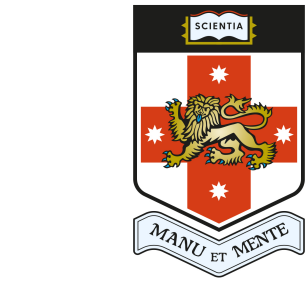

UNSW
SYDNEY

Statistical inference for renewal Hawkes self-exciting point processes

Tom Stindl
School of Mathematics and Statistics
University of New South Wales

A thesis in the fulfilment of the requirements for the degree of
Doctor of Philosophy
September 2019

Thesis/Dissertation Sheet

Surname/Family Name	:	**Stindl**
Given Name/s	:	**Thomas**
Abbreviation for degree as give in the University calendar	:	**PhD**
Faculty	:	**Science**
School	:	**Mathematics and Statistics**
Thesis Title	:	**Statistical inference for renewal Hawkes self-exciting point processes**

Abstract 350 words maximum:

The class of self-exciting point process evolve within a self-excitation mechanism that allows past events to contribute to the arrival rate of future events. The significant contributions this thesis introduces are techniques to conduct efficient statistical inferences for the recently proposed renewal Hawkes self-exciting point processes. By employing a substantial modification to the baseline arrival rate of the Hawkes process, the renewal Hawkes process provides superior versatility. The additional flexibility afforded to the renewal Hawkes process occurs by defining the immigration process in terms of a general renewal process rather than a homogenous Poisson process. The renewal Hawkes process has the potential to widen the application domains of self-exciting processes significantly. However, it was initially asserted that likelihood evaluation of the process demands exponential computational time and therefore is practically impossible. As a consequence, two Expectation-Maximization (E-M) algorithms were developed to compute the maximum likelihood estimator (MLE), a bootstrap procedure to estimate the variance-covariance matrix of the MLE and a Monte Carlo approach to compute a goodness-of-fit test statistic.

Considering the fundamental role played by the likelihood function in statistical inferences, a practically feasible method for likelihood evaluation is highly desirable. This thesis develops algorithms to evaluate the likelihood of the renewal Hawkes process in quadratic time, a drastic improvement from the exponential time initially claimed. Simulations will demonstrate the superior performance of the resulting MLEs of the model relative to the E-M estimators. This thesis will also introduce computationally efficient procedures to calculate the Rosenblatt residuals of the process for goodness-of-fit assessment and a simple yet efficient procedure for future event predictions. Faster fitting methods, and linear time algorithms to fit the process are also discussed. The computational efficiency of the methods developed facilitates the application of these algorithms to multi-event and marked point process models with renewal immigration. As such, this thesis proposes two additional models termed the multivariate renewal Hawkes process and the mark renewal Hawkes process. The additional computational challenges that arise in these frameworks are solved herein.

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed ……..

Date ……..

INCLUSION OF PUBLICATIONS STATEMENT

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in their thesis in lieu of a Chapter if:

- The student contributed greater than 50% of the content in the publication and is the "primary author", ie. the student was responsible primarily for the planning, execution and preparation of the work for publication
- The student has approval to include the publication in their thesis in lieu of a Chapter from their supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis

Please indicate whether this thesis contains published material or not.

☐ *This thesis contains no publications, either published or submitted for publication (if this box is checked, you may delete all the material on page 2)*

☒ *Some of the work described in this thesis has been published and it has been documented in the relevant Chapters with acknowledgement (if this box is checked, you may delete all the material on page 2)*

☐ *This thesis has publications (either published or submitted for publication) incorporated into it in lieu of a chapter and the details are presented below*

CANDIDATE'S DECLARATION

I declare that:

- I have complied with the Thesis Examination Procedure
- where I have used a publication in lieu of a Chapter, the listed publication(s) below meet(s) the requirements to be included in the thesis.

Name	**Signature**	**Date (dd/mm/yy)**

Postgraduate Coordinator's Declaration (to be filled in where publications are used in lieu of Chapters)

I declare that:

- the information below is accurate
- where listed publication(s) have been used in lieu of Chapter(s), their use complies with the Thesis Examination Procedure
- the minimum requirements for the format of the thesis have been met.

PGC's Name	**PGC's Signature**	**Date (dd/mm/yy)**

Abstract

The class of self-exciting point process evolve within a self-excitation mechanism that allows past events to contribute to the arrival rate of future events. The significant contributions this thesis introduces are techniques to conduct efficient statistical inferences for the recently proposed renewal Hawkes self-exciting point processes. By employing a substantial modification to the baseline arrival rate of the Hawkes process, the renewal Hawkes process provides superior versatility. The additional flexibility afforded to the renewal Hawkes process occurs by defining the immigration process in terms of a general renewal process rather than a homogenous Poisson process. The renewal Hawkes process has the potential to widen the application domains of self-exciting processes significantly. However, it was initially asserted that likelihood evaluation of the process demands exponential computational time and therefore is practically impossible. As a consequence, two Expectation-Maximization (E-M) algorithms were developed to compute the maximum likelihood estimator (MLE), a bootstrap procedure to estimate the variance-covariance matrix of the MLE and a Monte Carlo approach to compute a goodness-of-fit test statistic.

Considering the fundamental role played by the likelihood function in statistical inferences, a practically feasible method for likelihood evaluation is highly desirable. This thesis develops algorithms to evaluate the likelihood of the renewal Hawkes process in quadratic time, a drastic improvement from the exponential time initially claimed. Simulations will demonstrate the superior performance of the resulting MLEs of the model relative to the E-M estimators. This thesis will also introduce computationally efficient procedures to calculate the Rosenblatt residuals of the process for goodness-of-fit assessment and a simple yet efficient procedure for future event predictions. Faster fitting methods, and linear time algorithms to fit the process are also discussed. The computational efficiency of the methods developed facilitates the application of these algorithms to multi-event and marked point process models with renewal immigration. As such, this thesis proposes two additional models termed the multivariate renewal Hawkes process and the mark renewal Hawkes process. The additional computational challenges that arise in these frameworks are solved herein.

Acknowledgements

A significant piece of work such as this is never simply the efforts of one individual but rather a collection of people who have had considerable influence. I need to thank the following individuals for without their contribution this would not have been possible.

First and foremost, my appreciations and gratitude must go to my dedicated supervisor Dr. Feng Chen. For without his patience, commitment, and willingness to engage in the same discussions multiple times, this would not have been achievable. The time and effort that you devoted to our work and my development as a researcher, I will forever be indebted. Secondly, to Prof. William Dunsmuir for providing me with a first taste into research and showing the enjoyment and passion that can be enjoyed with ones work. Most importantly, thank you for suggesting Feng as a PhD supervisor. I will be forever grateful for your help in making one of the best decisions I've made.

To all my friends, you should know that your support and encouragement was worth more than what I can express on paper. However, here it is in on paper. My gratitude must go to my friends who have always been a significant source of support when things became discouraging and would listen to all my complaints and grievances. Thank you for being there for me, even when things got particularly trying.

Finally, my sincere and genuine appreciation to my family for their endless love, help, and support. I am grateful to my brother Ryan and sister Sonja for always supporting me. I am forever indebted to my parents Karl and Yvonne for giving me the opportunities and experiences that have made me who I am today.

This journey would not have been possible if not for all these individuals and their contributions. I dedicate this milestone to you all.

Thank you all,
Tom

Contents

Acknowledgements **x**

1 Introduction to self-exciting point processes **1**
1.1 Motivation . . . 1
1.2 Point process framework . . . 2
1.3 Hawkes self-exciting point processes . . . 3
1.4 Inference for Hawkes self-exciting point processes . . . 4
1.4.1 Estimation . . . 5
1.4.1.1 Maximum likelihood estimation . . . 5
1.4.1.2 Expectation-Maximization algorithm . . . 6
1.4.2 Model assessment . . . 7
1.4.3 Simulation algorithms . . . 8
1.5 Non-stationary self-exciting point processes . . . 9

2 Background on renewal Hawkes processes **11**
2.1 Model and notation . . . 11
2.2 Expectation-Maximization algorithms . . . 12
2.2.1 Expectation step . . . 14
2.2.2 Maximization step . . . 18
2.2.3 E-M algorithm with reduced set of missing data . . . 18
2.3 Statistical inferences . . . 19
2.3.1 Likelihood and variance-covariance matrix evaluation . . . 20
2.3.2 Model assessment . . . 22

3 Direct Likelihood Evaluation of the renewal Hawkes process **23**
3.1 Introduction . . . 23
3.2 Properties of the renewal Hawkes process model . . . 24
3.3 Maximum likelihood estimation . . . 26
3.4 Model assessment . . . 33
3.5 Model predictions . . . 34
3.5.1 Predictive density and hazard function . . . 34

3.5.2 Predictive simulations . . . 34
3.6 Simulations . . . 36
3.6.1 Simulation algorithm . . . 36
3.6.2 Simulation models . . . 36
3.6.3 Simulation results . . . 37
3.6.4 Comparison with the E-M algorithms of Wheatley et al. (2016) 39
3.7 Applications . . . 41
3.7.1 Earthquakes in the Japan Pacific Ring of Fire . . . 41
3.7.2 Mid-price changes of the AUD/USD currency exchange rate . 45

4 Fast fitting of the renewal Hawkes process 50
4.1 Introduction . . . 50
4.2 Estimation using Newton-Raphson optimization . . . 52
4.3 Estimation using approximate likelihood functions . . . 55
4.4 Simulations . . . 59
4.4.1 Simulation models . . . 59
4.4.2 Simulation results . . . 59
4.4.3 Accuracy and speed of the log-likelihood approximation methods 61
4.4.4 Influence of the tuning parameters on the speed and accuracy of log-likelihood approximation . . . 63
4.5 Application . . . 65
4.5.1 Mid-price changes of foreign currency exchange rates . . . 65

5 The multivariate renewal Hawkes process 71
5.1 Introduction . . . 71
5.2 Model and notation . . . 73
5.3 Maximum likelihood estimation . . . 75
5.4 Model assessment . . . 79
5.5 Model predictions . . . 80
5.5.1 Predictive density and hazard function . . . 80
5.5.2 Predictive simulations . . . 81
5.6 Simulations . . . 82
5.6.1 Simulation algorithm . . . 82
5.6.2 Simulation results . . . 82
5.6.3 Comparison with the modified likelihood evaluation algorithm 85
5.6.4 Assessment of the predictive performance . . . 86
5.7 Applications . . . 87
5.7.1 Analysis of earthquakes around Fiji and Vanuatu . . . 87
5.7.2 Modeling trade-throughs using bivariate RHawkes processes . 92

6 Modeling Extreme Negative returns using marked renewal Hawkes processes **98**
6.1 Introduction . . . 98
6.2 ASX stock data . . . 100
6.3 Model and methodologies . . . 101
6.3.1 Marked renewal Hakwes process . . . 101
6.3.2 Likelihood evaluation algorithm . . . 103
6.3.3 Excess modeling . . . 105
6.3.4 Model assessment . . . 106
6.3.5 Predicting exceedances . . . 107
6.3.6 Forecasting conditional risk measures . . . 108
6.4 Simulation study . . . 110
6.4.1 Simulation algorithm . . . 110
6.4.2 Simulation model . . . 111
6.4.3 Results . . . 111
6.5 Modeling extreme negative returns . . . 114

7 Conclusion **123**
7.1 Summary . . . 123
7.2 Perspective on future work . . . 125

Appendix A Derivatives of the most recent immigrant probabilities in the RHawkes process **127**
A.1 First derivative of the most recent immigrant probabilities . . . 127
A.2 Second derivative of the most recent immigrant probabilities . . . 128

References **130**

Chapter 1

Introduction to self-exciting point processes

The self-exciting point process is a natural and powerful statistical model to illustrate the temporal patterns of the occurrence times of certain events. This introductory chapter provides a formal introduction to these processes and defines a framework in which statistical inferences such as estimation, model assessment, and prediction can be implemented herein. Furthermore, this chapter provides a concise review of the important and influential self-exciting point processes that have been developed thus far, and the standard inferential methods that are used in practice.

1.1 Motivation

Point process models are stochastic processes that model the occurrence of points that occur in either time or space. A point process on the real line is generally interpreted as 'time' and the points as 'events'. The Hawkes point process was a significant advancement in the field of point processes since it provided a valuable avenue in which to describe sequences of events occurring at random times that demonstrate temporal clustering. Following its proposal by Hawkes (1971), this model and its diverse extensions have been applied to analyze data arising in an extensive range of fields such as seismology (Ogata, 1988), neuroscience (Chornoboy et al., 1988), finance (McNeil et al., 2005; Chavez-Demoulin et al., 2005; Embrechts et al., 2011; Filimonov and Sornette, 2012), social interaction modelling (Crane and Sornette, 2008), credit risk (Errais et al., 2010), genome analysis (Reynaud-Bouret and Schbath, 2010), criminology (Mohler et al., 2011), terrorist activity modeling (Porter and White, 2012) among many others.

1.2 Point process framework

This section primarily focuses on temporal point processes on the positive real line $\mathbb{R}^+$. The extensions to marked and multivariate processes are defined at the beginning of the appropriate chapter in this thesis. The following definitions and notations follow closely the presentation adopted in Daley and Vere-Jones (2003, 2008). Let $\mathcal{N}$ be the space of all boundedly finite integer-valued measures termed counting measures, and $\mathcal{B}(\mathcal{N})$ be the σ-algebra generated by the sets $\{m \in \mathcal{N} : m(A) = n\}$, $A \in \mathcal{B}(\mathbb{R}^+)$, $n \in \mathbb{N}_0$ where $\mathbb{N}_0$ denotes the natural numbers including zero. A point process on the positive real line $\mathbb{R}^+$ is a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into $(\mathcal{N}, \mathcal{B}(\mathcal{N}))$. Hence, any random variable of the form,

$$N : (\Omega, \mathcal{F}) \to (\mathcal{N}, \mathcal{B}(\mathcal{N})), \tag{1.2.1}$$

is a point process on $\mathbb{R}^+$. In particular, the function $N((a, b]) : (\Omega, \mathcal{F}) \to \mathbb{N}_0$ for $0 \leqslant a < b < \infty$, is a random variable that counts the number of events in the time interval $(a, b]$. A notable assumption often used is that only one point can be observed in any instant of time, that is, $\mathbb{P}\left(N(\{t\}) \in \{0, 1\}\right) = 1$. In this case, the point process N is termed a *simple* point process. In less general terms, a simple temporal point process consists of a sequence of event times in ascending order such that $\tau_1 < \tau_2 < \ldots$ The counting process $N(t), t \geqslant 0$ associated with the time series of occurrence times counts the number of events that have occurred up to and including time t,

$$N(t) = \sum_{i=1}^{\infty} 1\{\tau_i \leqslant t\}, \quad t \geqslant 0.$$

For instance, consider a simple point process N such that $N(A_i)$ has a Poisson distribution with mean $\int_{A_i} \lambda(t)\mathrm{d}t$, $i = 1, \ldots, n$, where each $N(A_i)$ are independent for all mutually disjoint sets $A_i \in \mathcal{B}(\mathbb{R}^+)$ and $\lambda : \mathbb{R}^+ \to \mathbb{R}^+$ is locally integrable. Suppose that λ is a constant function that does not depend on t, then this is known as a stationary process, and termed a homogenous Poisson process with rate $\lambda \in (0, \infty)$. The points of a homogenous Poisson process exhibit complete randomness in A_i. For example, if n points occur in $A_i \in \mathcal{B}(\mathbb{R}^+)$, then these n points are independent and uniformly distributed over the set A_i. This is the most well-known point process. However, it may not be sufficient for many real-world data applications, in which events cluster heavily over time or exhibit more regular temporal patterns. The clustering phenomenon of points has necessitated a generalization of the homogenous Poisson process. In this generalization, the arrival rate λ is no longer a constant but rather a deterministic function of time $\lambda(t)$. This is termed an inhomogeneous Poisson process and can model the seasonality patterns present in the points.

For both the homogenous and inhomogeneous Poisson process, the arrival rate of future points does not depend on the occurrence of points in the past. Since Poisson processes do not account for the history of the process, they might also not be adequate for particular applications. A common tool to allow the history to influence the future evolution of the process is to introduce a self-exciting mechanism that depends on the points in the past. Before discussing such an extension, the notion of history needs to be properly defined. The history at time t contains the cumulative information up to and including time t. More formally, let $\mathcal{F} = \{\mathcal{F}_t;\ t \geqslant 0\}$ with $\mathcal{F}_t = \sigma\{N(s); s \leqslant t\}$, denote the natural filtration of the point process, that is, the $\sigma-$algebra generated by the counting process N.

The dynamical evolution of history-dependent point process models are defined in terms of their (conditional) intensity process $\lambda(t), t \geqslant 0$ which takes the form,

$$\lambda(t) := \lim_{\delta \to 0} \frac{\mathbb{E}\left[N(t+\delta) - N(t)|\mathcal{F}_{t-}\right]}{\delta}, \qquad t \in \mathbb{R}^+, \tag{1.2.2}$$

where $\mathcal{F}_{t-}$ denotes the history just before time t. The intensity process indicates the instantaneous arrival rate of points at time t. The form of the intensity process affords versatility in the modeling capabilities of point process models and leads to different point processes. For instance, the intensity may not depend on $\mathcal{F}_{t-}$ such as the homogenous Poisson process in which the intensity process is just the constant λ, or it may depend on time, in which $\lambda(t)$ is a deterministic function as in the inhomogeneous Poisson process.

For the history-dependent point process, the history of the process, namely its past events, generate a self-exciting or autoregressive mechanism. The Hawkes (1971) process is the pioneering self-exciting point process of this type. The self-exciting mechanism allows the intensity to momentarily increases on the arrival of a point with this effect usually diminishing over time. The focus of this thesis surrounds self-exciting point processes and in the next section, a brief introduction to the Hawkes process is presented.

1.3 Hawkes self-exciting point processes

The class of self-exciting point process models provides a framework for modeling sequences of events whose arrival rate depends on the occurrence times of earlier events. The dependence on the history of the process takes the form of a self-exciting phenomenon in which the cumulative effects of previous events increase the intensity for future events. Such effects will decay over time, with the most recent events before time t having the most significant contribution to the intensity. The most notable self-exciting point process is that of Hawkes (1971). The Hawkes process is a

clustering process, in which one can interpret the points of the process as immigrant or offspring events. The immigrant events follow a homogenous Poisson process, and then when these individuals enter the population, whether by immigration or by birth, they begin to give birth to offspring of their own according to independent inhomogeneous Poisson processes with a common intensity function.

The branching process interpretation can be made more explicit by examination of the Hawkes (conditional) intensity process. Let $N(t)$, $t \geqslant 0$ be a simple point process on the positive half-line so that all its points are distinct and interpretable as event times. Let the event times in ascending order be denoted by $\tau_1 < \tau_2 < \dots$. The intensity process for the Hawkes process with respect to its natural filtration $\mathcal{F}_t$ takes the form,

$$\lambda(t) = \mu + \sum_{j=1}^{N(t-)} \eta h(t - \tau_j), \tag{1.3.1}$$

for a positive constant $\mu \in (0, \infty)$, a constant $\eta \in [0, 1)$, and a positive function $h(\cdot)$ on $\mathbb{R}^+$ such that $\int_0^\infty h(t)\mathrm{d}t = 1$. The constant μ is interpreted as the hazard function of the independent and identically distributed (*i.i.d.*) waiting times between immigration events, also known as the *baseline intensity* or *background rate*. The constant η is referred to as the *branching ratio* and indicates the mean number of offspring for an individual in the population. The function $h(\cdot)$ is a probability density function often termed the *offspring density*. The function $\eta h(\cdot)$ is also known as the *excitation function*, the *infectivity function*, the *fertility function*, or the *memory kernel*. The linear form of the intensity process in (1.3.1), indicates that at any time t, the instantaneous event rate of the Hawkes process is the immigrant arrival rate plus the sum of birthing rates of existing members of the population. The constant baseline intensity describes the arrival of exogenous events, which arrive independently of previous events.

1.4 Inference for Hawkes self-exciting point processes

This section outlines estimation for the Hawkes process based on observations over the interval $[0, T]$. The objective is to estimate the model parameters which specify the (conditional) intensity process $\lambda(t)$ of the Hawkes process and uniquely determine the distribution of the counting process N. This exposition focuses on the parametric estimation of the Hawkes model with offspring density function $h(\cdot; \theta_h)$ parametrized through a finite-dimensional parameter θ_h, with the aim to estimate the parameter vector $\theta = (\mu, \eta, \theta_h^\top)$. Following the discussion on estimation, this section addresses model assessment and simulation of the process, which are vital aspects to understanding the fitted model and its appropriateness.

1.4.1 Estimation

1.4.1.1 Maximum likelihood estimation

The likelihood function of the Hawkes process is straightforward to calculate, and this attractive feature implies that estimation using the method of maximum likelihood is simple to implement. For a point process realization $\tau_{1:n} = (\tau_1, \ldots, \tau_n)$ on the interval $[0, T]$ with $N(T) = n$, the likelihood of the point process with a parameter vector θ takes the form,

$$L(\theta|\tau_{1:n}) = p_\theta(\tau_{1:n})\mathbb{P}_\theta\left(N(T) - N(\tau_n) = 0|\tau_{1:n}\right). \tag{1.4.1}$$

The likelihood is simply the product of the joint density of all the event times and the probability that no events occur between the last event time τ_n and the censoring time T. The joint density is then further factored into a product of conditional marginal densities of which take the form,

$$p(t|\tau_{1:N(t-)}) = \lambda(t)\exp\left(-\int_{\tau_{N(t-)}}^{t}\lambda(s)\mathrm{d}s\right), \tag{1.4.2}$$

where $\lambda(t)$ takes on the form specified in (1.3.1) for the Hawkes process. The form of (1.4.2) is simply the product of the probability of observing an event at time t, and the probability that no events occur between the previous event time $\tau_{N(t-)}$ and time t, where both probabilities are conditional on the history at time t. Then substituting (1.4.2) into (1.4.1) leads to the classical expression for the likelihood of a point processes and takes the form (Daley and Vere-Jones, 2003, Proposition 7.3.III),

$$L(\theta|\tau_{1:n}) = \prod_{i=1}^{n}\lambda(\tau_i)\exp\left(-\int_0^T\lambda(s)\mathrm{d}s\right). \tag{1.4.3}$$

The maximum likelihood estimator (MLE) is computed by maximizing the likelihood in (1.4.3) with respect to the parameter vector θ. This optimization problem naturally attempts to maximize the value of the intensity at the observed event times while concurrently minimizing the intensity over the intervals where no events are observed. It is standard practice, due to its computational appeal, to use the log-likelihood when performing the optimization routine to compute the MLE, and by taking the logarithm of the likelihood in (1.4.3), the log-likelihood takes the following form,

$$\ell(\theta|\tau_{1:n}) = \sum_{i=1}^{n}\log\lambda(\tau_i) - \int_0^T\lambda(s)\mathrm{d}s. \tag{1.4.4}$$

Many optimization techniques are readily available, such as the simplex downhill method Nelder and Mead (1965). The log-likelihood in (1.4.4) above is provided in

sufficient generality to apply to a wide variety of point process models, such as the time-varying baseline intensity model (non-stationary Hawkes process) discussed in the next section. The evaluation, or more importantly, optimization of the Hawkes likelihood is rather simple to develop, but remains computationally demanding as it requires evaluation of the conditional intensities in (1.4.2) at all observed time points. Hence, the complexity of the likelihood calculation is of order n^2. However, this complexity can be reduced to order n (linear time complexity) when the offspring density function is exponential, and a recursion can be employed to compute the excitation function at each observed time point.

The preceding discussions rely on the intensity process with respect to its natural filtration having a readily computable expression. For the Hawkes process, this is readily available, but as will be shown later, this is not the case for the renewal Hawkes process. When direct evaluation of the intensity process with respect to the natural filtration is not feasible, the log-likelihood optimization cannot be performed using (1.4.4). This primarily occurs in circumstances in which the immigrant events are no longer Poisson, and the branching structure is not observable. The Expectation-Maximization (E-M) algorithm is a natural approach to overcome this hindrance since the unobserved branching structure can be treated probabilistically as a missing data problem.

1.4.1.2 Expectation-Maximization algorithm

The E-M algorithm of Dempster et al. (1977) is an iterative procedure that estimates the MLE in circumstances in which the observed data X is known, but the model depends on some latent or missing data Z. In these instances, the likelihood based only on the observed data $L(\theta|X)$ (herein termed the incomplete-data likelihood) may be challenging to compute to perform direct MLE calculations. Alternatively, it may be simpler to work with the likelihood based on the observed, and missing data $L(\theta|X, Z)$ (herein termed the complete-data likelihood) and account for the unobserved, or missing data Z probabilistically.

Suppose the E-M algorithm has a starting parameter estimate $\hat{\theta}^{[0]}$. The E-M algorithm proceeds to iterate between the following two steps at the $(m+1)$-th iteration:

1. The *Expectation step* computes the expected complete-data log-likelihood with respect to the conditional distribution of the missing data Z, conditional on the observed data X, and current parameter estimate $\hat{\theta}^{[m]}$ as follows,

$$Q(\theta|X, \hat{\theta}^{[m]}) = \mathbb{E}_{Z|X,\hat{\theta}^{[m]}}\Big[\log L(\theta|X, Z)\Big]. \tag{1.4.5}$$

2. The *Maximization step* maximizes the expected complete-data log-likelihood in (1.4.5) to compute the next iterations parameter estimate $\hat{\theta}^{[m+1]}$ by solving the following optimization problem,

$$\hat{\theta}^{[m+1]} = \underset{\theta}{\operatorname{argmax}} \; Q(\theta|X, \hat{\theta}^{[m]}). \tag{1.4.6}$$

The E-M algorithm continues to iterate between the Expectation and Maximization steps until the specified convergence criterion is satisfied. The parameter estimates are assured not to reduce the observed-data likelihood $L(\theta|X)$ at each consecutive iteration.

The branching process interpretation of immigrant and offspring events discussed previously facilitates the application of the E-M algorithm to the Hawkes process, in which the missing data is the branching structure such as used in Veen and Schoenberg (2008) and Lewis and Mohler (2011). By conditioning on the branching structure, the process is decoupled into identical (up to a shift in time), and independent homogenous Poisson processes, and this facilitates straightforward calculation of the complete-data log-likelihood. A more in-depth discussion of the E-M algorithm applied to the renewal Hawkes process is discussed in Chapter 2, and to avoid repetitive discussions of this concept, no further discussion of the algorithm is mentioned here.

1.4.2 Model assessment

After identifying the fitted Hawkes process, the adequacy of the fitted model to data should be assessed. For this purpose, a residual point process based on Papangelou's random time change theorem Daley and Vere-Jones (2003) can assess the appropriateness of the fitted Hawkes process. The point process model is correctly specified when the sequence of event times $\{\tau_i\}_{i=1,\dots,n}$ on $[0, T]$ follows a point process with the specified conditional intensity $\lambda(t)$. This assessment can be conducted by computing the integral transformed point pattern $\{\Lambda(\tau_i)\}_{i=1,\dots,n}$ and evaluating whether this transformed sequence follows a unit rate Poisson process on $[0, \Lambda(T)]$ where $\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s$ denotes the cumulative intensity process, also known as the compensator of the point process.

However, rather than assessing whether the transformed sequence conforms to a unit rate Poisson process, a test based on uniformity can be justified. A Poisson process is observed from time 0 until the censoring time T, and since the joint distribution of the ordered event times of the process is equal to that of the order statistics of an equal number of uniformly distributed event times on the interval $[0, T]$, the transformed sequence $\{\Lambda(\tau_i)\}_{i=1,\dots,n}$ on the interval $[0, \Lambda(T)]$, should be uniformly

distributed. By replacing $\Lambda(t)$ with an estimate of the cumulative intensity process by replacing the unknown parameters by their estimates $\hat{\Lambda}(t)$, the transformed sequence can be assessed using formal statistical tests such as the Kolmogorov-Smirnov (K-S) test or the Anderson-Darling (A-D) test, or it can be graphically assessed using a quantile-quantile plot (QQ plot). A large p-value emerging from these formal tests indicate that the transformed sequence conforms to a uniform sequence and suggests that the fitted model is adequate for the data.

1.4.3 Simulation algorithms

This section illustrates algorithms to simulate realizations of the Hawkes self-exciting point process, in which some slight difficulties occur due to the dependence on the history of events. To begin any discussion on simulation algorithms for self-exciting processes, a method to simulate a homogenous Poisson process demands attention. For a homogenous Poisson process with a constant arrival rate λ, the inter-event waiting times are independent and exponentially distributed and can be generated using an inversion sampling method. First, simulate u, a uniform random variable on the unit interval and then generate an inter-event waiting time by substituting u into the inverse cumulative distribution function $F^{-1}(u) = -\log(1-u)/\lambda$. Then proceed to simulate exponential inter-event waiting times until the cumulative sum of these times is greater than the censoring time T.

Simulating an inhomogeneous Poisson process is now possible, albeit slightly more challenging and depends on the form of the specified intensity function. One simplified approach is the thinning algorithm proposed by Lewis and Shedler (1979). It is an iterative procedure in which the points are simulated sequentially using the intensity process, which is updated at each consecutively simulated point. The thinning algorithm utilizes the thinning property of Poisson processes that states that the contribution to the intensity process from all independent sub-processes equals the total intensity process at any time t, implying that a homogeneous Poisson process with rate $\lambda_{\max}$ such that $\lambda_{\max} > \lambda(t), t \geqslant 0$ can be simulated and then thinned appropriately to obtain an inhomogeneous Poisson process with rate $\lambda(t)$. This is an efficient algorithm to simulate an inhomogeneous Poisson process with a specified intensity function over the interval $[0, T]$, and following the `simPois` function in the `IHSEP R` package works as follows:

1. Compute the maximum intensity of $\lambda(t)$ over the interval $[0, T]$ and denote it by $\lambda_{\max}$.

2. Simulate $\lambda_{\max}T + 1.96\lambda_{\max}T$ exponential random variables with rate $\lambda_{\max}$.

3. Sum up all the simulated exponential random variables, and if the cumulative sum is less than T, iteratively simulate additional exponential random variables in batches (of size approx. $\lambda_{\max}T$) with rate $\lambda_{\max}$.

4. Retain the cumulative sum of exponentials that are less than T as the event times of the homogeneous Poisson process.

5. Perform thinning based on the retention probability $\lambda(t)/\lambda_{\max}$ for each simulated event times to obtain the event times for the inhomogeneous Poisson process.

Strategies that exploit the branching process interpretation of Hawkes and Oakes (1974) usually provides greater efficiencies for simulating self-exciting processes. The simulation comprises of two components relating to the independent immigrant arrivals and the offspring generations. The algorithm works as follows. First, all of the immigrant events are simulated as a homogeneous (or inhomogeneous in the case of time-varying baseline self-exciting processes) Poisson process. Then, from each simulated point τ_i, simulate an inhomogeneous Poisson process with rate $\eta h(t)$ over the interval $[0, T - \tau_i]$, this includes both immigrants and any future offspring events that have been generated. Such an exploit of the branching structure will be utilized when simulating the renewal Hawkes process later on in Chapter 3.

1.5 Non-stationary self-exciting point processes

The classical Hawkes self-exciting point process of Hawkes assumes that the baseline intensity is constant. However, this unnecessarily restricts the application domain of self-exciting processes, as for many applications a constant baseline intensity would not be adequate. For instance, in seismological applications, the sequence of aftershocks following a mainshock display the self-exciting phenomenon, but the arrival rate of mainshocks typically decays over time (Utsu, 1961) and is not constant. In financial applications, the modeling of intra-day stock trading may also be modeled using a Hawkes process, but the baseline trading intensity may be inappropriate as the trading intensity during the open and close of the market exhibit drastically different features to the rest of the trading day (Engle and Russell, 1998).

In the many contexts in which a constant baseline intensity is unrealistic or inadequate for the data, a non-stationary self-exciting point process with time-varying baseline intensity might be an appropriate alternative (Chen and Hall, 2013). The time-varying baseline intensity self-exciting point process has a time dependent function $\mu(t)$ that replaces the constant μ in (1.3.1). More specifically, the intensity

process $\lambda(t)$ takes the form,

$$\lambda(t) = \mu(t) + \sum_{j=1}^{N(t-)} \eta h(t - \tau_j), \tag{1.5.1}$$

which incorporates the same history-dependent self-excitation mechanism that was introduced for the Hawkes process. The statistical inferences and model assessment for the non-stationary Hawkes process are very similar to the Hawkes process and are discussed in the work of Chen and Hall (2013).

In the non-stationary Hawkes process, the baseline rate is allowed to vary but only in a deterministic way. The renewal Hawkes process also allows the baseline rate to vary but does so stochastically and still maintains stationarity, at least in an asymptotic sense. The remainder of this thesis will be devoted to this class of self-exciting processes, and in the next chapter, a background on renewal Hawkes process and the recent statistical inferences that have proposed so far will be discussed.

Chapter 2

Background on renewal Hawkes processes

The classical process of Hawkes (1971) models the immigrant arrival times as a homogenous Poisson process. The estimation of model parameters for this class of model or its inhomogeneous Poisson generalizations can be performed using MLE. Recently Wheatley et al. (2016) introduced a nascent extension to the Hawkes process by allowing the immigrant arrival process to be a general renewal process, rather than a homogeneous Poisson process. Wheatley et al. (2016) termed the extension a renewal Hawkes process or RHawkes process for an abbreviation. They demonstrated that RHawkes processes are more versatile than Hawkes processes as such processes can feature dependence between clusters, where each cluster consists of an immigrant and its direct offspring of all generations. Wheatley et al. (2016) also claimed that the likelihood of the RHawkes processes is practically impossible to compute because the required computational time is an exponential function of the observed number of events of the process up to the censoring time. This claim will be proved to be incorrect in this thesis. To calculate the MLE of the RHawkes process, Wheatley et al. (2016) proposed two E-M type algorithms based on two different alternatives to the set of missing variables. They proposed to estimate the variance-covariance matrix of the computed MLE using bootstrap. They also claimed that the goodness-of-fit test statistic for the RHawkes process requires exponential computational time, and hence proposed a Monte Carlo approach for goodness-of-fit assessment.

2.1 Model and notation

Before introducing the RHawkes process model, several notations are introduced. Let M_i, $i = 1, 2, \ldots$ denote the unobservable event indicator, where $M_i = 0$ indicates that the i-th event is an immigrant and $M_i = 1$ indicates that the i-th event is an

offspring. Furthermore, let $I(t) = \max\{i|\, \tau_i < t, M_i = 0\}$ denote the (unobservable) index of the most recent immigration event before time t, with the convention that $I(t) := 0$ when $t < \tau_1$ and $\tau_0 := 0$. The point process $N(t)$ is called a *renewal Hawkes* process, or *RHawkes* process for short, if the intensity process $\lambda(t)$, $t \geqslant 0$ relative to the enlarged filtration $\tilde{\mathcal{F}}_t = \sigma\{N(s), I(s); s \leqslant t\}$, $t \geqslant 0$ takes the form,

$$\begin{aligned}\lambda(t) &= \frac{\mathbb{E}\left[\mathrm{d}N(t)|\tilde{\mathcal{F}}_{t-}\right]}{\mathrm{d}t} = \frac{\mathbb{E}\left[\mathrm{d}N(t)|\mathcal{F}_{t-}, I(t)\right]}{\mathrm{d}t} \\ &= \mu(t - \tau_{I(t)}) + \sum_{j:\tau_j<t} \eta h(t - \tau_j) \\ &= \mu(t - \tau_{I(t)}) + \phi(t), \end{aligned} \tag{2.1.1}$$

for a positive function $\mu(\cdot)$ on the positive half-line $\mathbb{R}^+$. The function $\mu(\cdot)$ is interpreted as the hazard function of the *i.i.d.* waiting times between the immigration events, which forms a renewal process. For the stability of the process, it is required that $\int_0^\infty e^{-\int_0^t \mu(s)\mathrm{d}s}\mathrm{d}t < \infty$, which ensures the expected waiting time between successive immigrants is finite. Interestingly, when the function $\mu(\cdot)$ is simply a constant, the RHawkes process reduces to a Hawkes process. The self-excitation mechanism is identical to that of the Hawkes process and its non-stationary extension where η and $h(\cdot)$ have the same interpretations as before. Furthermore, note the introduction of the notation $\phi(t)$, which denotes the cumulative contribution of offspring effects at time t to the total intensity process. The remainder of this chapter will be devoted to a review of the inferential methodologies proposed in Wheatley et al. (2016).

2.2 Expectation-Maximization algorithms

The intensity process for the Hawkes process with respect to its natural filtration defined in (1.3.1) can be computed at any time t, and consequently, MLE can be performed using the log-likelihood function in (1.4.4). However, the RHawkes process model requires knowledge of which event is the most recent immigrant event, and consequently, the intensity process defined in (2.1.1) cannot be directly computed and used to evaluate the log-likelihood function in (1.4.4). This led Wheatley et al. (2016) to claim that evaluation of the exact likelihood and hence, direct MLE is not feasible for the RHawkes process.

The tremendous success and well-established application of the E-M algorithm to point process models led Wheatley et al. (2016) to circumvent the direct calculation of the likelihood by formulating this problem using the E-M framework of Dempster et al. (1977). The approaches of Veen and Schoenberg (2008) and Lewis and Mohler (2011) were extended by Wheatley et al. (2016) to incorporate renewal immigration, which they termed the EM1 algorithm. In addition to the EM1 al-

gorithm, they introduced a further E-M algorithm with a reduced set of missing data (EM2). The EM2 algorithm is additionally applicable to the Hawkes process with inhomogeneous Poisson process immigration. The EM1 and EM2 algorithms allow for straightforward estimation of the RHawkes process when the functions $\mu(\cdot)$ and $\phi(\cdot)$ have separate parameters. Both the EM1 and EM2 algorithms can be tailored to multivariate, spatio-temporal, and marked point process data, but no such attempt has been made.

First, a discussion on why Wheatley et al. (2016) asserted that likelihood evaluation demands exponential time. The observed data consists of the observed event times $\tau_1, \ldots, \tau_n$ over the interval $[0, T]$ and $N(T) = n$, and the missing data consists of the immigrant or offspring indicators $M_{1:n} = (M_1, \ldots, M_n)$. The complete-data log-likelihood conditioned on $(\tau_{1:n}, M_{1:n})$ of the RHawkes process takes the form,

$$\log L(\theta|\tau_{1:n}, M_{1:n}) = \sum_{i=1}^{n}(1 - M_i)\log\mu(\tau_i - \tau_{I(\tau_i)}) - \int_0^T \mu(s - \tau_{I(s)})\mathrm{d}s + \sum_{i=1}^{n} M_i\phi(\tau_i) - \int_0^T \phi(s)\mathrm{d}s, \quad (2.2.1)$$

which has two separate components for the immigrant and offspring processes. However, the immigrant vector $M_{1:n}$ is not observable and must be treated as random. Consequently, the observed data log-likelihood is defined in terms of an expectation of the conditional likelihoods in (2.2.1) as follows,

$$\log L(\theta|\tau_{1:n}) = \log\left(\sum_{j\in\{0,1\}^n} L(\theta|\tau_{1:n}, M_{1:n} = j_{1:n})\mathbb{P}\left(M_{1:n} = j_{1:n}|\theta\right)\right), \quad (2.2.2)$$

where $\mathbb{P}\left(M_{1:n} = j_{1:n}|\theta\right)$ denotes the probability that the immigrant vector equals one of the possible immigrant/offspring combinations $j_{1:n} = (j_1, \ldots, j_n) \in \{0, 1\}^n$. It is clear that these probabilities depend on the model parameters and hence makes direct evaluation infeasible, let alone optimization of the log-likelihood in (2.2.2) over the parameter space. This is what directed Wheatley et al. (2016) to consider the use of the E-M algorithm. However, it should be remarked that the log-likelihood defined in (2.2.1) depends on the random number of events $N(T) = n$ but the log-likelihood in (2.2.2) treats n as a known quantity and this is not actually correct.

The remainder of this section outlines the E-M algorithms of Wheatley et al. (2016). Similar to Veen and Schoenberg (2008) and Lewis and Mohler (2011), the missing data for the first of the two E-M algorithms is the full branching structure, which consists of immigrants and offspring labels, denoted as $Z_{n\times n}$. The branching structure is a lower-triangular matrix $Z_{n\times n}$ whose diagonal elements $Z_{i,i}$ indicate

whether the i-th event is an immigrant with $Z_{i,i} = 1$ or an offspring with $Z_{i,i} = 0$. The off-diagonal elements indicate if the j-th event is the parent of the i-th event, in which case $Z_{i,j} = 1$ for $j < i$, otherwise it takes the value zero. It should be mentioned that each event can only be an immigrant or a direct offspring from a previous event. Consequently, each row of the matrix $Z_{n\times n}$ has exactly one element taking the value one and the remaining elements taking the value zero.

Rather than optimizing the log-likelihood directly in (2.2.2), the EM1 algorithm utilizes the complete-data log-likelihood, which is conditioned on the unobserved branching structure $Z_{n\times n}$ and the observed event times $\tau_{1:n}$. The complete-data log-likelihood can be computed by computing the product of the joint probability density function of the observed events and the branching structure $p(\tau_{1:n}, Z_{n\times n}) = p(\tau_{1:n}|Z_{n\times n})p(Z_{n\times n})$ and the probability that no events occur in the interval $(\tau_n, T]$, that is, $\mathbb{P}\left(N(T) - N(\tau_n) = 0|\tau_{1:n}, Z_{n\times n}\right)$. The joint density of the observed event times $\tau_{1:n}$ conditional on the branching structure $Z_{n\times n}$ takes the form,

$$p(\tau_{1:n}|Z_{n\times n}) = \prod_{i=1}^{n}\prod_{j=1}^{i-1}\left(\mu(\tau_i - \tau_j)e^{-\int_{\tau_j}^{\tau_i}\mu(s-\tau_j)\mathrm{d}s}\right)^{Z_{i,i}1\{I(\tau_i)=j\}} \times \prod_{i=1}^{n}\prod_{j=1}^{i-1}\left(\eta h(\tau_i - \tau_j)e^{-\int_{\tau_j}^{\tau_i}\eta h(s-\tau_j)\mathrm{d}s}\right)^{Z_{i,j}}, \quad (2.2.3)$$

where $Z_{i,i}1\{I(\tau_i) = j)\}$ indicates that the i-th event is an immigrant and the most recent immigrant prior to the i-th event was the j-th event and $Z_{i,j}$ indicates that the j-th event is a parent of the i-th event. Note that by convention when the product is indexed over the empty set, the product is defined to be one. For the time lag $\tau_i - \tau_j$, the immigrant waiting time density is evaluated at this lag when $Z_{i,i}1\{I(\tau_i) = j)\} = 1$, and the offspring inter-event waiting time density is evaluated at this lag when $Z_{i,j} = 1$.

2.2.1 Expectation step

By combining the conditional density function of the observed event times conditional on the branching structure in (2.2.3), and the distribution of the branching structure, the complete-data log-likelihood takes the form,

$$\log L(\theta|\tau_{1:n}, Z_{n\times n}) = \log p(Z_{n\times n}) + \left(\sum_{i=1}^{n}\sum_{j=1}^{i-1} Z_{i,j}\log\eta h(\tau_i - \tau_j) - \int_0^T \phi(s)\mathrm{d}s\right) + \left(\sum_{i=1}^{n}\sum_{j=1}^{i-1} Z_{i,i}1\{I(\tau_i) = j\}\log\mu(\tau_i - \tau_j) - \sum_{i=1}^{n+1}\sum_{j=1}^{i-1} 1\{I(\tau_i) = j\}\int_{\tau_j}^{\tau_i}\mu(s - \tau_j)\mathrm{d}s\right), \quad (2.2.4)$$

where for notational convenience $\tau_0 := 0$ and $\tau_{n+1} := T$, and they are not to be included as observed event times. The Expectation step computes the conditional expectation of the complete-data log-likelihood with respect to the branching structure $Z_{n\times n}$ conditional on the observed events $\tau_{1:n}$ and the current parameter estimate $\hat{\theta}^{[m]}$. Hence, using (1.4.5) and (2.2.4), the expression for the Expectation step takes the form,

$$\begin{aligned} Q_1(\theta|\tau_{1:n}, \hat{\theta}^{[m]}) &= \mathbb{E}_{Z_{n\times n}|\tau_{1:n},\hat{\theta}^{[m]}}\left[\log L(\theta|\tau_{1:n}, Z_{n\times n})\right] \\ &\propto \sum_{i=1}^{n}\sum_{j=1}^{i-1} \mathbb{P}\left(Z_{i,j} = 1|\tau_{1:i}, \hat{\theta}^{[m]}\right) \log \eta h(\tau_i - \tau_j) - \int_0^T \phi(s)\mathrm{d}s \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{i-1} \mathbb{P}\left(Z_{i,i}1\{I(\tau_i) = j\} = 1|\tau_{1:i}, \hat{\theta}^{[m]}\right) \log \mu(\tau_i - \tau_j) \\ &\quad - \sum_{i=1}^{n+1}\sum_{j=1}^{i-1} \mathbb{P}\left(I(\tau_i) = j|\tau_{1:i}, \hat{\theta}^{[m]}\right) \int_{\tau_j}^{\tau_i} \mu(s - \tau_j)\mathrm{d}s. \end{aligned} \tag{2.2.5}$$

The expression Q_1 in (2.2.5) depends on the distribution of the branching structure. Hence, before optimization can be implemented, a procedure to compute the branching structure probabilities requires attention. Let $\pi_{i,j}^{[m]}$ denote the conditional probability that the j-th event is the parent to the i-th event, that is,

$$\pi_{i,j}^{[m]} = \mathbb{P}\left(Z_{i,j} = 1|\tau_{1:i}, \hat{\theta}^{[m]}\right). \tag{2.2.6}$$

Next, define $\pi_i^{[m]} := \pi_{i,i}^{[m]}$ to be the conditional probability that the i-th event is an immigrant. It should be observed that for fixed i, these probabilities must sum to one because each event can only be an immigrant or an offspring event of only a single parent i.e., $\sum_{j=1}^{i} \pi_{i,j}^{[m]} = 1$, for all $i = 1, \ldots, n$. Furthermore, let $\pi_{i,j|k}^{[m]}$ denote the probability that the j-th event is the parent event of the i-th event conditional on the most recent immigrant being the k-th event, that is,

$$\pi_{i,j|k}^{[m]} = \mathbb{P}\left(Z_{i,j} = 1|\tau_{1:i}, I(\tau_i) = k, \hat{\theta}^{[m]}\right), \quad k, j < i.$$

Again, a very similar abbreviation for the immigrant events will be used as before with $\pi_{i|k}^{[m]} = \pi_{i,i|k}^{[m]}$.

In the next chapter, an important discovery facilitates the construction of an efficient recursive algorithm to compute the most recent immigrant probabilities, which is essential in developing an algorithm to compute the likelihood in quadratic time. For now, consider the procedure implemented in Wheatley et al. (2016) and let $w_{i,j}^{[m]}$ denote the conditional most recent immigrant probability. In their approach,

the probability that at time τ_i the j-th event is the most recent immigrant event is computed as the product of the probability that the j-th event is an immigrant and the probability that all subsequent events are not immigrant events (i.e., they are all offspring events), with all these probabilities conditioned on the j-th event being the most recent immigrant, that is,

$$\mathbb{P}\left(I(\tau_i) = j | \tau_{1:i}, \hat{\theta}^{[m]}\right) := \omega_{i,j}^{[m]} = \pi_j^{[m]} \bar{\pi}_{j+1|j}^{[m]} \dots \bar{\pi}_{i-1|j}^{[m]}, \tag{2.2.7}$$

where $\bar{\pi} = 1 - \pi$ denotes the complementary probability. Hence, the final probability needed to compute Q_1 in (2.2.5) is given by,

$$\mathbb{P}\left(Z_{i,i} 1\{I(\tau_i) = j\} = 1 | \tau_{1:i}, \hat{\theta}^{[m]}\right) = \omega_{i,j}^{[m]} \pi_{i|j}^{[m]}, \tag{2.2.8}$$

which indicates the probability that the i-th event is an immigrant and the j-th event is the most recent immigrant event before the i-th event.

The conditional probabilities $\pi_{i|k}^{[m]}$ and $\pi_{i,j|k}^{[m]}$ are immediately computable from the intensity process by employing the thinning property mentioned already, and leads to the following,

$$\begin{aligned}
\pi_{i|k}^{[m]} &= \frac{\mu(\tau_i - \tau_k; \hat{\theta}^{[m]})}{\mu(\tau_i - \tau_k; \hat{\theta}^{[m]}) + \phi(\tau_i; \hat{\theta}^{[m]})}, \qquad k < i = 2, \dots, n, \\
\pi_{i,j|k}^{[m]} &= \frac{\hat{\eta}^{[m]} \hat{h}^{[m]}(\tau_i - \tau_j)}{\mu(\tau_i - \tau_k; \hat{\theta}^{[m]}) + \phi(\tau_i; \hat{\theta}^{[m]})}, \qquad j, k < i = 2, \dots, n.
\end{aligned} \tag{2.2.9}$$

However, the conditional probabilities $\pi_i^{[m]}$ and $\pi_{i,j}^{[m]}$ are slightly more challenging to calculate since they depend on the distribution of the most recent immigrant. To this end, Wheatley et al. (2016) define the incomplete-data intensity process $\lambda_*(t), t \geqslant 0$ as follows,

$$\lambda_*(t) = \mu_*(t) + \phi(t), \tag{2.2.10}$$

where the incomplete-data intensity process for immigrants $\mu_*(t), t \geqslant 0$ is a weighted average of immigrant intensities with weights $\omega_{N(t),j}$ given by,

$$\mu_*(t) = \sum_{j=1}^{N(t)} \omega_{N(t),j} \ \mu(t - \tau_j). \tag{2.2.11}$$

The weights $\omega_{N(t),j}$ in the incomplete-data intensity process for immigrants in (2.2.11) depends only on $N(t)$ and the index of the most recent immigrant j. However, these weights are incorrect since they also need to depend on the current time t. For the correct expression for the weights to be used in the incomplete-data intensity process see (3.3.15).

The probabilities and the incomplete-data intensity process defined above enables the development of a recursive algorithm to compute the remaining probability weights. Let $\boldsymbol{\pi}_i = (\pi_{i,1}, \ldots, \pi_{i,i})$ and $\boldsymbol{\omega}_i = (\omega_{i,1}, \ldots, \omega_{i,i-1})$ denote a vector of probabilities for the i-th event. The probabilities $\pi_{i,j}^{[m]}$ and $\omega_{i,j}^{[m]}$ are then jointly computed using the subsequent recursion:

1. Let $i = 1$, $\pi_{1,1} = 1$ and $\omega_{2,1} = 1$ since the first event is an immigrant event.

2. Let $i = i + 1$ and then compute the following probabilities for $j = 1, \ldots, i-1$:

$$\pi_i^{[m]} = \frac{\mu_*(\tau_i; \hat{\theta}^{[m]})}{\mu_*(\tau_i; \hat{\theta}^{[m]}) + \phi(\tau_i; \hat{\theta}^{[m]})} \qquad \text{and} \qquad \pi_{i,j}^{[m]} = \frac{\hat{\eta}^{[m]} \hat{h}^{[m]}(\tau_i - \tau_j)}{\mu_*(\tau_i; \hat{\theta}^{[m]}) + \phi(\tau_i; \hat{\theta}^{[m]})}. \tag{2.2.12}$$

3. Then compute the most recent immigrant probabilities for the i-th event as follows,

$$\begin{aligned} \boldsymbol{\omega}_i &= \left(\pi_1 \bar{\pi}_{2|1} \ldots \bar{\pi}_{i-1|1}, \ldots, \pi_j \bar{\pi}_{j+1|j} \ldots \bar{\pi}_{i-1|j}, \ldots \pi_{i-1}\right), \\ &= \left((\boldsymbol{\omega}_{i-1}) \circ (\bar{\pi}_{i-1|1}, \ldots, \bar{\pi}_{i-1|i-2}), \pi_{i-1}\right), \end{aligned} \tag{2.2.13}$$

   where $\circ$ is the Hadamard product i.e., $(a, b) \circ (c, d) = (ac, bd)$.

4. Repeat steps 2 and 3 until $i = n$ and then stop.

When the recursion terminates, the probabilities $\pi_{i|k}^{[m]}$, $\pi_{i,j|k}^{[m]}$, $\pi_i^{[m]}$, $\pi_{i,j}^{[m]}$, and $\omega_{i,j}^{[m]}$ are substituted into Q_1 given in (2.2.5), which is then maximized with respect to θ to obtain the parameter vector $\hat{\theta}^{[m+1]}$ for the next iteration.

A closer inspection of the Expectation step of the E-M algorithms of Wheatley et al. (2016) reveals two pertinent issues. First, the conditional distribution of the missing data given the observed data $\{\tau_{1:n}, \tau_{n+1} > T\}$ is required. However, in calculating these distributions, Wheatley et al. implicitly assumed conditional independence of $\{M_{1:i}\}$ and $\tau_{i+1:n+1}$ given $\tau_{1:i}$, which is incorrect for general RHawkes processes, although it is true in the classical Hawkes process models. The second problem is that their calculation of the conditional distributions of M_i given $\tau_{1:i}$ is incorrect. For example, in the numerical experiments in the next chapter, it was observed that the $\pi_{i,j}^{[m]}$'s in (2.2.12) (eq. (16) of Wheatley et al. (2016)) do not sum to one as they should for fixed i. This mistake was apparently due to their treatment of the RHawkes process as if it were a non-stationary Hawkes process with a time-varying background event rate $\mu_*(t)$ given in (2.2.11) (eq. (15) of Wheatley et al. (2016)). This mistake can be corrected by using eq. (3.3.12), (3.3.13) and (3.3.14). However, with only this second issue fixed, their E-M algorithms still do not work as expected, due to the first issue mentioned above.

2.2.2 Maximization step

Now that a scheme has been developed to compute the expectation of the complete-data log-likelihood, the updated parameters $\hat{\theta}^{[m+1]}$ are calculated by maximizing Q_1 in (2.2.5) with respect to $\theta = (\theta_\mu^\top, \eta, \theta_h^\top)$. By conditioning on the branching structure $Z_{n\times n}$, the expression Q_1 can be divided into two distinct components that are estimated independently. These components consists of immigration process $\mu(\cdot)$ and self-exciting effects $\eta h(\cdot)$. The optimization of Q_1 with respect to θ_μ reduces to solving the problem,

$$\hat{\theta}_\mu^{[m+1]} = \underset{\theta_\mu}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{j=1}^{i-1} \omega_{i,j}^{[m]} \pi_{i|j}^{[m]} \left(\log \mu(\tau_i - \tau_j) - \int_0^{\tau_i - \tau_j} \mu(s)\mathrm{d}s \right), \tag{2.2.14}$$

where $\mu(\cdot)$ depends on θ_μ and the solution provides the next iterations parameter estimate for the immigration process.

The maximization of Q_1 with respect to η leads to the following analytical expression for the updated branching ratio parameter,

$$\hat{\eta}^{[m+1]} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{i-1} \pi_{i,j}^{[m]}}{\sum_{i=1}^{n} \hat{H}^{[m]}(T-\tau_i)} = \frac{n - \sum_{i=1}^{n} \pi_i^{[m]}}{\sum_{i=1}^{n} \hat{H}^{[m]}(T-\tau_i)}, \tag{2.2.15}$$

where $\hat{H}^{[m]}(s) := \int_0^t \hat{h}^{[m]}(s)\mathrm{d}s$ denotes the offspring distribution function estimated at the parameter $\hat{\theta}_h^{[m]}$. Finally, to estimate the offspring parameter θ_h, the optimization problem reduces to the following,

$$\hat{\theta}_h^{[m+1]} = \underset{\theta_h}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{j=1}^{i-1} \pi_{i,j}^{[m]} \log h(\tau_i - \tau_j), \tag{2.2.16}$$

where $h(\cdot)$ depends on θ_h and the solution gives the updated parameter for the offspring density. Note that throughout these optimization problems, the probabilities computed in the Expectation step all depend on the current iterations parameter estimate $\hat{\theta}^{[m]}$. The parameters θ_μ and θ_h which specify the functions $\mu(\cdot)$ and $h(\cdot)$ and the constant η are the variables which are estimated when maximizing the expression in (2.2.5).

2.2.3 E-M algorithm with reduced set of missing data

Further to the E-M algorithm discussed in the previous section, Wheatley et al. (2016) proposed an alternative algorithm by employing a reduced set of missing data. This modified algorithm reduces the memory requirements and applies to much larger datasets due to its enhanced computational efficiencies. In the so-

called, EM2 algorithm, the branching structure reduces to the diagonal elements of the full branching structure $Z_{1:n} := \{Z_{i,i}\}_{i=1,\dots,n}$, and simply indicates whether it is an immigrant or an offspring. Hence the memory requirement for storing the missing data in the EM2 algorithm is reduced from $O(n^2)$ to $O(n)$.

Conditional on the reduced branching structure, the semi-complete data log-likelihood takes on the form,

$$\begin{aligned}\log L(\theta|\tau_{1:n}, Z_{1:n}) \propto \sum_{i=1}^{n}(1-Z_i)\log\phi(\tau_i) - \int_0^T \phi(s)\mathrm{d}s \\ + \sum_{i=1}^{n}\sum_{j=1}^{i-1} Z_i 1\{I(\tau_i)=j\}\log\mu(\tau_i-\tau_j) \\ - \sum_{i=1}^{n+1}\sum_{j=1}^{i-1} 1\{I(\tau_i)=j\}\int_{\tau_j}^{\tau_i}\mu(s-\tau_j)\mathrm{d}s. \quad (2.2.17)\end{aligned}$$

Adopting a similar derivation to that employed in the EM1 algorithm to compute Q_1, the Expectation step for the EM2 algorithm evaluates the following,

$$\begin{aligned}Q_2(\theta|\tau_{1:n}, \hat{\theta}^{[m]}) = \mathbb{E}_{Z_{1:n}|\tau_{1:n},\hat{\theta}^{[m]}}\big[\log L(\theta|\tau_{1:n}, Z_{1:n})\big] \\ \propto \sum_{i=1}^{n}(1-\pi_i^{[m]})\log\phi(\tau_i) - \int_0^T\phi(s)\mathrm{d}s \\ + \sum_{i=1}^{n}\sum_{j=1}^{i-1}\pi_i^{[m]}\omega_{i,j}^{[m]}\log\mu(\tau_i-\tau_j) - \sum_{i=1}^{n+1}\sum_{j=1}^{i-1}\omega_{i,j}^{[m]}\int_{\tau_j}^{\tau_i}\mu(s-\tau_j)\mathrm{d}s, \quad (2.2.18)\end{aligned}$$

where the probabilities $\pi_i^{[m]}$ and $\omega_{i,j}^{[m]}$ are calculated using (2.2.12) and (2.2.13), and depend on the parameter estimate $\hat{\theta}^{[m]}$. Furthermore, observe that the reduced missing data implies that the calculation of the probabilities $\pi_{i,j}^{[m]}$ and $\pi_{i,j|k}^{[m]}$ can be avoided.

In the Maximization step, the immigrant and offspring components are again separable, and the optimization problem to determine the updated parameters for θ_μ is identical to (2.2.14) in the EM1 algorithm. Furthermore, the updated parameter for η is also identical to the EM1 algorithm given by (2.2.15). However, by conditioning on the reduced branching structure, the individual offspring processes are incapable of being decoupled. Hence, to compute the updated parameters of θ_h, the equation in (2.2.18) requires numerical optimization techniques.

2.3 Statistical inferences

This thesis is concerned with the statistical estimation of model parameters for the RHawkes process model. Following the estimation of the model parameters,

methods and procedures to perform statistical inferences such as model assessment, predictions, and simulations are a natural next objective. To perform statistical inferences, a method to calculate the likelihood value, a procedure to compute the variance-covariance matrix, and methods to assess the goodness-of-fit of the model to data are a necessity. In this section, methods to perform these inferences are discussed, with the objective to discuss superior inferential methods in the next chapter.

2.3.1 Likelihood and variance-covariance matrix evaluation

The E-M algorithms discussed so far should converge to a parameter estimate that optimizes the likelihood function. The direct calculation of the likelihood procedure suggested by Wheatley et al. (2016) demands evaluation of the intensity process in (2.1.1), and this is not possible without conditioning on the immigrant vector $Z_{1:n} \in \{0,1\}^n$. Since there is a total of 2^{n-1} possible immigrant combinations (as the first event is an immigrant event), direct calculation of the likelihood using this approach is computationally infeasible. However, simulation of all the possible immigrant vectors, which are denoted by $z_{1:n}^{(j)}$, $j = 1, \ldots, 2^{n-1}$ is possible for rather small samples sizes.

From each possible realization of the immigrant vector, the likelihood of the RHawkes process can be computed using the appropriate immigrant intensity process as follows,

$$\mu^{(j)}(t) = \mu(t - \tau_{I(\tau_{N(t)})}|z_{1:n}^{(j)}), \tag{2.3.1}$$

which is a deterministic function. Hence the classical likelihood formula for point processes can be applied in this instance, similar to that of time-varying self-exciting point processes. Using the likelihood function in (1.4.1) and the immigrant intensity process in (2.3.1), the complete-data likelihood conditioned on a particular simulated immigrant vector $z_{1:n}^{(j)}$, takes the form,

$$L(\theta|\tau_{1:n}, z_{1:n}^{(j)}) = \prod_{i=1}^{n} \left(\mu^{(j)}(\tau_i) + \phi(\tau_i)\right) \exp\left(-\int_0^T \mu^{(j)}(s) + \phi(s)\mathrm{d}s\right). \tag{2.3.2}$$

Hence, the incomplete-data likelihood is computed as a weighted average of the conditional incomplete-data likelihood in (2.3.2) as follows,

$$L(\theta|\tau_{1:n}) = \sum_{j=1}^{2^{n-1}} L(\theta|\tau_{1:n}, z_{1:n}^{(j)})\, \mathbb{P}\left(Z_{1:n} = z_{1:n}^{(j)}|\theta\right). \tag{2.3.3}$$

The probabilities $\mathbb{P}(Z_{1:n} = z_{1:n}^{(j)}|\theta)$ in (2.3.3) are computed during the Expectation step of the E-M algorithms but as pointed out above are not necessarily correct.

However, this procedure to compute the likelihood is computationally demanding even on moderate sample sizes due to the exponential computational time required for evaluation. To offset this demanding likelihood computation, a Monte Carlo approximation to the likelihood in (2.3.3) can be used. A Monte Carlo approximation computes the sample average of the likelihood and significantly reduces the computational burden. The branching structure probabilities from the last iteration of the E-M algorithm are stored and used to simulate the immigrant vector. Then the likelihood for each realization is computed. The computed average of all the likelihoods from the simulated realizations provides a very close approximation to the actual value of the likelihood.

First, to simulate the immigrant vector, Wheatley et al. (2016) used the acceptance-rejection thinning type algorithm of Lewis and Fieller (1979). Their algorithm to simulate a realization of the immigrant vector $z_{1:n} = (z_1, \ldots, z_n)$, works as follows;

1. Let $z_1 = 1$ and $I(\tau_2) = 1$ as the first event is an immigrant event. Then let $i = 2$.

2. Compute (or retrieve) the probability $\pi_{i|I(\tau_i)}$ from the E-M algorithm and then generate a uniform random variable u on $[0, 1]$.

3. If $u < \pi_{i|I(\tau_i)}$ set $I(\tau_{i+1}) = i$ and $z_i = 1$, otherwise let $z_i = 0$ and do not change the index of the most recent immigrant. Then let $i = i + 1$.

4. If $i < n$ return to step 2. Otherwise return the vector $z_{1:n}$ as the realization of the simulated immigrant vector.

This simple simulation procedure generates one particular realization of the immigrant vector $z_{1:n}$. By replicating the above procedure N times, a sample set of immigrant vectors $\left\{z_{1:n}^{(1)}, \ldots, z_{1:n}^{(N)}\right\}$ is obtained. Hence, to approximate the likelihood, the Monte Carlo approximation takes the form,

$$L(\theta|\tau_{1:n}) \approx \frac{1}{N} \sum_{i=1}^{N} L(\theta|\tau_{1:n}, z_{1:n}^{(i)}), \tag{2.3.4}$$

The approximation to the log-likelihood computes the logarithm of the computed average in (2.3.4). Since the exponential function might produce computational complications, it is suggested to calculate the log of the conditional incomplete-data likelihood in (2.3.2) and then compute the exponential when computing the weighted average in (2.3.3). Another strategy to overcome the numerical instability concerns is to compute the average of the log of the conditional likelihoods in (2.3.2), but then this may produce an underestimation of the Monte Carlo log-likelihood.

For variance-covariance estimation for the RHawkes process, a bootstrap procedure was the recommended strategy of Wheatley et al. (2016). This is because

there is no explicit closed-form solution. However, these methods perform poorly for small sample sizes and thus Monte Carlo methods are generally recommended to improve upon the approximation.

2.3.2 Model assessment

To address the problem of assessing the fitted model's adequacy to data, a procedure based on the residuals using the time-change property can be implemented (Ogata, 1988; Papangelou, 1972). The transformed event times according to $\tilde{\tau}_i = \int_0^{\tau_i} \lambda(s)\mathrm{d}s$ generate a sequence $\{\tilde{\tau}_i\}_{i\in\mathbb{N}}$ that forms a unit rate Poisson process when the model specification is correct. The model fit is then assessed by testing whether the transformed times conform to a unit rate Poisson process. The K-S distance is one such metric to assess this conformance, but it depends on the particular immigrant vector $z_{1:n}$ to allow the intensity process, or more specifically, the cumulative intensity process, to be a deterministic function over the interval $[0, T]$.

The K-S distance test statistic is a random variable that is defined in terms of the event times and immigrant vector $S(\tau_{1:n}, Z_{1:n})$. For the semi-complete data $\left\{\tau_{1:n}, z_{1:n}^{(j)}\right\}$ the null hypothesis $H_0^{(j)}$ is that the RHawkes model is adequate for the particular immigrant vector $Z_{1:n} = z_{1:n}^{(j)}$. Hence, the semi-complete-data p-values are given by,

$$p^{(j)} = \mathbb{P}\left(S > S(\tau_{1:n}, Z_{1:n})|H_0^{(j)}\right), \quad j = 1, \ldots, 2^{n-1}. \tag{2.3.5}$$

Furthermore, for the incomplete-data $\{\tau_{1:n}\}$ the null hypothesis H_0 is that the RHawkes model is adequate for the point process realization. The test statistic in this case is unobserved since the immigrant vector is unobservable. To overcome this, the incomplete-data p-value is computed by conditioning on the immigrant vector as follows,

$$p = \mathbb{P}\left(S > S(\tau_{1:n}, Z_{1:n})|H_0\right) = \sum_{j=1}^{2^{n-1}} p^{(j)} \mathbb{P}\left(Z_{1:n} = z_{1:n}^{(j)}|\theta\right), \tag{2.3.6}$$

where j denotes the index of all possible immigrant vectors from the set $\{0, 1\}^n$. Hence, by applying a similar Monte Carlo approximation that was used to calculate the likelihood, the Monte Carlo approximation to the p-value in (2.3.6) is computed as the average of the semi-complete data p-values as follows,

$$p \approx \frac{1}{N} \sum_{i=1}^{N} p^{(i)}. \tag{2.3.7}$$

where $\left\{z_{1:n}^{(1)}, \ldots, z_{1:n}^{(N)}\right\}$ is again a sampled set of immigrant vectors.

Chapter 3

Direct likelihood evaluation of the renewal Hawkes process[1]

3.1 Introduction

The fundamental role played by the likelihood in statistical inferences, such as estimation, hypothesis testing, and model selection implies that a lack of a feasible approach to evaluate the likelihood is a significant hurdle that hinders the application of the RHawkes process model. Therefore, this chapter undertakes this challenge by providing an efficient algorithm that computes the likelihood of the RHawkes process in quadratic time, a drastic improvement from the exponential time claimed by Wheatley et al. (2016). By overcoming the challenge of likelihood evaluation, this thesis enables maximum likelihood estimation and other likelihood-based inferences, such as goodness-of-fit testing and prediction for the RHawkes model computationally feasible.

The superior performance of the MLE computed by directly maximizing the log-likelihood relative to the estimators of Wheatley et al. based on E-M algorithms discussed in Chapter 2, both in computational and in statistical terms, is illustrated using a simulation study. Numerical evidence shows that the inverted observed information matrix provides satisfactory estimates of the variance of the MLE, and therefore, computationally expensive bootstrap procedures for variance estimation are avoided. As a by-product of the likelihood evaluation algorithm, the Rosenblatt transformation of the observed event times are efficiently computable in quadratic time and serve as the foundation for assessing the adequacy of the fitted model. A simulation-based procedure for future event prediction is also presented and was found to be able to predict the number of earthquakes in a region near Japan reasonably well.

[1]Most of the content shown in this chapter has been published in the *Journal of Computational and Graphical Statistics*; see Chen and Stindl (2018).

The remainder of this chapter is organized as follows. Section 3.2 investigates several properties of the RHawkes process model. Section 3.3 presents the procedure to compute the likelihood of the model efficiently in quadratic time. Section 3.4 discusses the model assessment. Section 3.5 briefly outlines procedures to make predictions using the fitted RHawkes model based on the observations up until the censoring time. Section 3.6 reports the results of two simulation studies to evaluate the numerical performance of the direct MLE estimator and compare it to the E-M algorithms based estimators of Wheatley et al. (2016). In Section 3.7, the proposed methodologies are applied to analyze real data arising in seismology and finance. An R package called `RHawkes` implementing the proposed methodologies and the R scripts used to perform the reported data analyses can be found in the online supplementary materials to the article Chen and Hall (2016), also this can be downloaded from the CRAN (https://CRAN.R-project.org/package=RHawkes).

3.2 Properties of the renewal Hawkes process model

This section investigates some of the useful properties of the RHawkes process. Recall the Poisson cluster or branching process interpretation of the Hawkes process (Hawkes and Oakes, 1974), which also applies to the RHawkes process, that allows the points of the point process to be interpreted as the arrival times of immigrants and birth times of offspring of all generations in a dynamic population pooled together. For the RHawkes process, the immigrants arrive according to a general renewal process and once these individuals enter the population, whether by immigration or by birth, they start to give birth to children of their own according to independent Poisson processes with a common intensity function.

By the linear form of the intensity of the RHawkes process, at any given time t, the instantaneous event rate is the immigrant arrival rate plus the sum of the birthing rates of existing members of the population. Statistically, the RHawkes process is equivalent to the independent sum of a renewal process and *i.i.d.* non-stationary self-exciting processes (Chen and Hall, 2013, 2016) starting at the event times of the renewal process, where the inter-arrival times of the renewal process have a common hazard function $\mu(\cdot)$, and the self-exciting processes have a common background event intensity function and excitation function both equal to $\eta h(\cdot)$. From this observation, useful properties of the RHawkes process can be derived. For example, by conditioning on the first event time, it can be observed that the mean of the process, $M(t) = \mathbb{E}[N(t)]$, uniquely satisfies the integral equation,

$$M(t) = \int_0^t \{1 + K(t-s) + M(t-s)\}\, \mu(s) \exp\left(-\int_0^s \mu(x)\mathrm{d}x\right) \mathrm{d}s, \qquad (3.2.1)$$

where $K(t)$ is the expected number of events of the non-stationary self-exciting process up to time t, and itself uniquely satisfies the integral equation,

$$K(t) = \int_0^t (1 + K(t-s))\, \eta h(s) \mathrm{d}s. \tag{3.2.2}$$

A derivation of (3.2.1) and (3.2.2) is detailed below.

Proof. The derivation is based upon conditioning on the first event time τ_1. If $\tau_1 > t$, then $N(t) = 0$, and if $\tau_1 \leqslant t$, $N(t) = 1 + N_0(t - \tau_1) + N_1(t - \tau_1)$, where $N_0(\cdot)$ denotes the point process formed by the offspring of the immigrant at time τ_1; and $N_1(\cdot)$ denotes the point process formed by the immigrants after time τ_1 and their offspring. Note that $N_0(\cdot)$ is a non-stationary self-exciting point process with background event intensity and excitation function both equal to $\eta h(\cdot)$, $N_1(\cdot)$ is an RHawkes point process equal to $N(\cdot)$ in distribution, and $N(\cdot)$, $N_0(\cdot)$, and $N_1(\cdot)$ are independent. Therefore, with $f_{\tau_1}(s) = \mu(s) \exp(-\int_0^s \mu(x)\mathrm{d}x)$ denoting the density of τ_1, and $K(\cdot) = \mathbb{E}\left[N_0(\cdot)\right]$ denoting the mean process of $N_0(\cdot)$,

$$\begin{aligned} M(t) = \mathbb{E}\left[N(t)\right] &= \int_0^\infty \mathbb{E}\left[N(t) | \tau_1 = s\right] f_{\tau_1}(s)\mathrm{d}s \\ &= \int_0^t \mathbb{E}\left[N(t) | \tau_1 = s\right] f_{\tau_1}(s)\mathrm{d}s \\ &= \int_0^t \left\{1 + K(t-s) + M(t-s)\right\} f_{\tau_1}(s)\mathrm{d}s. \end{aligned}$$

This proves (3.2.1), and it remains to show (3.2.2). Note the intensity process of $N_0(t)$ is $\lambda_0(t) = \eta h(t) + \int_0^t \eta h(t-u)\mathrm{d}N_0(u)$. Now, let $\Lambda_0(t) = \int_0^t \lambda_0(s)\mathrm{d}s$, and $H(t) = \int_0^t h(s)\mathrm{d}s$. By the definition of the intensity process, $N_0(t) - \Lambda_0(t)$ is a zero mean martingale, and therefore $\mathbb{E}\left[N_0(t)\right] = \mathbb{E}\left[\Lambda_0(t)\right]$. By Fubini's theorem and integration by parts,

$$\begin{aligned} \Lambda_0(t) &= \int_0^t \eta h(s)\mathrm{d}s + \int_0^t \int_0^s \eta h(s-u)\mathrm{d}N_0(u)\mathrm{d}s \\ &= \int_0^t \eta h(s)\mathrm{d}s + \int_0^t \int_u^t \eta h(s-u)\mathrm{d}s\mathrm{d}N_0(u) \\ &= \int_0^t \eta h(s)\mathrm{d}s + \int_0^t \eta H(t-u)\mathrm{d}N_0(u) \\ &= \int_0^t \eta h(s)\mathrm{d}s + \int_0^t N_0(u)\eta h(t-u)\mathrm{d}u \\ &= \int_0^t \eta h(s)\mathrm{d}s + \int_0^t N_0(t-u)\eta h(u)\mathrm{d}u. \end{aligned}$$

Taking expectations on both sides, and by Fubini's theorem, yields the following,

$$\begin{aligned} K(t) = \mathbb{E}\left[\Lambda_0(t)\right] &= \int_0^t \eta h(s)\mathrm{d}s + \int_0^t \mathbb{E}\left[N_0(t-u)\right]\eta h(u)\mathrm{d}u \\ &= \int_0^t \eta h(s)\mathrm{d}s + \int_0^t K(t-u)\eta h(u)\mathrm{d}u. \end{aligned}$$

This concludes the proof of (3.2.2). □

Before this chapter addresses how to calculate the likelihood efficiently in the next section, it is worth remarking that the RHawkes process is substantially more versatile than the original Hawkes process. For instance, depending on the choice of model parameters, the RHawkes process can demonstrate drastically different features than the Hawkes process. For example, Figure 3.2.1 shows five random realizations of the event times up to the censoring time 100, of each of 6 RHawkes processes labeled as (a)-(f) in the figure. The 6 RHawkes processes have Weibull distributed inter-immigration times with the same mean 1 and varying shape parameter $\kappa \in \{1/3, 1, 3\}$, and have the same branching ratio $\eta = 0.5$, and different offspring density functions h in the set $\{h_1(t) = (1/2)(1 + t/2)^{-2},\ h_2(t) = \exp(-t)\}$. Note that when $\kappa = 1/3$, the events of the RHawkes processes tend to occur in bursts compared to those of the Hawkes processes ($\kappa = 1$), while when $\kappa = 3$, the events of the RHawkes process seem to be more evenly distributed in comparison. Meanwhile, the RHawkes processes with the exponential offspring density tend to show stronger event clustering than the RHawkes with the same immigrant processes, but the polynomial offspring density function. Using the ratio of the average of the 25% longest inter-event waiting times of a realization of a point process to the average of the shortest 25%, called the B-index, as a measure of the burstiness/clustering, so that a larger B-index indicates stronger burstiness/clustering, then the average of the B-indexes of the 5 realizations of the RHawkes process in the 6 cases are given respectively by: (a) $B = 208.1$, (b) $B = 19.7$, (c) $B = 11.0$, (d) $B = 448.6$, (e) $B = 23.6$, and (f) $B = 13.2$, which confirms the visual impression. It deserves mention that, with $\Gamma(s, x) = \int_x^\infty t^{s-1}e^{-t}\mathrm{d}t$ denoting the incomplete gamma function, the B-index of homogeneous Poisson processes is $\Gamma(2, \log 4)/\{1 - \Gamma(2, \log(4/3))\} = 17.424$, which can serve as a benchmark when the B-index is used to assess burstiness.

3.3 Maximum likelihood estimation

This section considers the estimation of the RHawkes model based on observations over the interval $[0, T]$ using the maximum likelihood method. The likelihood is defined as the Radon-Nikodym density of the distribution of the RHawkes process relative to the distribution of the unit rate Poisson process on $[0, T]$, interpreted as a

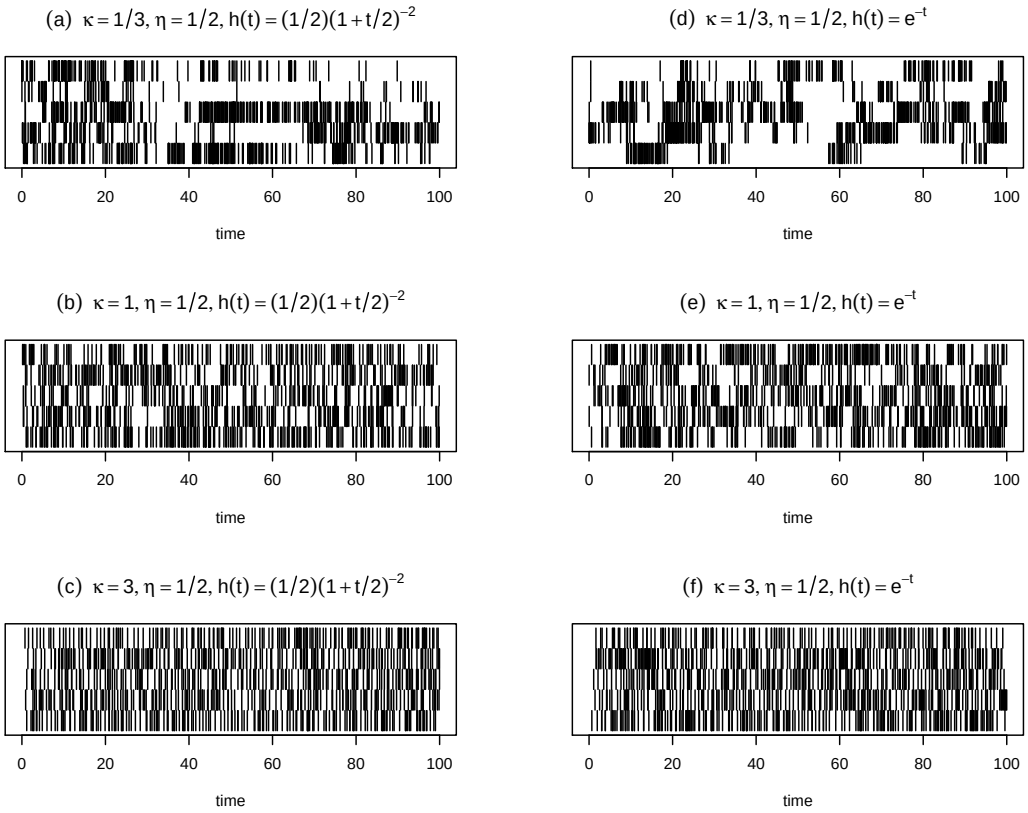

Figure 3.2.1: Barcode plot of simulated event times of 6 RHawkes processes, with five random realizations of each process. Each vertical bar indicates an event time, with bars in the same row in each graph indicating events belonging to the same realization of the process. The censoring time is 100 in all simulations. The branching ratio $\eta = 0.5$ in all 6 processes. The waiting times between immigrations in all 6 cases are Weibull distributed with the same mean 1 but different shape parameters κ shown in the title of each figure. The offspring density h in each case is either polynomial or exponential, as shown in the figure titles.

function of the model parameters (Andersen et al., 1993, Section II.7). The sample path of the RHawkes process on the interval $[0, T]$ is entirely specified by the number n of its jump discontinuities and the positions of these jumps $\tau_1 < \ldots < \tau_n$. Let $\mathbb{P}_\theta$ denote the distribution of the RHawkes process with parameters $\theta = (\mu(\cdot), \eta, h(\cdot))$, and $\mathbb{P}_{\text{Poi}}$ denote the distribution of the unit rate Poisson process, then the density of the RHawkes process is given by,

$$\begin{aligned}
&\frac{\mathbb{P}_\theta(\tau_1 \in \mathrm{d}\tau_1, \ldots, \tau_n \in \mathrm{d}\tau_n, \tau_{n+1} > T)}{\mathbb{P}_{\text{Poi}}(\tau_1 \in \mathrm{d}\tau_1, \ldots, \tau_n \in \mathrm{d}\tau_n, \tau_{n+1} > T)} \\
&= \frac{p_\theta(\tau_{1:n})\mathrm{d}\tau_1 \cdots \mathrm{d}\tau_n \mathbb{P}_\theta(\tau_{n+1} > T | \tau_{1:n})}{\exp(-\tau_1)\,\mathrm{d}\tau_1 \left\{\prod_{i=2}^n \exp(-(\tau_i - \tau_{i-1}))\,\mathrm{d}\tau_i\right\} \exp(-(T - \tau_n))} \\
&= p_\theta(\tau_{1:n}) \mathbb{P}_\theta(\tau_{n+1} > T | \tau_{1:n}) \exp(T),
\end{aligned}$$

where $p_\theta(\tau_{1:n})$ is the $\mathbb{P}_\theta$-density of $\tau_{1:n}$. Up to a constant free of θ, the likelihood is given by,

$$L(\theta) = p_\theta(\tau_{1:n})\mathbb{P}_\theta(\tau_{n+1} > T|\tau_{1:n}); \tag{3.3.1}$$

see also Wheatley et al. (2016, eq. 34). The subsequent theorem provides a computable expression of the likelihood, where for notational convenience, the subscript θ in p_θ and $\mathbb{P}_\theta$ is dropped, while the dependence of the relevant densities and probabilities on the parameter θ is silently understood. A description of the algorithm to compute the likelihood in pseudocode is displayed in Algorithm 1.

The theorem below utilizes the following notation for convenience; $U(t) = \int_0^t \mu(s)\mathrm{d}s$, $H(t) = \int_0^t h(s)\mathrm{d}s$, and $\Phi(t) = \int_0^t \phi(s)\mathrm{d}s = \eta \sum_{j:\tau_j<t} H(t-\tau_j)$. Also recall that $\phi(t) = \sum_{j:\tau_j<t} \eta h(t-\tau_j)$. A proof of the theorem is provided after the statement of the theorem.

Theorem 3.3.1. *The likelihood for the renewal Hawkes (RHawkes) process* (3.3.1) *can be written as,*

$$L(\theta) = \begin{cases} e^{-U(T)}, & n = 0, \\ \mu(\tau_1)e^{-U(\tau_1)-U(T-\tau_1)-\eta H(T-\tau_1)} & n = 1, \\ \mu(\tau_1)e^{-U(\tau_1)} \left\{\prod_{i=2}^n \sum_{j=1}^{i-1} p_{ij}d_{ij}\right\} \sum_{j=1}^n S_{n+1,j}p_{n+1,j}, & n \geqslant 2, \end{cases} \tag{3.3.2}$$

where

$$d_{ij} = \left(\mu(\tau_i - \tau_j) + \phi(\tau_i)\right) e^{-\{U(\tau_i-\tau_j)-U(\tau_{i-1}-\tau_j)\}-\{\Phi(\tau_i)-\Phi(\tau_{i-1})\}}, \tag{3.3.3}$$

$$S_{n+1,j} = e^{-\{U(T-\tau_j)-U(\tau_n-\tau_j)\}-\{\Phi(T)-\Phi(\tau_n)\}}, \tag{3.3.4}$$

and the p_{ij}, $i = 2 \ldots n+1$, $j = 1, \ldots i-1$, are given by $p_{21} = 1$ and the following recursion,

$$p_{ij} = \begin{cases} \frac{\phi(\tau_{i-1})}{\mu(\tau_{i-1}-\tau_j)+\phi(\tau_{i-1})} \frac{d_{i-1,j}p_{i-1,j}}{\sum_{j=1}^{i-2} p_{i-1,j}d_{i-1,j}}, & j = 1, \ldots, i-2 \\ 1 - \sum_{k=1}^{i-2} p_{ik}, & j = i-1, \end{cases} \tag{3.3.5}$$

for $i = 3, \ldots, n+1$.

Proof. When the total number of events $n \leqslant 1$, the theorem is trivially true. Now assume $n \geqslant 2$. To calculate the likelihood, first note that,

$$L(\theta) = p(\tau_1) \left\{\prod_{i=2}^n p(\tau_i|\tau_{1:i-1})\right\} \mathbb{P}(\tau_{n+1} > T|\tau_{1:n}). \tag{3.3.6}$$

By conditioning on the event index of the most recent immigrant, the following must

hold,

$$p(\tau_i|\tau_{1:i-1}) = \sum_{j=1}^{i-1} p(\tau_i|\tau_{1:i-1}, I(\tau_i) = j)\mathbb{P}(I(\tau_i) = j|\tau_{1:i-1}), \tag{3.3.7}$$

$$\mathbb{P}(\tau_{n+1} > T|\tau_{1:n}) = \sum_{j=1}^{n} \mathbb{P}(\tau_{n+1} > T|\tau_{1:n}, I(\tau_{n+1}) = j)\mathbb{P}(I(\tau_{n+1}) = j|\tau_{1:n}). \tag{3.3.8}$$

Note that given $\tau_{1:i-1}$ and $I(\tau_i) = j$, the conditional hazard function of the inter-event waiting time $\tau_i - \tau_{i-1}$ is $\mu(\cdot + \tau_{i-1} - \tau_j) + \phi(\cdot + \tau_{i-1})$. By the relations between the hazard, the density, and the survival functions (see e.g. Daley and Vere-Jones, 2003, Eqn. (1.1.1)-(1.1.3)), the conditional densities and survival probabilities in (3.3.7) and (3.3.8) are given by,

$$p(\tau_i|\tau_{1:i-1}, I(\tau_i) = j) = d_{ij}, \tag{3.3.9}$$

$$\mathbb{P}(\tau_{n+1} > T|\tau_{1:n}, I(\tau_{n+1}) = j) = S_{n+1,j}, \tag{3.3.10}$$

where d_{ij} and $S_{n+1,j}$ are as in (3.3.3) and (3.3.4).

Now it remains to show that $\mathbb{P}(I(\tau_i) = j|\tau_{1:i-1}) = p_{ij}$. When $i = 2$, it is clear that $\mathbb{P}(I(\tau_2) = 1|\tau_1) = 1 = p_{21}$ since the first event has to be an immigrant. When $i = 3, \ldots, n+1$, by conditioning on $I(\tau_{i-1})$ and the Bayes rule, the following recursion holds,

$$\begin{aligned}
&\mathbb{P}(I(\tau_i) = j|\tau_{1:i-1}) \\
&= \sum_{k=1}^{i-2} \mathbb{P}(I(\tau_i) = j|\tau_{1:i-1}, I(\tau_{i-1}) = k)\mathbb{P}(I(\tau_{i-1}) = k|\tau_{1:i-1}) \\
&= \sum_{k=1}^{i-2} \mathbb{P}(I(\tau_i) = j|\tau_{1:i-1}, I(\tau_{i-1}) = k)\frac{p(\tau_{i-1}|I(\tau_{i-1}) = k, \tau_{1:i-2})\mathbb{P}(I(\tau_{i-1}) = k|\tau_{1:i-2})}{p(\tau_{i-1}|\tau_{1:i-2})} \\
&= \sum_{k=1}^{i-2} \mathbb{P}(I(\tau_i) = j|\tau_{1:i-1}, I(\tau_{i-1}) = k)\frac{d_{i-1,k}\mathbb{P}(I(\tau_{i-1}) = k|\tau_{1:i-2})}{p(\tau_{i-1}|\tau_{1:i-2})}.
\end{aligned} \tag{3.3.11}$$

Now make the important observation that $I(\tau_i)$ can only be $I(\tau_{i-1})$ or $i-1$, according to whether $M_{i-1} = 1$ or $M_{i-1} = 0$. Therefore, when $j \leqslant i - 2$,

$$\begin{aligned}
&\mathbb{P}(I(\tau_i) = j|\tau_{1:i-1}, I(\tau_{i-1}) = k) \\
&= \begin{cases} 0, & k \in \{1, \ldots, i-2\}, k \neq j, \\ \mathbb{P}(M_{i-1} = 1|\tau_{1:i-1}, I(\tau_{i-1}) = j) = \frac{\phi(\tau_{i-1})}{\mu(\tau_{i-1} - \tau_j) + \phi(\tau_{i-1})}, & k = j. \end{cases}
\end{aligned} \tag{3.3.12}$$

On the other hand, when $j = i - 1$,

$$\begin{aligned}
&\mathbb{P}(I(\tau_i) = j | \tau_{1:i-1}, I(\tau_{i-1}) = k) \\
&= \mathbb{P}(M_{i-1} = 0 | \tau_{1:i-1}, I(\tau_{i-1}) = k) = \frac{\mu(\tau_{i-1} - \tau_k)}{\mu(\tau_{i-1} - \tau_k) + \phi(\tau_{i-1})}, \quad k \in \{1, \ldots, i-2\}.
\end{aligned} \tag{3.3.13}$$

By (3.3.12) and (3.3.13), the recursion (3.3.11) simplifies to the the following,

$$\mathbb{P}(I(\tau_i) = j | \tau_{1:i-1}) = \begin{cases} \frac{\phi(\tau_{i-1})}{\mu(\tau_{i-1} - \tau_j) + \phi(\tau_{i-1})} \frac{d_{i-1,j} \mathbb{P}(I(\tau_{i-1}) = j | \tau_{1:i-2})}{p(\tau_{i-1} | \tau_{1:i-2})}, & j = 1, \ldots, i-2 \\ 1 - \sum_{k=1}^{i-2} \mathbb{P}(I(\tau_i) = k | \tau_{1:i-1}), & j = i-1. \end{cases} \tag{3.3.14}$$

Since $p(\tau_{i-1} | \tau_{1:i-2}) = \sum_{j=1}^{i-2} \mathbb{P}(I(\tau_{i-1}) = j | \tau_{1:i-2}) d_{i-1,j}$, it can be observed by comparing (3.3.14) and (3.3.5) that, p_{ij} and $\mathbb{P}(I(\tau_i) = j | \tau_{1:i-1})$ satisfy exactly the same recursion, and therefore must be equal for all $i \geqslant 2$. This concludes the proof. □

Remark 3.3.1. *By the general formula for point process likelihood (e.g. Daley and Vere-Jones, 2003, Proposition 7.2.III), the likelihood is given by*

$$\prod_{i=1}^{n} \lambda_*(\tau_i) \exp\{-\int_0^T \lambda_*(s) \mathrm{d}s\},$$

where $\lambda_*(t),\ t \geqslant 0$ is the conditional intensity process of the RHawkes process relative to its natural filtration $\mathcal{F}_t,\ t \geqslant 0$, which is referred to as the incomplete-data conditional intensity function *by Wheatley et al. (2016). The expression for $\lambda_*(t)$ given in eq. (14)-(15) of Wheatley et al. (2016) is incorrect and if plugged into the general formula would lead to a wrong likelihood. The correct expression of $\lambda_*(t)$ can be computed by noticing $\lambda_*(t) = r(t - \tau_{N(t-)})$, where $r(\cdot)$ denotes the conditional hazard function of the random variable $\tau_{N(t-)+1} - \tau_{N(t-)}$ given $N(t-)$ and $\tau_1, \ldots, \tau_{N(t-)}$, and by calculating $r(\cdot)$ from the conditional density function of $\tau_{N(t-)+1} - \tau_{N(t-)}$ obtained in the proof of Theorem 3.3.1. With the correct expression of $\lambda_*(t)$, the general formula gives the same likelihood as Theorem 3.3.1.*

Remark 3.3.1 facilitates the application of the general formula for point process likelihood evaluation for the RHawkes process as the conditional intensity process with respect to its natural filtration is now readily available. Furthermore, Therorem 3.3.2 provides an explicit form for the intensity process to be employed when applying the log-likelihood formula in (1.4.4) with a derivation provided after the statement of the theorem.

Function likRHawkes $(n, \tau_{1:n}, T, \mu(\cdot), \eta, h(\cdot), U(\cdot), H(\cdot))$;
if $n = 0$ **then**
  `return` $e^{-U(T)}$;
end
if $n = 1$ **then**
  `return` $\mu(\tau_1)e^{-U(\tau_1)-U(T-\tau_1)-\eta H(T-\tau_1)}$;
end
`lik` $\leftarrow \mu(\tau_1)e^{-U(\tau_1)}$; `// likelihood initialized to` $p(\tau_1)$
vector `p1`$[1:n]$; `// to store the` p_{ij}`'s`
vector `p0`$[1:(n-1)]$; `// to store the` $p_{i-1,j}$`'s`
vector `d1`$[1:n]$; `// to store the` d_{ij}`'s/`$S_{n+1,j}$`'s`
vector `d0`$[1:(n-1)]$; `// to store the` $d_{i-1,j}$`'s`
scalar `ptau, ph0, ph1, Ph0, Ph1`; `// to store` $p(\tau_i|\tau_{1:i-1})$`,` $\phi(\tau_{i-1})$`,` $\phi(\tau_i)$`,` $\Phi(\tau_{i-1})$`,` $\Phi(\tau_i)$
vector `mu1`$[1:(n-1)]$; `// to store` $\mu(\tau_i - \tau_{1:i-1})$
vector `Mu1`$[1:(n-1)]$`// to store` $U(\tau_i - \tau_{1:i-1})$
vector `mu0`$[1:(n-2)]$; `// to store` $\mu(\tau_{i-1} - \tau_{1:i-2})$
vector `Mu0`$[1:(n-2)]$ `// to store` $U(\tau_{i-1} - \tau_{1:i-2})$
`mu1`$[1] \leftarrow \mu(\tau_2 - \tau_1)$; `Mu1`$[1] \leftarrow U(\tau_2 - \tau_1)$; `ph1` $\leftarrow \eta h(\tau_2 - \tau_1)$; `Ph1` $\leftarrow \eta H(\tau_2 - \tau_1)$;
`ptau` $\leftarrow (\texttt{mu1}[1] + \texttt{ph1}[1])\, e^{-\texttt{Mu1}[1]-\texttt{Ph1}[1]}$;
`lik` $\leftarrow$ `lik` $\times$ `ptau`;
`d0`$[1] \leftarrow$ `ptau`; `p0`$[1] \leftarrow 1$; `mu0`$[1] \leftarrow$ `mu1`$[1]$;
`Mu0`$[1] \leftarrow$ `Mu1`$[1]$; `ph0` $\leftarrow$ `ph1`; `Ph0` $\leftarrow$ `Ph1`;
$i \leftarrow 3$;
while $i \leqslant n$ **do**
  `mu1`$[1:(i-1)] \leftarrow \mu(\tau_i - \tau_{1:i-1})$; `Mu1`$[1:(i-1)] \leftarrow U(\tau_i - \tau_{1:i-1})$;
  `ph1` $\leftarrow$ `sum`$(\eta h(\tau_i - \tau_{1:i-1}))$; `Ph1` $\leftarrow$ `sum`$(\eta H(\tau_i - \tau_{1:i-1}))$;
  `d1`$[1:(i-1)] \leftarrow e^{-(\texttt{Mu1}[1:(i-1)]-(\texttt{Mu0}[1:(i-2)],0))-(\texttt{Ph1}-\texttt{Ph0})}\,(\texttt{mu1}[1:(i-1)] + \texttt{ph1})$;
  `p1`$[1:(i-2)] \leftarrow \frac{\texttt{ph0}}{\texttt{mu0}[1:(i-2)]+\texttt{ph0}} \frac{d0[1:(i-2)] \times p0[1:(i-2)]}{\texttt{ptau}}$;
  `p1`$[i-1] \leftarrow 1 -$ `sum` $(\texttt{p1}[1:(i-2)])$;
  `ptau` $\leftarrow$ `sum` $(\texttt{d1}[1:(i-1)] \times \texttt{p1}[1:(i-1)])$;
  lik $\leftarrow$ lik $\times$ `ptau`;
  `d0`$[1:(i-1)] \leftarrow$ `d1`$[1:(i-1)]$; `p0`$[1:(i-1)] \leftarrow$ `p1`$[1:(i-1)]$;
  `mu0`$[1:(i-1)] \leftarrow$ `mu1`$[1:(i-1)]$; `Mu0`$[1:(i-1)] \leftarrow$ `Mu1`$[1:(i-1)]$;
  `ph0` $\leftarrow$ `ph1`; `Ph0` $\leftarrow$ `Ph1`;
  $i \leftarrow i + 1$;
end
`Mu1`$[1:n] \leftarrow U(T - \tau_{1:n})$; `Ph1` $\leftarrow$ `sum`$(\eta H(T - \tau_{1:n}))$;
`d1`$[1:n] \leftarrow e^{-(\texttt{Mu1}[1:n]-(\texttt{Mu0}[1:(n-1)],0))-(\texttt{Ph1}-\texttt{Ph0})}$;
`p1`$[1:(n-1)] \leftarrow \frac{\texttt{ph0}}{\texttt{mu0}[1:(n-1)]+\texttt{ph0}} \frac{\texttt{d0}[1:(n-1)] \times \texttt{p0}[1:(n-1)]}{\texttt{ptau}}$;
`p1`$[n] \leftarrow 1 -$ `sum` $(\texttt{p1}[1:(n-1)])$;
`ptau` $\leftarrow$ `sum` $(\texttt{d1}[1:n] \times \texttt{p1}[1:n])$;
lik $\leftarrow$ lik $\times$ `ptau`;
`return lik`;

Algorithm 1: Algorithm to compute the likelihood of the RHawkes process.

Theorem 3.3.2. *Let $r(\cdot)$ denote the conditional hazard function of the random variable $\tau_{N(t-)+1} - \tau_{N(t-)}$ conditional on $N(t-)$ and $\tau_1, \ldots, \tau_{N(t-)}$. The intensity of the RHawkes process relative to the natural filtration $\mathcal{F} = \{\mathcal{F}_t, t \geqslant 0\}$ with $\mathcal{F}_t = \sigma\{N(s);\, s \leqslant t\}$ is then given by $\lambda^*(t) = r(t - \tau_{N(t)})$ where,*

$$r(t) = \sum_{j=1}^{N(t-)} w_j(t)\mu(t - \tau_j) + \phi(t), \tag{3.3.15}$$

and

$$w_j(t) = \frac{e^{-\{U(t-\tau_j) - U(\tau_{N(t-)} - \tau_j)\} - \{\Phi(t) - \Phi(\tau_{N(t-)})\}} p_{N(t-)+1,j}}{\sum_{k=1}^{N(t-)} e^{-\{U(t-\tau_k) - U(\tau_{N(t-)} - \tau_k)\} - \{\Phi(t) - \Phi(\tau_{N(t-)})\}} p_{N(t-)+1,k}}.$$

Proof. The hazard function is computed as the ratio of the density function and the survival function as follows,

$$r(t) = \frac{p(t|\tau_{1:N(t-)})}{S(t|\tau_{1:N(t-)})}. \tag{3.3.16}$$

Then by conditioning on the index of the most recent immigrant, the density takes the form,

$$\begin{aligned}
&p(t|\tau_{1:N(t-)}) \\
&= \sum_{j=1}^{N(t-)} p(t|\tau_{1:N(t-)}, I(t) = j)\mathbb{P}\left(I(t) = j|\tau_{1:N(t-)}\right) \\
&= \sum_{j=1}^{N(t-)} \{\mu(t - \tau_j) + \phi(t)\} e^{-\{U(t-\tau_j) - U(\tau_{N(t-)} - \tau_j)\} - \{\Phi(t) - \Phi(\tau_{N(t-)})\}} p_{N(t-)+1,j},
\end{aligned}$$

where the probability $\mathbb{P}\left(I(t) = j|\tau_{1:N(t-)}\right) = \mathbb{P}\left(I(\tau_{N(t-)+1})|\tau_{1:N(t-)}\right) = p_{N(t-)+1,j}$ when $\tau_{N(t-)} < t \leqslant \tau_{N(t-)+1}$. Using a similar argument, the survival function is given by,

$$\begin{aligned}
S(t|\tau_{1:N(t-)}) &= \sum_{j=1}^{N(t-)} \mathbb{P}(\tau_{N(t-)+1} > t|\tau_{1:N(t-)}, I(t) = j)\mathbb{P}\left(I(t) = j|\tau_{1:N(t-)}\right) \\
&= \sum_{j=1}^{N(t-)} e^{-\{U(t-\tau_j) - U(\tau_{N(t-)} - \tau_j)\} - \{\Phi(t) - \Phi(\tau_{N(t-)})\}} p_{N(t-)+1,j}.
\end{aligned}$$

By combing the above two equations, and noting the form of the hazard function in (3.3.16), the expression in (3.3.15) holds true. □

Remark 3.3.2. *From Algorithm 1, for each $i \in 2, \ldots, n+1$ there are $5(i-1)$ function calls of $\mu(\cdot)$, $U(\cdot)$, $h(\cdot)$, $H(\cdot)$, or $\exp(\cdot)$; at most $13(i-1)$ additions or subtractions; at most $7(i-1)$ multiplications or divisions; at most $9(i-1)+7$ assignments.*

Therefore the time complexity of the algorithm is of the order $O(17n(n+1)) = O(n^2)$. *Also, each of the 8 vectors has a length* $\leqslant n$, *and therefore the space (memory) complexity is* $O(n)$.

As in Wheatley et al. (2016), this chapter will focus on the parametric estimation of the RHawkes model. Suppose the functions $\mu(\cdot) = \mu(\cdot;\theta_\mu)$ and $h(\cdot) = h(\cdot;\theta_h)$ are parametrized through finite dimensional parameters θ_μ and θ_h respectively, and let $\theta = ({\theta_\mu}^\top, \eta, {\theta_h}^\top)$ denote the parameter vector of the model. The MLE of θ is formally defined as,

$$\hat{\theta} = \arg\max L(\theta) \equiv \arg\max L(\mu(\cdot;\theta_\mu), \eta, h(\cdot;\theta_h)). \tag{3.3.17}$$

Computation of the MLE $\hat{\theta}$ can be done by minimizing the negative log-likelihood function $-\log L(\theta)$ using general purpose optimization routines. The variance-covariance matrix of $\hat{\theta}$ can be estimated by inverting the observed information matrix, that is, the Hessian matrix of the negative log-likelihood function.

3.4 Model assessment

This section outlines a procedure to assess the suitability of the RHawkes model on point process data. Although the goodness-of-fit of point process models is often evaluated using the time change theorem (e.g. Daley and Vere-Jones, 2003, Section 7.4), in the current context, the lack of a simple expression for the intensity process relative to the natural filtration makes this approach cumbersome. Instead, an approach based on the Rosenblatt (1952) transforms shall be used.

The Rosenblatt transformation maps a random vector with a given (continuous) joint distribution into independent and uniformly distributed random variables on the unit interval. In the current context, the transformation is given by $U = (U_1, .., U_n)$, where, $U_1 = F_1(\tau_1)$, $U_2 = F_2(\tau_2|\tau_1)$, $\cdots$, $U_n = F_n(\tau_n|\tau_{1:n-1})$, with $F_i(t|\tau_{1:i-1})$ denoting the conditional distribution function of τ_i given $\tau_{1:i-1}$. By $F_i(t|\tau_{1:i-1}) = \int_{\tau_{i-1}}^{t} p(\tau_i|\tau_{1:i-1})\mathrm{d}\tau_i$, and the expression of $p(\tau_i|\tau_{1:i-1})$ derived earlier, the Rosenblatt residuals of the observed event times are given by,

$$U_i = F_i(\tau_i|\tau_{1:i-1}) = 1 - \sum_{j=1}^{i-1} p_{ij} S_{ij}, \tag{3.4.1}$$

with the p_{ij}'s being given previously in Theorem 3.3.1 and the S_{ij}'s being given similarly to (3.3.4) by the following,

$$S_{ij} = e^{-\{U(\tau_i - \tau_j) - U(\tau_{i-1} - \tau_j)\} - \{\Phi(\tau_i) - \Phi(\tau_{i-1})\}}, \quad j = 1, \ldots, i-1.$$

Hence the adequacy of the fitted RHawkes model is assessed using the Rosenblatt residuals given by (3.4.1). The residuals are checked for approximate independence and uniformity, with the parameters $\mu(\cdot)$, η, $h(\cdot)$, $U(\cdot)$, and $H(\cdot)$ involved in the p_{ij}'s and the S_{ij}'s replaced by their plug-in estimates $\hat{\mu}(\cdot) = \mu(\cdot; \hat{\theta}_\mu)$, $\hat{\eta}$, $\hat{h}(\cdot) = h(\cdot; \hat{\theta}_h)$, $\hat{U}(\cdot) = \int_0^{\cdot} \hat{\mu}(t)\mathrm{d}t$, $\hat{H}(\cdot) = \int_0^{\cdot} \hat{h}(t)\mathrm{d}t$ respectively. The independence can be visually examined using an autocorrelation function (ACF) plot, or formally tested by such tests as the Ljung-Box test or the Box-Pierce test. The uniformity can be visually checked by a QQ plot, or formally tested by the K-S test.

3.5 Model predictions

In this section, methods are presented to make predictions about future event occurrences using observations of the RHawkes process up to the censoring time, which form the foundation for assessing the predictive performance of the RHawkes model on the Japan earthquake data in Section 3.7. Here, two prediction problems are studied. The first is to predict the time of the first event after the censoring time. The second is to forecast the number of events from the censoring time until a future time point.

3.5.1 Predictive density and hazard function

The conditional predictive density of the time of the first event after the censoring time T, $\tau_{N(T)+1}$, can be straightforwardly computed from the conditional most recent immigrant probabilities $p_{n+1,j}$ computed in Section 3.3, which is given by,

$$p\left(\tau_{n+1} | \tau_{1:n}, \tau_{n+1} > T\right) = \frac{\sum_{j=1}^{n} p_{n+1,j} d_{n+1,j} 1\{\tau_{n+1} > T\}}{\mathbb{P}\left(\tau_{n+1} > T | \tau_{1:n}\right)}, \tag{3.5.1}$$

where the $p_{n+1,j}$'s are calculated using (3.3.5), the denominator is computed using (3.3.7)–(3.3.10) and (3.3.14), and the $d_{n+1,j}$'s are given in (3.3.3). Furthermore, it follows that the conditional hazard rate function of $\tau_{N(T)+1}$ is given by,

$$\mathrm{haz}\left(\tau_{n+1} | \tau_{1:n}, \tau_{n+1} > T\right) = \frac{\sum_{j=1}^{n} p_{n+1,j} d_{n+1,j} 1\{\tau_{n+1} > T\}}{\sum_{j=1}^{n} p_{n+1,j} \tilde{S}_{n+1,j}}, \tag{3.5.2}$$

where $\tilde{S}_{n+1,j} = e^{-\{U(\tau_{n+1}-\tau_j)-U(\tau_n-\tau_j)\}-\{\Phi(\tau_{n+1})-\Phi(\tau_n)\}}$.

3.5.2 Predictive simulations

A simulation-based approach can predict the number of events from the censoring time T to a later time $\tilde{T}$, similar to Section 7.5 of Daley and Vere-Jones (2003),

since the linear nature of the RHawkes process model facilitates straightforward and efficient simulations, as explained later in Section 3.6.1. The following steps outline an approach to simulate the RHawkes process over the interval $(T, \tilde{T}]$, conditional on $N(T) = n$ and the values of $\tau_{1:n}$.

1. Simulate the birth times $\{t_1^0, t_2^0, \dots\}$ of all the offspring of the individuals in the population by time T, as the event times over the interval $(0, \tilde{T} - T]$ of a non-stationary self-exciting point process (Chen and Hall, 2013) with baseline event intensity $\nu(t) = \sum_{j=1}^{n} \eta h(T + t - \tau_j),\ t \geqslant 0$ and excitation function $g(t) = \eta h(t)$, translated into the interval $(T, \tilde{T}]$.
2. Simulate $I(T)$, the event index of the most recent immigration before time T, according to its conditional distribution $\mathbb{P}(I(T) = j | \tau_{1:n}, \tau_{n+1} > T) = p_{n+1,j}$, $j = 1, \dots, n$, and denote the realized value by l.
3. Simulate the time t_1^1 of the first immigration after time T as $\tau_l + W$, where W is distributed as an inter-immigration waiting time W subject to the condition that $W > T - \tau_l$, and can be simulated using a rejective method.
4. If $t_1^1 \geqslant \tilde{T}$, collect all the event times in the interval $(T, \tilde{T}]$ simulated so far, and finish.
5. If $t_1^1 < \tilde{T}$, simulate the RHawkes process on the interval $(0, \tilde{T} - t_1^1]$ and translate the corresponding event times into the interval $(t_1^1, \tilde{T}]$; and simulate the birth times in the interval $(t_1^1, \tilde{T}]$ of all offspring of the first immigrant at t_1^1, as the event times in the interval $(0, \tilde{T} - t_1^1]$ of a non-stationary self-exciting process with baseline intensity function and excitation function both equal to $\eta h(\cdot)$, translated to the interval $(t_1^1, \tilde{T}]$. Then collect all the simulated event times in the interval $(T, \tilde{T}]$, and finish.

Simulation of the self-exciting processes described in the steps mentioned above can be performed by using an efficient cascading algorithm, also explained in Section 3.6.1. From here, point and interval prediction of the number of events in $(T, \tilde{T}]$ can be computed by simulating the RHawkes process over $(T, \tilde{T}]$ for a large number of times and extracting the median or mean and appropriate quantiles of the number of events. However, the above predictions rely on the model parameters. In practice, the parameters can be estimated from the data observed until the censoring time. However, the use of the estimated values of the parameters in place of the true value can potentially lead to overly confident predictions. These effects should not be detrimental when there is sufficient data to guarantee accurate estimates of the parameters. If there is concern about the prediction intervals being too narrow, the randomness in the parameter estimates can be accounted for by sampling the needed parameter vector from its sampling distribution, which might be approximated by

a normal distribution with mean and variance-covariance matrix equal to the MLE and the inverse of the observed information matrix respectively, or by a bootstrap distribution of the MLE.

3.6 Simulations

This section evaluates the numerical performance of the direct MLE of the RHawkes model and compares the direct MLE with the E-M algorithms of Wheatley et al. (2016) using simulated data.

3.6.1 Simulation algorithm

The RHawkes process with supplied parameters $\mu(\cdot)$, η, and $h(\cdot)$ can be efficiently simulated using a cascading algorithm motivated by the cluster process representation of the RHawkes process. Let T be a predetermined censoring time. Then, to simulate the event times of the RHawkes process up to time T, first simulate the immigrant arrival times up to time T, as the cumulative sums of *i.i.d.* positive random variables with hazard rate function $\mu(\cdot)$. Furthermore, denote the simulated immigrant arrival times by $\tau_1^0 < \tau_2^0 < \cdots < \tau_{n_0}^0 \leqslant T$. Then for each $i = 1, \ldots, n_0$, simulate the corresponding offspring birth times up to time T, which can be achieved by simulating a non-stationary self-exciting point process with baseline intensity function and excitation function both equal to $\eta h(\cdot)$ on the interval $(0, T - \tau_i^0]$, and translating the event times into the interval $(\tau_i^0, T]$.

The non-stationary self-exciting point process itself can be simulated using a similar cascading algorithm as follows (Møller and Rasmussen, 2005). First simulate the generation 0 individual arrival times of the non-stationary self-exciting point process according to a Poisson process on $(0, T]$ with intensity function $\nu(\cdot)$; then keep simulating generation i ($i = 1, 2, \ldots$) events in the interval $(0, T]$ according to Poisson processes with a common intensity function $\eta h(\cdot)$ as long as the number of generation $i - 1$ individuals in the interval $(0, T]$ is not zero. When this recursive process stops, collect event times of all generations as the event times of the non-stationary self-exciting point process on the interval $(0, T]$. See e.g., the programs `simHawkes0` and `simHawkes1` in the file "simHawkes.R" of the supplementary materials for Chen and Hall (2016), for implementations of the algorithm in the R language.

3.6.2 Simulation models

The simulation models used in this section and throughout this thesis are similar to those used by Wheatley et al. (2016). The renewal process for immigrant arrivals has

inter-renewal waiting times following a Weibull distribution with shape parameter κ, scale parameter β, with density function given by,

$$g(t) = \frac{\kappa}{\beta^\kappa} t^{\kappa-1} \exp\left(-\left(\frac{t}{\beta}\right)^\kappa\right), \quad t \geqslant 0, \tag{3.6.1}$$

and hazard function,

$$\mu(t) = \frac{\kappa}{\beta^\kappa} t^{\kappa-1}, \quad t \geqslant 0. \tag{3.6.2}$$

Notice that when κ is unity, the model corresponds to the classical Hawkes process with a constant background intensity $\mu = 1/\beta$. The offspring density $h(\cdot)$ is exponential with shape parameter (or mean) γ,

$$h(t) = \frac{1}{\gamma} \exp\left(-\frac{t}{\gamma}\right), \quad t \geqslant 0. \tag{3.6.3}$$

Two illustrations of the Weibull distributions were studied. In the first illustration, the shape and scale parameters were $\kappa = 3$ and $\beta = 1.2$ and in the second example $\kappa = 1/3$ and $\beta = 0.2$. These two situations correspond to highly bursty ($\kappa = 1/3$) and more evenly distributed ($\kappa = 3$) event times, respectively; cf. Figure 3.2.1. The scale parameter β was determined so that the expected waiting time between immigrations is close to unity. The shape parameter of the offspring density was always set to $\gamma = 1$. The branching ratio was selected to be either $\eta = 0.3$, or $\eta = 0.7$, corresponding to a low and high level of self-excitation effect, respectively. In each simulation model, two censoring times T were determined to ensure the expected number of events by the censoring time were approximately 400 and 800 respectively. For each combination of $\kappa, \beta, \gamma, \eta$, and T, the RHawkes process was simulated 1000 times. For each simulated data set, the MLE was directly computing by minimizing the negative log-likelihood function using the derivative-free Nelder-Mead simplex method, and the Hessian matrix by numerical differentiation. The computations were implemented using the R language (R Core Team, 2016), with the aid of the `optim` function.

3.6.3 Simulation results

The estimation results are reported in Table 3.6.1, which contains the mean of the 1000 parameter estimates (Est), the mean of the 1000 standard error estimates obtained by inverting the observed information matrix ($\widehat{\text{SE}}$), the empirical standard error of each estimator (SE), i.e. the standard deviation of the 1000 estimates, the empirical coverage probability (CP) of the 1000 approximate 95% confidence intervals (CIs) computed by assuming (asymptotic) normality of the estimators, and the average running time (RT) of the `optim` routine to compute the minimizer

of the negative log-likelihood function and the Hessian matrix at the minimizer, on Intel Xeon X5675 processors (12M cache, 3.06 GHz, 6.4GT/S QPI).

	$\kappa = 3$		$\beta = 1.2$		$\gamma = 1$		$\eta = 0.3$	
T	300	600	300	600	300	600	300	600
Est	3.028	3.010	1.197	1.200	1.319	1.134	0.294	0.300
SE	0.297	0.215	0.056	0.041	1.986	0.623	0.044	0.030
$\widehat{\text{SE}}$	0.278	0.196	0.052	0.038	0.720	0.416	0.041	0.029
CP	0.937	0.923	0.937	0.923	0.874	0.919	0.932	0.945
RT	56.6 secs. for $T = 300$				174.3 secs. for $T = 600$			
	$\kappa = 3$		$\beta = 1.2$		$\gamma = 1$		$\eta = 0.7$	
T	132	260	132	260	132	260	132	260
Est	3.130	3.066	1.172	1.182	1.034	1.018	0.683	0.691
SE	0.800	0.573	0.145	0.107	0.324	0.199	0.063	0.044
$\widehat{\text{SE}}$	0.703	0.506	0.127	0.093	0.278	0.186	0.058	0.040
CP	0.887	0.897	0.897	0.912	0.920	0.934	0.926	0.932
RT	61.4 secs. for $T = 132$				187.6 secs. for $T = 260$			
	$\kappa = 1/3$		$\beta = 0.2$		$\gamma = 1$		$\eta = 0.3$	
T	375	775	375	775	375	775	375	775
Est	0.327	0.326	0.225	0.215	1.016	0.997	0.306	0.310
SE	0.020	0.016	0.057	0.038	0.233	0.166	0.051	0.037
$\widehat{\text{SE}}$	0.018	0.013	0.049	0.033	0.235	0.158	0.050	0.035
CP	0.930	0.873	0.968	0.951	0.933	0.941	0.945	0.941
RT	55.1 secs. for $T = 375$				177.0 secs. for $T = 775$			
	$\kappa = 1/3$		$\beta = 0.2$		$\gamma = 1$		$\eta = 0.7$	
T	145	320	145	320	145	320	145	320
Est	0.334	0.327	0.231	0.228	1.025	1.001	0.686	0.700
SE	0.033	0.025	0.103	0.065	0.185	0.122	0.062	0.045
$\widehat{\text{SE}}$	0.032	0.022	0.084	0.056	0.179	0.121	0.062	0.043
CP	0.947	0.919	0.959	0.951	0.960	0.942	0.952	0.941
RT	52.5 secs. for $T = 145$				157.6 secs. for $T = 320$			

Table 3.6.1: Estimation results using the maximum likelihood method of the RHawkes processes with Weibull distributed inter-immigration waiting times and exponential offspring densities, based on 1000 simulated datasets in each case.

Table 3.6.1 indicates that the parameter estimates are all close to their respective true parameter values, relative to their standard errors, with the empirical biases and standard errors decreasing with the censoring time. The coverage probabilities of the 95% CIs are close to the nominal confidence level, notably with larger censoring times. However, this is not evident for the shape parameter κ. The significant variance in the waiting time distribution between immigrants when $\kappa = 1/3$ and $\beta = 0.2$, which is given by,

$$\beta^2 \left[\Gamma(1 + \frac{2}{\kappa}) - \Gamma\left(1 + \frac{1}{\kappa}\right)^2 \right] = 0.2^2 \left(\Gamma(7) - \Gamma(4)^2 \right) = 27.36,$$

implies that a larger sample would be required to employ any asymptotic results and as such, the coverage probability for the shape parameter κ was reduced with the larger censoring time due to the finite and small sample size. Using a much

longer censoring time, these effects can be shown not to be present and the coverage probability was very close to the nominal level 0.95. The means of the estimated standard errors approximately agree with the empirical standard errors in most circumstances, and the agreement improves as the censoring time increases. However, on some occasions the standard error estimate were biassed downwards slightly. The branching ratio η is well estimated even with shorter censoring times.

In the instances when the shape parameter for the inter-immigration waiting time distribution is large ($\kappa = 3$), the estimation of the offspring shape parameter γ seems quite tricky with the MLE showing a noticeably larger empirical bias and standard error when the branching ratio is small ($\eta = 0.3$). Furthermore, the estimation of the shape parameter κ for the inter-immigration waiting time distribution seems complicated when the branching ratio is large ($\eta = 0.7$). This seems to be expected as there tend to be fewer offspring events when η is small, and fewer immigration events when η is large, given the total number of events expected is more or less fixed. In the case when the shape parameter is small ($\kappa = 1/3$), the estimators are less affected by this phenomenon, which might be due to the contrast between the heavy-tailedness of inter-immigration times and the light-tailedness of the waiting times to offspring births, in this circumstance, which makes it easier to disentangle the two types of events. However, even in the more difficult cases, the empirical bias and standard error of the estimators reduce as the censoring time increases. Therefore, the MLE for the RHawkes process and the inverse Hessian matrix variance estimator have satisfactory finite sample performances.

The average time required to compute the MLE and the Hessian matrix with datasets containing about 400 events is approximately one minute, and the required time on datasets with about 800 events is about three minutes. The latter time is less than four times of the former, because with a larger dataset the log-likelihood surface experiences larger curvatures, and the Nelder-Mead algorithm manages to converge with a smaller number of likelihood function evaluations, although the time to compute the likelihood with a larger dataset was approximately four times that with the corresponding smaller dataset, as expected by the time complexity of the proposed method.

3.6.4 Comparison with the E-M algorithms of Wheatley et al. (2016)

This section conducts a comparison of the direct MLE and the estimators from the E-M algorithms, called EM1 and EM2, of Wheatley et al. (2016) discussed in Chapter 2. The simulations consist of 100 datasets from the simulation model discussed in Section 3.6.2 with parameters $(\kappa, \beta, \gamma, \eta) = (3, 1, 0.5, 0.5)$ and censoring

times $T = 100$, $T = 200$, and $T = 400$. The convergence criteria for the E-M algorithms was that the absolute difference between the maximizers in consecutive iterations was less than 1×10^{-8}. The simulation results are reported in Table 3.6.2 and contains the the empirical bias of the estimator (bias), the mean square error of the estimator (MSE), the average running time per iteration (RT), which is the average time required by one likelihood evaluation in the case of MLE, or the average time required by one iteration of the E-M algorithm, and the average number of iterations (Iter No.).

Wheatley et al. (2016) discovered that the EM1 and EM2 algorithms produce very comparable estimates of the parameters, which was further confirmed in the experiments on simulated datasets in this section, with the short censoring time $T = 100$. Therefore, to save computational time, the much slower EM1 algorithm was not run on the larger datasets simulated with the larger censoring times. Unlike the MLE whose empirical bias and MSE manage to shrink toward 0 as the censoring time increases, the empirical bias and MSE of the E-M algorithm estimators do not seem to converge to 0. This suggests that the estimators from the E-M algorithms of Wheatley et al. (2016) are different from the MLEs, and they do not appear to be consistent while the MLE does. Additionally note that, while the direct MLE and the E-M algorithms both seem to require quadratic computational time with increasing censoring times, the direct MLE method is considerably faster than the E-M algorithms, particularly on larger datasets.

		$\kappa = 3$		$\beta = 1$		$\gamma = 0.5$		$\eta = 0.5$		RT	Iter
T		bias	MSE	bias	MSE	bias	MSE	bias	MSE	(secs.)	No.
100	MLE	0.161	0.282	0.013	0.006	−0.019	0.028	−0.004	0.005	0.086	205
	EM1	−1.063	1.207	−0.158	0.032	−0.174	0.066	−0.128	0.021	1.280	101
	EM2	−1.064	1.210	−0.159	0.032	−0.174	0.067	−0.129	0.021	1.160	62
200	MLE	0.060	0.160	0.002	0.003	−0.023	0.016	0.001	0.002	0.272	196
	EM2	−1.083	1.216	−0.172	0.033	−0.201	0.060	−0.125	0.018	4.285	64
400	MLE	0.044	0.070	0.008	0.002	−0.015	0.009	−0.001	0.001	0.895	188
	EM2	−1.063	1.147	−0.165	0.029	−0.221	0.055	−0.127	0.017	16.798	62

Table 3.6.2: Estimation results comparing the three estimation methods MLE, EM1 and EM2, on simulated data.

Considering the well-documented success of the E-M algorithms in the classical Hawkes process and its multivariate and marked versions (Chornoboy et al., 1988; Mino, 2001; Veen and Schoenberg, 2008; Halpin, 2013; Olson and Carley, 2013), the discrepancy between the direct MLE and the E-M algorithm based estimators of Wheatley et al. (2016) observed here warrants investigation. The first issue of the E-M algorithms of Wheatley et al. (2016) was that when calculating the conditional distributions of the missing data given the observed data $\{\tau_{1:n}, \tau_{n+1} > T\}$, Wheatley et al. implicitly assumed conditional independence of $\{M_{1:i}\}$ and $\tau_{i+1:n+1}$. As an attempt to address this issue, one can derive the joint distribution of $M_{1:n}$ given all

the observations $\{\tau_{1:n}, \tau_{n+1} > T\}$ using a recursive algorithm similar to that used in the proof of Theorem 3.3.1, and integrate out the other dimensions to arrive at the marginal distribution of M_i. With the correct conditional distributions of the missing data used in the Expectation step, the E-M algorithms were found to produce the same estimates as the direct MLE (modulo numerical rounding error) on small datasets (with $n = 4$ or 5). However, this method is not feasible on data with realistic sizes, as the storage of the joint probability mass functions takes exponential space (or even factorial in the case of EM1), and the marginalization step takes exponential time (or even factorial in the case of EM1). The second issue as mentioned previously is that the conditional distributions of M_i given $\tau_{1:i}$ is incorrect as the probabilities $\pi_{i,j}^{[m]}$'s in (2.2.12) (eq. (16) of Wheatley et al. (2016)) do not sum to one as they should for fixed i.

3.7 Applications

3.7.1 Earthquakes in the Japan Pacific Ring of Fire

In this application, the Japan earthquake data previously analyzed by Ogata (1988) is investigated. The data consists of 483 earthquakes with magnitude 6 or greater that occurred in a polygonal region in the vicinity of Japan from 1885-1980. The polygonal region is part of the so-called *Pacific Ring of Fire*, and is defined by the sequence of vertex points (42°N, 142°E), (39°N,142°E), (38°N, 141°E), (35°N,140.5°E), (35°N, 144°E), (42°N,146°E), and (42°N, 142°E), as indicated by the polygon in Figure 3.7.1. The earthquake data shown in this graph was downloaded from the earthquake archive maintained by the United States Geological Survey (USGS).

This analysis only examines the occurrence times of the earthquakes. The B-index of waiting times between earthquakes was found to be 366.9, comparable to those of the RHawkes processes (a) and (d) illustrated at the end of Section 3.2. Therefore, an RHawkes process with Weibull inter-immigration waiting time distribution and exponential offspring density was fitted to the earthquake times. The resulting MLEs of the parameters were as follows, $\hat{\kappa} = 0.314(0.019)$, $\hat{\beta} = 22.2(5.47)$, $\hat{\gamma} = 1266(356)$, $\hat{\eta} = 0.512(0.047)$, with the numbers in brackets being the standard errors. The Rosenblatt residuals were calculated, and their uniform QQ plot and ACF plot are shown in the upper panels of Figure 3.7.2. The plots suggest good agreement between the empirical and theoretical quantiles and with minimal to no serial correlation among the residuals up to lag 26, which were respectively confirmed by the large p-values of the K-S test of uniformity ($P = 0.93$) and the Ljung-Box test of independence ($P = 0.22$).

In the seismological context, the immigrant events and offspring events have a natural interpretation as mainshocks and aftershocks, respectively. By the fit-

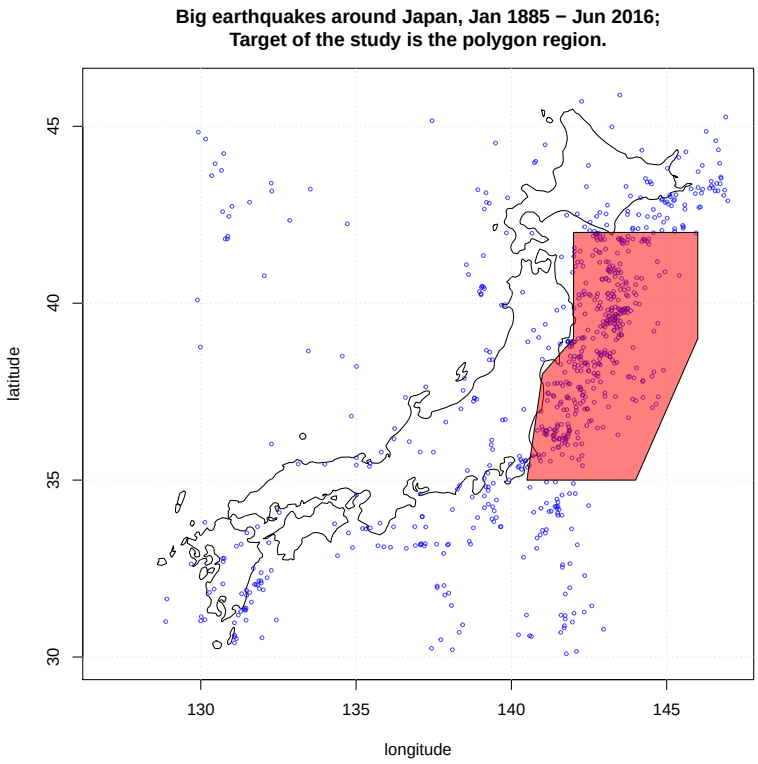

Figure 3.7.1: Big earthquakes (Magnitude $\geqslant 6$) around the Japan Pacific Ring of Fire. Each open dot indicates one earthquake. The analysis uses only the data for the polygonal region.

ted RHawkes process model, the mainshocks arrive according to a renewal process with mean waiting time between arrivals equal to $\hat{\beta}\Gamma(1 + 1/\hat{\kappa}) = 169$ days, and each (main or after) shock on average directly induces 0.51 aftershocks, and given a shock generates an aftershock, the waiting time between the shock and the first aftershock is exponentially distributed with mean equal to $\hat{\gamma} = 1266$ days. The estimated shape parameter $\hat{\kappa}$ of the Weibull distribution for the waiting times between successive mainshocks is significantly less than one, suggesting that the numbers of mainshocks in neighboring nonoverlapping time intervals are not independent, but positively correlated. Such a positive correlation could be due to some latent variable(s) underlying the mainshock intensity.

For comparison, the classical Hawkes process was fitted to the earthquake data with a constant immigration rate ν and an excitation function following the modified Omori's law (Utsu, 1961),

$$\eta h(t) = K\,(t + c)^{-p}\,; \tag{3.7.1}$$

see also Ogata (1988, eq. 14). The resulting MLE of the parameters are as follows, $\hat{\nu} = 0.00523(0.000934)$, $\hat{K} = 0.0448(0.00496)$, $\hat{c} = 0.0167(0.00886)$, and $\hat{p} = 0.996(0.0372)$. The estimated value of ν suggests the mean waiting time between mainshocks is $1/\hat{\nu} = 191$ days, similar to that suggested by the RHawkes model.

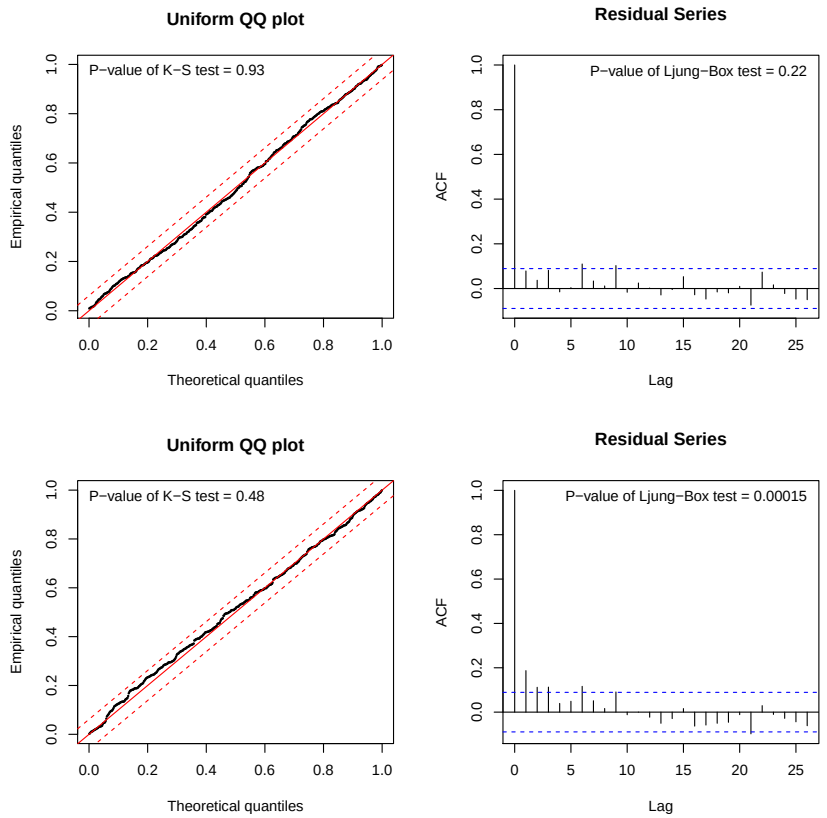

Figure 3.7.2: Graphical comparison of the goodness-of-fit assessment by employing the Rosenblatt residuals, with the upper two panels for the RHawkes process model and lower two panels for the Hawkes process model.

The estimated value of p agrees very well with $p = 1$ in the original Omori's law (Omori, 1894). However, with such a value of p, the corresponding Hawkes process model is not stable, as the integral of the excitation function diverges. A Hawkes process with an excitation function in the exponential decay form (Ozaki, 1979) was also fit to the data. However, the fit was worse than with the polynomial excitation function (3.7.1), and therefore not reported here.

The Rosenblatt residual plots are displayed in the lower two panels of Figure 3.7.2. Notice that the uniformity of the residuals is quite acceptable with the K-S test yielding a reasonably large p-value of 0.48. However, the ACF plot reveals rather strong serial correlation among the residuals, with the Ljung-Box test returning a very small p-value of 1.5×10^{-4}, which firmly rejects independence of the residuals. Thus, it can be concluded that the original Hawkes process with a polynomial decay excitation function in the form of the modified Omori's law is not a sufficient model for the Japan earthquake data considered here, while the RHawkes process seems to be an ideal model.

To evaluate the predictive performance of the RHawkes model identified here and to illustrate the prediction methods in Section 3.5, a simulation-based approach

to predict significant earthquake occurrences in the study region from 1/1/1981 to 30/6/2016, the 35.5 years from the censoring time to the time of performing this analysis will be implemented. First, 10,000 realizations of the identified RHawkes process model are simulated from the censoring time up to 30/6/2016, conditional on the earthquake times by the censoring time. The point-wise median and lower and upper 2.5 percentiles of the simulated sample paths of the RHawkes process are displayed in Figure 3.7.3, as well as the actual sample path based on data during the prediction window, downloaded from the earthquake archive of the United States Geological Survey (USGS). Notice that, as a point prediction, the median tracks the actual earthquake count reasonably well, and the point-wise 95% prediction intervals contain the actual counts during the whole prediction window. In particular, the actual number 179 of earthquakes during the prediction window is well within the 95% prediction interval $[18, 1011]$, and is not far from the median 149 or mean 211 of the simulated numbers of earthquakes in the prediction window.

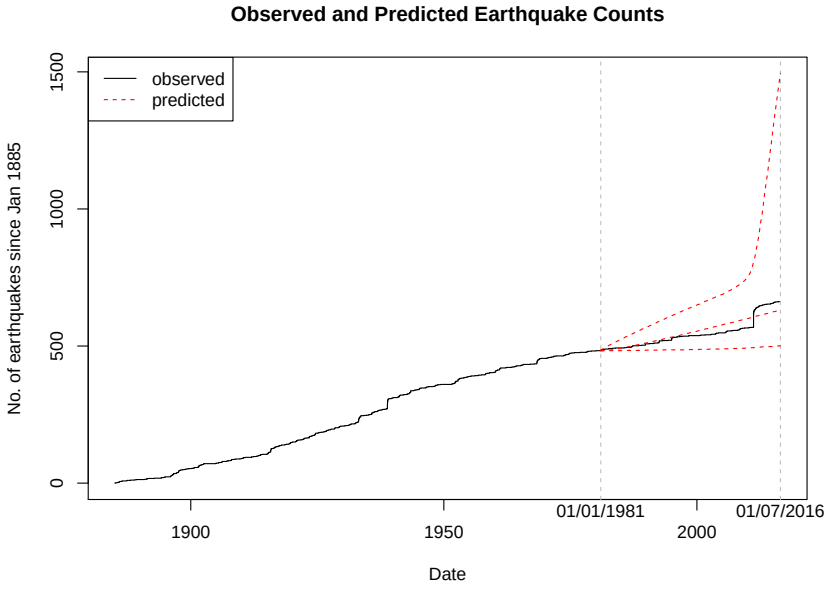

Figure 3.7.3: Actual and predicted earthquake occurrences. The solid curve shows the actual earthquake counts, and the dashed curves show the point prediction and 95% prediction intervals at different time points.

The predictive density of the waiting time until the first earthquake after the censoring time and the corresponding hazard rate function are displayed in Figure 3.7.4, as well as the density histogram of the sample of 10,000 waiting times extracted from the simulated sample paths of the RHawkes process. From the graph, good agreement between the predictive density and the density histogram can be observed. The graph infers that the probability of observing the first earthquake after the censoring time in the first, second, third and fourth 100-day period were roughly

30%, 20%, 13%, and 10% respectively. By the USGS data, the first big earthquake after the end of the year 1980 in the study region occurred at 1981-01-18 18:11:28 GMT, during the first 100-day period. The predictive hazard function provides an estimate of the conditional probability of observing the first big earthquake on any day given that it has not occurred by the previous day. For example, by Figure 3.7.4, the conditional probability of seeing the first magnitude 6 or higher earthquake in the region on the first day of 1981 is about 0.36%, and this conditional probability decreases gradually over time, and halves in 3 years.

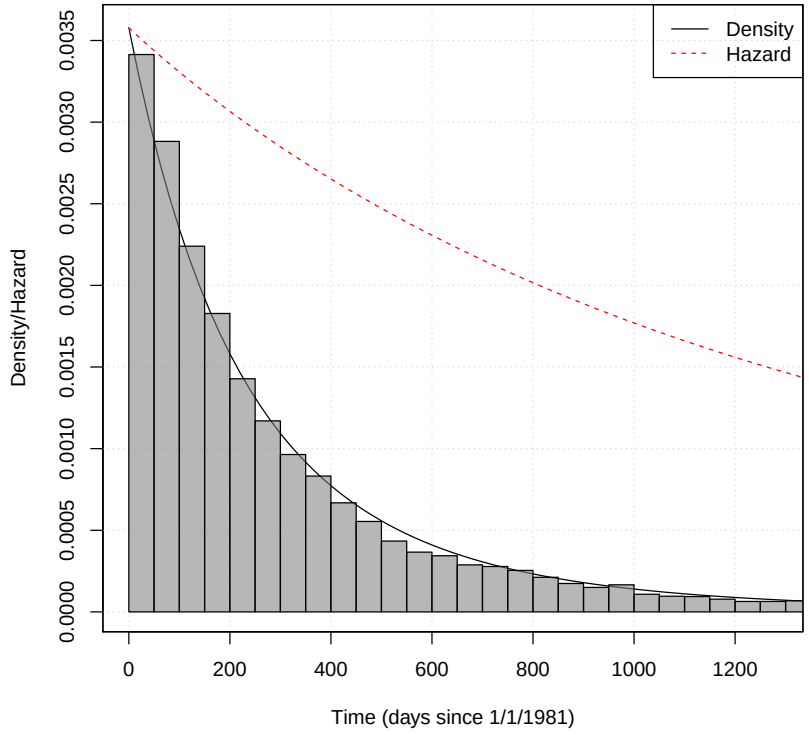

Figure 3.7.4: Predictive density and hazard functions of the waiting time to the first earthquake after the censoring time (end of the year 1980). The solid line is the density, and the dashed line is the hazard function.

3.7.2 Mid-price changes of the AUD/USD currency exchange rate

The fluctuations of the prices of financial assets such as stocks and foreign currencies can be due to factors external or internal to the specific markets in which the assets are traded. Introduction of high-frequency and algorithmic trading has furthered the need to understand the relationship between exogenous and endogenous effects in the financial markets. Recently, Filimonov and Sornette (2012) used the

Hawkes process to quantify the endogeneity or reflexivity of the mid-quote price movements of the E-mini S&P500 futures contracts. More recently, Wheatley et al. (2016) used the renewal Hawkes process for the same purpose. By analyzing hundreds of datasets obtained from 20-minute time intervals, Wheatley et al. reported that the RHawkes model was able to provide an adequate fit to the majority (about 80%) of the datasets. When the Hawkes or RHawkes process is used to model the times of price changes, the branching ratio η has the interpretation as the proportion of price changes that are not caused by external information but due to the reactions of market participants to previous price changes, and hence quantifies market reflexivity.

Motivated by these previous works, this analysis aims to quantify the level of endogeneity in the foreign exchange (forex) market using the RHawkes process. Specifically, the mid-quote price changes of the AUD/USD foreign exchange rate during the trading week from 20:00:00 Greenwich Mean Time (GMT) on Sunday 19/7/2015, when the New Zealand forex market officially opened on Monday at 8 AM local time, to 21:00:00 GMT Friday 24/7/2015, when the New York forex market

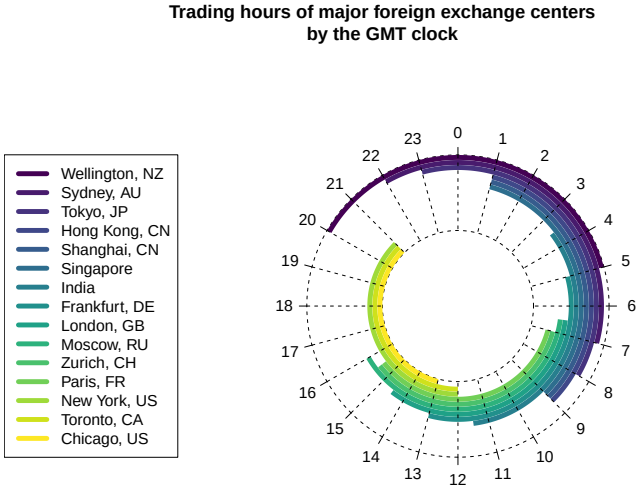

Figure 3.7.5: GMT clock displaying the trading hours of the major foreign exchange centres around the world.

officially closed on Friday at 5 PM local time; see Figure 3.7.5 for the trading hours of major forex trading centers around the world will be investigated. The mid-quote price is defined as the average of the best bid and best ask prices. By Figure 3.7.6, which shows the hourly number of mid-quote price changes in the whole trading week, the price change exhibits a distinct diurnal pattern, and thus the point process of price changes for the whole period is not stationary. Therefore, similar to Filimonov and Sornette (2012) and Wheatley et al. (2016), this analysis will be performed on sequences of mid-price changes that occurred within short time

windows of fixed length, although this work uses non-overlapping one-hour time windows, rather than the overlapping 20-minute windows used in their works, so that there is sufficient data to fit the model in each time window while avoiding the issue of non-stationarity to some extent. This yields 121 point process observations with the total number of events ranging from a minimum of 108 to a maximum of 3961. Although the timestamps in this data set are accurate down to microseconds (10^{-6} seconds), on a number of occasions, two or even three price changes were recorded to occur simultaneously, and so a small random noise was added to these event times to break the ties while maintaining the time order of the price moves.

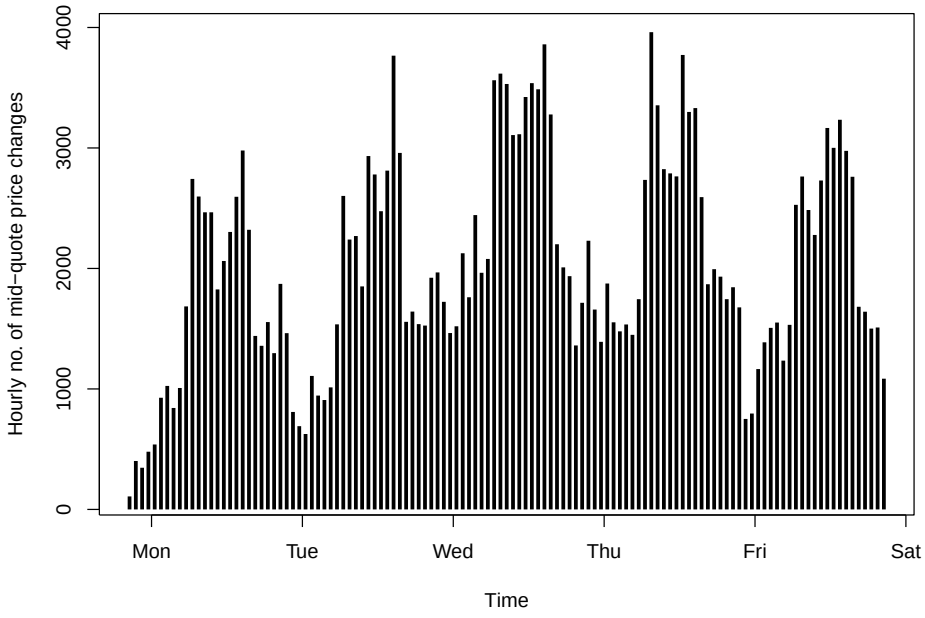

Figure 3.7.6: Hourly number of mid-quote price changes of the AUD/USD exchange rate during the trading week 19/7/2015 – 24/7/2015.

The RHawkes process model was fitted to the hourly data with Weibull or gamma inter-immigration waiting time distributions and an exponential offspring density. Figure 3.7.7 presents the goodness-of-fit test results, which suggest that both models produce very similar fits to the data, although the fit by the Weibull model is slightly better. Therefore, this section will only report the results of the Weibull model. At the 1% level, the Weibull model passed both the K-S test of uniformity and the Ljung-Box test of independence of the Rosenblatt residuals on 100 (82.6%) hourly datasets, indicating the Weibull RHawkes model can fit the hourly data well in the majority of cases. The cases where the RHawkes model fit was inadequate are likely due to non-stationarity of the price change process in the corresponding hourly intervals. For example, on four of the five trading days, the fit to the data of the 6:00

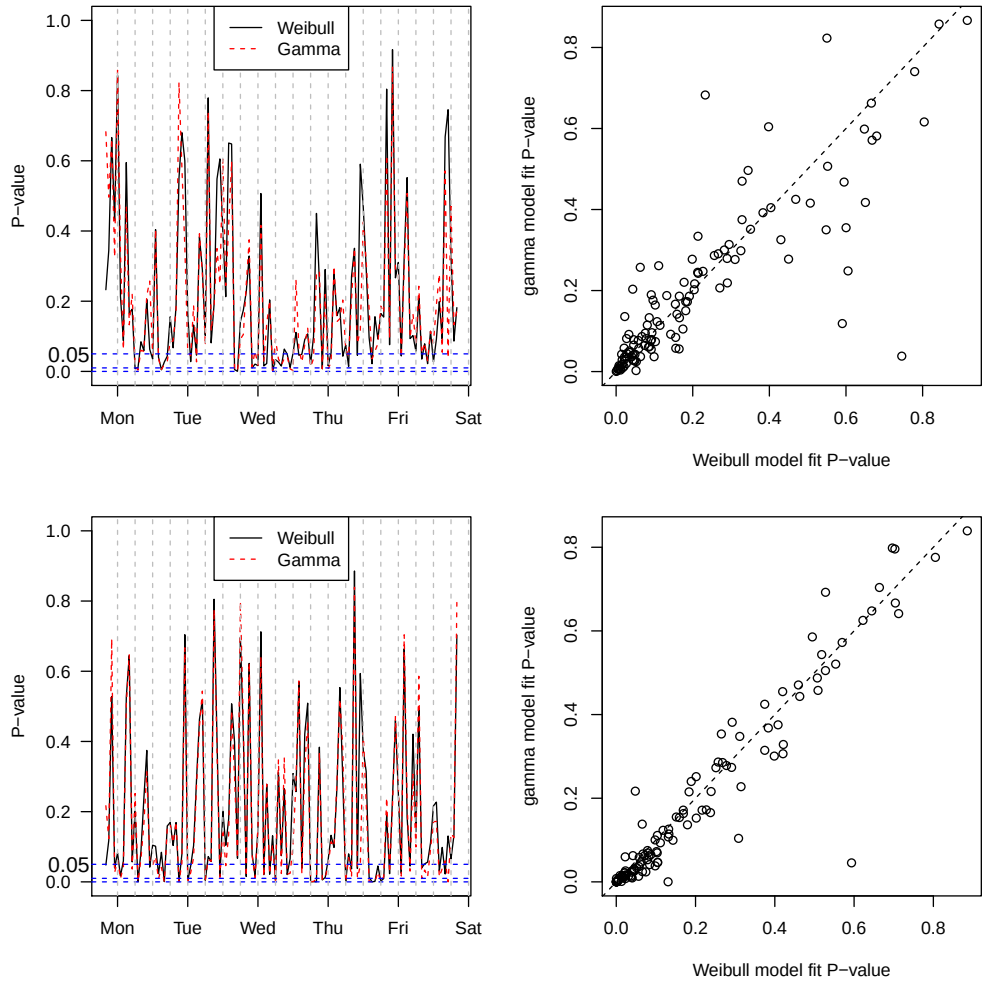

Figure 3.7.7: Goodness-of-fit tests of the RHawkes models with either Weibull or gamma inter-immigration waiting times and exponential offspring density, on the hourly AUD/USD exchange rate data in the trading week 19/7/2015 –24/7/2015. Top panels: K-S tests of the uniformity of the Rosenblatt residuals; bottom panels: Ljung-Box test of the independence of the Rosenblatt residuals.

hour was inadequate, which is not surprising as the 6:00 hour contains the opening at 6:30 of the London forex center, the largest forex trading center (by turnover) of the world, which has caused substantial non-stationarity of the price change process, with price changes in the hour occurring substantially more frequently after 6:30.

Figure 3.7.8 displays the time series of MLEs of the model parameters. The Weibull shape parameter estimate $\hat{\kappa}$ exhibits a fairly clear diurnal pattern with peaks occurring around midnight GMT, and troughs around midday GMT. The value of $\hat{\kappa}$ is mostly less than one, with the 95% confidence intervals below one 94.2% of the time. The Weibull scale parameter estimate $\hat{\beta}$ The offspring density parameter estimate $\hat{\gamma}$ is considerably sporadic with several peaks and troughs occurring throughout the day. However, it is still noticeable that the highest peaks tend to occur around midday GMT, while the lowest trough occurs around midnight

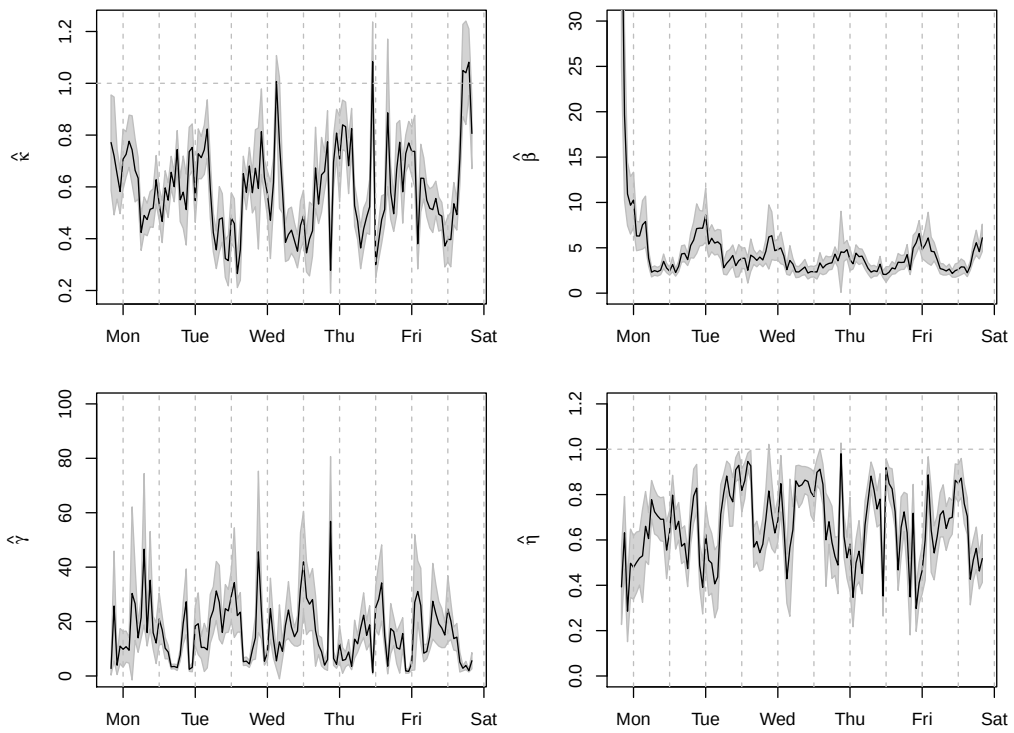

Figure 3.7.8: Time series plot of the MLEs of the parameters based on hourly data of the RHawkes process, over the trading week 19/7/2015 – 25/07/2015. Solid curve: MLE; shaded region: point-wise 95% confidence intervals.

GMT. These phenomena suggest that during the sessions when the US markets are open and overlap with the European markets (12:00 - 14:00 GMT), the exogenous effects are more potent than when the Asian markets begin to open (23:00 - 1:00 GMT), where there tend to be more endogenous transactions occurring. The branching ratio estimate $\hat{\eta}$ is quite volatile throughout the week, although it still exhibits highs and lows occurring at midday and midnight GMT respectively. The estimated values range from 0.29 to 0.98 with median 0.66 and quartiles $(0.53, 0.80)$, and this medium to high but still sub-critical ($\eta < 1$) values suggest that substantial high-frequency self-excitation is occurring in the mid-price changes. These trends and regular trading patterns during a particular day suggest that the parameters might be predictable in time. For market participants, this might be useful when deciding on trades in future time periods. The time series of estimates could be fit with a time series model and then used to forecast the parameters during the next one-hour window.

Chapter 4

Fast fitting of the renewal Hawkes process[1]

4.1 Introduction

The likelihood evaluation algorithm developed in the preceding chapter makes it viable to fit the RHawkes process model in practice, however the algorithm requires the calculation, at each event time, of the probabilities of the most recent immigrant being equal to every past event in the history of the process, and therefore its time-complexity is still quadratic. This implies that it can be intolerably slow when applied on large datasets. However, in financial applications, the Hawkes process and its extensions have been applied extensively to model data that occur in very high frequencies and have an abundance of observations. Financial data of this nature motivate the need to conduct inferences on large datasets expeditiously. This leads to the proposal of faster methods for fitting RHawkes processes which will enable the process to be applicable in real-world trading applications or other applied domains. The developments are based on two distinct frameworks. The first approach implements the Newton-Raphson method to optimize the log-likelihood function, which generally takes a much smaller number of iterations to converge than the derivative-free method, such as the Nelder-Mead downhill simplex method. The second approach employs an approximation to the likelihood function by truncating the distribution of the most recent immigrant, which is similar in spirit to the approach used by Halpin (2013) to speed up the E-M algorithm for the classical Hawkes process with non-exponential excitation function.

The choice of optimization procedure profoundly impacts the computational time required for estimation. Previously, the derivative-free Nelder-Mead simplex method was employed, which requires a large number of likelihood evaluations until convergence. To overcome this, the Newton-Raphson method which requires fewer it-

[1]Most of the content shown in this chapter has been submitted for publication.

erations until convergence is a suitable alternative. To this end, an algorithm to compute the gradient and Hessian of the log-likelihood for the RHawkes process is derived. Once these have been determined, implementation of the Newton-Raphson algorithm is rather simple. The small number of iterations and exact Hessian computation reduce the computational time required for estimation significantly on larger datasets. As a by-product, the exact Hessian matrix calculated at the last cycle of Newton-Raphson iterations is automatically available and can be used to compute estimators of the variance of the MLE, thereby avoiding the need to compute a numerical approximation to the Hessian in a separate step after finding the maximizer of the log-likelihood function.

The second approach to fast fitting of the RHawkes model is to optimize approximations to the likelihood, rather than the exact value because it is often possible to compute accurate approximations to the RHawkes process likelihood in a much shorter time than the exact likelihood. This chapter proposes simple modifications to the iterative algorithm developed in Chapter 3 that achieves significant gains in computational efficiency and memory requirements at the small cost of a minor loss in accuracy of the estimation, which enables fast fitting of the RHawkes process using the maximum likelihood method on much larger datasets. The likelihood approximation works by truncating the possible candidates for the most recent immigrant event to events in the recent past. This truncation is justified in quite general conditions since most of the past events, particularly those in the distant past have negligibly small probabilities of being the most recent immigrant at the current event time. Computing and storing these most recent immigrant probabilities unnecessarily slows down the likelihood evaluation algorithm. The possible candidates for the most recent immigrant at the i-th event time will be restricted to the most recent B_i events.

The tuning parameter B_i embodies a trade-off between computational complexity and memory requirements against the accuracy of the estimation. However, an appropriately determined sequence B_i can reduce the computational time significantly, and doing so without degrading the accuracy of the estimates. Two different methods are used to determine the tuning parameter B_i, with the first method merely fixing $B_i = B$ before likelihood evaluation and only computing the most recent immigrant probabilities for the most recent B events at each step of the likelihood evaluation algorithm. In a more dynamic approach, B_i can fluctuate at each iteration of the likelihood evaluation algorithm according to the computed most recent immigrant probabilities at the previous iteration. The B_i at each iteration depends on the waiting time distribution between successive immigrant events, the waiting time distribution between an event and its direct offspring event, and the level of self-excitation.

The computational and statistical efficiency of the proposed estimation methods for the renewal Hawkes process will be examined using a simulation study. The newly proposed estimation methods will be compared to the derivative-free Nelder-Mead algorithm used in Chapter 3. The simulations will confirm that the newly proposed estimation methods accomplish comparable accuracy but require a much shorter running time to compute estimates and standard errors. Furthermore, to highlight the applicability of the methods presented in this chapter, an analysis of the mid-price for several pairs of USD exchanges rates for four trading weeks will be investigated.

The rest of this chapter is organized as follows. Section 4.2 details the Newton Raphson algorithm for faster estimation of RHawkes processes and describe their algorithms. Next, in Section 4.3, the approximate likelihood is discussed, and in 4.4 statistical evidence will be provided using a simulation study, of the improvement in computational efficiency while preserving accurate estimation results. In the last section, the mid-price analysis of USD currency pairs is illustrated.

4.2 Estimation using Newton-Raphson optimization

Chapter 3 makes straightforward and direct evaluation of the likelihood feasible and allows MLE to be performed using standard numerical optimization procedures, such as the derivative-free Nelder-Mead simplex method by optimizing the log-likelihood function, which was shown to take the form,

$$\ell(\theta) = \log \mu(\tau_1) - U(\tau_1) + \sum_{i=2}^{n} \log \left(\sum_{j=1}^{i-1} p_{ij}\, d_{ij} \right) + \log \left(\sum_{j=1}^{n} p_{n+1,j} S_{n+1,j} \right), \quad (4.2.1)$$

where the notations used here were introduced in Chapter 3. Since the Nelder-Mead algorithm only utilizes the value of the log-likelihood at each iteration and does not explicitly account for the shape information, such as slope and the curvature, of the log-likelihood surface, it often requires many more iterations to converge than the derivative-based optimization method. Therefore, this section will consider the well-known derivative-based Newton-Raphson method.

This section will begin by briefly introducing the Newton-Raphson method and then follow with a procedure to compute the gradient vector and Hessian matrix. Let $\nabla(\theta)$ and $H(\theta)$ denote the gradient vector and Hessian matrix of the log-likelihood for the RHawkes process evaluated at the parameter vector θ, respectively. The Newton-Raphson method is an iterative procedure where at each iteration t, the following operations are computed,

$$\theta^{[t+1]} = \theta^{[t]} - H(\theta^{[t]})^{-1} \nabla(\theta^{[t]}), \quad (4.2.2)$$

then set $t \leftarrow t+1$ and repeat until a convergence criterion is satisfied. The optimization procedure is initialized at $\theta^{[0]}$. It may prove beneficial to obtain an initial estimate using the approximate likelihood approach (see Section 4.3), which would provide an initial starting point close to the true MLE. This approach can potentially reduce the computational time required for the implementation of the Newton-Raphson algorithm significantly as typically only a minimal number of iterations will be required until convergence.

For the implementation of the Newton-Raphson algorithm, a means to compute the gradient vector, and the Hessian matrix is necessary. Recall that the log-likelihood for the RHawkes process is computed using a recursive algorithm because the most recent immigrant probabilities are computed recursively. As such, the gradient vector and Hessian matrix also require a recursion to be evaluated. In the sequel, the recursive algorithm to be used when computing the gradient vector and the Hessian matrix is outlined. As mentioned earlier, an additional advantage of this method is that the Hessian matrix needed for variance estimation is automatically obtained as a by-product at the end of the iterations.

Let $\partial_\theta = \partial/\partial\theta$ and $\partial^2_{\theta\theta^\top} = \partial^2/\partial\theta\partial\theta^\top$ denote the first and second order partial derivatives with respect to the model parameters. By computing the partial derivatives of the log-likelihood in (4.2.1) with respect to the parameter vector θ, the gradient vector $\nabla(\theta) = \partial_\theta \ell(\theta)$ takes the form,

$$\begin{aligned}\nabla(\theta) = \frac{\partial_\theta \mu(\tau_1)}{\mu(\tau_1)} - \partial_\theta U(\tau_1) + \sum_{i=2}^{n} \left\{ \frac{\sum_{j=1}^{i-1} d_{ij}\partial_\theta p_{ij} + p_{ij}\partial_\theta d_{ij}}{\sum_{j=1}^{i-1} p_{ij} d_{ij}} \right\} \\ + \frac{\sum_{j=1}^{n} S_{n+1,j}\partial_\theta p_{n+1,j} + p_{n+1,j}\partial_\theta S_{n+1,j}}{\sum_{j=1}^{n} p_{n+1,j} S_{n+1,j}},\end{aligned} \tag{4.2.3}$$

where

$$\begin{aligned}\partial_\theta d_{ij} &= \Psi_{ij}\Big[\partial_\theta \mu(\tau_i - \tau_j) + \partial_\theta \phi(\tau_i) + \left(\mu(\tau_i - \tau_j) + \phi(\tau_i)\right)\partial_\theta \psi_{ij}\Big], \\ \partial_\theta S_{n+1,j} &= \Psi_{n+1,j}\partial_\theta \psi_{n+1,j}, \\ \Psi_{ij} &= e^{-\{U(\tau_i-\tau_j)-U(\tau_{i-1}-\tau_j)\}-\{\Phi(\tau_i)-\Phi(\tau_{i-1})\}}, \\ \psi_{ij} &= -\left\{U(\tau_i - \tau_j) - U(\tau_{i-1} - \tau_j)\right\} - \left\{\Phi(\tau_i) - \Phi(\tau_{i-1})\right\}.\end{aligned}$$

Here, direct equations to compute most of the terms in (4.2.3) are available. However, a method is still required to compute the derivatives of the most recent immigrant probabilities $\partial_\theta p_{ij}$ with respect to each parameter of the model. A recursive procedure to compute the derivatives of p_{ij} is derived in Appendix A.1. The derivation follows directly by applying the chain rule and product rule for differentiation multiple times.

Next, to calculate the Hessian matrix $H(\theta) = \partial^2_{\theta\theta^\top}\ell(\theta)$, the partial derivatives of the (transposed) gradient vector $\nabla(\theta_1)^\top$ in (4.2.3) with respect to θ are computed as follows,

$$\begin{aligned}
H(\theta) = \frac{\partial^2_{\theta\theta^\top}\mu(\tau_1)}{\mu(\tau_1)} - \frac{\{\partial_\theta \mu(\tau_1)\}^{\otimes 2}}{\mu(\tau_1)^2} - \partial^2_{\theta\theta^\top} U(\tau_1) \\
+ \sum_{i=2}^{n} \Bigg\{ \frac{\sum_{j=1}^{i-1} d_{ij}\partial^2_{\theta\theta^\top} p_{ij} + 2\partial_\theta d_{ij} \odot \partial_\theta p_{ij} + p_{ij}\partial^2_{\theta\theta^\top} d_{ij}}{\sum_{j=1}^{i-1} p_{ij} d_{ij}} \\
+ \frac{\left(\sum_{j=1}^{i-1} d_{ij}\partial_\theta p_{ij} + p_{ij}\partial_\theta d_{ij}\right)^{\otimes 2}}{\left(\sum_{j=1}^{i-1} p_{ij} d_{ij}\right)^2} \Bigg\} \\
+ \Bigg\{ \frac{\sum_{j=1}^{n} S_{n+1,j}\partial^2_{\theta\theta^\top} p_{n+1,j} + 2\partial_\theta S_{n+1,j} \odot \partial_\theta p_{n+1,j} + p_{n+1,j}\partial^2_{\theta\theta^\top} S_{n+1,j}}{\sum_{j=1}^{n} p_{n+1,j} S_{n+1,j}} \\
+ \frac{\left(\sum_{j=1}^{i-1} S_{n+1,j}\partial_\theta p_{n+1,j} + p_{n+1,j}\partial_\theta S_{n+1,j}\right)^{\otimes 2}}{\left(\sum_{j=1}^{i-1} p_{n+1,j} S_{n+1,j}\right)^2} \Bigg\}, \qquad (4.2.4)
\end{aligned}$$

where and henceforth $x^{\otimes 2} := xx^\top$ denotes the outer product of a vector x with itself, and $x \odot y := \frac{1}{2}(xy^\top + yx^\top)$ denotes the symmetrized outer product of two vectors x and y of the same dimension, the first derivatives are as in (4.2.3), and

$$\begin{aligned}
\partial^2_{\theta\theta^\top} d_{ij} = \Psi_{ij} \Big[\{\mu(\tau_i - \tau_j) + \phi(\tau_i)\} \left\{(\partial_\theta \psi_{ij})^{\otimes 2} + \partial^2_{\theta\theta^\top}\psi_{ij}\right\} \\
+ 2\partial_\theta \psi_{ij} \odot \partial_\theta \{\mu(\tau_i - \tau_j) + \phi(\tau_i)\} \\
+ \partial^2_{\theta\theta^\top} \{\mu(\tau_i - \tau_j) + \phi(\tau_i)\} \Big], \\
\partial^2_{\theta\theta^\top} S_{n+1,j} = \Psi_{n+1,j} \Big[(\partial_\theta \psi_{n+1,j})^{\otimes 2} + \partial^2_{\theta\theta^\top}\psi_{n+1,j} \Big].
\end{aligned}$$

Similar to the gradient vector computation, a recursion to compute the second derivatives of the most recent immigrant probabilities $\partial^2_{\theta\theta^\top} p_{ij}$ are needed. These can be computed by applying the recursive procedure as derived and outlined in Appendix A.2.

Remark 4.2.1. *The recursive algorithms to compute the gradient vector and the Hessian matrix both have quadratic computational time. The storage requirements are linear, in that an RHawkes model with m parameters requires m vectors of dimension n to store the first derivatives and a total of $m(m+1)/2$ vectors of dimension n to store the second derivatives. There is no need to store m^2 vectors due to the symmetry of the second-order partial derivatives.*

Using the expressions in (4.2.3) and (4.2.4), the Newton-Raphson method is rather simple to implement. The recursion in (4.2.2) generates a sequence of estimates $\theta^{[t]}$ that converge to the true MLE quite rapidly, particularly when the initial starting point $\theta^{[0]}$ is not far from the estimate $\hat{\theta}$.

4.3 Estimation using approximate likelihood functions

In Section 4.2, the estimation procedure was enhanced by improving the computational efficiency of the optimization procedure using the Newton-Raphson method. However, there are situations where a good starting value for the Newton-Raphson iterations is difficult to come by, and the evaluation of the gradient and Hessian of the log-likelihood still demands quadratic time. Therefore, an alternative approach to speeding up the fitting of the RHawkes process by maximizing an approximate rather than the exact log-likelihood of the RHawkes process will be beneficial. Two methods to approximate the log-likelihood are considered. In both methods, a truncation of the distribution of the most recent immigrant and the range of the excitation effect is implemented. The potential candidates for the most recent immigrant at an event time are restricted to a relatively small number of past events, instead of considering all the past events, and the self-excitation effect of an event is restricted to a small number of events in the future rather than let it last forever. The two methods differ in that the first method uses fixed ranges in the truncations at different event times, while the second uses varying ranges of truncation at different event times. The second is somewhat similar to the approach used by Halpin (2013) to approximate the complete data log-likelihood of the classical Hawkes process. It might also be worth noting that the first method is similar to the use of a fixed number of particles in the sequential Monte Carlo approximation to the likelihood of hidden Markov models.

First, a discussion on the truncation to the distribution of the most recent immigrant. Specifically, at each event time τ_i, this method only considers at most B_i most recent events as the possible candidates for the most recent immigrant. This method effectively assumes that the events which occur before the B_i-th most recent event have negligible probabilities of being the most recent immigrant event. There are two choices of B_i considered in this section. The first assumes that $B_i \equiv B$ for a large integer B and the second allows the B_i to depend on the event time τ_i. In the case of dynamic B_i's, at each step of the likelihood evaluation algorithm, the B_i is the value such that the sum of the most recent immigrant probabilities sum up to q or larger, where the threshold probability $q = 1 - \epsilon$ for a small $\epsilon > 0$. Let $c_{i,j}$ denote the cumulative most recent immigrant probability of the most recent j events at time τ_i, so that $c_{i,j} = \sum_{k=1}^{j} p_{i,i-k}$, then the dynamically chosen B_i is given

by, $B_i = \min\{j \geqslant 1 : c_{ij} \geqslant q\}$. Next, the probabilities p_{ij}, $j = i - B_i, \ldots, i-1$, are renormalized to sum to 1, and the p_{ij} for $j = 1, \ldots, i - B_i - 1$ are all set to 0. By a slight misuse of notation, these slightly modified probabilities are still denoted by p_{ij}. The inner summation terms in the log-likelihood function (4.2.1) are then approximated as follows at each iteration,

$$\sum_{j=1}^{i-1} p_{ij} d_{ij} \approx \sum_{j: i-B_i \leqslant j \leqslant i-1} p_{ij} d_{ij}, \tag{4.3.1}$$

$$\sum_{j=1}^{n} p_{n+1,j} S_{n+1,j} \approx \sum_{j: n-B_n \leqslant j \leqslant n} p_{n+1,j} S_{n+1,j}, \tag{4.3.2}$$

and the most recent immigrant probabilities $p_{i+1,j}$ at the beginning of the next iteration are still calculated using the recursion (3.3.5), before they are subsequently truncated.

At this point, it appears already that the approximation to the RHawkes process log-likelihood is linear in computational time even without the truncation to the excitation effect, at least when $B_i \equiv B$, since at each iteration, at most B computations are needed to calculate the most recent immigrant probabilities p_{ij}, $j = i-B, \ldots, i-1$, at most B computations to calculate the conditional densities d_{ij} or the conditional survival probabilities $S_{n+1,j}$, $j = i - B, \ldots, i-1$, given in (3.3.3) and (3.3.4), and a final summation of at most B terms in (4.3.1) or (4.3.2). However, this is not true in general, because the computation of p_{ij}, d_{ij} and S_{ij} involves $\phi(\tau_i) = \sum_{j=1}^{i-1} g(\tau_i - \tau_j)$ and $\Phi(\tau_i) = \sum_{j=1}^{i-1} G(\tau_i - \tau_j)$, both of which require linearly growing time to compute in general, due to their dependence on all past event times τ_j, $j = 1, \ldots, i-1$. Note that the notation $g(t) = \eta h(t)$ and $G(t) = \eta H(t)$ has been used here, and will be used for the remainder of this chapter.

For genuine linear-time approximation algorithm, one more approximation is needed to make sure that the computation time required for the d_{ij}'s and S_{ij}'s for each i stays bounded. To this end, observe that their dependence on $\Phi(\tau_i) - \Phi(\tau_{i-1})$ can be circumvented. Specifically, note that from the definition of d_{ij} and S_{ij} in (3.3.3) and (3.3.4) and the expression of the log-likelihood (4.2.1) that the likelihood can be expressed in this alternative form,

$$\begin{aligned}\ell(\theta) = {} & \log \mu(\tau_1) - U(\tau_1) + \sum_{i=2}^{n} \log \left(\sum_{j=1}^{i-1} p_{ij}\, \tilde{d}_{ij} \right) + \log \left(\sum_{j=1}^{n} p_{n+1,j} \tilde{S}_{n+1,j} \right) \\ & + \Phi(T),\end{aligned} \tag{4.3.3}$$

where $\tilde{d}_{ij}$ and $\tilde{S}_{ij}$ are free from the $\Phi(\tau_i)$ or $\Phi(\tau_{i-1})$ and are as follows,

$$\tilde{d}_{ij} = (\mu(\tau_i - \tau_j) + \phi(\tau_i))e^{-\{U(\tau_i-\tau_j)-U(\tau_{i-1}-\tau_j)\}},$$
$$\tilde{S}_{ij} = e^{-\{U(\tau_i-\tau_j)-U(\tau_{i-1}-\tau_j)\}}.$$

Furthermore, the recursion to compute the most recent immigrant probabilities p_{ij} in (A.1.1) can also be expressed in terms of $\tilde{d}_{ij}$'s and $\tilde{S}_{ij}$'s as follows,

$$\begin{aligned} p_{ij} &= \begin{cases} \frac{\phi(\tau_{i-1})}{\mu(\tau_{i-1}-\tau_j)+\phi(\tau_{i-1})} \frac{p_{i-1,j}\tilde{d}_{i-1,j}}{\sum_{k=1}^{i-2} p_{i-1,k}\tilde{d}_{i-1,k}}, & j = 1, \ldots, i-2, \\ 1 - \sum_{k=1}^{i-2} p_{ik}, & j = i-1, \end{cases} \\ &= \begin{cases} \frac{p_{i-1,j}\phi(\tau_{i-1})\tilde{S}_{i-1,j}}{\sum_{k=1}^{i-2} p_{i-1,k}\tilde{d}_{i-1,k}}, & j = 1, \ldots, i-2, \\ 1 - \sum_{k=1}^{i-2} p_{ik}, & j = i-1. \end{cases} \end{aligned} \tag{4.3.4}$$

Note also that, with the truncation on the most recent immigrant distribution applied, the p_{ij}'s at the start of the ith iteration, before the truncation and renormalization at the current iteration happens, takes the form,

$$p_{ij} = \begin{cases} \frac{p_{i-1,j}\phi(\tau_{i-1})\tilde{S}_{i-1,j}}{\sum_{k=i-1-B_{i-1}}^{i-2} p_{i-1,k}\tilde{d}_{i-1,k}}, & j = i-1-B_{i-1}, \ldots, i-2 \\ 1 - \sum_{k=i-1-B_{i-1}}^{i-2} p_{ik}, & j = i-1. \end{cases}$$

The computation of the $\Phi(T)$ in (4.3.3) still takes linear time. In considering the approximate likelihood inference for the classical Hawkes process, some authors (Lewis and Mohler, 2011; Veen and Schoenberg, 2008) have suggested approximating $\Phi(T)$ by $nG(\infty)$, which is a close approximation when the parameters of the excitation kernel are such that the kernel decays rapidly. However, this approximation might be too crude for some parameter values and cause efficiency loss on the parameters of the excitation kernel. Therefore, calculation of its exact value without an approximation will be used hereafter, which will not cause concern here, since the computation needs to happen only once, and the goal, after all, is to find a linear-time algorithm.

The computational time needed to evaluate $\phi(\tau_i)$ can be reduced by assuming that the excitation effects due to events in the distant past are negligible. This is justifiable since the integrability condition on the kernel g implies $g(\tau_i - \tau_j) \approx 0$ when τ_i and τ_j are far apart. Specifically, the approximation,

$$\phi(\tau_i) \approx \sum_{j=i-F_i}^{i-1} g(\tau_i - \tau_j),$$

is used for large values of F_i. Again, two choices of F_i are discussed. The first

one assumes $F_i = \min(F, i-1)$ for a fixed (large) positive integer F. With the second approach, F_i is selected to be the smallest integer $j \leqslant i-1$ such that $G(\tau_i - \tau_{i-j}) \geqslant (1-\delta)G(\infty)$, for a small $\delta > 0$. That is,

$$\begin{aligned} F_i &= \min\left\{j \leqslant i-1; G(\tau_i - \tau_{i-j}) \geqslant (1-\delta)G(\infty)\right\} \\ &= \min\left\{j \leqslant i-1; \tau_i - \tau_{i-j} \geqslant \tilde{G}^{-1}(1-\delta)\right\}, \end{aligned}$$

where $\tilde{G}^{-1}(1-\delta)$ denotes the $(1-\delta)$-quantile of the normalised excitation kernel $\tilde{g}(\cdot) = g(\cdot)/G(\infty)$, and $F_i := i-1$, when the the set is empty.

Remark 4.3.1. *The dynamic approximation to $\phi(\tau_i)$ discussed above is similar to, but different than, that used by Halpin (2013), which approximates $\phi(\tau_i)$ by discarding the terms that are smaller than or equal to a small value δ from the summation:*

$$\phi(\tau_i) \approx \sum_{j \in W_i} g(\tau_i - \tau_j), \text{ with } W_i = \left\{j \leqslant i-1;\, g(\tau_i - \tau_j) > \delta\right\}.$$

Compared to Halpin's approach, the approach considered here seems easier to implement, as determining F_i from $\tilde{G}^{-1}(1-\delta)$ is simple using a binary search, while Halpin's approach requires the determination of the set W_i, which is not straightforward when the excitation kernel is not monotonically decreasing.

Remark 4.3.2. *When $B_i \equiv B$ and $F_i \equiv F$, it is clear that the likelihood approximation takes linear time to compute. On the other hand, when the B_i's and F_i's are dynamically determined using the aforementioned approach, it is not instantly clear that the likelihood approximation algorithm is still a linear time one. However, there is strong numerical evidence that the sequences $B_1, \ldots, B_n$ and $F_1, \ldots, F_n$ are both asymptotically stationary, suggesting that the algorithm is a linear time one in a stochastic sense.*

Remark 4.3.3. *When the excitation kernel is an exponential function, or a finite linear combination of exponential functions, then the approximation to the $\phi(\tau_i)$'s are not necessary, since they can be evaluated exactly in linear time. Specifically, if $g(t) = ae^{-bt}$ for some a, b, then the $\phi(\tau_i)$'s can be calculated using this recursion, starting with $\phi(\tau_i) = 0$:*

$$\phi(\tau_i) = \left\{\phi(\tau_{i-1}) + a\right\} e^{-b(\tau_i - \tau_{i-1})}, \quad i = 2, \ldots, n.$$

Now if $g(t) = \sum_{k=1}^{K} a_k e^{-b_k t}$ for some K and a_k, b_k, $k = 1, \ldots, K$, then from $\phi(\tau_i) = \sum_{k=1}^{K} \phi_k(\tau_i)$ with $\phi_k(\tau_i) := \sum_{j=1}^{i-1} a_k e^{-b_k(\tau_i - \tau_j)}$, and the fact that each of the K sequences $\left\{\phi_k(\tau_i),\, i = 1, \ldots, n\right\}$, $k = 1, \ldots, K$, can be evaluated in linear time, it can be seen that $\phi(\tau_i)$, $i = 1, \ldots, n$ can also be evaluated in linear time.

4.4 Simulations

4.4.1 Simulation models

This section performs simulations to compare the performance of the Nelder-Mead simplex, the Newton-Raphson, and approximate likelihood methods. The simulations consist of $N = 1000$ realizations from the RHawkes process with Weibull distributed inter-renewal waiting times with hazard function given in (3.6.2) with shape parameter $\kappa = 3$ or $1/3$ and scale $\beta = 1.2,$ or 0.2 such that the mean waiting time between successive immigrant events is close to one (1.07 and 1.2 respectively). The excitation function takes an exponential form, given in (3.6.3) with mean waiting time between offspring events $\gamma = 1$ and branching ratio $\eta = 0.5$. The chosen branching ratio value implies that approximately half of the events are immigrants and the other half are offspring events. The censoring times T are 550 and 700 so that the mean number of events is close to 1000 in both cases.

4.4.2 Simulation results

In this section, the performance of the full likelihood and approximate likelihood estimation are investigated by comparing the following four methods:

- A full likelihood optimization using the Nelder-Mead simplex method (NM);
- A full likelihood optimization using the Newton-Raphson algorithm (NR);
- An approximate likelihood optimization with fixed B and F values using the Nelder-Mead simplex method (AL1), where $B = F$ are both set to 10, or the number of events divided by 20 and then rounded up, whichever is smaller;
- An approximate likelihood optimization consisting of a dynamically chosen B_i and F_i with $\epsilon = 10^{-6}$ and $\delta = 10^{-6}$, using the Nelder-Mead simplex method (AL2).

The optimization routines were considered converged when they were unable to decrease the value of the log-likelihood by $10^{-8}(|\ell| + 10^{-8})$ at an iteration, that is, `reltol` was set at `1e-8` when the `optim` function in R (R Core Team, 2016) was called. Table 4.4.1 reports the bias, empirical standard error (SE), mean standard error estimate ($\hat{\text{SE}}$) for each of the estimated parameters using the four estimation methods described above. The average running time of the R code [on Intel Xeon Gold 6130 (22M Cache, 2.1GHz) Skylake processors] to perform the optimization procedure and compute the approximate Hessian matrix (exact Hessian matrix in the case of NR) and the total number of iterations until convergence are also displayed.

		$\kappa = 3$	$\beta = 1.2$	$\gamma = 1$	$\eta = 0.5$	Iter	Time (s)
	Bias	0.0180	-0.0034	0.0147	-0.0033		
NM	SE	0.2622	0.0505	0.2132	0.0302	169.96	95.4
	$\widehat{\text{SE}}$	0.2538	0.0489	0.2021	0.0304		
	Bias	0.0242	-0.0022	0.0256	-0.0025		
NR	SE	0.2590	0.0496	0.2056	0.0296	4.26	69.2
	$\widehat{\text{SE}}$	0.2545	0.0489	0.2041	0.0303		
	Bias	0.0179	-0.0034	0.0145	-0.0033		
AL1	SE	0.2622	0.0505	0.2128	0.0302	170.3	23.1
	$\widehat{\text{SE}}$	0.2537	0.0489	0.2019	0.0304		
	Bias	0.0180	-0.0034	0.0147	-0.0033		
AL2	SE	0.2622	0.0505	0.2132	0.0302	169.62	19.3
	$\widehat{\text{SE}}$	0.2526	0.0488	0.2011	0.0303		
		$\kappa = 1/3$	$\beta = 0.2$	$\gamma = 1$	$\eta = 0.5$	Iter	Time (s)
	Bias	-0.0073	0.0155	0.0088	0.0043		
NM	SE	0.0178	0.0418	0.1171	0.0391	143.36	85.9
	$\widehat{\text{SE}}$	0.0141	0.0351	0.1115	0.0353		
	Bias	-0.0073	0.0155	0.0092	0.0042		
NR	SE	0.0178	0.0419	0.1167	0.0391	3.89	68.0
	$\widehat{\text{SE}}$	0.0141	0.0351	0.1116	0.0353		
	Bias	-0.0072	0.0170	-0.0123	0.0045		
AL1	SE	0.0177	0.0440	0.1138	0.0391	144.60	19.8
	$\widehat{\text{SE}}$	0.0142	0.0355	0.1054	0.0353		
	Bias	-0.0073	0.0155	0.0088	0.0043		
AL2	SE	0.0178	0.0418	0.1171	0.0391	143.39	19.8
	$\widehat{\text{SE}}$	0.0141	0.0351	0.1114	0.0353		

Table 4.4.1: Estimation results for $N = 1000$ realizations from two simulation models using the Nelder-Mead (NM) based on the exact likelihood, Newton-Raphson (NR) and Nelder-Mead based on the two likelihood approximation likelihoods (AL1, AL2) methods. The NR method needs about four iterations to converge, while the Nelder-Mead method with exact or approximate log-likelihoods needs substantially more (about 30 times as many) iterations to converge. The estimates using different methods are mostly identical. In terms of speed, AL1 and AL2 are roughly five times as fast as the NM, while the Newton-Raphson is about 30% faster than the NM.

For all the methods NM, NR, AL1, and AL2, the estimation methods give comparable results. The biases for each parameter using the two full likelihood approaches with the NM and NR methods are very close to zero, suggesting that these estimates are approximately unbiased and this is expected due to the large sample size. The average standard errors and empirical standard error estimates are close for all methods and in particular, the results from NM and AL1 and AL2 are nearly identical, suggesting the log-likelihood approximations are highly accurate. An important observation that is beneficial and of practical importance is the similarly small bias for the approximate likelihood methods AL1 and AL2.

In terms of computation time required, the method NR is about 30% faster than NM, despite the number of iterations required to achieve convergence being substantially (over 30 times) smaller than that of the NM (with exact or approximate log-likelihood). The approximate likelihood methods AL1 and AL2 are the fastest,

both about five times as fast as the NM method, and about three times as fast as the NR method. On larger datasets, the speed gains by using the two approximate likelihood methods should be more impressive due to their linear time complexity in contrast to the quadratic time complexity of the NM and the NR methods.

4.4.3 Accuracy and speed of the log-likelihood approximation methods

This section examines the accuracy and speed of the two log-likelihood approximation methods. To this end, 100 realizations of the sample path of the second simulation model in Section 4.4.1, where $\kappa = 1/3$, $\beta = 0.2$, $\gamma = 1$, and $\eta = 0.5$, up to the censoring time of 8000 are generated. The exact and the two approximate log-likelihoods of four parameter vectors relative to each of the 100 sample paths up to four censoring times are evaluated, to see the influence of the parameter and the amount of data on the accuracy and speed of the approximation. The four parameter vectors are $\theta_1 = (0.1, 0.1, 0.1, 0.1)$, $\theta_2 = (0.3, 0.2, 0.5, 0.5)$, $\theta_3 = (1, 1, 1, 0.8)$, and $\theta_4 = (3, 2, 10, 0.9)$, which are chosen from different regions of the parameter space. The four censoring times are $T_1 = 1000$, $T_2 = 2000$, $T_3 = 4000$, and $T_4 = 8000$. For the tuning parameters, $B = F = 50$ in AL1, and in AL2 the B_i's and F_i's were dynamically determined using the threshold probabilities $1-\epsilon = 1-\delta = 1-0.001$. The average running time in seconds and the median absolute relative error (MARE) of the log-likelihood approximation are shown in Table 4.4.2 together with the average number of events, and the mean of the average value of B_i and F_i in AL2. The average running time and MARE of the two log-likelihood approximation methods against the average number of events are graphed in Figure 4.4.1.

From Table 4.4.2 and Figure 4.4.1 it can be observed that the running times of the two log-likelihood approximation methods increase approximately linearly with the number of events, which is to be expected, since in AL1 the B and F values are preset, and the average B_i and F_i values are also remarkably stable over time, as shown in the table. From the table, note that the time required to evaluate the exact log-likelihood increases roughly quadratically with the amount of data. The speed of the method AL1 does not seem to depend the parameter vector, while the speed of AL2 does seem to depend on the parameter vector, because in method AL2 the B_i and F_i values need to adapt to the parameter vectors under consideration to make sure the preset requirements on the truncations to the most recent immigrant distribution and the excitation kernel are met. For example, for parameter vector θ_4, where the mean of the offspring birth time distribution γ is 10, the dynamically selected F_i values are around 130 on average to make sure the truncation to the excitation kernel accounts for 99.9% ($= 1 - 0.001$) of the excitation effect.

			T	1000	2000	4000	8000
			$\mathbb{E}[N(T)]$	1676.1	3309.0	6478.0	11385.1
θ_1 :	$\kappa = 0.1$ $\beta = 0.1$ $\gamma = 0.1$ $\eta = 0.1$	Time (s.)	Exact	1.01	3.57	13.61	41.52
			AL1	0.15	0.31	0.63	0.95
			AL2	0.26	0.31	0.76	1.51
		MARE	AL1	0.00e+00	1.24e-15	1.95e-14	4.78e-14
			AL2	1.56e-04	1.58e-04	1.53e-04	1.48e-04
		$E[B]$ in AL2		5.12	5.10	5.11	5.08
		$E[F]$ in AL2		4.99	4.95	4.96	4.93
θ_2 :	$\kappa = 0.3$ $\beta = 0.2$ $\gamma = 0.5$ $\eta = 0.5$	Time (s.)	Exact	1.05	3.74	14.26	42.78
			AL1	0.15	0.31	0.62	0.95
			AL2	0.15	0.34	0.82	1.61
		MARE	AL1	1.99e-06	4.67e-06	7.52e-06	9.24e-06
			AL2	3.85e-04	3.95e-04	3.93e-04	4.04e-04
		$E[B]$ in AL2		11.50	11.46	11.55	11.49
		$E[F]$ in AL2		14.74	14.61	14.66	14.55
θ_3 :	$\kappa = 1$ $\beta = 1$ $\gamma = 1$ $\eta = 0.8$	Time (s.)	Exact	0.99	3.61	13.64	39.3
			AL1	0.15	0.31	0.61	0.93
			AL2	0.16	0.37	0.87	1.69
		MARE	AL1	2.00e-03	3.78e-03	4.70e-03	2.97e-03
			AL2	3.99e-03	5.17e-03	6.33e-03	4.52e-03
		$E[B]$ in AL2		18.39	18.23	18.30	18.18
		$E[F]$ in AL2		23.20	22.96	23.10	22.95
θ_4 :	$\kappa = 3$ $\beta = 2$ $\gamma = 10$ $\eta = 0.9$	Time (s.)	Exact	0.99	3.67	13.87	38.71
			AL1	0.15	0.32	0.63	0.95
			AL2	0.16	0.37	0.89	1.72
		MARE	AL1	1.91e-03	1.48e-03	1.33e-03	4.46e-04
			AL2	2.26e-05	1.60e-05	1.37e-05	4.85e-06
		$E[B]$ in AL2		14.33	14.21	14.26	14.17
		$E[F]$ in AL2		127.22	128.74	131.00	130.40

Table 4.4.2: The average running time (Time) and the median absolute relative error (MARE) of the two log-likelihood approximation methods AL1 and AL2 for four different parameter vectors at four increasingly large censoring times. The average time required to calculate the exact log-likelihood is also reported. In AL1, $B = F = 50$; In AL2, the B_i's and the F_i's are dynamically determined from cut-off probabilities $1 - \epsilon$ and $1 - \delta$ respectively, with $\epsilon = \delta = 0.001$.

The MARE of the two approximation methods does not necessarily grow with the amount of data. Both approximation methods are reasonably accurate in that the median absolute errors in all cases are well below 1% of the exact log-likelihood value. The accuracy of both approximation methods seems to be affected by the parameter vector under consideration, although it seems more so for AL1 than for AL2, as in AL1 the fixed B and F values can not adapt to the parameter vector under consideration and therefore might be too small for specific parameter values to produce accurate approximations.

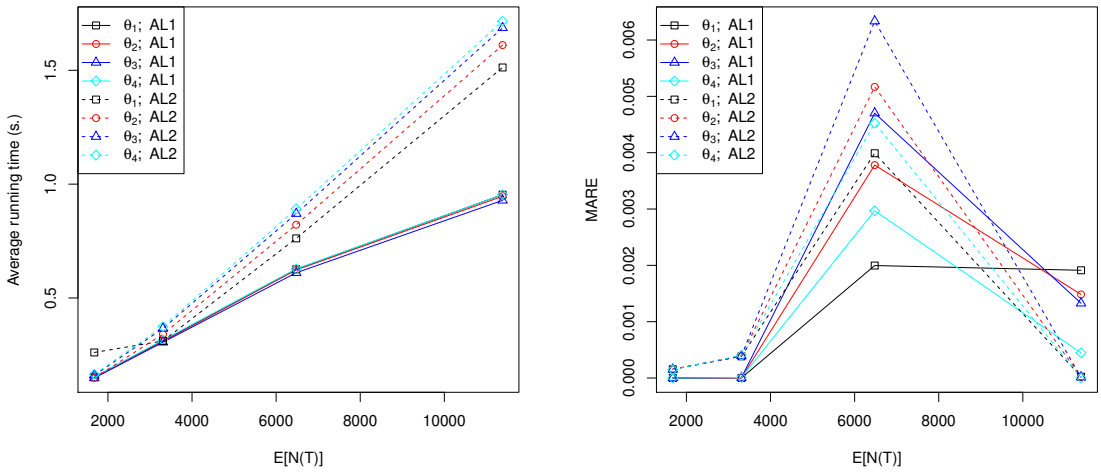

Figure 4.4.1: The average running times and median absolute relative errors (MAREs) of the two log-likelihood methods (AL1 and AL2) on different parameters at different censoring times. Left: running times; right: MAREs. Solid lines for AL1; dashed lines for AL2. Different points indicate different parameter vectors.

4.4.4 Influence of the tuning parameters on the speed and accuracy of log-likelihood approximation

The speed and accuracy of the likelihood approximation parameters depend on how the tuning parameters are determined. With large B_i and F_i values, the likelihood approximations will be more accurate, but the running times will be longer, and similarly, smaller B_i and F_i values mean faster but less accurate approximations. This section explores to what extent the choice of the tuning parameters influences the speed and accuracy of the log-likelihood. To this end, the log-likelihood approximation methods AL1 and AL2 are applied with varying tuning parameters to find the log-likelihoods of the parameter vectors θ_1, θ_2, θ_3 and θ_4 relative to the 100 simulated sample paths up to censoring time $T = 8000$ of the RHawkes model. The tuning parameter values used in AL1 are $B = F = 10, 20, \ldots, 160$; and the tuning parameter values used in AL2 are $\epsilon = \delta = 10^{-1}, 10^{-2}, \ldots, 10^{-16}$. The average running times of the R code on the 100 datasets, the MARE of the log-likelihood approximations, for each of the four-parameter vectors in Table 4.4.2 with different values of the tuning parameters are shown in Figure 4.4.2.

The average of the mean values of the $\mathbb{E}[B]$ and $\mathbb{E}[F]$ in AL2 on the 100 datasets are also reported. From Figure 4.4.2, note that the average running time of AL1 increases roughly linearly with the value of the tuning parameter B and F. The running times of AL2 also tend to increase as ϵ and δ shrinks, or equivalently the approximations to the most recent immigrant distribution and the excitation kernel function become more and more accurate. It is interesting to note that although AL2

seems slower than AL1 for the tuning parameters considered, the running time of AL2 increases much slower than AL1 and plateaus when ϵ and δ are smaller enough, despite the average number of back-looking lag $\mathbb{E}[B]$ for the most recent immigrant and the average of forward-looking lag $\mathbb{E}[F]$ for the range of the excitation effect

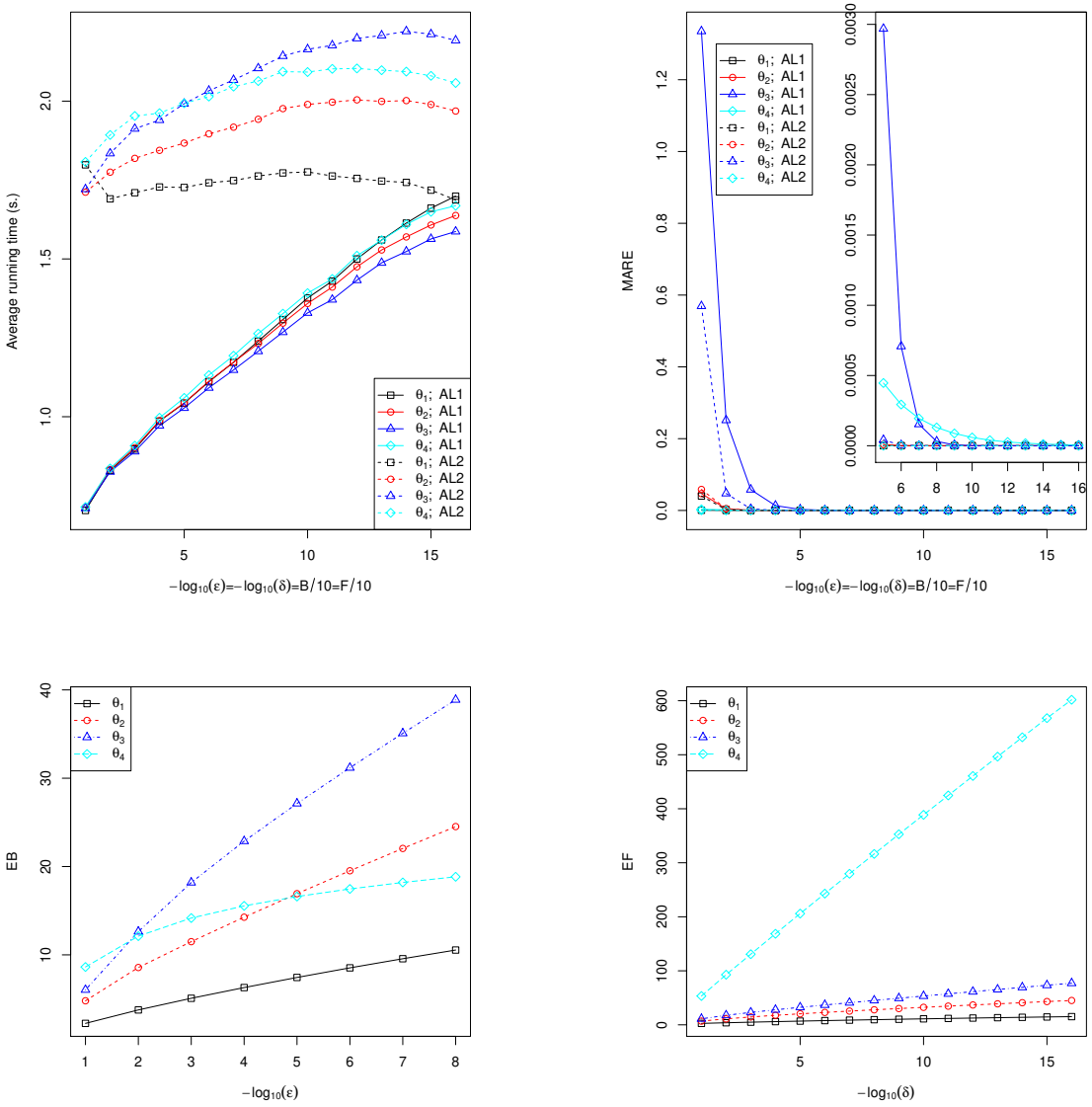

Figure 4.4.2: The average running time (top left), and the median absolute relative error (MARE, top right) of AL1 and AL2 at four different parameter vectors, and the average of the means of B_i in AL2 ($\mathbb{E}[B]$, bottom left), and the average of the means of the F_i in AL2 ($\mathbb{E}[F]$, bottom right) against the (transformed) tuning parameter. In the top right panel, the smaller inset graph is a zoom-in of the right part of the larger graph.

both increasing linearly (cf. lower panels of Figure 4.4.2) as their deterministic counterparts in AL1. This might be because in AL2 the computational overhead to determine the values of B_i and F_i dominate the computational cost in each

iteration when the values of B_i and F_i are small relative to the total sample size. It should also be mentioned that AL2 seems to be more accurate than AL1, with the relative approximation error of AL2 on all the four parameters considered being practically zero when ϵ and δ are 10^{-6} or smaller. This is also the reason why the simulation experiments to compute the MLE of the parameters (cf. Table 4.4.1), the average number of iterations of the Nelder-Mead optimization routine applied with the exact log-likelihood and is nearly identical to that with the AL2 approximate log-likelihood. The implication for fitting RHawkes processes (therefore including classical Hawkes processes) in practice is that one can simply apply any derivative-free optimization routines on the approximate log-likelihood calculated using AL2 with very small values of ϵ and δ, e.g. 10^{-6} or smaller, to achieve significant gains in computational speed without having to worry about loss in statistical efficiency. Another practically useful strategy of fitting the RHawkes process on big datasets is to obtain an initial estimate of the parameters quickly using AL1 with smaller values of B and F, and then use the initial estimate as the starting value in a sub-sequential optimization using the Newton-Raphson method, or using a derivative-free method with a more accurate log-likelihood approximation.

4.5 Application

4.5.1 Mid-price changes of foreign currency exchange rates

This analysis intends to quantify the level of endogeneity (self-exciting effects) in the foreign exchange market (forex) using RHawkes processes. For this purpose, the mid-price changes of the following currency pairs : AUD/USD (Australian Dollar against US Dollar), USD/CAD (USD against Canadian Dollar), USD/CHF(USD against Swiss Franc), EUR/USD (euro against USD), GBP/USD (British Pound against USD), USD/HKD (USD against Hong Kong Dollar), USD/JPY (USD against Japanese Yen) and USD/SEK (USD against Swedish Krona) are studied. These currency pairs are analyzed as they represent some of the most traded currency pairs in terms of value traded. This analysis examines the mid-price changes for these pairs, for the four trading weeks (excluding weekends) from 1st July 2019 until 26th July 2019 and only the hours between 12:00:00 GMT until 21:00:00 GMT, the official operating hours for the New York forex market are considered. The different currency pairs are compared by analyzing the level of endogenous and exogenous trading activities and their impact on the mid-price changes.

The mid-price is defined as the mean of the best bid and ask prices, and a mid-price change happens when the value of this mid-price changes. This change occurs when either the bid, ask, or a combination thereof, changes the price. It is well observed that trading activity does not happen in a stationary manner during

the trading day. Extremely evident is the drastically different features of trades exhibited during the opening and closing times of international forex markets. To see this, in Figure 4.5.1, the expected duration between mid-price changes conditional on

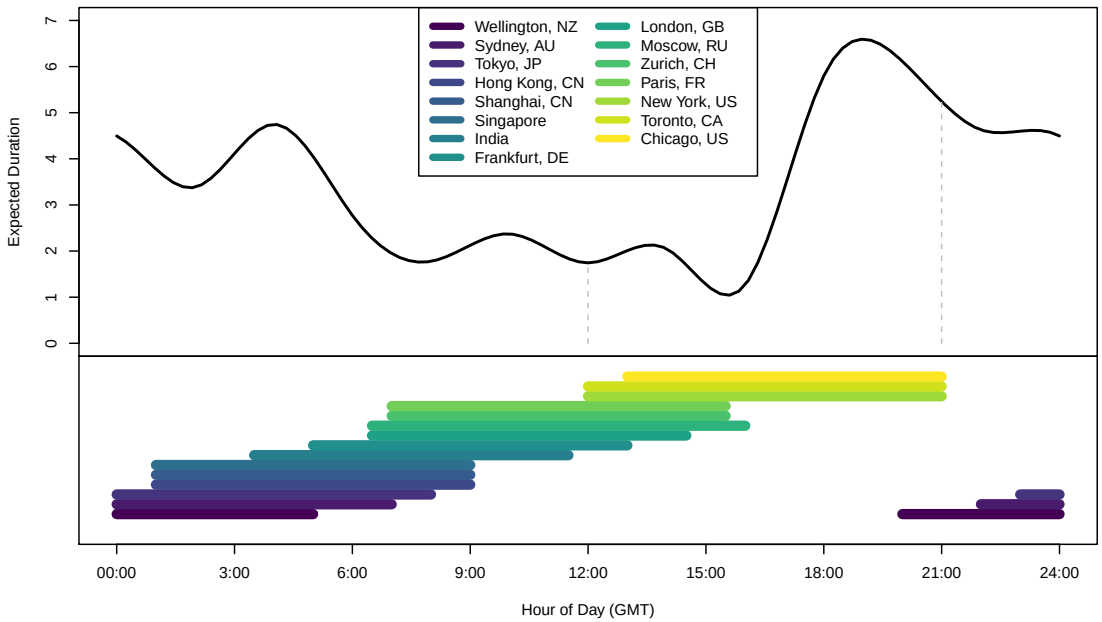

Figure 4.5.1: A non-parametric estimate of the daily pattern of the duration between mid-price changes, and in the bottom panel, the operating hours of the major forex centers around the world.

the time of day that the mid-price change transpired is displayed. The expectation is estimated using a cubic regression spline, where the knots were chosen every two hours throughout the 24 hour day. This procedure has previously been employed in the work of Engle and Russell (1998). The figure shows a clear diurnal pattern with the New York forex hours having a shorter duration during the opening hours than the closing hours. It is apparent, that when the majority of the forex markets are in operation, the expected waiting time between mid-price changes is much shorter, than when fewer markets are in operation. The non-stationary arrival time regime of the mid-price changes over any particular trading day necessitates a transformation to remove the intra-day patterns and hence obtain stationarity.

Following the work of Engle and Russell (1998), the observed durations are discounted by a factor proportional to the corresponding expected duration subjected to the requirement that the sum of the modified durations in a trading day is equivalent to the sum of the original durations. The purpose is to supply less weight to mid-price changes occurring in the opening of the US forex markets when activity is high as the majority of the international forex markets are open; cf. Figure 4.5.1. Also note that, although the timestamps in the dataset are accurate down

to microseconds (10^{-6} seconds), on several occasions two or more price changes are recorded to have transpired at identical times. On these occasions, small random noise was added to these event times to break the ties while still preserving the time order of the price moves.

For each of the currency pairs, the mid-price changes for the twenty trading days (four trading weeks) are each fitted with a renewal Hawkes processes with Weibull immigration. The motivation for the Weibull distribution follows from the analysis in Chapter 3, in which the Weibull renewal distribution was able to provide a superior fit than the gamma and exponential distributions. Also, an exponential offspring density function was utilized. The model fitting tasks were accomplished by employing the AL2 approximate likelihood method with $\epsilon = \delta = 10^{-20}$. The pa-

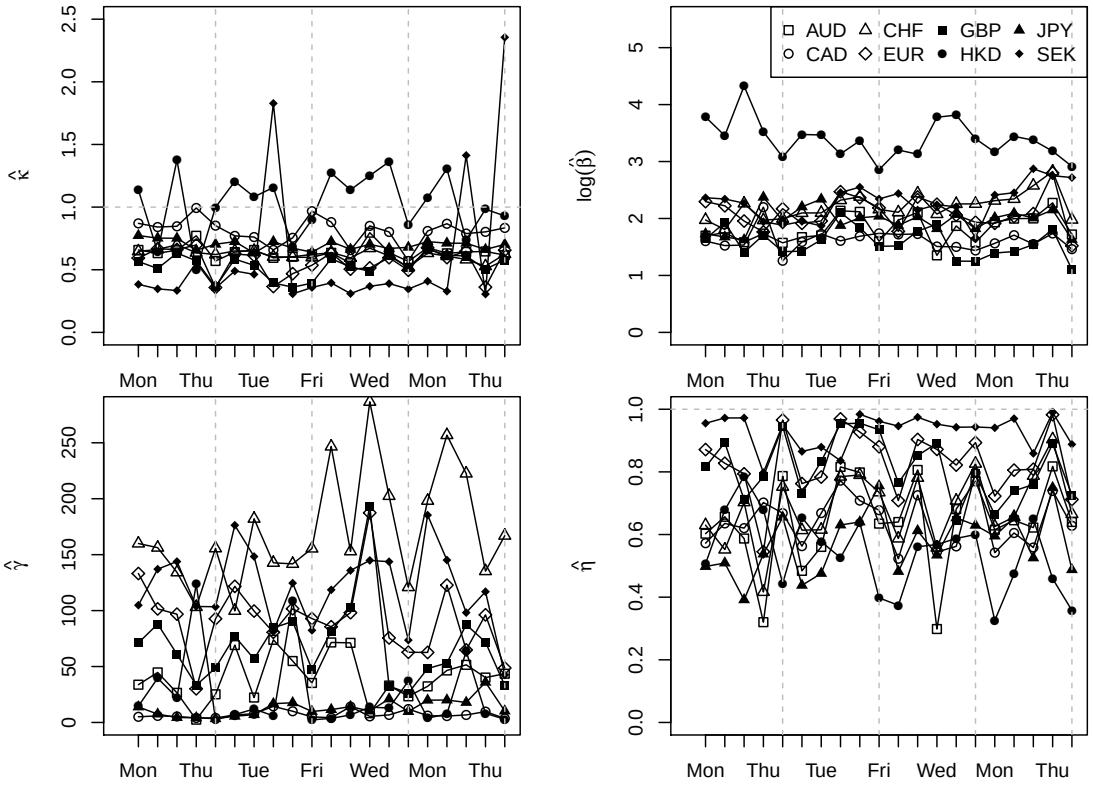

Figure 4.5.2: Time series plot of the estimated parameters of the RHawkes process as they evolve over the four trading weeks from 1st July 2019 until 26th July 2019 between the hours of 12:00:00 GMT until 21:00:00 GMT when the New York forex market is open, for all seven exchanges rates with the USD.

rameter estimates for each of the currency pairs are plotted in Figure 4.5.2, whereby the time series of estimates evolve over the four trading weeks. Note that, for the scale parameter of the Weibull distribution β, the results are presented on the log-scale for better visualization. Furthermore, notice that this analysis only deals with a fixed window of time and does not take into account the mid-price changes before

the start of the observation period. As such, edge effects may influence the parameter estimates, and the results presented in Figure 4.5.2, but the large sample size should guarantee that these effects are inconsequential.

First, the exogenous features of the mid-price arrival regime are analyzed before investigating the endogenous components of the process after that. From Figure 4.5.2, observed that the arrival process of exogenously driven mid-price changes, has a Weibull shape parameter that persists below one for the majority of the currency pairs. This implies that exogenously driven mid-price changes occur in heavier bursts than would be suggested by a Poisson process. For instance, consider the pair GBP/USD, the estimated shape parameter $\hat{\kappa}$ has the following five-number summary for the four trading weeks: $(0.36, 0.50, 0.55, 0.59, 0.66)$ with mean 0.53 and standard deviation 0.09. This implies a heavy-tailed distribution for the arrival process of exogenous mid-price changes, and therefore price changes due to due to external influences tend to occur in a more bursty fashion rather than uniformly through time. However, for the pair USD/HKD, this is not evident, and the arrival process of exogenous mid-price changes seem to exhibit complete randomness and does not deviate far from a Poisson process, as the estimated shape parameter $\hat{\kappa}$ is not very different from one.

It is evident that the scale parameter for the currency pair USD/HKD have a significantly larger estimated value and deviate significantly from the other currency pairs. This is because the total number of mid-price changes for this pair is extremely small compared to the other currency pairs. For instance, the mean number of mid-price changes for a particular day is $2,399$, and this is three times smaller than the next smallest, which is the currency pair USD/CHF with $8,299$ mid-price changes. For each trading week, there tends to be a trough that happens on Fridays for the estimated parameters of $\hat{\beta}$ by looking at Figure 4.5.1. Again, this can be attributed to the high level of trading activity that tends to transpire on Fridays in comparison to the rest of the trading week, and this phenomenon is present for the majority of the currency pairs, and hence, the mean waiting time between exogenous mid-price changes are much smaller on these days.

Market participants generally take a longer time to react to mid-price movements than to external factors or exogenous mid-price changes when executing currency trades. For instance, in Figure 4.5.3, the expected waiting time between exogenously driven mid-price changes is displayed in solid lines and endogenously driven mid-price changes in dashed lines. It can be observed that the waiting time between endogenous mid-price changes take much longer to transpire than exogenous ones, implying that market participants take a prolonged time to react to previous price movements than to external market news or events. Note that the mean waiting time between exogenous events is computed using $\hat{\beta}\Gamma(1 + 1/\hat{\kappa})$, the mean of a

Weibull distribution, and endogenous events using $\hat{\gamma}$, the mean of an exponential distribution. These two waiting times tend to move in the same direction, although on some occasions, the two waiting times move in opposite directions as can been seen in Figure 4.5.3.

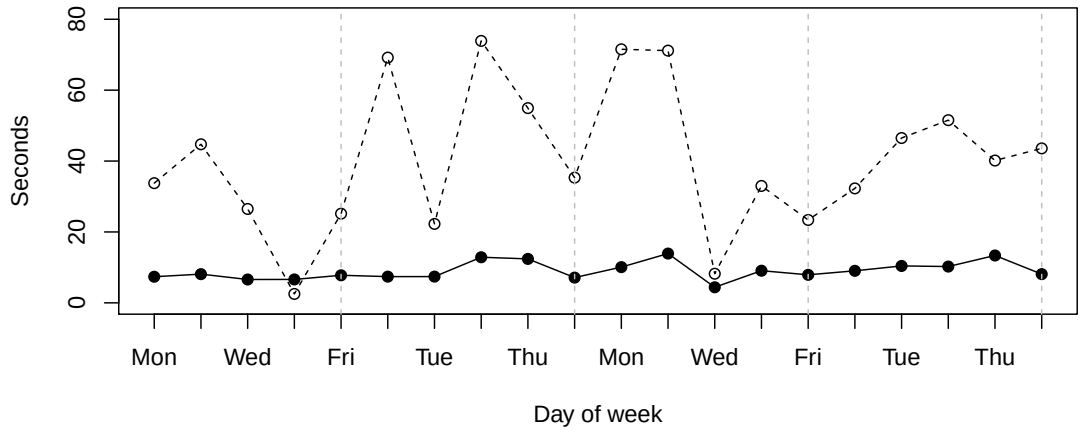

Figure 4.5.3: Mean waiting time between exogenously driven mid-price changes in solid lines, and the mean waiting time between endogenously driven mid-price changes in dashed lines, for the currency pair AUD/USD.

This analysis concludes by analyzing the level of endogeneity in the forex markets. As seen in Figure 4.5.2, the estimated branching ratio $\hat{\eta}$ shows varying levels of self-excitation with estimated branching ratios ranging between 0.54 for the pair USD/HKD and 0.93 for the pair USD/SEK. This suggests that a significant portion of the price movements are endogenously generated rather than related to the arrival of news in the market place. The overall weaker endogeneity in the price movements of the HKD relative to the USD compared to the other currency pairs could be because the price of the HKD is pegged to the USD and therefore traders of this currency pair might be less sensitive to its price movements. Another critical aspect to acknowledge is the waiting time between endogenous mid-price changes. It can be observed that the USD/CHF exchange rate exhibit highly dispersed values of $\hat{\gamma}$ compared to the other currency pairs. For instance, the average of the mean waiting times for the pair USD/CAD is 7.14 seconds, and for the pair USD/CHF, it is significantly longer with 171.03 seconds.

The two-step procedure of transforming the observed event times and then fitting a model to these transformed times could potentially be handled within a single framework. Instead of transforming the observed event times to account for the diurnal patterns exhibited during the trading day, Zhuang and Mateu (2019) have implemented a semi-parametric approach to allow the background intensity to de-

pend on three separate temporal components relating to a long term trend, daily periodicity, and weekly periodicity as well as a separable spatial component. They generalize the non-parametric stochastic reconstruction method to estimate each component in the background intensity and apply this model to describe the occurrences of violence or robbery in Castellón, Spain. This approach might be more natural to account for the daily variation in trading activity that is present during the operating hours, and the different levels of activity that occur throughout a trading week.

Chapter 5

The multivariate renewal Hawkes process[1]

5.1 Introduction

In many areas, researchers encounter multi-type event sequence data. For example, in earthquake modeling, the data may contain earthquakes from several neighboring regions. In finance, the tick history data on stocks records both the times of trades and the times of quotes; and order book data records the arrival times and other features of limit and market orders such as the side of the trade. In these applications, it is of interest to study not only the interactions within events of the same type but also the interactions between events of different types. Therefore, multivariate point processes, where different components are allowed to interact with each other, are needed. A multivariate point process model for this purpose is the multivariate Hawkes process (Hawkes, 1971). Bowsher (2007) modeled the timing of trades and mid-quote price changes for an NYSE stock using a generalized bivariate Hawkes process that allows the baseline event rate to vary with time. Embrechts et al. (2011) fit the bivariate Hawkes process to daily data on the negative and positive exceedances of certain threshold levels for the Dow Jones Industrial Average index. Bacry et al. (2013) showed that the multivariate Hawkes process could demonstrate the Epps effect and lead-lag effect observed in financial data. When interpreted as branching Poisson processes, both the multivariate Hawkes process and the generalization considered by Bowsher (2007) assume the arrival processes of immigrants to be Poisson and therefore do not allow over- or under-dispersion of the numbers of immigrants, or serial correlation of the numbers of immigrants in non-overlapping time intervals, even events of the same type. Such assumptions restrict the modeling capabilities of the multivariate Hawkes process unnecessarily.

[1]Most of the content shown in this chapter has been published in the *Computational Statistic and Data Analysis*; see Stindl and Chen (2018).

This chapter introduces a point process model which extends on the RHawkes process by allowing the events of the process to be of different types and, in addition to the self-excitation effect among events of the same type, allowing events of each type to affect the future occurrence rates of events of other types through the mutual excitation mechanism adopted in the multivariate Hawkes processes. The model additionally extends the multivariate Hawkes process in that the immigrant events of different types can arrive according to general renewal processes, rather than Poisson processes in the classical multivariate Hawkes process. This implies that the numbers of immigrant events of the same type in non-overlapping time intervals are allowed to have serial correlation and to be over- or under-dispersed relative to the Poisson distribution. The extension is naturally called the multivariate renewal Hawkes process or termed MRHawkes process for short.

Similar to the RHawkes process, the MRHawkes process can be efficiently simulated by utilizing the branching process interpretation. Moreover, similar to the RHawkes process, the MRHawkes process does not have an easy to evaluate likelihood function. This chapter will derive an algorithm to efficiently evaluate the likelihood of the MRHawkes process model, using an approach analogous to that of Chapter 3. The feasibility of fitting the MRHawkes process model is demonstrated by applying the model to data by likelihood maximization, on simulated data, and real-life data. The time and space complexities of the algorithm for MRHawkes process likelihood evaluation are both polynomial in the number of events observed, and therefore, the algorithm can be relatively slow on large datasets. To overcome this issue, a modification to the algorithm will be proposed, which can yield a good approximation of the likelihood in quadratic time and linear storage space. This chapter will provide an approach to assess two aspects of the goodness-of-fit of the MRHawkes model, the temporal patterns of the events and the event type distribution using the Rosenblatt residuals (Rosenblatt, 1952) and the Universal residuals (Brockwell, 2007) respectively. A simulation-based approach to predict future event occurrences will also be proposed.

The remainder of the chapter is structured as follows. Section 5.2 introduces the MRHawkes process model. Section 5.3 derives an exact algorithm for evaluation of the likelihood of the MRHawkes process. A method to evaluate the goodness-of-fit is presented in Section 5.4 and methods for future events prediction in Section 5.5. Results of the simulation studies are reported in Section 5.6 as well as methods to simulate the process and the assessment of the predictive performance of the model. Applications in seismology and finance are presented in Section 5.7 with an analysis of earthquakes arising in two Pacific island countries Fiji and Vanuatu and a dataset of trade-throughs for the stock BNP Paribas, traded on the Euronext Paris stock exchange. An R package called `MRHawkes` implementing the

proposed methodologies can be downloaded from the CRAN (https://CRAN.R-project.org/package=MRHawkes).

5.2 Model and notation

Let $\{(\tau_i, z_i),\, i = 1, 2, \dots\}$ be a realization of a multivariate point process where $\tau_1 < \tau_2 < \dots$ are all distinct and interpretable as the occurrence time of the i-th event and $z_i \in \{1, \dots, M\}$ indicates the i-th event type. Let the associated M-variate counting process be $\boldsymbol{N}_t = (N_1(t), \dots, N_M(t))$, where $N_m(t)$ is the number of type-m events up to time t. Analogous to Chapter 3, let M_i denote the unobservable immigrant or offspring status indicator, where again $M_i = 0$ if the event is an immigrant otherwise it is an offspring and $M_i = 1$. The intensity process of the MRHawkes process requires knowledge of each of the unobservable indexes of the most recent immigrant events for each component m before time t denoted by $I_m(t) = \max\{i | \tau_i < t, M_i = 0, z_i = m\}$. Then collect these to form the M-dimensional vector $I(t)$ that contains the most recent immigrant index for all M components at time t. The natural filtration of the multivariate point process is denoted by $\mathcal{F} = \{\mathcal{F}_t;\, t \geqslant 0\}$, so that $\mathcal{F}_t = \sigma\{\boldsymbol{N}_s;\, s \leqslant t\}$.

The intensity process for the m-th component of the multivariate renewal Hawkes (MRHawkes) process $\lambda_m(t)$, $t \geqslant 0$ relative to the enlarged (non-natural) filtration $\tilde{\mathcal{F}}_t = \sigma\{\boldsymbol{N}_s, I(s);\, s \leqslant t\}$, $t \geqslant 0$ takes the following form,

$$\begin{aligned}
\lambda_m(t) &= \frac{\mathbb{E}\left[\mathrm{d}N_m(t) | \tilde{\mathcal{F}}_{t-}\right]}{\mathrm{d}t} \\
&= \mu_m(t - \tau_{I_m(t)}) + \sum_{j\,:\,\tau_j < t} \eta_{m,z_j} h_{m,z_j}(t - \tau_j) \\
&=: \mu_m(t - \tau_{I_m(t)}) + \phi_m(t), \qquad (5.2.1)
\end{aligned}$$

where $\mu_m(t - \tau_{I_m(t)})$ is the immigrant arrival rate that renews on the arrival of a type-m immigrant, with $\mu_m(\cdot) > 0$ being the hazard rate function of the *i.i.d.* waiting times between successive type-m immigrants, the constant $\eta_{m,z_j} \geqslant 0$ is termed the branching ratio and indicates the average number of type-m children due to an event of type z_j, the function $h_{m,z_j}(\cdot) > 0$ is termed the offspring density function and indicates the density of the birth times of type-m children given there is at least one type-z_j child. The function $\eta_{m,z_j} h_{m,z_j}(\cdot)$ is known as the excitation function and indicates the excitation effect for component m due to component z_j. It is a requirement that the functions $\mu_m(\cdot)$ integrate to infinity, and the branching ratios η_{m,z_j} are non-negative and strictly smaller than unity. Furthermore, it is assumed that the branching matrix, defined as $\boldsymbol{H} := (\eta_{j,k};\, j, k \in \{1, \cdots, M\})$, has eigenvalues that are strictly smaller than unity. This chapter deals with the estimation of

model parameters when the functions $\mu_m(\cdot)$'s and $h_{m,z_j}(\cdot)$'s are given a parametric form up to a finite-dimensional parameter.

The *i.i.d.* waiting times between immigrant arrivals form a renewal process for the m-th component, whereas, in the multivariate Hawkes process, type-m immigrants arrive according to a Poisson process. As a consequence, the multivariate Hawkes process can be viewed as a particular sub-model of the MRHawkes process when the immigrant renewal process is specified to have an exponential inter-renewal distribution, and then the hazard functions $\mu_m(\cdot)$ are merely constants. Again, similar to before, a commonly used form of the inter-renewal hazard function is,

$$\mu_m(t) = \frac{\kappa_m}{\beta_m}\left(\frac{t}{\beta_m}\right)^{\kappa_m - 1}, \ t \geqslant 0,$$

in which case the corresponding distributions are Weibull, and the κ_m and β_m are referred to as the shape and scale parameters respectively. When the shape parameters are unity, the hazard functions are constant, and the immigrant arrival processes are Poisson. A hypothesis test on the shape parameters can be conducted to assess whether any of these parameters are statistically different from unity, in which case the MRHawkes process deviates from the multivariate Hawkes process.

Figure 5.2.1 displays a simulated realization of a bivariate RHawkes process, with Weibull inter-immigration waiting time distributions for both component processes. The shape parameters for the components are $\kappa_1 = 3$ and $\kappa_2 = 1/3$, and the scale parameters are $\beta_1 = 50$ and $\beta_2 = 20$. The offspring densities are both exponential with mean 8. The branching ratios are $\eta_{1,1} = 0.2$, $\eta_{2,1} = 0.2$, $\eta_{1,2} = 0.3$, and $\eta_{2,2} = 0.1$. The top panel of Figure 5.2.1 displays the component intensities in (5.2.1) for each individual component, $\lambda_1(t)$ and $\lambda_2(t)$, and the total intensity $\lambda(t) = \lambda_1(t) + \lambda_2(t)$. Note that due to the self-and mutual excitation effects, each of the intensity curves has a jump discontinuity whenever there is an event. Moreover, the intensity $\lambda_2(t)$ and the total intensity $\lambda(t)$ is unbounded at times. This is because the shape parameter for component two $\kappa_2 = 1/3$ is less than unity, and therefore $\mu_2(t) = \frac{1/3}{20}(\frac{t}{20})^{-2/3}$ tends to infinity when t approaches 0 from the right, which causes $\mu_2(t - \tau_{I_2(t)})$ to tend to infinity when t approaches any immigrant arrival time of component two. The bottom panel of Figure 5.2.1 displays the barcode plot of the event times for each component and also the pooled event times. The first component appears to exhibit more evenly distributed event times due to the relatively large shape parameter $\kappa_1 = 3 > 1$ despite the clustering of events due to the excitation effects, while the second component exhibits much stronger clustering among the event times, due to a small shape parameter $\kappa_1 = 1/3 < 1$ causing the spiking of the event intensity after each immigration event of component 2. The multivariate Hawkes (1971) process can not easily accommodate such widely varying clustering features.

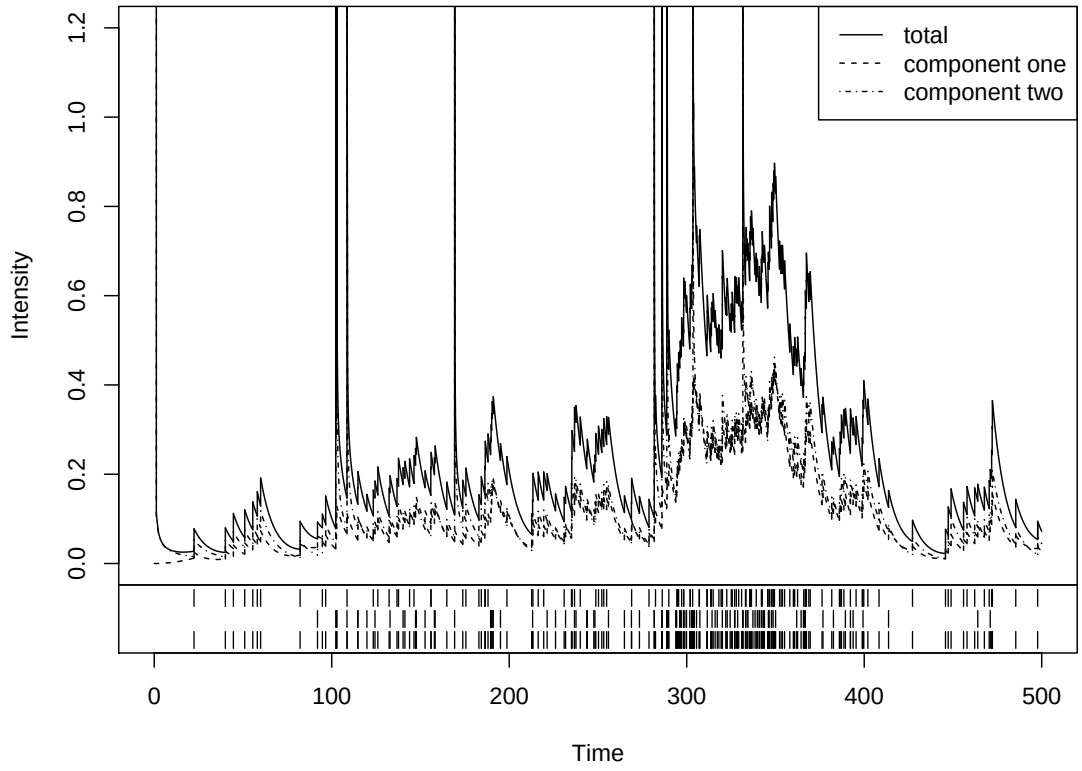

Figure 5.2.1: A simulated realization of a bivariate renewal Hawkes process. The figure displays the intensity function $\lambda(t)$ and the associated component intensities $\lambda_1(t)$ and $\lambda_2(t)$. The figure also presents a barcode plot of the event times of the component processes as well as the overall process, where the bars in the first row indicate events of the first component process, the second row the second component process, and the last row indicates the pooled event times.

5.3 Maximum likelihood estimation

This section develops a recursive algorithm to compute the likelihood of the MRHawkes process model based on the observed data over the interval $(0, T]$, which consists of the event times $\tau_{1:n}$, event types $z_{1:n}$ and $N(T) = n$. It is natural to evaluate the likelihood by conditioning upon the history of the process up until each event time, that is, condition on the previous event times and types. Thus, the likelihood can be decomposed as a product of conditional densities by employing the chain rule as follows,

$$L(\tau_{1:n}, z_{1:n}|\theta) = p(\tau_1, z_1) \left\{ \prod_{i=2}^{n} p(\tau_i, z_i|\tau_{1:i-1}, z_{1:i-1}) \right\} \mathbb{P}(\tau_{n+1} > T|\tau_{1:n}, z_{1:n}), \quad (5.3.1)$$

where for notation convenience $\tau_{1:i}$ is short for $(\tau_1, \ldots, \tau_i)$ as before and $z_{1:i}$ denotes $(z_1, \ldots, z_i)$. The form of the conditional intensity given in (5.2.1) means that the most recent immigrant index vector must be conditioned upon. This allows the

intensity and indeed the inter-event waiting time distribution between events to have an easily computable expression. Under this condition, the expression $\mu_m(t-\tau_{I_m(t)})$ can be evaluated and thus by conditioning on the index of the most recent immigrant events for each component $m = 1, \ldots, M$, denoted by $\boldsymbol{j} = (j_1, \ldots, j_M)$ taking values in $\{0, 1, \ldots, , i-1\}^M$, the following holds,

$$p(\tau_i, z_i|\tau_{1:i-1}, z_{1:i-1}) = \sum_{\forall \boldsymbol{j}} d_i(\boldsymbol{j}) \times p_i(\boldsymbol{j}), \qquad i = 1, 2, \ldots, n, \tag{5.3.2}$$

$$\mathbb{P}\left(\tau_{n+1} > T|\tau_{1:n}, z_{1:n}\right) = \sum_{\forall \boldsymbol{j}} S_{n+1}(\boldsymbol{j}) \times p_{n+1}(\boldsymbol{j}), \tag{5.3.3}$$

where

$$d_i(\boldsymbol{j}) := p(\tau_i, z_i \mid \tau_{1:i-1}, z_{1:i-1}, I(\tau_i) = \boldsymbol{j}), \tag{5.3.4}$$

$$S_{n+1}(\boldsymbol{j}) := \mathbb{P}\left(\tau_{n+1} > T|\tau_{1:n}, z_{1:n}, I(\tau_{n+1}) = \boldsymbol{j}\right), \tag{5.3.5}$$

$$p_i(\boldsymbol{j}) := \mathbb{P}(I(\tau_i) = \boldsymbol{j}|\tau_{1:i-1}, z_{1:i-1}). \tag{5.3.6}$$

The most recent immigrant index vector $\boldsymbol{j}$ has additional requirements to be adequately defined. The elements of $\boldsymbol{j}$ must be unique unless they take the value zero, and this exception only occurs when no immigrant from a particular component has arrived by time τ_i. This constraint implies that the most recent immigrant for different components do not coincide.

In the sequel, further notations that will follow closely to those introduced in Chapter 3 are adjusted to accommodate the multivariate context. Lets denote the cumulative immigrant hazard function for type-m events as $U_m(t) = \int_0^t \mu_m(s)\mathrm{d}s$, the offspring distribution function for individuals of type-m given the parent is a type-n event as $H_{m,n}(t) = \int_0^t h_{m,n}(s)\mathrm{d}s$ and the cumulative offspring effects for type-m events as $\Phi_m(t) = \int_0^t \phi_m(s)\mathrm{d}s = \sum_{j\,:\,\tau_j<t} \eta_{m,z_j} H_{m,z_j}(t-\tau_j)$. The cumulative offspring effects for the entire process is given by $\Phi(t) = \sum_{m=1}^M \Phi_m(t)$. Then by conditioning on the previous event times, event types and most recent immigrant index vector $I(\tau_{i-1}) = \boldsymbol{j}$, the conditional hazard rate function of the inter-event waiting time $\tau_i - \tau_{i-1}$ is given by,

$$\mathrm{haz}(t) = \sum_{m=1}^{M} \Big\{\mu_m(t + \tau_{i-1} - \tau_{j_m}) + \phi_m(t + \tau_{i-1})\Big\}.$$

The model implicitly assumes that the first event is an immigrant event, and thus the joint density of the first event time τ_1 and event type z_1 is given by $p(\tau_1, z_1) = e^{-\sum_{m=1}^M U_m(\tau_1)}\mu_{z_1}(\tau_1)$. For the other event times $i > 2$, the conditional densities and survival probabilities given in (5.3.4) and (5.3.5) are directly computable and given

by the following,

$$d_i(\boldsymbol{j}) = (\mu_{z_i}(\tau_i - \tau_{j_{z_i}}) + \phi_{z_i}(\tau_i))e^{-\sum_{m=1}^{M}\{U_m(\tau_i-\tau_{j_m})-U_m(\tau_{i-1}-\tau_{j_m})\}-\{\Phi(\tau_i)-\Phi(\tau_{i-1})\}}, \tag{5.3.7}$$

$$S_{n+1}(\boldsymbol{j}) = e^{-\sum_{m=1}^{M}\{U_m(T-\tau_{j_m})-U_m(\tau_n-\tau_{j_m})\}-\{\Phi(T)-\Phi(\tau_n)\}}. \tag{5.3.8}$$

It still remains to calculate the conditional probabilities $p_i(\boldsymbol{j})$ given in (5.3.6). By conditioning on the most recent immigrant index vector $I(\tau_{i-1})$ and Bayes' theorem, the following recursion is obtained,

$$\begin{aligned}
p_i(\boldsymbol{j}) &= \sum_{\boldsymbol{j'}} \mathbb{P}\left(I(\tau_i) = \boldsymbol{j} |\ \tau_{1:i-1}, z_{1:i-1}, I(\tau_{i-1}) = \boldsymbol{j'}\right) \mathbb{P}\left(I(\tau_{i-1}) = \boldsymbol{j'} | \tau_{1:i-1}, z_{1:i-1}\right) \\
&= \sum_{\boldsymbol{j'}} \mathbb{P}\left(I(\tau_i) = \boldsymbol{j} |\ \tau_{1:i-1}, z_{1:i-1}, I(\tau_{i-1}) = \boldsymbol{j'}\right) \\
&\qquad \times \frac{p(\tau_{i-1}, z_{i-1} | \tau_{1:i-2}, z_{1:i-2}, I(\tau_{i-1}) = \boldsymbol{j'})}{p\left(\tau_{i-1}, z_{i-1} | \tau_{1:i-2}, z_{1:i-2}\right)} \mathbb{P}\left(I(\tau_{i-1}) = \boldsymbol{j'} | \tau_{1:i-2}, z_{1:i-2}\right) \\
&= \sum_{\boldsymbol{j'}} \mathbb{P}\left(I(\tau_i) = \boldsymbol{j} |\ \tau_{1:i-1}, z_{1:i-1}, I(\tau_{i-1}) = \boldsymbol{j'}\right) \times \frac{d_{i-1}(\boldsymbol{j'}) \times p_{i-1}(\boldsymbol{j'})}{p(\tau_{i-1}, z_{i-1} | \tau_{1:i-2}, z_{1:i-2})},
\end{aligned} \tag{5.3.9}$$

where the summation index $\boldsymbol{j'}$ takes values in $\{0, \ldots, i-2\}^M$. One important observation to make is that at most one component of the most recent immigrant index vector $I(\tau_i)$ can equal $i-1$ while the remaining components must remain the same as in $I(\tau_{i-1})$, according to whether $M_{i-1} = 0$ or $M_{i-1} = 1$ and the event type z_{i-1}. Now define $\boldsymbol{e}_m \in \mathbb{R}^M$ to be the unit vector with the m-th element taking the value one and other elements taking the value zero, then the following Markov type property holds,

$$I(\tau_i) = \begin{cases} I(\tau_{i-1}) & \text{if} \quad M_{i-1} = 1 \\ \delta_{z_{i-1}}(I(\tau_{i-1})) & \text{if} \quad M_{i-1} = 0, \end{cases} \tag{5.3.10}$$

where $\delta_m(\boldsymbol{v}) = \boldsymbol{v} + ((i-1) - \boldsymbol{e}_m^T\boldsymbol{v})\boldsymbol{e}_m$ defines a function that returns the same input vector $\boldsymbol{v}$ except the m-th component v_m is replaced with the value $i-1$. The function $\delta_m(\cdot)$ hence updates the most recent immigrant index vector to indicate that a type-m immigrant has arrived at time τ_{i-1}.

The conditional distribution of the indexes of the most recent immigrants for each component process in (5.3.10) are computed by employing the following property of Poisson processes which states that, for two independent non-stationary Poisson processes, the probabilities of the first event of their superposition (or sum) process belonging to each constituent process are proportional to the intensity functions of the constituent processes evaluated at the time of the first event. Then observe that,

on the arrival of an offspring event, the most recent immigrant index vector remains unchanged and hence $\boldsymbol{j} = \boldsymbol{j}'$. However, when a type-$m$ immigrant arrives, observe that $\boldsymbol{j} = \delta_m(I(t))$. Now by conditioning on the most recent immigrant index vector at time τ_{i-1} the following holds,

$$\begin{aligned}
&\mathbb{P}\left(I(\tau_i) = \boldsymbol{j} | \tau_{1:i-1}, z_{1:i-1}, I(\tau_{i-1}) = \boldsymbol{j}'\right) \\
&\qquad = \begin{cases} \mathbb{P}\left(M_{i-1} = 1 | \tau_{1:i-1}, z_{1,i-1}, I(\tau_{i-1}) = \boldsymbol{j}'\right) & \boldsymbol{j} = \boldsymbol{j}' \\ \mathbb{P}\left(M_{i-1} = 0 | \tau_{1:i-1}, z_{1:i-1}, I(\tau_{i-1} = \boldsymbol{j}')\right) & \boldsymbol{j} = \delta_{z_{i-1}}(I(\tau_{i-1})) \\ 0 & \text{else,} \end{cases}
\end{aligned}$$

where

$$\mathbb{P}\left(M_{i-1} = 1 | \tau_{1:i-1}, z_{1,i-1}, I(\tau_{i-1}) = \boldsymbol{j}'\right) = \frac{\phi_{z_{i-1}}(\tau_{i-1})}{\mu_{z_{i-1}}(\tau_{i-1} - \tau_{j'_{z_{i-1}}}) + \phi_{z_{i-1}}(\tau_{i-1})}, \quad (5.3.11)$$

and

$$\mathbb{P}\left(M_{i-1} = 0 | \tau_{1:i-1}, z_{1:i-1}, I(\tau_{i-1}) = \boldsymbol{j}'\right) = \frac{\mu_{z_{i-1}}(\tau_{i-1} - \tau_{j'_{z_{i-1}}})}{\mu_{z_{i-1}}(\tau_{i-1} - \tau_{j'_{z_{i-1}}}) + \phi_{z_{i-1}}(\tau_{i-1})}. \quad (5.3.12)$$

Then by combining (5.3.9), (5.3.11) and (5.3.12), the calculation of the most recent immigrant probabilities in (5.3.6) reduces to the following recursion,

$$p_i(\boldsymbol{j}) = \begin{cases} \dfrac{\phi_{z_{i-1}}(\tau_{i-1})}{\mu_{z_{i-1}}(\tau_{i-1} - \tau_{j_{z_{i-1}}}) + \phi_{z_{i-1}}(\tau_{i-1})} \times \dfrac{d_{i-1}(\boldsymbol{j})\, p_{i-1}(\boldsymbol{j})}{p\left(\tau_{i-1}, z_{i-1} | \tau_{1:i-1}, z_{1:i-1}\right)}, & \boldsymbol{j} = \boldsymbol{j}' \\ \displaystyle\sum_{j'_{z_{i-1}}=0}^{i-2} \dfrac{\mu_{z_{i-1}}(\tau_{i-1} - \tau_{j'_{z_{i-1}}})}{\mu_{z_{i-1}}(\tau_{i-1} - \tau_{j'_{z_{i-1}}}) + \phi_{z_{i-1}}(\tau_{i-1})} \times \dfrac{d_{i-1}(\boldsymbol{j}')\, p_{i-1}(\boldsymbol{j}')}{p\left(\tau_{i-1}, z_{i-1} | \tau_{1:i-1}, z_{1:i-1}\right)}, & \boldsymbol{j} = \delta_{z_{i-1}}(\tau_{i-1}) \end{cases} \quad (5.3.13)$$

for $i = 3, \ldots, n+1$.

The likelihood function can then be evaluated at some supplied parameter values by computing the conditional density $p(\tau_i, z_i | \tau_{1:i-1}, z_{1:i-1})$ and the most recent immigrant probabilities $p_i(\boldsymbol{j})$ using the bivariate recursion developed in (5.3.2), (5.3.13), and the expression for $d_i(\boldsymbol{j})$ given by (5.3.7). The initial conditions for the recursion are $p_2(\boldsymbol{e}_m)$ equals one if $z_1 = m$ otherwise it equals zero, for all $m = 1, \ldots, M$. The survival probability $\mathbb{P}(\tau_{n+1} > T | \tau_{1:n}, z_{1:n})$ is calculated using (5.3.3), (5.3.8), and $p_{n+1}(\boldsymbol{j})$. Following the evaluation of all these terms, they can be substituted into (5.3.1) to compute the likelihood. The parametric forms for the immigrant hazard rate functions and offspring densities imply that the likelihood can be directly evaluated and maximized with general-purpose optimization routines to obtain the MLEs.

Remark 5.3.1. *The computational time required for likelihood evaluation of the MRHawkes process is a polynomial function of the number of events n, or $O(n^{M+1})$. The storage of the most recent immigrant probabilities requires an M-dimension matrix of size n^M. Therefore, computation of the likelihood is practically infeasible in applications with a large n. However, the most recent immigrant probabilities become insignificant for very distant event times and therefore, to speed up the likelihood evaluation algorithm, these probabilities are assumed to be insignificant and are thus truncated. The modified algorithm only considers the most recent B events to be possible immigrants, and hence the storage is always reduced to an M-dimensional matrix of size B^M. With this truncation, the time required to compute the likelihood is also reduced to $O(n^2)$ in general, or $O(n)$ when the offspring distribution is exponential. The tuning parameter B signifies a trade-off between computational time and computational accuracy. In practice, it is advisable to use several B values to make sure the truncation is not having a material effect on the final parameter estimates.*

5.4 Model assessment

A natural next step is to assess how well the model fits the data. There are two aspects of the model that need to be examined, the temporal patterns of the events and the distribution of the event types. For the former, the Rosenblatt (1952) residuals, similar to that used in Chapter 3 will be employed. For the latter, the universal residuals introduced by Brockwell (2007), a generalized version of the Rosenblatt residuals to accommodate distributions with discontinuities is implemented. When the model specification is correct, the residuals form an *i.i.d.* sequence of uniform random variables on the unit interval. Therefore, to assess model fit, one can examine the residual sequence for uniformity and independence.

More specifically, the Rosenblatt residuals for the observed event times are defined as $W_i = F_i(\tau_i|\tau_{1:i-1}, z_{1:i-1})$, where $F_i(t|\tau_{1:i-1}, z_{1:i-1})$ is the conditional distribution function of τ_i given $\tau_{1:i-1}$ and $z_{1:i-1}$. Analogous to (3.4.1) in Chapter 3, the W_i are given by,

$$W_i = F_i(\tau_i|\tau_{1:i-1}, z_{1:i-1}) = 1 - \sum_{\forall \boldsymbol{j}} p_i(\boldsymbol{j}) S_i(\boldsymbol{j}), \tag{5.4.1}$$

where $p_i(\boldsymbol{j})$ are the most recent immigrant probabilities which are computed in the likelihood evaluation in (5.3.13) and $S_i(\boldsymbol{j})$ are given similarly to (5.3.8) by the

following,

$$S_i(\boldsymbol{j}) = \exp\left\{-\sum_{m=1}^{M}\{U_m(\tau_i - \tau_{j_m}) - U_m(\tau_{i-1} - \tau_{j_m})\} - \{\Phi(\tau_i) - \Phi(\tau_{i-1})\}\right\}, \quad \forall \boldsymbol{j}.$$

Furthermore, the universal residuals for the observed event types $z_{1:n}$ are defined as follows,

$$V_i = (1 - U_i)G_i(z_i - |z_{1:i-1}, \tau_{1:i}) + U_i G_i(z_i|z_{1:i-1}, \tau_{1:i}),\ i = 1, \ldots, n,$$

where $U_{1:n}$ is an auxiliary sequence of *i.i.d.* uniform random variables on the unit interval, independent of $N(T)$, $z_{1:n}$ and $\tau_{1:n}$; $G_i(z|z_{1:i-1}, \tau_{1:i})$ is the (discrete) distribution function of the event type z_i given previous event types $z_{1:i-1}$ and previous and current event times $\tau_{1:i}$; and $G_i(z-|z_{1:i-1}, \tau_{1:i})$ denote the left limit of $G_i(\cdot|z_{1:i-1}, \tau_{1:i})$ at z. To compute the conditional distribution function G_i, the conditional probabilities of the event types are necessary. From the calculations in Section 5.3, note that,

$$\mathbb{P}(z_i = m|z_{1:i-1}, \tau_{1:i}) = \frac{\sum_{\forall \boldsymbol{j}} d(\boldsymbol{j}, m)p_i(\boldsymbol{j})}{\sum_{z=1}^{M}\sum_{\forall \boldsymbol{j}} d(\boldsymbol{j}, z)p_i(\boldsymbol{j})},$$

where $d(\boldsymbol{j}, z_i) = d_i(\boldsymbol{j})$ as given previously in (5.3.4).

5.5 Model predictions

This section considers the problem of predicting the occurrence time and event type of the next event after the censoring time T based on the observations up until the censoring time.

5.5.1 Predictive density and hazard function

The plug-in predictive density function provides the first solution to this problem. Using the conditional probabilities derived in the evaluation of the likelihood, the joint conditional predictive density of the next occurrence time and event type $(\tau_{N(T)+1}, z_{N(T)+1})$ with respect to the product measure $\mathcal{L} \otimes \mathcal{C}$, where $\mathcal{L}$ and $\mathcal{C}$ are the Lebesgue and counting measures respectively, is given by,

$$p\left(\tau, z|\tau_{1:n}, z_{1:n}, \tau > T\right) = \frac{\sum_{\boldsymbol{j}} p_{n+1}(\boldsymbol{j})\, d_{n+1}(\boldsymbol{j})}{\mathbb{P}\left(\tau_{n+1} > T|\tau_{1:n}, z_{1:n}\right)}, \quad \tau > T,\ z \in \{1, \ldots, M\}, \quad (5.5.1)$$

where $p_{n+1}(\boldsymbol{j})$ are calculated using (5.3.13) as before, the denominator is computed in (5.3.3) and $d_{n+1}(\boldsymbol{j})$ are given similar to (5.3.7) by the following,

$$d_{n+1}(\boldsymbol{j}) = (\mu_z(\tau - \tau_{j_z}) + \phi_z(\tau))e^{-\sum_{m=1}^{M}\{U_m(\tau - \tau_{j_m}) - U_m(\tau_n - \tau_{j_m})\} - \{\Phi(\tau) - \Phi(\tau_n)\}}. \quad (5.5.2)$$

The predictive density for the next events occurrence time is obtained by computing the marginal density from the joint density in (5.5.1) by taking the summation over all possible event types z to obtain the following,

$$p\left(\tau|\tau_{1:n}, z_{1:n}, \tau > T\right) = \frac{\sum_{\boldsymbol{j}} \left\{p_{n+1}(\boldsymbol{j}) \sum_z d_{n+1}(\boldsymbol{j})\right\}}{\mathbb{P}\left(\tau_{n+1} > T|\tau_{1:n}, z_{1:n}\right)}, \quad \tau > T. \tag{5.5.3}$$

The predictive density depends on the unknown parameters of the model and to overcome this, the unknown parameters are replaced by the MLEs obtained in Section 5.3, which then give rise to the plug-in predictive density.

5.5.2 Predictive simulations

A second approach to future event prediction is to simulate the number of events from the censoring T to a future time point $\tilde{T}$. The predictive simulations account for the actual observations up to time T by conditioning on $N(T) = n$ and the values of $\tau_{1:n}$ and $z_{1:n}$. The algorithm works as follows. First, simulate the M-dimensional event index vector of the most recent immigrants at the censoring time, according to the conditional distribution $\mathbb{P}\left(I(T) = \boldsymbol{j}|\tau_{1:n}, z_{1:n}, \tau_{n+1} > T\right) = p_{n+1}(\boldsymbol{j})$. Second, simulate the next immigrant arrival time for each component according to the appropriate conditional inter-renewal distribution given that it is greater than the duration between the simulated most recent immigrant arrival time and the censoring time. Third, simulate the future arrival times of immigrants of different components by time $\tilde{T}$ according to the respective inter-renewal distributions. Fourth, simulate the offspring processes up to time $\tilde{T}$ for each of the immigrants in the interval $(T, \tilde{T}]$ according to the algorithm to be described in Section 5.6.1 below. Last, simulate the arrival times in $(T, \tilde{T}]$ of the offspring descending from events prior to the censoring time T according to a non-stationary multivariate Hawkes process (NSMHP) with baselines intensity functions $\nu_j(\cdot) = \phi_j(T + \cdot)$, $j = 1, \ldots, M$ and excitation functions $g_{m,n}(\cdot) = \eta_{m,n} h_{m,n}(\cdot)$, $m, n = 1, \ldots, M$, using the algorithm described below in Section 5.6.1.

This procedure enables the future to be simulated many times, given the model parameters, and thus leads to another method to the first prediction problem. From the predictive simulations, the next event time and event type can be extracted. In fact, any quantity of interest can be extracted from these simulations. For example, the construction of prediction intervals for the number of events in a prediction window is obtained by extracting the appropriate quantiles of the simulated numbers of events. It should be remarked that this method does not take into account the uncertainty in the parameter estimates and could lead to overly confident predictions. However, Section 5.6.4 will demonstrate through simulations that this effect is

inconsequential when there is sufficient data to ensure accurate estimation of model parameters.

5.6 Simulations

This section conducts simulation studies to assess the numerical performance of the MLE of the MRHawkes process model introduced in Section 5.3 and assesses the predictive performance of the model using simulated data. This section also explains how to efficiently simulate the process up to a predetermined censoring time T by utilizing the linear nature of the intensity.

5.6.1 Simulation algorithm

Simulation of the MRHawkes process model can be efficiently implemented using the cascading algorithm motivated by the cluster process representation of the MRHawkes process. To simulate the occurrence times and event types to a predetermined censoring time T, first, simulate the immigrant arrival times up to time T for each event type as the cumulative sums of *i.i.d.* positive random variables with the appropriate hazard rate function. For each event type $z \in \{1, \dots, M\}$, denote the simulated immigrant arrival times by $\tau^0_{z,1} < \tau^0_{z,2} < \cdots < \tau^0_{z,n_z} \leqslant T$. Then simulate the corresponding offspring for each of the immigrants $i = 1, \dots, n_z$ up to time T, by simulating a non-stationary multivariate Hawkes process (NSMHP) with baseline intensity functions $\nu_j(t) = \eta_{j,z} h_{j,z}(t)$, $j = 1, \dots, M$, and excitation functions $g_{m,n}(t) = \eta_{m,n} h_{m,n}(t)$, $m, n \in \{1, \dots, M\}$, on the interval $(0, T - \tau^0_{z,i}]$, and then translate the event times into the interval $(\tau^0_{z,i}, T]$.

The NSMHP on the interval $(0, T]$ with baseline intensity functions $\nu_j(\cdot)$, $j = 1, \dots, M$ and excitation functions $g_{m,n}(\cdot)$, $m, n = 1, \dots M$, can be simulated with a cascading algorithm as follows. First, simulate the generation 0 events of types $j = 1, \dots M$ on $(0, T]$ according to independent Poisson processes with intensity functions $\nu_j(\cdot)$, $j = 1, \dots, M$; then keep simulating generation i ($i = 1, 2, \dots$) events as long as the number of generation $i - 1$ events of any type is non-zero. For each event type $n = 1, 2, \dots, M$, simulate the generation i events of types $m = 1, \dots, M$ according to M independent Poisson processes with respective intensity functions $g_{m,n}(\cdot)$. When this recursive process stops, return events of all generations with their respective event type labels as the events of the NSMHP on the interval $(0, T]$.

5.6.2 Simulation results

This section assesses the numerical performance of the statistical inferential methods developed in Section 5.3. The simulations performed in this section are similar

to those discussed in Section 3 with Weibull renewal immigrant inter-event waiting times with shape parameter κ_m and scale parameter β_m. The offspring densities are chosen to be exponential with shape parameter (or mean) $\gamma_{m,n}$. The bivariate version of the MRHawkes process model is analyzed with $M = 2$. For the first Weibull renewal distribution, the shape and scale parameter are $\kappa_1 = 3$ and $\beta_1 = 1.2$ and for the second $\kappa_2 = 1/3$ and $\beta_2 = 0.2$. These two processes correspond to evenly distributed immigrant arrivals and high levels of burstiness and clustering, as mentioned previously. The scale parameters for the renewal immigrant distributions are selected so that the expected waiting time between immigrants of the same type is close to one. For the endogenous aspects of the process, the exponential offspring densities are chosen to have a mean waiting time in the set $\gamma_{m,n} \in \{0.5, 1, 3\}$. This parameter selection is chosen to exhibit offspring waiting times that are shorter than, equal to, and longer than the expected immigrant inter-event waiting times. Furthermore, this simulation study assumes that the offspring densities for offspring events of the same type have a common shape parameter, that is, $\gamma_{1,1} = \gamma_{1,2}$ and $\gamma_{2,2} = \gamma_{2,1}$ which are henceforth denoted by γ_1 and γ_2 respectively. The branching ratios for the self-excitation effects are chosen to be either $\eta_s = 0.3$ or $\eta_s = 0.7$, corresponding to low and high levels of self-excitation respectively. The cross-excitation effects have branching ratios $\eta_c = 0.1$ or $\eta_c = 0.2$, implying that the branching matrix $\boldsymbol{H}$ has a spectral radius less than one, ensuring the stability of the process.

For each combination of the chosen parameters, the MRHawkes process was simulated 500 times with varying censoring times T indicated in the table to ensure the mean length of the realizations was about 1000. For each simulated dataset, the MLE was computed by directly maximizing the negative log-likelihood function using the quasi-Newton method, BFGS (Broyden-Fletcher-Goldfarb-Shanno). The computations were implemented using the `R` language (R Core Team, 2016), with the aid of the `optim` function. The computations are conducted on Intel Xeon X5675 processors (12M cache, 3.06 GHz, 6.4GT/S QPI). The results of the simulations are reported in Table 5.6.1 in which it reports; the mean of the parameter estimates (Est), the empirical standard error of the parameter estimates (SE), the average of the standard errors obtained by inverting the approximate Hessian matrix ($\hat{\text{SE}}$), the average length of the realizations (AL), the average running time for the optimization and computation of the approximate Hessian matrix (RT) and the empirical coverage probability (CP) of the approximate 95% confidence intervals. The results in Table 5.6.1 suggest that the maximum likelihood parameter estimates display consistency as the estimates show minimal bias. The standard error estimates capture the true variance of the estimates considerably well as they are very close to the empirical standard errors.

	κ_1	β_1	κ_2	β_2	γ_1	γ_2	$\eta_{1,1}$	$\eta_{1,2}$	$\eta_{2,1}$	$\eta_{2,2}$
True	3	1.2	1/3	0.2	1	1	0.7	0.2	0.1	0.3
Est	3.314	1.173	0.331	0.232	1.028	1.097	0.686	0.212	0.101	0.291
SE	1.113	0.169	0.0288	0.0870	0.248	0.814	0.0568	0.0911	0.0257	0.0826
$\hat{\text{SE}}$	0.803	0.132	0.0294	0.0751	0.206	0.435	0.0517	0.0838	0.0251	0.0789
CP	0.944	0.950	0.948	0.972	0.950	0.986	0.952	0.948	0.950	0.964
	RT = 20.2 hrs $T = 170$ Spr($\boldsymbol{H}$) = 0.75 AL = 1022									
True	3	1.2	1/3	0.2	1	1	0.3	0.1	0.2	0.7
Est	3.119	1.200	0.332	0.231	1.151	1.030	0.282	0.107	0.213	0.683
SE	0.522	0.0909	0.0356	0.137	0.592	0.193	0.0700	0.0310	0.0827	0.0587
$\hat{\text{SE}}$	0.479	0.0836	0.0334	0.0899	0.465	0.177	0.0663	0.0293	0.0799	0.0576
CP	0.950	0.948	0.952	0.984	0.952	0.950	0.942	0.940	0.940	0.950
	RT = 20.8 hrs $T = 170$ Spr($\boldsymbol{H}$) = 0.75 AL = 1000									
True	3	1.2	1/3	0.2	0.5	3	0.3	0.1	0.1	0.3
Est	3.027	1.197	0.326	0.221	0.527	3.304	0.294	0.103	0.103	0.302
SE	0.278	0.0519	0.0218	0.0641	0.169	2.640	0.0378	0.0330	0.0398	0.0608
$\hat{\text{SE}}$	0.262	0.0484	0.0195	0.0515	0.145	1.155	0.0357	0.0298	0.0360	0.0588
CP	0.952	0.928	0.950	0.976	0.948	0.990	0.948	0.956	0.948	0.954
	RT = 19.2 hrs $T = 360$ Spr($\boldsymbol{H}$) = 0.40 AL = 1003									
True	3	1.2	1/3	0.2	3	0.5	0.3	0.1	0.1	0.3
Est	3.021	1.195	0.326	0.218	3.932	0.506	0.287	0.113	0.102	0.306
SE	0.311	0.0597	0.0216	0.0554	3.509	0.101	0.0558	0.0624	0.0215	0.0490
$\hat{\text{SE}}$	0.294	0.0542	0.0195	0.0497	1.955	0.0973	0.0507	0.0481	0.0223	0.0485
CP	0.952	0.960	0.934	0.962	0.970	0.960	0.952	0.966	0.954	0.954
	RT = 20.8 hrs $T = 360$ Spr($\boldsymbol{H}$) = 0.40 AL = 1016									

Table 5.6.1: Results of the maximum likelihood estimation of the MRHawkes processes with Weibull renewal immigration and exponential offspring densities, based on 500 simulated datasets in each case.

When the spectral radius of the branching matrix $\boldsymbol{H}$ is high, that is, when $\text{Spr}(\boldsymbol{H}) = 0.75$, the immigration scale parameters tend to have a much larger empirical bias and standard error compared to when the spectral radius is lower, that is, when $\text{Spr}(\boldsymbol{H}) = 0.4$. This observation is to be expected since the total number of immigrants is much smaller in this circumstance due to the higher levels of excitation effects, and the length of the realizations remain comparatively fixed. The expected number of type-1 immigrants is 159, and the expected number of type-2 immigrants is 142 when the spectral radius is 0.75. When the spectral radius is only 0.40, the expected number of type-1 immigrants is 336 and 300 for type-2 immigrants. When the arrival of offspring events occur more frequently relative to immigrants, there tends to be an overestimation of the shape parameter κ_m while the scale parameter β_m exhibits quite a significant bias. The branching ratios $\eta_{m,n}$ tend to be well estimated for the range of situations considered.

The top two panels presented in Table 5.6.1 have mean waiting times for offspring generation γ_m well estimated with the estimates showing minimal bias. This is a result of a large number of total offspring events present in the realizations. In

the lower two panels, observe that the estimates of the mean offspring waiting time parameter tend to have a much larger empirical bias when the offspring mean waiting time parameter is larger ($\gamma_m = 3$) compared to when the parameter is smaller ($\gamma_m = 0.5$). The reason for this phenomenon is that a larger shape parameter value causes the likelihood surface to become much flatter near the true parameter value in the dimension of the shape parameter.

5.6.3 Comparison with the modified likelihood evaluation algorithm

This section performs a simulation study to compare the performance of the modified likelihood evaluation algorithm discussed in Remark 5.3.1 against the full likelihood evaluation algorithm. The simulation model considered in this comparison is the first simulation model discussed in Table 5.6.1. The estimates of the parameters are computed using the modified likelihood evaluation algorithm with values of B in the set $\{100, 200, 300, 400, \infty\}$. The same set of simulations as in Table 5.6.1 are used, and the results using the modified algorithm are reported in Table 5.6.2. The case in which immigrants arrive more uniformly across time ($\kappa_1 = 3$) only demands a relatively small value of B. It can be observed that looking back only 100 events is reasonably sufficient for accurate estimation of the parameters governing the first component as κ_1, β_1, γ_1, $\eta_{1,1}$ and $\eta_{1,2}$ are all very close to the true MLEs and are their standard errors and coverage probabilities. However, when immigrants arrive in burst or cluster heavily ($\kappa_2 = 1/3$), more distant events are required in the approximation, approximately 400 events in this particular simulation. The type-2 immigrant inter-event waiting times in this simulation model can be considerably large in comparison to the offspring waiting times and also the type-1 immigrant inter-event waiting times. As a result, more distant events are needed in the approximation to obtain an accurate account of all possible most recent immigrants. The component two parameters κ_2, β_2, γ_2, $\eta_{2,1}$ and $\eta_{2,2}$ gradually get closer to the true MLEs as the tuning parameter B becomes larger, with quite good agreement when $B = 400$.

From this simulation study, it is evident that the choice of B depends on the immigrant inter-event waiting times relative to the offspring waiting times as well as the level of self- and cross-excitation, as discussed in detail in Chapter 4. A larger number of events must be considered as possible most recent immigrants when the immigrant renewal distribution exhibits over-dispersion relative to the Poisson process. This is because many occasions occur when the offspring waiting times are much shorter than the waiting time between successive immigrant arrivals, and so one needs to look further back into the past, and hence a more considerable value

	κ_1	β_1	κ_2	β_2	γ_1	γ_2	$\eta_{1,1}$	$\eta_{1,2}$	$\eta_{2,1}$	$\eta_{2,2}$
True	3	1.2	1/3	0.2	1	1	0.7	0.2	0.1	0.3
Est	3.314	1.173	0.331	0.232	1.028	1.097	0.686	0.212	0.101	0.291
SE	1.113	0.169	0.0288	0.0870	0.248	0.814	0.0568	0.0911	0.0257	0.0826
$\hat{\text{SE}}$	0.803	0.132	0.0294	0.0751	0.206	0.435	0.0517	0.0838	0.0251	0.0789
CP	0.944	0.950	0.948	0.972	0.950	0.986	0.952	0.948	0.950	0.964
				RT = 20.2 hrs		B = Inf (MLE)				
Est	3.315	1.173	0.348	0.279	1.028	1.122	0.686	0.212	0.0869	0.282
SE	1.113	0.169	0.0277	0.128	0.248	0.511	0.0568	0.0912	0.0264	0.0777
$\hat{\text{SE}}$	0.804	0.132	0.0292	0.0796	0.206	0.425	0.0517	0.0837	0.0247	0.0776
CP	0.944	0.950	0.952	0.970	0.950	0.962	0.952	0.948	0.944	0.956
				RT = 0.407 hrs		B = 100				
Est	3.314	1.173	0.335	0.247	1.028	1.120	0.686	0.212	0.0986	0.287
SE	1.113	0.169	0.0276	0.109	0.248	1.077	0.0568	0.0911	0.0261	0.0853
$\hat{\text{SE}}$	0.803	0.132	0.0294	0.0778	0.206	0.444	0.0517	0.0838	0.0250	0.0790
CP	0.944	0.950	0.946	0.972	0.950	0.992	0.952	0.948	0.950	0.974
				RT = 1.185 hrs		B = 200				
Est	3.314	1.173	0.332	0.238	1.028	1.100	0.686	0.212	0.101	0.290
SE	1.113	0.169	0.0281	0.0986	0.248	0.827	0.0568	0.0911	0.0257	0.0823
$\hat{\text{SE}}$	0.803	0.132	0.0294	0.0766	0.206	0.437	0.0517	0.0838	0.0251	0.0789
CP	0.944	0.950	0.946	0.972	0.950	0.988	0.952	0.948	0.948	0.966
				RT = 2.63 hrs		B = 300				
Est	3.314	1.173	0.331	0.235	1.028	1.098	0.686	0.212	0.101	0.291
SE	1.113	0.169	0.0284	0.0941	0.248	0.819	0.0568	0.0911	0.0257	0.0825
$\hat{\text{SE}}$	0.803	0.132	0.0294	0.0762	0.206	0.435	0.0517	0.0838	0.0251	0.0789
CP	0.944	0.950	0.942	0.978	0.950	0.986	0.952	0.948	0.946	0.964
				RT = 4.988 hrs		B = 400				

Table 5.6.2: Results of the maximum likelihood estimation with the modified likelihood evaluation algorithm that only considers the previous B events as candidates for the most recent immigrant event.

of B needs to be selected. However, when immigrants occur quite regularly through time, and the immigration process exhibits under-dispersion, only recent events need to be considered as possible most recent immigrants, and B can be much smaller in this case. The choice of tuning parameter value B represents a trade-off between the accuracy of the parameter estimates compared to the true MLE and the time required for estimation. One possible method to determine an appropriate value of B is to study the difference between parameter estimates for different choices of B, and when the difference is immaterial for that particular application, then that value of B would be appropriate.

5.6.4 Assessment of the predictive performance

The predictive performance of the simulation-based prediction procedure discussed in Section 5.5 will now be examined. Consider making predictions using the first set of simulated data in Table 5.6.1. The aim is to predict the number of events

that occur in the prediction window $(170, 283]$, which is two thirds the length of the observation period for the 500 simulated datasets from the top panel in Table 5.6.1. To assess the predictive performance, the predicted sample paths based on simulating the future with the estimated model parameters using the method of maximum likelihood are compared to the true parameter simulated path. For each simulated dataset, the future is simulated 500 times. The resulting 95% prediction interval for the 500 sample paths contains the true number of events in 86.52% of all cases. This is slightly lower than the expected 95%, but the randomness of the parameter estimates has not been taken into consideration. The length of the observed sample paths are on average 1000 events, but the large number of parameters and the large standard errors can hinder the predictive performance. When the realizations are long enough, this should not be overly detrimental to the accuracy of the predictions.

The randomness inherent in the parameter estimates could be accounted for by using the sampling distribution for θ, where θ contains the vector of parameters. For each simulated path, a new parameter $\hat{\theta}_j$ is simulated from the multivariate normal distribution with mean $\hat{\theta}$ (the estimated parameters) and variance-covariance matrix obtained from the approximate Hessian matrix from the numerical optimization procedure. The simulation study suggests that the estimates are relatively normal as the empirical coverage probability is relatively consistent at the 95% level suggesting asymptotic normality. However, it should be observed that when the shape parameter of the offspring density is large, this assumption might not be reasonable. As the number of observations of the multivariate point process increases, the randomness in the parameter estimates are substantially reduced and will have a modest impact on the prediction interval.

5.7 Applications

5.7.1 Analysis of earthquakes around Fiji and Vanuatu

This section analyzes the arrival times of earthquakes occurring in two Pacific island countries Fiji and Vanuatu. The study considers magnitude 5.5 or higher earthquakes measured on the Richter scale for the 25 years from 01/01/1991 to 31/12/2015. The data for this analysis was obtained from the earthquakes archive from the United States Geological Survey (USGS), which consists of 1076 earthquakes occurrences. During this period 646 earthquakes occur in the area of Fiji, from hereon in denoted as type-1 events and 428 occur in the area of Vanuatu which will be denoted as type-2 events. Figure 5.7.1 presents a plot of earthquake occurrences surrounding Fiji and Vanuatu during this period where Fiji is on the right, and Vanuatu is on the left, where the solid circles indicate earthquake in Fiji and circles for earthquake occurring in Vanuatu.

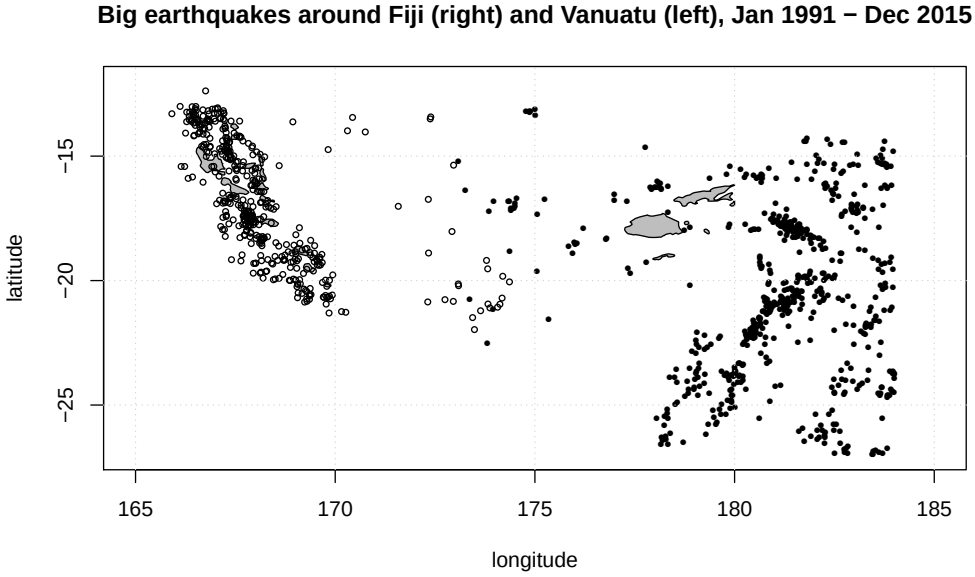

Figure 5.7.1: Large magnitude earthquakes occurring in Fiji and Vanuatu during 1991 to 2015, where circles indicate locations of earthquakes in Vanuatu and solid circles indicate the locations of earthquakes in Fiji.

The objective is to understand the interactions between these two neighboring countries and their propensity for earthquakes. The earthquake data is modeled with an MRHawkes process model with two components. The renewal immigration inter-event waiting time distributions are Weibull, and the offspring excitation function is exponential with a common shape parameter γ_m, which does not depend on the country (location) of the igniting earthquake. The model incorporates interactions between the two neighboring countries and allows for ease of interpretation between immigrant arrivals (mainshocks) and their descendants (aftershocks) as well as location. It is plausible that some useful information is lost when attempting to model this relationship in the univariate context. The MLEs for the MRHawkes model are as follows,

$$\underset{(0.0654)}{\hat{\kappa}_1 = 0.470}, \quad \underset{(3.47)}{\hat{\beta}_1 = 19.97}, \quad \underset{(0.0221)}{\hat{\kappa}_2 = 0.342}, \quad \underset{(2.51)}{\hat{\beta}_2 = 10.36}, \quad \underset{(155)}{\hat{\gamma}_1 = 394},$$
$$\underset{(269)}{\hat{\gamma}_2 = 566}, \quad \underset{(0.152)}{\hat{\eta}_{1,1} = 0.428}, \quad \underset{(0.245)}{\hat{\eta}_{1,2} = 0.367}, \quad \underset{(0.111)}{\hat{\eta}_{2,1} = 0.382}, \quad \underset{(0.164)}{\hat{\eta}_{2,2} = -0.0375}.$$

where the standard errors are in brackets. The multivariate Hawkes process was also fit to the data for comparison with the same exponential offspring generation and resulted in the following parameter estimates,

$$\underset{(0.61)}{\hat{\mu}_1 = 14.84}, \quad \underset{(1.64)}{\hat{\mu}_2 = 27.26}, \quad \underset{(0.0198)}{\hat{\gamma}_1 = 0.0600}, \quad \underset{(0.0930)}{\hat{\gamma}_2 = 0.320},$$
$$\underset{(0.0111)}{\hat{\eta}_{1,1} = 0.0536}, \quad \underset{(0.00502)}{\hat{\eta}_{1,2} = -0.00332}, \quad \underset{(0.00563)}{\hat{\eta}_{2,1} = 0.00128}, \quad \underset{(0.0288)}{\hat{\eta}_{2,2} = 0.221}.$$

The Rosenblatt residuals are calculated to assess the models' ability at modeling the temporal patterns of the events. The uniform quantile plot and ACF plot are displayed in Figure 5.7.2. The uniformity of the residuals seems to be satisfied as the theoretical and empirical quantiles have a good agreement. This is reinforced with a large p-value of 0.76 and 0.83 for the Anderson-Darling (A-D) and K-S tests of uniformity respectively. The residuals also exhibit insignificant serial correlation up to lag 30, with a p-value for the Ljung-Box test of independence of 0.47. However, the Rosenblatt residuals for the multivariate Hawkes process also satisfy the uniformity assumption with a p-value of 0.22 and 0.72 for the A-D and K-S tests respectively but fails the Ljung-Box independence test with a p-value of only 0.01. The Universal residuals are also computed to assess how well the model captures the distribution of location. For the MRHawkes model, the computed p-value for the A-D and K-S test are 0.69 and 0.85 respectively, and the Ljung-Box test returns a p-value of 0.41, which suggest the model is capturing the distribution of location between the two countries well.

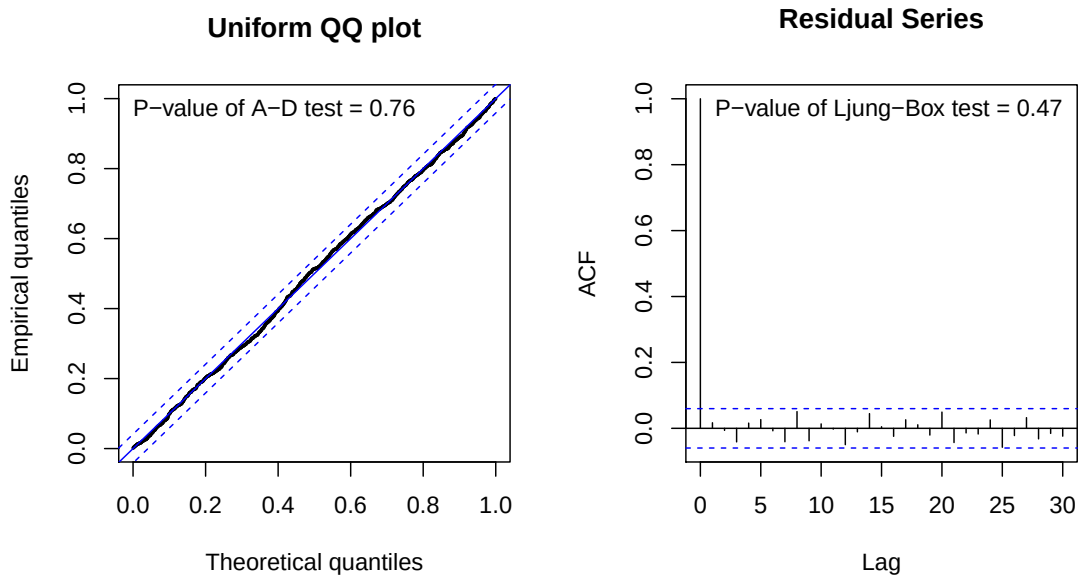

Figure 5.7.2: Uniform quantile plot and ACF plot for the computed Rosenblatt transform residuals for the Fiji and Vanuatu earthquake dataset for the MRHawkes model.

Next, this analysis assesses how well the model can capture both aspects of the point process, the temporal patterns of the events, and the event type distribution. The combined series of the Rosenblatt and universal residuals together are assessed for uniformity and independence. For the combined residuals, the uniformity and independence is well satisfied for the MRHawkes model with p-values of 0.91 for the A-D test, 0.60 for the K-S test and 0.32 for the Ljung-Box test, while the Hawkes

model returns p-values of 0.30 for the A-D test, 0.40 for the K-S test and only 0.01 for the Ljung-Box test and thus fails independence at the 5% level. Several auxiliary uniform random variables were used to compute the universal residuals, and the MRHawkes model was able to pass the test at 5% significance level in the majority of cases. However, the multivariate Hawkes model fails the Ljung-Box test on quite a few of the residual series with relatively low p-values. The Akaike information criterion (AIC) for the two models are 7775.5 for the MRHawkes model and 7820.3 for the Hawkes model. Furthermore, a likelihood ratio test gives a test statistic of 40.8 and a p-value of 1.38×10^{-9} which again indicates that the MRHawkes model is to be prefered over the multivariate Hawkes model. The AIC criterion, the likleihood ratio test, as well as the assessment of the residual series, suggests that the MRHawkes process is outperforming the multivariate Hawkes process and provides a superior quality of fit. Indeed the Hawkes model fails to capture any interaction between the two locations since the cross-exciting branching ratios are not significantly different from zero when taking into account their standard errors.

In the seismological context, immigrants are interpreted as mainshocks and offspring events as the aftershocks induced by a main or aftershock. The MRHawkes process model suggest that main shocks occur in Fiji on average every $\hat{\beta}_1\Gamma(1 + 1/\hat{\kappa}_1) = 45.08$ days and in Vanuatu it is every $\hat{\beta}_2\Gamma(1 + 1/\hat{\kappa}_2) = 56.49$ days. This interpretation differs significantly from the Hawkes model, in which, immigrants arrive on average every 14.84 and 27.26 days respectively, and occur more frequently. When a earthquake occurs in Fiji it directly induces on average $\hat{\eta}_{1,1} = 0.428$ aftershocks in Fiji and $\hat{\eta}_{2,1} = 0.382$ aftershocks in Vanuatu. Each earthquake in Vanuatu directly induces on average $\hat{\eta}_{1,2} = 0.367$ earthquakes in Fiji, although, with a standard error of 0.245, this effect is not significant at the 5% level. Somewhat surprisingly, the earthquakes in Vanuatu do not seem to generate aftershocks in Vanuatu. The offspring shape parameters $\hat{\gamma}_m$ are interpreted as the expected waiting time for a directly induced aftershock to occur with $\hat{\gamma}_1 = 394$ (days) in Fiji and $\hat{\gamma}_2 = 566$ (days) in Vanuatu.

Future earthquake occurrences are predicted by utilizing the fitted model and the prediction procedures developed in Section 5.5. The performance of the predictive simulations is assessed by comparing the predictions with the observed earthquake occurrences over the prediction interval under consideration. The simulation-based approach is used to predict considerable-sized earthquakes in the areas of Fiji and Vanuatu from 01/01/2016 until 30/06/2017. Using the identified model with $\eta_{2,2}$ set equal to zero, as it is not significantly different from zero and non-negative branching ratios are required for the simulation procedure employed in this thesis, 10,000 realizations of earthquake occurrences conditional on the earthquake times and types by the censoring time are simulated. The pointwise median and lower

and upper 2.5 percentiles of the simulated paths of the process as well as the actual count is presented in Figure 5.7.3 for Fiji, Vanuatu and also the combined count. It is observed that the sample paths for Fiji fall well within the prediction interval for the entire period, and the median tends to track the observed path well. In Vanuatu, a large number of earthquakes occur over this period. However, it still falls within our prediction interval by the end of the prediction window. The total number of earthquakes is comfortably within the prediction interval for the entire eighteen months.

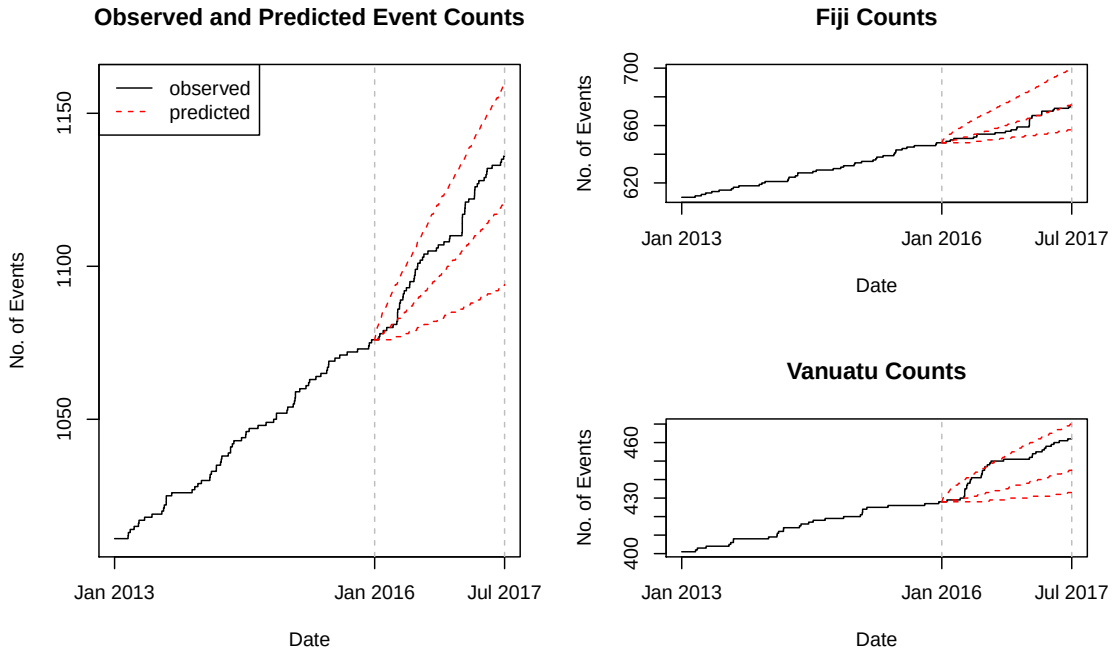

Figure 5.7.3: Actual and predicted earthquake occurrences. The solid curve is the actual earthquake counts, and the dashed curves show the point predictions and 95% prediction intervals at different time points.

The waiting time until the next earthquake occurrence can be studied using the plug-in predictive density function. The probability that an earthquake occurs within the next 20, 40, 60 and 80 day period is given by 65.72%, 87.18%, 94.91% and 97.88% respectively, with the next earthquake occurring in Fiji in only 17.77 days. Another prediction of interest is to predict the location of the next earthquake. Extracting the location of the first earthquake from the 10,000 simulations reveals that the first earthquake occurred in Fiji on 60.68% of all realizations which suggest that the next earthquake is more likely to occur in Fiji which again agrees with the actual data.

The primary focus of this analysis was on the temporal patterns of earthquake arrivals, with spatial aspects only taken into account by assigning the earthquake to either Vanuatu or Fiji. The assignment procedure of an earthquake to a particular

country only used the description data from the USGS database. Therefore, a limitation of this approach is the arbitrary classification of earthquakes in the Pacific Ocean and in particular, the boundaries between the two countries. As seen in Figure 5.7.1, some locations could be seen to have either classification. To fully account for the spatial aspects of the earthquake occurrences, spatio-temporal versions of the RHawkes process model similar to the ETAS (Epidemic Type Aftershock-Sequences) model of Ogata (1998) should be developed. A comprehensive comparison of the MRHawkes process to other alternative temporal point process models and their multivariate extensions, such as the trigger process model (Vere-Jones and Davies, 1966; Vere-Jones, 1970; Adamopoulos, 1976; Türkyilmaz et al., 2013), would also be interesting.

5.7.2 Modeling trade-throughs using bivariate RHawkes processes

Market participants generally attempt to hide or minimize their market impact by submitting orders based on the liquidity available in the order book. Instead of executing large orders and revealing their intentions to the market, traders typically split and restrict the size of their order to the quantity available at the best limit price. This assures that the price does not change unfavorably against them and thereby controls to some extent, the market impact of their order. However, in some instances, the speed of execution exceeds the cost of the market impact, and large orders are submitted with quantities greater than what is available at the first limit. Such transactions are termed trade-throughs. A trade-through is a transaction that occurs at least at the second level of limit orders in an order book and hence provides valuable information about price dynamics and market microstructure.

Empirical studies indicate that trade-throughs occur in clusters, and thus, self-exciting processes are a natural choice to model this phenomenon. Pomponio and Abergel (2013) examine the clustering effect of trade-throughs by comparing the waiting time between successive trade-throughs for the stock BNP Paribas. A clustering effect is evident when the next trade-through arrives at a faster rate after a trade-through than after any regular trade. To see this, they computed the empirical arrival time distribution of the next trade-through by conditioning on whether the current trade is a trade-through or any regular trade. The waiting time distribution until the next trade-through had a higher peak for shorter waiting times when the current trade is a trade-through.

Furthermore, Muni Toke and Pomponio (2011) computed the mean waiting time between trade-throughs and found that by conditioning on the current trade being a trade-through, the mean waiting time was only 36.9 seconds compared to

51.8 seconds for any regular trade. This suggests that trade-throughs are generally more likely to be followed by another trade-through and occur more closely in time. Muni Toke and Pomponio (2011) further revealed that there is no asymmetrical effect for the side of the book that the trade-through occurred. That is, irrespective of the sign of the trade, the mean waiting time was shorter if a trade-through occurred rather than a regular trade. They also show that the cross-excitation effects of trade-throughs are rather weak compared to the self-excitation effects. The mean waiting time for a trade-through on the same side of the order book is smaller than a trade-through on the opposite side of the order book.

Earlier attempts to model trade-throughs were conducted by Muni Toke and Pomponio (2011) in which they analyze the Thomson-Reuters tick-by-tick data of the Euronext-traded limit order book for the stock BNP Paribas (BNPP.PA) for the 109 trading days from 1st June 2010 to 29th October 2010. The data contains the timestamps, volume, and price of the trades and the volume, price, and side of the order book for the quotes. The Euronext Paris is open from 9 am to 5:30 pm local time (07:00 to 15:30 GMT). For each trading day, they extract the series of timestamps $(\tau_i^A)_{i\geqslant 1}$ and $(\tau_i^B)_{i\geqslant 1}$ of trade-throughs for the ask and bid side of the limit book. The non-stationarity of trading throughout the day requires Muni Toke and Pomponio (2011) only to consider trades that occur between 9:30 am to 11:30 am the local time where the number of trade-throughs during this period remains relatively constant throughout the period of analysis. They show that the bivariate Hawkes process with an exponentially decaying kernel can fit the majority of the two hour trading periods and that the cross-influence of the bid and ask trade-throughs is particularly weak.

Motivated by their work, the analysis herein aims to model the same trade-throughs data for the stock BNP Paribas by using bivariate RHawkes processes. Rather than only analyzing the two-hour window, this analysis considers the entire trading day. However, generally opening trades exhibit drastically different features than the rest of the trading day and for this reason, transactions that occur during the first half an hour of the day from 9 am to 9:30 am are removed from the analysis. The analysis will consider trade-throughs occurring during the trading day from 9:30 am to 5:30 pm. For the 109 trading days, the mean number of trade-throughs over this period was 756 with an average and standard deviation of 367 and 217 for the ask side and 389 and 204 for the bid side. The first and third quantiles are reasonably comparable with 126 and 859 for the ask side and 140 and 860 for the bid side.

Figure 5.7.4 displays the expected inter-event waiting time between trade-throughs conditioned on the time of day that the trade-through occurred. The expectation is estimated using a cubic regression spline approach used by Engle and Russell (1998),

where the knots are set at each hour of the trading day, with an extra knot at the middle of the last hour to account for the quickly changing level of trading activity near market close. Figure 5.7.4 displays a clear diurnal pattern with the opening of the market being quite active with trade-throughs occurring roughly every 20 seconds. The activity then reduces in the middle of the day with trade-throughs only occurring every 60 to 70 seconds. The activity then picks up again before the close with trade-throughs occurring roughly every 20 seconds again. The non-stationary nature of the arrival times of trade-throughs over a trading day is clearly evident, and therefore a data transformation is applied, similar to that used by Engle and Russell (1998) to account for the level of trading activity, by discounting the observed duration by a factor proportional to the corresponding expected duration subject to the constraint that the sum of the adjusted durations in a day is the same as the original durations.

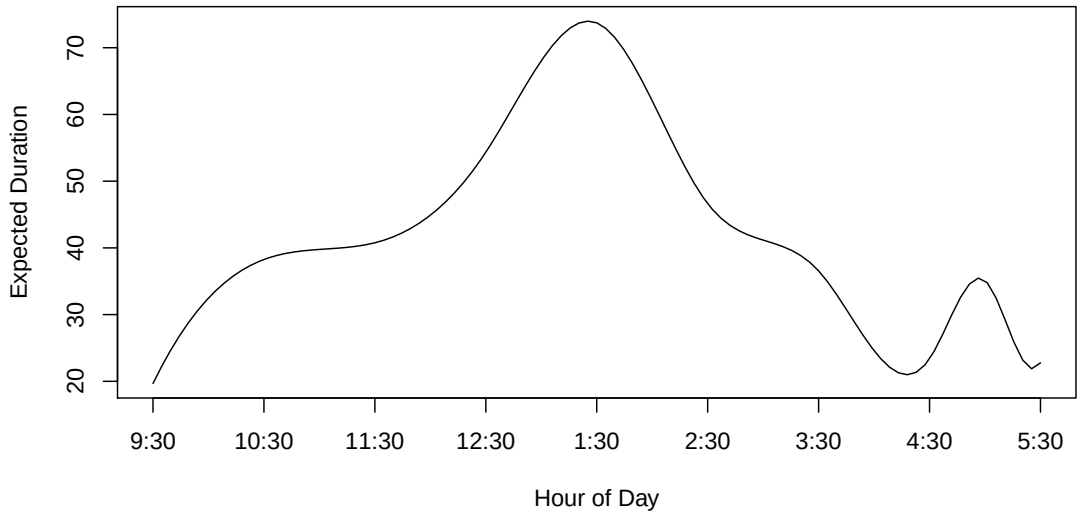

Figure 5.7.4: A nonparametric estimate of the daily pattern for trade-through durations conditional on the time of day of the trade.

The adjusted point processes are modeled with Weibull renewal processes for the immigrant arrivals, and the offspring densities are chosen to be exponential with those for offspring events of the same type having a common shape parameter. The parameter estimates for the 109 trading days are computed by directly minimizing the negative log-likelihood function, and the results are presented in Table 5.7.1. The table reports the mean, median, lower 2.5 percentile $Q_{0.025}$, upper 2.5 percentile $Q_{0.975}$, and the standard deviation of the estimates. The parameters governing the bid and ask side of the process are relatively similar, and both sides of trade-throughs are displaying very similar features. The immigration process for both the bid and ask trade-throughs tend to exhibit strong clustering and over-dispersion relative to a Poisson process with the estimated shape parameter of the Weibull distribution $\hat{\kappa}$

remaining below one on most trading days. By taking the median as an estimate for the true parameter, the mean waiting time between exogenous ask trade-through is 109.3 seconds while for the bid side trade-through is 106.4 seconds. The branching ratio parameters for the self- and cross-excitation are reasonably similar on either side of the book with the median value for the cross-exciting branching ratio being appropriately one-third of the self-exciting branching ratio.

	Mean	Median	$Q_{0.025}$	$Q_{0.975}$	StdDev
κ_A	0.801	0.816	0.320	1.006	0.148
β_A	108.8	97.55	18.82	251.8	61.47
κ_B	0.797	0.806	0.474	0.990	0.130
β_B	95.91	86.99	23.70	220.7	50.74
γ_A	22.75	0.0257	0.00919	294.4	126.7
γ_B	18.17	0.0228	0.00903	119.1	134.2
$\eta_{A,A}$	0.120	0.113	0.0293	0.220	0.0674
$\eta_{A,B}$	0.0543	0.0342	0.00544	0.329	0.0965
$\eta_{B,A}$	0.0446	0.0269	0.00633	0.229	0.0711
$\eta_{B,B}$	0.131	0.121	0.0688	0.247	0.0569

Table 5.7.1: Statistics summary for the MLEs for the diurnally adjusted BNP Paribas trade-throughs data.

The quality of fit to the data is evaluated using the Rosenblatt residuals. At the 1% level, the bivariate RHawkes model passed the A-D test and K-S test of uniformity on 78.90% and 81.65% of the trading days, respectively. For comparison, the bivariate Hawkes process was also fit to the transformed data with exponential offspring densities. At the 1% level, the A-D and K-S test of uniformity were passed on only 25.69% and 36.70% of the trading days, respectively. Furthermore, the goodness-of-fit of the event type distribution was assessed with the aid of the universal residuals. The residuals were assessed for uniformity on the unit interval and are found to pass 81 (74.31%) and 83 (76.15%) of all trading days at the 1% level for the A-D and K-S test respectively, which suggest the model is able to adequately model the distribution of the side on which a trade-through occurs. Again, similar to the earthquake case study the combined residuals series are assessed as well. For the MRHawkes model, the A-D and K-S test passes 80 (73.39%) and 82 (75.23%) of all trading days respectively, while the multivariate Hawkes model only passes 43 days (39.45%) for the A-D test and 50 days (45.87%) for the K-S test. Thus it can be concluded that the Weibull MRHawkes model is providing a better fit to the data than the classical multivariate Hawkes model.

To determine the necessity for bivariate RHawkes processes, two aspects of the fitted model require consideration, that is, the need for cross-excitation effects and whether a departure from Poisson immigration exists. The first question is solved by computing the z-score for the cross-exciting branching ratios $\eta_{A,B}$ and $\eta_{B,A}$ under the assumption that the cross-exciting effects are zero. Assuming the parameter estimates are asymptotically normal, which is suggested by the simulation study, the z-score can be compared to the value of 1.96. The parameters $\eta_{A,B}$ and $\eta_{B,A}$ are statistically different from zero for 92 (84.40%) and 87 (79.82%) trading days. Figure 5.7.5 displays the z-scores across the 109 trading days where the top panel is the influence of bid trade-throughs on ask trade-throughs $\eta_{A,B}$ and the bottom panel is the opposite effect. Although the cross-excitation exist, similar to Muni Toke and Pomponio (2011), the cross-influence effect between the two sides of the market tends to be relatively small compared to the self-exciting effects, although this is not always evident.

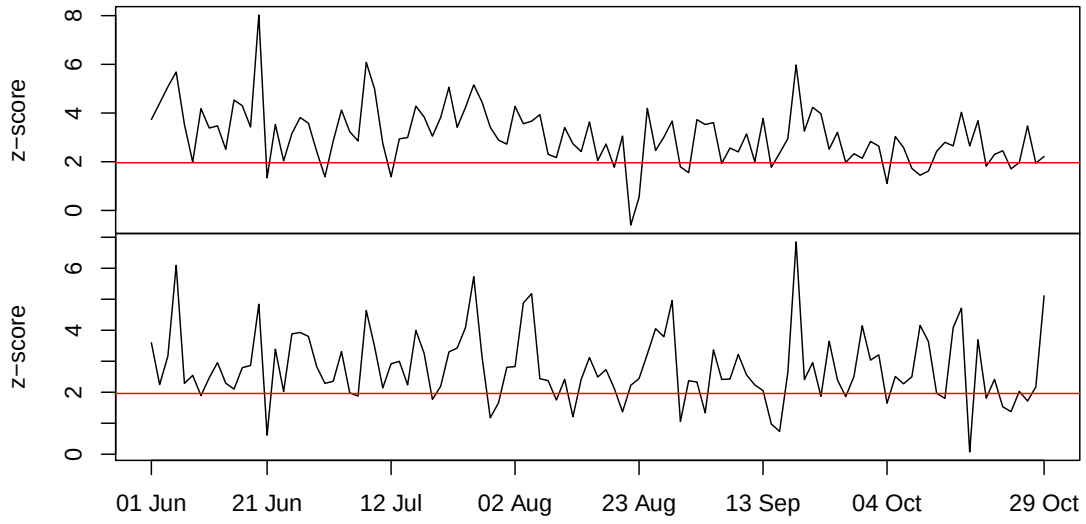

Figure 5.7.5: Time series plot of the z-scores for the cross-exciting branching ratios across the 109 trading days for the stock BNP Paribas. Top panel: the influence of bid trade-throughs on the ask side of the market. Bottom panel: the influence of ask trade-throughs on the bid side of the market.

The second question is solved by examining the shape parameter of the Weibull renewal distribution for both sides of the market. Figure 5.7.6 presents the time series plot of the shape parameter $\hat{\kappa}$ for both sides of the market together with a shaded 95% confidence interval. The value of κ_A and κ_B are mostly different from one, with the 95% confidence intervals not containing the value one 76.15% and 80.73% of all trading days. This suggests that departure from Poisson immigration for the arrival of the bid and ask trade-throughs exists.

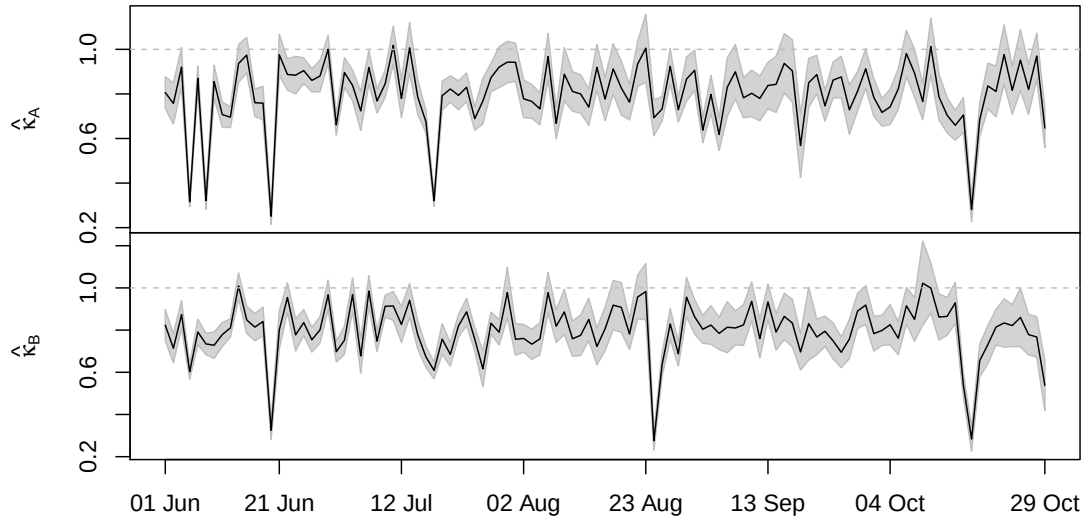

Figure 5.7.6: Time series plot of the MLEs for the shape parameters of the two Weibull immigration parameters over the 109 trading days 01/06/10 -29/10/10. Solid curve: MLE; shaded region: point-wise 95% confidence intervals.

Chapter 6

Modeling extreme negative returns using marked renewal Hawkes processes[1]

6.1 Introduction

Modeling extreme financial returns has important applications, such as in the estimation of risk measures. However, like many other financial time series, the series of extreme returns are challenging to model due to the presence of heavy temporal clustering of extremes and intense bursts of return volatility. To address this challenge, Chavez-Demoulin et al. (2005) and Chavez-Demoulin and McGill (2012) proposed the marked Hawkes process model, and reported sufficient fits to extreme negative return data on a share price and on the Dow Jones Industrial Average index, while the traditional peaks over threshold (POT) model was shown not to be suitable for the data considered. The use of marked Hawkes processes to forecast market risk measures has also been applied in the work of McNeil et al. (2005, pp 306-311) and Herrera and Schipp (2009). Furthermore, Embrechts et al. (2011) considered multitype event sequence data in which the multivariate version of the marked Hawkes process was used to model the interaction between positive and negative extreme returns for the Dow Jones Industrial Average index.

Extreme returns are fundamental to the risk management of financial institutions such as investment banks, insurers, and pension funds as they are often required to demonstrate their financial stability under extreme market conditions. A meaningful measure of risk for extreme loss outcomes is given by the quantile of the loss distribution of a given asset or portfolio over a predefined period, and this is known as the value at risk (VaR). Many approaches to estimate the unconditional VaR often as-

[1]Most of the content shown in this chapter has been published in the *Extremes*; see Stindl and Chen (2019).

sume that the return distribution is normally distributed and then forecast volatility using the exponential-weighted moving average method as in Mina and Xiao (2001). Other unconditional approaches often rely on generalized autoregressive conditional heteroskedasticity (GARCH) models with either normal or t innovations.

In other approaches to estimate the VaR, the conditional return distribution, which accounts for the current financial environment in which the asset is traded, are often used. McNeil and Frey (2000) introduced a conditional approach using a two-stage procedure by combining GARCH models to forecast volatility and then applying techniques from extreme value theory (EVT) to the residuals from the GARCH analysis. Although this method circumvents the use of unconditional return distributions, it introduces a new problem that relates to the sensitivity of the EVT analysis on the GARCH model fit. Another conditional approach was developed by Chavez-Demoulin et al. (2005), in which they apply the POT model from EVT to the excesses (return above a given threshold), which are treated as *i.i.d.* observations and model the temporal patterns of exceedances (days when an excess occurs), using a marked Hawkes (1971) self-exciting process. This approach models the serial dependence present in returns and provides a convenient method to estimate the conditional VaR and other risk measures of interest, such as the expected shortfall (ES).

However, despite their success, marked Hawkes processes are not always able to provide an adequate fit to data. On these occasions, added flexibility in the specification of the background arrival rate may be required. For instance, the background arrival rate may be allowed to depend on some covariates, but this approach requires appropriate external covariates to be available. This chapter proposes that the marked renewal Hawkes process model can provide this flexibility without the need to find suitable covariates. A readily implementable recursive algorithm to evaluate the likelihood of the model in linear storage space and quadratic computational time, which can be optimized to obtain estimates of model parameters and their standard errors is developed. A procedure to assess the goodness-of-fit for both aspects of the model, the temporal patterns of exceedances and the distribution of excesses, by calculating the Rosenblatt residuals (see Rosenblatt (1952)) and testing the residuals for uniformity and independence is also discussed. As by-products of the direct likelihood evaluation algorithm, estimates of the two risk measures, conditional VaR and conditional ES can be obtained. Furthermore, methods are provided to make predictions about future extreme negative returns, and in particular, the waiting time until the next exceedance and compare these predictions with actual observations.

In the next section, the ASX stock data is introduced. Section 6.3 introduces the marked RHawkes process model, which includes its estimation and inferential meth-

ods such as goodness-of-fit assessment, prediction and estimation of risk measures. Numerical illustrations will follow in Section 6.4 with a simulation study. The focus of Section 6.5 is on applying the proposed methods to extreme negative returns for each of the five ASX stocks.

6.2 ASX stock data

This section introduces five commonly traded stocks on the ASX (Australian Securities Exchange) that are studied in the analysis conducted in Section 6.5. The data was obtained from the Yahoo! Finance database and contains the date, open, high, low and close price for the following stocks traded on the ASX from 1 January 2006 to 31 December 2016; JB Hi-Fi Limited (JBH), Adelaide Brighton Limited (ABC), Computershare Limited (CPU), Downer EDI Limited (DOW) and James Hardie Industries plc (JHX).

The RHawkes model's performance at forecasting market risk and in particular, estimating the conditional VaR is assessed by backtesting. Therefore, the period under consideration is divided into two non-overlapping periods, which are termed the in-sample and out-of-sample period. The in-sample data are used to estimate the model parameters, and then the estimated model is used to make a forecast of market risk measures during the out-of-sample period and the actual data in the out-of-sample period is used to assess the forecasted risk measures. The period from 1 January 2006 to 31 December 2015 is the in-sample period and the following year from 1 January 2016 to 31 December 2016 is used as the out-of-sample period. Numerous descriptive statistics for the daily log-losses for the in-sample period are reported in Table 6.2.1, which contains the number of observations, minimum, maximum, mean, standard deviation, and kurtosis. Notice that the kurtosis for all stocks are larger than three, which suggest that the return distributions are leptokurtic rather than normal.

Stock	JBH	ABC	CPU	DOW	JHX
n	2528	2527	2528	2524	2528
min	-16.03	-13.31	-14.51	-13.35	-20.15
max	16.57	14.61	11.20	36.38	12.84
mean	-0.0624	-0.0323	-0.0213	0.0260	-0.0262
std dev	2.316	2.049	1.893	2.611	2.359
kurtosis	8.125	7.430	8.011	30.212	8.320

Table 6.2.1: Numerous descriptive statistics for the in-sample daily log-losses in percent for each of the five ASX stocks.

Now, denote the percentage log-loss from day $t-1$ to day t by $r_t = -100 \times \log(s_t/s_{t-1})$ where s_t is the closing price. This analysis considers extreme negative

returns, which exceed a high threshold denoted by u. If the loss on day t exceeds the threshold value u, then an exceedance has occurred and provided that an exceedance has occurred, the excess loss is given by $w_t = r_t - u$. The choice of the threshold value of u requires special attention. In this analysis, a threshold equal to the 90% quantile of the log-losses in the in-sample period was used for each stock, so that the 10% largest losses are considered as extreme negative returns. The choice of threshold value u is to some extent rather arbitrary but follows the convention used in the work of Chavez-Demoulin et al. (2005) and it can be argued that using a lower threshold would question the validity of EVT while using a higher threshold would reduce the sample size considerably. With the current choice, the (not shown) mean-excess plots (cf. Embrechts et al., 1997, p. 355) do not indicate that a violation of the assumptions is apparent. The threshold value u for all five stocks are shown in Table 6.2.2 as well as some descriptive statistics, including the number of exceedances, mean excess and median excess for the in-sample period.

Stock	JBH	ABC	CPU	DOW	JHX
threshold	2.425	2.281	2.082	2.662	2.645
no. of exceedances	253	252	253	253	253
mean excess	1.608	1.456	1.255	1.873	1.401
median excess	1.074	0.898	0.743	0.935	0.887

Table 6.2.2: Numerous descriptive statistics for the in-sample loss excesses for each of the five ASX stocks with a threshold value u chosen as the 90% quantile.

Figure 6.2.1 visualizes the transformation from raw price data into exceedance data with threshold $u = 2.425$ for the stock JBH. The top panel displays a time series of the daily closing asset price s_t. The daily closing asset prices are then transformed to daily losses on the log scale r_t excluding weekends and non-trading weekdays aggregated and the time series is displayed in the middle panel. The plot shows clear signs of intense bursts of loss volatility. Next, the excess loss above the threshold u given that the log-loss r_t exceeds the threshold is computed. The bottom panel shows the times of exceedances and size of excesses w_t. The presence of substantial temporal clustering of extremes is evident and generally occur near periods of significant losses.

6.3 Model and methodologies

6.3.1 Marked renewal Hakwes process

Let the arrival times of exceedances be denoted by $\{\tau_i\}_{i\geqslant 1} \subset \mathbb{R}^+$, $\tau_i < \tau_{i+1}$ and denote the associated excesses by $\{w_i\}_{i\geqslant 1}$. Let $N(t)$ be a simple point process

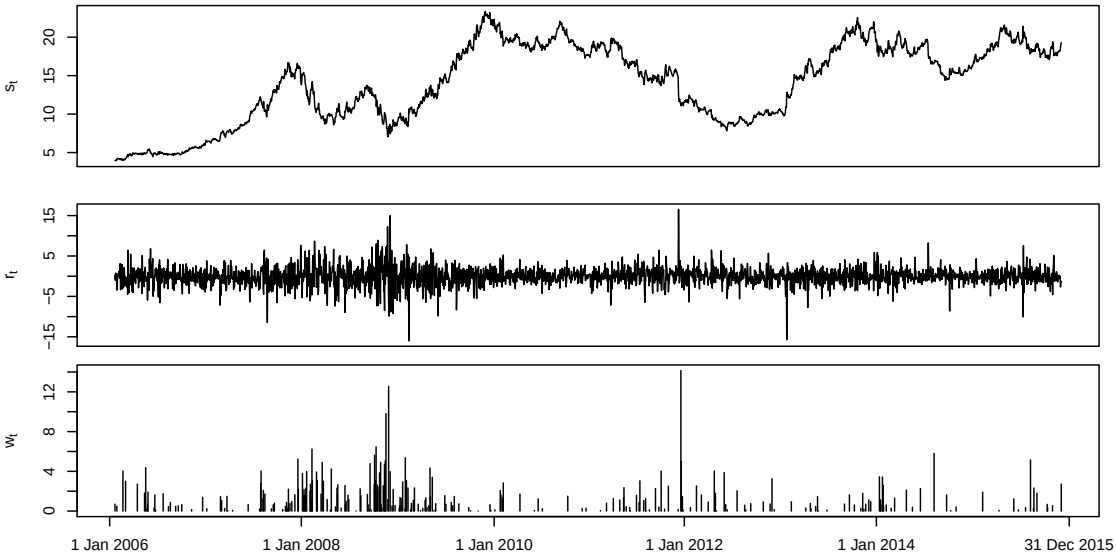

Figure 6.2.1: JB Hi-Fi Limited (JBH) stock data from 1 January 2006 to 31 of December 2015. Top panel: time series plot of the daily closing stock price. Middle panel: time series plot of the negative daily log returns. Bottom panel: time of exceedances and size of excesses over the threshold $u = 2.425$.

on $\mathbb{R}^+$ that counts the number of exceedances by time t. There are two types of exceedance events, namely exogenously and endogenously driven ones. The type is identified by a further (unobservable) mark $M_i \in \{0, 1\}$, where $M_i = 0$ indicates the arrival of an exogenously driven exceedance (immigrant) and $M_i = 1$ indicates an endogenously driven exceedance (offspring event). Furthermore, let $I(t) := \max\{i|, \tau_i < t, M_i = 0\}$ denote the index of the most recent immigrant, with the convention that $I(t) := 0$ when $t < \tau_1$ and $\tau_0 := 0$. This is identical to the conventions that were introduced for the RHawkes process introduced in Chapter 2.

The exceedance times and loss excesses will be modeled using the marked RHawkes process, in which, the waiting times between successive exogenously driven exceedances are assumed to be *i.i.d.* and arrive according to a general renewal process, and the exciting mechanism among the events is the same as in the classical marked Hawkes process model. That is, the ground intensity process $\lambda(t), t \geqslant 0$ relative to the enlarged filtration $\tilde{\mathcal{F}}_t = \sigma\left\{N(s), w_{1:N(s)}, I(s); s \leqslant t\right\}$, $t \geqslant 0$ takes the form,

$$\lambda(t) = \frac{\mathbb{E}\left[\mathrm{d}N(t)|\tilde{\mathcal{F}}_{t-}\right]}{\mathrm{d}t} = \mu(t - \tau_{I(t)}) + \sum_{j=1}^{N(t-)} \eta h(t - \tau_j) g(w_j) \tag{6.3.1}$$
$$=: \mu(t - \tau_{I(t)}) + \phi(t).$$

The function $\mu : \mathbb{R}^+ \to \mathbb{R}^+$ is the hazard rate function of the waiting times between successive exogenously driven exceedances. For the stability of the process, it is required that $\int_0^\infty e^{-\int_0^t \mu(s)\mathrm{d}s}\mathrm{d}t < \infty$, which ensures the expected waiting time between

successive immigrants is finite. The mark's influence on the conditional intensity is governed by the impact function $g : \mathbb{R}^+ \to \mathbb{R}^+$. The constant $\eta \geqslant 0$ is a normalizing constant, and for stability, it is required that $\eta \mathbb{E}[g(w_i)] < 1$ so that the expected number of children of an event is less than one. If the impact function is normalized so that $\mathbb{E}[g(w_i)] = 1$, the parameter $\eta \in [0, 1)$ has the interpretation of a branching ratio. The function $h : \mathbb{R}^+ \to \mathbb{R}^+$ is the offspring density function. As in Chapter 3, the process $\phi(t)$ describes the total excitation effect of past events on current event intensity.

From the intensity process specification in (6.3.1), observe that the background intensity depends on when the most recent immigrant arrives, and it resets to the function $\mu(\cdot)$ upon the arrival of an immigrant. When the hazard function $\mu(\cdot)$ is a constant, the waiting times are exponentially distributed, and the immigrants arrive according to a Poisson process, and therefore the model reduces to the marked Hawkes process in Chavez-Demoulin et al. (2005) and Chavez-Demoulin and McGill (2012). However, in general, the marked RHawkes process is substantially more flexible than the marked Hawkes process because the event counts of the marked RHawkes process in regular time intervals can be over- or under-dispersed relative to the Poisson process, while the counts in a marked Hawkes process can only be over-dispersed.

The conditional intensity function in (6.3.1) is the time-intensity and only describes the dynamics of the ground process. It does not account for the distribution of the marks. For full specification of the intensity process of the marked point process, one also need to specify the distribution of the event mark given an event happens at a certain time t and all the information before time t. In this chapter, a conditional independence assumption is imposed, so that the mark w_i is independent of the event time τ_i conditional on the previous exceedance times $\tau_{1:i-1} := (\tau_1, \ldots, \tau_{i-1})$ and excesses $w_{1:i-1} := (w_1, \ldots, w_{i-1})$. This conditional independence assumption makes parametric methods for modeling marks simple to implement. In this instance, optimization of the log-likelihood function can be divided into two separate optimization problems, and therefore the MLEs for parameters in the ground process model and the mark distribution can be separately.

6.3.2 Likelihood evaluation algorithm

This section develops an algorithm to compute the likelihood function of the marked RHawkes process. The likelihood function can be represented as a product of the conditional joint densities of the event time and mark, conditional on all previous

event times and marks as follows,

$$L(\theta|\tau_{1:n}, w_{1:n}) = \\ p_\theta(\tau_1, w_1) \left\{ \prod_{i=2}^{n} p_\theta(\tau_i, w_i|\tau_{1:i-1}, w_{1:i-1}) \right\} \mathbb{P}_\theta \left(\tau_{n+1} > T|\tau_{1:n}, w_{1:n}\right), \quad (6.3.2)$$

In what succeeds, the subscript θ in p_θ and $\mathbb{P}_\theta$ is dropped for notational convenience, while the dependence of the relevant densities and probabilities on the parameter θ is silently understood. The conditional independence assumption allows the log-likelihood function in (6.3.2) to be divided into two separate components, and inferences can be conducted independently for the temporal patterns of exceedances and the loss excesses. The log-likelihood function then takes the form,

$$\begin{aligned} & l(\tau_{1:n}, w_{1:n}|\theta) \\ = & \left[\log p(\tau_1) + \sum_{i=2}^{n} \log p(\tau_i|\tau_{1:i-1}, w_{1:i-1}) + \log \mathbb{P}\left(\tau_{n+1} > T|\tau_{1:n}, w_{1:n}\right) \right] \quad (6.3.3) \\ & + \left[\log p(w_1) + \sum_{i=2}^{n} \log p(w_i|\tau_{1:i-1}, w_{1:i-1}) \right] \\ =: & \, l_\tau + l_w, \quad (6.3.4) \end{aligned}$$

where l_τ and l_w denotes the temporal component and the mark component of the log-likelihood respectively. In the next section, the modeling of the marks and the evaluation of l_w will be discussed. The remainder of this section is devoted to the evaluation of l_τ.

The ground intensity function $\lambda(t)$ depends on the index of the most recent immigrant and hence to compute the conditional densities required in (6.3.3), the distribution of the most recent immigrant is required. By defining the following terms,

$$d_{ij} := p\left(\tau_i|\tau_{1:i-1}, w_{1:i-1}, I(\tau_i) = j\right), \quad (6.3.5)$$

$$S_{n+1,j} := \mathbb{P}\left(\tau_{n+1} > T|\tau_{1:n}, w_{1:n}, I(\tau_{n+1}) = j\right), \quad (6.3.6)$$

$$p_{ij} := \mathbb{P}\left(I(\tau_i) = j|\tau_{1:i-1}, w_{1:i-1}\right), \quad (6.3.7)$$

and by conditioning on the index of the most recent immigrant, the following holds,

$$p\left(\tau_i|\tau_{1:i-1}, w_{1:i-1}\right) = \sum_{j=1}^{i-1} d_{ij}p_{ij}, \qquad i = 2, \ldots, n, \tag{6.3.8}$$

$$\mathbb{P}\left(\tau_{n+1} > T|\tau_{1:n}, w_{1:n}\right) = \sum_{j=1}^{n} S_{n+1,j}p_{n+1,j}. \tag{6.3.9}$$

The functions $U(t)$, $H(t)$ and $\Phi(t)$ that were defined prior to Theorem 3.3.1 will be used throughout this chapter except that $\Phi(t) = \int_0^t \phi(s)\mathrm{d}s = \eta \sum_{j=1}^{N(t-)} H(t - \tau_j)g(w_j)$ now includes a contribution from the marks. The process assumes that the first event is an immigrant with event time density $p(\tau_1) = e^{-U(\tau_1)}\mu(\tau_1)$. The conditional densities and survival probabilities in (6.3.5) and (6.3.6) are computed using,

$$d_{ij} = e^{-\{U(\tau_i-\tau_j)-U(\tau_{i-1}-\tau_j)\}-\{\Phi(\tau_i)-\Phi(\tau_{i-1})\}}\left(\mu(\tau_i - \tau_j) + \phi(\tau_i)\right), \tag{6.3.10}$$

$$S_{n+1,j} = e^{-\{U(T-\tau_j)-U(\tau_n-\tau_j)\}-\{\Phi(T)-\Phi(\tau_n)\}}. \tag{6.3.11}$$

Next, the conditional probabilities p_{ij} in (6.3.7) are computed using the following forward recursion with initial conditions $p_{21} = 1$ and $p(\tau_2|\tau_1, w_1) = d_{21}$,

$$p_{ij} = \begin{cases} \frac{\phi(\tau_{i-1})}{\mu(\tau_{i-1}-\tau_j)+\phi(\tau_{i-1})} \frac{d_{i-1,j}p_{i-1,j}}{p(\tau_{i-1}|\tau_{1:i-2}, w_{1:i-2})}, & j = 1, \ldots, i-2, \\ 1 - \sum_{k=1}^{i-2} p_{ik}, & j = i-1, \end{cases} \tag{6.3.12}$$

for $i = 3, \ldots, n+1$. The derivation of (6.3.12) is similar to the derivation used in Chapter 3.

The direct evaluation of the likelihood is now practically feasible at a given parameter vector θ. To evaluate the conditional densities $p(\tau_i|\tau_{1:i-1}, w_{1:i-1})$ and the most recent immigrant probabilities p_{ij}, the bivariate recursion given in (6.3.8) and (3.3.14) is implemented as well as computing d_{ij} given by (6.3.10). The survival probability $\mathbb{P}(\tau_{n+1} > T|\tau_{1:n}, w_{1:n})$ is computed using (6.3.9), (6.3.11), and the $p_{n+1,j}$. The above terms are then substituted into the first pair of square brackets in (6.3.3), to calculate the part of the log-likelihood needed for the estimation of parameters of the ground process model, that is, l_τ.

6.3.3 Excess modeling

The generalized Pareto distribution (GPD) is commonly used in EVT and has been applied to model excesses of extreme negative returns (e.g. Chavez-Demoulin et al., 2005), and shall also be used in this chapter. The use of GPD is justified by the

following result from EVT (cf. Embrechts et al., 1997, Theorem 3.4.5): If the random variable X has a distribution function $F(\cdot)$ belonging to the maximum domain of attraction of the standard generalized extreme value distribution $H_\xi(\cdot)$,

$$H_\xi(x) = \begin{cases} 1 - \exp\left\{-(1+\xi x)^{-1/\xi}\right\}, & \xi \neq 0; \\ 1 - \exp\left\{-e^{-x}\right\}, & \xi = 0, \end{cases}$$

then there exists a positive function $a(\cdot)$ such that,

$$\mathbb{P}(X - u \leqslant x | X > u) \to G_{\xi,a(u)}(x), \quad \text{as } u \to x_F,$$

where $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$, and $G_{\xi,\sigma}$ is the *generalized Pareto distribution* function with shape parameter ξ and scale parameter σ, defined by,

$$G_{\xi,\sigma}(x) = \begin{cases} 1 - (1 + \xi x/\sigma)_+^{-1/\xi}, & \xi \neq 0; \\ 1 - \exp(-x/\sigma), & \xi = 0. \end{cases}$$

Following the work of Chavez-Demoulin et al. (2005), the loss excesses are modeled using generalized Pareto distributions with a common shape parameter and scale parameters evolving according to a first-order Markov process. More specifically,

$$w_i | w_{1:i-1} \sim G_{\xi,a+bw_{i-1}}, \tag{6.3.13}$$

with parameters $\xi > 0$, $a > 0$ and $b > 0$. For stability, it is also require that $\xi + b < 1$. It can then be shown using Theorem 3.2 in Cline and Pu (2002) that the Markov process (6.3.13) is geometrically ergodic. These parameters are then estimated by maximizing the part of the log-likelihood in the second pair of brackets in (6.3.3), that is,

$$l_w = \sum_{i=1}^{n} \left\{ -(\frac{1}{\xi} + 1) \log\left(1 + \frac{\xi w_i}{a + bw_{i-1}}\right) - \log(a + bw_{i-1}) \right\}, \tag{6.3.14}$$

where $w_0 := a/(1 - \xi - b)$ is set to be the mean of the loss excesses.

6.3.4 Model assessment

Two aspects of the model are considered in the goodness-of-fit assessment; the temporal patterns of exceedances and the distribution of excesses. For the former, a similar procedure to Chapter 3 is applied using the Rosenblatt (1952) residuals. The basis of the method is to transform the exceedance times using the Rosenblatt (1952) transformation to produce residuals which should be independent and uni-

formly distributed on the unit interval when the model is correctly specified. The residuals are given by, $U_1 = \hat{F}_1(\tau_1) = 1 - e^{-\hat{U}(\tau_1)}$ and

$$U_i = \hat{F}_i(\tau_i|\tau_{1:i-1}, w_{1:i-1}) = 1 - \sum_{j=1}^{i-1} \hat{p}_{ij}\hat{S}_{ij}, \quad i = 2, \ldots, n, \tag{6.3.15}$$

where $\hat{F}_i(t|\tau_{1:i-1}, w_{1:i-1})$ is the estimated conditional distribution function of τ_i conditional on $\tau_{1:i-1}$ and $w_{1:i-1}$, $\hat{p}_{ij}$ are the estimated most recent immigrant probabilities in (6.3.12) and $\hat{S}_{ij}$ are given by,

$$\hat{S}_{ij} = e^{-\left\{\hat{U}(\tau_i - \tau_j) - \hat{U}(\tau_{i-1} - \tau_j)\right\} - \left\{\hat{\Phi}(\tau_i) - \hat{\Phi}(\tau_{i-1})\right\}}, \quad j = 1, \ldots, i-1.$$

Note that $\hat{U}(t)$ and $\hat{\Phi}(t)$ are the plug-in estimates of the cumulative hazard function $U(t)$ and cumulative excitation effect $\Phi(t)$ defined on Page 105 above Eq. (6.3.10).

For a consistent approach to the goodness-of-fit assessment of the model, the Rosenblatt (1952) transformation is also applied to the excesses. This chapter only considers marks with the GPD model in (6.3.13), although the procedure is general enough to apply to most choices of marked distributions. In this instance, the residuals are given by,

$$V_i = \hat{G}_i(w_i|\tau_{1:i-1}, w_{1:i-1}) = 1 - \left[1 + \frac{\hat{\xi} w_i}{\hat{a} + \hat{b} w_{i-1}}\right]^{-1/\hat{\xi}}, \tag{6.3.16}$$

where $\hat{G}_i(w|\tau_{1:i-1}, w_{1:i-1})$ is the estimated conditional distribution function of w_i conditional on $\tau_{1:i-1}$ and $w_{1:i-1}$. However, for the GPD model considered in this chapter, the conditional distribution function $\hat{G}_i(\cdot|\tau_{1:i-1}, w_{1:i-1})$ only depends on w_{i-1}.

The two residual series $\{U_i\}$ and $\{V_i\}$ then serve as the basis for assessing the model's ability to model the temporal patterns of exceedances and the distribution of excesses. Both residual series should be approximately *i.i.d.* uniformly on (0,1) if both aspects of the model are adequate. The uniformity and independence can be assessed graphical with techniques such as uniform Q-Q plot, and the ACF plot, or more formal statistical tests, such as the K-S test and the Ljung-Box (L-B) test respectively.

6.3.5 Predicting exceedances

Predictions using point process models often rely on simulations as explicit algorithms are generally not available (Daley and Vere-Jones, 2003, pp. 274). The distribution of quantities of interests such as the time until the next exceedance or

the number of exceedances in a given time interval can be extracted from predictive simulations. The algorithm works by sequentially simulating event times until the censoring time, as in Section 6.4.1. This procedure requires a computable expression for the hazard function (or cumulative hazard function) and hence the index of the most recent immigrant before time T must be simulated using the conditional probabilities $p_{n+1,j} = \mathbb{P}\left(I(\tau_{n+1}) = j | \tau_{1:n}, w_{1:n}\right)$ so that $\mu(t - \tau_{I(T)})$ can be computed for all $t > T$. For each realization of the future, one can extract any quantity of interest that one wants to predict, and use its empirical distribution obtained from a large number of realizations as the basis for prediction.

A particularly important prediction for financial stakeholders is the time until the next extreme loss. For this purpose, predictive simulations are not necessary, as direct calculation of the predictive density and hazard function for the waiting time until the next exceedance after the censoring time is readily available, and are given respectively by,

$$p(\tau_{n+1} | \tau_{1:n}, w_{1:n}, \tau_{n+1} > T) = \frac{\sum_{j=1}^{n} p_{n+1,j} d_{n+1,j}}{\mathbb{P}\left(\tau_{n+1} > T | \tau_{1:n}, w_{1:n}\right)}, \qquad \tau_{n+1} > T, \tag{6.3.17}$$

and

$$\text{haz}(\tau_{n+1} | \tau_{1:n}, w_{1:n}, \tau_{n+1} > T) = \frac{\sum_{j=1}^{n} p_{n+1,j} d_{n+1,j}}{\sum_{j=1}^{n} p_{n+1,j} \tilde{S}_{n+1,j}}, \qquad \tau_{n+1} > T, \tag{6.3.18}$$

where the $p_{n+1,j}$'s are calculated using (6.3.12), the denominator in (6.3.17) is computed using (6.3.9), the $d_{n+1,j}$'s are given as in (6.3.10) and

$$\tilde{S}_{n+1,j} = e^{-\{U(\tau_{n+1} - \tau_j) - U(\tau_n - \tau_j)\} - \{\Phi(\tau_{n+1}) - \Phi(\tau_n)\}}.$$

The estimated parameters are then be substituted into (6.3.17) and (6.3.18) to compute the estimated predictive density and hazard function, which assists in making predictions regarding the time until the next exceedance.

6.3.6 Forecasting conditional risk measures

Conditional risk measures such as VaR and ES are significant quantities used by many financial institutions, and as such, a method to estimate their value is of particular importance. The proposed algorithm to evaluate the likelihood of the marked RHawkes process implies a procedure to compute the predictive distribution of the loss excesses conditional on the history of the process by time t, $\mathcal{F}_t = \sigma\left\{N(t), \tau_{1:N(t)}, w_{1:N(t)}\right\}$, and therefore estimates of conditional VaR and ES can be readily obtained. For the remainder of this chapter, the conditional VaR and conditional ES will be referred to as VaR and ES, recognizing that these quantities

are conditioned upon all priorly observed data. For the out-of-sample period, the MLEs obtained from the in-sample period are used to obtain the plug-in predictive loss distribution.

Let R_{t+1} denote the daily log-loss on day $t+1$, then the VaR at level q on day $t+1$ is given by,

$$\mathrm{VaR}_{t+1}^{q} = \inf\left\{r \in \mathbb{R} :\ F_{R_{t+1}|\mathcal{F}_t}(r) \geqslant q\right\}.$$

In financial applications, attention is generally directed to extreme outcomes with the quantile levels $q = 0.95$ or $q = 0.99$ typically used. Now by conditioning on the index of the most recent immigrant before time $t+1$, the survival function for the daily log-loss R_{t+1} becomes,

$$\mathbb{P}\left(R_{t+1} > r|\mathcal{F}_t\right) = \sum_{k=1}^{N(t)} \mathbb{P}\left(R_{t+1} > r|\mathcal{F}_t,\ I(t) = k\right) \mathbb{P}\left(I(t) = k|\mathcal{F}_t\right). \tag{6.3.19}$$

As this analysis is only interested with extreme returns, by conditioning upon the return being greater than the threshold u, the following holds for $r > u$,

$$\begin{aligned}\mathbb{P}\left(R_{t+1} > r|\mathcal{F}_t,\ I(t) = k\right) = \mathbb{P}\left(R_{t+1} - u > r - u|R_{t+1} > u, \mathcal{F}_t, I(t) = k\right) \\ \times\, \mathbb{P}\left(R_{t+1} > u|\mathcal{F}_t, I(t) = k\right).\end{aligned} \tag{6.3.20}$$

A discrete-time process is approximated using a continuous-time process to compute (6.3.20). The second term on the right of (6.3.20) can be approximated using the probability that at least one exceedance event occurs in the interval $(t, t+1]$ conditional on index k being the most recent immigrant. This sort of approximation has previously been applied in the work of Chavez-Demoulin et al. (2005) and Chavez-Demoulin and McGill (2012). In this context, the approximation takes the form,

$$\mathbb{P}\left(R_{t+1} > u|\mathcal{F}_t, I(t) = k\right) \approx 1 - \exp\left(-\int_t^{t+1} \hat{\mu}(s - \tau_k) + \hat{\phi}(s)\mathrm{d}s\right). \tag{6.3.21}$$

The first term on the right hand side of (6.3.20) is computed using the fitted GPD for the excesses, that is,

$$R_{t+1} - u|\mathcal{F}_t;\ R_{t+1} > u \sim G_{\hat{\xi},\hat{a}+\hat{b}w_{N(t)}}.$$

A forecast of the VaR at level q can be obtained by solving,

$$\mathbb{P}\left(R_{t+1} > \mathrm{VaR}_{t+1}^{q}|\mathcal{F}_t\right) = 1 - q,$$

with the solution given by,

$$\widehat{\mathrm{VaR}}_{t+1}^{q} = \frac{\hat{a} + \hat{b} w_{N(t)}}{\hat{\xi}} \left[\left(\frac{C_{t+1}}{1-q} \right)^{\hat{\xi}} - 1 \right] + u, \tag{6.3.22}$$

where

$$C_{t+1} = \sum_{k=1}^{N(t)} \left[1 - \exp\left\{ - \int_{t}^{t+1} \hat{\mu}(s - \tau_k) + \hat{\phi}(s) \mathrm{d}s \right\} \right] \hat{p}_{N(t)+1,k},$$

is an approximation to the probability that the return on day $t+1$ is greater than the threshold u, i.e. $\mathbb{P}\left(R_t > u | \mathcal{F}_t\right)$. The expression in (6.3.22) is only valid when $C_{t+1}/(1-q) > 1$, or more elegantly put $q \geqslant \mathbb{P}\left(R_{t+1} < u | \mathcal{F}_t\right)$. When this does not hold, a conservative approach is applied and the VaR estimate to defined to be equal the threshold value u.

Although the VaR is a beneficial tool for measuring risk, it does not indicate the size of an extreme loss. This deficiency has led to the consideration of alternative risk measures. One such alternative is the ES, which is an attractive risk measure as it provides a measure of the size of the loss given that it exceeds the VaR level. The conditional ES for day $t+1$ at quantile level q is defined as follows,

$$\mathrm{ES}_{t+1}^{q} = \frac{\int_{q}^{1} \mathrm{VaR}_{t+1}^{\alpha} \mathrm{d}\alpha}{1-q}.$$

Based on this definition, a forecast of the conditional ES on day $t+1$ is given by (cf. Chavez-Demoulin and McGill, 2012),

$$\widehat{\mathrm{ES}}_{t+1}^{q} = \frac{\hat{\mathrm{VaR}}_{t+1}^{q}}{1-\hat{\xi}} + \frac{\hat{a} + \hat{b} w_{N(t)} - u\hat{\xi}}{1-\hat{\xi}}. \tag{6.3.23}$$

6.4 Simulation study

6.4.1 Simulation algorithm

The algorithm to simulate the process requires a sequential approach as the event marks might be autocorrelated. The algorithm works as follows. First, simulate the initial immigrant arrival time according to the specified inter-renewal distribution. Then each event is simulated by first simulating the corresponding waiting time since the last event according to an appropriate hazard function, and then simulating the event type (immigrant or offspring) according to an appropriate Bernoulli distribution, and finally simulating the event mark according to previously simulated events marks (and event times, depending on model specification) and the specified dependence structure. Events are simulated sequentially until the next simulated

event time exceeds the censoring time T. The realization of the marked RHawkes process consists of the time-mark pairs corresponding to all the simulated events by the censoring time. Note that the event types are not retained.

6.4.2 Simulation model

The rest of this section reports numerical evidence of the finite sample performance of the MLEs in a simulation study. The models chosen to perform the simulations are motivated by the model choices in Section 6.5 and consist of gamma inter-renewal waiting times with hazard function,

$$\mu(t) = \frac{1}{\Gamma(t/\beta, \kappa)\beta^{\kappa}} t^{\kappa-1} e^{-t/\beta}, \tag{6.4.1}$$

where κ is the shape parameter, β is the scale parameter and $\Gamma(x, k) = \int_x^{\infty} s^{k-1}e^{-s}\mathrm{d}s$ is the upper incomplete gamma function. The offspring density is exponential $h(t) = e^{-t/\gamma}/\gamma$ with mean waiting time parameter γ. The event marks are conditionally generalised Pareto distributed, and follows the first order Markov process (6.3.13) with parameters ξ, a and b. The impact function $g(w)$ is the normalized version of the affine function $1 + \delta w$, that is,

$$g(w) = \frac{1 + \delta w}{\mathbb{E}\left[1 + \delta w_i\right]} = \frac{1 + \delta w}{1 + \delta a/(1 - \xi - b)}. \tag{6.4.2}$$

The simulations consist of 1000 realizations of the marked RHawkes process up to a predetermined censoring time T for a variety of parameter values specified in Table 6.4.1. The censoring time T is determined so that the expected numbers of events by T are approximately 500 and 1000 respectively. For each realization, the parameters of the mark distribution ξ, a and b are estimated by directly minimizing $-l_w$, and the parameters of the RHawkes process κ, β, γ, δ and η are estimated by directly minimizing $-l_\tau$.

6.4.3 Results

All computations were performed on Intel Xeon X5675 processors (12M cache, 3.06 GHz, 6.4GT/S QPI) using the R language (R Core Team, 2016). Likelihood maximization was performed by direct calls to the R function `optim`. As the log-likelihood function is relatively flat along the parameters ξ and δ, a reparametrization $\theta = e^{\theta'}$ improves convergence speed and estimation accuracy. Table 6.4.1 reports the estimation results, which contains the true value for each parameter (True), the mean of the 1000 parameter estimates (Est), the empirical standard error of each estimator (SE), i.e. the standard deviation of the 1000 estimates, the mean of the 1000 stan-

dard error estimates by inverting the approximate Hessian matrix ($\hat{\text{SE}}$), the mean squared error (MSE), the censoring time (T), the mean number of events (ML) and the mean running time (RT) to perform the optimization procedure and compute the Hessian matrix. Due to the heavy-tailed nature of the generalized Pareto distribution, substantial mark values occasionally occur. The finite and small sample size permits these extreme mark values to have a substantial influence on some parameter estimates, especially on the estimates of δ. Therefore in summarizing the estimation results in Table 6.4.1, the extreme estimates are trimmed by removing a percentage (2.5% when the mean number of events is 500 and 1% when the mean number of events is 1000) of the smallest and largest estimates for each parameter.

The bias and standard errors are decreasing as the censoring time increases in the majority of the model scenarios considered. The standard errors are also decreasing and approximately at a rate of $1/\sqrt{T}$ as the standard errors reduce by a factor of approximately $1/\sqrt{2}$ when the censoring time T is doubled. The mean of the standard errors are relatively comparable with the empirical standard errors as they agree in most cases, and again, this improves as the censoring time increases. For the majority of the parameters, the MSE is reasonably close to zero with the only notable exceptions being γ and δ. However, the MSE drops substantially as the censoring time increases, e.g., the MSE for δ decreases from 5.996 to 0.029 in the first simulation model considered and most decrease by a factor of two.

The estimated mean waiting time $\hat{\gamma}$ between an event and its direct offspring and the estimated impact parameters for the marks $\hat{\delta}$ lead to reasonably large standard errors when compared to the standard errors of the other parameter estimates. However, the standard errors of $\hat{\gamma}$ and $\hat{\delta}$ are still shrinking as the censoring time gets longer. The comparably significant standard errors for the mean offspring waiting time is also evident in the simulation study conducted in Chapters 3 for the RHawkes process. The estimates for γ and impact function parameter δ have significantly less bias when the immigrant arrivals exhibit heavy clustering ($\kappa = 1/2$) compared to when the immigrants exhibit more evenly distributed arrival times ($\kappa = 2$). The parameter γ is well estimated when the branching ratio η is large, i.e., when the level of self-excitation is high, as the expected number of offspring events present in those realizations is larger. The significant bias evident in the estimation of the parameter δ relates to the heavy-tailed nature inherent in the generalized Pareto distribution, and a reasonably large sample would be needed to reduce the bias to a reasonable level. The branching ratio parameter η is generally adequately estimated with minimal bias for the variety of censoring times, level of self-excitation, and variability of immigrant inter-event waiting times. This leads to the conclusion that the MLE obtained by directly maximizing the log-likelihood function has satisfactory finite sample performances.

	Immigration		Offspring		Mark			Impact
	κ	β	γ	η	ξ	a	b	δ
True	2	0.5	1	0.3	0.1	1	0.1	0.1
Est.	2.039	0.495	1.138	0.289	0.094	1.009	0.098	0.366
SE	0.328	0.059	0.627	0.069	0.044	0.076	0.045	2.434
$\widehat{\text{SE}}$	0.359	0.067	0.588	0.074	0.050	0.087	0.049	4.695
MSE.	0.109	0.004	0.413	0.005	0.002	0.006	0.002	5.996
	RT = 325.89 secs		$T = 350$		$ML = 498.71$			
Est.	2.023	0.500	1.081	0.296	0.097	1.005	0.099	0.146
SE	0.241	0.048	0.440	0.051	0.034	0.058	0.033	0.162
$\widehat{\text{SE}}$	0.254	0.049	0.509	0.055	0.035	0.061	0.035	0.195
MSE.	0.058	0.002	0.200	0.003	0.001	0.003	0.001	0.029
	RT = 1031.71 secs		$T = 700$		$ML = 996.75$			
True	2	0.5	1	0.7	0.1	1	0.1	0.1
Est.	2.263	0.475	1.030	0.674	0.092	1.005	0.097	0.155
SE	1.045	0.151	0.260	0.072	0.043	0.074	0.041	0.170
$\widehat{\text{SE}}$	0.899	0.155	0.278	0.074	0.050	0.087	0.050	0.204
MSE.	1.161	0.024	0.068	0.006	0.002	0.006	0.002	0.032
	RT = 330.55 secs		$T = 150$		$ML = 494.77$			
Est.	2.099	0.496	1.008	0.683	0.098	1.003	0.099	0.123
SE	0.689	0.124	0.183	0.054	0.032	0.057	0.034	0.109
$\widehat{\text{SE}}$	0.618	0.122	0.189	0.053	0.035	0.061	0.035	0.117
MSE.	0.484	0.015	0.034	0.003	0.001	0.003	0.001	0.012
	RT = 1055.37 secs		$T = 300$		$ML = 990.62$			
True	0.5	2	1	0.3	0.1	1	0.1	0.1
Est.	0.506	2.014	1.026	0.299	0.094	1.011	0.097	0.259
SE	0.038	0.298	0.352	0.061	0.043	0.073	0.041	0.438
$\widehat{\text{SE}}$	0.042	0.337	0.378	0.070	0.050	0.087	0.049	0.599
MSE.	0.001	0.089	0.124	0.004	0.002	0.005	0.002	0.217
	RT = 369.06 secs		$T = 350$		$ML = 500.03$			
Est.	0.503	2.007	1.016	0.298	0.096	1.005	0.098	0.184
SE	0.029	0.219	0.253	0.047	0.033	0.058	0.033	0.246
$\widehat{\text{SE}}$	0.029	0.234	0.268	0.049	0.035	0.061	0.035	0.280
MSE.	0.001	0.048	0.064	0.002	0.001	0.003	0.001	0.068
	RT = 1227.40 secs		$T = 700$		$ML = 998.89$			
True	0.5	2	1	0.7	0.1	1	0.1	0.1
Est.	0.533	1.858	0.997	0.671	0.093	1.014	0.095	0.157
SE	0.076	0.503	0.200	0.065	0.043	0.075	0.044	0.179
$\widehat{\text{SE}}$	0.083	0.578	0.223	0.074	0.051	0.089	0.050	0.230
MSE.	0.007	0.274	0.040	0.005	0.002	0.006	0.002	0.035
	RT = 348.17 secs		$T = 150$		$ML = 490.36$			
Est.	0.516	1.941	1.001	0.687	0.095	1.006	0.099	0.129
SE	0.053	0.409	0.143	0.048	0.032	0.058	0.034	0.123
$\widehat{\text{SE}}$	0.056	0.434	0.151	0.051	0.035	0.061	0.035	0.138
MSE.	0.003	0.171	0.021	0.002	0.001	0.003	0.001	0.016
	RT = 1140.52 secs		$T = 300$		$ML = 996.05$			

Table 6.4.1: Results of the maximum likelihood estimation based on 1000 simulated datasets of the marked RHawkes process with gamma distributed inter-immigration waiting times, exponential offspring densities, normalized affine linear impact function, and event marks following the one-step Markov model in (6.3.13).

Figure 6.4.1 displays the normal Q-Q plots for the estimated parameters in the last simulation model considered in Table 6.4.1. For most parameters, the standard

normal quantiles and the empirical quantiles of the estimator align reasonably well. However, as the actual value for $\delta = 0.1$ is very close to the lower bound 0, and because of the significant variance in the estimator for δ, the empirical distribution for the estimator of δ is heavily skewed to the right, causing severe deviation of the quantile points from the Q-Q line in the Q-Q plot for $\hat{\delta}$.

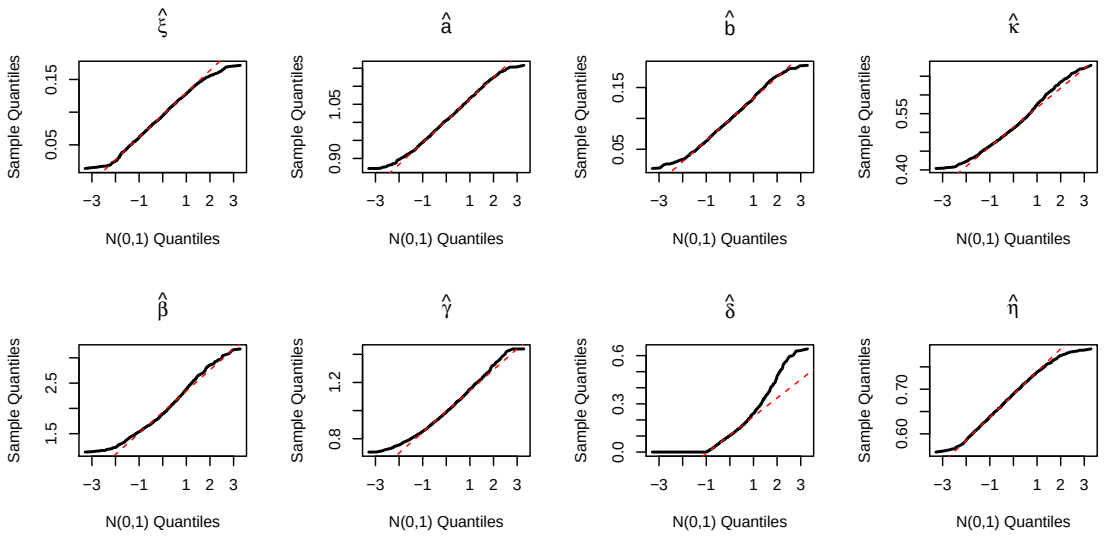

Figure 6.4.1: Normal Q-Q plots for the estimated parameters in the case where $\theta = (\kappa, \beta, \gamma, \delta, \eta) = (0.5, 2, 1, 0.1, 0.7)$ and a mean numbers of events close to 1000.

6.5 Modeling extreme negative returns

The versatility of the marked RHawkes process for modeling extreme negative returns will be illustrated on five stocks traded on the ASX introduced in Section 6.2. The marked RHawkes process model was fit to the data with three different choices of inter-renewal distributions; exponential, gamma, and Weibull. Recall that the exponential inter-renewal distribution is equivalent to the competitor approach based on the classical marked Hawkes process for which the marked RHawkes process will be backtested. The gamma model has an inter-renewal hazard function given by (6.4.1) and the Weibull model with shape parameter κ and scale parameter β has hazard function given in (3.6.2). For all models, a normalized affine impact function as in (6.4.2) is used, where δ reflects the strength of the excesses on the ground intensity process. The offspring density is selected to be exponential $h(t) = e^{-t/\gamma}/\gamma$ where γ represents the mean waiting time between an exceedance event and any exceedance events directly excited by it. The estimated parameters for each model on the ASX stocks were found by directly optimizing the log-likelihoods and the standard errors by inverting the Hessian matrix. The Rosenblatt residuals for goodness-of-fit assessment were also calculated.

Following Chavez-Demoulin et al. (2005), the loss excesses follow the order-1 Markov process model in (6.3.13). The contribution to the likelihood in (6.3.3) based only on the excesses is optimized first. The results are reported in Table 6.5.1, which reports the estimates, the standard errors (in parentheses), and the p-values of the K-S tests and L-B tests on the Rosenblatt residuals. The significant p-values suggest the model for the excesses is adequate for most of the stocks, except in the case of the JHX and CPU stocks, where there is still significant serial correlation among the residuals. A higher-order Markov process might be considered for these stocks. The estimated parameter $\hat{b}$ is positive for all the stocks, suggesting that an excessively large loss is likely to be followed by another substantial loss, although the result is only significant for the stocks JBH and CPU. The estimated shape parameter $\hat{\xi}$ is positive for all the stocks, although it is statistically significant only for the stock DOW, suggesting the loss excesses on this stock is heavier-tailed than on the other stocks, which agrees with the kurtosis statistics shown in Table 6.2.1.

	$\hat{\xi}$	$\hat{a}$	$\hat{b}$	K–S	L–B
JBH	0.113 (0.0672)	1.213 (0.152)	0.133 (0.0675)	0.7720	0.3189
ABC	0.118 (0.0706)	1.124 (0.137)	0.0883 (0.0697)	0.7737	0.4681
CPU	0.129 (0.0687)	0.748 (0.108)	0.286 (0.0883)	0.8115	0.0058
DOW	0.337 (0.0755)	1.192 (0.134)	0.0071 (0.0351)	0.4579	0.0649
JHX	0.0976 (0.0720)	1.089 (0.128)	0.122 (0.0657)	0.9627	2.656×10^{-7}

Table 6.5.1: Results of the maximum likelihood estimation of excesses using the GPD with common shape parameter ξ and scale parameters evolving according to a first order Markov process $\sigma_j = a + bw_{j-1}$ for each five ASX stocks.

Table 6.5.2 contains the estimates of the model parameters for the temporal component, their standard errors (in parentheses), the mean waiting time between exogenously driven exceedances (WT) and the p-values of the K-S and L-B tests on the residuals calculated using the in-sample data. The marked RHawkes process with gamma or Weibull inter-renewal distributions fits to the data better than the Hawkes process model (the exponential case), with uniformity of residuals for the RHawkes process models passing the K-S test at the 5% level on all five stocks. However, the residuals for the Hawkes process are always smaller and mostly fail at the 5% level. Furthermore, the results on the L-B tests of independence of residuals for the Hawkes and RHawkes processes are very similar, except for the stock DOW. The goodness-of-fit results of the gamma and Weibull RHawkes processes are similar, although the fit by the gamma RHawkes model is slightly better. Figure 6.5.1 displays the uniform Q-Q plots and the ACF plots of the Rosenblatt residuals U_i (6.3.15) for the gamma RHawkes model fitted to the different stocks. Graphically, the uniform Q-Q plots indicate good agreement between empirical and theoretical quantiles, although

they manage to depart in the upper quantiles slightly but remain within the 95% confidence intervals. The autocorrelations among the residuals are mostly negligible as seen in the ACF plots, and this is confirmed by the large p-values of the L-B tests for the majority of the stocks.

	$\hat{\kappa}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\delta}$	$\hat{\eta}$	WT	K-S	L-B
JBH								
Exponential	$\hat{\mu} = 22.67$ (5.73)		42.82 (12.67)	1.643 (3.118)	0.564 (0.115)	22.67	0.049	0.015
Weibull	1.544 (0.336)	28.07 (6.10)	29.30 (9.47)	0.694 (0.712)	0.609 (0.0906)	25.25	0.146	0.019
Gamma	1.983 (0.906)	12.44 (4.28)	29.87 (10.28)	0.772 (0.908)	0.599 (0.101)	24.67	0.189	0.019
ABC								
Exponential	$\hat{\mu} = 23.37$ (6.33)		51.84 (16.97)	0.722 (0.870)	0.575 (0.120)	23.37	0.008	0.185
Weibull	1.384 (0.220)	24.50 (4.85)	35.37 (12.71)	0.571 (0.548)	0.555 (0.0911)	22.37	0.044	0.187
Gamma	1.606 (0.374)	13.21 (3.26)	37.37 (13.96)	0.700 (0.733)	0.530 (0.0948)	21.21	0.064	0.177
CPU								
Exponential	$\hat{\mu} = 27.30$ (8.82)		41.77 (15.22)	0.170 (0.250)	0.639 (0.123)	27.30	0.067	0.340
Weibull	1.322 (0.242)	29.74 (8.19)	32.61 (10.86)	0.152 (0.201)	0.639 (0.103)	27.38	0.116	0.314
Gamma	1.704 (0.616)	15.66 (4.94)	32.13 (10.77)	0.160 (0.208)	0.630 (0.106)	26.68	0.131	0.314
DOW								
Exponential	$\hat{\mu} = 19.90$ (4.29)		20.44 (6.91)	0.110 (0.137)	0.498 (0.109)	19.90	0.074	0.287
Weibull	1.768 (0.578)	30.49 (8.57)	15.49 (4.46)	0.0844 (0.0787)	0.633 (0.108)	27.14	0.230	0.139
Gamma	7.129 (4.409)	4.95 (2.60)	17.51 (4.10)	0.0791 (0.0671)	0.719 (0.0656)	35.26	0.244	0.071
JHX								
Exponential	$\hat{\mu} = 21.82$ (6.40)		47.92 (15.61)	0.800 (1.168)	0.549 (0.139)	21.82	0.019	0.332
Weibull	1.374 (0.234)	25.39 (6.65)	36.19 (10.77)	0.552 (0.613)	0.575 (0.113)	23.21	0.065	0.345
Gamma	1.705 (0.513)	12.76 (2.98)	36.12 (11.10)	0.686 (0.856)	0.546 (0.122)	21.75	0.109	0.355

Table 6.5.2: Results of the maximum likelihood estimation of the marked RHawkes process with exponential (mean μ), Weibull and gamma (shape κ and scale β) distributed inter-immigration waiting times, exponential offspring densities (mean γ) and linear impact function (δ describes the strength of the excesses) and branching ratio η for each of the five ASX stocks.

The additional versatility of the RHawkes process introduces one extra parameter, and to conduct an appropriate comparison between the RHawkes process, and the less flexible Hawkes process, the Akaike information criterion (AIC) for each model on each of the five stocks is also computed and reported in Table 6.5.3. For each stock, the Weibull and gamma renewal models outperform the Hawkes model. The AIC for the gamma and Weibull models are again comparable. Therefore, by observing the goodness-of-fit test results and the AIC values, to correctly model the temporal patterns of exceedances, the more flexible RHawkes model is preferred to the classical Hawkes process. In the sequel, when talking about RHawkes processes, only the case with gamma distributed inter-renewal times will be discussed.

The stock JBH will be used to illustrate the interpretation of the model fit since the results on the remaining stocks and can be understood similarly. The estimated shape parameter of the gamma inter-renewal waiting time, $\hat{\kappa} = 1.983$, suggests that exogenously driven extreme losses occur more evenly through time than suggested by the classical Hawkes process (i.e., $\hat{\kappa} = 1$). The mean waiting time between successive immigrants is $\hat{\kappa}\hat{\beta} = 24.67$ days, which is slightly larger than suggested

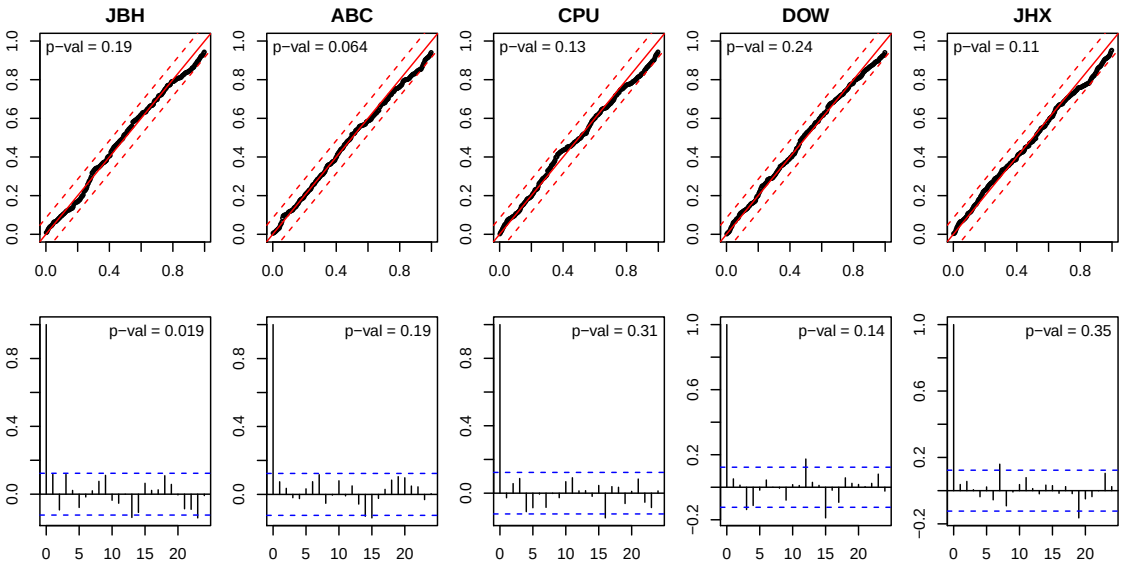

Figure 6.5.1: Graphical goodness-of-fit test of the Rosenblatt residuals for the RHawkes process with gamma distributed inter-immigration waiting times for each of the five ASX stocks. The top panels are the uniform Q-Q plots and the lower panels are the ACF plots.

by the Hawkes model (22.67 days). The mean waiting time from an extreme loss to an extreme loss directly generated by it is 29.87 days, while the Hawkes model suggests this mean waiting time is considerably longer at 42.82 days. Another interesting comparison is between the mean waiting time between exogenously and endogenously driven exceedances suggested by the RHawkes model. By comparing the WT and the estimated $\hat{\gamma}$ values, it is clear that exogenously driven exceedances occur more rapidly than endogenously driven ones, and this is consistent with all but one of the stocks. The value $\hat{\delta} = 0.772$ reflects a moderate impact of the excess values on the propensity for future exceedances. For example, an excess of 2% leads to an increase in propensity contribution from that exceedance by 13.44% while an excess of 5% leads to an increase of 16.7%. The estimated $\hat{\delta}$ value decreases when switching from exponential inter-renewals to gamma inter-renewals, and this is consistent for all the stocks. Using gamma inter-renewals thus reduces the impact that the loss excesses has on the intensity for future exceedances as compared to the exponential inter-renewals. The relatively large branching ratio $\hat{\eta} = 0.599$ suggests a high degree of self-excitation, with the model interpreting slightly more exceedances to be endogenous rather than exogenous.

Next, the procedure developed in Section 6.3.6 are used to estimate the risk measures on the five ASX stocks. The performance of the estimation is assessed by backtesting, where the number of VaR exceptions expected by the estimated risk measure is compared to the actual number of VaR exceptions. A VaR exception occurs whenever the actual log-loss on a particular day exceeds the estimated value

Stock	JBH	ABC	CPU	DOW	JHX
Exponential	1613.36	1625.59	1634.97	1644.57	1627.29
Weibull	1609.37	1622.72	1634.06	1640.93	1623.64
Gamma	1609.60	1622.37	1633.00	1640.48	1622.41

Table 6.5.3: Akaike information criterion (AIC) for each of the five ASX stock with exponential, Weibull and gamma distributed inter-renewals.

of the VaR. A large number of exceptions implies that the model is underestimating the risk. Figure 6.5.2 displays the time series of the log-losses and the estimated 95% and 99% VaR based on the RHawkes process and based on the classical Hawkes process for the ASX stocks over the period from 1 January 2012 to 31 December 2016. Here, a more extended period for backtesting was used, which contains part of the in-sample period as well as the entire out-of-sample period, for the comparison between the expected and actual numbers of VaR exceptions to be meaningful.

One method to formally assess how well the VaR estimator performs is to test whether the observed proportion of VaR exceptions $\hat{p}$ agrees with the expected proportion of exceptions $p = 1 - q$, where q is the quantile level used in the VaR calculation. The null hypothesis states that the model correctly forecasts the VaR, while the alternate hypothesis states that the model underestimates the VaR, since financial applications typically give more importance to not underestimating risk. For the stock JBH, the percentage of actual exceptions based on the RHawkes model estimate of the 95% VaR is 4.49%, and is 0.528% based on the 99% VaR estimate. Both these values agree well with the respective expected proportions, with the p-values of the one-side exact binomial tests equal to 0.8345 and 0.9839, respectively. The corresponding p-values on the other four stocks are all much more substantial than 5%, suggesting the RHawkes model-based VaR estimator has satisfactory performance. The Hawkes model-based VaR estimator also passes the test on all five stocks, although typically with smaller p-values. Therefore, the VaR estimators based on the RHawkes and Hawkes models have similar performances.

However, Figure 6.5.2 reveals that the VaR estimate based on the RHawkes model can drop following an exceedance, but the VaR estimate by the Hawkes model always jumps up following an exceedance. The rationalization for this is that the estimated shape parameter $\hat{\kappa}$ of the inter-renewal distribution is larger than one, implying that the estimated hazard function μ given in (6.4.1) is monotonically increasing from 0, and therefore $\mu(t-\tau_k)+\phi(t)$ can drop following an immigrant event, which in turn causes the approximation (6.3.21) used in the calculation of the VaR in (6.3.22) to go down following an event with a small excess, but a high probability of being an immigrant. In contrast, the hazard function μ in the Hawkes model is a constant, so the intensity $\mu+\phi(t)$ always jumps up when the excitation effect enters

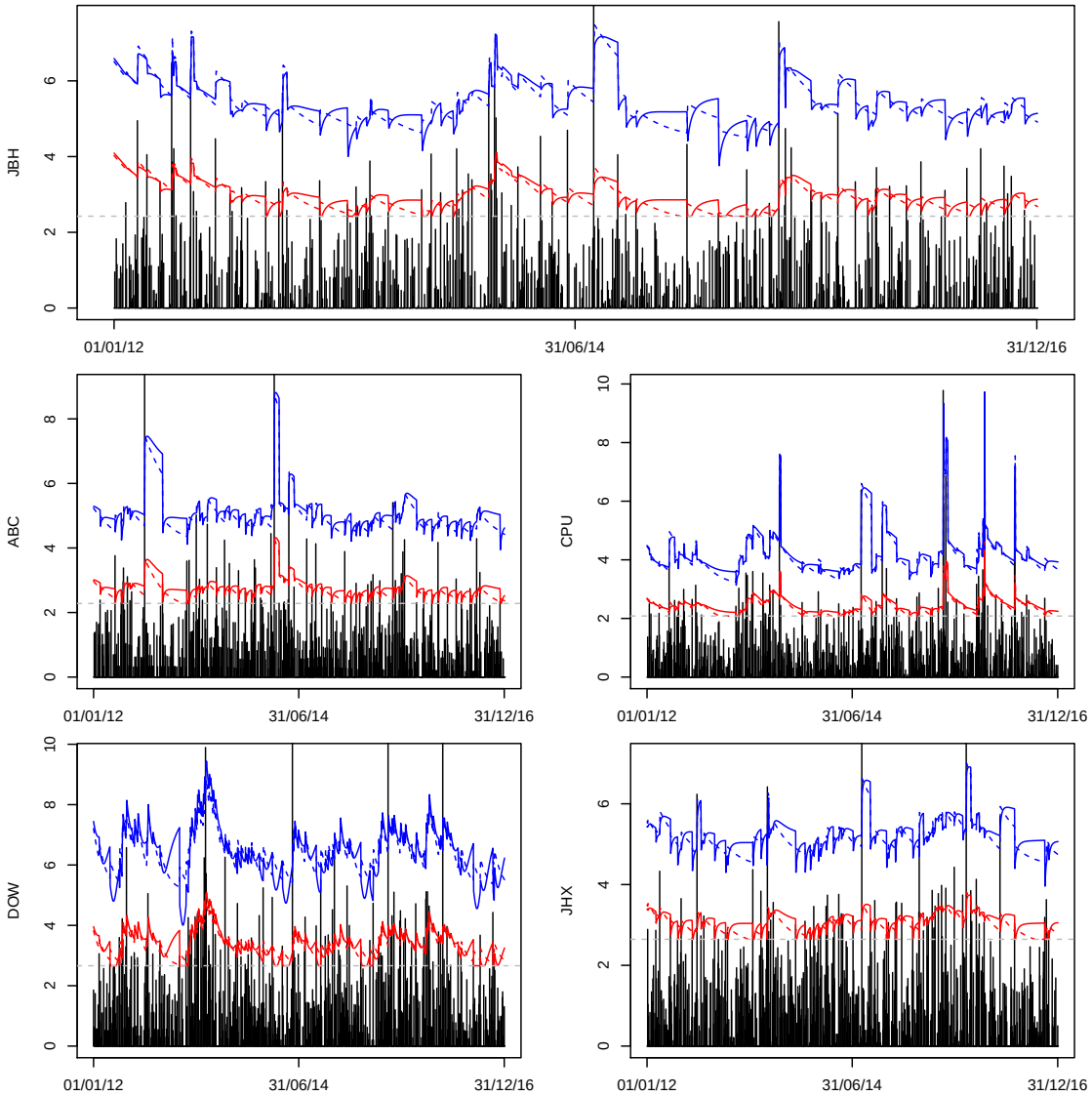

Figure 6.5.2: Time series plot of the daily log-losses for each of the five ASX stocks for the period 1 January 2012 to 31 December 2016 together with the 95% and 99% estimates of VaR. The solid lines are based on estimates for the gamma RHawkes model, and the dashed lines are based on the Hawkes model.

$\phi(t)$ following each exceedance, which causes the VaR estimate in (6.3.22) to jump up. The estimated ES at the 95% level for the five stocks are shown in Figure 6.5.3, from which the estimates using both the RHawkes and the Hawkes models are again similar to each other. However, similar to the VaR estimate, the ES estimate by the RHawkes model can drop momentarily following an exceedance with a small excess, suggesting that the chance and the size of an extreme loss right after a small exceedance can both be smaller than before the exceedance.

Another interesting forecast to make is the number of days until the next extreme loss. The developments in Section 6.3.5 are used to make this forecast for each stock and then compared with actual observations. The waiting time until the next

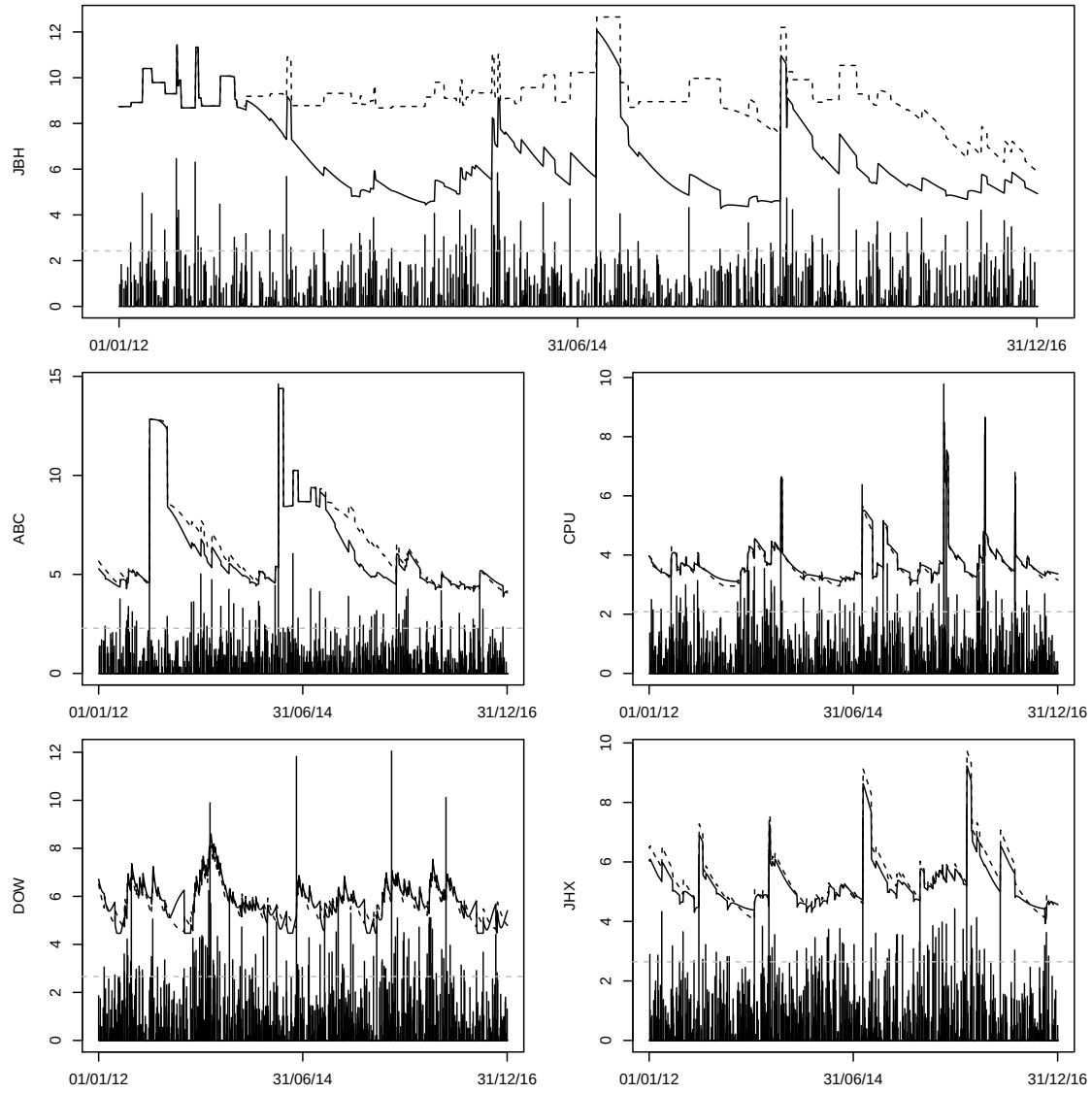

Figure 6.5.3: Time series plot of the daily log-losses for each of the five ASX stocks for the period 1 January 2012 to 31 December 2016 together with the 95% estimates of ES. The solid line is based on estimates from the gamma RHawkes model, and the dashed line is the Hawkes model.

extreme loss is predicted by conditioning upon all priorly available information at the censoring time, and computing the predictive density and hazard function using both the RHawkes and Hawkes models. Figure 6.5.4 plots the predictive densities using solid lines and the dashed lines for the predictive hazard functions. The black lines indicate the RHawkes model, and the grey lines indicate the Hawkes model.

For the stock JBH, the probability that an exceedance occurs in the first, second, third and fourth 10-day period implied by the RHawkes model are 54.12%, 24.53%, 11.29%, and 5.28% respectively, with the actual exceedance occurring during the first ten-day period. The predicted probabilities by the Hawkes model are similar, although the predicted probability of having the first extreme loss in the first 10-day

period is slightly smaller. The predictive hazard function provides an estimate of the hazard or conditional probability of having an extreme loss on a given day conditional on that it has not occurred by the previous day. Observe from Figure 6.5.4 that by the RHawkes model the hazard of seeing an extreme loss on 1 January 2016 is 7.86%, while by the Hawkes model, it is only 7.09% . The predictions for the other four stocks by the two models can be interpreted similarly. Note that the hazard function for the stock DOW increases over time and this is a result of the significant estimated shape parameter $\hat{\kappa} = 7.129$ and a recent exceedance having a high probability of being exogenously driven.

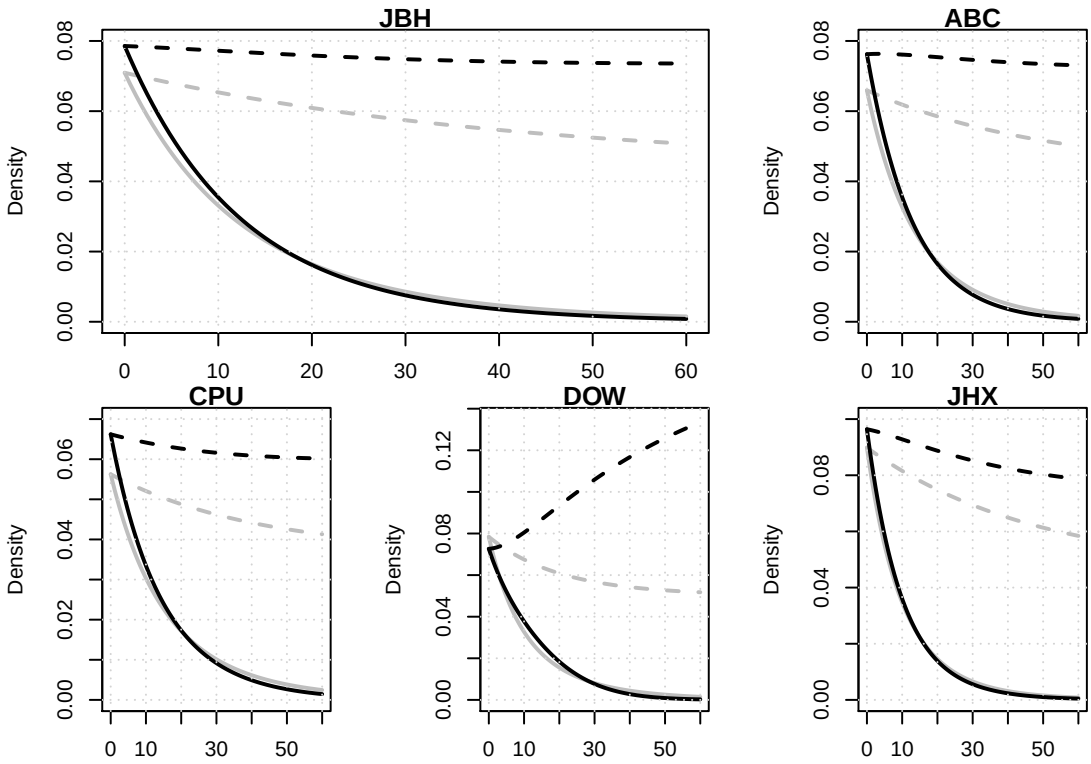

Figure 6.5.4: Solid lines display the predictive densities, and the dashed lines display the predictive hazard functions for the waiting time (in days) until the next exceedance after the censoring time. The black lines indicate the RHawkes model, and the shaded lines indicate the Hawkes model.

The significance of incorporating a renewal distribution for exogenously driven negative returns lies in the flexible properties that the conditional risk measures such as VaR and ES can exhibit while still preserving a straightforward estimation procedure. For instance, these risk measures are permitted to drop following an extreme loss and this is a trait not held by the classical Hawkes process. This suggests that after an extreme loss occurs the chances of observing another large loss could potentially decrease. Furthermore, the RHawkes process can produce a hazard function for the next extreme loss that actually increase over time which suggest that the pressure for the stock price to drop substantially gains momentum from

exogenous factors. Again, this is another property that the classical Hawkes process cannot exhibit. Furthermore, the Hawkes process is unable to adequately fit the five ASX stocks as the p-values for the tests of independence and uniformity mostly fail at the five percent level. Hence, we can conclude that the RHawkes process is the superior model of choice as it is able to successfully pass these tests at the five percent level and provides more flexibility when performing risk quantification.

Chapter 7

Conclusion

7.1 Summary

Stochastic point processes model the temporal patterns of event arrival times. Hawkes processes, as opposed to conventional Poisson processes, allow events in the past to determine the future arrival rate. Since the intensity process of a Hawkes process increases temporarily when an event occurs, a heavier clustering of the event times is achievable as a consequence of the self-exciting mechanism. These modeling capabilities of the Hawkes process were extended by Wheatley et al. (2016), in which they modified the arrival process of immigrants, by allowing the inter-immigrant waiting times to depend on the most recent immigrant arrival time, that is, a renewal immigration process was introduced.

However, Wheatley et al. (2016) insisted that the computation of the likelihood for the RHawkes process demanded exponential computational time and therefore, was practically infeasible on any meaningful datasets. Due to the perceived intractability of finding the MLE directly to make statistical inferences, and relying on the branching process representation of self-exciting process, they implemented two E-M type algorithms to compute the MLE of the model parameters. Furthermore, a bootstrap procedure was used to estimate the variance-covariance matrix of the MLE and a Monte Carlo approach to compute a goodness-of-fit test statistic, but these methods are still computationally expensive.

This thesis contributes by providing superior methods to conduct statistical inferences for the RHawkes process, and overcome the computationally expensive inferential procedures currently available. Since the likelihood function plays a fundamental role in statistical inferences, a practically feasible method for likelihood evaluation is highly desirable. Chapter 2 solved this by describing an efficient algorithm to calculate the likelihood of the RHawkes process in quadratic time, which was a significant improvement from the exponential time claimed by Wheatley et al. (2016). This chapter furthermore discussed methods to simulate the model, perform

goodness-of-fit assessments and predictions, which are efficient and straightforward to implement and illustrated the applicability of the methods discussed to real data from seismology and finance.

Nevertheless, fitting the renewal Hawkes process to data still remains a challenging task, particularly on bigger datasets. Chapter 3 undertook this challenge by developing two approaches that significantly reduce the time required to fit RHawkes processes. Since the derivative-based methods for optimization, in general, converge faster than the derivative-free methods, the first approach derived algorithms to evaluate the gradient and Hessian of the log-likelihood function and then applied the derivative-based Newton-Raphson method in maximizing the likelihood, instead of the derivative-free method used in Chapter 2. The second approach sought linear time algorithms that produced accurate approximations to the likelihood function by truncating the most recent immigration distribution. Simulation experiments showed that the Newton-Raphson method reduced the computational time significantly and in some simulates, halved the computational time, even on moderately large datasets. Furthermore, the approximate likelihood methods that have linear computational time produced comparably accurate estimates. The methods presented therein were readily applicable to data, and as an illustration, an analysis of mid-price changes on several currencies relative to the US Dollar was presented.

The computational efficiency gains from the likelihood evaluation algorithms detailed in Chapters 2 and 3 facilitated the application of these algorithms to the estimation of multivariate and marked point process models with renewal immigration. In Chapter 4, a multivariate extension to the RHawkes process was introduced, in which, different event types interact with self- and cross-excitation effects. A similar recursive algorithm was formulated to directly calculate the likelihood of the model, which established the basis of statistical inferences. Furthermore, to overcome the high computational demands required for estimation in the high dimensional settings, a modified algorithm that reduced the computational time significantly was also discussed. The likelihood evaluation algorithm implied a procedure to assess the goodness-of-fit for both the temporal patterns of the events and the distribution of the event types. The plug-in predictive density function for the next event time and methods to make future predictions using simulations were also addressed. The simulation studies showed that the likelihood evaluation algorithms and the prediction procedures were performing as expected. The proposed methodologies were illustrated on two datasets; the first was earthquakes occurring in two Pacific island countries Fiji and Vanuatu and the second on trade-through data for the stock BNP Paribas on the Euronext Paris stock exchange.

Chapter 6 analyzed extreme financial returns using RHawkes processes with marks. Extreme return financial time series are often challenging to model due to

the presence of heavy temporal clustering of extremes and strong bursts of return volatility. In this chapter, a model for extreme financial returns was introduced, which provided additional flexibility in the specification of the background arrival rate compared to the Hawkes process. The model is a marked version of the renewal Hawkes process discussed in Chapter 2 and 3. A procedure is developed to evaluate the likelihood of the model, which can be optimized to obtain estimates of model parameters and their standard errors. The proposed model was applied to extreme negative returns for five stocks traded on the Australian Securities Exchange. The models identified for the stocks using in-sample data were found to be able to successfully forecast the out-of-sample risk measures such as the value at risk and expected shortfall, and provided a better quality of fit than the competing Hawkes model.

7.2 Perspective on future work

The scope of this thesis has detailed computational aspects to fit, assess, and make predictions for the RHawkes process model. Although inference for the RHawkes model, using the maximum likelihood method, is straightforward to implement, the results concerning the properties of the MLEs are yet to be discussed. This lack of asymptotic theory for maximum likelihood-based inference for RHawkes processes presents an avenue for future work. It would be anticipated that under some stationarity and ergodicity conditions, that the maximum likelihood estimator $\hat{\theta}$ would display consistency and asymptotic normality (as the observation time $T \to \infty$), as suggested by the simulation studies conducted herein. Additionally of interest is the asymptotic distribution of the gaps between successive events. This would help design approximations to the likelihood and have an understanding of the bound of the relative error in the approximations.

In the seismological applications illustrated herein, the proximity of the earthquake epicenters or the more generally, the spatial aspects of earthquake occurrences are either disregarded entirely or accounted for by using an event type indicator for the different geographical regions. This unnecessarily limits the modeling capabilities of the RHawkes process in terms of spatial interactions. A potential extension that has not received attention as yet is a spatio-temporal extension to the RHawkes process. To this end, the intensity would depend on the location of the earthquakes similar to the space-time ETAS (Epidemic Type Aftershock Sequence)) (see, e.g., Ogata and Zhuang (2006)).

An autoregressive-moving-average (ARMA) point process is another recent exciting modification to the Hawkes process and was introduced by Wheatley et al. (2018). The ARMA point process incorporates the classical Hawkes process but is

driven by a Neyman-Scott process with Poisson immigration. The process is a natural analog to the ARMA time series model for integer-values times series. It provides an alternative generalization to the baseline rate of a Hawkes process, but rather than allow the baseline intensity function to renew upon the arrival of an immigrant like in the RHawkes process, the ARMA point process introduces a short-noise type burst at the immigrant times. For instance, consider high-frequency mid-price changes in financial markets, such as in Chapters 3 and 4 herein, the ARMA point process would then allow exogenous shocks and clustering due to multiple mechanisms, such as order splitting and the near-simultaneous independent actions of multiple market participants in response to an exogenous shock. The approach undertaken in this thesis may be applicable in evaluating the likelihood of the ARMA point process, or at least an approximation to the likelihood, to facilitate likelihood inferences rather than the EM algorithm employed by Wheatley et al. (2018).

Another possible direction for future work could be to modulate the hazard function for immigrant arrivals of the renewal Hawkes process so that it depends on both the absolute time t, and on the time since the most recent immigrant $\tau_{I(t)}$. The modulated intensity function would take the form,

$$\lambda(t) = \mu(t - \tau_{I(t)}, t) + \sum_{j=1}^{N(t-)} \eta h(t - \tau_j),$$

where $\mu(\cdot, \cdot)$ is now a hazard function that takes two arguments that relate to the waiting time since the most recent immigrant $t - \tau_{I(t)}$ and the absolute time t. This provides a convenient method to introduce non-stationarity into the renewal process for immigrants. This would provide a flexible specification that accounts for long terms trends in the intensity. For instance, in the data analysis of currency exchange rates and trade-throughs, the modulated renewal RHawkes process might be convenient to handle the natural variation in intra-day trading activities in a single framework rather than applying the two-step procedure of transforming the observed event times, and then fitting a stationary RHawkes process to the transformed times.

Appendix A

Derivatives of the most recent immigrant probabilities in the RHawkes process

A.1 First derivative of the most recent immigrant probabilities

In this appendix, the derivation to compute the derivatives of the most recent immigrant probabilities with respect to the parameter θ using a recursive algorithm is presented. For any fixed $i \in \{3, \ldots, n\}$ and for $j \leqslant i-2$, the most recent immigrant probabilities $p_{ij} = \mathbb{P}\left(I(\tau_i) = j | \tau_{1:i-1}\right)$ take the form,

$$p_{ij} = \frac{\phi(\tau_{i-1})\Psi_{i-1,j}p_{i-1,j}}{\sum_{k=1}^{i-2}\left(\mu(\tau_{i-1}-\tau_k)+\phi(\tau_{i-1})\right)\Psi_{i-1,k}p_{i-1,k}}. \tag{A.1.1}$$

where the notations are as in Chapter 4. The derivatives of p_{ij} are obtained by application of the quotient rule. The derivative of the p_{ij} numerator in (A.1.1) with respect to parameter vector θ of the model is given by,

$$\Psi_{i-1,j}\Big[p_{i-1,j}\partial_\theta\phi(\tau_{i-1}) + \phi(\tau_{i-1})\partial_\theta p_{i-1,j} + \phi(\tau_{i-1})p_{i-1,j}\partial_\theta\psi_{i-1,j}\Big], \tag{A.1.2}$$

and the derivative of the denominator of p_{ij} is

$$\sum_{k=1}^{i-2}\Psi_k(\tau_{i-1})\Big[\left\{\partial_\theta\mu(\tau_{i-1}-\tau_k)+\partial_\theta\phi(\tau_{i-1})\right\}p_{i-1,k} + \left\{\mu(\tau_{i-1}-\tau_k)+\phi(\tau_{i-1})\right\}\left(p_{i-1,k}\partial_\theta\psi_{i-1,k}+\partial_\theta p_{i-1,k}\right)\Big]. \tag{A.1.3}$$

Now by applying the quotient rule with the aid of (A.1.2) and (A.1.3), the following recursion for the derivative of the most recent immigrant probabilities $\partial_\theta p_{ij}$ holds

for $i \in \{3, \ldots, n+1\}$,

$$\partial_\theta p_{ij} = \frac{A}{C} - \frac{B}{C^2}, \qquad j = 1, \ldots, i-2, \tag{A.1.4}$$

$$\partial_\theta p_{i,i-1} = -\sum_{j=1}^{i-2} \partial_\theta p_{ij}, \tag{A.1.5}$$

where

$$A := \Psi_{i-1,j}\Big[p_{i-1,j}\partial_\theta\phi(\tau_{i-1}) + \phi(\tau_{i-1})p_{i-1,j}\partial_\theta\psi_{i-1,j} + \phi(\tau_{i-1})\partial_\theta p_{i-1,j}\Big],$$

$$B := \phi(\tau_{i-1})\Psi_{i-1,j}p_{i-1,j}\sum_{k=1}^{i-2}\Psi_{i-1,k}\Bigg[\left\{\partial_\theta\mu(\tau_{i-1}-\tau_k) + \partial_\theta\phi(\tau_{i-1})\right\}p_{i-1,k}$$

$$+ \left\{\mu(\tau_{i-1}-\tau_k) + \phi(\tau_{i-1})\right\}\left(p_{i-1,k}\partial_\theta\psi_{i-1,k} + \partial_\theta p_{i-1,k}\right)\Bigg],$$

and

$$C := \sum_{k=1}^{i-2}\left(\mu(\tau_{i-1}-\tau_k) + \phi(\tau_{i-1})\right)\Psi_k(\tau_{i-1})p_{i-1,k}.$$

This recursive procedure is initialized with the condition that $\partial_\theta p_{21} = 0$.

A.2 Second derivative of the most recent immigrant probabilities

Now for the second derivative of the most recent immigrant probabilities, the terms in (A.1.4) is differentiated, again using the quotient rule, and then the following is obtained,

$$\partial^2_{\theta\theta^\top} p_{ij} = \frac{\partial_\theta A^\top}{C} - \frac{A\partial_{\theta^\top}C + \partial_\theta B^\top}{C^2} + 2\frac{B\partial_{\theta^\top}C}{C^3}, \quad j = 1, \ldots, i-2, \tag{A.2.1}$$

where

$$\partial_\theta A^\top = \Psi_{i-1,j}\Bigg[\phi(\tau_{i-1})\partial^2_{\theta\theta^\top}p_{i-1,j} + \left(\partial^2_{\theta\theta^\top}\phi(\tau_{i-1})\right)p_{i-1,j} + 2\partial_\theta\phi(\tau_{i-1}) \odot \partial_\theta p_{i-1,j}$$

$$+ 2\left\{\left(\partial_\theta\phi(\tau_{i-1})\right)p_{i-1,j} + \phi(\tau_{i-1})\partial_\theta p_{i-1,j}\right\} \odot \partial_\theta\psi_{i-1,j}$$

$$+ \phi(\tau_{i-1})p_{i-1,j}\left\{\left(\partial_\theta\psi_{i-1,j}\right)^{\otimes 2} + \partial^2_{\theta\theta^\top}\psi_{i-1,j}\right\}\Bigg],$$

$$A\partial_{\theta^\top}C+\partial_\theta B^\top = 2\left\{\Psi_{i-1,j}\left[p_{i-1,j}\partial_\theta\phi(\tau_{i-1})+\phi(\tau_{i-1})p_{i-1,j}\partial_\theta\psi_{i-1,j}+\phi(\tau_{i-1})\partial_\theta p_{i-1,j}\right]\right\}$$
$$\odot\left\{\sum_{k=1}^{i-2}\Psi_{i-1,k}\Big[\left\{\partial_\theta\mu(\tau_{i-1}-\tau_k)+\partial_\theta\phi(\tau_{i-1})\right\}p_{i-1,k}\right.$$
$$\left.+\left\{\mu(\tau_{i-1}-\tau_k)+\phi(\tau_{i-1})\right\}\left(p_{i-1,k}\partial_\theta\psi_{i-1,k}+\partial_\theta p_{i-1,k}\right)\Big]\right\},$$

and

$$B\partial_{\theta^\top}C = \phi(\tau_{i-1})\Psi_{i-1,j}p_{i-1,j}\left\{\sum_{k=1}^{i-2}\Psi_{i-1,k}\Big[\left\{\partial_\theta\mu(\tau_{i-1}-\tau_k)+\partial_\theta\phi(\tau_{i-1})\right\}p_{i-1,k}\right.$$
$$\left.+\left\{\mu(\tau_{i-1}-\tau_k)+\phi(\tau_{i-1})\right\}\left(p_{i-1,k}\partial_\theta\psi_{i-1,k}+\partial_\theta p_{i-1,k}\right)\Big]\right\}^{\otimes 2}.$$

The recursive procedure is again initialized with $\partial^2_{\theta\theta^\top}p_{21}=0$. Then for each consecutive $i\in\{3,\ldots,n+1\}$ compute (A.2.1) for $j\leqslant i-2$, and then when $j=i-1$ simply use,

$$\partial^2_{\theta\theta^\top}p_{i,i-1} = -\sum_{k=1}^{i-2}\partial^2_{\theta\theta^\top}p_{ik}.$$

References

Adamopoulos, L. (1976). Cluster models for earthquakes: Regional comparisons. *Journal of the International Association for Mathematical Geology*, 8(4):463–475.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Bacry, E., Delattre, S., Hoffmann, M., and Muzy, J. (2013). Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Processes and their Applications*, 123(7):2475 – 2499. A Special Issue on the Occasion of the 2013 International Year of Statistics.

Bowsher, C. G. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912.

Brockwell, A. (2007). Universal residuals: A multivariate transformation. *Statistics & probability letters*, 77(14):1473–1478.

Chavez-Demoulin, V., Davison, A. C., and McNeil, A. J. (2005). Estimating value-at-risk: a point process approach. *Quantitative Finance*, 5(2):227–234.

Chavez-Demoulin, V. and McGill, J. (2012). High-frequency financial data modeling using Hawkes processes. *Journal of Banking & Finance*, 36(12):3415 – 3426. Systemic risk, Basel III, global financial stability and regulation.

Chen, F. and Hall, P. (2013). Inference for a non-stationary self-exciting point process with an application in ultra-high frequency financial data modeling. *Journal of Applied Probability*, 50(4):1006–1024.

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.

Chen, F. and Stindl, T. (2018). Direct likelihood evaluation for the renewal Hawkes process. *Journal of Computational and Graphical Statistics*, 27(1):119–131.

Chornoboy, E., Schramm, L., and Karr, A. (1988). Maximum likelihood identification of neural point process systems. *Biological Cybernetics*, 59(4):265–275.

Cline, D. B. and Pu, H. H. (2002). A note on a simple markov bilinear stochastic process. *Statistics & Probability Letters*, 56(3):283 – 288.

Crane, R. and Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105:15649–15653.

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes Volume I: Elementary Theory and Methods*. Springer-Verlag, New York, 2nd edition.

Daley, D. J. and Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes Volume II: General Theory and Structure*. Springer-Verlag, New York, 2nd edition.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Embrechts, P., Liniger, T., and Lin, L. (2011). Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability*, 48:367–378.

Embrechts, P., Mikosch, T., and Kluppelberg, Claudia, . (1997). *Modelling extremal events for insurance and finance*. New York : Springer. Formerly published in series: Applications of mathematics v 34.

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, 66(5):1127–1162.

Errais, E., Giesecke, K., and Goldberg, L. R. (2010). Affine point processes and portfolio credit risk. *SIAM Journal on Financial Mathematics*, 1:642–665.

Filimonov, V. and Sornette, D. (2012). Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5):056108.

Halpin, P. F. (2013). A scalable EM algorithm for hawkes processes. In Millsap, R. E., van der Ark, L. A., Bolt, D. M., and Woods, C. M., editors, *New Developments in Quantitative Psychology: Presentations from the 77th Annual Psychometric Society Meeting*, pages 403–414. Springer New York, New York, NY.

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

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):pp. 493–503.

Herrera, R. and Schipp, B. (2009). Self-exciting extreme value models for stock market crashes. In Schipp, B. and Kräer, W., editors, *Statistical Inference, Econometric Analysis and Matrix Algebra: Festschrift in Honour of Götz Trenkler*, pages 209–231. Physica-Verlag HD, Heidelberg.

Lewis, E. and Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. In *Joint Statistical Meetings*, Miami, Florida.

Lewis, P. A. W. and Shedler, G. S. (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413.

Lewis, T. and Fieller, N. R. J. (1979). A recursive algorithm for null distributions for outliers: I. gamma samples. *Technometrics*, 21(3):371–376.

McNeil, A., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton Series in Finance. Princeton University Press.

McNeil, A. J. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, 7(3):271 – 300. Special issue on Risk Management.

Mina, J. and Xiao, J. Y. (2001). Return to riskmetrics: The evolution of a standard. *RiskMetrics Group*.

Mino, H. (2001). Parameter estimation of the intensity process of self-exciting point processes using the EM algorithm. *IEEE Transactions on Instrumentation & Measurement*, 50(50):658–664.

Mohler, G. O., Short, M. B., Brantingham, P. J., Tita, G. E., and Schoenberg, F. P. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108.

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

Muni Toke, I. and Pomponio, F. (2011). Modelling trades-through in a limited order book using Hawkes processes. *Economics: The Open-Access, Open-Assessment E-Journal*, 6:1–23.

Nelder, J. A. and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313.

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

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

Ogata, Y. and Zhuang, J. (2006). Space–time etas models and an improved extension. *Tectonophysics*, 413(1):13 – 23. Critical Point Theory and Space-Time Pattern Formation in Precursory Seismicity.

Olson, J. F. and Carley, K. M. (2013). Exact and approximate EM estimation of mutually exciting hawkes processes. *Statistical Inference for Stochastic Processes*, 16(1):63–80.

Omori, F. (1894). On the aftershocks of earthquakes. *Journal of the College of Science, Imperial University of Tokyo*, 7:111–200.

Ozaki, T. (1979). Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155.

Papangelou, F. (1972). Integrability of expected increments of point processes and a related random change of scale. *Transactions of the American Mathematical Society*, 165:483–506.

Pomponio, F. and Abergel, F. (2013). Multiple-limit trades: empirical facts and application to lead–lag measures. *Quantitative Finance*, 13(5):783–793.

Porter, M. D. and White, G. (2012). Self-exciting hurdle models for terrorist activity. *The Annals of Applied Statistics*, 6(1):106–124.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472.

Stindl, T. and Chen, F. (2018). Likelihood based inference for the multivariate renewal Hawkes process. *Computational Statistics & Data Analysis*, 123:131 – 145.

Stindl, T. and Chen, F. (2019). Modeling extreme negative returns using marked renewal Hawkes processes. *Extremes*.

Türkyilmaz, K., van Lieshout, M. N. M., and Stein, A. (2013). Comparing the Hawkes and trigger process models for aftershock sequences following the 2005 kashmir earthquake. *Mathematical Geosciences*, 45(2):149–164.

Utsu, T. (1961). A statistical study of the occurrence of aftershocks. *Geophysical Magazine*, 30:521–605.

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

Vere-Jones, D. (1970). Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(1):1–62.

Vere-Jones, D. and Davies, R. B. (1966). A statistical survey of earthquakes in the main seismic region of New Zealand. *New Zealand Journal of Geology and Geophysics*, 9(3):251–284.

Wheatley, S., Filimonov, V., and Sornette, D. (2016). The Hawkes process with renewal immigration & its estimation with an EM algorithm. *Computational Statistics & Data Analysis*, 94:120 – 135.

Wheatley, S., Schatz, M., and Sornette, D. (2018). The ARMA Point Process and its Estimation. *arXiv e-prints*, page arXiv:1806.09948.

Zhuang, J. and Mateu, J. (2019). A semiparametric spatiotemporal hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(3):919–942.