Gaussian process modulated Cox processes under linear inequality constraints

Andrés F. López-Lopera, St John, Nicolas Durrande

Gaussian Process Modulated Cox Processes under Linear Inequality Constraints

Andrés F. López-Lopera
Mines Saint-Étienne*

ST John
PROWLER.io

Nicolas Durrande
PROWLER.io

Abstract

Gaussian process (GP) modulated Cox processes are widely used to model point patterns. Existing approaches require a mapping (link function) between the unconstrained GP and the positive intensity function. This commonly yields solutions that do not have a closed form or that are restricted to specific covariance functions. We introduce a novel finite approximation of GP-modulated Cox processes where positiveness conditions can be imposed directly on the GP, with no restrictions on the covariance function. Our approach can also ensure other types of inequality constraints (e.g. monotonicity, convexity), resulting in more versatile models that can be used for other classes of point processes (e.g. renewal processes). We demonstrate on both synthetic and real-world data that our framework accurately infers the intensity functions. Where monotonicity is a feature of the process, our ability to include this in the inference improves results.

1 INTRODUCTION

Point processes are used in a variety of real-world problems for modelling temporal or spatiotemporal point patterns in fields such as astronomy, geography, and ecology (Baddeley et al., 2015; Møller and Waagepetersen, 2004). In reliability analysis, they are used as renewal processes to model the lifetime of items or failure (hazard) rates (Cha and Finkelstein, 2018).

Poisson processes are the foundation for modelling point patterns (Kingman, 1992). Their extension to stochastic intensity functions, known as doubly stochastic Poisson processes or Cox processes (Cox, 1955), enables non-parametric inference on the intensity function and allows expressing uncertainties (Møller and Waagepetersen, 2004). Moreover, previous studies have shown that other classes of point processes may also be seen as Cox processes. For example, Yannaros (1988) proved that Gamma renewal processes are Cox processes under non-increasing conditions. A similar analysis was made later for Weibull processes (Yannaros, 1994).

Gaussian processes (GPs) form a flexible prior over functions, and are widely used to model the intensity process $\Lambda(\cdot)$ (Møller et al., 2001; Adams et al., 2009; Teh and Rao, 2011; Gunter et al., 2014; Lasko, 2014; Lloyd et al., 2015; Fernandez et al., 2016; Donner and Opper, 2018). However, to ensure positive intensities, this commonly requires link functions between the intensity process and the GP $g(\cdot)$. Typical examples of mappings are $\Lambda(x) = \exp(g(x))$ (Møller et al., 2001; Diggle et al., 2013; Flaxman et al., 2015) or $\Lambda(x) = g(x)^2$ (Lloyd et al., 2015; Kozachenko et al., 2016). The exponential transformation has the drawback that there is no closed-form expression for some of the integrals required to compute the likelihood. Although the square inverse link function allows closed-form expressions for certain kernels, it leads to models exhibiting "nodal lines" with zero intensity due to the non-monotonicity of the transformation (see John and Hensman, 2018, for a discussion). Furthermore, current approaches to Cox process inference cannot be used in applications such as renewal processes that require both positivity and monotonicity constraints.

Here, we introduce a novel approximation of GP-modulated Cox processes that does not rely on a mapping to obtain the intensity. In our approach we impose the constraints (e.g. non-negativeness or monotonicity) directly on $\Lambda(\cdot)$ by sampling from a truncated Gaussian vector. This has the advantage that the likelihood can be computed in closed form.

*Part of this work was completed during an internship of A. F. López-Lopera at PROWLER.io.

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

Moreover, our approach can ensure any type of linear inequality constraint everywhere, which allows modelling of a broader range of point processes.

This paper is organised as follows. In Section 2, we briefly describe inhomogeneous Poisson processes and some of their extensions. In Sections 3 and 4, we introduce a finite representation of GP-modulated Cox processes and the corresponding Cox process inference under inequality constraints. In Section 5, we apply our framework to 1D and 2D inference examples under different inequality conditions. We also test its performance in reliability applications with hazard rates exhibiting monotonic behaviours. Finally, in Section 6, we summarise our results and outline potential future work.

2 POISSON POINT PROCESSES

A Poisson process X is a random countable subset of $\mathcal{S} \subseteq \mathbb{R}^d$ where points occur independently (Baddeley et al., 2006). Let $N \in \mathbb{N}$ be a random variable (r.v.) denoting the number of points in X. Let $X_1, \cdots, X_n$ be a set of n independent and identically distributed (i.i.d.) r.v.'s on $\mathcal{S}$. The likelihood of $(N = n, X_1 = \mathbf{x}_1, \cdots, X_n = \mathbf{x}_n)$ under an inhomogeneous Poisson process with non-negative intensity $\lambda(\cdot)$ is given by (Møller and Waagepetersen, 2004)

$$f_{(N,X_1,\cdots,X_n)}(n, \mathbf{x}_1, \cdots, \mathbf{x}_n) = \frac{\exp(-\mu)}{n!} \prod_{i=1}^{n} \lambda(\mathbf{x}_i), \quad (1)$$

where

$$\mu = \int_{\mathcal{S}} \lambda(\mathbf{s}) \, \mathrm{d}\mathbf{s} \quad (2)$$

is the intensity measure or overall intensity.

When $\mathcal{S}$ is the real line, the distance (inter-arrival time) between consecutive points of a Poisson process follows an exponential distribution. Renewal processes are a generalisation of Poisson processes where inter-arrival times are i.i.d. but not necessarily exponentially distributed. An example is the Weibull process where inter-arrival times are distributed following $\lambda(x) = \alpha\beta x^{\beta-1}$ (Cha and Finkelstein, 2018).

Cox processes (Cox, 1955) are a natural extension of inhomogeneous Poisson processes where $\lambda(\cdot)$ is sampled from a non-negative stochastic process $\Lambda(\cdot)$. Previous studies have shown that many classes of point processes can be seen as Cox processes under certain conditions (Møller and Waagepetersen, 2004; Yannaros, 1988, 1994). For example, Weibull renewal processes are Cox processes for $\beta \in (0, 1]$ (Yannaros, 1994). This motivates the construction of GPs with non-negative and monotonic constraints, so that they can be used as intensities $\Lambda(\cdot)$ of Cox processes.

3 APPROXIMATION OF GP MODULATED COX PROCESSES

In this work, we approximate the intensity $\Lambda(\cdot)$ of the Cox process by a finite-dimensional GP $\Lambda_m(\cdot)$ subject to some inequality constraints (e.g. boundedness, monotonicity, convexity). Since positiveness constraints are imposed directly on $\Lambda_m(\cdot)$, a link function is no longer necessary. This has two main advantages. First, the likelihood (1) can be computed analytically. Second, as our approach ensures any linear inequality constraint, it can be used for modelling a broader range of point processes.

3.1 Finite Approximation of 1D GPs

Let $\Lambda(\cdot)$ be a zero-mean GP on $\mathbb{R}$ with arbitrary covariance function k. Consider $x \in \mathcal{S}$, with compact space $\mathcal{S} = [0, 1]$, and a set of knots $t_1, \cdots, t_m \in \mathcal{S}$. Here we consider equispaced knots $t_j = (j-1)\Delta_m$ with $\Delta_m = 1/(m-1)$. We define $\Lambda_m(\cdot)$ as the finite-dimensional approximation of $\Lambda(\cdot)$ consisting of its piecewise-linear interpolation at knots $t_1, \cdots, t_m$, i.e.,

$$\Lambda_m(x) = \sum_{j=1}^{m} \phi_j(x)\xi_j, \quad (3)$$

where $\xi_j := \Lambda(t_j)$ for $j = 1, \cdots, m$, and $\phi_1, \cdots, \phi_m$ are hat basis functions given by

$$\phi_j(x) := \begin{cases} 1 - \left|\frac{x-t_j}{\Delta_m}\right| & \text{if } \left|\frac{x-t_j}{\Delta_m}\right| \leq 1, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

Similarly to spline-based approaches (e.g., Sleeper and Harrington, 1990), we assume that $\Lambda(\cdot)$ is piecewise defined by (first-order) polynomials. The striking property of this basis is that satisfying the inequality constraints (e.g. boundedness, monotonicity, convexity) at the knots implies that the constraints are satisfied everywhere in the input space (Maatouk and Bay, 2017). Although it is tempting to generalise the above construction to smoother basis functions, it makes this property difficult to enforce.

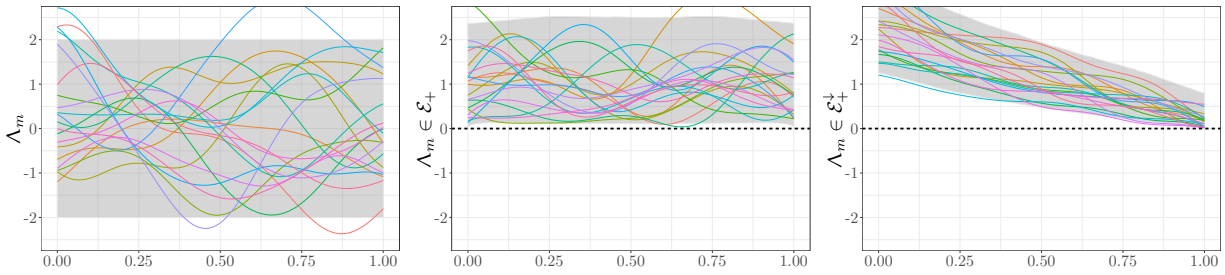
We aim at computing the distribution of $\Lambda_m(\cdot)$ under the condition that it belongs to a convex set of functions $\mathcal{E}$ defined by some inequality constraints (e.g. positivity). This piecewise-linear representation has the benefit that satisfying $\Lambda_m(\cdot) \in \mathcal{E}$ is equivalent to satisfying only a finite number of inequality constraints. More precisely,

$$\Lambda_m(\cdot) \in \mathcal{E} \;\Leftrightarrow\; \boldsymbol{\xi} \in \mathcal{C}, \quad (5)$$

where $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_m]^\top$, and $\mathcal{C}$ is a convex set on $\mathbb{R}^m$.

For non-negativeness conditions $\mathcal{E}_+$, $\mathcal{C}$ is given by

$$\mathcal{C}_+ := \{c \in \mathbb{R}^m;\ \forall\ j = 1, \cdots, m\ :\ c_j \geq 0\}, \quad (6)$$



(a) unconstrained GP (b) GP with $\mathcal{C}_+$ constraints (c) GP with $\mathcal{C}_+^{\downarrow}$ constraints

Figure 1: Samples from the prior $\Lambda_m(\cdot)$ under (a) no constraints, (b) non-negativeness constraints, (c) both non-negativeness and non-increasing constraints. The grey region shows the 95% confidence interval.

and for non-increasing conditions $\mathcal{E}_{\downarrow}$, $\mathcal{C}$ is given by

$$\mathcal{C}_{\downarrow} := \{c \in \mathbb{R}^m;\ \forall\ j = 2, \cdots, m\ :\ c_{j-1} \geq c_j\}. \quad (7)$$

Constraints can be composed, e.g. the convex set of non-negativeness and non-increasing conditions is given by $\mathcal{C}_+^{\downarrow} = \mathcal{C}_+ \cap \mathcal{C}_{\downarrow}$.

Assuming that $\boldsymbol{\xi}$ is zero-mean Gaussian-distributed with covariance matrix $\boldsymbol{\Gamma} = (k(t_i, t_j))_{1\leq i,j\leq m}$, then the distribution of $\boldsymbol{\xi}$ conditioned on $\boldsymbol{\xi} \in \mathcal{C}$ is a truncated Gaussian distribution. Then, quantifying uncertainty on Λ_m relies on sampling $\boldsymbol{\xi} \in \mathcal{C}$ (see López-Lopera et al., 2018, for further discussion).

The effect of different constraints on samples from the prior $\Lambda_m(\cdot)$ can be seen in Figure 1. Here we set $m = 100$ and use a squared-exponential (SE) covariance function[1] with covariance parameters $\sigma^2 = 1$, $\ell = 0.2$. The samples were generated via Hamiltonian Monte Carlo (HMC) (Pakman and Paninski, 2014).

3.2 Application to 1D GP-Modulated Cox Processes

The key challenge in building GP-modulated Cox processes is the evaluation of the integral in the intensity measure. By considering $\Lambda_m(\cdot)$ as the intensity of the Cox process, the intensity measure (2) becomes

$$\mu_m = \int_0^1 \Lambda_m(x)\,\mathrm{d}x = \int_0^1 \sum_{j=1}^m \phi_j(x)\xi_j\,\mathrm{d}x = \sum_{j=1}^m c_j\xi_j,$$

where $c_1 = c_m = \frac{\Delta_m}{2}$ and $c_j = \Delta_m$ for $1 < j < m$. The likelihood of $(N = n, X_1 = x_1, \cdots, X_n = x_n)$ is

$$f_{(N,X_1,\cdots,X_n)|\{\xi_1,\cdots,\xi_m\}}(n, x_1, \cdots, x_n) = \frac{1}{n!}\exp\left(-\sum_{j=1}^m c_j\xi_j\right)\prod_{i=1}^n\sum_{j=1}^m \phi_j(x_i)\xi_j. \quad (8)$$

[1] SE covariance function: $k(t,t') = \sigma^2\exp(-\frac{(t-t')^2}{2\ell^2})$.

Since (8) depends on r.v.'s $\xi_1, \cdots, \xi_m$, it can be approximated using samples of $\boldsymbol{\xi}$. To estimate the covariance parameters $\boldsymbol{\theta}$ of the vector $\boldsymbol{\xi}$, we can use stochastic global optimisation (Jones et al., 1998).

3.3 Extension to Higher Dimensions

The approximation in (3) can be extended to grids in d dimensions by tensorisation. For ease of notation, we assume the same number of knots m and knot-spacing Δ_m in each dimension, but the generalisation to different $m_1, \cdots, m_d$ or $\Delta_{m_1}, \cdots, \Delta_{m_d}$ is straightforward. Consider $\mathbf{x} = (x_1, \cdots, x_d) \in [0,1]^d$, and a set of knots per dimension $(t_1^1, \cdots, t_m^1), \cdots, (t_1^d, \cdots, t_m^d)$. Then Λ_m is given by

$$\Lambda_m(\mathbf{x}) = \sum_{j_1,\cdots,j_d=1}^m \left[\prod_{i=\{1,\cdots,d\}} \phi_{j_i}^i(x_i)\right]\xi_{j_1,\cdots,j_d}, \quad (9)$$

where $\xi_{j_1,\cdots,j_d} := \Lambda(t_{j_1}, \cdots, t_{j_d})$ and $\phi_{j_i}^i$ are the hat basis functions defined in (4). Inequality constraints can be imposed as in López-Lopera et al. (2018). By substituting (9) in (2), we obtain

$$\mu_m = \int_0^1 \Lambda_m(\mathbf{x})\,\mathrm{d}\mathbf{x} = \sum_{j_1,\cdots,j_d=1}^m \left[\prod_{i=\{1,\cdots,d\}} c_{j_i}\right]\xi_{j_1,\cdots,j_d},$$

with c_{j_i} defined as in 1D, and the likelihood is

$$\begin{aligned} &f_{(N,X_1,\cdots,X_n)|\boldsymbol{\xi}}(n, \mathbf{x}_1, \cdots, \mathbf{x}_n) \\ &= \frac{1}{n!}\exp\left(-\sum_{j_1,\cdots,j_d=1}^m \left[\prod_{i=\{1,\cdots,d\}} c_{j_i}\right]\xi_{j_1,\cdots,j_d}\right) \\ &\quad\times \prod_{i=1}^n \sum_{j_1,\cdots,j_d=1}^m \left[\prod_{k=\{1,\cdots,d\}} \phi_{j_i}(x_{i,k})\right]\xi_{j_1,\cdots,j_d}, \end{aligned} \quad (10)$$

where $x_{i,k}$ is the k-th component of the point $\mathbf{x}_i$.

Due to the tensor structure of the finite representation, it becomes costly as the dimension d increases. The

HMC sampler for truncated multivariate Gaussians from Pakman and Paninski (2014) follows the same dynamics as a classical HMC sampler, but the particle "bounces" on the boundaries if its trajectory reaches one of the inequality constraints. The computational complexity of each iteration scales linearly with the number of inequality conditions (e.g. m^d for positiveness constraints) if the iteration does not require any reflection, but also increases with each bounce. Hence, in the best case, the computational complexity is $\mathcal{O}(m^d)$. However, this drawback could be mitigated by using sparse representations of the constraints (Pakman and Paninski, 2014), or using other types of designs of the knots (e.g. sparse designs).

4 COX PROCESS INFERENCE

Having introduced the model, we now establish an inference procedure for $\Lambda(\cdot)$ using the approximation $\Lambda_m(\cdot)$. For readability, we only assume non-negativeness constraints, i.e. $\boldsymbol{\xi} \geq \mathbf{0}$, but the extension to other types of constraints can be made by constructing a set of linear inequalities of the form $\boldsymbol{l} \leq \boldsymbol{A}\boldsymbol{\xi} \leq \boldsymbol{u}$, where $\boldsymbol{A}$ is a full-rank matrix encoding the linear operations, and $\boldsymbol{l}$ and $\boldsymbol{u}$ are the lower and upper bounds. In that case, results for $\boldsymbol{A}\boldsymbol{\xi}|\{\boldsymbol{l} \leq \boldsymbol{A}\boldsymbol{\xi} \leq \boldsymbol{u}\}$ are similar as for $\boldsymbol{\xi}|\{\mathbf{0} \leq \boldsymbol{\xi} < \boldsymbol{\infty}\}$, and samples of $\boldsymbol{\xi}$ can be recovered from samples of $\boldsymbol{A}\boldsymbol{\xi}$, by solving a linear system.

Consider the non-negative Gaussian vector $\boldsymbol{\xi}$ and its sample $\boldsymbol{\chi}$. The posterior distribution of $\boldsymbol{\xi}$ conditioned on a point pattern $(N = n, X_1 = x_1, \cdots, X_n = x_n)$ is

$$f_{\boldsymbol{\xi}|\{N=n,X_1=x_1,\cdots,X_n=x_n\}}(\boldsymbol{\chi}) \propto f_{(N,X_1,\cdots,X_n)|\{\boldsymbol{\xi}=\boldsymbol{\chi}\}}(n, x_1, \cdots, x_n)\, f_{\boldsymbol{\xi}}(\boldsymbol{\chi}), \tag{11}$$

where the likelihood is defined in (8) and $f_{\boldsymbol{\xi}}(\boldsymbol{\chi})$ is the (truncated) Gaussian density given by

$$f_{\boldsymbol{\xi}}(\boldsymbol{\chi}) = \frac{\exp\left\{-\frac{1}{2}\boldsymbol{\chi}^\top \boldsymbol{\Gamma}^{-1}\boldsymbol{\chi}\right\}}{\int_0^\infty \exp\left\{-\frac{1}{2}\mathbf{s}^\top \boldsymbol{\Gamma}^{-1}\mathbf{s}\right\} \mathrm{d}\mathbf{s}}, \quad \text{for} \quad \boldsymbol{\chi} \geq \mathbf{0}. \tag{12}$$

Since the posterior distribution (11) can be approximated using samples of $\boldsymbol{\xi}$, it is possible to infer $\Lambda_m(\cdot)$ via Metropolis-Hastings.

4.1 Metropolis-Hastings Algorithm with Truncated Gaussian Proposals

The implementation of the Metropolis-Hastings algorithm requires a proposal distribution q for the next step in the Markov chain. In practice, Gaussian proposals are often used, leading to the famous random-walk Metropolis algorithm (Murphy, 2012). However, since inequality constraints are not necessarily satisfied using (non-truncated) Gaussian proposals, the standard random walk can suffer from small acceptance rates due to constraint violations. We propose as an alternative a constrained version of the random-walk Metropolis algorithm where inequality conditions are ensured when sampling from the proposal q. As $\boldsymbol{\xi}$ is (non-negative) truncated Gaussian-distributed (with covariance matrix $\boldsymbol{\Gamma}$), we suggest the truncated Gaussian proposal q given by

$$q(\boldsymbol{\chi}^{k+1}|\boldsymbol{\chi}^k) = \frac{\exp\left\{-\frac{1}{2}[\boldsymbol{\chi}^{k+1} - \boldsymbol{\chi}^k]^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{\chi}^{k+1} - \boldsymbol{\chi}^k]\right\}}{\int_0^\infty \exp\left\{-\frac{1}{2}[\mathbf{s} - \boldsymbol{\chi}^k]^\top \boldsymbol{\Sigma}^{-1}[\mathbf{s} - \boldsymbol{\chi}^k]\right\} \mathrm{d}\mathbf{s}}, \tag{13}$$

where $\boldsymbol{\chi}^{k+1}, \boldsymbol{\chi}^k \geq \mathbf{0}$ are samples of $\boldsymbol{\xi}$ and $\boldsymbol{\Sigma}$ is the covariance matrix. Sampling from q can then be performed via MCMC (Pakman and Paninski, 2014). We use $\boldsymbol{\Sigma} = \eta\boldsymbol{\Gamma}$, where η is a scale factor. This has the benefit that we are sampling from a distribution with similar structure to the true one, while η controls the step size of the Metropolis-Hastings procedure and can be manually tuned to obtain a trade-off between mixing speed and acceptance rate of the algorithm. The acceptance probability is given by

$$\alpha_k = \frac{\widetilde{f}_{\boldsymbol{\xi}|\{N=n,X_1=x_1,\cdots,X_n=x_n\}}(\boldsymbol{\chi}^{k+1})}{\widetilde{f}_{\boldsymbol{\xi}|\{N=n,X_1=x_1,\cdots,X_n=x_n\}}(\boldsymbol{\chi}^k)} \times \beta_k, \tag{14}$$

where $\beta_k = q(\boldsymbol{\chi}^k|\boldsymbol{\chi}^{k+1})/q(\boldsymbol{\chi}^{k+1}|\boldsymbol{\chi}^k)$, and

$$\widetilde{f}_{\boldsymbol{\xi}|\{N=n,X_1=x_1,\cdots,X_n=x_n\}}(\boldsymbol{\chi}) = \exp\Big(-\frac{1}{2}\boldsymbol{\chi}^\top \boldsymbol{\Gamma}^{-1}\boldsymbol{\chi} - \boldsymbol{c}^\top \boldsymbol{\chi}\Big) \prod_{i=1}^n \boldsymbol{\phi}^\top(x_i)\boldsymbol{\chi} \tag{15}$$

is the (unnormalised) posterior distribution. $\boldsymbol{\phi}(\cdot) = [\phi_1(\cdot), \cdots, \phi_m(\cdot)]^\top$ and $\boldsymbol{c} = [c_1, \cdots, c_m]^\top$ are defined in (4) and (8). We now focus on the term β_k. Since the truncated Gaussian density has the same functional form as the non-truncated one, apart from the differing support and normalising constants, this yields

$$\beta_k = \frac{\int_0^\infty \exp\left\{-\frac{1}{2}[\mathbf{s} - \boldsymbol{\chi}^k]^\top \boldsymbol{\Sigma}^{-1}[\mathbf{s} - \boldsymbol{\chi}^k]\right\} \mathrm{d}\mathbf{s}}{\int_0^\infty \exp\left\{-\frac{1}{2}[\mathbf{s} - \boldsymbol{\chi}^{k+1}]^\top \boldsymbol{\Sigma}^{-1}[\mathbf{s} - \boldsymbol{\chi}^{k+1}]\right\} \mathrm{d}\mathbf{s}}. \tag{16}$$

The orthants $\int_0^\infty \exp\left\{-\frac{1}{2}[\mathbf{x} - \boldsymbol{\mu}]^\top \boldsymbol{\Sigma}^{-1}[\mathbf{x} - \boldsymbol{\mu}]\right\} \mathrm{d}\mathbf{x}$ cannot be computed in closed form, but they can be estimated via MC (Genz, 1992; Botev, 2017). Algorithm 1 summarises the implementation of the Metropolis-Hastings algorithm for the Cox process inference using the finite approximation of Section 3.

Algorithm 1 Metropolis-Hastings algorithm for Cox process inference with truncated Gaussian proposals

1: Input: $\boldsymbol{\chi}^{(0)} \in (\mathbb{R}^m)^+$, $\boldsymbol{\Gamma}$ (covariance matrix of $\boldsymbol{\xi}$), η (scale factor).
2: **for** $k = 0, 1, 2, \cdots$ **do**
3: Sample $\boldsymbol{\chi}' \sim \mathcal{N}\left(\boldsymbol{\chi}^{(k)}, \eta\boldsymbol{\Gamma}\right)$ such that $\boldsymbol{\chi}' \in \mathcal{C}_+$.
4: Compute α_k as in (14).
5: Sample $u_k \sim \text{uniform}(0, 1)$.
6: Set new sample to
7: $$\boldsymbol{\chi}^{(k+1)} = \begin{cases} \boldsymbol{\chi}', & \text{if } \alpha_k \geq u_k \\ \boldsymbol{\chi}^{(k)}, & \text{if } \alpha_k < u_k \end{cases}.$$
8: Compute $\lambda_m^{(k)}(x) = \sum_{j=1}^m \phi_j(x)\chi_j^{(k)}$ at location x with $\phi_j(\cdot)$ defined in (4).

4.2 Inference with Multiple Observations

For N_o independent observations $(X_{\nu,1}, \cdots, X_{\nu,n_\nu})$ with $\nu = 1, \cdots, N_o$, the acceptance probability follows

$$\alpha_k = \frac{\prod_{\nu=1}^{N_o} f_{\boldsymbol{\xi}|\{N_\nu = n_\nu, \cdots, X_{\nu,n_\nu} = x_{\nu,n_\nu}\}}(\boldsymbol{\chi}^{k+1})}{\prod_{\nu=1}^{N_o} f_{\boldsymbol{\xi}|\{N_\nu = n_\nu, \cdots, X_{\nu,n_\nu} = x_{\nu,n_\nu}\}}(\boldsymbol{\chi}^{k})} \, \beta_k, \quad (17)$$

with posterior $f_{\boldsymbol{\xi}|\{N_\nu = n_\nu, X_{\nu,1} = x_{\nu,1}, \cdots, X_{\nu,n_\nu} = x_{\nu,n_\nu}\}}$ and β_k given by (11) and (16). Then, Algorithm 1 can be used with (17).

5 EMPIRICAL RESULTS

We test the performance of the finite approximation of GP-modulated Cox process on 1D and 2D applications. In the following, we use the squared-exponential covariance for the Gaussian vector $\boldsymbol{\xi}$ so that we can compare to Lloyd et al. (2015). We estimate the covariance parameters $\boldsymbol{\theta} = (\sigma^2, \ell)$ by maximising the likelihood (8). For all numerical experiments, we fix m such that we obtain accurate resolutions of the finite representations while minimising the cost of MCMC (see Bay et al., 2016; Maatouk and Bay, 2017, for discussion about the convergence of the finite-dimensional approximation of GPs).[2] For simulating $\boldsymbol{\xi}$, we use the exact HMC sampler proposed by Pakman and Paninski (2014). To approximate the Gaussian orthant probabilities from (16), we use the estimator proposed by Botev (2017) using 200 MC samples. We run Algorithm 1 with a scale factor η between 10^{-3} and 10^{-4} for a good trade-off between the mixing speed and the acceptance rate for each experiment.[3] The number of discarded burn-in samples until the Markov chains became stationary varied between 10^3 and 10^4 samples. The code was implemented in the R programming language based on the package `lineqGPR` (López-Lopera, 2018).

[2]We tested our model for various values of m, observing that, after a certain value, inference results are unchanged.

[3]We observed convergence of Algorithm 1 for a wide range of values of $\eta \in [10^{-5}, 10^{-2}]$. Fine-tuning of η can help the experiment run faster by gradually increasing η until the sampling mixes well.

5.1 Examples with Multiple Observations

Here, we test our approach using the three toy examples proposed by Adams et al. (2009),

$$\lambda_1(x) = 2\exp\{-x/15\} + \exp\{-[(x-25)/10]^2\},$$
$$\lambda_2(x) = 5\sin(x^2) + 6,$$
$$\lambda_3(x) = \text{piecewise linear through } (0,2), (25,3), (50,1), (75,2.5) \text{ and } (100,3).$$

The domains for λ_1, λ_2 and λ_3 are $\mathcal{S}_1 = [0, 50]$, $\mathcal{S}_2 = [0, 5]$ and $\mathcal{S}_3 = [0, 100]$, respectively.

Figure 2 shows the inference results using $N_o = 1, 10, 100$ observations sampled from the ground truth. With increasing number of observations the inferred intensity converges to the ground truth. Here, we fixed $m = 100$ and $\eta = 10^{-3}$.

In Table 1, we assess the performance of our approach under non-negativeness constraints (cGP-c_+). We compare our inference results to the ones obtained with a log-Gaussian process (log-GP) modulated Cox process (Møller et al., 2001) and Variational Bayes for Point Processes (VBPP) (Lloyd et al., 2015) using the Q^2 criterion. This criterion is defined as $Q^2 = 1 - \text{SMSE}(\lambda(\cdot), \widehat{\lambda}(\cdot))$, where SMSE is the standardised mean squared error (Rasmussen and Williams, 2005). Q^2 is equal to one if the inferred $\widehat{\lambda}(\cdot)$ is exactly equal to the true $\lambda(\cdot)$, zero if it is equal to the average intensity $\overline{\lambda}$, and negative if it performs worse than $\overline{\lambda}$. We compute the Q^2 indicator on a regular grid of 1000 locations in $\mathcal{S}$. Then, we compute the mean μ and one standard deviation σ of the Q^2 results across 20 different replicates. Table 1 shows that our approach outperforms its competitors, with consistently higher means of the Q^2 results and lesser dispersion σ.

We assess the computational cost of our approach using the third toy example λ_3 for $N_o = 100$ (which has the largest number of events with on average 22500 events in total). Obtaining one sample using our approach takes around 60 milliseconds, and generating all 10^4 samples takes 10 minutes in total (in contrast to the 18 minutes required by VBPP).[4] The multivariate effective sample size (ESS) (Flegal et al., 2017) was estimated at 322, corresponding to an effective sampling rate of 0.536 s^{-1}.

[4]These experiments were executed on a single core of an Intel® Core™ i7-6700HQ CPU.

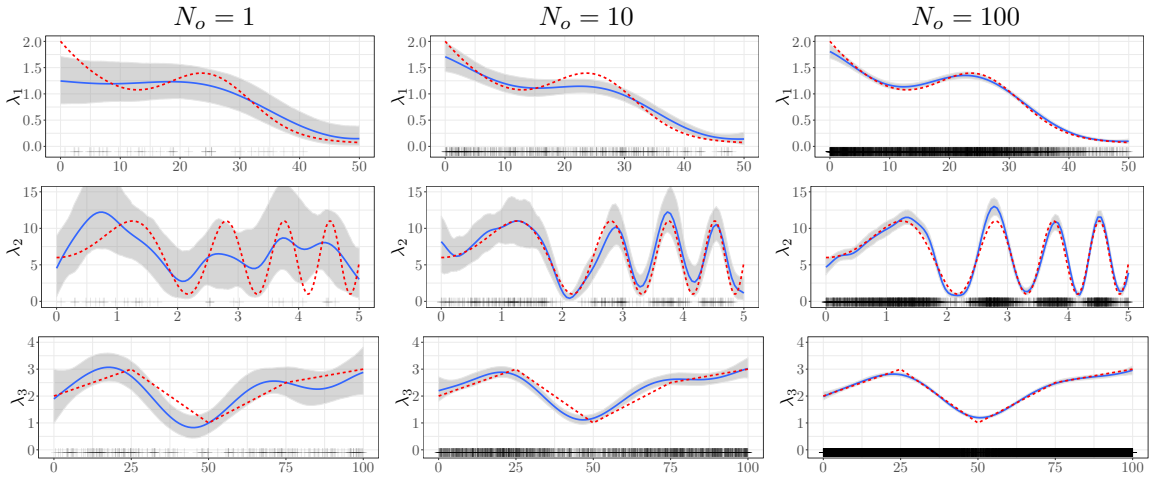


Figure 2: Inference results with multiple observations ($N_o = 1, 10, 100$) using the toy examples from Adams et al. (2009). Each panel shows the point patterns (black crosses), the true intensity λ (red dashed lines) and the intensity inferred by the finite approximation of GP-modulated Cox processes (blue solid lines). The estimated 90% confidence intervals of the finite approximation are shown in grey.

Table 1: Q^2 results for the toy examples of Figure 2, averaged over 20 (†10) replicates. Our results (cGP-c_+) are compared to results for Møller et al. (2001) (log-GP) and Lloyd et al. (2015) (VBPP).

Toy	N_o	Q^2 ($\mu \pm \sigma$) [%] log-GP	VBPP	cGP-c_+
λ_1	1	51.2±30.1	51.9±26.1	**65.7±14.3**
	10	95.1± 3.9	94.6± 3.7	**95.4± 2.3**
	100	**99.5± 0.2**	**99.5± 0.3**	**99.5± 0.3**
λ_2	1	-35.2±43.4	-1.1±28.8	**0.7±24.0**
	10	72.6± 9.1	71.7±10.4	**81.9± 7.4**
	100	95.4± 0.7	92.1± 3.9	**97.8± 0.6**
λ_3	1	49.2±22.6	49.5±29.9	**58.1±21.4**
	10	91.7± 4.4	93.8± 2.8	**94.3± 2.5**
	100	98.4± 0.4	**98.9±0.3**†	**98.8± 0.3**

5.2 Modelling Hazard Rates in Renewal Processes

Poisson processes have been extended to model renewal processes where intensity functions are seen as hazard rates defining the probability that an operating object fails (Serfozo, 2009; Cha and Finkelstein, 2018). However, in many application, e.g. reliability engineering and survival analysis, hazard rates exhibit monotonic behaviours describing the degradation of items or lifetime of organisms. For example, the hazard functions for the failure of many mechanistic devices and the mortality of adult humans tend to exhibit monotonic behaviours. Thus, taking monotonicity constraints into account in renewal processes is crucial for the study of many applications. Moreover, it is known that introducing monotonicity information in GPs can lead to more realistic uncertainties (Riihimäki and Vehtari, 2010; Maatouk and Bay, 2017).

As discussed in Section 2, some renewal processes can be seen as Cox processes under certain conditions. In order to demonstrate that we can model other types of point patterns, here we use two toy examples where hazard rates are known to be monotonic. Both examples are inspired by two classical renewal process: Weibull process and Gamma process.

For the first class, the Weibull hazard function is

$$\lambda^W(x) = \alpha\beta x^{\beta-1} \quad \text{for} \quad x \geq 0, \tag{18}$$

where α and β are the scale and shape parameters, respectively. Depending on β, λ^W can be either non-increasing ($0 < \beta < 1$), constant ($\beta = 1$), or non-decreasing ($\beta > 1$). Moreover, for $\beta \in (0, 1]$, the Weibull renewal process can be seen as a Cox process (Yannaros, 1988). For numerical experiments, we consider the case of non-increasing conditions in the domain $\mathcal{S} = [0, 100]$ by fixing $\alpha = 1$ and $\beta = 0.7$ (see Figure 3). We test our framework using $N_o = 100$ observations from λ^W, and we consider non-negativeness conditions, with (cGP-$c_+^\downarrow$) or without (cGP-c_+) taking into account the non-increasing constraint. We also consider the case where λ^W is non-increasing and convex (cGP-$c_+^{\downarrow\downarrow}$).

For the Gamma class, the hazard function is given by

$$\lambda^G(x) = \frac{\alpha\, x^{\beta-1} e^{-x}}{\Gamma(\beta) - \Gamma_x(\beta)}, \quad \text{for} \quad x \geq 0, \tag{19}$$

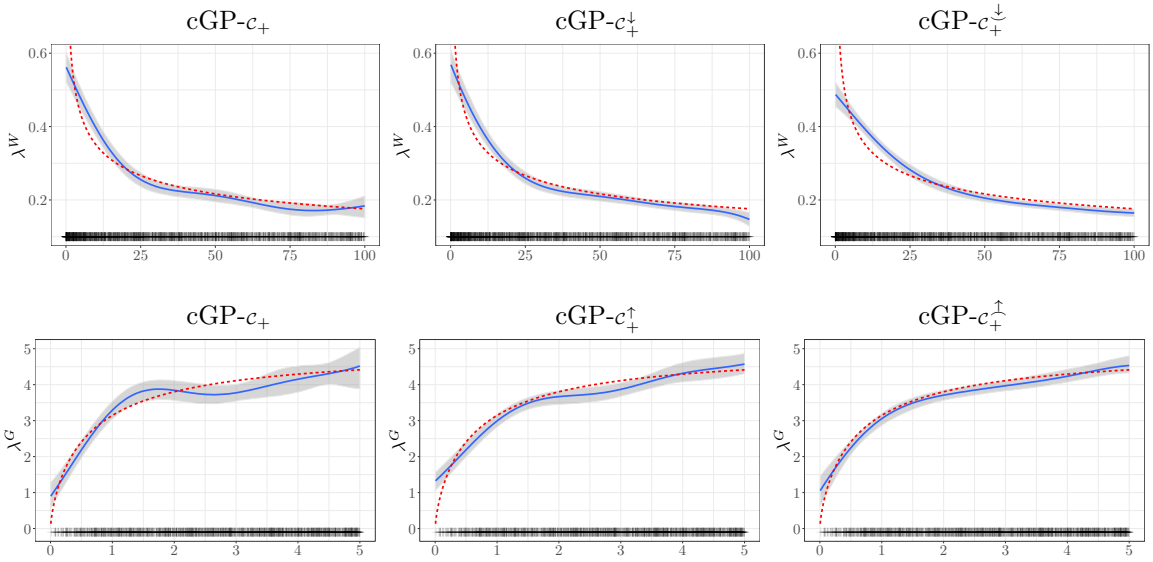


Figure 3: Renewal inference examples under different inequality constraints using $N_o = 100$ and $m = 100$. Inference results are shown for (top row) a Weibull renewal process with $\alpha = 1$ and $\beta = 0.7$, and (bottom row) a Gamma renewal process with $\alpha = 5$ and $\beta = 1.7$. The panel description is the same as in Figure 2.

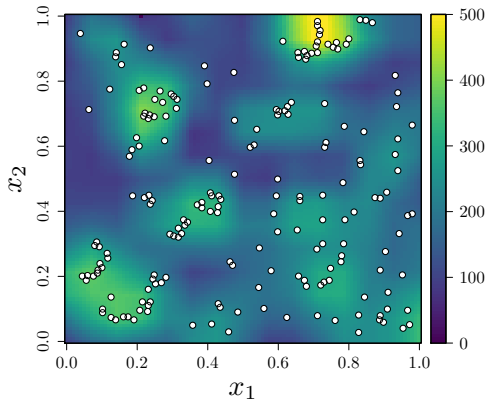
where $\Gamma(\cdot)$ and $\Gamma_x(\cdot)$ are the Gamma function and the incomplete Gamma function, respectively (Cha and Finkelstein, 2018), and α and β are the scale and shape parameters. As for the Weibull process, different behaviours can be obtained using different values of β. Since similar profiles are obtained for $\beta \in (0, 1]$, here we are interested in the case where λ^G exhibits non-decreasing constraints ($\beta > 1$). We fix $\mathcal{S} = [0, 5]$, $\alpha = 5$ and $\beta = 1.7$ obtaining a non-decreasing profile as shown in Figure 3. Here, we consider non-decreasing (cGP-$c_+^{\uparrow}$), and non-decreasing and concave (cGP-$c_+^{\overset{\uparrow}{\frown}}$) constraints. Since $\lambda^G(x) < \alpha$ for $x \in \mathcal{S}$, we add the constraint $\lambda^G \in [0, \alpha]$.

Figure 3 shows the inferred intensities of λ^W and λ^G under the different conditions previously discussed. In both experiments, we fixed $m = 100$ and $\eta = 10^{-4}$. For the Weibull class λ^W, the performance of all three models, cGP-c_+, cGP-$c_+^{\downarrow}$ and cGP-$c_+^{\overset{\downarrow}{\frown}}$, tends to be similar. However, the model without monotonicity constraint exhibits undesired oscillations, whereas the other two approaches provide more realistic decreasing profiles and more accurate inference results for $x > 50$. We can also observe that the three models cannot learn the singularity at $x = 0$. Note that the proposed methodology does not make any assumption on the kernel, and it would be possible to consider a covariance function such as $k(x, y)/(xy)$ in order to improve the model behaviour for small and large values of x. For the Gamma hazard function λ^G, one may clearly observe the benefits of adding the non-decreasing and concave constraints, obtaining absolute improvements between 0.8% and 3.5% of the Q^2 indicator. Both examples of Figure 3 show that the monotonicity and convexity conditions found in certain point processes can be difficult to learn directly from the data. This suggests that including those constraints in the GP prior is necessary to get accurate models with more realistic uncertainties.

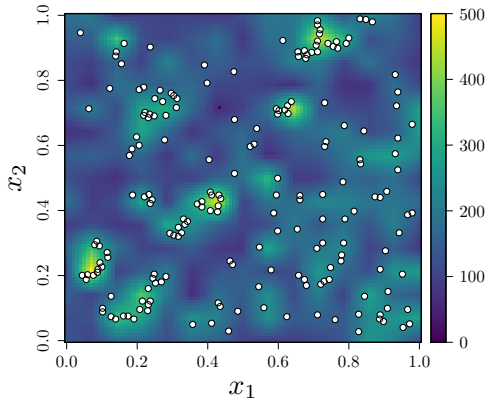
5.3 2D Redwoods Data

We now assess the performance of the proposed approach for a 2D spatial problem. We use the dataset provided by Ripley (1977) which describes the locations of redwood trees. The dataset contains $n = 195$ events scaled to the unit square (see Figure 4). Here we choose $m = 15$, obtaining 225 knots in total, to obtain a good trade-off between resolution and computational cost. We use the product of two SE kernels with covariance parameters $\boldsymbol{\theta} = (\sigma^2, \ell_1, \ell_2)$ as the covariance function of the Gaussian vector $\boldsymbol{\xi}$, and we choose $\eta = 10^{-4}$ in Algorithm 1. Following the burn-in step, we keep 10^5 samples for the inference of $\lambda(\cdot)$, yielding a total running time of 7.6 hours (i.e. a sampling rate of approximately 4 s^{-1}).

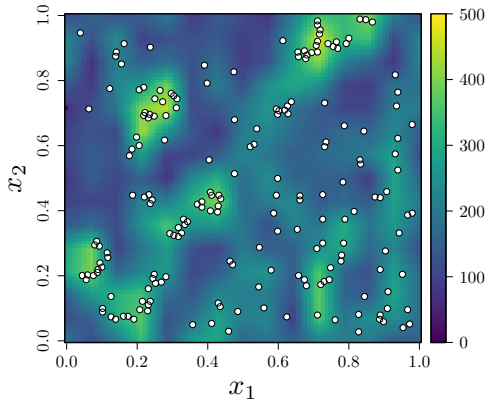
Figure 4 shows the normalised inference results for the redwood dataset for different values of the lengthscale parameters. Since in our approach we directly impose the inequality conditions on the Gaussian vector $\boldsymbol{\xi}$ instead of using a link function, the interpretation of the lengthscale parameters (ℓ_1, ℓ_2) are the same as for standard GPs: one can find a trade-off between fidelity and regularity by tuning ℓ. One can note, from Figures 4(a) and 4(b), that both profiles tend to



(a) $\ell_1 = \ell_2 = 0.1$



(b) $\ell_1 = \ell_2 = 0.01$



(c) $\widehat{\ell}_1 = 0.055$, $\widehat{\ell}_2 = 0.084$

Figure 4: Inference results of the redwoods data from Ripley (1977); Baddeley et al. (2015). Each panel shows the point pattern (white dots) and the estimated intensity $\lambda(\cdot)$.

properly learn the point patterns but more regularity is exhibited when $\ell_1 = \ell_2 = 10^{-1}$. For the case $\ell_1 = \ell_2 = 10^{-2}$, although the model follows the point patterns, one may observe noisy behaviour in regions without points, e.g. around $(x_1, x_2) = (0.30, 0.85)$, as small values of ℓ lead to more oscillatory Gaussian random fields. Finally, we infer $\lambda(\cdot)$ when the covariance parameters $\boldsymbol{\theta}$ are estimated via maximum likelihood using (10). According to the estimated lengthscales ($\widehat{\ell}_1 = 0.055$, $\widehat{\ell}_2 = 0.084$), one can conclude that the estimated intensity $\lambda(\cdot)$ is smoother along the second dimension x_2. This is in agreement with the inference results by Adams et al. (2009), where more variations of $\lambda(\cdot)$ were exhibited across x_1.

6 CONCLUSIONS

The proposed model for GP-modulated Cox processes is based on a finite-dimensional approximation of a GP that is constrained to be positive. This approach shows several advantages. First of all, it is based on general linear inequality constraints so it allows us to incorporate more information, such as monotonicity and convexity, in the prior. As seen in the experiments, this appears to be particularly helpful when few data are available. Second, imposing directly the positivity constraint on the GP makes the use of a link function unnecessary. Both the likelihood and the intensity measure can be computed analytically, which is not always the case when using a link function. Finally, the fact that our model is based on a finite-dimensional representation ensures that the computational burden grows linearly with the number of observations.

There are two key elements that make the method work: (a) the finite-dimensional representation of the GP that ensures that the constraints are satisfied everywhere, and (b) the dedicated MCMC proposal distribution based on a truncated normal distribution which allows us to have high acceptance rates compared to a naive multivariate Gaussian proposal.

The main limitation regarding the scaling of the proposed method lies in the dimension of the input space. This is due to the construction by tensorisation of the basis functions used to obtain the finite-dimensional representation. Moreover, our model is also sensitive to three parameters: the dimensionality of the space in which we perform HMC, the number of constraints, and the number of times the HMC particles violate a constraint. However, we believe that these limitations are not inherent to the proposed model and that other types of designs of the knots (e.g. sparse designs) could be used in high dimensions.

Acknowledgements

This work was supported by the Chair in Applied Mathematics OQUAIDO (`oquaido.emse.fr`, France) and by PROWLER.io (`www.prowler.io`, UK). We thank O. Roustant (EMSE) and D. Rullière (ISFA) for their advice throughout this work.

References

Adams, R. P., Murray, I., and MacKay, D. J. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *ICML*, pages 9–16.

Baddeley, A., Gregori, P., Mahiques, J., Stoica, R., and Stoyan, D. (2006). *Case Studies in Spatial Point Process Modeling*. Lecture Notes in Statistics. Springer, New York.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press, Boca Raton, FL.

Bay, X., Grammont, L., and Maatouk, H. (2016). Generalization of the Kimeldorf–Wahba correspondence for constrained interpolation. *Electronic Journal of Statistics*, 10(1):1580–1595.

Botev, Z. I. (2017). The normal law under linear restrictions: Simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B*, 79(1):125–148.

Cha, J. and Finkelstein, M. (2018). *Point Processes for Reliability Analysis: Shocks and Repairable Systems*. Springer Series in Reliability Engineering. Springer, New York.

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B*, 17(2):129–164.

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

Donner, C. and Opper, M. (2018). Efficient Bayesian inference of sigmoidal Gaussian Cox processes. *Journal of Machine Learning Research*, 19(67):1–34.

Fernandez, T., Rivera, N., and Teh, Y. W. (2016). Gaussian processes for survival analysis. In *NIPS*, pages 5021–5029.

Flaxman, S., Wilson, A., Neill, D., Nickisch, H., and Smola, A. (2015). Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. In *ICML*, pages 607–616.

Flegal, J. M., Hughes, J., Vats, D., and Dai, N. (2017). `mcmcse`: Monte Carlo standard errors for MCMC. `https://cran.r-project.org/web/packages/mcmcse/`.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–150.

Gunter, T., Lloyd, C. M., Osborne, M. A., and Roberts, S. J. (2014). Efficient Bayesian nonparametric modelling of structured point processes. In *UAI*, pages 310–319.

John, S. T. and Hensman, J. (2018). Large-scale Cox process inference using variational Fourier features. In *ICML*, pages 2362–2370.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492.

Kingman, J. (1992). *Poisson Processes*. Oxford Studies in Probability. Clarendon Press, New York.

Kozachenko, Y., Pogorilyak, O., Rozora, I., and Tegza, A. (2016). Simulation of Cox random processes. In *Simulation of Stochastic Processes with Given Accuracy and Reliability*, pages 251–304. Elsevier, Amsterdam.

Lasko, T. A. (2014). Efficient inference of Gaussian-process-modulated renewal processes with application to medical event data. In *UAI*, pages 469–476.

Lloyd, C. M., Gunter, T., Osborne, M. A., and Roberts, S. J. (2015). Variational inference for Gaussian process modulated Poisson processes. In *ICML*, pages 1814–1822.

López-Lopera, A. F. (2018). `lineqGPR`: Gaussian process regression models with linear inequality constraints. `https://cran.r-project.org/web/packages/lineqGPR/`.

López-Lopera, A. F., Bachoc, F., Durrande, N., and Roustant, O. (2018). Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1224–1255.

Maatouk, H. and Bay, X. (2017). Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, 49(5):557–582.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (2001). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

Møller, J. and Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective (Adaptive Computation And Machine Learning)*. The MIT Press, Cambridge.

Pakman, A. and Paninski, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. MIT Press, Cambridge.

Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. In *AISTATS*, pages 645–652.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B*, 39(2):172–212.

Serfozo, R. (2009). *Renewal and Regenerative Processes*, pages 99–167. Springer, New York.

Sleeper, L. A. and Harrington, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, 85(412):941–949.

Teh, Y. W. and Rao, V. (2011). Gaussian process modulated renewal processes. In *NIPS*, pages 2474–2482.

Yannaros, N. (1988). On Cox processes and Gamma renewal processes. *Journal of Applied Probability*, 25(2):423–427.

Yannaros, N. (1994). Weibull renewal processes. *Annals of the Institute of Statistical Mathematics*, 46(4):641–648.