

# GP-ETAS: Semiparametric Bayesian inference for the spatio-temporal Epidemic Type Aftershock Sequence model

Christian Molkenthin<sup>\*1</sup>, Christian Donner<sup>\*2</sup>, Sebastian Reich<sup>1</sup>, Gert Zöller<sup>1</sup>,  
Sebastian Hainzl<sup>3</sup>, Matthias Holschneider<sup>1</sup>, and Manfred Opper<sup>4</sup>

<sup>1</sup>University of Potsdam, Institute of Mathematics

<sup>2</sup>ETH Zürich, Swiss Data Science Center

<sup>3</sup>Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences

<sup>4</sup>TU Berlin, Institute of Software Engineering and Theoretical Computer Science

May 27, 2020

## Abstract

The spatio-temporal Epidemic Type Aftershock Sequence (ETAS) model is widely used to describe the self-exciting nature of earthquake occurrences. While traditional inference methods provide only point estimates of the model parameters, we aim at a full Bayesian treatment of model inference, allowing naturally to incorporate prior knowledge and uncertainty quantification of the resulting estimates. Therefore, we introduce a highly flexible, non-parametric representation for the spatially varying ETAS background intensity through a Gaussian process (GP) prior. Combined with classical triggering functions this results in a new model formulation, namely the GP-ETAS model. We enable tractable and efficient Gibbs sampling by deriving an augmented form of the GP-ETAS inference problem. This novel sampling approach allows us to assess the posterior model variables conditioned on observed earthquake catalogues, i.e., the spatial background intensity and the parameters of the triggering function. Empirical results on two synthetic data sets indicate that GP-ETAS outperforms standard models and thus demonstrate the predictive power for observed earthquake catalogues including uncertainty quantification for the estimated parameters. Finally, a case study for the l’Aquila region, Italy, with the devastating event on 6 April 2009, is presented.

## 1 Introduction

Point process models are often used in statistical seismology for describing the occurrence of earthquakes (point data) in a spatio-temporal setting. The most widely used one is the Epidemic Type Aftershock Sequence (ETAS) model, first introduced as a temporal point process model (Ogata, 1988), and later enhanced to the currently predominantly employed spatio-temporal version (Ogata, 1998). Main applications are seismic forecasting or the characterisation of earthquake clustering in a particular geographical region and topics alike (e.g., Jordan et al., 2011). The ETAS model is an example of a self-exciting, spatio-temporal, marked point process, which is a particular Hawkes process model, extending the temporal Hawkes process proposed by Hawkes (1971). Self-excitation means that one event can trigger a series of subsequent follow-up events (offspring), as in the case of earthquakes, main shocks and aftershocks. The ETAS model assigns the earthquake magnitude as an additional mark to each event. Besides its primary application in seismology, the Hawkes process is utilised in several other domains, e.g. finance (Bacry et al., 2015; Filimonov and Sornette, 2015), crime (Mohler et al., 2011; Porter and White, 2010), neuronal activities (Gerhard et al., 2017), social networks (Zhao et al., 2015; Zhou et al., 2013), genomes (Reynaud-Bouret and Schbath, 2010), transportation (Hu and Jin, 2017).

---

<sup>\*</sup>Contributed equally, e-mail: [molkenth@uni-potsdam.de](mailto:molkenth@uni-potsdam.de), [christian.research@mailbox.org](mailto:christian.research@mailbox.org)

The ETAS model is characterised by its conditional intensity function, that is, the rate of arriving events conditioned on the history of previous events. This time-dependent conditional intensity function itself consists of two parts, (i) a stationary background intensity  $\mu$  of a Poisson process, which models the arrival of spontaneous (exogenous) events, and (ii) a time-dependent triggering function  $\varphi$  which encodes the form of self-excitation by adding a positive impulse response for each event, that is, a spontaneous jump which decays gradually at time progresses. An alternative approach interprets the stationary Hawkes process (i.e., the ETAS model) as a Poisson cluster process or branching process (Hawkes and Oakes, 1974), which leads to the concept of a non-observable, underlying branching structure (latent variable). Each event has either a direct parent from which it was generated or is background; this yields an ordered branching structure useful for designing simulation and inference algorithms, e.g. (Veen and Schoenberg, 2008; Zhuang et al., 2002).

The fitting of an ETAS model to data entails learning the conditional intensity function. Most currently used ETAS models employ a *parametric* form for the background  $\mu$  and the triggering function  $\varphi$ . The parameters are then calibrated via maximum-likelihood estimation (MLE), maximising the classical likelihood function for point processes. Unfortunately, MLE has no simple analytical form. Alternatively, different numerical optimisation methods are employed involving, e.g., an Expectation-Maximisation (EM, Dempster et al. (1977)) algorithm using the latent branching structure (Lippiello et al., 2014; Lombardi, 2015; Ogata, 1998; Veen and Schoenberg, 2008).

*Non-parametric* methods have also been suggested previously to fit the conditional intensity function (or parts of it). For example, Zhuang et al. (2002) and Adelfio and Chiodi (2014) fit simultaneously a non-parametric background intensity via kernel density estimation and a classical parametric triggering kernel; Marsan and Lengliné (2008) consider a constant background intensity combined with a non-parametric histogram estimator of the triggering kernel; Mohler et al. (2011) suggest non-parametric kernel density estimators for both the components, background  $\mu$  and offspring  $\varphi$ ; and Fox et al. (2016) propose a non-parametric kernel density estimator for the background and non-parametric histogram estimation for the triggering kernel. Furthermore, Bacry and Muzy (2016) suggest a non-Bayesian, non-parametric way of estimating the triggering function of a Hawkes process based on Wiener Hopf integral equation; Kirchner (2017) presents a non-Bayesian non-parametric estimation procedure for a multivariate Hawkes process based on an integer-valued autoregressive model.

Uncertainty quantification of the ETAS model remains challenging. Most estimation techniques deliver a point estimate for its conditional intensity function and uncertainty quantification is usually achieved by relying on standard errors of estimated ETAS parameters, based on the Hessian (Ogata, 1978; Rathbun, 1996; Wang et al., 2010). This approach requires that the observational window is long enough (sufficiently large sample size), otherwise it may lead to an underestimation of parameter uncertainties. Moreover, standard errors based on Hessians cannot be obtained in the non-parametric case. Another approach to uncertainty quantification relies on various bootstrap techniques based on many forward simulations, e.g., Fox et al. (2016). Ad hoc variants for quantifying uncertainty have also been devised, e.g., by the solutions of multiple optimisation runs of the MLE, e.g. Lombardi (2015).

None of the aforementioned uncertainty quantification methods are fully satisfactory and we believe that a fully semi-parametric Bayesian framework is worthwhile pursuing, which allows one to incorporate prior knowledge. The posterior distribution effectively encodes the uncertainty of the quantities arising from data and a prior distribution. However, this poses a challenge for a spatio-temporal ETAS model, as there is no known conjugate structure, that is, the posterior can not be obtained in closed-form. One way to deal with this problem is to employ Monte Carlo sampling techniques, e.g. via Markov chain Monte Carlo (MCMC). However implementing MCMC remains challenging for *non- or semi-parametric* conditional intensity functions. Several studies have suggested Bayesian methods for the temporal, or multivariate Hawkes process, either based on parametric forms of the conditional intensity function (Rasmussen, 2013; Ross, 2018) or for non-parameteric versions (Donnet et al., 2018; Linderman and Adams, 2015; Zhang et al., 2019a,b; Zhou et al., 2019). But these studies rarely consider the spatio-temporal ETAS model and only with strong simplifications, e.g., a constant background intensity  $\mu$  (Rasmussen, 2013). Recently, however, (Kolev and Ross, 2020) considered an inhomogeneous background intensity modelled via a Dirichlet process.

It is desirable to estimate the spatially dependent background  $\mu$  of an ETAS model fully *non-parametrically* as it is often difficult to specify an appropriate functional form a priori. The background intensity (also called long-term component) is of particular importance for seismic hazard assessment and seismic forecasting. It is often preferred to maintain a specific *parametric* triggering function  $\varphi$  (e.g., modified Omori law [Omori \(1894\)](#); [Utsu \(1961\)](#)) as there is a long tradition for interpreting and comparing this particular parametric form in different settings, regions, etc. Thus, one faces two main issues for the development of a suitable Bayesian inference approach:

- (i) providing a Bayesian non-parametric way of modelling the background intensity  $\mu$ , and
- (ii) creating a fully Bayesian inference algorithm for the resulting ETAS models including its parametric triggering component  $\varphi$ .

We address these two issues in this paper by first formulating a Bayesian non-parametric approach to the estimation of the background intensity  $\mu$  via a Gaussian process (GP) prior. Secondly, we propose and implement a computationally tractable approach for the implied Bayesian inference problem by introducing auxiliary variables: a latent branching structure, a latent Poisson process, and latent Pólya–gamma random variables. More specifically, we suggest to model the background intensity  $\mu$  *non-parametrically* by sigmoid transformed realisations of a GP prior, i.e., as a Sigmoid-Gauss-Cox-Process (SGCP, [Adams et al. \(2009\)](#)), which is a doubly stochastic Poisson process. No specific functional form has to be chosen for the intensity function, and the prior fully specifies the chosen GP. [Adams et al. \(2009\)](#) proposed a Bayesian inference scheme via MCMC for SGCPs. However, the suggested scheme is computationally demanding and convergence is slow. Our paper relies instead on the work of [Donner and Opper \(2018\)](#) who recently enhanced Bayesian inference for SGCPs substantially by data augmentation with Pólya–gamma random variables ([Polson et al., 2013](#)). The triggering function  $\varphi$  is modelled in a classical *parametric* way, which together with the SGCP model for  $\mu$  leads to a novel semi-parametric ETAS model formulation, which we denote as GP-ETAS. In order to implement such an approach, we need to address a number of computational challenges:

- (i) the background intensity  $\mu$  and the triggering function  $\varphi$  are not directly separable in the likelihood;
- (ii) intractable integrals for the posterior computation when  $\mu$  is modelled as SGCP, and
- (iii) handling a non-Gaussian point process likelihood while using a Gaussian process prior.

We show how these challenges can be resolved by data augmentation (introducing auxiliary variables), which strongly simplifies the Bayesian inference problem. It effectively allows us to construct an efficient MCMC sampling scheme for the posterior involving an overall Gibbs sampler ([Geman and Geman, 1984](#)) consisting of three main steps, each conditioned on the previous:

- (a) conditionally sampling the latent branching structure which factorises the likelihood function into background and triggering component;
- (b) conditionally sampling the posterior of the background intensity  $\mu$  from explicit conditional densities easy to sample from; and
- (c) conditionally sampling the parameters of the triggering function  $\varphi$  by employing Metropolis-Hastings (MH) ([Hastings, 1970](#)) steps.

The remainder of this paper is structured as follows: First we describe the classical spatio-temporal ETAS model; secondly we introduce our GP-ETAS model including a simulation algorithm; thirdly the Bayesian inference approach is presented; fourthly empirical results based on synthetic and real data illustrate practical aspects of the framework. The paper concludes with a discussion and some final remarks.

## 2 Background

We start with a review of the classical spatio-temporal ETAS model, which we will use as a benchmark for comparison.

### 2.1 Classical ETAS model

The ETAS model (Ogata, 1998), describes a stochastic process, which generates point pattern over some domain  $\mathcal{X} \times \mathcal{T} \times \mathcal{M}$ , where  $\mathcal{T} \times \mathcal{X}$  is the time-space window and  $\mathcal{M}$  the mark space of the process. Realisations of this point process are denoted by  $\mathcal{D} = \{(t_i, \mathbf{x}_i, m_i)\}_{i=1}^{N_D}$ , which in seismology can be interpreted as an earthquake catalog consisting of  $N_D$  observed events.  $\mathcal{D}$  is usually ordered in time (time series),  $t_i \in \mathcal{T} \subseteq \mathbb{R}_{>0}$  is the time of the  $i$ th event (time of the earthquake),  $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^2$  is the corresponding location (longitude and latitude of the epicenter), and  $m_i \in \mathcal{M} \subseteq \mathbb{R}$  the corresponding mark (the magnitude of the earthquake).

#### 2.1.1 Interpretations

There are two equivalent interpretations of the ETAS model (Hawkes process). We briefly discuss both.

**Conditional intensity function** One way to define the ETAS model is by a conditional intensity function, which models the infinitesimal rate of expected arrivals around  $(t, \mathbf{x})$  given the history  $H_t = \{(t_i, \mathbf{x}_i, m_i) : t_i < t\}$  of the process until time  $t$ . The earthquake magnitudes  $m_i \geq m_0$  are not influenced by  $H_t$  and are modeled as independent each following an exponential distribution  $p_M(m|\beta) = \beta e^{-\beta(m-m_0)}$ ,  $\beta > 0$ , and  $m_0$  is the magnitude of completeness, (cut-off magnitude) a threshold above which all events are observed (complete data). The conditional ETAS intensity function can be written as (Ogata, 1998) with  $\boldsymbol{\theta} = (\boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\varphi)$

$$\lambda(t, \mathbf{x}|H_t, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\varphi) = \mu(\mathbf{x}|\boldsymbol{\theta}_\mu) + \sum_{i: t_i < t} \varphi(t - t_i, \mathbf{x} - \mathbf{x}_i|m_i, \boldsymbol{\theta}_\varphi), \quad (1)$$

a set of parameters. Here the background intensity  $\mu(\mathbf{x}|\boldsymbol{\theta}_\mu) : \mathbb{R}^2 \rightarrow [0, \infty)$  defines a non-homogeneous Poisson process in space but stationary in time with  $\boldsymbol{\theta}_\mu$  as the required parameters, while  $\varphi(t - t_i, \mathbf{x} - \mathbf{x}_i|m_i, \boldsymbol{\theta}_\varphi) : \mathbb{R}^4 \rightarrow [0, \infty)$  is the triggering function, modeling the rate of aftershocks (self-exciting process) following an event at  $(t_i, \mathbf{x}_i)$  with magnitude  $m_i$ , controlled by the parameters  $\boldsymbol{\theta}_\varphi$ . Specific parametric representations of  $\mu(\cdot)$  and  $\varphi(\cdot)$  for the ETAS model will be discussed in Section 2.1.2.

**Latent branching structure** Another interpretation of a Hawkes process (with the ETAS model being a particular example) is as Poisson cluster- or branching process (Hawkes and Oakes, 1974), leading to the concept of an underlying branching structure, that is, a non observable latent random variable  $z_i$  for each event  $i$ . Events are structured in an ensemble of trees, either having a parent, which is one of the previous events or being spontaneous, called background. The latent variable is typically modelled as taking integer values in a discrete set  $z_i \in \{0, 1, \dots, i-1\}$ , where

$$z_i = \begin{cases} 0 & \text{event } i \text{ is background} \\ j > 0 & \text{event } i \text{ is direct offspring (aftershock) of event } j \text{ at } t_j < t_i. \end{cases}$$

Background events  $z_i = 0$  occur according to a Poisson process with intensity  $\mu(\mathbf{x})$  and form cluster centres, i.e., initial points for branching trees. Within each branching tree, an existing event at  $t_j$  can produce direct offspring at  $t > t_j$  according to an inhomogeneous Poisson process with rate  $\lambda_j(t|t_j, \mathbf{x}_j, m_j) = \varphi(t - t_j, \mathbf{x} - \mathbf{x}_j|m_j, \boldsymbol{\theta}_\varphi)$ . The overall intensity  $\lambda(t, \mathbf{x}|H_t)$  is the sum of all the offspring Poisson processes  $\sum_j \lambda_j$  with  $t_j < t$  and the background Poisson process  $\mu(\mathbf{x})$  (Poisson superposition), as given in (1).

The latent branching structure cannot be observed. However, by its construction (superposition of  $i$  Poisson processes at  $t_i$ ) the probability  $p_{i0} = p(z_i = 0)$  (background event) is (see,

e.g., Zhuang et al., 2002),

$$p_{i0} = \frac{\mu(\mathbf{x}_i | \boldsymbol{\theta}_\mu)}{\lambda(t_i, \mathbf{x}_i | H_{t_i}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\varphi)}, \quad (2)$$

while the probability  $p_{ij} = p(z_i = j)$  (event  $j$  triggered event  $i$ ,  $j > 0$ ) is,

$$p_{ij} = \frac{\varphi(t_i - t_j, \mathbf{x}_i - \mathbf{x}_j | m_j, \boldsymbol{\theta}_\varphi)}{\lambda(t_i, \mathbf{x}_i | H_{t_i}, \boldsymbol{\theta}_\mu, \boldsymbol{\theta}_\varphi)}, \quad (3)$$

with  $p_{i0} + \sum_j p_{ij} = 1$ .

### 2.1.2 Components of the ETAS model

This section sketches the components (background and triggering function) as given in (1).

**Background intensity** The background intensity  $\mu(\mathbf{x})$  is usually modelled either as piecewise constant function over a rectangular grid (or specific polygons, seismo-tectonic units) with  $L$  cells (e.g., in Lombardi (2015); Veen and Schoenberg (2008)),

$$\mu(\mathbf{x} | \boldsymbol{\theta}_\mu) = \mu_l \quad (4)$$

if  $\mathbf{x}$  is in grid cell  $l$ ,  $l = 1, \dots, L$ ; or via a weighted kernel density estimator with variable bandwidth, as suggested by Zhuang et al. (2002),

$$\mu_{\text{kde}}(\mathbf{x}) = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{N_D} p_{i0} k_{d_i}(\mathbf{x} - \mathbf{x}_i). \quad (5)$$

Here,  $|\mathcal{T}|$  is the length of the observational time window,  $p_{i0}$  is the probability that event  $i$  is background as defined in (2),  $d_i = \max\{d_{\min}, r_{i,n_p}\}$  is the variable bandwidth determined for event  $i$  corresponding to the distance  $r_{i,n_p}$  of its number of nearest neighbours  $n_p$ , where  $d_{\min}$  is some minimal bandwidth, and  $k_d(\cdot)$  is an isotropic, bivariate Gaussian kernel function. There are different suggestions to select  $n_p$ ; Zhuang et al. (2002) propose to choose  $n_p$  between 10 and 100, and state that estimated parameters only change slightly if  $n_p$  is changed in the range of 15–100; Zhuang (2011) suggests based on cross-validation experiments, that an optimal  $n_p$  is in the range 3 ~ 6 for Japan. The minimal bandwidth is commonly chosen as  $d_{\min} \in [0.02, 0.05]$  degrees, which is in the range of the localisation error (Zhuang et al., 2002).

Background parameters to be estimated are  $\boldsymbol{\theta}_\mu = (\mu_1, \mu_2, \dots, \mu_L)$  in the first case and the scaled kernel density estimator  $\mu_{\text{kde}}(\mathbf{x})$  given through estimated background probabilities  $\{p_{i0}\}_{i=1}^{N_D}$  in the second case, respectively. For non-parametric models of  $\mu$  as in (5) we neglect the explicit dependency on  $\boldsymbol{\theta}_\mu$  in our notation, but the reader should keep in mind, that in such cases  $\mu$  depends on a varying (potentially infinite) number of parameters.

**Parametric triggering function** The triggering function  $\varphi(t - t_i, \mathbf{x} - \mathbf{x}_i | m_i, \boldsymbol{\theta}_\varphi)$  of the ETAS model is usually a non-negative parametric function, which is separable in space and time, and depends on  $m_i$  and  $\boldsymbol{\theta}_\varphi$ . There are numerous suggested parameterisations. See, for example, Console et al. (2003); Ogata (1998); Ogata and Zhuang (2006); Zhuang et al. (2002). One of the most common parametrisations is provided by

$$\varphi(t - t_i, \mathbf{x} - \mathbf{x}_i | m_i, \boldsymbol{\theta}_\varphi) = \kappa(m_i | K_0, \alpha) g(t - t_i | c, p) s(\mathbf{x} - \mathbf{x}_i | m_i, \boldsymbol{\theta}_s). \quad (6)$$

The first term  $\kappa(\cdot)$  is proportional to the aftershock productivity (or Utsu law, Utsu (1970)) of event  $i$  with  $m_i$ ,

$$\kappa(m_i | K_0, \alpha) = K_0 e^{\alpha(m_i - m_0)}, \quad (7)$$

and  $K_0$  is called productivity coefficient. The second term  $g(\cdot)$  describes the temporal distribution of aftershocks (offspring); a power law decay proportional to the modified Omori Utsu

law (Omori, 1894; Utsu, 1961), and  $t - t_i > 0$  is the elapsed time since the parent event (main shock), that is,

$$g(t - t_i | c, p) = (t - t_i + c)^{-p}. \quad (8)$$

Finally, the third term  $s(\cdot)$  is a probability density function for the spatial distribution of the direct aftershocks (offspring) around the triggering event at  $\mathbf{x}_i$ . Often, one of the following probability density functions are employed. One distinguishes between a short range decay, which uses an isotropic Gaussian distribution with covariance  $d_1^2 e^{\alpha(m_i - m_0)} \mathbf{I}$  (Ogata, 1998; Zhuang et al., 2002); and a long range decay following a Pareto distribution (Kagan, 2002; Ogata and Zhuang, 2006),

$$s(\mathbf{x} - \mathbf{x}_i | m_i, d, \gamma, q) = \frac{q-1}{\pi \sigma_m(m_i)} \left[ 1 + \frac{(\mathbf{x} - \mathbf{x}_i)^\top (\mathbf{x} - \mathbf{x}_i)}{\sigma_m(m_i)} \right]^{-q}, \quad (9)$$

where  $\sigma_m(m_i) = d^2 10^{2\gamma(m_i)}$ .

The unknown parameters to be estimated are  $\boldsymbol{\theta}_\varphi = (K_0, \alpha, c, p, d_1)$ , or  $\boldsymbol{\theta}_\varphi = (K_0, \alpha, c, p, d, \gamma, q)$  depending on which version of  $s(\cdot)$  is used. Note that  $q > 1$  and the rest of the parameters are strictly positive.

### 2.1.3 Parameter estimation via MLE

The likelihood function observing  $\mathcal{D}$  under the spatio-temporal ETAS model is given in (14); it is usually analytically intractable for simple direct optimisation. Numerical optimisation methods (e.g., quasi-Newton methods as in (Ogata, 1988, 1998), using an EM algorithm (Veen and Schoenberg, 2008) or simulated annealing (Lippiello et al., 2014; Lombardi, 2015)) are usually employed. Often the integral term related to the triggering function in (14) using (1) is approximated as  $\int_{\mathcal{T}_i} \int_{\mathcal{X}} \sum_{i:t_i < t} \varphi(t - t_i, \mathbf{x} - \mathbf{x}_i | m_i, \boldsymbol{\theta}_\varphi) d\mathbf{x} dt \leq \int_{\mathcal{T}_i} \int_{\mathbb{R}^2} \sum_{i:t_i < t} \varphi(t - t_i, \mathbf{x} - \mathbf{x}_i | m_i, \boldsymbol{\theta}_\varphi) d\mathbf{x} dt$ , by integrating over  $\mathbb{R}^2$  in space instead of an arbitrary  $\mathcal{X}$  (Schoenberg, 2013). The introduced bias is small and often negligible (Lippiello et al., 2014; Schoenberg, 2013) while the computations are greatly simplified as  $\int_{\mathbb{R}^2} s(\mathbf{x} - \mathbf{x}_i | m_i) d\mathbf{x} = 1$ . We also use this approximation. Computational and numerical details of MLE using (14) are given in Ogata (1998). Instead of directly maximising (14), one can augment the likelihood function by a the latent branching structure  $Z$  and apply an EM algorithm for MLE (Mohler et al., 2011; Veen and Schoenberg, 2008), which is supposed to be advantageous, e.g. regarding stability and convergence (Veen and Schoenberg, 2008).

## 3 Bayesian GP-ETAS model

Our goal is to improve the inference of the spatio-temporal ETAS model in order to allow for comprehensive uncertainty quantification. Despite the availability of powerful MLE based inference methods (see, e.g., Lippiello et al., 2014; Lombardi, 2015; Ogata, 1998; Veen and Schoenberg, 2008), we believe that a Bayesian framework can complement existing methods and will provide a more reliable quantification of uncertainties.

### 3.1 GP-ETAS model specification

We introduce a novel formulation of the spatio-temporal ETAS model, which models the background rate  $\mu(\mathbf{x})$  in a Bayesian *non-parametric* way via a GP (Williams and Rasmussen, 2006), while the triggering function  $\varphi(\cdot)$  assumes still a classical parametric form (modified Omori law (6)). As we will see subsequently, we are able to perform Bayesian inference for this model via Monte Carlo sampling despite its complex form.

While the conditional intensity function of the GP-ETAS model is still given by (1), the background intensity is a priori defined by

$$\mu(\mathbf{x}) = \bar{\lambda} \sigma(f(\mathbf{x})) = \frac{\bar{\lambda}}{1 + e^{-f(\mathbf{x})}}, \quad (10)$$

where  $\sigma(\cdot)$  is the logistic sigmoid function,  $\bar{\lambda}$  a positive scalar, and  $f(\mathbf{x})$  an arbitrary scalar function mapping  $\mathbf{x} \in \mathcal{X}$  to the real line  $\mathbb{R}$ . Since  $\sigma : \mathbb{R} \rightarrow [0, 1]$  the background intensity of the GP-ETAS model is bounded from above by  $\bar{\lambda}$ , i.e.,  $\mu(\mathbf{x}) \in [0, \bar{\lambda}]$  for any  $\mathbf{x} \in \mathcal{X}$ .

For the function  $f(\mathbf{x})$  the GP-ETAS model assumes a Gaussian process prior, which implies that the prior over any discrete set of  $J$  function values  $\mathbf{f} = \{f(\mathbf{x}_i)\}_{i=1}^J$  at positions  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_J\}$  is a  $J$  dimensional Gaussian distribution  $\mathcal{N}(\mathbf{f}|\boldsymbol{\mu}_f, \mathbf{K}_{f,f})$ , where  $\boldsymbol{\mu}_f$  is the prior mean and  $\mathbf{K}_{f,f} \in \mathbb{R}^{J \times J}$  is the covariance matrix between function values at positions  $\mathbf{x}_i$ . The matrix  $\mathbf{K}_{f,f}$  is built from the covariance function (kernel)  $k(\mathbf{x}, \mathbf{x}'|\boldsymbol{\nu})$  such that  $\mathbf{K}_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\nu})$ , where  $\boldsymbol{\nu}$  are hyperparameters. We set  $\boldsymbol{\mu}_f = 0$  and employ a Gaussian covariance function

$$k(\mathbf{x}, \mathbf{x}'|\boldsymbol{\nu}) = \nu_0 \prod_{i=1}^2 e^{-\frac{(\mathbf{x}-\mathbf{x}')^2}{2\nu_i^2}}, \quad (11)$$

where  $\nu_0$  is the so called amplitude and  $(\nu_1, \nu_2)$  are the length scales, representing a distance in input space over which the function values become weakly correlated. Note that the parameter  $\bar{\lambda}$  and the hyperparameters  $\boldsymbol{\nu}$  are also to be inferred from the data. For an in-depth treatment of GPs we refer to [Williams and Rasmussen \(2006\)](#).

The complete specification of the prior model of GP-ETAS including the hyperparameters is now as follows:

$$\boldsymbol{\nu} \sim p_{\boldsymbol{\nu}}, \text{ a prior on } \boldsymbol{\nu} \text{ (exponential distribution)} \quad (12a)$$

$$f \sim \mathcal{GP} \text{ prior with zero mean and a covariance function} \quad (12b)$$

$$\bar{\lambda} \sim p_{\bar{\lambda}}, \text{ a prior on } \bar{\lambda} \text{ (gamma distribution)} \quad (12c)$$

$$\mu | \bar{\lambda}, f, \boldsymbol{\nu} \sim \text{prior model on } \mu \text{ as defined in (10)} \quad (12d)$$

$$\boldsymbol{\theta}_\varphi \sim p_{\boldsymbol{\theta}_\varphi}, \text{ a prior on } \boldsymbol{\theta}_\varphi \text{ of a triggering function } \varphi(\cdot) \text{ (uniform distribution)} \quad (12e)$$

The corresponding observational model is

$$\mathcal{D} | \mu, \boldsymbol{\theta}_\varphi \sim \text{Hawkes process with ETAS intensity function given by GP-ETAS in (1), (10), (13)}$$

where  $\mathcal{D}$  is the data. Note that some quantities are independent by construction, e.g.,  $\boldsymbol{\nu}$  and  $\bar{\lambda}$ ,  $f$  and  $\bar{\lambda}$ .

Without the triggering function in the intensity function (1) the GP-ETAS model would be equivalent to the SGCP model which is used to describe an inhomogeneous Poisson process ([Adams et al., 2009](#)) because of its favourable statistical properties ([Kirichenko and Van Zanten, 2015](#)).

In the following we sketch how to generate data from the GP-ETAS model. A full description of the Bayesian inference problem is provided in Section 4.

### 3.2 Simulating the GP-ETAS model

Data  $\mathcal{D} = \{(t_i, \mathbf{x}_i, m_i)\}_{i=1}^{N_{\mathcal{D}}}$  can be easily simulated from the GP-ETAS model using the latent branching structure of the point process. We propose a procedure which consists of two parts:

1. Generate all *background* events  $\mathcal{D}_0 = \{(t_i, \mathbf{x}_i, m_i, z_i = 0)\}_{i=1}^{N_{\mathcal{D}_0}}$  from a SGCP in equation (10) as explained in [Adams et al. \(2009\)](#).
2. Sample all *aftershock* events (offspring) given  $\mathcal{D}_0$  in possibly several generations denoted as  $\mathcal{D}_\varphi = \{(t_i, \mathbf{x}_i, m_i, z_i \neq 0)\}_{i=1}^{N_{\mathcal{D}_\varphi}}$  and add them to obtain  $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_\varphi$ .

The above procedure can be implemented based on the *thinning* algorithm ([Lewis and Shedler, 1976](#)); a variant of rejection sampling for point processes.

After choosing  $\bar{\lambda}$ ,  $\boldsymbol{\nu}$ ,  $\boldsymbol{\theta}_\varphi$  and a mark distribution  $p(m)$ , the simulation procedure of  $\mathcal{D} \in \mathcal{X} \times \mathcal{T} \times \mathcal{M}$  can be summarised as follows: *First part*: One uses the upper bound  $\bar{\lambda}$  to generate positions  $\{\mathbf{x}_j\}_{j=1}^J$  of events from a homogeneous Poisson process with mean  $|\mathcal{X}| |\mathcal{T}| \bar{\lambda}$  which provide candidate background events (Figure 1 a). Subsequently a Gaussian process  $\mathbf{f}$  is sampled

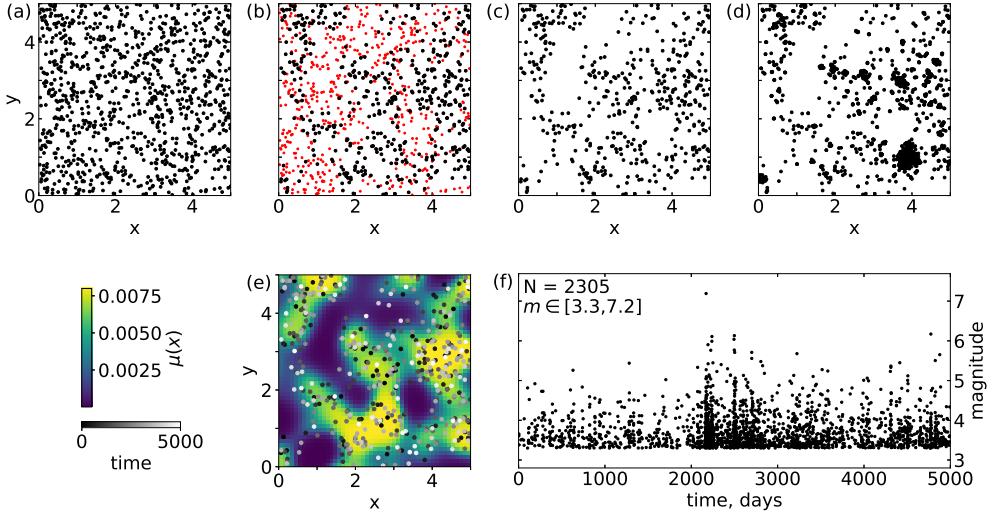


Figure 1: The Figure depicts the different steps of a forward simulation of the generative GP-ETAS model. (a) Events of a homogeneous Poisson process with intensity  $\bar{\lambda}$  are generated ( $\bar{\lambda} = 0.008, N = 988$ ). (b) One retains events according to an inhomogeneous Poisson process with the desired intensity  $\mu(\mathbf{x}) = \bar{\lambda}\sigma(f(\mathbf{x}))$  by randomly deleting events (red dots) via *thinning*. (c) The background events (black dots from (b)) are denoted by  $\mathcal{D}_0$  ( $N_{\mathcal{D}_0} = 481$ ), (d) After adding aftershocks (offspring events)  $\mathcal{D}_\varphi$  to  $\mathcal{D}_0$  in accordance with the triggering function  $\varphi(\cdot)$  one obtains finally the simulated data  $\mathcal{D}$  ( $N_{\mathcal{D}} = 2305$ ) of the spatio-temporal GP-ETAS. (e) Shows the background intensity  $\mu(\mathbf{x}) = \bar{\lambda}\sigma(f(\mathbf{x}))$  together with the generated background events. Gray scaling of the dots refers to the event times. (f) Depicts the simulated data as a synthetic earthquake catalogue in time.

from the prior  $\mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_{\mathbf{f}, \mathbf{f}})$  based on  $\{\mathbf{x}_j\}_{j=1}^J$  using (11). The values  $\mu(\mathbf{x}_j)$  can be computed using (10). Afterwards events, which do not follow an inhomogeneous Poisson process with intensity  $\mu(\mathbf{x})$  as given by (10), are randomly deleted via *thinning* (Figure 1 b). The remaining  $N_{\mathcal{D}_0}$  events are background events (Figure 1 c). The event times  $\{t_i\}_{i=1}^{N_{\mathcal{D}_0}}$  are sampled from a uniform distribution  $\mathcal{U}(|\mathcal{T}|)$  and the marks  $\{m_i\}_{i=1}^{N_{\mathcal{D}_0}}$  from an exponential distribution, e.g., Gutenberg-Richter relation. Finally one obtains  $\mathcal{D}_0$ . *Second part:* Given the background events  $\mathcal{D}_0$ , the aftershock events (offsprings) of all generations are added to  $\mathcal{D}_0$  in accordance with the triggering function  $\varphi(\cdot)$  and using the mark distribution which yields  $\mathcal{D}$  (Figure 1 d).

The overall simulation algorithm is described in detail in the Appendix A, and is visualised in Figure 1.

## 4 Bayesian inference

In this section, we address the Bayesian inference problem of our spatio-temporal GP-ETAS model. The objective is to estimate the random conditional intensity function (1) in a Bayesian way including uncertainties, i.e. the joint posterior  $p(\mu, \boldsymbol{\theta}_\varphi | \mathcal{D})$ , where  $\mu$  denotes the entire random field of the background intensity as in (12d).

The likelihood of observing a point pattern  $\mathcal{D} = \{(t_i, \mathbf{x}_i, m_i)\}_{i=1}^{N_{\mathcal{D}}}$  under the GP-ETAS model (10) is given by the point process likelihood

$$p(\mathcal{D} | \mu, \boldsymbol{\theta}_\varphi) = \prod_{i=1}^{N_{\mathcal{D}}} \lambda(t_i, \mathbf{x}_i | \mu(\mathbf{x}_i), \boldsymbol{\theta}_\varphi) \exp \left( - \int_{\mathcal{T}} \int_{\mathcal{X}} \lambda(t, \mathbf{x} | \mu(\mathbf{x}), \boldsymbol{\theta}_\varphi) d\mathbf{x} dt \right), \quad (14)$$

where the intensity  $\lambda(\cdot)$  is given by (10), and the dependencies on  $H_t$ ,  $H_{t_i}$  are omitted for notational convenience.

Assuming a joint prior distribution denoted here by  $p(\mu, \boldsymbol{\theta}_\varphi)$  for simplicity, the posterior

distribution becomes

$$p(\mu, \boldsymbol{\theta}_\varphi | \mathcal{D}) \propto p(\mathcal{D} | \mu, \boldsymbol{\theta}_\varphi) p(\mu, \boldsymbol{\theta}_\varphi). \quad (15)$$

This posterior is intractable in practice and hence standard inference techniques are not directly applicable. More precisely, the following three main challenges arise:

- (i) The background intensity  $\mu$  and triggering function  $\varphi(\cdot | \boldsymbol{\theta}_\varphi)$  cannot be treated separately in the likelihood function (14).
- (ii) The likelihood (14) includes an intractable integral inside the exponential term. Furthermore, normalisation of (15) requires an intractable marginalisation over  $\mu$  and  $\boldsymbol{\theta}_\varphi$ . Thus, the posterior distribution is *doubly intractable* (Murray et al., 2006).
- (iii) We assume a Gaussian process prior for modelling the background rate. However, the point process likelihood (14) is non-Gaussian, which makes the functional form of the posterior nontrivial to treat in practice.

We approach these challenges by data augmentation based on the work of Adams et al. (2009); Donner and Opper (2018); Polson et al. (2013). We will find that this augmentation simplifies the inference problem substantially. The following three auxiliary random variables are introduced:

- (1.) A *latent branching structure*  $Z$ , as described in Section 2.1.1, decouples  $\mu$  and  $\boldsymbol{\theta}_\varphi$  in the likelihood function (e.g., Veen and Schoenberg, 2008).
- (2.) A *latent Poisson process*  $\Pi$  enables an unbiased estimation of the integral term in the likelihood function that depends on  $\mu$ .
- (3.) We make use of the fact, that the logistic sigmoid function can be written as an infinite scale mixture of Gaussians using latent *Pólya–gamma random variables*  $\boldsymbol{\omega} \sim p_{\text{PG}}(\omega)$  (Polson et al., 2013), defined in Appendix B. This leads to a likelihood representation, which is conditional conjugate to all the priors including the Gaussian process prior for the background component of the likelihood function (Donner and Opper, 2018).

These three augmentations allow one to implement a Gibbs sampling procedure (Geman and Geman, 1984) that produces samples from the posterior distribution in (15). More precisely, random samples are generated in a Gibbs sampler by drawing one variable (or a block of variables) from the conditional posterior given all the other variables. Hence, we need to derive the required conditional posterior distributions as outlined next.

The suggested sampler consists of three modules using the solutions (data augmentations) sketched above: *sampling the latent branching structure*, *inference of the background*  $\mu$ , and *inference of the triggering*  $\boldsymbol{\theta}_\varphi$ . Our overall *Gibbs sampling algorithm* of the posterior distribution is summarised in Algorithm 1. After an initial burn-in (a sufficiently long run of the three modules (Section 4.1 – 4.3), the generated samples converge to the desired joint posterior distribution  $p(\mu, \boldsymbol{\theta}_\varphi | \mathcal{D})$ .

In the following, we discuss some important aspects of the three modules of the Gibbs sampler which the sampler runs repeatedly through.

## 4.1 Sampling the latent branching structure

**Augmentation by the latent branching structure.** We consider an auxiliary variable  $z_i$  for each data point  $i$ , which represents the latent branching structure as defined in Section 2.1.1. Recall that it gives the time index of the parent event. If  $z_i = 0$  then the event is a spontaneous background event. The likelihood  $p(\mathcal{D}, Z | \mu, \boldsymbol{\theta}_\varphi)$  of the augmented model can be written as

$$\begin{aligned} p(\mathcal{D}, Z | \mu, \boldsymbol{\theta}_\varphi) &= \underbrace{\prod_{i=1}^{N_D} \mu(\mathbf{x}_i)^{\mathbb{I}(z_i=0)} \exp \left( -|\mathcal{T}| \int_{\mathcal{X}} \mu(\mathbf{x}) d\mathbf{x} \right)}_{(a)=p(\mathcal{D}_0 | Z, \mu)} \\ &\times \underbrace{\prod_{i=1}^{N_D} \prod_{j=1}^{i-1} \varphi_{ij}(\boldsymbol{\theta}_\varphi)^{\mathbb{I}(z_i=j)} \exp \left( - \int_{\mathcal{T}_i} \int_{\mathcal{X}} \varphi_i(\boldsymbol{\theta}_\varphi) d\mathbf{x} dt \right) p(Z)}_{(b)=p(\mathcal{D} | Z, \boldsymbol{\theta}_\varphi)}, \end{aligned} \quad (16)$$

where  $\mathbb{I}(\cdot)$  denotes the indicator function, i.e.,  $\mathbb{I}(z_i = j)$  takes the value 1 for all  $z_i = j$  and 0 otherwise,  $\varphi_{ij}(\boldsymbol{\theta}_\varphi) = \varphi(t_i - t_j, \mathbf{x}_i - \mathbf{x}_j | m_j, \boldsymbol{\theta}_\varphi)$ ,  $\varphi_i(\boldsymbol{\theta}_\varphi) = \varphi(t - t_i, \mathbf{x} - \mathbf{x}_i | m_i, \boldsymbol{\theta}_\varphi)$ ,  $\mathcal{T}_i = [t_i, |\mathcal{T}|] \subset \mathcal{T}$ , and all possible branching structures are equally likely, i.e.  $p(Z) = \text{const.}$  Furthermore,  $\mathcal{D}_0 = \{\mathbf{x}_i\}_{i:z_i=0}$  denotes the set of  $N_{\mathcal{D}_0}$  background events. Note, that marginalizing over  $Z$  in (16) recovers (14), because  $\sum_{z_i=0}^{i-1} \mu(\mathbf{x}_i)^{\mathbb{I}(z_i=0)} \prod_{j=1}^{i-1} \varphi_{ij}(\boldsymbol{\theta}_\varphi)^{\mathbb{I}(z_i=j)} = \lambda(t_i, \mathbf{x}_i | \mu(\mathbf{x}_i), \boldsymbol{\theta}_\phi)$ . The augmented likelihood factorises into two independent components, (a) a likelihood component for the background intensity which depends on  $\mu$  (first two terms on the rhs of (16)) and (b) a likelihood component of the triggering function which depends on  $\boldsymbol{\theta}_\varphi$  (last two terms on the rhs of (16)).

From (16) one can derive the conditional distribution of  $z_i$  given all the other variables. Note that all  $z_i$ 's are independent. The conditional distribution is proportional to a categorical distribution,

$$p(z_i | \mathcal{D}, \mu(\mathbf{x}_i), \boldsymbol{\theta}_\varphi) \propto [\mu(\mathbf{x}_i)]^{\mathbb{I}(z_i=0)} \prod_{j=1}^{i-1} [\varphi_{ij}(\boldsymbol{\theta}_\varphi)]^{\mathbb{I}(z_i=j)} = \prod_{j=0}^{i-1} p_{ij}^{\mathbb{I}(z_i=j)}, \quad (17)$$

with the probabilities  $p_{ij}$  given by (2) and (3) and which we collect in a vector  $\mathbf{p}_i \in \mathbb{R}^i$ .

From (17) one can see that the latent branching structure at the  $k$ th iteration of the Gibbs sampler is sampled from a categorical distribution,

$$\forall i = 1, \dots, N_{\mathcal{D}} \quad z_i^{(k)} | \mathcal{D}, (\mu(\mathbf{x}_i), \boldsymbol{\theta}_\varphi)^{(k-1)} \sim \text{Categorical}(\mathbf{p}_i). \quad (18)$$

Here  $(\mu(\mathbf{x}_i), \boldsymbol{\theta}_\varphi)^{(k-1)}$  denotes the values of  $\mu(\mathbf{x}_i)$  and  $\boldsymbol{\theta}_\varphi$  from the previous iteration.

## 4.2 Inference for the background intensity

Given an instance of a branching structure  $Z$ , the background intensity in (16) depends on events  $i$  for which  $z_i = 0$  only. One finds that the resulting term is a Poisson likelihood of the form

$$p(\mathcal{D}_0 | f, \bar{\lambda}, Z) = \prod_{i=1:z_i=0}^{N_{\mathcal{D}}} \bar{\lambda} \sigma(f_i) \exp \left( -|\mathcal{T}| \int_{\mathcal{X}} \bar{\lambda} \sigma(f(\mathbf{x})) d\mathbf{x} \right), \quad (19)$$

where  $\mu(\mathbf{x})$  has been replaced by (10) and  $f_i = f(\mathbf{x}_i)$  has been used for notational convenience.

Because of the aforementioned problems in Section 4, sampling the conditional posterior  $p(f, \bar{\lambda} | \mathcal{D}_0, Z)$  is still non trivial and require further augmentations which we describe next.

**Augmentation by a latent Poisson process.** We can resolve issue (ii) from Section 4 by introducing an independent latent Poisson process  $\Pi = \{\mathbf{x}_l\}_{l=N_{\mathcal{D}}+1}^{N_{\mathcal{D}\cup\Pi}}$  on the data space with rate  $\hat{\lambda}(\mathbf{x}) = \bar{\lambda}(1 - \sigma(f(\mathbf{x}))) = \bar{\lambda}(\sigma(-f(\mathbf{x})))$  using  $1 - \sigma(z) = \sigma(-z)$ . The points in  $\mathcal{D}$ ,  $\Pi$  form the joint set  $\mathcal{D} \cup \Pi$  with cardinality  $N_{\mathcal{D}\cup\Pi}$ . Note, that the number of elements in  $\Pi$ , i.e.  $N_\Pi$ , is also a random variable. The joint likelihood of  $\mathcal{D}_0$  and the new random variable  $\Pi$  is,

$$p(\mathcal{D}_0, \Pi | f, \bar{\lambda}, Z) = \prod_{i=1:z_i=0}^{N_{\mathcal{D}}} \bar{\lambda} \sigma(f_i) \prod_{l=N_{\mathcal{D}}+1}^{N_{\mathcal{D}\cup\Pi}} \bar{\lambda} \sigma(-f_l) \exp(-|\mathcal{X}| |\mathcal{T}| \bar{\lambda}), \quad (20)$$

where  $f_l = f(\mathbf{x}_l)$ . Thus, by introducing the latent Poisson process  $\Pi$ , we obtain a likelihood representation of the augmented system, where the former intractable integral inside the exponential term disappears, i.e. reduces to a constant.

Let us provide further intuition for this augmentation. Due to the fact that  $\bar{\lambda} \sigma(f(\mathbf{x}))$  is bounded, one can simply use the superposition property of Poisson processes to construct the latent  $\Pi$ . The basic idea is to introduce an independent  $\Pi$  with bounded intensity  $\hat{\lambda}(\mathbf{x})$  such that its superposition with the inhomogeneous Poisson process of the background results in a homogeneous Poisson process with intensity  $\bar{\lambda}$ . Hence,  $\hat{\lambda}(\mathbf{x}) + \bar{\lambda} \sigma(f(\mathbf{x})) = \bar{\lambda}$ , and one obtains  $\hat{\lambda}(\mathbf{x}) = \bar{\lambda}(1 - \sigma(f(\mathbf{x}))) = \bar{\lambda} \sigma(-f(\mathbf{x}))$  for the intensity of the latent Poisson process. Writing the joint likelihood for events in  $i$  with  $z_i = 0$  and events from the latent  $\Pi$  we get (20).

More rigorously one can derive the latent Poisson process  $\Pi$  following Donner and Opper (2018). Note that (19) implies

$$\exp\left(-|\mathcal{T}|\int_{\mathcal{X}} \bar{\lambda}\sigma(f(\mathbf{x}))d\mathbf{x}\right) = \exp\left(\int_{\mathcal{T}}\int_{\mathcal{X}} \bar{\lambda}(\sigma(-f(\mathbf{x})) - 1)d\mathbf{x}dt\right) = \mathbb{E}_{\bar{\lambda}}\left[\prod_{\mathbf{x}_l \in \Pi} \sigma(-f(\mathbf{x}_l))\right], \quad (21)$$

where the expectation is over random sets  $\Pi$  with respect to a Poisson process measure with rate  $\bar{\lambda}$  on the space-time window of the data  $\mathcal{T} \times \mathcal{X}$ . Here, one uses Campbell's theorem (Kingman, 1993). Writing the likelihood parts depending on  $f$  and  $\bar{\lambda}$  in (16) in terms of the new random variable  $\Pi$  we get (20). Note that marginalisation over the augmented variable  $\Pi$  leads back to the background likelihood in (19) conditioned on the branching structure  $Z$ .

**Augmentation by Pólya–gamma random variables.** In order to resolve issue (iii) from Section 4 we substitute the sigmoid function by an infinite scale mixture of Gaussians using latent Pólya–gamma random variables (Polson et al., 2013), that is,

$$\sigma(z) = \frac{e^{\frac{z}{2}}}{2 \cosh(\frac{z}{2})} = \frac{1}{2} e^{\frac{z}{2}} \int_0^\infty e^{-\frac{z^2}{2}\omega} p_{PG}(\omega|1,0)d\omega, \quad (22)$$

where the new random variable  $\omega$  is distributed according to the Pólya–gamma density  $p_{PG}(\omega|1,0)$ , see Appendix B. Inserting the Pólya–gamma representation of the sigmoid function (22) into (20) yields

$$\begin{aligned} p(\mathcal{D}_0, \boldsymbol{\omega}_{\mathcal{D}}, \Pi, \boldsymbol{\omega}_{\Pi} | \mathbf{f}, \bar{\lambda}, Z) &= \prod_{i:z_i=0}^{N_{\mathcal{D}}} \frac{\bar{\lambda}}{2} e^{\frac{f_i}{2} - \frac{f_i^2}{2}\omega_i} p_{PG}(\omega_i|1,0) \\ &\times \prod_{l=N_{\mathcal{D}}+1}^{N_{\mathcal{D} \cup \Pi}} \frac{\bar{\lambda}}{2} e^{-\frac{f_l}{2} - \frac{f_l^2}{2}\omega_l} p_{PG}(\omega_l|1,0) \exp(-\bar{\lambda}|\mathcal{X}|T), \end{aligned} \quad (23)$$

where we set the Pólya-gamma variables of all events  $\boldsymbol{\omega}_{\mathcal{D}} = (\omega_1, \dots, \omega_{N_{\mathcal{D}}})$  to  $\omega_i = 0$  if  $z_i \neq 0$ . For the latent Poisson process the Pólya-gamma variables are denoted by  $\boldsymbol{\omega}_{\Pi} = (\omega_{N_{\mathcal{D}}+1}, \dots, \omega_{N_{\mathcal{D} \cup \Pi}})$ . The likelihood representation of the augmented system (23) has a Gaussian form with respect to  $\mathbf{f}$  (that is, only linear or quadratic terms of  $\mathbf{f}$  appear in the exponential function) and is therefore conditionally conjugate to the GP prior denoted by  $p(\mathbf{f})$ . Hence, we can implement an efficient Gibbs sampler for the background intensity function.

Employing a Gaussian process prior over  $\mathbf{f}$  and a Gamma distributed prior over  $\bar{\lambda}$ , one gets from (23) the following conditional posteriors for the  $k$ th Gibbs iteration:

$$\Pi^{(k)} | (\bar{\lambda}, \mathbf{f})^{(k-1)} \sim \text{PP}(\bar{\lambda}(\sigma(-f(\mathbf{x})))) \quad (24a)$$

$$\forall l : N_{\mathcal{D}} + 1, \dots, N_{\mathcal{D} \cup \Pi} \quad \omega_l^{(k)} | f_l^{(k-1)}, \Pi^{(k)} \sim p_{PG}(1, |f_l|) \quad (24b)$$

$$\forall i : z_i = 0 \quad \omega_i^{(k)} | f_i^{(k-1)}, \mathcal{D}, Z^{(k)} \sim p_{PG}(1, |f_i|) \quad (24c)$$

$$\bar{\lambda}^{(k)} | Z^{(k)}, \Pi^{(k)} \sim \text{Gamma}(N_{\mathcal{D}_0 \cup \Pi} + \alpha_0, |\mathcal{X}||\mathcal{T}| + \beta_0) \quad (24d)$$

$$\mathbf{f}^{(k)} | \mathcal{D}, (\boldsymbol{\omega}_{\mathcal{D}}, \Pi, \boldsymbol{\omega}_{\Pi}, Z)^{(k)} \sim \mathcal{N}((\boldsymbol{\Omega} + \mathbf{K}^{-1})^{-1}\mathbf{u}, (\boldsymbol{\Omega} + \mathbf{K}^{-1})^{-1}) \quad (24e)$$

where  $\mathbf{f} = (\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\Pi}) \in \mathbb{R}^{N_{\mathcal{D} \cup \Pi}}$  is the Gaussian process at the data locations  $\mathcal{D}$  and  $\Pi$ ; and  $\text{PP}(\cdot)$  denotes an inhomogeneous Poisson process with intensity  $\bar{\lambda}(\sigma(-f(\mathbf{x})))$ ;  $\boldsymbol{\Omega}$  is a diagonal matrix with  $(\boldsymbol{\omega}_{\mathcal{D}}, \boldsymbol{\omega}_{\Pi})$  as diagonal entries.  $\mathbf{K} \in \mathbb{R}^{N_{\mathcal{D} \cup \Pi} \times N_{\mathcal{D} \cup \Pi}}$  is the covariance matrix of the Gaussian process prior at positions  $\mathcal{D}$  and  $\Pi^{(k)}$ . It can be shown that, the vector  $\mathbf{u}$  is 1/2 for all entries in  $\mathcal{D}_0$ , zero for all entries of the remaining data  $\mathcal{D} \setminus \mathcal{D}_0$ , and -1/2 for the corresponding entries of  $\Pi$ .  $\text{Gamma}(\cdot)$  is a Gamma distribution, where the Gamma prior has shape and rate parameters  $\alpha_0, \beta_0$ . We used  $e^{-\frac{c^2}{2}\omega} p_{PG}(\omega|1,0) \propto p_{PG}(\omega|1,c)$  due to the definition of a tilted Pólya–gamma density (32) as given in (Polson et al., 2013), see Appendix B. Note that one does not need an explicit form of the Pólya–gamma density for our inference approach since it is sampling based. In other words, we only need an efficient way to sample from the tilted  $p_{PG}$  density (32) which was provided by Polson et al. (2013); Windle et al. (2014). Several  $p_{PG}$  samplers are freely available for different computer languages.

A detailed step-by-step derivation of the conditional distributions is given in the Appendix C.

**Hyperparameters** The Gaussian process covariance kernel given in (11) depends on the hyperparameters  $\boldsymbol{\nu}$ . Compare Section 3. We use exponentially distributed priors on  $p(\nu_i) = p_{\nu_i}$ , and we sample  $\boldsymbol{\nu}$  using a standard MH algorithm as there is no closed form for the conditional posterior available. The only terms where  $\boldsymbol{\nu}$  enter are in the Gaussian process prior and hence the relevant terms are

$$\ln p(\boldsymbol{\nu} | \mathbf{f}, \mathcal{D}, \Pi, \boldsymbol{\omega}_{\mathcal{D}}, \boldsymbol{\omega}_{\Pi}) = -\frac{1}{2} \mathbf{f}^\top \mathbf{K}_{\boldsymbol{\nu}}^{-1} \mathbf{f} - \frac{1}{2} \ln \det \mathbf{K}_{\boldsymbol{\nu}} + \ln p(\boldsymbol{\nu}) + \text{const.}, \quad (25)$$

where  $\mathbf{K}_{\boldsymbol{\nu}}$  is the Gaussian process prior covariance matrix depending on  $\boldsymbol{\nu}$  via (11).

#### 4.2.1 Conditional predictive posterior distribution of the background intensity

Given the  $k$ th posterior sample  $(\bar{\lambda}^{(k)}, \mathbf{f}^{(k)}, \boldsymbol{\nu}^{(k)})$ , the background intensity  $\mu(\mathbf{x}^*)^{(k)}$  at any set of positions  $\{\mathbf{x}_i^*\} \in \mathcal{X}$  (predictive conditional posterior) can be obtained in the following way, see (12d). Conditioned on  $\mathbf{f}^{(k)}$  and hyperparameters  $\boldsymbol{\nu}^{(k)}$  the latent function values  $\mathbf{f}^*$  can be sampled via the conditional prior  $p(\mathbf{f}^* | \mathbf{f}^{(k)}, \boldsymbol{\nu}^{(k)})$  using (41) with covariance function given in (11) (Williams and Rasmussen, 2006). Using (10) one gets  $\mu(\mathbf{x}^*)^{(k)} = \bar{\lambda}^{(k)} \sigma(\mathbf{f}^*)$ .

### 4.3 Inference for the parameters of the triggering function

Given an instance of a branching structure  $Z$ , the likelihood function in (16) factorises in terms involving  $\mu$  and terms involving  $\boldsymbol{\theta}_{\varphi}$ . The relevant terms related to  $\boldsymbol{\theta}_{\varphi}$  are

$$p(\mathcal{D}|Z, \boldsymbol{\theta}_{\varphi}) = \prod_{i=1:z_i \neq 0}^{N_{\mathcal{D}}} \varphi(t_i - t_{z_i}, \mathbf{x}_i - \mathbf{x}_{z_i} | m_{z_i}, \boldsymbol{\theta}_{\varphi}) \\ \times \prod_{i=1}^{N_{\mathcal{D}}} \exp \left( - \int_{\mathcal{T}_i} \int_{\mathcal{X}} \varphi(t - t_i, \mathbf{x} - \mathbf{x}_i | m_i, \boldsymbol{\theta}_{\varphi}) d\mathbf{x} dt \right). \quad (26)$$

The conditional posterior  $p(\boldsymbol{\theta}_{\varphi} | \mathcal{D}, Z) \propto p(\mathcal{D}|Z, \boldsymbol{\theta}_{\varphi})p(\boldsymbol{\theta}_{\varphi})$  with prior  $p(\boldsymbol{\theta}_{\varphi})$  has no closed form. The dimension of  $\boldsymbol{\theta}_{\varphi}$  is usually small ( $\leq 7$ ). We employ MH sampling (Hastings, 1970), which can be considered a nested step within the overall Gibbs sampler. We use a random walk MH where proposals are generated by a Gaussian in log space. The acceptance probability of  $\boldsymbol{\theta}_{\varphi}^{(k)}$  based on (26) is given by

$$p_{\text{accept}} = \min \left\{ 1, \frac{p(\mathcal{D}|Z^{(k)}, \boldsymbol{\theta}_{\varphi}^{\text{proposed}})p(\boldsymbol{\theta}_{\varphi}^{\text{proposed}})}{p(\mathcal{D}|Z^{(k)}, \boldsymbol{\theta}_{\varphi}^{(k-1)})p(\boldsymbol{\theta}_{\varphi}^{(k-1)})} \right\}. \quad (27)$$

We take 10 proposals before we return to the overall Gibbs sampler, that is, to step in Section 4.1.

## 5 Experiments and results

We consider two kinds of experiments where we evaluate the performance of our proposed Bayesian approach GP-ETAS (see Section 3 and 4). First we look at synthetic data, with known conditional intensity  $\lambda(t, \mathbf{x})$ , i.e. with known background intensity  $\mu(\mathbf{x})$  and known parameters  $\boldsymbol{\theta}_{\varphi}$  of the triggering function. Here, we investigate if GP-ETAS can recover the model underlying the data well. Secondly, we apply our method to observational earthquake data.

**Comparison:** We compare our approach with the current standard spatio-temporal ETAS model which uses MLE. This classical ETAS model is based on kernel density estimation with variable bandwidths for the background intensity  $\mu(\mathbf{x})$  as described in Section 2.1.2. Two variations are considered: (1) ETAS model with standard choice of the minimal bandwidth (0.05 degrees) and  $n_p = 15$  the number of nearest neighbors used for obtaining the individual bandwidths (ETAS-classical; Zhuang et al. (2002)), and (2) ETAS model with a minimal bandwidth given by Silverman's rule (Silverman, 1986) and  $n_p = 15$  (ETAS-Silverman).

**Evaluation metrics:** Two metrics are used to evaluate the performances. The first metric is the test likelihood, which evaluates the likelihood (14) for a data test set  $\mathcal{D}^*$  (unseen

---

**Algorithm 1** Gibbs Sampler for the posterior distribution of spatio-temporal GP-ETAS

---

```

1: Initialise randomly  $\bar{\lambda}^{(0)}, \mathbf{f}^{(0)}, \boldsymbol{\theta}_\varphi^{(0)}$  from the priors
2: for  $k = 1$  to  $K$  do
3:   Factorisation of the likelihood (Section 4.1)
4:    $\forall i = 1, \dots, N_D$  Sample latent branching structure  $z_i^{(k)} | \mathcal{D}, (\mu(\mathbf{x}_i), \boldsymbol{\theta}_\varphi)^{(k-1)}$  (18)
5:   Inference of the background intensity  $p(\bar{\lambda}, \mathbf{f} | \mathcal{D}_0, \boldsymbol{\omega}_\mathcal{D}, \Pi, \boldsymbol{\omega}_\Pi, Z)$  (Section 4.2)
6:   Sample latent Poisson process  $\Pi^{(k)} | (\bar{\lambda}, \mathbf{f})^{(k-1)}$  (24a)
7:    $\forall l = N_D + 1, \dots, N_{\mathcal{D} \cup \Pi}$  Sample Pólya–gamma variables  $\omega_l^{(k)} | f_l^{(k-1)}, \Pi^{(k)}$  (24b)
8:    $\forall i : z_i = 0$  Sample Pólya–gamma variables  $\omega_i^{(k)} | \mathcal{D}, f_i^{(k-1)}, Z^{(k)}$  (24c)
9:   Set  $\omega_i^{(k)} = 0$  otherwise.
10:  Sample upper bound  $\bar{\lambda}^{(k)} | (Z, \Pi)^{(k)}$  (24d)
11:  Sample Gaussian process  $\mathbf{f}^{(k)} | \mathcal{D}, (\boldsymbol{\omega}_\mathcal{D}, \Pi, \boldsymbol{\omega}_\Pi, Z)^{(k)}$  (24e)
12:  Sample hyperparameters  $\boldsymbol{\nu}^{(k)}$  using MH step (25)
13:   $\forall i = 1, \dots, N_D$  compute  $(\mu(\mathbf{x}_i))^{(k)}$  (10)
14:  Inference of the triggering function  $p(\boldsymbol{\theta}_\varphi | \mathcal{D}, Z^{(k)})$  (Section 4.3)
15:  Sample  $\boldsymbol{\theta}_\varphi^{(k)}$  using MH steps (27)
16: end for

```

---

data during the inference) given the inferred model on training data  $\mathcal{D}$ , which is  $p(\mathcal{D}^* | \mathcal{D}) = \mathbb{E}_{p(\mu, \boldsymbol{\theta}_\varphi | \mathcal{D})} [p(\mathcal{D}^* | \mu, \boldsymbol{\theta}_\varphi)]$ , where the expectation is over the inferred model posterior. The test likelihood reflects the predictive power of the different modelling approaches. In the case of GP-ETAS we obtain  $K$  posterior samples  $\{(\mu^k, \boldsymbol{\theta}_\varphi^k)\}_{k=1}^K$  and we evaluate the log expected test likelihood,  $\ell_{\text{test}} = \ln p(\mathcal{D}^* | \mathcal{D}) \approx \ln \frac{1}{K} \sum_{k=1}^K p(\mathcal{D}^* | \mu^{(k)}, \boldsymbol{\theta}^{(k)})$ . In the case of ETAS–classical and ETAS–Silverman we use the MLE point estimate for evaluating  $\ell_{\text{test}}$ . The involved spatial integral in (14) is approximated by Riemann sums on a  $50 \times 50$  point grid. The second metric is the  $\ell_2$  norm between true background intensity  $\mu$  and the predicted  $\hat{\mu}$ ,  $\ell_2 = \sqrt{\int_{\mathcal{X}} (\mu(\mathbf{x}) - \hat{\mu}(\mathbf{x}))^2 d\mathbf{x}}$ . This is only possible for the experiments with synthetic data.

## 5.1 Synthetic data

**General experimental setups:** We simulate synthetic data from two different conditional intensity functions which differ in  $\mu(\mathbf{x})$  on a spatio-temporal domain  $\mathcal{X} \times \mathcal{T} = [0, 5] \times [0, 5] \times [0, 5000]$ . In the first case we consider  $\mu_1(\mathbf{x})$  to be constant over large spatial regions, e.g. large area sources (*Case 1*, Figure 2 first row). In a second experiment we consider another particular setting where  $\mu_2(\mathbf{x})$  is concentrated mainly on small fault-type areas (*Case 2*, Figure 2 second row). These two settings (area sources and faults) are important, typical limiting cases in analysing seismicity pattern, both used in seismic hazard assessment. The two chosen intensity functions are,

$$\mu_1(\mathbf{x}) = \begin{cases} 0.005 & \mathbf{x} \in [0, 3] \times [1.5, 5] \\ 0.001 & \mathbf{x} \in [3, 5] \times [1.5, 5] \\ 0.0005 & \mathbf{x} \in [0, 5] \times [0, 1.5] \end{cases} \quad (28)$$

$$\mu_2(\mathbf{x}) = \begin{cases} 0.07035 & \mathbf{x} \in [1, 3] \times [1.4, 1.5] \\ 0.07035 & \mathbf{x} \in [1, 4] \times [2.4, 2.5] \\ 0.03535 & \mathbf{x} \in [2, 3] \times [3.9, 4] \\ 0.00035 & \text{else} \end{cases} \quad (29)$$

The triggering function is given in (6–8) with spatial kernel (9) in both cases. The magnitudes are simulated following an exponential distribution  $p_M(m_i) = \frac{1}{\beta} e^{-\beta(m_i - m_0)}$  with  $\beta = \ln(10)$  which corresponds to a Gutenberg–Richter relation with b-value of 1;  $m_0 = 3.36$  in the first case and  $m_0 = 3$  in the second case. The test likelihood is computed for twelve unseen data sets, simulated from the generative model and averaged. The simulations are done on the same spatial domain  $\mathcal{X}_{\text{sim}} = [0, 5] \times [0, 5]$ . The time window is  $\mathcal{T}_{\text{sim}} = [0, 1500]$  and  $\ell_{\text{test}}$  is evaluated using events with event times  $t_i \in [500, 1500]$ , all previous events are taken into account in the history  $H_t$ .

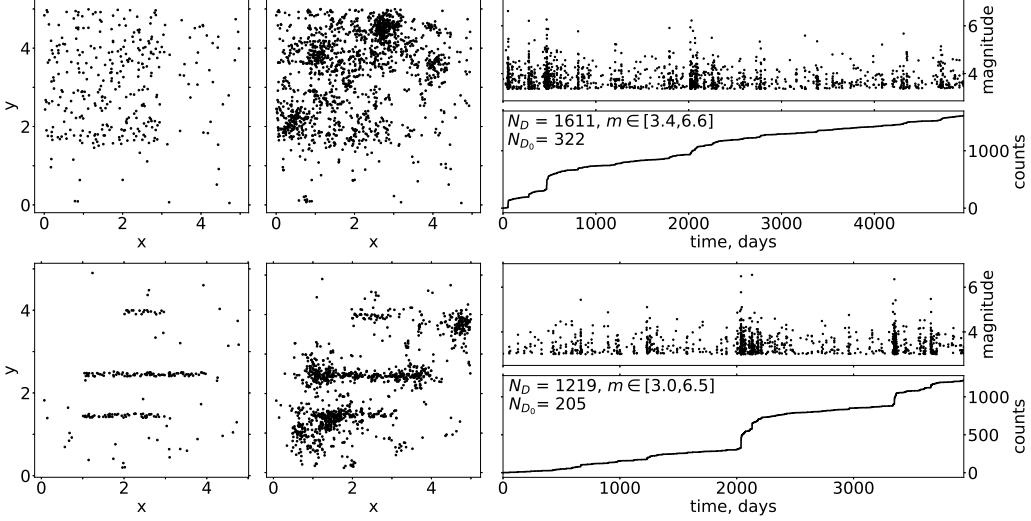


Figure 2: Setup of synthetic data experiments, Case 1 (first row) and Case 2 (second row): From left to right: background events, data set including background and offspring events, visualisation of the data as earthquake sequence over time.

Table 1: GP-ETAS setup, prior choice.

variable, symbol	prior	values
latent function $f$	Gaussian process prior	zero mean function; cov function (11)
upper bound, $\bar{\lambda}$	$\text{Gamma}(\alpha_0, \beta_0)$	$\alpha_0 = 1/c_s^2, \beta_0 = 1/c_s^2/\mu_{\bar{\lambda}}$ with $c_s = 1, \mu_{\bar{\lambda}} = 2N_D  \mathcal{X} $
hyper parameters of cov function, $\nu$	exponential distribution	$\beta\nu_0 = 1/5, \beta\nu_1 = \beta\nu_2 = 5/2$
parameters of the triggering function, $\theta_\varphi$	uniform distribution	$K_0 \in (0, 10), c \in (0, 10), p \in (0, 10), \alpha \in (0, 10), d \in (0, 10), \gamma \in (0, 10), q \in (1, 10)$
number of posterior samples, K		5000
burn-in (number of discarded initial iteration)		2000
MH proposal distribution for $\nu$ : Gaussian		$\sigma_p = 0.05$ in log units
MH proposal distribution for $\theta_\varphi$ : Gaussian		$\sigma_p = 0.01$ in log units
number of MH steps per iteration for $\theta_\varphi$		10

**GP-ETAS setup:** In GP-ETAS we need to set priors, and parameters of the Gibbs sampler, these are given in Table 1, see Section (3).

**Findings and interpretations:** The ground truth and inferred results of  $\mu(\mathbf{x})$  and  $\theta_\varphi$  are given in Figure 3–4 and Table 2–3. Performance metrics: the averaged  $\ell_{\text{test}}$  of twelve unseen data sets and the numerically approximated error of the estimated background intensity  $\ell_2$  are shown in Table 4–5. Here we describe a few noteworthy aspects. First of all, GP-ETAS recovers well the assumptions both the background intensity  $\mu(\mathbf{x})$  and the parameters of the triggering function  $\theta_\varphi$ . GP-ETAS outperforms the standards models for both metrics  $\ell_{\text{test}}$  and  $\ell_2$ . The latter fact is of particular importance, as it is common practice to use the declustered background intensity  $\mu(\mathbf{x})$  for seismic hazard assessment. One may appreciate that ETAS–classical occasionally tends to strongly overshoot the true  $\mu(\mathbf{x})$  (see Figure 3); in regions with many aftershocks, e.g. near (2.8,4.5), (1.2,3.8). This effect is less pronounced for ETAS–Silverman, where the minimum bandwidth is broader. In our approach no bandwidth selection has to be made in advance, it is obtained via sampling the hyperparameters. One also observes, that ETAS–classical and ETAS–Silverman suffer more strongly from edge effects than GP-ETAS, which seems to be fairly unaffected.

The parameters of the triggering function  $\theta_\varphi$  are roughly correctly identified in all cases and methods. All the values are close to those of the generative model. All the methods overestimate  $c$  and  $K_0$  (in Case 2), however the true values are still included in the uncertainty band (credible band) of GP-ETAS. GP-ETAS has the advantage that it provides the whole distribution of the parameters instead of only a point estimate. The median of the obtained upper bound  $\bar{\lambda}$  on  $\mu(\mathbf{x})$  using GP-ETAS overestimates in Case 1 (underestimates in Case 2) the true upper bound, however, it is fairly close to the true value, which is contained in the uncertainty band around  $\bar{\lambda}$ . Therefore,  $\bar{\lambda}$  can be considered in seismic hazard assessment as conservative choice.

Table 2: Case1: Parameter values  $\theta_\varphi$  of the triggering function.

model	quantiles	$K_0$	$c$	$p$	$\alpha$	$d$	$\gamma$	$q$
generative		0.018	0.006	1.20	1.690	0.015	0.20	2.00
ETAS-classical		0.0180	0.0075	1.23	1.667	0.018	0.18	2.06
ETAS-Silverman		0.0184	0.0068	1.21	1.666	0.018	0.18	2.07
GP-ETAS	median 0.05 0.95	0.0184 0.0164 0.0203	0.0068 0.0056 0.0085	1.21 1.19 1.24	1.662 1.595 1.734	0.017 0.014 0.022	0.19 0.17 0.21	2.07 1.93 2.23

Table 3: Case 2: Parameter values  $\theta_\varphi$  of the triggering function.

model	quantiles	$K_0$	$c$	$p$	$\alpha$	$d$	$\gamma$	$q$
generative		0.018	0.006	1.20	1.690	0.015	0.20	2.00
ETAS classical		0.0193	0.0081	1.21	1.646	0.015	0.20	2.08
ETAS silverman		0.0214	0.0058	1.15	1.616	0.015	0.20	2.10
GP-ETAS	median 0.05 0.95	0.0194 0.0175 0.0216	0.0084 0.0064 0.0103	1.21 1.19 1.24	1.648 1.586 1.703	0.014 0.011 0.017	0.20 0.18 0.22	2.00 1.89 2.14

Table 4: Averaged test likelihood  $\ell_{\text{test}}$  of twelve unseen data sets (higher is better).

experiment	generative model	ETAS-classical	ETAS-Silverman	GP-ETAS
Case 1	-344.9	-417.0	-404.2	<b>-347.8</b>
Case 2	-140.1	-201.8	-210.2	<b>-171.8</b>

Table 5: Comparison on  $\ell_2$  norm to the true background intensity (smaller is better).

experiment	criterium	ETAS-classical	ETAS-Silverman	GP-ETAS
Case 1	$\ell_2$	0.0482	0.0297	<b>0.0190</b>
Case 1	normalised	2.53	1.56	1
Case 2	$\ell_2$	0.1950	0.2410	<b>0.1282</b>
Case 2	normalized	1.52	1.88	1

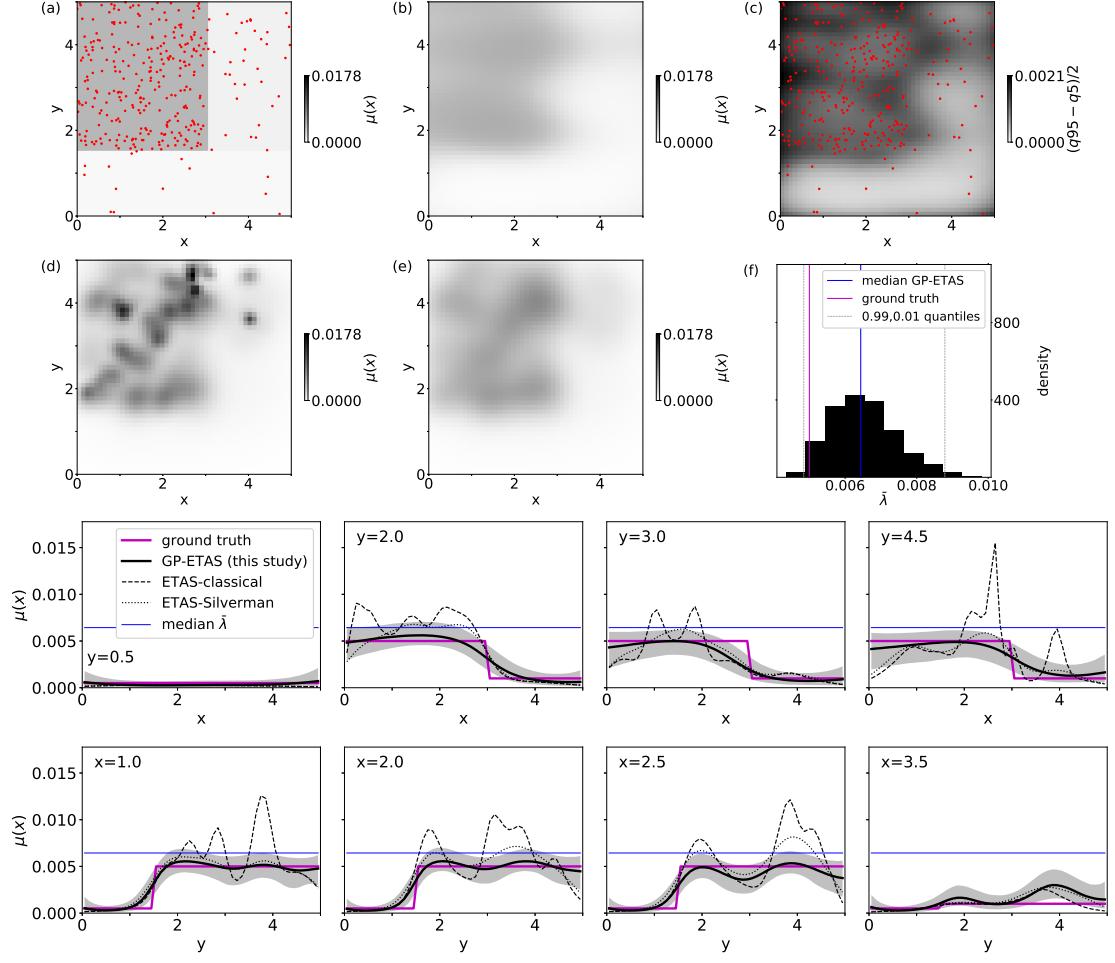


Figure 3: Experimental results of background intensity  $\mu_1(\mathbf{x})$  for the synthetic data of Case 1: *First and second row*: (a) generative model, (b) median GP-ETAS, (c) uncertainty GP-ETAS as semi inter quantile 0.05, 0.95 distance (d) ETAS-classical MLE, (e) ETAS-Silverman MLE, (f) normalised histogram of the sampled upper bound  $\bar{\lambda}$ . Dots are the background events of the realisation. *Third and fourth row*: One dimensional profiles of  $\mu_1(\mathbf{x})$  (ground truth) and inferred results are shown. The profiles are at  $y \in \{0.5, 2, 3, 4.5\}$  and  $x \in \{1, 2, 2.5, 3.5\}$ .

**Computational costs:** GP-ETAS encounters approximately a complexity of  $\mathcal{O}((N_{\mathcal{D} \cup \Pi})^3)$  for the inference of  $\mu$ . This is due to the matrix inversions involved in Gaussian process modelling. The estimation of  $\boldsymbol{\theta}_\varphi$  is less expensive and approximately of  $\mathcal{O}(N_{\mathcal{D}}^2)$ , where  $N_{\mathcal{D}}$  is the number of data points. In addition, the number of required samples in order to obtain a valuable approximation of the posterior distribution depends on the mixing properties of the Markov chain. From our experience based on the performed experiments one needs  $> 10^3$  samples after a burn-in phase of  $> 10^3$  iterations. Therefore, our proposed method in its current implementation is computationally expensive but still feasible for small to intermediate data sets with approximately  $N_{\mathcal{D}} \lesssim 10^4$  events; which seems sufficient for many situations where site specific seismic analysis takes place.

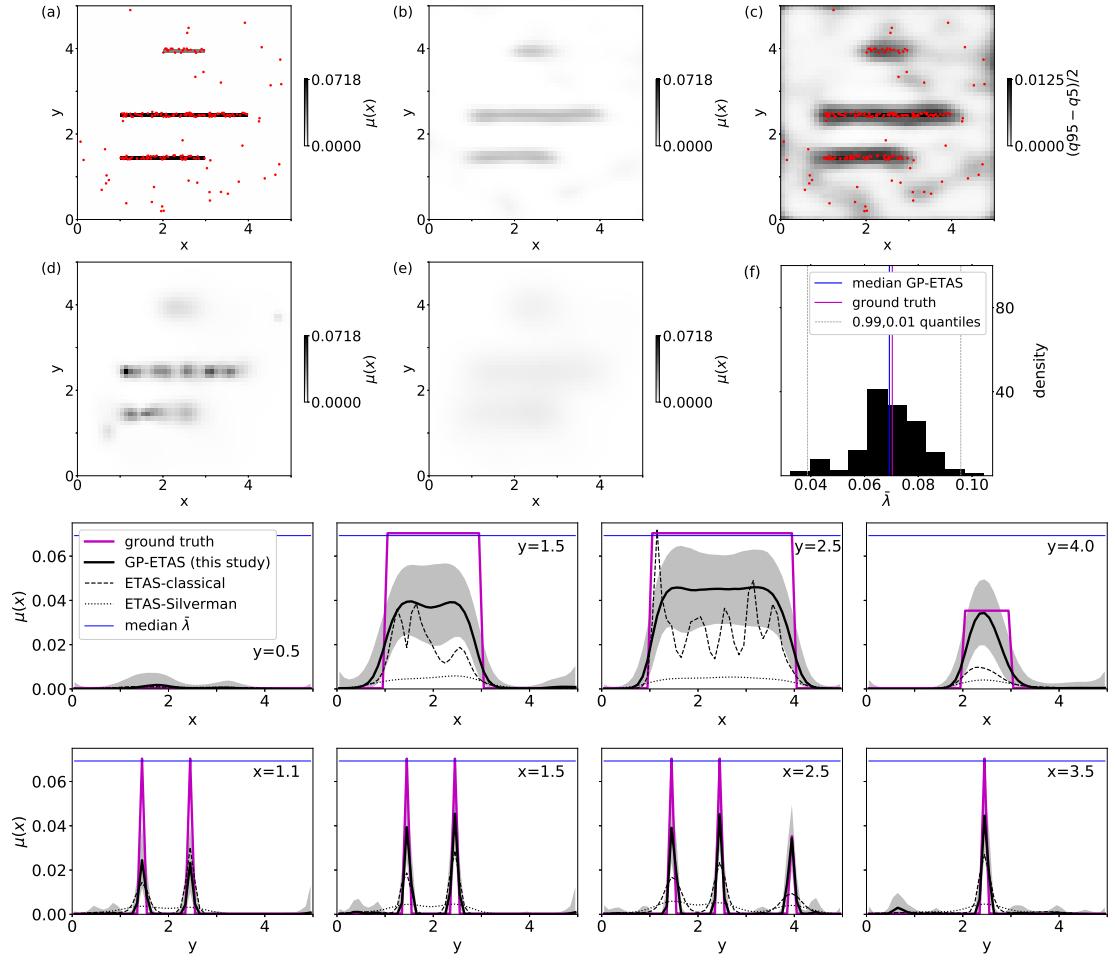


Figure 4: Same as Figure 3 but now results of background intensity  $\mu_2(\mathbf{x})$  for the synthetic data of Case 2. See Figure 3 for the description of plots and lines. The profiles are at  $y \in \{0.5, 1.5, 2.5, 4\}$  and  $x \in \{1.1, 1.5, 2.5, 3.5\}$ .

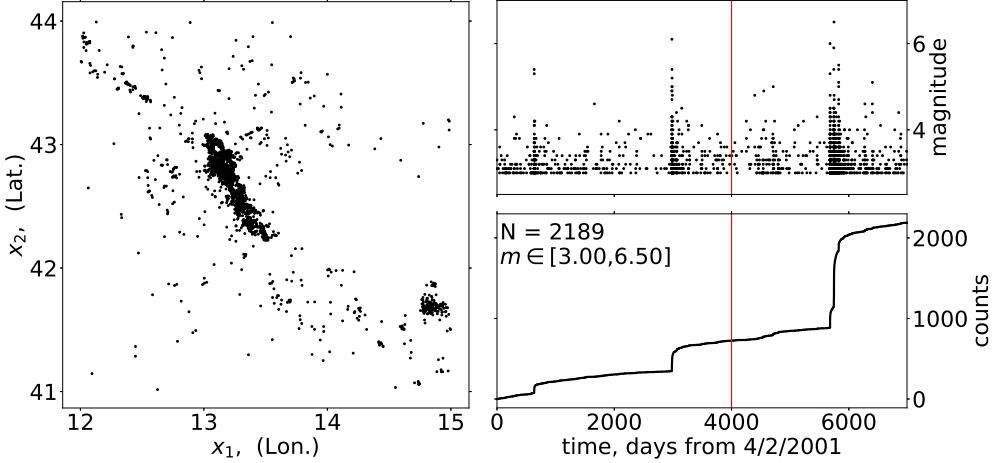


Figure 5: Earthquake data from central Italy: Epicentre plot (left) and visualisation of the data as earthquake sequence over time (right).

## 5.2 Case study: L’Aquila, Italy

Now we apply GP-ETAS to real data and compare the performance with the other models ETAS–classical and ETAS–Silverman.

The L’Aquila region in central Italy is seismically active and experiences from time to time severe earthquakes. The most famous example is the  $M_w = 6.2$  earthquake on 6 April 2009, which occurred directly below the City of L’Aquila, and caused large damage and more than 300 deaths (Marzocchi et al., 2014). This event was followed by a seismic sequence with a largest earthquake of  $M_w = 4.2$ , latter occurred almost one year later on 30 March 2010 (Marzocchi et al., 2014).

The L’Aquila data set comprises  $N = 2189$  events which occurred in a time period from 04/02/2001 to the 28/3/2020, on a spatial domain  $\mathcal{X} = [12^\circ E, 15^\circ E] \times [41^\circ N, 44^\circ N]$  with earthquake magnitudes  $3.0 \leq m \leq 6.5$ . The data was obtained from the website of the National Institute of Geophysics and Vulcanology of Italy (<http://terremoti.ingv.it/>), Istituto Nazionale della Geofisica e Vulcanologia, INGV). We split the data set into training data, all events with event times  $t_i \leq 4000$  days ( $N_{\text{training}} = 723$  events,  $\mathcal{T}_{\text{training}} = [0, 4000]$  days), and *test data*, all events with  $t_i > 4000$  days ( $N_{\text{test}} = 1466$ ,  $\mathcal{T}_{\text{test}} = [4000, 6992]$  days), as shown in Figure 5. The training data is used for the inference and the test data is used to evaluate the performance of the different models.

The inference setup is the same as described for the synthetic data. We simulate 15000 posterior samples after a burn in of 2000. The priors are – as for the synthetic data – given in Table 1. The inference results for  $\mu(\mathbf{x})$  and  $\boldsymbol{\theta}_\varphi$  are shown in Figure 6 and Table 6. The performance metric  $\ell_{\text{test}}$  for different unseen data sets (next 30 days, one year in future, five years in future, and the total test data  $\approx 8$  years) is shown in Table 7. The posterior distribution of the upper bound of the background intensity is shown in Figure 7.

For the l’Aquila data set GP-ETAS performs slightly better than the other models in terms of  $\ell_{\text{test}}$ . Note, that ETAS–classical estimates fairly large values for  $\mu(\mathbf{x})$  in regions with many aftershocks (Figure 6). This is similar to Case 1 for the synthetic experiments. Hence, as for the synthetics one may assume that ETAS–classical overshoots in these regions; the posterior of  $\bar{\lambda}$  supports this hypothesis, see Figure 7. The estimated  $\boldsymbol{\theta}_\varphi$  are similar for  $K_0, c, p, \alpha$ , differ for  $d, \gamma, q$ . Recall, that the latter describe the spatial distribution of the aftershocks. The discrepancies are to be seen in the context of (almost) linear trade-offs between the parameters  $d, \gamma$ ; for  $d, \gamma$  this can be discerned from Figure 8. The spatial kernels of the three models are shown in Figure 9 for the mean magnitude and a large magnitude.

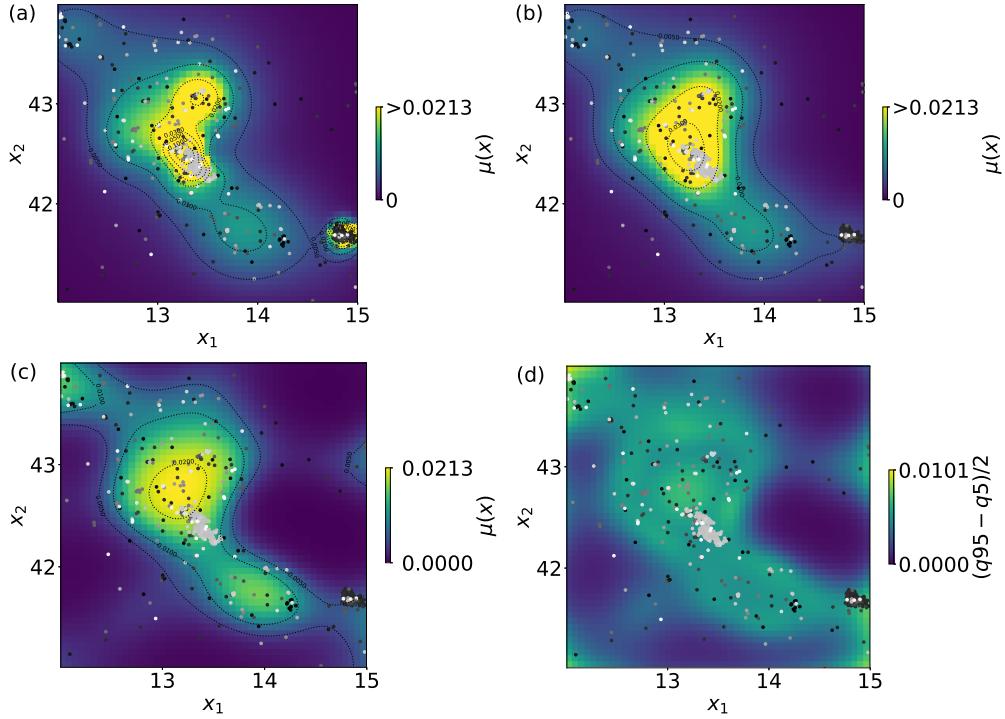


Figure 6: Results real data, L'Aquila data set: background intensity  $\mu(\mathbf{x})$  [number of shocks with  $m \geq 3$  /day/degree $^2$ ] (a) ETAS-classical MLE, (b) ETAS-Silverman MLE, (c) median GP-ETAS, (d) uncertainty GP-ETAS: semi inter quantile 0.05, 0.95 distance, and dots are the events of the training data, where the grey scaling depicts the event times, from black (older events) to white (current events). Note, (a–c) have the same scale.

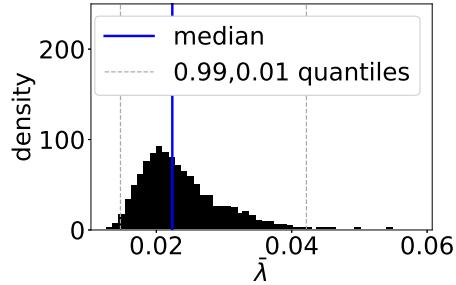


Figure 7: Normalised histogram of the sampled posterior of the upper bound  $\bar{\lambda}$

Table 6: L'Aquila data set: Parameter values  $\theta_\varphi$  of the triggering function.

model	quantiles	$K_0$	$c$	$p$	$\alpha$	$d$	$\gamma$	$q$
ETAS-classical		0.0293	0.0400	1.21	1.801	0.0008	0.30	1.95
ETAS-Silverman		0.0300	0.0352	1.18	1.773	0.0005	0.34	1.91
GP-ETAS	median	0.0269	0.0276	1.16	1.780	0.0044	0.19	2.57
	0.05	0.0224	0.0164	1.11	1.660	0.0029	0.16	2.17
	0.95	0.0321	0.0451	1.20	1.887	0.0063	0.23	3.27

Table 7: Test likelihood  $\ell_{\text{test}}$  of unseen test data sets.

testing period	$N_{\text{test}}$	ETAS-classical	ETAS-Silverman	GP-ETAS
30.0 days	2	-13.4	-13.3	<b>-12.4</b>
1.0 years	18	-79.8	-80.1	<b>-77.4</b>
5.0 years	1116	5058.8	5050.5	<b>5076.1</b>
total test period ( $\approx$ 8.2 years)	1466	5748.3	5735.5	<b>5749.0</b>

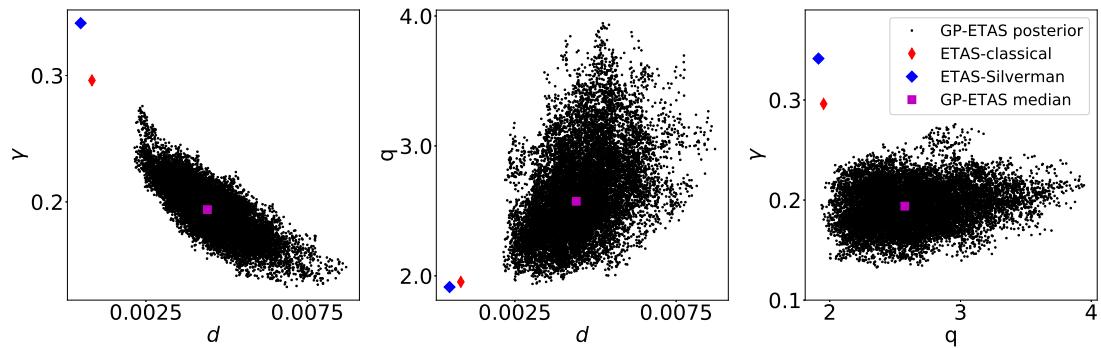


Figure 8: Scatter plot of the posterior samples of  $d, \gamma, q$

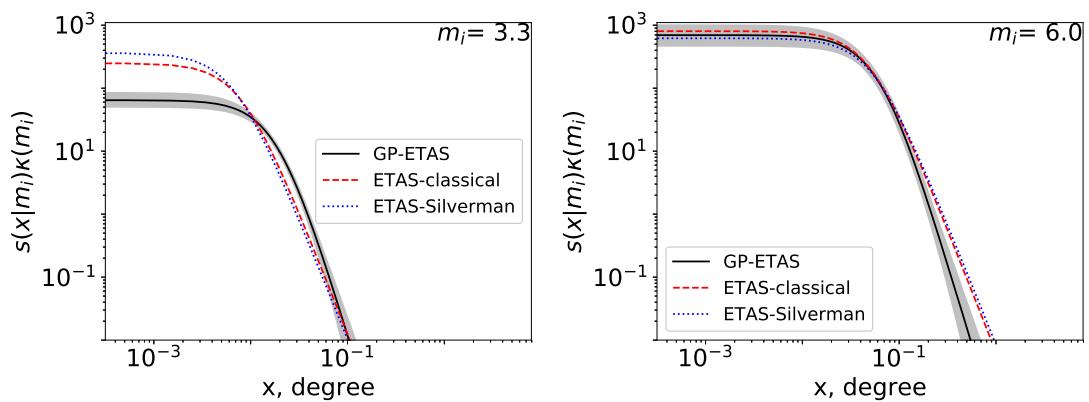


Figure 9: Estimated spatial kernel (9) multiplied by the productivity (7) for the mean magnitude  $m = 3.3$  and for a large  $m = 6$  mainshock at position  $(0,0)$ . Shaded area is 0.05, 0.95 percentile of GP-ETAS.

## 6 Discussion and Conclusions

We have demonstrated that the proposed GP-ETAS model in combination with augmentation techniques (Adams et al., 2009; Donner and Opper, 2018; Hawkes and Oakes, 1974) provides the means of assessing the Bayesian posterior of a semi-parametric spatio-temporal ETAS model. We have shown for three examples that the predictive performance improves over classical methods. In addition, we can quantify parameter uncertainties via their empirical posterior density. The developed framework is flexible and allows for several extensions that deserve consideration in future research, e.g. a time depending background rate. Another obvious extension of our work would be a Bayesian non-parametric treatment of the triggering function  $\varphi$ . See also (Zhang et al., 2019b).

Future research will deal with a more geology informed choice of the prior GP. Sometimes a given catalog comes with information about, e.g. fault locations, which are not straightforward to incorporate in traditional treatments of the spatio-temporal ETAS model. For the GP-ETAS model, however, incorporation of informative priors is possible within our Bayesian setting. For example, spatial information about fault zones can be incorporated by an adequate choice of the mean of the GP, which was chosen to be 0 throughout this work. While we restricted ourselves to the squared exponential (11) as covariance function for the GP prior of  $f$  the framework is not restricted to this either and other covariance function can be used to incorporate prior information, e.g. from the Matérn class, or any other function that ensures that the covariance matrix is positive definite.

Another important issue is the computational effort. Having to sample the GP at all observed events in  $\mathcal{D}$  and at positions of the latent Poisson process  $\Pi$ , resulting in a cubic complexity of  $\mathcal{O}((N_{\mathcal{D} \cup \Pi})^3)$ , implies an undesirable computational complexity of the current GP-ETAS Gibbs sampler. There are however several possibilities to mitigate this complexity via model approximations and/or alteration. For example, one could resort to approximations to the posterior distribution in order to be able to scale the GP-ETAS model to larger catalogues ( $N_D \gg 10^3$ ). While those always come with the sacrifice of asymptotic exactness, some approaches are likely to provide good estimates in the large data regime. One of such approximations is provided by variational inference, which was already proposed for the SGCP by Donner and Opper (2018) utilising sparse GPs (Titsias, 2009). This approach makes use of the same model augmentations utilised in this work. The variational posterior of the triggering parameters could be inferred, e.g., via black-box variational inference (Ranganath et al., 2014). Alternatively, one could restrict the calculations to finding the MAP estimate of the GP-ETAS model. For the background intensity this can be efficiently done by an expectation-maximisation algorithm based on the model augmentations presented here and sparse GPs (Donner and Opper, 2018). This can be combined with a Laplace approximation to provide an approximate Gaussian posterior. The limiting factor under such approximations will most likely arise from the branching structure for which the required computations scale like  $\mathcal{O}(N_{\mathcal{D}}(N_{\mathcal{D}} - 1)/2)$ . Finally, one could also investigate gradient-free affine invariant sampling methods as proposed by Garbuno-Inigo et al. (2019, 2020); Reich and Weissmann (2019).

We conclude by reemphasising the importance of semi-parametric Bayesian approaches to spatio-temporal statistical earthquake modelling and the need for developing efficient tools for their computational inference. Within this work we have followed the Gibbs sampling approach in combination with data augmentation and have demonstrated its applicability to realistic earthquake catalogs.

## Acknowledgments

This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG, German Science Foundation) - SFB 1294/1 - 318763901.

## References

- R. P. Adams, I. Murray, and D. J. MacKay. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16, 2009.
- G. Adelfio and M. Chiodi. Alternated estimation in semi-parametric space-time branching-type point processes with application to seismic catalogs. *Stochastic Environmental Research and Risk Assessment*, 29(2):443–450, 2014.
- E. Bacry and J.-F. Muzy. First- and second-order statistics characterization of Hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62:2184–2202, 2016.
- E. Bacry, I. Mastromatteo, and J.-F. Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015.
- R. Console, M. Murru, and A. M. Lombardi. Refining earthquake clustering models. *Journal of Geophysical Research: Solid Earth*, 108(B10):1–9, 2003.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.
- C. Donner and M. Opper. Efficient Bayesian inference of sigmoidal Gaussian cox processes. *Journal of Machine Learning Research*, 19(1998):1–34, 2018.
- S. Donnet, V. Rivoirard, and J. Rousseau. Nonparametric Bayesian estimation of multivariate Hawkes processes. *arXiv preprint arXiv:1802.05975*, 2018.
- V. Filimonov and D. Sornette. Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data. *Quantitative Finance*, 15(8):1293–1314, 2015.
- E. W. Fox, F. P. Schoenberg, and J. S. Gordon. Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *Annals of Applied Statistics*, 10(3):1725–1756, 2016.
- A. Garbuno-Inigo, N. Nüsken, and S. Reich. Affine invariant interacting Langevin dynamics for Bayesian inference. Technical Report arXiv:1912.02859, *SIAM J. Dyn. Syst.* in press, 2019.
- A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. Stuart. Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. *SIAM J. Appl. Dyn. Syst.*, 19:412–441, 2020.
- S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.
- F. Gerhard, M. Deger, and W. Truccolo. On the stability and dynamics of stochastic spiking neuron models: Nonlinear Hawkes process and point process GLMs. *PLOS Computational Biology*, 13(2):1–31, 2017.
- W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.
- A. G. Hawkes and D. Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974.
- W. Hu and P. J. Jin. An adaptive Hawkes process formulation for estimating time-of-day zonal trip arrivals with location-based social networking check-in data. *Transportation Research Part C: Emerging Technologies*, 79:136–155, 2017.

- T. H. Jordan, Y.-T. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, G. Papadopoulos, G. Sobolev, K. Yamaoka, and J. Zschau. Operational earthquake forecasting. State of knowledge and guidelines for utilization. *Annals of Geophysics*, 54(4), 2011.
- Y. Y. Kagan. Aftershock zone scaling. *Bulletin of the Seismological Society of America*, 92(2):641–655, 2002.
- J. F. C. Kingman. *Poisson processes*. Oxford University Press, 1993. ISBN 9780198536932.
- M. Kirchner. An estimation procedure for the Hawkes process. *Quantitative Finance*, 17(4):571–595, 2017.
- A. Kirichenko and H. Van Zanten. Optimality of Poisson processes intensity learning with Gaussian processes. *The Journal of Machine Learning Research*, 16(1):2909–2919, 2015.
- A. A. Kolev and G. J. Ross. Semiparametric Bayesian forecasting of spatial earthquake occurrences. *arXiv preprint arXiv:2002.01706*, 2020.
- P. A. Lewis and G. S. Shedler. Simulation of nonhomogeneous Poisson processes with log linear rate function. *Biometrika*, 63(3):501–505, 1976.
- S. W. Linderman and R. P. Adams. Scalable Bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*, 2015.
- E. Lippiello, F. Giacco, L. de Arcangelis, W. Marzocchi, and C. Godano. Parameter estimation in the ETAS model: Approximations and novel methods. *Bulletin of the Seismological Society of America*, 104(2):985–994, 2014.
- A. M. Lombardi. Estimation of the parameters of ETAS models by simulated annealing. *Scientific Reports*, 5(8417):1–11, 2015.
- D. Marsan and O. Lengliné. Extending earthquakes’ reach through cascading. *Science*, 319(5866):1076–1079, 2008.
- W. Marzocchi, A. M. Lombardi, and E. Casarotti. The establishment of an operational earthquake forecasting system in Italy. *Seismological Research Letters*, 85(5):961–969, 2014.
- G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.
- I. Murray, Z. Ghahramani, and D. J. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence, UAI 2006*, pages 359–366, 2006.
- Y. Ogata. Asymptotic behavoir of maximum likelihood. *Annals of the Institute of Statistical Mathematics*, 30:243–261, 1978.
- Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83:9–27, 1988.
- Y. Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998.
- Y. Ogata and J. Zhuang. Space-time ETAS models and an improved extension. *Tectonophysics*, 413(1-2):13–23, 2006.
- F. Omori. On the after-shocks of earthquakes. *Journal of the College of Science*, 7:111–120, 1894.
- N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.
- M. D. Porter and G. White. Self-exciting hurdle models for terrorist activity. *Annals of Applied Statistics*, 4(1):106–124, 2010.

- R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014.
- J. G. Rasmussen. Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642, 2013.
- S. L. Rathbun. Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *Journal of Statistical Planning and Inference*, 51(1):55 – 74, 1996.
- S. Reich and S. Weissmann. Fokker–Planck particle systems for Bayesian inference: Computational approaches. Technical Report arXiv:1911.10832, University of Potsdam, 2019.
- P. Reynaud-Bouret and S. Schbath. Adaptive estimation for Hawkes processes; Application to genome analysis. *Annals of Statistics*, 38(5):2781–2822, 2010.
- G. Ross. Bayesian estimation of the ETAS model for earthquake occurrences. *Preprint*, 2018.
- F. P. Schoenberg. Facilitated estimation of ETAS. *Bulletin of the Seismological Society of America*, 103(1):601–605, 2013.
- B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986. ISBN 0412246201.
- M. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.
- T. Utsu. A statistical study on the occurrence of aftershocks. *Geophys. Mag.*, 30:521–605, 1961.
- T. Utsu. Aftershocks and earthquake statistics (1): Some parameters which characterize an aftershock sequence and their interrelations. *Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics*, 3(3):129–195, 1970.
- A. Veen and F. P. Schoenberg. Estimation of space-time branching process models in seismology using an EM-type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008.
- Q. Wang, F. P. Schoenberg, and D. D. Jackson. Standard errors of parameter estimates in the ETAS model. *Bulletin of the Seismological Society of America*, 100(5 A):1989–2001, 2010.
- C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006. ISBN 0-262-18253-X.
- J. Windle, N. G. Polson, and J. G. Scott. Sampling pólya-gamma random variates: Alternate and approximate techniques. *arXiv preprint arXiv:1405.0506*, 2014.
- R. Zhang, C. Walder, and M.-A. Rizoiu. Variational inference for sparse Gaussian process modulated Hawkes process. *arXiv preprint arXiv:1905.10496*, 2019a.
- R. Zhang, C. Walder, M.-A. Rizoiu, and L. Xie. Efficient non-parametric Bayesian Hawkes processes. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4299–4305, 2019b.
- Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1513–1522, 2015.
- F. Zhou, Z. Li, X. Fan, Y. Wang, A. Sowmya, and F. Chen. Scalable inference for nonparametric Hawkes process using Pólya-Gamma augmentation. *arXiv:1910.13052v1*, 2019.
- K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 641–649, 2013.
- J. Zhuang. Next-day earthquake forecasts for the Japan region generated by the ETAS model. *Earth, Planets and Space*, 63(3):207–216, 2011.
- J. Zhuang, Y. Ogata, and D. Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380, 2002.

# Appendices

## A GP-ETAS generative model

The generative model of GP-ETAS consists of two parts given in Algorithm 2 – 3, and requires several inputs. Here, we chose a GP with zero mean and covariance function  $k(\mathbf{x}, \mathbf{x}')|\boldsymbol{\nu})$  given in (11) with hyperparameters  $\boldsymbol{\nu}$ , a triggering function  $\varphi(\cdot|\boldsymbol{\theta}_\varphi)$  given in (6 – 8) with spatial kernel (9) and therefore  $\boldsymbol{\theta}_\varphi = (K_0, c, p, \alpha, d, \gamma, q)$ , and a mark distribution, that is an exponential distribution  $m_i - m_0 \sim \text{Exponential}(\beta)$  (Gutenberg–Richter relation) with parameters  $\beta, m_0$ . The simulation algorithm can be easily adjusted for other choices.

---

**Algorithm 2 :** GP-ETAS generative model Part 1: generating background events  $\mathcal{D}_0$

---

**Input:** Spatio-temporal domain  $\mathcal{X} \times \mathcal{T}$ ; GP mean function and covariance function with hyperparameters  $\boldsymbol{\nu}$ ; upper bound  $\bar{\lambda}$ ; parameters of the mark density  $\beta, m_0$

**Output:** Background events  $\mathcal{D}_0 = \{t_i, \mathbf{x}_i, m_i, z_i = 0\}_{i=1}^{N_{\mathcal{D}_0}}$

```

1:  $N \sim \text{Poisson}(\bar{\lambda}|\mathcal{X}||\mathcal{T}|)$             $\triangleright$  Sample number of candidate events from Poisson distribution
2:  $\{\mathbf{x}_i\}_{i=1}^N \sim \mathcal{U}(\mathcal{X})$             $\triangleright$  Distribute candidate events uniformly in  $\mathcal{X}$ 
3:  $\{f(\mathbf{x}_i)\}_{i=1}^N \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_\boldsymbol{\nu})$   $\triangleright$  Draw function values from the GP with  $\mathbf{K}_\boldsymbol{\nu} = \{k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\nu})\}_{i,j}^N$ 
4:  $\mathcal{D}_0 \leftarrow \emptyset$                           $\triangleright$  Initialise the set of background events  $\mathcal{D}_0$ 
5:  $N_{\mathcal{D}_0} \leftarrow 0$                           $\triangleright$  Initialise number of background events
6: for  $i \leftarrow 1, \dots, N$  do
7:    $r_i \sim \mathcal{U}(0, \bar{\lambda})$             $\triangleright$  Draw a uniform random variable on the interval  $[0, \bar{\lambda}]$ 
8:   if  $r_i < \bar{\lambda}\sigma(f(\mathbf{x}_i))$  then            $\triangleright$  Acceptance criteria
9:      $\mathbf{x}_i$  is accepted
10:     $t_i \sim \mathcal{U}(\mathcal{T})$             $\triangleright$  Distribute background event uniformly in  $\mathcal{T}$ 
11:     $m_i - m_0 \sim \text{Exponential}(\beta)$   $\triangleright$  Sample a mark from a shifted exponential distribution
12:     $z_i \leftarrow 0$             $\triangleright$  Assign branching variable  $z_i$ , index of parent event
13:     $\mathcal{D}_0 \leftarrow \mathcal{D}_0 \cup \{t_i, \mathbf{x}_i, m_i, z_i\}$             $\triangleright$  Add event to accepted background events  $\mathcal{D}_0$ 
14:     $N_{\mathcal{D}_0} \leftarrow N_{\mathcal{D}_0} + 1$             $\triangleright$  Count number of background events
15:   end if
16: end for
17: Sort  $\mathcal{D}_0$  by event times  $t_i$ 
18: return  $\mathcal{D}_0$ 

```

---

---

**Algorithm 3** : GP-ETAS generative model Part 2: generating and adding offspring events

---

**Input:**  $\mathcal{D}_0$ ,  $N_{\mathcal{D}_0}$  from Algorithm (2); spatio-temporal domain  $\mathcal{X} \times \mathcal{T}$ ; triggering function  $\varphi(\cdot)$  with defining parameters  $\boldsymbol{\theta}_\varphi = (K_0, c, p, \alpha, d, \gamma, q)$ ; parameters of the mark distribution  $\beta$ ,  $m_0$

**Output:** Background events  $\mathcal{D} = \{t_i, \mathbf{x}_i, m_i, z_i\}_{i=1}^{N_{\mathcal{D}}}$

```

1:  $\mathcal{D} \leftarrow \mathcal{D}_0$             $\triangleright$  Initialise the set of simulated events  $\mathcal{D}$  with the background events  $\mathcal{D}_0$ 
2:  $N_{\mathcal{D}} \leftarrow N_{\mathcal{D}_0}$      $\triangleright$  Initialise number of simulated events with the number of background events
3:  $j \leftarrow 1$                    $\triangleright$  Initialise the index of potential parent event
4: while  $j < N_{\mathcal{D}}$  do       $\triangleright$  Consider all events in  $\mathcal{D}$  for producing potentially offspring
5:    $\{t_j, \mathbf{x}_j, m_j, z_j\} \leftarrow \mathcal{D}[j]$            $\triangleright$  Obtain entries of the  $j$ th simulated event in  $\mathcal{D}$ 
6:    $\lambda_{\max,j} \leftarrow \max(\kappa(m_j)g(t - t_j)) = K e^{\alpha(m_j - m_0)} c^{-p}$   $\triangleright$  Get upper bound  $\lambda_{\max,j}$  of  $j$ th PP
7:    $|\mathcal{T}_j| \leftarrow |\mathcal{T}| - t_j$            $\triangleright$  Compute the size of the time window  $|\mathcal{T}_j|$  of direct offspring
8:    $N_j \leftarrow \text{Poisson}(\lambda_{\max,j} |\mathcal{T}_j|)$            $\triangleright$  Sample number of candidate offspring events
9:   if  $N_j > 0$  then           $\triangleright$  Check if there are candidate offspring events
10:     $\{t_i\}_{i=1}^{N_j} \sim \mathcal{U}(0, |\mathcal{T}_j|) + t_j$        $\triangleright$  Distribute candidate events uniformly in  $[t_j, t_{\max}]$ 
11:    for  $i \leftarrow 1, \dots, N_j$  do           $\triangleright$  Thinning procedure
12:       $r_i \sim \mathcal{U}(0, \lambda_{\max,j})$        $\triangleright$  Draw a uniform random variable on the interval  $[0, \lambda_{\max,j}]$ 
13:      if  $r_i < \lambda_{\max,j}(t_i - t_j + c)^{-p}$  then           $\triangleright$  Acceptance criteria using (7,8)
14:         $t_i$  is accepted
15:         $\mathbf{x}_i \sim s(\mathbf{x} - \mathbf{x}_j | m_j, d, \gamma, q)$        $\triangleright$  Sample position of the offspring event (9)
16:         $m_i - m_0 \sim \text{Exponential}(\beta)$            $\triangleright$  Sample marks, see above
17:         $z_i \leftarrow j$            $\triangleright$  Assign branching variable  $z_i$ , index of parent event
18:         $\mathcal{D} \leftarrow \mathcal{D} \cup \{t_i, \mathbf{x}_i, m_i, z_i\}$            $\triangleright$  Add offspring event to  $\mathcal{D}$ 
19:         $N_{\mathcal{D}} \leftarrow N_{\mathcal{D}} + 1$            $\triangleright$  Count number of simulated events  $D$ 
20:    end if
21:  end for
22:   $j \leftarrow j + 1$            $\triangleright$  Advances to next possible parent event
23: end if
24: end while
25: Sort  $\mathcal{D}$  by event times  $t_i$ , while taking care of properly mapping the branching variables
26: return  $\mathcal{D}$ 

```

---

## B Definition of the Pólya–gamma density

Here, we briefly define the Pólya-gamma density (Polson et al., 2013). First we define the  $p_{\text{PG}}(\omega|b, 0)$ , which is completely defined through its Laplace transform

$$\int_0^\infty e^{-\omega t} p_{\text{PG}}(\omega|b, 0) d\omega = \cosh^{-b}(\sqrt{t/2}). \quad (30)$$

With this definition it can be shown that

$$\omega \stackrel{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2}, \quad (31)$$

where  $g_k \sim \text{Gamma}(b, 1)$  and the equality is in distribution. With this result one can then define a *tilted* Pólya–gamma density given by

$$p_{\text{PG}}(\omega|b, c) \propto e^{-\frac{c^2}{2}\omega} p_{\text{PG}}(\omega|b, 0), \quad (32)$$

where  $b \in \mathbb{R}^+$  and  $c \in \mathbb{R}$ , and the normalisation can be straightforwardly obtained with (30). Also for this the tilted density we can derive the Laplace transform

$$\int_0^\infty e^{-\omega t} p_{\text{PG}}(\omega|b, c) d\omega = \frac{\cosh^b(c/2)}{\cosh^b\left(\frac{\sqrt{c^2/2+t}}{2}\right)}. \quad (33)$$

From (33) all moments of the Pólya–gamma density can be derived analytically and furthermore an acceptance-rejection algorithm with high acceptance rate was derived by Polson et al. (2013).

## C Conditional posteriors for the background intensity

Here we derive the conditional posterior distributions given in (24a)–(24e). For each of those our starting point is the augmented likelihood of the background intensity in (23) with the prior of interest. Note, that for the augmented variables  $\Pi, \boldsymbol{\omega}_{\mathcal{D}}, \boldsymbol{\omega}_\Pi$  there are no additional priors, and hence their conditionally distribution will only be determined by (23).

**The latent Poisson process  $\Pi$**  To derive the conditional posterior for  $\Pi$ , we consider all terms in (23) that depend on  $\Pi = \{x_l\}_{l=1+N_{\mathcal{D}}}^{N_{\mathcal{D} \cup \Pi}^{(k)}}$  and marginalise over the  $\boldsymbol{\omega}_\Pi$ . This results in

$$p(\Pi|\bar{\lambda}, \mathbf{f}) \propto \prod_{l=N_{\mathcal{D}}+1}^{N_{\mathcal{D} \cup \Pi}} \bar{\lambda} \sigma(-f_l) \exp\left(-|\mathcal{T}| \int_{\mathcal{X}} \bar{\lambda} \sigma(-f(\mathbf{x})) d\mathbf{x}\right), \quad (34)$$

which we identify as an unnormalised Poisson process density with rate  $\bar{\lambda} \sigma(-f_l)$ . Again note, that the process is defined over  $\mathcal{T} \times \mathcal{X}$ . To sample  $\Pi^{(k)}$  in the  $k^{\text{th}}$  iteration, we can utilise the *thinning* procedure (Lewis and Shedler, 1976), where we first sample a homogeneous point process with intensity  $\bar{\lambda}^{(k-1)}$ . At the resulting points we draw the GP  $f$  given the previous sample  $\mathbf{f}^{(k-1)}$  from the predictive distribution (see below). Then, we keep all events  $l$  with probability  $\sigma(-f_l)$  which yields  $\Pi^{(k)}$  according to (24a).

**The Pólya–gamma random variables  $\boldsymbol{\omega}$ :** Next we sample the Pólya–gamma random variables at  $\mathcal{D}_0$ . From (23) we see, that the components in  $\boldsymbol{\omega}_{\mathcal{D}_0}$  factorise, meaning that the conditional posteriors are independent. Hence, we get for each  $i : z_i = 0$

$$p(\omega_i|\mathbf{f}, \mathcal{D}, Z) \propto e^{-\frac{(f_i)^2}{2}} \omega_i p_{\text{PG}}(\omega_i|1, 0) \propto p_{\text{PG}}(\omega_i|1, |f_i|), \quad (35)$$

and hence  $\boldsymbol{\omega}_{\mathcal{D}}^{(k)}$  can be sampled independently from a tilted Pólya-gamma distribution where  $i : z_i = 0$ , given branching structure  $Z^{(k)}$  and GP  $\mathbf{f}^{(k-1)}$ . For  $i : z_i \neq 0$  we set  $\omega_i = 0$ . The last equivalence follows from a property of an *tilted* Pólya-gamma distribution, see (32). In effect, we get (24c).

The conditional posterior  $\omega_{\Pi}$  at the positions of the latent events  $\Pi$  also factorises in all components and hence we get for each  $l = N_{\mathcal{D}} + 1, \dots, N_{\mathcal{D} \cup \Pi}$

$$p(\omega_l | \mathbf{f}, \Pi) \propto e^{-\frac{(f_l)^2}{2}\omega_l} p_{\text{PG}}(\omega_l | 1, 0) \propto p_{\text{PG}}(\omega_l | 1, |f_l|), \quad (36)$$

and also  $\omega_{\Pi}^{(k)}$  can be sampled independently from a tilted Pólya–gamma distribution (24b) given the samples of  $\Pi^{(k)}$  and  $\mathbf{f}^{(k-1)}$ ; we get (24b).

**The upper bound on the intensity  $\bar{\lambda}$ :** For  $\bar{\lambda}$  we assume a Gamma prior  $p(\bar{\lambda} | \alpha_0, \beta_0)$  with shape parameter  $\alpha_0$  and rate parameter  $\beta_0$ . Together with (23) we derive the conditional posterior being,

$$p(\bar{\lambda} | \mathcal{D}_0, \Pi, Z) \propto \bar{\lambda}^{N_{\mathcal{D}_0 \cup \Pi}} e^{-\bar{\lambda}|\mathcal{X}| |\mathcal{T}|} p(\bar{\lambda} | \alpha_0, \beta_0) \propto \text{Gamma}(\bar{\lambda} | \alpha_1, \beta_1). \quad (37)$$

where  $\alpha_1 = N_{\mathcal{D}_0 \cup \Pi} + \alpha_0$  and  $\beta_1 = |\mathcal{T}| |\mathcal{X}| + \beta_0$ . Hence, given  $\Pi^{(k)}$  and  $D_0^{(k)}$  we can sample  $\bar{\lambda}^{(k)}$ , and one gets (24d).

**The posterior Gaussian Process  $f$ :** For the conditional posterior of the Gaussian process  $f$  we rewrite (23) with the terms depending on  $f$  as follows

$$\begin{aligned} p(\mathcal{D}_0, \omega_{\mathcal{D}}, \Pi, \omega_{\Pi} | \mathbf{f}, \bar{\lambda}, Z) &\propto \prod_{i:z_i=0}^{N_{\mathcal{D}}} e^{f_i u_i - \frac{f_i^2}{2}\omega_i} \prod_{i:z_i \neq 0}^{N_{\mathcal{D}}} e^{f_i u_i - \frac{f_i^2}{2}\omega_i} \prod_{l=N_{\mathcal{D}}+1}^{N_{\mathcal{D} \cup \Pi}} e^{f_l u_l - \frac{f_l^2}{2}\omega_l} \\ &= e^{-\frac{1}{2}\mathbf{f}^\top \Omega \mathbf{f} + \mathbf{u}^\top \mathbf{f}}, \end{aligned} \quad (38)$$

where we define  $u_i = \frac{1}{2}$  if  $z_i = 0$ ,  $u_i = 0$  if  $z_i \neq 0$ , and  $u_l = -\frac{1}{2}$  for  $l = N_{\mathcal{D}} + 1, \dots, N_{\mathcal{D} \cup \Pi}$ . It follows that  $\Omega = \text{diag}(\omega_{\mathcal{D}}, \omega_{\Pi})$ . The GP prior  $\mathbf{f}$  at a finite set  $\mathcal{D}, \Pi$  of points is given by

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f} | \mathbf{0}, \mathbf{K}_{\mathbf{f}, \mathbf{f}}) \quad (39)$$

where the entry of row  $i$  and column  $j$  of  $\mathbf{K}_{\mathbf{f}, \mathbf{f}}$  is given by the covariance function (11)  $k(\mathbf{x}_i, \mathbf{x}_j | \boldsymbol{\nu})$ . Together with (38) we identify the conditional posterior

$$p(\mathbf{f} | \mathcal{D}, \omega_{\mathcal{D}}, \Pi, \omega_{\Pi}, Z) \propto e^{-\frac{1}{2}\mathbf{f}^\top \Omega \mathbf{f} + \mathbf{u}^\top \mathbf{f}} \mathcal{N}(\mathbf{f} | \mathbf{0}, \mathbf{K}_{\mathbf{f}, \mathbf{f}}) \quad (40a)$$

$$\propto \mathcal{N}\left(\mathbf{f} | \left[\Omega + \mathbf{K}_{\mathbf{f}, \mathbf{f}}^{-1}\right]^{-1} \mathbf{u}, \left[\Omega + \mathbf{K}_{\mathbf{f}, \mathbf{f}}^{-1}\right]^{-1}\right), \quad (40b)$$

which defines the conditional posterior at  $\mathbf{f}$ , and it is easy to sample  $\mathbf{f}^{(k)}$  given instances of  $\omega_{\mathcal{D}}, \Pi, \omega_{\Pi}$  from previous samples. However, the algorithm requires us to sample the posterior also at other points  $\mathbf{x}^* \notin \mathcal{D} \cup \Pi$ , e.g. for sampling the next instance of  $\Pi$  or for visualisation of the background intensity  $\mu(\mathbf{x})$  on a grid. Here, we denote the GP at all additional points  $\mathbf{f}^*$ . The GP prior defines a predictive distribution (Williams and Rasmussen, 2006) of any set  $\mathbf{f}^*$  given  $\mathbf{f}$

$$p(\mathbf{f}^* | \mathbf{f}, \boldsymbol{\nu}) = \mathcal{N}\left(\mathbf{f}^* | \mathbf{K}_{\mathbf{f}^*, \mathbf{f}} \mathbf{K}_{\mathbf{f}, \mathbf{f}}^{-1} \mathbf{f}, \mathbf{K}_{\mathbf{f}^*, \mathbf{f}^*} - \mathbf{K}_{\mathbf{f}^*, \mathbf{f}} \mathbf{K}_{\mathbf{f}, \mathbf{f}}^{-1} \mathbf{K}_{\mathbf{f}, \mathbf{f}^*}\right), \quad (41)$$

where  $\mathbf{K}_{\mathbf{f}^*, \mathbf{f}^*}$  contains the covariances between the points  $\mathbf{x}^*$  of  $\mathbf{f}^*$  and  $\mathbf{K}_{\mathbf{f}^*, \mathbf{f}} = \mathbf{K}_{\mathbf{f}^*, \mathbf{f}}^\top$  between the points  $\mathbf{x}^*$  of  $\mathbf{f}^*$  and  $\mathbf{f}$  at  $\mathbf{x}$ . Note, that the posterior of  $\mathbf{f}^*$  given  $\mathbf{f}$  is equal to the conditional prior, since (23) does not depend on  $\mathbf{f}^*$ .