Exact Bayesian inference in spatio-temporal Cox processes driven by multivariate Gaussian processes

Flávio B. Gonçalves[a], Dani Gamerman[b]

[a] Departamento de Estatística, Universidade Federal de Minas Gerais, Brazil
[b] Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Brazil

Abstract

In this paper we present a novel inference methodology to perform Bayesian inference for spatio-temporal Cox processes where the intensity function depends on a multivariate Gaussian process. Dynamic Gaussian processes are introduced to allow for evolution of the intensity function over discrete time. The novelty of the method lies on the fact that no discretisation error is involved despite the non-tractability of the likelihood function and infinite dimensionality of the problem. The method is based on a Markov chain Monte Carlo algorithm that samples from the joint posterior distribution of the parameters and latent variables of the model. A particular choice of the dominating measure to obtain the likelihood function is shown to be crucial to devise a valid MCMC. The models are defined in a general and flexible way but they are amenable to direct sampling from the relevant distributions, due to careful characterisation of its components. The models also allow for the inclusion of regression covariates and/or temporal components to explain the variability of the intensity function. These components may be subject to relevant interaction with space and/or time. Simulated examples illustrate the methodology, followed by concluding remarks.

Key Words: Cox process, spatio-temporal, dynamic Gaussian process, exact inference, MCMC.

[1]Address: Av. Antônio Carlos, 6627 - DEST/ICEx/UFMG - Belo Horizonte, Minas Gerais, 31270-901, Brazil. E-mail: fbgoncalves@est.ufmg.br

1 Introduction

A Cox process is an inhomogeneous Poisson process where the intensity function evolves stochastically. It is also referred to as doubly stochastic process. Cox processes (Cox, 1955) have been extensively used in a variety of areas to model point process phenomena. Effects in Cox processes may present spatio-temporal variation to reflect the possibility of interaction between space-time and other model components. They can be traced back to log-Gaussian Cox processes (Møller et al., 1998), where a Gaussian Process (GP) representation is used for the log-intensity (see also Diggle, 2014, and references within).

The application of (GP driven) Cox processes is closely related to two main problems: simulation and inference. These are hard problems due to the infinite dimensionality of the process and the intractability of the likelihood function. Simulation is one of the main tools to tackle the inference problem which primarily consists of estimating the unknown intensity function (IF) and potential unknown parameters. However, prediction is often a concern, i.e., what should one expect in a future realisation of the same phenomenon.

Solutions for the inference problem have required, until recently, the use of discrete approximations (see, for example, Møller et al., 1998; Brix and Diggle, 2001; Reis et al., 2013). These represent a considerable source of error and, therefore, ought to be used with care. Moreover, quantifying and controlling this error may be hard and expensive. This motivates the development of exact methodologies, i.e. free from discretisation errors. Exact solutions for inference on infinite-dimensional processes with intractable likelihood can be found, for example, in Beskos et al. (2006), Sermaidis et al. (2013) and Gonçalves et al. (2015).

One non-parametric exact approach to the analysis of spatial point patterns was proposed in Adams et al. (2009). They consider a univariate Gaussian process to describe the IF dynamics and an augmented model for the data and latent variables that simplifies the likelihood function. Another non-parametric exact approach was adopted in Kottas and Sansó (2007), where a particular factorisation of the IF was proposed and Dirichlet processes priors were used. Their work was extended to the spatio-temporal context by Xiao et al. (2015).

The aim of this work is to propose an exact inference methodology for spatio-temporal Cox processes in which the intensity function dynamics is driven by a Gaussian process. The exactness stems from an augmented model approach as in Adams et al. (2009). However, we generalise their point pattern models by firstly considering spatio-temporal models and, secondly, by using multivariate (possibly dynamic) Gaussian processes to allow the inclusion of different model components (regression and temporal effects) in a flexible manner. Space and time may be considered continuous or discrete. In this paper we ultimately consider the general formulation of continuous space and discrete time. This is actually the most general formulation, since its continuous time version can be seen as a continuous space process where time is one of the dimensions.

Our methodology also introduces a particularly suited MCMC algorithm that enables direct simulation from the full conditional distributions of the Gaussian process and of other relevant latent variables. A particular choice of dominating measure to obtain the likelihood function is crucial to derive these sampling steps. Moreover, estimation of (possibly intractable) functionals of the IF and prediction based on the output of the

MCMC is straightforward. We also provide formal proofs of the validity of the MCMC algorithm.

This paper is organised as follows. Section 2 presents the class of models to be considered and the augmented model to derive the MCMC. Section 3 presents the general Bayesian approach and addresses some identifiability issues. Section 4 describes the MCMC algorithm for the spatial model and Section 5 presents its extension to the spatio-temporal case. Section 6 presents simulated examples to illustrate the proposed methodology. Final remarks and possible directions for future work are presented in Section 7.

2 Model specification

In this section we present the complete probabilistic model for spatio-temporal point processes with Gaussian process driven intensities. We break the presentation in parts considering the different levels and generalisations of the model.

2.1 The general Cox process model

We consider a Poisson process (PP) $Y = \{Y_t;\ t \in \mathcal{T}\}$ in $S \times \mathcal{T}$, where S is some compact region in $\mathbb{R}^d$ and $\mathcal{T}$ is a finite set of $\mathbb{N}$. It can be seen as a Poisson process in a region S that evolves in time. We assume the Poisson process has an intensity function $\lambda_t(s) : S \times \mathcal{T} \to \mathbb{R}^+$. This implies, for example, that the number of points $N_t(A)$ in $A \subseteq S$ - a compact region in S, follows a Poisson distribution with mean $\int_A \lambda_t(s)ds$, at time t. Moreover, from standard properties of Poisson process, given $\lambda_{S,\mathcal{T}} := \{\lambda_t(s),\ s \in S, t \in \mathcal{T}\}$, for each $t \in \mathcal{T}$, Y_t is a Poisson process in S with intensity function $\lambda_{S,t} := \{\lambda_t(s),\ s \in S\}$ and the Y_t's are mutually independent.

We assume that the IF is a function of a (multivariate spatio-temporal) Gaussian process and covariates. A Gaussian process β is a stochastic process in some space such that the joint distribution of any finite collection of points in this space is Gaussian. This space may be defined so that we have spatial or spatio-temporal processes.

A detailed presentation of (dynamic) Gaussian processes is given in Section 2.3. For now, let $\beta := \{\beta_0, \beta_1, \ldots, \beta_p\}$ be a collection of $p+1$ independent GP's (a multivariate GP) in $S \times \mathcal{T}$, for $\mathcal{T} = \{0, \ldots, T\}$. We assume the following model for Y:

$$
\begin{aligned}
(Y_t|\lambda_{S,t}) &\sim PP(\lambda_{S,t}),\ \ \forall\, t \in \mathcal{T}, && (1)\\
\lambda_t(s) &= \lambda_t^* \Phi(f(\beta_t(s), W_t(s))),\ \ \forall\, t \in \mathcal{T}, && (2)\\
(\beta|\theta) &\sim GP_\theta, && (3)\\
(\lambda_{\mathcal{T}}^*, \theta) &\sim prior, && (4)
\end{aligned}
$$

where $\lambda_{\mathcal{T}}^* = (\lambda_0^*, \ldots, \lambda_T^*)$, Φ is the distribution function of the standard Gaussian distribution, GP_θ is a Gaussian process indexed by (unknown) parameters θ and $W = \{W_t(s)\}$ is a set of covariates. We assume f to be linear in the coordinates of β and write W in a way such that $f(\beta_t(s), W_t(s)) = W_t(s)\beta_t(s)$. Any distribution function of a continuous r.v. or any general bounded function may be used instead of Φ. For example, a common

choice is the logistic function, which is used by Adams et al. (2009) and is very similar to Φ (the largest difference is 0.0095). The particular choice in (2) contributes to the construction of an efficient MCMC algorithm as discussed in Section 4. We are interested in estimating not only the overall rate $\lambda_{S,\mathcal{T}}$ but also the univariate GP's, separately, given their meaningful interpretation.

Parameter λ_t^* ought to represent the supremum of the intensity function at time t. One extreme possibility is to assume $\lambda_t^* = \lambda^*$, which is a reasonable assumption in the case of purely spatial processes or processes whose maximum intensity is time invariant. In this case, a common choice for the prior distribution of λ^* is $\mathcal{G}(\alpha_\lambda, \beta_\lambda)$ - a Gamma distribution. At the other extreme, unrelated parameters vary independently over time according to independent $\mathcal{G}(\alpha_{\lambda_t}, \beta_{\lambda_t})$ prior distributions. In between them, models allow for temporal dependence between (successive) λ_t^*'s. One such formulation with attractive features is described in Section 5.

The Gaussian processes may represent a number of relevant model features such as the effect of covariates and/or spatio-temporal components such as the spatially-varying trend components or seasonality of the baseline intensity. They may also be space and/or time invariant. One common example is the model with p covariates:

$$f(\beta_t(s), W_t(s)) = \beta_{0,t}(s) + \beta_{1,t}(s)W_{1,t}(s) + \ldots + \beta_{q,t}(s)W_{q,t}(s). \tag{5}$$

This approach allows the use of extra prior information through covariates. The spatio-temporal variation of the effects is particularly relevant in applications where covariates present significant interaction with space and/or time. Examples are provided in Pinto Jr et al. (2015).

The first advantage of the formulation in (1)-(4) is that it allows exact simulation of data from the model which is the key to develop exact inference methods. Exact simulation of the model is based on a key result from Poisson processes called Poisson thinning. This is a variant of rejection sampling for point processes proposed by Lewis and Shedler (1979) and is given in Algorithm 1 below:

Algorithm 1

1. Simulate a Poisson process $(s_1, \ldots, s_K)$ with constant intensity function λ_t^* on S:
   (a) Simulate $K_t \sim Poisson(\lambda_t^* \mu(S))$, where $\mu(S)$ is the volume of S;
   (b) Distribute the K_t points uniformly on S.
2. Simulate β_t and observe W_t at points $\{s_1, \ldots, s_K\}$;
3. Keep each of the K_t points with probability $\lambda_t(s_k)/\lambda_t^*$;
4. OUTPUT the points kept at the previous step.

To simulate the process at additional times $t \in \mathcal{T}$, it is enough to perform the algorithm above for each t and simulate the Gaussian process conditional on the points previously simulated. The idea of Poisson thinning is applied in related contexts by Gonçalves and Roberts (2014).

2.2 The augmented model

Performing exact inference for Cox processes is a challenging problem mainly because of the intractability of the likelihood function, which is given by

$$L(\lambda_{S,\mathcal{T}}, y) = \exp\left\{-\sum_{t=0}^{T}\int_S \lambda_t(s)ds\right\}\prod_{t=1}^{T}\prod_{n=1}^{N_t}\lambda_t(s_{t,n}), \tag{6}$$

where $s_{n,t}$ is the location of the n-th event of Y_t.

The crucial step to develop exact methods is to avoid dealing with the likelihood above. One possible solution is to define an augmented model for Y and some additional variable X, such that the joint (pseudo-)likelihood based on (X, Y) is tractable. This poses the problem as a missing data problem and allows us to use standard methods. The augmented model is constructed based on the Poisson thinning presented in Algorithm 1.

Firstly, define $X = \{X_t;\ t \in \mathcal{T}\}$ where each X_t is a homogeneous PP with intensity λ_t^* on S and the X_t's are mutually independent. Now let $\{s_{t,k}\}_{k=1}^{K_t}$ be the locations of the K_t events of X_t. We also define T vectors Z_t, $t = 1, \ldots, T$, with each coordinate taking values in $\{0, 1\}$ such that $(Z_t|X, \beta_{K_t}, W_{K_t})$ is a random vector $(Z_{t,1}, \ldots, Z_{t,K_t})$, where the $Z_{t,k}$'s are all independent with $Z_{t,k} \sim Ber(\Phi(W_t(s_{t,k})\beta_t(s_{t,k})))$ and (β_{K_t}, W_{K_t}) is (β, W) at the points from X_t. Finally, define $Y_t = h(Z_t, X_t)$ as the non-zero coordinates of the vector $(Z_{t,1}s_{t,1}, \ldots, Z_{t,K_t}s_{t,K_t})$, which leads to the model (1).

Namely, the augmented model defines Y as the events remaining from performing the Poisson thinning to a PP X. It is important to note, however, that only Y is observed. We define $\{s_{t,n}\}_{n=1}^{N_t}$ as the N_t events of Y_t and $\{s_{t,m}\}_{m=1}^{K_t-N_t}$ as the $M_t := K_t - N_t$ thinned events. Most importantly, this approach leads to a tractable likelihood when the joint distribution of X and Y is considered, as it is shown in Section 3.1.

The spatial model is a particular case where $T = 1$ which implies that X and Y are Poisson processes on S with intensity functions λ^* and $\lambda(s)$, respectively. We observe $\{s_n\}_{n=1}^{N}$ from Y and simplify the notation above accordingly. Note that the space model for unidimensional S is generally seen as the commonly used Cox process in time.

2.3 Dynamic Gaussian processes

Gaussian processes are a very flexible component to handle spatial variation, specially when smooth processes are expected. We say that β follows a stationary Gaussian process in S if $\beta(s) \sim N(\mu, \sigma^2)$ and $Cov(\beta(s), \beta(s')) = h(s, s')$, for $s, s' \in S$, constants μ and σ^2 and a (almost everywhere) differentiable function h. Further simplification is obtained if isotropy can be assumed, leading to $h(s, s') = \rho(|s - s'|)$. In this case, the process is denoted by $\beta \sim GP(\mu, \sigma^2, \rho)$ and ρ is referred to as the correlation function.

Typical choices for h belong to the γ-exponential family of covariance functions:

$$h(s, s') = \sigma^2 \exp\left\{-1/(2\tau^2)|s - s'|^\gamma\right\}, \quad 0 < \gamma \leq 2. \tag{7}$$

The special case when $\gamma = 2$ leads to almost-surely differentiable paths (surfaces).

GP's can be extended in many directions. The most important ones here are extensions to handle multivariate GP's and extensions to cope with space and time. There are a

number of different ways to allow for multivariate responses. The main ones are reviewed in Gamerman et al. (2007) and include independent GP's, dependent processes with a common correlation function, or linear mixtures of independent GP's.

Extensions to cope with space and time were introduced by Gelfand et al. (2005). A process β follows a dynamic Gaussian process in discrete time if it can be described by a difference equation

$$\beta_{t'}(\cdot) = G_{t',t}\beta_t(\cdot) + w_{t',t}(\cdot) \quad , \quad w_{t',t} \sim GP, \tag{8}$$

where the multivariate Gaussian process disturbances $w_{t',t}(\cdot)$ are zero mean and time-independent; they are also taken as identically distributed in the equidistant case $t' = t+1$. The law of the process is completed with a Gaussian process specification for $\beta_0(\cdot)$. Similar processes were proposed in continuous-time by Brix and Diggle (2001).

A number of options are available for the temporal transition matrix G, including the identity matrix. If additionally the disturbance processes w consist of independent GP's then the resulting process consist of independent univariate dynamic GP's.

We present some alternatives to model trend and seasonality of the IF. Typically, one ought to consider a baseline process β_0 that evolves according to (8), that is

$$\beta_{0,t+1}(s) = \alpha_{t+1,t}\beta_{0,t}(s) + w_{t+1,t}(s). \tag{9}$$

The simplest temporal model one can think of for the IF would then be:

$$f(\beta_t(s), W_t(s)) = \beta_{0,t}(s), \tag{10}$$

with $\alpha_{t+1,t} = 1, \ \forall\, t$. In the case with q covariates, one may consider other q independent DGP's as in (5).

The α parameters can be used to model trend, for example when the IF is subject to a (local) growth process. Seasonality may be modelled by considering a multivariate process (β_0, β_1) such that

$$(\beta_{0,t+1}(s), \beta_{1,t+1}(s))' = G_{t+1,t}(\beta_{0,t}(s), \beta_{1,t}(s))' + (w_{0;t+1,t}(s), w_{1;t+1,t}(s))', \tag{11}$$

and

$$f(\beta_t(s), W_t(s)) = \beta_{0,t}(s) + \beta_{1,t}(s)\cos\left(2\pi t/p + \phi\right), \tag{12}$$

where p is the period and ϕ is the harmonic phase angle. For example, for quarterly data with annual cycles we have $p = 4$. A simple but useful choice would be $G = I$ and $w_{1;t+1,t} = 0, \ \forall\, t$. These modelling ideas are illustrated in Section 6.

2.4 Using non-spatiotemporal covariates

The use of non-spatio-temporal covariates to explain the intensity function variation is appealing when such information is available. For example, individual covariates may carry important information about the spatial distribution of the mortality due to some disease (see, for example, Pinto Jr et al., 2015). This approach, however, requires some adaptations in the original model presented in Section 2.1. Firstly, we define $\{\nu_1, \ldots, \nu_{q_t}\}$ as the set of all the configurations of the covariates appearing in the data at time t. Now,

for each t, the PP Y_t is decomposed into q_t independent PP's $(Y_{t,1} \ldots, Y_{t,q_t})$ such that each of them has a intensity function $\lambda_t(s,\nu) = \lambda_t^* \Phi(f(\beta_t(s),\nu))$. This allows the use of the inference methodology proposed in this paper and the prediction for non-observed configurations.

3 Inference for the spatial model

We now focus on the inference problem of estimating the intensity function $\lambda_{S,\mathcal{T}}$, parameter λ^* and potential unknown parameters θ from the (multivariate) Gaussian process, based on observations from the Poisson process Y. We shall also discuss how to make prediction. In order to make the presentation of the methodology as clear as possible, we consider first the (purely) spatial process and then the generalisation for the spatio-temporal case. The extension for non-spatio-temporal covariates is omitted but it is straightforwardly devised from the original methodology.

3.1 Posterior distribution

Let us first establish some notation. We have that $\{s_k\}_{k=1}^K$, $\{s_n\}_{n=1}^N$ are the points from X and Y, respectively, and $\{s_m\}_{m=1}^{K-N}$ are the thinned events. Naturally, $\{s_k\}_{k=1}^K = \{s_n\}_{n=1}^N \bigcup \{s_m\}_{m=1}^{K-N}$. Furthermore, β_K, β_N, and β_M are β at $\{s_k\}_{k=1}^K$, $\{s_n\}_{n=1}^N$ and $\{s_m\}_{m=1}^{K-N}$, respectively. Naturally $\beta_K = (\beta_N, \beta_M)$. Analogously, W_K, W_N and W_M are the respective subsets of W. Keeping in mind the existent redundancies, the vector of all random components of the model is given by $\big(\{s_n\}_{n=1}^N\ ,\ \{s_m\}_{m=1}^{K-N}\ ,\ \beta\ ,\{s_k\}_{k=1}^K\ ,\ K\ ,\ \lambda^*\ ,\ \theta,\ W\big)$.

Initially, assume W to be deterministic and consider only β_K instead of β. Thus, define $\psi := \big(\{s_n\}_{n=1}^N\ ,\ \{s_m\}_{m=1}^{K-N}\ ,\ \beta_K\ ,\{s_k\}_{k=1}^K\ ,\ K\ ,\ \lambda^*\ ,\ \theta\big)$. This strategy simplifies the problem (making it finite-dimensional) but still makes it possible to estimate the infinite-dimensional remainder of β - to be discussed further ahead. Note that, due to redundance issues, the components of the model could be specified in other ways.

We now specify the joint distribution of all the random components of the model - $(\psi|W,S)$. Note that the joint posterior density we aim is proportional to this. It is important to make an appropriate choice of a dominating measure w.r.t. which we write the density of ψ. This choice depends on the specification of the components and is not unique. For example, one may choose to write the density of $(K, \{s_k\}_{k=1}^K|\lambda^*)$ w.r.t. the measure of a unit rate Poisson process, resulting in

$$\exp\left(-(\lambda^*-1)\mu(S)\right)(\lambda^*)^K \propto \exp\left(-\lambda^*\mu(S)\right)(\lambda^*)^K, \tag{13}$$

which is the usual Poisson process likelihood function. However, note that this is a valid likelihood for λ^* only, since the dominating measure is independent of this parameter. It is not a valid likelihood function for K, which, in our case, is unknown and must be estimated. This gives good intuition to why this is not a good choice for the dominating measure. Although it is one possibility, it makes the derivation of the full conditional distributions (or the acceptance probability of potential MH steps) more difficult.

We choose to write the density of $(\psi|W,S)$ w.r.t. the dominating measure given by the product measure $\mathbb{Q} := \delta^K \otimes \mathbb{L}^K \otimes \mathbb{L}^K \otimes \delta \otimes \mathbb{L} \otimes \mathbb{L}^{d_\theta}$, where δ is the counting measure, $\mathbb{L}^d$

is the d-dimensional Lebesgue measure and d_θ is the dimension of the parameter vector θ. This choice is related to the factorisation we choose. If we let $\mathbb{P}$ be the probability measure of our full model, the density π of $(\psi|W,S)$, defined as the Radon-Nikodym derivative of $\mathbb{P}$ w.r.t. $\mathbb{Q}$, is given by

$$\begin{aligned}
& \pi(\psi|W,S) = \pi(\{s_n\},\{s_m\}|\{s_k\},\beta_K,W)\pi(\beta_K|\{s_k\},\theta)\pi(\{s_k\}|K,S)\pi(K|\lambda^*,S)\pi(\lambda^*)\pi(\theta) \\
= & \left[\prod_{n=1}^{N}\Phi(W(s_n)\beta(s_n))\prod_{m=1}^{K-N}\Phi(-W(s_m)\beta(s_m))\right]\pi_{GP}(\beta_K|\theta) \\
& [\mu(S)]^{-K}\left[e^{-\lambda^*\mu(S)}\left(\lambda^*\mu(S)\right)^K\frac{1}{K!}\right]\pi(\lambda^*)\pi(\theta) \\
= & \left[\Phi_N(W_N\beta_N;I_N)\Phi_{K-N}(-W_M\beta_M;I_{K-N})\right]\pi_{GP}(\beta_K|\theta)\left[e^{-\lambda^*\mu(S)}\left(\lambda^*\right)^K\frac{1}{K!}\right]\pi(\lambda^*)\pi(\theta)
\end{aligned} \tag{14}$$

where $\{s_n\} = \{s_n\}_{n=1}^N$, $\{s_m\} = \{s_m\}_{m=1}^{K-N}$, $\{s_k\} = \{s_k\}_{k=1}^K$. $\Phi_k(\cdot;I_k)$ is the distribution function of the k-dimensional Gaussian distribution with mean vector zero and covariance matrix I_k (k-dimensional identity matrix) and $\beta_N = (\beta_0(s_1)\dots\beta_0(s_N)\dots\beta_q(s_1)\dots\beta_q(s_N))'$. Also, $W_N = (I_N\ W_1\ \dots W_q)$, where W_i is a $N\times N$ diagonal matrix with the (n,n)-entry being $W_i(s_n)$ - the i-th covariate at location s_n. Furthermore, $\pi_{GP}(\beta_K|\theta)$ is the density of the multivariate Gaussian process at locations $\{s_k\}_{k=1}^K$ w.r.t. $\mathbb{L}^K$. Finally, $\pi(\lambda^*)$ and $\pi(\theta)$ are the prior densities of λ^* and θ, respectively, w.r.t. $\mathbb{L}$ and $\mathbb{L}^{d_\theta}$.

3.2 Estimation of the intensity function

The MCMC algorithm to be proposed in Section 4.2 outputs samples from the posterior distribution of the intensity function λ_S at the observed locations $\{s_n\}_{n=1}^N$ and at another finite collection of locations $\{s_m\}_{m=1}^{K-N}$ which varies among the iterations of the algorithm. Nevertheless, we need to have posterior estimates of λ_S over the whole space S. It is quite simple to sample from this posterior. This may be equivalently done by adding an extra step to the MCMC algorithm or by a sampling procedure after the MCMC runs. Both schemes may suffer from high computational cost but the former is considerably cheaper if well-designed.

Firstly note that $\lambda_S = \{\lambda(s),\ s\in S\}$ and $\lambda(s) = \lambda^*\Phi(W(s)\beta(s))$, which means that to sample from the posterior of λ_S at a fixed location s we need to sample from the posterior of $(\lambda^*,\beta(s))$. In order to have a practical algorithm we choose a fine (squared) grid determined by locations $S_0 = \{\tilde{s}_1,\dots,\tilde{s}_G\}$ and use the discrete field $\lambda_{S_0} = \{\lambda(s),\ s\in S_0\}$ to access the posterior of the intensity function. We need the following lemma.

Lemma 1. *Given (β_K,θ), β_{S_0} is independent of $\left(\{s_n\}_{n=1}^N, W\right)$ and its posterior distribution is given by*

$$\pi(\beta_{S_0}|\{s_n\}_{n=1}^N,W,S) = \int \pi(\beta_{S_0}|\beta_K,\theta)\pi(\beta_K,\theta|\{s_n\}_{n=1}^N,W,S)d\beta_K d\theta \tag{15}$$

Proof. *See Appendix A.*

This implies that a sample from the posterior of λ_{S_0} can be obtained by sampling from a multivariate Gaussian as it will be discussed in details in Section 4.3.

3.3 Model identifiability and practical implementation

As we have mentioned before, our model may suffer from identifiability problems concerning parameter λ^*. The natural way to identify it is to have this parameter as the supremum of the intensity function which, under the Bayesian approach, should be achieved by an appropriate specification of the prior distribution. This means that, under all possible most likely configurations of (λ^*, β), we are looking for the one that minimises λ^*.

A reasonable choice for the prior of λ^* is an Exponential distribution for which the hyperparameter could be specified through an empirical analysis of the data set. More specifically, by obtaining an empirical estimate of the intensity in a small area with the highest concentration of points. This area should not be too small nor too large. Note that the data is being used only to identify the model, which is different from using the data twice in a model which is already identified. Moreover, note that having λ^* as the supremum also optimises the computational cost as it minimises M - the smaller is M the smaller the dimension of the covariance matrix that needs to be inverted to simulate β.

Generally speaking, identifiability is an important issue when estimating the intensity function of a non-homogeneous PP. It is well known that the reliability of the estimates rely on the amount of data available. In this sense, the higher is the actual function the better. In a Bayesian framework, in particular, the prior on the intensity function plays an important role on the identification and estimation of this function. This is related to the fact that the data does not contain much information about the hyperparameters of GP priors. The simulated examples from Section 6 explore this issue and provides some insight on how to proceed in general.

Another important issue is the computational cost from dealing with GP's. Despite their great flexibility on a variety of statistical modelling problems, Gaussian processes have a considerable practical limitation when it comes to computational cost. More specifically, simulating a n-dimensional GP has a cost which is typically on the order of n^3. This means that in our case the cost would be of order $(q+1)K^3$, without involving the procedures in Sections 3.2. Given nowadays computational resources, this cost is reasonable for quite large values of n. For the cases where the cost is too high, several approximating solutions have been proposed in the literature (see Banerjee et al., 2013; Carlin et al., 2007).

There are other model-based solutions to reduce the computational cost without the need of approximating solutions. Instead of considering one parameter λ^* for the whole space S, one may consider a partition $\{S_1, \ldots, S_L\}$ of S and assign one λ_l^* to each sub-space S_l. This means that the general model is now $\lambda(s) = \lambda^*_{l(s)}\sigma(s)$, where $l(s)$ corresponds to the sub-space that contains s. Inference is carried out analogously to the case where no partition of the space is considered. The advantage of this approach is that it will generate less thinned events s_m and, consequently, reduce the computational cost. Naturally, the choice of the partition determines the cost reduction. One general rule is that the rate $\lambda(s)$ should be as close as possible to being homogeneous in each sub-space. Furthermore, L should not be very large and, equivalently, each sub-space should have a sizeable number of points to ensure informative estimates of the λ_l^* parameters.

4 Computation for the spatial model

In this Section, we present the computational details to perform inference in the spatial model. The proposed methodology consists of a MCMC algorithm which has the exact joint posterior distribution of the unknown components of the model as its invariant distribution. More specifically, the algorithm is a Gibbs sampler. The derivation of the full conditional distributions is not straightforward due to several reasons: intractability issues; the redundancy among some of the components; the hierarchical structure of the model, specially the fact that the observations are not (explicitly) on the first level, due to thinning. In order to sample efficiently from the full conditional distributions, it is essential to be able to simulate from a general class of multivariate skew-normal distributions. We define such class and propose an efficient algorithm to sample from it.

4.1 A general class of multivariate skew-normal distributions

We consider a general class of skew-normal distributions originally proposed in Arellano-Valle and Azzalini (2006) and present it here in a particularly useful way for the context of our work. Equally important, we also propose an algorithm to sample from this distribution.

For a d-dimensional vector ξ, a $m \times d$ matrix W and a $d \times d$ matrix Σ, we define

$$U = \begin{pmatrix} U_0 \\ U_1 \end{pmatrix} \sim N_{m+d}(0, \Sigma^*) \text{ and } \Sigma^* = \begin{pmatrix} \Gamma & \Delta' \\ \Delta & \Sigma \end{pmatrix}, \tag{16}$$

where $\Gamma = I_m + W\Sigma W'$ and $\Delta' = W\Sigma$. Let $a = (a_1, \ldots, a_r) > b = (b_1, \ldots, b_r)$ mean that $a_i > b_i$, $\forall i$ and define $\gamma = W\xi$. We say that $(U_1 + \xi | U_0 > -\gamma)$ has a $SN(\xi, \Sigma, W)$ distribution whose density is given in the following proposition.

Proposition 1. *The density of* $(U_1 + \xi | U_0 > -\gamma)$ *is given by*

$$f(z) = \frac{1}{\Phi_m(\gamma; \Gamma)} \phi_d(z - \xi; \Sigma) \Phi_m(Wz; I_m), \tag{17}$$

where $\phi_d(\cdot; \Omega)$ *and* $\Phi_d(\cdot; \Omega)$ *are the density and distribution function, respectively, of the* d*-dimensional Gaussian distribution with mean vector zero and covariance matrix* Ω*.*

Proof. *See Appendix A.*

We propose the following algorithm to sample from the density in (17). Define $U_0^* = A^{-1}U_0$, where A is the lower diagonal matrix obtained from the Cholesky decomposition of Γ, i.e. $\Gamma = AA'$. This implies that $U_0^* \sim N_m(0, I_m)$ and $U_0 = AU_0^*$. We use the following results to construct our algorithm.

$$f(U_1, U_0 | U_0 > -\gamma) = f(U_1 | U_0, U_0 > -\gamma) f(U_0 | U_0 > -\gamma). \tag{18}$$

Proposition 2. $(AU_0^* | U_0^* \in B)$ *has the same distribution as* $(U_0 | U_0 > -\gamma)$*, where* $B = \{u_0^* : Au_0^* > -\gamma\}$*.*

Proof. *See Appendix A.*

The decomposition in (18) suggests that simulation from (1) may be performed by firstly simulating $(U_0|U_0 > -\gamma)$ and then using this value to simulate from $(U_1|U_0)$. Moreover, the simulation of U_0 is more efficient (as described below) if we first simulate U_0^* and then apply the appropriate transformation, as suggested by Proposition 2. The algorithm to simulate from (1) is the following.

Algorithm 4.1

1. Simulate a value u^* from $(U_0^*|U_0^* \in B)$;
2. Obtain $u = Au^*$;
3. Simulate a value z^* from $(U_1|U_0 = u) \sim \mathcal{N}(\Delta\Gamma^{-1}u, \Sigma - \Delta\Gamma^{-1}\Delta')$;
4. Obtain $z = z^* + \xi$;
5. OUTPUT z;

The simulation of step 3 is trivial. Step 1 consists of the simulation of a truncated (by linear constraints) multivariate Normal and cannot be performed directly. The simulation from this distribution is described in Appendix B.

4.2 The Gibbs sampling algorithm

Notice that, given the data $\{s_k\}$, the remaining unknown quantities are $(\{s_m\}, \beta_K, K, \lambda^*, \theta)$. We block these quantities as:

$$(\{s_m\}, \beta_M, K) \quad , \quad \beta_K \quad , \quad \theta \quad , \quad \lambda^*.$$

Note that β_M is sampled twice. That is mainly because updating B_K instead of only β_N significantly improves the mixing of the chain (by reducing the correlation among blocks) and because sampling the first block without β_M is virtually impossible. All full conditional densities are proportional to $\pi(\psi|W, S)$. Thus, elimination of constant terms leads to

$$\pi(\{s_m\}, \beta_M, K|\cdot) \propto \Phi_{K-N}(-W_M\beta_M; I_{K-N})\pi_{GP}(\beta_M|\beta_N, \theta)\left[(\lambda^*)^K \frac{1}{K!}\right]\mathbb{1}(K \geq N), \quad (19)$$

$$\pi(\beta_K|\cdot) \propto \Phi_N(W_N\beta_N; I_N)\Phi_{K-N}(-W_M\beta_M; I_{K-N})\pi_{GP}(\beta_K|\theta), \quad (20)$$

$$\pi(\lambda^*|\cdot) \propto \left[e^{-\lambda^*\mu(S)}\,(\lambda^*)^K\right]\pi(\lambda^*), \quad (21)$$

$$\pi(\theta|\cdot) \propto \pi_{GP}(\beta_K|\theta)\pi(\theta). \quad (22)$$

The four densities above are written with respect to the dominating measures: $\mathbb{L}^{K-N} \otimes \mathbb{L}^{K-N} \otimes \delta$, $\mathbb{L}^K$, $\mathbb{L}$ and $\mathbb{L}^{d_\theta}$, respectively, in accordance with the dominating measure used to write (14).

Define π_0 as a $Poisson(\lambda^*\mu(S))$ distribution truncated to $\{N, N+1, \ldots\}$. We propose the following rejection sampling (RS) algorithm to sample from (19):

Algorithm 4.2

1. Simulate $\dot{K}$ from π_0;
2. IF $\dot{K} = N$, make $\{\dot{s}_m\}_{m=1}^{\dot{K}-N} = \dot{\beta}_M = \emptyset$ and GOTO 8, ELSE GOTO 3;
3. Make $m = 1$ and $\dot{\beta}_{1:m-1} = \emptyset$;
4. Make $r_m = 1$;
5. Simulate $\dot{s}_{r_m} \sim \mathcal{U}(S)$ and $\dot{\beta}_{r_m}(\dot{s}_{r_m})$ from $\pi_{GP}(\dot{\beta}_{r_m}(\dot{s}_{r_m})|\beta_N, \dot{\beta}_{1:m-1}, \theta)$;
6. Simulate $Z_{r_m} \sim Ber(\Phi(-W(\dot{s}_{r_m})\beta(\dot{s}_{r_m})))$;
7.
   - IF $Z_{r_m} = 1$ and $m < K - N$, set $\dot{s}_m = \dot{s}_{r_m}$, $\dot{\beta}(\dot{s}_m) = \dot{\beta}_{r_m}(\dot{s}_{r_m})$, $\dot{\beta}_{1:m-1} = \dot{\beta}_{1:m-1} \bigcup \dot{\beta}_{r_m}(\dot{s}_{r_m})$, $m = m + 1$ and GOTO 4;
   - IF $Z_{r_m} = 1$ and $m = K - N$, set $\dot{s}_m = \dot{s}_{r_m}$, $\dot{\beta}(\dot{s}_m) = \dot{\beta}_{r_m}(\dot{s}_{r_m})$ and GOTO 8;
   - IF $Z_{r_m} = 0$, set $r_m = r_m + 1$ and GOTO 5;
8. OUTPUT $\left(\dot{K}, \{\dot{s}_m\}_{m=1}^{\dot{K}-N}, \dot{\beta}_M\right)$.

Note that $\dot{\beta}_M = \left(\dot{\beta}(\dot{s}_1), \ldots, \dot{\beta}(\dot{s}_{\dot{K}-N})\right)$.

Lemma 2. *The output of algorithm 4.2 is an exact draw from the full conditional distribution in (19).*

Proof. *See Appendix A.*

Note that Algorithm 4.2 takes advantage of the factorisation of the global acceptance probability to perform the accept/reject procedure pointwise and avoid a much higher cost. The straightforward version of this algorithm would propose and accept/reject the variables all at once, resulting in a possibly very small acceptance probability. Firstly, K is sampled from π_0 then, for each of the $K - N$ locations, a pair $(s, \beta(s))$ is proposed from a $\mathcal{U}(S)$ and the prior GP and accepted with probability $\Phi(-W(s)\beta(s))$. MH alternatives may sound like an attractive possibility because of the lower computational cost but the usual choices for the proposal distribution may lead to slower convergence. This option performed poorly even for simple examples in some simulated studies considering both dependent and independent proposals.

The choice of the Gaussian c.d.f. in (2) is justified by the fact that it makes it possible to sample directly from the full conditional distribution in (20). This means that we have an algorithm with a reasonable computational cost and good convergence properties. The algorithm is the following.

Algorithm 4.3

1. Obtain W_K from (W_N, W_M) such that $\Phi_K(W_K\beta_K; I_K) = \Phi_N(W_N\beta_N; I_N)\Phi_{K-N}(-W_M\beta_M; I_{K-N})$;
2. Sample $\beta_K \sim SN(\mu_K, \Sigma_K, W_K)$ using Algorithm 4.1, where μ_K and Σ_K are the mean vector and covariance matrix, respectively, of $\pi_{GP}(\beta_K|\theta)$;
3. OUTPUT β_K.

Lemma 3. *The output of Algorithm 4.3 is an exact draw from the full conditional distribution in (20).*

Proof. *Simply note that (20) is proportional to the density of a* $SN(\mu_K, \Sigma_K, W_K)$.

The algorithms above are expected to contribute to the efficiency of the MCMC algorithm because of the blocking scheme with high-dimensional blocks that help controlling the autocorrelation of the chain which in turn improves its convergence properties.

Algorithm 4.2 may suggest a high computational cost as every try of the RS algorithm requires the simulation of the GP at a location given the existent ones. However, the most expensive part of this simulation is the computation of the inverse covariance matrix of the existing points which can be computed only once for every location. Moreover, once a location is accepted and we move to the next one, the new inverse covariance matrix may be obtained from the previous one at a very low cost using Schur complement.

The MCMC algorithm from Adams et al. (2009) uses the same blocking scheme considered here but with great differences in each update step. Whilst in here the two most crucial steps (first and second blocks) are performed by sampling directly from their respective full conditional distribution, they are performed via MH and Hamiltonian MC, respectively, in Adams et al. (2009). Their first block has a proposal distribution that, at each iteration of the chain, proposes a removal or an insertion of a single thinned event. This indicates that our algorithm will produce larger moves at every iteration and is, therefore, bound to have better mixing and convergence properties. Moreover, sampling the second block directly from its full conditional is expected to be more efficient than using a Hamiltonian update.

The next step of the Gibbs sampler draws θ from its full conditional distribution. This task may be carried out ordinarily - using direct simulation when possible or via an appropriately tuned MH step. There is also the option of breaking θ into smaller blocks if that is convenient for computational reasons. One attractive option is to use an adaptive Gaussian random walk Metropolis-Hastings step where the covariance matrix of the proposal is based on the empirical covariance matrix of the previous steps, as proposed by Roberts and Rosenthal (2009).

The forth and last step from the Gibbs sampler draws λ^* from its full conditional distribution. This can be obtained by routine calculations: if a conjugated Gamma prior $\mathcal{G}(\alpha_\lambda, \beta_\lambda)$ is adopted for λ^*, its full conditional is $\mathcal{G}(\alpha_\lambda + K, \beta_\lambda + \mu(S))$.

4.3 Estimating functionals of the intensity function

One of the purposes of fitting a Cox process to an observed point pattern is to estimate functionals of the intensity function. These functionals may include the intensity itself, the mean number of points at some subregion, etc. The estimation is perform by sampling such functionals to obtain MC estimates. The sampling step is performed based on the result in Lemma 1. Basically, it states that, in order to sample the Gaussian process at some arbitrary location from its posterior distribution, it is enough to sample from the GP prior conditional on the GP sample from the MCMC at locations $\{s_k\}_{k=1}^K$. The most efficient way to do this is by adding a sampling step to each iteration of the MCMC.

For example, in order to obtain estimates of the intensity function in a finite subset S_0 of S, we need posterior samples of (λ^*, β_{S_0}). That is achieved by sampling $(\lambda^*, \beta_K, \theta)$ from $\pi(\lambda^*, \beta_K, \theta|\{s_n\}_{n=1}^N, W, S)$ and then β_{S_0} from $\pi(\beta_{S_0}|\beta_K, \theta)$ at each step of the Markov chain after convergence is assumed to hold. This way, at iteration j of the chain, a draw from the posterior of λ_{S_0} is given by $\left\{\lambda^{*(j)}\Phi(W(\tilde{s})\beta^{(j)}(\tilde{s})),\ \tilde{s} \in S_0\right\}$. In order to reduce the computational cost, one may, for example, store the sum of β_{S_0} over the iterations for each location and output its posterior mean. Moreover, note that there is no particular reason to store all the draws from β_K, which also reduces the computational cost considerably. Another option to reduce computational cost is to use the thinned events as part of the grid and add extra locations to S_0 as necessary. If S_0 is a small set, it is computationally feasible to store the whole posterior sample of λ_{S_0} and compute, for example, credibility intervals and/or posterior marginal densities.

Another interesting functional to be estimated is the integral $I = \int_R \lambda(s)ds$ for some region $R \subseteq S$. This is the mean number of points in R. Monte Carlo estimates of $E[I|y]$ may be obtained without any discretisation error by introducing a r.v. $U \sim \mathcal{U}(R)$ and noting that

$$E_U[\lambda(U)] = \frac{1}{\mu(R)} \int_R \lambda(s)ds, \tag{23}$$

thus suggesting the estimator

$$\hat{E} = \mu(R)\ \frac{1}{J}\ \sum_{j=1}^{J} \lambda^{(j)}(U^{(j)}), \tag{24}$$

which is a strongly consistent estimator of $E[I|y]$ by the SLLN (see also Beskos et al., 2006). The samples of λ come from the posterior distribution. The accuracy of the estimator may be improved defining a partition of R and using one uniform to approximate the integral from each subregion of the partition.

5 Spatio-temporal model

It is straightforward to generalise the MCMC algorithm from Section 4.2 to the spatio-temporal case. We remind that $(X_0, \ldots, X_T)$ are conditionally mutually independent homogeneous PP's on S, given λ_T^*. The temporal dependence of the model is defined through β (and λ_T^*), as shown in Figure 1 (when considering the dashed arrows).

We now write the density of $(\psi|W, S)$ with respect to the dominating measure given by the product measure of the counting measure and the Lebesgue measure with corresponding dimensions and get

$$
\begin{aligned}
\pi(\psi|W,S) &= \prod_{t=0}^{T} \left[\pi(\{s_{t,n}\},\{s_{t,m}\}|\{s_{t,k}\},\beta_{K_t},W_t)\right] \pi(\beta_{K_\mathcal{T}}|\{s_{\mathcal{T},K}\},\theta) \\
&\quad \times \prod_{t=0}^{T} \left[\pi(\{s_{t,k}\}|K_t)\pi(K_t|\lambda^*)\pi(\lambda_t^*)\right] \pi(\theta) \\
&= \prod_{t=0}^{T} \left[\Phi_{N_t}(W_{N_t}\beta_{N_t};I_{N_t})\Phi_{K_t-N_t}(-W_{M_t}\beta_{M_t};I_{K_t-N_t})\right] \pi_{GP}(\beta_{K_\mathcal{T}}|\theta) \\
&\quad \times \prod_{t=0}^{T} \left[\frac{e^{-\lambda_t^*\mu(S)}\,(\lambda_t^*)^{K_t}}{K_t!}\pi(\lambda_t^*)\right] \pi(\theta) \qquad (25)
\end{aligned}
$$

where the new notation has a natural interpretation and π_{GP} is the density of the dynamic GP in (8).

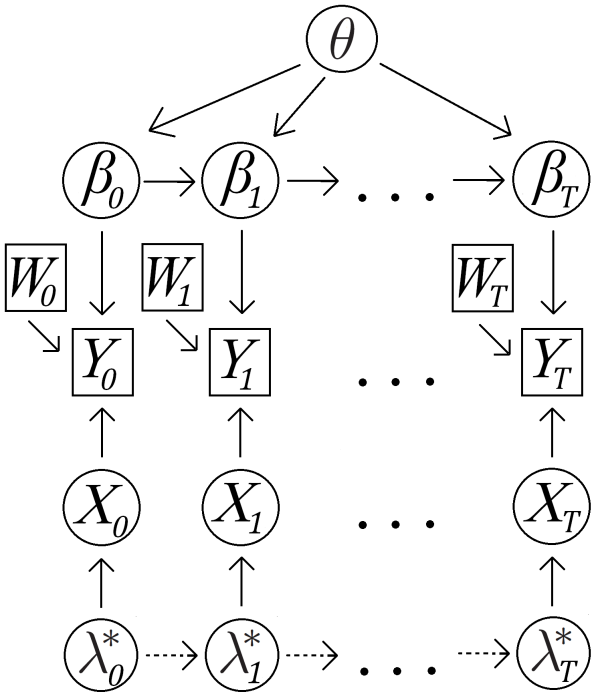

Figure 1: Graphical model for the spatio-temporal approach.

We have at least two options for the blocking scheme. The first one samples $\left(K_t, \{s_{t,m}\}, \beta_{M_t}\right)$ and β_{K_t} separately, for each time. This algorithm may, however, lead to a chain with poor mixing properties if T is large due to the temporal dependence of β (see Carter and Kohn, 1994; Fruhwirth-Schnatter, 1994; Gamerman, 1998). This problem motivates the second blocking scheme which makes $\left\{\left(K_t, \{s_{t,m}\}, \beta_{M_t}\right)\right\}_{t=0}^{T}$ and $\left\{\beta_{K_t}\right\}_{t=0}^{T}$ one block each. This choice eliminates the mixing problem mentioned above

and is particularly appealing in the DGP context. To sample $\left\{\left(K_t, \{s_{t,m}\}, \beta_{M_t}\right)\right\}_{t=0}^{T}$, Algorithm 4.2 is applied for each time t and step 5 has to consider the temporal dependence of the GP. The same idea extends Algorithm 4.3 - step 4 now considers the temporal dependence of the GP. The one-at-a-time simulation is possible due to the following factorisation of the full conditional distribution:

$$\pi\left(\left\{\beta_{K_t}\right\}_{t=0}^{T}|\cdot\right) \propto \prod_{t=0}^{T}\left[\Phi_{N_t}(W_{N_t}\beta_{N_t}; I_{N_t})\Phi_{K_t-N_t}(-W_{M_t}\beta_{M_t}; I_{K_t-N_t})\pi_{GP}(\beta_{K_t}|\beta_{K_{(t-1)}}, \theta)\right], \tag{26}$$

where $\beta_{K_{(t-1)}} := (\beta_{K_0}, \ldots, \beta_{K_{t-1}})$ and $\beta_{K_{-1}} = \emptyset$.

The full conditional distribution of θ is carried out as before and particular blocking schemes may be motivated by the spatio-temporal structure. Finally, for a prior $\mathcal{G}(\alpha_{\lambda_t}, \beta_{\lambda_t})$ - which may be the same for every t, the full conditional of each λ_t^* is $\mathcal{G}(\alpha_{\lambda_t} + K_t, \beta_{\lambda_t} + \mu(S))$. In the case $\lambda_t^* = \lambda^*$, $\forall t$, the full conditional of this parameter is $\mathcal{G}(\alpha_{\lambda_t} + \sum_{t=1}^{T} K_t, \beta_{\lambda_t} + T\mu(S))$.

Extensions of the spatio-temporal model above can be proposed by adding a temporal dependence structure to $\lambda_{1:T}^*$. This is particularly useful if prediction towards future times is required. One interesting possibility is the Markov structure proposed by Gamerman et al. (2013) (in a state-space model context) where $\lambda_1^* \sim \mathcal{G}(wa_0, wb_0)$, $\lambda_t^*|K_{1:t-1}, \lambda_{t-1}^* = w^{-1}\lambda_{t-1}^*\varsigma_t$ and $\varsigma_t \sim Beta(wa_t, (1-w)a_t)$. This structure is represented in Figure 1 by the dashed arrows. The full conditional distribution of $\lambda_{1:T}^*$ is available for direct sampling using the results in Gamerman et al. (2013).

5.1 Prediction

A Bayesian analysis usually includes the task of prediction, i.e. what should we expect from a new observation of the model being consider after we have updated our knowledge about this with the current data. That is performed through the predictive distribution, which is the conditional distribution of the future observations given the current data and is obtained by integrating out the unknown parameters and other components of the model. In our context, this means to integrate out the joint posterior of future data and {intensity function, hyperparameters} over the latter.

Samples from the predictive distribution can be trivially obtained in a MCMC context. We simply add one extra step to each iteration of the MCMC after convergence is assumed. In this step we firstly sample $\lambda_{\mathcal{T}^*}^*$, which may be drawn from the prior (for independent λ_t^*'s), from the posterior (for a common λ^*) or from an evolution equation (if a Markov structure is adopted for the λ_t^*'s). Secondly, we perform Algorithm 1 to sample a new realisation of the process. Note that this will include simulating β at the candidate points (to be thinned), which is performed the same way β_{S_0} is sampled.

Suppose that we want to predict the process at future times $\mathcal{T}^* = (T+1, \ldots, T+J)$. The algorithm to sample from the predictive distribution of $Y_{\mathcal{T}^*}$ - the Cox process on $S \times \mathcal{T}^*$, proceeds iteratively on time from $T+1$ onwards. Firstly, we sample λ_t^* (which depends on the structure that has been adopted), then apply Algorithm 1 with the Gaussian process being simulated from $\pi_{GP}(\beta_{K_t}|\beta_{K_{\mathcal{T}}}, \beta_{K_{T+1:t-1}}, \theta)$. This algorithm is

supported by the following result:

$$\pi(y_{\mathcal{T}^*}|y_{\mathcal{T}}) \propto \int \pi(y_{\mathcal{T}^*}, \lambda^*_{\mathcal{T}^*}, \beta_{\mathcal{T}^*}, \lambda^*_T, \beta_{\mathcal{T}}, \theta|y_{\mathcal{T}}) d\lambda^*_{\mathcal{T}^*} d\beta_{\mathcal{T}^*} d\lambda^*_T d\beta_{\mathcal{T}} d\theta$$
$$= \int \prod_{t=T+1}^{T+J} \left[\pi(y_t|\lambda^*_t, \beta_t)\pi(\lambda^*_t|\lambda^*_{t-1}, y_{t-1})\pi(\beta_t|\beta_{t-1}, \theta)\right] \pi(\lambda^*_T, \beta_{\mathcal{T}}, \theta|y_{\mathcal{T}}) d\lambda^*_{\mathcal{T}^*} d\beta_{\mathcal{T}^*} d\lambda^*_T d\beta_{\mathcal{T}} d\theta, \tag{27}$$

where $y_{\mathcal{T}}$ are the observed data at times $\mathcal{T}$, $\beta_{\mathcal{T}}$ is the GP at the locations of $y_{\mathcal{T}}$ and $(y_{\mathcal{T}^*}, \beta_{\mathcal{T}^*})$ represent these components at a finite collection of locations at times $\mathcal{T}^*$.

The predictive distribution may be explored in different ways, specially in a point process context, by choosing convenient functions of the observations to analyse. Note that the same algorithm provides prediction of the intensity function in future times $\mathcal{T}^*$.

6 Simulated examples

The methodology proposed in this paper is now used to perform inference in synthetic data sets. We present three examples, the first two examples consider spatial models in one and two dimensions, respectively, and the third one considers a spatio-temporal model with seasonal effect.

6.1 Spatial models

The data was simulated from a Poisson process with IF $\lambda(s) = 2\exp(-s/15) + \exp(-(s-25)^2/100)$, for $s \in [0, 50]$. We apply the inference methodology proposed in this paper to study its efficiency and robustness under different prior specifications. We assume that β is a Gaussian process with constant mean function μ and the covariance function given in (7) with $\gamma = 3/2$.

An extensive analysis indicated that the posterior distribution of the intensity function is sensitive to the prior specification of the GP, as expected. Also, data does not contain much information about the hyperparameters of the GP. As a consequence, non-informative priors lead to high variance posterior for these and unstable estimates of the intensity function. Efficient estimation requires the hyperparameters to be fixed or have highly informative priors. Reasonable choices of the hyperparameters' values can be obtained with some knowledge about the behavior of Gaussian processes. In particular the values should reflect the smoothness expected for the intensity function. Parameter λ^* is also sensitive to prior specification but good results are obtained following the guidelines of Section 3.3.

We consider four different prior specifications. All of them consider a $\mathcal{G}(2.2, 1.5)$ prior for λ^*. The first three specifications fix the parameter vector $\theta = (\mu, \sigma^2, \tau^2)$ at $(0, 1, 20)$, $(0, 1, 10)$ and $(0, 1, 5)$, respectively. The last case fixes $\mu = 0$ and estimates (σ^2, τ^2) with uniform priors $\sigma^2 \sim \mathcal{U}(0.25, 4)$ and $\tau^2 \sim \mathcal{U}(1, 30)$. The estimated intensity function in each case is shown in Figure 2. Results are similar for all prior specifications which reinforces the idea that reasonable choices for the hyperparameters lead to good results.

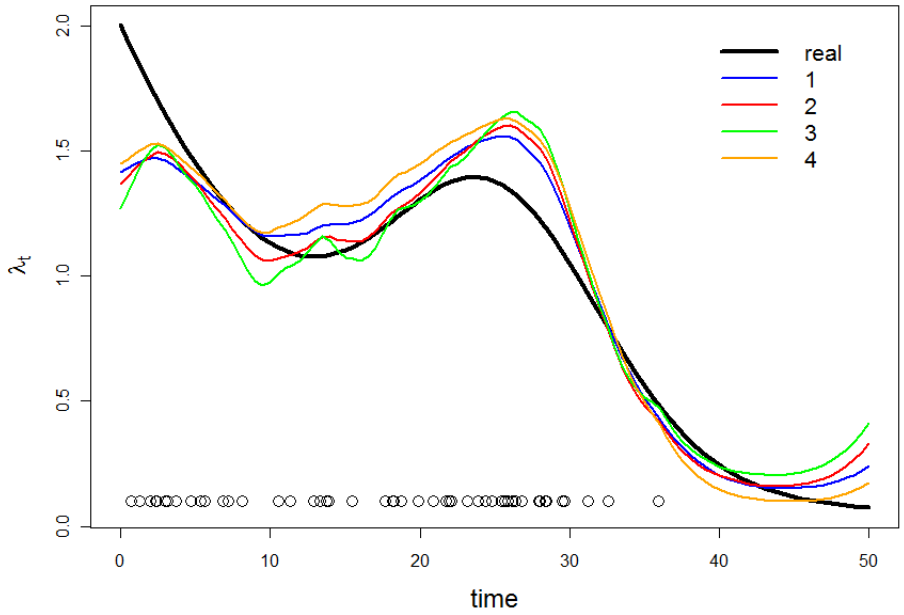

Figure 2: Real and posterior mean intensity function and realisation of the unidimensional Poisson process.

The influence of the prior is clearly seen as the estimated function is smoother for higher choices of τ^2. The trace plots of λ^* and (σ^2, τ^2) (in case 4) suggest fast convergence of the chain. The posterior distribution of λ^* is very similar in all cases, with mean around 2.05 and s.d. 0.90. The parameter vector (σ^2, τ^2) in case 4 has large posterior variance indicating that the data does not have much information about it. The posterior mean and standard deviation of (σ^2, τ^2) are $(2.81, 22.85)$ and $(0.79, 5.43)$, respectively.

Data was also simulated from a bidimensional Poisson process on $[0, 10] \times [0, 10]$ with IF $\lambda(s) = 3\Phi(\beta^{(0)}(s))$, where $\beta^{(0)}(s) = (8/3)\exp\{-s_{(1)}^2/30\} + (4/3)\exp\{-(s_{(2)} - 7)^2/12\} - 2$. We assume the same covariance function as in the unidimensional example and set the values (0,4,10) for the mean, variance and correlation parameters of $\beta^{(0)}$. Figure 3 shows real and estimated intensity function and the realisation of the process. It is clear that the intensity function is well estimated.

6.2 Spatio-temporal model with a seasonal component

We consider a spatio-temporal example with a seasonal component whose effect varies in space. More specifically, we consider the model specified in (11)-(12) with $p = 4$, $G = I$ and $w_{1,t+1,t} = 0$, $\forall\, t$, to simulate four annual cycles with quarterly data ($T = 15$). We consider Gaussian processes with the same covariance structure used in the previous section, for $\beta_{0,0}$, w_t and β_1. More specifically, we set their hyperparameters $(0, 2^2, 5)$, $(0, 0.7^2, 10)$ and $(1, 1.5^2, 5)$, respectively. The data is simulated considering $(-0.2, 1.8^2, 15)$ and $(0, 0.5^2, 20)$ for $\beta_{0,0}$ and w_t and a deterministic structure for β_1 - $2.4\exp\{-s_{(1)}^2/25\} +$

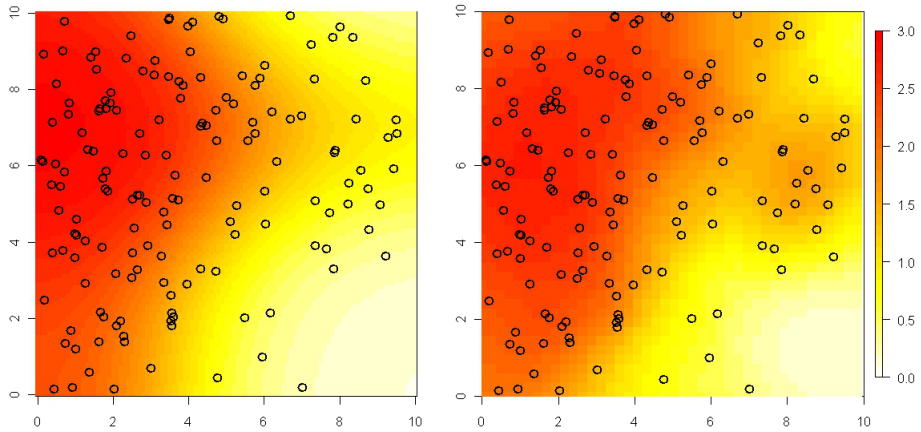

Figure 3: Real (left) and posterior mean (right) intensity function and realisation of the bidimensional Poisson process.

$0.6\exp\{-(s_{(2)}-7)^2/36\}-0.288$. Furthermore, we set $\lambda_t^*=1.5$, $\forall t$ and $\phi=\pi/2$ (for the generation and for the analysis). The results from this section are based on first-order approximations to simulate the Gaussian processes.

Figure 4 shows the estimation of the space-varying seasonal coefficient to be satisfactory, capturing the dip in the (south)eastern portion of the area of study. Figure 5 shows that the IF is very well estimated at a wide selection of times. Figure 6 shows prediction results. Prediction for the (latent) IF and for the (observed) number of points in $[0,2]\times[0,2]$ for future times is also good. Finally, we also predict the average number of points in the whole space at time $T=16$ using estimator (24), which returned the value 114.69. The expected value based on the true model is 121.58.

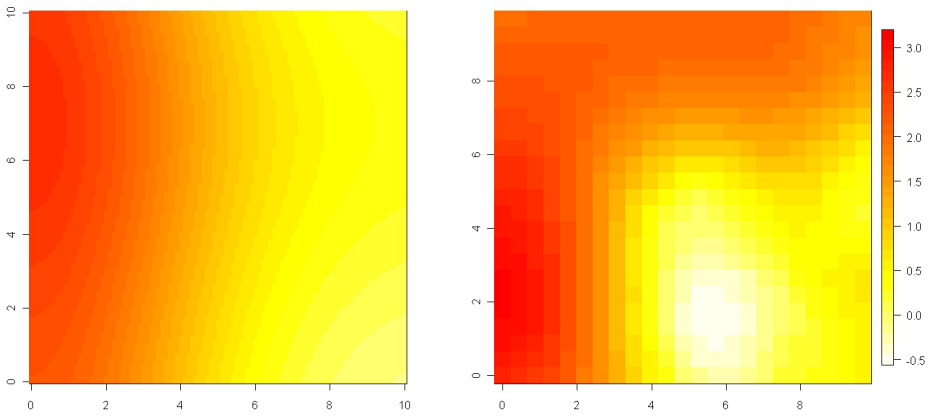

Figure 4: Seasonal effect β_1. Real (left) and posterior mean (right).

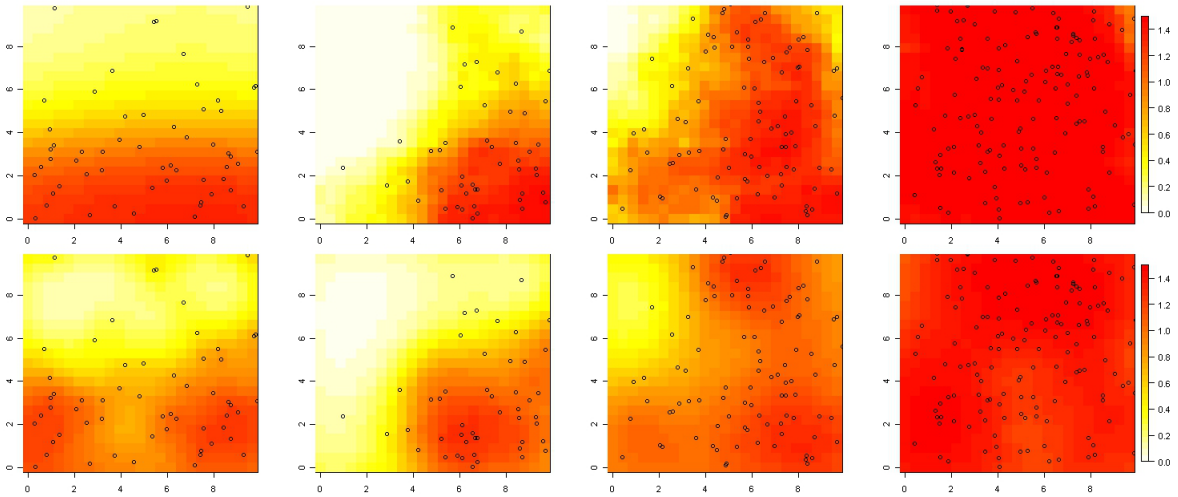

Figure 5: True (top) and posterior mean (bottom) intensity function for times 0, 5, 10 and 15. Circles represent the data.

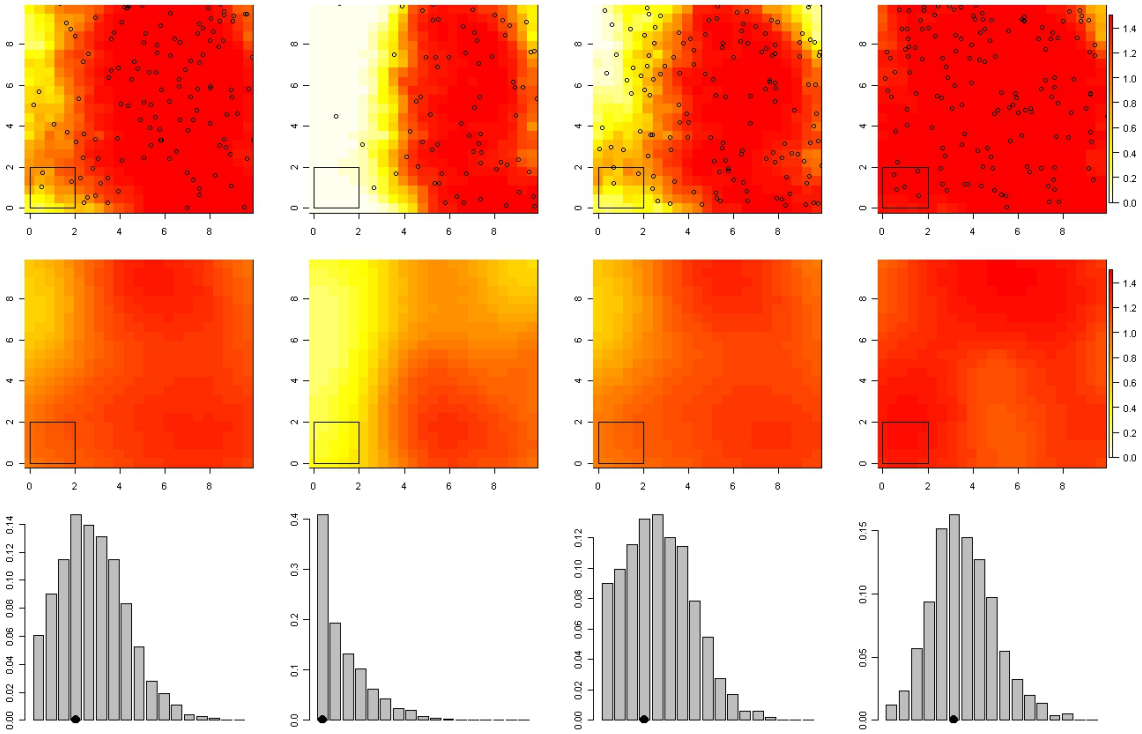

Figure 6: Prediction for times 16, 17, 18 and 19. Top: true IF and realisation of the Y process; middle: predictive mean of the IF; bottom: predictive distribution for the number of points in $[0, 2] \times [0, 2]$, the black dots represent the respective (future) true values.

7 Final remarks

This paper proposes a novel methodology to perform exact Bayesian inference in spatio-temporal Cox processes in which the intensity function dynamics is described by a multivariate Gaussian process. We showed how usual components of spatio-temporal point patterns such as trend, seasonality and covariates can be incorporated, with flexibility of their effects warranted by the Gaussian process prior.

The methodology is exact in the sense that no discrete approximation of the process is used and Monte Carlo is the only source of inaccuracy. Inference is performed via MCMC, more specifically, a Gibbs sampler whose particular choice of blocking and sampling scheme leads to fast convergence. The validity of the methodology is established through the proofs of the main results. Finally, simulated studies illustrate the methodology and provide empirical evidence of its efficiency.

This work may give rise to new problems and possibilities that may be considered in future work. For example, computational developments are required to deal with very large data sets, which is a general problem when working with Gaussian processes. An immediate extension is to consider stochastic marks to the Poisson events. These marks may be described with a variety of components, whose effects are allowed to vary smoothly, in line with the models used for the IF.

Acknowledgements

The authors would like to thank Gareth Roberts and Krzysztof Łatuszyński for insightful discussions about MCMC and Piotr Zwiernik for insightful discussions about matrices. The second author would like to thank CNPq Brazil for financial support.

Appendix A - Proofs

Proof of Lemma 1

The first part of the Lemma comes of the fact that $\pi(\beta_{S_0}, \psi|W, S) = \pi(\beta_{S_0}|\beta_K, \theta)\pi(\psi|W, S)$.

For the second part we have

$$\begin{aligned}
\pi(\beta_{S_0}|\{s_n\}, W, S) &= \int \pi(\beta_{S_0}, \beta_K, \theta|\{s_n\}, W, S) d\beta_K d\theta \\
&= \int \pi(\beta_{S_0}|\beta_K, \theta, \{s_n\}, W, S)\pi(\beta_K, \theta|\{s_n\}, W, S) d\beta_K d\theta \\
&= \int \pi(\beta_{S_0}|\beta_K, \theta)\pi(\beta_K, \theta|\{s_n\}, W, S) d\beta_K d\theta
\end{aligned}$$

To go from the second to the third row we use the first part of the Lemma.

Proof of Proposition 1

Firstly, we prove that Γ and Σ^* are positive definite matrices, for any $m \times d$ real matrix W and positive definite $d \times d$ matrix Σ. Let X and ε be r.v.'s such that $X \sim N(0, \Sigma)$ and $\varepsilon \sim$

$N(0, I_m)$. Now define a r.v. $Y = WX + \varepsilon$. This implies that $Cov(Y, Y) = I_m + W\Sigma W' = \Gamma$ and $\Gamma \succ 0$. Now, the Schur complement of Σ^* is given by $S = \Gamma - \Delta'\Sigma^{-1}\Delta = \Gamma - W\Sigma W' = I_m$. The fact that $S \succ 0$ and $\Sigma \succ 0$ implies that $\Sigma^* \succ 0$.

Now, by standard properties of the multivariate normal distribution we have that $(U_0|U_1 = z^*) \sim N_m(\Delta'\Sigma^{-1}z^*, \Gamma - \Delta'\Sigma^{-1}\Delta)$, where $\Gamma - \Delta'\Sigma^{-1}\Delta = I_m$. Therefore, by the symmetry of the standard Gaussian cdf and the Bayes Theorem, we have that the density of $(U_1|U_0 > -\gamma)$ is given by

$$f(z^*) = \frac{f_{U_1}(z^*)P(U_0 > -\gamma|U_1 = z^*)}{P(U_0 > -\gamma)} = \frac{\phi(z^*, \Sigma)\Phi_m(\gamma + \Delta'\Sigma^{-1}z^*; I_m)}{\Phi_m(\gamma; \Gamma)}.$$

We now apply the transformation theorem to find the density of $(U_1 + \xi|U_0 > -\gamma)$.

Proof of Proposition 2

The density f of $(U_0^*|U_0^* \in B)$ is given by

$$f(u) = \frac{1}{P(U_0^* \in B)}\phi_m(u; I_m)\mathbb{1}(u \in B).$$

Applying the transformation theorem to find the density f^* of $(AU_0^*|U_0^* \in B)$, we get

$$f^*(u) = \frac{1}{\Phi_m(\gamma; \Gamma)}\phi_m(u; \Gamma)\mathbb{1}(u > -\gamma).$$

Proof of Lemma 2

Firstly note that $\pi_0(K_0) = \dfrac{P(\dot{K} = K_0)\mathbb{1}(K_0 \geq N)}{P(\dot{K} \geq N)} \propto [\lambda^*\mu(S)]^{K_0}\dfrac{1}{K_0!}\mathbb{1}(K_0 \geq N)$.

Now consider the case $\dot{K} > N$. The density of the algorithms's output w.r.t. the dominating measure $\delta \otimes \mathbb{L}^{\dot{K}-N} \otimes \mathbb{L}^{\dot{K}-N}$ is given by

$$\begin{aligned}
&\pi(\dot{K}, \{\dot{s}_m\}_{m=1}^{\dot{K}-N}, \dot{\beta}_M) \propto \pi_0(\dot{K})\pi(\dot{s}_{r_1}, \dot{\beta}_{r_1}, \ldots, \dot{s}_{r_{K-N}}, \dot{\beta}_{r_{K-N}}|Z_{r_1} = 1, \ldots, Z_{r_{K-N}} = 1) \\
&\propto \pi_0(\dot{K})P(Z_{r_1} = 1, \ldots, Z_{r_{K-N}} = 1|\dot{s}_{r_1}, \dot{\beta}_{r_1}, \ldots, \dot{s}_{r_{K-N}}, \dot{\beta}_{r_{K-N}})\pi(\dot{s}_{r_1}, \dot{\beta}_{r_1}, \ldots, \dot{s}_{r_{K-N}}, \dot{\beta}_{r_{K-N}}) \\
&= \pi_0(\dot{K})\left[\prod_{m=1}^{\dot{K}-N}(\Phi(W(\dot{s}_{r_m})'\beta(\dot{s}_{r_m})))[\mu(s)]^{K-N}\pi_{GP}(\dot{\beta}_{r_1}, \ldots, \dot{\beta}_{r_{K-N}}|\beta_N, \theta)\right] \propto (19),
\end{aligned}$$

where $\dot{\beta}_{r_m} = \dot{\beta}_{r_m}(\dot{s}_{r_m})$. The case where $\dot{K} = N$ is straightforward - just compare π_0 with $\pi(\psi|W, S)$, when $M = 0$.

Appendix B

Algorithm to sample from a truncated multivariate Normal distribution

Firstly note that our aim is to simulate from a multivariate Normal distribution with independent coordinates restricted to a region R defined by linear constraints, more specifically

$R := \{u_0^*,\ Au_0^* > -\gamma\}$. Moreover, note that A is the lower triangular matrix obtained from applying the Cholesky decomposition to Γ.

The most obvious way to simulated exactly from a truncated multivariate Normal distribution is via a rejection sampling algorithm that proposes from the untruncated distribution. In this case, the global acceptance probability of this algorithm is equal to the probability of the truncated region. However, this probability is typically going to be very small in high dimensions, making this algorithm very inefficient.

A more efficient alternative is provided by MCMC, more specifically, a Gibbs sampling. This method could be applied directly to U_0 but the resulting chain would have much higher correlation among the blocks of the Gibbs sampler, which would considerably slow down its convergence (see Rodriguez-Yam et al., 2004).

The Gibbs sampler alternates among the simulation of $(U_{0,i}^*|U_{0,-i}^*),\ \ i = 1, \ldots, m$, where $U_{0,-i}^* = U_0^* \setminus \{U_{0,i}^*\}$, which are all univariate standard Gaussians restricted to R. Basically, for a given i, R consists of $(m - i + 1)$ linear inequalities and has the form $(\max\{l_j\}, \min\{L_j\})$, where l_j and L_j are the lower and upper limits, respectively, of each inequality. Note that the diagonal of A is strictly positive (Γ is positive definite) which means that the lower limit will always be a real number whereas the upper limit may be $+\infty$.

In order to favor a faster convergence, we choose an initial value that already belongs to R. This is trivially obtained by taking advantage of the triangular form of A and simulating U_0^* recursively from $U_{0,1}^*$ onwards. Simulation experiments suggest that m iterations of the chain are enough to obtain good results. The algorithm is as follows:

1. Simulate the initial value of the chain and make $k = 1$;
2. For i in $1 : m$ do:
   2.1. Simulate ${u_{0,i}^*}^{(k)}$ from $({U_{0,i}^*}^{(k)}|{U_{0,1}^*}^{(k)}, \ldots, {U_{0,i-1}^*}^{(k)}, {U_{0,i+1}^*}^{(k-1)}, \ldots, {U_{0m}^*}^{(k-1)})$;
3. If k has reached the desired number of iterations, STOP and OUTPUT the last sampled value of U_0^*; ELSE, GOTO 2;

References

Adams, R. P., Murray, I., and Mackay, D. J. C. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*.

Arellano-Valle, R. B. and Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33:561–574.

Banerjee, A., Dunson, D. B., and Tokdar, S. T. (2013). Efficient Gaussian process regression for large datasets. *Biometrika*, 100:75–89.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based inference for discretely observed diffusion pro-

cesses (with discussion). *Journal of the Royal Statistical Society, Series B*, 68(3):333–382.

Brix, A. and Diggle, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society, Series B*, 63:823–841.

Carlin, B. P., Banerjee, S., and Finley, A. O. (2007). spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19:1–24.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81:541–533.

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

Diggle, P. J. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Chapman & Hall, London, 3rd edition.

Fruhwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15:183–202.

Gamerman, D. (1998). Markov chain monte carlo for dynamic generalised linear models. *Biometrika*, 85:215–227.

Gamerman, D., Salazar, E., and Reis, E. A. (2007). *Bayesian Statistics 8*, chapter Dynamic Gaussian process priors with applications to the analysis of space-time data (with discussion), pages 149–174. Oxford Univeristy Press, Oxford.

Gamerman, D., Santos, T. R., and Franco, G. C. (2013). A non-Gaussian family of state-space models with exact marginal likelihood. *Journal of Time Series Analysis*, 34:625–645.

Gelfand, A. E., Banerjee, S., and Gamerman, D. (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *EnvironMetrics*, 16:465–479.

Gonçalves, F. B. and Roberts, G. O. (2014). Exact simulation problems for jump-diffusions. *Methodology and Computing in Applied Probability*, 16:907–930.

Gonçalves, F. B., Roberts, G. O., and Łatuszyński, K. G. (2015). Exact monte carlo likelihood-based inference for jump-diffusion processes. *Submitted*.

Kottas, A. and Sansó, B. (2007). Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference*, 137:3151–3163.

Lewis, P. A. W. and Shedler, G. S. (1979). Simulation of a nonhomogeneous Poisson process by thinning. *Naval Research Logistics Quarterly*, 26:403–413.

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

Pinto Jr, J. A., Gamerman, D., Paez, M. S., and Alves, R. H. F. (2015). Point pattern analysis with spatially varying covariate effects, applied to the study of cerebrovascular deaths. *Statistics in Medicine*, 34:1214–1226.

Reis, E. A., Gamerman, D., Paez, M. S., and Martins, T. G. (2013). Bayesian dynamic models for space-time point processes. *Computational Statistics & Data Analysis*, 60:146–156.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18:349–367.

Rodriguez-Yam, G., Davis, R. A., and Scharf, L. L. (2004). Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression, unpublished manuscript.

Sermaidis, G., Papaspiliopoulos, O., Roberts, G. O., Beskos, A., and Fearnhead, P. (2013). Markov chain Monte Carlo for exact inference for diffusions. *Scandinavian Journal of Statistics*, 40:294–321.

Xiao, S., Kottas, A., and Sansó, B. (2015). Modeling for seasonal marked point processes: An analysis of evolving hurricane occurrences. *To appear in Annals of Applied Statistics*.