# A SELF-LIMITING HAWKES PROCESS: INTERPRETATION, ESTIMATION, AND USE IN MODELING

A Dissertation
Presented to
The Academic Faculty

By

John Garnier Olinde

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mathematics
College of Sciences

Georgia Institute of Technology

May 2022

**A SELF-LIMITING HAWKES PROCESS: INTERPRETATION, ESTIMATION, AND USE IN MODELING**

Thesis committee:


Dr. Martin Short
School of Mathematics
*Georgia Institute of Technology*

Dr. Wenjing Liao
School of Mathematics
*Georgia Institute of Technology*

Dr. Sung Ha Kang
School of Mathematics
*Georgia Institute of Technology*

Dr. Karen Yan
School of Economics
*Georgia Institute of Technology*

Dr. Haomin Zhou
School of Mathematics
*Georgia Institute of Technology*



Date approved: 1 April 2022

I think crime pays. The hours are good, you meet a lot of interesting people, you travel a lot.

*Woody Allen*

For my family

## ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Dr. Martin Short for his guidance, patience, and kindness throughout my time in graduate school. I especially appreciate Dr. Short always taking the time to clearly answer all of my questions and making sure that I was following the right path. I also appreciated his positive attitude towards this project even when things weren't working or our results weren't as good as we were expecting.

In addition to my advisor, I would like to thank the rest of my committee: Dr. Sung Ha Kang, Dr. Haomin Zhou, Dr. Wenjing Liao, and Dr. Karen Yan. In addition to serving as my committee members, I would like to thank Dr. Kang, Dr. Zhou, and Dr. Liao for showing me how interesting and useful numerical mathematics can be, and Dr. Yan for introducing me to the world of econometrics. Without them and the classes they taught, I never would have realized how much I loved these topics.

Lastly, I would like to thank my family: my mom, Elaine, for always being there when I feel sad or stressed or just need someone to talk to; my dad, Jay, for always making time to spend with me and my family around his busy schedule, even if it was just a quick lunch between surgeries; my brother, Nicholas, for keeping me grounded by reminding me that I'm "nothin' but a weenie!"; my brother, Hunter, for showing me that more of the world's problems can be solved with a walk in the woods than on a whiteboard; and my brother, Matthew, for being a good roommate, a good friend, and most importantly, always down to go get ice cream regardless of the time of day or weather. I especially want to thank my grandparents, Grammy and Papa, for believing that education is the most important gift that you can give. I greatly appreciate all the support, both emotional and financial, that they provided to me throughout my college and graduate school years.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## SUMMARY

Many real life processes that we would like to model have a self-exciting property, i.e. the occurrence of one event causes a temporary spike in the probability of other events occurring nearby in space and time. Examples of processes that have this property are earthquakes, crime in a neighborhood, or emails within a company. In 1971, Alan Hawkes first used what is now known as the Hawkes process to model such processes. Since then much work has been done on estimating the parameters of a Hawkes process given a data set and creating variants of the process for different applications.

In this thesis, we will be proposing a new variant of a Hawkes process, called a self-limiting Hawkes process, that takes into account the effect of police activity on the underlying crime rate and an algorithm for estimating its parameters given a crime data set. We show that the self-limiting Hawkes process fits real crime data just as well, if not better, than the standard Hawkes model. We also show that the self-limiting Hawkes process fits real financial data at least as well as the standard Hawkes model.

# CHAPTER 1
# INTRODUCTION AND BACKGROUND

## 1.1 Introduction

Many real-world stochastic systems appear to exhibit a self-exciting tendency, a phenomenon whereby the occurrence of these stochastic events seems to cause an increase in the rate of occurrence of subsequent events, at least locally in time and potentially in space. Some examples include earthquakes [1], financial markets [2, 3, 4, 5], and various forms of communication [6]. One common model used to describe these systems is the Hawkes process [7], a linear model that is particularly amenable to fitting to potentially self-exciting datasets.

Another self-exciting system that has been modeled by the Hawkes process is urban crime. Various criminological theories and studies [8, 9] note the existence of "repeat victimization", whereby criminals have a tendency to commit their crimes at or against places or people who have previously been victimized. In [10], this basic phenomenon was cast in the form of a Hawkes process to describe repeat victimization in burglary data; other studies have followed [11].

But in the case of crime, there is another factor at play - the actions of police, one of whose goals is to prevent crimes from occurring in the first place. Indeed, in [11], the Hawkes process fit to up-to-date crime data was used in conjunction with police forces to inform police patrols, with measurable success. However, there is a subtle issue involved here that has not previously been addressed: given that past crime data was presumably influenced in some way by the past actions of the police, but the Hawkes process model does not explicitly capture this interaction, any estimates of the Hawkes process using past crime data will also implicitly include prior police effects. Using these fits to inform future police actions is therefore questionable, even if we have a good model for how police might

influence true crime rates, as our estimates of the stochastic crime rates already include in some unknown way the affect of police. This effect is compounded by the fact that police actions themselves are typically influenced by those crimes that do occur, such that a feedback loop exists in the crime-police system, which should alter the estimated Hawkes process in some non-trivial way.

Within the point-process literature, there are models variously termed as self-correcting [12]. These models differ from a standard self-exciting Hawkes process in that events are typically modeled as decreasing the intensity of the process via multiplication by some positive factor less than unity. A common feature of these models is an exogenous rate of increase of the stochastic intensity over time, to offset the intensity decreases accompanying the events themselves. While these models have the flavor of what we want to capture in our crime example – a police-like effect limiting growth of the event rate – they don't explicitly capture the tension between self-excitation and self-correction that we believe the crime-police system ought to exhibit.

For this reason, we introduce here what we refer to as a self-limiting Hawkes process. The specific motivation is, as discussed, the crime-police system, but the model, and the methods we show to estimate it from data, could be of potential interest in other domains whereby control is often exercised or desired over the occurrence of self-exciting events.

In chapter 2, we review how to simulate and estimate the parameters of a standard Hawkes process as well as a standard spatio-temporal Hawkes process. In chapter 3, we introduce a self-limiting Hawkes process and a self-limiting spatio-temporal Hawkes process as well as give algorithms for simulating and estimating the parameters of both models. In chapter 4, we compare the performance of the standard and the self-limiting Hawkes models in both the temporal and spatio-temporal cases using simulated data. In chapter 5, we compare the performance of the standard and the self-limiting Hawkes models in the temporal case on real crime data from the city of Chicago. Finally, in chapter 6, we argue that the self-limiting Hawkes model is a good fit to model some systems other than crime.

Here, we also use the standard and self-limiting Hawkes models to find the parameters of and compare the two models' performances on cryptocurrency data.

## 1.2 Background

The simplest point process that we might use to model the occurrence of random events is the one-dimensional homogeneous Poisson process. In a one-dimensional homogeneous Poisson process, the time differences between events are exponential random variables with mean $\frac{1}{\lambda}$. More formally, we can define a Poisson process with intensity $\lambda$ as $\{N(t) : t \geq 0\}$, where $N(t)$ counts the number of events that have occurred in the interval $[0, t]$, if $N(0) = 0$, the interarrival times are independent, and

$$\mathbb{P}\{N(t) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

While the Poisson process has many useful properties and is easy to work with, it is limited to modelling only events that are independent. For modelling systems such as urban crime where there is believed to be a high degree of dependence amongst some events, the Poisson process will not capture any of this dependence. For this reason, Hawkes processes are often used to model systems such as these. A Hawkes process can be thought of as an extension of a Poisson process that allows some dependence between events. In particular, a Hawkes process allows individual events to temporarily increase the intensity of the process.

Before we define a Hawkes process, we will first define the conditional intensity of a point process as

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbb{E}[N(t, t + dt \mid H_t)]}{dt},$$

where $H_t$ is the history of the process up until time $t$ and $N$ counts the number of points in the interval $[t, t + dt)$ given $H_t$. By allowing $N$ to be dependent on the history of the

process up until $t$, we are able to capture the dependence between events in the types of systems discussed above.

This leads us to the definition of a Hawkes process, which has conditional intensity

$$\lambda(t) = \mu(t) + \sum_{i:t_i<t} g(t - t_i), \tag{1.1}$$

where $\mu$ is the background intensity of the process and $g$, called the self-exciting kernel, is a function that describes the self-exciting property of the process [7].

The conditional probability given in Equation 1.1 could be easily modified to yield a spatio-temporal Hawkes process on the interval $[0, T] \times [0, L] \times [0, L]$:

$$\lambda(t, \vec{x}) = \mu(t, \vec{x}) + \sum_{i:t_i<t} g(t - t_i, \vec{x} - \vec{x_i}). \tag{1.2}$$

One way to conceptualize a Hawkes process over an interval of time $[0, T]$ is as a sum of individual Poisson processes: $\lambda_0(t) = \mu(t), \lambda_1(t) = g(t-t_1), \ldots, \lambda_n(t) = g(t-t_n)$. Each Poisson process creates a generation of points upon which the following Poisson processes are based. Intensity $\lambda_0 = \mu(t)$ has no conditions, so it defines a Poisson process on the whole interval $[0, T]$. Events that arise from intensity $\lambda_0$ are referred to as background events. Intensities $\lambda_i,\ i > 0$ do not activate until $t > t_i$, where $t_i$ is the $i^{\text{th}}$ point in the overall process. So $\lambda_i$ defines a Poisson process on the interval $[t_i, T]$. Events arising from intensity $\lambda_i,\ i > 0$ are called daughter events, and the parent event of each of these daughters is event $t_i$.

For chapter 4, chapter 5, and chapter 6, we will only be considering Hawkes processes where $\mu(t) = \mu(t, \vec{x}) = \mu \in \mathbb{R}^+$, $g(t - t_i) = k\omega e^{-\omega(t-t_i)}$ in the temporal case, and $g(t - t_i, \vec{x} - \vec{x_i}) = \frac{k\omega}{4s^2} e^{-\omega(t-t_i)} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}}$ in the spatio-temporal case. In chapter 2 and chapter 3, we introduce methods for simulating Hawkes processes with both general and exponential excited kernels, but all of our methods for estimating the parameters of the processes require the kernels given above.

# CHAPTER 2

# SIMULATION AND PARAMETER ESTIMATION OF A HAWKES PROCESS

In this chapter we introduce algorithms for simulating and estimating the parameters of a Hawkes process. We include simulation algorithms because it is useful to be able to generate instances of Hawkes processes with known parameters to verify that the parameter estimation algorithms are working correctly. Additionally, an instance of a Hawkes process with known parameters can be used to compare the performance of two different models. This plays a role in chapter 4 where we compare the estimation performance of the standard and self-limiting Hawkes models on hypothetical Hawkes datasets.

## 2.1 Simulating a Hawkes Process

One way to simulate a Hawkes process is by using what is known as the thinning method. This method was first proposed as a way to simulate non-homogeneous point processes [13], but has since been modified to simulate Hawkes processes [14]; it is especially useful if the excited kernel is not exponential. (If the excited kernel is exponential, there are more efficient simulation algorithms that we will discuss later in this section.) The algorithm is described in Figure 2.1.

**Input:** $\lambda(t)$ - The intensity function, $T$ - the final simulation time
**Output:** A realization of the point process, $\{t_1, \ldots, t_n\}$
1: Define $M = \max\{\lambda(t) : t \in [0, T]\}$ and let $N \sim \text{Pois}(MT)$.
2: Place $N$ points uniformly at random in the interval $[0, T]$.
3: For $i = 1, \ldots, N$, delete point $t_i$ with probability $1 - \frac{\lambda(t_i)}{M}$.
4: Return all the points that were not deleted.

Figure 2.1: Thinning Method for Simulating a Non-homogeneous Point Process

The intuition behind this algorithm is that, for an intensity function $\lambda(t)$, we have $M = \lambda(t) + (M - \lambda(t))$, where $M = \max\{\lambda(t) : t \in [0, T]\}$. If we simulate a process using $M$

as the intensity, then each simulated event is attributable to either $\lambda(t)$ or $(M - \lambda(t))$. The probability that an event $t_i$ is attributable to $\lambda(t)$ is $\frac{\lambda(t_i)}{M}$. Since we are only interested in the events that are attributable to $\lambda(t)$, we will keep each event with probability $\frac{\lambda(t_i)}{M}$. This is equivalent to deleting each event with probability $1 - \frac{\lambda(t_i)}{M}$.

When applied to the Hawkes process, the thinning method first simulates background events from intensity $\mu$. These events constitute the first or background generation of the process. Then, it simulates all of the direct offspring of the background events through the various $g$ kernels that the background events produced. These new events constitute the second generation of the process. Then the offspring of these offspring are simulated through the various $g$ kernels that the offspring events produced. These new events constitute the third generation of the process. This is repeated until the excited kernels of one generation simulate no events in the following generation. The resulting Hawkes process consists of the union of the generations. This method essentially relies on the fact noted above that a Hawkes process can be thought of as a sum of many Poisson processes. We can run the algorithm given in Figure 2.1 to simulate each of the component Poisson processes and then take the union of all the generated points [14].

Though this method for simulating a Hawkes process is intuitive and can be implemented relatively easily, there are more efficient methods of simulation available, especially when the excited kernel $g$ is an exponential. One particular example is the method of Dassios and Zhao [15]. This method assumes that $\mu(t) = \mu \in \mathbb{R}^+$ and the function $g(t - t_i) = k\omega e^{-\omega(t-t_i)}$ for $t > t_i$, $g(t - t_i) = 0$ for $t \leq t_i$. The algorithm is given in Figure 2.2.

Rather than simulating the process layer-by-layer as is done when using the thinning method, the method given by Dassios and Zhao starts at time $t = 0$ and jumps forward by simulating the time interval $\Delta t$ until the next event occurs. Each such $\Delta t$ is found by randomly generating two possible values: one from the background rate $\mu$ and one from the full summation of the excited kernels, which is itself simply a decaying exponential.

**Input:** $T$ - The final simulation time; $\mu$, $k$, $\omega$ - Hawkes parameters
**Output:** A realization of a Hawkes process, $\{t_1, \ldots, t_n\}$

$ts = -\frac{\ln(U(0,1))}{\mu}$
$g = k$
times $= ts$
**while** $ts < T$ **do**
    $tb = -\frac{\ln(U(0,1))}{\mu}$
    $\zeta = \frac{\ln(U(0,1))}{g} + 1$
    **if** $\zeta > 0$ **then**
        $td = -\frac{\ln(\zeta)}{\omega}$
    **else**
        $td = tb$
    **end if**
    $\Delta t = \min(tb, td)$
    $ts = ts + \Delta t$
    $g = ge^{-\omega \Delta t}$
    $g = g + k$
    times $= [\text{times}, ts]$
**end while**
times $=$ times that are less than $T$

Figure 2.2: Hawkes Process Simulation Algorithm by Dassios and Zhao

The smaller of these two times is then chosen as $\Delta t$, time is incremented by this value, and the fully summed excited kernel is updated via exponential decay and an increase by $k$, and the algorithm continues.

This algorithm will become especially useful once we introduce a self-limiting Hawkes process in chapter 3.

## 2.2 Simulating A Spatio-temporal Hawkes Process

Suppose we want to simulate a spatio-temporal Hawkes process on the interval $[0, T] \times [0, L] \times [0, L]$. Further, suppose that the self-exciting kernel can be written as the product of a function of time and a function of a spatial vector, i.e.

$$g(t - t_i, \vec{x} - \vec{x_i}) = f(t - t_i)p(\vec{x} - \vec{x_i}),$$

where $p$ is the probability density function that describes the spatial distribution of daughter events around their parents.

We can use the same thinning algorithm given in Figure 2.1 except when it comes time to simulate a process with intensity $g(t-t_i, \vec{x}-\vec{x_i})$, we will use $f(t-t_i)$ [14]. We will then assign a location to each newly created event according to the probability density function $p(\vec{x}-\vec{x_i})$, where $\vec{x_i}$ is the location of the parent of the events being generated.

If $\mu(t, \vec{x}) = \mu \in \mathbb{R}^+$ and the function $g(t-t_i, \vec{x}-\vec{x_i}) = g(t-t_i, x-x_i, y-y_i) = k\omega e^{-\omega(t-t_i)} p(x-x_i, y-y_i)$ for $t > t_i$, $g(t-t_i) = 0$ for $t \leq t_i$, where $p$ is a two-dimensional probability density function, we can modify the method of Dassios and Zhou to more efficiently simulate a spatio-temporal Hawkes process [15]. This modified algorithm is given in Figure 2.3.

This method works similarly to the algorithm in Figure 2.2, the only difference being that now we have to consider the locations of the events that are created. When the next $\Delta t$ is determined, if the new event came from the background process, the location of the new event is selected uniformly at random in the region $[0, L] \times [0, L]$ [13]. If the new event came from the full summation of the excited kernels, then we first select a parent for this event.

Note that in this version of the algorithm, $g$ is a vector whose entries are the values of the self-exciting kernels from all of the events that have occurred up until the new event. We can think of this vector as recording the intensities due to each of the previous events at time $ts$. Since each of these entries are non-negative, if we normalize $g$ so that the entries sum to one, we can use $g$ as a probability distribution over the possible parents of the new event. So, we select a parent for our new event according to this distribution.

Once a parent is chosen, the location of the new event is drawn randomly according to the probability distribution function $p$, centered at the location of the parent.

**Input:** $T$ - The final simulation time; $\mu$, $k$, $\omega$ - Hawkes parameters; $p$ - a two-dimensional probability distribution function

**Output:** A realization of a spatio-temporal Hawkes process, $\{(t_1, x_1, y_1), \ldots, (t_n, x_n, y_n)\}$

$ts = -\frac{\ln(U(0,1))}{\mu}$
$xs = U(0, L)$
$ys = U(0, L)$
$g = k$
events $= (ts, xs, ys)$
**while** $ts < T$ **do**
    $tb = -\frac{\ln(U(0,1))}{\mu}$
    $\zeta = \frac{\ln(U(0,1))}{\text{sum}(g)} + 1$
    $xs = U(0, L)$
    $ys = U(0, L)$
    **if** $\zeta > 0$ **then**
        $td = -\frac{\ln(\zeta)}{\omega}$
    **else**
        $td = tb$
    **end if**
    $\Delta t = \min(tb, td)$
    $ts = ts + \Delta t$
    $g = ge^{-\omega \Delta t}$
    **if** $td == \Delta t$ **then**
        $parentX = x$-value of location of parent chosen according to $g$
        $parentY = y$-value of location of parent chosen according to $g$
        $xs$ and $ys$ are chosen according to $p(x - parentX, y - parentY)$
    **end if**
    $g = [g, k]$
    events $=$ [events, $(ts, xs, ys)$]
**end while**
events $=$ events where $t < T$ and $(x, y) \in [0, L] \times [0, L]$

Figure 2.3: Hawkes Process Simulation Algorithm by Dassios and Zhao Modified to Include Spatial Component

## 2.3 Estimating the Parameters of a Hawkes Process

Here we present a review of the Expectation-Maximization (E-M) method [16] to estimate the parameters $\mu$, $k$, and $\omega$ of a Hawkes process, with intensity $\lambda$, from data, where

$$\lambda(t) = \mu + \sum_{i:t_i<t} k\omega e^{-\omega(t-t_i)}.$$

First, we need the likelihood function of the parameters given the data $\{t_1, \ldots, t_n\}$. This is given by

$$\begin{aligned} L &= e^{-\int_0^{t_1} \lambda(t)dt}\lambda(t_1)dt \times e^{-\int_{t_1}^{t_2} \lambda(t)dt}\lambda(t_2)dt \times \cdots \times e^{-\int_{t_{n-1}}^{t_n} \lambda(t)dt}\lambda(t_n)dt \\ &= e^{-\int_0^T \lambda(t)dt} dt^n \prod_i \lambda(t_i). \end{aligned} \tag{2.1}$$

Each term $e^{-\int_{t_{i-1}}^{t_i} \lambda(t)dt}\lambda(t_i)dt$ approximates the probability of no events in the interval $(t_{i-1}, t_i)$ followed by an event occurring in a small interval of width $dt$ around $t_i$.

To make this easier to work with, we will take the natural log of this equation. This is valid since we will be maximizing $L$, and the maxima of $\ln(L)$ correspond to those of $L$. We will denote $\ln(L)$ as $\mathscr{L}$. Then we have the following log-likelihood function:

$$\mathscr{L} = -\int_0^T \lambda(t)dt + n\ln(dt) + \sum_i \ln(\lambda(t_i)). \tag{2.2}$$

Since we will be maximizing $\mathscr{L}$ with respect to $\mu$, $k$, and $\omega$, we can drop any terms that don't contain these parameters (i.e., the term $n\ln(dt)$). From now on, $\mathscr{L}$ will refer to the log-likelihood function without the term $n\ln(dt)$.

Suppose we knew the true branching structure of the process: which events were background events and which were daughters, along with which event was the parent of each daughter. Then we could rewrite $\mathscr{L}$ as

$$\mathscr{L} = \sum_{i \in B} \ln(\mu) - \int_0^T \mu dt + \sum_{i \in D} \ln\left(k\omega e^{-\omega(t_i - t_{p(i)})}\right) - \int_0^T \sum_{i:t_i<t} k\omega e^{-\omega(t-t_i)} dt,$$

where $B$ and $D$ are the sets of background and daughter events, respectively, and $p(i)$ is defined as the parent event of event $t_i$. Here, the first two terms only depend on $\mu$ and can be thought of as the log-likelihood of the background process of the Hawkes process. Likewise, the last two terms only depend on $k$ and $\omega$ and can be thought of as the log-likelihood of the self-exciting part of the Hawkes process.

Though we generally don't know the true branching structure of the process, we will assume we can still generate a probabilistic branching structure $P$, where

$$P_{ij} = \begin{cases} \text{prob. that } i \text{ is a background event} & , i = j \\ \text{prob. that } i \text{ is a daughter of } j & , j < i \end{cases}. \tag{2.3}$$

Taking the expectation of $\mathscr{L}$ with respect to $P$ gives us what is called the complete data log-likelihood [16]:

$$\mathbb{E}[\mathscr{L}] = \sum_i P_{ii} \ln(\mu) - \int_0^T \mu dt + \sum_{j<i} P_{ij} \ln\left(k\omega e^{-\omega(t_i - t_j)}\right) - \int_0^T \sum_{i:t_i<t} k\omega e^{-\omega(t-t_i)} dt,$$

which can be simplified to

$$\begin{aligned} \mathbb{E}[\mathscr{L}] = {} & \ln(\mu) \sum_i P_{ii} + \ln(k\omega) \sum_{j<i} P_{ij} - \omega \sum_{j<i} P_{i,j}(t_i - t_j) - \mu T \\ & - k \sum_i \left(1 - e^{-\omega(T-t_i)}\right). \end{aligned} \tag{2.4}$$

For more details on this simplification, see Appendix A.

We can then maximize $\mathbb{E}[\mathscr{L}]$ with respect to $\mu$, $k$, and $\omega$ by taking the respective partial derivatives and setting them equal to $0$. This gives us the following formulas:

$$\mu = \frac{\sum_i P_{ii}}{T}, \tag{2.5}$$

$$k = \frac{\sum_{j<i} P_{ij}}{\sum_i \left(1 - e^{-\omega(T-t_i)}\right)}, \tag{2.6}$$

and

$$0 = \sum_{j<i} P_{ij} - \omega \sum_{j<i} P_{ij}(t_i - t_j) - k\omega \sum_i \left[(T - t_i)e^{-\omega(T-t_i)}\right]. \tag{2.7}$$

We can then simultaneously solve these equations to get estimates for the parameters.

Of course, we must still specify $P_{ij}$ in order to use these formulas. But, since a Hawkes Process can be thought of as a sum of Poisson processes, we have

$$P_{ij} = \begin{cases} \frac{\mu}{\lambda(t_i)} & i = j \\ \frac{k\omega e^{-\omega(t_i - t_j)}}{\lambda(t_i)} & j < i \end{cases}. \tag{2.8}$$

From here, we can see that we need $\mu$, $k$, and $\omega$ to calculate $P_{ij}$ and $P_{ij}$ to calculate $\mu$, $k$, and $\omega$. This leads us to the iterative Expectation-Maximization (E-M) method given in Figure 2.4.

**Input:** $\mu$, $k$, $\omega$ - An initial guess for these parameters, $\epsilon$ - The tolerance of convergence
**Output:** $\mu$, $k$, $\omega$ - The estimated Hawkes parameters

1: For each event pair $j \leq i$, calculate $P_{ij}$ using the current values of $\mu$, $k$, and $\omega$ using Equation 2.8. This is the Expectation step of the E-M algorithm.
2: Update our values of $\mu$, $k$, and $\omega$ using these $P_{ij}$ by maximizing Equation 2.4. This is the Maximization step of the E-M algorithm.
3: Repeat steps 2 and 3 until some measure of convergence, given the desired tolerance $\epsilon$, is achieved.

Figure 2.4: E-M Algorithm for Estimating Hawkes Process Parameters

## 2.4 Estimating the Parameters of a Spatio-temporal Hawkes Process

Here we present a review of a variant of the above Expectation-Maximization (E-M) method [16] to estimate the parameters $\mu$, $k$, $\omega$, and $s$ of a spatio-temporal Hawkes process, with intensity $\lambda$, from data, where

$$\lambda(t,x,y) = \mu + \frac{k\omega}{4s^2}\sum_{i:t_i<t} e^{-\omega(t-t_i)}e^{\frac{-(|x-x_i|+|y-y_i|)}{s}}.$$

Estimating the parameters of a spatio-temporal Hawkes process follows a very similar procedure. Just like before, we need the likelihood function of the data $\{(t_1,x_1,y_1),\ldots,(t_n,x_n,y_n)\}$. We use a version of Equation 2.1 that is modified to include a spatial component. This is given by

$$\begin{aligned}
L &= e^{-\int_0^{t_1}\int_0^L\int_0^L \lambda(t,x,y)dxdydt}\lambda(t_1,x_1,y_1)dtdxdy \\
&\times e^{-\int_{t_1}^{t_2}\int_0^L\int_0^L \lambda(t,x,y)dxdydt}\lambda(t_2,x_2,y_2)dtdxdy \times \cdots \\
&\times e^{-\int_{t_{n-1}}^{t_n}\int_0^L\int_0^L \lambda(t,x,y)dxdydt}\lambda(t_n,x_n,y_n)dtdxdy \\
&= e^{-\int_0^T\int_0^L\int_0^L \lambda(t,x,y)dxdydt}(dtdxdy)^n\prod_i \lambda(t_i,x_i,y_i).
\end{aligned}$$

This time, each term $e^{-\int_{t_1}^{t_2}\int_0^L\int_0^L \lambda(t,x,y)dxdydt}\lambda(t_2,x_2,y_2)dtdxdy$ approximates the probability of no events in the interval $(t_{i-1},t_i)$ anywhere in space followed by an event occurring in a small interval of width $dt$ around $t_i$ with a location in a small $dx \times dy$ rectangle around $(x_i,y_i)$.

Again, we will take the natural log of this equation to give the following log-likelihood function:

$$\mathscr{L} = -\int_0^T\int_0^L\int_0^L \lambda(t,x,y)dxdydt + n\ln(dtdxdy) + \sum_i \ln(\lambda(t_i,x_i,y_i)). \tag{2.9}$$

Dropping any terms that don't contain $\mu$, $k$, or $\omega$ and then taking the expectation of $\mathscr{L}$ with respect to $P$ as defined in Equation 2.3, gives us the complete data log-likelihood of the spatio-temporal Hawkes process [16]:

$$\begin{aligned}\mathbb{E}[\mathscr{L}] = \ln(\mu)\sum_i P_{ii} + \ln\left(\frac{k\omega}{4s^2}\right)\sum_{j<i} P_{ij} - \omega\sum_{j<i} P_{ij}(t_i - t_j) \\ - \frac{1}{s}\sum_{j<i} P_{ij}(|x_i - x_j| + |y_i - y_j|) - \int_0^T\int_0^L\int_0^L \lambda(t,x,y)dxdydt,\end{aligned}$$

which can be simplified to

$$\begin{aligned}\mathbb{E}[\mathscr{L}] &= \ln(\mu)\sum_i P_{ii} + \ln\left(\frac{k\omega}{4s^2}\right)\sum_{j<i} P_{ij} - \omega\sum_{j<i} P_{ij}(t_i - t_j) \\ &- \frac{1}{s}\sum_{j<i} P_{ij}(|x_i - x_j| + |y_i - y_j|) - \mu TL^2 \\ &+ \frac{k}{4}\sum_i \left(2 - e^{\frac{-x_i}{s}} - e^{\frac{-(L-x_i)}{s}}\right)\left(2 - e^{\frac{-y_i}{s}} - e^{\frac{-(L-y_i)}{s}}\right)\left(e^{-\omega(T-t_i)} - 1\right).\end{aligned} \tag{2.10}$$

For more details on this simplification, see Appendix A.

Note that in this case, $P$ is given by

$$P_{ij} = \begin{cases} \frac{\mu}{\lambda(t_i,x_i,y_i)} & i = j \\ \frac{k\omega e^{-\omega(t_i-t_j)}p(x_i-x_j,y_i-y_j)}{\lambda(t_i,x_i,y_i)} & j < i \end{cases}. \tag{2.11}$$

To maximize $\mathbb{E}[\mathscr{L}]$ with respect to $\mu$, $k$, $\omega$, and $s$, we can once again take the respective partial derivatives and set them equal to $0$. This gives us the following formulas:

$$\mu = \frac{\sum_i P_{ii}}{TL^2},$$

$$k = \frac{-4\sum_{j<i} P_{ij}}{\sum_i \left(2 - e^{\frac{-x_i}{s}} - e^{\frac{-(L-x_i)}{s}}\right)\left(2 - e^{\frac{-y_i}{s}} - e^{\frac{-(L-y_i)}{s}}\right)(e^{-\omega(T-t_i)} - 1)},$$

$$\begin{aligned} 0 = &\sum_{j<i} P_{ij} - \omega \sum_{j<i} P_{ij}(t_i - t_j) \\ &+ \frac{k\omega}{4} \sum_i \left(2 - e^{\frac{-x_i}{s}} - e^{\frac{-(L-x_i)}{s}}\right)\left(2 - e^{\frac{-y_i}{s}} - e^{\frac{-(L-y_i)}{s}}\right)(t_i - T)e^{-\omega(T-t_i)}, \end{aligned}$$

and

$$\begin{aligned} 0 = &\frac{-2}{s} \sum_{j<i} P_{ij} + \frac{1}{s^2} \sum_{j<i} P_{ij}(|x_i - x_j| + |y_i - y_j|) \\ &+ \frac{k}{4} \sum_i \left(e^{-\omega(T-t_i)} - 1\right)\Big(\frac{-2y_i}{s^2} e^{\frac{-y_i}{s}} - \frac{2x_i}{s^2} e^{\frac{-x_i}{s}} + \frac{2(y_i - L)}{s^2} e^{\frac{-(L-y_i)}{s}} \\ &+ \frac{2(x_i - L)}{s^2} e^{\frac{-(L-x_i)}{s}} + \frac{x_i - y_i + L}{s^2} e^{\frac{-(x_i - y_i + L)}{s}} + \frac{y_i - x_i + L}{s^2} e^{\frac{-(y_i - x_i + L)}{s}} \\ &+ \frac{x_i + y_i}{s^2} e^{\frac{-(x_i + y_i)}{s}} + \frac{2L - x_i - y_i}{s^2} e^{\frac{-(2L - x_i - y_i)}{s}}\Big). \end{aligned}$$

Instead of numerically solving $\frac{\partial \mathbb{E}[\mathscr{L}]}{\partial \omega} = 0$ and $\frac{\partial \mathbb{E}[\mathscr{L}]}{\partial s} = 0$, which can be computationally expensive, we rewrote these equations in the form $\omega = f_1(k, \omega, s)$ and $s = f_2(k, \omega, s)$ and then evaluate these two functions at the last iteration's approximation for $k$, $\omega$, and $s$ to update these parameters. Rewriting $\frac{\partial \mathbb{E}[\mathscr{L}]}{\partial \omega} = 0$ and $\frac{\partial \mathbb{E}[\mathscr{L}]}{\partial s} = 0$ in this way gives

$$\omega = \frac{\sum_{j<i} P_{ij}}{\sum_{j<i} P_{ij}(t_i - t_j) - \frac{k}{4}\sum_i \left(2 - e^{\frac{-x_i}{s}} - e^{\frac{-(L-x_i)}{s}}\right)\left(2 - e^{\frac{-y_i}{s}} - e^{\frac{-(L-y_i)}{s}}\right)(t_i - T)e^{-\omega(T-t_i)}},$$

$$s = \frac{\sum_{j<i} P_{ij}(|x_i - x_j| + |y_i - y_j|) + \frac{k}{4}\sum_i h_i(\omega, s)}{2\sum_{j<i} P_{ij}},$$

where

$$h_i(\omega, s) = \left(e^{-\omega(T-t_i)} - 1\right)\Big( - 2y_i e^{\frac{-y_i}{s}} - 2x_i e^{\frac{-x_i}{s}} + 2(y_i - L)e^{\frac{-(L-y_i)}{s}} + 2(x_i - L)e^{\frac{-(L-x_i)}{s}} + (x_i - y_i + L)e^{\frac{-(x_i-y_i+L)}{s}} + (y_i - x_i + L)e^{\frac{-(y_i-x_i+L)}{s}} + (x_i + y_i)e^{\frac{-(x_i+y_i)}{s}} + (2L - x_i - y_i)e^{\frac{-(2L-x_i-y_i)}{s}}\Big).$$

In practice, updating $\omega$ and $s$ in this way resulted in the same estimated values of the parameters.

We can now use an E-M algorithm similar to the one given in Figure 2.4 to estimate the parameters of the process [16].

# CHAPTER 3

# A SELF-LIMITING HAWKES PROCESS

We now turn to the development of our model for a self-limiting Hawkes process. Recall that the overall goal is to model a stochastic process with two competing properties: 1) the process should have self-excitation, for which a Hawkes process can serve as a baseline, and 2) the model should incorporate a mechanism by which the intensity of the process can also be reduced by the occurrence of events, to represent potentially exogenous influences such as police activity. To capture this second, self-limiting effect, we introduce two new parameters, $\alpha$ and $\beta$, and define $N(\alpha, t)$, which counts the number of events that occurred through the process on interval $[t-\alpha, t)$. Then our model of a self-limiting Hawkes process intensity is

$$\lambda(t) = \left( \mu + \sum_{i:t_i<t} g(t - t_i) \right) e^{-\beta N(\alpha,t)}. \tag{3.1}$$

Parameter $\beta$ therefore represents the strength of self-limiting, with greater values decreasing the intensity more than smaller values, and $\alpha$ represents a time-window over which any given event can contribute to self-limiting of the overall process. If the process is being used to model criminal events, then we can think of $\alpha$ as the memory of the police and $\beta$ as the increase in police deterrent activity for each additional crime that occurs in the interval $[t - \alpha, t)$.

We also introduce a self-limiting spatio-temporal Hawkes process. Our model for the intensity of this process is

$$\lambda(t, x, y) = \left( \mu + \sum_{i:t_i<t} g(t - t_i) p(x - x_i, y - y_i) \right) q(t, x, y, \alpha, \beta), \tag{3.2}$$

where

$$q(t, x, y, \alpha, \beta) = \begin{cases} e^{\frac{-\beta N(\alpha, t)}{|\text{box}(t)|}} & \text{if } (x, y) \in \text{box}(t) \\ 1 & \text{else} \end{cases}$$

and box$(t)$ is a box centered on the mean location of the events in the interval $[t-\alpha, t)$. Its width is twice the standard deviation in the $x$ component of the locations of the events in the interval $[t-\alpha, t)$. Its height is twice the standard deviation in the $y$ component of the locations of the events in the interval $[t-\alpha, t)$. Just like in the intensity function of the standard spatio-temporal Hawkes process, $p$ is a two-dimensional probability density function.

Here, we can think of the parameters $\alpha$ and $\beta$ just as we did for the (temporal) self-limiting Hawkes process: $\beta$ represents the strength of self-limiting and $\alpha$ represents a time-window over which any given event can contribute to self-limiting of the overall process. Note that the self-limiting at time $t$ only occurs inside of box$(t)$; outside of box$(t)$, the process behaves exactly like a standard spatio-temporal Hawkes process with no self-limiting component.

### 3.1 Simulating a Self-limiting Hawkes Process

Equation 3.1 can be interpreted in the following mechanistic way, which aids in simulating. In the absence of self-limiting, the process would behave as a standard Hawkes process, and would generate some sequence of hypothetical events. However, the self-limiting effect is such that each event $t_i$ that does in fact occur via the process causes every subsequent hypothetical event within the period $(t_i, t_i + \alpha]$ to be probabilistically "blocked" from occurring, with probability $p = 1 - e^{-\beta}$. If we assume that multiple overlapping blockings of a single hypothetical event are probabilistically independent, then the probability of a hypothetical event at time $t_j$ not being blocked is $e^{-\beta N(\alpha, t_j)}$. Hence, the intensity of Equation 3.1 tells us that events occur only when the underlying Hawkes process would hypothetically cause them to occur, and only if they are not probabilistically blocked by some of the prior events

that did in fact occur (weren't blocked themselves).

Using this interpretation, one could create a straightforward Poisson-thinning type algorithm to simulate the self-limiting Hawkes process. Specifically, first simulate the underlying Hawkes process by itself, without any self-limiting effect, being sure to retain the true branching structure of the process. Then, starting with the first event and working sequentially, retain each event with probability $e^{-\beta N(\alpha,t_j)}$, where $t_j$ is the time of the event. If an event $t_i$ is retained, continue to the next event. If event $t_i$ is not retained, remove it from the list of event times and also remove all subsequent events that are descendants (either directly or indirectly) of $t_i$ in the branching process, then proceed to the next event.

While the above process is straightforward to describe, it is not very computationally efficient. Hence, we also provide a more efficient algorithm, which is a modified version of the algorithm of Dassios and Zhao [15], and which incorporates the preventative action right into the generation of the Hawkes process. Just as is the case in the standard algorithm of Dassios and Zhao, this algorithm does have the drawback of only being valid for the exponential excited kernel $g(t-t_i) = k\omega e^{-\omega(t-t_i)}$. Therefore, we will use this choice of $g$ for our self-limiting Hawkes process.

To modify this method to account for the self-limiting aspect, we simply add a step where each event to be added is only added with probability $e^{-\beta N(\alpha,t_j)}$, where $t_j$ is the time of the potential new event. If $t_j$ is added, the algorithm continues just as in the Dassios and Zhao method. If $t_j$ is not added, time is incremented by $\Delta t$ and the exponential decay of the excited kernel is updated, but the excited kernel is not incremented as one normally would. This gives the algorithm in Figure 3.1.

In Figure 3.1, each $U(0,1)$ is a uniform random variable on $[0,1]$ and $B(1,p)$ is a Bernoulli random variable with probability of success $p$.

We note here that, unlike a standard Hawkes process, our self-limiting process can still remain bounded even if $k > 1$. In a standard Hawkes process, $k > 1$ means that each event on average gives rise to more than one daughter event, generally causing the intensity

**Input:** $T$ - The final simulation time; $\mu$, $k$, $\omega$ - Hawkes parameters; $\alpha$, $\beta$ - Self-limiting parameters

**Output:** A realization of a self-limiting Hawkes process, $\{t_1, \ldots, t_n\}$

$ts = -\frac{\ln(U(0,1))}{\mu}$
$g = k$
times $= ts$
**while** $ts < T$ **do**
    $tb = -\frac{\ln(U(0,1))}{\mu}$
    $\zeta = \frac{\ln(U(0,1))}{g} + 1$
    **if** $\zeta > 0$ **then**
        $td = -\frac{\ln(\zeta)}{\omega}$
    **else**
        $td = tb$
    **end if**
    $\Delta t = \min(tb, td)$
    $ts = ts + \Delta t$
    $g = ge^{-\omega \Delta t}$
    $p = e^{-\beta N(\alpha, ts)}$
    **if** $B(1, p) == 1$ **then**
        $g = g + k$
        times $= [\text{times}, ts]$
    **end if**
**end while**
times $=$ times that are less than $T$

Figure 3.1: Self-limiting Hawkes Process Simulation Algorithm

to grow exponentially in time. However, the self-limiting process avoids this through the $e^{-\beta N(\alpha,t)}$ term. If $k > 1$ starts to cause $\lambda$ to grow very large, then the number of events $N$ will also grow, and the exponential dampening will force the value of $\lambda$ back down. In Figure 3.2, we illustrate two realizations of a self-limiting Hawkes process, one of which has $k > 1$.

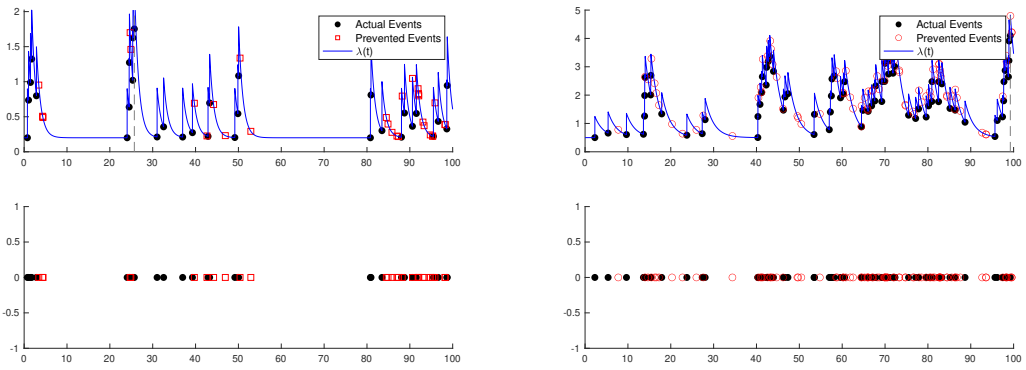


Figure 3.2: Examples of simulated data from a self-limiting Hawkes model and the intensity function $\lambda(t)$. For the subfigure on the left, the parameters were $\mu = 0.15$, $k = 0.6$, $\omega = 1$, $\alpha = 5$, and $\beta = 0.3$. For the subfigure on the right, the parameters were $\mu = 0.5$, $k = 1.5$, $\omega = 0.5$, $\alpha = 10$, and $\beta = 0.1$.

## 3.2 Simulating a Self-limiting Spatio-temporal Hawkes Process

Simulating a self-limiting spatio-temporal Hawkes process can be done a mechanistic way similar to what was described at the beginning of the previous section. In the absence of self-limiting, the process would behave as a standard spatio-temporal Hawkes process, and would generate some sequence of hypothetical events. However, the self-limiting effect is such that each event $(t_i, x_i, y_i)$ that would have occurred in the underlying Hawkes process with no self-limiting is probabilistically "blocked" with probability $p = 1 - e^{\frac{-\beta}{|\text{box}(t_i)|}}$ by each of the events that occurred and were not blocked in the interval $(t_i - \alpha, t_i]$. If we assume that multiple overlapping blockings of a single hypothetical event are probabilistically independent, then the probability that the event $(t_i, x_i, y_i)$ is not blocked is $e^{\frac{-\beta N(\alpha, t_i)}{|\text{box}(t_i)|}}$. Hence, the intensity of Equation 3.2 tells us that events occur only when the underlying Hawkes process would hypothetically cause them to occur, and only if they are not probabilistically blocked by some of the prior events that did in fact occur (weren't blocked themselves).

Under this interpretation, one could again create a Poisson-thinning type algorithm to simulate the self-limiting spatio-temporal Hawkes process. Though the implementation of such an algorithm is not quite as straightforward as it was in the previous cases.

Once again, simulation of a spatio-temporal Hawkes process can be done more efficiently by further modifying the algorithm of Dassios and Zhao [15]. To use this algorithm, we need to use the self-exciting kernel $g(t - t_i) = k\omega e^{-\omega(t-t_i)}$. This gives the algorithm in Figure 3.3

In Figure 3.3, each $U(0, 1)$ is a uniform random variable on $[0, 1]$ and $B(1, q)$ is a Bernoulli random variable with probability of success $q$.

**Input:** $T$ - The final simulation time; $\mu$, $k$, $\omega$, $s$ - Hawkes parameters; $p$ - a two-dimensional probability distribution function

**Output:** A realization of a spatio-temporal Hawkes process, $\{(t_1, x_1, y_1), \ldots, (t_n, x_n, y_n)\}$

$ts = -\frac{\ln(U(0,1))}{\mu}$
$xs = U(0, L)$
$ys = U(0, L)$
$g = k$
events $= (ts, xs, ys)$
**while** $ts < T$ **do**
    $tb = -\frac{\ln(U(0,1))}{\mu}$
    $\zeta = \frac{\ln(U(0,1))}{\text{sum}(g)} + 1$
    $xs = U(0, L)$
    $ys = U(0, L)$
    **if** $\zeta > 0$ **then**
        $td = -\frac{\ln(\zeta)}{\omega}$
    **else**
        $td = tb$
    **end if**
    $\Delta t = \min(tb, td)$
    $ts = ts + \Delta t$
    $g = ge^{-\omega \Delta t}$
    **if** $td == \Delta t$ **then**
        $parentX = x$-value of location of parent chosen according to $g$
        $parentY = y$-value of location of parent chosen according to $g$
        $xs$ and $ys$ are chosen according to $p(x - parentX, y - parentY)$
    **end if**
    **if** $(xs, ys) \in \text{box}(ts)$ **then**
        $q = e^{\frac{-\beta N(\alpha, ts)}{|\text{box}(t)|}}$
    **else**
        $q = 1$
    **end if**
    **if** $B(1, q) == 1$ **then**
        $g = [g, k]$
        events $= [\text{events}, (ts, xs, ys)]$
    **end if**
**end while**
events $=$ events where $t < T$ and $(x, y) \in [0, L] \times [0, L]$

Figure 3.3: Self-limiting Spatio-temporal Hawkes Process Simulation Algorithm

### 3.3 Estimating the Parameters of a Self-limiting Hawkes Process

To estimate the parameters of a self-limiting Hawkes process, we modify the Expectation-Maximization procedure described above. Plugging the intensity from Equation 3.1 into Equation 2.2 and then taking the expectation with respect to the probabilistic branching structure $P$ as defined by Equation 2.3 yields the following equation:

$$\begin{aligned}\mathbb{E}[\mathscr{L}] = \ln(\mu)\sum_i P_{ii} - \beta\sum_i P_{ii}N(\alpha,t_i) + \ln(k\omega)\sum_{j<i} P_{ij} - \omega\sum_{j<i} P_{ij}(t_i - t_j) \\ - \beta\sum_{j<i} P_{ij}N(\alpha,t_i) - \mu\int_0^T e^{-\beta N(\alpha,t)}dt - k\omega\int_0^T e^{-\beta N(\alpha,t)}\sum_{t_i<t} e^{-\omega(t-t_i)}dt.\end{aligned} \tag{3.3}$$

Here $P$ is defined as

$$P_{ij} = \begin{cases} \frac{\mu e^{-\beta N(\alpha,t_i)}}{\lambda(t_i)} & i = j \\ \frac{k\omega e^{-\beta N(\alpha,t_i)} e^{-\omega(t_i - t_j)}}{\lambda(t_i)} & j < i, \end{cases} \tag{3.4}$$

which is actually equivalent to the definition of $P$ in Equation 2.8 since $e^{-\beta N(\alpha,t_i)}$ is a factor in the numerator and denominator of each fraction, and thus, can be cancelled.

First, we note that on a fixed time interval from $[0,T]$, $N(\alpha,t)$ is just a piece-wise constant function

$$N(\alpha,t) = \begin{cases} n_1 & t \in [\tau_0 = 0, \tau_1] \\ n_2 & t \in (\tau_1, \tau_2] \\ \vdots & \vdots \\ n_l & t \in (\tau_{l-1}, \tau_l = T]. \end{cases} \tag{3.5}$$

The pairs $(n_i, \tau_i)$ can be easily computed by realizing that $N(\alpha,t)$ increases by exactly one at each event time $t_i$ and decreases by exactly one at each time $t_i + \alpha$. The set $\{\tau_i\}$ is

therefore constructed by taking the union of the two sets $\{t_i\}$ and $\{t_i + \alpha\}$, sorting it, and removing any entries with values greater than $T$.

Using this, we can rewrite our complete data log-likelihood as

$$\begin{aligned}\mathbb{E}[\mathscr{L}] = &\ln(\mu)\sum_i P_{ii} - \beta\sum_i P_{ii}n_{s(t_i)} + \ln(k)\sum_{j<i} P_{ij} + \ln(\omega)\sum_{j<i} P_{ij} \\ &- \omega\sum_{j<i} P_{ij}(t_i - t_j) - \beta\sum_{j<i} P_{ij}n_{s(t_i)} - \mu\sum_{i=1}^{l} e^{-\beta n_i}(\tau_i - \tau_{i-1}) \\ &+ k\sum_{i=1}^{n}\sum_{j=1}^{l} e^{-\beta n_j}\left[e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1} - t_i)}\right]\mathbb{1}_E,\end{aligned} \tag{3.6}$$

where $\mathbb{1}_E$ is the indicator function for the event $E = \{t_i < \tau_j\}$ and $s(t_i)$ is the index of $t_i$ in $\{\tau_0, \ldots, \tau_l\}$.

For more details on this calculation, see Appendix B.

Our log-likelihood function now contains five unknowns ($\mu$, $k$, $\omega$, $\alpha$, and $\beta$). For $\mu$, $k$, $\omega$, and $\beta$, we can find the respective partial derivatives of $\mathbb{E}[\mathscr{L}]$ and set them equal to $0$ in order to maximize the log-likelihood. This gives the following formulas:

$$\mu = \frac{\sum_{i=1}^{n} P_{i,i}}{\sum_{i=1}^{l} e^{-\beta n_i}(\tau_i - \tau_{i-1})}, \tag{3.7}$$

$$k = \frac{-\sum_{j<i} P_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{l} e^{-\beta n_j}\left[e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1} - t_i)}\right]\mathbb{1}_{\{t_i<\tau_j\}}}, \tag{3.8}$$

$$\begin{aligned}0 = &\frac{1}{\omega}\sum_{j<i} P_{i,j} - \sum_{j<i} P_{i,j}(t_i - t_j) + k\sum_{j=1}^{l} e^{-\beta n_j}\sum_{i=1}^{n}(t_i - \tau_j)e^{-\omega(\tau_j - t_i)}\mathbb{1}_E \\ &- k\sum_{j=1}^{l} e^{-\beta n_j}\sum_{i=1}^{n}(t_i - \tau_{j-1})e^{-\omega(\tau_{j-1} - t_i)}\mathbb{1}_E,\end{aligned} \tag{3.9}$$

and

$$
\begin{aligned}
0 = &- \sum_{i} P_{i,i} n_{s(t_i)} - \sum_{j<i} P_{i,j} n_{s(t_i)} + \mu \sum_{i=1}^{l} n_i e^{-\beta n_i} (\tau_i - \tau_{i-1}) \\
&- k \sum_{i=1}^{n} \sum_{j=1}^{l} n_j e^{-\beta n_j} \left[ e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1} - t_i)} \right] \mathbb{1}_E.
\end{aligned} \tag{3.10}
$$

However, as $N(\alpha, t)$ is not differentiable with respect to $\alpha$, we must maximize over $\alpha$ using some other method. For now, assume $\alpha$ is given, in which case the remaining parameters $\mu$, $k$, $\omega$, and $\beta$ could be found using the same basic E-M algorithm given in Figure 2.4, with the maximization step done using Equation 3.7 - Equation 3.10. In practice, however, simultaneously solving Equation 3.9 and Equation 3.10 can be quite computationally demanding. Hence, in our implementations throughout the remainder of this paper, we have chosen instead to perform a parameter sweep over $\beta$ values, numerically solving only Equation 3.9 for each of the $\beta$ values swept over, thereby also obtaining $\mu$ and $k$ from Equation 3.7 and Equation 3.8, then simply choosing the parameter combination that resulted in the highest value for $\mathscr{L}$.

To estimate $\alpha$, we also perform a parameter sweep, noting the maximal log-likelihood obtained for each test value of $\alpha$ and then simply selecting that $\alpha$, and its accompanying $\mu$, $k$, $\omega$, and $\beta$, with the greatest overall log-likelihood.

### 3.4 Estimating the Parameters of Self-limiting Spatio-temporal Hawkes Process

Just as we have done before, we will use $g(t - t_i) = k\omega e^{-\omega(t-t_i)}$ in Equation 3.2. We will also use $p(x - x_i, y - y_i) = \frac{1}{4s^2} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}}$ in the same equation.

To estimate the parameters of a self-limiting spatio-temporal Hawkes process, we will once again use a version of the Expectation-Maximization algorithm given in Figure 2.4

[16]. Plugging in the intensity from Equation 3.2 into Equation 2.9 and simplifying gives

$$\begin{aligned}\mathbb{E}[\mathscr{L}] = \ln(\mu)\sum_i P_{ii} + \sum_i P_{ii}q(t_i,x_i,y_i,\alpha,\beta) + \ln\left(\frac{k\omega}{4s^2}\right)\sum_{j<i} P_{ij} - \omega\sum_{j<i} P_{ij}(t_i-t_j) \\ - \frac{1}{s}\sum_{j<i} P_{ij}(|x_i-x_j|+|y_i-y_j|) + \sum_{j<i} P_{ij}q(t_i,x_i,y_i,\alpha,\beta) \\ - \mu\left[TL^2 + \sum_{j=1}^{l} b_j(\tau_j-\tau_{j-1})\left(e^{\frac{-\beta n_j}{b_j}}-1\right)\right] \\ + \frac{k}{4}\sum_{j=1}^{l}\sum_{i=1}^{n}\left[\left(e^{\frac{-\beta n_j}{b_j}}\frac{S_{ij}}{s^2} + \frac{S'_{ij}}{s^2}\right)\left(e^{-\omega(\tau_j-t_i)} - e^{-\omega(\tau_{j-1}-t_i)}\right)\mathbb{1}_{\{t_i<\tau_j\}}\right],\end{aligned} \tag{3.11}$$

where $\{n_1,\ldots,n_l\}$ are the discrete values of the function $N(\alpha,t)$, $\{\tau_1,\ldots,\tau_l\}$ are the times where $N(\alpha,t)$ changes value, and $b_j$ is the area of box$(t)$ in the interval $[\tau_{i-1},\tau_i)$. For more information on $\{n_1,\ldots,n_l\}$ and $\{\tau_1,\ldots,\tau_l\}$, see the previous section.

Before we define $S_{ij}$ and $S'_{ij}$, let

$$S_i(t) = \iint_{\text{box}(t)} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}}dxdy$$

and

$$S'_i(t) = \iint_{\text{box}(t)^C} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}}dxdy.$$

Since box$(t)$ and box$(t)^C$ are constants in the interval $[\tau_{j-1},\tau_j)$ and each integrand does not depend on $t$, each of these integrals is a constant in the interval $[\tau_{j-1},\tau_j)$. $S_{ij}$ is the value of $S_i(t)$ in the interval $[\tau_{j-1},\tau_j)$ and $S'_{ij}$ is the value of $S'_i(t)$ in the interval $[\tau_{j-1},\tau_j)$.

To define $S_{ij}$ and $S'_{ij}$ explicitly, let us say that box$(t) = [\text{left}_j,\text{right}_j]\times[\text{bottom}_j,\text{top}_j]$ in the interval $[\tau_{j-1},\tau_j]$. For convenience, we give explicit definition of $\frac{S_{ij}}{s^2}$ and $\frac{S'_{ij}}{s^2}$ instead of $S_{ij}$ and $S'_{ij}$. These are given by

$$\frac{S_{ij}}{s^2} = \begin{cases} \left(e^{\frac{x_i - \text{left}_j}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(e^{\frac{y_i - \text{bottom}_j}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 1} \\ \left(2 - e^{\frac{\text{left}_j - x_i}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(e^{\frac{y_i - \text{bottom}_j}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 2} \\ \left(e^{\frac{\text{right}_j - x_i}{s}} - e^{\frac{\text{left}_j - x_i}{s}}\right)\left(e^{\frac{y_i - \text{bottom}_j}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 3} \\ \left(e^{\frac{x_i - \text{left}_j}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(2 - e^{\frac{\text{bottom}_j - y_i}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 4} \\ \left(2 - e^{\frac{\text{left}_j - x_i}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(2 - e^{\frac{\text{bottom}_j - y_i}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 5} \\ \left(e^{\frac{\text{right}_j - x_i}{s}} - e^{\frac{\text{left}_j - x_i}{s}}\right)\left(2 - e^{\frac{\text{bottom}_j - y_i}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 6} \\ \left(e^{\frac{x_i - \text{left}_j}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(e^{\frac{\text{top}_j - y_i}{s}} - e^{\frac{\text{bottom}_j - y_i}{s}}\right) & \text{for Case 7} \\ \left(2 - e^{\frac{\text{left}_j - x_i}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(e^{\frac{\text{top}_j - y_i}{s}} - e^{\frac{\text{bottom}_j - y_i}{s}}\right) & \text{for Case 8} \\ \left(e^{\frac{\text{right}_j - x_i}{s}} - e^{\frac{\text{left}_j - x_i}{s}}\right)\left(e^{\frac{\text{top}_j - y_i}{s}} - e^{\frac{\text{bottom}_j - y_i}{s}}\right) & \text{for Case 9,} \end{cases}$$

where

$$\text{Case 1} = \{(x_i, y_i) : x_i \leq \text{left}_j, y_i \leq \text{bottom}_j\}$$

$$\text{Case 2} = \{(x_i, y_i) : \text{left}_j < x_i < \text{right}_j, y_i \leq \text{bottom}_j\}$$

$$\text{Case 3} = \{(x_i, y_i) : x_i \geq \text{right}_j, y_i \leq \text{bottom}_j\}$$

$$\text{Case 4} = \{(x_i, y_i) : x_i \leq \text{left}_j, \text{bottom}_j < y_i < \text{top}_j\}$$

$$\text{Case 5} = \{(x_i, y_i) : \text{left}_j < x_i < \text{right}_j, \text{bottom}_j < y_i < \text{top}_j\}$$

$$\text{Case 6} = \{(x_i, y_i) : x_i \geq \text{right}_j, \text{bottom}_j < y_i < \text{top}_j\}$$

$$\text{Case 7} = \{(x_i, y_i) : x_i \leq \text{left}_j, y_i \geq \text{top}_j\}$$

$$\text{Case 8} = \{(x_i, y_i) : \text{left}_j < x_i < \text{right}_j, y_i \geq \text{top}_j\}$$

$$\text{Case 9} = \{(x_i, y_i) : x_i \geq \text{right}_j, y_i \geq \text{top}_j\}.$$

Then we have

$$\frac{S'_{ij}}{s^2} = \left(2 - e^{\frac{-x_i}{s}} - e^{\frac{-(L-x_i)}{s}}\right)\left(2 - e^{\frac{-y_i}{s}} - e^{\frac{-(L-y_i)}{s}}\right) - \frac{S_{ij}}{s^2}.$$

For more details on the derivation of Equation 3.11, see Appendix B.

To maximize $\mathbb{E}[\mathscr{L}]$ with respect to $\mu$, $k$, $\omega$, and $s$, we can once again take the respective partial derivatives and set them equal to 0. This gives us the following formulas:

$$\mu = \frac{\sum_i P_{ii}}{TL^2 + \sum_{j=1}^{l} b_j(\tau_j - \tau_{j-1})\left(e^{\frac{-\beta n_j}{b_j}} - 1\right)},$$

$$k = \frac{-4\sum_{j<i} P_{ij}}{\sum_{j=1}^{l}\sum_{i=1}^{n}\left[\left(e^{\frac{-\beta n_j}{b_j}}\frac{S_{ij}}{s^2} + \frac{S'_{ij}}{s^2}\right)\left(e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1}-t_i)}\right)\mathbb{1}_{\{t_i<\tau_j\}}\right]},$$

$$0 = \sum_{j<i} P_{ij} - \omega\sum_{j<i} P_{ij}(t_i - t_j) + \frac{k\omega}{4}\sum_{j=1}^{l}\sum_{i=1}^{n}\left[\left(e^{\frac{-\beta n_j}{b_j}}\frac{S_{ij}}{s^2} + \frac{S'_{ij}}{s^2}\right)\left((t_i - \tau_j)e^{-\omega(\tau_j - t_i)} - (t_i - \tau_{j-1})e^{-\omega(\tau_{j-1}-t_i)}\right)\mathbb{1}_{\{t_i<\tau_j\}}\right],$$

and

$$0 = -2s\sum_{j<i} P_{ij} + \sum_{j<i} P_{ij}(|x_i - x_j| + |y_i - y_j|) + \frac{k}{4}\sum_{j=1}^{l}\sum_{i=1}^{n}\left[\left(e^{\frac{-\beta n_j}{b_j}} s^2\frac{d}{ds}\left(\frac{S_{ij}}{s^2}\right) + s^2\frac{d}{ds}\left(\frac{S'_{ij}}{s^2}\right)\right)\left(e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1}-t_i)}\right)\mathbb{1}_{\{t_i<\tau_j\}}\right].$$

Here $\frac{d}{ds}\left(\frac{S_{ij}}{s^2}\right)$ and $\frac{d}{ds}\left(\frac{S'_{ij}}{s^2}\right)$ are again defined piecewise using the same nine cases as before:

$$\text{Case 1: } \frac{d}{ds}\left(\frac{S_{ij}}{s^2}\right) = \frac{1}{s^2}\Bigg[ -(-\text{left}_j - \text{bottom}_j + x_i + y_i)e^{\frac{-\text{left}_j - \text{bottom}_j + x_i + y_i}{s}}$$
$$+ (-\text{left}_j - \text{top}_j + x_i + y_i)e^{\frac{-\text{left}_j - \text{top}_j + x_i + y_i}{s}}$$
$$+ (-\text{right}_j - \text{bottom}_j + x_i + y_i)e^{\frac{-\text{right}_j - \text{bottom}_j + x_i + y_i}{s}}$$
$$- (-\text{right}_j - \text{top}_j + x_i + y_i)e^{\frac{-\text{right}_j - \text{top}_j + x_i + y_i}{s}}\Bigg],$$

$$\text{Case 2: } \frac{d}{ds}\left(\frac{S_{ij}}{s^2}\right) = \frac{1}{s^2}\Bigg[ -2(y_i - \text{bottom}_j)e^{\frac{y_i - \text{bottom}_j}{s}} + 2(y_i - \text{top}_j)e^{\frac{y_i - \text{top}_j}{s}}$$
$$+ (\text{left}_j - \text{bottom}_j - x_i + y_i)e^{\frac{\text{left}_j - \text{bottom}_j - x_i + y_i}{s}}$$
$$- (\text{left}_j - \text{top}_j - x_i + y_i)e^{\frac{\text{left}_j - \text{top}_j - x_i + y_i}{s}}$$
$$+ (-\text{right}_j - \text{bottom}_j + x_i + y_i)e^{\frac{-\text{right}_j - \text{bottom}_j + x_i + y_i}{s}}$$
$$- (-\text{right}_j - \text{top}_j + x_i + y_i)e^{\frac{-\text{right}_j - \text{top}_j + x_i + y_i}{s}}\Bigg],$$

$$\text{Case 3: } \frac{d}{ds}\left(\frac{S_{ij}}{s^2}\right) = \frac{1}{s^2}\Bigg[ -(\text{right}_j - \text{bottom}_j - x_i + y_i)e^{\frac{\text{right}_j - \text{bottom}_j - x_i + y_i}{s}}$$
$$+ (\text{right}_j - \text{top}_j - x_i + y_i)e^{\frac{\text{right}_j - \text{top}_j - x_i + y_i}{s}}$$
$$+ (\text{left}_j - \text{bottom}_j - x_i + y_i)e^{\frac{\text{left}_j - \text{bottom}_j - x_i + y_i}{s}}$$
$$- (\text{left}_j - \text{top}_j - x_i + y_i)e^{\frac{\text{left}_j - \text{top}_j - x_i + y_i}{s}}\Bigg],$$

$$\text{Case 4: } \frac{d}{ds}\left(\frac{S_{ij}}{s^2}\right) = \frac{1}{s^2}\Bigg[ -2(x_i - \text{left}_j)e^{\frac{x_i-\text{left}_j}{s}} + 2(x_i - \text{right}_j)e^{\frac{x_i-\text{right}_j}{s}}$$
$$+ (\text{bottom}_j - \text{left}_j - y_i + x_i)e^{\frac{\text{bottom}_j-\text{left}_j-y_i+x_i}{s}}$$
$$- (\text{bottom}_j - \text{right}_j - y_i + x_i)e^{\frac{\text{bottom}_j-\text{right}_j-y_i+x_i}{s}}$$
$$+ (-\text{top}_j - \text{left}_j + y_i + x_i)e^{\frac{-\text{top}_j-\text{left}_j+y_i+x_i}{s}}$$
$$- (-\text{top}_j - \text{right}_j + y_i + x_i)e^{\frac{-\text{top}_j-\text{right}_j+y_i+x_i}{s}}\Bigg],$$

$$\text{Case 5: } \frac{d}{ds}\left(\frac{S_{ij}}{s^2}\right) = \frac{1}{s^2}\Bigg[2(\text{bottom}_j - y_i)e^{\frac{\text{bottom}_j-y_i}{s}} + 2(y_i - \text{top}_j)e^{\frac{y_i-\text{top}_j}{s}}$$
$$+ 2(\text{left}_j - x_i)e^{\frac{\text{left}_j-x_i}{s}} - (\text{left}_j + \text{bottom}_j - x_i - y_i)e^{\frac{\text{left}_j+\text{bottom}_j-x_i-y_i}{s}}$$
$$- (\text{left}_j - \text{top}_j - x_i + y_i)e^{\frac{\text{left}_j-\text{top}_j-x_i+y_i}{s}} + 2(x_i - \text{right}_j)e^{\frac{x_i-\text{right}_j}{s}}$$
$$- (-\text{right}_j + \text{bottom}_j + x_i - y_i)e^{\frac{-\text{right}_j+\text{bottom}_j+x_i-y_i}{s}}$$
$$- (-\text{right}_j - \text{top}_j + x_i + y_i)e^{\frac{-\text{right}_j-\text{top}_j+x_i+y_i}{s}}\Bigg],$$

$$\text{Case 6: } \frac{d}{ds}\left(\frac{S_{ij}}{s^2}\right) = \frac{1}{s^2}\Bigg[ -2(\text{right}_j - x_i)e^{\frac{\text{right}_j-x_i}{s}} + 2(\text{left}_j - x_i)e^{\frac{\text{left}_j-x_i}{s}}$$
$$+ (\text{bottom}_j + \text{right}_j - x_i - y_i)e^{\frac{\text{bottom}_j+\text{right}_j-x_i-y_i}{s}}$$
$$- (\text{bottom}_j + \text{left}_j - x_i - y_i)e^{\frac{\text{bottom}_j+\text{left}_j-x_i-y_i}{s}}$$
$$+ (\text{right}_j - \text{top}_j + y_i - x_i)e^{\frac{\text{right}_j-\text{top}_j+y_i-x_i}{s}}$$
$$- (\text{left}_j - \text{top}_j + y_i - x_i)e^{\frac{\text{left}_j-\text{top}_j+y_i-x_i}{s}}\Bigg],$$

$$\text{Case 7: } \frac{d}{ds}\left(\frac{S_{ij}}{s^2}\right) = \frac{1}{s^2}\Bigg[ -(\text{top}_j - \text{left}_j - y_i + x_i)e^{\frac{\text{top}_j - \text{left}_j - y_i + x_i}{s}} + (\text{top}_j - \text{right}_j - y_i + x_i)e^{\frac{\text{top}_j - \text{right}_j - y_i + x_i}{s}} + (\text{bottom}_j - \text{left}_j - y_i + x_i)e^{\frac{\text{bottom}_j - \text{left}_j - y_i + x_i}{s}} - (\text{bottom}_j - \text{right}_j - y_i + x_i)e^{\frac{\text{bottom}_j - \text{right}_j - y_i + x_i}{s}}\Bigg],$$

$$\text{Case 8: } \frac{d}{ds}\left(\frac{S_{ij}}{s^2}\right) = \frac{1}{s^2}\Bigg[ -2(\text{top}_j - y_i)e^{\frac{\text{top}_j - y_i}{s}} + 2(\text{bottom}_j - y_i)e^{\frac{\text{bottom}_j - y_i}{s}} + (\text{left}_j + \text{top}_j - y_i - x_i)e^{\frac{\text{left}_j + \text{top}_j - y_i - x_i}{s}} - (\text{left}_j + \text{bottom}_j - y_i - x_i)e^{\frac{\text{left}_j + \text{bottom}_j - y_i - x_i}{s}} + (\text{top}_j - \text{right}_j + x_i - y_i)e^{\frac{\text{top}_j - \text{right}_j + x_i - y_i}{s}} - (\text{bottom}_j - \text{right}_j + x_i - y_i)e^{\frac{\text{bottom}_j - \text{right}_j + x_i - y_i}{s}}\Bigg],$$

and

$$\text{Case 9: } \frac{d}{ds}\left(\frac{S_{ij}}{s^2}\right) = \frac{1}{s^2}\Bigg[ -(\text{right}_j + \text{top}_j - x_i - y_i)e^{\frac{\text{right}_j + \text{top}_j - x_i - y_i}{s}} + (\text{right}_j + \text{bottom}_j - x_i - y_i)e^{\frac{\text{right}_j + \text{bottom}_j - x_i - y_i}{s}} + (\text{left}_j + \text{top}_j - x_i - y_i)e^{\frac{\text{left}_j + \text{top}_j - x_i - y_i}{s}} - (\text{left}_j + \text{bottom}_j - x_i - y_i)e^{\frac{\text{left}_j + \text{bottom}_j - x_i - y_i}{s}}\Bigg].$$

Just like in the standard spatio-temporal Hawkes case, we can rewrite $\frac{\partial \mathbb{E}[\mathscr{L}]}{\partial \omega} = 0$ and $\frac{\partial \mathbb{E}[\mathscr{L}]}{\partial s} = 0$ in the form $\omega = f_1(k, \omega, s)$ and $s = f_2(k, \omega, s)$ and then evaluate these two functions at the values of $k$, $\omega$, and $s$ from the previous iteration to update these parameters.

Rewriting $\frac{\partial \mathbb{E}[\mathscr{L}]}{\partial \omega} = 0$ and $\frac{\partial \mathbb{E}[\mathscr{L}]}{\partial s} = 0$ in this way gives

$$\omega = \frac{\sum\limits_{j<i} P_{ij}}{\sum\limits_{j<i} P_{ij}(t_i - t_j) - A(k, \omega, s)}$$

and

$$s = \frac{\sum\limits_{j<i} P_{ij}(|x_i - x_j| + |y_i - y_j|) + A'(k, \omega, s)}{2\sum\limits_{j<i} P_{ij}},$$

where

$$A(k, \omega, s) = \frac{k}{4} \sum_{j=1}^{l} \sum_{i=1}^{n} \left[ \left( e^{\frac{-\beta n_j}{b_j}} \frac{S_{ij}}{s^2} + \frac{S'_{ij}}{s^2} \right) \right.$$
$$\left. \left( (t_i - \tau_j) e^{-\omega(\tau_j - t_i)} - (t_i - \tau_{j-1}) e^{-\omega(\tau_{j-1} - t_i)} \right) \mathbb{1}_{\{t_i < \tau_j\}} \right]$$

and

$$A'(k, \omega, s) = \frac{k}{4} \sum_{j=1}^{l} \sum_{i=1}^{n} \left[ \left( e^{\frac{-\beta n_j}{b_j}} s^2 \frac{d}{ds} \left( \frac{S_{ij}}{s^2} \right) + s^2 \frac{d}{ds} \left( \frac{S'_{ij}}{s^2} \right) \right) \right.$$
$$\left. \left( e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1} - t_i)} \right) \mathbb{1}_{\{t_i < \tau_j\}} \right].$$

# CHAPTER 4

# TESTING PARAMETER ESTIMATION

## 4.1 Hawkes E-M Algorithm Vs. Self-limiting Hawkes E-M Algorithm

We tested our method over the following sets of parameters for the preventative action:

$$\alpha \in \{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5\},$$

$$\beta \in \{0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045\}.$$

We tested each value in each of these sets by choosing one parameter to vary while holding the other parameter constant at the fifth value in its set. So while we were testing the different values of $\alpha$, $\beta$ was fixed at $0.025$. Likewise, while we were testing the different values of $\beta$, $\alpha$ was fixed at $2.5$. For each test, $\mu$, $k$, and $\omega$ were fixed at $0.65$, $0.65$, and $50$, respectively.

For a particular set of values of $\alpha$ and $\beta$, $100$ realizations of a Hawkes process on time interval $[0, 1000]$ were created with the given parameters. When simulating these processes, we employed the thinning algorithm. This was done so that each realization would yield two datasets: one a standard Hawkes process, which we will refer to as the hypothetical dataset, which could be interpreted as the set of crimes that might have occurred with no police intervention; the other, which we will refer to as the self-limiting dataset, a corresponding subset of the hypothetical dataset representing the full self-limiting process, which can be thought of as the set of crimes that might have actually occurred when police deterrence was in place.

For each realization, the parameters $\mu$, $k$, and $\omega$ were then estimated using the standard Hawkes process E-M algorithm on both resulting datasets, and using our self-limiting E-M

algorithm on the self-limiting dataset only. For this test, when using our self-limiting E-M algorithm on the self-limiting dataset, we used the true values of $\alpha$ and $\beta$ to estimate the other parameters. This allows us to determine the extent to which the self-limiting aspect of the process affects parameter estimation. This is an important point to consider, given that current applications to crime data do not explicitly consider the effects of police activity, and therefore may have systematically biased estimates for parameter values.

To determine how well parameters have been estimated in each of these various scenarios, we consider relative error metrics for each of the estimated parameters. For example, if $\mu_e^{(i)}$ is the estimated value of $\mu$ for the $i^{\text{th}}$ Hawkes process in one of the three scenarios, then our average relative error over the 100 realizations for that scenario is

$$\frac{1}{100}\sum_{i=1}^{100}\frac{|\mu_e^{(i)}-\mu|}{\mu};$$

corresponding values are computed for the other parameters.

Results are shown in Figure 4.1 and Figure 4.2. Each figure consists of nine different plots. Each row shows the relative errors in estimates of $\mu$, $k$, and $\omega$ under one of the three scenarios. The top row shows the errors in estimation for the standard Hawkes datasets using the standard Hawkes E-M algorithm, and serves as a control. The middle row shows the errors when the standard Hawkes E-M algorithm is used on the self-limiting datasets. The bottom row shows the errors in estimation when using our self-limiting E-M algorithm on the self-limiting datasets.

Across $\alpha$ and $\beta$ values, we find that estimation error for each of the three parameters is roughly the same when comparing the standard Hawkes datasets estimated via standard Hawkes E-M (top rows) to the self-limiting datasets estimated via our self-limiting E-M (bottom rows). This shows that our algorithm is able to estimate the parameters of a self-limiting Hawkes process as well as can be done for a standard Hawkes process of the same parameters. The middle rows show how mis-specification of the model – using standard Hawkes as the model when the process is in reality self-limiting – can lead to significant

errors in parameter estimation, in a systematic way. Recall that $\alpha$ can be thought of as the memory of the police force and $\beta$ can be thought of as the intensity of the police force. So as $\alpha$ and $\beta$ increase, the number of events prevented within the hypothetical datasets increases. As more events are prevented, we should expect that the accuracy of the estimation via standard Hawkes E-M should decrease as we lose more information about the underlying Hawkes process, which is precisely what we find for parameters $\mu$ and $k$. However, the error in estimating parameter $\omega$ is not very sensitive to the precise value of $\alpha$ or $\beta$ used.

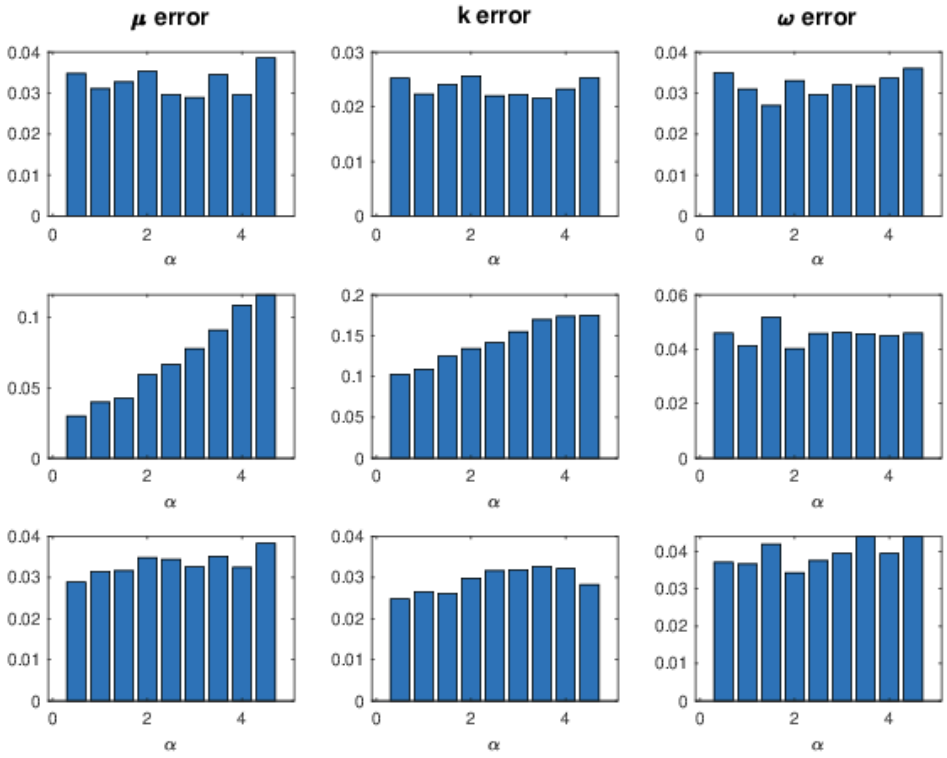


Figure 4.1: Parameter estimation error when $\alpha$ is varied under various testing conditions described in the text.

Next, we tested how well the parameters are estimated when using the parameter sweep method for estimating $\alpha$ and $\beta$. To do this, we first chose two sets of self-limiting Hawkes parameters on which to test the estimation:

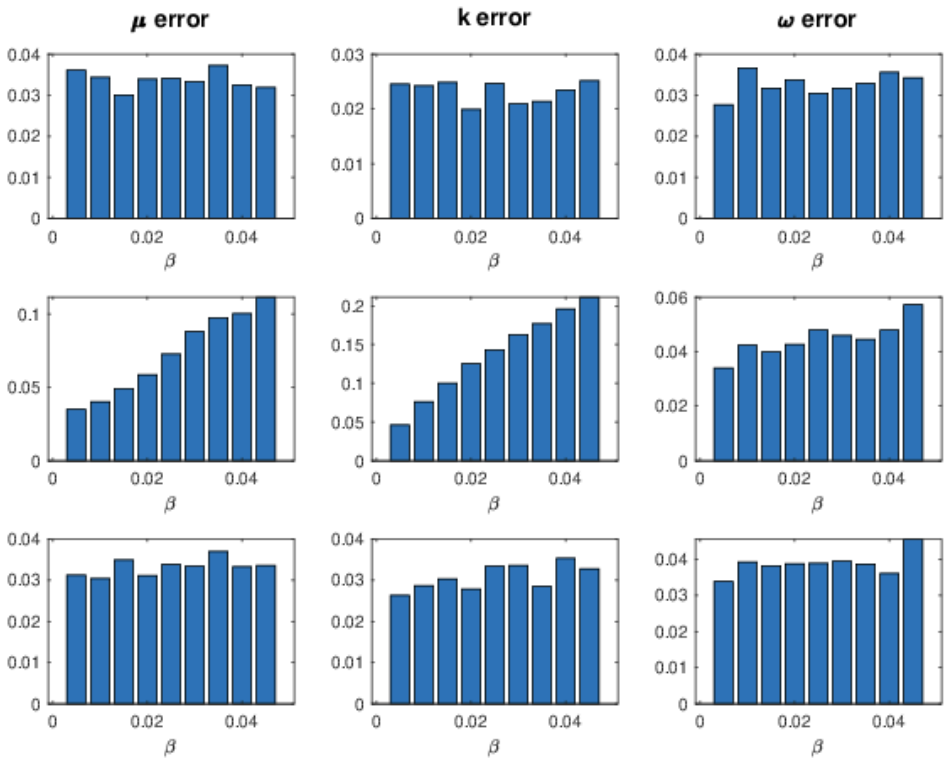


Figure 4.2: Parameter estimation error when $\beta$ is varied under various testing conditions described in the text.

$$\{\mu = 0.65, k = 0.65, \omega = 50, \alpha = 2, \beta = 0.05\},$$

$$\{\mu = 0.65, k = 0.65, \omega = 50, \alpha = 5, \beta = 0.01\},$$

We then generated $100$ self-limiting Hawkes processes for each set of parameters. For each process, we used the self-limiting E-M algorithm along with the parameter sweep method to estimate the parameters. Table 4.1 and Table 4.2 show the average values of each of the five parameters estimated using this method as well as the percent error between the average estimated and true values of the parameters for both sets of true parameters. As we can see, even though the estimation of $\alpha$ and $\beta$ leads to much higher errors than is found with the other parameters, the values found are still reasonable and lead to accurate estimation of $\mu$, $k$, and $\omega$.

Table 4.1: The average estimated parameters compared with the true parameters for the first set of true parameters.

| Parameter | True Values | Average Estimated Values | Percent Error (%) |
|---|---|---|---|
| $\mu$ | 0.65 | 0.6637 | 2.11 |
| $k$ | 0.65 | 0.6449 | -0.78 |
| $\omega$ | 50 | 51.5862 | 3.17 |
| $\alpha$ | 2 | 3.15 | 57.50 |
| $\beta$ | 0.05 | 0.0429 | -14.20 |

Table 4.2: The average estimated parameters compared with the true parameters for the second set of true parameters.

| Parameter | True Values | Average Estimated Values | Percent Error (%) |
|---|---|---|---|
| $\mu$ | 0.65 | 0.6955 | 7.00 |
| $k$ | 0.65 | 0.7138 | 9.82 |
| $\omega$ | 50 | 50.6129 | 1.23 |
| $\alpha$ | 5 | 3.745 | -25.10 |
| $\beta$ | 0.01 | 0.0261 | 161.00 |

## 4.2 Spatio-temporal Hawkes E-M Algorithm Vs. Self-limiting Spatio-temporal Hawkes E-M Algorithm

To test our self-limiting spatio-temporal Hawkes method, we repeated the same analysis as was done in section 4.1. Just like before, we tested over the following sets of parameters:

$$\alpha \in \{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5\},$$

$$\beta \in \{0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045\},$$

with $\mu$, $k$, $\omega$, and $s$ fixed at $0.65$, $0.65$, $50$, and $0.1$, respectively. In this analysis, for a particular set of values of $\alpha$ and $\beta$, $10$ realizations of a Hawkes process on time interval $[0, 100]$ and the spatial region $[0, 10] \times [0, 10]$ were created with the given parameters. Recall from section 4.1, that in the prior analysis we used $100$ realizations of a Hawkes process, each on the time interval $[0, 1000]$, for each pair of parameters. Due to the com-

putational expense of our spatio-temporal methods, repeating the prior analysis was not feasible without reducing the number of realizations as well the size of the time interval.

Results are shown in Figure 4.3 and Figure 4.4. Each subplot shows how the error in the estimation of one of the parameters changes as either $\alpha$ varies (Figure 4.3) or $\beta$ varies (Figure 4.4). The top row shows the errors in estimation for the standard Hawkes datasets using the standard Hawkes E-M algorithm, and serves as a control. The middle row shows the errors when the standard Hawkes E-M algorithm is used on the self-limiting datasets. The bottom row shows the errors in estimation when using our self-limiting E-M algorithm on the self-limiting datasets. For more information about the testing conditions described above, refer back to section 4.1.

Across $\alpha$ and $\beta$ values, we find that estimation error for each of the four parameters is roughly the same when comparing the standard spatio-temporal Hawkes datasets estimated via standard spatio-temporal Hawkes E-M (top rows) to the self-limiting datasets estimated via our self-limiting spatio-temporal E-M (bottom rows). This shows that our algorithm is able to estimate the parameters of a self-limiting spatio-temporal Hawkes process as well as can be done for a standard spatio-temporal Hawkes process of the same parameters.

While the errors displayed in the bottom rows are comparable to those in the top rows, we should note that the errors in estimating $\mu$, $k$, $\omega$, and $s$ are increasing as $\alpha$ and $\beta$ increase, a phenomenon that we did not see to this degree in our previous analysis. This could be caused by the fact that in a realization of a spatio-temporal Hawkes process, preventing a single event causes a loss of information larger than the prevention of a single event in a temporal Hawkes process. We suspect that if we redid the analysis in section 4.1 with larger values of $\alpha$ and $\beta$, we would eventually see the same behavior. Another possible cause for the experiment in which we allowed $\alpha$ to vary is that in this analysis the $\alpha$ values used were a greater percentage of $T$ than they were in the previous analysis. Thus, a higher percentage of events were blocked.

The middle rows show how mis-specification of the model – using standard spatio-

temporal Hawkes as the model when the process is in reality self-limiting – can lead to significant errors in parameter estimation, in a systematic way. Recall that $\alpha$ can be thought of as the memory of the police force and $\beta$ can be thought of as the intensity of the police force. So as $\alpha$ and $\beta$ increase, the number of events prevented within the hypothetical datasets increases. As more events are prevented, we should expect that the accuracy of the estimation via standard Hawkes E-M should decrease as we lose more information about the underlying Hawkes process, which is precisely what we find for parameters $\mu$ and $\omega$. This time, the error in estimating parameter $k$ and $s$ does not seem very sensitive to the precise value of $\alpha$ or $\beta$ used, however, I expect that $k$ would exhibit similar behavior to $\mu$ and $\omega$ given a greater number of iterations.

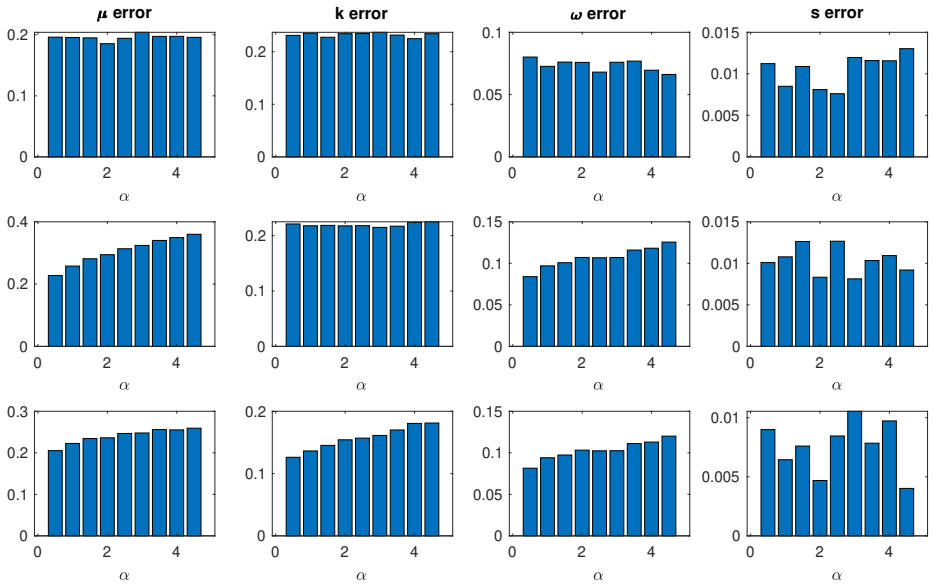


Figure 4.3: Parameter estimation error when $\alpha$ is varied under various testing conditions described in the text.

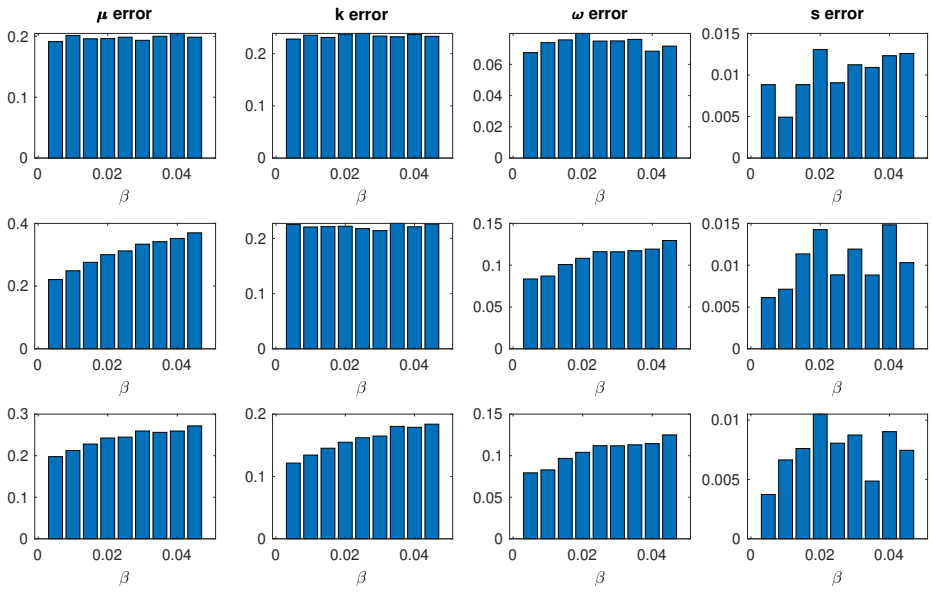


Figure 4.4: Parameter estimation error when $\beta$ is varied under various testing conditions described in the text.

# CHAPTER 5
# RESULTS USING REAL CRIME DATA

## 5.1 Hawkes Model Vs. Self-limiting Hawkes Model

Here, we employ our self-limiting Hawkes process on crime data from Chicago, obtained via their open access portal [17]. We stress here that the main purpose of this analysis is not to establish that the self-limiting Hawkes process is superior to the standard Hawkes process in describing real world crime data; indeed, as we will show, neither model seems especially well fitting to the data analyzed here, for reasons we hypothesize below. Instead, this analysis is meant only to establish the plausibility of the model with respect to an often-used alternative (the standard Hawkes with exponentially decaying excited kernel), and show that one can readily fit the self-limiting Hawkes to real world datasets.

The dataset considered contains the times and locations of all burglaries in Chicago from 2001 to 2020. We selected burglaries for the tests in this section because the "repeat victimization" of burglary targets has been studied in depth and is widely accepted [9, 18]. It is suspected that many other crime types show this repeat victimization as well [8]. While this is good evidence that pure crime data of these crime types (with no police activity) will have the self-exciting tendency that we see in a Hawkes process, without knowing more about the police response to each particular crime type, we can't say whether or not either the standard Hawkes model or our self-limiting Hawkes model will accurately represent actual data (with any police activity already baked in) of these crime types. For example, if we have a crime type where the police response to each individual event is highly variable and depends upon the circumstances of the event, then it is unlikely that either model will accurately represent actual data of this crime type. This is true even if the crime type exhibits strong repeat victimization. The response of police to burglaries likely varies less

than other types of crime, making burglary a good candidate for our model.

Though location information is provided in the dataset, we are only considering purely temporal processes here. Prior work [18] has shown that parent-daughter crime pairs are often separated by relatively small distances. To allow our temporal processes to account for this, we have binned our data into squares with sides $1500$ feet long, considering each such bin separately. In a square this size, any crime could conceivably be the daughter of any crime that occurred before it.

### 5.1.1 Residual Analysis

Having spatially binned our data, we only consider the ten squares with the highest total crime counts. For each such bin, we first divide the events into two sets, a training set (the first half of the events) and a testing set (the second half of the events). We then estimate the parameters of the training set using both the standard and self-limiting Hawkes models. As we mentioned in chapter 2 and chapter 3, when estimating the parameters of the self-limiting Hawkes model, we will be using the E-M algorithm to estimate $\mu$, $k$, and $\omega$, and these estimates will be done independently for each square. For the estimation of $\alpha$ and $\beta$, we use the same values for all squares; this seems plausible, as the response of police to crime numbers is likely more consistent across space than crime rates themselves. After performing the sweeps, we choose the values of $\alpha$ and $\beta$ that maximize the number of squares where the self-limiting Hawkes process outperformed the standard Hawkes process on the square's training set. Here, we say that model A outperforms model B if the parameters estimated using model A result in a higher log-likelihood value and lower Akeike information criteron value (to be defined later) than the parameters estimated using model B.

We choose this metric for selecting $\alpha$ and $\beta$ this way since the standard Hawkes is our baseline model, and we want to determine if the self-limiting Hawkes could potentially be a better fit than it. We note that, given this procedure, the actual best fitting self-limiting

Hawkes parameters would likely do better than what we present below, were we to allow $\alpha$ and $\beta$ to vary from square to square, and choose the parameters for each square that maximized that square's log-likelihood.

The values of the parameters found in this way are given in Table 5.1 and Table 5.2.

Table 5.1: The values of the parameters using the standard Hawkes model.

| Square | $\mu$ (days$^{-1}$) | $k$ | $\omega$ (days$^{-1}$) |
|---|---|---|---|
| 1 | 0.0917 | 0.3933 | 0.0890 |
| 2 | 0.0590 | 0.6032 | 0.0204 |
| 3 | 0.0948 | 0.1439 | 0.4550 |
| 4 | 0.0513 | 0.5644 | 0.0287 |
| 5 | 0.1111 | 0.0461 | 11.1671 |
| 6 | 0.0987 | 0.2162 | 0.3034 |
| 7 | 0.0874 | 0.3115 | 0.1268 |
| 8 | 0.0734 | 0.4469 | 0.0531 |
| 9 | 0.0737 | 0.4428 | 0.0553 |
| 10 | 0.0720 | 0.4608 | 0.0620 |

Table 5.2: The values of the parameters using the self-limiting Hawkes model using average best fit values $\alpha = 1.124$ days and $\beta = 0.03$.

| Square | $\mu$ (days$^{-1}$) | $k$ | $\omega$ (days$^{-1}$) |
|---|---|---|---|
| 1 | 0.0821 | 0.4703 | 0.0798 |
| 2 | 0.0581 | 0.6152 | 0.0275 |
| 3 | 0.0689 | 0.3901 | 0.0712 |
| 4 | 0.0585 | 0.4981 | 0.0463 |
| 5 | 0.1110 | 0.0485 | 10.8960 |
| 6 | 0.0599 | 0.5416 | 0.0460 |
| 7 | 0.0823 | 0.3621 | 0.1170 |
| 8 | 0.0692 | 0.4905 | 0.0592 |
| 9 | 0.0657 | 0.5177 | 0.0539 |
| 10 | 0.0710 | 0.4761 | 0.0724 |

To determine goodness of fit for each of these estimates, we compute the residuals $\{r_1, \ldots, r_n\}$ for each of the two models in each testing set, where

$$r_i = \int_0^{t_i} \lambda(t)dt.$$

Here, $\{t_1, \ldots, t_n\}$ are the times of the crimes in the testing set that occurred in the

current square and $\lambda$ uses our best fit parameters from the training set for that square. Note that the times in the testing set are shifted so that the final event of the training set is time $t = 0$ for the testing set.

If a model correctly represents the dataset, then the residuals should be distributed in a way consistent with a unit rate homogeneous Poisson process. Graphically, this means that when plotted as points $(i,\ r_i)$, the resulting curve should lie close to the line $y = x$. We measure the goodness of fit of each model using the Kolmogorov–Smirnov test statistic

$$\text{KS} = \max_{1 \le i \le n} |r_i - i|;$$

results are given in Table 5.3. We found that the self-limiting Hawkes model has a smaller KS test statistic than the standard Hawkes model in seven of the ten squares tested, meaning that the self-limiting model is statistically significant at a higher confidence level than the standard Hawkes process for these seven squares. However, neither was model statistically significant at the 95% confidence level in any of the squares. There are several possible reasons for this finding. One possibility is that the excited kernel for real data is not exponentially decaying; this can be observed in [10], where a non-parametric method is used to estimate the kernel $g$ of the Hawkes process, and the results are clearly not exponential decay. Another possibility is that the true values of the parameters are changing over the time period spanned by the dataset. In fact, this is especially likely since the dataset spans a very long time period (20 years) and since we trained our model solely on the first 10 years while testing solely on the last 10 years. Additionally, by spatially grouping the data into small squares with no interaction, we are introducing some error, since it is possible that families of events could straddle the border of two or more squares.

Figure 5.1 shows a graphical example of the residual analysis. Here the two solid lines on either side of the line $y = x$ represent the boundaries of the regions in which $95\%$ of Poisson processes with the same number of events fall. Since the residuals of both models go outside this region, it is not likely that they are consistent with a Poisson process with

Table 5.3: The Kolmogorov–Smirnov test statistics of the residuals using both models. Square numbers written in green designate squares where the self-limiting model outperformed the standard model while numbers in red designate the opposite.

| Square | KS (Standard) | KS (S-L) |
|---|---|---|
| 1 | 96.1088 | 91.6981 |
| 2 | 69.5929 | 71.0726 |
| 3 | 71.6870 | 49.8222 |
| 4 | 41.1858 | 45.4628 |
| 5 | 65.4766 | 65.3632 |
| 6 | 81.0281 | 55.3470 |
| 7 | 80.7324 | 79.6442 |
| 8 | 88.6134 | 87.9864 |
| 9 | 99.2195 | 93.5650 |
| 10 | 101.9143 | 103.3003 |

rate 1. Therefore, it is not likely that the crimes in this data set follow a standard Hawkes model or a self-limiting Hawkes model with the estimated values of $\alpha$ and $\beta$. In seven of the ten squares, the self-limiting model outperformed the Hawkes model even though neither fell completely in the 95% confidence region.

### 5.1.2 Log-likelihood and Akeike Information Criterion

Another way to measure the goodness of fit between models is to compare the log-likelihood and Akeike information criterion (AIC) values for the two models. The AIC is defined as

$$\text{AIC} = 2(p - \mathscr{L}),$$

where $p$ is the number of estimated parameters in the model and $\mathscr{L}$ is the log-likelihood value of the estimated parameters. The set of parameters that minimizes the AIC is the more likely model.

For this analysis, we divide the data set up into the same squares, training sets, and testing sets as we used for the residual analysis above. Once again, we estimate the parameters of the training sets using both models. We then compute the log-likelihood and AIC values

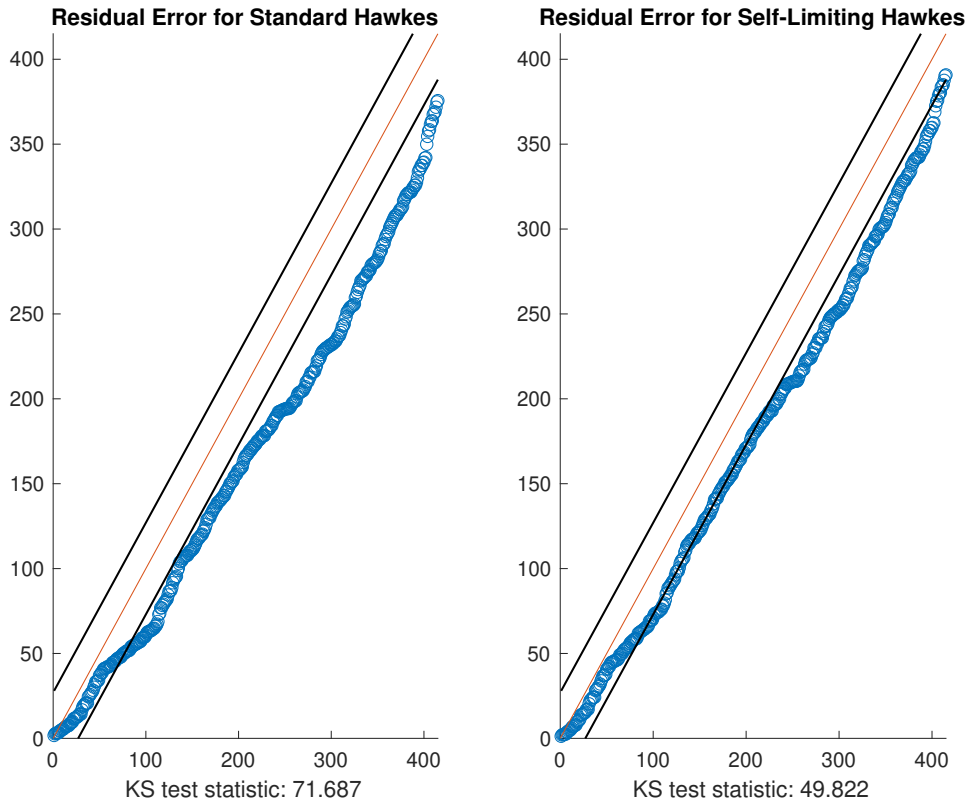


Figure 5.1: Residual analysis for square number 3.

of each testing set using the parameters estimated using the corresponding training sets. The log-likelihood values are given in Table 5.4 and the AIC values are given in Table 5.5. The right-most column of Table 5.5 lists the relative likelihoods of the better performing models, defined as

$$\text{relative likelihood} = e^{\frac{\text{AIC}_{\min} - \text{AIC}_{\max}}{2}}.$$

This measure represents how probable the higher AIC model is to minimize the information loss relative to the lower AIC model.

For example, if we look at square 3, the relative likelihood of the self-limiting Hawkes model over the standard Hawkes model is approximately $4.9547 \times 10^{-5}$. So, in this square the standard Hawkes model is $4.9547 \times 10^{-5}$ times as likely to minimize the information

loss as the self-limiting Hawkes model.

Table 5.4: The log-likelihood values for each square using both models. Square numbers written in green designate squares where the self-limiting model outperformed the standard model while numbers in red designate the opposite.

| Square | Standard Hawkes | Self-Limiting Hawkes |
|---|---|---|
| 1 | -6636.9262 | -6637.0029 |
| 2 | -6306.8672 | -6305.7260 |
| 3 | -5990.5639 | -5978.6513 |
| 4 | -5852.7513 | -5854.3436 |
| 5 | -5756.4231 | -5756.2455 |
| 6 | -5784.7157 | -5769.6630 |
| 7 | -5724.0204 | -5721.3741 |
| 8 | -5705.6152 | -5705.5996 |
| 9 | -5560.0839 | -5554.7970 |
| 10 | -5509.8785 | -5510.6495 |

Table 5.5: The AIC values for each square using both models and the relative likelihood of the better performing model. Square numbers written in green designate squares where the self-limiting model outperformed the standard model while numbers in red designate the opposite.

| Square | Standard Hawkes | Self-Limiting Hawkes | Relative Likelihood |
|---|---|---|---|
| 1 | 13279.8524 | 13284.0058 | 0.1253 |
| 2 | 12619.7345 | 12621.4520 | 0.4237 |
| 3 | 11987.1279 | 11967.3027 | $4.9547 \times 10^{-5}$ |
| 4 | 11711.5026 | 11718.6872 | 0.0275 |
| 5 | 11518.8463 | 11522.4911 | 0.1616 |
| 6 | 11575.4314 | 11549.3261 | $2.1444 \times 10^{-6}$ |
| 7 | 11454.0408 | 11452.7481 | 0.5239 |
| 8 | 11417.2304 | 11421.1992 | 0.1375 |
| 9 | 11126.1679 | 11119.5940 | 0.0374 |
| 10 | 11025.7570 | 11031.2991 | 0.0626 |

As we can see in Table 5.5, the self-limiting Hawkes model resulted in lower AIC values in four of the ten squares. Thus, in these four squares, it is more likely that the data follows a self-limiting Hawkes process rather than a standard Hawkes process. Moreover, in three of these squares, the relative likelihood values indicate that the probability of the standard Hawkes model resulting in a smaller information loss than the self-limiting Hawkes model

is effectively zero. Hence, the self-limiting Hawkes process based on an exponentially decaying excited kernel is in some circumstances a better fitting model to our crime data than the standard Hawkes process with the same form of kernel.

### 5.1.3 Receiver Operating Characteristic (ROC) Curve

Measuring the goodness of fit between the two models can also be done by measuring the predictive power of the models. One way to do this is to measure the area under the receiver operating characteristic curve.

Before we define the receiver operating characteristic curve, first recall that

$$\text{true positive rate (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

and

$$\text{false positive rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}},$$

where TP is the number of true positive results, i.e. the test correctly indicates the presence of some characteristic; TN is the number of true negative results, i.e. the test correctly indicates the absence of the characteristic; FP is the number of false positive results, i.e. the test incorrectly indicates the presence of the characteristic; and FN is the number of false negative results, i.e. the test incorrectly indicates the absence of the characteristic.

A true positive rate close to 1 means that when a characteristic is truly present, the test indicates that it is present most of the time. A false positive rate close to 0 means that when a characteristic is truly absent, the test indicates it is absent most of the time. Therefore, the closer the the true positive rate is to 1 and the the closer the false positive rate is to 0, the better the test.

The ROC curve is a way to visualize the diagnostic ability of a binary classifier using the true and false positive rates of the classifier. Given two possible states, state 1 (positive)

and state 2 (negative), and a vector where entry $i$ is the probability that element $i$ belongs to state 1, the ROC curve is a parametric plot of the true positive rate and false positive rate for each possible threshold to consider an element as belonging to state 1.

For example, suppose that we set a threshold of 0. Then the classifier will classify all results as positive, resulting in a true positive rate of 1 and a false positive rate of 1. Thus, the point $(1, 1)$ belongs to all ROC curves. Likewise, suppose we set a threshold of 1. Then the classifier will classify all results as negative, resulting in a true positive rate of 0 and a false positive rate of 0. Thus, the point $(0, 0)$ also belongs to all ROC curves. For thresholds between 0 and 1, the classifier will classify all elements with probabilities greater than the chosen threshold to be positive and all other elements to be negative. In general, this results in true and false positive rates between 0 and 1. The point $(\text{FPR}, \text{TPR})$ corresponding to our chosen threshold will then be another point on the ROC curve. By repeating this process for many different threshold values, we can approximate the ROC curve.

To compare two classifiers using their ROC curves, we can simply compare the areas under the two curves. This is known as the area under the receiver operating characteristic curve (AUROC). The better the classifier, the closer the AUROC will be to 1. The worse the classifier is, the closer the AUROC will be to 0.5 (a classifier that randomly guesses will result in a ROC curve near the line $y = x$). Therefore, a classifier with a higher AUROC is more likely to be a better classifier than one with a lower AUROC.

To apply this analysis to our standard and self-limiting Hawkes models, we first divided the Chicago burglary dataset up into the same squares as before. For each square, we started by considering the data from the first $75\%$ of the days (e.g. if a square contained 100 days worth of data, we would start by considering just the events that occurred in the first 75 days). We then estimated the parameters of each square of these truncated datasets using both the standard Hawkes and self-limiting Hawkes models. Next, we evaluated each intensity function at the values of the estimated parameters. Using this, we estimated the probability that at least one event would occur during the next day. Then the truncated

datasets for each square were updated with the next day's actual data (not predicted) and the process was repeated until we reached the final day in the dataset. This algorithm is summarized in Figure 5.2.

**Input:** A dataset with the times and locations of crimes, $\alpha$, $\beta$
**Output:** Two daily crime probability vectors for each square, one using the standard Hawkes algorithm and one using the self-limiting Hawkes model

Divide the dataset into squares with the desired side length.
**for** each square **do**
    $tempSet =$ the beginning $75\%$ of the data in each square
    **for** each day of the prediction period **do**
        $[\mu_H, k_H, \omega_H] =$ the parameters of $tempSet$ using the standard Hawkes model
        $[\mu_S, k_S, \omega_S] =$ the parameters of $tempSet$ using the self-limiting Hawkes model
        $p_H =$ the probability of a crime occurring on the current day using the Hawkes intensity function evaluated at $\mu_H, k_H, \omega_H$
        $p_S =$ the probability of a crime occurring on the current day using the self-limiting Hawkes intensity function evaluated at $\mu_S, k_S, \omega_S, \alpha, \beta$
        $tempSet = tempSet \cup$ the current day's actual data
    **end for**
**end for**

Figure 5.2: Algorithm for using both the standard Hawkes and self-limiting Hawkes model to estimate the probabilities of crimes occurring each day.

After completing this algorithm, we are left with the estimated probabilities of at least one crime occurring each on day in the last $25\%$ of the days in each square using both models. With this, we can we can create two ROC curves for each square; one for the standard Hawkes model and one for the self-limiting model.

In addition to the ten squares from the Chicago burglary dataset, we also repeated this analysis on two hypothetical datasets: one containing a realization of a standard Hawkes process with parameters similar to what we found in our real squares and one containing a realization of a self-limiting Hawkes process also with parameters similar to what we found in our real squares. On the Hawkes dataset, we created an ROC curve using only the standard Hawkes model, whereas on the self-limiting dataset, we created two ROC curves: one using the standard Hawkes model and another using the self-limiting model. This gave us a baseline of the predictability of each type of Hawkes model as well as what we might

expect to see if the model is mis-specified.

As graphical examples of ROC curves, Figure 5.3 shows the ROC curves from the two hypothetical datasets and Figure 5.4 shows the two ROC curves from square 3 of the Chicago dataset.

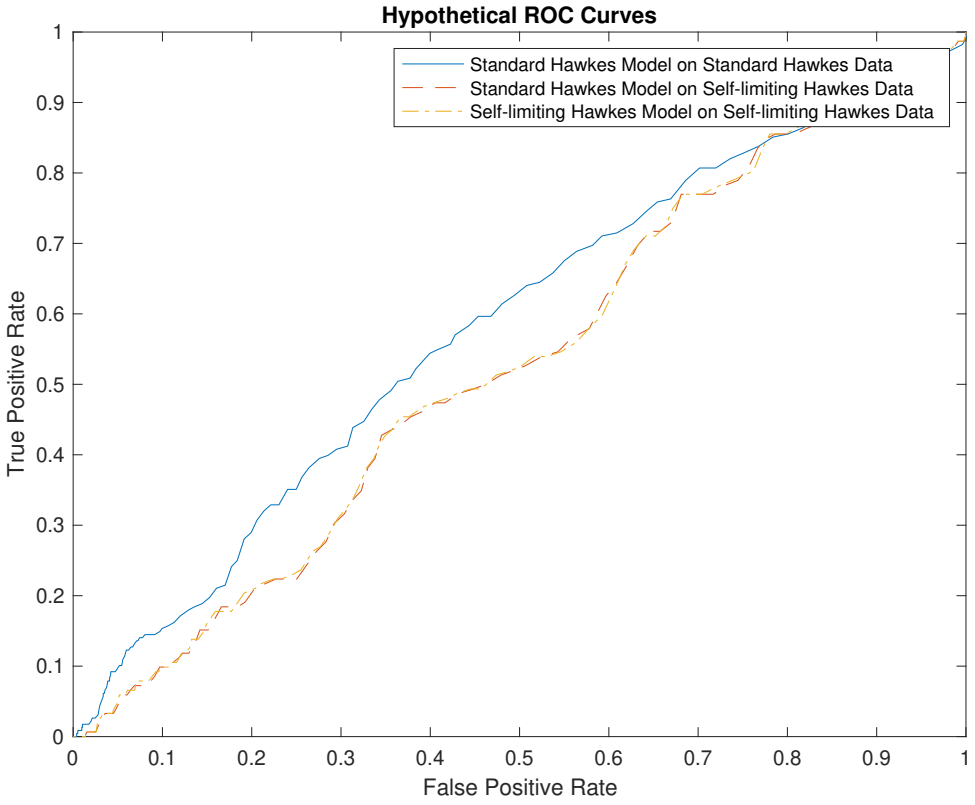


Figure 5.3: The ROC curves for the hypothetical datasets described in the text.

Additionally, the AUROC values for each of the three ROC curves created using the hypothetical datasets are given in Table 5.6 and the AUROC values from each square in the Chicago dataset are given in Table 5.7.

Table 5.6: The area under the receiver operating characteristic curve (AUROC) for hypothetical standard Hawkes dataset using the standard Hawkes model and the hypothetical self-limiting dataset using both models.

| AUROC (Standard on Standard Dataset) | AUROC (Standard on S-L Dataset) | AUROC (S-L on S-L Dataset) |
|---|---|---|
| 0.5812 | 0.5267 | 0.5279 |

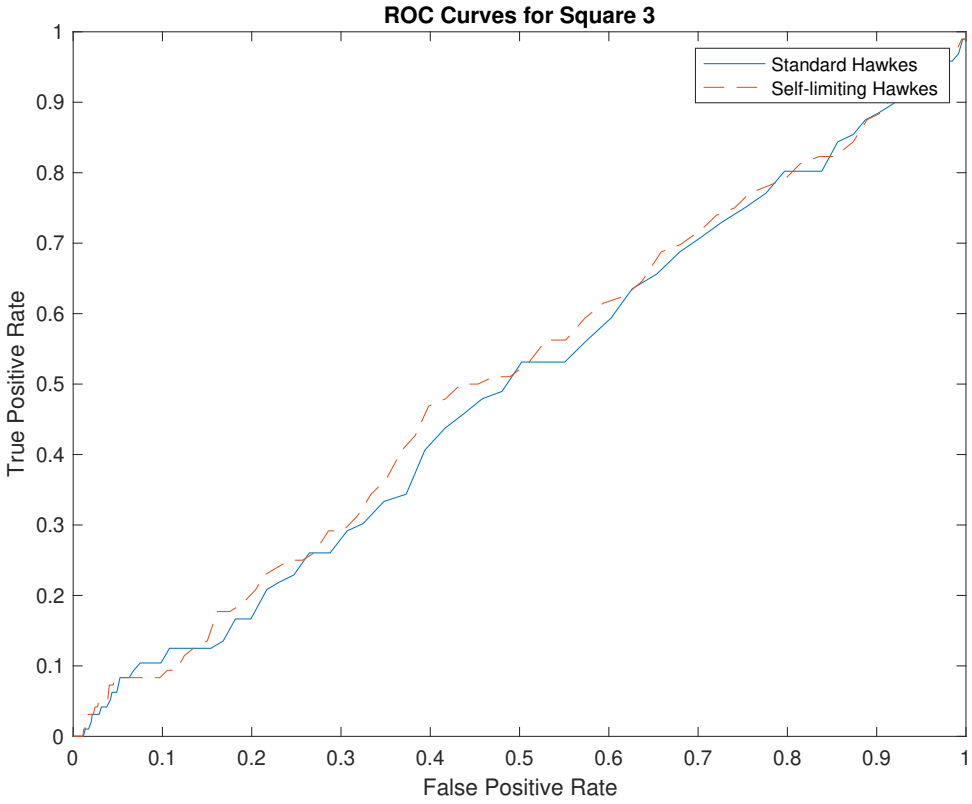


Figure 5.4: The ROC curves for the square with the third most crimes using both the standard and self-limiting Hawkes models.

Let us first examine the results from the hypothetical datasets. Intuitively, the AUROC obtained from the hypothetical standard Hawkes dataset should be the highest of the three. This will become clear once we realize what makes a stochastic process predictable. A stochastic process is predictable if its intensity function has a high degree in dependence on the history of the process up until that point. So, a process such as the Poisson process, whose intensity function has no dependence on the history of the process should be completely unpredictable leading to an AUROC of around 0.5. Since the background part of a Hawkes process is a Poisson process, the self-exciting aspect of the Hawkes process is the only part that lends itself to predictability. So, the self-limiting Hawkes process, which has a damped self-exciting effect due to the self-limiting tendency, should be less predictable than the standard Hawkes process. This is exactly what we see in Table 5.6.

Table 5.7: The area under the receiver operating characteristic curve (AUROC) for each of the squares. Square numbers written in green designate squares where the self-limiting model outperformed the standard model while numbers in red designate the opposite.

| Square | AUROC (Standard) | AUROC (S-L) |
|---|---|---|
| 1 | 0.5418 | 0.5399 |
| 2 | 0.5243 | 0.5245 |
| 3 | 0.4956 | 0.5100 |
| 4 | 0.5760 | 0.5746 |
| 5 | 0.5495 | 0.5478 |
| 6 | 0.5882 | 0.5859 |
| 7 | 0.5633 | 0.5648 |
| 8 | 0.4674 | 0.4691 |
| 9 | 0.5711 | 0.5751 |
| 10 | 0.5034 | 0.5006 |

Additionally, since each AUROC value is between 0.5 and 0.6, neither hypothetical dataset is very predictable. This means that most of the time, the background intensity, which is a Poisson process, outweighs the self-exciting part of the intensity for the Hawkes parameters that are typical of our real crime data squares. So, we shouldn't expect the prediction results in any of the squares to be much better than these hypothetical values. The purpose of this analysis is not to actually be able to predict crimes in the real world (though maybe this is possible with a change in methodology!), but to establish that the self-limiting Hawkes model is at least as likely as the standard Hawkes model to be the true model of the burglary crime data.

Now, let's examine the results from the real crime data. As predicted from our hypothetical datasets, neither model in any of the squares performed particularly well. In fact, in square 8, both models performed worse than chance! This means that they would have performed better by simply predicting a crime when they did not expect one and predicting no crime when they did expect one. In five of the ten squares, the self-limiting Hawkes model out-performed the standard Hawkes model, which is on par with the other results given in this chapter. Also, in the squares in which the standard model performed better, the AUROC of the self-limiting model was no more than $0.0028$ or $0.56\%$ less than the

AUROC of the standard model. So, according to this analysis, the self-limiting Hawkes process fits the crime data at least as well, if not slightly better, than the standard Hawkes model.

# CHAPTER 6

# RESULTS USING NON-CRIME DATASETS

Throughout this thesis, we have focused primarily on the application of self-limiting Hawkes processes to urban crime data. In chapter 1, we mention that the standard Hawkes process has been used to accurately model many systems that exhibit self-exciting tendencies. Similarly, we can use the self-limiting Hawkes process to model systems other than urban crime if we suspect that the system exhibits a self-limiting tendency alongside the self-exciting tendency.

It is well accepted that financial markets exhibit a self-exciting tendency [2, 3, 4, 5] and can be modelled with a standard Hawkes process. Most of these studies use a Hawkes process to model the arrival of orders of a particular security. We believe that price jumps of some securities exhibits both the self-exciting and self-limiting behavior typical of self-limiting Hawkes processes.

This idea relies on the assumption that there are two large groups of investors: those who buy into a security as its price rises and those who either sell or short the security as the price rises. Under this assumption, as the price of a security begins to rise, the first group of investors will start buying more of the security nudging the price higher. This causes the self-exciting tendency of the positive price jumps. At the same time, the second group of investors, believing the security to be overvalued, either sell what they already own of the security or short it, nudging the price lower. This causes the self-limiting tendency of the positive price jumps. If we consider the negative price jumps, the roles are reversed.

We believe that for some securities, this assumption is reasonable. In particular, this assumption is reasonable for high profile securities that are experiencing what most investors believe to be a bubble. Examples of this are the GamesStop and AMC stocks in early 2021 as well as the cryptocurrency Dogecoin in 2021.

## 6.1 The Data

For our analysis, we will be using the minute-by-minute close price of Dogecoin from August 1, 2021 to August 13, 2021 [19]. Before we can begin our analysis, we need to decide what constitutes an event. For the Chicago crime dataset, this was straightforward since the dataset already consisted of burglary events. The Dogecoin dataset is effectively a sampling of the Dogecoin price every minute. So, we need to decide which of these samples should be considered an event.

For this, we divide the dataset up into minutes when the price rose and minutes when the price fell. Then we only consider the minutes where the rise or fall is larger than certain thresholds. For each choice of threshold, this gives us two sequences of events, one containing the minutes where the price rose by more than the threshold and one containing the minutes where the price fell by more than threshold. We will refer to the sequence containing the price rises as dataset 1 and the sequence containing the price falls as dataset 2. We can then model each dataset of events with both the standard and self-limiting Hawkes processes and compare their performances.

For this analysis, we will be using the thresholds $0.2\%$, $0.3\%$, and $0.4\%$. Thresholds larger than around $0.4\%$ resulted in datasets without enough events to obtain meaningful results. Conversely, thresholds less than around $0.2\%$ resulted in datasets with too many events to be computationally feasible for this analysis.

## 6.2 Residual Analysis

Recall from subsection 5.1.1, that one way to determine the goodness of fit of a model is to use residual analysis. Just as we did before, we will divide each dataset into a training set (the first half of the events) and a testing set (the second half of the events). We will then estimate the parameters of each training set and use the estimated parameters to calculate the residuals on the corresponding testing sets using both the standard Hawkes and self-

limiting Hawkes models.

The values of the parameters found using the standard Hawkes model are given in Table 6.1 and the values of the parameters found using the self-limiting Hawkes model are given in Table 6.2.

Table 6.1: The values of the parameters using the standard Hawkes model. The threshold value is given in parentheses next to the dataset number.

| Dataset | $\mu$ (minutes$^{-1}$) | $k$ | $\omega$ (minutes$^{-1}$) |
|---|---|---|---|
| $1(0.2\%)$ | $0.0105$ | $0.8658$ | $0.0217$ |
| $2(0.2\%)$ | $0.0094$ | $0.8805$ | $0.0186$ |
| $1(0.3\%)$ | $0.0038$ | $0.9026$ | $0.0189$ |
| $2(0.3\%)$ | $0.0042$ | $0.8855$ | $0.0205$ |
| $1(0.4\%)$ | $0.0017$ | $0.9167$ | $0.0181$ |
| $2(0.4\%)$ | $0.0025$ | $0.8626$ | $0.0200$ |

Table 6.2: The values of the parameters using the self-limiting Hawkes model.

| Dataset | $\mu$ (minutes$^{-1}$) | $k$ | $\omega$ (minutes$^{-1}$) | $\alpha$ (minutes) | $\beta$ |
|---|---|---|---|---|---|
| $1(0.2\%)$ | $0.0115$ | $1.1336$ | $0.0426$ | $6.9$ | $0.25$ |
| $2(0.2\%)$ | $0.0098$ | $1.0941$ | $0.0311$ | $6.9$ | $0.19$ |
| $1(0.3\%)$ | $0.0040$ | $1.0095$ | $0.0256$ | $6.9$ | $0.15$ |
| $2(0.3\%)$ | $0.0044$ | $0.9438$ | $0.0244$ | $6.9$ | $0.08$ |
| $1(0.4\%)$ | $0.0017$ | $0.9319$ | $0.0187$ | $6.9$ | $0.03$ |
| $2(0.4\%)$ | $0.0025$ | $0.8626$ | $0.0200$ | $0$ | $0$ |

Note the unexpected result that in five of the six datasets, $\alpha = 6.9$ minutes. This implies that $\alpha$, which can be thought of as the "memory" of the self-limiting component, is an inherent property in the Dogecoin price dataset and is therefore not very sensitive to either our choice of threshold or choice of positive or negative price jumps. Also, note that in dataset 2 at the $0.4\%$ threshold, our estimated values of $\alpha$ and $\beta$ are both 0. According to the self-limiting model, there is no self-limiting going on in this dataset! So, we should expect that the values of the other three parameters are the same as when they are estimated with the standard Hawkes model, which is exactly what we see. Because of this, the two models will behave nearly identically throughout this analysis. We will still include this dataset in the analysis to compare with the other datasets.

Recall, that given events $\{t_1, \ldots, t_n\}$ and intensity function $\lambda$, the residuals $\{r_1, \ldots, r_n\}$ are defined as

$$r_i = \int_0^{t_i} \lambda(t)dt.$$

Note that we shifted the times in the testing sets so that the final event of each training set is time $t = 0$ for the corresponding testing set before we computed the residuals.

If a model correctly represents a dataset, we should expect that, graphically, the points $(i, r_i)$ should fall near the line $y = x$. So, we can compare multiple models by plotting the residuals obtained from each model and determining which model's residuals lie closest to the line $y = x$ under some metric. To quantify this distance from $y = x$, we used the Kolmogorov-Smirnov test statistic and the sum of the squared errors which are defined as

$$\text{KS} = \max_{1 \leq i \leq n} |r_i - i|$$

and

$$\text{Errors} = \sum_{i=1}^{n} (r_i - i)^2.$$

For more information on residual analysis and the KS test statistic, refer back to subsection 5.1.1.

The KS values for each of the datasets are given in Table 6.3. The sums of the squared errors are given in Table 6.4. As you can see, at a threshold of $0.2\%$, the standard Hawkes model outperformed the self-limiting Hawkes model on both datasets using both metrics. At a threshold of $0.3\%$, the self-limiting Hawkes model outperformed the standard Hawkes model on dataset 1 using the KS test statistic and on both datasets using the sum of squared errors. At a threshold of $0.4\%$, the self-limiting Hawkes model outperformed the standard Hawkes model on dataset 1 using both metrics and, of course, both models performed equally well on dataset 2.

Table 6.3: The Kolmogorov–Smirnov test statistics of the residuals using both models. Datasets where the self-limiting model outperformed the standard model are written in green. Datasets written in red designate the opposite. The threshold value is given in parentheses next to the dataset number.

| Dataset | KS (Standard) | KS (S-L) |
|---|---|---|
| 1(0.2%) | 54.4358 | 78.1165 |
| 2(0.2%) | 49.7013 | 65.2982 |
| 1(0.3%) | 22.2464 | 21.9434 |
| 2(0.3%) | 27.2578 | 27.9174 |
| 1(0.4%) | 12.9956 | 12.8733 |
| 2(0.4%) | 16.7154 | 16.7154 |

Table 6.4: The sums of the squared errors of the residuals using both models. Datasets where the self-limiting model outperformed the standard model are written in green. Datasets written in red designate the opposite. The threshold value is given in parentheses next to the dataset number.

| Dataset | Error (Standard) | Error (S-L) |
|---|---|---|
| 1(0.2%) | $9.0841 \times 10^5$ | $1.6692 \times 10^6$ |
| 2(0.2%) | $8.4575 \times 10^5$ | $1.2076 \times 10^6$ |
| 1(0.3%) | 72525 | 59723 |
| 2(0.3%) | $1.1283 \times 10^5$ | $1.1067 \times 10^5$ |
| 1(0.4%) | 9379 | 8945 |
| 2(0.4%) | 17016 | 17016 |

Figure 6.1 and Figure 6.2 show a graphical example of the residual analysis for the two datasets at the $0.3\%$ threshold. Just as before, the two solid lines on either side of the line $y = x$ represent the boundaries of the regions in which $95\%$ of Poisson processes with the same number of events fall.

At the $0.2\%$ threshold, in neither model did the residuals stay within this region on either dataset. So, it is not likely that they are consistent with a Poisson process with rate $1$. Therefore, it is not likely that either dataset at this threshold follow a standard Hawkes model or a self-limiting Hawkes model.

At the $0.3\%$ threshold, the residuals for dataset $1$ stay inside of this region for both models. So, both models were statistically significant at the $95\%$ confidence level for this

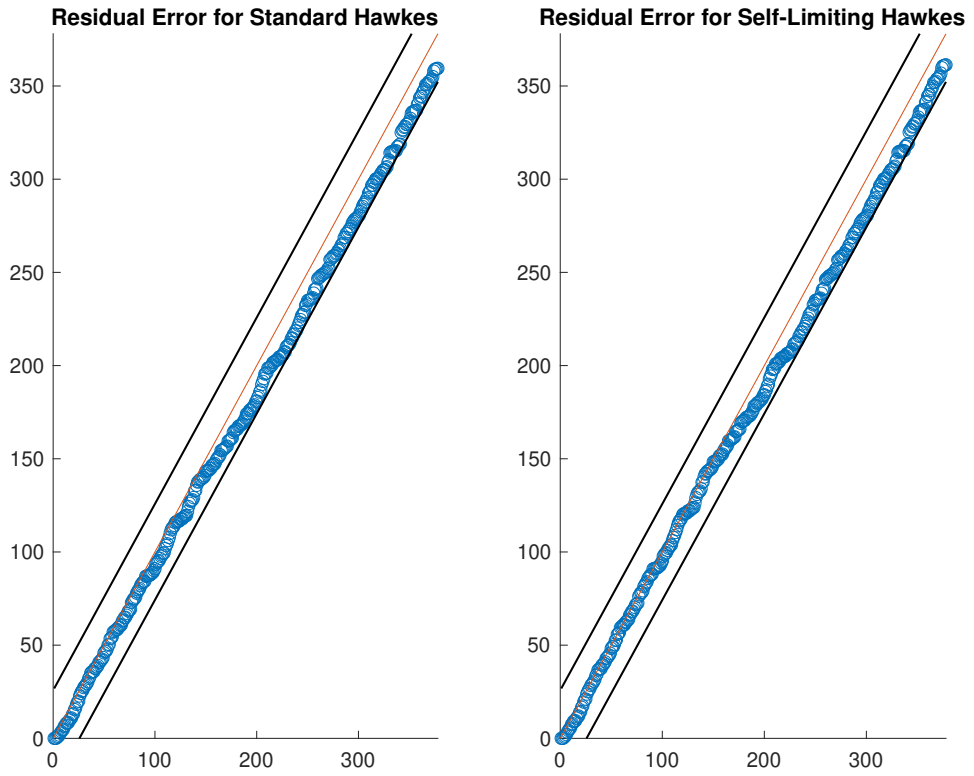


Figure 6.1: Residual analysis for dataset 1 at a threshold of $0.3\%$.

dataset, and it is plausible that the positive price jumps follow a standard Hawkes model or a self-limiting Hawkes model, though the self-limiting Hawkes model is more likely. Since the residuals for dataset 2 go outside this region, neither model was statistically significant at the $95\%$ confidence level for this dataset. Therefore, it is not likely that the negative price jumps in this data set follow a standard Hawkes model or a self-limiting Hawkes model.

At the $0.4\%$ threshold, dataset 1 stays inside of this region for both models. So, both models were statistically significant at the $95\%$ confidence level for this dataset, and it is plausible that the positive price jumps follow a standard Hawkes model or a self-limiting Hawkes model, though the self-limiting Hawkes model is more likely.

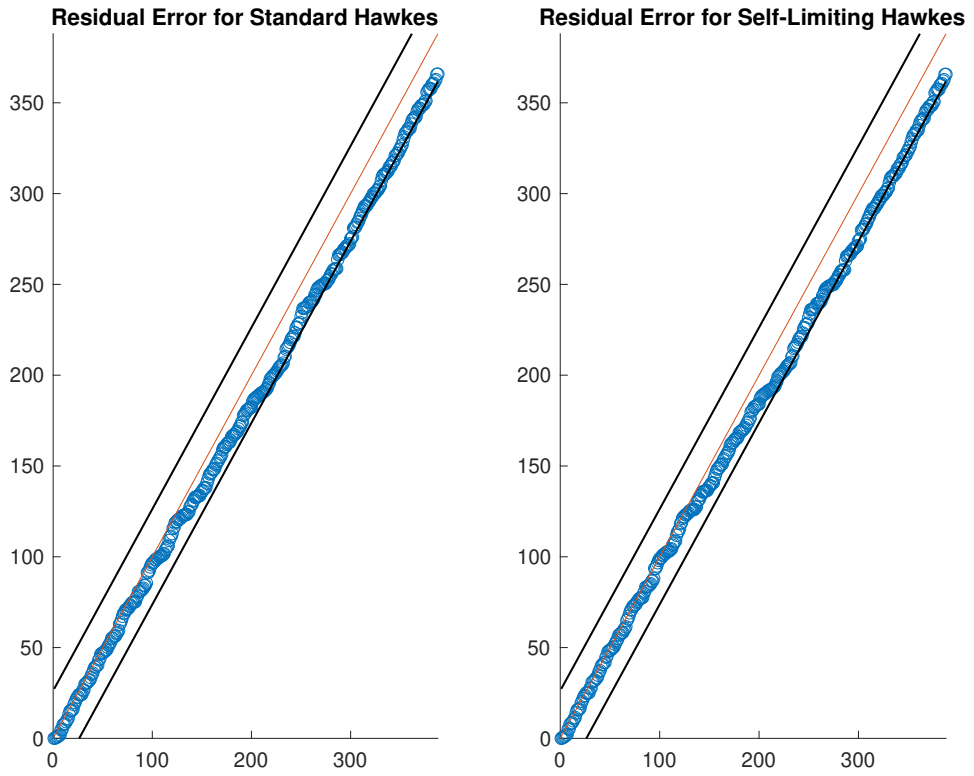


Figure 6.2: Residual analysis for dataset 2 at a threshold of $0.3\%$.

## 6.3 Log-likelihood and Akeike Information Criterion

Another way to measure the goodness of fit between models is to compare the log-likelihood and Akeike information criterion (AIC) values for the two models. The AIC is defined as

$$\text{AIC} = 2(p - \mathscr{L}),$$

where $p$ is the number of estimated parameters in the model and $\mathscr{L}$ is the log-likelihood value of the estimated parameters. We can then use the AIC values to calculate relative likelihoods of the two models on each dataset:

$$\text{relative likelihood} = e^{\frac{\text{AIC}_{\text{min}} - \text{AIC}_{\text{max}}}{2}}.$$

For more information on these metrics, refer back to subsection 5.1.2.

The log-likelihood values are given in Table 6.5 and the AIC values are given in Table 6.6. The right-most column of Table 6.6 lists the relative likelihoods of the better performing models.

Table 6.5: The log-likelihood values for each dataset using both models. Datasets where the self-limiting model outperformed the standard model are written in green. Datasets written in red designate the opposite. The threshold value is given in parentheses next to the dataset number.

| Dataset | Standard Hawkes | Self-Limiting Hawkes |
|---|---|---|
| $1(0.2\%)$ | -2411.0 | -2396.6 |
| $2(0.2\%)$ | -2421.9 | -2412.9 |
| $1(0.3\%)$ | -1383.8 | -1381.0 |
| $2(0.3\%)$ | -1384.2 | -1383.7 |
| $1(0.4\%)$ | -795.223 | -795.205 |
| $2(0.4\%)$ | -809.4 | -809.4 |

Table 6.6: The AIC values for each dataset using both models and the relative likelihood of the better performing model. Datasets where the self-limiting model outperformed the standard model are written in green. Datasets written in red designate the opposite. The threshold value is given in parentheses next to the dataset number.

| Dataset | Standard Hawkes | Self-Limiting Hawkes | Relative Likelihood |
|---|---|---|---|
| $1(0.2\%)$ | 4827.9 | 4803.2 | $4.2840 \times 10^{-6}$ |
| $2(0.2\%)$ | 4849.8 | 4835.7 | $8.6422 \times 10^{-4}$ |
| $1(0.3\%)$ | 2773.6 | 2772.0 | 0.4324 |
| $2(0.3\%)$ | 2774.4 | 2777.5 | 0.2172 |
| $1(0.4\%)$ | 1596.4 | 1601.2 | 0.0934 |
| $2(0.4\%)$ | 1624.8 | 1628.8 | 0.1353 |

As we can see in Table 6.5 and Table 6.6, at the $0.2\%$ threshold, the self-limiting model resulted in higher log-likelihood values and a lower AIC values for both datasets. Furthermore, the relative likelihood values indicate that the probability of the standard Hawkes model resulting in a smaller information loss than the self-limiting Hawkes model is effectively zero for both datasets. Hence, the self-limiting Hawkes process is in some circumstances a better fitting model to our Dogecoin price data at this threshold than the standard Hawkes process.

At the $0.3\%$ threshold, in both datasets the self-limiting Hawkes model resulted in a higher log-likelihood, but the slightly higher log-likelihood of dataset 2 was not enough overcome the extra two parameters estimated in the self-limiting model. This resulted in a higher AIC values for dataset 2. This further confirms our results from section 6.2 that a self-limiting Hawkes model is a plausible model for the positive price jumps, but it is unlikely that the negative price jumps also follow a self-limiting Hawkes model at this threshold.

At the $0.4\%$ threshold, for dataset 1, the self-limiting Hawkes model resulted in a higher log-likelihood, but a lower AIC value due to the extra two parameters estimated in the self-limiting model. Notice that for dataset 2, the AIC value due the standard Hawkes model is lower than the AIC value due to the self-limiting model even though they have identical log-likelihoods. This is because the AIC punishes a model for estimating extra parameters. Since the self-limiting model estimates two more parameters than the standard Hawkes model, we see a lower AIC value here. So, it is unlikely that the positive or negative price jumps follow a self-limiting Hawkes model at this threshold.

### 6.4 Receiver Operating Characteristic (ROC) Curve

Just as we did in subsection 5.1.3, we then measured the goodness of fit of the two models using each of their predictive powers. In particular, we compared the areas under the receiver operating characteristic curves (AUROCs) of each dataset using both models. Recall that the receiver operating characteristic curve (ROC) is a curve consisting of points of the form $(\text{FPR}, \text{TPR})$, where FPR and TPR are the false positive and true positive rates, respectively, for a given threshold for being classified as positive. We can compare the predictive power of different models by comparing the areas under their respective ROC curves; models resulting in higher AUROC values are more likely to be better fitting models to the data than models with lower AUROC values.

For more information on ROC curves and the area under them, refer back to subsec-

tion 5.1.3.

In addition to our real Dogecoin datasets, we repeated this analysis on two hypothetical datasets: one containing a realization of a standard Hawkes process with parameters similar to what we found in our real datasets and one containing a realization of a self-limiting Hawkes process also with parameters similar to what we found in our real datasets. On the Hawkes dataset, we created an ROC curve using only the standard Hawkes model, whereas on the self-limiting dataset, we created two ROC curves: one using the standard Hawkes model and another using the self-limiting model. This gave us a baseline of the predictability of each type of Hawkes model as well as what we might expect to see if the model is mis-specified.

As graphical examples of ROC curves, Figure 6.3 shows the ROC curves from the two hypothetical datasets, Figure 6.4 shows the two ROC curves from dataset 1 at the $0.3\%$ threshold, and Figure 6.5 shows the two ROC curves from dataset 2 at the $0.3\%$ threshold.

Additionally, the AUROC values for each of the three ROC curves created using the hypothetical datasets are given in Table 6.7 and the AUROC values from each Dogecoin dataset are given in Table 6.8.

Table 6.7: The area under the receiver operating characteristic curve (AUROC) for the hypothetical standard Hawkes dataset using the standard Hawkes model and the hypothetical self-limiting dataset using both models.

| AUROC (Standard on Standard Dataset) | AUROC (Standard on S-L Dataset) | AUROC (S-L on S-L Dataset) |
|---|---|---|
| 0.7191 | 0.7068 | 0.7102 |

Let us first examine the results from the hypothetical datasets. Intuitively, the AUROC obtained from the hypothetical standard Hawkes dataset should be the highest of the three. An explanation for this is given in subsection 5.1.3. This is exactly what we see in Table 5.6, though the effect is less pronounced here using Dogecoin-like parameters than it was in the previous chapter using crime-like parameters. This could be a result of a smaller self-limiting effect than was seen in the hypothetical datasets using the crime-like parameters.

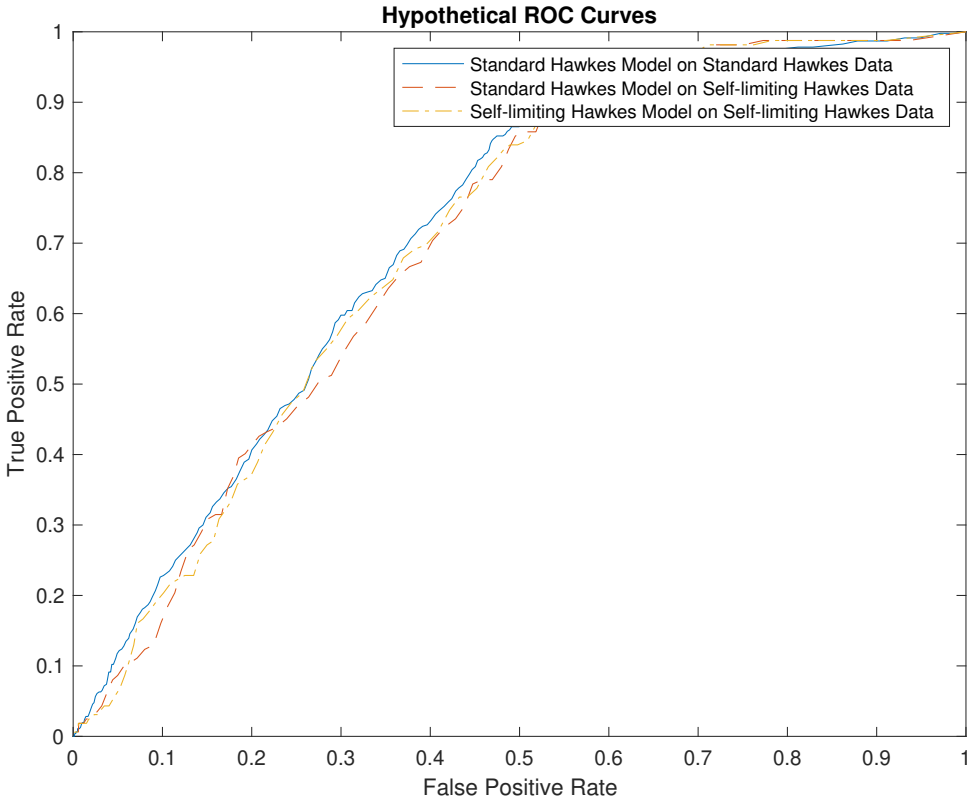


Figure 6.3: The ROC curves for the hypothetical datasets described in the text.

This would cause the self-limiting Hawkes dataset to be relatively close to the standard Hawkes dataset.

Additionally, since each AUROC value is above 0.7, both hypothetical datasets are quite predictable. This means that most of the time, the self-exciting part of the intensity, which relies on the history of the process, outweighs the background intensity for the Hawkes parameters that are typical of our Dogecoin datasets. So, we should expect the prediction results for both datasets to be similar to these hypothetical values.

Now, let's examine the results from the real Dogecoin data. On both datasets at each threshold, the standard Hawkes model outperformed the self-limiting Hawkes model. Though in dataset 1 at the $0.3\%$ threshold, where the self-limiting model performed the worst relative to the standard model, the AUROC value for the self-limiting model was only $0.0072$ or $1.13\%$ lower than the AUROC value for the standard model. So, even though according

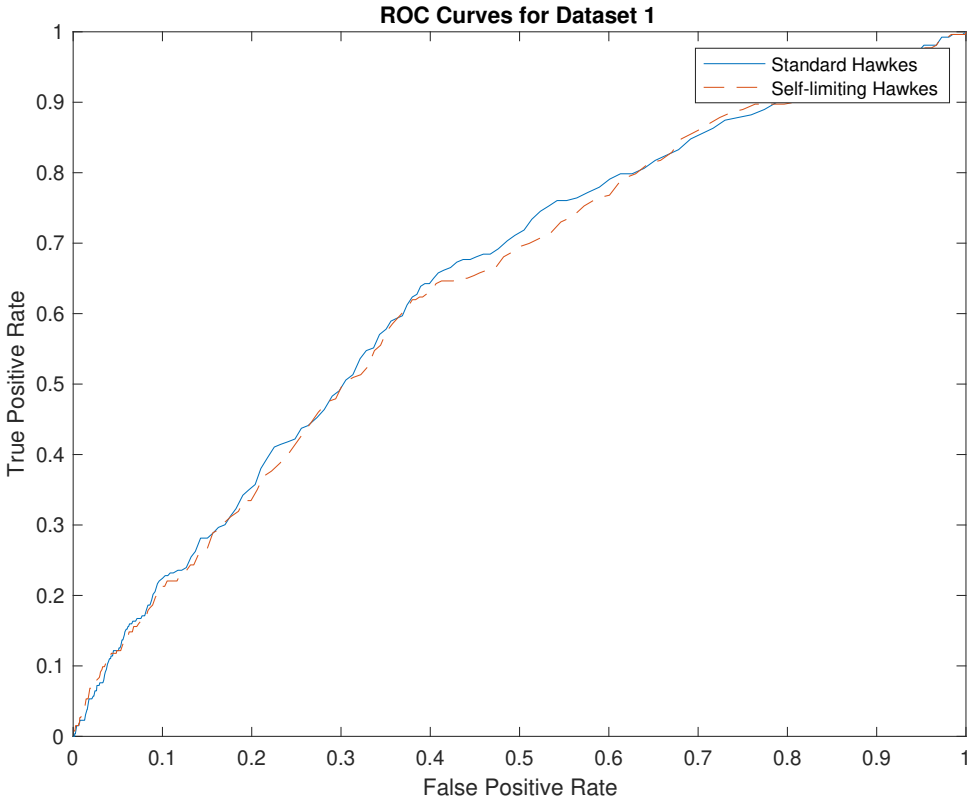


Figure 6.4: The ROC curves for dataset 1 at the $0.3\%$ threshold using both the standard and self-limiting Hawkes models.

to this analysis, the standard Hawkes model is a better fit to the Dogecoin data at each threshold, it is not much more likely than the self-limiting model.

Note that as the threshold increases, so does the AUROC value. This could be due to the fact that small price jumps in the Dogecoin data are more governed by random fluctuations in the market than the large price jumps. As the threshold increases, we are excluding more and more of these small price jumps leaving us with jumps that are significant enough to be governed by something other than random fluctuations in the market.

All analyses considered, we think that both the standard Hawkes model and the self-limiting Hawkes models are both plausible models for the Dogecoin datasets.

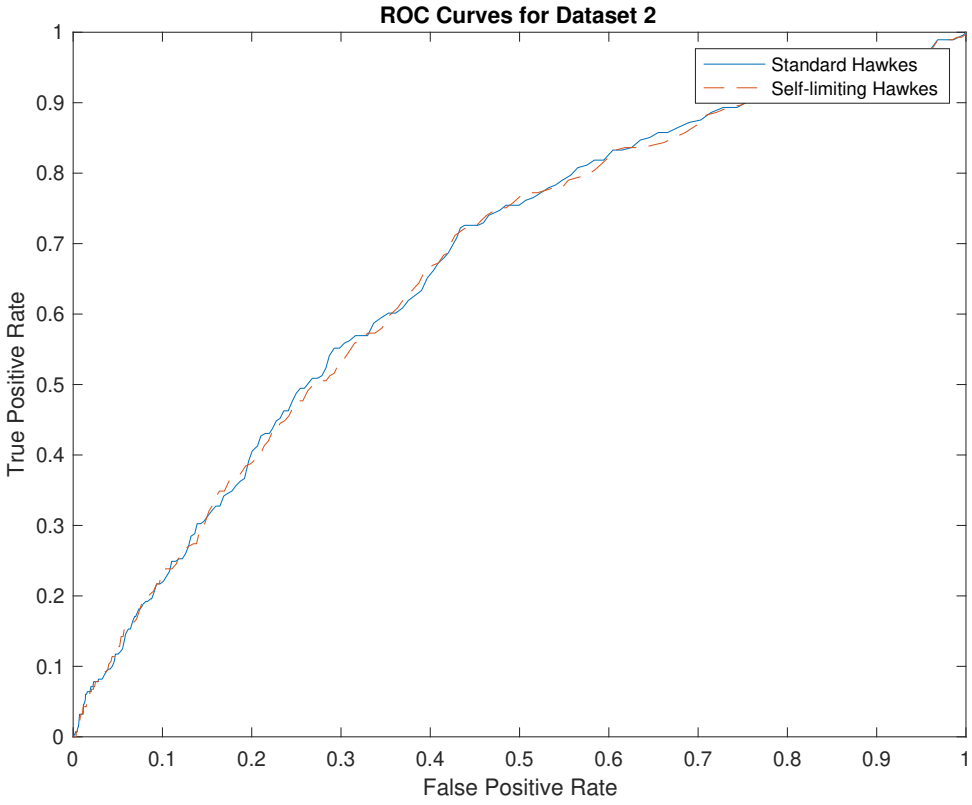


Figure 6.5: The ROC curves for dataset 2 at the $0.3\%$ threshold using both the standard and self-limiting Hawkes models.

Table 6.8: The area under the receiver operating characteristic curve (AUROC) for each dataset. Datasets where the self-limiting model outperformed the standard model are written in green. Datasets written in red designate the opposite. The threshold value is given in parentheses next to the dataset number.

| Dataset | AUROC (Standard) | AUROC (S-L) |
|---|---|---|
| $1(0.2\%)$ | 0.5746 | 0.5698 |
| $2(0.2\%)$ | 0.6038 | 0.6023 |
| $1(0.3\%)$ | 0.6432 | 0.6360 |
| $2(0.3\%)$ | 0.6666 | 0.6645 |
| $1(0.4\%)$ | 0.6903 | 0.6853 |
| $2(0.4\%)$ | 0.6803 | 0.6782 |

# CHAPTER 7
# CONCLUSION AND FUTURE WORK

In this thesis, we introduced a self-limiting Hawkes process, a variant of the Hawkes process where the self-exciting component is counteracted by a self-limiting component. We also introduced a self-limiting spatio-temporal Hawkes process, which is analogous to a self-limiting Hawkes process except that it contains a spatial component in addition to the temporal component.

In the context of modelling crime data, the self-exiting component represents the likelihood that crime at a point in time will likely lead to more crime in the near future. The self-limiting component represents the efforts of a police force in preventing crime events from happening. In the context of modelling financial data, the self-exciting component represents investors who buy into a security as its price rises. The self-limiting component represents investors who either sell or short a security believing it to be overvalued. More generally, we can use a self-limiting Hawkes process to model any system where we can reasonably identify likely self-exciting and self-limiting tendencies as we have done above.

We provide methods for simulating the self-limiting Hawkes and self-limiting spatio-temporal Hawkes processes, as well as methods for estimating the parameters of the self-limiting Hawkes process and the self-limiting spatio-temporal Hawkes process given a dataset of event times. Using maximum likelihood estimation, it has been shown that the parameters of a standard Hawkes process can be estimated with high accuracy [16]. Using a variation of this method that takes into account the preventative action of the self-limiting Hawkes process, we show that one can still estimate the parameters of the underlying Hawkes process with high accuracy. This is true for both the temporal and spatio-temporal models.

Further, using real crime data, we were able to show that the self-limiting Hawkes

process is a plausible alternative to the standard Hawkes process, though neither of the two processes were very likely fits to the data. Using real financial data, the self-limiting Hawkes process is a plausible alternative to the standard Hawkes process, though, by some measures, both the standard Hawkes and self-limiting Hawkes models were better fits for the financial data than they were for the crime data.

Due to computational constraints, we were only able to demonstrate the the performance of our self-limiting spatio-temporal Hawkes methods on hypothetical data, not on the real crime data. Future work in this area could involve making the spatio-temporal algorithms more efficient so that they could be applied to real-world data. One way this could be done is by relaxing some of the constraints on the algorithm. For example, at a particular location, we might be able to ignore the effects of events that are sufficiently far away since they are unlikely to generate any daughter events near the current location. As new stochastic processes are developed and analyzed that better model real-world data, we expect that the demand for fast, efficient estimation algorithms will greatly increase.

Another avenue of inquiry would be testing self-limiting Hawkes models with excited kernels $g$ that are not decaying exponentials, which would also likely enhance the ability of the model to fit real-world data.

# Appendices

# APPENDIX A

# CH. 2 CALCULATIONS

## A.1 Derivation of Equation 2.4

Recall that the complete data log-likelihood for the Hawkes process was first given by

$$\mathbb{E}[\mathscr{L}] = \sum_i P_{ii} \ln(\mu) - \int_0^T \mu dt + \sum_{j<i} P_{ij} \ln\left(k\omega e^{-\omega(t_i - t_j)}\right) - \int_0^T \sum_{i:t_i<t} k\omega e^{-\omega(t-t_i)} dt.$$

The first and third term can be simplified using the properties of logarithms. Since the second term is the integral of a constant, $\mu$, over the interval $[0, T]$, this is just $\mu T$.

Now, let's take a look at the fourth term. Since the excited kernel $g(t - t_i)$ doesn't take effect until time $t_i$, we can switch the integral and sum if we adjust the limits on each:

$$\begin{aligned}
\int_0^T \sum_{i:t_i<t} k\omega e^{-\omega(t-t_i)} dt &= \sum_{i=1}^n \int_{t_i}^T k\omega e^{-\omega(t-t_i)} dt \\
&= \sum_{i=1}^n \left[-k e^{-\omega(t-t_i)}\right]_{t_i}^T \\
&= \sum_{i=1}^n \left(-k e^{-\omega(T-t_i)} - k\right) \\
&= k \sum_{i=1}^n \left(1 - e^{-\omega(T-t_i)}\right).
\end{aligned}$$

Putting this all together gives us Equation 2.4:

$$\mathbb{E}[\mathscr{L}] = \ln(\mu) \sum_i P_{ii} + \ln(k\omega) \sum_{j<i} P_{ij} - \omega \sum_{j<i} P_{i,j}(t_i - t_j) - \mu T$$
$$- k \sum_i \left(1 - e^{-\omega(T-t_i)}\right).$$

## A.2 Derivation of Equation 2.10

Recall that the complete data log-likelihood for the spatio-temporal Hawkes process was first given by

$$\mathbb{E}[\mathscr{L}] = \ln(\mu) \sum_i P_{ii} + \ln\left(\frac{k\omega}{4s^2}\right) \sum_{j<i} P_{ij} - \omega \sum_{j<i} P_{ij}(t_i - t_j)$$
$$- \frac{1}{s} \sum_{j<i} P_{ij}(|x_i - x_j| + |y_i - y_j|) - \int_0^T \int_0^L \int_0^L \lambda(t, x, y) dx dy dt.$$

We can find a closed-form expression of the integral term as follows:

$$
\begin{aligned}
&\int_0^T \int_0^L \int_0^L \lambda(t,x,y)dxdydt \\
&= \int_0^T \int_0^L \int_0^L \left(\mu + \frac{k\omega}{4s^2} \sum_{i:t_i<t} e^{-\omega(t-t_i)} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}}\right) dxdydt \\
&= \mu T L^2 + \frac{k\omega}{4s^2} \sum_{i=1}^{n} \int_{t_i}^T \int_0^L \int_0^L e^{-\omega(t-t_i)} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}} dxdydt \\
&= \mu T L^2 + \frac{k\omega}{4s^2} \sum_{i=1}^{n} \int_{t_i}^T e^{-\omega(t-t_i)} \int_0^L e^{\frac{-|y-y_i|}{s}} \int_0^L e^{\frac{-|x-x_i|}{s}} dxdydt \\
&= \mu T L^2 + \frac{k\omega}{4s^2} \sum_{i=1}^{n} \left[2s - s\left(e^{\frac{-x_i}{s}} + e^{\frac{-(L-x_i)}{s}}\right)\right] \int_{t_i}^T e^{-\omega(t-t_i)} \int_0^L e^{\frac{-|y-y_i|}{s}} dydt \\
&= \mu T L^2 \\
&\quad + \frac{k\omega}{4s^2} \sum_{i=1}^{n} \left[2s - s\left(e^{\frac{-x_i}{s}} + e^{\frac{-(L-x_i)}{s}}\right)\right] \left[2s - s\left(e^{\frac{-y_i}{s}} + e^{\frac{-(L-y_i)}{s}}\right)\right] \int_{t_i}^T e^{-\omega(t-t_i)} dt \\
&= \mu T L^2 - \frac{k}{4s^2} \sum_{i=1}^{n} \left[2s - s\left(e^{\frac{-x_i}{s}} + e^{\frac{-(L-x_i)}{s}}\right)\right] \left[2s - s\left(e^{\frac{-y_i}{s}} + e^{\frac{-(L-y_i)}{s}}\right)\right] \left[e^{-\omega(t-t_i)}\right]_{t_i}^T \\
&= \mu T L^2 - \frac{k}{4} \sum_{i=1}^{n} \left(2 - e^{\frac{-x_i}{s}} - e^{\frac{-(L-x_i)}{s}}\right) \left(2 - e^{\frac{-y_i}{s}} - e^{\frac{-(L-y_i)}{s}}\right) \left(e^{-\omega(T-t_i)} - 1\right).
\end{aligned}
$$

This gives us Equation 2.10:

$$
\begin{aligned}
\mathbb{E}[\mathscr{L}] = \ln(\mu) \sum_i P_{ii} + \ln\left(\frac{k\omega}{4s^2}\right) \sum_{j<i} P_{ij} - \omega \sum_{j<i} P_{ij}(t_i - t_j) \\
- \frac{1}{s} \sum_{j<i} P_{ij}(|x_i - x_j| + |y_i - y_j|) - \mu T L^2 \\
+ \frac{k}{4} \sum_i \left(2 - e^{\frac{-x_i}{s}} - e^{\frac{-(L-x_i)}{s}}\right) \left(2 - e^{\frac{-y_i}{s}} - e^{\frac{-(L-y_i)}{s}}\right) \left(e^{-\omega(T-t_i)} - 1\right).
\end{aligned}
$$

# APPENDIX B

# CH. 3 CALCULATIONS

## B.1 Derivation of Equation 3.6

Recall that the complete data log-likelihood for the self-limiting Hawkes process was first given by

$$\begin{aligned}\mathbb{E}[\mathscr{L}] = \ln(\mu)\sum_i P_{ii} - \beta\sum_i P_{ii}N(\alpha,t_i) + \ln(k\omega)\sum_{j<i} P_{ij} - \omega\sum_{j<i} P_{ij}(t_i - t_j) \\ - \beta\sum_{j<i} P_{ij}N(\alpha,t_i) - \mu\int_0^T e^{-\beta N(\alpha,t)}dt - k\omega\int_0^T e^{-\beta N(\alpha,t)}\sum_{t_i<t} e^{-\omega(t-t_i)}dt.\end{aligned}$$

Replacing $N(\alpha,t)$ with the values given in Equation 3.5, we have

$$\begin{aligned}\mathbb{E}[\mathscr{L}] = \ln(\mu)\sum_i P_{ii} - \beta\sum_i P_{ii}n_{s(t_i)} + \ln(k\omega)\sum_{j<i} P_{ij} - \omega\sum_{j<i} P_{ij}(t_i - t_j) \\ - \beta\sum_{j<i} P_{ij}n_{s(t_i)} - \mu\sum_{i=1}^{l}\int_{\tau_{i-1}}^{\tau_i} e^{-\beta n_i}dt - k\omega\sum_{j=1}^{l}\int_{\tau_{j-1}}^{\tau_j} e^{-\beta n_j}\sum_{t_i<t} e^{-\omega(t-t_i)}dt,\end{aligned}$$

where $s(t_i)$ is the index of $t_i$ in $\{\tau_0,\ldots,\tau_l\}$. Since the intervals $[\tau_0,\tau_1],\ldots,[\tau_{l-1},\tau_l]$ don't overlap and their union is $[0,T]$, each of the integrals over $[0,T]$ can be rewritten as the sum of integrals over each of the above intervals. It is more convenient to do this since $N(\alpha,t)$ is a constant in each of these intervals.

We will handle each of the integral terms separately:

$$\mu \sum_{i=1}^{l} \int_{\tau_{i-1}}^{\tau_i} e^{-\beta n_i} dt = \mu \sum_{i=1}^{l} e^{-\beta n_i} \int_{\tau_{i-1}}^{\tau_i} dt$$
$$= \mu \sum_{i=1}^{l} e^{-\beta n_i} (\tau_i - \tau_{i-1}) dt$$

and

$$k\omega \sum_{j=1}^{l} \int_{\tau_{j-1}}^{\tau_j} e^{-\beta n_j} \sum_{t_i<t} e^{-\omega(t-t_i)} dt = k\omega \sum_{j=1}^{l} e^{-\beta n_j} \int_{\tau_{j-1}}^{\tau_j} \sum_{t_i<t} e^{-\omega(t-t_i)} dt$$
$$= k\omega \sum_{j=1}^{l} e^{-\beta n_j} \int_{\tau_{j-1}}^{\tau_j} \sum_{i=1}^{n} e^{-\omega(t-t_i)} \mathbb{1}_{\{t_i<t\}} dt$$
$$= k\omega \sum_{i=1}^{n} \sum_{j=1}^{l} e^{-\beta n_j} \int_{\tau_{j-1}}^{\tau_j} e^{-\omega(t-t_i)} \mathbb{1}_{\{t_i<t\}} dt$$
$$= k\omega \sum_{i=1}^{n} \sum_{j=1}^{l} e^{-\beta n_j} \int_{\tau_{j-1}}^{\tau_j} e^{-\omega(t-t_i)} \mathbb{1}_{\{t_i<\tau_j\}} dt$$
$$= k\omega \sum_{i=1}^{n} \sum_{j=1}^{l} e^{-\beta n_j} \mathbb{1}_{\{t_i<\tau_j\}} \int_{\tau_{j-1}}^{\tau_j} e^{-\omega(t-t_i)} dt$$
$$= k\omega \sum_{i=1}^{n} \sum_{j=1}^{l} e^{-\beta n_j} \mathbb{1}_{\{t_i<\tau_j\}} \frac{-1}{\omega} \left[ e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1} - t_i)} \right]$$
$$= -k \sum_{i=1}^{n} \sum_{j=1}^{l} e^{-\beta n_j} \left[ e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1} - t_i)} \right] \mathbb{1}_{\{t_i<\tau_j\}}.$$

Notice that between the third and fourth lines, $\mathbb{1}_{\{t_i<t\}}$ was replaced with $\mathbb{1}_{\{t_i<\tau_j\}}$. This follows from the definition of the $\tau_i$'s. Recall that the set $\{\tau_i\}$ is constructed by taking the union of the two sets $\{t_i\}$ and $\{t_i + \alpha\}$, sorting it, and removing any entries with values greater than $T$.

In the third line, we are integrating $e^{-\omega(t-t_i)} \mathbb{1}_{\{t_i<t\}}$ over the interval $[\tau_{j-1}, \tau_j]$. So, we have three cases: $t_i \leq \tau_{j-1}$, $t_i \in (\tau_{j-1}, \tau_j)$, or $t_i \geq \tau_j$. Since $t_i \in \{\tau_0, \ldots, \tau_l\}$, $t_i \notin (\tau_{j-1}, \tau_j)$. So, there are really only two cases: $t_i \leq \tau_{j-1}$ or $t_i \geq \tau_j$. If $t_i \leq \tau_{j-1}$,

then $e^{-\omega(t-t_i)}\mathbb{1}_{\{t_i<t\}}$ just becomes $e^{-\omega(t-t_i)}$ on the interval $[\tau_{j-1}, \tau_j]$. If $t_i \geq \tau_j$, then $e^{-\omega(t-t_i)}\mathbb{1}_{\{t_i<t\}}$ just becomes 0 on the interval $[\tau_{j-1}, \tau_j]$. So $\mathbb{1}_{\{t_i<t\}}$ can be replaced with either $\mathbb{1}_{\{t_i\leq\tau_{j-1}\}}$ or $\mathbb{1}_{\{t_i<\tau_j\}}$.

Putting this all together gives us Equation 3.6:

$$\begin{aligned}\mathbb{E}[\mathscr{L}] = \ln(\mu)\sum_i P_{ii} - \beta\sum_i P_{ii}n_{s(t_i)} + \ln(k)\sum_{j<i} P_{ij} + \ln(\omega)\sum_{j<i} P_{ij} \\ - \omega\sum_{j<i} P_{ij}(t_i - t_j) - \beta\sum_{j<i} P_{ij}n_{s(t_i)} - \mu\sum_{i=1}^{l} e^{-\beta n_i}(\tau_i - \tau_{i-1}) \\ + k\sum_{i=1}^{n}\sum_{j=1}^{l} e^{-\beta n_j}\left[e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1}-t_i)}\right]\mathbb{1}_E,\end{aligned}$$

where $E = \{t_i < \tau_j\}$.

## B.2 Derivation of Equation 3.11

Plugging the intensity function for a spatio-temporal Hawkes process into Equation 2.9 and removing all terms that don't include the parameters $\mu$, $k$, $\omega$, or $s$ gives

$$\begin{aligned}\mathscr{L} = -\int_0^T\int_0^L\int_0^L\left(\mu + \frac{k\omega}{4s^2}\sum_{i:t_i<t} e^{-\omega(t-t_i)}e^{\frac{-(|x-x_i|+|y-y_i|)}{s}}\right)q(t,x,y,\alpha,\beta)dxdydt \\ + \sum_i \ln\left(\left(\mu + \frac{k\omega}{4s^2}\sum_{j:t_j<t_i} e^{-\omega(t_i-t_j)}e^{\frac{-(|x_i-x_j|+|y_i-y_j|)}{s}}\right)q(t_i,x_i,y_i,\alpha,\beta)\right),\end{aligned} \tag{B.1}$$

where

$$q(t,x,y,\alpha,\beta) = \begin{cases} e^{\frac{-\beta N(\alpha,t)}{|\text{box}(t)|}} & \text{if } (x,y)\in\text{box}(t) \\ 1 & \text{else}\end{cases}$$

and box$(t)$ is a box centered on the mean location of the events in the interval $[t-\alpha, t)$. Its width is twice the standard deviation in the $x$ component of the locations of the events in the interval $[t-\alpha, t)$. Its height is twice the standard deviation in the $y$ component of the locations of the events in the interval $[t-\alpha, t)$.

We will begin by finding a simplified expression for the integral term in Equation B.1. Define

$$I_1 = \int_0^T \int_0^L \int_0^L \mu q(t, x, y, \alpha, \beta) dx dy dt$$

and

$$I_2 = \int_0^T \int_0^L \int_0^L \frac{k\omega}{4s^2} q(t, x, y, \alpha, \beta) \sum_{i:t_i<t} e^{-\omega(t-t_i)} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}} dx dy dt.$$

Then

$$\int_0^T \int_0^L \int_0^L \lambda(t, x, y) dx dy dt = I_1 + I_2.$$

Now

$$
\begin{aligned}
I_1 &= \mu \int_0^T \int_0^L \int_0^L q(t,x,y,\alpha,\beta)dxdydt \\
&= \mu \left[ \int_0^T \iint_{\text{box}(t)} e^{\frac{-\beta N(\alpha,t)}{|\text{box}(t)|}} dxdydt + \int_0^T \iint_{\text{box}(t)^C} dxdydt \right] \\
&= \mu \left[ \sum_{j=1}^{l} \int_{\tau_{j-1}}^{\tau_j} \iint_{\text{box}(t)} e^{\frac{-\beta n_j}{b_j}} dxdydt + \int_0^T (L^2 - |\text{box}(t)|)dt \right] \\
&= \mu \sum_{j=1}^{l} \left( e^{\frac{-\beta n_j}{b_j}} \int_{\tau_{j-1}}^{\tau_j} b_j dt + \int_{\tau_{j-1}}^{\tau_j} (L^2 - b_j)dt \right) \\
&= \mu \sum_{j=1}^{l} \left( e^{\frac{-\beta n_j}{b_j}} \cdot b_j(\tau_j - \tau_{j-1}) + (L^2 - b_j)(\tau_j - \tau_{j-1}) \right) \\
&= \mu \left[ TL^2 + \sum_{j=1}^{l} b_j(\tau_j - \tau_{j-1}) \left( e^{\frac{-\beta n_j}{b_j}} - 1 \right) \right].
\end{aligned}
$$

and

$$
\begin{aligned}
I_2 &= \int_0^T \int_0^L \int_0^L \frac{k\omega}{4s^2} q(t,x,y,\alpha,\beta) \sum_{i:t_i<t} e^{-\omega(t-t_i)} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}} dxdydt \\
&= \frac{k\omega}{4s^2} \left[ \int_0^T \iint_{\text{box}(t)} e^{\frac{-\beta N(\alpha,t)}{|\text{box}(t)|}} \sum_{i:t_i<t} e^{-\omega(t-t_i)} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}} dxdydt \right. \\
&+ \left. \int_0^T \iint_{\text{box}(t)^C} \sum_{i:t_i<t} e^{-\omega(t-t_i)} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}} dxdydt \right] \\
&= \frac{k\omega}{4s^2} \left[ \sum_{j=1}^{l} \sum_{i=1}^{n} \int_{\tau_{j-1}}^{\tau_j} \iint_{\text{box}(t)} e^{\frac{-\beta n_j}{b_j}} e^{-\omega(t-t_i)} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}} \mathbb{1}_{\{t_i<\tau_j\}} dxdydt \right. \\
&+ \left. \sum_{j=1}^{l} \sum_{i=1}^{n} \int_{\tau_{j-1}}^{\tau_j} \iint_{\text{box}(t)^C} e^{-\omega(t-t_i)} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}} \mathbb{1}_{\{t_i<\tau_j\}} dxdydt \right] \\
&= \frac{k\omega}{4s^2} \sum_{j=1}^{l} \sum_{i=1}^{n} \left[ e^{\frac{-\beta n_j}{b_j}} \int_{\tau_{j-1}}^{\tau_j} e^{-\omega(t-t_i)} S_i(t) \mathbb{1}_{\{t_i<\tau_j\}} dt + \int_{\tau_{j-1}}^{\tau_j} e^{-\omega(t-t_i)} S_i'(t) \mathbb{1}_{\{t_i<\tau_j\}} dt \right],
\end{aligned}
$$

where

$$S_i(t) = \iint_{\text{box}(t)} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}} dxdy$$

and

$$S_i'(t) = \iint_{\text{box}(t)^C} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}} dxdy.$$

Since $S_i(t)$ and $S_i'(t)$ are a constants in the interval $[\tau_{j-1}, \tau_j)$, we will call these constants $S_{ij}$ and $S_{ij}'$, respectively. Then we have

$$\begin{aligned}
I_2 &= \frac{k\omega}{4s^2} \sum_{j=1}^{l} \sum_{i=1}^{n} \left[ e^{\frac{-\beta n_j}{b_j}} S_{ij} \int_{\tau_{j-1}}^{\tau_j} e^{-\omega(t-t_i)} \mathbb{1}_{\{t_i<\tau_j\}} dt + S_{ij}' \int_{\tau_{j-1}}^{\tau_j} e^{-\omega(t-t_i)} \mathbb{1}_{\{t_i<\tau_j\}} dt \right] \\
&= \frac{k\omega}{4s^2} \sum_{j=1}^{l} \sum_{i=1}^{n} \left[ \left( e^{\frac{-\beta n_j}{b_j}} S_{ij} + S_{ij}' \right) \int_{\tau_{j-1}}^{\tau_j} e^{-\omega(t-t_i)} \mathbb{1}_{\{t_i<\tau_j\}} dt \right] \\
&= \frac{-k}{4s^2} \sum_{j=1}^{l} \sum_{i=1}^{n} \left[ \left( e^{\frac{-\beta n_j}{b_j}} S_{ij} + S_{ij}' \right) \left( e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1} - t_i)} \right) \mathbb{1}_{\{t_i<\tau_j\}} \right]. \\
&= \frac{-k}{4} \sum_{j=1}^{l} \sum_{i=1}^{n} \left[ \left( e^{\frac{-\beta n_j}{b_j}} \frac{S_{ij}}{s^2} + \frac{S_{ij}'}{s^2} \right) \left( e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1} - t_i)} \right) \mathbb{1}_{\{t_i<\tau_j\}} \right].
\end{aligned}$$

In the above calculations, we use several constants which have not been defined yet. $\{\tau_0, \ldots, \tau_l\}$ are the times when $N(\alpha, t)$ changes. Therefore, $N(\alpha, t)$ and $|\text{box}(t)|$ are both constants in the interval $[\tau_{j-1}, \tau_j]$, and so we call these constants $n_j$ and $b_j$, respectively.

To find expressions for $S_{ij}$ and $S_{ij}'$, we first need to define the boundaries of the box in which preventative action takes place. $\text{box}(t) = [c_x(t) - \sigma_x(t), c_x(t) + \sigma_x(t)] \times [c_y(t) - \sigma_y(t), c_y(t) + \sigma_y(t)]$, where $c_x(t)$, $c_y(t)$, $\sigma_x(t)$, and $\sigma_y(t)$ are the means and standard deviations described above. Since $c_x(t)$, $c_y(t)$, $\sigma_x(t)$, and $\sigma_y(t)$ are all constants in the interval $[\tau_{j-1}, \tau_j]$, we will define $\text{box}(t) = [\text{left}(t), \text{right}(t)] \times [\text{bottom}(t), \text{top}(t)] = [\text{left}_j, \text{right}_j] \times [\text{bottom}_j, \text{top}_j]$ in that interval. Then

$$
\begin{aligned}
S_{ij} &= \iint_{\text{box}_j} e^{\frac{-(|x-x_i|+|y-y_i|)}{s}} dxdy \\
&= \iint_{\text{box}_j} e^{\frac{-|x-x_i|}{s}} e^{\frac{-|y-y_i|}{s}} dxdy \\
&= \int_{\text{bottom}_j}^{\text{top}_j} e^{\frac{-|y-y_i|}{s}} \int_{\text{left}_j}^{\text{right}_j} e^{\frac{-|x-x_i|}{s}} dxdy \\
&= \int_{\text{left}_j}^{\text{right}_j} e^{\frac{-|x-x_i|}{s}} dx \int_{\text{bottom}_j}^{\text{top}_j} e^{\frac{-|y-y_i|}{s}} dy.
\end{aligned}
$$

If $x_i \leq \text{left}_j$, then

$$
\int_{\text{left}_j}^{\text{right}_j} e^{\frac{-|x-x_i|}{s}} dx = \int_{\text{left}_j}^{\text{right}_j} e^{\frac{x_i-x}{s}} dx = -se^{\frac{x_i-\text{right}_j}{s}} + se^{\frac{x_i-\text{left}_j}{s}}.
$$

If $\text{left}_j \leq x_i \leq \text{right}_j$, then

$$
\int_{\text{left}_j}^{\text{right}_j} e^{\frac{-|x-x_i|}{s}} dx = \int_{\text{left}_j}^{x_i} e^{\frac{x-x_i}{s}} dx + \int_{x_i}^{\text{right}_j} e^{\frac{x_i-x}{s}} dx = 2s - se^{\frac{\text{left}_j-x_i}{s}} - se^{\frac{x_i-\text{right}_j}{s}}.
$$

If $x_i \geq \text{right}_j$, then

$$
\int_{\text{left}_j}^{\text{right}_j} e^{\frac{-|x-x_i|}{s}} dx = \int_{\text{left}_j}^{\text{right}_j} e^{\frac{x-x_i}{s}} dx = se^{\frac{\text{right}_j-x_i}{s}} - se^{\frac{\text{left}_j-x_i}{s}}.
$$

The $y$ integrals will be very similar. Then wherever $(x_i, y_i)$ is with relation to the self-limiting box, we can combine these integrals to get

$$\frac{S_{ij}}{s^2} = \begin{cases} \left(e^{\frac{x_i - \text{left}_j}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(e^{\frac{y_i - \text{bottom}_j}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 1} \\ \left(2 - e^{\frac{\text{left}_j - x_i}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(e^{\frac{y_i - \text{bottom}_j}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 2} \\ \left(e^{\frac{\text{right}_j - x_i}{s}} - e^{\frac{\text{left}_j - x_i}{s}}\right)\left(e^{\frac{y_i - \text{bottom}_j}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 3} \\ \left(e^{\frac{x_i - \text{left}_j}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(2 - e^{\frac{\text{bottom}_j - y_i}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 4} \\ \left(2 - e^{\frac{\text{left}_j - x_i}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(2 - e^{\frac{\text{bottom}_j - y_i}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 5} \\ \left(e^{\frac{\text{right}_j - x_i}{s}} - e^{\frac{\text{left}_j - x_i}{s}}\right)\left(2 - e^{\frac{\text{bottom}_j - y_i}{s}} - e^{\frac{y_i - \text{top}_j}{s}}\right) & \text{for Case 6} \\ \left(e^{\frac{x_i - \text{left}_j}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(e^{\frac{\text{top}_j - y_i}{s}} - e^{\frac{\text{bottom}_j - y_i}{s}}\right) & \text{for Case 7} \\ \left(2 - e^{\frac{\text{left}_j - x_i}{s}} - e^{\frac{x_i - \text{right}_j}{s}}\right)\left(e^{\frac{\text{top}_j - y_i}{s}} - e^{\frac{\text{bottom}_j - y_i}{s}}\right) & \text{for Case 8} \\ \left(e^{\frac{\text{right}_j - x_i}{s}} - e^{\frac{\text{left}_j - x_i}{s}}\right)\left(e^{\frac{\text{top}_j - y_i}{s}} - e^{\frac{\text{bottom}_j - y_i}{s}}\right) & \text{for Case 9,} \end{cases}$$

where

$$\text{Case 1} = \{(x_i, y_i) : x_i \leq \text{left}_j, y_i \leq \text{bottom}_j\}$$

$$\text{Case 2} = \{(x_i, y_i) : \text{left}_j < x_i < \text{right}_j, y_i \leq \text{bottom}_j\}$$

$$\text{Case 3} = \{(x_i, y_i) : x_i \geq \text{right}_j, y_i \leq \text{bottom}_j\}$$

$$\text{Case 4} = \{(x_i, y_i) : x_i \leq \text{left}_j, \text{bottom}_j < y_i < \text{top}_j\}$$

$$\text{Case 5} = \{(x_i, y_i) : \text{left}_j < x_i < \text{right}_j, \text{bottom}_j < y_i < \text{top}_j\}$$

$$\text{Case 6} = \{(x_i, y_i) : x_i \geq \text{right}_j, \text{bottom}_j < y_i < \text{top}_j\}$$

$$\text{Case 7} = \{(x_i, y_i) : x_i \leq \text{left}_j, y_i \geq \text{top}_j\}$$

$$\text{Case 8} = \{(x_i, y_i) : \text{left}_j < x_i < \text{right}_j, y_i \geq \text{top}_j\}$$

$$\text{Case 9} = \{(x_i, y_i) : x_i \geq \text{right}_j, y_i \geq \text{top}_j\}.$$

Since $S'_{ij}$ are the spatial integrals outside of the self-limiting box, this is the same as the

spatial integral over all of $[0, L] \times [0, L]$ minus the part that is inside the self-limiting box. Thus

$$\begin{aligned} S'_{ij} &= \int_0^L \int_0^L e^{\frac{-(|x-x_i|+|y-y_i|)}{s}} dxdy - S_{ij} \\ &= s^2 \left(2 - e^{\frac{-x_i}{s}} - e^{\frac{-(L-x_i)}{s}}\right)\left(2 - e^{\frac{-y_i}{s}} - e^{\frac{-(L-y_i)}{s}}\right) - S_{ij}. \end{aligned}$$

So,

$$\frac{S'_{ij}}{s^2} = \left(2 - e^{\frac{-x_i}{s}} - e^{\frac{-(L-x_i)}{s}}\right)\left(2 - e^{\frac{-y_i}{s}} - e^{\frac{-(L-y_i)}{s}}\right) - \frac{S_{ij}}{s^2}.$$

Now we will turn to the summation term in Equation B.1.

$$\begin{aligned} &\sum_i \ln\left(\left(\mu + \frac{k\omega}{4s^2}\sum_{j:t_j<t_i} e^{-\omega(t_i-t_j)} e^{\frac{-(|x_i-x_j|+|y_i-y_j|)}{s}}\right) q(t_i, x_i, y_i, \alpha, \beta)\right) \\ &= \sum_i \left[\ln\left(\mu + \frac{k\omega}{4s^2}\sum_{j:t_j<t_i} e^{-\omega(t_i-t_j)} e^{\frac{-(|x_i-x_j|+|y_i-y_j|)}{s}}\right) + \ln\left(q(t_i, x_i, y_i, \alpha, \beta)\right)\right] \qquad \text{(B.2)} \\ &= \sum_i \ln\left(\mu + \frac{k\omega}{4s^2}\sum_{j:t_j<t_i} e^{-\omega(t_i-t_j)} e^{\frac{-(|x_i-x_j|+|y_i-y_j|)}{s}}\right) + \sum_i \ln\left(q(t_i, x_i, y_i, \alpha, \beta)\right). \end{aligned}$$

Assume that we knew which events were background events and which were daughter events. Let $B$ be the set of indices of the background events and $D$ the set of indices of the daughter events. Since background events are attributable entirely to the $\mu$ part of the intensity function and daughter events are attributable entirely to the self-limiting part of the intensity function, Equation B.2 can be rewritten as

$$\sum_{i\in B}\ln(\mu)+\sum_{i\in D}\ln\left(\frac{k\omega}{4s^2}\sum_{j:t_j<t_i}e^{-\omega(t_i-t_j)}e^{\frac{-(|x_i-x_j|+|y_i-y_j|)}{s}}\right)+\sum_{i\in B\cup D}\ln\left(q(t_i,x_i,y_i,\alpha,\beta)\right)$$

$$=\sum_{i\in B}\ln(\mu)+\sum_{i\in D}\ln\left(\frac{k\omega}{4s^2}\right)+\sum_{i\in D}\ln\left(\sum_{j:t_j<t_i}e^{-\omega(t_i-t_j)}e^{\frac{-(|x_i-x_j|+|y_i-y_j|)}{s}}\right)$$

$$+\sum_{i\in B\cup D}\ln\left(q(t_i,x_i,y_i,\alpha,\beta)\right)$$

Since, in practice, we don't actually know what $B$ and $D$ are, we can instead take the expectation of Equation B.1 with respect to the probabilistic branching structure $P$, where

$$P_{ij}=\begin{cases}\text{prob. that } i \text{ is a background event} & ,i=j\\ \text{prob. that } i \text{ is a daughter of } j & ,j<i\end{cases}.$$

More specifically,

$$P_{ij}=\begin{cases}\frac{\mu q(t_i,x_i,y_i,\alpha,\beta)}{\lambda(t_i)} & i=j\\ \frac{\frac{k\omega}{4s^2}e^{-\omega(t_i-t_j)}e^{\frac{-(|x_i-x_j|+|y_i-y_j|)}{s}}q(t_i,x_i,y_i,\alpha,\beta)}{\lambda(t_i)} & j<i\end{cases}.$$

After taking the expectation, Equation B.2 becomes

$$
\begin{aligned}
&\sum_i P_{ii} \ln(\mu) + \sum_{j<i} P_{ij} \ln\left(\frac{k\omega}{4s^2}\right) + \sum_{j<i} P_{ij} \ln\left(e^{-\omega(t_i - t_j)} e^{\frac{-(|x_i - x_j| + |y_i - y_j|)}{s}}\right) \\
&+ \sum_i P_{ii} \ln\left(q(t_i, x_i, y_i, \alpha, \beta)\right) + \sum_{j<i} P_{ij} \ln\left(q(t_i, x_i, y_i, \alpha, \beta)\right) \\
&= \ln(\mu) \sum_i P_{ii} + \ln\left(\frac{k\omega}{4s^2}\right) \sum_{j<i} P_{ij} + \sum_{j<i} P_{ij} \ln\left(e^{-\omega(t_i - t_j)}\right) \\
&+ \sum_{j<i} P_{ij} \ln\left(e^{\frac{-(|x_i - x_j| + |y_i - y_j|)}{s}}\right) + \sum_i P_{ii} \ln\left(q(t_i, x_i, y_i, \alpha, \beta)\right) \\
&+ \sum_{j<i} P_{ij} \ln\left(q(t_i, x_i, y_i, \alpha, \beta)\right) \\
&= \ln(\mu) \sum_i P_{ii} + \ln\left(\frac{k\omega}{4s^2}\right) \sum_{j<i} P_{ij} - \omega \sum_{j<i} P_{ij}(t_i - t_j) \\
&- \frac{1}{s} \sum_{j<i} P_{ij}(|x_i - x_j| + |y_i - y_j|) + \sum_i P_{ii} \ln\left(q(t_i, x_i, y_i, \alpha, \beta)\right) \\
&+ \sum_{j<i} P_{ij} \ln\left(q(t_i, x_i, y_i, \alpha, \beta)\right).
\end{aligned}
$$

Putting all this together with the integral term gives the complete data log-likelihood

$$
\begin{aligned}
\mathbb{E}[\mathscr{L}] = &\ln(\mu) \sum_i P_{ii} + \sum_i P_{ii} q(t_i, x_i, y_i, \alpha, \beta) + \ln\left(\frac{k\omega}{4s^2}\right) \sum_{j<i} P_{ij} - \omega \sum_{j<i} P_{ij}(t_i - t_j) \\
&- \frac{1}{s} \sum_{j<i} P_{ij}(|x_i - x_j| + |y_i - y_j|) + \sum_{j<i} P_{ij} q(t_i, x_i, y_i, \alpha, \beta) \\
&- \mu \left[ TL^2 + \sum_{j=1}^{l} b_j(\tau_j - \tau_{j-1}) \left( e^{\frac{-\beta n_j}{b_j}} - 1 \right) \right] \\
&+ \frac{k}{4} \sum_{j=1}^{l} \sum_{i=1}^{n} \left[ \left( e^{\frac{-\beta n_j}{b_j}} \frac{S_{ij}}{s^2} + \frac{S'_{ij}}{s^2} \right) \left( e^{-\omega(\tau_j - t_i)} - e^{-\omega(\tau_{j-1} - t_i)} \right) \mathbb{1}_{\{t_i < \tau_j\}} \right],
\end{aligned}
$$

which is exactly what appears in Equation 3.11.

## REFERENCES


[1] L. Adamopoulos, "Cluster models for earthquakes: Regional comparisons," *Journal of the International Association for Mathematical Geology*, vol. 8, no. 4, pp. 463–475, 1976.

[2] S. J. Hardiman, N. Bercot, and J.-P. Bouchaud, "Critical reflexivity in financial markets: A hawkes process analysis," *The European Physical Journal B*, vol. 86, no. 10, p. 442, 2013.

[3] V. Filimonov and D. Sornette, "Quantifying reflexivity in financial markets: Toward a prediction of flash crashes," *Physical Review E*, vol. 85, no. 5, p. 056 108, 2012.

[4] P. Hewlett, "Clustering of order arrivals, price impact and trade path optimisation," in *Workshop on Financial Modeling with Jump processes, Ecole Polytechnique*, 2006, pp. 6–8.

[5] M. Rambaldi, P. Pennesi, and F. Lillo, "Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach," *Physical Review E*, vol. 91, no. 1, p. 012 819, 2015.

[6] E. W. Fox, M. B. Short, F. P. Schoenberg, K. D. Coronges, and A. L. Bertozzi, "Modeling e-mail networks and inferring leadership using self-exciting point processes," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 564–584, 2016.

[7] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[8] S. D. Johnson, K. Bowers, and A. Hirschfield, "New insights into the spatial and temporal distribution of repeat victimization," *The British Journal of Criminology*, vol. 37, no. 2, pp. 224–241, 1997.

[9] S. D. Johnson and K. J. Bowers, "The burglary as clue to the future: The beginnings of prospective hot-spotting," *European Journal of Criminology*, vol. 1, no. 2, pp. 237–255, 2004.

[10] 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*, vol. 106, no. 493, pp. 100–108, 2011.

[11] G. O. Mohler, M. B. Short, S. Malinowski, M. Johnson, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham, "Randomized controlled field trials of predictive policing,"

*Journal of the American statistical association*, vol. 110, no. 512, pp. 1399–1411, 2015.

[12] V. Isham and M. Westcott, "A self-correcting point process," *Stochastic processes and their applications*, vol. 8, no. 3, pp. 335–347, 1979.

[13] P. W. Lewis and G. S. Shedler, "Simulation of nonhomogeneous poisson processes by thinning," *Naval research logistics quarterly*, vol. 26, no. 3, pp. 403–413, 1979.

[14] Y. Ogata, "On lewis' simulation method for point processes," *IEEE transactions on information theory*, vol. 27, no. 1, pp. 23–31, 1981.

[15] A. Dassios and H. Zhao, "Exact simulation of hawkes process with exponentially decaying intensity," *Electronic Communications in Probability*, vol. 18, no. 62, 2013.

[16] J. F. Olson and K. M. Carley, "Exact and approximate em estimation of mutually exciting hawkes processes," *Statistical Inference for Stochastic Processes*, vol. 16, no. 1, pp. 63–80, 2013.

[17] C. of Chicago. (2001). "Crimes - 2001 to present," (visited on 08/31/2020).

[18] S. D. Johnson, "Repeat burglary victimisation: A tale of two theories," *Journal of Experimental Criminology*, vol. 4, no. 3, pp. 215–240, 2008.

[19] F. Data. (2021). "Historical intraday dogecoin (doge) data," (visited on 01/12/2022).

# VITA

John Garnier Olinde Jr. was born in November of 1994 in New Orleans, LA. He spent most of his youth with his family in Melbourne, FL after moving there in 2000. He graduated high school in 2013 and attended Georgia Tech for his undergraduate studies. After graduating with Highest Honors in 2017 with a B.S. in Applied Mathematics, he entered the Mathematics PhD program at Georgia Tech.

Besides math, he greatly enjoys reading, hiking, chess, shooting, weightlifting, mountain biking, and above all else eating ice cream.