Markov Modulated Gaussian Cox Processes for Semi-Stationary Intensity Modeling of Events Data

Minyoung Kim [1 2]

Abstract

The Cox process is a flexible event model that can account for uncertainty of the intensity function in the Poisson process. However, previous approaches make strong assumptions in terms of time stationarity, potentially failing to generalize when the data do not conform to the assumed stationarity conditions. In this paper we bring up two most popular Cox models representing two extremes, and propose a novel semi-stationary Cox process model that can take benefits from both models. Our model has a set of Gaussian process latent functions governed by a latent stationary Markov process where we provide analytic derivations for the variational inference. Empirical evaluations on several synthetic and real-world events data including the football shot attempts and daily earthquakes, demonstrate that the proposed model is promising, can yield improved generalization performance over existing approaches.

1. Introduction

Accurately modeling the underlying generative process of complex events is an important problem in statistical machine leaning and many related areas. Although events could be spatial and/or high-dimensional, in this paper we exclusively focus on event modeling in the temporal setup due to its dominance in real-world applications. The Poisson process is a de facto standard for its simplicity in mathematical analysis and flexibility in representing the intensity function (i.e., the event occurring rate) $\lambda(t)$. Unlike traditional treatments via adopting a fixed parametric form of $\lambda(t)$ (e.g., piecewise constant or the Weibull), several extensions have been introduced. The nonparametric modeling of $\lambda(t)$ (e.g., the recent RKHS formulation (Flaxman et al., 2017)) can reduce the burden of deciding an appropriate form of $\lambda(t)$. Another is to regard $\lambda(t)$ as a random process, known as the Cox process (Cox, 1955), which is useful for accounting for uncertainty in the intensity function.

In this paper, we are particularly interested in the Cox process where two most popular ones are: the Markov modulated Poisson process (MMPP) and the Gaussian process modulated Cox process (GPCox). Popular in statistics, the MMPP considers $\lambda(t)$ as a random sample (trajectory) from a continuous-time Markov chain. The model has a finite set of intensity levels where the latent state at each time determines which intensity level is used at that moment. The GPCox is a nonparametric Bayesian model formed by placing a Gaussian process (GP) prior on $\lambda(t)$. The GPCox has received significant attention in the machine learning community for the last decade for its flexible nonparametric modeling with principled uncertainty treatment.

The MMPP is good at modeling highly different intensity phases: bursty events for some intervals and rare events for others. However, there can be abrupt intensity changes between these regimes which may be unnatural in certain situations. Furthermore, for the interval under a given latent state, the model follows a constant intensity (i.e., a homogeneous process), which may limit its representational capacity. On the other hand, the GPCox encourages smooth intensity changes over time. However, the disadvantage is that the drastic intensity changes are not properly dealt with unless a highly non-smooth kernel is adopted, which can usually happen when a large amount of data is available.

So the main idea in this paper is to devise a novel model that takes benefits from both models. Similar to the MMPP, we consider an underlying continuous-time Markov chain (CTMC) that generates a latent state trajectory (taking say, r different states). We incorporate r latent functions with their own GP priors, each of which serves as the intensity responsible for each of the r states. This model is thus able to model major intensity regime (possibly abrupt) changes via the CTMC dynamics, and at the same time, it also enjoys the GP's smooth intensity modeling, non-constant within the interval under a given latent state. Our model, referred to as the *Markov modulated Gaussian Cox Process* (MMGCP), has richer representational power than previous two models.

[1]Seoul National University of Science & Technology, Korea [2]Rutgers University, Piscataway, NJ, USA. Correspondence to: Minyoung Kim <mikim21@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Indeed, it subsumes both models as special cases: i) if the GP priors put all their masses to constant functions, then we end up with the MMPP, and ii) if $r = 1$ (single-state), then the model reduces to the GPCox.

In terms of time-stationarity[1], the previous two models exhibit extreme characteristics. The MMPP makes $\lambda(t)$ *fully stationary* (i.e., time independent) since the CTMC is stationary and the intensity under a given state is a constant, invariant of t. On the other hand, the GPCox builds a *fully non-stationary* (time-variant) $\lambda(t)$ on top of the kernel function defined over t. Our MMGCP somehow aims to model a so-called *semi-stationary* intensity function in that the macro-scale intensity regime change is governed by the stationary CTMC dynamics, while within each regime, the intensity is modeled as a smooth time-variant function. In this sense, an ideal scenario for our model is as follows: there are r underlying candidate intensity functions $\{\lambda^i(t)\}_{i=1}^r$ where at a given time t, which of these candidates is active is determined by the stationary r-state Markov process $X(t)$, that is, $\lambda^{X(t)}(t)$. Our model further imposes the GP prior on these candidate functions to account for uncertainty and grant more modeling flexibility. In the evaluations, we not only implement this scenario as a synthetic setup, but we also demonstrate on some real datasets that our MMGCP significantly outperforms the previous models.

We provide an efficient variational inference for the model which is also analytic by adopting the squared link function for the intensity, similar to that of (Lloyd et al., 2015). However, the posterior expectation over the latent state trajectory required in the variational inference, is carefully analyzed within our model to derive closed-form formulas. The paper is organized as follows. After briefly discussing some background and reviewing previous two models in Sec. 2, our model is introduced in Sec. 3 with the variational inference fully described in Sec. 4. The empirical evaluation on some synthetic and real-world event datasets follows in Sec. 5.

2. Background

We are interested in modeling events that can occur over the fixed time horizon $[0, T]$. We basically assume that the events are generated by the (inhomogeneous) Poisson process, which is fully specified by the non-negative intensity function $\lambda : [0, T] \to \mathbb{R}_+$. It defines the event occurring rate (i.e., the probability of event occurring during the infinitesimal interval $[t, t + dt)$ is $\lambda(t)dt$). Then the log-likelihood of observing the event data $\mathcal{D} = \{t_1, \ldots, t_N\}$ ($\subset [0, T]$) can be written as:

$$\log P(\mathcal{D}|\lambda(\cdot)) = \sum_{n=1}^{N} \log \lambda(t_n) - \int_0^T \lambda(t)\, dt. \quad (1)$$

It is common in statistics to assume a specific parametric form for $\lambda(t)$, then estimate it by the maximum likelihood criterion with (1). Instead, the Cox process further regards $\lambda(t)$ as a random process. Two most popular ones are briefly described below.

2.1. Markov Modulated Poisson Process (MMPP)

This model basically forms piecewise constant $\lambda(t)$. Specifically there are r constant intensity levels $\{\overline{\lambda}_1, \ldots, \overline{\lambda}_r\}$, but which level is used at a given moment is determined by the latent Markov process $X : [0, T] \to \{1, \ldots, r\}$ governed by a continuous-time Markov chain (CTMC). An r-state CTMC is specified by the initial state probability $\pi_i = P(X(0) = i)$ for $i = 1, \ldots, r$, and the transition rate matrix Q whose off-diagonal Q_{ij} ($i \neq j$) defines the probability rate of state change from i to j, namely

$$Q_{ij} = \lim_{\Delta t \to 0} \frac{P(X(t + \Delta t) = j | X(t) = i)}{\Delta t}. \quad (2)$$

Defining diagonal entries as $Q_{ii} := -\sum_{j \neq i} Q_{ij}$ lets the probability of staying at state i for duration h be $e^{hQ_{ii}}$. Note that the model has no time-variant component, thus adequate for modeling stationary event data. There are well-known EM learning algorithms (Asmussen et al., 1996; Ryden, 1996) for estimating the parameters of the model.

2.2. Gaussian Cox Process (GPCox)

The GPCox model has a latent function $f(t)$ distributed by a Gaussian process a priori, which determines the intensity function as $\lambda(t) = \rho(f(t))$ where $\rho(\cdot)$ is a non-negative link function, for instance, sigmoid, exponential or square function. The posterior inference $P(f(\cdot)|\mathcal{D})$ is challenging mainly due to the integration in the likelihood function (1). Let alone the computational overhead of evaluating the integral, one has to deal with latent function values at *all inputs* $t \in [0, T]$, not just those t_n's in the data $\mathcal{D}$ as in most conventional GP models (Rasmussen & Williams, 2006). Accordingly some previous approaches had to resort to discretizing the time domain (Rathbun & Cressie, 1994; Møller et al., 1998; Cunningham et al., 2008).

Recently, several sophisticated inference methods have been proposed to address this difficulty. (Adams et al., 2009) formed a tractable MCMC dynamics by exploiting the idea of thinning-based sampling in the Poisson process. However, its time complexity is cubic in the data size, which is often prohibitive for large-scale problems. To deal with the scalability, (Gunter et al., 2014) proposed an alternative

[1] For clarity, the term *stationarity* is used in the following sense: if there is no time dependency in the data generating model (eg, MMPP), we say that it is stationary; if the data generation process is solely dependent on the time index (eg, a non-constant deterministic intensity function), then it is non-stationary. As our model contains both components, we call it semi-stationary.

thinning strategy by sampling from a non-uniform intensity process, while (Samo & Roberts, 2015) introduced inducing points within the MCMC sampler. (Lasko, 2014) used a positively transformed intensity function for direct numerical integration and interpolation. In parallel, (Lloyd et al., 2015) derived analytic formulation for the scalable variational inference using the square link function and the pseudo input treatment (Titsias, 2009; Dezfouli & Bonilla, 2015).

3. Markov Modulated Gaussian Cox Process

In this section we describe our model that can take benefits from previous models in Sec. 2. We consider that there are r underlying latent functions devoted for representing different characteristics of the intensity function. Denoted by $\mathbf{f}(\cdot) := \{f^i(\cdot)\}_{i=1}^r$, they are assumed to be independently GP distributed a priori. That is,

$$P(\mathbf{f}(\cdot)) = P(f^1(\cdot), \dots, f^r(\cdot)) = \prod_{i=1}^r P(f^i(\cdot)) \quad (3)$$
$$\text{where} \quad f^i(\cdot) \sim \mathcal{GP}\big(m^i(\cdot), k^i(\cdot,\cdot)\big),\ i = 1, \dots, r.$$

To determine which of these r functions is responsible for the intensity at each time, we introduce a latent Markov process $X(t)$, similarly as the MMPP, generated from a r-state CTMC (Q, π). The intensity at time t is then determined by $f^{X(t)}$, and we use the square link function similarly as (Lloyd et al., 2015), which leads to:

$$\lambda(t) \mid \mathbf{f}(\cdot), X(\cdot) = (f^{X(t)})^2. \quad (4)$$

The full joint distribution of the model can be written as:

$$P(\mathcal{D}, X(\cdot), \mathbf{f}(\cdot)|\Theta, \Omega) = \quad (5)$$
$$P(\mathbf{f}(\cdot)|\Theta) \times P(X(\cdot)|\Omega) \times P(\mathcal{D}|X(\cdot), \mathbf{f}(\cdot)),$$

where $\Theta = \{\boldsymbol{\theta}_m, \boldsymbol{\theta}_k\}$ is the parameters of the mean and covariance functions of the prior GP (e.g., $\boldsymbol{\theta}_k = \{\theta_k^i\}_{i=1}^r$ with θ_k^i denoting the parameters of the covariance function $k^i(\cdot,\cdot)$ for $f^i(\cdot)$). The CTMC parameters are denoted by $\Omega = \{Q, \pi\}$. Thus Θ and Ω constitute the model parameters of the MMGCP. The last two terms in the RHS of (5) correspond to the likelihood of the state trajectory under the CTMC and the data likelihood given the state trajectory and the latent functions. To formally derive these likelihoods, it is convenient to partition the horizon $[0, T]$ according to a realized state trajectory $X(\cdot)$. Suppose that a realization $X(\cdot)$ undergoes $(L-1)$ state changes during $[0, T]$. We denote by u_l the time epoch when the l-th state change occurs ($l = 1, \dots, L-1$) with $u_0 = 0$ and $u_L = T$ for convenience. We let $s_l \in \{1, \dots, r\}$ be the state during the interval $[u_{l-1}, u_l)$, and Δu_l be the length of the interval (i.e., $\Delta u_l = u_l - u_{l-1}$). Note that these variables $\{u_l, s_l, \Delta u_l\}_l$ are determined by the realization $X(\cdot)$, and vice versa, in a one-to-one manner.

Looking into the likelihood of X restricted to each interval $(u_{l-1}, u_l]$, it is composed of two steps: i) no state change during (u_{l-1}, u_l) and ii) state change from s_l to s_{l+1} right at the moment u_l. For the last interval ($l = L$), it only involves the step i). From the well-known theorems of the CTMC[2], the probability of the first step is $\exp(\Delta u_l Q_{s_l s_l})$, while the likelihood of the second step is $Q_{s_l s_{l+1}}$. Combining these over $l = 1, \dots, L$ and including the initial state probability $P(X(0) = s_1) = \pi_{s_1}$, we have the likelihood of the state trajectory as follows.

$$P(X(\cdot)|\Omega) = \pi_{s_1} \times \prod_{l=1}^{L} e^{\Delta u_l Q_{s_l s_l}} \times \prod_{l=1}^{L-1} Q_{s_l s_{l+1}}. \quad (6)$$

To derive the likelihood of observing $\mathcal{D}$ given $X(\cdot)$ and $\mathbf{f}(\cdot)$, we let $\{t_1^l, \dots, t_{k_l}^l\}$ be the event times in $\mathcal{D}$ that fall into the interval $\mathcal{I}_l := [u_{l-1}, u_l)$. Within $\mathcal{I}_l$, the intensity is fixed as $\lambda^{s_l}(t) := (f^{s_l}(t))^2$, and applying the Poisson process likelihood gives: $\lambda^{s_l}(t_1^l) \cdots \lambda^{s_l}(t_{k_l}^l) \exp(-\int_{\mathcal{I}_l} \lambda^{s_l}(t) dt)$. Multiplying them over $l = 1, \dots, L$ yields:

$$P(\mathcal{D}|X, \mathbf{f}) = \prod_{\substack{1 \le l \le L, \\ n: t_n \in \mathcal{I}_l}} (f^{s_l}(t_n))^2 \times e^{-\int_{\mathcal{I}_l} (f^{s_l}(t))^2 dt}. \quad (7)$$

4. Variational Inference and Learning

In this section we provide inference for the posterior distribution in our MMGCP model, specifically

$$P(X(\cdot), \mathbf{f}(\cdot)|\mathcal{D}, \Theta, \Omega). \quad (8)$$

This inference is analytically intractable, however, we do it approximately using the recent scalable variational inference technique[3] (Titsias, 2009; Dezfouli & Bonilla, 2015; Lloyd et al., 2015; Matthews et al., 2016). It is based on the pseudo inputs which especially plays a crucial role of making inference tractable even if one has to deal with function values for all inputs $t \in [0, T]$. So we begin with the introduction of our GP notations regarding pseudo inputs.

We often use the superscript for indicating a specific function among the r latent functions, while the subscript for a specific time epoch or a set of time epochs at which the functions are evaluated. For instance, for a set of p inputs $\mathcal{T} = \{\tilde{t}_1, \dots, \tilde{t}_p\} \subset [0, T]$, we denote by $f_{\mathcal{T}}^i = [f^i(\tilde{t}_1), \dots, f^i(\tilde{t}_p)]^\top$, the p-dim vector of the i-th function

[2] The full derivation can be found in standard textbooks on Markov chains or stochastic ODEs such as (Anderson, 2011). We also provide some brief derivations in Appendix A in the supplemental materials.

[3] We specifically follow the variational free energy approach. But we would like to note that there exist other approximation techniques where the readers are encouraged to refer to the recent work on comparison of different approaches (Bauer et al., 2016) and some unified view (Bui et al., 2017).

values. The boldfaced $\mathbf{f}_{\mathcal{T}}$ indicates the collection of the function values for all r functions, that is, $\mathbf{f}_{\mathcal{T}} = \{f^1_{\mathcal{T}}, \ldots, f^r_{\mathcal{T}}\}$. For the GP prior mean and covariance functions (3), we follow the similar convention: $m^i_{\mathcal{T}} = [m^i(\tilde{t}_1), \ldots, m^i(\tilde{t}_p)]^\top$ is the p-dim vector of the i-th mean function values on $\mathcal{T}$. For two input sets $\mathcal{T}$ and $\mathcal{S}$, $K^i_{\mathcal{T},\mathcal{S}}$ denotes the $(|\mathcal{T}| \times |\mathcal{S}|)$ kernel matrix by applying $k^i(\cdot,\cdot)$ on $(\mathcal{T} \times \mathcal{S})$.

For each i-th GP ($i = 1, \ldots, r$), we assume that there are M_i ($\ll N$) pseudo inputs denoted by $\mathcal{Z}^i = \{z^i_1, \ldots, z^i_{M_i}\} \subset [0, T]$. We also let $\mathcal{Z} = \bigcup_{i=1}^r \mathcal{Z}^i$. These pseudo inputs can be thought of as representative points in that knowing the function values at $\mathcal{Z}$ has significant impacts on inferring function values at the other input points. But further insights can be found in the nice survey (Quiñonero-Candela & Rasmussen, 2005). The pseudo inputs can also be learned from data along with the model parameters, but for the time being we assume that they are fixed[4].

We denote the whole state trajectory and the function values of all r latent functions for the entire set $[0, T]$ as (infinite dimensional) X and $\mathbf{f}$, respectively. We define a tractable form of the variational density $q(X, \mathbf{f})$, and optimize it to approximate the true posterior (8) as much as possible. In defining $q(\cdot)$, we impose independence between X and $\mathbf{f}$ for computational tractability. First, we let the posterior distribution of X follows a CTMC, which allows analytic derivations feasible as will be shown below. Also the posterior of the latent functions $\mathbf{f}$ are assumed to be Gaussians factorized over $i = 1, \ldots, r$. Furthermore, we force the conditional density $q(\mathbf{f}|\mathbf{f}_{\mathcal{Z}})$ to coincide with the prior $P(\mathbf{f}|\mathbf{f}_{\mathcal{Z}})$ exactly, which is crucial to have some difficult terms canceled out in the KL divergence objective, making the inference scalable (Titsias, 2009; Lloyd et al., 2015). In summary, our variational density is defined as:

$$q(X, \mathbf{f}) = q(X; C, \alpha) \times \int q(\mathbf{f}_{\mathcal{Z}}) P(\mathbf{f}|\mathbf{f}_{\mathcal{Z}}) \, d\mathbf{f}_{\mathcal{Z}} \quad (9)$$

where C is the $(r \times r)$ transition rate matrix and α is the $(1 \times r)$ initial state probabilities for the CTMC $q(X)$. Also,

$$q(\mathbf{f}_{\mathcal{Z}}) = \prod_{i=1}^r \mathcal{N}(f^i_{\mathcal{Z}^i}; \mu^i, \Sigma^i). \quad (10)$$

Here μ^i is the M_i-dim mean vector and Σ^i is the $(M_i \times M_i)$ covariance matrix. The variational parameters are denoted as $\Lambda := (C, \alpha, \boldsymbol{\mu} := \{\mu^i\}_{i=1}^r, \boldsymbol{\Sigma} := \{\Sigma^i\}_{i=1}^r)$.

We aim to minimize the KL divergence between $q(\cdot)$ and the posterior (8), which can be written as:

$$\mathrm{KL}\big(q(X, \mathbf{f})||P(X, \mathbf{f}|\mathcal{D})\big) = \log P(\mathcal{D}) - \mathrm{ELBO}(\Theta, \Omega, \Lambda), \quad (11)$$

where the ELBO (evidence lower-bound) is defined as:

$$\mathrm{ELBO}(\Theta, \Omega, \Lambda) = \mathbb{E}_{q(X,\mathbf{f})}\big[\log P(\mathcal{D}|X, \mathbf{f})\big] - \mathrm{KL}\big(q(X)||P(X)\big) - \mathrm{KL}\big(q(\mathbf{f}_{\mathcal{Z}})||P(\mathbf{f}_{\mathcal{Z}})\big). \quad (12)$$

[4] We often use the uniformly sampled points from $[0, T]$.

From (11) and the fact that KL divergence is non-negative, the ELBO is the lower bound of the log-evidence, namely

$$\log P(\mathcal{D}|\Theta, \Omega) \geq \mathrm{ELBO}(\Theta, \Omega, \Lambda). \quad (13)$$

Note that the bounding gap in (13) is exactly the KL divergence between $q(\cdot)$ and the posterior. Thus increasing $\mathrm{ELBO}(\Theta, \Omega, \Lambda)$ wrt Λ leads to a better variational density (closer to the posterior), whereas increasing it wrt the model parameters (Θ, Ω) can *potentially*[5] improve the data evidence score of the model. Hence, maximizing the ELBO wrt all the parameters can achieve both variational inference (i.e., $q(\cdot)$ optimization) and model selection (i.e, learning prior model parameters) simultaneously.

In what follows, we provide full derivations for evaluating each term comprising (12). The gradients are also required for the optimization of the ELBO, and can be found in Appendix C in the supplemental material.

4.1. $\mathrm{KL}\big(q(\mathbf{f}_{\mathcal{Z}})||P(\mathbf{f}_{\mathcal{Z}})\big)$

It is not difficult to see that due to the fully factorized $q(\mathbf{f}_{\mathcal{Z}})$ and $P(\mathbf{f}_{\mathcal{Z}})$ over individual latent functions $i = 1, \ldots, r$, the KL divergence is the sum of the individual Gaussian KL divergences. More specifically,

$$\mathrm{KL}\big(q(\mathbf{f}_{\mathcal{Z}})||P(\mathbf{f}_{\mathcal{Z}})\big) = \sum_{i=1}^r \frac{1}{2}\left[\log \frac{|K^i_{\mathcal{Z}^i,\mathcal{Z}^i}|}{|\Sigma^i|} - M_i + \mathrm{Tr}\Big(\big(K^i_{\mathcal{Z}^i,\mathcal{Z}^i}\big)^{-1}\Sigma^i\Big) + (\mu^i - m^i_{\mathcal{Z}^i})^\top \big(K^i_{\mathcal{Z}^i,\mathcal{Z}^i}\big)^{-1}(\mu^i - m^i_{\mathcal{Z}^i})\right] \quad (14)$$

The gradients with respect to the related parameters are derived in Appendix C.1.

4.2. $\mathrm{KL}\big(q(X)||P(X)\big)$

This term involves computing the expectations of the log-likelihoods of the CTMC models (both the prior $\log P(X)$ and the variational posterior $\log q(X)$) with respect to $q(X)$. We describe how to compute $E_{q(X)}[P(X)]$ analytically ($E_{q(X)}[q(X)]$ done similarly). For this purpose, we rephrase the CTMC likelihood in (6) using some total statistics from the realization $X(\cdot)$. With $X(\cdot)$ fixed, let n_{ij} be the number of transitions from state i to j where $j \neq i$, and Δ_i be the sojourn time at state i, that is, $\Delta_i = \sum_{l:s_l=i} \Delta u_l$. Note that (n_{ij}, Δ_i) are the functions of $X(\cdot)$. Then we have:

$$\log P(X) = \sum_{i=1}^r \left(\mathbb{I}_{\{X(0)=i\}} \log \pi_i + \Delta_i Q_{ii} + \sum_{j \neq i} n_{ij} \log Q_{ij}\right), \quad (15)$$

[5] However, this does not guarantee to improve the evidence $\log P(\mathcal{D})$ since the inequality (13) is not tight.

where $\mathbb{I}_{\{p\}}$ is 1 (0) if the predicate p is true (false).

Thus the expectation of (15) requires: $E_q[n_{ij}]$ and $E_q[\Delta_i]$. For the latter, we first note that $\Delta_i = \int_0^T \mathbb{I}_{\{X(t)=i\}} dt$. Using $q(X(t)=i) = [\alpha e^{tC}]_i$ from the CTMC theorems (see (4) in Appendix A), we have:

$$E_{q(X)}[\Delta_i] = [\alpha J_C]_i, \tag{16}$$

where $J_C = \int_0^T e^{tC} dt$, is the $(r \times r)$ matrix by integrating the matrix exponential over $[0, T]$, and $[v]_i$ indicates the i-th element of the vector v. As the number of transitions $n_{ij} = \int_0^T \mathbb{I}_{\{X(t)=i \text{ AND } X(t+dt)=j\}}$, and using $q(X(t) = i, X(t+dt) = j) = [\alpha e^{tC}]_i C_{ij} dt$ ((5) in Appendix A),

$$E_{q(X)}[n_{ij}] = [\alpha J_C]_i C_{ij}. \tag{17}$$

By applying these to (15), we finally have:

$$\mathrm{KL}\big(q(X)||P(X)\big) = \sum_{i=1}^{r} \Bigg\{ \alpha_i \log \frac{\alpha_i}{\pi_i} \; + \; [\alpha J_C]_i \Big(C_{ii} - Q_{ii} + \sum_{j \neq i} C_{ij} \log \frac{C_{ij}}{Q_{ij}} \Big) \Bigg\}. \tag{18}$$

The remaining thing is how to compute J_C. It can be done analytically once the matrix C is diagonalized. The details are found in Appendix B of the supplemental material. Since r is usually small (e.g., 2 or 3), diagonalization must not incur any computational or numerical issues. When we compute the gradients of (18) with respect to C, special techniques of taking derivatives of matrix exponentials such as (Kalbfleisch & Lawless, 1985) can be used. The technical details are described in Appendix B and Appendix C.2.

4.3. $\mathbb{E}_{q(X,\mathbf{f})}\big[\log P(\mathcal{D}|X,\mathbf{f})\big]$

This is the conditional log-likelihood given the state trajectory and the latent functions, expected with respect to the variational posterior $q(X, \mathbf{f})$. From (7), after slight rephrasing, the conditional log-likelihood can be written as:

$$\log P(\mathcal{D}|X,\mathbf{f}) = \sum_{i=1}^{r} \sum_{n=1}^{N} \mathbb{I}_{\{X(t_n)=i\}} \log \left(f^i(t_n)\right)^2 - \sum_{i=1}^{r} \int_0^T \mathbb{I}_{\{X(t)=i\}} (f^i(t))^2 dt. \tag{19}$$

Exploiting the factorization $q(X, \mathbf{f}) = q(X)q(\mathbf{f})$, we take the expectation of (19) wrt $q(X)$ first, followed by $q(\mathbf{f})$. Using $q(X(t) = i) = [\alpha e^{tC}]_i$ from the previous section,

$$\mathbb{E}_{q(X,\mathbf{f})}\big[\log P(\mathcal{D}|X,\mathbf{f})\big] = \sum_{i=1}^{r} \big(\mathrm{ELL}_i - \mathrm{ENO}_i\big) \text{ where } (20)$$

$$\mathrm{ELL}_i = \sum_{n=1}^{N} [\alpha e^{t_n C}]_i \mathbb{E}_{q(f^i(t_n))}\big[\log \left(f^i(t_n)\right)^2\big], \tag{21}$$

$$\mathrm{ENO}_i = \int_0^T [\alpha e^{tC}]_i \mathbb{E}_{q(f^i(t))}\big[(f^i(t))^2\big] dt. \tag{22}$$

Note that (21) and (22) are very similar to those in the variational inference of the GPCox model proposed in (Lloyd et al., 2015). However, we have the weighted expected log-likelihood by the weights $[\alpha e^{tC}]_i$ over $i = 1, \ldots, r$, determined by the latent state posterior probabilities $q(X(t))$. Computing these weights for each t can be done analytically when we have a diagonalization of C (See Appendix B in the supplemental material). The expectations in (21) and (22) are with respect to Gaussians, more specifically, $q(f^i(t))) = \int q(f^i_{\mathcal{Z}^i}) P(f^i(t)|f^i_{\mathcal{Z}^i}) df^i_{\mathcal{Z}^i} = \mathcal{N}(\tilde{\mu}_i(t), \tilde{\sigma}^2_i(t))$ where

$$\tilde{\mu}_i(t) = m^i(t) + K^i_{t,\mathcal{Z}^i} (K^i_{\mathcal{Z}^i,\mathcal{Z}^i})^{-1} (\mu^i - m^i_{\mathcal{Z}^i}), \tag{23}$$

$$\tilde{\sigma}^2_i(t) = K^i_{t,t} - K^i_{t,\mathcal{Z}^i} (K^i_{\mathcal{Z}^i,\mathcal{Z}^i})^{-1} K^i_{\mathcal{Z}^i,t} + K^i_{t,\mathcal{Z}^i} (K^i_{\mathcal{Z}^i,\mathcal{Z}^i})^{-1} \Sigma^i (K^i_{\mathcal{Z}^i,\mathcal{Z}^i})^{-1} K^i_{\mathcal{Z}^i,t}. \tag{24}$$

Then the expectation in (22) equals $(\tilde{\mu}_i(t))^2 + \tilde{\sigma}^2_i(t)$, which allows us to compute the integral analytically for a certain kernel form (e.g., squared exponential or polynomial kernel) as shown in (Lloyd et al., 2015). With the additional weight term $[\alpha e^{tC}]_i$ multiplied to the integrand in our model, by rewriting the weight as a sum of *scalar* exponentials after diagonalization of C, we can still derive a closed-form expression for ENO_i. However, it is highly complicated, which becomes even worse when evaluating its gradients (e.g., wrt C). For the expectation of the log-squared term of ELL_i in (21), some confluent hyper-geometric function was adopted in (Lloyd et al., 2015), however, it is either numerically unstable or based on certain interpolation.

Instead, we employ fairly straightforward strategies for computing ELL_i and ENO_i. First, the expectation of the log-squared term in (21) is done by the Monte-Carlo estimation. This must not incur much computational overhead since it is *univariate* sampling. Considering that we have to take derivatives of ELL_i wrt the parameters related to $q(f^i(t_n))$, we also adopt the re-parametrized Gaussian sampling technique as suggested in (Kingma & Welling, 2014). The idea is to express the random samples from $q(f^i(t_n))$, denoted by $f_n^{i\,(s)}$ for $s = 1, \ldots, S$, as:

$$f_n^{i\,(s)} = \tilde{\mu}_i(t_n) + (\tilde{\sigma}^2_i(t_n))^{1/2} \epsilon_{in}^{(s)}, \quad \epsilon_{in}^{(s)} \sim \mathcal{N}(0,1). \tag{25}$$

After sampling $\epsilon_{in}^{(s)}$, we fix them, and ELL_i is estimated as:

$$\sum_{n=1}^{N} \frac{[\alpha e^{t_n C}]_i}{S} \sum_{s=1}^{S} \log \Big(\tilde{\mu}_i(t_n) + (\tilde{\sigma}^2_i(t_n))^{1/2} \epsilon_{in}^{(s)} \Big)^2. \tag{26}$$

As it separates randomness ($\epsilon_{in}^{(s)}$) from the parameters, the gradient of (26) can be computed straightforwardly while yielding an unbiased estimate of the gradient of the original (21). See Appendix C.3 for the full derivations. Furthermore, to reduce the variance of the estimate, one can use the Rao-Blackwellization technique (Casella & Robert, 1996).

For the integration in (22), we do this numerically by uniform grid sampling. Specifically, by having G uniform grid points $\{\tilde{t}_g\}_{g=1}^G$ over $[0, T]$ with $\Delta t = \tilde{t}_{g+1} - \tilde{t}_g$, we define the following statistics:

$$\Psi_0^i = \sum_{g=1}^G w_g^i K_{\mathcal{Z}^i,\tilde{t}_g}^i K_{\tilde{t}_g,\mathcal{Z}^i}^i, \quad \Psi_1^i = \sum_{g=1}^G w_g^i m^i(\tilde{t}_g) K_{\tilde{t}_g,\mathcal{Z}^i}^i,$$
$$\Psi_2^i = \sum_{g=1}^G w_g^i (m^i(\tilde{t}_g))^2, \quad \Psi_3^i = \sum_{g=1}^G w_g^i K_{\tilde{t}_g,\tilde{t}_g}^i, \quad (27)$$

where $w_g^i = [\alpha e^{\tilde{t}_g C}]_i \Delta t$. We then numerically compute ENO_i as:

$$(\mu^i - m_{\mathcal{Z}^i}^i)^\top (K_{\mathcal{Z}^i,\mathcal{Z}^i}^i)^{-1} \Psi_0^i (K_{\mathcal{Z}^i,\mathcal{Z}^i}^i)^{-1} (\mu^i - m_{\mathcal{Z}^i}^i) +$$
$$+ \ 2\Psi_1^i (K_{\mathcal{Z}^i,\mathcal{Z}^i}^i)^{-1} (\mu^i - m_{\mathcal{Z}^i}^i) - \text{Tr}\big((K_{\mathcal{Z}^i,\mathcal{Z}^i}^i)^{-1} \Psi_0^i\big)$$
$$+ \ \text{Tr}\big((K_{\mathcal{Z}^i,\mathcal{Z}^i}^i)^{-1} \Sigma^i (K_{\mathcal{Z}^i,\mathcal{Z}^i}^i)^{-1} \Psi_0^i\big) + \Psi_2^i + \Psi_3^i. \quad (28)$$

In this way the gradient of ENO_i can be derived fairly easily, which is summarized in Appendix C.3.

When the optimization is done by first-order gradient methods, the computational complexity of the variational inference for our model is no more than r (the number of latent GP functions) times that of the variational inference of the GPCox as in (Lloyd et al., 2015), which can be seen as a special MMGCP model with $r = 1$.

4.4. Model Selection and Test Prediction

We discuss how to determine the optimal value of r. The ELBO objective, the lower bound of the data log-likelihood, tends to increase as we increase r since models with higher r naturally subsume those with lower. However, it would incur higher chance of overfitting and worse generalization on unseen test data. We need to trade off between the model complexity and the goodness of data fitting, and along this line one can employ certain information criteria such as the Bayesian criterion (Schwarz, 1978). When specifying the model complexity, we take into consideration all related parameters as well as the inducing points. Alternatively, we can choose r by cross validation, measuring performance on a validation set, randomly held-out portion of the training data. Once the model and the variational parameters are learned, we can estimate the predictive likelihood for an unseen test data $\mathcal{D}_*$. We see that the predictive log-likelihood, $\log P(\mathcal{D}_*|\mathcal{D}, \Theta, \Omega)$ is lower-bounded by $\mathbb{E}_{q(X,\mathbf{f})}\big[\log P(\mathcal{D}_*|X, \mathbf{f}, \Theta, \Omega)\big]$, which can be computed by the exactly same procedures as in Sec. 4.3 with $\mathcal{D}_*$.

5. Evaluations

In this section we demonstrate the performance of the proposed MMGCP model. We mainly compare our model with two existing extreme stationarity models, MMPP and GPCox, since our model is motivated from both. For the GPCox, among several inference strategies, we opt for the latest variational inference method proposed in (Lloyd et al., 2015), which exhibits comparable generalization performance to other approaches while being significantly faster than sampling-based methods such as (Adams et al., 2009). As a baseline, we also consider: i) the classical kernel smoothing (KS) approach (Diggle, 1985), specifically the Gaussian kernel density estimator, and ii) the log Gaussian Cox process (LGCP) (Rathbun & Cressie, 1994; Møller et al., 1998), which approximates the problem as a standard GP inference with Poisson-likelihood iid data via event counting through discretization of the time horizon.

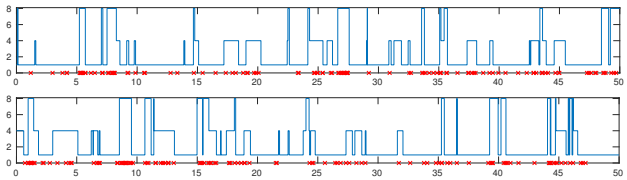

Figure 1. Two event sequences from the synthetic `Full-Stn` data. The X-axis is time. The (blue) curves depict the realized intensity functions from the true MMPP model. The generated events are marked as (red) crosses on the X-axis.

5.1. Synthetic data

To demonstrate the effectiveness of the proposed MMGCP model, we devise three different synthetic data setups that exhibit highly different aspects in terms of time stationarity.

The first setup simulates a fully stationary scenario (denoted by `Full-Stn`), where we generate data from a $r = 3$-state MMPP model with highly different intensity levels $\{1.0, 4.0, 8.0\}$. Within the time horizon $T = 50$, we generate 10 event sequences from the model (Fig. 1 for two exemplar sequences), from which we randomly take 5 sequences as training data while the rest as a test set. For the MMPP and our MMGCP models, we choose the model order by cross validation, which both correctly recovered $r = 3$ hidden states. For the GPCox and our model, we use the squared exponential kernels, and the variational inference in both models uses the same $M = 10$ pseudo inputs (also the same across $i = 1, \ldots, r$ for the MMGCP).

The average test log-likelihoods are shown in Table 1(A). As expected, the MMPP model attains the best performance since the model structure exactly matches that of the data generating one. Our MMGCP, although a generalization of MMPP, performs worse than the MMPP due to the use of smooth kernels. However, the MMGCP significantly outperforms the GPCox and other non-stationary models. The figures in the parentheses indicate the p-values from the paired sample t-test for the competing models against our MMGCP. Thus the differences between existing models and ours are all statistically significant (p-values less than 0.05).

Table 1. Average test log-likelihoods for the three synthetic data setups. The boldfaced figures indicate the best performing ones with statistical significance. From the paired sample t-test, the p-values of the competing models against our MMGCP are shown in the parentheses.

(A) FULL-STN

MMGCP	GPCox	MMPP	KS	LGCP
2.50	1.14	**15.45**	−27.12	−36.61
(–)	(0.029)	(0.014)	(0.0004)	(0.0017)

(B) NON-STN

MMGCP	GPCox	MMPP	KS	LGCP
−42.19	**−42.36**	−45.00	−45.54	**−44.16**
(–)	(0.89)	($< 10^{-4}$)	(0.0004)	(0.29)

(C) SEMI-STN

MMGCP	GPCox	MMPP	KS	LGCP
−87.66	−97.64	−90.66	−102.18	−160.88
(–)	(0.0019)	(0.001)	(0.0003)	(0.0001)

The second synthetic dataset represents fully non-stationary intensity setup (denoted by `Non-Stn`). From (Adams et al., 2009) we take $\lambda(t) = 2\exp(-t/15) + \exp(-((t-25)/10)^2)$ as the true (deterministic) intensity function over $[0, 50]$, and generate data from the inhomogeneous Poisson process. See Fig. 2 for the true intensity function and a sample event sequence. With the similar experimental setups as the first dataset, we run the five models and report the test scores in Table 1(B). The MMPP, with $r = 3$ hidden states chosen, now underperforms the non-stationary time-dependent intensity modeling methods with statistical significance. The GPCox and our MMGCP perform comparably well. The MMGCP selects $r = 2$ hidden states by the cross validation although $r = 1$ (i.e., GPCox) yields a slightly smaller but very close validation score than that of $r = 2$. This implies that the smooth intensity change is properly represented by the covariance functions of the Gaussian processes.

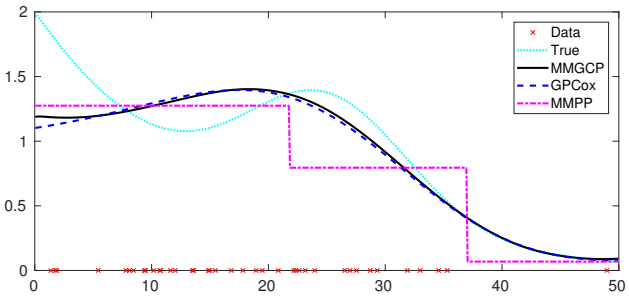

Figure 2. The synthetic `Non-Stn` data. The true intensity function, a sample event sequence, and the estimated (expected) intensity functions of the competing models are shown.

In this `Non-Stn` dataset, since we have the true intensity function available, we can measure the distance between the estimated (expected) intensity functions and the true one. We use the L2 error defined as $\int_0^T (\lambda^{\text{true}}(t) - \overline{\lambda}(t))^2\, dt$, where $\overline{\lambda}(t) = \mathbb{E}[\lambda(t)|\mathcal{D}]$ is the posterior-expected intensity function. In our MMGCP model, as we have the posterior approximation $q(X, \mathbf{f})$, using $\overline{\lambda}(t) \approx \mathbb{E}_q[(f^{X(t)})^2]$ we have:

$$\overline{\lambda}(t) \approx \sum_{i=1}^{r} [\alpha e^{tC}]_i \big((\tilde{\mu}_i(t))^2 + \tilde{\sigma}_i^2(t)\big). \qquad (29)$$

Table 2. Average L2 errors for the `Non-Stn` setup.

MMGCP	GPCox	MMPP	KS	LGCP
2.20	2.59	19.44	7.22	8.85

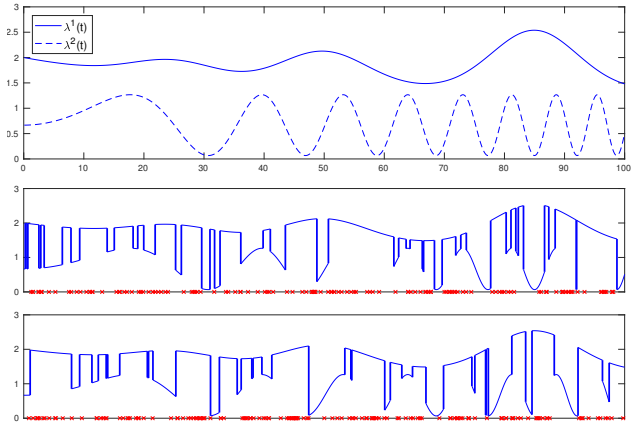

Figure 3. Two event sequences from the synthetic `Semi-Stn` data. The X-axis is time. The top panel depicts two candidate intensity functions $\lambda^1(t)$ and $\lambda^2(t)$. The other two panels show two event sequences generated from the model: the curve indicates the realized intensity function by selecting each of two candidates according to the underlying Markov process, while the generated events are marked as (red) crosses on the X-axis.

The L2 errors of the competing methods are reported in Table 2. See also Fig. 2 for the estimated intensity functions. Our MMGCP and the GPCox exhibit the best performance.

For the last synthetic setup, we aim to simulate the semi-stationary scenario (denoted by `Semi-Stn`). We consider two underlying candidate intensity functions as follows:

$$\lambda^1(t) = \frac{2e^{-t/30} + G_{25}(t) + 2G_{50}(t) + 3.5G_{85}(t) + 4}{3},$$
$$\lambda^2(t) = \frac{1}{3}(1.8\sin(0.005t^2) + 2), \qquad (30)$$

where $G_a(t) = e^{-((t-a)/10)^2}$. As shown in Fig. 3, they exhibit highly different patterns and levels from each other. We also incorporate a 2-state CTMC so that which of the two candidate functions is active at each moment is determined stochastically by the latent Markov process. We generate event sequences over the horizon $T = 100$ from the model.

We follow the experimental setup similar to the previous two datasets. The test log-likelihood scores of the compet-

Table 3. Average test log-likelihoods on the `Football` and `Earthquakes` datasets. For both sets, the boldfaced figures indicate the best performing ones with statistical significance (i.e., the p-values, from the paired sample t-test of the competing models against our MMGCP, all less than 10^{-4}).

(A) FOOTBALL

MMGCP	GPCox	MMPP	KS	LGCP
$\mathbf{-69.48}$	-72.26	-70.34	-71.26	-76.56

(B) ITALY'S EARTHQUAKES

MMGCP	GPCox	MMPP	KS	LGCP
$\mathbf{-101.19}$	-109.73	-117.17	-186.76	-130.78

ing approaches are summarized in Table 1(C). In this case, our MMGCP is outstanding for this semi-stationary data. The superiority of the MMGCP to competing methods is statistically significant whereas the fully stationary MMPP and the time-dependent inhomogeneous models like GPCox and kernel smoother suffer from the heterogeneity of data: globally undergoing stationary regime switching but being time-dependent within each regime.

Overall, our MMGCP is viable consistently across all different time stationarity setups, ranging from fully stationary to non-stationary as well as semi-stationary in between. In the following sections, we also demonstrate the effectiveness of our model on some real-world event datasets.

5.2. Football Data

We test on the football events dataset[6] from the Kaggle open data platform. There are 9074 football games as a whole collected from major European leagues for 5 years (from 2011/12 season to 2016/17). For each game, the major events (e.g., shot attempts, goals, corners, fouls, etc.) are marked in the minute scale. The types and times of the events are obtained from various sources, mainly text commentary and web scraping.

From the dataset, we consider the events of *shot attempts* only, and focus on those games which contain 30 or more events, which comprise about 2000 games. Each game is represented as a sequence of events, and we regard each sequence as an iid sample from an unknown process within the horizon $[0, T]$ with $T = 90+\alpha$ minutes where α amounts to the random extra time which is usually less than 5 (minutes). The average number of events per sequence is 33.4 with standard deviation 3.4.

Among these sequences, we randomly select 500 sequences for training and 100 as a test set. The test likelihood scores are summarized in Table 3(A). It shows that the proposed MMGCP outperforms the competing models with statistical significance (the p-values with regard to our MMGCP are all less than 10^{-4}). The improvement achieved by the proposed approach can be attributed to the semi-stationary nature of the data in some sense: the event rates can be time dependent in certain regimes (e.g., there are often more active attack attempts during the beginning/end of the game or the half time than in the middle of the game), but overall intensities tend to be stationary, exhibiting highly different aspects from game to game.

[6] https://www.kaggle.com/secareanualin/football-events

5.3. Italy's Earthquakes Data

We next demonstrate the performance of the proposed approach on the daily earthquake data publicly available from the Kaggle open data platform. The dataset[7] is obtained by real-time collections of the earthquake events from the Italian Earthquakes National Center, which contains earthquake records of various magnitudes that hit the center of Italy for three months, from August to November in 2016. As we are interested in the daily patterns, we group them on a daily basis, and regard the events for each day as an iid sequence sample. There are 99 (daily) event sequences for which we split them randomly into $60/39$ training/test sets.

We consider all the events with the Richter magnitude no less than 2.0, where the magnitude 2.0 corresponds to earthquakes that are minor, but felt by some people. The number of events per sequence is highly varying across sequences: the mean is 81.7 and the standard deviation 107.5. We also scale the event times from the original data down to $[0, 100]$. The test results are shown in Table 3(B). The MMGCP again exhibits significantly better generalization capability than models based on the extreme time stationarity assumptions. Considering the complexity of the underlying event generating process for this data (e.g., time-sensitive factors as well as stationary changes of states), it signifies that the MMGCP's added flexibility attained by combining inhomogeneous Poisson process with the latent regime switching to account for major trend changes, can be highly effective for representing a complex event process.

6. Conclusion

In this paper we have proposed a novel Markov modulated Gaussian Cox process model that incorporates both the GP-based smooth intensity changes along with major regime switches through a hidden Markov process. While subsuming existing stationary and non-stationary Cox process models as special cases, the proposed model is especially suitable for representing semi-stationary event data. Through empirical evaluations on both synthetic and real-world datasets, we have demonstrated that the model is promising, yielding better generalization for complex event data modeling than existing approaches.

[7] https://www.kaggle.com/blackecho/italy-earthquakes/data

Acknowledgements

This research is supported by National Research Foundation of Korea (NRF-2016R1A1A1A05921948). The author thanks Mark Schmidt for helpful discussions where part of this work was conducted when the author was at UBC.

References

Adams, R. P., Murray, I., and MacKay, D. J. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities, 2009. International Conference on Machine Learning.

Anderson, W. J. *Continuous-Time Markov Chains: An Applications-Oriented Approach*. Springer, 2011.

Asmussen, S., Nerman, O., and Olsson, M. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23(4):419–441, 1996.

Bauer, M., van der Wilk, M., and Rasmussen, C. E. Understanding probabilistic sparse Gaussian process approximations, 2016. In Advances in Neural Information Processing Systems.

Bui, T. D., Yan, J., and Turner, R. E. A unifying framework for Gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research*, 18:1–72, 2017.

Casella, G. and Robert, C. P. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, 1996.

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

Cunningham, J. P., Shenoy, K. V., and Sahani, M. Fast Gaussian process methods for point process intensity estimation, 2008. International Conference on Machine Learning.

Dezfouli, A. and Bonilla, E. V. Scalable inference for Gaussian process models with black-box likelihoods, 2015. In Advances in Neural Information Processing Systems.

Diggle, P. A kernel method for smoothing point process data. *Applied Statistics*, 34:138–147, 1985.

Flaxman, S., Teh, Y. W., and Sejdinovic, D. Poisson intensity estimation with reproducing kernels, 2017. International Conference on Artificial Intelligence and Statistics.

Gunter, T., Lloyd, C., Osborne, M. A., and Roberts, S. J. Efficient Bayesian nonparametric modelling of structured point processes, 2014. Uncertainty in Artificial Intelligence.

Kalbfleisch, J. D. and Lawless, J. F. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80:863–871, 1985.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes, 2014. In Proceedings of the Second International Conference on Learning Representations.

Lasko, T. A. Efficient inference of Gaussian process modulated renewal processes with application to medical event data, 2014. Uncertainty in Artificial Intelligence.

Lloyd, C., Gunter, T., Osborne, M. A., and Roberts, S. J. Variational inference for Gaussian process modulated Poisson processes, 2015. International Conference on Machine Learning.

Matthews, A., Hensman, J., Turner, R. E., and Ghahramani, Z. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes, 2016. International Conference on Artificial Intelligence and Statistics.

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

Quiñonero-Candela, J. and Rasmussen, C. E. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Rathbun, S. L. and Cressie, N. Asymptotic properties of estimators for the parameters of spatial inhomogeneous Poisson point processes. *Advances in Applied Probability*, 26(1):122–154, 1994.

Ryden, T. An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics & Data Analysis*, 21(4):431–447, 1996.

Samo, Y.-L. K. and Roberts, S. Scalable nonparametric Bayesian inference on point processes with Gaussian processes, 2015. International Conference on Machine Learning.

Schwarz, G. E. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

Titsias, M. K. Variational learning of inducing variables in sparse Gaussian processes, 2009. In Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics.