On the Intensity Estimation of Poisson Process

Yichen Ruan
Andrew ID: yichenr

1 Introduction

Poisson process, homogeneous or inhomogeneous, plays a fundamental role in the theory and application of stochastic process. However, the intensity estimation of inhomogeneous Poisson process remains difficult. Since 80s, tons of works have emerged to tackle this problem. Ramlau-Hansen (1983) and Diggle (1985) propose to use the smoothing kernel method that was originally designed for density estimation (Rosenblatt, 1956). Another idea is proposed by Bartoszynski et al. (1981) and Flaxman et al. (2017) to formulate the problem in a Reproducing Kernel Hilbert Space (RKHS). More recent works are focusing on direct modeling of the underlying Cox process from the Bayesian perspective (Adams et al. 2009, Illian et al. 2013).

In this paper, we will be focusing on three estimation methods: smoothing kernel method, Mercer kernel method, as well as Bayesian method. The remaining part of this section introduces the general background of Poisson process and Cox process. Section 2 presents some general notations. Key results of the three methods as well as the corresponding assumptions are listed in Section 3. The proof for relevant theorems are presented in Section 4.

1.1 Poisson process

Consider a point process over domain $S \subset \mathbb{R}^d$. For a Lebesgue measurable subset $\Gamma \subset S$, denote by $N(\Gamma)$ the number of points inside T. Define the first-order intensity (aka local intensity) as:

$$\lambda(s) = \lim_{|ds|\to 0} \frac{\mathbb{E}[N(ds)]}{|ds|} \tag{1.1.1}$$

The general (inhomogeneous) Poisson process is a point process given by

$$N(\Gamma) \sim \text{Poisson}(\int_\Gamma \lambda(s)ds) \tag{1.1.2}$$

When $\lambda(s)$ is constant over S, the process is called a homogeneous Poisson process. In that case the number of points is only depending on the volume of the subspace, i.e.

$$N(\Gamma) \sim \text{Poisson}(\lambda|\Gamma|) : \lambda(s) \equiv \lambda \tag{1.1.3}$$

Given n iid samples $x_1, ..., x_n$ drawn from some Poisson process over domain $S \subset \mathbb{R}^d$, our objective is to estimate the underlying intensity $\lambda(x)$. Condition on a intensity function, the likelihood function for the observations is given by

$$\mathcal{L}(x_1, ..., x_n|\lambda(x)) = \prod_{i=1}^n \lambda(x_i)e^{-\int_S \lambda(x)dx} \tag{1.1.4}$$

The log-likelihood function is thus

$$\ell(x_1, ..., x_n|\lambda(x)) = \sum_{i=1}^n \log(\lambda(x_i)) - \int_S \lambda(x)dx \tag{1.1.5}$$

For applications involving time series data, e.g. financial data and medical data, the one-dimension Poisson process where $d = 1$ is of special interest. For many methods to be discussed in this paper, the analysis of one-dimension Poisson process is typically easier and yields tractable solutions.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

1.2 Cox process

The Cox process, also known as doubly stochastic Poisson process, is a Poisson process whose intensity function (or surface function) $\lambda(s)$ is itself drawn from a stochastic process $\Lambda(s)$. We are interested in estimating a realization of the underlying process $\Lambda(s)$.

Under a one-dimensional scenario, we call a Cox process a stationary process if it has a stationary rate process. A formal definition is as follows:

- $\Lambda(s) : s \in S \subset \mathbb{R}$ is stationary, non-zero valued random process.
- Conditional on a realization $\lambda(s)$ of $\Lambda(s)$, the point process is an inhomogeneous Poisson process with intensity $\lambda(s)$.

For two-dimensional scenario, a well-known Cox process is the log-Gaussian Cox process (LGCP, Moller et al., 1998), for which the logarithm of the intensity function is a Gaussian process over $\mathbb{R}^2$. I.e.

$$\lambda(s) = \exp(Y(s)) \tag{1.2.1}$$

where $Y = \{Y(s)|s \in \mathbb{R}^2\}$ is a real-valued Gaussian process. In other words, the joint distribution of any finite vector $(Y(s_1), ..., Y(s_n))$ is Gaussian. Under very weak conditions, the Log-Gaussian Cox process exhibits pretty good properties. The process can also be extended to higher dimensions, but is mostly used in $\mathbb{R}^2$.

Another important Cox process is the sigmoidal Gaussian Cox process (SGCP). For which the intensity function is defined by:

$$\lambda(s) = \lambda^* \sigma(Y(s)) \tag{1.2.2}$$

where $Y = \{Y(s)|s \in S\}$ is a Gaussian process, λ^* is a hyper-parameter determines the upper bound of the intensity function, $\sigma(\cdot)$ is the sigmoid function.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1.2.3}$$

2 Notations and assumptions

Here we list some notations that are consistently used throughout the paper. Different methods may have specific assumptions and notations. Those are introduced in respective sections.

- S: The domain of the point process of interest. Generally $S \subset \mathbb{R}^d$. For one-dimension cases we sometimes write $S = [0, T] \subset \mathbb{R}$.
- $\lambda(s)$: A realization of the intensity process $\Lambda(s)$ over the domain S.
- $Y(s)$: A real-valued Gaussian process over S.
- K: Kernel. Depending on the scenario, it could be smoothing kernel function $k(\cdot)$, Mercer kernel function $k(\cdot, \cdot)$, or the Gram matrix of the Mercer kernel $K_{ij} = k(x_i, x_j)$.
- h: Bandwidth of smoothing kernels. The smoothing kernel estimators are typically subscripted with h.
- $\mathcal{H}$: Some reproducing kernel Hilbert space (RKHS).
- η_j, e_j: The j^{th} eigenvalue and eigenvector of some RKHS.

3 Key results

3.1 Smoothing kernel method

Given observations $X_1, ..., X_n \in S \subset \mathbb{R}^d$ from a Cox process with intensity function $\lambda(x)$. For a given kernel function k, define the estimator of the intensity as

$$\hat{\lambda}_h(x) \triangleq (\sum_{i=1}^{n} \frac{1}{h} k(\frac{\|x - x_i\|}{h}))/p_h(x) \tag{3.1.1}$$

where h is bandwidth, $p_h(x) = \int_S \frac{1}{h} k(\frac{\|x-u\|}{h}) du$ is an end correction.

Define the risk, or mean square error as

$$R_k(h) \triangleq \mathbb{E}[(\hat{\lambda}_h(x) - \Lambda(x))^2] \tag{3.1.2}$$

It is well known that the choice of the kernel function is of very secondary importance as compared to the selection of bandwidth. For general kernel functions, close-form risks are hard to find. The optimal bandwidth is chosen via cross validation.

However, for a one-dimension stationary Cox process, we can directly analyze the risk function and derive the optimal bandwidth therefrom. The idea is to first study the risk of uniform kernels, then extend the result to other kernels. Suppose $S = [0, T] \subset \mathbb{R}$, we have the following theorem:

Theorem 1 (Diggle, 1985)**.** For a uniform kernel $k_u(x) \equiv \frac{1}{2}, -1 \leq x \leq 1$, the risk is given by

$$R_{k_u}(h) = v(0) + \mu \frac{1 - 2\mu M(h)}{2h} + (\frac{\mu}{2h})^2 \int_0^{2h} M(z) dz \tag{3.1.3}$$

where $v(|x-y|) = \mathbb{E}[\Lambda(x)\Lambda(y)], \mu = \mathbb{E}[\Lambda(x)], M(h) = 2\mu^{-2} \int_0^h v(u) du$.

In order to find the optimal bandwidth that minimizes $R_{k_u}(h)$, we need to estimate μ and $M(h)$. Diggle (1985) suggests use $\hat{\mu} = n/T$ and $\hat{M}(h) = Tn^{-2} \sum_i \sum_{j \neq i} Q(x_i, x_j, h)$, where Q is an indicator function.

$$Q(x, y, h) = \begin{cases} 0 : |x-y| > h \\ 1 : |x-y| \leq h, |x-y| \leq \min(x, T-x) \\ 2 : |x-y| \leq h, |x-y| > \min(x, T-x) \end{cases}$$

For general kernels other than uniform kernel, people usually choose bandwidth $h_k = (3c)^{-1/2} h_u$, where $c = \int \frac{x^2}{h} k(\frac{x}{h})$, h_u is the optimal bandwidth using uniform kernel. Easy to check for uniform kernel supported on $[-1, 1]$, $c = \frac{1}{3}$, thus $h_{k_u} = h_u$. Simulation results show that such approximation yields very close bandwidths.

3.2 RKHS method (I)

Let $k : S \times S \to \mathbb{R}$ be a Mercer kernel supported on a compact set $S \subset \mathbb{R}^d$. Denote by $\mathcal{H}_k$ its associated RKHS. Model the intensity function of an inhomogeneous Poisson process as follows:

$$\hat{\lambda}(x) \triangleq a f^2(x) : f \in \mathcal{H}_k \tag{3.2.1}$$

Formulate the problem as the minimization of the negative log-likelihood (1.1.5) penalized by squared Hilbert space norm:

$$\min_{f \in \mathcal{H}_k} J(f) = -\sum_{i=1}^n \log(a f^2(x_i)) + a \int_S f^2(x) dx + \gamma \|f\|^2_{\mathcal{H}_k} \tag{3.2.2}$$

where $a, \gamma > 0$ are tuning parameters. Recall the two important properties of RKHS:

Lemma 1 (Mercer Expansion)**.** See Section 5.7 of lecture note (Nonparametric). For a kernel k, suppose $\sup_{x,y} k(x, y) < \infty$. Define eigenvalues η_j and orthonormal eigenfunctions $e_j(x)$ by

$$\int k(x, y) e_j(y) dy = \eta_j e_j(x)$$

Then we have Mercer's expansion for functions in $\mathcal{H}_k$: $f(x) = \sum_i \alpha_i k(x_i, x) = \sum_{j=1}^\infty \beta_j e_j(x)$

Furthermore, for $f(x) = \sum_j a_j e_j(x)$, $g(x) = \sum_j b_j e_j(x)$, we have

$$\langle f, g \rangle = \sum_{j=1}^\infty \frac{a_j b_j}{\eta_j}, \|f\|^2_{\mathcal{H}_k} = \sum_j \frac{a_j^2}{\eta_j} \tag{3.2.3}$$

Lemma 2 (Representer Theorem)**.** See Section 5.8 of lecture note (Nonparametric). Let ℓ be a loss function depending on $(X_1, Y_1), ..., (X_n, Y_n)$ and on $f(X_1), ..., f(X_n)$. Let $\hat{f}$ minimize

$$\ell + g(\|f\|^2_{\mathcal{H}_k})$$

where g is any monotone increasing function. Then $\hat{f}$ has the form

$$\hat{f}(x) = \sum_i^n \alpha_i k(x_i, x)$$

Due to the integral term of f^2, the representer theorem does not apply directly to the objective function $J(f)$ in 3.2.2. We will show that we are able to construct a new kernel $\tilde{k}$ such that the objective function under the new kernel satisfies the requirements of representer theorem.

Expand f as in Lemma 1 for $J(f)$ as follows:

$$\begin{aligned} J(f) &= -\sum_{i=1}^n \log(af^2(x_i)) + a\|f\|^2_{L_2(S)} + \gamma\|f\|^2_{\mathcal{H}_k} \\ &= -\sum_{i=1}^n \log(af^2(x_i)) + a\sum_{j=1}^\infty b_j^2 + \gamma\sum_{j=1}^\infty \frac{b_j^2}{\eta_j} \\ &= -\sum_{i=1}^n \log(af^2(x_i)) + \sum_{j=1}^\infty \frac{b_j^2}{\eta_j(a\eta_j + \gamma)^{-1}} \end{aligned} \tag{3.2.4}$$

Define the new kernel as

$$\tilde{k}(x, x') \triangleq \sum_{j=1}^\infty \frac{\eta_j}{a\eta_j + \gamma} e_j(x) e_j(x') \tag{3.2.5}$$

Function 3.2.4 can be written as

$$J(f) = -\sum_{i=1}^n \log(af^2(x_i)) + \|f\|^2_{\mathcal{H}_{\tilde{k}}} \tag{3.2.6}$$

Theorem 2 (Flaxman et al., 2017)**.** There exists a minimizer of problem 3.2.2 for observations $x_1, ..., x_n$, which takes the form $f^*(\cdot) = \sum_{i=1}^n \alpha_i \tilde{k}(x_i, \cdot)$

The problem then comes down to the computation of $\tilde{k}$. For kernels whose explicit Mercer expansions are known, we can directly apply the adjustment. In the case when no Mercer expansion is available, one can approximate it numerically by using the Nystrom method (Rasmussen and Williams, 2006).

Example 1 (Sobolev Space)**.** A kernel on the s-order Sobolev space on $[0, 1]$ has Mercer expansion

$$k(x, y) = 1 + \sum_{j=1}^\infty \frac{2\cos(2\pi j(x - y))}{(2\pi j)^{2s}}$$

the adjusted kernel $\tilde{k}$ is given by

$$\tilde{k}(x, y) = \frac{1}{1 + \gamma} + \sum_{j=1}^\infty \frac{2\cos(2\pi j(x - y))}{1 + \gamma(2\pi j)^{2s}}$$

The penalized risk in Eq. 3.2.4 can be minimized with gradient descent. Let $\alpha = [\alpha_1, ..., \alpha_N]^T$, $\tilde{K}$ be the Gram matrix such that $\tilde{K}_{ij} = \tilde{k}(x_i, x_j)$. Therefore, $[f(x_1), ..., f(x_N)]^T = \tilde{K}\alpha$.

The gradient of the objective function is thus given by

$$\begin{aligned}\nabla_\alpha J &= -\nabla_\alpha \sum_i \log(af^2(x_i)) + \gamma \nabla_\alpha \|f\|^2_{\mathcal{H}_{\tilde{k}}} \\ &= -\nabla_\alpha \sum_i \log(a(\sum_j \tilde{k}_{ij}\alpha_j)^2) + \gamma \nabla_\alpha \alpha^T \tilde{K}\alpha \\ &= -\sum_i \frac{2a(\sum_j \tilde{k}_{ij}\alpha_j)\nabla_\alpha \sum_j \tilde{k}_{ij}\alpha_j}{a(\sum_j \tilde{k}_{ij}\alpha_j)^2} + 2\gamma\tilde{K}\alpha \\ &= -2\sum_i (\tilde{K}_{\cdot i}./(\tilde{K}\alpha)) + 2\gamma\tilde{K}\alpha\end{aligned} \tag{3.2.7}$$

where ./ denotes the element-wise division.

To select the hyperparameter γ and a, it is recommended to crossvalidate on the negative log-likelihood of the inhomogeneous Poisson process from Eq. 1.1.5. However, the integral $\int_S f^2(u)du$ is usually intractable. As an approximation, one can use a set of locations $u = (u_1, ..., u_m)$, and let

$$\int_S f^2(u)du \approx \frac{1}{m}\sum_i f^2(u_i)$$

It is worth noting that despite the representer's theorem does not hold for the unadjusted kernel k (Eq. 3.2.2), a solution of the form $\sum_{j=1}^N \alpha_j k(x_j, \cdot)$ can still work as an approximation. Fig. 3.2 compares over a synthetic dataset the performance of the smoothing kernel method, the RKHS method, and the approximated method.

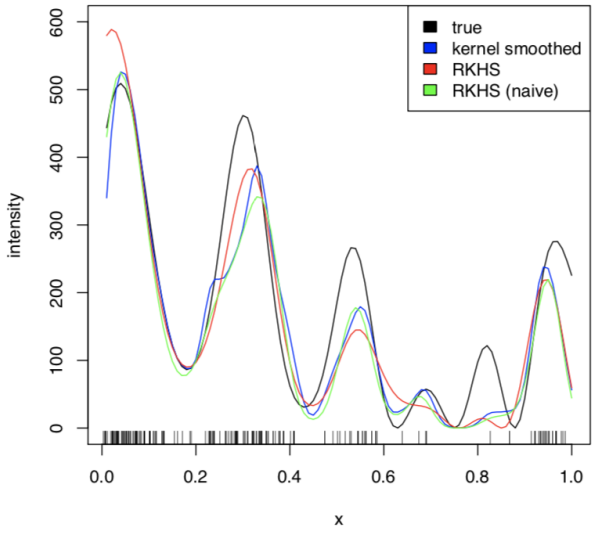

Figure 1: A synthetic dataset, comparing the smoothing kernel method (purple), the RKHS method with adjusted squared exponential kernel (red), the RKHS method with unadjusted squared exponential kernel (green). The rug plot at bottom shows the location of data points.

3.3 RKHS method (II)

In many cases, we may have multiple draws from the Poisson process. Generally, two different observations may come from different subsets of S. It turns out that the RKHS method can be very useful for data of this type.

Consider there are n draws from some one-dimension Poisson process with intensity function $\lambda(x)$. The i^{th} observation is a set of T_i data points $x_i = \{x_{i1}, ..., x_{iT_i}\}$. As per Eq. 1.1.4, the likelihood of

the i^{th} observation is given by:

$$\mathcal{L}(x_i|\lambda(x)) = \prod_{t=1}^{T_i} \lambda(x_{it}) e^{-\int_0^{T_i} \lambda(x)dx} \tag{3.3.1}$$

Thus for all the n observations

$$\mathcal{L}(x_1, ..., x_n|\lambda(x)) = \prod_{i=1}^{n} \prod_{t=1}^{T_i} \lambda(x_{it}) e^{-\sum_{i=1}^{n} \int_0^{T_i} \lambda(x)dx} \tag{3.3.2}$$

Similar to Section 3.2, we can formulate a maximum penalized likelihood problem in a RKHS:

$$\max_{\hat{\lambda} \in \mathcal{H}^s(0,T} J = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \log \hat{\lambda}(x_{it}) - \sum_{i=1}^{n} \int_0^{T_i} \hat{\lambda}(x)dx - \|\hat{\lambda}\|^2_{\mathcal{H}^s} \tag{3.3.3}$$

where $T = \max T_i$, and

$$H^s(0,T) = \{f | f^{(j)} \in L^2(0,L); j = 0, 1, ..., s \geq 1\} \tag{3.3.4}$$

We then have the following important conclusions:

Lemma 3 (Lions and Magenes, 1972)**.** $\mathcal{H}^s(0,T)$ is a reproducing kernel Hibert space.

Theorem 3 (Bartoszynski et al., 1981)**.** The maximum problem 3.3.3 has a unique maximizer in $\mathcal{H}^s(0,T) \cap \{\lambda | \lambda \geq 0\}$.

Despite the existence of an unique maximizer, simply solving problem 3.3.3 with an ordinary optimization algorithm could be highly storage demanding. An efficient approximation is to divide the range $(0,T)$ into a set of intervals. It can be shown that if the optimal intensity function λ^* is nondecreasing, then λ^* falls precisely in such a set.

3.4 Bayesian method

The basic strategy of Bayesian approaches is to put a prior on the underlying Cox process. The inference of Bayesian estimators, however, is typically intractable and often relies on techniques like MCMC and Laplace approximation. For this part, we will briefly introduce two Bayesian approache and their inference schemes.

I. Log-Gaussian Cox process (LGCP) approach

Let $X = \{x_1, ..., x_n\} \subset \Gamma \subset S$ be the observations, suppose the intensity function have the form

$$\log \lambda(x|\theta, z(x)) = Q(x)\beta + z(x) \tag{3.4.1}$$

where $Q(x)$ is a covariate vector with β an associated coefficient vector. $z(X) = \{z(x_1), ..., z(x_n)\}$ denotes the n-variate zero mean Gaussian distributed variables with covariance matrix $C_\xi(X,X) = (C_\xi(x_i, x_j))_{i,j=1,...,n}$, with ξ the vector for hyperparameters for the covariance function. Write $\theta = (\beta, \xi)$ as a collection of all the parameters. The likelihood of LGCP is therefore given by

$$\mathcal{L}(X|\theta, z) = \exp(|\Gamma| - \int_\Gamma \lambda(u|\theta, z(u))du) \prod_i \lambda(x_i|\theta, z(x_i)) \tag{3.4.2}$$

To apply MCMC, we will have to cope with the intractable integral over Γ. A common method is approximating with finite grid. Pick G points $u_1, ...u_G$, the integral can be approximated by $\sum_{g=1}^{G} \lambda(u_g|\theta, z(u_g))\Delta_g$, where Δ_g is the area of the g^{th} grid. To implement the approximation, we need to sample G variables from the Gaussian process, which involves computing the inverse of the covariance matrix within each MCMC iteration.

II. Sigmoidal Gaussian Cox process (SGCP) approach

Define the intensity function as

$$\lambda(s) = \lambda^* \sigma(g(s)) \tag{3.4.3}$$

where $g(s) : \Gamma \to \mathbb{R}$ is a random scalar function with Gaussian process prior.

Apply the Bayes' theorem, the posterior can be written as:

$$p(\mathbf{g}|X_1, ..., X_n) = \frac{\mathcal{GP}(\mathbf{g})\mathcal{L}_n}{\int \mathcal{GP}(\mathbf{g})\mathcal{L}_n d\mathbf{g}} \tag{3.4.4}$$

where L_n is the likelihood function (Eq. 1.1.4).

$$\mathcal{L}_n = \exp\left(-\int_\Gamma \lambda^* \sigma(g(s))ds\right) \prod_{i=1}^{n} \lambda^* \sigma(g(x_i)) \tag{3.4.5}$$

Standard MCMC are unable to cope with the posterior due to its doubly-intracbility: the intractable integral over Γ and the intractable integral over $\mathbf{g}$. A possible fix is to augment the Markov chain with latent variables. We can then use Markov transitions to sample from the joint distribution of observations and latent variables.

4 Proof outlines

4.1 Smoothing kernel method

Proof of Theorem 1. For a uniform kernel supported on $[-1, 1]$, ignoring the end-correction, the smoothing kernel estimator is given by

$$\tilde{\lambda}_h(x) = (2h)^{-1} N([x-h, x+h]) \tag{4.1.1}$$

where $N(\Gamma)$ denotes the number of points of the underlying Cox process located inside interval Γ (see Eq. 1.1.2).

By stationarity, $R_k(h)$ does not depend on x. Also, conditional on the realization of the rate process $\Lambda(x)$, $N([x-h, x+h])$ follows a Poisson distribution with mean $\int_h^h \Lambda(x)dx$. Thus,

$$\begin{aligned} &\mathbb{E}[(\tilde{\lambda}_h(0) - \Lambda(0))^2 | \Lambda(x)] = (2h)^{-2} \int_{-h}^{h} \Lambda(x)dx + ((2h)^{-1} \int_{-h}^{h} \Lambda(x)dx - \Lambda(0))^2 \\ &= (2h)^{-2} (\int_{-h}^{h} \Lambda(x)dx + \int_{-h}^{h}\int_{-h}^{h} \Lambda(x)\Lambda(y)dxdy) - h^{-1}\Lambda(0) \int_{-h}^{h} \Lambda(x)dx + (\Lambda(0))^2 \end{aligned} \tag{4.1.2}$$

Take expectations with respect to $\Lambda(x)$:

$$R_{k_u}(h) = (2h)^{-1}\mu + (2h)^{-2} \int_{-h}^{h}\int_{-h}^{h} v(|x-y|)dxdy - h^{-1} \int_{-h}^{h} v(|x|)dx + v(0) \tag{4.1.3}$$

Substitute $x - y$ by z and use the definition of $M(h)$ we can get Eq. 3.1.3. □

4.2 RKHS method (I)

Proof of Theorem 2. Since $\sum_j \frac{b_j^2}{\eta_j} < \infty$ if and only if $\sum_j \frac{b_j^2}{\eta_j (a\eta_j + \gamma)^{-1}} < \infty$, we have $f \in \mathcal{H}_k$ if and only if $f \in \mathcal{H}_{\tilde{k}}$. In other words, the two Hilbert spaces correspond to exactly the same set of functions. Thus, optimizing over $\mathcal{H}_k$ is equivalent to optimizing over $\mathcal{H}_{\tilde{k}}$. The proof is done by applying the representer theorem (see Lemma 2). □

Proof of Example 1. The kernel function of the Sobolev space can be written as

$$\begin{aligned} k(x, y) &= 1 + \sum_{j=1}^{\infty} \frac{2\cos(2\pi j(x-y))}{(2\pi j)^{2s}} \\ &= 1 + \sum_{j=1}^{\infty} \frac{2}{(2\pi j)^{2s}} [\cos(2\pi jx)\cos(2\pi jy) + \sin(2\pi jx)\sin(2\pi jy)] \\ &= 1 + \frac{(-1)^{s-1}}{(2s)!} B_{2s}(\{x-y\}) \end{aligned} \tag{4.2.1}$$

where $B_{2s}(\{x-y\})$ is the Bernoulli polynomial of degree $2s$ applied to the fractional part of $x-y$. Note that

$$\begin{aligned}
&\int_0^1 2\cos(2\pi jx)\sin(2\pi j'x)dx = 0 \\
&\int_0^1 2\cos(2\pi jx)\cos(2\pi j'x)dx = \delta(j-j') \\
&\int_0^1 2\sin(2\pi jx)\sin(2\pi j'x)dx = \delta(j-j')
\end{aligned} \tag{4.2.2}$$

Thus, the Mercer expansion of the kernel function is given by

$$k(x,y) = \sum_j \eta_j e_j(x) e_j(y) \tag{4.2.3}$$

where $\eta_j = (2\pi j)^{(-2s)}$, $e_j(x) = \sqrt{2}\cos(2\pi jx)$

Apply Eq. 3.2.5, the adjusted kernel function is given by

$$\begin{aligned}
\tilde{k}(x,y) &= \sum_j \frac{\eta_j}{\eta_j+\gamma} e_j(x)e_j(y) \\
&= \frac{1}{1+\gamma} + \sum_j \frac{2\cos(2\pi j(x-y))}{1+\gamma(2\pi j)^{(2s)}}
\end{aligned} \tag{4.2.4}$$

□

4.3 RKHS method (II)

Proof of Theorem 3. We will show the problem is convex. The set $H^s(0,T)\cap\{\lambda|\lambda\geq 0\}$ is clearly closed and convex. Since $H^s(0,T)$ is a reproducing kernel Hilbert space (Lemma 3), the pointwise evaluation function is continuous. On the other hand, as per the property of norms, for any sequence such that $\lambda_m \to \lambda_0$ in the $H^s(0,T)$ norm, we have $\int_0^t \lambda_n(\tau)d\tau \to \int_0^t \lambda_0(\tau)d\tau$ holds. As a result, the objective function J is continuous in $H^s(0,T)\cap\{\lambda|\lambda\geq 0\}$. Finally, it can be shown that the second Gateaux variation of J at λ in the direction ξ is given by

$$J^{''}(\lambda)(\xi,\xi) = -\sum_{i=1}^n \sum_{t=1}^{T_i} \frac{\xi^2(x_{ij})}{\lambda^2(x_{ij})} - 2\|\xi\|^2 < \|\xi\|^2 \tag{4.3.1}$$

As a result, J is uniformly negative definite on $H^s(0,T)\cap\{\lambda|\lambda\geq 0\}$. □

5 Conclusion

In this paper, we summarize three popular technologies for the intensity estimation of Poisson process. The smoothing kernel method has elegant results for one-dimension stationary data. The RKHS method works by adjusting the kernel functions such that the representer's theorem can be applied. Similar method also works for data consists of multiple draws from some one-dimension Poisson process. Two Bayesian approached based on the LGCP model and the SGCP model are also discussed.

References

[1] Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, 453-466.

[2] Diggle, P. (1985). A kernel method for smoothing point process data. *Applied statistics*, 138-147.

[3] Flaxman, S., Teh, Y. W., & Sejdinovic, D. (2017). Poisson intensity estimation with reproducing kernels. *Electronic Journal of Statistics*, 11(2), 5081-5104.

[4] Bartoszynski, R., Brown, B. W., McBride, C. M., & Thompson, J. R. (1981). Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary Poisson process. *The Annals of Statistics*, **9**(5), 1050-1060.

[5] Adams, R. P., Murray, I., & MacKay, D. J. (2009, June). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 9-16). ACM.

[6] Møller, J., Syversveen, A. R., & Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics*, **25**(3), 451-482.

[7] Shirota, S., & Gelfand, A. E. (2016). Inference for log Gaussian Cox processes using an approximate marginal posterior. arXiv preprint arXiv:1611.10359.

[8] Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning. 2006. The MIT Press, Cambridge, MA, USA, 38, 715-719.

[9] Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, 795-810.

[10] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 832-837.