Imperial College London

Imperial College London

Department of Mathematics

Identifiability of Hawkes Processes

Author:
Andrew Connell

Supervisor:
Dr. Ed Cohen

A thesis submitted for the degree of

MSc Statistics (Theory and Methods)

September 25, 2020

Anti-Plagiarism Statement

I certify that this dissertation is my own work, based on my personal research and that I have acknowledged all material and sources used in its preparation, whether they be books, articles, reports, lecture notes, and any other kind of document, electronic or personal communication. I also certify that this dissertation has not previously been submitted for any other assessment. Further to this, I have not copied or otherwise plagiarised the work of others.

Acknowledgements

I would like to thank my supervisor Dr. Ed Cohen, and Leigh Shlomovich for their support. The M-dimensional derivation of the Hessian given by Shlomovich was a necessary step in deriving the Fisher information. In addition, the Imperial College Mathematics Department for their commitment in continuing research in these challenging times. I also wish to thank Dr. Rebecca Killick for her support throughout the duration of my academic studies.

Furthermore, I would like to thank my family and friends for supporting me throughout the years. I wish to pay special thanks to my parents without whom, I would never have studied mathematics.

Abstract

Hawkes processes are a class of stochastic point processes that model self-exciting and mutually exciting event data. The occurrence of an event can trigger further events of the same type and events of a different type. Hawkes processes have gained much attention due to their real-world applications in areas such as: medicine, seismology, finance and counter terrorism.

A vital aspect of model fitting and parameter estimation is one of identifiability. Identifiability can be defined as the existence of a unique maximum likelihood estimator for any arbitrary data set. When models become unidentifiable, there are multiple parameter values that are equally likely to have given rise to the data. This can cause problems as the fitted model can diverge significantly from the true data generating process, rendering inference and prediction essentially useless.

Understanding when Hawkes processes become unidentifiable has yet to be fully explored; however, in much of the current literature identifiability is either assumed or ignored entirely. Despite identifiability not being a concern exclusive to Hawkes processes, without a set of interpretable conditions, problems arise more readily in Hawkes processes. Furthermore, these issues are more evident when multidimensional Hawkes processes are considered.

This dissertation investigates under what circumstances Hawkes processes become unidentifiable: providing an interpretation of how the structure of Hawkes processes affect the identifiability of the process itself. Theoretical results, such as the Fisher information matrix, will be given alongside empirical evidence to provide a methodology to choose appropriate identifiable models for an arbitrary data set and highlight situations where this is not possible. This novel theory provides a much needed interpretable set of conditions that can be applied to achieve reproducible and rigorous predictions.

Contents

Notation

$\mathbb{R}^+$	Positive real numbers $\{x : x \in \mathbb{R}, x \geq 0\}$.
$N(t)$	A counting process indexed at time t.
$\mathbb{1}_{\{\}}$	An indicator function where the condition is in $\{\}$.
$\lambda(t)$	The conditional intensity, representing the excitement of the process at time t.
$\mathcal{H}(t)$	History of process up to time t.
α	Instantaneous jump size of the process.
β	Decay rate of a process.
μ	Base or background intensity.
ν	General kernel that denotes any kernel from the set of possible choices.
η	Branching ratio of a Hawkes process.
$\{t_1, t_2, ..., t_n\}$	Observed event times indexed from 1 up to the n^{th} event.
Λ	The compensator defined as the integral of the conditional intensity function.
$f^*(t)$	The conditional probability density function.
$F^*(t)$	The conditional cumulative distribution function.
$[0, T]$	The interval over which all event times occur between. That is, $0 \leq t_1 < ... < t_n \leq T$.
$\ell(\cdot)$	The log-likelihood function.
$\boldsymbol{H}$	The Hessian matrix.
$I(\cdot)$	The Fisher information.
$\partial f/\partial\theta$	The partial derivative of the function f with respect to θ.
$\xrightarrow{\mathbb{P}}$	Tends to in probability.

Chapter 1

Introduction

Hawkes processes were introduced in 1971 in the context of modelling earthquakes [Hawkes, 1971]. They mathematically model self–exciting processes; that is, an event occurring increases the probability of another event occurring. Hawkes self-exciting processes are a special type of point process where, due to the history of the process effecting the likelihood, information of past events is required. Hawkes introduced a mathematically tractable point process model that incorporated the self-exciting structure of the data [Hawkes, 1971, Oakes and Hawkes, 1974].

Hawkes processes are a type of branching point process model that are often used in modelling clustered phenomena in many real-world settings. It is common when collecting event data to notice clustering of events arising. The versatility of Hawkes processes has resulted in numerous real-world applications. Examples include: in financial data, to model the propagation of stock crashes and surges [Embrechts et al., 2011]; in social media, for modelling 'twitter cascades' [Kobayashi and Lambiotte, 2016]; the modelling of earthquake events [Ogata, 1988]; the monitoring and prediction of terrorist activity [Lewis and Mohler, 2011]; and for medical applications such as neuron firing times [Gerhard et al., 2017]. Note, that an extension to the simplistic Hawkes process model can relax the assumption that the background event intensity is constant and therefore, allow for a wide range of further applications including modelling pandemics [Chiang et al., 2020].

Identifiability is a statistical property that concerns itself with the ability to distinguish between multiple parameter values and recover the true parameter values from the observed data. Identifiability analysis is of major importance in all statistical studies due to the property's implication. Identifiability directly allows for precise inference to be made on a model meaning that multiple independent groups must draw the exact same inference. Many methods for parameter estimation require that a model is identifiable, despite the methods themselves not resolving the issue. Significant extensions to the original Hawkes processes have been made in recent years, however, identifiability in the Hawkes process setting is yet to be fully explored. This dissertation extends the limited theory available on the conditions that lead to Hawkes processes becoming unidentifiable. The identification problem has only ever been examined after a specific Hawkes process has been fitted [Chen and Hall, 2016]. Furthermore, even discussion in this context is rare and it is more common for identifiability to be assumed or ignored entirely. This dissertation investigates under what circumstances Hawkes processes become unidentifiable: providing an interpretation of how the structure of Hawkes processes affect the identifiability of the process itself. Theoretical results, such as the Fisher information matrix, will be given alongside empirical evidence to provide a methodology to deduce whether an arbitrary data set corresponds to an appropriate identifiable Hawkes process.

The remainder of the dissertation is organised as follows. Chapter 2 introduces the basic mathematics of Hawkes processes, discussing both the univariate and multivariate cases. Identifiability and its importance for Hawkes processes will be introduced in Chapter 3. The literature review is given in Chapter 4 and discusses the existing theory on Hawkes processes and identifiability research. Chapter 5 then outlines some novel methodology with proofs available in the Appendix. Chapter 6 outlines the applications of this methodology, and through simulated data, applies them and provides some numerical analysis. Finally, concluding remarks can be found in Chapter 7.

Chapter 2

Background to Hawkes Processes

2.1 Point and Counting Processes

Before Hawkes self-exciting processes can be fully explored, the fundamental aspects of point processes to Poisson processes need to be explained.

A stochastic process is defined as a family of random variables $\{T_i, i \in \mathbb{N}_0\}$ which are indexed over time and are from some probability space Ω. Here, i is the index set of the process. In statistics, point processes are sequences of points that all occur on an underlying space and are monotonically increasing. More formally, let $\{T_i, i \in \mathbb{N}_0\}$ be a sequence of non-negative random variables such that $\forall i \in \mathbb{N}_0$, $T_i < T_{i+1}$. It is often assumed that T_i is the time at which an event has occurred. Note, while not necessary for the definition to hold, it is often assumed that $T_0 = 0$.

A counting process is a logical extension to a point process. To gain some basic intuition of counting processes, consider a simple stochastic process defined as $\{N(t), t \geq 0\}$. This is called a counting process if $N(t)$ is the function that outputs the total number of events that have occurred up to and including time t. As counting processes occur in several areas of mathematics, they have many marginally different definitions depending on the context in which they are discussed. Here, counting processes are formally defined to meet the following conditions:

$$N(0) = 0, \qquad (2.1) \qquad\qquad N(t) \in \mathbb{N}_0, \qquad (2.2)$$

$$t_1 < t_2 \Rightarrow N(t_1) \leq N(t_2), \qquad (2.3) \qquad\qquad \lim_{h \to 0}\{N(t+h)\} = N(t), \qquad (2.4)$$

$$N(t) - N(t-) \in \{0, 1\}, \qquad (2.5) \qquad\qquad \mathbb{E}[N(t)] < \infty. \qquad (2.6)$$

Note that all six conditions must be held to consider $N(t)$ a counting process, though conversely, all hold true and can be utilised if it is known $N(t)$ is a counting process. Equations 2.4 and 2.5 denote right-continuity and the left limit respectively.

Given some point process $\{T_i, i \in \mathbb{N}_0\}$, then the right-continuous process may be expressed as

$$N(t) = \sum_{i \in \mathbb{N}_0} \mathbb{1}_{\{T_i \leq t\}},$$

which is the corresponding counting process to $\{T_i, i \in \mathbb{N}_0\}$ and hence, the relationship can be seen.

These two processes relate to one another, and while each has some unique properties, both have similar overarching features. In each case, three properties of interest arise when these processes are used for statistical modelling: independence, stationarity, and homogeneity. These properties are commonplace in most statistical studies, especially in time series analysis; however in the context of counting processes are defined slightly differently. A counting process $\{N(t), t \geq 0\}$ is said to meet the independent increment property if the numbers of events that occur in disjoint time intervals are independent. The counting process would be said to meet the stationarity increment property if for any $t, k \geq 0$, then $N(t-k) \sim N(t) - N(k)$. This mathematically expresses that the distribution of the number of events in a given interval is only dependent on the length of said interval. Finally, a counting process would be said to meet the homogeneity increment property if the transition probability between any two given states, $t, k \geq 0$ is only dependent on the difference between those two states. That is, for some $t \geq 0$ then,

$$p_{t,t+1} = \mathbb{P}(N(t+1) = k|N(t) = l) = \mathbb{P}(N(t+n+1) = k|N(t+n) = l) = p_{t+n,t+n+1},$$

for any $n \in \mathbb{N}_0$. It can also be concluded that this condition must hold for the m-step transition probabilities, $p_{i,j}^{(m)}$.

The stationarity and independence increment properties are always satisfied when using a special example of a counting process: the Poisson process.

2.2 Poisson Processes

The Poisson process is one of the most commonly used counting processes due to it satisfying a large number of modelling assumptions. Poisson processes are one of the most basic forms of continuous-time stochastic processes. Intuitively, Poisson processes are a counting process that start at zero and then count events as they occur during a specific time period, t. More formally, a homogeneous Poisson process $\{N(t), t \geq 0\}$ with rate $\lambda > 0$ is a stochastic process that satisfies three conditions. The first is, $N(0) = 0$, which is Equation 2.1. The increments are independent, that is, given any choice of $N \in \mathbb{N}_0$ and $0 \leq t_0 < t_1 < \cdots < t_n$, the random variables $N_{t_0}, N_{t_1} - N_{t_0}, N_{t_2} - N_{t_1}, ..., N_{t_n} - N_{t_{n-1}}$ are independent. Finally, for any $0 \leq s < t, k \in \mathbb{N}_0$,

$$\mathbb{P}(N_t - N_s = k) = \frac{(\lambda(t-s))^k e^{-\lambda(t-s)}}{k!}.$$

This directly states that the number of events on the interval $[s, t]$ is a Poisson random variable with rate $\lambda(t-s)$. The proof of this statement follows directly from the moment generating function of a Poisson random variable. This last property implies that $\mathbb{E}[N(t)] = \lambda t$.

Another definition commonly used, and one more comparable to the Hawkes process definition, is that a homogeneous Poisson process is a point process $\{N(t), t \geq 0\}$, with rate $\lambda \in \mathbb{R}^+$, if the properties

$$N(0) = 0, \qquad (2.7) \qquad\qquad \{N(t), t \geq 0\} \text{ is independent,} \qquad (2.8)$$

$$\mathbb{P}(N(t+h) - N(t) = 1) = \lambda h + o(h), \quad (2.9) \qquad \mathbb{P}(N(t+h) - N(t) \geq 2) = o(h), \quad (2.10)$$

are satisfied. This definition shows that intensity does not depend on the process' history, and therefore, the probability of an event occurring in the interval $[t, t+h]$ is also independent from the history. There are a large number of other extensions to the Poisson process that follow directly from this definition including the superposition of independent Poisson processes being itself a Poisson process. However, in this setting the development of the inhomogeneous Poisson process is of particular interest.

2.2.1 Inhomogeneous Poisson Processes

The Poisson process often has a prefix dropped; there are two alternative versions of the process: the homogeneous and inhomogeneous Poisson process. The inhomogeneous Poisson process is a more generalised version, where the rate varies as a function with time expressed as $\lambda(t)$. The definition remains the same except in the final condition where the constant λ is replaced with the deterministic rate function $\lambda(t)$. Often, rather than referring to this as a rate function, it is known as the intensity function.

In the case of the inhomogeneous Poisson process, it is still true that the number of events that occur in any interval, $[t, t+s]$, is a Poisson random variable. However, previously where the value was the constant λ, now the distribution must be written as

$$N(t+s) - N(t) \sim \text{Poisson}(\Lambda_{t,t+s}),$$

where $\Lambda_{t,t+s} = \int_t^{t+s} \lambda(x) \ \mathrm{d}x$. The term $\Lambda_{t,t+s}$ is often referred to as the compensator of the inhomogeneous Poisson process where the subscripts are dropped. The definition of the compensator can be generalised for counting processes, given that Λ is a non-decreasing function [Laub et al., 2015].

Note further, that the density of an inhomogeneous Poisson process on Ω for a given set of observations, $Y = (y_1, ..., y_n)$, is

$$f(Y) = \left(\prod_{i=1}^{n} \lambda_\theta(y_i) \right) e^{\{\int_{x \in \Omega} \lambda_\theta(x) \ \mathrm{d}x\}},$$

where λ_θ is the intensity function parameterised by θ. It can be noted that the maximum likelihood estimator, $\hat{\theta}$, is intractable, meaning it can only be found through numerical methods. This can cause issues with identifiability, which will be further explored in the Hawkes process setting.

2.3 An Introduction to Hawkes Processes

A major development of point processes in the statistics' literature occurred during the twentieth century. In 1963, Bartlett introduced several new statistical methods for point processes using a process' spectral density [Bartlett, 1963]. By the late 1960s point processes had been extended to many time series applications [Cox and Lewis, 1966]. This rapid development of point processes led to the development of the Hawkes process. Hawkes processes are statistical models for self-exciting processes, gaining their name from Alan G. Hawkes. The original Hawkes process described in 1971 was a one-dimensional point process that was also extended to M-dimensional cases, where $M \geq 1$ [Hawkes, 1971]. Moreover, it was a counting process that modelled the occurrence of earthquake events where, as each event occurred, it excited the process, increasing the probability of subsequent events occurring for a given period until the effect of the original event decayed. It is, therefore, possible to show that the Hawkes process is a non-Markovian extension of the Poisson process except in a few special cases [Laub et al., 2015]. This condition implies that Hawkes processes have a dependency structure that is formed from the whole history of the process. Hawkes processes can therefore also be referred to as self-exciting long memory processes [Karabash, 2012].

Hawkes processes can be completely characterised either by their intensity functions or through a Poisson cluster process [Lim et al., 2016, Møller and Rasmussen, 2005]. The major focus in the next two sections will be based on characterising the processes through their intensity functions. Consider a counting process $\{N(t), t \geq 0\}$ with some associated history $\{\mathcal{H}(t), t \geq 0\}$. This is a Hawkes process if for some $t, h \in \mathbb{N}_0$ then,

$$\mathbb{P}(N(t+h) - N(t) = k | \mathcal{H}(t)) = \begin{cases} \lambda(t|\mathcal{H}(t))h + o(h) & \text{for } k = 1, \\ o(h) & \text{for } k > 1, \\ 1 - \lambda(t|\mathcal{H}(t))h + o(h) & \text{for } k = 0, \end{cases} \tag{2.11}$$

where $\lambda(t|\mathcal{H}(t)) = \lambda(t|N(s), s < t)$ is the conditional intensity function. The history, $\mathcal{H}(t)$ is a filtration on the underlying probability space $(\Omega, \mathcal{H}, \mathbb{P})$ generated by the counting process [Daw and Pender, 2018]. For simplicity the intensity function is often denoted $\lambda(t)$. This counting process considered would be a Hawkes process provided that $\lambda(t)$ is a random intensity function. Hawkes presented this definition in his original paper, though many slightly altered forms have since been suggested [Laub et al., 2015].

Before proceeding to define the generalised Hawkes process, it is worthwhile consolidating the previous definitions to define what it means for a counting process to be self-exciting. A counting process $\{N(t), t \geq 0\}$ is said to be self-exciting if

$$\lambda(t) = \mu(t) + \int_{-\infty}^{t} \nu(t-s) \ \mathrm{d}N(s), \qquad \text{or} \tag{2.12a}$$

$$\lambda(t) = \mu(t) + \sum_{t_i \leq t} \nu(t - t_i). \tag{2.12b}$$

Further note, that $\nu(t) \geq 0$ for all t and $\int_0^\infty \nu(s)\ \mathrm{d}s < 1$. Equations 2.12a and 2.12b sometimes are expressed using different notation and it is commonplace in literature to see $\lambda_0(t)$ instead of $\mu(t)$. In the case of both equations, $\mu : \mathbb{R} \to \mathbb{R}^+$ is the base, or background, intensity and $\nu : \mathbb{R}^+ \to \mathbb{R}^+$ is the kernel which expresses the positive influence of past events t_i on the current value of the intensity process. Figure 2.1 is a realisation of a counting process that meets the requirements outlined here and in this case, corresponds to a stationary one-dimensional Hawkes process.

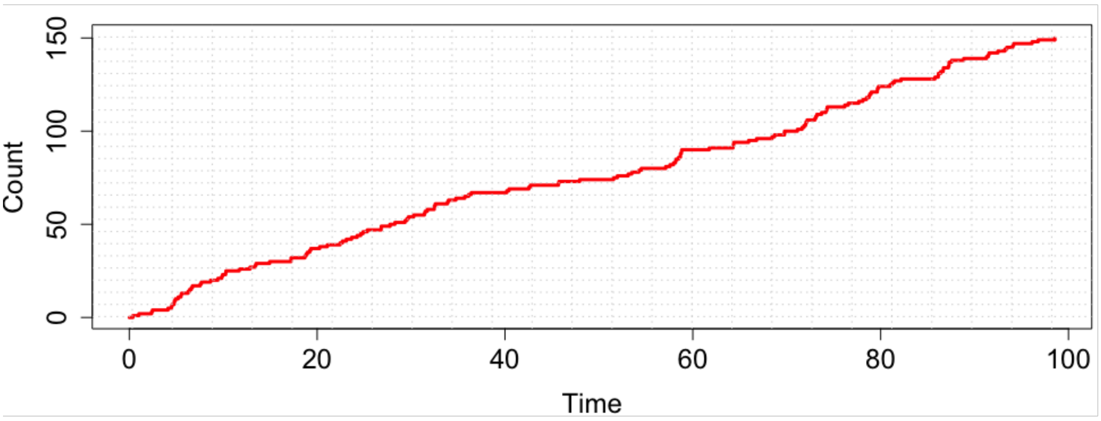

Figure 2.1: Example of a Counting Process for a One-Dimensional Hawkes Process.

Hawkes processes are a specific type of self-exciting counting process which were originally introduced with the exponential kernel. The simplest form the exponential kernel can take is $\nu(t) = \alpha e^{-\beta t}$ where $\alpha, \beta \in \mathbb{R}^+$. The parameters α and β are respectively thought of as the instantaneous increase, or jump, in the arrival intensity when an event occurs, and the decay rate of the new arrival intensity after the event's occurrence.

Alternative conditional intensity functions follow from Equations 2.12a and 2.12b where different kernels may be used to define a Hawkes process and though not used here, are worth briefly noting. A power law function is an acceptable choice with the formula

$$\lambda(t) = \mu(t) + \int_{-\infty}^{t} \frac{k}{(c + (t-s))^p}\ \mathrm{d}N(s) = \mu(t) + \sum_{t_i \leq t} \frac{k}{(c + (t - t_i))^p},$$

where $c, k, p \in \mathbb{R}^+$ [Laub et al., 2015]. Other alternatives include the use of piecewise linear functions for numerical computing of the intensity function [Chatalbashev et al., 2007].

The general definition in Equation 2.11 combined with the first order exponential kernel proposed by Hawkes, where the the exponential kernel was alternatively called the excitation function [Hawkes, 1971], is the definition used throughout. The Hawkes process models discussed in the following sections directly follow from Hawkes' seminal work but allow for higher order exponential kernels.

2.4 The One-Dimensional Hawkes Process

For the one-dimensional Hawkes process, the self-exciting process is of order P with the exponential kernel $\nu(t) = \sum_{j=1}^{P} \alpha_j e^{-\beta_j t}$, yielding the conditional intensity to be,

$$\lambda(t) = \mu(t) + \int_0^t \sum_{j=1}^{P} \alpha_j e^{-\beta_j (t-s)}\ \mathrm{d}N(s), \tag{2.13}$$

$$= \mu(t) + \sum_{t_i \leq t} \sum_{j=1}^{P} \alpha_j e^{-\beta_j (t - t_i)},$$

where t_i denotes the i^{th} event or 'arrival' and t the last event time. Note, that μ denotes the base, or background, intensity of the process, α_j is the j^{th} non-negative jump size of the intensity function and β_j is the j^{th} decay rate. In the simplest case when $P = 1$ and $\mu(t) = \mu$ for all t, then the intensity simplifies to

$$\lambda(t) = \mu + \int_0^t \alpha e^{-\beta(t-s)} \, \mathrm{d}N(s) = \mu + \sum_{t_i \leq t} \alpha e^{-\beta(t-t_i)},$$

yielding constant values of both α and β.

In the one-dimensional setting, the process can easily be verified to be stationary, by noting that $\mathbb{E}[\lambda(t)] = c$, where c is some constant, if the process is stationary. Assuming this, when $P = 1$ the expected value of the intensity of a stationary process is constant and can be seen to be

$$\mathbb{E}[\lambda(t)] = \frac{\mu}{1 - \alpha/\beta}. \tag{2.14}$$

The derivation of this can be seen in Appendix A.1. A simple example of the intensity function for a one-dimensional stationary Hawkes process can be seen below in Figure 2.2 where the expected intensity has also been plotted.

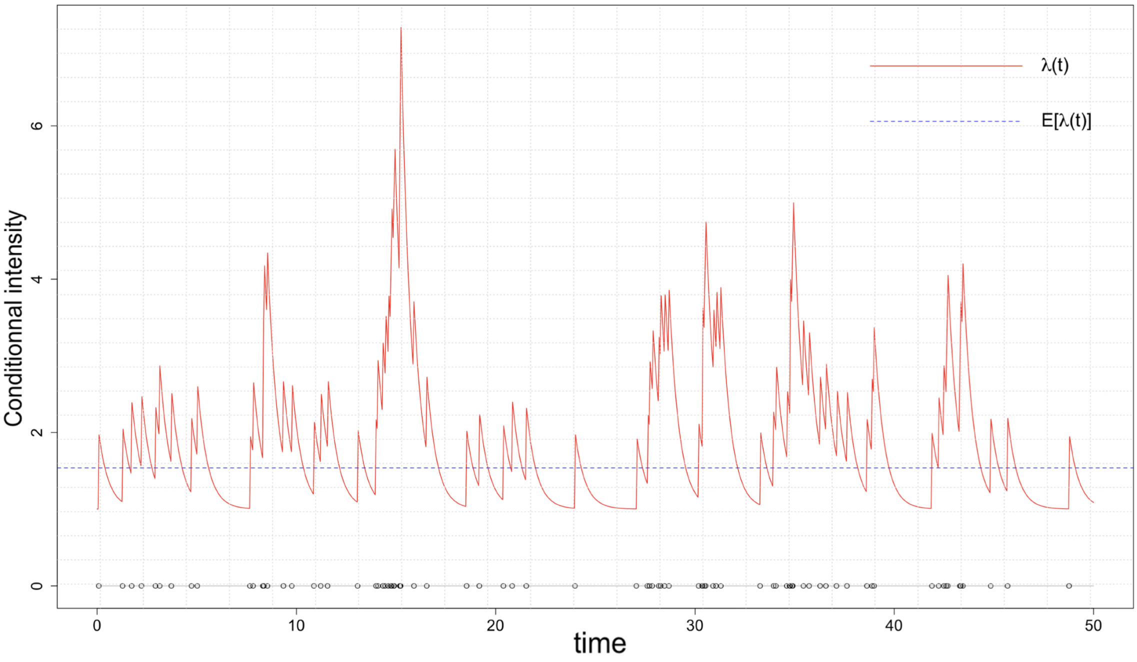

Figure 2.2: Intensity of Hawkes Process with $\mu = 1, \alpha = 0.7$ and $\beta = 2$.

Using Equation 2.13 it can be seen that the one-dimensional Hawkes process is stationary provided,

$$\sum_{j=1}^{P} \frac{\alpha_j}{\beta_j} < 1. \tag{2.15}$$

The term branching ratio is often used to describe the relationship between the α and β parameters presented in Equation 2.15 [Laub et al., 2015]. Literature exists on estimating the branching ratio for Hawkes processes, however little is known about the effect this ratio has on identifiability [Hardiman and Bouchaud, 2014]. The branching ratio, η, is found by integrating the kernel chosen. In the simplest case,

$$\eta = \int_0^\infty \nu(s) \, \mathrm{d}s = \int_0^\infty \alpha e^{-\beta s} \, \mathrm{d}s = \frac{\alpha}{\beta}.$$

Note, that when η is greater than one, the process explodes. Conceptually, the process is considered explosive when each event causes more events than itself to occur. The condition stated in Equation 2.15, also states that explosive values of η are non-stationary.

2.4.1 Likelihood of a One-Dimensional Hawkes Process

The conditional intensity function expressed in Equations 2.12a and 2.12b may also be expressed for a point process as

$$\lambda(t) = \frac{f^*(t)}{1 - F^*(t)}, \tag{2.16}$$

where $f^*(t)$ and $F^*(t)$ are respectively the conditional probability density and cumulative distribution functions. These may be defined as

$$f^*(t) = \lambda(t)e^{-\int_{t_i}^{t} \lambda(s)\ \mathrm{d}s}, \quad (2.17) \qquad F^*(t) = 1 - e^{-\int_{t_i}^{t} \lambda(s)\ \mathrm{d}s}, \quad (2.18)$$

where t_i denotes the event prior to t [Shlomovich et al., 2020]. Following from these definitions, it can be seen that the likelihood of a point process with an intensity function $\lambda(s|\theta)$ and observations $t_1, t_2, ..., t_n$ may be written as

$$L(t_1, ..., t_n, |\theta) = L(t|\theta) = \prod_{i=1}^{n} f^*(t)(1 - F^*(t)) = \left\{ \prod_{i=1}^{n} \lambda(t_i) \right\} e^{-\int_0^t \lambda(s)\ \mathrm{d}s}, \tag{2.19}$$

and therefore,

$$\ell(t|\theta) = \sum_{i=1}^{n} \log(\lambda(t_i)) - \int_0^t \lambda(s|\theta)\ \mathrm{d}s = \int_0^t \log\{\lambda(s|\theta)\}\ \mathrm{d}N(s) - \int_0^t \lambda(s|\theta)\ \mathrm{d}s. \tag{2.20}$$

Note here, the process is over the time interval $[0, t]$ and $t \geq t_n$ [Rubin, 1972]. Therefore, substituting in Equation 2.13, it can be shown that the log-likelihood of a one-dimensional Hawkes process is

$$\ell(t|\theta) = \int_0^t \log\left(\mu(s) + \sum_{t_i \leq t} \sum_{j=1}^{P} \alpha_j e^{-\beta_j (t - t_i)}\right) \mathrm{d}N(s) - \int_0^t \left(\mu(s) + \sum_{t_i \leq t} \sum_{j=1}^{P} \alpha_j e^{-\beta_j (t - t_i)}\right) \mathrm{d}s, \tag{2.21}$$

where $\theta = (\mu, \alpha_1, ..., \alpha_P, \beta_1, ..., \beta_P)$. Equation 2.19 can be used to derive an expression for the observed information [Ogata, 1988]. Calculating the log-likelihood in this manner is computationally costly and leads to $\mathcal{O}(n^2)$ complexity [Laub et al., 2015]. However, this complexity may be reduced to $\mathcal{O}(n)$ by a recursive formula [Ogata, 1981]. Using this recursion, the log-likelihood can be re-expressed as

$$\ell(t|\theta) = \left(\sum_{i=1}^{n} \log\left(\mu(t_i) + \sum_{j=1}^{P} \alpha_j R_j(i)\right)\right) - \int_0^t \mu(s)\ \mathrm{d}s - \sum_{i=1}^{n} \sum_{j=1}^{P} \alpha_j e^{-\beta_j (t - t_i)}, \tag{2.22}$$

where,

$$R_j(i) = \sum_{t_k < t_i} e^{-\beta_j (t_i - t_k)} = e^{-\beta_j (t_i - t_{i-1})}(1 + R_j(i-1)),$$

and for all j, $R_j(1) = 0$. The full proof of this process can be found in Appendix A.2 for the general M-dimensional process. Again, it is worth noting the special case when $P = 1$ and $\mu(t) = \mu$ for all t yields the simpler log-likelihood

$$\ell\Big(t|\theta = (\mu, \alpha, \beta)\Big) = \int_0^t \log\left(\mu + \alpha \sum_{t_j \leq t} e^{-\beta(t - t_j)}\right) \mathrm{d}N(s) - \int_0^t \left(\mu + \alpha \sum_{t_j \leq t} e^{-\beta(t - t_j)}\right) \mathrm{d}s. \tag{2.23}$$

For the stationary one-dimensional Hawkes process, with the likelihood stated above, the maximum likelihood estimator $\hat{\theta}^T = (\hat{\mu}, \hat{\alpha}, \hat{\beta})$ has been shown to be consistent, asymptotically normal and asymptotically efficient [Ogata, 1978].

The maximum likelihood estimator is of particular interest in the discussion of parameter identifiability, which is discussed in Chapter 3. However, for Hawkes processes the estimator does not have a closed form and therefore must be found through numerical methods. There are several methods available to simulate the one-dimensional Hawkes process however, a simulation method for M-dimensional Hawkes processes based on the superposition theory of point processes will be utilised [Lim et al., 2016]. This method allows for the trivial M-dimensional case when $M = 1$ and, for convenience, this is also the way in which the one-dimensional process will be simulated.

2.5 The M-Dimensional Hawkes Process

In the M-dimensional setting, the Hawkes process can again be defined to be of order P with an exponential kernel, but it is also necessary to denote the background intensities μ_m with $m = 1, ..., M$ and M being the total number of dimensions. Thus, the general M-dimensional Hawkes process has conditional intensities λ_m expressed as

$$\lambda_m(t) = \mu_m(t) + \sum_{i=1}^{M} \int_0^t \sum_{j=1}^{P} \alpha_{i,j,m} e^{-\beta_{i,j,m}(t-s)} \, \mathrm{d}N_i(s) \ \text{ or} \tag{2.24a}$$

$$\lambda_m(t) = \mu_m(t) + \sum_{i=1}^{M} \sum_{t_{k,i}<t} \sum_{j=1}^{P} \alpha_{i,j,m} e^{-\beta_{i,j,m}(t-t_{k,i})}. \tag{2.24b}$$

Note here, that μ_m is the base intensity of the m^{th} process and the definitions of α and β follow similarly to those presented in Equation 2.13. Extending the definition for M-dimensional processes, the α and β parameters are presented as matrices, which express both the behaviour and the cross-behaviour between the dimensions. Again it is worth noting the simple case when $P = 1$ and $\mu_m(t) = \mu_m$ for all t, then the intensity simplifies to

$$\lambda_m(t) = \mu_m + \sum_{i=1}^{M} \int_0^t \alpha_{i,m} e^{-\beta_{i,m}(t-s)} \, \mathrm{d}N_i(s) = \mu_m + \sum_{i=1}^{M} \sum_{t_{k,i}<t} \alpha_{i,m} e^{-\beta_{i,m}(t-t_{k,i})}. \tag{2.25}$$

The stationarity condition can easily be found here, though for simplicity will be given for when $P = 1$. The condition can be visualised by rewriting Equation 2.25 using vector notation [Hawkes, 1971]. Recall first, that for a given dimension, when $P = 1$ and the background intensity is constant,

$$\lambda_m(t) = \mu_m + \sum_{i=1}^{M} \int_{-\infty}^t \nu_{i,m}(t-s) \, \mathrm{d}N_i(s).$$

It is therefore possible to denote the vector of intensities as

$$\boldsymbol{\lambda}(t) = \boldsymbol{\mu} + \sum_{i=1}^{M} \int_0^t \boldsymbol{G}(t-s) \, \mathrm{d}\boldsymbol{N}_s, \tag{2.26}$$

where $\boldsymbol{G}(t) = \{\alpha_{i,m} e^{\beta_{i,m}(t-s)}\}_{i,m=1}^{M}$ [Cordi et al., 2018]. It is then possible to note that if the process is stationary then $\mathbb{E}[\boldsymbol{\lambda}(t)] = \boldsymbol{c}$, where $\boldsymbol{c}$ is a constant vector. Therefore, for the order $P = 1$ M-dimensional Hawkes process, the expectation is

$$\boldsymbol{c} = \frac{\boldsymbol{\mu}}{\left(\boldsymbol{1} - \sum_{i=1}^{M} \int_0^\infty \boldsymbol{G}(x) \, \mathrm{d}x\right)}, \tag{2.27}$$

where $\boldsymbol{c}$ denotes the constant vector that satisfies $\mathbb{E}[\boldsymbol{\lambda}(t)] = \boldsymbol{c}$ and $\boldsymbol{1}$ is a vector of length M with each element equal to one. Note, that the formula divides element-wise. The condition to satisfy that any M-dimensional Hawkes process is stationary is

$$\boldsymbol{\Gamma} = \int_0^\infty \boldsymbol{G}(x) \, \mathrm{d}x = \left\{ \frac{\alpha_{i,m}}{\beta_{i,m}} \right\}_{i,m=1}^{M} < 1. \tag{2.28}$$

This integral is sometimes referred to as the spectral radius. Further to this, as discussed in the literature [Cordi et al., 2018], the spectral radius of the matrix $\boldsymbol{G}$ is equivalent to the maximum absolute eigenvalue of $\boldsymbol{G}$. That is,

$$\rho(\boldsymbol{G}) = \max_{\omega \in \delta(\boldsymbol{G})} |\omega|,$$

where $\delta(\boldsymbol{G})$ is all the eigenvalues of the matrix $\boldsymbol{G}$. Note that $\boldsymbol{c}$ is a vector of length M, while $\boldsymbol{G}$ is an $M \times M$ matrix which is then summed across each row. The M-dimensional likelihood for a Hawkes process must now be introduced in order to perform inference.

2.5.1 Likelihood of an M-Dimensional Hawkes Process

The log-likelihood of a multidimensional Hawkes process can be computed as the sum of the likelihood of each dimension. That is, given some realisation that contains all the points in each dimension denoted $\{t_{i,m}\}_{i=1}^{N}$ for each of the $1, ..., M$ dimensions on the interval $[0, t]$, the full log-likelihood may be expressed as

$$\ell(t|\boldsymbol{\theta}) = \sum_{m=1}^{M} \ell_m(t|\theta_m), \tag{2.29}$$

where $\boldsymbol{\theta}$ denotes all parameters necessary for the model. The formula for the likelihood of the m^{th} dimension is

$$\ell_m(t|\theta_m) = \int_0^t \log\{\lambda_m(s|\theta_m)\}\ \mathrm{d}N_m(s) - \int_0^t \lambda_m(s|\theta_m)\ \mathrm{d}s, \tag{2.30}$$

where the similarity to Equation 2.20 can be seen. The log-likelihood for the m^{th} dimension in this general case can be written tractably as

$$\begin{aligned}\ell_m(t|\theta_m) = \Bigg(\sum_{k=1}^{N} \log\Big(\mu_m(t_{k,m}) + \sum_{i=1}^{M}\sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k)\Big)\Bigg) - \Bigg(\sum_{k=1}^{N} \mu_m(t_{k,m})\Bigg) \\ - \Bigg(\sum_{i=1}^{M}\sum_{j=1}^{P}\sum_{k=1}^{N} \frac{\alpha_{i,j,m}}{\beta_{i,j,m}}\Big(1 - e^{-\beta_{i,j,m}(t-t_{k,i})}\Big)\Bigg),\end{aligned} \tag{2.31}$$

where $R_{i,j,m}(k)$ is the recursive formula defined as

$$R_{i,j,m}(k) = \begin{cases} e^{-\beta_{i,j,m}(t_{k,m}-t_{k-1,m})} R_{i,j,m}(k-1) + \sum_{t_{k-1,m} \leq t_{d,i} < t_{k,m}} e^{-\beta_{i,j,m}(t_{k,m}-t_{d,i})} & \text{if } i \neq m, \\ e^{-\beta_{i,j,m}(t_{k,m}-t_{k-1,m})}\Big(1 + R_{i,j,m}(k-1)\Big) & \text{if } i = m, \end{cases}$$

and $R_{i,j,m}(0) = 0$. The full derivation of the log-likelihood can be found in Appendix A.2. In the simpler case when $P = 1$, Equation 2.31 simplifies to

$$\ell_m(t|\theta_m) = \sum_{k=1}^{N} \log\Big(\mu_m + \sum_{i=1}^{M} \alpha_{i,m} R_{i,m}(k)\Big) - \mu_m t - \sum_{i=1}^{M}\sum_{k=1}^{N} \frac{\alpha_{i,m}}{\beta_{i,m}}\Big(1 - e^{-\beta_{i,m}(t-t_{k,i})}\Big), \tag{2.32}$$

where $R_{i,m}(k)$ is defined as

$$R_{i,m}(k) = \sum_{d:t_{d,i}<t_{k,m}} e^{-\beta_{i,m}(t_{k,m}-t_{d,i})},$$

for $k \geq 2$ and $R_{i,m}(0) = 0$. This can then be written recursively as

$$R_{i,m}(k) = e^{-\beta_{i,m}(t_{k,m}-t_{k-1,m})} R_{i,m}(k-1) + \sum_{d:t_{k-1,m} \leq t_{d,i} < t_{k,m}} e^{-\beta_{i,m}(t_{k,m}-t_{d,i})}.$$

The derivation of the M-dimensional process allows for an alternate form for the one-dimensional log-likelihood with $P = 1$ [Laub et al., 2015]. That is,

$$\ell(t|\theta) = \sum_{k=1}^{N} \log\Big(\mu + \alpha R(k)\Big) - \mu t - \frac{\alpha}{\beta}\sum_{k=1}^{N}\Big(1 - e^{-\beta(t-t_k)}\Big), \tag{2.33}$$

with $R(k) = \sum_{t_i<t_k} e^{-\beta(t_k-t_i)} = e^{-\beta(t_k-t_{k-1})}(1 + R(k-1))$.

In the context of identifiability, having a closed form likelihood is paramount. The next chapter presents the broader topic of parameter identifiability and explores the different issues that can result in a process being non-identifiable.

Chapter 3

Identifiability

Identifiability is a property that must be satisfied in order to ascertain that the inference from a model is precise. Any model can be considered identifiable if the underlying structure, and therefore the true parameter values, can be found given an infinite set of observations. In most cases, the discussion simplifies to parameter identifiability where the parameters are considered identifiable when the maximum likelihood estimator of each parameter can be found.

Model identifiability occurs if a unique bijective mapping from the parameter space, Θ, to the space of distributions for the data, sometimes thought of as the likelihood, exists then the model is identifiable [Patel et al., 2019]. More practically, assume that some arbitrary data, Y, has log-likelihood $\ell(Y;\theta) = \log(L(Y;\theta))$ where $L(\cdot,\cdot)$ is the likelihood function and $\theta \in \Theta$. The model is identifiable if for any $\theta, \phi \in \Theta$,

$$\ell(Y;\theta) = \ell(Y;\phi) \Rightarrow \theta = \phi.$$

Note that this definition holds if the log-likelihood is replaced with the probability distribution of the data [Huang, 2005]. The identifiability property is often not held due to poor model specification; however, identifiability can never be satisfied if the number of parameters in the model is greater than the number of observations. Due to the importance of identifiability in making precise inference, if the model is unidentifiable, the application of additional constraints to meet the identifiability property can be justified.

3.1 Structural and Practical Identifiability

When discussing identifiability, two major types of identifiability should be considered: structural and practical.

Structural identifiability is the name given to the exact definition where the model is identifiable when the parameters can be inferred from infinite perfect data. This definition is rigid and dictates that a model is structurally unidentifiable if the same predictions can be made for multiple sets of parameter values [Bellman and Åström, 1970]. In this case, the mapping is not injective and therefore, cannot be bijective. A model that is unable to meet this condition is considered an unfavourable choice even when compared to an identifiable model with lower predictive capabilities [Goodrich and Caines, 1979]. Situations where unidentifiable models outperform identifiable models can occur, however unidentifiable parameters often lead to imprecise predictions.

In most real world problems the strictness of structural identifiability makes it difficult to verify. It is possible for a model to be structurally identifiable yet without an infinitely large data set impossible to prove. Therefore, the more applicable definition of practical identifiability can be used. This stipulates the need for the bijective mapping to exist on the real data set available. Practical identifiability is a stricter definition than that of structural identifiability as all models that are practically identifiable are also structurally identifiable however, the inverse is not true.

Identifiability analysis, the field of research that deals with identifiability issues, provides a set of methods that can be applied to test if the property is held. The profile likelihood is often utilised as an identifiability test: providing interpretable results on parameter dependency and how a model may be reduced to be made identifiable. This is a common method however, it is worth noting other tests exist such as the Identifiability-Test by Radial Penalisation (ITRP) that also may be used to verify the identifiability of a model [Kreutz, 2018]. This method verifies that only one parameter set attains the maximum of some penalised objective function, which is often a modified log-likelihood.

3.2 Local Identifiability

When identifiability is considered, it is often assumed to mean global identifiability. A model is globally identifiable if the maximum likelihood estimator, $\hat{\theta}$, for the model is unique. Note, that this condition must be satisfied for any M-dimensional model and each parameter in the model must be the maximum for this to hold.

Nevertheless, it is not always possible for such a point to occur and usually, in these settings, it is due to multiple parameter sets attaining the maximum likelihood. When this is not possible, a weaker form of identifiability is utilised: local identifiability. For local identifiability, the parameter vector θ is defined as locally identifiable if there exists a neighbourhood around it such that there is no other ϕ in that neighbourhood for which $\ell(Y;\theta) = \ell(Y;\phi)$ almost everywhere [Patel et al., 2019]. Due to the nature of local identifiability, it is therefore possible to have multiple interpretations, unlike in the global case. However, when restricted only to the neighbourhood with a given unique solution, θ, it is possible to state that there exists no other solutions in this region. This therefore shifts the problem from locating a unique maximum to finding neighbourhoods within which local identifiability holds, and finding the corresponding maxima in these given neighbourhoods.

A significant benefit of local identifiability is the ability to determine whether it holds given some basic statistical conditions. One such condition is to verify that the columns of the Jacobian matrix are independent of one another and hence θ is locally identifiable . Alternatively, the Fisher information matrix $I(\theta)$ is non-singular if and only if θ is locally identifiable [Rothenberg, 1971]. In other words, all the eigenvalues of the Fisher information matrix are positive. In most practical settings, the complexity of the model means the Fisher information matrix would take too long to compute and instead identifiability is verified through the observed Fisher information matrix, $J(\hat{\theta}) = -\nabla\nabla^T \ell(Y;\theta)|_{\theta=\hat{\theta}}$ evaluated at the maximum likelihood estimate $\hat{\theta}$ of θ (note also that ∇ is the vector of partial derivatives) [Colquhoun et al., 2003]. Generally, when it is not possible to verify identifiability it is due to a flat ridge within the likelihood surface meaning there is not a unique maximum value of θ.

3.3 Identifiability of Point Processes

3.3.1 The Hessian Matrix

In the case of identifiability for the Hawkes process, very little is currently known. The current research in the area predominantly concerns itself with searching for global maxima within the likelihood surface. In this setting, several random starting values may be chosen and a globally identifiable model would converge to the same maximum likelihood estimate in the possible parameter space Ω. A common technique for identifying the maximum likelihood is maximisation algorithms that implement quasi-Newton methods [Goodrich and Caines, 1979]. Quasi-Newton methods use an algorithm that relies on the calculation of the Hessian matrix and therefore, for most processes a closed tractable form has been found. Section 4.2 provides the Hessian matrix for the simple one-Dimensional Hawkes process.

The Hessian matrix is a square symmetric matrix containing all combinations of second-order partial derivatives of the likelihood for a given parameter set. For the simplest Hawkes process model this would yield a 3×3 matrix. In the most generic mathematical notation, if the log-likelihood is denoted l and the models parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ then the Hessian is

$$\boldsymbol{H} = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_p^2} \end{pmatrix}.$$

The more condensed definition uses indices to express the Hessian as

$$\boldsymbol{H}_{i,j} = \frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 \ell}{\partial \theta_j \partial \theta_i}. \tag{3.1}$$

The Hessian matrix describes the likelihood surface's curvature. The measure of the curvature of a likelihood surface can provide useful information on the likelihood function and may prove useful in verifying whether a Hawkes process is identifiable. As proven in 1979 for a general point process, provided the likelihood surface is smooth continuous, a Hessian matrix observed at θ is asymptotically non-singular if and only if θ is locally practically identifiable [Goodrich and Caines, 1979]. The implication of this theorem is practically useful as any non-singular Hessian matrix immediately leads to identifiability issues. However, more exploration is still needed as the assumptions and conditions are restrictive in many applications and moreover, make no claims about identifiability as the Hessian approaches non-singularity.

3.3.2 First and Second Order Derivatives

It is necessary to calculate all the second-order derivatives to find the Hessian. The recursive formula introduced allows all the derivatives to be calculated with $\mathcal{O}(n)$ complexity. The exact derivatives for the Hawkes process are outlined in Chapter 4 but first, the general derivatives for a point process with the intensity function given in Equation 2.12a will be given as first outlined in 1977 [Ozaki, 1977]. Recall

$$\lambda(s|\theta) = \mu + \int_{-\infty}^{s} \nu(s-u|\theta)\ \mathrm{d}N(u).$$

Given some observations $t_1, t_2, ..., t_N$ on the interval $[0, t]$ where $t \geq t_N$ then the log-likelihood may be expressed as in Equation 2.20. That is,

$$\ell(t_1, t_2, ..., t_N|\theta) = \int_0^t \log\{\lambda(s|\theta)\}\ \mathrm{d}N(s) - \int_0^t \lambda(s|\theta)\ \mathrm{d}s, \tag{3.2}$$

where the full parameters are expressed as $\theta = (\theta_1, \theta_2, ..., \theta_r)$. Now, it can be noted that the general first order derivative, the gradient, for any $i \in \{1, ..., r\}$ is equivalent to

$$\frac{\partial \ell}{\partial \theta_i} = \int_0^t \left(\frac{\partial \lambda(s|\theta)}{\partial \theta_i} \Big/ \lambda(s|\theta) \right) \mathrm{d}N(s) - \int_0^t \frac{\partial \lambda(s|\theta)}{\partial \theta_i}\ \mathrm{d}s. \tag{3.3}$$

This was also further extended to find the general second order derivative which forms the basis of the Hessian matrix. That is, each element of the Hessian matrix may be expressed as

$$\boldsymbol{H}_{i,j} = \frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} = \int_0^t \left(\frac{\frac{\partial^2 \lambda(s|\theta)}{\partial \theta_i \partial \theta_j} \lambda(s|\theta) - \frac{\partial \lambda(s|\theta)}{\partial \theta_i} \frac{\partial \lambda(s|\theta)}{\partial \theta_j}}{(\lambda(s|\theta))^2} \right) \mathrm{d}N(s) - \int_0^t \frac{\partial^2 \lambda(s|\theta)}{\partial \theta_i \partial \theta_j}\ \mathrm{d}s, \tag{3.4}$$

for all $i, j \in \{1, ..., r\}$. The permutation of all is and js forms the complete Hessian matrix. The Hessian matrix can then be easily implemented in many maximisation algorithms and the determinant of the matrix can be calculated to determine whether the Hessian is non-singular.

3.3.3 Fisher Information Matrix

As discussed by Rothenberg, identifiability is closely linked to the singularity of the Fisher information matrix [Rothenberg, 1971]. This is practically more useful than comparing the Hessian due to its interpretability. In statistics, the Fisher information is a metric that measures the information that a random variable gives about an unknown parameter set θ of the distribution that models the random variable of interest. Asymptotic theory of maximum likelihood estimation often incorporates the Fisher information and in the specific case of Hawkes process, this will be fully discussed.

The similarity between the Hessian and the Fisher information matrix, which provides the information available about the parameters, can be seen by looking at their respective definitions. The Fisher information is generically defined as

$$I(\boldsymbol{\theta}) = \mathbb{E}\Bigg[- \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}^2} \Bigg],$$

and by looking at pairs of elements from $\boldsymbol{\theta}$, may be rewritten as

$$I_{i,j}(\theta) = \mathbb{E}\Bigg[- \boldsymbol{H}_{i,j} \Bigg].$$

The Fisher information matrix is necessary to solve many statistical problems. To the knowledge of this author, no closed-form solution exists for Hawkes processes. A tractable form of the Fisher information is therefore given in Section 5.3 alongside some empirical results. The Fisher information is vital to resolve the identifiability issue in Hawkes processes. In order to derive the Fisher information, first the full Hessian must be written tractably. In the next chapter, current literature discussing the Hessian matrix and identifiability for Hawkes processes will be reviewed.

Chapter 4

Statistical Theory of Hawkes Processes

The features and structure of Hawkes processes have been extensively researched by Hawkes, Ogata and Ozaki. The methodology they introduced forms the familiar aspects of the literature, however, in more recent years there have been new developments in aspects of the Hawkes process such as the moment generating function of the conditional intensity. As the literature available is broad, this chapter will focus only on the theory that has direct implications for the discussion of identifiability.

4.1 Asymptotic Theory of Hawkes Process

Before discussing the asymptotic behaviour of a Hawkes process, the conditional intensity function must be recalled. The notation provided in Equations 2.12a and 2.12b has been primarily used so far, though this is not the only accepted definition. More generally, consider an indexed stationary point process, $P(\cdot)$, such that $\omega = \{t_i; i = 0, \pm 1, \pm 2, ...\} \in \Omega$ and $... < t_{-1} < 0 \leq t_0 < t_1 < ....$ where for all i we have $t_i \in \mathbb{R}$ [Ogata, 1978]. In this setting, by considering $\mathcal{H}_{0,t}$, the history of the process, the intensity function may be expressed as

$$\lambda(t, \omega) = \lim_{\delta \to 0} \frac{1}{\delta} P[N([t, t+\delta)) > 0 | \mathcal{H}_{0,t}], \tag{4.1}$$

which is equivalent to $\mathbb{E}[\lambda(t, \omega) | \mathcal{H}_{0,t}]$. Similarly to the previous definition, it is common to drop ω from the intensity and instead write $\lambda(t, \omega)$ more simply as $\lambda(t)$. Note here, that by applying the law of total expectation, sometimes called the Tower property, to Equation 4.1, the results of the expectation of the intensity function given in Equations 2.14 and 2.27 may be recovered. Furthermore, this definition can also be used to form the general log-likelihood given in Equation 2.20. That is,

$$\ell(t|\theta) = \sum_{i=1}^{n} \log(\lambda(t_i)) - \int_0^t \lambda(s|\theta) \, \mathrm{d}s = \int_0^t \log\{\lambda(s|\theta)\} \, \mathrm{d}N(s) - \int_0^t \lambda(s|\theta) \, \mathrm{d}s,$$

where the observations are denoted $0 < t_1 < t_2 < ... < t_n \leq t$ and exist on the interval $[0, t]$. The asymptotic properties of the maximum likelihood estimator are based upon a log-likelihood conditional on the infinite past which, given a large enough interval, is equivalent to the exact log-likelihood.

These definitions can be extended to show that the maximum likelihood estimator is consistent, asymptotically normal, and efficient. It will be shown in Chapter 5, especially on large intervals, that the asymptotic properties have a direct implication on the identifiability of the Hawkes process. The asymptotic properties described by Ogata required a lengthy set of assumptions however, since 1977 further research has allowed some of these assumptions to be relaxed. The assumptions still necessary for the theory to hold are outlined in the next section.

4.1.1 Assumptions of Existing Asymptotic Theory

The assumptions made on the point processes, or more specifically the Hawkes processes, concern the behaviour the observations must obey and the regularity conditions for the parametric family of the intensity process denoted in Equation 4.1 [Ogata, 1978].

Observational assumptions are made about the data and therefore, are beyond the control of any model choice. Fortunately, in the case of the necessary asymptotic assumptions, they are relatively weak. Firstly, the point process must be stationary, ergodic and absolutely continuous on any finite interval [Ogata, 1978]. In the context of Hawkes processes, to ensure this assumption is met, the branching ratio must be less than 1. This is the stationary condition given in Equations 2.15 and 2.28. Further to this, the process must be indexed and orderly. Mathematically, this is expressed as

$$\lim_{\delta \to 0} \frac{1}{\delta} P[N([0, \delta)) \geq 2] = 0, \tag{4.2}$$

though more intuitively it may be considered that the counting process corresponding to the Hawkes process, must remain the same or increase by unit size 1 as the time moves from t to $t+1$. This is a slight restriction to the model, meaning even in the M-dimensional setting an event can only occur in one dimension at once. Despite events occurring instantaneously and only one event occurring at once, poorly chosen sampling rates can mask this behaviour.

Assumptions must also be made on the parameter space in which the maximum likelihood estimators exists. Firstly, the parameter space, θ, is a compact metric space. Further to this, the intensity process is predictable and continuous for all θ and at time 0 must be non-negative. Identifiability also requires the choice of an injective intensity process. It is not necessary for the process to be surjective though it reduces the complexity of the calculations at later steps if an appropriate bijective function is chosen. Another assumption made, is that all of the partial derivatives up to and including those of order three exist. This will be used in the following section to dictate constraints on the moments of the intensity function. Finally the most important assumption that identifiability is concerned with is the matrix $K(\theta)$, with parameter space $\theta = (\theta_1, ..., \theta_p)$, and can be described as

$$K(\theta) = \{K_{i,j}(\theta)\}_{i,j=1,\dots,p} \quad \text{with} \quad K_{i,j}(\theta) = \mathbb{E}\left[\Big(\frac{1}{\lambda(t)}\Big)\Big(\frac{\partial \lambda(t)}{\partial \theta_i}\Big)\Big(\frac{\partial \lambda(t)}{\partial \theta_j}\Big)\right]. \tag{4.3}$$

If $K(\theta)$ exists, then it is non-singular and each element has a finite second moment [Ogata, 1978].

The final set of conditions are used to note that the intensity function over the history $\mathcal{H}_{0,t}$ can be used as an approximation for the infinite intensity process, which is

$$\lambda_{-\infty}(t, \omega) = \lambda_{-\infty}(t) \coloneqq \lim_{\delta \to 0} \frac{1}{\delta} P[N([t, t+\delta)) > 0 | \mathcal{H}_{-\infty,t}]. \tag{4.4}$$

These assumptions are necessary to prove the consistency and asymptotic normality of the estimator, though, more importantly, they also allow for a fundamental understanding of the Hessian matrix.

Given some true parameter set, $\theta \in \Omega$, there exists a neighbourhood ϕ of θ such that the following properties hold:

$$\lim_{t \to \infty} \Big\{ \sup_{\theta' \in \phi} |\lambda_{-\infty,\theta'}(t) - \lambda_{\theta'}(t)| \Big\} \xrightarrow{\mathbb{P}} 0, \tag{4.5}$$

and $\sup_{\theta' \in \phi} |\log \lambda(t)|$ has at least finite second moment with a uniform bound. To add to this, for any $\theta \in \Omega$,

$$\frac{\lambda_{-\infty}(t)}{\lambda(t)}, \quad \frac{1}{\lambda(t)} \frac{\partial \lambda(t)}{\partial \theta_i} \frac{\partial \lambda(t)}{\partial \theta_j} \quad \text{and} \quad \frac{\partial^2 \lambda(t)}{\partial \theta_i \partial \theta_j},$$

all have at least finite second moments uniformly bounded. Moreover, it is also known that for any $\theta \in \Omega$,

$$\lim_{t\to\infty}\Big\{\lambda_{-\infty}(t)-\lambda(t)\Big\}\xrightarrow{\mathbb{P}}0,\ \ \lim_{t\to\infty}\Big\{\frac{\partial\lambda_{-\infty}(t)}{\partial\theta_i}-\frac{\partial\lambda(t)}{\partial\theta_i}\Big\}\xrightarrow{\mathbb{P}}0,\ \text{ and }\lim_{t\to\infty}\Big\{\frac{\partial^2\lambda_{-\infty}(t)}{\partial\theta_i\partial\theta_j}-\frac{\partial^2\lambda(t)}{\partial\theta_i\partial\theta_j}\Big\}\xrightarrow{\mathbb{P}}0,$$

for all $i, j = 1, 2, ..., p$. A similar definition holds for the third moment though is omitted for brevity. Finally, as $t \to \infty$ for the interval of events $[0, t]$

$$\mathbb{E}\Bigg[\frac{1}{\sqrt{t}}\int_0^t\Big|\frac{\partial\lambda_{-\infty}(s)}{\partial\theta_i}-\frac{\partial\lambda(s)}{\partial\theta_i}\Big|\,\mathrm{d}s\Bigg]\to 0, \tag{4.6a}$$

$$\mathbb{E}\Bigg[\frac{1}{\sqrt{t}}\int_0^t|\lambda_{-\infty}(s)-\lambda(s)|\Big|\frac{1}{\lambda(s)}\frac{\partial\lambda(s)}{\partial\theta_i}\Big|\,\mathrm{d}s\Bigg]\to 0. \tag{4.6b}$$

Now all the assumptions have been outlined, the necessary asymptotic properties to derive the identifiability conditions can be discussed.

4.1.2 Statistical Theory of the MLE and Hessian

In the following section, it will be assumed that the parameter space is Θ, the true parameters within the parameter space are denoted θ_0 and the maximum likelihood estimates are $\hat{\theta}$. In the simplest case, the true parameters of a one-dimensional Hawkes process would be $\theta_0 = (\mu_0, \alpha_0, \beta_0)$ however, the following properties hold for any self-exciting process. Under the assumptions given in the previous section, it can be stated that

$$\lim_{t\to\infty}\hat{\theta}\xrightarrow{\mathbb{P}}\theta_0. \tag{4.7}$$

That is, given an infinite positive interval, $[0, t]$, the maximum likelihood estimator will converge in probability to the true parameters. In Ogata's original paper some properties of the partial derivatives were found that prove useful in the discussion of identifiability. These are, for some process on the finite interval $[0, t]$, and using a model with parameters $\theta = (\theta_1, \theta_2, ..., \theta_p)$, then for all $i, j = 1, 2, ..., p$

$$\mathbb{E}\Bigg[\frac{\partial\ell(\theta)}{\partial\theta_i}\Bigg]_{\theta=\theta_0}=0, \tag{4.8}$$

$$\mathbb{E}\Bigg[\frac{\partial\ell(\theta)}{\partial\theta_i}\frac{\partial\ell(\theta)}{\partial\theta_j}\Bigg]_{\theta=\theta_0}=-\mathbb{E}\Bigg[\frac{\partial^2\ell(\theta)}{\partial\theta_i\partial\theta_j}\Bigg]_{\theta=\theta_0}=t\mathbb{E}\Bigg[\frac{1}{\lambda(t)}\frac{\partial\lambda(t)}{\partial\theta_i}\frac{\partial\lambda(t)}{\partial\theta_j}\Bigg]. \tag{4.9}$$

Combined with the full derivation of the Hessian matrix given in the next section, these properties can be used to derive the full Fisher information matrix. These equations are theoretically correct even in the M-dimensional case; however, it has never proven useful in obtaining a closed form Fisher information. This is due to the lack of the existence of a rigorous moment generating function for the conditional intensity, particularly one that may be used to derive negative moments. The moment generating function for the intensity will be discussed in Section 4.4.

A corollary from the proof of Equations 4.8 and 4.9, is that the Hessian matrix is asymptotically negative-definite in some neighbourhood ϕ of θ_0. This, in turn, directly yields a recognisable asymptotic property that is often of concern. Ogata showed

$$\frac{1}{\sqrt{t}}\Bigg(\frac{\partial\ell(\theta_0)}{\partial\theta}\Bigg)\to\mathrm{N}\Big(0, I(\theta_0)\Big), \tag{4.10}$$

as $t \to \infty$. This may be extended to a multivariate normal distribution when considering M-dimensional Hawkes processes. Here, $I(\theta_0)$ denotes the Fisher information of the true parameters. Finally, provided the maximum likelihood estimator, $\hat{\theta}$, satisfies $\partial\ell(\theta)/\partial\theta = 0$, then as $t \to \infty$,

$$\sqrt{t}(\hat{\theta}-\theta_0)\to\mathrm{N}\Big(0, I^{-1}(\theta_0)\Big), \tag{4.11}$$

$$2\{\ell(\hat{\theta})-\ell(\theta_0)\}\to\chi^2_p. \tag{4.12}$$

This concludes all the necessary properties to derive the conditions that lead to identifiability issues.

4.2 The Full One-Dimensional Hessian Matrix

After Hawkes' paper in the twentieth century, several other statisticians introduced new methodology in the field of Hawkes processes. For efficiency, the Hessian was derived so as to compute the maximum likelihood estimator more quickly. Much of this work was possible due to the full derivation of the one-dimensional Hawkes processes' Hessian matrix in 1977 [Ozaki, 1977]. The full Hessian matrix in this simplistic case can be expressed by the 3 × 3 matrix

$$\boldsymbol{H} = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \alpha} & \frac{\partial^2 \ell}{\partial \mu \partial \beta} \\ \frac{\partial^2 \ell}{\partial \alpha \partial \mu} & \frac{\partial^2 \ell}{\partial \alpha^2} & \frac{\partial^2 \ell}{\partial \alpha \partial \beta} \\ \frac{\partial^2 \ell}{\partial \beta \partial \mu} & \frac{\partial^2 \ell}{\partial \beta \partial \alpha} & \frac{\partial^2 \ell}{\partial \beta^2} \end{pmatrix}. \tag{4.13}$$

The parameters of interest are $\theta = (\mu, \alpha, \beta)$ as discussed in Section 2.3. Therefore, for the set of observations $t_1, t_2, ..., t_N$ on the interval $[0, t]$, the first order derivatives for the log-likelihood using the recursive formula presented in Equation 2.31 are the following:

$$\frac{\partial \ell}{\partial \mu} = \sum_{k=1}^{N} \left(\frac{1}{\mu + \alpha R(k)} \right) - t, \tag{4.14a}$$

$$\frac{\partial \ell}{\partial \alpha} = \sum_{k=1}^{N} \left(\frac{R(k)}{\mu + \alpha R(k)} \right) + \sum_{k=1}^{N} \frac{1}{\beta} (e^{-\beta(t-t_k)} - 1), \tag{4.14b}$$

$$\frac{\partial \ell}{\partial \beta} = \sum_{k=1}^{N} \left(\frac{\alpha R'(k)}{\mu + \alpha R(k)} \right) - \alpha \sum_{k=1}^{N} \left(\frac{1}{\beta} (t - t_k) e^{-\beta(t-t_k)} + \frac{1}{\beta^2} (1 - e^{-\beta(t-t_k)}) \right), \tag{4.14c}$$

where $R'(k) = -\sum_{t_i < t_k} (t_k - t_i) e^{-\beta(t_k - t_i)}$ for $k \geq 2$ and $R'(1) = 0$. These derivatives can then be used to derive the simplest Hessian [Ozaki, 1977]. The partial derivatives for the Hessian follow.

$$\frac{\partial^2 \ell}{\partial \alpha^2} = -\sum_{k=1}^{N} \left(\frac{R(k)}{\mu + \alpha R(k)} \right)^2, \tag{4.15}$$

$$\frac{\partial^2 \ell}{\partial \mu^2} = -\sum_{k=1}^{N} \frac{1}{(\mu + \alpha R(k))^2}, \tag{4.16}$$

$$\frac{\partial^2 \ell}{\partial \mu \partial \alpha} = -\sum_{k=1}^{N} \frac{R(k)}{(\mu + \alpha R(k))^2}, \tag{4.17}$$

$$\frac{\partial^2 \ell}{\partial \mu \partial \beta} = -\sum_{k=1}^{N} \frac{\alpha R'(k)}{(\mu + \alpha R(k))^2}, \tag{4.18}$$

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \alpha \partial \beta} = & \left(\sum_{k=1}^{N} -\frac{1}{\beta} (t - t_k) e^{-\beta(t-t_k)} + \frac{1}{\beta^2} (1 - e^{-\beta(t-t_k)}) \right) \\ & + \left(\sum_{k=1}^{N} \frac{R'(k)}{\mu + \alpha R(k)} - \frac{\alpha R(k) R'(k)}{(\mu + \alpha R(k))^2} \right), \end{aligned} \tag{4.19}$$

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \beta^2} = \alpha & \left(\sum_{k=1}^{N} \frac{1}{\beta} (t - t_k)^2 e^{-\beta(t-t_k)} + \frac{2}{\beta^2} (t - t_k) e^{-\beta(t-t_k)} + \frac{2}{\beta^3} (e^{-\beta(t-t_k)} - 1) \right) \\ & + \left(\sum_{k=1}^{N} \frac{\alpha R''(k)}{\mu + \alpha R(k)} - \left(\frac{\alpha R'(k)}{\mu + \alpha R(k)} \right)^2 \right). \end{aligned} \tag{4.20}$$

Note, that $R''(k) = \sum_{t_i < t_k} (t_k - t_i)^2 e^{-\beta(t_k - t_i)}$. It is obvious, due to the behaviour of partial derivatives, that this completes all the unique cases of the simple 3 × 3 Hessian. Further note, that for computational methods it is assumed that $t = t_N$; similar to when calculating the likelihood.

The extension to this derivation for the M-dimensional Hawkes process is discussed later as an extension to Ozaki's original work. Section 3.3.3 discussed the link between the Hessian and the Fisher information. The Fisher information has no tractable form in any of the literature. This thesis, in later chapters, explores empirical and analytical methods to find the Fisher information matrix.

4.3 The Score Function and Fisher Information

In the past year, there has been some development in the theory of the Fisher information for the Hawkes process [Wang et al., 2020]. The new literature provides an empirical approximation and closed-form solutions to several of the terms within the Fisher information. The methodology introduced relies on the Score function of the Hawkes process. In the M-dimensional setting, the full parameter set is defined as $\boldsymbol{\theta}$, while the parameters relating to a specific dimension are denoted $\boldsymbol{\theta}_m = (\theta_{1,m}, ..., \theta_{M,m})$. A specific parameter can be identified using the notation $\boldsymbol{\theta}_{i,j}$. The Score function for the m^{th} dimension, on the interval $[0, t]$, may be defined as

$$S_m(\boldsymbol{\theta}_m) = \frac{\partial \ell(\boldsymbol{\theta}_m)}{\partial \boldsymbol{\theta}_m} = -\int_0^t \frac{\partial \lambda_m(s)}{\partial \boldsymbol{\theta}_m}\,\mathrm{d}s + \int_0^t \frac{1}{\lambda_m(s)} \frac{\partial \lambda_m(s)}{\partial \boldsymbol{\theta}_m}\,\mathrm{d}N_m(s), \tag{4.21}$$

$$= \int_0^t \frac{1}{\lambda_m(s)} \frac{\partial \lambda_m(s)}{\partial \boldsymbol{\theta}_m}(\,\mathrm{d}N_m(s) - \lambda_m(s)\,\mathrm{d}s). \tag{4.22}$$

This allows each row of the M-dimensional Hessian matrix to be expressed as

$$\boldsymbol{H}_m(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta}_m)}{\partial \boldsymbol{\theta}_m \partial \boldsymbol{\theta}_m^\top} = -\int_0^t \frac{1}{\lambda_m^2(s)} \frac{\partial \lambda_m(s)}{\partial \boldsymbol{\theta}_m} \frac{\partial \lambda_m(s)}{\partial \boldsymbol{\theta}_m^\top}\,\mathrm{d}N_m(t). \tag{4.23}$$

Combining this result with Equation 4.9, it is possible to denote a row of the Fisher information matrix as

$$I(\boldsymbol{\theta}_m) = \mathbb{E}\left[\frac{1}{\lambda(t)} \frac{\partial \lambda_m(s)}{\partial \boldsymbol{\theta}_m} \frac{\partial \lambda_m(s)}{\partial \boldsymbol{\theta}_m^\top}\right].$$

The result for the Fisher information, in this form, has been known since the late twentieth century [Ogata, 1978, Wang et al., 2020]. Without a tractable formula for the first negative moment, it is not possible to have the exact Fisher information. However, it is possible to form a bound on each of the Fisher information matrix's elements from this definition. It is necessary to note, that since

$$\lambda(t) = \mu + \alpha \int_{-\infty}^t e^{-\beta(t-s)}\,\mathrm{d}N(s) \geq \alpha \int_{-\infty}^t e^{-\beta(t-s)}\,\mathrm{d}N(s),$$

it is possible to state that $1/\lambda(t) \leq 1/\mu$. A bound can then be placed on the Fisher information. That is,

$$I(\boldsymbol{\theta}_m) \leq \frac{1}{\mu_m} \mathbb{E}\left[\frac{\partial \lambda_m(s)}{\partial \boldsymbol{\theta}_m} \frac{\partial \lambda_m(s)}{\partial \boldsymbol{\theta}_m^\top}\right]. \tag{4.24}$$

It is also possible to generally express the expectation in Equation 4.28 as

$$\mathbb{E}\left[\frac{\partial \lambda_m(s)}{\partial \boldsymbol{\theta}_m} \frac{\partial \lambda_m(s)}{\partial \boldsymbol{\theta}_m^\top}\right] = \Lambda\Lambda^\top + \frac{1}{2}\boldsymbol{\beta}\Sigma + \frac{1}{4}\boldsymbol{\beta}A(\mathbb{I} - A)^{-1}\Sigma + \frac{1}{4}\boldsymbol{\beta}\Sigma A^\top(\mathbb{I} - A)^{-1}, \tag{4.25}$$

where $\mathbb{I}$ is the identity matrix, $A = \boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_M)$, $\Lambda = (\mathbb{I} - A)^{-1}\boldsymbol{\mu}$ and $\Sigma = \text{diag}(\Lambda)$.

Wang et al also showed that the Score function is asymptotically normal and the empirical Hessian converges to the Fisher information for large t [Wang et al., 2020]. From these properties, a non-asymptotic confidence set can then be constructed for the parameters.

These results, while useful, are not a complete solution due to an inability to solve the first negative moment. Chapter 5 looks directly to extend this work.

4.4 Moment Generating Functions

In Hawkes' original paper the theoretical properties introduced were the Bartlett spectrum and the corresponding covariance density function [Hawkes, 1971]. These proved useful in model selection, however this did not fully describe what is now known about the self-exciting processes. More recently, there has been much discussion on the moments of a Hawkes process [Daw and Pender, 2018]. The moment generating functions outlined here are given with respect to the intensity function used in Equation 2.13 however, can be easily extended to other kernel choices. This kernel is used so as to avoid intricate algebra that more complex kernels would produce. This section outlines the method for calculating the moments. More specifically, the general moments derived include: $\mathbb{E}[N^m(t)], \mathbb{E}[\lambda^k(t)]$ and $\text{Cov}[N(t), \lambda(t)]$ [Cui et al., 2020].

Recalling the definition of an orderly point process, $P(\cdot)$, given in Section 4.1, and noting that all stationary self-exciting point processes with a finite intensity function can be expressed as Poisson cluster processes, then counting properties of $P(\cdot)$ may be derived using its corresponding probability generating function [Oakes and Hawkes, 1974]. Equation 4.1 showed that $\lambda(t)$ is equivalent to $\mathbb{E}[\lambda(t)|\mathcal{H}_{0,t}]$ and as outlined by Vere-Jones [Daley and Vere-Jones, 1971], this can be used to show

$$\mathbb{E}[N([a,b))] = \int_a^b \lambda(s)\ \mathrm{d}s, \tag{4.26}$$

where in the Hawkes process setting, $\lambda(t)$ is expressed by Equation 2.12a. In the case of the exponential kernel, this gives the stationary condition, the branching ratio, but also notes that η is the integral that satisfies the inequality

$$0 < \int_0^\infty \nu(s)\ \mathrm{d}s < 1. \tag{4.27}$$

This extended definition directly implies that given a stationary Poisson cluster process exists, then the rate of the process is

$$\mathbb{E}[\lambda(t)] = \frac{\mu}{1-\eta}.$$

If the branching ratio is $\eta = \alpha/\beta$ then this is equivalent to Equation 2.14. Oakes and Hawkes in turn showed that when this condition and Equation 2.12a are satisfied then there exists precisely one stationary orderly point process with finite rate [Oakes and Hawkes, 1974]. Noting this and Equation 4.26 it can be shown that

$$\mathbb{E}[N(t)] = \frac{\mu t}{1-\eta} - \frac{\eta}{(1-\eta)^2}\frac{\mu}{\beta}(1 - e^{-\beta(1-\eta)t}). \tag{4.28}$$

A procedural method exists to recursively calculate the moments for both the intensity and counting process building upon Equations 2.14, 4.26 and 4.28. This is explored in the next section.

4.4.1 Derivation of Moments - An Elemental Approach

The simple procedure outlined in this section is fundamental in defining the Fisher information matrix. There are six steps to the process and they may be applied recursively to find any non-negative moment. Some exact examples will also be given.

Initialising Step - First an objective function, $y(t)$, must be set. This can be any combination of counting processes and intensities and is generally expressed as:

$$y(t) = \mathbb{E}[g(N(t), \lambda(t), t)] = \mathbb{E}[N^p(t)\lambda^q(t)],$$

where p and q can be any non-negative integers. Negative values will be discussed later in Section 5.1. For the following steps to hold, it is assumed that the partial derivatives of $g(\cdot)$ are all uniformly continuous.

Step 2 - The conditional probabilities must be calculated. These are based on those described in Equation 2.11 though explicitly, for a simple Hawkes process

$$\mathbb{P}(N(t+h)-N(t)=0|\mathcal{H}(t))=1-h\lambda(t)+o(h), \tag{4.29a}$$

$$\mathbb{P}(N(t+h)-N(t)=1|\mathcal{H}(t))=h\lambda(t)+o(h). \tag{4.29b}$$

Step 3 - It is necessary to find the intensity functions corresponding to the two events described in Equations 4.29a and 4.29b. To simplify the notation in the next steps, $N(t+h)-N(t)$ will be defined as $\phi(t+h)$, or just ϕ. For the first of the two outcomes, when $\{\phi=0\}$ the corresponding intensity is

$$\lambda_{\phi=0}(t+h)=\mu+\sum_{t_i\leq t}\alpha e^{-\beta(t+h-t_i)}=(1-\beta h)\lambda(t)+\beta\mu h+o(h), \tag{4.30}$$

which follows from Equation 2.13. In the second case, when $\{\phi=1\}$ the corresponding intensity is,

$$\lambda_{\phi=1}(t+h)=\mu+\left(\sum_{t_i\leq t}\alpha e^{-\beta(t+h-t_i)}\right)+\alpha e^{t+h-t_{N(t)+1}}=\lambda_{\phi=0}(t+h)+\alpha(1-\beta k)+o(k), \tag{4.31}$$

where k is some value that satisfies the inequality $0<k<h$.

Step 4 - Enough information is now available to calculate the conditional expectation of the target function, $g(\cdot)$. The conditional expectation to calculate is

$$\mathbb{E}[g(N(t+h),\lambda(t+h),t+h)|\mathcal{H}_t],$$

where again $\mathcal{H}_t$ is the filtration or history of the process up to time t. Again, for a Hawkes process,

$$\begin{aligned}\mathbb{E}[N^p(t+h)\lambda^q(t+h)|\mathcal{H}_t]=N^p(t)\lambda^q(t)+N^p(t)\sum_{i=0}^{q-1}\binom{q}{i}\Big(-\beta(\lambda(t)-\mu)h\Big)^{q-i}\lambda^i(t)-\\ N^p(t)\sum_{i=0}^{q}\binom{q}{i}\Big(-\beta(\lambda(t)-\mu)h\Big)^{q-i}\lambda^{i+1}(t)h+\sum_{j=0}^{p}\binom{p}{j}N^j(t)\sum_{i=0}^{q}\binom{q}{i}\alpha^{q-i}\lambda^{i+1}(t)h+o(h),\end{aligned} \tag{4.32}$$

Step 5 - By taking expectations on both sides of Equation 4.32 the final result can be achieved in the form of a differential equation by finding the limit as $h\to 0$. Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[N^p(t)\lambda^q(t)]=\lim_{h\to 0}\frac{\mathbb{E}[N^p(t+h)\lambda^q(t+h)-\mathbb{E}[N^p(t)\lambda^{q+1}(t)]}{h}, \tag{4.33}$$

which yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[N^p(t)\lambda^q(t)]=q\beta\mu\mathbb{E}[N^p(t)\lambda^{q-1}(t)]-q\beta\mathbb{E}[N^p(t)\lambda^q(t)]\\ +\sum_{j=0}^{p-1}\binom{p}{j}\mathbb{E}[N^j(t)\lambda^q(t)]+\sum_{j=0}^{p}\sum_{i=0}^{q-1}\binom{p}{j}\binom{q}{i}\alpha^{q-i}\mathbb{E}[N^j(t)\lambda^{i+1}(t)].\end{aligned} \tag{4.34}$$

This step relies on the Tower property to remove the conditionality introduced in step 4. It is also possible to calculate negative moments under this procedure; however, it is given in the original paper that $\sum_{i=0}^{z}[\cdot]=0$ for any value of z that satisfies $z<0$. This will be modified later to allow for negative integers.

Step 6 - The differential equation presented in Equation 4.34 can then be solved provided the boundary conditions are known. For most Hawkes processes it is acceptable to assume $\mathbb{E}[N^z(0)]=0$

and $\mathbb{E}[\lambda^z(0)] = \mu^z$ for all non-negative values of z.

This completes the full procedure outlined by Cui and Hawkes [Cui et al., 2020], and is necessary to derive the full Fisher information matrix.

4.4.2 Mean, Variance and Covariance of a Hawkes Process

The first moments of the counting process and intensity have had expressions since Hawkes' initial paper [Hawkes, 1971]. The method outlined in the previous section however, is given so as to ensure there is agreement with these original results.

There are several definitions and formulae that follow immediately from the general procedure given, though of most interest is when $p = 0$ or $q = 0$. By setting $q = 0$, it can be seen

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[N^p(t)] = \sum_{j=0}^{p-1}\binom{p}{j}\mathbb{E}[N^j(t)\lambda(t)], \tag{4.35}$$

for all values of $p \geq 1$. Furthermore, a similar differential equation can be formed for the moments of the intensity. This may be written as,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\lambda^q(t)] = q\beta\mu\mathbb{E}[\lambda^{q-1}(t)] - q\beta\mathbb{E}[\lambda^q(t)] + \sum_{i=0}^{q-1}\binom{q}{i}\alpha^{q-i}\mathbb{E}[\lambda^{i+1}(t)], \tag{4.36}$$

for all $q \geq 1$. Equations 4.35 and 4.36 can therefore be used to derive the expectations and variances. However, to calculate the covariance a further differential equation must be set up:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[N(t)\lambda(t)] = \beta\mu\mathbb{E}[N(t)] + \alpha\mathbb{E}[\lambda(t)] + (\alpha-\beta)\mathbb{E}[N(t)\lambda(t)] + \mathbb{E}[\lambda^2(t)]. \tag{4.37}$$

Equation 4.37 also allows, through recursion, all moments of the intensity process to be derived. It is then possible to write the first moments in a tractable form. The first moment of the intensity with expectation of the initial intensity of $\mathbb{E}[\lambda(0)] = \mu$, the base intensity, can be expressed as

$$\mathbb{E}[\lambda(t)] = \begin{cases} \frac{\beta\mu}{\beta-\alpha} - \frac{\alpha\mu}{\beta-\alpha}e^{(\alpha-\beta)t}, & \text{if } \alpha \neq \beta, \\ \mu + \mu\alpha t, & \text{if } \alpha = \beta. \end{cases} \tag{4.38}$$

It is not immediate how this formula relates back to Equation 2.14. However, by applying the stationary condition, $\alpha < \beta$, and setting the limit as $t \to \infty$ the result follows. Furthermore, recalling Equation 4.26, the first moment of the counting process is

$$\mathbb{E}[N(t)] = \begin{cases} \frac{\beta\mu t}{\beta-\alpha} + \frac{\alpha\mu}{(\beta-\alpha)^2}\left[e^{(\alpha-\beta)t} - 1\right], & \text{if } \alpha \neq \beta, \\ \mu t + \frac{1}{2}\mu\alpha t^2, & \text{if } \alpha = \beta. \end{cases} \tag{4.39}$$

Using the first moment of the intensity in Equation 4.39, it is then possible to derive the second moment. Let the initial conditions be $\mathbb{E}[\lambda^2(0)] = \mu^2$ and $\mathbb{E}[N^2(0)] = 0$. Then,

$$\mathbb{E}[\lambda^2(t)] = \begin{cases} \left[\mu^2 + (\alpha^2 + 2\beta\mu)\mu\left(\frac{\delta_1}{\beta-\alpha} - \frac{\delta_2}{\beta-\alpha}\right)\right]e^{2(\beta-\alpha)t}, & \text{if } \alpha \neq \beta, \\ \mu^2 + \mu(\alpha^2 + 2\beta\mu)\left(\frac{1}{2}t^2 + t\right), & \text{if } \alpha = \beta, \end{cases} \tag{4.40}$$

where

$$\delta_1 = \frac{\beta}{2(\beta-\alpha)} - \frac{\alpha}{3(\beta-\alpha)}, \qquad \text{and} \qquad \delta_2 = \frac{\beta e^{-2(\beta-\alpha)t}}{2(\beta-\alpha)} - \frac{\alpha e^{-3(\beta-\alpha)t}}{3(\beta-\alpha)}.$$

It is then possible to derive the variance of the intensity process. That is,

$$\mathrm{Var}[\lambda(t)] = \begin{cases} \left[\mu^2 + (\alpha^2 + 2\beta\mu)\mu\left(\frac{\delta_1-\delta_2}{\beta-\alpha}\right) - \frac{\alpha^2\mu^2}{(\beta-\alpha)^2}\right]e^{2(\beta-\alpha)t} + \frac{2\alpha\beta\mu^2}{(\beta-\alpha)^2}e^{(\alpha-\beta)t} - \frac{\beta^2\mu^2}{(\beta-\alpha)^2}, & \text{if } \alpha \neq \beta, \\ \mu\left[(\alpha^2 + 2\beta\mu)(\frac{1}{2}t^2 + t) - (2 + \alpha t)\mu\alpha t\right], & \text{if } \alpha = \beta. \end{cases}$$

This concludes the derivation for the intensity function. To derive the second moment and the variance of the counting process as well as the covariance between the counting process and the intensity, $\mathbb{E}[N(t)\lambda(t)]$ must be found. It can be found noting,

$$\mathbb{E}[N(t)\lambda(t)] = \begin{cases} e^{-(\beta-\alpha)t}\Big(\int_0^t \beta\mu\mathbb{E}[N(s)] + \mathbb{E}[\lambda(s)] + \mathbb{E}[\lambda^2(s)]\ \mathrm{d}s\Big), & \text{if } \alpha \neq \beta, \\ \int_0^t \beta\mu\mathbb{E}[N(s)] + \mathbb{E}[\lambda(s)] + \mathbb{E}[\lambda^2(s)]\ \mathrm{d}s, & \text{if } \alpha = \beta, \end{cases} \tag{4.41}$$

and then using the initial conditions,

$$\mathbb{E}[N^2(t)] = \int_0^t \mathbb{E}[\lambda(s)] + 2\mathbb{E}[N(s)\lambda(s)]\ \mathrm{d}s, \tag{4.42}$$

that these expectations are also tractable. From these equations, the derivation of the variance and covariance can be found. The general method to solve any moment for the counting process is to first find the moment of the same order of the intensity function, i.e. $\mathbb{E}[\lambda^p(t)]$, then find $\mathbb{E}[N(t)\lambda^{p-1}(t)]$ and then recursively solve until reaching the solution for $\mathbb{E}[N^{p-1}(t)\lambda(t)]$. From here, the final step always follows to give $\mathbb{E}[N^p(t)]$ for any arbitrary integer p.

By combining the theory outlined here and in the previous sections, it is possible to explore and understand how a Hawkes process is structured and where identifiability issues will arise.

Chapter 5

Identifiability of Hawkes Processes

The following chapter outlines some novel methodology for approaching identifiability issues in the Hawkes process setting. Some theoretical results will be derived in Sections 5.1, 5.2 and 5.3 that are then verified in the following chapter.

Identifiability issues arise in a wide variety of Hawkes processes; however, of particular interest in this section is when these issues occur in stationary Hawkes process. Figure 5.1 shows a one-dimensional Hawkes process simulated with parameters $\theta = (\mu, \alpha, \beta) = (2, 2, 2.2)$. The plot shows how the log-likelihood varies with respect to the α and β parameters.

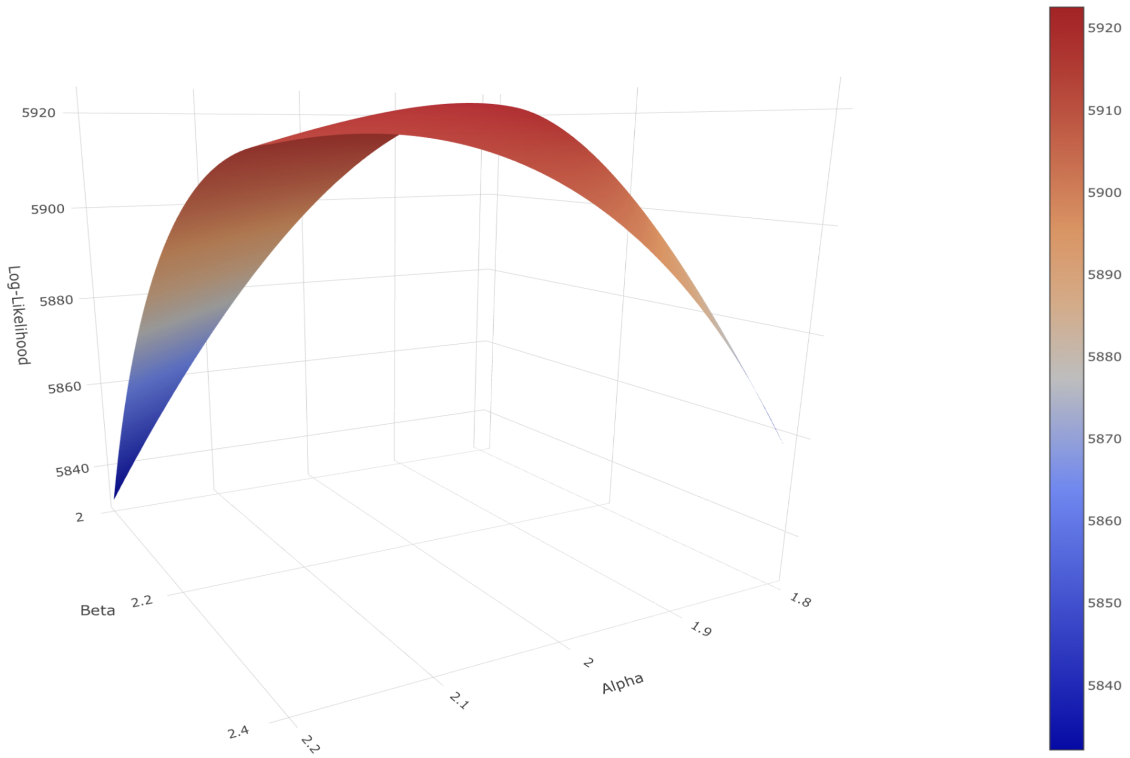

Figure 5.1: Log-Likelihood of a Non-Identifiable One-Dimensional Hawkes Process.

Figure 5.1 does not include the μ parameter in the plot as it has marginal effect on the overall shape of the log-likelihood surface. The exact shape, and the curvature of the surface, will be further discussed in Chapter 6. The likelihood, in this case, is not identifiable and the maximum likelihood is attained by multiple sets of parameter values. Intriguingly, these sets have a relationship to one another. This relationship is visible when viewing the diagonal ridge that forms along the maximum log-likelihood values. On closer inspection, all values that attain the maximum have approximately the same branching ratio. The effect that this ratio has is explored in later sections.

5.1 Negative Moments of the Intensity Function

Empirically, it can be seen that the branching ratio has a relationship with the determinant of the Fisher information matrix. However, when analysing the behaviour of the Fisher information as the moments vary, it can be seen that a close relationship exists. As discussed in Section 4.4, the non-negative moment generating function has been fully derived [Cui et al., 2020]. A closed-form expression for these moments is useful, not just in the context of identifiability but for Hawkes processes in general. Unfortunately, the original papers did not explore the negative moments [Daw and Pender, 2018, Cui et al., 2020], something essential when checking for identifiability in Hawkes processes. The exact reason for this is explored in Section 5.4.

It is acknowledged that $\mathbb{E}[1/\lambda(t)]$ is not necessarily equal to $1/\mathbb{E}[\lambda(t)]$ which follows from Jensen's inequality [Zabandan and Kılıçman, 2012]. The expectation of $\mathbb{E}[1/\lambda(t)]$ is the first negative moment of the intensity function, $\lambda(t)$. Practically, it is possible to use Jensen's inequality to find a bound for each negative moment. That is, given a random variable X and a convex function ϕ, the inequality states

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)], \tag{5.1}$$

while for a concave function ψ, the inequality becomes

$$\psi(\mathbb{E}[X]) \geq \mathbb{E}[\psi(X)], \tag{5.2}$$

where the difference between each side of the inequality, $|\phi(\mathbb{E}[X]) - \mathbb{E}[\phi(X)]|$, is known as the Jensen gap [Zabandan and Kılıçman, 2012]. As $\lambda(t)$ is a concave function [Cui et al., 2020], and it is possible to state that for any $n \in \mathbb{N}$,

$$\mathbb{E}[\lambda^{-n}(t)] \leq \frac{1}{\mathbb{E}[\lambda^{n}(t)]}. \tag{5.3}$$

Furthermore, it is also known for all $z \in \mathbb{Z}$ that $\mathbb{E}[\lambda^{z}(t)] \geq 0$ for stationary Hawkes processes. This provides a practical bound on the negative moments. Provided small branching ratios, the upper bound will be small and result in only a small interval of possible values for the negative moments. This is due to all non-negative moments including the divisor $(\beta - a)$ to some arbitrary order. Empirical methods can be employed to approximate the negative moments given a specific parameter set and data. This has proven useful while no alternatives have existed, however a small extension to Hawkes' moment generating procedure, given in Section 4.4.1 can achieve tractable solutions for these negative moments.

As opposed to Equation 4.36, which is the first-order differential equation for all positive moments, for any $q < 0$ it may be stated that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\lambda^{q+1}(t)] = q\beta\mathbb{E}[\lambda^{q+1}(t)] - q\beta\mu\mathbb{E}[\lambda^{q}(t)] + \sum_{i=q+1}^{0}\binom{-q}{-i}\alpha^{i-q}\mathbb{E}[\lambda^{i}(t)], \tag{5.4}$$

by noting that the original definition can be applied provided the summation is inverted. This equation may then be applied recursively to find all negative moments in a similar step-by-step process given in Section 4.4.1. That is, to find the n^{th} negative moment, the first must be found to calculate the second and so forth. Similarly, the first-order differential equation for the negative moments of the counting process may be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[N^{p}(t)] = \sum_{j=p+1}^{0}\binom{-p}{-j}\mathbb{E}[N^{j}(t)\lambda(t)], \tag{5.5}$$

for $p < 0$. These two equations alongside Equations 4.35 and 4.36 make it possible for all moments to be derived. The first negative moment will be derived as an example.

5.1.1 The First Negative Moment of the Intensity

The derivation given here assumes the Hawkes process is stationary for conciseness. As the algorithm is recursive, the negative moment that must be derived first is $\mathbb{E}[\lambda^{-1}(t)]$. It is known that $\mathbb{E}[\lambda^{0}(t)] = \mathbb{E}[1] = 1$ and is the necessary initialising moment to derive $\mathbb{E}[\lambda^{-1}(t)]$. Using Equation 5.4,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\lambda^0(t)] = (-1)\beta\mathbb{E}[\lambda^0(t)] - (-1)\beta\mu\mathbb{E}[\lambda^{-1}(t)] + \sum_{i=(-1)+1}^{0}\binom{-(-1)}{-i}\alpha^{i-(-1)}\mathbb{E}[\lambda^i(t)],$$

and therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\lambda^0(t)] = -\beta\mathbb{E}[\lambda^0(t)] + \beta\mu\mathbb{E}[\lambda^{-1}(t)] + \binom{1}{0}\alpha\mathbb{E}[\lambda^0(t)].$$

Rearranging,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\lambda^0(t)] + (\beta-\alpha)\mathbb{E}[\lambda^0(t)] = \beta\mu\mathbb{E}[\lambda^{-1}(t)],$$

and as $\mathbb{E}[\lambda^0(t)] = 1$, the above equation may be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\lambda^0(t)] + (\beta-\alpha) = \beta\mu\mathbb{E}[\lambda^{-1}(t)].$$

Furthermore, as $\mathbb{E}[\lambda^0(t)]$ is a constant value, its derivative is 0 and this yields the solution for the first negative moment as

$$\mathbb{E}[\lambda^{-1}(t)] = \frac{\beta-\alpha}{\beta\mu}. \tag{5.6}$$

On closer inspection, it is possible to see that $\mathbb{E}[\lambda^{-1}(t)] = (\mathbb{E}[\lambda(t)])^{-1}$. Therefore, the first negative moment attains the maximum bound of the interval on which it may exist, given by Jensen's inequality. This is due to the formulae in Equation 4.36 and 5.4 only differing by their symmetry around 0. This means the two formulae are the inverse of one another.

5.1.2 The Relationship between the Positive and Negative Moments

It is possible to derive all negative moments however, the algebra quickly becomes complex. The exact calculation of further negative moments can therefore be omitted and instead the values can be easily found through simulation. Table 5.1 shows the Monte Carlo estimations for the first, second and third negative moments in addition to the inverse of the first three positive moments. Note, the parameter space is $\theta = (\mu, \alpha, \beta)$ and η is the branching ratio.

Parameters (θ)	η	$\mathbb{E}[\lambda^{-1}(t)]$	$(\mathbb{E}[\lambda(t)])^{-1}$	$\mathbb{E}[\lambda^{-2}(t)]$	$(\mathbb{E}[\lambda^2(t)])^{-1}$	$\mathbb{E}[\lambda^{-3}(t)]$	$(\mathbb{E}[\lambda^3(t)])^{-1}$
$(5.50, 0.10, 11.16)$	0.00896	0.180	0.180	0.0325	0.0325	0.00585	0.00585
$(1.00, 0.10, 2.31)$	0.0433	0.956	0.956	0.910	0.909	0.867	0.869
$(1.00, 5.05, 11.16)$	0.453	0.548	0.549	0.535	0.539	0.495	0.531
$(5.50, 10.00, 20.00)$	0.500	0.0901	0.0903	0.01343	0.01521	0.00198	0.00477
$(10.00, 5.05, 8.94)$	0.565	0.0437	0.0436	0.00266	0.00314	0.000161	0.000233
$(10.00, 5.05, 6.73)$	0.750.	0.0253	0.0254	0.00107	0.00178	0.0000457	0.0000731
$(10.00, 10.00, 11.16)$	0.896	0.0104	0.0106	0.000482	0.000693	0.0000213	0.0000506

Table 5.1: Monte Carlo Estimation of Moments.

The values of (μ, α, β) were arbitrarily chosen to generate the results in Table 5.1. Several branching ratios have been provided across the stationary range. It is possible to see that the results for the first negative moment and the inverse of the first moment are equivalent. Moreover, it is possible to show mathematically that

$$\lim_{\eta \to 0} \mathbb{E}[\lambda^{-q}(t)] = \frac{1}{\mathbb{E}[\lambda^{q}(t)]},$$

for all values of q. Further to this, since all moments of the intensity are non-negative by definition of the intensity function [Cui et al., 2020], and each moment's size relates to its neighbours', it is possible to find the interval of values that all moments must be within. Provided the Hawkes process is stationary and $\mu \geq 1$ then,

$$0 \leq \mathbb{E}[\lambda^{-q}(t)] \leq ... \leq \mathbb{E}[\lambda^{-2}(t)] \leq \mathbb{E}[\lambda^{-1}(t)] \leq \mathbb{E}[\lambda^{0}(t)] = 1, \tag{5.7a}$$

$$\mathbb{E}[\lambda(t)] \geq \mathbb{E}[\lambda^{2}(t)] \geq ... \geq \mathbb{E}[\lambda^{q}(t)], \tag{5.7b}$$

and therefore, for a stationary Hawkes process all the negatives moments must exist on the interval $[0, 1]$, similar to the branching ratio. While this may be utilised to derive the Fisher information, first the full Hessian matrix must be derived.

5.2 The Full Hessian Matrix

Before identifiability can be fully explored, the Hessian matrix of a Hawkes process must be generalised for an M-dimensional process with an exponential kernel of order P. In this setting, the parameter set may be expressed as

$$\boldsymbol{\theta} = \left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_M \end{pmatrix}^T, \left\{ \begin{pmatrix} \alpha_{1,j,1} & \alpha_{1,j,2} & \cdots & \alpha_{1,k,M} \\ \alpha_{2,j,1} & \alpha_{2,j,2} & \cdots & \alpha_{2,k,M} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{M,j,1} & \alpha_{M,j,2} & \cdots & \alpha_{M,k,M} \end{pmatrix} \right\}_{j=1}^{P}, \left\{ \begin{pmatrix} \beta_{1,j,1} & \beta_{1,j,2} & \cdots & \beta_{1,k,M} \\ \beta_{2,j,1} & \beta_{2,j,2} & \cdots & \beta_{2,k,M} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{M,j,1} & \beta_{M,j,2} & \cdots & \beta_{M,k,M} \end{pmatrix} \right\}_{j=1}^{P} \right).$$

The full extension of this derivation for the Hessian with a P^{th} order kernel can be found in Appendix A.3 alongside the derivatives. This derivation is provided for completeness and may be used for any M-dimensional empirical calculations of the Fisher information.

Here, a brief discussion will highlight the main differences between the one-dimensional and M-dimensional Hessian matrices. In the one-dimensional $P = 1$ case the matrix is 3×3 however, by two dimensions the size increases to 10×10 and continues to increase with each additional dimension. The general formula for the square dimension of the matrix is $M + 2M^2$. While the matrix increases in size quadratically, the calculation does not. This is due to the sparsity of large Hessian matrices. The second order derivatives are equal to zero when there is no cross-behaviour between parameters. Furthermore, the calculation of the diagonal elements remains the same whether in one dimension or many. That is, since the diagonal elements only consider behaviour within their own dimension, the derivatives may be treated as M one-dimensional processes.

It is possible to calculate the determinant of the Hessian matrix for any M-dimensional process. However, it is necessary for any $M > 2$ that the Hessian matrix is row reduced before attempting to calculate the determinant.

As is the case for the one-dimensional process, the first and second derivatives of the recursive algorithm must exist for the computational complexity of a single element to stay at $\mathcal{O}(n)$. The use of the recursive algorithm is necessary for empirical calculations and proves useful when deriving the Fisher information.

5.3 The Fisher Information Matrix

The previous sections outlined the moment generating functions and the Hessian matrix necessary to derive the full Fisher information matrix. The elements of the matrix may be derived by first recalling Equation 4.9.

$$I(\boldsymbol{\theta}_{i,j}) = \mathbb{E}\Bigg[-\frac{\partial^2 \ell(\theta)}{\partial\theta_i \partial\theta_j}\Bigg]_{\theta=\theta_0} = t\Bigg(\mathbb{E}\Bigg[\frac{1}{\lambda(t)}\frac{\partial\lambda(t)}{\partial\theta_i}\frac{\partial\lambda(t)}{\partial\theta_j}\Bigg]\Bigg).$$

This definition always holds. However, due to the variability of Hawkes processes that can be generated from extreme parameter sets, it is best always to use a t of at least 100 units. Extreme parameter sets are those with branching ratios close to 1 or with a large expected intensity. It is advisable to consider a sampling window that allows for these conditions to be met, though is not necessary for the results to hold. When too few events are provided, it is not possible for the dataset to exhibit its full structure. In these settings, a single unique parameter set is unlikely and therefore, identifiability issues are immediate regardless of the true underlying structure.

5.3.1 One-Dimensional Fisher Information Matrix

Similarly to the one-dimensional Hessian, the one-dimensional Fisher information matrix is 3 × 3. Therefore, as the matrix is symmetric, there are 6 unique cases. By utilising Equation 4.9, the Fisher information can be denoted

$$I(\boldsymbol{\theta}) = \begin{pmatrix} \frac{(\beta-\alpha)t}{\beta\mu} & \frac{t}{\beta} & -\frac{\alpha t}{\beta^2} \\ \frac{t}{\beta} & \frac{\mu t}{\beta(\beta-\alpha)} + \frac{t}{2(\beta+\alpha)} & -\frac{\alpha\mu t}{\beta^2(\beta-\alpha)} \\ -\frac{\alpha t}{\beta^2} & -\frac{\alpha\mu t}{\beta^2(\beta-\alpha)} & \frac{\alpha^2\mu t}{\beta^3(\beta-\alpha)} \end{pmatrix}. \tag{5.8}$$

The full derivation of the process can be found in Appendix A.4. This matrix then gives a determinant of

$$\det(I(\boldsymbol{\theta})) = \frac{\alpha^2 t^3}{\beta^6}\Bigg(\frac{\beta^5 + (2\mu-\alpha)\beta^4 + (2\alpha\mu - 4\mu - 1)\beta^3 + (2+\alpha-4\alpha\mu)\beta^2 - 2\alpha^2}{2(\alpha^2-\beta^2)}\Bigg). \tag{5.9}$$

These calculations require the closed-form of $\mathbb{E}[\lambda^{-1}(t)]$, the Hessian and $\mathbb{E}[\lambda(t)]$. However, there is also a general formula of the Fisher information available in the Appendix. By rearranging the determinant it is possible to show that the value must always be non-negative. This is practically useful as the identifiability condition need only discuss the interval $[0,\infty)$.

The derivation of the Fisher information provides useful insight especially when considered alongside the branching ratio or intensity. The determinant of the Fisher information is more sensitive to the value that β takes. As μ behaves as an additive term, it is unsurprising that it can be shown it has the least influence in the value the determinant takes.

Furthermore, by recalling Equation 2.14, it is possible to rearrange the determinant into the form

$$\det(I(\boldsymbol{\theta})) = t^3\mathbb{E}[\lambda(t)]\Bigg(\frac{\alpha^2}{2\mu\beta^5}(\beta^2+2\mu-1) - \frac{\alpha^2}{2\mu\beta^4(\beta+\alpha)}(\beta^2 - \mu(\alpha+\beta) - 1) + \frac{\alpha^3}{\beta^6}\Bigg). \tag{5.10}$$

A linear trend between the determinant of the Fisher information and the intensity would be expected given that $\mathbb{E}[\lambda(t)]$ is multiplying each term. It is notable that the right hand term must always be positive, provided that μ, α and β are also positive. Therefore, it is only possible for the determinant of the Fisher information matrix to be zero when $\mathbb{E}[\lambda(t)] = 0$.

5.3.2 M-Dimensional Fisher Information Matrix

As discussed in Section 5.2, the Fisher information matrix, for any M, will be of size $M+2M^2$. The size of the matrix therefore increases in $\mathcal{O}(M^2)$ with each additional dimension. However, similarly to the Hessian matrix, there is symmetric behaviour reducing the number of unique elements to 15 when $P=1$ and 30 for any $P>1$ where P is the order of the kernel. For the simple exponential kernel, when $P=1$, the full M-dimensional Fisher information has been provided in Appendix A.5.

It is therefore possible to calculate the determinant of the Fisher information for any M-dimensional Hawkes process. This is primarily a computational task due to the number of derivatives that need to be calculated. However, an interesting feature arises from calculating any M-dimensional determinant: for any number of dimensions the determinant may always be expressed as

$$\det(I(\boldsymbol{\theta})) = t^{M+2M^2}\left(\gamma\sum_{i=1}^{M}\mathbb{E}[\lambda_i(t)] + \delta\right). \tag{5.11}$$

Here, δ is some constant value that incorporates the remaining parameters. Note, δ can take any real value and is often negative. Initially it may seem that it would be logical to assume that as the dimensionality increases the determinant of the Fisher information explodes. However, this is not the case. As M increases, the first term, that incorporates t, begins to increase while δ decreases. The parameter γ cannot be equal to zero for any M. Deriving closed-form solutions for γ and δ is possible, though it is more efficient numerically to find the determinant of the Fisher information by substituting in the true parameter values and computing this.

5.3.2.1 2-Dimensional Fisher Information Matrix

The 2-dimensional, or bivariate, Hawkes process' Fisher information can easily be derived from Appendix A.5. In this setting, the Fisher information is a 10×10 matrix. The full set of elements of this matrix will be omitted; however, the determinant will be provided. Note, the parameters for the 2-dimensional Hawkes process are

$$\boldsymbol{\theta} = \left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}^T, \begin{pmatrix}\alpha_{1,1} & \alpha_{1,2}\\ \alpha_{2,1} & \alpha_{2,2}\end{pmatrix}, \begin{pmatrix}\beta_{1,1} & \beta_{1,2}\\ \beta_{2,1} & \beta_{2,2}\end{pmatrix}\right).$$

Due to the size of the matrix, calculating the determinant is computationally expensive. Row reduction can be used to transform the Fisher information into a row reduced echelon form upper triangular matrix. It is then possible to find the determinant by taking the product of the diagonal elements of the upper triangular matrix. Under this procedure, the 2-dimensional Fisher information determinant can be yielded. In its simplest form, the determinant may be expressed as

$$\begin{aligned}\det(I(\boldsymbol{\theta})) = t^{10}\Bigg(&\frac{\beta_{1,1}^7\beta_{1,2}^5\beta_{2,1}^5\beta_{2,2}^7}{\alpha_{1,1}^2\alpha_{1,2}\alpha_{2,1}\alpha_{2,2}^2(\alpha_{1,1}\beta_{1,1}+\alpha_{1,2}\beta_{1,2}+\alpha_{2,1}\beta_{2,1}+\alpha_{2,2}\beta_{2,2})}\sum_{i=1}^{2}\mathbb{E}[\lambda_i(t)]\\ &-\frac{\alpha_{1,1}\beta_{1,2}+\alpha_{1,2}\beta_{1,1}+\alpha_{2,1}\beta_{2,2}+\alpha_{2,2}\beta_{2,1}}{\alpha_{1,1}\beta_{1,1}+\alpha_{1,2}\beta_{1,2}+\alpha_{2,1}\beta_{2,1}+\alpha_{2,2}\beta_{2,2}}\Bigg).\end{aligned} \tag{5.12}$$

The 2-dimensional example has been given in a formula that corresponds to Equation 5.11; however, as closed-forms for the expectation of the intensity exist, the determinant can be expressed more simply. Further to this, as both the δ and γ terms share a divisor, rearranging this to form an identifiable condition is simple. The next section provides inequalities that must be satisfied for identifiability to hold.

5.4 Identifiability and the Fisher Information

As can be easily seen in the one-dimensional case, the Fisher information's determinant will always be non-negative. That is,

$$\det(I(\boldsymbol{\theta})) \geq 0.$$

Regardless of the dimensionality of the process this remains true due to the Fisher information being a square symmetric matrix. This property may be utilised to derive a clear identifiability condition. First, note that another alternative form of the determinant may be expressed in the one-dimensional setting as

$$\det(I(\boldsymbol{\theta})) = \frac{\alpha^2 t^3}{\beta^6}\left(\frac{\beta(-2+\beta-\beta^3)-2\alpha}{2(\alpha+\beta)} + (\beta^2-2\beta)\mathbb{E}[\lambda(t)]\right). \tag{5.13}$$

Utilising the determinant's non-negative property it is immediate that

$$\frac{\alpha^2 t^3}{\beta^6}\left(\frac{\beta(-2+\beta-\beta^3)-2\alpha}{2(\alpha+\beta)} + (\beta^2-2\beta)\mathbb{E}[\lambda(t)]\right) \geq 0. \tag{5.14}$$

By rearranging, the one-dimensional identifiability condition is yielded as,

$$\mathbb{E}[\lambda(t)] \geq \frac{2\alpha+\beta(\beta^2-\beta+2)}{2(\alpha+\beta)(\beta^2-2\beta)}. \tag{5.15}$$

If the Hawkes processes is non-stationary then the expectation of the intensity, given in Equation 4.39 must satisfy the inequality for all t. However, for a stationary Hawkes process it is possible to extend the definition and check the 3 parameters satisfy

$$\mu \geq \frac{(2\alpha+\beta(\beta^2-\beta+2))(\beta-\alpha)}{2\beta(\alpha+\beta)(\beta^2-2\beta)}. \tag{5.16}$$

In the M-dimensional setting it is possible to again state that the Fisher information's determinant must be non-negative. Therefore, simply note that for the process to be identifiable

$$\sum_{i=1}^{M}\mathbb{E}[\lambda_i(t)] \geq -\frac{\delta}{\gamma}. \tag{5.17}$$

Note, that both γ and δ are specific values for a given M and these may be derived using Appendix A.5. It is possible for -δ/γ to be negative. In these cases, the choice of α and β leads to an identifiable process regardless of the choice of a non-negative μ or intensity.

In the case of the 2-dimensional Hawkes process, the identifiability condition may be expressed as

$$\sum_{i=1}^{2}\mathbb{E}[\lambda_i(t)] \geq \frac{\alpha_{1,1}^2\alpha_{1,2}\alpha_{2,1}\alpha_{2,2}^2(\alpha_{1,1}\beta_{1,2}+\alpha_{1,2}\beta_{1,1}+\alpha_{2,1}\beta_{2,2}+\alpha_{2,2}\beta_{2,1})}{\beta_{1,1}^7\beta_{1,2}^5\beta_{2,1}^5\beta_{2,2}^7}. \tag{5.18}$$

Chapter 6

Analysis

The methods discussed in Chapter 5 will now be employed to check for identifiability using simulated data. Simulated data must be used to ascertain the effectiveness of the theoretical results. The data is simulated using Algorithm 2, provided in Appendix A.6, and the identifiability of the process is highlighted in advance. Both identifiable and non-identifiable processes have bee simulated using Algorithm 2.

There are many methods for simulating Hawkes processes including Ogata's modified thinning algorithm [Ogata, 1981] though most have restrictions that make them less desirable in the identifiability setting. Therefore, the simulation method employed is one that requires only weak assumptions and allows for asymmetric parameter matrices [Lim et al., 2016]. The work presented used a generalised stationary condition; however, the results still hold under the stationary conditions given in Chapter 2. The pseudo-code has been provided in Algorithm 2 which modifies the notation from the original paper to keep in-line with the notation used in this dissertation. Note, the initial jumps are denoted $\alpha_{i,m}(0)$ and $\alpha_{i,m}$ denotes the subsequent distribution of the jumps, $N_m(0)$ denotes the initial count of the m^{th} process and $t_{\max}$ is the horizon, the time up to which events should be simulated. In most circumstances, the distribution of subsequent jumps is set to be equal to $\alpha_{i,m}(0)$ such that all jump sizes for a given process are constant. The output of the simulation algorithm returns the event times in all processes, r, the event times corresponding to the m^{th} process, t_m, the counting process corresponding to the m^{th} process, N_m, and finally, the intensity processes corresponding to the m^{th} process, λ_m.

Unlike many previous methods, this novel method put forward by Lim et al produces an exact simulation by employing the superposition property of point processes [Daley and Vere-Jones, 2003]. This extension allows for a wider set of parameter choices and will be useful when comparing the theoretical results to practical examples of non-identifiable processes.

6.1 Simulated Results

It is possible to find empirical estimates for all the theory derived in Chapter 5 through Algorithm 2. All simulations made in this section, assume a constant background intensity, jump size and decay rate.

Before simulating Hawkes processes, it is necessary to verify the derivation of the moments are correct. The Fisher information's derivation relies intrinsically on $\mathbb{E}[\lambda^{-1}(t)]$ being equivalent to $1/\mathbb{E}[\lambda(t)]$ as discussed in Section 5.1.1. Algorithm 2 outputs the empirical intensity for a given parameter set. Using these outputs, the equality $1/\mathbb{E}[\lambda(t)] = \mathbb{E}[\lambda^{-1}(t)]$ may be verified empirically. Provided the equality holds, the determinant of the full Fisher information matrix can be simplified to Equation 5.10.

6.1.1 Negative Moments of the Conditional Intensity

Figure 6.1 plots the relationship between the empirical estimate of $1/\mathbb{E}[\lambda(t)]$ against $\mathbb{E}[\lambda^{-1}(t)]$. It is reassuring that the empirical estimate confirms the theoretical result derived in Section 5.1.1. This would be expected; however, due to its importance in all simulated results, it was necessary to verify.

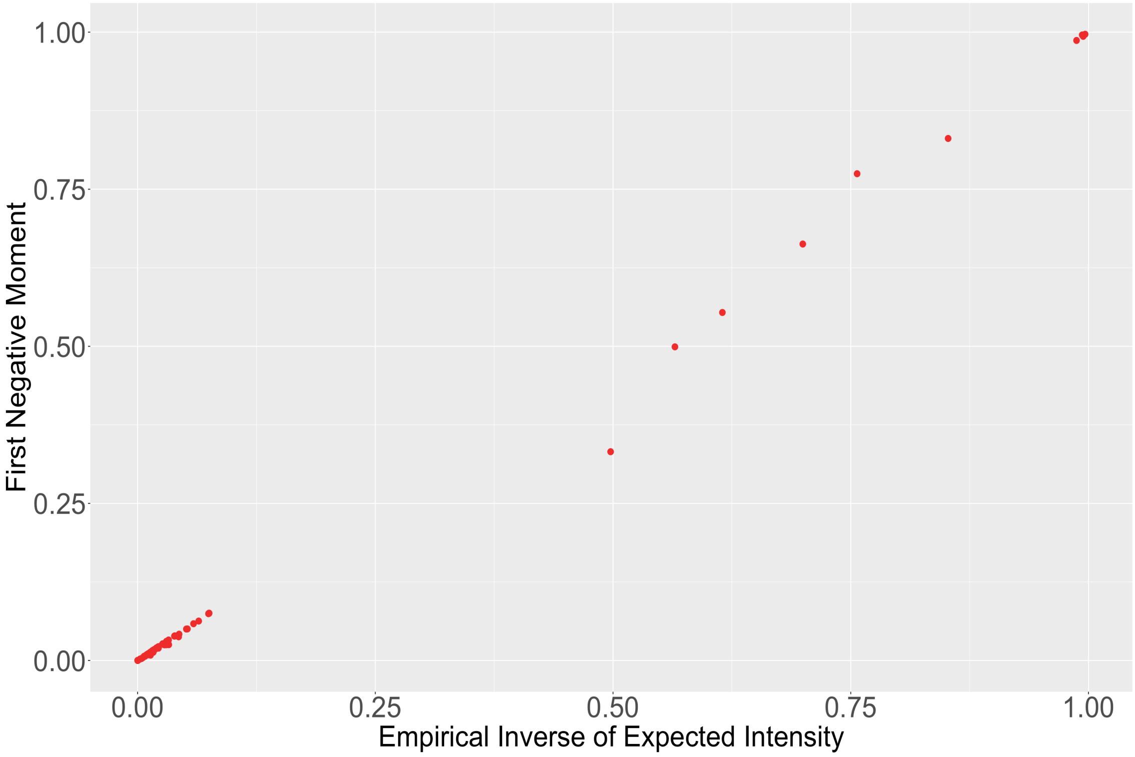

Figure 6.1: Empirical Estimate of $1/\mathbb{E}[\lambda(t)]$ against $\mathbb{E}[\lambda^{-1}(t)]$.

While there is a linear trend, there is some small variations from the expected result. The empirical estimates are found through Monte Carlo estimation. The results were generated over $10,000$ runs for all intensities; however, it can be seen that as the intensity approaches zero, or alternatively as the inverse of the intensity tends to one, more runs are necessary to reach convergence.

6.2 The Fisher Information Matrix in Practice

In this section, the Fisher information matrix will be compared to empirical results. It is possible to generate the true values from the theory outlined in Section 5.2. To compare the true results however, first an empirical method for finding the Fisher information must be derived.

6.2.1 Empirical Method for finding the Fisher Information Matrix

This thesis provides a closed-form for the Fisher information that current research has been unable to identify. A closed-form is presented in Section 5.3; however, it is also possible to utilise Monte Carlo estimation to derive the elements within the matrix and the determinant. The steps to implement this are given in Algorithm 1. The algorithm presents the method to calculate the Fisher information for a given parameter set, though in practice is repeated for many different parameters.

Note, it is possible to have a more efficient Monte Carlo estimation procedure by recalling that the Fisher information matrix is symmetric. Furthermore, it is possible to see the empirical determinant is equal to the determinant of the empirical Fisher information matrix for large enough N.

Algorithm 1 Monte Carlo Estimation of Fisher Information.

```
procedure GENERATE FISHER INFORMATION MATRIX
INPUT: Parameter set, $\theta = (\mu, \alpha, \beta)$, event horizon, $t$ and number of runs, $N$.
Initialise Generate an empty matrix, $\boldsymbol{G}$, for all parameter set combinations of size $M \times M$
            and empty vector $V$.
    for $i = 1, ..., N$
        Simulate a Hawkes process for given parameters $\theta$ using Algorithm 2.
        Generate empty Hessian matrix, $\boldsymbol{H}$, of size $M \times M$.
        for $j = 1, ..., M$
            for $k = 1, ..., M$
                Calculate $\boldsymbol{H}_{j,k}$ from the derivatives given in Appendix A.3.
                $\boldsymbol{G}_{j,k} = \boldsymbol{G}_{j,k} + \boldsymbol{H}_{j,k}$.
            END for
        END for
        $V_i = \det(\boldsymbol{H}_{j,k})$.
    END for
    Calculate the empirical Fisher information for all elements: $\boldsymbol{F}_{j,k} = \boldsymbol{G}_{j,k}/N$.
    Calculate empirical determinant: $D = \sum_{i=1}^{N} V_i/N$.
OUTPUT: $\boldsymbol{F}$ and $D$.
end procedure
```

6.2.2 The Elements of the Fisher Information

Given the closed-form of the Fisher information, it is possible to compare the true values of the elements to their empirical estimates. Algorithm 1 has been employed to find the estimates of each element. The Fisher information can be calculated provided the true parameter values are given.

Table 6.1 presents the error between the empirical estimates and the true values of the six unique elements in the one-dimensional Fisher information. These results have been generated by implementing Algorithm 1 with 10,000 different parameter sets. Furthermore, the simulation of each empirical Fisher information matrix used 10,000 distinct runs. The average errors were then calculated by taking the sum of absolute value of the differences between the true element value and the estimate, and divided by the total number of sets. More formally, let $\boldsymbol{F}$ be the true Fisher information and $\boldsymbol{G}$ be the empirical Fisher. The method employed finds

$$\bar{x}_{i,j} = \frac{1}{N} \sum_{k=1}^{N} |\boldsymbol{F}_{i,j} - \boldsymbol{G}_{i,j}|, \tag{6.1}$$

where N is the total number of runs and is the matrix $\boldsymbol{x}$ of the absolute average error. It is then possible to generate the percentage error of each element in the Fisher information.

Element	$\mathbb{E}\Big[-\frac{\partial^2\ell}{\partial\mu^2}\Big]$	$\mathbb{E}\Big[-\frac{\partial^2\ell}{\partial\alpha^2}\Big]$	$\mathbb{E}\Big[-\frac{\partial^2\ell}{\partial\beta^2}\Big]$	$\mathbb{E}\Big[-\frac{\partial^2\ell}{\partial\mu\partial\alpha}\Big]$	$\mathbb{E}\Big[-\frac{\partial^2\ell}{\partial\mu\partial\beta}\Big]$	$\mathbb{E}\Big[-\frac{\partial^2\ell}{\partial\alpha\partial\beta}\Big]$
Average Error (%)	0.1197	0.7308	1.3748	0.4259	0.8182	0.8563

Table 6.1: Error between Theoretical and Simulated Results.

The results in Table 6.1 suggest agreement between the empirical estimates and the theoretical results. It is notable that the third column, corresponding to the second derivative with respect to β, has the greatest error. This is not surprising if the relative size of the elements are considered. This element is of the scale 10^{-5} and is calculated using the formula given in Equation 5.8. The empirical estimate is sensitive to the number of runs used. By increasing the number of runs to $100,000$ and then repeating the process given in Equation 6.1, the average error drastically decreases. This procedure has not been repeated for each derivative due to its computational cost. It is acceptable to state that the empirical and theoretical results agree and the differences, due to the error decreasing as the runs increase, are inaccuracies in the Monte Carlo estimate [Goodman, 1976].

6.2.3 The Determinant of the Fisher Information Matrix

Previously, the empirical estimates were shown to closely correspond with the true values of the Fisher information. This is useful however, in most practical applications, and especially in the case of identifiability, interest lies in the determinant of the information matrix. It is therefore important to compare the empirical and theoretical determinants.

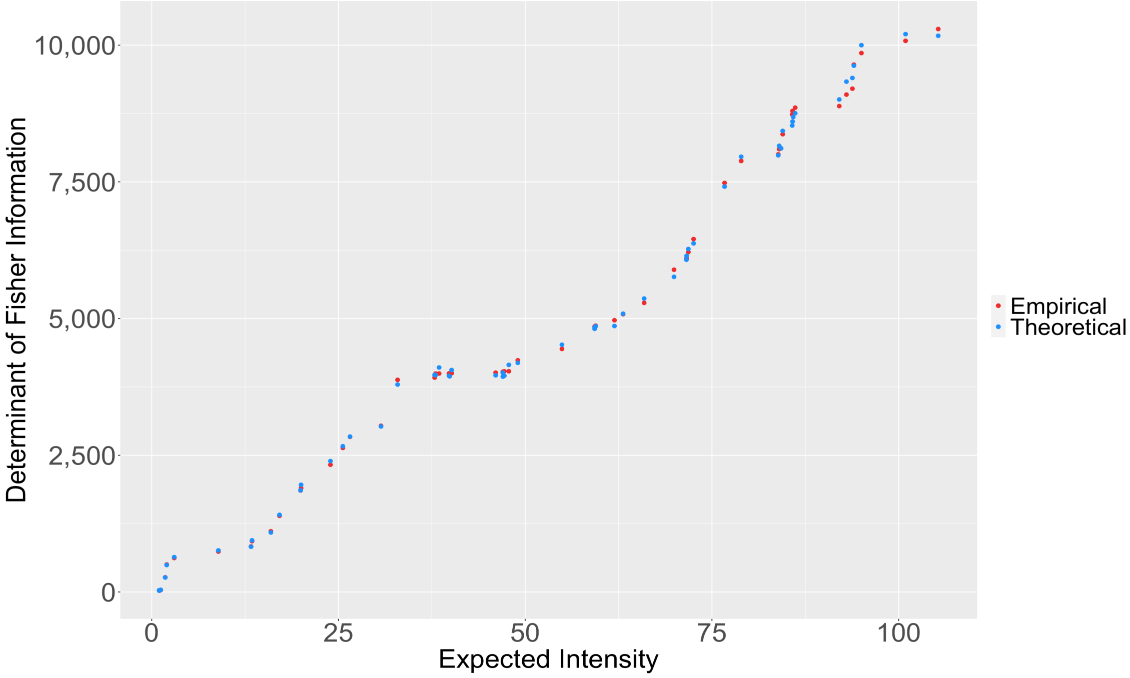

Figure 6.2: Comparison of Empirical and Theoretical Results of the Fisher's Determinant.

Figure 6.2 provides the empirical determinant in red and the theoretical determinant in blue. The results agree to such a precise level, certain theoretical results seem superimposed on their corresponding Monte Carlo estimates. The previous section discussed the error of the empirical result from the theoretical for each element within the information matrix. While an average percentage error was given, one aspect of the elements not discussed was each of their relative sizes. Larger values of $\mathbb{E}[\lambda(t)]$ yielded larger diagonal elements and therefore, greater determinant values. This explains why there are greater differences between the empirical and theoretical determinants for large $\mathbb{E}[\lambda(t)]$. The determinant is sensitive to small changes of the diagonal elements irrespective of the process' dimensionality.

It is important to note that for volatile Hawkes processes, more than 10,000 runs are necessary. The convergence rate to the true values is heavily dependent on the branching ratio. Therefore, it is best to attempt as many runs as computationally possible.

6.3 Visualising Identifiability in Hawkes Processes

There are many practical uses of the identifiability condition given in Equations 5.16, 5.17 and 5.18. It is important to note that there is pay-off between the branching ratio and the background intensity, μ. That is, if a particularly small branching ratio is necessary, then a larger background intensity must be used to ensure the process remains identifiable. Conversely, if a small background intensity is needed, then a branching ratio close to 1 must be used. This condition allows a user to observe metrically how non-identifiable a process is given the true parameters.

The next sections explore specific examples and visualise the identifiability of Hawkes processes.

6.3.1 Parameter Effect on Identifiability

Rather than consider α and β as two parameters it is possible to model their behaviour in one term: the branching ratio, γ. Figure 6.3 is a heatmap of the determinant of the Fisher information. By taking a range of values for μ and γ, it is possible to visualise the determinant of the Fisher information's behaviour. The heatmap is divided into clear segments, with each segment corresponding to a specific value of $\mathbb{E}[\lambda(t)]$.

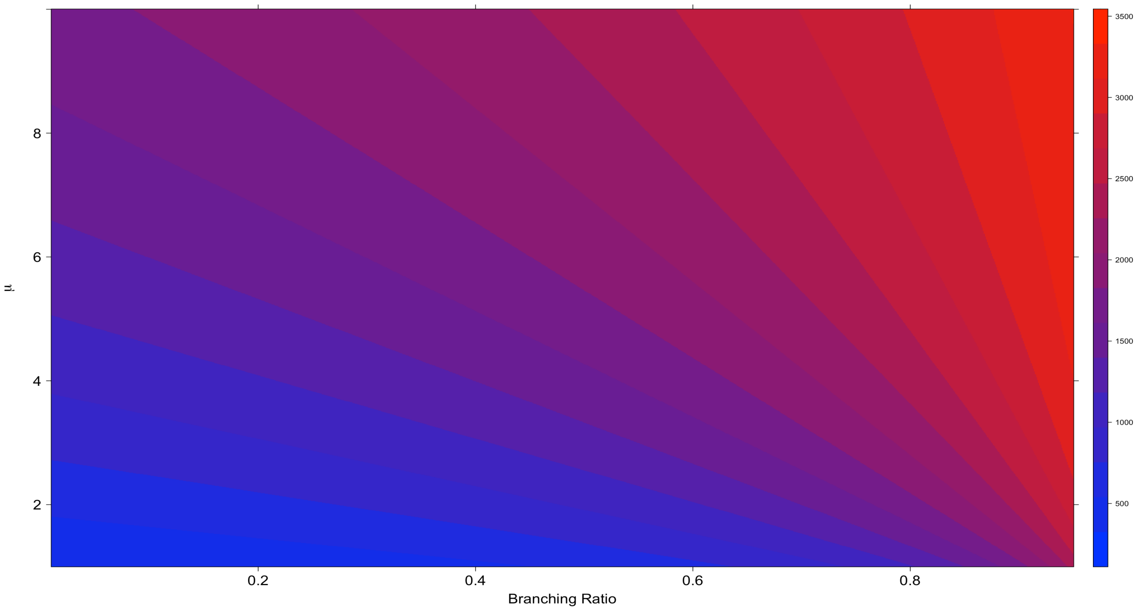

Figure 6.3: Heatmap of Determinant of the Fisher Information for a Range of Values of μ and γ.

Similar to Figure 6.2, it can be seen in Figure 6.3 that large values of $\mathbb{E}[\lambda(t)]$ yield large determinants. The banding visible in figure 6.3 however, shows that it is possible to dictate sections of the parameter space that will always be non-identifiable. Due to visualisation constraint, the structure can only be shown for the one-dimensional Hawkes process.

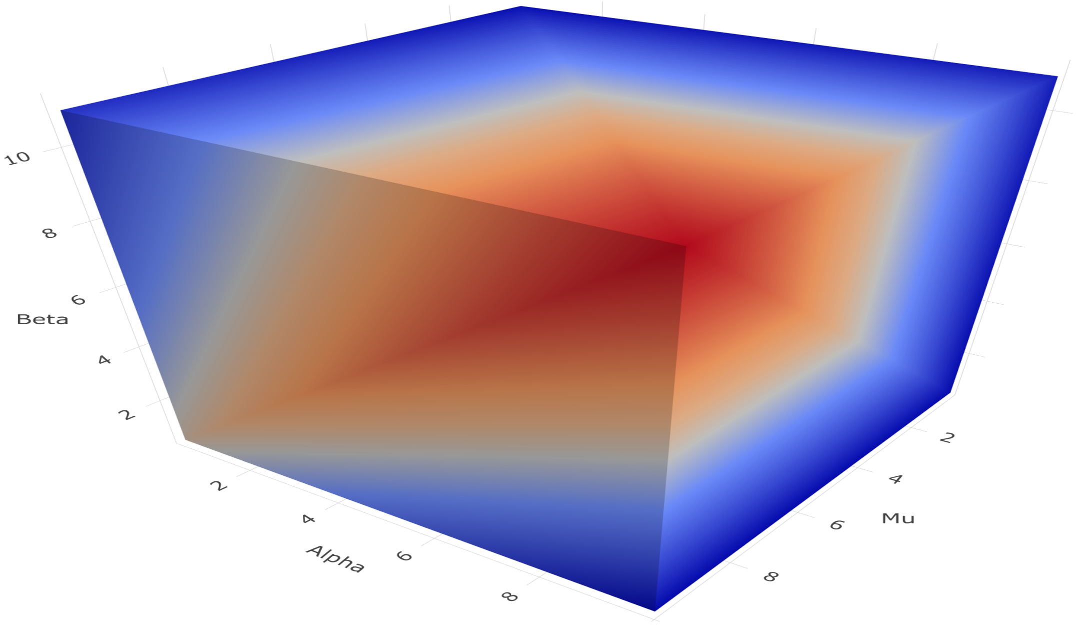

Figure 6.4: Identifiability of a One-Dimensional Hawkes Process: Blue regions dictate non-identifiable parameter sets, White regions are parameter sets on the boundary of the identifiability condition, and Red regions are identifiable parameter sets.

Identifiable regions of the one-dimensional Hawkes process can be seen in Figure 6.4. A gradient of each colour has been employed that provides a visual interpretation of distance between the terms in the inequality in Equation 5.16. As the red and blue colours get lighter, $\mathbb{E}[\lambda(t)]$ is closer to the boundary condition. It is of note that there are regions, shown in white, where

$$\mathbb{E}[\lambda(t)] = \frac{2\alpha + \beta(\beta^2 - \beta + 2)}{2(\alpha+\beta)(\beta^2 - 2\beta)}.$$

Figure 6.4 shows the identifiability of a Hawkes process as the three parameters vary. A smaller cube like shape can almost be observed for values that yield the larger intensities. However, the cube's shape is distorted in the $\alpha - \beta$ plane due to the branching ratio's effect on $\mathbb{E}[\lambda(t)]$. The figure shows that processes with branching ratios closer to 1 are more likely to be identifiable as can be seen by the diagonal dark red line in the $\alpha - \beta$ plane.

6.3.2 Simulations of Non-identifiable Hawkes Processes

In the one-dimensional setting, identifiability will break because of one of two possibilities. The issues arise from either a small branching ratio, which can be seen in Figure 6.5, or a small μ value, exhibited in Figure 6.6.

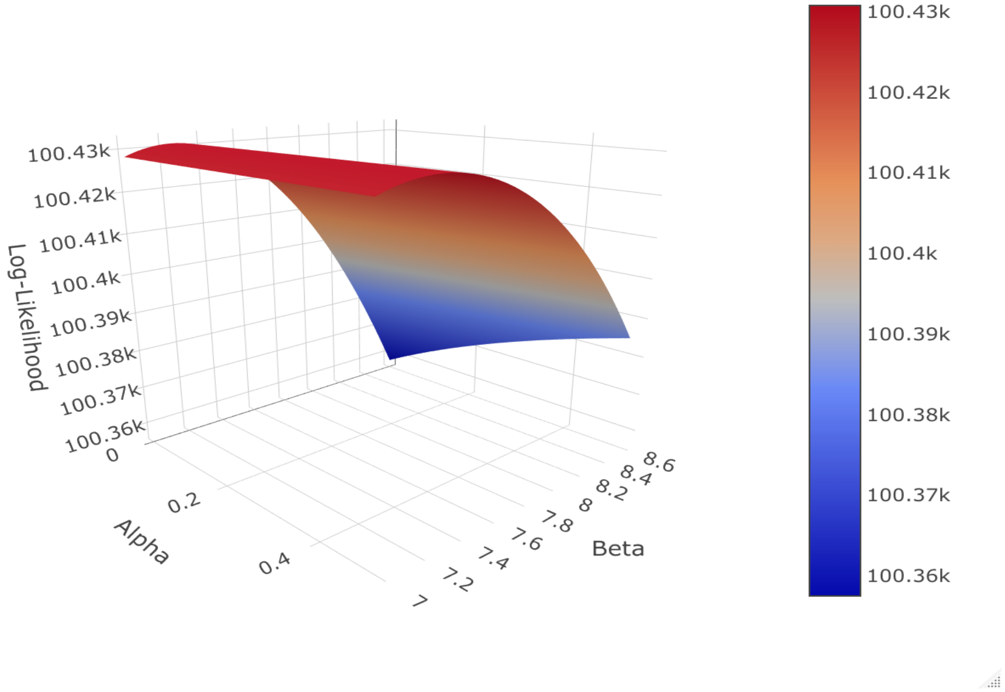

Figure 6.5: Non-identifiable Hawkes Process with true parameters $\theta = (0.5, 0.3, 7.5)$.

Figure 6.5 exhibits the behaviour already seen in Figure 5.1. That is, the process is unidentifiable due to a maximum likelihood ridge forming. However, unlike Figure 5.1 the ridge has formed along the true value of β. The relative size of β compared to α allows for a large set of possible true combinations. Note, the process in Figure 6.5 has expected intensity

$$\mathbb{E}[\lambda(t)] = \frac{0.5}{1 - 0.3/7.5} = \frac{0.5}{0.96} \approx 0.5208.$$

When this value is substituted into Equation 5.16, it can be noted that μ must be at least 0.569. It is therefore possible to state this process is unidentifiable and multiple parameter sets give rise to the same maximum likelihood.

Figure 6.6 generates a ridge-like surface. This is due to the branching ratio, which is of value 0.5. Similarly to the process for Figure 6.5, the identifiability condition can be checked. In this case, μ must be at least 12.205 which is much greater than the true value of 2.

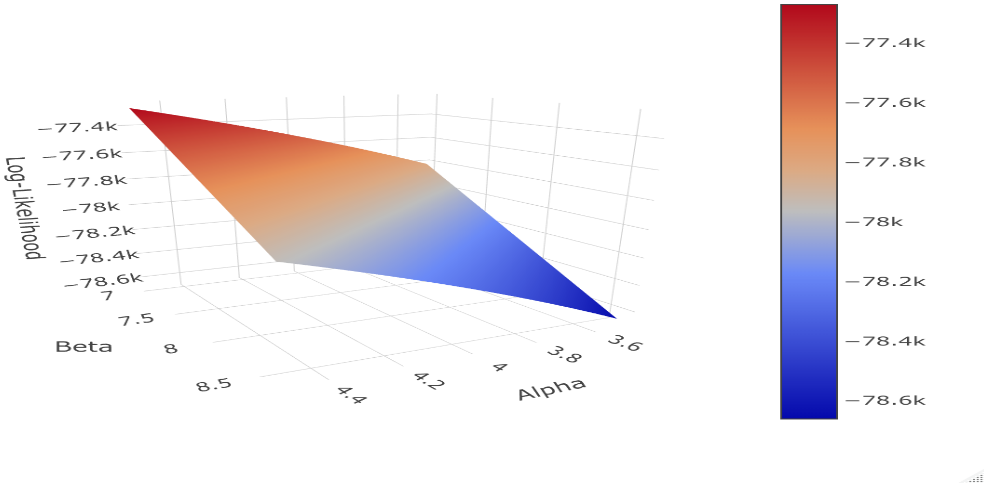

Figure 6.6: Non-identifiable Hawkes Process with true parameters $\theta = (2, 4, 8)$.

In the M-dimensional setting, issues arise when the spectral radius is small or the μ vector is small. However, it is also of note that if any given element within either of the vectors is small, issues may occur. It is sometimes possible for a large branching ratio or background intensity in one dimension to placate the identifiability issue caused by small values in another.

6.3.3 Simulations of Identifiable Hawkes Processes

Figure 6.7 presents a slightly altered process to the one given in Figure 6.5. In this case, by adjusting the μ parameter by a small amount it can be seen that the likelihood's surface has a more appropriate shape. Note that it is not necessary to adjust the μ parameter as much as has been done here. The choice of 0.7 is for visual purposes; however, a choice of $\mu = 0.57$ would suffice.

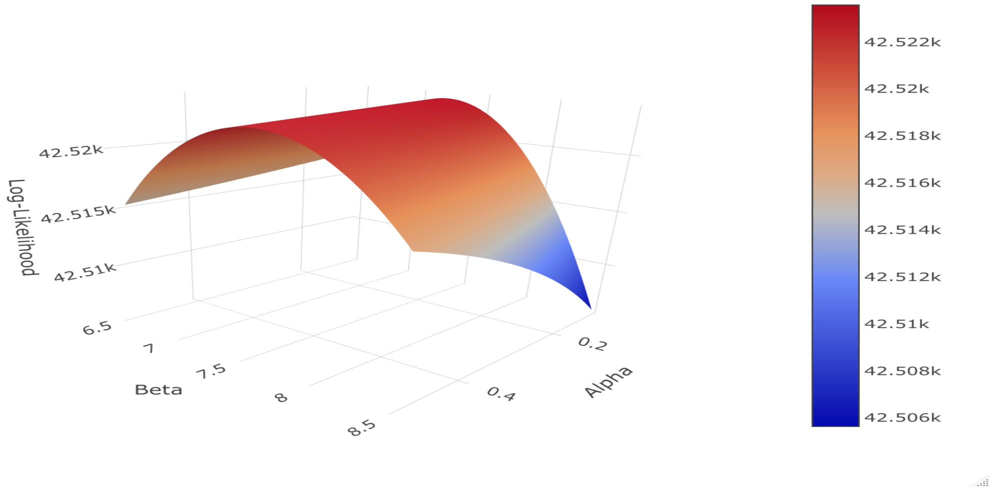

Figure 6.7: Identifiable Hawkes Process with true parameters $\theta = (0.7, 0.3, 7.5)$.

Figure 6.7 exhibits a likelihood surface when the Hawkes process is identifiable. By checking the inequality given in Equation 5.17 holds, it is possible to comment on the identifiability of any stationary Hawkes process regardless of its dimensionality.

Chapter 7

Conclusion

Hawkes processes have become an increasingly popular model choice in the last decade. Their non-Markovian properties allow for a broad range of real-world applications [Laub et al., 2015]. Nonetheless the reliance of Hawkes processes on their full history has often made for cumbersome statistical theory. There have been many useful additions to the literature, from Ogata's recursive algorithm which improves computational efficiency to Ozaki's derivation of the Hessian matrix. The research undertaken in the 1970s has allowed the development of identification methods for point processes.

Identifiability is an ever-present issue in most statistical problems. Hawkes processes are a complex model where research into the problem has mostly been avoided. This thesis has provided three extensions to the literature. Firstly, a full derivation of the M-dimensional Hessian matrix, which has been utilised to find empirical estimates of the Fisher information. Secondly, an extension has been made to Hawkes' work that, using Dynkin's formula, allows for the calculation of negative moments of both the intensity and the counting process. Finally, a full derivation of the one-dimensional and M-dimensional Fisher information has been provided. This may be easily applied to determine whether a given Hawkes process will be identifiable.

Identifiability is a unique problem. In many cases, ignoring the problem entirely can have no overall effect on the inference made. However, there are a wide range of examples that show when non-identifiability does occur inference is not possible. The research presented here has given some initial findings about the identifiability of Hawkes processes; however, there is further behaviour to explore. While an identifiability condition has been derived, this continues to rely on the calculation of a large Fisher information matrix. In more general terms, it is possible to analyse the branching ratio of a process or the intensity and make some assumptions about whether a process is likely to be identifiable. In future work, it would be beneficial to attach a given probability of identifiability for a process based solely on its intensity. Furthermore, this work explored identifiability for the stationary Hawkes process and therefore, an extension for locally-stationary processes could be derived in the future. Another possible extension to this work would be to analyse the behaviour of identifiability in aggregated Hawkes processes. Parameter estimation methods have been developed for aggregated Hawkes processes; however, there are no methods to verify identifiability in this setting [Shlomovich et al., 2020].

Hawkes processes provide an alternative and interesting model type to many other current choices. With the addition of the identifiability criteria, it is no longer necessary to assume the property holds. The inequality given in Equation 5.17 may be verified to ensure that the results generated by the data are unique and therefore, the inference is precise.

References

[Bartlett, 1963] Bartlett, M. S. (1963). The Spectral Analysis of Point Processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 25(2):264–281.

[Bellman and Åström, 1970] Bellman, R. and Åström, K. (1970). On Structural Identifiability. *Mathematical Biosciences*, 7:329–339.

[Chatalbashev et al., 2007] Chatalbashev, V., Liang, Y., Officer, A., and Trichakis, N. (2007). *Exciting Times for Trade Arrivals.* Available at http://users.iems.northwestern.edu/ armbruster/2007msande444/report1a.pdf.

[Chen and Hall, 2016] Chen, F. and Hall, P. (2016). Nonparametric Estimation for Self-Exciting Point Processes — A Parsimonious Approach. *Journal of Computational and Graphical Statistics*, 25(1):209–224.

[Chiang et al., 2020] Chiang, W. H., Liu, X., and Mohler, G. (2020). Hawkes Process Modelling of COVID-19 with Mobility Leading Indicators and Spatial Covariates. medRxiv:2020.06.06.20124149.

[Colquhoun et al., 2003] Colquhoun, D., Hatton, C. J., and Hawkes, A. G. (2003). The Quality of Maximum Likelihood Estimates of Ion Channel Rate Constants. *The Journal of Physiology*, 547(3):699—728.

[Cordi et al., 2018] Cordi, M., Challet, D., and Toke, I. M. (2018). Testing the Causality of Hawkes Processes with Time Reversal. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(3):033408.

[Cox and Lewis, 1966] Cox, D. R. and Lewis, P. A. (1966). *The Statistical Analysis of Series of Events.* Springer.

[Cui et al., 2020] Cui, L., Hawkes, A., and Yi, H. (2020). An Elementary Derivation of Moments of Hawkes Processes. *Advances in Applied Probability*, 52(1):102–137.

[Daley and Vere-Jones, 1971] Daley, D. J. and Vere-Jones, D. (1971). *A Summary of the Theory of Point Processes.* Springer.

[Daley and Vere-Jones, 2003] Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods.* Springer.

[Daw and Pender, 2018] Daw, A. and Pender, J. (2018). Queues Driven by Hawkes Processes. *Stochastic Systems*, 8(3):192–229.

[Embrechts et al., 2011] Embrechts, P., Liniger, T., and Lin, L. (2011). Multivariate Hawkes Processes: An Application to Financial Data. *Journal of Applied Probability*, 48(A):367–378.

[Gerhard et al., 2017] Gerhard, F., Deger, M., and Truccolo, W. (2017). On the Stability and Dynamics of Stochastic Spiking Neuron Models: Nonlinear Hawkes Process and Point Process GLMs. *PLOS Computational Biology*, 13(2):1–31.

[Goodman, 1976] Goodman, J. (1976). *Accuracy and Efficiency of Monte Carlo Method.* International Atomic Energy Agency.

[Goodrich and Caines, 1979] Goodrich, R. and Caines, P. (1979). Necessary and Sufficient Conditions for Local Second-Order Identifiability. *IEEE Transactions on Automatic Control*, 24(1):125–127.

[Hardiman and Bouchaud, 2014] Hardiman, S. J. and Bouchaud, J.-P. (2014). Branching-Ratio Approximation for the Self-Exciting Hawkes Process. *American Physical Society Physical Review E*, 90(6).

[Hawkes, 1971] Hawkes, A. G. (1971). Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58(1):83–90.

[Huang, 2005] Huang, G. H. (2005). Model Identifiability. *Encyclopedia of Statistics in Behavioral Science*, 3(1):1249–1251.

[Karabash, 2012] Karabash, D. (2012). On Stability of Hawkes Process. arXiv:1201.1573.

[Kobayashi and Lambiotte, 2016] Kobayashi, R. and Lambiotte, R. (2016). TiDeH: Time-Dependent Hawkes Process for Predicting Retweet Dynamics. *Proceedings of the 10th International Conference on Web and Social Media.*

[Kreutz, 2018] Kreutz, C. (2018). An Easy and Efficient Approach for Testing Identifiability. *Bioinformatics*, 34(11):1913–1921.

[Laub et al., 2015] Laub, P. J., Taimre, T., and Pollett, P. K. (2015). Hawkes Processes. arXiv:1507.02822.

[Lewis and Mohler, 2011] Lewis, E. and Mohler, G. (2011). A Nonparametric EM Algorithm for Multiscale Hawkes Processes. *Journal of Nonparametric Statistics*, 1(1):1–20.

[Lim et al., 2016] Lim, K. W., Young, L., Hanlen, L., and Hongbiao, Z. (2016). Simulation and Calibration of a Fully Bayesian Marked Multidimensional Hawkes Process with Dissimilar Decays. *Journal of Machine Learning Research*, 63:238–253.

[Møller and Rasmussen, 2005] Møller, J. and Rasmussen, J. G. (2005). Perfect simulation of Hawkes processes. *Advances in Applied Probability*, 37(3):629–646.

[Oakes and Hawkes, 1974] Oakes, D. and Hawkes, A. G. (1974). A Cluster Process Representation of a Self-Exciting Process. *Journal of Applied Probability*, 11(3):493–503.

[Ogata, 1978] Ogata, Y. (1978). The Asymptotic Behaviour of Maximum Likelihood Estimators for Stationary Point Processes. *Annals of the Institute of Statistical Mathematics*, 30(2):243–261.

[Ogata, 1981] Ogata, Y. (1981). On Lewis' Simulation Method for Point Processes. *IEEE Transactions on Information Theory*, 27(1):23–31.

[Ogata, 1988] Ogata, Y. (1988). Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *Journal of the American Statistical Association*, 83(401):9–27.

[Ozaki, 1977] Ozaki, T. (1977). Maximum Likelihood Estimation of Hawkes' Self-Exciting Point Processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155.

[Patel et al., 2019] Patel, L., Gustafsson, N., Lin, Y., Ober, R., Henriques, R., and Cohen, E. (2019). A Hidden Markov Model Approach to Characterizing the Photo-Switching Behaviour of Fluorophores. *Annals of Applied Statistics*, 13(3):1397–1429.

[Rothenberg, 1971] Rothenberg, T. J. (1971). Identification in Parametric Models. *Econometrica*, 39:577–591.

[Rubin, 1972] Rubin, I. (1972). Regular Point Processes and Their Detection. *IEEE Transactions on Information Theory*, 18(5):547–557.

[Shlomovich et al., 2020] Shlomovich, L., Cohen, E., Adams, N., and Patel, L. (2020). A Monte Carlo EM Algorithm for the Parameter Estimation of Aggregated Hawkes Processes. arXiv:2001.07160.

[Wang et al., 2020] Wang, H., Xie, L., Cuozzo, A., Mak, S., and Xie, Y. (2020). Uncertainty Quantification for Inferring Hawkes Networks. arXiv:2006.07506.

[Zabandan and Kılıçman, 2012] Zabandan, G. and Kılıçman, A. (2012). A New version of Jensen's inequality and Related Results. *Journal of Inequalities and Applications*, 2012(1):238.

Appendix A

Supplementary Material

The appendices provide all supplementary proofs. Please note, all code can be found in the GitHub repository `HawkesProcesses`.

A.1 Derivation of Expectation of the Intensity Function

Equation 2.12a is the general formula for the conditional intensity function and can be recalled to be

$$\lambda(t) = \mu(t) + \int_{-\infty}^{t} \nu(t-s)\ \mathrm{d}N(s).$$

Assuming that the background intensity is constant, and taking expectation of both sides, it can be seen that

$$\begin{aligned}
\mathbb{E}[\lambda(t)] &= \mathbb{E}\left[\mu + \int_{-\infty}^{t} \nu(t-s)\ \mathrm{d}N(s)\right], \\
&= \mu + \mathbb{E}\left[\int_{-\infty}^{t} \nu(t-s)\ \mathrm{d}N(s)\right], \\
&= \mu + \int_{-\infty}^{t} \nu(t-s)\mathbb{E}\left[\ \mathrm{d}N(s)\right].
\end{aligned}$$

It is known from Hawkes' original paper, that $\mathbb{E}\left[\ \mathrm{d}N(s)\right] = \mathbb{E}[\lambda(s)]\ \mathrm{d}s$. If the process is assumed to be stationary, then $\mathbb{E}[\lambda(t)] = c$ where c is some constant. Therefore,

$$\begin{aligned}
c &= \mu + \int_{-\infty}^{t} \nu(t-s)c\ \mathrm{d}s = \mu + c\int_{-\infty}^{t} \nu(t-s)\ \mathrm{d}s, \\
&= \mu + c\int_{0}^{\infty} \nu(s)\ \mathrm{d}s.
\end{aligned}$$

Rearranging in terms of c it is therefore possible to see the stationary condition is

$$\mathbb{E}[\lambda(t)] = c = \frac{\mu}{1 - \int_0^\infty \nu(s)\ \mathrm{d}s}.$$

Now, in the case of the simplest kernel, $\nu(t) = \alpha e^{-\beta t}$, this may be written as

$$\lambda(t) = \mu + \int_{-\infty}^{t} \alpha e^{-\beta(t-s)}\ \mathrm{d}N(s)$$

and therefore,

$$\mathbb{E}[\lambda(t)] = \frac{\mu}{1 - \int_0^\infty \alpha e^{-\beta s}\ \mathrm{d}s} = \frac{\mu}{1 - \frac{\alpha}{\beta}}.$$

A.2 Derivation of M-dimensional Log-Likelihood

It is known that,

$$\ell(t|\boldsymbol{\theta}) = \sum_{m=1}^{M} \ell_m(t|\theta_m), \tag{A.1}$$

and therefore, interest lies in finding a tractable form of $\ell_m(t|\theta_m)$. To do this, first recall,

$$\ell_m(t|\theta_m) = \int_0^t \log\{\lambda_m(s|\theta_m)\}\ \mathrm{d}N_m(s) - \int_0^t \lambda_m(s|\theta_m)\ \mathrm{d}s, \tag{A.2}$$

for an M-dimensional Hawkes process, denote the ordered set of events as $\{t_i\}_{i=1}^N$ with all events denoted $\{t_{i,m}\}_{m=1}^M$ on the interval $[0,t]$. Now, in Equation A.2 the right hand term on the right hand side of the equation is often called the compensator. The compensator, or integrated intensity, is denoted Λ. The compensator of the m^{th} dimension of an M-dimensional Hawkes process between two consecutive events $t_{n-1,m}$ and $t_{n,m}$ is

$$\Lambda_m(t_{n-1,m}, t_{n,m}) = \int_{t_{n-1,m}}^{t_{n,m}} \lambda_m(s|\theta_m)\ \mathrm{d}s. \tag{A.3}$$

By substituting Equation 2.24b into Equation A.3, it can be shown that

$$\begin{aligned}\Lambda_m(t_{n-1,m}, t_{n,m}) &= \int_{t_{n-1,m}}^{t_{n,m}} \mu_m(s) + \sum_{i=1}^{M} \sum_{t_{k,i}<s} \sum_{j=1}^{P} \alpha_{i,j,m} e^{-\beta_{i,j,m}(s-t_{k,i})}\ \mathrm{d}s,\\ &= \int_{t_{n-1,m}}^{t_{n,m}} \mu_m(s)\ \mathrm{d}s + \int_{t_{n-1,m}}^{t_{n,m}} \sum_{i=1}^{M} \sum_{t_{k,i}<s} \sum_{j=1}^{P} \alpha_{i,j,m} e^{-\beta_{i,j,m}(s-t_{k,i})}\ \mathrm{d}s,\\ &= \int_{t_{n-1,m}}^{t_{n,m}} \mu_m(s)\ \mathrm{d}s + \int_{t_{n-1,m}}^{t_{n,m}} \sum_{i=1}^{M} \sum_{t_{k,i}<t_{n-1,m}} \sum_{j=1}^{P} \alpha_{i,j,m} e^{-\beta_{i,j,m}(s-t_{k,i})}\ \mathrm{d}s\\ &+ \int_{t_{n-1,m}}^{t_{n,m}} \sum_{i=1}^{M} \sum_{t_{n-1,m}\le t_{k,i}<s} \sum_{j=1}^{P} \alpha_{i,j,m} e^{-\beta_{i,j,m}(s-t_{k,i})}\ \mathrm{d}s.\end{aligned}$$

Now, by taking terms not dependent outside of the integral, it can be seen

$$\begin{aligned}\Lambda_m(t_{n-1,m}, t_{n,m}) &= \int_{t_{n-1,m}}^{t_{n,m}} \mu_m(s)\ \mathrm{d}s + \sum_{i=1}^{M} \sum_{t_{k,i}<t_{n-1,m}} \sum_{j=1}^{P} \alpha_{i,j,m} \int_{t_{n-1,m}}^{t_{n,m}} e^{-\beta_{i,j,m}(s-t_{k,i})}\ \mathrm{d}s\\ &+ \sum_{i=1}^{M} \sum_{t_{n-1,m}\le t_{k,i}<s} \sum_{j=1}^{P} \alpha_{i,j,m} \int_{t_{n-1,m}}^{t_{n,m}} e^{-\beta_{i,j,m}(s-t_{k,i})}\ \mathrm{d}s,\\ &= \int_{t_{n-1,m}}^{t_{n,m}} \mu_m(s)\ \mathrm{d}s + \sum_{i=1}^{M} \sum_{t_{k,i}<t_{n-1,m}} \sum_{j=1}^{P} \frac{\alpha_{i,j,m}}{\beta_{i,j,m}} \Big(e^{-\beta_{i,j,m}(t_{n-1,m}-t_{k,i})} - e^{-\beta_{i,j,m}(t_{n,m}-t_{k,i})}\Big)\\ &+ \sum_{i=1}^{M} \sum_{t_{n-1,m}\le t_{k,i}<t_{n,m}} \sum_{j=1}^{P} \frac{\alpha_{i,j,m}}{\beta_{i,j,m}} \Big(1 - e^{-\beta_{i,j,m}(t_{n,m}-t_{k,i})}\Big).\end{aligned}$$

Therefore, on the interval $[0,t]$ the compensator is equivalent to

$$\Lambda_m(0,t) = \int_0^t \mu_m(s)\ \mathrm{d}s + \sum_{i=1}^{M} \sum_{0\le t_{k,i}<t} \sum_{j=1}^{P} \frac{\alpha_{i,j,m}}{\beta_{i,j,m}} \Big(1 - e^{-\beta_{i,j,m}(t-t_{k,i})}\Big). \tag{A.4}$$

However, since all $t_{k,i} \in [0,t]$ then,

$$\Lambda_m(0,t) = \int_0^t \mu_m(s)\ \mathrm{d}s + \sum_{i=1}^{M}\sum_{k=1}^{N}\sum_{j=1}^{P} \frac{\alpha_{i,j,m}}{\beta_{i,j,m}}\Big(1 - e^{-\beta_{i,j,m}(t-t_{k,i})}\Big), \tag{A.5}$$

$$= \sum_{k=1}^{N} \mu_m(t_{k,m}) + \sum_{i=1}^{M}\sum_{k=1}^{N}\sum_{j=1}^{P} \frac{\alpha_{i,j,m}}{\beta_{i,j,m}}\Big(1 - e^{-\beta_{i,j,m}(t-t_{k,i})}\Big). \tag{A.6}$$

This solves the term to the furthest right of Equation A.2, while the remaining term in Equation A.2 can be expressed as

$$\int_0^t \log\{\lambda_m(s|\theta_m)\}\ \mathrm{d}N_m(s) = \int_0^t \log\Big(\mu_m(s) + \sum_{i=1}^{M}\sum_{t_{k,i}<s}\sum_{j=1}^{P} \alpha_{i,j,m} e^{-\beta_{i,j,m}(s-t_{k,i})}\Big)\ \mathrm{d}N_m(s). \tag{A.7}$$

Before Equation A.7 can be simplified, a recursive relationship must be defined [Ogata, 1981]. Let

$$\begin{aligned}
R_{i,j,m}(d) &= \sum_{t_{k,i}<t_{d,m}} e^{-\beta_{i,j,m}(t_{d,m}-t_{k,i})}, \\
&= \sum_{t_{k,i}<t_{d-1,m}} e^{-\beta_{i,j,m}(t_{d,m}-t_{k,i})} + \sum_{t_{d-1,m}\le t_{k,i}<t_{d,m}} e^{-\beta_{i,j,m}(t_{d,m}-t_{k,i})}, \\
&= e^{-\beta_{i,j,m}(t_{d,m}-t_{d-1,m})} \sum_{t_{k,i}<t_{d-1,m}} e^{-\beta_{i,j,m}(t_{d-1,m}-t_{k,i})} + \sum_{t_{d-1,m}\le t_{k,i}<t_{d,m}} e^{-\beta_{i,j,m}(t_{d,m}-t_{k,i})}, \\
&= e^{-\beta_{i,j,m}(t_{d,m}-t_{d-1,m})} R_{i,j,m}(d-1) + \sum_{t_{d-1,m}\le t_{k,i}<t_{d,m}} e^{-\beta_{i,j,m}(t_{d,m}-t_{k,i})},
\end{aligned}$$

and therefore,

$$R_{i,j,m}(d) = \begin{cases} e^{-\beta_{i,j,m}(t_{d,m}-t_{d-1,m})} R_{i,j,m}(d-1) + \sum_{t_{d-1,m}\le t_{k,i}<t_{d,m}} e^{-\beta_{i,j,m}(t_{d,m}-t_{k,i})} & \text{if } i \neq m, \\ e^{-\beta_{i,j,m}(t_{d,m}-t_{d-1,m})}\Big(1 + R_{i,j,m}(d-1)\Big) & \text{if } i = m, \end{cases}$$

Now using this recursive formula, Equation A.7 may be rewritten as

$$\int_0^t \log\{\lambda_m(s|\theta_m)\}\ \mathrm{d}N_m(s) = \sum_{d:t_{d,m}<t} \log\Big(\mu_m(t_{d,m}) + \sum_{i=1}^{M}\sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(d)\Big), \tag{A.8}$$

noting that $R_{i,j,m}(0) = 0$. Also, the use of log should be taken to mean the natural logarithm. Again, regardless of dimension $t_{d,m} < t$ for all d and therefore,

$$\int_0^t \log\{\lambda_m(s|\theta_m)\}\ \mathrm{d}N_m(s) = \sum_{k=1}^{N} \log\Big(\mu_m(t_{k,m}) + \sum_{i=1}^{M}\sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k)\Big). \tag{A.9}$$

Finally, substituting Equations A.6 and A.9 into Equation A.2, it can be seen that

$$\begin{aligned}
\ell_m(t|\theta_m) = \Big(\sum_{k=1}^{N} \log\Big(\mu_m(t_{k,m}) + \sum_{i=1}^{M}\sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k)\Big)\Big) - \Big(\sum_{k=1}^{N} \mu_m(t_{k,m})\Big) \\
- \Big(\sum_{i=1}^{M}\sum_{j=1}^{P}\sum_{k=1}^{N} \frac{\alpha_{i,j,m}}{\beta_{i,j,m}}\Big(1 - e^{-\beta_{i,j,m}(t-t_{k,i})}\Big)\Big),
\end{aligned} \tag{A.10}$$

thus providing the required solution. Note that a simplified form of this proof for the one-dimensional case has been derived previously [Laub et al., 2015].

A.3 Hessian Derivation

Recall from Equation 2.29 that the full log-likelihood for observations on the interval $[0, t]$ is

$$\ell(t|\boldsymbol{\theta}) = \sum_{m=1}^{M} \ell_m(t|\theta_m).$$

Furthermore, Equation 2.31 gives

$$\begin{aligned}\ell_m(t|\theta_m) = \left(\sum_{k=1}^{N} \log\left(\mu_m(t_{k,m}) + \sum_{i=1}^{M}\sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k)\right)\right) - \left(\sum_{k=1}^{N} \mu_m(t_{k,m})\right) \\ - \left(\sum_{i=1}^{M}\sum_{j=1}^{P}\sum_{k=1}^{N} \frac{\alpha_{i,j,m}}{\beta_{i,j,m}}\left(1 - e^{-\beta_{i,j,m}(t-t_{k,i})}\right)\right),\end{aligned}$$

with $R_{i,j,m}(k) = \sum_{t_{d,i}<t_{k,m}} e^{-\beta_{i,j,m}(t_{k,m}-t_{d,i})}$. Therefore, the gradient may be expressed as

$$\frac{\partial \ell_m}{\partial \mu_m} = \left(\sum_{k=1}^{N} \frac{1}{\mu_m + \sum_{i=1}^{M}\sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k)}\right), \tag{A.11}$$

$$\frac{\partial \ell_m}{\partial \alpha_{i,j,m}} = \left(\sum_{k=1}^{N} \frac{R_{i,j,m}(k)}{\mu_m + \sum_{i=1}^{M}\sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k)}\right) - \frac{1}{\beta_{i,j,m}}\left(\sum_{k=1}^{N} 1 - e^{-\beta_{i,j,m}(t-t_{k,i})}\right), \tag{A.12}$$

$$\begin{aligned}\frac{\partial \ell_m}{\partial \beta_{i,j,m}} = \frac{\alpha_{i,j,m}}{\beta_{i,j,m}^2}\left(\sum_{k=1}^{N} 1 - e^{-\beta_{i,j,m}(t-t_{k,i})}\right) - \frac{\alpha_{i,j,m}}{\beta_{i,j,m}}\left(\sum_{k=1}^{N} (t - t_{k,i}) e^{-\beta_{i,j,m}(t-t_{k,i})}\right) \\ - \left(\sum_{k=1}^{N} \frac{\alpha_{i,j,m} R^{'}_{i,j,m}(k)}{\mu_m + \sum_{i=1}^{M}\sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k)}\right),\end{aligned} \tag{A.13}$$

where $R^{'}_{i,j,m}(k) = \sum_{t_{d,i}<t_{k,m}} (t_{k,m} - t_{d,i}) e^{-\beta_{i,j,m}(t_{k,m}-t_{d,i})}$ and $R^{'}_{i,j,m}(0) = 0$.

Then, the Hessian matrix, $\boldsymbol{H}$, can be computed for the continuous M-dimensional Hawkes process likelihood. There are 30 separate cases to consider.

For μ_m,

$$\frac{\partial^2 \ell_m}{\partial \mu_m^2} = -\left(\sum_{k=1}^{N} \frac{1}{(\mu_m + \sum_{i=1}^{M}\sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2}\right), \quad \text{and} \quad \frac{\partial^2 \ell_m}{\partial \mu_m \mu_{m'}} = 0 \text{ for } m \neq m^{'}.$$

For $\alpha_{i,j,m}$,

$$\frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m}^2} = -\sum_{k=1}^{N} \left(\frac{R_{i,j,m}(k)}{\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k)} \right)^2,$$

$$\frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \alpha_{i',j,m}} = -\sum_{k=1}^{N} \frac{R_{i,j,m}(k) R_{i',j,m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2},$$

$$\frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \alpha_{i,j',m}} = -\sum_{k=1}^{N} \frac{R_{i,j,m}(k) R_{i,j',m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2},$$

$$\frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \alpha_{i',j',m}} = -\sum_{k=1}^{N} \frac{R_{i,j,m}(k) R_{i',j',m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2},$$

and

$$\frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \alpha_{i,j,m'}} = \frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \alpha_{i',j,m'}} = \frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \alpha_{i,j',m'}} = \frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \alpha_{i',j',m'}} = 0,$$

where, $i \neq i^{'}, j \neq j^{'}$ and $m \neq m^{'}$.

For $\beta_{i,j,m}$,

$$\begin{aligned} \frac{\partial^2 \ell_m}{\partial \beta_{i,j,m}^2} = \frac{-2\alpha_{i,j,m}}{\beta_{i,j,m}^3} \left(\sum_{k=1}^{N} 1 - e^{-\beta_{i,j,m}(t - t_{k,i})} \right) + \frac{2\alpha_{i,j,m}}{\beta_{i,j,m}^2} \left(\sum_{k=1}^{N} (t - t_{k,i}) e^{-\beta_{i,j,m}(t - t_{k,i})} \right) \\ + \frac{\alpha_{i,j,m}}{\beta_{i,j,m}} \left(\sum_{k=1}^{N} (t - t_{k,i})^2 e^{-\beta_{i,j,m}(t - t_{k,i})} \right) \\ + \left(\sum_{k=1}^{N} \frac{\alpha_{i,j,m} R^{''}_{i,j,m}(k)}{\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k)} - \left(\frac{\alpha_{i,j,m} R^{'}_{i,j,m}(k)}{\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k)} \right)^2 \right), \end{aligned}$$

where $R^{''}_{i,j,m}(k) = \sum_{t_{d,i} < t_{k,m}} (t_{k,m} - t_{d,i})^2 e^{-\beta_{i,j,m}(t_{k,m} - t_{d,i})}$ and $R^{''}_{i,j,m}(0) = 0$. Also,

$$\frac{\partial^2 \ell_m}{\partial \beta_{i,j,m} \partial \beta_{i',j,m}} = -\left(\sum_{k=1}^{N} \frac{\alpha_{i,j,m} \alpha_{i',j,m} R^{'}_{i,j,m}(k) R^{'}_{i',j,m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2} \right),$$

$$\frac{\partial^2 \ell_m}{\partial \beta_{i,j,m} \partial \beta_{i,j',m}} = -\left(\sum_{k=1}^{N} \frac{\alpha_{i,j,m} \alpha_{i,j',m} R^{'}_{i,j,m}(k) R^{'}_{i,j',m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2} \right),$$

$$\frac{\partial^2 \ell_m}{\partial \beta_{i,j,m} \partial \beta_{i',j',m}} = -\left(\sum_{k=1}^{N} \frac{\alpha_{i,j,m} \alpha_{i',j',m} R^{'}_{i,j,m}(k) R^{'}_{i',j',m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2} \right),$$

and

$$\frac{\partial^2 \ell_m}{\partial \beta_{i,j,m} \partial \beta_{i,j,m'}} = \frac{\partial^2 \ell_m}{\partial \beta_{i,j,m} \partial \beta_{i',j,m'}} = \frac{\partial^2 \ell_m}{\partial \beta_{i,j,m} \partial \beta_{i,j',m'}} = \frac{\partial^2 \ell_m}{\partial \beta_{i,j,m} \partial \beta_{i',j',m'}} = 0,$$

where, $i \neq i^{'}, j \neq j^{'}$ and $m \neq m^{'}$.

For μ_m and $\alpha_{i,j,m}$,

$$\frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \mu_m} = -\left(\sum_{k=1}^{N} \frac{R_{i,j,m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2} \right), \quad \text{and} \quad \frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \mu_{m'}} = 0 \text{ for } m \neq m^{'}.$$

For μ_m and $\beta_{i,j,m}$,

$$\frac{\partial^2 \ell_m}{\partial \beta_{i,j,m} \partial \mu_m} = \Bigg(\sum_{k=1}^{N} \frac{\alpha_{i,j,m} R^{'}_{i,j,m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2} \Bigg) \quad \text{and} \quad \frac{\partial^2 \ell_m}{\partial \beta_{i,j,m} \partial \mu_{m'}} = 0 \text{ for } m \neq m^{'}.$$

Finally, for $\alpha_{i,j,m}$ and $\beta_{i,j,m}$,

$$\begin{aligned} \frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \beta_{i,j,m}} = & \frac{-1}{\beta_{i,j,m}} \Bigg(\sum_{k=1}^{N} (t - t_{k,i}) e^{-\beta_{i,j,m}(t - t_{k,i})} \Bigg) + \frac{1}{\beta^2_{i,j,m}} \Bigg(\sum_{k=1}^{N} 1 - e^{-\beta_{i,j,m}(t - t_{k,i})} \Bigg) \\ & - \Bigg(\sum_{k=1}^{N} \frac{R^{'}_{i,j,m}(k)}{\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k)} \Bigg) + \Bigg(\sum_{k=1}^{N} \frac{\alpha_{i,j,m} R_{i,j,m}(k) R^{'}_{i,j,m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2} \Bigg), \end{aligned}$$

$$\frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \beta_{i',j,m}} = \sum_{k=1}^{N} \frac{\alpha_{i,j,m} R_{i,j,m}(k) R^{'}_{i',j,m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2},$$

$$\frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \beta_{i,j',m}} = \sum_{k=1}^{N} \frac{\alpha_{i,j,m} R_{i,j,m}(k) R^{'}_{i,j',m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2},$$

$$\frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \beta_{i',j',m}} = \sum_{k=1}^{N} \frac{\alpha_{i,j,m} R_{i,j,m}(k) R^{'}_{i',j',m}(k)}{(\mu_m + \sum_{i=1}^{M} \sum_{j=1}^{P} \alpha_{i,j,m} R_{i,j,m}(k))^2},$$

and

$$\frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \beta_{i,j,m'}} = \frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \beta_{i',j,m'}} = \frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \beta_{i,j',m'}} = \frac{\partial^2 \ell_m}{\partial \alpha_{i,j,m} \partial \beta_{i',j',m'}} = 0,$$

where, $i \neq i^{'}$, $j \neq j^{'}$ and $m \neq m^{'}$.

This illustrates all cases required for the full Hessian.

A.4 Derivation of the One-Dimensional Fisher Information

Similar to the one-dimensional Hessian matrix, the Fisher information matrix is a 3 × 3 square symmetric matrix and therefore, has six unique elements to derive. The elements will be found case by case, however, first recall Equation 4.9.

$$I(\boldsymbol{\theta}) = \mathbb{E}\Bigg[-\frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j} \Bigg]_{\theta = \theta_0} = t \Bigg(\mathbb{E}\Bigg[\frac{1}{\lambda(t)} \frac{\partial \lambda(t)}{\partial \theta_i} \frac{\partial \lambda(t)}{\partial \theta_j} \Bigg] \Bigg).$$

Note here, that the observations have occurred on the interval $[0, t]$. In using this method to derive the Fisher information, it is necessary also to recall the conditional intensity function given in Equation 2.13. For a $P = 1$ order exponential kernel,

$$\lambda(t) = \mu + \int_{-\infty}^{t} \alpha e^{-\beta(t-s)} \, \mathrm{d}N(s). \tag{A.14}$$

In the one-dimensional setting, $\theta = (\mu, \alpha, \beta)$. Therefore, before calculating the six elements, the partial derivatives of the intensity function need to be derived. These are,

$$\frac{\partial \lambda(t)}{\partial \mu} = 1, \tag{A.15a}$$

$$\frac{\partial \lambda(t)}{\partial \alpha} = \int_{-\infty}^{t} e^{-\beta(t-s)} \, \mathrm{d}N(s) = \frac{\lambda(t) - \mu}{\alpha}, \tag{A.15b}$$

$$\frac{\partial \lambda(t)}{\partial \beta} = \int_{-\infty}^{t} -\alpha(t-s)e^{-\beta(t-s)}\, \mathrm{d}N(s) = t(\mu - \lambda(t)) + \int_{-\infty}^{t} \alpha s e^{-\beta(t-s)}\, \mathrm{d}N(s). \tag{A.15c}$$

It is also known that the expectation of the partial derivative of the log-likelihood is 0 with respect to any parameter. Throughout this derivation, the linearity of expectation will be utilised. Furthermore, it is necessary in most steps to use the definitions of the first positive and negative moment of the intensity function given in Equation A.14. These are,

$$\mathbb{E}[\lambda(t)] = \frac{\beta\mu}{\beta-\alpha}, \tag{A.16}$$

$$\mathbb{E}[\lambda^{-1}(t)] = \frac{\beta-\alpha}{\beta\mu}. \tag{A.17}$$

These formulae can then be used to find the expectation of Equations A.15a, A.15b and A.15c.

$$\mathbb{E}\left[\frac{\partial \lambda(t)}{\partial \mu}\right] = 1, \tag{A.18a}$$

$$\mathbb{E}\left[\frac{\partial \lambda(t)}{\partial \alpha}\right] = \mathbb{E}\left[\frac{\lambda(t)-\mu}{\alpha}\right] = \frac{1}{\alpha}\mathbb{E}\Big[\lambda(t)\Big] - \frac{\mu}{\alpha} = \frac{\mu}{\beta-\alpha}, \tag{A.18b}$$

$$\mathbb{E}\left[\frac{\partial \lambda(t)}{\partial \beta}\right] = \mu t - t\Big(\mathbb{E}\Big[\lambda(t)\Big]\Big) + \int_{-\infty}^{t} \alpha s e^{-\beta(t-s)}\mathbb{E}\Big[\, \mathrm{d}N(s)\Big] = \frac{-1}{\beta(\beta-\alpha)}. \tag{A.18c}$$

The solutions here utilise Hawkes' original work that showed $\mathbb{E}[\lambda(s)]\ \mathrm{d}s = \mathbb{E}[N(s)]$. This allows the integral solution,

$$\begin{aligned}
\int_{-\infty}^{t} \alpha s e^{-\beta(t-s)}\mathbb{E}\Big[\, \mathrm{d}N(s)\Big] &= \int_{-\infty}^{t} \alpha s e^{-\beta(t-s)}\mathbb{E}\Big[\lambda(s)\Big]\ \mathrm{d}s = \alpha\mathbb{E}\Big[\lambda(s)\Big]\int_{-\infty}^{t} s e^{-\beta(t-s)}\ \mathrm{d}s, \\
&= \frac{\alpha\beta\mu}{\beta-\alpha}\int_{-\infty}^{t} s e^{-\beta(t-s)}\ \mathrm{d}s = \frac{\alpha\beta\mu}{\beta-\alpha}\left(\frac{\beta t - 1}{\beta^2}\right), \\
&= \frac{\alpha\mu(\beta t - 1)}{\beta(\beta-\alpha)}.
\end{aligned}$$

This completes the necessary set of formulae that then allows for the Fisher information derivation. Now it is possible to derive the unique cases. Firstly for $\partial^2\ell/\partial\mu^2$,

$$\begin{aligned}
\mathbb{E}\left[-\frac{\partial^2\ell}{\partial\mu^2}\right] &= t\left(\mathbb{E}\left[\frac{1}{\lambda(t)}\frac{\partial\lambda(t)}{\partial\mu}\frac{\partial\lambda(t)}{\partial\mu}\right]\right) = t\left(\mathbb{E}\left[\frac{1}{\lambda(t)}\left(\frac{\partial\lambda(t)}{\partial\mu}\right)^2\right]\right), \\
&= t\left(\mathbb{E}\left[\frac{1}{\lambda(t)}(1)^2\right]\right) = t\left(\mathbb{E}\left[\frac{1}{\lambda(t)}\right]\right), \\
&= t\left(\frac{\beta-\alpha}{\beta\mu}\right),
\end{aligned}$$

where the last step follows by recalling the closed-form of the first negative moment of the intensity function given in Equation A.17. Next, it is possible to derive the term for $\partial^2\ell/\partial\mu\partial\alpha$,

$$\begin{aligned}
\mathbb{E}\left[-\frac{\partial^2\ell}{\partial\mu\partial\alpha}\right] &= t\left(\mathbb{E}\left[\frac{1}{\lambda(t)}\frac{\partial\lambda(t)}{\partial\mu}\frac{\partial\lambda(t)}{\partial\alpha}\right]\right), \\
&= t\left(\mathbb{E}\left[\frac{1}{\lambda(t)}(1)\left(\frac{\lambda(t)-\mu}{\alpha}\right)\right]\right) = t\left(\mathbb{E}\left[\left(\frac{\lambda(t)-\mu}{\alpha\lambda(t)}\right)\right]\right), \\
&= t\left(\mathbb{E}\left[\frac{1}{\alpha} - \frac{\mu}{\alpha\lambda(t)}\right]\right) = t\left(\frac{1}{\alpha} - \frac{\mu}{\alpha}\mathbb{E}\left[\frac{1}{\lambda(t)}\right]\right), \\
&= \frac{t}{\alpha}\left(1 - \mu\frac{\beta-\alpha}{\beta\mu}\right) = \frac{t}{\alpha}\left(1 - \frac{\beta-\alpha}{\beta}\right), \\
&= \frac{t}{\beta}.
\end{aligned}$$

This derivation again requires Equation A.17.

Next, it is logical to look at $\partial^2 \ell / \partial \alpha^2$.

$$\begin{aligned}
\mathbb{E}\Big[-\frac{\partial^2 \ell}{\partial \alpha^2}\Big] &= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\frac{\partial \lambda(t)}{\partial \alpha}\frac{\partial \lambda(t)}{\partial \alpha}\Big]\bigg) = t\bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big(\frac{\partial \lambda(t)}{\partial \alpha}\Big)^2\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big(\frac{\lambda(t)-\mu}{\alpha}\Big)^2\Big]\bigg) = t\bigg(\mathbb{E}\Big[\frac{\lambda^2(t)-2\mu\lambda(t)+\mu^2}{\alpha^2\lambda(t)}\Big]\bigg),\\
&= \frac{t}{\alpha^2}\bigg(\mathbb{E}\Big[\lambda(t)-2\mu+\frac{\mu^2}{\lambda(t)}\Big]\bigg) = \frac{t}{\alpha^2}\bigg(\mathbb{E}\Big[\lambda(t)\Big]-2\mu+\mu^2\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big]\bigg),\\
&= \frac{t}{\alpha^2}\bigg(\frac{\beta\mu}{\beta-\alpha}-2\mu+\mu^2\frac{\beta-\alpha}{\beta\mu}\bigg) = \frac{\mu t}{\alpha^2}\bigg(\frac{\beta}{\beta-\alpha}-2+\frac{\beta-\alpha}{\beta}\bigg),\\
&= \frac{\mu t}{\alpha^2}\bigg(\frac{2\beta^2-2\alpha\beta+\alpha^2}{\beta(\beta-\alpha)}-2\bigg) = \frac{\mu t}{\alpha^2}\bigg(\frac{2\beta^2-2\alpha\beta+\alpha^2-2\beta(\beta-\alpha)}{\beta(\beta-\alpha)}\bigg),\\
&= \frac{t}{\alpha^2}\bigg(\frac{2\mu\alpha^2\beta+2\mu\alpha^3+\alpha^2\beta^2-\alpha^3\beta}{2\beta(\beta-\alpha)(\beta+\alpha)}\bigg) = \frac{t}{\alpha^2}\bigg(\frac{2\mu\alpha^2(\beta+\alpha)+\alpha^2\beta(\beta-\alpha)}{2\beta(\beta-\alpha)(\beta+\alpha)}\bigg),\\
&= \frac{t}{\alpha^2}\bigg(\frac{\mu\alpha^2}{\beta(\beta-\alpha)}+\frac{\alpha^2}{2(\beta+\alpha)}\bigg),\\
&= \frac{\mu t}{\beta(\beta-\alpha)}+\frac{t}{2(\beta+\alpha)}.
\end{aligned}$$

Then looking to solve $\partial^2 \ell / \partial \mu \partial \beta$.

$$\begin{aligned}
\mathbb{E}\Big[-\frac{\partial^2 \ell}{\partial \mu \partial \beta}\Big] &= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\frac{\partial \lambda(t)}{\partial \mu}\frac{\partial \lambda(t)}{\partial \beta}\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}(1)\Big(\int_{-\infty}^{t} -\alpha(t-s)e^{-\beta(t-s)}\ \mathrm{d}N(s)\Big)\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big(\int_{-\infty}^{t} -\alpha(t-s)e^{-\beta(t-s)}\ \mathrm{d}N(s)\Big)\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big]\Big(\int_{-\infty}^{t} -\alpha(t-s)e^{-\beta(t-s)}\ \mathbb{E}\Big[\mathrm{d}N(s)\Big]\Big)\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big]\int_{-\infty}^{t} -\alpha(t-s)e^{-\beta(t-s)}\ \mathbb{E}[\lambda(s)]\ \mathrm{d}s\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{\lambda(s)}{\lambda(t)}\Big]\int_{-\infty}^{t} -\alpha(t-s)e^{-\beta(t-s)}\ \mathrm{d}s\bigg),\\
&= t\int_{-\infty}^{t} -\alpha(t-s)e^{-\beta(t-s)}\ \mathrm{d}s = t\Big(-\frac{\alpha}{\beta^2}\Big),\\
&= -\frac{\alpha t}{\beta^2}.
\end{aligned}$$

This is possible since the expectation of the intensity is itself a constant and it is therefore, possible to separate the terms as described above.

The next two elements are slightly more complicated and it is worth recalling the rules of squaring integrals. Now,

$$
\begin{aligned}
\mathbb{E}\Big[-\frac{\partial^2 \ell}{\partial \beta^2}\Big] &= t\Bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\frac{\partial \lambda(t)}{\partial \beta}\frac{\partial \lambda(t)}{\partial \beta}\Big]\Bigg) = t\Bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big(\frac{\partial \lambda(t)}{\partial \beta}\Big)^2\Big]\Bigg),\\
&= t\Bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big(\int_{-\infty}^{t} -\alpha(t-s)e^{-\beta(t-s)}\ \mathrm{d}N(s)\Big)^2\Big]\Bigg),\\
&= \alpha^2 t\Bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big(\int_{-\infty}^{t} (t-s)e^{-\beta(t-s)}\ \mathrm{d}N(s)\Big)^2\Big]\Bigg),\\
&= \alpha^2 t\Bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\int_{-\infty}^{t}\int_{-\infty}^{t} (t-s)(t-r)e^{-\beta(t-s)}e^{-\beta(t-r)}\ \mathrm{d}N(s)\ \mathrm{d}N(r)\Big]\Bigg),\\
&= \alpha^2 t\Bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big]\int_{-\infty}^{t}\int_{-\infty}^{t} (t-s)(t-r)e^{-\beta(t-s)}e^{-\beta(t-r)}\ \mathbb{E}[\lambda(s)]\ \mathrm{d}s\ \mathbb{E}[\lambda(r)]\ \mathrm{d}r\Bigg),\\
&= \alpha^2 t\Bigg(\mathbb{E}\Big[\frac{\lambda(r)\lambda(s)}{\lambda(t)}\Big]\int_{-\infty}^{t}\int_{-\infty}^{t} (t-s)(t-r)e^{-\beta(t-s)}e^{-\beta(t-r)}\ \mathrm{d}s\ \mathrm{d}r\Bigg),\\
&= \alpha^2 t\Bigg(\mathbb{E}\Big[\frac{\lambda(r)\lambda(s)}{\lambda(t)}\Big]\int_{-\infty}^{t} (t-s)e^{-\beta(t-s)}\Big(\int_{-\infty}^{t} (t-r)e^{-\beta(t-r)}\ \mathrm{d}r\Big)\ \mathrm{d}s\Bigg),\\
&= \frac{\alpha^2 t}{\beta^2}\Bigg(\mathbb{E}\Big[\frac{\lambda(r)\lambda(s)}{\lambda(t)}\Big]\int_{-\infty}^{t} (t-s)e^{-\beta(t-s)}\ \mathrm{d}s\Bigg),\\
&= \frac{\alpha^2 t}{\beta^4}\mathbb{E}\Big[\frac{\lambda(r)\lambda(s)}{\lambda(t)}\Big] = \frac{\alpha^2 t}{\beta^4}\Big(\frac{\beta\mu}{\beta-\alpha}\Big),\\
&= \frac{\alpha^2\mu t}{\beta^3(\beta-\alpha)}.
\end{aligned}
$$

Finally, it is possible to see,

$$
\begin{aligned}
\mathbb{E}\Big[-\frac{\partial^2 \ell}{\partial \alpha\partial \beta}\Big] &= t\Bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\frac{\partial \lambda(t)}{\partial \alpha}\frac{\partial \lambda(t)}{\partial \beta}\Big]\Bigg),\\
&= t\Bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big(\int_{-\infty}^{t} e^{-\beta(t-s)}\ \mathrm{d}N(s)\Big)\Big(\int_{-\infty}^{t} -\alpha(t-s)e^{-\beta(t-s)}\ \mathrm{d}N(s)\Big)\Big]\Bigg),\\
&= -\alpha t\Bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\int_{-\infty}^{t}\int_{-\infty}^{t} (t-r)e^{-\beta(t-r)}e^{-\beta(t-s)}\ \mathrm{d}N(r)\ \mathrm{d}N(s)\Big]\Bigg),\\
&= -\alpha t\Bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big]\int_{-\infty}^{t}\int_{-\infty}^{t} (t-r)e^{-\beta(t-r)}e^{-\beta(t-s)}\mathbb{E}[\mathrm{d}N(r)]\mathbb{E}[\mathrm{d}N(s)]\Bigg),\\
&= -\alpha t\Bigg(\mathbb{E}\Big[\frac{1}{\lambda(t)}\Big]\int_{-\infty}^{t}\int_{-\infty}^{t} (t-r)e^{-\beta(t-r)}e^{-\beta(t-s)}\mathbb{E}[\lambda(r)]\mathbb{E}[\lambda(s)]\ \mathrm{d}r\ \mathrm{d}s\Bigg),\\
&= -\alpha t\Bigg(\mathbb{E}\Big[\frac{\lambda(r)\lambda(s)}{\lambda(t)}\Big]\int_{-\infty}^{t}\Big(\int_{-\infty}^{t} (t-r)e^{-\beta(t-r)}\ \mathrm{d}r\Big)e^{-\beta(t-s)}\ \mathrm{d}s\Bigg),\\
&= -\alpha t\Bigg(\mathbb{E}\Big[\frac{\lambda(r)\lambda(s)}{\lambda(t)}\Big]\int_{-\infty}^{t} \frac{1}{\beta^2}e^{-\beta(t-s)}\ \mathrm{d}s\Bigg),\\
&= -\alpha t\Bigg(\mathbb{E}\Big[\frac{\lambda(r)\lambda(s)}{\lambda(t)}\Big]\Big(\frac{1}{\beta^3}\Big)\Bigg) = -\frac{\alpha t}{\beta^3}\Bigg(\mathbb{E}\Big[\frac{\lambda(r)\lambda(s)}{\lambda(t)}\Big]\Bigg),\\
&= -\frac{\alpha t}{\beta^3}\Big(\frac{\beta\mu}{\beta-\alpha}\Big),\\
&= -\frac{\alpha\mu t}{\beta^2(\beta-\alpha)}.
\end{aligned}
$$

This concludes all the unique cases and thus completes the derivation of the Fisher information for a one-dimensional Hawkes process.

The Fisher information matrix can therefore be expressed as,

$$I(\boldsymbol{\theta}) = \begin{pmatrix} \frac{(\beta-\alpha)t}{\beta\mu} & \frac{t}{\beta} & -\frac{\alpha t}{\beta^2} \\ \\ \frac{t}{\beta} & \frac{\mu t}{\beta(\beta-\alpha)} + \frac{t}{2(\beta+\alpha)} & -\frac{\alpha\mu t}{\beta^2(\beta-\alpha)} \\ \\ -\frac{\alpha t}{\beta^2} & -\frac{\alpha\mu t}{\beta^2(\beta-\alpha)} & \frac{\alpha^2\mu t}{\beta^3(\beta-\alpha)} \end{pmatrix} = \frac{t}{\beta}\begin{pmatrix} \frac{\beta-\alpha}{\mu} & 1 & -\frac{\alpha}{\beta} \\ \\ 1 & \frac{\mu}{\beta-\alpha} + \frac{\beta}{2(\beta+\alpha)} & -\frac{\alpha\mu}{\beta(\beta-\alpha)} \\ \\ -\frac{\alpha}{\beta} & -\frac{\alpha\mu}{\beta(\beta-\alpha)} & \frac{\alpha^2\mu}{\beta^2(\beta-\alpha)} \end{pmatrix}. \tag{A.19}$$

with determinant

$$\det(I(\boldsymbol{\theta})) = \frac{\alpha^2 t^3}{\beta^6}\left(\frac{\beta^5 + (2\mu-\alpha)\beta^4 + (2\alpha\mu - 4\mu - 1)\beta^3 + (2+\alpha-4\alpha\mu)\beta^2 - 2\alpha^2}{2(\alpha^2-\beta^2)}\right). \tag{A.20}$$

The Fisher can also be shown to follow for a non-stationary Hawkes process,

$$I(\boldsymbol{\theta}) = \frac{t}{\beta^2}\begin{pmatrix} \beta^2\mathbb{E}[\lambda^{-1}(t)] & \beta & -\alpha \\ \\ \beta & \mathbb{E}[\lambda(t)] + \frac{\beta^2}{2(\beta+\alpha)} & -\eta\mathbb{E}[\lambda(t)] \\ \\ -\alpha & -\eta\mathbb{E}[\lambda(t)] & \eta^2\mathbb{E}[\lambda(t)] \end{pmatrix}, \tag{A.21}$$

where $\eta = \alpha/\beta$. This can then yield the general determinant,

$$\begin{aligned} \det(I(\boldsymbol{\theta})) = \frac{\alpha^2 t^3}{\beta^6}\Bigg(\mathbb{E}[\lambda^{-1}(t)](\mathbb{E}[\lambda(t)])^2 + \left(\frac{\beta^2}{2(\beta+\alpha)} - 1\right)\mathbb{E}[\lambda^{-1}(t)]\mathbb{E}[\lambda(t)] \\ - (\beta^2 - 2\beta + 1)\mathbb{E}[\lambda(t)] - \frac{\beta^4}{2(\beta+\alpha)}\Bigg). \end{aligned} \tag{A.22}$$

and in the one-dimensional stationary case,

$$\det(I(\boldsymbol{\theta})) = t^3\mathbb{E}[\lambda(t)]\left(\frac{\alpha^2}{2\mu\beta^5}\left(\beta^2 + 2\mu - 1\right) - \frac{\alpha^2}{2\mu\beta^4(\alpha+\beta)}\left(\beta^2 - \mu(\alpha+\beta) - 1\right) + \frac{\alpha^3}{\beta^6}\right).$$

Rewriting the equation in this form, clearly illustrates how the determinant relates to $\mathbb{E}[\lambda(t)]$. Furthermore, as α and β must always be positive, the determinant must always be non-negative.

A.5 Derivation of the M-Dimensional Fisher Information

The derivation of the M-dimensional Fisher information will follow a similar procedure to above. In this case, the log-likelihood should be recalled as,

$$\ell(t|\boldsymbol{\theta}) = \sum_{m=1}^{M} \ell_m(t|\theta_m),$$

where $\ell_m(t|\theta_m)$ may be expressed as,

$$\ell_m(t|\theta_m) = \sum_{k=1}^{N} \log\left(\mu_m + \sum_{i=1}^{M} \alpha_{i,m} R_{i,m}(k)\right) - \mu_m t - \sum_{i=1}^{M}\sum_{k=1}^{N} \frac{\alpha_{i,m}}{\beta_{i,m}}\left(1 - e^{-\beta_{i,m}(t-t_{k,i})}\right).$$

The recursive formula employed here is

$$R_{i,m}(k) = \sum_{d:t_{d,i}<t_{k,m}} e^{-\beta_{i,m}(t_{k,m}-t_{d,i})}, \text{ for } k \geq 2,$$

with $R_{i,m}(0) = 0$. Further, note that the intensity for any of the m dimensions may be expressed as

$$\lambda_m(t) = \mu_m + \sum_{i=1}^{M} \int_{-\infty}^{t} \alpha_{i,m} e^{-\beta_{i,m}(t-s)} \, \mathrm{d}N_i(s). \tag{A.23}$$

This choice of intensity makes no assumptions about the parameters μ, α and β. More importantly, this choice allows for asymmetric parameter matrices which will be utilised in keeping the derivation as general as possible.

It is assumed throughout that M is the number of dimensions with some $i, m \in \{1, 2..., M\}$. It is possible to see there are fifteen unique cases for the M-dimensional Fisher information. Before deriving any cases, it is possible to use the derivation of the M-dimensional Hessian matrix to avoid any further calculation of specific derivatives. Note that,

$$\mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \mu_i \partial \mu_{i'}}\right] = 0, \tag{A.24}$$

$$\mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \alpha_{i,m} \partial \mu_{i'}}\right] = 0, \tag{A.25}$$

$$\mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \alpha_{i,m} \partial \alpha_{i',m'}}\right] = 0, \tag{A.26}$$

$$\mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \beta_{i,m} \partial \mu_{i'}}\right] = 0, \tag{A.27}$$

$$\mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \alpha_{i,m} \partial \beta_{i',m'}}\right] = 0, \tag{A.28}$$

$$\mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \beta_{i,m} \partial \beta_{i',m'}}\right] = 0, \tag{A.29}$$

where $i \neq i'$ and $m, m' \in \{1, 2, .., M\}$. It is now possible to look at the remaining nine cases. Consider any of the m dimensions on the interval $[0, t]$.

$$\begin{aligned}
\mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \mu_i^2}\right] &= t\left(\mathbb{E}\left[\frac{1}{\lambda_i(t)} \frac{\partial \lambda_i(t)}{\partial \mu_i} \frac{\partial \lambda_i(t)}{\partial \mu_i}\right]\right) = t\left(\mathbb{E}\left[\frac{1}{\lambda_i(t)} \left(\frac{\partial \lambda_i(t)}{\partial \mu_i}\right)^2\right]\right), \\
&= t\left(\mathbb{E}\left[\frac{1}{\lambda_i(t)} (1)^2\right]\right) = t\left(\mathbb{E}[\lambda_i^{-1}(t)]\right), \\
&= t\left(\frac{1 - \sum_{j=1}^{M} \frac{\alpha_{j,i}}{\beta_{j,i}}}{\mu_i}\right).
\end{aligned}$$

Next,

$$
\begin{aligned}
\mathbb{E}\Big[-\frac{\partial^2 \ell}{\partial \alpha_{i,m}^2}\Big] &= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\frac{\partial \lambda_i(t)}{\partial \alpha_{i,m}}\frac{\partial \lambda_i(t)}{\partial \alpha_{i,m}}\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\Big(\frac{\partial \lambda_i(t)}{\partial \alpha_{i,m}}\Big)^2\Big]\bigg) = t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\Big(\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}\ \mathrm{d}N_i(s)\Big)^2\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\int_{-\infty}^{t}\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}e^{-\beta_{i,m}(t-r)}\ \mathrm{d}N_i(s)\ \mathrm{d}N_i(r)\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\Big]\int_{-\infty}^{t}\Big(\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}\mathbb{E}[\lambda_i(s)]\ \mathrm{d}s\Big)e^{-\beta_{i,m}(t-r)}\mathbb{E}[\lambda_i(r)]\ \mathrm{d}r\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{\lambda_i(s)\lambda_i(r)}{\lambda_i(t)}\Big]\int_{-\infty}^{t}\Big(\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}\ \mathrm{d}s\Big)e^{-\beta_{i,m}(t-r)}\ \mathrm{d}r\bigg),\\
&= t\bigg(\mathbb{E}[\lambda_i(t)]\int_{-\infty}^{t}\Big(\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}\ \mathrm{d}s\Big)e^{-\beta_{i,m}(t-r)}\ \mathrm{d}r\bigg),\\
&= t\bigg(\mathbb{E}[\lambda_i(t)]\int_{-\infty}^{t}\Big(\frac{1}{\beta_{i,m}}\Big)e^{-\beta_{i,m}(t-r)}\ \mathrm{d}r\bigg) = \frac{t}{\beta_{i,m}^2}\bigg(\mathbb{E}[\lambda_i(t)]\bigg),\\
&= \frac{t}{\beta_{i,m}^2}\bigg(\frac{\mu_i}{1-\sum_{j=1}^{M}\frac{\alpha_{j,i}}{\beta_{j,i}}}\bigg).
\end{aligned}
$$

It is necessary to recall that the notation i' means any value in $\{1, 2, ..., M\}$ which is not i.

$$
\begin{aligned}
\mathbb{E}\Big[-\frac{\partial^2 \ell}{\partial \alpha_{i,m}\partial \alpha_{i,m'}}\Big] &= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\frac{\partial \lambda_i(t)}{\partial \alpha_{i,m}}\frac{\partial \lambda_i(t)}{\partial \alpha_{i,m'}}\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\int_{-\infty}^{t}\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}e^{-\beta_{i,m'}(t-r)}\ \mathrm{d}N_i(s)\ \mathrm{d}N_i(r)\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\Big]\int_{-\infty}^{t}\Big(\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}\mathbb{E}[\lambda_i(s)]\ \mathrm{d}s\Big)e^{-\beta_{i,m'}(t-r)}\mathbb{E}[\lambda_i(r)]\ \mathrm{d}r\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{\lambda_i(s)\lambda_i(r)}{\lambda_i(t)}\Big]\int_{-\infty}^{t}\Big(\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}\ \mathrm{d}s\Big)e^{-\beta_{i,m'}(t-r)}\ \mathrm{d}r\bigg),\\
&= t\bigg(\mathbb{E}[\lambda_i(t)]\int_{-\infty}^{t}\Big(\frac{1}{\beta_{i,m}}\Big)e^{-\beta_{i,m'}(t-r)}\ \mathrm{d}r\bigg) = \frac{t}{\beta_{i,m}\beta_{i,m'}}\bigg(\mathbb{E}[\lambda_i(t)]\bigg),\\
&= \frac{t}{\beta_{i,m}\beta_{i,m'}}\bigg(\frac{\mu_i}{1-\sum_{j=1}^{M}\frac{\alpha_{j,i}}{\beta_{j,i}}}\bigg).
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\Big[-\frac{\partial^2 \ell}{\partial \alpha_{i,m}\partial \mu_i}\Big] &= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\frac{\partial \lambda_i(t)}{\partial \alpha_{i,m}}\frac{\partial \lambda_i(t)}{\partial \mu_i}\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}(1)\Big(\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}\ \mathrm{d}N_i(s)\Big)\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\Big]\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}\mathbb{E}[\ \mathrm{d}N_i(s)]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\Big]\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}\mathbb{E}[\lambda_i(t)]\ \mathrm{d}s\bigg),\\
&= t\bigg(\int_{-\infty}^{t} e^{-\beta_{i,m}(t-s)}\ \mathrm{d}s\bigg) = t\bigg(\frac{1}{\beta_{i,m}}\bigg),\\
&= \frac{t}{\beta_{i,m}}.
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\Big[-\frac{\partial^2 \ell}{\partial \beta_{i,m}\partial \mu_i}\Big] &= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\frac{\partial \lambda_i(t)}{\partial \beta_{i,m}}\frac{\partial \lambda_i(t)}{\partial \mu_i}\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}(1)\Big(-\alpha_{i,m}\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m}(t-s)}\,\mathrm{d}N_i(s)\Big)\Big]\bigg),\\
&= -\alpha_{i,m}t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\Big]\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m}(t-s)}\mathbb{E}[\,\mathrm{d}N_i(s)]\bigg),\\
&= -\alpha_{i,m}t\bigg(\mathbb{E}\Big[\frac{\lambda_i(s)}{\lambda_i(t)}\Big]\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m}(t-s)}\,\mathrm{d}s\bigg),\\
&= -\alpha_{i,m}t\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m}(t-s)}\,\mathrm{d}s,\\
&= -\frac{\alpha_{i,m}t}{\beta_{i,m}^2}.
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\Big[-\frac{\partial^2 \ell}{\partial \alpha_{i,m}\partial \beta_{i,m}}\Big] &= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\frac{\partial \lambda_i(t)}{\partial \alpha_{i,m}}\frac{\partial \lambda_m(t)}{\partial \beta_{i,m}}\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\Big(\int_{-\infty}^{t}e^{-\beta_{i,m}(t-r)}\,\mathrm{d}N_i(r)\Big)\Big(-\alpha_{i,m}\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m}(t-s)}\,\mathrm{d}N_i(s)\Big)\Big]\bigg),\\
&= -\alpha_{i,m}t\bigg(\mathbb{E}\Big[\frac{\lambda_i(r)\lambda_i(s)}{\lambda_i(t)}\Big]\Big(\int_{-\infty}^{t}\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m}(t-s)}e^{-\beta_{i,m}(t-r)}\,\mathrm{d}s\,\mathrm{d}r\Big)\bigg),\\
&= -\alpha_{i,m}t\bigg(\mathbb{E}[\lambda_i(t)]\Big(\int_{-\infty}^{t}\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m}(t-s)}e^{-\beta_{i,m}(t-r)}\,\mathrm{d}s\,\mathrm{d}r\Big)\bigg),\\
&= -\alpha_{i,m}t\bigg(\frac{\mu_i}{1-\sum_{j=1}^{M}\frac{\alpha_{j,i}}{\beta_{j,i}}}\bigg)\bigg(\frac{1}{\beta_{i,m}^3}\bigg) = -\frac{\alpha_{i,m}t}{\beta_{i,m}^3}\bigg(\frac{\mu_i}{1-\sum_{j=1}^{M}\frac{\alpha_{j,i}}{\beta_{j,i}}}\bigg).
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\Big[-\frac{\partial^2 \ell}{\partial \alpha_{i,m}\partial \beta_{i,m'}}\Big] &= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\frac{\partial \lambda_i(t)}{\partial \alpha_{i,m}}\frac{\partial \lambda_i(t)}{\partial \beta_{i,m'}}\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\Big(\int_{-\infty}^{t}e^{-\beta_{i,m}(t-r)}\,\mathrm{d}N_i(r)\Big)\Big(-\alpha_{i,m'}\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m'}(t-s)}\,\mathrm{d}N_i(s)\Big)\Big]\bigg),\\
&= -\alpha_{i,m'}t\bigg(\mathbb{E}\Big[\frac{\lambda_i(r)\lambda_i(s)}{\lambda_i(t)}\Big]\Big(\int_{-\infty}^{t}\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m'}(t-s)}e^{-\beta_{i,m}(t-r)}\,\mathrm{d}s\,\mathrm{d}r\Big)\bigg),\\
&= -\alpha_{i,m'}t\bigg(\frac{\mu_i}{1-\sum_{j=1}^{M}\frac{\alpha_{j,i}}{\beta_{j,i}}}\bigg)\bigg(\frac{1}{\beta_{i,m'}^2\beta_{i,m}}\bigg) = -\frac{\alpha_{i,m'}t}{\beta_{i,m'}^2\beta_{i,m}}\bigg(\frac{\mu_i}{1-\sum_{j=1}^{M}\frac{\alpha_{j,i}}{\beta_{j,i}}}\bigg).
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\Big[-\frac{\partial^2 \ell}{\partial \beta_{i,m}^2}\Big] &= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\frac{\partial \lambda_i(t)}{\partial \beta_{i,m}}\frac{\partial \lambda_i(t)}{\partial \beta_{i,m}}\Big]\bigg) = t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\Big(\frac{\partial \lambda_i(t)}{\partial \beta_{i,m}}\Big)^2\Big]\bigg),\\
&= t\bigg(\mathbb{E}\Big[\frac{1}{\lambda_i(t)}\Big(-\alpha_{i,m}\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m}(t-s)}\,\mathrm{d}N_i(s)\Big)^2\Big]\bigg),\\
&= \alpha_{i,m}^2t\bigg(\mathbb{E}\Big[\frac{\lambda_i(r)\lambda_i(s)}{\lambda_i(t)}\Big]\Big(\int_{-\infty}^{t}\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m}(t-s)}(t-r)e^{-\beta_{i,m}(t-r)}\,\mathrm{d}s\,\mathrm{d}r\Big)\bigg),\\
&= \alpha_{i,m}^2t\bigg(\frac{\mu_i}{1-\sum_{j=1}^{M}\frac{\alpha_{j,i}}{\beta_{j,i}}}\bigg)\bigg(\frac{1}{\beta_{i,m}^4}\bigg),\\
&= \frac{\alpha_{i,m}^2t}{\beta_{i,m}^4}\bigg(\frac{\mu_i}{1-\sum_{j=1}^{M}\frac{\alpha_{j,i}}{\beta_{j,i}}}\bigg).
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\bigg[-\frac{\partial^2 \ell}{\partial \beta_{i,m}\partial \beta_{i,m'}}\bigg] &= t\Bigg(\mathbb{E}\bigg[\frac{1}{\lambda_i(t)}\frac{\partial \lambda_i(t)}{\partial \beta_{i,m}}\frac{\partial \lambda_i(t)}{\partial \beta_{i,m'}}\bigg]\Bigg),\\
= t\Bigg(\mathbb{E}\bigg[\frac{1}{\lambda_i(t)}\bigg(-\alpha_{i,m}\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m}(t-s)}\ \mathrm{d}N_i(s)\bigg)&\bigg(-\alpha_{i,m'}\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m'}(t-s)}\ \mathrm{d}N_i(s)\bigg)\bigg]\Bigg),\\
&= \alpha_{i,m}\alpha_{i,m'}t\Bigg(\mathbb{E}\bigg[\frac{\lambda_i(r)\lambda_i(s)}{\lambda_i(t)}\bigg]\bigg(\int_{-\infty}^{t}\int_{-\infty}^{t}(t-s)e^{-\beta_{i,m}(t-s)}(t-r)e^{-\beta_{i,m'}(t-r)}\ \mathrm{d}s\ \mathrm{d}r\bigg)\Bigg),\\
&= \alpha_{i,m}\alpha_{i,m'}t\bigg(\frac{\mu_i}{1-\sum_{j=1}^{M}\frac{\alpha_{j,i}}{\beta_{j,i}}}\bigg)\bigg(\frac{1}{\beta_{i,m}^2\beta_{i,m'}^2}\bigg),\\
&= \frac{\alpha_{i,m}\alpha_{i,m'}t}{\beta_{i,m}^2\beta_{i,m'}^2}\bigg(\frac{\mu_i}{1-\sum_{j=1}^{M}\frac{\alpha_{j,i}}{\beta_{j,i}}}\bigg).
\end{aligned}
$$

This completes all the unique cases for the M-dimensional Fisher information matrix.

A.6 Simulation of Hawkes Processes

The algorithm provided in this section, modifies and simplifies the original notation given when it was first presented [Lim et al., 2016].

Algorithm 2 Exact Simulation of M-Dimensional Hawkes Process

```
procedure GENERATE EVENT TIMES
INPUT: $\mu_m$, $\alpha_{i,m}(0)$, $\beta_{i,m}$, $N_m(0)$, $t_{\max}$ and the distribution of $\alpha_{i,m}(j)$.
Initialise $r_0 = 0$.
    for $m = 1, ..., M$
        for $i = 1, ..., M$
            $\lambda_{i,m}(r_0) = \alpha_{i,m}(0)$.
        END for
    END for
Repeat for $j = 1, 2, ...$
    for $m = 1, ..., M$
        Sample $a_{0,m} \sim \mathrm{Exp}(\mu_m)$.
        for $i = 1, ..., M$
            Generate $u \sim \mathrm{Unif}(0, 1)$.
            if $u < 1 - \exp\{-\frac{1}{\delta_{i,m}}\lambda_{i,m}(r_{j-1})\}$
                $a_{i,m} = -\frac{1}{\delta_{i,m}}\log(1 + \frac{\delta_{i,m}}{\lambda_{i,m}(r_{j-1})}\log(1-u))$,
            else
                $a_{i,m} = \infty$.
                END if
        END for
    END for
    $d_j = \min_{i,m} a_{i,m}$.
    $r_j = r_{j-1} + d_j$.
    $(Z_j, X_j) = (m^*, i^*) = \arg\min_{i,m} a_{i,m}$.
    for $m = 1, ..., M$
        Sample $\alpha_{m^*,m}(j)$. If distribution constant then $\alpha_{m^*,m}(j) = \alpha_{m^*,m}(0)\ \ \forall j$.
        $\lambda_{i,m}(r_j) = \lambda_{i,m}(r_{j-1})\exp\{-\delta_{i,m}d_j\} + \alpha_{m^*,m}(j)\mathbb{1}_{i=m^*}$.
        $N_m(r_j) = N_m(r_{j-1}) + \mathbb{1}_{m=m^*}$.
    END for
    $t_{k,m^*} = r_j$ for $k = N_{m^*}(r_j)$.
Until: $r_j > t_{\max}$.
Discard the last $r_j$.
OUTPUT: $r$, $N_m$, $t_m$, and $\lambda_m$.
end procedure
```