ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Master Thesis Econometrics

Non-parametric Bayesian inference in multidimensional marked Hawkes processes

Thijs de Vries
Student ID: 482528

Supervisor: Dr. M. D. Zaharieva
Second assessor: Dr. A. M. Schnucker

24 september 2020

The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

Non-parametric Bayesian inference in multidimensional marked Hawkes processes

Thijs de Vries

Abstract

In this thesis, I expand the Gibbs-Hawkes algorithm, a non-parametric Bayesian inference algorithm that can be used to estimate the kernel function of a uni-dimensional unmarked Hawkes process, for the estimation of kernel functions of multidimensional marked Hawkes processes. The new expanded algorithm (called Multidimensional Gibbs-Hawkes) allows for flexible estimation of kernel functions for Hawkes kernels for multivariate Hawkes processes. It also accounts for marked processes, where the marks can have influence on the offspring intensity. I show how the Multidimensional Gibbs-Hawkes is expanded from the original Gibbs-Hawkes process and show empirically using two simulated datasets that it is able to estimate flexible triggering kernels. I also show that it performs similar or better to two benchmarks: a parametric approach and a non-parametric non-Bayesian approach.

Contents

1 Introduction

Point processes are used to describe random events triggering in some dimension (e.g. time). A particular class of point processes is the self-exciting point process, where an event triggering can increase the likelihood of more events triggering. The Hawkes process is a common type of self-exciting point processes and can be used to model these self-exciting point processes. Hawkes processes (first described by Hawkes (1971)) are used in a variety of fields, including modelling earthquake occurrences (Ogata (1998)), social interactions on Twitter (Simma & Jordan (n.d.)), systemic risk in finance (Aït-Sahalia *et al.* (2015)), civilian deaths during wars (Lewis *et al.* (2012)) and transcriptional regulatory events in biology (Carstensen *et al.* (2010)).

The Poisson point process (or simply Poisson process) is often used for describing point processes (of all kinds, not necessarily self-exiting point processes) due to its mathematical properties. Hawkes processes are typically described as Poisson processes. Poisson processes can be homogeneous or inhomogeneous. This relates to the parameter λ, called the *rate* or *intensity*. The intensity λ describes how often events happen and is always positive (i.e. $\lambda \geq 0$). If λ is a constant, the Poisson process is homogeneous. In the inhomogeneous case, the intensity is described by some locally integrable positive function $\lambda(x)$, where x is some variable in the underlying parameter space.

Estimating the intensity on the homogeneous Poisson process is less complicated than estimating the intensity on the inhomogeneous Poisson process, but homogeneous Poisson processes lack the ability to change the intensity, hence Hawkes processes are described as a type of inhomogeneous Poisson processes. This is done due to the convenient and well understood properties of the inhomogeneous Poisson process.

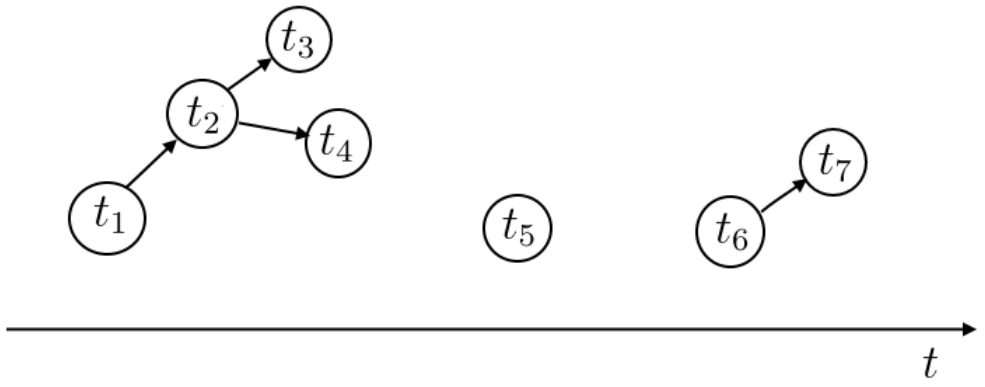


Figure 1: Schematic of a branching structure

In a Hawkes process, events are categorized as either immigrants or offspring. Immigrants are new arrivals; events that could happen any time. Offspring events follow from immigrants, and are what differentiates the Hawkes process from the Poisson process. This creates a branching structure which describes the relation between the immigrants and their offspring. The branching structure is unknown, but can be retrieved from the data. A simple example is shown in figure 1. Here, t_1 is an immigrant, with t_2, t_3 and t_4 as its offspring. Immigrants do not need to have offspring (t_5).

The Hawkes process can be extended into a marked Hawkes process (e.g. Embrechts *et al.* (2011)). In a marked Hawkes process, the events also have some variables associated to that event. Earthquakes can be considered as a marked Hawkes process (Ogata (1998)). Here, the magnitude and coordinates of the epicenter can be considered as marks to the earthquake. Any offspring events can be dependent on the marks. In the earthquake example, if an earthquake has a high magnitude, we can expect more aftershocks compared to an earthquake with low magnitude. Similarly, we expect aftershocks to be close to the epicenter of the earthquake.

A multidimensional Hawkes process (also called a multivariate Hawkes process) is a point process where events are generated in multiple dimensions. Here, an event triggering in one dimension can create offspring in another dimension. A simple example is that of earthquakes and volcanic eruptions. Both can be seen as a Hawkes process (in the case of volcanic eruptions, large eruptions causing smaller eruptions), and there has been some research on volcanic eruptions causing earthquakes (and vice versa, e.g. Hill *et al.* (2002)). In a multidimensional Hawkes process, these interactions are dimensions onto themselves with their

own offspring intensity.

Hawkes processes can be used to describe a wide variety of events and are useful in studying the interactions between these events. Multidimensional Hawkes processes allow for investigation into the correlation between events happening, such as spillover effects in finance (Embrechts *et al.* (2011)) and the interactions between orders of futures and their size (Rambaldi *et al.* (2017)).

Although defining the Hawkes process theoretically is straightforward, estimation using maximum likelihood (the most common frequentist method) can prove difficult (Veen & Schoenberg (2008)). Likelihood functions for Hawkes processes tend to be non-linear and (in case of multidimensional Hawkes processes) complex, leading to multimodal or flat likelihood functions, which are difficult and computationally intensive to numerically optimize. A non-parametric EM-algorithm has been proposed by Lewis & Mohler (2011) that uses maximum penalized likelihood estimation (MPLE) to make a computationally tractable problem. This MPLE algorithm is computationally fast and works well and I use it in the simulation study to compare my proposed algorithm against.
Bayesian inference using samplers can provide alternative methods in analyzing Hawkes Processes. Bayesian inference has some advantages over frequentist methods. They allow the researcher to include prior information. Inference on data are exact and not asymptotic approximations, and it is generally easier to interpret the parameters in a Bayesian setting compared to a frequentist setting. Thus, Bayesian methods for Hawkes processes are useful and desired (see O'Hagan (2004) for a more in depth review of Bayesian methods).

Rasmussen (2013) was the first to use a Bayesian approach to estimate parameters for a marked Hawkes process. The research used a Metropolis-within-Gibbs method to sample the parameters. Linderman & Adams (2015) then proposed a discrete-time formulation and developed a scalable and computationally efficient algorithm for Bayesian inference on multidimensional unmarked Hawkes processes. Donnet *et al.* (2019) showed a non-parametric Bayesian approach using Markov Chain Monte Carlo sampling, also for multidimensional unmarked Hawkes processes. Their non-parametric approach, based on an infinite-dimensional parameter space, remains theoretical and the Markov Chain Monte Carlo sampling algorithm is not easily scalable.

Non-parametric approaches allow for inference in statistical processes without relying on some assumed structure of the process. They are thus less dependent on the assumptions

of the researcher, and rely more on the data. They usually sacrifice some accuracy over parametric models, but are better suited in cases where the structure of the underlying data is unknown. Zhang *et al.* (2019) proposed a Bayesian non-parametric algorithm for Hawkes process by modeling the Hawkes process with a Laplace Bayesian Poisson process (an inhomogeneous Poisson process, Walder & Bishop (2017)) and use Gibbs sampling, but limited themselves to an univariate unmarked Hawkes process. They also show in their study that their proposed methods have a linear time complexity, making it more efficient than Donnet *et al.* (2019). For this reason, in this thesis I examine an extension of Zhang *et al.* (2019) by proposing non-parametric Bayesian inference for multidimensional marked Hawkes process. The proposed algorithm, called Multidimensional Gibbs-Hawkes, extends the work of Zhang *et al.* (2019) by conditioning the Laplace Bayesian Poisson process on the marks of the events and defining how to apply the process to a multidimensional Hawkes process.

In this thesis I explain the methodology behind the proposed algorithm. I detail how the multidimensional Hawkes process is split up for Bayesian inference into separate conditional posterior distributions. Using the conditional posterior distributions I describe a computational algorithm similar to the *Gibbs-Hawkes* algorithm proposed by Zhang *et al.* (2019). I then describe the simulation study I performed to evaluate my proposed algorithm, and finally discuss the results. The simulations show that Multidimensional Gibbs-Hawkes performs sometimes better and sometimes worse compared to two non-Bayesian benchmarks.

Specifically, in some dimensions, Multidimensional Gibbs-Hawkes performed better than the benchmarks, and in others one of the benchmarks performed better. Multidimensional Gibbs-Hawkes is unable to detect when there is no triggering mechanism from one dimension to another, similar to one of the benchmarks. It is able to account for different types of offspring intensity kernels. The computation time remains a barrier to the practical use of Multidimensional Gibbs-Hawkes, as computation time is too long, especially compared to the benchmarks. Multidimensional Gibbs-Hawkes took approximately eight hours to complete, whereas the benchmarks completed near instantly.

2 Methodology

In this section, I first describe the Hawkes process on a mathematical level. Then, following the approach of Zhang *et al.* (2019), I first propose a computational algorithm called Multidimensional Gibbs-Hawkes. I describe the distribution of the branching structure conditional on the immigrant and offspring intensities. Then I detail the posterior distributions

of immigrant and offspring intensities conditional on the branching structure and marks.

2.1 Hawkes processes

In the Hawkes process, the intensity $\lambda(t)$ depends on time t (that is, time t is some real number; $t \in \mathbf{R}$) of the events and is given as

$$\lambda(t) = \mu(t) + \sum_{t_i<t} \phi(t - t_i). \tag{1}$$

Here $\mu(t)$ is the base intensity and dictates how often new events (often called immigrants) happen. In the general case μ can be inhomogeneous and thus time dependent, but it is often assumed as a constant. With i the index for a previous event (that is, $t_i < t$), $\phi(t - t_i)$ is some function (often called the kernel function of the Hawkes process) that causes the self-exciting behaviour of a Hawkes process, and influences how often events trigger following an immigrant event with arrival time t_i (which are often called offspring).

In a marked Hawkes process every event has some variables $x \in \mathbf{R}^d$ associated with that event, such that each event (t, x) contains the time of the event t and some variable or a d-dimensional vector of variables x that are associated with that event. For example, t is the time of the earthquake and x could be the magnitude of an earthquake. The probability density function of x is model specific and can depend on the time of the event, the previous events, all previous events or be fully independent. Let $\mathcal{H}_t$ be the collection of all points (t_i, x_i) until time t (that is, $(t_i, x_i) \in \mathcal{H}_t \forall t_i < t$), then the most general probability density function of x_i at the event time t_i is described as

$$\rho(x_i|t_i, \mathcal{H}_{t_i}), \tag{2}$$

although it is generally not necessary to include the all events from $\mathcal{H}_{t_i}$. In fact, the distribution of x is often modelled differently depending on whether the event is an immigrant or an offspring. Embrechts *et al.* (2011) describe the distribution of the marks as only dependent on the dimension and independent of the past. The marks themselves can also have an effect on the generation of offspring; in the epidemic type aftershock sequences model (Ogata (1988)), a higher magnitude can increase the likelihood of aftershocks). Thus, we adapt the intensity $\lambda(t)$ to include the marks in the most general case (Rasmussen (2013)):

$$\lambda(t) = \mu(t) + \sum_{t_i<t} \phi(t - t_i|x_i). \tag{3}$$

A multidimensional Hawkes process (also called a multivariate Hawkes process) is a point process where points are generated in multiple dimensions d. Note that in a multidimensional

Hawkes process where the dimensions are independent of each other one can simply separate the dimensions and consider a Hawkes process for each dimension individually. Embrechts *et al.* (2011) describe two notations for a multidimensional marked Hawkes process; the *scalar-valued* notation and the *vector-valued* notation. In the *scalar-valued* notation, which I use from now on, each point is noted in the form (t, i, x) where t is the time of the event, i the dimension of the event and x the mark of the event. In this case, each dimension has its individual intensity $\lambda_d(t)$, given in the most general form as

$$\lambda_i(t) = \mu_i(t) + \sum_{j=1}^{d} \sum_{t_i<t} \phi_{ij}(t - t_i|x_i), \tag{4}$$

where i is the dimension of interest, d is the number of dimensions and j the parent dimension. such that each dimension has its own base intensity $\mu_i(t)$ and offspring intensity $\phi_{ij}(t-t_i|x_i)$. Here the intensity in one dimension is dependent on the offspring intensity of all other dimensions. The offspring intensities can be dependent on the direction of the dimension. For example, in a 2-dimensional Hawkes process, the offspring intensity from dimension 2 to dimension 1 is not necessary equal to the offspring intensity from dimension 1 to dimension 2 (i.e. $\phi_{12} \neq \phi_{21}$). Thus, for d dimensions there are (at most) d^2 different offspring intensities. The density of the marks follows similarly. Let $\mathcal{H}_{d,t}$ describe the the collection of all points in dimension d until time t, then the probability density of the mark in dimension d is described in the most general case as

$$\rho_i(x_i|t_i, \sum_{j=1}^{d} \mathcal{H}_{j,t_i}). \tag{5}$$

In this thesis, I will use the following model:

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{d} \sum_{t_i<t} \phi_{ij}(t - t_i|x_i), \tag{6}$$

that is, the immigrant intensity is homogeneous and specific to its dimension, and the offspring intensity is only dependent on the mark of the to be examined event.

Non-parametric Bayesian models are different from a non-parametric frequentist model. Non-parametric frequentist models try to make as few assumptions on the form of the actual model (in contrast to a parametric model, where the form of the model is set and the parameters are fitted to the data). The Bayesian method inherently requires model assumptions in the prior distribution, so a non-parametric Bayesian model avoids assumptions on the data

by modelling the data on an infinite-dimensional parameter space for a given problem. As this is infeasible in reality, practical Bayesian non-parametric models only use a finite subset of the infinite-dimensional parameter space. This infinite-dimensional parameter space consists of a number of random functions or random measures (also called stochastic processes). Typical processes include Gaussian processes, Dirichlet processes and beta processes, although mixtures are also found. Orbanz & Teh (2010) provide a (somewhat dated) overview of non-parametric Bayesian models. In this thesis I make use of a Laplace Bayesian Poisson process, which can be seen as a specific type of Gaussian process.

2.2 Computational algorithm Multidimensional Gibbs-Hawkes

I propose the following Gibbs sampling algorithm, called Multidimensional Gibbs-Hawkes (MGH). First, begin by initializing $\mu_{i,0}$, $\phi_{ij,0}$, $\mathbf{\Sigma}_{i,imm,0}$, $\mathbf{\Sigma}_{ij,0}$ and $\mathbf{\Sigma}_{i,off,0}$ for all $i, j \in D$ randomly. Then, for h_{max} iterations, repeat:

- Sample a branching structure $\mathcal{B}_h$ from $p(\mathcal{B}|\mathcal{S}, \mu_{h-1}, \phi_{h-1})$ (equation (7))
- Calculate $\hat{\omega}$ and $\hat{Q^{-1}}$ from equations (22) and (23) for all $i, j \in D$
- Sample $\mu_{i,h}$ and $\phi_{ij,h}$ from $p(\mu_{i,h}|\mathcal{B}_h, \mathcal{S})$ (equation (11)) and $p(\phi_{ij}|\mathcal{B}_h, \mathcal{S})$ (equation 15) for all $i, j \in D$
- Sample $\beta_{i,h}$, $\mathbf{\Pi}_{ij,h}$ and $\mathbf{\Pi}_{i,h}$ from $p(\beta_i|\mathbf{\Sigma}_{i,imm,h-1}, \mathbf{x}_i)$ (equation (26)), $p(\mathbf{\Pi}_{ij}|\mathbf{\Sigma}_{ij,h-1}, \mathbf{x}_i)$ (equation (31)) and $p(\mathbf{\Pi}_i|\mathbf{\Sigma}_{i,off,h-1}, \mathbf{x}_i)$ (equation (35)) for all $i, j \in D$
- Sample $\mathbf{\Sigma}_{i,imm,h}$, $\mathbf{\Sigma}_{ij,h}$ and $\mathbf{\Sigma}_{i,off,h}$ from $p(\mathbf{\Sigma}_{i,imm}|\beta_{i,h}, \mathbf{x}_i)$ (equation (27)), $p(\mathbf{\Sigma}_{ij}|\mathbf{\Pi}_{ij,h}, \mathbf{x}_i)$ (equation (32)) and $(\mathbf{\Sigma}_{i,off}|\mathbf{\Pi}_{i,h}, \mathbf{x}_i)$ (equation (36)) for all $i, j \in D$

The parameters can then be estimated using (for example) a quadratic loss functions by taking the means of the samples after the sampler has reached convergence. Similar to Zhang *et al.* (2019) I expect the estimated parameters to converge to the true parameters.

2.3 Conditional distribution of the branching structure

The branching structure (see figure 1 for an example) is unknown from the data and must be estimated. I do this in a similar manner to Zhang *et al.* (2019)). First I denote the branching structure as $\mathcal{B}$ (that is, $\mathcal{B}$ is defined as a set structured like in figure 1, with offspring events starting from an immigrant) and the collection of all events as $\mathcal{S}$ (that is, $\mathcal{S}$ is defined as the set of all events). Second, I assume that the triggering events are independent, allowing

the the probability of the full branching structure to be the product of the probabilities of triggering events:

$$p(\mathcal{B}|\mathcal{S},\mu,\phi) = \prod_{i,j\in\mathcal{S}} p_{ij}, \forall t_j < t_i \tag{7}$$

where p_{ij} is the probability of event j triggering event i. With $\mu_j(t_i)$ and $\phi_{ij}(t_i - t_j|x_j)$ given, p_{ij} is explained as the ratio between the offspring intensity and the full intensity, i.e.

$$p_{ij} = \frac{\phi_{ij}(t_i - t_j|x_j)}{\mu_i(t_i) + \sum_{k=1}^{d}\sum_{t_l<t_j}\phi_{ik}(t_i - t_l|x_l)}, j \leq i. \tag{8}$$

To clarify, $\phi_{ij}(t_i - t_j$ is the offspring intensity between event i and event j. The denominator is the sum of all intensities; $\mu_i(t_i)$ is the immigrant intensity of event i and the sum of all offspring intensities across all dimensions d between event i and all events that happened before event j. In short, it is the sum of all relevant intensities. The larger the intensity ϕ_{ij} relative to all other intensities, the more probable that event i was triggered by event j.

Similarly the probability of event i being an immigrant (called p_{i0}) is the ratio between the immigrant intensity and all other relevant intensities:

$$p_{i0} = \frac{\mu_i(t_i)}{\mu_i(t_i) + \sum_{k=1}^{d}\sum_{t_l<t_j}\phi_{ik}(t_i - t_l|x_l)}, j \leq i \tag{9}$$

Thus, to sample a branching structure we sample for each t_i whether it has a parent or is an immigrant according to the probabilities p_{i0} and $p_{ij}\forall j < i$ (where it should be noted that $p_{i0} + \sum_{j<i} p_{ij} = 1$). Doing this for all events gives us a branching structure conditional on the intensities.

2.4 Conditional posterior of the immigrant intensity

The posterior distribution of the immigrant intensity is estimated similar to Zhang *et al.* (2019) and adapted for a multidimensional case. First, let $\mathcal{S}_{\mu,d}$ denote the set of all real immigrants in dimension d. Second, assume that $\mathcal{S}_{\mu,d}$ is generated by a homogeneous Poisson process with intensity μ_d (the immigrant intensity in dimension d). Since the immigrants are independently generated in each dimension, the Poisson likelihood for each dimension is separate. Thus, given a set of events with event times t_i over parameter space $\Omega_d = [0, T]$, the Poisson likelihood for generating immigrants in dimension d is given as the following probability:

$$p(\mathcal{S}_{\mu,d}|\mu_d,\Omega_d) = e^{-\mu_d T}\frac{(\mu_d T)^N_{\mu,d}}{N_{\mu,d}!}, \tag{10}$$

where $N_{\mu,d}$ is the number of immigrant events in $\mathcal{S}_{\mu,d}$. Next, we place a conjugate Gamma prior (which causes the posterior to be Gamma distributed as well), such that $\mu_d T \sim Gamma(\alpha_d, \beta_d)$, which gives the posterior distribution of $\mu_d \sim Gamma(\alpha_d + N_{\mu,d}, \beta_d + 1)$ (see e.g. Fink (1997)). Similar to Zhang *et al.* (2019), by choosing $\alpha_d = N_{\mu,d}$ and $\beta_d = 1$ such that the mean of the posterior is equal to $N_{\mu,d}$ and the variance as 2, we obtain the posterior for the immigrant intensity:

$$p(\mu_d T|\mathcal{S}_{\mu,d}, \alpha, \beta) = Gamma(2N, 2) \tag{11}$$

2.5 Conditional posterior of the offspring intensity

The posterior distribution of the offspring intensity is more involved than the immigrant intensity due to the non-parametric setting. I model the posterior distribution of $\phi_{ij}()$ as a Laplace Bayesian Poisson process (similar to Zhang *et al.* (2019)). The Laplace Bayesian Poisson process is detailed by Walder & Bishop (2017) and uses a Gaussian process prior. Specifically, like Zhang *et al.* (2019) I use their covariance function for thin-plate semi-norms on the hyper-cube. By choosing the Gaussian process (GP) as the prior, we can specify a non-parametric setting. In a Gaussian process all sets of events across some finite parameter space have a multivariate normal distribution. The Gaussian process then is the joint distribution of the collection of all these multivariate normal distribution. In short, the Gaussian process is a distribution of functions and measures the similarity between sets of events. This is convenient since it doesn't rely on assumining a model beforehand, thus avoiding misspecifying the model.

Below is a general description of the Laplace Bayesian Poisson process, as given by Zhang *et al.* (2019). As the Hawkes process require that the general intensity λ is non-negatively valued, a deterministic link function is added, such that the prior over λ is defined as the function composition $\lambda = g \circ f$ (i.e, a deterministic link function $g \circ f = g(f(x))$), where $f \sim GP(k)$ (a Gaussian process) and k is the covariance function for the Gaussian process. Walder & Bishop (2017) use the permanental process for g, that is $g(z) = \frac{1}{2}z^2$ for its computational and analytical advantages over the more common exponential function for $g(z)$.

The covariance function $k(x, y) = \text{Cov}(f(x), f(y))$ (with x and y some dependent variables) can be written as a Mercer expansion (Mercer (1909))

$$k(x, y) = \sum_{i=1}^{K} \xi_i e_i(x) e_i(y), \tag{12}$$

where $K = \infty$ for non-degenerate kernels, ξ_i is some scalar and e_i are chosen as orthonormal eigenfunctions. This Mercer expansion has some convenient convergence properties for positive semi-definite functions. This lets f be represented as a linear combination of the eigenfunctions e_i: $f() = \omega' \mathbf{e}()$, where ω has a Gaussian prior (that is, $\omega \sim \mathcal{N}(0, \Xi)$, with $\Xi = diag(\xi_1, \xi_2, \ldots, \xi_K)$ as a diagonal covariance matrix) and $\mathbf{e}() = [e_1(), e_2(), \ldots, e_K()]'$ is a vector of basis functions.

The posterior distribution of the intensity $\lambda()$ is then equivalent to the posterior distribution of ω in the Laplace Bayesian Poisson process and is approximated by a normal distribution (called a Laplace approximation (Rasmussen (2003))):

$$p(\omega|X, \Omega, k) \approx \mathcal{N}(\omega|\hat{\omega}, Q), \tag{13}$$

where $X \in \mathbf{R}^d$ is a dataset containing points in d dimensions, Ω is the d-dimensional sample space, k is the kernel of the Gaussian process, $\hat{\omega}$ is selected as the mode of the true posterior and Q is the negative inverse Hessian of the true posterior. Both $\hat{\omega}$ and Q are estimates. Thus, to estimate λ, we can use this Laplace approximation, which is computationally simple. Since the $f()$ is ω multiplied by the eigenfunctions $\mathbf{e}()$, the approximate posterior distribution of $f(t)$ is normally distributed, i.e.:

$$f(t) \sim \mathcal{N}(\hat{\omega}' \mathbf{e}(t), \mathbf{e}(t)' Q \mathbf{e}(t)). \tag{14}$$

Then, since $\lambda = g \circ f$ with $g(z) = \frac{1}{2} z^2$, the posterior distribution of is given as

$$\lambda(t) \sim Gamma(\alpha, \beta) \tag{15}$$

with hyper-parameters

$$\alpha = \frac{(\nu^2 + \sigma^2)^2}{4\nu^2\sigma^2 + 2\sigma^4} \tag{16}$$

and

$$\beta = \frac{\nu^2 + \sigma^2}{2\nu^2\sigma^2 + \sigma^4}, \tag{17}$$

where ν is the mean of $f(t)$ (i.e. $\nu = \hat{\omega}' \mathbf{e}(t)$) and σ^2 is the variance of $f(t)$ (i.e. $\sigma^2 = \mathbf{e}(t)' Q \mathbf{e}(t)$).

The above process is a general description using the Laplace Bayesian Poisson process from Walder & Bishop (2017). For the posterior distribution of $\phi_{ij}()$ I apply the above process as follows. Similar to Zhang *et al.* (2019), I assume that $\phi_{ij}(t) = \frac{1}{2} f(t)^2$ and consider for the GP distribution (over the sample domain $[0, \pi]$) the so-called *cosine kernel* by Zhang *et al.*

(2019) (which is described in Walder & Bishop (2017) as a series expansion of the so-called m-th order thin-plate spline semi-norm):

$$k(x, y) = \sum_{\gamma \geq \mathbf{0}} \xi_\gamma e_\gamma(x) e_\gamma(y), \tag{18}$$

$$\xi_\gamma = \frac{1}{a(\gamma^2)^m + b} \tag{19}$$

$$e_\gamma(x) = \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \sqrt{1/2}^{I_{\gamma=0}} \cos(\gamma x) \tag{20}$$

where γ is a multi-index with non-negative (integer) values and $I_\gamma = 0$ is an indicator function, whose value is 1 if an index of γ is 0, and is valued 0 for any non-zero indices of γ. a and b are parameters controlling smoothness and I set $m = 2$. In the experiments from Zhang *et al.* (2019), a and b are set to 0.002. Furthermore they also recommend using 32 basis functions for γ, as this gives a suitable trade off between fitting accuracy and speed. The above is a general description of the Laplace Bayesian Poisson process.

As I am examining a multi-dimensional marked Hawkes process, this next part is an extension of the Gibbs-Hawkes algorithm from Zhang *et al.* (2019). First I condition on the branching structure by considering the *aligned sequences*. An aligned sequence is a set of offspring events in dimension j that originate from some immigrant t_i. Let $S_{t_i,j}$ denote the aligned sequence in dimension j generated by event t_i. An aligned sequence is similar to a branch from figure 1, but moves in only one direction dimensionally. For example, if immigrant t_i is in dimension 1, then $S_{t_i,2}$ is the set that contains all offspring in dimension 2 that can be traced back to immigrant t_i. Using figure 1, suppose that events t_1 and t_3 are in dimension 1 and t_2 and t_4 are in dimension 2, then the aligned sequence $S_{t_1,2}$ contains events t_1, t_2 and t_4. These aligned sequences are used in the log-likelihood of ω.

Secondly, to introduce a dependency on the marks, I define $s_{t_i} = (t_i + \frac{1}{||x_i||})$, a scalar with the time of the event and the magnitude of its mark x_i. As the unidimensional likelihood from Zhang *et al.* (2019) was only dependent on time, this new scalar adds dependency on the marks for the offspring intensity. Since the time between events is inversely related to the offspring intensity, and that a larger magnitude should increase the offspring intensity, I add the inverse of the magnitude so that a larger magnitude causes a larger intensity. I take the magnitude to compensate for the (possible) different dimensions of the marks across dimensions (since the immigrant might spawn events in other dimensions). The joint distribution of ω and the set of all aligned sequences starting in dimension l, $\{S_{t_i,j}\}_l$ (note that

the dimension of all immigrants t_i in this set is l), is given by the log-likelihood

$$\begin{aligned}\log p(\omega, \{S_{t_i,d}\}_l|\Omega_j, k) = \sum_{\{S_{t_i,j}\}} \Big\{ \sum_{\Delta s \in S_{t_i}} \log \frac{1}{2}(\omega' \mathbf{e}(\Delta s))^2 - \frac{1}{2}\int_0^{T-t_i-\frac{1}{||x_i||}} (\omega'\mathbf{e}(s))^2 \mathrm{d}s \Big\} \\ - \frac{1}{2}\log[(2\pi)^K|\Xi|] - \frac{1}{2}\omega'\Xi^{-1}\omega,\end{aligned} \tag{21}$$

where Δs is $(t_i + \frac{1}{||x_i||}) - (t_j + \frac{1}{||x_j||})$ with t_i the immigrant event of sequence $\{S_{t_i,j}\}_l$ and t_j an offspring in the aligned sequence $\{S_{t_i,j}\}_l$. The summation over $\{S_{t_i,j}$ is similar to the likelihood used in Hawkes processes (see for example, equation (2) in Lewis & Mohler (2011)). and the last two terms come from the normal distribution (see equation (7) in Walder & Bishop (2017)). This joint distribution is then used to calculate the approximate log-posterior of ω by using

$$\hat{\omega} = \underset{\omega}{\operatorname{argmax}} \log p(\omega, \{S_{t_i,d}\}_l|\Omega_j, k) \tag{22}$$

and

$$Q_{il}^{-1} = -\sum_{\{S_{t_i,j}\}} \Big\{ \sum_{\Delta s \in S_{t_i}} \frac{2\mathbf{e}(\Delta s)\mathbf{e}(\Delta s)'}{(\hat{\omega}'\mathbf{e}(\Delta s))^2} - \int_0^{T-t_i-\frac{1}{||x_i||}} \mathbf{e}(s)\mathbf{e}(s)'\mathrm{d}s \Big\} + \Xi^{-1}, \tag{23}$$

with Ξ as a matrix with ξ_k on the diagonal. Optimizing equation (22) can be done using L-BFGS (which was developed by Byrd *et al.* (1994), and is proposed by Zhang *et al.* (2019)). An analytical expression of the integral in equation (23) can be found in appendix A.

With these equations we can then draw from equation (14) and estimate a conditional intensity for $\phi_{ij}()$ from the function composition $\lambda = g \circ f$, that is, $\phi_{ij}(t) = \frac{1}{2}f(t)^2$.

2.6 Conditional posterior of the marks

As the distribution of equation (5) is the most general case I make some assumptions for simplicity, without reducing the practicality of the multidimensional marked Hawkes process. The first assumption is that the immigrant and offspring distribution of the marks are different. Specifically, the distributions of the marks of immigrants are independent (i.e. $\rho_d \sim$ I.I.D.). Secondly, I assume that the distributions of the marks of offspring only depends on its parent (i.e. $\rho_d(x_i|t_i, (x_j, t_j))$, where (x_j, t_j) is the parent of (x_i, t_i)), similar to Rasmussen (2013).

In a practical application these assumptions are reasonable as some knowledge on the structure of the marks is usually present. For example, consider the ETAS earthquake models where aftershocks are modeled as offspring of the earthquake (with the magnitude of the

earthquake as mark). It is reasonable to assume that the magnitude of the aftershock is only dependent on the earthquake that came before it. These assumptions on the structure of the marks are not necessary, but the focus of this work is not on the most general case of distributions of the marks.

2.6.1 Conditional posterior of the immigrant marks

For simplicity, I assume that the immigrant mark x_d (a $[m \times 1]$ vector, similar to e.g. Embrechts *et al.* (2011)) has a standard normal distribution with unknown mean and variance. Let $\mathbf{x}_d = [x_{1d}, x_{2d}, \ldots, x_{nd}]$ be the collection of all N immigrant marks in dimension d. The distribution is then $\mathbf{x}_d \sim \mathcal{N}(\mu_d, \mathbf{\Sigma}_{d,imm})$. I use a diffuse prior specification, where

$$p(\mu_d|\mathbf{\Sigma}_{d,imm}) \propto 1, \tag{24}$$

$$p(\mathbf{\Sigma}_{d,imm}) \propto |\mathbf{\Sigma}_{d,imm}|^{-(m+1)/2}, \tag{25}$$

which is a degenerate inverted-Wishart prior. Then we have the conditional posterior distributions for β_d and $\mathbf{\Sigma}_{d,imm}$ given as

$$p(\mu_d|\mathbf{\Sigma}_{d,imm}, \mathbf{x}_i) \sim \mathcal{N}\big(\bar{\mathbf{x}}_d, \mathbf{\Sigma}_{d,imm}\big), \tag{26}$$

$$p(\mathbf{\Sigma}_{d,imm}|\mathbf{x}_i) \sim \mathcal{W}^{-1}\big(\mathbf{S}, N\big), \tag{27}$$

where $\bar{\mathbf{x}}_d = 1/N \sum_{i=1}^{N} \mathbf{x}_{di}$ and $\mathbf{S} = 1/N \sum_{i=1}^{N} (\mathbf{x}_{di} - \bar{\mathbf{x}}_d)(\mathbf{x}_{di} - \bar{\mathbf{x}}_d)'$.

2.6.2 Conditional posterior of the offspring marks

First, I distinguish two cases: when an event has a parent in the same dimension, and when an event has a parent in a different dimension. Since the amount of parameters can be different across dimensions this distinction must be made. In the first case, I will use a multivariate regression model, and in the second case I will use a vector autoregressive model (or VAR model). Other models are possible and can be used when more is known about the data generating process, but these two models are quite general. The first model is described in Greenberg (2012), the second model is described in Schorfheide & Song (2015).

In the first case, let x_i be a $[l \times 1]$ dimensional mark of the offspring in dimension i and x_j be a $[m \times 1]$ dimensional mark of the parent in dimension j. The multivariate regression model is given by

$$\mathbf{x}_i = \mathbf{x}_j \Pi_{ij} + \epsilon_i \tag{28}$$

where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ni})'$ is a $[n \times m]$ matrix, $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})'$ is a $[n \times l]$ matrix, $\epsilon_\mathbf{i} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$ is a $[n \times m]$ matrix and Π_{ij} is a $[l \times m]$ matrix of parameters. Further, the error vector has a multivariate normal distribution (i.e. $\epsilon_i \sim \mathcal{N}(0, \mathbf{\Sigma}_{ij})$, where $\mathbf{\Sigma}_{ij}$ is a $[m \times m]$ covariance matrix). Then $\mathbf{x}_i \sim \mathcal{MN}(\mathbf{x}_j \mathbf{\Pi}_{ij}, \mathbf{\Sigma}_{ij} \otimes I_n)$, which is a matricvariate normal distribution.

Without assuming prior knowledge I use a diffuse prior specification, that is

$$p(\mathbf{\Pi}_{ij}) \propto 1 \tag{29}$$

and

$$p(\mathbf{\Sigma}_{ij}) \propto |\mathbf{\Sigma}_{ij}|^{-(k+1)/2}, \tag{30}$$

which is a degenerate inverted-Wishart prior. These give the following full conditional distributions:

$$p(\mathbf{\Pi}_{ij}|\mathbf{\Sigma}_{ij}, \mathbf{x}_i) \sim \mathcal{MN}\big((\mathbf{x}_j'\mathbf{x}_j)^{-1}\mathbf{x}_j'\mathbf{x}_i, \mathbf{\Sigma}_{ij} \otimes (\mathbf{x}_j'\mathbf{x}_j)^{-1}\big), \tag{31}$$

$$p(\mathbf{\Sigma}_{ij}|\mathbf{\Pi}_{ij}, \mathbf{x}_i) \sim \mathcal{W}^{-1}\big((\mathbf{x}_i - \mathbf{x}_j\mathbf{\Pi}_{ij})'(\mathbf{x}_i - \mathbf{x}_j\mathbf{\Pi}_{ij}), n\big), \tag{32}$$

For the second case where the parent and offspring events are in the same dimension I use a simple VAR(1) model. It can be written in a similar format as equation (28), where x_d is the $[m \times 1]$ dimensional offspring mark of an event in dimension d, and $x_{d,-1}$ is the $[m \times 1]$ dimensional mark of the parent (i.e. the lags). Then the VAR(1) model is given as

$$\mathbf{x}_d = \mathbf{x}_{d,-1}\mathbf{\Pi}_d + \epsilon_\mathbf{i}, \tag{33}$$

where $\mathbf{x}_d = (x_{1d}, x_{2d}, \ldots, x_{nd})'$ is a $[n \times m]$ matrix with all the offspring marks, $\mathbf{x}_{d,-1} = (1, x_{1d,-1}, x_{2d,-1}, \ldots, x_{nd,-1})'$ is a $[n \times (1+m)]$ matrix with all the parent marks, $\epsilon_\mathbf{i} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$ is a $[n \times m]$ matrix and $\Pi_d = (\alpha_d', \phi_d')'$ is a $[(1+m) \times m]$ matrix of parameters. Here α' is a $[m \times 1]$ vector of intercepts and ϕ_d is a $[m \times m]$ matrix of autoregressive parameters. The error vector has a multivariate normal distribution (i.e. $\epsilon_i \sim \mathcal{N}(0, \mathbf{\Sigma}_{d,off})$, where $\mathbf{\Sigma}_{d,off}$ is a $[m \times m]$ covariance matrix). Assuming no prior information, I use a diffuse priors, that is a uniform distribution for $\mathbf{\Pi}_d$ and Jeffreys' prior for $\mathbf{\Sigma}_{d,off}$:

$$p(\mathbf{\Pi}_d, \mathbf{\Sigma}_{d,off}) \propto |\mathbf{\Sigma}_{d,off}|^{-(m+1)/2}. \tag{34}$$

With these priors the parameters $\mathbf{\Pi}_d$ have a matricvariate normal distribution and the variance $\mathbf{\Sigma}$ has an inverted Wishart distribution:

$$p(\mathbf{\Pi}_d|\mathbf{x}_d, \mathbf{\Sigma}_{d,off}) \sim \mathcal{MN}\big((\mathbf{x}_{d,-1}'\mathbf{x}_{d,-1})^{-1}\mathbf{x}_{d,-1}'\mathbf{x}_d, \mathbf{\Sigma}_{d,off} \otimes (\mathbf{x}_{d,-1}'\mathbf{x}_{d,-1})^{-1}\big), \tag{35}$$

$$p(\mathbf{\Sigma}_{d,off}|\mathbf{\Pi}_d, \mathbf{x}_d) \sim \mathcal{W}^{-1}\big((\mathbf{x}_d - \mathbf{x}_{d,-1}\mathbf{\Pi}_d)'(\mathbf{x}_d - \mathbf{x}_{d,-1}\mathbf{\Pi}_d), n - m\big). \tag{36}$$

More details can be found in Schorfheide & Song (2015).

3 Simulation study

In this section I describe the simulation study I run to study the proposed algorithm. First I describe the data generating processes (DGP) I use. Then I describe the two non-Bayesian algorithms (a parametric algorithm and a non-parametric algorithm) that I use to compare the proposed algorithm to. Lastly I describe how I will evaluate the proposed algorithm.

3.1 Data Generating Processes

I simulate two data generating processes (DGP) to test the proposed algorithm on. The first DGP uses exponential decays functions for the offspring intensities, and is a common decay function for Hawkes processes. The second DGP I use power decay functions fot the offspring intensities. To simulate the DGP I use a multivariate extension of *Ogata's modified thinning algorithm* (Daley & Vere-Jones (2003)), which is described in Liniger (2009).

For both simulations I consider a simple two-dimensional DGP. The DGP is sampled on the time-frame $t \in [0, 10]$. One of the processes has a 3-dimensional mark, the other has a 2-dimensional mark. The parameters for the DGPs are chosen such that the DGP generates between 100 and 200 events. The first DGP uses exponential decay functions for the offspring intensities, which have the form of

$$\phi(t) = \alpha \exp(-\beta t).$$

The parameter α is given by the norm of the mark of the parent event (i.e. $\alpha_{ij} = ||\mathbf{x}_{j,-1}||$ where $\mathbf{x}_{j,-1}$ is the mark of the parent). The marks for immigrants are drawn from a multivariate normal distribution. The marks for offspring are drawn using an autoregressive model:

$$\mathbf{x}_i = \gamma \mathbf{x}_j + \epsilon,$$

where the regressor is the parent event and the dependent variable the child event and the error ϵ is standard normally distributed (i.e. $\epsilon \sim \mathcal{N}(0, I)$. The exact parameters are as follows:

$$
\begin{aligned}
\mu_1(t) =& 2 \\
\mu_2(t) =& 1.5 \\
\phi_{11}(t) =& ||\mathbf{x}_{1,-1}|| \exp(-5t) \\
\phi_{12}(t) =& ||\mathbf{x}_{2,-1}|| \exp(-2t) \\
\phi_{21}(t) =& 0 \\
\phi_{22}(t) =& ||\mathbf{x}_{2,-1}|| \exp(-8t) \\
\gamma_1 \sim& \mathcal{N}((1,2,3)', I) \\
\gamma_2 \sim& \mathcal{N}(0, 2I) \\
\gamma_{11} : \mathbf{x}_1 =& 0.8\mathbf{x}_{1,-1} + \epsilon \\
\gamma_{12} : \mathbf{x}_1 =& 0.5\mathbf{x}_2 + \epsilon \\
\gamma_{21} =& 0 \\
\gamma_{22} : \mathbf{x}_2 =& 0.3\mathbf{x}_{2,-1} + \epsilon,
\end{aligned}
$$

where in γ_{11}, γ_{12} and γ_{22}, $\mathbf{x}_1, \mathbf{x}_1$ and $\mathbf{x}_2$ are the marks of offspring and $\mathbf{x}_{1,-1}$ and $\mathbf{x}_{2,-1}$ are the marks of their respective parents, $\epsilon \sim \mathcal{N}(0, I)$ and I is the identity matrix. Note that the offspring intensity for $\phi_{12} = 0$, meaning that there are no offspring events in dimension 2 whose parents are in dimension 1.

For the second DGP, I use power decay functions for the offspring intensities, which have the form of

$$\phi(t) = \alpha(1 + b * t)^{-\beta}.$$

The idea for using power decay functions is that the parametric algorithm is expected to have more trouble with this DGP. Similar to the first DGP, I choose α as the norm of the mark of the parent event (i.e. $\alpha_{ij} = ||\mathbf{x}_{j,-1}||$ where $\mathbf{x}_{j,-1}$ is the mark of the parent), and b

and β are chosen. The exact parameters are as follows:

$$\begin{aligned}
\mu_1(t) =& 2 \\
\mu_2(t) =& 1.5 \\
\phi_{11}(t) =& ||\mathbf{x}_{1,-1}||(1+2t)^{-2} \\
\phi_{12}(t) =& ||\mathbf{x}_{2,-1}||(1+2t)^{-3} \\
\phi_{21}(t) =& ||\mathbf{x}_{1,-1}||(1+2t)^{-1.5} \\
\phi_{22}(t) =& ||\mathbf{x}_{2,-1}||(1+8t)^{-2} \\
\gamma_1 \sim& \mathcal{N}((1,2,3)', I) \\
\gamma_2 \sim& \mathcal{N}(0, 2I) \\
\gamma_{11} : \mathbf{x}_1 =& 0.8\mathbf{x}_{1,-1} + \epsilon \\
\gamma_{12} : \mathbf{x}_1 =& 0.5\mathbf{x}_2 + \epsilon \\
\gamma_{21} : \mathbf{x}_2 =& 2\mathbf{x}_1 + \epsilon \gamma_{22} : \mathbf{x}_2 = \quad 0.3\mathbf{x}_{2,-1} + \epsilon,
\end{aligned}$$

where in $\gamma_{11}, \gamma_{12}, \gamma_{21}$ and γ_{22}, $\mathbf{x}_1$ and $\mathbf{x}_2$ are the marks of offspring, $\mathbf{x}_{1,-1}$ and $\mathbf{x}_{2,-1}$ are the marks of their respective parents, $\epsilon \sim \mathcal{N}(0, I)$ and I is the identity matrix.

3.2 Algorithm evaluation

Since there are no equivalent Bayesian estimation methods for multi-dimensional marked Hawkes processes, I will compare my algorithm to two non-Bayesian algorithms suited for multidimensional Hawkes processes: a parametric exponential Hawkes model that is estimated using maximum likelihood (that I will call Exp-ML and is similar to what can be found in Embrechts *et al.* (2011)) and the MPLE algorithm from Lewis & Mohler (2011) for non-parametric models. Both these algorithms are available in the `tick` package for `Python` (Bacry *et al.* (2017)). The `tick` package is a library for machine learning techniques. I specifically use the `tick.hawkes` module, which is a collection of algorithms for Hawkes processes.
The MGH algorithm is programmed in `Python`. To optimize the log-likelihood, I use the L-BFGS algorithm implemented in the `SciPy` package. The `SciPy` package is a `Python` library with a large collection of modules for scientific and technical computing. The L-BFGS algorithm is found in the optimization module.

The Exp-ML algorithm uses the following model for the offspring intensities:

$$\phi_{ij}(t) = \alpha^{ij}\beta^{ij}\exp\left(-\beta^{ij}t\right). \tag{37}$$

It is important to note that in the implementation in `tick` requires that β is specified beforehand, and that the algorithm estimates α. I expect the Exp-ML algorithm to be thus quite accurate for DGP 1 and less so for DGP 2. Since α is dependent on the norm of the mark of the parent event, estimation of α is not straightforward however.

The MPLE algorithm is a non-parametric EM-algorithm from Lewis & Mohler (2011), that specifies the following intensity for the Hawkes process:

$$\lambda_i = \mu_i + \sum_{j=1}^{D} \int \phi_{ij} dN_j \tag{38}$$

where D is the number of dimensions, μ_i is the immigrant intensity of dimension i, ϕ_{ij} is the offspring intensity of dimension j going to dimension i add N_j is the sum of events in dimension j. The algorithm uses a maximum penalized likelihood estimation (MPLE) for the maximization step. This maximum penalized likelihood maximizes the following equations:

$$\sum_{i-1}^{n} p_{ii} \log(\mu(t_i)) + \int_0^T \mu(t)dt + \alpha_1 ||(\mu^{1/2})'||_2^2 \tag{39}$$

for the immigrant intensity with ≥ 0 and α_1 a penalty parameter, and

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} \big(p_{ij} \log(\phi(t_i - t_j)) - \int_{t_j}^{T} \phi(t - t_j)dt \big) + \alpha_2 ||(^{1/2})'||_2^2 \tag{40}$$

for the offspring intensity with $\phi \geq 0$. This is done by solving an Euler-Lagrange equation which maximizes the penalized likelihood which works quite efficient.

First, a dataset is simulated from the DGP. Next, the parameters are estimated using MGH, the MPLE algorithm and the Exp-ML method. To compare the frequentist models to my Bayesian model, and since the true DGP is known, I can take the squared distance over the true parameters and estimated parameters by adapting the l_2 distance from Zhang *et al.* (2019):

$$d(g(t), \hat{g}(t)) = \Big(\int_{\Omega} (g(t) - \hat{g}(t))^2 dt \Big)^{1/2}, \tag{41}$$

where $g(t)$ is a function generating an intensity (i.e. $g(t)$ is some $\phi_{ij}(t)$), $\hat{g}(t)$ is the estimated $\hat{\phi}_{ij}$ and Ω is the parameter space). In this case, $\Omega = [0, 10]$. Additionally I plot the estimated intensities of ϕ_{ij} of all the algorithms against the true offspring intensity.

4 Results simulation study

This section discusses the results of the simulation study. First I present discuss the key findings of the simulation study on the MGH algorithm. Then I discuss for both simulations a comparison between the true and estimated offspring intensity by MGH with regard to both time since previous event and the mark of the previous event. I compare the the true and estimated offspring intensities of both MGH and the benchmarks next for both simulations. Then I discuss the l_2 distances for both simulations. Then I discuss the estimates on the parameters on the marks for both simulations.

4.1 Key findings

For the proposed algorithm to work, I must set some parameters. The first are for ξ_γ in the *cosine kernel* (from equation (19)), where I set $m = 2, a = b = 0.002$. To improve computation speed, I set 8 basis functions for γ. This is lower than the 32 basis functions Zhang *et al.* (2019) used, but computation time proved quite significant, hence my choice for less basis functions. The basis functions are $0, 1, \ldots, 7$. I also choose to run 5000 iterations, of which the first 1000 are discarded as burn-ins. A full list of all events can be found in appendix D.

One of the main results of Zhang *et al.* (2019) is that the Gibbs-Hawkes algorithm has linear computational complexity (meaning that the computation time increases linearly with the number of datapoints). It is important to note that this is due to using Halpin's procedure (in Halpin (2012)). Halpin's procedure is used to reduce the computational complexity of Expectation-Maximization type algorithms for Hawkes processes. It does so by dividing the data into branches and treating each branch as an independent Poisson process. They then introduce a missing value to denote to which specific branch an event belongs to and estimate this missing value using an Expectation-Maximization algorithm. As Halpin's procedure is for Expectation-Maximization type algorithm, I did not use it, thus I cannot guarantee the linear computational complexity of MGH.

The first conclusion from the simulation study is that the algorithm works well in estimating the offspring intensities compared to the benchmarks. Especially going from the l_2 distances, the MGH algorithm performs sometimes better and sometimes worse than the benchmarks. The graphs also show that the MGH algorithm is able to give feasible estimates. A solution to obtain better estimates is to use more basis functions for γ (i.e. 32 instead of the 8 used in the simulation study), but this increases computation time.

Secondly, Gibbs-Hawkes and MGH rely on numerically optimizing the likelihood to find $\hat{\omega}$ from equation (22). This is very time consuming, and took 99.98% of computation time of a single iteration. Although runtime cannot be directly compared across systems, running the datasets of the simulation study took eight hours, whereas runtime on the benchmarks was near instantaneous.

Lastly, in the simulation setup there were not enough datapoints to find good estimates for the mark parameters. Although they are not the focus of this thesis, I will note that more datapoints in each dimension than used in this thesis should be used for improved results.

4.2 Relation between time and marks offspring intensity

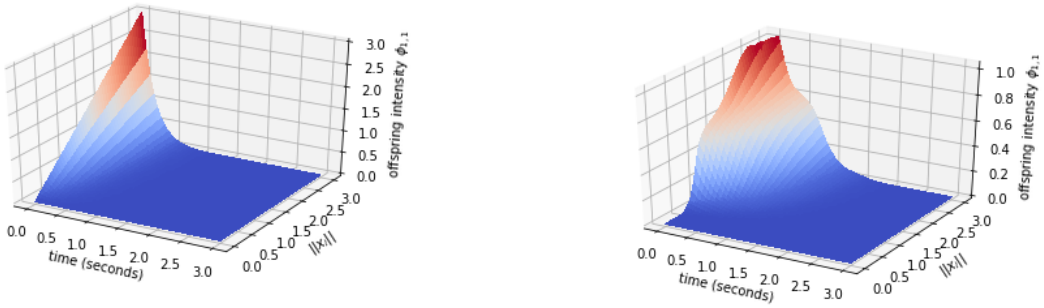
In this section I discuss the relation between the offspring intensity, time (in seconds) and the norm of the mark of the previous event, which I show in a 3-d surface plot. I compare this relation for the true offspring intensity to the estimated offspring intensity, first for dataset 1 and then for dataset 2.

4.2.1 Dataset 1

Figure 2: Relation between offspring intensity ϕ_{11}, time in seconds and the norm of the mark of the previous event for dataset 1

(a) Surface plot of the true offspring intensity ϕ_{11}.

(b) Surface plot of the MGH estimated offspring intensity ϕ_{11}.



Note: surface plots of the relation between time, the marks and the offspring intensities On the x-axis is the time in seconds, on the y-axis the norm of the mark of the parent, and on the z-axis the offspring intensity ϕ_{11}

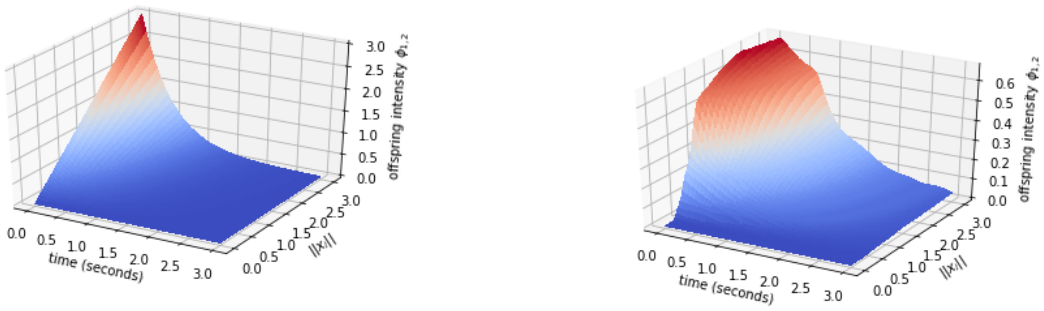
The first dataset is generated using exponential decay functions for the offspring inten-

sities. The offspring intensity is dependent on both the norm of the mark of the parent and the β-parameter. First, in figure 2b the MGH estimate somewhat follows the curved shape of the true offspring intensity in figure 2a. The true offspring intensity shows a linear relation between the norm of the mark of the parent event, which the MGH algorithm does not show. A reason for this could be the lack of data on smaller marks (whith norms < 0.5). Another more likely reason could be that in the log-likelihood from equation (21) I use s instead of t, and the norm of the mark means that there is never a small time difference.
Another difference is the height of the peak of the true offspring intensity, which is three times higher than the estimated intensity. The same explanation is possible here; the estimated intensities are not high enough for small time differences, but this becomes lesser with larger time differences. For example, the estimated offspring intensity in figure 2b at $t = 0.5$ is about equal to the true offspring intensity in 2a at $t = 0.5$. It is again possible that there are not enough events with such small time differences.

Figure 3: Relation between offspring intensity ϕ_{12}, time in seconds and the norm of the mark of the previous event for dataset 1

(a) Surface plot of the true offspring intensity ϕ_{12}.

(b) Surface plot of the MGH estimated offspring intensity ϕ_{12}.



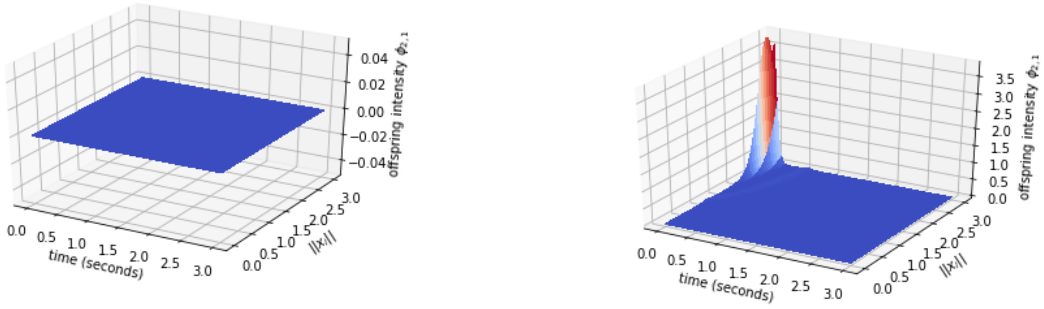
Note: surface plots of the relation between time, the marks and the offspring intensities On the x-axis is the time in seconds, on the y-axis the norm of the mark of the parent, and on the z-axis the offspring intensity ϕ_{12}

A similar thing can be seen in in figure 3b. The relation to the norms is not quite linear, and the estimated offspring intensity is much lower compared to the true offspring intensity at smaller time differences. This figure estimate is worse compared to the estimate for ϕ_{11}.

Figure 4: Relation between offspring intensity ϕ_{21}, time in seconds and the norm of the mark of the previous event for dataset 1

(a) Surface plot of the true offspring intensity ϕ_{21}.

(b) Surface plot of the MGH estimated offspring intensity ϕ_{21}.



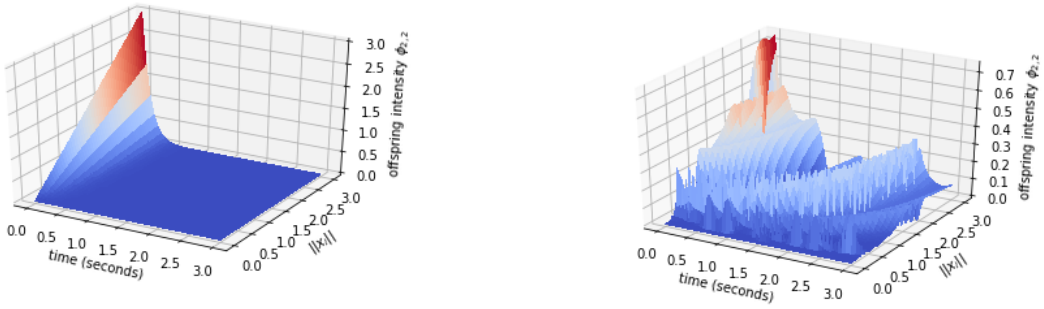
Note: surface plots of the relation between time, the marks and the offspring intensities On the x-axis is the time in seconds, on the y-axis the norm of the mark of the parent, and on the z-axis the offspring intensity ϕ_{21}

MGH was unable to adequately detect that the true offspring intensity ϕ_{21} is 0. Although the estimates go to zero quickly with larger time differences, it is not exactly zero, as is seen in figure 4b.

Figure 5: Relation between offspring intensity ϕ_{22}, time in seconds and the norm of the mark of the previous event for dataset 1

(a) Surface plot of the true offspring intensity ϕ_{11}.

(b) Surface plot of the MGH estimated offspring intensity ϕ_{22}.



Note: surface plots of the relation between time, the marks and the offspring intensities On the x-axis is the time in seconds, on the y-axis the norm of the mark of the parent, and on the z-axis the offspring intensity ϕ_{22}

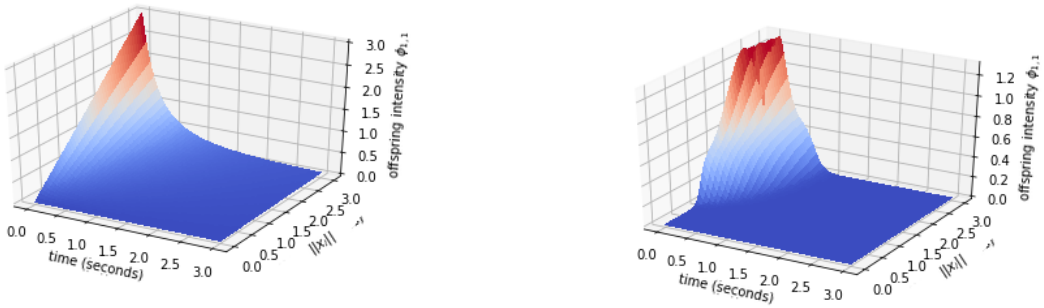
The shape in figure 5b is very different from the true shape in figure 5a. The true relation is a smooth curve, but 5a behaves more wavelike. Here the non-near-zero estimates are also off for larger time differences (say $t > 1$). Again, the MGH estimates are lower than the true offspring intensities for small time differences.

4.2.2 Dataset 2

Figure 6: Relation between offspring intensity ϕ_{11}, time in seconds and the norm of the mark of the previous event for dataset 2

(a) Surface plot of the true offspring intensity ϕ_{11}.

(b) Surface plot of the MGH estimated offspring intensity ϕ_{11}.



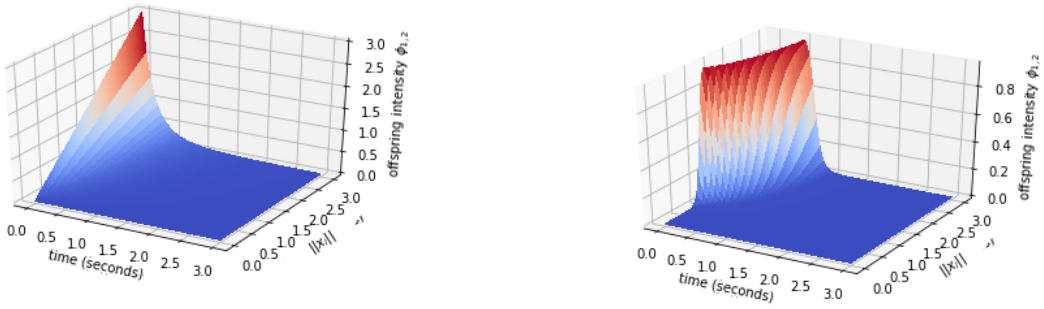
Note: surface plots of the relation between time, the marks and the offspring intensities On the x-axis is the time in seconds, on the y-axis the norm of the mark of the parent, and on the z-axis the offspring intensity ϕ_{11}

The estimated results are similar to those in dataset 1. The clearly linear relation between offspring intensity and the norm of the mark of the previous event in figure 6a is not seen in the estimate in figure 6b, and the height is also about three times lower for small time differences and large norms of the marks of the previous event.

Figure 7: Relation between offspring intensity ϕ_{12}, time in seconds and the norm of the mark of the previous event for dataset 2

(a) Surface plot of the true offspring intensity ϕ_{12}.

(b) Surface plot of the MGH estimated offspring intensity ϕ_{12}.



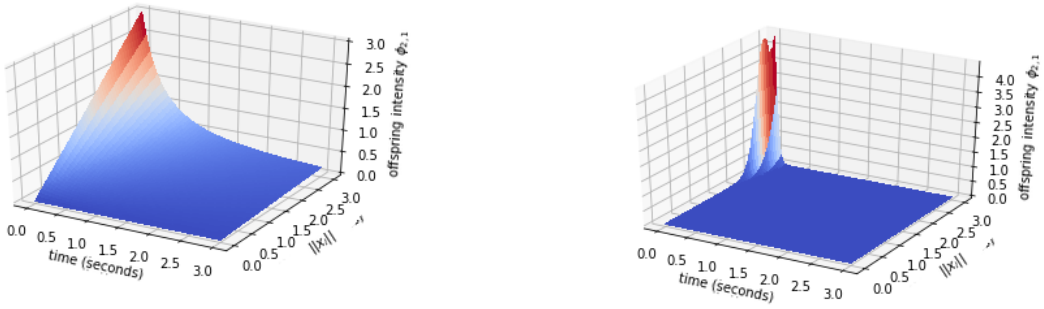
Note: surface plots of the relation between time, the marks and the offspring intensities On the x-axis is the time in seconds, on the y-axis the norm of the mark of the parent, and on the z-axis the offspring intensity ϕ_{12}

The linear relation between the norm of the mark of the previous event is even less present in figure 7a compared to figure 7b, and there is a much sharper drop in the relation between time and the offspring intensity. A reason could be due to the shape of power decay functions having a sharper drop, but difference is quite large.

Figure 8: Relation between offspring intensity ϕ_{21}, time in seconds and the norm of the mark of the previous event for dataset 2

(a) Surface plot of the true offspring intensity ϕ_{21}.

(b) Surface plot of the MGH estimated offspring intensity ϕ_{21}.



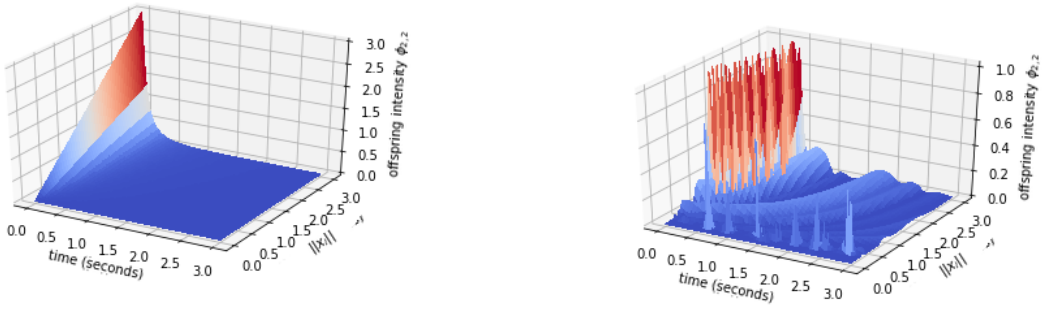
Note: surface plots of the relation between time, the marks and the offspring intensities On the x-axis is the time in seconds, on the y-axis the norm of the mark of the parent, and on the z-axis the offspring intensity ϕ_{21}

This time the estimated offspring intensity in figure 8a is closer in height in figure 8b compared to previous figures. Instead of a curved shape the estimate is more a peak. It is possible that the estimated offspring events in this dimension only had large norms and little time between previous events, and that more data could solve this.

Figure 9: Relation between offspring intensity ϕ_{22}, time in seconds and the norm of the mark of the previous event for dataset 2

(a) Surface plot of the true offspring intensity ϕ_{22}.

(b) Surface plot of the MGH estimated offspring intensity ϕ_{22}.



Note: surface plots of the relation between time, the marks and the offspring intensities On the x-axis is the time in seconds, on the y-axis the norm of the mark of the parent, and on the z-axis the offspring intensity ϕ_{22}

For ϕ_{22} the true offspring intensity has a large drop when time increases (as seen in 9b), which is not captured by the estimates from MGH. Again the shape of the surface in 9a shows a wave-like pattern that does not approach the shape of the true offspring intensity.

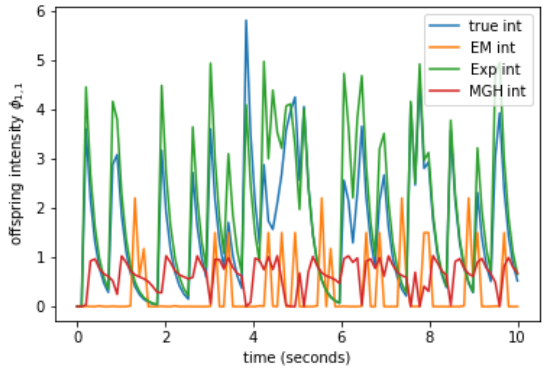
4.3 Comparison to benchmarks

To compare the benchmarks against the MGH algorithm I plot the true and estimated offspring intensities for each case. Each plot shows the offspring intensity of ϕ_{ij} over time in seconds on the interval $[0, 10]$. First I discuss the first dataset and then I discuss the second dataset.

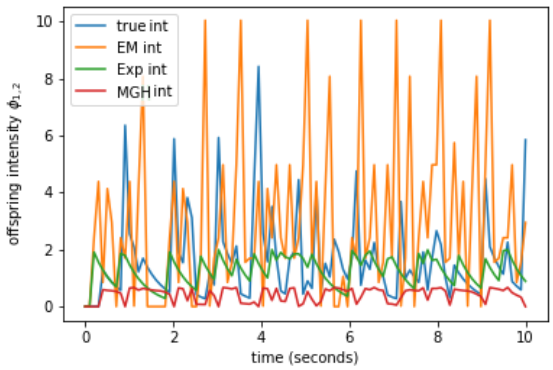
4.3.1 Dataset 1

Figure 10: Offspring intensities over time (seconds) on $[0, 10]$ for dataset 1

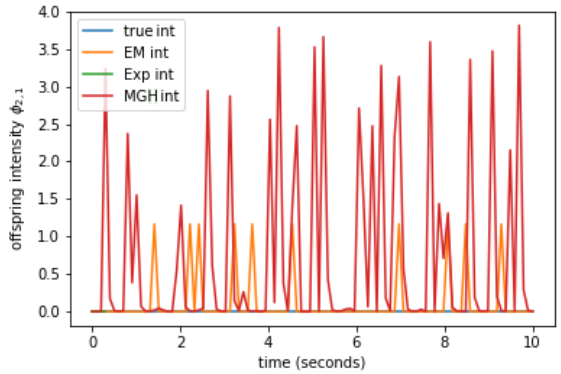
(a) True and estimated offspring intensities ϕ_{11} on $[0, 10]$.



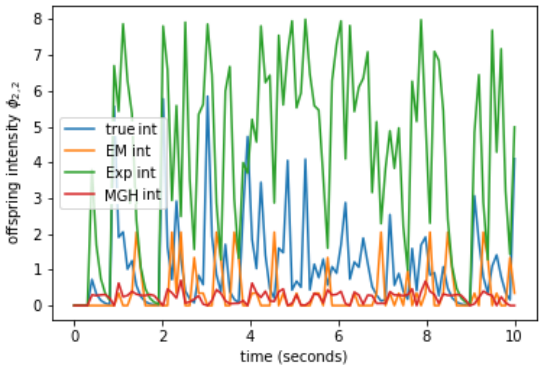
(b) True and estimated offspring intensities ϕ_{12} on $[0, 10]$.



(c) True and estimated offspring intensities ϕ_{21} on $[0, 10]$.



(d) True and estimated offspring intensities ϕ_{22} on $[0, 10]$.



The offspring intensities over time. For each figure, the blue line is the true offspring intensity, the orange line is the estimated offspring intensity by the MPLE algorithm (labeled EM in the legend), the green line is the estimated offspring intensity by the Exp-ML algorithm and the red line is the estimated offspring intensity by the MGH algorithm. On the x-axis is the time in seconds and on the y-axis is the offspring intensity.

In figure 10 the offspring intensities are plotted over time. It can here clearly be seen that the Exp-ML algorithm performs very similar to the true algorithm, especially in figure 10a. Similarly, due to the model in equation (37), the estimates when $\beta = 0$ is also straightforward. It does underestimate the offspring intensity in figure 10b, and overestimates the offspring intensity in figure 10d.

The MPLE algorithm performed worse in all figures. The main reason is that it is not

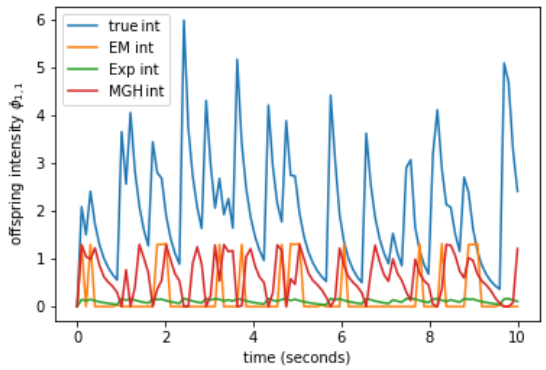
accurately able to detect the parent-child combinations (i.e. when events happen in the graphs). For example, in figure 10a, two events happen before $t = 1$ (as seen by the blue peaks), but the MPLE estimate is 0. It does however give better estimates in the heights of the offspring intensity compared to MGH. It also overestimated the number of events from dimension 1 to dimension 2 (seen in figure 10c)

The MGH algorithm is generally able to detect the correct parent-child combinations except from dimension 1 to dimension 2 (figure 10c), and is worse than the MPLE algorithm. Additionally, the offspring intensity estimates are too low for all figures. One possible reason could be that MGH overestimated the immigrant intensity (see table 3 in the next section), which could cause the offspring intensities to be too low since this will overestimate the number of immigrants in that dimension. This will in turn cause the number of parent-child combinations to be lower, hence the worse estimates.

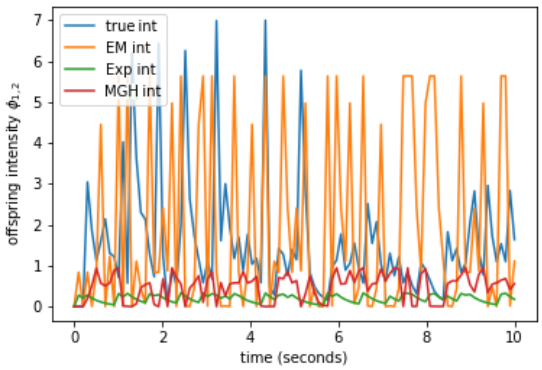
4.3.2 Dataset 2

Figure 11: Offspring intensities over time (seconds) on $[0, 10]$ for dataset 2

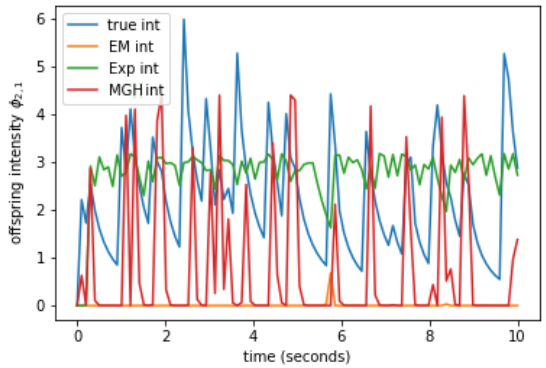
(a) True and estimated offspring intensities ϕ_{11} on $[0, 10]$.



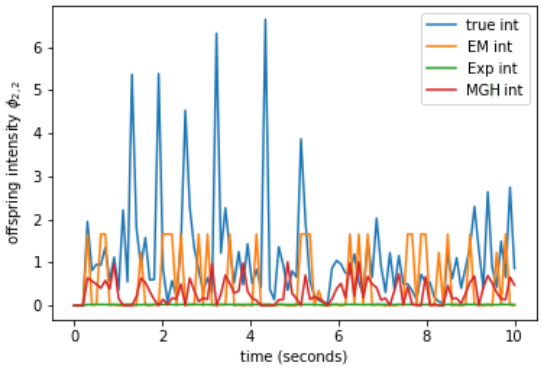
(b) True and estimated offspring intensities ϕ_{12} on $[0, 10]$.



(c) True and estimated offspring intensities ϕ_{21} on $[0, 10]$.



(d) True and estimated offspring intensities ϕ_{22} on $[0, 10]$.



Note: the offspring intensities over time. For each figure, the blue line is the true offspring intensity, the orange line is the estimated offspring intensity by the MPLE algorithm (labeled EM in the legend), the green line is the estimated offspring intensity by the parametric ML algorithm and the red line is the estimated offspring intensity by the MGH algorithm. On the x-axis is the time in seconds and on the y-axis is the offspring intensity.

The comparison of offspring intensities to the benchmarks is given in figure 11. In this dataset, where the true intensity is given by power decay functions, you can see that Exp-ML clearly performs much worse. It underestimates the height of the offspring intensity in figures 11a, figure 11b and in figure 11d it estimated no events. Furthermore it seems unable to respond correctly to events happening.

Similar performance as in the first dataset is seen for the MPLE algorithm. It estimated no events in figure 11c. It gave better estimates for the offspring intensity in figure 11b,

but gave similar estimates in the other figures. It also did not seem to correctly guess when events happened.

The effect of the shape of the estimated surface in figure 8b can be seen in figure 11c. Here MGH correctly estimates the height when events happen, but the offspring intensity drops off too quick. It clearly performs better than MPLE and Exp-ML. MGH does again underestimate the offspring intensity in the other figures. Again MGH overestimated the immigrant intensity for dimension 2 as seen in table 4 in the next section. It does show a much better response to events happening than the benchmark algorithms.

4.4 l_2 distances of the algorithms

In this section I compare the l_2 distances from equation (41) for each algorithm. First for dataset 1, then for dataset 2.

4.4.1 Dataset 1

Table 1: l_2 distance of offspring intensities for dataset 1. The lowest distances are in bold.

ϕ_{ij}	MGH	MPLE	Exp-ML
ϕ_{11}	**5.802**	6.929	6.494
ϕ_{12}	7.545	12.874	**4.303**
ϕ_{21}	4.285	13.203	**0**
ϕ_{22}	6.211	**5.430**	12.417

Despite that the height of the peaks in figure 10 is too low for MGH, looking at the l_2 distances from equation (41) shows that MGH does perform quite well. The results are given in table 1. The MGH algorithm shows good performance, especially compared to the MPLE algorithm, and there are no cases where the l_2 distance is particularly high.

4.4.2 Datset 2

Table 2: l_2 distance of offspring intensities for dataset 2. The lowest distances are in bold.

ϕ_{ij}	MGH	MPLE	Exp-ML
ϕ_{11}	**6.104**	7.704	7.979
ϕ_{12}	**7.028**	8.425	7.155
ϕ_{21}	6.699	7.892	**4.204**
ϕ_{22}	5.564	**4.541**	4.596

From table 2 the performance of MGH is reflected in the l_2 distances. MGH outperforms both MPLE and Exp-ML in the first two dimensions, whilst Exp-ML performs better for ϕ_{21} and MPLE performs better for ϕ_{22}. MGH give similar values for the l_2 distances.

4.5 Posterior means of the marks

In this section I discuss the posterior means of the marks, again first for dataset 1 and second for dataset 2.

4.5.1 Dataset 1

The lack of events in any of the categories for the offspring marks is a problem for estimating the parameters for the marks. Posterior means for $\mathbf{\Pi}_{21}$ are missing because there were not enough parent-child combinations for this dimension, and thus not enough events could be used to make a prediction. The posterior variance for all mark parameters is high and most posterior means are either insignificant or unreasonable. Hence tables with the posterior means of the mark parameters can be found in appendix B.1.

The exception to this is table 3. This contains the MGH posterior means for the immigrant intensities μ_1 and μ_2. From the posterior The posterior means for the parameters of the marks are given in tables 5 to 14. The immigrant mark parameters are close for dimension 1, but are overestimated for dimension 2. This could also affect the estimation of the offspring intensities, as discusses earlier in section 4.3.
12 to 14)

Table 3: Posterior mean and variance μ_1 and μ_2

	True parameter	Posterior mean	Posterior variance
μ_1	2	2.761	0.421
μ_2	1.5	16.044	0.979

4.6 Dataset 2

The same problems for estimations for the parameters for the marks in dataset 1 are present in the estimates for dataset 2. Again the variance for the offspring marks is too high to make any conclusions on the estimates. All the estimates are given in appendix B.2. Table 4 gives the posterior means of the immigrant intensities μ_1 and μ_2. The posterior mean for both μ_1 and μ_2 are worse than for dataset 1. Similarly, the posterior mean for μ_1 is much closer to the true value than μ_2. It might also affect the estimation of the offspring intensities similar to dataset 1..

Table 4: Posterior mean and variance μ_1 and μ_2

	True parameter	posterior mean	posterior variance
μ_1	2	3.156	0.586
μ_2	1.5	17.932	1.024

5 Conclusion

In this thesis I expanded the *Gibbs-Hawkes* algorithm made by Zhang *et al.* (2019) to allow non-parametric Bayesian inference for multidimensional marked Hawkes processes. I adapted the algorithm (called Multidimensional Gibbs-Hawkes) by adapting the log-likelihood to use the norm of the marks in order to account for the influence of the marks on the offspring intensity.

Simulations have shown that the Multidimensional Gibbs-Hawkes algorithm performs sometimes better and sometimes worse to two benchmark algorithms. It is able to account for different offspring intensity kernels due to the non-parametric nature as well when the offspring intensity kernel is related to the marks of the events.

However, due to the unreasonably long computation time compared to the benchmarks,

it is not a practical algorithm to use. Especially when using more basis functions, the computation time is not really worth the slightly better results compared to the benchmark. Either a faster numerical optimization scheme or a different way to optimize ω is needed to make this algorithm better for practical use. This would be the best direction for improving the MGH algorithm in further research.

Another problem is that the algorithm consistently underestimates the offspring intensity and overestimates the immigrant intensity. This causes the algorithm to assign more events as immigrants in sampling the branching structure, which in turn causes an increase in the immigrant intensity. It is possible that more data improves these estimates, but further research could go into the relation between the marks and the time. More specifically, how s and Δs are calculated for the log-likelihood in equation (21) can be an interesting avenue for further research into different forms of this relation. This could also help improve situations with a very small time difference and large norms of the parent marks.

References

Aït-Sahalia, Yacine, Cacho-Diaz, Julio, & Laeven, Roger JA. 2015. Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics*, **117**(3), 585–606.

Bacry, Emmanuel, Bompaire, Martin, Deegan, Philip, Gaïffas, Stéphane, & Poulsen, Søren V. 2017. Tick: a Python library for statistical learning, with an emphasis on hawkes processes and time-dependent models. *The Journal of Machine Learning Research*, **18**(1), 7937–7941.

Byrd, Richard H, Nocedal, Jorge, & Schnabel, Robert B. 1994. Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming*, **63**(1-3), 129–156.

Carstensen, Lisbeth, Sandelin, Albin, Winther, Ole, & Hansen, Niels R. 2010. Multivariate Hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics*, **11**(1), 456.

Daley, Daryl J, & Vere-Jones, David. 2003. *An introduction to the theory of point processes: volume 1.* Springer Science & Business Media.

Donnet, Sophie, Rivoirard, Vincent, & Rousseau, Judith. 2019. Nonparametric Bayesian estimation of multivariate Hawkes processes. *To appear in Annals of Statistics*, **arXiv:1802.05975**.

Embrechts, Paul, Liniger, Thomas, & Lin, Lu. 2011. Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability*, **48**(A), 367–378.

Fink, Daniel. 1997. A compendium of conjugate priors. *http://www.people.cornell.edu/pages/df36/CONJINTRnew%20TEX.pdf*, **46**.

Greenberg, Edward. 2012. *Introduction to Bayesian econometrics*. Cambridge University Press.

Halpin, Peter F. 2012. An EM algorithm for Hawkes process. *Psychometrika*, **2**.

Hawkes, Alan G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, **58**(1), 83–90.

Hill, David P, Pollitz, Fred, & Newhall, Christopher. 2002. Earthquake-volcano interactions. *Physics Today*, **55**(11), 41–47.

Lewis, Erik, & Mohler, George. 2011. A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, **1**(1), 1–20.

Lewis, Erik, Mohler, George, Brantingham, P Jeffrey, & Bertozzi, Andrea L. 2012. Self-exciting point process models of civilian deaths in Iraq. *Security Journal*, **25**(3), 244–264.

Linderman, Scott W, & Adams, Ryan P. 2015. Scalable Bayesian inference for excitatory point process networks. *arXiv Preprint*.

Liniger, Thomas Josef. 2009. *Multivariate Hawkes processes*. Ph.D. thesis, ETH Zurich.

Mercer, James. 1909. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, **209**(441-458), 415–446.

Ogata, Yosihiko. 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, **83**(401), 9–27.

Ogata, Yosihiko. 1998. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, **50**(2), 379–402.

Orbanz, Peter, & Teh, Yee Whye. 2010. Bayesian Nonparametric Models. *Encyclopedia of machine learning*.

O'Hagan, Anthony. 2004. Bayesian statistics: principles and benefits. *Frontis*, 31–45.

Rambaldi, Marcello, Bacry, Emmanuel, & Lillo, Fabrizio. 2017. The role of volume in order book dynamics: a multivariate Hawkes process analysis. *Quantitative Finance*, **17**(7), 999–1020.

Rasmussen, Carl Edward. 2003. Gaussian processes in machine learning. *Pages 63–71 of: Summer School on Machine Learning*. Springer.

Rasmussen, Jakob Gulddahl. 2013. Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, **15**(3), 623–642.

Schorfheide, Frank, & Song, Dongho. 2015. Real-time forecasting with a mixed-frequency VAR. *Journal of Business & Economic Statistics*, **33**(3), 366–380.

Simma, Aleksandr, & Jordan, Michael I. Modeling events with cascades of Poisson processes. *Proceedings of the Twenty-Sixth Conference on Unvertainty in Artificial Intelligence*, 546–555.

Veen, Alejandro, & Schoenberg, Frederic P. 2008. Estimation of space–time branching process models in seismology using an EM–type algorithm. *Journal of the American Statistical Association*, **103**(482), 614–624.

Walder, Christian J, & Bishop, Adrian N. 2017. Fast Bayesian intensity estimation for the permanental process. *Pages 3579–3588 of: Proceedings of the 34th International Conference on Machine Learning-Volume 70*. Journal of Machine Learning Research.

Zhang, Rui, Walder, Christian, Rizoiu, Marian-Andrei, & Xie, Lexing. 2019. Efficient Non-parametric Bayesian Hawkes processes. *Pages 4299–4305 of: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*. ICJAI.

A Integral of the log-likelihood

The computation of the integral in the log-likelihood in equation (21) is similar to the derivation found in appendix A of Zhang *et al.* (2019).The notation is slightly different for t_i and $||x_i||$:

$$\begin{aligned}
&= -\frac{1}{2}\sum_{i=1}^{N}\int_0^{(T)} f^2(s)d(s) \\
&= -\frac{1}{2}\sum_{i=1}^{N}\int_0^{(T-t_i-\frac{1}{||x_i||})} \big(\sum_{k=1}^{K}\omega_k e_k(s)\big)^2 d(s) \\
&= -\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{k'=1}^{K}\omega_k\omega_{k'}\int_0^{(T-t_i,||x_i||)'} e_k(s)e_{k'}(s)d(s) \\
&= -\frac{1}{2}\sum_{i=1}^{N}\omega' U_{kk'}\omega
\end{aligned}$$

With $e_k(x) = \left(\frac{2}{\pi}\right)^{\frac{1}{2}}\sqrt{1/2}^{I_{\gamma=0}}\cos(\gamma x)$ we get

$$\begin{aligned}
e_k(x)e_{k'}(x) =& \frac{1}{\pi}\sqrt{1/2}^{I_{k-1=0}}\sqrt{1/2}^{I_{k'-1=0}} \times \big[\cos\big((k-1)x-(k'-1)x\big) \\
&+ \cos\big((k-1)x+(k'-1)x\big)\big].
\end{aligned}$$

The matrix $U_{kk'}$, with $k = 1, 2, \ldots,$ has following elements:

$$\begin{aligned}
U_{1,1} =& \int_0^{T-t_i-\frac{1}{||x||}} \frac{1}{\pi}d(s) = \frac{T-t_i-\frac{1}{||x_i||}}{\pi} \\
U_{k>1,1} =& U_{1,k'>1} = \frac{\sqrt{2}}{\pi(k-1)}\sin\Big[(k-1)(T-t_i-\frac{1}{||x_i||})\Big] \\
U_{k,k(k>1)} =& \frac{1}{\pi}\Big[T-t_i-\frac{1}{||x_i||}+\frac{1}{2(k-1)}\sin\big(2(k-1)(T-t_i-\frac{1}{||x_i||})\big)\Big] \\
U_{k,k'(k\neq k')} =& \big[\pi\big((k-1)^2-(k'-1)^2\big)\big]^{-1} \\
&\times\Big[(k-1)\sin\big((k-1)(T-t_i-\frac{1}{||x_i||})\big)\cos\big((k'-1)(T-t_i-\frac{1}{||x_i||})\big) \\
&-(k'-1)\cos\big((k-1)(T-t_i-\frac{1}{||x_i||})\big)\sin\big((k'-1)(T-t_i-\frac{1}{||x_i||})\big)\Big]
\end{aligned}$$

Similarly the integral in equation (23) is the same as calculating $-\sum_{i=1}^{N}U_{kk'}^{(i)}$.

B Posterior means for the marks

The posterior means for the marks are given in this appendix.

B.1 dataset 1

Posterior means immigrant marks mean	0.645	2.102	3.059

Table 5: Posterior means dataset 1: mean immigrant marks dimension 1

Posterior means immigrant marks mean	1.169	1.159

Table 6: Posterior means dataset 1: mean immigrant marks dimension 2

	0.271	0.0167	0.0399
Posterior means variance immigrant marks dimension 1	0.0167	0.235	0.033
	0.0399	0.033	0.188

Table 7: Posterior means dataset 1:s variance immigrant marks dimension 1

Posterior means variance immigrant marks dimension 1	0.0646	0.047
	0.047	0.062

Table 8: Posterior means dataset 1: variance immigrant marks dimension 2

Posterior means $\mathbf{\Pi_1}$	0.829	-0.180	-0.130
	0.811	-0.0678	-0.737
	0.074	0.180	0.911

Table 9: Posterior means dataset 1: for $\mathbf{\Pi}_1$ from equation (31)

B.2 Dataset 2

Posterior means for $\mathbf{\Pi_{12}}$	-0.389	0.059	0.100
	0.773	0.556	0.656

Table 10: Posterior means dataset 1: for $\mathbf{\Pi}_{12}$ from equation (35)

Posterior means for $\mathbf{\Pi 2}$	-0.346	0.107
	0.054	0.108

Table 11: Posterior means dataset 1: for $\mathbf{\Pi}_2$ from equation (31)

Posterior means for $\mathbf{\Sigma}_1$	6543.94	304.071	2385.809
	304.071	17288.929	10123.555
	2385.809	10123.555	8584.739

Table 12: Posterior means dataset 1: for $\mathbf{\Sigma}_1$ from equation (32)

Posterior means for $\mathbf{\Sigma}_{12}$	3.993	2.901	4.326
	2.901	9.587	12.949
	4.326	12.949	21.552

Table 13: Posterior means dataset 1: for $\mathbf{\Sigma}_{12}$ from equation (36)

Posterior means for $\mathbf{\Sigma}_2$	7.140	6.007
	6.007	7.087

Table 14: Posterior means dataset 1: for $\mathbf{\Sigma}_1$ from equation (32)

Posterior means immigrant marks mean	1.083	1.531	2.437

Table 15: Posterior means dataset 2: mean immigrant marks dimension 1

Posterior means immigrant marks mean	1.079	1.144

Table 16: Posterior means dataset 2: mean immigrant marks dimension 2

Posterior means variance immigrant marks dimension 1	0.148	0.023	0.030
	0.023	0.234	0.083
	0.030	0.083	0.167

Table 17: Posterior means dataset 2: variance immigrant marks dimension 1

Posterior means variance immigrant marks dimension 1	0.047 0.036 0.036 0.0486

Table 18: Posterior means dataset 2: variance immigrant marks dimension 2

Posterior means $\mathbf{\Pi_1}$	0.454 0.602 -0.862 0.489 2.085 -0.855 0.166 -0.305 0.941

Table 19: Posterior means dataset 2: for $\mathbf{\Pi}_1$ from equation (31)

Posterior means for $\mathbf{\Pi_{12}}$	-0.387 0.090 0.141 0.769 0.534 0.629

Table 20: Posterior means dataset 2: for $\mathbf{\Pi}_{12}$ from equation (35)

Posterior means for $\mathbf{\Pi 2}$	-0.331 0.128 0.037 0.099

Table 21: Posterior means dataset 2: for $\mathbf{\Pi}_2$ from equation (31)

Posterior means for $\mathbf{\Sigma}_1$	9080.584 6570.586 9619.308 6570.586 11140.455 11847.319 9619.308 11847.319 21190.766

Table 22: Posterior means dataset 2: for $\mathbf{\Sigma}_1$ from equation (32)

Posterior means for $\mathbf{\Sigma}_{12}$	4.099 2.904 4.314 2.904 9.658 12.955 4.314 12.955 21.482

Table 23: Posterior means dataset 2: for $\mathbf{\Sigma}_{12}$ from equation (36)

Posterior means for $\mathbf{\Sigma}_2$	7.187 6.049 6.049 7.128

Table 24: Posterior means dataset 2: for $\mathbf{\Sigma}_1$ from equation (32)