{|ds|\to 0} \mathbb{E}[N(ds))]/|ds|. \quad (1)$$

The inhomogeneous Poisson process is driven solely by the intensity function $\lambda(\cdot)$:

$$N(T) \sim \text{Poisson}(\int_T \lambda(x)dx). \quad (2)$$

In the homogeneous Poisson process, $\lambda(x) = \lambda$ is constant, so the number of points in any region T simply depends on the volume of T, which we denote $|T|$:

$$N(T) \sim \text{Poisson}(\lambda|T|). \quad (3)$$

For a given intensity function $\lambda(\cdot)$, the likelihood of a set of $N = N(S)$ points $x_1, \ldots, x_N$ observed over a domain S is given by:

$$\mathcal{L}(x_1, \ldots, x_N|\lambda(\cdot)) = \prod_{i=1}^{N} \lambda(x_i) e^{-\int_S \lambda(x)dx} \quad (4)$$

2.2 Reproducing Kernel Hilbert Spaces

Given a non-empty domain S and a positive definite kernel function $k : S \times S \to \mathbb{R}$, there exists a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$. An RKHS is a space of functions $f : S \to \mathbb{R}$, in which evaluation is a continuous functional, meaning it can be represented by an inner product $f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}_k}$ for all $f \in \mathcal{H}_k, x \in S$ (this is known as the reproducing property), cf. Berlinet and Thomas-Agnan (2004). While $\mathcal{H}_k$ is in most interesting cases an infinite-dimensional space of functions, due to the classical representer theorem (Kimeldorf and Wahba, 1971), (Schölkopf and Smola, 2002, Section 4.2), optimization over $\mathcal{H}_k$ is typically a tractable finite-dimensional problem. In particular, if we have a set of N observations $x_1, \ldots, x_N$, $x_i \in S$ and consider the problem:

$$\min_{f\in\mathcal{H}_k} \left\{R\left(f(x_1), \ldots, f(x_N)\right) + \Omega\left(\|f\|_{\mathcal{H}_k}\right)\right\}. \quad (5)$$

where $R\left(f(x_1), \ldots, f(x_N)\right)$ depends on f through its evaluations on the set of observations only, and Ω is a non-decreasing function of the RKHS norm of f, there exists a solution to Eq. (5) of the form $f^*(\cdot) = \sum_{i=1}^{N} \alpha_i k(x_i, \cdot)$, and the optimization can thus be cast in terms of $\alpha \in \mathbb{R}^N$. This formulation is widely used in the framework of regularized Empirical Risk Minimization (ERM) for supervised learning, where $R\left(f(x_1), \ldots, f(x_N)\right) = \frac{1}{N}\sum_{i=1}^{N} L(f(x_i), y_i)$ is the empirical risk corresponding to a loss function L.

If domain S is compact and kernel k is continuous, one can assign to k its integral kernel operator $T_k : L_2(S) \to L_2(S)$, given by $T_k g = \int_S k(x, \cdot)g(x)dx$, which is positive, self-adjoint and compact. There thus exists an orthonormal set of eigenfunctions $\{e_j\}_{j=1}^{\infty}$ of T_k, and the corresponding eigenvalues $\{\eta_j\}_{j=1}^{\infty}$. This spectral decomposition of T_k leads to Mercer's representation of kernel function k (Schölkopf and Smola, 2002, Section 2.2):

$$k(x, x') = \sum_{j=1}^{\infty} \eta_j e_j(x) e_j(x'), \qquad x, x' \in S \quad (6)$$

with uniform convergence on $S \times S$. Any function $f \in \mathcal{H}_k$ can then be written as $f = \sum_j b_j e_j$ where $\|f\|^2_{\mathcal{H}_k} = \sum_j b_j^2/\eta_j < \infty$.

2.3 Related work

The classic approach to nonparametric intensity estimation is based on smoothing kernels (Ramlau-Hansen, 1983; Diggle, 1985) and has a form closely related to the kernel density estimator:

$$\hat{\lambda}(x) = \sum_{i=1}^{N} \kappa(x_i - x) \quad (7)$$

where κ is a smoothing kernel (related but a distinct notion from that of an RKHS kernel), that is, any bounded function integrating to 1. Early work in this area focused on edge-corrections and methods for choosing the bandwidth (Diggle, 1985; Berman and Diggle, 1989; Brooks and Marron, 1991). Connections with RKHS have been considered by, for example, Bartoszynski et al. (1981) who use a maximum penalized likelihood approach based on Hilbert spaces to estimate the intensity of a Poisson process. There is long literature on maximum penalized likelihood approaches to density estimation, which also contain interesting connections with RKHS, e.g. Silverman (1982).

Much recent work on estimating intensities for point processes has focused on Bayesian approaches to modeling Cox processes. The log Gaussian Cox Process (Møller et al., 1998) and related parameterizations of Cox (doubly stochastic) Poisson processes in terms of Gaussian processes have been proposed, along with Monte Carlo (Adams et al., 2009; Diggle et al., 2013; Teh and Rao, 2011), Laplace approximate (Illian et al., 2012; Cunningham et al., 2008; Flaxman et al., 2015) and variational (Lloyd et al., 2015; Kom Samo and Roberts, 2015) inference schemes.

3 PROPOSED METHOD AND KERNEL TRANSFORMATION

Let S be a compact domain of observations, e.g. the interval $[0, T]$ for a time series dataset observed between

times 0 and T. Let $k : S \times S \to \mathbb{R}$ be a continuous positive definite kernel, and $\mathcal{H}_k$ its corresponding RKHS of functions $f : S \to \mathbb{R}$. We model the intensity function $\lambda(\cdot)$ of an inhomogeneous Poisson process as:

$$\lambda(x) := af^2(x), \quad x \in S, \tag{8}$$

which is parametrized by $f \in \mathcal{H}_k$ and an additional scale parameter $a > 0$. Note that we have squared f to ensure that the intensity is non-negative on S, a pragmatic choice that has previously appeared in the literature (e.g. Lloyd et al. (2015)). The rationale for including a is that it allows us to decouple the overall scale and units of the intensity (e.g. number of points per hour versus number of points per year) from the penalty on the complexity of f which arises from the classical regularized Empirical Risk Minimization framework (and which should depend only on how complex, i.e. "wiggly" f is).

We use the inhomogeneous Poisson process likelihood from Eq. (4) to write the log-likelihood of a Poisson process corresponding to the observations $\{x_1, \ldots, x_N\}$, for $x_i \in S$, and intensity $\lambda(\cdot)$:

$$\ell(x_1, \ldots, x_N | \lambda) = \sum_{i=1}^{N} \log(\lambda(x_i)) - \int_S \lambda(x) dx. \tag{9}$$

We will consider the problem of minimization of the penalized negative log likelihood, where the regularization term corresponds to the squared Hilbert space norm of f in parametrization Eq. (8):

$$\min_{f \in \mathcal{H}_k} \left\{ -\sum_{i=1}^{N} \log(af^2(x_i)) + a \int_S f^2(x) dx + \gamma \|f\|^2_{\mathcal{H}_k} \right\}. \tag{10}$$

This objective is akin to a classical regularized empirical risk minimization framework over RKHS: there is a term that depends on evaluations of f at the observed points $x_1, \ldots, x_N$ as well as a term corresponding to the RKHS norm. However, the representer theorem does not apply directly to Eq. (10): since there is also a term given by the L_2-norm of f, there is no guarantee that there is a solution of Eq. (10) that lies in $\text{span}\{k(x_i, \cdot)\}_{i=1}^N$. We will show that Eq. (10) fortunately still reduces to a finite-dimensional optimization problem corresponding to a different kernel function $\tilde{k}$ which we define below.

Using the Mercer expansion of k in Eq. (6), we can write the objective Eq. (10) as follows:

$$J[f] = -\sum_{i=1}^{N} \log(af^2(x_i)) + a\|f\|^2_{L_2(S)} + \gamma \|f\|^2_{\mathcal{H}_k} \tag{11}$$

$$= -\sum_{i=1}^{N} \log(af^2(x_i)) + a \sum_{j=1}^{\infty} b_j^2 + \gamma \sum_{j=1}^{\infty} \frac{b_j^2}{\eta_j}. \tag{12}$$

The last two terms can now be merged together, giving

$$a \sum_{j=1}^{\infty} b_j^2 + \gamma \sum_{j=1}^{\infty} \frac{b_j^2}{\eta_j} = \sum_{j=1}^{\infty} b_j^2 \frac{a\eta_j + \gamma}{\eta_j} = \sum_{j=1}^{\infty} \frac{b_j^2}{\eta_j (a\eta_j + \gamma)^{-1}}.$$

Now, if we define kernel $\tilde{k}$ to be the kernel corresponding to the integral operator $T_{\tilde{k}} := T_k(aT_k + \gamma I)^{-1}$, i.e., $\tilde{k}$ is given by:

$$\tilde{k}(x, x') = \sum_{j=1}^{\infty} \frac{\eta_j}{a\eta_j + \gamma} e_j(x) e_j(x'), \qquad x, x' \in S,$$

we see that:

$$J[f] = -\sum_{i=1}^{N} \log(af^2(x_i)) + \|f\|^2_{\mathcal{H}_{\tilde{k}}}. \tag{13}$$

Thus, we have merged the two squared norm terms into a squared norm in a new RKHS. We note that a similar idea has previously been used to modify Gaussian process priors in Csató et al. (2001). We are now ready to state the representer theorem in terms of kernel $\tilde{k}$.

Theorem 1. *There exists a solution of Eq.* (10) *for observations $x_1, \ldots, x_N$, which takes the form $f^*(\cdot) = \sum_{i=1}^N \alpha_i \tilde{k}(x_i, \cdot)$.*

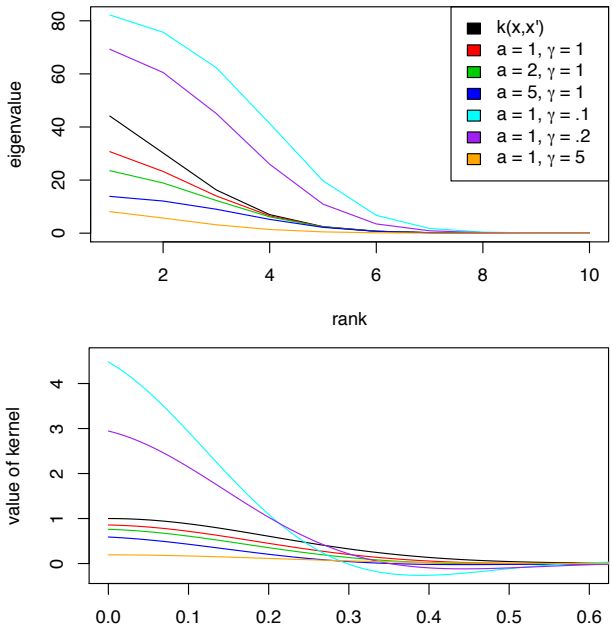
Proof. Since $\sum_j \frac{b_j^2}{\eta_j} < \infty$ if and only if $\sum_j \frac{b_j^2}{\eta_j (a\eta_j + \gamma)^{-1}} < \infty$, i.e. $f \in \mathcal{H}_k \iff f \in \mathcal{H}_{\tilde{k}}$, we have that the two spaces correspond to exactly the same set of functions. Optimization over $\mathcal{H}_k$ is therefore equivalent to optimization over $\mathcal{H}_{\tilde{k}}$. The proof now follows by applying the classical representer theorem in $\tilde{k}$ to the representation of the objective function in Eq. (13). For completeness, this is given in Appendix D. □

Remark 1. The notions of the inner product in $\mathcal{H}_k$ and $\mathcal{H}_{\tilde{k}}$ are different and thus in general $\text{span}\{k(x_i, \cdot)\} \neq \text{span}\{\tilde{k}(x_i, \cdot)\}$.

Remark 2. Notice that unlike in a standard ERM setting, $\gamma = 0$ does not recover the unpenalized risk, because γ appears in $\tilde{k}$. Notice further that the overall scale parameter a also appears in $\tilde{k}$. This is important in practice, because it allows us to decouple the scale of the intensity (which is controlled by a) from its complexity (which is controlled by γ).

Illustration. The eigenspectrum of $\tilde{k}$ where k is a squared exponential kernel is shown below for various settings of a and γ. Reminiscent of spectral filtering studied by Muandet et al. (2014), in the top plot we see that depending on the settings of a and γ, eigenvalues are shrunk or inflated as compared to $k(x, x')$

which is shown in black. In the bottom plot, the values of $k(0, x)$ are shown for the same set of kernels.



4 COMPUTATION OF $\tilde{k}$

In this section, we consider first the case in which an explicit Mercer expansion is known, and then we consider the more commonly encountered situation in which we only have access to the parametric form of the kernel $k(x, x')$, so we must approximate $\tilde{k}$. We show experimentally that our approximation is very accurate by considering the Sobolev kernel, which can be expressed in both ways.

4.1 Explicit Mercer Expansion

We start by assuming that we have a kernel k with an explicit Mercer expansion, so we have eigenvectors $\{e_j(x)\}_{j\in J}$ and eigenvalues $\{\eta_j\}_{j\in J}$:

$$k(x, x') = \sum_{j\in J} \eta_j e_j(x) e_j(x'), \tag{14}$$

with an at most countable index set J. Given a and γ we can calculate:

$$\tilde{k}(x, x') = \sum_{j\in J} \frac{\eta_j}{a\eta_j + \gamma} e_j(x) e_j(x') \tag{15}$$

up to a desired precision as informed by the spectral decay in $\{\eta_j\}_{j\in J}$. We consider a kernel on the Sobolev space on $[0, 1]$ with a periodic boundary condition, proposed by Wahba (1990, chapter 2) and recently used in Bach (2015):

$$k(x, x') = 1 + \sum_{j=1}^{\infty} \frac{2\cos\left(2\pi j\left(x - x'\right)\right)}{(2\pi j)^{2s}} \tag{16}$$

where $s = 1, 2, \ldots$ denotes the order of the Sobolev space (larger s means existence of a larger number of square-integrable derivatives). We will return to this kernel in the experiments and use it to model point process data on periodic domains, including dihedral angles in protein structures. The Mercer expansion is given by:

$$k(x, x') = \sum_{j\in\mathbb{Z}} \eta_j e_j(x) e_j(x') \tag{17}$$

where the eigenfunctions are $e_0(x) = 1$ and $e_j(x) = \sqrt{2}\cos(2\pi j x)$, $e_{-j}(x) = \sqrt{2}\sin(2\pi j x)$ for $j = \{1, 2, \ldots\}$ with the corresponding eigenvalues $\eta_0 = 1$, $\eta_j = \eta_{-j} = (2\pi j)^{-2s}$. Further details are in the Appendix in Section C.1. We derive:

$$\tilde{k}(x, x') = \frac{1}{1 + c} + \sum_{j=1}^{\infty} \frac{2\cos\left(2\pi j\left(x - x'\right)\right)}{a + \gamma(2\pi j)^{2s}}. \tag{18}$$

We discuss a Mercer expansion of the squared exponential kernel in the Appendix in Section C.2 and extensions of the Mercer expansion to multiple dimensions using a tensor product formulation in the Appendix in Section C.4. Although not practical for large datasets, we can use the Mercer expansion with summing terms up to $j > 50$ (for which the error is less than 10^{-5}) to evaluate the further approximations where Mercer expansion is not available, which we develop next.

4.2 Numerical Approximation

We propose an approximation to $\tilde{k}$ given access only to a kernel k for which we do not have an explicit Mercer expansion with respect to Lebesgue measure. We only assume that we can form Gram matrices corresponding to k and calculate their eigenvectors and eigenvalues. As a side benefit, this representation will also enable scalable computations through Toeplitz / Kronecker algebra or primal reduced rank approximations.

Let us first consider the one-dimensional case and construct a uniform grid $\mathbf{u} = (u_1, \ldots, u_m)$ on $[0, 1]$. Then the integral kernel operator T_k can be approximated with the (scaled) kernel matrix $\frac{1}{m}K_{\mathbf{uu}} : \mathbb{R}^m \to \mathbb{R}^m$, where $[K_{\mathbf{uu}}]_{ij} = k(u_i, u_j)$, and thus $\tilde{K}_{\mathbf{uu}}$ is approximately $K_{\mathbf{uu}}\left(\frac{a}{m}K_{\mathbf{uu}} + \gamma I\right)^{-1}$. Note that for the general case of multidimensional domains S, the kernel matrix would have to be multiplied by $\mathrm{vol}(S)$. Without loss of generality we assume $\mathrm{vol}(S) = 1$ below.

We are not primarily interested in evaluations of $\tilde{k}$ on this grid, but on the observations $x_1, \ldots, x_N$. Simply adding the observations into the kernel matrix is not an option however, as it changes the base measure with respect to which the integral kernel operator is to be

computed (Lebesgue measure on $[0, T]$). Thus, we consider the relationship between the eigendecomposition of $K_{\mathbf{uu}}$ and the eigenvalues and eigenfunctions of the integral kernel operator T_k.

Let $\lambda_i^u, \mathbf{e}_i^u$ be the eigenvalue/eigenvector pairs of the matrix $K_{\mathbf{uu}}$, i.e., its eigendecomposition is given by $K_{\mathbf{uu}} = Q\Lambda Q^\top = \sum_{i=1}^m \lambda_i^u \mathbf{e}_i^u (\mathbf{e}_i^u)^\top$. Then the estimates of the eigenvalues/eigenfunctions of the integral operator T_k are given by the Nyström method (see Rasmussen and Williams (2006, Section 4.3) and references therein, especially Baker (1977)):

$$\hat{\eta}_i = \frac{1}{m}\lambda_i^u, \qquad \hat{e}_i(x) = \frac{\sqrt{m}}{\lambda_i^u} K_{x\mathbf{u}} \mathbf{e}_i^u, \tag{19}$$

with $K_{x\mathbf{u}} = [k(x, u_1), \ldots, k(x, u_m)]$, leading to:

$$\begin{aligned}
\widehat{\tilde{k}}(x, x') &= \sum_{i=1}^m \frac{\hat{\eta}_i}{a\hat{\eta}_i + \gamma} \hat{e}_i(x)\hat{e}_i(x') \qquad (20)\\
&= \sum_{i=1}^m \frac{\frac{1}{m}\lambda_i^u}{\frac{a}{m}\lambda_i^u + \gamma} \cdot \frac{m}{(\lambda_i^u)^2} K_{x\mathbf{u}} \mathbf{e}_i^u (\mathbf{e}_i^u)^\top K_{\mathbf{u}x'}\\
&= K_{x\mathbf{u}} \left\{ \sum_{i=1}^m \frac{1}{\left(\frac{a}{m}\lambda_i^u + \gamma\right)\lambda_i^u} \mathbf{e}_i^u (\mathbf{e}_i^u)^\top \right\} K_{\mathbf{u}x'}.
\end{aligned}$$

For an estimate of the whole matrix $\tilde{K}_{\mathbf{xx}}$ we thus have

$$\begin{aligned}
\widehat{\tilde{K}}_{\mathbf{xx}} &= K_{\mathbf{xu}} \left\{ \sum_{i=1}^m \frac{1}{\left(\frac{a}{m}\lambda_i^u + \gamma\right)\lambda_i^u} \mathbf{e}_i^u (\mathbf{e}_i^u)^\top \right\} K_{\mathbf{ux}}\\
&= K_{\mathbf{xu}} Q \left(\frac{a}{m}\Lambda^2 + \gamma\Lambda \right)^{-1} Q^\top K_{\mathbf{ux}}. \qquad (21)
\end{aligned}$$

The above is reminiscent of the Nyström method (Williams and Seeger, 2001) proposed for speeding up Gaussian process regression. It has computational cost $O(m^3 + N^2 m)$. A reduced rank representation for Eq. (21) is straightforward by considering only the top p eigenvalues/eigenvectors of $K_{\mathbf{uu}}$. Furthermore, a primal representation with the features corresponding to kernel $\tilde{k}$ is readily available and is given by

$$\tilde{\phi}(x) = \left(\frac{a}{m}\Lambda^2 + \gamma\Lambda \right)^{-1/2} Q^\top K_{\mathbf{u}x}, \tag{22}$$

which allows linear computational cost in the number N of observations.

For $D > 1$ dimensions, one can exploit Kronecker and Toeplitz algebra approaches. Assuming that the $K_{\mathbf{uu}}$ matrix corresponds to a Cartesian product structure of the one-dimensional grids of size m, one can write $K_{\mathbf{uu}} = K_1 \otimes K_2 \cdots \otimes K_D$. Thus, the eigenspectrum can be efficiently calculated by eigendecomposing each of the smaller $m \times m$ matrices $K_1, \ldots, K_D$ and then applying standard Kronecker algebra, thereby avoiding ever having to form the prohibitively large $m^D \times m^D$ matrix $K_{\mathbf{uu}}$. For regular grids and stationary kernels, each small matrix will be Toeplitz structured, yielding further efficiency gains (Wilson et al., 2015). The resulting approach thus scales linearly in dimension D.

An even simpler alternative to the above is to sample the points $u_1, \ldots, u_m$ uniformly from the domain S using Monte Carlo or Quasi-Monte Carlo (see Oates and Girolami (2016) for a discussion in the context of RKHS). We found this approach to work well in practice in high-dimensions ($D = 15$), even when m was fixed, meaning that the scaling was effectively independent of the dimension D.

We compared the exact calculation of $\tilde{K}_{\mathbf{uu}}$ with $s = 1$, $a = 10$, and $\gamma = .5$ to our approximate calculation. For illustration we tried a coarse grid of size 10 on the unit interval (left) to a finer grid of size 100. The RMSE was 2E-3 for the coarse grid and 1.6E-5 for the fine grid, as shown in the Appendix in Fig. A9. In the same figure we compared the exact calculation of $\tilde{K}_{\mathbf{xx}}$ with $s = 1$, $a = 10$, and $\gamma = .5$ to our Nyström-based approximation, where $x_1, \ldots, x_{400} \sim \text{Beta}(.5, .5)$ distribution. The RMSE was 0.98E-3. A low-rank approximation using only the top 5 eigenvalues gives the RMSE of 1.6E-2.

5 INFERENCE

The penalized risk can be readily minimized with gradient descent. Let $\alpha = [\alpha_1, \ldots, \alpha_N]^\top$ and $\tilde{K}$ be the Gram matrix corresponding to $\tilde{k}$ such that $\tilde{K}_{ij} = \tilde{k}(x_i, x_j)$. Then $[f(x_1), \ldots, f(x_N)]^\top = \tilde{K}\alpha$ and the gradient of the objective function J from (13) is given by

$$\begin{aligned}
\nabla_\alpha J &= -\nabla_\alpha \sum_i \log(a f^2(x_i)) + \gamma \nabla_\alpha \|f\|^2_{\mathcal{H}_{\tilde{k}}}\\
&= -\nabla_\alpha \sum_i \log(a(\sum_j \tilde{k}_{ij}\alpha_j)^2) + \gamma \nabla_\alpha \alpha^\top \tilde{K}\alpha\\
&= -\sum_i \frac{2a(\sum_j \tilde{k}_{ij}\alpha_j)\nabla_\alpha \sum_j \tilde{k}_{ij}\alpha_j}{a(\sum_j \tilde{k}_{ij}\alpha_j)^2} + 2\gamma\tilde{K}\alpha\\
&= -\sum_i \frac{2\tilde{K}_{\cdot i}}{\sum_j \tilde{k}_{ij}\alpha_j} + 2\gamma\tilde{K}\alpha\\
&= -2\sum_i (\tilde{K}_{\cdot i}./(\tilde{K}\alpha)) + 2\gamma\tilde{K}\alpha
\end{aligned}$$

where ./ denotes element-wise division. Computing $\tilde{K}$ requires $\mathcal{O}(N^2)$ time and memory, and each gradient and likelihood computation requires matrix-vector multiplications which are also $\mathcal{O}(N^2)$. Overall, the running time is $\mathcal{O}(qN^2)$ for q iterations of the gradient descent method, where q is usually very small in practice.

6 NAÏVE RKHS MODEL

In this section, we compare the proposed approach, which uses the representer theorem in the transformed kernel $\tilde{k}$, to the naïve one, where a solution to Eq. (10)

of the form $f(\cdot) = \sum_{j=1}^{N} \alpha_j k(x_j, \cdot)$ is sought even though the representer theorem in k need not hold. Despite being theoretically suboptimal, this is a natural model to consider, and it might perform well in practice. The corresponding optimization problem is:

$$\min_{f \in \mathcal{H}_k} \left\{ -\sum_{i=1}^{N} \log(a f^2(x_i)) + a \int_S f^2(x) dx + \gamma \|f\|^2_{\mathcal{H}_k} \right\}$$

While the first and the last term are straightforward to calculate for any $f(\cdot) = \sum_j \alpha_j k(x_j, \cdot)$, $\int_S f^2(x)dx$ needs to be estimated. As before, we construct a uniform grid of fineness h, $\mathbf{u} = (u_1, \ldots, u_n)$ covering the domain. Then

$$\int_S f^2(u)du = \int_S \left(\alpha^\top K_{\mathbf{x}u}\right)^2 du = \alpha^\top \left\{ \int_S K_{\mathbf{x}u} K_{u\mathbf{x}} du \right\} \alpha$$
$$\approx h \alpha^\top K_{\mathbf{xu}} K_{\mathbf{ux}} \alpha,$$

and the optimization problem reads:

$$\min_{\alpha \in \mathbb{R}^N} \{ -\sum_{i=1}^{N} \log(a(\alpha^\top K_{\mathbf{x}x_i})^2) +$$
$$\alpha^\top \left(ahK_{\mathbf{xu}} K_{\mathbf{ux}} + \gamma K_{\mathbf{xx}} \right) \alpha \}.$$

As in the previous section, the gradient of this objective can be readily calculated, and optimized with gradient descent.

7 EXPERIMENTS

We use cross-validation to choose the hyperparameters in our methods: a, the fixed intensity, γ, the roughness penalty, and the length-scale of the kernel k. For a held-out set of points, we calculate the original unpenalized log-likelihood, which is given in Eq. (9) and requires the calculation of an integral over the domain. We follow the same strategy as we use for calculating $\tilde{k}$, reusing the same grid or uniform sample $u_1, \ldots, u_n$ as integration points. To calculate RMSE, we either make predictions at a grid of locations and calculate RMSE compared to the true intensity at that grid or for the high-dimensional synthetic example we pick a new uniform sample of locations over the domain and calculate the RMSE at these locations. We used limited memory BFGS in all experiments involving optimization, and found that it converged very quickly and was not sensitive to initial values. Code for our experiments is available at `https://bitbucket.org/flaxter/kernel-poisson`.

1-d synthetic Example. We generated a synthetic intensity using the Mercer expansion of a SE kernel with lengthscale 0.5, producing a random linear combination of 64 basis functions, weighted with iid draws $\alpha \sim \mathcal{N}(0, 1)$. In Fig. 1 we compare ground truth to estimates made with: our RKHS method with SE kernel, the naïve RKHS approach with SE kernel, and classical kernel intensity estimation with bandwidth selected by crossvalidation. The results are typical of what we observed on 1D and 2D examples: given similar kernel choices, each method performed similarly, and numerically there was not a significant difference in terms of the RMSE compared to the true underlying intensity.

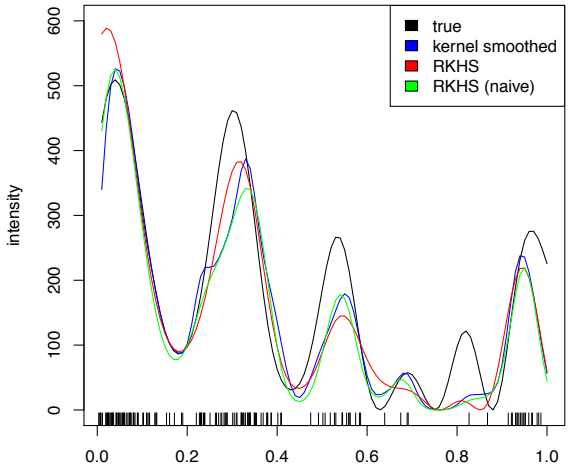
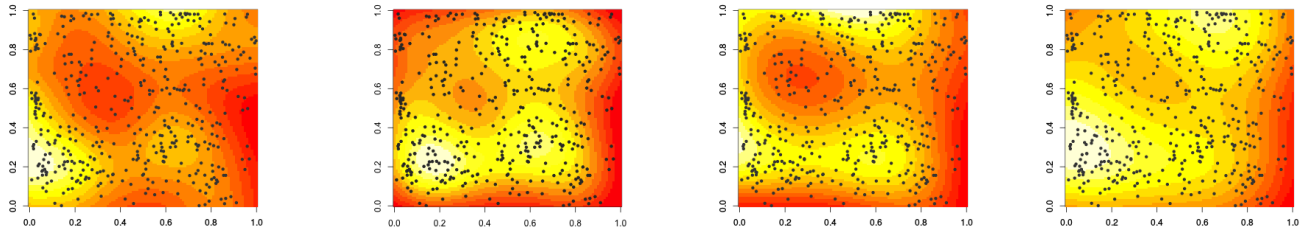


Figure 1: A synthetic dataset, comparing our RKHS method, the naïve model, and kernel smoothing to a synthetic intensity "true". The rug plot at bottom gives the location of points in the realized point pattern. The RMSE for each method was similar.

Environmental datasets. Next we demonstrate our method on a collection of two-dimensional environmental datasets giving the locations of trees. Intensity estimation is a standard first step in both exploratory analysis and modelling of these types of datasets, which were obtained from the R package `spatstat`. We calculated the intensity using various approaches: our proposed RKHS method with $\tilde{k}$ with a squared exponential kernel, the naïve RKHS method with squared exponential kernel, and classical kernel intensity estimation (KIE) with edge correction. Each method used a squared exponential kernel. We report average held-out cross-validated likelihoods in Table 1. With the exception of our method performing better on the Red oak dataset, each method had comparable performance. It is interesting to note, however, that our method does not require any explicit edge correction[1], because we are optimizing a likelihood which explicitly takes into account the window. A plot of the resulting intensity surfaces for each method and the effect of edge correction are shown in Fig. 2 for the Black oak dataset.

High dimensional synthetic examples. We generated random intensity surfaces in the unit hypercube

[1]Because no points are observed outside the window S, intensity estimates near the edge are biased downwards (Jones, 1993).



(a) KIE with edge correction (b) KIE without edge correction (c) Our RKHS method with $\tilde{k}$ (d) Naïve RKHS method

Figure 2: Location of white oak trees in Lansing, Michigan, smoothed with various approaches. Squared exponential kernels are used throughout. Edge correction makes a noticeable difference for classical kernel intensity estimation. Comparing (a) and (c) it is clear that our method is automatically performing edge correction.

Table 1: Tree Point Patterns from R Package `spatstat`

Dataset	Kernel intensity estimation	Naïve approach	Our approach with $\tilde{k}$
Lansing: Black oak (n = 135)	234	233	227
Hickory (n = 703)	1763	1746	1757
Maple (n = 514)	1239	1228	1233
Misc (n = 105)	179	177	172
New Zealand (n = 86)	119	119	119
Red oak (n = 346)	726	726	**739**
Redwoods in California (n = 62)	79	84	77
Spruces in Saxonia (n = 134)	215	212	212
Swedish pines (n = 71)	91	89	90
Waka national park (n = 504)	1142	1141	1144
White oak (n = 448)	992	992	996

for dimensions $D = 2, \ldots, 15$. The intensity was given by a constant multiplied by the square of the sum of 20 multivariate Gaussian pdfs with random means and covariances. The constant was automatically adjusted so that the number of points in the realizations would be held close to constant, around 200. We expected this to be a relatively simple synthetic example for kernel intensity estimation with a Gaussian kernel in low dimensions, but not in high dimensions. From each random intensity, we generated two random realizations, and trained our model using 2-fold crossvalidation with these two datasets. We predicted the intensity at a randomly chosen set of points and calculated the mean squared error as compared to the true intensity. For each dimension we repeated this process 100 times comparing kernel intensity estimation, the naïve approach, and our approach with $\tilde{k}$. As shown in Fig. 3(a) once we reach dimension 7 and above, our RKHS method with $\tilde{k}$ begins to outperform kernel intensity estimation, where performance is measured as MSE across 100 random datasets. Our method also significantly outperforms the naïve RKHS method as shown in Fig. 3(b). For high dimensions the difference between the two RKHS methods is not significant. This is most likely due to the fact that the number of points in the point pattern remains fixed, so the problem becomes very hard in high dimensions.[2] Finally, as shown in the Appendix in Fig. A7, kernel intensity estimation is almost always better than the naïve RKHS approach, although the difference is not significant in high dimensions.

Computational complexity. Using the synthetic data experimental setup, we evaluated the time complexity of our method with respect to dimensionality d, number of points in the point pattern dataset n, and number of points s used to estimate $\tilde{k}$ (Fig. A6), confirming our theoretical analysis. Further discussion and Figures are in the Appendix in Section A.

Spatiotemporal point pattern of crimes. To demonstrate the ability to use domain specific kernels and learn interpretable hyperparameters, we used 12 weeks (84 days) of geocoded, date-stamped reports of theft obtained from Chicago's data portal

[2]Note that our experiments are sensitive to the overall number of points in the synthetic point patterns; since kernel density estimation is a consistent method (Wied and Weißbach, 2012), we should expect kernel intensity estimation to become more accurate as the number of points grows. However, consistency in the sense of classical statistics is not necessarily useful in point processes, because our observations are not iid; the number of points that we observe is in fact part of the dataset since it reflects the underlying intensity.

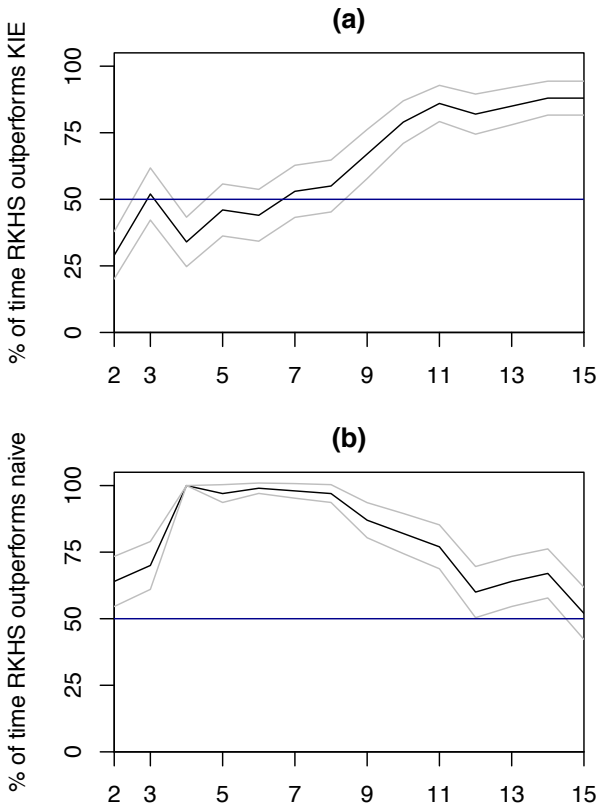


Figure 3: In (a): comparison of our RKHS method versus kernel intensity estimation (KIE) as the dimensionality grows, with 95% CIs shown based on 100 random surfaces for each dimension. In (b), our RKHS method versus the naïve RKHS method. Our method significantly outperforms kernel intensity estimation as the dimension increases, and outperforms the naïve method throughout.

(data.cityofchicago.org) starting January 1, 2004, a relatively large spatiotemporal point pattern consisting of 18,441 events. We used the following kernel: $\exp(-.5s^2/\lambda_s^2)(\exp(-2\sin^2(t\pi p))+1)(\exp(-.5t^2/\lambda_t^2))$ which is the product of a separable squared exponential space and decaying periodic time kernel (with frequency p in a time domain normalized to range from 0 to 1) plus a separable squared exponential space and time kernel. After finding reasonable values for the lengthscales and other hyperparameters of $\tilde{k}$ through exploratory data analysis, we used 2-fold cross-validation and calculated average test log-likelihoods for the number of cycles varying $p = 1, 2, \ldots, 14$ or equivalently a period of length 12 weeks (meaning no cycle), 6 weeks, ..., 6 days. These log-likelihoods are shown in Fig. 4; we found that the most likely frequency is 12, or equivalently a period lasting 1 week. This makes sense given known day-of-week effects on crime.

Dihedral angles as point process on a torus. We consider a novel application of Poisson process estimation, suited to the periodic Sobolev kernel in Eq. (16). The tensor product construction in two dimensions

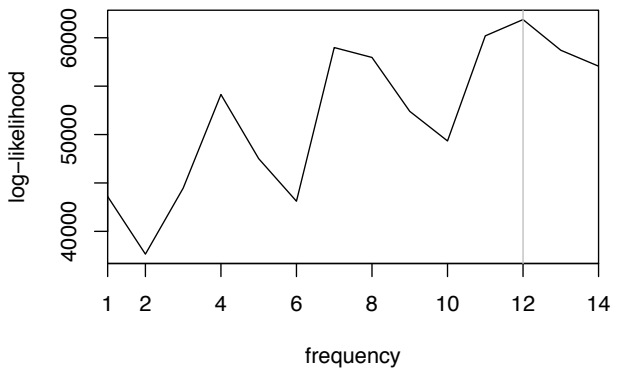


Figure 4: Log-likelihood for various frequencies in a dataset of 18,441 geocoded, date-stamped theft events from Chicago. The dataset is for 12 weeks starting January 1, 2004, and the maximum log-likelihood is attained when the frequency is 12, meaning that there is a weekly pattern in the data.

(cf. Appendix C.4) gives a periodic boundary condition appropriate for data observed on a torus. An example from protein bioinformatics is shown in Fig. A8 using data included with the R package `MDplot`, visualizing the dihedral torsion angles $[\psi, \phi]$ of amino acids in proteins (Ramachandran et al., 1963). Classically, datasets of observed angle pairs have been binned using two-dimensional histograms or they have been modelled using bivariate Von Mises mixtures fitted using an EM algorithm (Mardia, 2013), where a selection of the number of mixture components can be a challenge. We propose to treat a set of observed angles as an inhomogeneous Poisson process, thus enabling flexible nonparametric intensity estimation as shown, which directly captures the appropriate boundary conditions.

8 CONCLUSION

We presented a novel approach to inhomogeneous Poisson process intensity estimation using a Representer Theorem formulation in an appropriately transformed RKHS, providing a scalable approach giving strong performance on synthetic and real-world datasets. In future work, we will consider marked Poisson processes and other more complex point process models, as well as Bayesian extensions akin to Cox process modeling. A comparison to existing inference methods for Cox processes would also be worthwhile.

9 Acknowledgments

This work was supported by ERC (FP7/617071) and EPSRC (EP/K009362/1).

References

Adams, R. P., Murray, I., and MacKay, D. J. Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16. ACM, 2009.

Bach, F. On the equivalence between quadrature rules and random features. *arXiv:1502.06800*, 2015.

Baker, C. *The Numerical Treatment of Integral Equations*. Monographs on Numerical Analysis Series. Oxford : Clarendon Press, 1977. ISBN 9780198534068.

Bartoszynski, R., Brown, B. W., McBride, C. M., and Thompson, J. R. Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary poisson process. *The Annals of Statistics*, pages 1050–1060, 1981.

Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

Berman, M. and Diggle, P. Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 81–92, 1989.

Brooks, M. M. and Marron, J. S. Asymptotic optimality of the least-squares cross-validation bandwidth for kernel estimates of intensity functions. *Stochastic Processes and their Applications*, 38(1):157–165, 1991.

Cressie, N. and Wikle, C. *Statistics for spatio-temporal data*, volume 465. Wiley, 2011.

Csató, L., Opper, M., and Winther, O. TAP Gibbs Free Energy, Belief Propagation and Sparsity. In *Advances in Neural Information Processing Systems*, pages 657–663, 2001.

Cunningham, J. P., Shenoy, K. V., and Sahani, M. Fast gaussian process methods for point process intensity estimation. In *ICML*, pages 192–199. ACM, 2008.

Diggle, P. A kernel method for smoothing point process data. *Applied Statistics*, pages 138–147, 1985.

Diggle, P. J., Moraga, P., Rowlingson, B., Taylor, B. M., et al. Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.

Fasshauer, G. E. and McCourt, M. J. Stable evaluation of gaussian radial basis function interpolants. *SIAM Journal on Scientific Computing*, 34(2):A737–A762, 2012.

Flaxman, S. R., Wilson, A. G., Neill, D. B., Nickisch, H., and Smola, A. J. Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. *International Conference on Machine Learning*, 2015.

Illian, J. B., Sørbye, S. H., Rue, H., et al. A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (inla). *The Annals of Applied Statistics*, 6(4):1499–1530, 2012.

Jones, M. C. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3): 135–146, 1993.

Kimeldorf, G. and Wahba, G. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82 – 95, 1971. ISSN 0022-247X.

Kingman, J. F. C. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1993. ISBN 0-19-853693-3. Oxford Science Publications.

Kom Samo, Y.-L. and Roberts, S. Scalable nonparametric bayesian inference on point processes with gaussian processes. In *ICML*, pages 2227–2236, 2015.

Lloyd, C., Gunter, T., Osborne, M., and Roberts, S. Variational inference for gaussian process modulated poisson processes. In *ICML*, pages 1814–1822, 2015.

Mardia, K. V. Statistical approaches to three key challenges in protein structural bioinformatics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(3):487–514, 2013.

Møller, J., Syversveen, A., and Waagepetersen, R. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

Muandet, K., Sriperumbudur, B., and Schölkopf, B. Kernel mean estimation via spectral filtering. In *Advances in Neural Information Processing Systems*, 2014.

Oates, C. J. and Girolami, M. A. Control functionals for quasi-monte carlo integration. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, pages 56–65, 2016.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *Journal of molecular biology*, 7 (1):95–99, 1963.

Ramlau-Hansen, H. Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, 11 (2):453–466, 06 1983.

Rasmussen, C. E. and Williams, C. K. *Gaussian processes for machine learning*. MIT Press, 2006.

Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization and beyond*. MIT Press, 2002.

Silverman, B. W. On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.*, 10(3):795–810, 09 1982.

Teh, Y. W. and Rao, V. Gaussian process modulated renewal processes. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2011.

Wahba, G. *Spline models for observational data*, volume 59. Siam, 1990.

Wied, D. and Weißbach, R. Consistency of the kernel density estimator: a survey. *Statistical Papers*, 53

(1):1–21, 2012.

Williams, C. and Seeger, M. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688, 2001.

Wilson, A. G., Dann, C., and Nickisch, H. Thoughts on massively scalable gaussian processes. *arXiv:1511.01870*, 2015.

Zhu, H., Williams, C. K., Rohwer, R., and Morciniec, M. Gaussian regression and optimal finite dimensional linear models. 1997.