Hawkes processes and some financial applications

Brendon M. Lapham
University of Cape Town

April 14, 2014

Dissertation submitted in fulfilment of the requirements for the degree of Master of Business Science in Actuarial Science

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Declaration

I, Brendon Michael Lapham, hereby declare that the work on which this dissertation is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being or is to be submitted for another degree in this or any other University. I empower the University of Cape Town to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

_____

BM Lapham
April 2014

Abstract

The self-exciting point process, which is now more commonly known as the Hawkes process, is a model for a point process on the real line introduced by Hawkes (1971). The distinguishing feature of such processes is that they allow all past 'events' to affect the intensity function at the current time.

Over the years such processes have been applied in seismology and neurophysiology in particular, and in more recent years there have been significant financial applications. In almost all of these applications, the route used to find the maximum likelihood estimates (MLEs) is direct numerical maximisation (DNM) of the likelihood. An EM algorithm, which makes use of the Poisson cluster process interpretation of the Hawkes process, is an alternative route to the MLEs. This particular EM algorithm has received attention in the literature and has been claimed to have advantages over DNM of the likelihood. We carry out a simulation study for a simple Hawkes process to clarify statements made in the literature about these advantages. For the simple Hawkes process models that we consider, DNM of the likelihood is the preferable route to finding the MLEs.

We then use DNM of the likelihood to fit marked Hawkes process models to South African asset data. These applications to South African data include the modelling of extreme asset returns and the forecasting of conditional value-at-risk (VaR) and expected shortfall (ES). The models investigated include mostly models found in the literature, but also include some variations introduced here. In a backtesting exercise, we compare the conditional VaR and ES forecasts found by using the marked Hawkes process models with those found via some nonstandard stochastic volatility (SV) models. We find that the marked Hawkes process models give mostly competitive forecasts of conditional VaR and ES when compared with the nonstandard SV models.

Acknowledgements

I would like to thank my supervisor, Associate Professor Iain L. MacDonald. Our regular discussions about research, and other matters, are stimulating and have inspired much of the work presented here. I would also like to thank my colleagues in the Actuarial Science Section who have provided me with the time to complete this work.

A special thanks to my family, particularly my parents, who have provided me with unconditional support and encouragement.

The financial assistance of the National Research Foundation (NRF) and Institute of Applied Statistics towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF or the Institute of Applied Statistics.

Contents

Notation and abbreviations

Notation

Symbol	Meaning	Page
$\mathbb{R}$	real line	
$\mathbb{R}_+$	nonnegative numbers	
$\mathcal{H}_t$	history of the (marked) point process on the interval $(-\infty, t)$	6
$\widetilde{\mathcal{H}}_t$	history of the (marked) point process on the interval $[0, t)$	7
$\lambda^\dagger(\cdot\|\mathcal{H})$	complete intensity function	6
$\lambda(\cdot\|\widetilde{\mathcal{H}})$	conditional intensity function	7
$L(\cdot)$	likelihood function	8
$\ell(\cdot)$	log-likelihood function	8
$N(A)$	number of points in A	4
$N(s,t]$	number of points in half-open interval $(s,t]$	5
N_g	ground process of the marked point process N	9
$\boldsymbol{\theta}$	parameter vector	8
$\boldsymbol{\Theta}$	parameter space	8
$\widehat{\boldsymbol{\theta}}$	maximum likelihood estimate of $\boldsymbol{\theta}$	20
$F(\cdot)$	is the general symbol for a distribution function. A subscript is included when it is necessary to clarify which random variable the distribution function relates to.	10
$\boldsymbol{\Gamma}$	transition probability matrix	80

Abbreviations

BCBS	Basel Committee on Banking Supervision
CDLL	complete data log-likelihood
DNM	direct numerical maximisation
etc.	*et cetera*
e.g.	*exempli gratia*
ES	expected shortfall
EVT	extreme value theory
GPD	generalised Pareto distribution
HMM	hidden Markov model
i.e.	*id est*
iid	independent, identically distributed
mag.	magnitude
MLE	maximum likelihood estimate
ODLL	observed data log-likelihood
POT	peaks-over-threshold
SV	stochastic volatility
VaR	value-at-risk

CHAPTER 1

Introduction

This dissertation is concerned with the application of univariate Hawkes processes, in particular linear marked univariate Hawkes processes, on the nonnegative real line. The intensity representation of Hawkes processes is the principal approach adopted, but we do also consider the Poisson cluster process representation, as it facilitates estimation of Hawkes process models via an EM algorithm.

The main application that we investigate is the modelling of extreme asset returns; this is presented in Chapter 7. The preceding chapters present mainly the underlying theory used in this application. The exception is Chapter 3, where we discuss maximum likelihood estimation via an EM algorithm and via direct numerical maximisation (DNM) of the log-likelihood. Chapter 3 contains a simulation study and an application to some earthquake data. The simulation study, which involves estimating parameters from simulated data via an EM algorithm, an approximate EM algorithm, and DNM of the log-likelihood, validates the estimation methods, and attempts to clarify statements made in the literature about the advantages of the EM algorithm as applied to Hawkes processes. The application to the earthquake data is intended to be illustrative.

The application of marked Hawkes processes to extreme asset returns is not new, and our application adds to the growing literature on such applications. Marked Hawkes process models have been shown to be effective at modelling extreme asset returns; for example, in the work of Chavez-Demoulin *et al.* (2005), Chavez-Demoulin and McGill (2012), and Herrera (2013). We investigate applications of marked Hawkes process models to

extreme asset returns from several South African assets.

In our application, we compare the effectiveness of marked Hawkes process models and some nonstandard stochastic volatility (SV) models in forecasting conditional market risk measures, namely conditional value-at-risk (VaR) and expected shortfall (ES). The objectives of this application are to investigate whether models based on marked Hawkes processes are suitable for modelling extreme returns from South African assets, and to investigate whether such models produce competitive forecasts of conditional risk measures when compared to those found by using some nonstandard SV models. The SV models that we consider are models proposed by Langrock *et al.* (2012), who demonstrate that the models can produce suitable forecasts of conditional VaR. In addition to these objectives, we also consider here a wide range of marked Hawkes process models. Most of the models that we use appear in the literature, but we also introduce some specialisations and generalisations of these models. The intention is to identify the model, or general form of model, which appears to perform best.

In the final chapter we provide some overall remarks and suggest areas for further work.

CHAPTER 2

Introduction to Hawkes processes

The self-exciting point process was originally introduced by Hawkes (1971) and is now commonly referred to as the Hawkes process. It is a point process model for point events on the real line which display overdispersion relative to the Poisson process, and it comes close to fulfilling the role that autoregressive models fill for time series (Daley and Vere-Jones, 2003, pp. 180, 183). In this chapter we give a brief introduction to the probability theory of marked Hawkes processes. The organisation of this chapter is as follows. In Section 2.1 we introduce unmarked and marked point processes. Then in Section 2.2, we introduce unmarked and marked Hawkes processes. The description of unmarked Hawkes processes includes the Poisson cluster process interpretation which we make use of in Chapter 3.

The material presented in this chapter is largely based on that of Daley and Vere-Jones (2003), who provide a thorough treatment of point processes and their general theory.

2.1 Point processes on the nonnegative real line

An unmarked point process on the nonnegative real line, where the nonnegative line is taken to represent time, is a random process whose realisations consist of the times $t_1, t_2, \ldots$ of point events scattered along the line. A realisation of such a process can be illustrated by using a time line as in Figure 2.1. Each of the t_is in Figure 2.1 is the time of a point event of interest; for example, the time of an earthquake in seismology applications or the time of an extreme asset return in financial applications. A point process models

the time epochs $\{t_i\}$, where i is in some suitable index set, for example $\mathbb{N}$. We will primarily be concerned with point processes on the nonnegative real line, i.e. $t_i \in \mathbb{R}_+$, but occasionally we will depart from this. We will also assume that the point events are ordered, i.e. $t_i < t_{i+1}$.

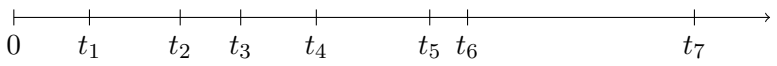


Figure 2.1: *Realisation of an unmarked point process, where t_i gives the time of the ith point event.*

In addition to the times of the point events, there may be additional variables that are of interest associated with each point event. For example, the magnitudes of the earthquakes, or the magnitudes of the extreme financial returns, may be of interest. The recorded realisations of the extended processes, i.e. the timing of the point events and their magnitudes, would consist of the points $(t_1, m_1), (t_2, m_2), \ldots$. The m_is give the magnitudes of the earthquakes, or the extreme returns, and are referred to as marks, and such processes are called marked point processes. We formally introduce marked point processes in Section 2.1.2.

2.1.1 Unmarked point processes

The stochastic process and the sample paths of a point process may be represented in several ways. The counts of point events in subsets of the nonnegative real line is one way. We define the counting measure N such that for a subset A of the nonnegative real line, the number of point events in that subset is given by $N(A)$. More precisely,

$$\begin{aligned} N(A) &= \text{number of indices } i \text{ for which } t_i \text{ lies in } A \\ &= \#\{i : t_i \in A\}. \end{aligned} \tag{2.1}$$

If the set A is expressed as the union of a finite number of disjoint sets $A_1, A_2, \ldots, A_k$, that is

$$A = \bigcup_{i=1}^{k} A_i, \quad \text{where } A_i \cap A_j = \emptyset \text{ for } i \neq j,$$

then it follows from Equation (2.1) that

$$N(A) = \sum_{i=1}^{k} N(A_i).$$

If A is the half-open interval $(u, v]$ for $0 < u < v$, for example, we write $N((u, v])$ for the count of points in that interval and immediately abbreviate this by $N(u, v]$. Similarly, $N(t) = N((0, t])$ for all $0 < t$, and $N(\mathrm{d}t) = N(t, t + \mathrm{d}t]$, where $\mathrm{d}t$ is positive and small. The counting measures $N(u, v]$ and $N(t)$ are nonnegative integer-valued random variables, and $N(t)$ is a nondecreasing function of t. The possibility of 'too many' points is excluded by requiring that $N(A)$ be finite for any bounded set A. The count $N(u, v]$ can be written as

$$\begin{aligned} N(u, v] &= \int_{(u,v]} N(\mathrm{d}s) \\ &= \sum_{j:t_j \in (u,v]} 1, \end{aligned}$$

where $N(\mathrm{d}s)$ has unit value when there is a point in the infinitesimal interval $(s, s + \mathrm{d}s]$ and is zero otherwise. More generally,

$$\int_{(u,v]} g(s)\, N(\mathrm{d}s) = \sum_{j:t_j \in (u,v]} g(t_j),$$

where in general $g : \mathbb{R}_+ \mapsto \mathbb{R}$.

Another method of representing the sample path of a point process on the nonnegative real line is by the durations between consecutive point events. The times of the observed point events $t_1, t_2, \ldots, t_n$ can be used to find the lengths of the intervals between consecutive points as

$$r_i = t_i - t_{i-1} \quad i = 1, 2, \ldots, n,$$

where $t_0 = 0$. The r_is are referred to as inter-arrival times, and given t_0 and a set of the ordered inter-arrival times $r_1, r_2, \ldots, r_n$, the times of the point events can be recovered.

For the history of the point process, we define $\boldsymbol{\mathcal{H}}$ to be a family of nested, increasing σ-algebras $\mathcal{H}_t$, which give the entire history of the point process prior to time t, i.e. $\mathcal{H}_t$ specifies the times of all point events in the interval $(-\infty, t)$. This history may include point events in the distant past which are

not observed. We refer to $\mathcal{H}_t$ as the complete history to distinguish it from the observed history (defined below). We assume that a suitable probability space exists such that $N(A)$, r_i, and t_i, are well-defined random variables.

For a point process with counting measure N, the complete intensity is defined as

$$\lambda^\dagger(t|\mathcal{H}_t) = \lim_{h\to 0^+} h^{-1} \Pr\{N[t, t+h) > 0|\mathcal{H}_t\}.$$

We assume that the limit exists for all t and for all possible $\mathcal{H}_t$. The complete intensity provides an important way of specifying a point process on the real line (Cox and Isham, 1980, p. 66). As an example, consider the archetypal point process, the homogeneous Poisson process. The homogeneous Poisson process has a constant complete intensity $\lambda > 0$, and can be defined for all t and $h \to 0^+$ by

$$\Pr\{N[t, t+h) = 1|\mathcal{H}_t\} = \lambda h + o(h), \tag{2.2}$$

$$\Pr\{N[t, t+h) > 1|\mathcal{H}_t\} = o(h), \tag{2.3}$$

whereby

$$\Pr\{N[t, t+h) = 0|\mathcal{H}_t\} = 1 - \lambda h + o(h). \tag{2.4}$$

This is the 'intensity specification' of the Poisson process. The simplicity of the homogeneous Poisson process belies its importance; it fulfils a role for point processes similar to that of the normal distribution for random variables (Cox and Isham, 1980, p. 45). An array of results can be proved for the homogeneous Poisson process; see Section 3.1 of Cox and Isham (1980) and Chapter 2 of Daley and Vere-Jones (2003). Two specific properties that we use in later chapters are that the inter-arrival times are iid exponential random variables with mean $1/\lambda$, and that, given the number of point events in an interval, those point events are independently and uniformly distributed over the interval.

A (general) point process N on the real line, for which (2.3) is true, is referred to as orderly. This can be thought of as effectively not allowing multiple occurrences at the same time point (Cox and Isham, 1980, pp. 3–4, 25). In the applications that we consider, the observed (marked) point processes mostly do not have multiple occurrences in the time domain as a result of the way that they are observed, and the (marked) Hawkes process models we use to model them are orderly.

Definitions of stationarity can be given for point processes in a similar manner to the definitions for other stochastic processes. A point process on the nonnegative real line is said to be stationary if, for every $k = 1, 2, \ldots$ and all bounded subsets of the nonnegative real line $A_1, A_2, \ldots, A_k$, the joint distribution of

$$\{N(A_1 + t), N(A_2 + t), \ldots, N(A_k + t)\} \tag{2.5}$$

is independent of $t > 0$. It is usually unnecessary to assume a point process is stationary when investigating the properties of that process; assuming that the particular property is stationary is usually sufficient (Cox and Isham, 1980, p. 24). Definitions for stationarity of the inter-arrival times can also be given; see, for example, Definition 3.2.II. of Daley and Vere-Jones (2003, p. 45).

The conditional intensity of a point process is defined as

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \lim_{h \to 0^+} h^{-1} \Pr\{N[t, t+h) > 0|\widetilde{\mathcal{H}}_t\}. \tag{2.6}$$

We assume that the limit exists for all t and for all possible $\widetilde{\mathcal{H}}_t$. This definition is similar to that for the complete intensity, but here the conditioning involves the 'observed' history of the process over the interval $[0, t)$, i.e. the history consistent with an observation on the process. To make the distinction clear we use $\widetilde{\mathcal{H}}_t$ for the observed history. The conditional intensity is a nonnegative piecewise continuous function which is taken to be left-continuous at any discontinuities. Intuitively, the conditional intensity at t gives the conditional 'risk' of a point event occurring at that instant in time, given the observed history of the process prior to time t. In general, the complete and conditional intensities can be functions of time, the history of the point process and, more generally, other external variables or processes.

Daley and Vere-Jones (2003, pp. 211–212, 229) note that for point processes described as having an evolutionary character, their conditional intensities and likelihoods are relatively simple. A point process is said to have evolutionary character if its conditional intensity can be expressed in terms of the observed history of the process, and more generally, in terms of the observed histories of other external variables or processes. The evolutionary

character of such point processes allows the likelihood to be found by successively conditioning on the past. Explicitly, the likelihood of a realisation $t_1, \ldots, t_{N(T)}$ over the finite interval $[0, T]$ of such a (regular) point process is given by

$$L(\boldsymbol{\theta}) = \left[\prod_{i=1}^{N(T)} \lambda(t_i | \widetilde{\mathcal{H}}_{t_i}) \right] \exp\left(- \int_0^T \lambda(u | \widetilde{\mathcal{H}}_u) \, \mathrm{d}u \right), \tag{2.7}$$

where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is the vector of parameters for $\lambda(\cdot | \widetilde{\mathcal{H}})$ (Rubin, 1972; Daley and Vere-Jones, 2003, Proposition 7.2.III., pp. 232–233). See Rubin (1972) and Daley and Vere-Jones (2003, pp. 213, 229–233) for derivations of the likelihood and the associated regularity conditions. As usual, the likelihood is treated as a function of the parameter vector and the point process realisation is taken as given. The notation $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ will be used for the log-likelihood.

Loosely, the factor in the square brackets on the right-hand side of (2.7) is the contribution to the likelihood from observing point events at the times $t_1, \ldots, t_{N(T)}$, and the second factor, which is an amalgam of conditional survivor probabilities, is the contribution from not observing point events at all of the intervening times in $[0, T]$. The conditional survivor probability for t_{i+1}, the time to the $(i+1)$st point event, given the observed history $\widetilde{\mathcal{H}}_t$ and $N(t) = i$, is given by

$$\Pr\{u < t_{i+1} | \widetilde{\mathcal{H}}_t, N(t) = i\} = \exp\left(- \int_t^u \lambda(s | \widetilde{\mathcal{H}}_t) \, \mathrm{d}s \right), \tag{2.8}$$

where $u > t$. This conditional survivor probability is the conditional probability of observing the next point event after time u given the observed history up to time t. It can be related to the counting measure as follows:

$$\Pr\{u < t_{i+1} | \widetilde{\mathcal{H}}_t, N(t) = i\} = \Pr\{N(t, u] = 0 | \widetilde{\mathcal{H}}_t, N(t) = i\}.$$

2.1.2 Marked point processes

A marked point process is a point process with a random variable or vector of random variables attached to each point (Cox and Isham, 1980, p. 15). Figure 2.2 depicts a realisation of a marked point process. Each of the t_is in Figure 2.2 has an m_i, called a mark, associated with it, and the realisation

of a marked point process is a collection of points $(t_1, m_1), (t_2, m_2), \ldots$ on the space $\mathbb{R}_+ \times \mathcal{M}$. The mark space is $\mathcal{M}$, and we will consider nonnegative marks in our applications, i.e. $\mathcal{M} = \mathbb{R}_+$. The definitions of the complete and observed histories, $\mathcal{H}_t$ and $\widetilde{\mathcal{H}}_t$, are extended for marked point processes to record both the marks and the times of the point events.

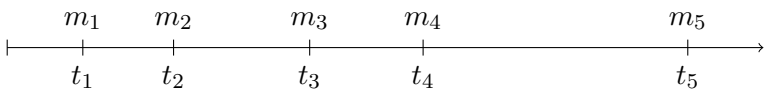


Figure 2.2: *Realisation of a marked point process, where t_i gives the time of the ith point event and m_i is its associated mark.*

The counting process associated with the point events in the time domain only, i.e. the t_is, is referred to as the ground process and the notation N_g is used to identify it. The notation N is used to refer to the marked point process as a whole. More formally, a marked point process N, with point event times in $\mathbb{R}_+$ and marks in $\mathcal{M}$, is a point process $\{(t_i, m_i)\}$ on $\mathbb{R}_+ \times \mathcal{M}$ with the additional property that the process N_g associated with times $t_1, t_2, \ldots$ is itself a point process on $\mathbb{R}_+$.

The generalisation of unmarked point processes to marked point processes subsumes several important point processes. For example, a marked point process can be used to define a point process with multiple occurrences; the marks in this case would give the number of occurrences at each point event. A marked point process may also be used to define a multitype point process with the marks identifying the type of a point event, e.g. $\mathcal{M} = \{1, 2, \ldots, k\}$ for a multitype point process with k types of point event.

A marked point process can be defined by using the so called time-space conditional intensity on $\mathbb{R}_+ \times \mathcal{M}$; see, for example, Daley and Vere-Jones (2003, pp. 249, 254). However, we specify a particular marked point process by defining the conditional intensity $\lambda(\cdot|\widetilde{\mathcal{H}})$ of the ground process N_g, and then, for a given point event and observed history at time t, we define the conditional distribution function for the marks. The conditional intensity of the ground process $\lambda(\cdot|\widetilde{\mathcal{H}})$ will be conditioned on the observed history of the marked point process and not just the observed point event times. We assume that the conditional distribution of the marks can be expressed in terms of the observed history of the marked point process and so it may also

be described as having an evolutionary character. In addition, the marks are assumed to be conditionally independent given the history of the marked point process. The conditional distribution function for the marks will be referred to as the conditional mark distribution, and for a point event at time t, it will be denoted by $F_{M|\widetilde{\mathcal{H}}_t,t}(\cdot)$ which is abbreviated by $F(\cdot)$. For a given marked point process N on $\mathbb{R}_+ \times \mathcal{M}$, the process N is said to have unpredictable marks if the distribution of the mark at t_i is independent of all previous point event times and marks, i.e. the distribution of m_i is independent of $\{(t_j, m_j)\}$ for all $t_j < t_i$.

The likelihood for a realisation $(t_1, m_1), \ldots, (t_{N_g(T)}, m_{N_g(T)})$ over the finite interval $[0, T]$ of such a (regular) marked point process N on $[0, T] \times \mathcal{M}$ is given by

$$L(\boldsymbol{\theta}) = \left[\prod_{i=1}^{N_g(T)} \lambda(t_i|\widetilde{\mathcal{H}}_{t_i})\right] \exp\left(-\int_0^T \lambda(u|\widetilde{\mathcal{H}}_u)\,\mathrm{d}u\right) \left[\prod_{i=1}^{N_g(T)} f(m_i)\right], \quad (2.9)$$

where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ now includes the parameters of the conditional mark density $f(\cdot)$ (Daley and Vere-Jones, 2003, Proposition 7.3.III., p. 251). See Daley and Vere-Jones (2003, pp. 246–256) for a development of the likelihood as well as the associated regularity conditions. The third factor on the right-hand side of (2.9) is the contribution to the likelihood from the observed marks.

2.2 Hawkes processes

In this section marked Hawkes processes are introduced. Maximum likelihood estimation of Hawkes processes is discussed in Chapter 3 and goodness-of-fit tests are discussed in Chapter 4. Simulation of marked Hawkes processes is discussed in Appendix A.1.

2.2.1 Unmarked Hawkes processes

The univariate Hawkes process N with complete intensity $\lambda^\dagger(\cdot|\mathcal{H}_t)$ can be defined for all t and $h \to 0^+$ by

$$\Pr\{N[t, t+h) = 1|\mathcal{H}_t\} = \lambda^\dagger(t|\mathcal{H}_t)h + o(h), \quad (2.10)$$

$$\Pr\{N[t, t+h) > 1|\mathcal{H}_t\} = o(h). \quad (2.11)$$

The complete intensity is defined for all t as

$$\lambda^{\dagger}(t|\mathcal{H}_t) = \tau + \int_{(-\infty,t)} \omega(t-u)\, N(\mathrm{d}u) \tag{2.12}$$

$$= \tau + \sum_{j:t_j<t} \omega(t-t_j), \tag{2.13}$$

where $\tau > 0$, $\omega(s) \geq 0$ for $s \geq 0$ and zero otherwise, and if the Hawkes process is assumed to be stationary, $\int_0^\infty \omega(s)\,\mathrm{d}s < 1$. The complete intensity is a stochastic process and can be thought of as a 'shot-noise process' (cf. Cox and Isham, 1980, p. 74), where all of the past point events can contribute to the current value of the complete intensity (Hawkes, 1971).

This is the original specification of the univariate Hawkes process as given by Hawkes (1971). There have been generalisations in the literature in several different directions; these generalisations include: the nonlinear Hawkes process (Brémaud and Massoulié, 1996; Daley and Vere-Jones, 2003, pp. 252–253), the inclusion of the Hawkes process intensity in a more general intensity to allow for self-excitement (Ogata and Akaike, 1982), the space-time self-exciting point process (Ogata, 1998; Veen and Schoenberg, 2008; Balderama *et al.*, 2012), the dynamic contagion process of Dassios and Zhao (2011), and the Markov-modulated Hawkes process of Wang *et al.* (2012).

The self-exciting nature of the Hawkes process arises via the integral in Equation (2.12). The contribution from a point event at time t_i ($< t$) to the complete intensity at time t is $\omega(t-t_i)$, and all points prior to time t contribute in such a way to the complete intensity at time t. The function $\omega(\cdot)$ governs the effect that past point events have on the intensity, and $\omega(\cdot)$ is often assumed to be a monotonically decreasing function so that the latest point events have the greatest influence on the current value of the intensity. For a monotonically decreasing $\omega(\cdot)$, the intensity will increase immediately after a point event, and as time passes the effect from the point event dies off. As a result, the risk of further point events occurring increases immediately following a point event and this increased risk dies off as time passes. The term 'response function' will be used to refer to $\omega(\cdot)$ in general, and in the case $\omega(\cdot)$ is monotonically decreasing, we will refer to it as a 'decay function'. The function $\omega(\cdot)$ does not have to be monotonically decreasing, as noted by Hawkes (1971). For example, it may be humped so as to allow for delayed effects from past point events on the intensity.

A popular decay function in the literature, and one which was originally used by Hawkes (1971), is the exponential decay function. The exponential decay function has the form

$$\omega(s) = \psi \exp(-\gamma s), \tag{2.14}$$

where $\psi \geq 0$, $\gamma > 0$, and $\psi < \gamma$ if the process is assumed to be stationary. Oakes (1975) identified that the complete intensity is a Markov process when the decay function is exponential. In the case $\omega(s) = \psi \exp(-\gamma s)$, if we write the complete intensity for $t > s$ as

$$\lambda^\dagger(t|\mathcal{H}_t) = \tau \left(1 - e^{-\gamma(t-s)}\right) + e^{-\gamma(t-s)}\lambda^\dagger(s|\mathcal{H}_s) + \psi \sum_{i:t_i \in [s,t)} e^{-\gamma(t-t_i)}, \tag{2.15}$$

it can be seen that, given the value of $\lambda^\dagger(s|\mathcal{H}_s)$, the value of $\lambda^\dagger(t|\mathcal{H}_t)$ can be found by using the times of the point events in $[s, t)$. The pair $(N(t), \lambda^\dagger(t|\mathcal{H}_t))$ also form a Markov process in this case, and such Hawkes processes have been termed Markovian self-exciting point processes by Oakes (1975).

The power(-law) decay function is another popular decay function in the literature, in particular the statistical seismology literature. It has the form

$$\omega(s) = \frac{\psi}{(s+\gamma)^{\eta+1}}, \tag{2.16}$$

where $\psi \geq 0$, $\gamma, \eta > 0$, and $\psi < \eta\gamma^\eta$ if the process is assumed to be stationary. This power decay function has the same form as the modified Omori formula for earthquake aftershock frequency over time. The original Omori formula, which has $\eta = 0$ in (2.16), was suggested by Omori (1894), and (2.16) (without the constraint on the value of η) was referred to as the modified Omori formula by Utsu (1961) (as cited by Utsu *et al.* (1995)). The modified Omori formula is an empirical function which describes how the frequency of earthquake aftershocks evolve over time. The ordinary epidemic type after-shock (ETAS) model, a point process model for earthquake occurrence times in the statistical seismology literature, is a Hawkes process with a power decay function and the impact function $e^{\delta m}$ (defined below). The ETAS model was originally introduced by Ogata (1988) and subsequently has received much attention.

For both the exponential and power decay functions, setting $\psi = 0$ returns us to the homogeneous Poisson process.

Liniger (2009, pp. 32–33) states that if one has no preferences for a particular decay function, then one should use the exponential decay function. The reason given by Liniger is that, for the exponential decay function, there are numerically efficient methods for computing the intensity. This efficiency can be important when the intensity is repeatedly evaluated — which is the case when estimating parameters via DNM of the likelihood.

As an example, consider a Hawkes process with complete intensity

$$\lambda^{\dagger}(t|\mathcal{H}_t) = 0.075 + 0.1 \sum_{i:t_i<t} \exp(-0.2(t - t_i)). \tag{2.17}$$

Figure 2.3 depicts a simulated realisation of the complete intensity and the associated point events for this Hawkes process. The simulated realisation is for the period $[0, 100)$ and is simulated by using Ogata's modified thinning algorithm; see Appendix A.1 and the references there for details. For the simulation, the Hawkes process is assumed to have no point events in the interval $(-\infty, 0)$.

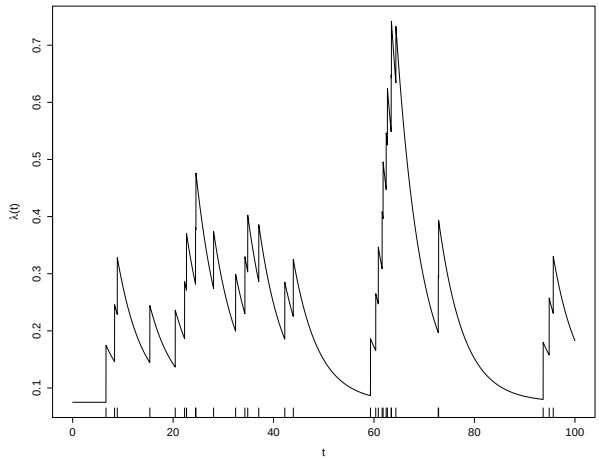


Figure 2.3: *A simulated realisation of the Hawkes process with complete intensity given by* (2.17)*. The curve is the complete intensity and the vertical lines in the lower portion of the panel indicate the times of the simulated point events.*

In the figure, the simulated times of the point events are indicated by the vertical lines in the lower portion of the panel. As there are no point events in the interval $(-\infty, 0)$, the depicted complete intensity initially equals 0.075. The intensity increases immediately after each point event occurs, and this

increases the likelihood of further point events occurring. The rapid decay of the intensity between point events means that the increased likelihood of further point events occurring only lasts for a short period of time. As a result, the simulated point events display visible clustering, e.g. the point events close to time $t = 60$.

The conditional intensity of the Hawkes process introduced in Equations (2.10)–(2.13) is given by

$$\begin{aligned} \lambda(t|\widetilde{\mathcal{H}}_t) &= \tau + \int_{(0,t)} \omega(t-u)\, N(\mathrm{d}u) \\ &= \tau + \sum_{i:t_i \in (0,t)} \omega(t-t_i). \end{aligned} \tag{2.18}$$

Note that the above conditional intensity ignores contributions from point events occurring before time 0. In practice, one rarely observes a point process from its origin and so point events occurring before the start of the observation period may affect the initial conditional intensity; one will need to decide how to treat $\lambda(0|\widetilde{\mathcal{H}}_0)$ and this is discussed in Section 3.2.1. The conditional intensity can be thought of as an approximation to the complete intensity, and in the case $\int_0^\infty \omega(s)\,\mathrm{d}s < 1$, the conditional intensity $\lambda(t|\widetilde{\mathcal{H}}_t)$ approaches $\lambda^\dagger(t|\mathcal{H}_t)$ as $t \to \infty$ (Daley and Vere-Jones, 2003, p. 234).

The Hawkes process has a conditional intensity that is expressed in terms of the past of the process, and so it can be described as having an evolutionary character. As a result, the likelihood function for an unmarked Hawkes process has the same form as given by (2.7).

2.2.2 Poisson cluster process interpretation

An alternative and theoretically important interpretation of the Hawkes process is as a Poisson cluster process. The interpretation as a Poisson cluster process was identified, and its equivalence to the original representation proved, by Hawkes and Oakes (1974). This interpretation of the Hawkes process is described below; the description is based on that presented by Hawkes and Oakes (1974) and Daley and Vere-Jones (2003, pp. 183–184).

The point events of a Hawkes process can be separated into two types; a point event is either an 'immigrant' point event or an 'offspring' point event. The offspring point events are produced by existing point events, and all existing point events can produce offspring point events. The immigrant point

events are the 'cluster centres' and do not have existing 'parent' point events; they arrive according to a homogeneous Poisson process with intensity τ, the baseline rate in the Hawkes process intensity (2.13). Associated with each immigrant point event is a cluster of offspring point events.

The immediate offspring point events produced by an immigrant point event arrive after the immigrant point event according to a nonhomogeneous Poisson process with intensity $\omega(\cdot)$, where $\omega(\cdot)$ is the response function in the Hawkes process intensity (2.13). All of the offspring point events from an immigrant point event, and the offspring of these offspring and so on, have their own offspring point events which also arrive according to nonhomogeneous Poisson processes each with intensity $\omega(\cdot)$. More specifically, consider the (immigrant or offspring) point event at time s, say. This point event triggers the nonhomogeneous Poisson process which generates the offspring point events associated with the point event at time s. The nonhomogeneous Poisson process has intensity $\omega(t-s)$ for $t \geq s$ and mean $\nu = \int_{\mathbb{R}_+} \omega(s)\, \mathrm{d}s$, and the point events it generates are distributed after time s. All of the point events trigger nonhomogeneous Poisson processes in such a manner. Given the point events that trigger them, the nonhomogeneous Poisson processes generating the offspring point events are mutually independent and are independent of the homogeneous Poisson process generating the immigrant point events.

All of the immediate offspring from an immigrant point event, and the offspring of these offspring, and so on, form the cluster of point events associated with the immigrant point event. The observed Hawkes process consists of all of the immigrant point events and their clusters of offspring, i.e. all of the point events.

Figure 2.4 illustrates a fictional realisation of a Hawkes process. The observed Hawkes process consists of the point events on the time axis. The unobserved relationships between the point events of the Hawkes process, also referred to as the branching structure, are illustrated in the upper portion of the figure. Each of the nodes in the illustration is associated with a particular point event and this association is indicated by the vertical dotted line. The nodes associated with the immigrant point events are labelled with zeroes, and all of the other nodes are associated with offspring point events. The numbering indicates the generation to which a point event belongs.

The immigrant point events are the point events which arrive according to the homogeneous Poisson process with intensity τ. The immediate offspring point events associated with a particular point event are indicated by using arrows to join the nodes of the parent and offspring point events, e.g. t_3 and t_6 are the immediate offspring of t_2. The immediate offspring point events associated with a particular point event arrive according to a nonhomogeneous Poisson process with intensity $\omega(\cdot)$, e.g. t_3 and t_6 are realisations from a nonhomogeneous Poisson process with intensity $\omega(s - t_2)$ for $s \geq t_2$. The point event which produces a particular offspring point event is described as the immediate ancestor of that offspring point event, e.g. in Figure 2.4, t_7 is the immediate ancestor of t_8. The point events with nodes which are directly and indirectly connected by arrows to the node associated with an immigrant point event form the cluster of offspring point events associated with that immigrant point event, e.g. in Figure 2.4, the point t_1 is an immigrant point event and the points t_2, t_3, t_4, t_5, and t_6 form the cluster of offspring point events associated with t_1.

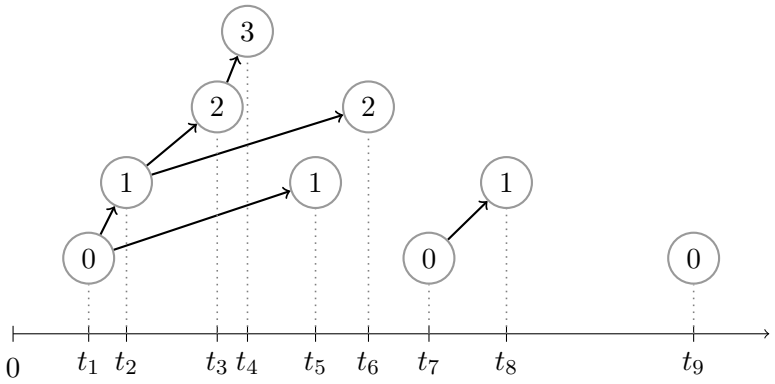


Figure 2.4: *Fictional realisation of a Hawkes process. The unobserved relationships between the point events, also referred to as the branching structure, are illustrated in the upper portion of the figure. The t_is give the observed times of the point events on the time axis, and each of the vertical dotted lines is used to indicate the node associated with a particular point event. Figure adapted from Møller and Rasmussen (2005, Figure 1, p. 630).*

As the branching structure is unobserved, this interpretation of the Hawkes process suggests an EM algorithm to find the MLEs (Veen and Schoenberg, 2008). EM algorithms based on this structure have received

attention in the literature recently. This work includes that of Veen and Schoenberg (2008), Lewis and Mohler (2011), Halpin and De Boeck (2013), the follow-up paper by Halpin (2013), and Olson and Carley (2013). Estimation via an EM algorithm is discussed in Chapter 3.

2.2.3 Marked Hawkes processes

The marked Hawkes process is a generalisation of the unmarked Hawkes process where each point event time now has a mark associated with it. When specifying a marked Hawkes process, we will define the conditional intensity of the ground process and then define the conditional mark distribution, as described above.

The intensity of the ground process of a marked Hawkes process can be extended to include influences from the observed mark values. This extension is common to applications in seismology and finance; see, for example, Ogata (1988) and Chavez-Demoulin *et al.* (2005). The extended complete intensity is defined for all t by

$$\lambda^{\dagger}(t|\mathcal{H}_t) = \tau + \sum_{j:t_j<t} \omega(t - t_j, m_j),$$

where $\omega(s, m) \geq 0$ for $s \geq 0$, $m \geq 0$ and zero otherwise. The increase in the complete intensity following a point event is now affected by the observed mark. Allowing the marks to affect the complete intensity value has intuitive appeal in some applications. The conditional intensity can be defined in similar manner to the complete intensity, but with the conditioning on the observed history.

The form of $\omega(t, m)$ will depend on the required effect of a point event on the intensity. It is common for

$$\omega(t, m) = g(m)\omega^*(t),$$

where $\omega^*(\cdot)$ is the response function as given earlier, and $g : \mathcal{M} \mapsto \mathbb{R}_+$. The function $g(\cdot)$ controls the effect of the observed marks on the intensity and will be referred to as the impact function. Note that this impact function is different from the impact function defined by Liniger (2009, p. 19). Impact functions which have the exponential form $g(m) = e^{\delta m}$ are popular in the literature. The degenerate case where $g(m) = 1$ results in an intensity that is independent of the observed marks.

The likelihood function for a marked Hawkes process has the same form as (2.9). As an example, consider an observation from a marked Hawkes process with an exponential decay function, an exponential impact function, and iid exponential marks with mean $\mu^{-1} > 0$. The conditional intensity of the ground process is given by

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi \sum_{j:t_j \in (0,t)} \exp\left(\delta m_j - \gamma(t - t_j)\right),$$

where $\psi, \delta \geq 0$, $\tau, \gamma > 0$, and effects from point events occurring before time 0 have been ignored. Suppose that the observation consists of the points $(t_1, m_1), \ldots, (t_{N_g(T)}, m_{N_g(T)})$ from the finite interval $[0, T]$. Then, upon substituting in the conditional intensity and the conditional mark density, the log-likelihood is given by

$$\begin{aligned}\ell(\boldsymbol{\theta}) = &\sum_{i=1}^{N_g(T)} \log\left(\tau + \psi \sum_{j:t_j<t_i} e^{\delta m_j - \gamma(t_i - t_j)}\right) - \tau T \\ &- \frac{\psi}{\gamma} \sum_{i=1}^{N_g(T)} e^{\delta m_i}\left(1 - e^{-\gamma(T-t_i)}\right) + N_g(T)\log\mu - \mu \sum_{i=1}^{N_g(T)} m_i. \quad (2.19)\end{aligned}$$

As is suggested by the nature of (2.19), analytic expressions for all of the MLEs are typically not available for marked Hawkes process models, and DNM, or some other iterative technique, will have to be used to find the estimates. This is the topic of the next chapter.

CHAPTER 3

Estimation

3.1 Introduction

The focus of this chapter is on maximum likelihood estimation of Hawkes process models. Both DNM of the log-likelihood and an EM algorithm are discussed.

Maximum likelihood estimation of Hawkes process models is typically carried out via DNM of the log-likelihood function. Early use of this approach to finding MLEs can be found in the work of Vere-Jones (1978) and Ozaki (1979). Since then, there have been numerous applications in the literature that have used DNM of the log-likelihood to find MLEs; examples include the work of Ogata (1988), Embrechts *et al.* (2011), and Chavez-Demoulin and McGill (2012). DNM of the log-likelihood is discussed in the second section of this chapter along with several practical considerations. The considerations discussed are: choosing the initial conditions, the computational burden of evaluating the log-likelihood, and multiple maxima in the log-likelihood surface. Some of these considerations also apply to estimation via an EM algorithm.

The EM algorithm is an alternative means of finding MLEs. The EM algorithm presented here uses the Poisson cluster process interpretation under which the unobserved branching structure is treated as the missing data. This particular structuring of the EM algorithm appears to be relatively recent and is advocated by Veen and Schoenberg (2008) and Olson and Carley (2013). Veen and Schoenberg (2008) demonstrate that an EM-type algorithm may be suited to Hawkes process models with nearly flat log-likelihood

surfaces, and Olson and Carley (2013) argue that an EM algorithm may be suited to 'complex' Hawkes process models. The EM algorithm for unmarked Hawkes processes is discussed, and to some extent investigated, in the third section of this chapter. The investigation involves a simulation study. The intention of this is to validate the estimation routines and to clarify statements made in the literature about the advantages of the EM algorithm as applied to Hawkes process models. The particular statements concern the accuracy of the estimates found via DNM of the log-likelihood relative to those found via the EM algorithm. In the simulation study, we also demonstrate several techniques for finding confidence intervals for the parameter estimates. We do not consider here the EM algorithm for marked Hawkes processes, as in most cases it is a straightforward extension to the EM algorithm presented for unmarked Hawkes processes.

The last section of this chapter presents an illustrative example where three unmarked Hawkes process models are fitted via DNM of the log-likelihood to the earthquake data considered by Ogata (1988). Two of the models we investigate are similar to models investigated by Ogata (1988), and this allows for a comparison of results. The other model that we consider has a response function related to the gamma density, and it performs reasonably well.

All of the computations in this chapter, and in the remainder of this dissertation, are performed by using R (R Core Team, 2012) and C.

3.2 Estimation via DNM of the likelihood

The general form of the likelihood function for a marked Hawkes process N was given in (2.9), and the associated log-likelihood function is given by

$$\ell(\boldsymbol{\theta}) = \underbrace{\sum_{i=1}^{N_g(T)} \log \lambda(t_i|\widetilde{\mathcal{H}}_{t_i}) - \int_0^T \lambda(u|\widetilde{\mathcal{H}}_u)\,\mathrm{d}u}_{\text{Part 1}} + \underbrace{\sum_{i=1}^{N_g(T)} \log f(m_i)}_{\text{Part 2}}. \tag{3.1}$$

The MLE of $\boldsymbol{\theta}$ can be found by maximising the likelihood function, or more typically the log-likelihood function, with respect to $\boldsymbol{\theta}$ over the parameter space $\boldsymbol{\Theta}$. The MLE $\widehat{\boldsymbol{\theta}}$ is defined as $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta})$. In the remainder of this section we discuss some practical considerations that need to be borne in mind when using DNM of the log-likelihood to find MLEs.

3.2.1 Problems: initial conditions

A point process is typically only observed for a finite interval $[0, T]$, where time 0 is some time after the origin of the process, and for point processes with evolutionary character, there may be effects from point events occurring before time 0 on the conditional intensity during the observation period. Such effects are referred to as edge or boundary effects; see, for example, Daley and Vere-Jones (2003, p. 235). In the case of the Hawkes process, there may be unobserved point events occurring before time 0 which give rise to offspring point events in the observation period. If this is the case, these unobserved point events will increase the conditional intensity at the start of the observation period, but the increase will be unknown. As such, before estimating the parameters, one has to specify the initial conditions for the conditional intensity (Daley and Vere-Jones, 2003, pp. 212, 234–235). This has to be done when using DNM of the log-likelihood and the EM algorithm.

The simplest choice for the initial conditions would be to assume that the initial value of the conditional intensity equals τ and ignore effects from point events occurring before the start of the observation period (Daley and Vere-Jones, 2003, p. 234). This is an approach often taken in applications in the literature. If one ignores the effects from point events occurring before the start of the observation period, the conditional intensity can be regarded as approximate for some initial part of the observation period, and as such, there is likely to be some effect on the estimated model. For example, Rasmussen (2011) highlights, amongst other effects, that the estimate of τ is likely to be too high. However, it is also noted by Rasmussen (2011), that the effects on the estimated model will be negligible if the data set being used is large. In our applications we set $\lambda(0|\widetilde{\mathcal{H}}_0) = \tau$ and ignore the effects from point events occurring before time 0, but we also outline below some alternative methods of handling the initial conditions.

A method which can be used to improve the estimate of the initial value of the conditional intensity involves splitting the observation period into two parts, $[0, S)$ and $[S, T]$. The point events in the first part of the observation period are used to calculate a value for the conditional intensity at time S. Then, by using this value for the conditional intensity at S and allowing for

the appropriate decay, the log-likelihood is maximised given the observations in the second period $[S, T]$. The idea is that by using the data in the first period $[0, S)$, a better value for the conditional intensity at the start of the effective observation period $[S, T]$ can be found, and this will reduce the effects on the estimated model. This is an idea suggested by Bebbington and Harte (2001) in the context of the linked stress release model.

Alternatively one may 'wrap' the observation period $[0, T]$ on itself, so that the point events towards the end of the observation period are used to obtain an initial value for the conditional intensity. The rationale behind this method is similar to that of the method just described, except here we have not reduced the size of the data set used to find the MLEs. This method is described in the literature as introducing 'periodic boundary conditions'. This, and the method above, are remedies used in the spatial statistics literature for edge effects; see, for example, Baddeley and Turner (2000).

Daley and Vere-Jones (2003, p. 234) identify another approach which is available if the Hawkes process is assumed to have reached equilibrium. In such a case, an 'averaged likelihood' can be calculated by integrating the likelihood function over the equilibrium distribution of $\lambda(0|\widetilde{\mathcal{H}}_0)$, provided the equilibrium distribution of $\lambda(0|\widetilde{\mathcal{H}}_0)$ can be found. This averaged likelihood can then be maximised to find the MLEs. See Daley and Vere-Jones (2003, p. 234) and the reference there for the meaning of equilibrium.

3.2.2 Problems: computational difficulties

Evaluating the log-likelihood, or more specifically the repeated evaluation of the conditional intensity function when evaluating the log-likelihood, can be computationally intensive. Liniger (2009, p. 7) notes that the computational burden, and the lack of available computing power, may have hindered early applications of Hawkes process models to real-world data sets. The abundance of relatively powerful personal computers in recent years may be one of the reasons for the increased number of applications. Even so, methods of coping with or reducing the computational burden of parameter estimation are present in recent literature. We discuss several of these methods here, as well as some methods which appeared early in the Hawkes process literature. We implement some of the methods in our applications.

Before discussing methods particular to Hawkes process models, it should

be noted that when maximising (3.1), the two parts of the log-likelihood function can be maximised separately if they do not have parameters in common. This is mentioned by both Ogata (1988, pp. 13–16) and Daley and Vere-Jones (2003, pp. 238–239). By maximising each of the parts separately, two simpler independent maximisation problems are produced. This can speed up the parameter-estimation routines.

The computational burden of evaluating (3.1) arises primarily from a nested sum. The nested sum is the first term of Part 1 of Equation (3.1) and can be expanded to

$$\sum_{i=1}^{N_g(T)} \log \lambda(t_i|\widetilde{\mathcal{H}}_{t_i}) = \sum_{i=1}^{N_g(T)} \log \left(\tau + \sum_{j:t_j<t_i} \omega(t_i - t_j, m_j) \right). \tag{3.2}$$

The number of operations required to evaluate this nested sum is of order $N_g(T)^2$ for most marked Hawkes process models, and its evaluation usually determines the order of operations for the entire log-likelihood evaluation. As a result, estimating the parameters can be slow for large $N_g(T)$, and this may be compounded if explicit loops in the evaluation of (3.2) cannot be avoided, as loops in R can be slow to evaluate. There are several methods available for reducing the computational burden of evaluating (3.2).

In the case of a marked Hawkes process model with an exponential decay function, the number of operations required to evaluate (3.2) can be reduced to the order of $N_g(T)$ by using a recursive formula. The recursive formula is used to evaluate the conditional intensity at each of the observed t_is. Such recursive formulae for unmarked Hawkes process models were presented by Ogata (1981), and more recently Liniger (2009, pp. 42–44) presented a recursive formula for marked multitype Hawkes process models. As an example, consider a marked Hawkes process model with a decay function of the form $\omega(s, m) = \psi e^{-\gamma s} g(m)$; i.e. a marked Hawkes process model with conditional intensity

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi \sum_{i:t_i \in (0,t)} e^{-\gamma(t-t_i)} g(m_i).$$

Then one particular recursive formula, which is useful when evaluating the log-likelihood, is given by

$$A(i) = \tau + e^{-\gamma(t_i - t_{i-1})}(A(i-1) - \tau) + \psi e^{-\gamma(t_i - t_{i-1})} g(m_{i-1}) \quad \text{for } i = 2, 3, \ldots,$$

where $A(i) = \lambda(t_i|\widetilde{\mathcal{H}}_{t_i})$ and $A(1) = \tau$. This recursive formula reduces the computational burden of evaluating the log-likelihood, as $\lambda(t_i|\widetilde{\mathcal{H}}_{t_i})$ can be found without having to sum over the entire observed history up to time t_i, if $\lambda(t_{i-1}|\widetilde{\mathcal{H}}_{t_{i-1}})$ is known, i.e. the nested sum (3.2) is reduced to a single sum. Such recursive formulae can be constructed because the intensity is a Markov process.

Ogata *et al.* (1993) present an involved strategy for reducing the computational burden of evaluating the log-likelihood function of a Hawkes process model with conditional intensity

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi \sum_{i:t_i\in(0,t)} \frac{e^{\delta m_i}}{(t - t_i + \gamma)^{\eta+1}}.$$

This is the conditional intensity of the ETAS model introduced by Ogata (1988). The method presented by Ogata *et al.* (1993) involves using two transformations and numerical integration. The result is an approximation for (3.2) which is of order $N_g(T)$. Ogata *et al.* (1993) demonstrate that there is a significant reduction in the time taken to estimate the parameters when using their approximation, and they also show that the resulting parameter estimates are close to the MLEs.

A general method for reducing the computational burden of evaluating the conditional intensity in the case $\omega(\Delta t, m) \approx 0$ for $\Delta t > t - q(t)$ is suggested by Lomnitz (1974, pp. 98–99) and described in detail by Liniger (2009, pp. 41–42). The method reduces the computational burden by truncating the summation involved in evaluating the conditional intensity. This is done by ignoring summands smaller than some threshold. In detail, the method involves calculating the approximate conditional intensity

$$\hat{\lambda}(t|\widetilde{\mathcal{H}}_t) = \tau + \sum_{j:t_j\in[q(t),t)} \omega(t - t_j, m_j), \tag{3.3}$$

where $q(t)$ is chosen such that the likely contribution to the conditional intensity at time t from point events occurring before time $q(t)$ is small. Liniger (2009, p. 42) notes that the summation in (3.3) should be run from the latest to the earliest point events, and stop once the contributions to the conditional intensity are sufficiently small. This avoids unnecessary checking of the condition for including a point event in the summation.

Liniger (2009, pp. 44–45) also suggests an approximation for evaluating the integral in Part 1 of Equation (3.1). From our experience, this approximation is not necessary, as evaluating the integral does not result in a bottleneck which slows the evaluation of the log-likelihood.

The method that we use here, and that may be used in conjunction with most of the methods described above, takes advantage of the ease with which C subroutines can be incorporated into R. Specifically, when evaluating the log-likelihood of a marked Hawkes process model, we use a compiled C subroutine to evaluate the inner summations of (3.2), and the outer summation is evaluated in R. The C subroutine is called into R via the `.C` interface in R. The advantage of using C subroutines is that evaluating loops in C is significantly faster than in R. Strictly speaking, this method alone does not reduce the computational burden, but is a means of efficiently coping with it. The absolute time gains of using C subroutines in R may be large, especially for computations which repeatedly evaluate the log-likelihood function, e.g. DNM of the log-likelihood and parametric bootstrap routines.

This is not a new method of coping with computational burden in R, and the speed advantages of calling C subroutines into R are well documented; see, for example, the list of points in favour of calling C subroutines into R given by Chambers (2008, pp. 412–413). Two R packages that make use of C subroutines to estimate marked and unmarked Hawkes process models are the `QRM` and `ETAS` packages. The `QRM` package (Pfaff and McNeil, 2012) uses the method described in the paragraph above to evaluate (3.2) when evaluating the log-likelihood. The models in the `QRM` package are marked Hawkes process models defined similarly to those of McNeil *et al.* (2005, pp. 306–311)[1]. The `ETAS` package (Jalilian, 2012) contains a parameter-estimation routine that makes use of C subroutines to estimate the space-time ETAS model of Ogata (1998).

[1]The models implemented in the `QRM` package are slightly different from those defined by McNeil *et al.* (2005, pp. 306–311), even though, from its description, the `QRM` package is intended to 'accompany the book *Quantitative Risk Management: Concepts, Techniques and Tools* by McNeil *et al.* (2005)' (Pfaff and McNeil, 2012). For example, the models implemented in the `QRM` package use a linear impact function of the form $(1+\delta m)$, instead of the exponential impact function $e^{\delta m}$ used by McNeil *et al.* (2005, p. 306). This, and other differences, can be seen in the package's C subroutines.

There would also be speed advantages to evaluating the nested sum in (3.2) by using Fortran or C++ subroutines instead of C subroutines (Chambers, 2008, p. 413). Embrechts *et al.* (2011) make use of R and C++ to implement their multitype marked Hawkes process models, but do not report any speed advantages of using C++. We present R and C++ code in Appendix A.1 which implements the log-likelihood used in Section 3.4. The time taken to find the MLEs for the simulated data in Section 3.4 when using this code is comparable to an implementation in R which makes use of a C subroutine.

In recent work, Guo *et al.* (2013) use reconfigurable computer hardware to accelerate the evaluation of the conditional intensity functions of a multitype Hawkes process. The 'speedup' achieved by their strategy is significant, with an increase in speed of up to 94 times.

3.2.3 Problems: multiple maxima in the likelihood

The minus log-likelihood function of a marked Hawkes process can be a complicated function of the parameters and as a result it may be nonconvex. The objective of maximum likelihood estimation is to find the global maximum of the log-likelihood function. As the log-likelihood may be nonconvex, a DNM routine could converge to a merely local maximum as opposed to the global maximum. This problem is also faced when using an EM algorithm to find MLEs.

A common strategy used to try identify the global maximum involves using several sets of different starting values for the DNM routine or EM algorithm. This strategy does not solve the problem entirely, and a merely local maximum may still be identified incorrectly as the global maximum. This strategy can be taken further when using DNM to find MLEs; two or more different DNM routines, with very different optimisation methods, may be used in conjunction with several sets of starting values to find the parameter estimates. If the different DNM routines identify the same point as the potential global maximum, more confidence can be placed in the identified point being the actual global maximum. This is the strategy employed in the applications in Chapter 7.

The numerical optimisation routines `constrOptim` and `DEoptim`, which are available in R (R Core Team, 2012; Mullen *et al.*, 2011; Ardia *et al.*,

2013), are two of the optimisation routines which we use. The `constrOptim` routine, as a default when gradient functions are not supplied, uses the Nelder–Mead method, and can enforce linear inequality constraints on parameters. The `DEoptim` routine is an implementation of the differential evolution algorithm of Storn and Price (1997). The `DEoptim` routine does not require starting values to be supplied, but requires that box constraints be given for parameter values. Alternative routines and methods can be used; for example, `nlm` is an alternative routine which we use in our simulation study in Section 3.4 and for parts of the applications in Chapter 7.

In passing, it is worth noting that there are Hawkes process models with convex minus log-likelihoods. In such cases, the log-likelihood has at most one maximum. An example of such a Hawkes process model is the model with conditional intensity of the form

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \sum_{k=1}^{K} \theta_k Q_k(t|\widetilde{\mathcal{H}}_t),$$

where $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_K\}$ is the parameter vector to be estimated, and $Q_k(t|\widetilde{\mathcal{H}}_t)$ for $k = 1, \ldots, K$ are known functions, i.e. their form and parameter values are known. For this example, the Hessian of the log-likelihood is negative semidefinite, and so, provided that $\boldsymbol{\Theta}$ is convex, the minus log-likelihood is convex; see Ogata (1978, p. 255). This is one of two examples presented by Ogata (1978, 1999), and it has received attention in the literature; for example, Ogata and Akaike (1982) and Chornoboy *et al.* (1988) use point process models with linear intensity functions of this form. The Hawkes process model with an exponential decay function $\omega(s) = \psi e^{-\gamma s}$, where γ is treated as known, is an example of such a model.

3.3 Estimation via an EM algorithm

The EM algorithm is a general iterative algorithm, or rather class of algorithms, that can be used to find the MLEs when the observations may be regarded as incomplete (Dempster *et al.*, 1977, McLachlan and Krishnan, 2008, p. 1). In the case of Hawkes process models, the unobserved branching structure under the Poisson cluster process interpretation is treated as the missing data and can be used to construct an EM algorithm. This structuring of the EM algorithm appeared in the work of Veen and Schoenberg

(2008), and is similar to the EM algorithm of Tsukakoshi and Shimazaki (2006). It is this structuring that we focus on. An alternative means of constructing an EM algorithm for Hawkes process models can be found in the work of Mino (2001).

The general structure of the EM algorithm we consider is presented in Section 3.3.1. This particular EM algorithm can be computationally intensive and several approximations have appeared in the literature which reduce this computational burden. An approximate EM algorithm is discussed in Section 3.3.2. The EM algorithm and the approximate version are investigated and compared to DNM of the log-likelihood in a simulation study presented in Section 3.4.

The log-likelihood for the observed points $t_1, t_2, \ldots, t_{N(T)}$ will be referred to as the observed-data log-likelihood (ODLL) to distinguish it from the complete data log-likelihood (CDLL) introduced below.

The EM algorithm described below is similar to those presented by Lewis and Mohler (2011), Hegemann *et al.* (2013), Halpin and De Boeck (2013), and Olson and Carley (2013).

3.3.1 EM algorithm for unmarked Hawkes process models

Suppose that we observe a Hawkes process with conditional intensity

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \sum_{j:t_j \in (0,t)} \omega(t - t_j) \tag{3.4}$$

on the interval $[0, T]$, and that the observation consists of the times of the point events $t_1, t_2, \ldots, t_{N(T)}$. Similar to Veen and Schoenberg (2008), we define the variable u_i, associated with the ith point event t_i, as follows:

$u_i = j$ if the immediate ancestor of point event i is point event j, and

$u_i = i$ if point event i is an immigrant point event.

The u_is describe the unobserved branching structure of the Hawkes process, and the complete data is $(t_1, u_1), (t_2, u_2), \ldots, (t_{N(T)}, u_{N(T)})$.

If the branching structure is assumed to be known, the complete data likelihood is given by

$$L_{\mathrm{CD}}(\boldsymbol{\theta}, \boldsymbol{u}) = \underbrace{\left[e^{-\tau T} \prod_{i:u_i=i} \tau \right]}_{\text{Part 1}} \times \underbrace{\prod_{i=1}^{N(T)} \left[\exp\left(- \int_{t_i}^{T} \omega(s - t_i)\, \mathrm{d}s \right) \prod_{j:u_j=i, j\neq i} \omega(t_j - t_{u_j}) \right]}_{\text{Part 2}}. \quad (3.5)$$

Part 1 on the right-hand side is the likelihood for the immigrant point events which arrive according to a homogeneous Poisson process with intensity τ. Part 2 on the right-hand side is a product of likelihoods. Each of the likelihoods in this product is for the offspring point events generated by the nonhomogeneous Poisson process triggered by the arrival of a particular point event. The complete data likelihood is constructed by using the fact that, given the point events that trigger them, the nonhomogeneous Poisson processes giving rise to the offspring point events are mutually independent and are independent of the homogeneous Poisson process generating the immigrant point events.

The CDLL is then given by

$$\ell_{\mathrm{CD}}(\boldsymbol{\theta}, \boldsymbol{u}) = \sum_{i:u_i=i} \log \tau - \tau T - \sum_{i=1}^{N(T)} \int_{t_i}^{T} \omega(s - t_i)\, \mathrm{d}s + \sum_{i:u_i\neq i} \log \omega(t_i - t_{u_i}). \quad (3.6)$$

Given this CDLL and treating the branching structure as missing, an EM algorithm can be constructed as follows.

The E step

The expectation step (E step) of the EM algorithm involves taking the conditional expectation of the CDLL with respect to the u_is, given $\boldsymbol{\theta}^{(k)}$, the parameter estimates at the kth iteration of the algorithm, and the observed point process $\widetilde{\mathcal{H}}_T$. The conditional expected value of the CDLL can be

written as

$$\begin{aligned} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) &= \mathrm{E}\left[\ell_{\mathrm{CD}}(\boldsymbol{\theta}, \boldsymbol{u}) \,\middle|\, \widetilde{\mathcal{H}}_T, \boldsymbol{\theta}^{(k)}\right] \\ &= \mathrm{E}\left[\sum_{i=1}^{N(T)} I_{\{u_i=i\}} \log \tau - \tau T - \sum_{i=1}^{N(T)} \int_{t_i}^{T} \omega(s - t_i)\,\mathrm{d}s \right. \\ &\qquad \left. + \sum_{i=1}^{N(T)} \sum_{j \neq i} I_{\{u_i=j\}} \log \omega(t_i - t_j) \,\middle|\, \widetilde{\mathcal{H}}_T, \boldsymbol{\theta}^{(k)}\right], \end{aligned} \tag{3.7}$$

where the indicator random variables $I_{\{u_i=i\}}$ and $I_{\{u_i=j\}}$ have unit value when the equality in the subscript is true, and are zero otherwise.

The following probabilities can then be used to find an expression for the conditional expected CDLL:

$$\Pr\left\{u_i = j \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\} = \begin{cases} \frac{\omega(t_i - t_j|\boldsymbol{\theta}^{(k)})}{\tau^{(k)} + \sum_{n:t_n<t_i} \omega\left(t_i - t_n|\boldsymbol{\theta}^{(k)}\right)} & \text{for } j < i, \\ \frac{\tau^{(k)}}{\tau^{(k)} + \sum_{n:t_n<t_i} \omega\left(t_i - t_n|\boldsymbol{\theta}^{(k)}\right)} & \text{for } j = i, \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

These probabilities are analogous to the probabilities used by Ogata (1981, p. 24) to perform the thinning in his simulation algorithm. That is, if they are used to thin a realisation of the Hawkes process with conditional intensity (3.4), the resulting thinned realisations would be equivalent to realisations from the nonhomogeneous Poisson processes and the homogeneous Poisson process under the Poisson cluster process interpretation of the Hawkes process. See Ogata (1981, p. 24) for the justification. We refer to (3.8) as the probability distribution of the branching structure.

The identity $\mathrm{E}(I_{\{u_i=j\}}|\widetilde{\mathcal{H}}_T, \boldsymbol{\theta}^{(k)}) = \Pr\left\{u_i = j \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}$ can be used to show

$$\begin{aligned} &Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) \\ &\quad = \log(\tau) \sum_{i=1}^{N(T)} \Pr\left\{u_i = i \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\} - \tau T - \sum_{i=1}^{N(T)} \int_{t_i}^{T} \omega(s - t_i)\,\mathrm{d}s \\ &\qquad + \sum_{i=2}^{N(T)} \sum_{j=1}^{i-1} \log\left(\omega(t_i - t_j)\right) \Pr\left\{u_i = j \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}. \end{aligned} \tag{3.9}$$

The M step

The maximisation step (M step) of the EM algorithm involves maximising $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ to obtain estimates $\boldsymbol{\theta}^{(k+1)}$. An analytic solution exists for the baseline rate $\tau^{(k+1)}$, which we give in Section 3.4, but analytic solutions for all of the remaining elements of $\boldsymbol{\theta}^{(k+1)}$ do not typically exist. To find the remaining elements of $\boldsymbol{\theta}^{(k+1)}$, which are the parameters of the response function, will at worst involve a numerical maximisation over all of the remaining parameters.

The EM algorithm

Starting from an initial estimate $\boldsymbol{\theta}^{(0)} \in \boldsymbol{\Theta}$, the EM algorithm is iterated through the E and M steps to find the MLEs. The algorithm is stopped once it is deemed to have converged, e.g. when the difference $\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)})$ is suitably small. Algorithm 3.1 presents the EM algorithm for an unmarked Hawkes process model.

begin
- Choose $\boldsymbol{\theta}^{(0)} \in \boldsymbol{\Theta}$, a small $\epsilon \in (0, 1)$, set $c \leftarrow 1$, and set $k \leftarrow 0$;
- **while** $c > \epsilon$ **do**
  - **E step:** Calculate $\Pr\left\{u_i = j \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}$ for all $j \leq i$ and all i by using $\boldsymbol{\theta}^{(k)}$ and $\widetilde{\mathcal{H}}_T$;
  - **M step:** Find $\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left[Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})\right]$;
  - Set $k \leftarrow k + 1$ and $c \leftarrow \ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)})$;
- **return** $\boldsymbol{\theta}^{(k)}$.

Algorithm 3.1: The EM algorithm for an unmarked Hawkes process model.

3.3.2 Estimation via an approximate EM algorithm

A numerical maximisation at each M step is computationally expensive. As a result, methods have been developed to try decrease the complexity of the maximisation problem at the M step, and so reduce the overall computational burden of the above EM algorithm.

A strategy which Veen and Schoenberg (2008) appear to use to reduce the complexity of the maximisation at the M step is to change the bounds of the integrals in (3.6). The strategy can be described as follows. Suppose that $\omega(s) = \zeta\omega^\dagger(s)$, where $\zeta \geq 0$ and $\omega^\dagger(s)$ is a probability density function defined for $s \geq 0$. Then if one lets the upper bounds of the integrals in Equation (3.6) be infinity, each of the integrals equals ζ and is not dependent on the actual observed time of the point event. For some Hawkes processes, e.g. the Hawkes process with exponential decay function (Lewis and Mohler, 2011; Olson and Carley, 2013), this results in analytic solutions at the M step. However, as ζ and $\log\omega^\dagger(s)$ may be complicated functions of the parameters for some Hawkes processes, analytic solutions do not always exist. In cases where analytic solutions do not exist, the approximation may help to reduce the complexity of the CDLL and potentially reduce the complexity of the maximisation at the M step.

The approximation described above is good when

$$\zeta \sum_{i=1}^{N(T)} \int_{t_i}^{T} \omega^\dagger(u - t_i)\,\mathrm{d}u \approx \zeta N(T). \tag{3.10}$$

This occurs when the effects on the intensity function from the point events in the observation period have 'died off' by the end of the observation period, i.e. $\int_{t_i}^{T} \omega^\dagger(u - t_i)\,\mathrm{d}u \approx 1$ for each i. More specifically, this may be the case when: the decay function 'dies off' quickly over time, there are few point events in the observation period, and there are few point events near to the end of the observation period (Olson and Carley, 2013).

In a similar manner, the E and M steps described above can be used to construct an estimation algorithm. The resulting algorithm is strictly not an EM algorithm, and is referred to here as the approximate EM algorithm.

Another approximation proposed by Halpin (2013), which is not investigated here, involves truncating some of the summations involved in calculating the conditional expected CDLL (3.9). Unlike the approximation described above, this approximation aims to reduce the computational burden of evaluating the conditional expected CDLL and does not explicitly reduce the complexity of the maximisation problem at the M step. In addition, the error introduced by truncating the summations can be controlled by choosing where to truncate the summations. Halpin (2013) investigates

this approximation and the error introduced. Olson and Carley (2013) also investigate a similar approximation.

3.4 A simulation study

A simulation study is presented in this section. The purpose of this simulation study is to investigate the three estimation routines presented, clarify some points made in the literature, and validate the parameter-estimation routines. The parameter-estimation routines are: DNM of the ODLL, an exact EM algorithm, and an approximate EM algorithm. The study involves estimating a Hawkes process model by using these three parameter-estimation routines for data that are simulated from a known Hawkes process. The estimates found by the three routines are then compared to the parameters used to simulate the data.

The Hawkes process used in the study has the conditional intensity

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi \sum_{j:t_j \in (0,t)} \exp(-\gamma(t - t_j)),$$

and the parameter values used are:

$$\tau = 0.05, \quad \psi = 0.035, \quad \text{and} \quad \gamma = 0.07.$$

These parameter values are referred to as the true parameter values. The data for the study are simulated by using Ogata's modified thinning algorithm; see Ogata (1981), Daley and Vere-Jones (2003, p. 271), and Appendix A.1. The simulation consists of 976 point events over the interval $[0, 10\,000)$. The algorithm and R code used to simulate the data are presented in Appendix A.1.

Lewis and Mohler (2011) carry out a simulation study using an EM algorithm and a similar Hawkes process model to that defined above. The focus of their simulation study is the estimation of the Hawkes process model for data that are simulated by using different values of γ. They find that the variances of the parameter estimates, and the number of iterations required for their EM algorithm to converge, both increase as γ^{-1} increases.

3.4.1 Estimation via DNM of the ODLL

The ODLL for this Hawkes process model is given by

$$\ell(\boldsymbol{\theta}) = -\tau T + \frac{\psi}{\gamma}\sum_{i=1}^{N(T)}\left(e^{-\gamma(T-t_i)}-1\right) + \sum_{i=1}^{N(T)}\log\left(\tau + \psi\sum_{j:t_j<t_i} e^{-\gamma(t_i-t_j)}\right). \tag{3.11}$$

The minus ODLL is minimised numerically by using the `nlm` routine in R. Constraints on the parameter values are enforced via log-transformations.

To find the MLEs, the `nlm` routine, with default convergence criteria, took 14 iterations to converge from the starting values $\tau^{(0)} = 0.08$, $\psi^{(0)} = 0.025$, and $\gamma^{(0)} = 0.035$. The relevant R code to find the MLEs is presented in Figures A.4 and A.5 in Appendix A.2. An implementation which makes use of a C++ subroutine, and which can be run entirely within R, is presented in Figure A.6 in Appendix A.2.

3.4.2 Estimation via an EM algorithm

The CDLL for this Hawkes process model is given by

$$\ell_{\mathrm{CD}}(\boldsymbol{\theta}, \boldsymbol{u}) = \sum_{i:u_i=i}\log\tau - \tau T - \sum_{i=1}^{N(T)}\int_{t_i}^{T}\psi\exp(-\gamma(s-t_i))\,\mathrm{d}s + \sum_{i:u_i\neq i}\left(\log\psi - \gamma(t_i - t_{u_i})\right). \tag{3.12}$$

E step

The conditional expected CDLL is given by

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}) = \log(\tau)\sum_{i=1}^{N(T)}\Pr\left\{u_i = i \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\} - \tau T + \frac{\psi}{\gamma}\sum_{i=1}^{N(T)}\left(e^{-\gamma(T-t_i)}-1\right) + \sum_{i=2}^{N(T)}\sum_{j=1}^{i-1}\left(\log\psi - \gamma(t_i - t_j)\right)\Pr\left\{u_i = j \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}.$$

M step

To maximise the conditional expected CDLL, we take the partial derivative of $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)})$ with respect to each of the parameters and set each derivative equal to zero. The resulting system of equations is

$$\begin{aligned} \tau^{(k+1)} &= \frac{\sum_{i=1}^{N(T)} \Pr\left\{u_i = i \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}}{T}, \\ \psi^{(k+1)} &= \frac{\gamma^{(k+1)} \sum_{i=2}^{N(T)} \sum_{j=1}^{i-1} \Pr\left\{u_i = j \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}}{\sum_{i=1}^{N(T)} \left(1 - e^{-\gamma^{(k+1)}(T-t_i)}\right)}, \end{aligned} \tag{3.13}$$

and

$$\gamma^{(k+1)} = \frac{\psi^{(k+1)} \left[\sum_{i=1}^{N(T)} \left(1 - e^{-\gamma^{(k+1)}(T-t_i)}\right) / \gamma^{(k+1)} - \sum_{i=1}^{N(T)} (T - t_i) e^{-\gamma^{(k+1)}(T-t_i)}\right]}{\sum_{i=2}^{N(T)} \sum_{j=1}^{i-1} (t_i - t_j) \Pr\left\{u_i = j \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}}. \tag{3.14}$$

The solution for $\tau^{(k+1)}$ is analytic. Analytic solutions do not exist for $\psi^{(k+1)}$ and $\gamma^{(k+1)}$. The parameter $\psi^{(k+1)}$ appearing on the right-hand side of Equation (3.14) can be substituted with the expression for $\psi^{(k+1)}$ given in Equation (3.13), and a solution for $\gamma^{(k+1)}$ can be found by using a root-finding algorithm. We use the root-finding algorithm `multiroot`, from the R package `rootSolve`, to do this (Soetaert and Herman, 2009; Soetaert, 2013). Once a solution for $\gamma^{(k+1)}$ has been found, a solution for $\psi^{(k+1)}$ can be found by using Equation (3.13).

To estimate the parameters for the simulated data, the exact EM algorithm was run for 193 iterations from the starting values $\tau^{(0)} = 0.08$, $\psi^{(0)} = 0.025$, and $\gamma^{(0)} = 0.035$.

3.4.3 Estimation via an approximate EM algorithm

The CDLL function can be simplified by making the approximation described above. That is, let the upper bounds of the integrals in Equation (3.12) be ∞ rather than T, so that the ith integral is over the range (t_i, ∞). The resulting approximate CDLL is given by

$$\widetilde{\ell_{\mathrm{CD}}}(\boldsymbol{\theta}, \boldsymbol{u}) = \sum_{i:u_i=i} \log \tau - \tau T - \frac{\psi N(T)}{\gamma} + \sum_{i:u_i \neq i} \left(\log \psi - \gamma(t_i - t_{u_i})\right).$$

The approximation results in the following relationship between the exact and approximate CDLL functions $\widetilde{\ell_{\mathrm{CD}}}(\boldsymbol{\theta}, \boldsymbol{u}) < \ell_{\mathrm{CD}}(\boldsymbol{\theta}, \boldsymbol{u})$.

By taking the conditional expectation of $\widetilde{\ell_{\mathrm{CD}}}(\boldsymbol{\theta}, \boldsymbol{u})$ with respect to the u_is, given the estimate $\boldsymbol{\theta}^{(k)}$ and the observed point process, and maximising the resulting conditional expected approximate CDLL, the following solutions for the elements of $\boldsymbol{\theta}^{(k+1)}$ can be found,

$$\tau^{(k+1)} = \frac{\sum_{i=1}^{N(T)} \Pr\left\{u_i = i \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}}{T}, \tag{3.15}$$

$$\psi^{(k+1)} = \frac{\left(\sum_{i=2}^{N(T)} \sum_{j=1}^{i-1} \Pr\left\{u_i = j \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}\right)^2}{N(T) \sum_{i=2}^{N(T)} \sum_{j=1}^{i-1} (t_i - t_j) \Pr\left\{u_i = j \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}}, \tag{3.16}$$

and

$$\gamma^{(k+1)} = \frac{\sum_{i=2}^{N(T)} \sum_{j=1}^{i-1} \Pr\left\{u_i = j \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}}{\sum_{i=2}^{N(T)} \sum_{j=1}^{i-1} (t_i - t_j) \Pr\left\{u_i = j \,\middle|\, \widetilde{\mathcal{H}}_{t_i}, \boldsymbol{\theta}^{(k)}, t_i\right\}}. \tag{3.17}$$

For this particular Hawkes process model, the approximation results in analytic solutions at the M step.

To estimate the parameters for the simulated data, the approximate EM algorithm was run for 198 iterations from the starting values $\tau^{(0)} = 0.08$, $\psi^{(0)} = 0.025$, and $\gamma^{(0)} = 0.035$.

3.4.4 Results and discussion

Table 3.1 presents the parameter estimates found for the simulated data, among other results. The reported estimates found via DNM and the exact EM algorithm are identical, and are close to the true parameter values. The approximate EM algorithm finds estimates which are larger than the MLEs, but which are also close to the true parameter values. The reason the approximate EM algorithm finds estimates which are quite close to the MLEs is due to a reasonable approximation; in this case there is only a small difference between

$$\sum_{i=1}^{N(T)} \int_{t_i}^{T} \widehat{\psi} \exp(-\widehat{\gamma}(u - t_i)) \, \mathrm{d}u = 477.2 \quad \text{and} \quad \frac{\check{\psi}}{\check{\gamma}} N(T) = 474.5,$$

where '$\widehat{\ }$' identifies the MLEs and '$\check{\ }$' identifies the estimates found via the approximate EM algorithm. If the parameter γ used for the simulation was

larger, the self-exciting effects would die off more quickly and it is likely that the approximate EM algorithm would find estimates even closer to the MLEs.

On the point of the approximate EM algorithm finding estimates which are not the MLEs, an interesting feature was identified when checking the approximate EM algorithm. For a smaller set of simulated data, the details of which are provided in Appendix A.3, it was noticed that the sequence of ODLL values from the approximate EM algorithm can be nonmonotonic. Figure 3.1 presents a plot of the ODLL value at each iteration of the approximate EM algorithm (the dashed line) for this smaller set of data. The ODLL values initially increase rapidly and then decrease after about 12 iterations as the approximate EM algorithm moves away from apparently better estimates. An EM algorithm is expected to find estimates at each iteration such that the sequence of ODLL values is nondecreasing, i.e. $\ell(\boldsymbol{\theta}^{(k+1)}) \geq \ell(\boldsymbol{\theta}^{(k)})$ (Dempster *et al.*, 1977). The approximate EM algorithm clearly does not satisfy this inequality, and as Meng (1997) writes,

> When an iterative algorithm is not monotone we know it cannot be an EM algorithm.

This reiterates the point about the approximate EM algorithm not being an EM algorithm. The plot in Figure 3.1 also contains the ODLL values for the exact EM algorithm (the grey dotted line) — reassuringly these are nondecreasing. Further results for this smaller simulation study are presented in Table A.1 in Appendix A.3.

The fact that the EM algorithm and DNM of the ODLL here find the same estimates is encouraging, and contrasts with results presented in some research investigating the EM algorithm as based on the Poisson cluster process interpretation of the Hawkes process. For instance, Halpin and De Boeck (2013, Table 1, p. 802) present results from a simulation study in which the estimates found via DNM of the ODLL, and those found via an EM algorithm, are different. They then draw conclusions about the relative sizes of the 'estimation error' of DNM and EM. In another simulation study, Olson and Carley (2013, p. 79) conclude that their EM algorithm finds estimates which are more 'statistically accurate' than those found via DNM of the ODLL.

Table 3.1: *Results from the simulation study. The estimates (est.) found via the three parameter-estimation methods are presented in the table along with the median time taken to find the estimates in seconds. Confidence intervals found by using the approximate Hessian matrix, profile likelihoods, and bootstrap (btst.) routines are also presented in the table. For the bootstrap routines, the confidence intervals are found by using the BC_a method; see Efron and Tibshirani (1994, pp. 184–188) for details. Most of the results in line three can be replicated by running the* R *code presented in Figures A.1–A.5 in Appendix A.1.*

Parameter	τ	ψ	γ		
True value	0.05000	0.03500	0.07000	$-\ell(\widehat{\boldsymbol{\theta}})$	Time (s)
DNM est. (`nlm`)	0.04988	0.03465	0.07082	3172.8106	2
Exact EM est.	0.04988	0.03465	0.07082	3172.8106	60
App. EM est.	0.05015	0.03482	0.07162	3172.8137	53
Wald-type 95% CI[1]	(0.040, 0.059)	(0.025, 0.044)	(0.049, 0.093)		
s.e.[1]	0.00484	0.00485	0.01114		
Likelihood-based 95% CI[2]	(0.041, 0.060)	(0.026, 0.045)	(0.052, 0.098)		
Btst. 95% CI[3]	(0.040, 0.060)	(0.024, 0.045)	(0.048, 0.097)		
Bootstrap mean[3]	0.05040	0.03495	0.07308		
Bootstrap s.e.[3]	0.00519	0.00538	0.01334		
Btst. 95% CI[4]	(0.040, 0.060)	(0.024, 0.045)	(0.048, 0.097)		
Bootstrap mean[4]	0.05040	0.03494	0.07307		
Bootstrap s.e.[4]	0.00519	0.00538	0.01334		
Btst. 95% CI[5]	(0.040, 0.060)	(0.024, 0.045)	(0.047, 0.094)		
Bootstrap mean[5]	0.05095	0.03540	0.07500		
Bootstrap s.e.[5]	0.00521	0.00534	0.01331		

[1] Found by using the approximate Hessian matrix supplied by the `nlm` routine in R.

[2] Found by using the profile likelihood functions and DNM; see Figure 3.2.

[3,4,5] Found by using a bootstrap routine where the parameter estimation was carried out via: (3) DNM of the ODLL function, (4) the exact EM algorithm, and (5) the approximate EM algorithm.

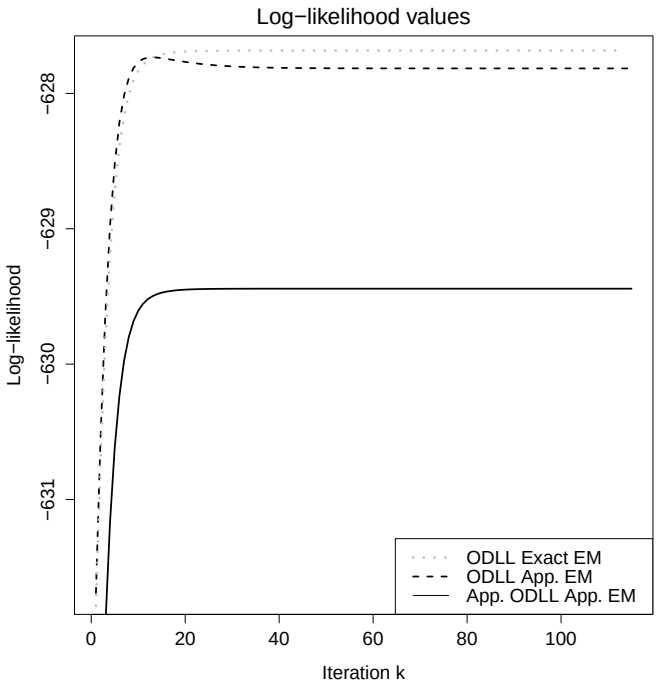


Figure 3.1: *Results from a small simulation study for the exact and approximate EM algorithms. The panel is a plot of the ODLL values at each iteration of the exact (grey dotted line) and approximate EM algorithms (dashed line), as well as the approximate ODLL values for the approximate EM algorithm (solid line). The starting values used to generate the plot are:* $\tau^{(0)} = 0.08$, $\psi^{(0)} = 0.025$, *and* $\gamma^{(0)} = 0.035$.

In both of these works, the results and conclusion suggest that the estimates found via the EM algorithm and DNM of the ODLL are different, when in fact they should both be the MLEs. It may be that the simulation study carried out by Halpin and De Boeck did little to ensure that the parameter-estimation routines converged to the global maximum, i.e. there was little oversight to ensure that MLEs were actually found. This could explain why their estimates found via EM and DNM are different. Olson and Carley (2013, p. 77) acknowledge explicitly that their DNM routine 'failed to converge in almost every instance it was tested with', which suggests that the estimates that they found via attempted DNM of the ODLL are simply not the MLEs.

We see DNM of the ODLL and the EM algorithm as alternative routes to the same estimates, the MLEs, and not as competing estimation methods. A conclusion that Halpin and De Boeck can perhaps reasonably draw from their simulation results is that, for the model that they consider, EM is more likely to converge to the MLEs than is DNM of the ODLL. We suspect that this conclusion may well be valid for the model they consider.

Table 3.1 also contains several sets of confidence intervals for each parameter. The first set are 95% confidence intervals of Wald type, and are found by using the approximate Hessian matrix supplied by `nlm`. They are calculated as MLE $\pm$ 1.96 $\times$ $\widehat{\text{s.e.}}$, where $\widehat{\text{s.e.}}$ is found from the approximate Hessian matrix. The second set of confidence intervals are approximate 'likelihood-based' 95% confidence intervals, and are calculated by using the profile likelihood of each parameter; see Pawitan (2001, pp. 35–41) for a description of likelihood-based confidence intervals. Figure 3.2 presents the profile likelihood for each of the parameters with the 95% confidence intervals indicated by the solid vertical lines. The likelihood-based confidence intervals may be preferable to the Wald-type confidence intervals as the approximate chi-squared distributions, upon which the likelihood-based CIs are constructed, are often quite accurate even when the normal approximations are unsatisfactory (McCullagh and Nelder, 1989, p. 473). Moreover, Pawitan (2001, p. 42) writes that if the two types of interval are not similar, the likelihood-based intervals are preferable. The remaining confidence intervals in Table 3.1 are found by using three different parametric bootstrap routines, one for each of the estimation methods. The bootstrap sample size

in each case is 2 500, and the confidence intervals are constructed by using the bias-corrected and accelerated (BC_a) method; see Efron and Tibshirani (1994, pp. 184–188) for details of the BC_a method. The MLEs and the same seed for the pseudo-random number generator are used for each of the three parametric bootstrap routines.

The true parameter values all fall within the confidence intervals. The confidence intervals presented for $\widehat{\tau}$ and $\widehat{\psi}$ are all very similar and any of the intervals would be suitable. The confidence intervals for $\widehat{\gamma}$ show some inconsistency. The likelihood-based confidence intervals and the bootstrap confidence intervals found by using DNM of the ODLL and the exact EM algorithm are asymmetric; see the likelihood-based interval for γ in Figure 3.2. It can also be seen that the likelihood-based confidence interval has a larger lower bound than the other intervals. This suggests that the Wald-type interval is not suitable for $\widehat{\gamma}$. In addition, the Wald-type interval is narrow and appears to understate the variability of the estimate of γ. The likelihood-based and bootstrap confidence intervals are preferable to the Wald-type intervals for $\widehat{\gamma}$. However, it should be noted that finding the bootstrap confidence intervals is substantially more computationally intensive than finding the likelihood-based confidence intervals.

The means of the bootstrap samples are also presented in Table 3.1. Those found for the bootstrap routines using DNM of the ODLL and the exact EM algorithm are close. The (very small) differences between the means are probably due to differences in the convergence criteria of the DNM routine and the EM algorithm, or due to the DNM routine and EM algorithm converging to different points for some of the bootstrap samples. From our experience, this last problem does not appear to be significant for the simple model we are considering. However, for the Hawkes process models considered by Halpin and De Boeck (2013) and Olson and Carley (2013), it is suspected that their DNM estimation routines tended to converge to the MLEs less frequently than did their EM algorithms.

The final set of results presented in Table 3.1 is the median time taken in seconds by each parameter-estimation routine to find the estimates. The times are for 50 replications of the particular parameter-estimation routine. Each of the replications used the starting values $\tau^{(0)} = 0.08$, $\psi^{(0)} = 0.025$, and $\gamma^{(0)} = 0.035$. The `microbenchmark` routine, in the R package of the

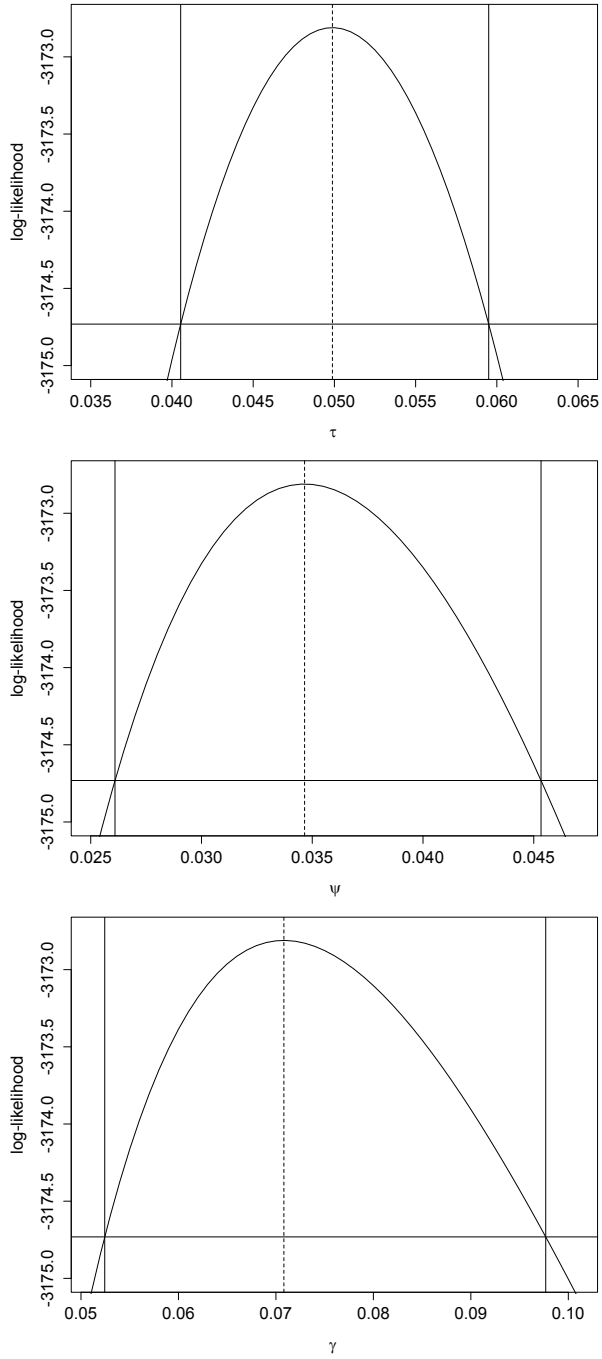


Figure 3.2: *Profile likelihoods for τ, ψ, and γ. The likelihood-based 95% confidence intervals are indicated by the solid vertical lines in each plot. The MLEs are indicated by the dashed vertical lines. The horizontal line in each plot is at the level $\ell(\widehat{\boldsymbol{\theta}}) - \chi^2_{1,0.95}/2 = \ell(\widehat{\boldsymbol{\theta}}) - 1.92$.*

same name, was used to measure the times (Mersmann, 2013). The absolute values of the times taken are not important as these will depend on the computer used to carry out the parameter estimation[2]. It is the relative times taken by the parameter-estimation routines which are of interest. It is clear that DNM of the ODLL is significantly faster than the exact and approximate EM algorithms, and that the approximate EM algorithm is marginally faster than the exact EM algorithm. These conclusions are different from those presented by Olson and Carley (2013, p. 78), who find that their exact and approximate EM algorithms are consistently and significantly faster than DNM of the ODLL. The model we are using here is much simpler than those investigated by Olson and Carley (2013), and this may be a likely reason why here DNM of the ODLL is faster.

The parameter-estimation method used for the remainder of this dissertation is DNM of the ODLL. One of the reasons is that it takes much less time to implement. Specifically, it does not require that the E and M steps be derived and programmed. In addition, the models investigated in our applications are unlikely to have M steps with analytic solutions, and so would either require a numerical maximisation at each iteration or some modification to the EM algorithm presented here.

3.5 Example: earthquake data

To illustrate parameter estimation via DNM of the ODLL for actual data, a small application to earthquake data is presented here. The data for this application are originally from Utsu (1982), and are provided by Ogata (1988, pp. 14–15). The data consist of the dates and magnitudes of large earthquakes that occurred in a region east of the main Japanese island of Honshū over a period of 96 years. The recorded earthquakes are those with magnitudes greater than or equal to six and with epicentres at a depth of less than 100km. Ogata considers several models for these data, and we compare our results to those presented by him.

A copy of the earthquake times, measured in days from the start of

[2]The computations here were carried out on a laptop computer with an Intel Core i7-2630QM CPU @ 2.00GHz, and 8GB of RAM, running the Windows 7 64-bit operating system.

1 January 1885, and their magnitudes is available from the R package `SMPracticals` (Davison, 2013), and is used here[3]. This package accompanies the practicals for the book by Davison (2003). There are 483 earthquakes recorded in this data set, which is roughly 5 earthquakes per year. The total length of the observation period is taken to be 35 063 days[4]. The data are plotted in Figure 3.3.

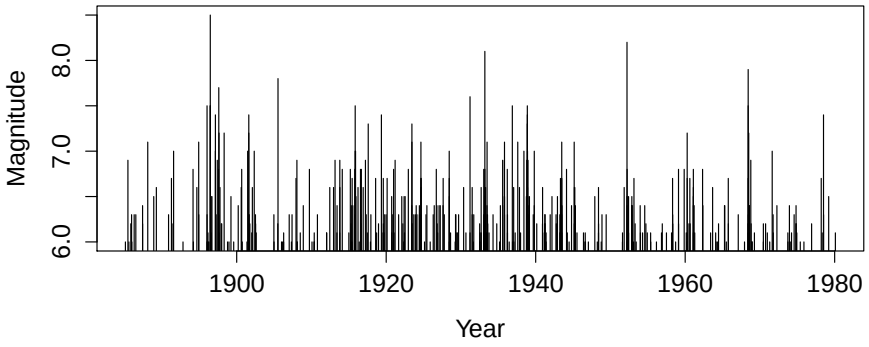


Figure 3.3: *Earthquake dates and magnitudes for the data taken from Ogata (1988, pp. 14–15).*

Three unmarked Hawkes process models are fitted to these data. Each of these models allows the observed earthquake magnitudes to affect the conditional intensity, but the earthquake magnitudes are treated as known and are not modelled. As such, this is may be viewed as a partial likelihood problem where we seek only to maximise the first part of the log-likelihood (3.1).

[3]Earthquakes 213 and 214 have the same recorded time in Ogata (1988, p. 14) and in `SMPracticals`. Earthquake 214 is treated as happening 1 minute after earthquake 213. A copy of this data set is also available from the R package `STAR` (Pouzat, 2012), but the earthquake times provided appear to be calculated from the start of the 9th of February 1885 and not the 1st of January 1885 as stated in the help documentation available at `http://cran.at.r-project.org/web/packages/STAR/STAR.pdf` (Accessed: 17 November 2013). Pouzat (2013) confirmed that the help documentation for the `STAR` package is inaccurate.

[4]Davison (2003, p. 288) uses 35 175 days for the length of the observation period. It is not clear where this number comes from. The length of the period that we use is calculated as follows: $96 \times 365 + 96/4 - 1 = 35\,063$.

The first model has the conditional intensity

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi \sum_{j:t_j\in(0,t)} \exp(\delta m_j - \gamma(t-t_j)),$$

where $\psi, \delta \geq 0$, $\tau, \gamma > 0$, and m_j is the magnitude of the jth earthquake. This model is referred to here as the exponential model.

The second model is a generalisation of the first model and has conditional intensity

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi \sum_{j:t_j\in(0,t)} (t-t_j)^{\zeta-1} \exp(\delta m_j - \gamma(t-t_j)), \tag{3.18}$$

where $\psi, \delta \geq 0$, τ, ζ and $\gamma > 0$. In this case, $\omega(s, m)$ has a form related to the gamma density and is equivalent to the first model when $\zeta = 1$. The form of $\omega(s, m)$ is similar to the decay function proposed by Otsuka (1985, 1987) for earthquake aftershock activity over time (as cited by Utsu *et al.* (1995)). Models with response functions related to the gamma density have also been considered by Halpin and De Boeck (2013) in their study of email communication. This model is referred to here as the gamma model.

The third model has conditional intensity

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi \sum_{j:t_j\in(0,t)} \frac{e^{\delta m_j}}{\gamma + t - t_j},$$

where $\psi, \delta \geq 0$, τ and $\gamma > 0$. This model is referred to here as the ETAS model. The first and third models are similar to models considered by Ogata for the data described above.

The models are fitted via DNM of the ODLL by using the `nlm` routine. The parameter estimates are reported in Table 3.2 along with the adjusted estimates for the ETAS model fitted by Ogata (1988, p. 18). The estimate of ψ from Ogata is adjusted by multiplying his estimate by $\exp(-6\widehat{\delta})$ to reflect that here the full magnitude of each earthquake is considered and not the truncated magnitude $m_i - 6$ as he considered.

The estimates found here for the ETAS model are close to the adjusted estimates of Ogata (1988, p. 18). Our minus ODLL for the exponential model is smaller than the value of 2248.0 reported by Ogata (1988, p. 17). The reason for this is not clear, and as parameter estimates for the exponential model are not reported by Ogata, a fuller comparison is difficult.

Table 3.2: *Parameter estimates for the exponential, gamma, and ETAS models.*

Model	$\widehat{\tau}$	$\widehat{\psi} \times 10^4$	$\widehat{\delta}$	$\widehat{\gamma}$	$\widehat{\zeta}$	$-\ell(\widehat{\boldsymbol{\theta}})$
Exponential	0.00979	0.03632	1.63932	0.62390	–	2243.4
Gamma	0.00776	0.01582	1.54612	0.01521	0.30351	2198.9
ETAS	0.00536	0.01077	1.61398	0.01969	–	2185.2
ETAS (Ogata, 1988)	0.00536	0.01077	1.61385	0.01959	–	2185.2

The ETAS model performs best on the basis of minus ODLL values, which is the same conclusion drawn by Ogata (1988, p. 17) who uses AIC for the models he considers. The gamma model performs better than the exponential model on the basis of minus ODLL values, which is due to the increased flexibility, and would, on the basis of AIC, outperform most of the models considered by Ogata (1988), but it does not outperform the ETAS model.

CHAPTER 4

Model selection and checking

4.1 Model selection

Once several families of models have been fitted to the data, the 'best' model has to be chosen by some criterion. This criterion needs to balance how well the models fit the data with the number of parameters in the models — the best model should be the simplest model that fits the data well. The introduction to model selection by Zucchini (2000) provides an overview of the theory underlying the criteria presented here.

The model selection criterion most often used in applications of point process models is the Akaike information criterion (AIC, Akaike, 1974, Guttorp and Thorarinsdottir, 2010). Examples of AIC being used in applications of Hawkes process models include the work of Ogata (1988), Wang *et al.* (2012), and Herrera (2013). In some applications of Hawkes process models other criteria are used to decide on the best model. For example, in an application to extreme asset returns and the forecasting of risk measures, Chavez-Demoulin and McGill (2012) use the results from a backtesting exercise to compare the models that they consider. In our applications to extreme asset returns, we use AIC to select the models used to forecast risk measures, and then, after performing some goodness-of-fit tests, we use a backtesting exercise to identify whether the models are suitable for forecasting risk measures. The goodness-of-fit tests that we use are described in the next section of this chapter, and the backtesting methods are presented in Section 6.2.

The AIC for a particular model provides a measure of the expected

discrepancy between the true underlying model which generates the observed data and the fitted model. As such, the model with the lowest AIC value is deemed the best model. The AIC value for a particular model and MLE $\widehat{\boldsymbol{\theta}}$ is calculated as

$$\text{AIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + 2p,$$

where p is the number of parameters in the model. The first term above gives a measure of the fit of the proposed model, and the second term penalises the inclusion of more parameters.

Of the three models fitted to the earthquake data at the end of Chapter 3, the ETAS model has the lowest AIC value at 4378.4. The AIC values for the exponential and gamma models are 4494.8 and 4407.8, respectively. It is clear that the increased flexibility of the gamma model is worthwhile when compared to the exponential model, but ultimately the ETAS model is the best of the three models considered according to AIC.

The use of AIC is a frequentist approach to model selection. The Bayes information criteria (BIC) is a model selection criteria which arises under the Bayesian framework (Schwarz, 1978). The BIC value for a particular model and MLE $\widehat{\boldsymbol{\theta}}$ is calculated as

$$\text{BIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + p\log(N_g(T)),$$

where p is as above and $N_g(T)$ is the number of observations. When using BIC, the best model is that with the lowest BIC value. BIC penalises the inclusion of further parameters more heavily than AIC when $N_g(T) > e^2$, and so in most cases it will favour simpler models. The BIC values for the models that we consider in Chapter 7 are not used for model selection, but are reported in Appendix B.1.

4.2 Model checking

Once the best model has been identified, its fit to the data needs to be checked. This is because the best model may still fail to capture important features of the data and a yet to be considered model may be better suited to the data (Ogata, 1988). The tests used to check the fit of the models are referred to as goodness-of-fit tests.

The goodness-of-fit tests that we describe are separated into tests for the 'temporal component' and tests for the conditional mark distribution of each model. The temporal component is the ground process of the marked point process model, the component which models the timing of the point events.

4.2.1 Goodness-of-fit tests for the temporal component

In the literature, the goodness-of-fit tests used for the temporal component of marked Hawkes process models predominantly make use of the random time change theorem. The random time change theorem involves changing the time index in such a way as to transform a point process N_g with conditional intensity $\lambda(\cdot|\widetilde{\mathcal{H}})$ into a unit-rate Poisson process. Goodness-of-fit tests can then be constructed to check whether or not various properties of the unit-rate Poisson process are present. Ogata (1988) presents several such goodness-of-fit tests, and we describe some of these tests here after first outlining the random time change theorem.

Random time change theorem

First we define the compensator function $\Lambda(t|\widetilde{\mathcal{H}}_t)$ of a point process N_g as

$$\Lambda(t|\widetilde{\mathcal{H}}_t) = \int_0^t \lambda(u|\widetilde{\mathcal{H}}_u)\,\mathrm{d}u.$$

Then according to Theorem 7.4.I. of Daley and Vere-Jones (2003, p. 258), for a point process N_g with conditional intensity $\lambda(\cdot|\widetilde{\mathcal{H}})$, if the time variable is rescaled by using the random time transformation $t \mapsto \Lambda(t|\widetilde{\mathcal{H}}_t)$, the transformed point process

$$\widetilde{N_g}(t) = N_g(\Lambda^{-1}(t|\widetilde{\mathcal{H}}_t))$$

is a unit-rate Poisson process. The implication of this theorem is that for a point process N_g with compensator $\Lambda(\cdot|\widetilde{\mathcal{H}})$ and observed point events at times $t_1, t_2, \ldots, t_{N_g(T)}$, the transformed point event times, $s_i = \Lambda(t_i|\widetilde{\mathcal{H}}_{t_i})$ for $i = 1, \ldots, N_g(T)$, will be a realisation from a unit-rate Poisson process. We refer to the transformed point process $s_1, s_2, \ldots, s_{N_g(T)}$ as the 'residual process', as is done by Ogata (1988).

The arguments underlying the goodness-of-fit tests presented by Ogata (1988) proceed along the following lines. The true underlying conditional intensity $\lambda(\cdot|\widetilde{\mathcal{H}})$ is unknown, and one will have to use the conditional intensity of the fitted model $\widehat{\lambda}(\cdot|\widetilde{\mathcal{H}})$ to perform the transformation. If the fitted model is a good approximation to the true underlying process, then the residual process is expected to closely resemble a unit-rate Poisson process. If the properties of the residual process depart significantly from those of a unit-rate Poisson process, then this is an indication that the fitted model is a poor approximation to the true underlying process. Tests can be constructed to assess whether the properties of the residual process differ significantly from those of a unit-rate Poisson process.

Description of some goodness-of-fit tests

Graphical goodness-of-fit tests can be constructed from the residual process. These tests are based on various quantities calculated from the residual process and give a qualitative indication of how well the model fits the data.

The inter-arrival times of the residual process are calculated as follows

$$\epsilon_{i+1} = s_{i+1} - s_i \quad \text{for } i = 1, \ldots, N_g(T) - 1. \tag{4.1}$$

These inter-arrival times should resemble realisations of iid exponential random variables with unit-mean if the model fits the data well. An exponential quantile-quantile plot (QQ-plot) can be constructed to assess whether this is the case.

Figure 4.1 presents two exponential QQ-plots for data simulated from a Hawkes process. The left-hand QQ-plot relates to the quantiles of the original inter-arrival times, and these are plotted against the quantiles of a reference exponential distribution. The right-hand QQ-plot relates to the quantiles of the inter-arrival times of the residual process, and these are plotted against the quantiles of a reference exponential distribution with unit-mean. The residual process is found by using the conditional intensity from which the data were simulated. The right-hand plot is the plot that we would be concerned with when checking the fit of a proposed model, and, unsurprisingly, it gives a strong indication that the inter-arrival times of the residual process are realisations from an exponential distribution with

unit-mean. The left-hand QQ-plot gives evidence against the original inter-arrival times being from an exponential distribution.

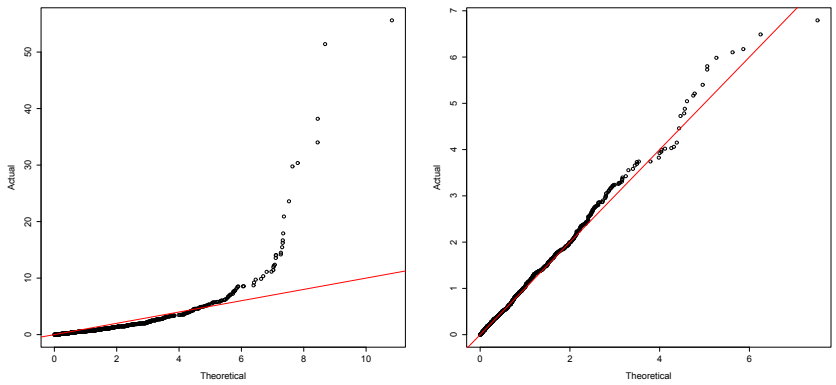


Figure 4.1: *Exponential QQ-plots for data simulated from a Hawkes process. The left-hand QQ-plot is for the inter-arrival times of the original simulated data, and the right-hand QQ-plot is for the inter-arrival times of the corresponding residual process.*

Ogata (1988) describes a test, attributed to Berman (1983), which can be used to test for independence between the inter-arrival times of the residual process. This test requires the quantities

$$U_i = 1 - \exp(-\epsilon_i),$$

for $i = 2, \ldots, N_g(T)$. The U_is should be iid uniform random variables on $(0, 1)$. Berman (1983) (as cited by Ogata (1988)) suggests that a plot of the points (U_i, U_{i+1}) can be constructed to check the independence the intervals, arguing that if there is any serial correlation, it is likely to be evident in consecutive intervals. If the fitted model is a good approximation to the true underlying process, then the plot of the points (U_i, U_{i+1}) should be a random scatter lying in the $(0, 1) \times (0, 1)$ space.

Goodness-of-fit tests can also be constructed directly from the residual process. For example, plots of the points (s_i, i), or $(s_i/\Lambda(T|\widetilde{\mathcal{H}}_T), i/N_g(T))$, are common in the literature. Examples of their use can be found in the work of Ogata (1988), Davison (2003, p. 290), Liniger (2009, p. 56), and Embrechts *et al.* (2011). Liniger (2009, pp. 55–56) and Davison (2003, pp. 327–328) describe the following test which is based on a plot of the points $(s_i/\Lambda(T|\widetilde{\mathcal{H}}_T), i/N_g(T))$.

If the fitted model is a good approximation to the true model, then, given the total number of point events, the scaled times from the residual process, $c_i = s_i/\Lambda(T|\widetilde{\mathcal{H}}_T)$ for $i = 1, \ldots, N_g(T)$, should be uniformly distributed over the interval $(0, 1)$. This is because, given the total number of point events from a homogeneous Poisson process in an interval, these point events are independently and uniformly distributed over the interval. We can use a Kolmogorov–Smirnov test to formally test whether the observed c_is are uniformly distributed on (0,1). Alternatively, we can plot and join the points $(c_i, i/N_g(T))$. If the fitted model is suitable, the resulting curve should go from $(0, 0)$ to $(1, 1)$ along the identity line. The confidence lines $y = x \pm d_{N_g(T),\alpha}$, where $d_{N_g(T),\alpha}$ is the $1 - \alpha$ quantile of the Kolmogorov–Smirnov statistic, can be added to the plot. For large n, the Kolmogorov–Smirnov statistic has 0.95 and 0.99 quantiles $d_{n,0.05} = 1.358/\sqrt{n}$ and $d_{n,0.01} = 1.628/\sqrt{n}$. If the curve breaches the confidence lines, this is interpreted as evidence against the null hypothesis that the model fits the data well.

Figure 4.2 gives an example of a plot based on the scaled times of the residual process for simulated Hawkes process data. The curve closely follows the identity line and does not breach the confidence lines, which gives us no reason to reject the model.

4.2.2 Goodness-of-fit tests for the conditional mark distributions

To investigate whether a proposed conditional mark distribution fits the observed data well, a graphical goodness-of-fit test can be constructed by using the estimated conditional mark distribution $\widehat{F}(\cdot)$. If the estimated conditional mark distribution $\widehat{F}(\cdot)$ is a close approximation to the true underlying conditional mark distribution, then the observed $\widehat{F}(m_i)$s, for $i = 1, 2, \ldots, N_g(T)$, should closely resemble realisations from iid uniform $(0, 1)$ random variables (Rosenblatt, 1952; Berkowitz, 2001). The goodness-of-fit test involves constructing a QQ-plot for the quantiles of the observed $\widehat{F}(m_i)$s, which are plotted against the quantiles of a reference uniform $(0, 1)$ distribution. The plot should be a straight line from $(0, 0)$ to $(1, 1)$, and any significant departures from the identity line indicate that the estimated conditional mark distribution is not suitable. A Kolmogorov–Smirnov test

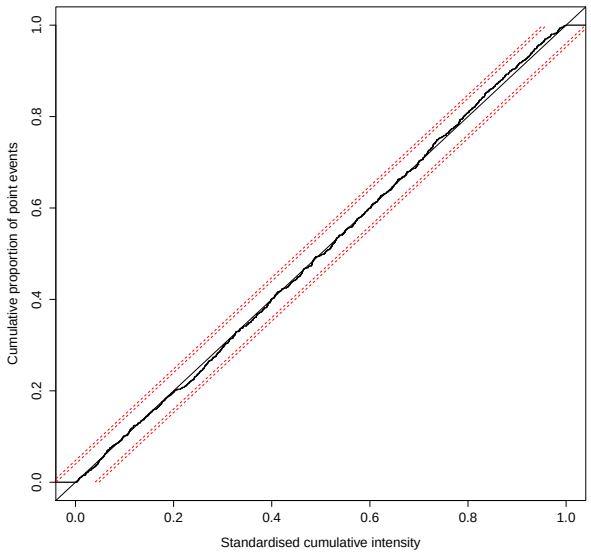


Figure 4.2: *A plot of the points* $(c_i, i/N_g(T))$ *(solid curve) for simulated Hawkes process data with 95% and 99% confidence lines (red dotted lines).*

can be performed to formally test the hypothesis that the $\widehat{F}(m_i)$s are from a uniform $(0, 1)$ distribution.

CHAPTER 5

Modelling extreme asset returns

5.1 Introduction

Marked Hawkes processes can be used to model extreme asset returns and this is an application that has been investigated in the literature. These applications include the work of Chavez-Demoulin *et al.* (2005), McNeil *et al.* (2005, pp. 306–311), Liniger (2009, pp. 48–56), Herrera and Schipp (2009, pp. 209–231), Embrechts *et al.* (2011), Chavez-Demoulin and McGill (2012), and Herrera (2013). The applications of Liniger (2009, pp. 48–56) and Embrechts *et al.* (2011) are mainly illustrative, and demonstrate how marked multitype Hawkes processes may be used to model extreme asset returns. The applications of McNeil *et al.* (2005, pp. 306–311), Chavez-Demoulin *et al.* (2005), Herrera and Schipp (2009, pp. 209–231), Chavez-Demoulin and McGill (2012), and Herrera (2013) are more practical. Most of their work investigates the suitability of marked Hawkes process models for forecasting market risk measures. The applications we present in Chapter 7 are along similar lines. We model extreme returns from three South African assets by using several marked Hawkes process models, and then investigate the suitability of the models to forecasting market risk measures by using a backtesting routine. The results from the backtesting routine are compared to the backtesting results for some nonstandard SV models to assess whether the marked Hawkes process models are competitive.

The motivation given in the literature for using particular marked Hawkes processes to model extreme asset returns is linked to the peaks-over-threshold (POT) method in extreme value theory (EVT). This motivation was origi-

nally presented by Chavez-Demoulin *et al.* (2005), and for completeness, a brief overview is presented in Section 5.2. Applications in the literature of marked Hawkes process models have mostly investigated models that have the same general form. The form of this general model, and the informal arguments used to motivate it, are presented in Section 5.3.

In Section 5.4 we define market risk and outline how the marked Hawkes process models can be used to forecast conditional VaR and ES. The marked Hawkes process models used in Chapter 7 are described in the final section of this chapter.

5.2 Motivation

5.2.1 Extracting a marked point process

We will be concerned with extreme negative returns, and for the remainder of this dissertation, we will mainly refer to losses and work with the right-hand tail of the loss distribution. A series of daily losses, which we will refer to as a loss series, is a time series and does not constitute a marked point process realisation to which we could reasonably fit a marked Hawkes process model. To extract a marked point process from a loss series, the extreme losses which exceed some high threshold are considered.

The notation X_t is used for the random (log-)loss on day t, where $t = 1, 2, \ldots$. The loss is calculated as $X_t = 100 \log(S_{t-1}/S_t)$, where S_t is the asset price at the end of day t. The threshold is denoted by u, and the excess random variable is denoted by $M_i = X_t - u$, given that $X_t > u$ and this is the ith exceedance. The convention of referring to the losses which exceed the threshold u as exceedances and referring to the differences between the exceedances and the threshold as excesses is adopted. The exceedances will also be referred to as extreme losses. The times of these exceedances constitute the times of the point events and the magnitudes of the excesses constitute the observed mark values.

For example, if the observed loss x_t exceeds the threshold u, and it is the ith observed exceedance, then the recorded mark value is $m_i = x_t - u$ and the time of the ith point event is t. Figure 5.1 depicts a marked point process realisation being extracted from a fictional observed loss series. The

time of the ith exceedance is indicated on the horizontal axis by t_i. The mark associated with the ith point event is denoted by m_i in the figure. The choice of the threshold u is discussed in the following subsection.

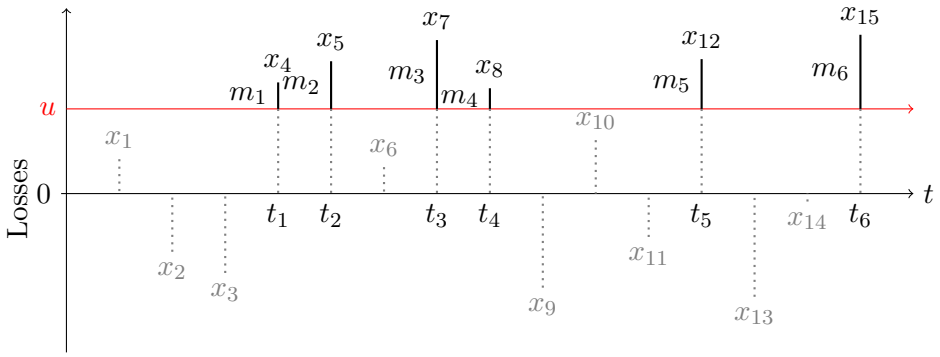


Figure 5.1: *An illustration of the observed loss series* $x_1, x_2, \ldots, x_{15}$ *and the threshold* u*. The threshold is used to extract the marked point process realisation* $(t_1, m_1), (t_2, m_2), \ldots, (t_6, m_6)$ *from the observed losses* $x_1, x_2, \ldots, x_{15}$*.*

The extracted marked point process realisation is a realisation from a discrete-time marked point process, as all of the t_is will be integers. This discrete-time marked point process will be modelled as if it were a continuous-time marked point process.

5.2.2 Overview of extreme value theory

EVT arguments are used to motivate the particular form of the marked Hawkes process models used to model extreme asset losses. These arguments were originally outlined by Chavez-Demoulin *et al.* (2005), and have been restated in many of the subsequent applications of marked Hawkes processes to extreme asset losses. The arguments are presented here for completeness; much of the material is standard in the EVT literature and is based on that presented by Embrechts *et al.* (1997) and Coles (2001).

Limit distribution for maxima

One of the main results in EVT is a limit distribution for the maxima of sequences of iid random variables. Consider a sequence $X_1, X_2, \ldots, X_n$ of independent random variables with an unknown common distribution function $F(\cdot)$. Suppose that this sequence represents the sequence of daily

losses. Let

$$M_n = \max\{X_1, X_2, \ldots, X_n\},$$

and suppose that x^F is the right endpoint of $F(\cdot)$, i.e. $x^F = \sup\{x \in \mathbb{R} : F(x) < 1\}$. Then for $z \in \mathbb{R}$,

$$\begin{aligned} \Pr\{M_n \leq z\} &= \Pr\{X_i \leq z, i = 1, \ldots, n\} \\ &= (F(z))^n. \end{aligned}$$

This is not helpful as $F(\cdot)$ is unknown and $(F(z))^n$ will degenerate as n becomes large,

$$\text{i.e. } (F(z))^n \to \begin{cases} 0 & \text{for } z < x^F, \\ 1 & \text{for } z \geq x^F, x^F < \infty, \end{cases} \quad \text{as } n \to \infty.$$

By normalising M_n, this type of degeneration can be avoided, i.e. for a suitable distribution function $F(\cdot)$, the distribution

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\},$$

where $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ are suitably chosen sequences of constants, will not degenerate as n increases.

The Fisher–Tippett theorem provides the limit distribution for the normalised maxima of sequences of iid random variables. Specifically, if sequences of constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ exist such that

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z) \quad \text{as } n \to \infty, \tag{5.1}$$

where $G(\cdot)$ is a non-degenerate distribution function, then $G(\cdot)$ is a member of the generalised extreme value (GEV) family of distributions. The GEV family of distributions have distribution functions which are given by

$$G(z) = \begin{cases} \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\beta}\right)\right]^{-1/\xi}\right\} & \text{if } \xi \neq 0, \\ \exp\left\{-\exp\left(-\frac{z-\mu}{\beta}\right)\right\} & \text{if } \xi = 0, \end{cases} \tag{5.2}$$

where μ, $\xi \in \mathbb{R}$ and $\beta > 0$. The support of $G(z)$ is

$$\begin{aligned} z > -\frac{\beta}{\xi} + \mu \quad &\text{for} \quad \xi > 0, \\ z < -\frac{\beta}{\xi} + \mu \quad &\text{for} \quad \xi < 0, \text{ and} \\ z \in \mathbb{R} \quad &\text{for} \quad \xi = 0. \end{aligned}$$

Embrechts *et al.* (1997, p. 122) provide a sketch of the proof for this result. A distribution function $F(\cdot)$ which satisfies (5.1), where $G(\cdot)$ is non-degenerate, is said to belong to the 'maximum domain of attraction' of the GEV family of distributions. The characterisation of the maximum domain of attraction of the standard GEV distribution is given by Embrechts *et al.* (1997, p. 158).

Peaks-over-threshold method

Instead of modelling the maxima, we can model the magnitudes of the excesses of a series of independent random variables above a high threshold. A limit distribution for the excesses exists and is outlined below.

Let $X_1, X_2, \ldots, X_n$ be a sequence of iid random variables with common distribution function $F(\cdot)$. Again the sequence $X_1, X_2, \ldots, X_n$ represents the sequence of daily losses. Suppose that $F(\cdot)$ belongs to the maximum domain of attraction of $G(\cdot)$. Then the approximate distribution of $M = X_t - u$, given $X_t > u$ and u is large, is given by

$$H(m) = \begin{cases} 1 - \left(1 + \xi \frac{m}{\tilde{\beta}}\right)^{-1/\xi} & \text{if } \xi \neq 0, \\ 1 - \exp\left(-\frac{m}{\beta}\right) & \text{if } \xi = 0, \end{cases} \tag{5.3}$$

for

$$\begin{aligned} m &\geq 0 && \text{if } \xi \geq 0, \\ 0 \leq m &\leq -\tilde{\beta}/\xi && \text{if } \xi < 0, \end{aligned}$$

where $\tilde{\beta} = \beta + \xi(u - \mu)$. The distribution (5.3) can be justified as the limiting distribution with $u \uparrow x^F$; see Embrechts *et al.* (1997, pp. 158–160, Theorem 3.4.13 (b), pp. 165–166). Informally, the above result states that if the distribution function $F(\cdot)$ satisfies the conditions for (5.1), where $G(\cdot)$ is non-degenerate, then the approximate distribution for the magnitudes of the excesses, given a suitably high threshold, is the generalised Pareto distribution (GPD) given by (5.3). The parameter ξ is referred to as the shape parameter and $\tilde{\beta}$ as the scale parameter. The GPD can also have a location parameter μ similar to the GEV distribution. Theorem 3.4.13 of Embrechts *et al.* (1997, p. 165) provides useful properties of the GPD.

The modelling of the excesses can be framed as a marked Poisson process. This particular method is called the POT method. For a high threshold, the timing of the exceedances on a rescaled time domain $(0, 1]$ can be shown

to approximately follow a homogeneous Poisson process with intensity

$$\lambda = \begin{cases} \left(1 + \xi \frac{u-\mu}{\beta}\right)^{-1/\xi} & \text{if } \xi \neq 0, \\ \exp\left(-\frac{u-\mu}{\beta}\right) & \text{if } \xi = 0. \end{cases} \tag{5.4}$$

The magnitudes of the excesses, the marks, are iid GPD random variables with density function

$$f(m) = \begin{cases} \frac{1}{\tilde{\beta}} \left(1 + \xi \frac{m}{\tilde{\beta}}\right)^{-1/\xi - 1} & \text{if } \xi \neq 0, \\ \beta^{-1} e^{-m/\beta} & \text{if } \xi = 0, \end{cases} \tag{5.5}$$

for

$$\begin{array}{ll} m \geq 0 & \text{if } \xi \geq 0, \\ 0 \leq m \leq -\tilde{\beta}/\xi & \text{if } \xi < 0. \end{array}$$

The result used to justify (5.4) is a limit result; see Theorem 5.3.2 of Embrechts *et al.* (1997, pp. 238–240). Coles (2001, pp. 128–132) and McNeil *et al.* (2005, pp. 298–302) provide justification for this marked Poisson process model for the timing and magnitudes of the excesses. This model is the marked point process model upon which the marked Hawkes process models are based. The case where $\xi < 0$ is not considered as it is convenient to allow the losses to potentially be unbounded.

Before developing this marked point process further, the choice of the threshold is discussed briefly. One means of choosing the threshold is by using a plot of the empirical mean excess function; see Embrechts *et al.* (1997, pp. 355–356) and Coles (2001, pp. 78–80, 83). The mean excess function is defined as

$$e(u) = \mathrm{E}(X - u | X > u),$$

and the empirical mean excess function is defined as

$$e_n(u) = \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \quad \text{for } 0 < u < \tilde{x},$$

where $x_{(1)}, x_{(2)}, \ldots, x_{(n_u)}$ are the n_u observed losses that exceed u and $\tilde{x}$ is the largest observed loss. For a GPD with shape parameter ξ and scale parameter β,

$$e(u) = \frac{\beta + \xi u}{1 - \xi},$$

provided $\xi < 1$. As the mean excess function is linear in u for such a GPD, a threshold can be chosen by identifying a value of u such that the plotted empirical mean excess function is approximately linear above that value. The appropriate choice of u is unlikely to be unique, as interpreting plots of the empirical mean excess function and the meaning of 'approximately linear' can be difficult, especially as for large values of u, the empirical mean excess function can be sensitive to changes in u (Embrechts *et al.*, 1997, pp. 355–356). The number of exceedances must also be considered when choosing the threshold; if the threshold is set too high, there will be few exceedances and the variances of the parameter estimates will be large; if the threshold is set too low, the above limit distributions will not apply and the estimated model will be biased (Coles, 2001, p. 78). In our applications, we set the threshold equal to a high percentile of the observed losses, and then examine a plot of the empirical mean excess function.

5.3 Marked Hawkes process models for extreme asset returns

The marked Poisson process model given by (5.4) and (5.5) is based on the assumption that each of the random variables in the observed sequence is iid. Observed asset loss series, and equivalently asset return series, such as those from a share, typically display characteristics which lead us to conclude that the individual losses are not iid. Stochastic volatility, and the resulting clusters of extreme losses, is one such characteristic that asset loss series typically display (McNeil *et al.* 2005, pp. 117–123, Taylor 2005, p. 14). The clustering of extreme losses would invalidate modelling the timing of the exceedances by a homogeneous Poisson process. This is a point made by Chavez-Demoulin *et al.* (2005), and is also made in other applications of marked Hawkes processes to extreme losses.

Figure 5.2 provides evidence against the use of a homogeneous Poisson process to model the timing of extreme losses on the South African All Share Index (ALSI). The panel presented is a QQ-plot for the inter-arrival times between extreme daily losses. The quantiles of the inter-arrival times are plotted against the quantiles of a reference exponential distribution. The losses are for the period 8 December 1999 to 12 June 2012, and the extreme

losses are those that exceed the 90th percentile of the daily losses for this period. If the exceedance times were a realisation from a homogeneous Poisson process, then we would expect the inter-arrival times to be realisations of iid exponential random variables. The QQ-plot suggests that the inter-arrival times are not exponentially distributed. There are too many short inter-arrival times relative to the reference exponential distribution, suggesting that the exceedances tend to cluster together. The use of a homogeneous Poisson process is not appropriate in this case. Applications in the literature of marked Hawkes process models to extreme asset losses typically present similar evidence when demonstrating that a homogeneous Poisson process is not suitable. See, for example, the QQ-plots presented by Chavez-Demoulin *et al.* (2005, p. 228) and Herrera (2013, p. 66).

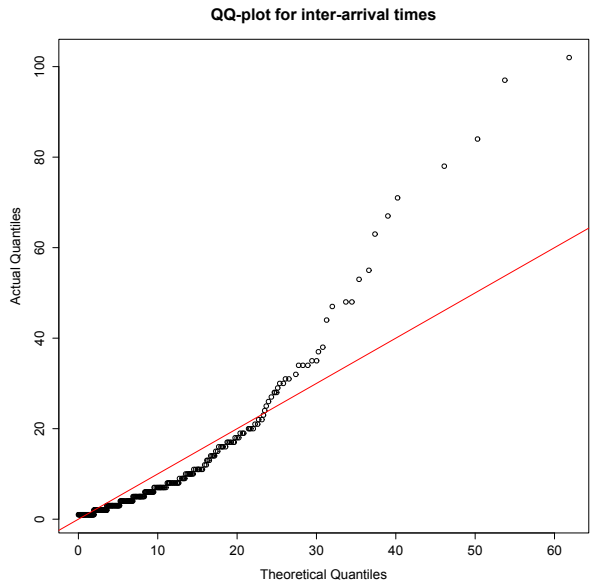


Figure 5.2: *A QQ-plot for the inter-arrival times between extreme daily losses on the ALSI. The quantiles of the inter-arrival times are plotted against the quantiles of a reference exponential distribution.*

The approach taken in the marked Hawkes process literature to handle the clustering of extreme losses is to use a Hawkes process to model the temporal behaviour of the exceedances instead of a homogeneous Poisson process. In this way, the self-exciting nature of the Hawkes process can be used to model the observed clustering of extreme losses. Chavez-Demoulin

et al. (2005) were the first to propose such a model, and they demonstrated that a Hawkes process can adequately mimic the observed clustering of extreme losses. The models proposed by Chavez-Demoulin *et al.* (2005) have subsequently been developed further in the literature. Most of the models used in the literature have forms similar to the model with predictable marks proposed by McNeil *et al.* (2005, pp. 306–309). For example, Herrera and Schipp (2009, pp. 209–231), Chavez-Demoulin and McGill (2012), and Herrera (2013) all use models similar to that proposed by McNeil *et al.* (2005, pp. 306–309).

The form of the model proposed by McNeil *et al.* (2005, pp. 306–309) is as follows. The conditional intensity of the ground process N_g used to model the timing of the exceedances is given by

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi \sum_{j:t_j \in (0,t)} \omega(t - t_j, m_j), \tag{5.6}$$

where $\tau > 0$, $\psi \geq 0$, and $\omega(s, m) \geq 0$ for $s \geq 0$, $m \geq 0$ and zero otherwise. Note that we have made a small change to $\omega(s, m)$. The function $\omega(s, m)$ no longer includes the factor ψ, which we now explicitly included in the conditional intensity. Explicit forms for $\omega(s, m)$ are discussed in Section 5.5.1. The conditional intensity (5.6) effectively replaces the intensity (5.4) of the homogeneous Poisson process.

The conditional mark distribution proposed by McNeil *et al.* (2005, pp. 306–309) is a generalisation of the GPD given in (5.5). The conditional density function of this GPD, given an event at time t and the history of the marked point process $\widetilde{\mathcal{H}}_t$, is given by

$$f(m) = \frac{1}{\beta + \alpha v(t|\widetilde{\mathcal{H}}_t)} \left(1 + \xi \frac{m}{\beta + \alpha v(t|\widetilde{\mathcal{H}}_t)}\right)^{-1/\xi - 1} \quad \text{for } m \geq 0, \tag{5.7}$$

where $\beta > 0$, $\alpha, \xi \geq 0$, and

$$v(t|\widetilde{\mathcal{H}}_t) = \sum_{j:t_j \in (0,t)} \omega(t - t_j, m_j).$$

The scale parameter of this conditional GPD is dependent on the history of the marked point process, and so this marked Hawkes process can be described as having predictable marks. The model consisting of conditional intensity (5.6) and conditional mark distribution (5.7) has some intuitive

appeal in the context of modelling extreme losses which we discuss below. This model is sometimes referred to as the self-exciting POT model in the literature, but for brevity we refer to it as Model h. It is the general form of the models that we consider in our applications in Chapter 7.

Model h has intuitive appeal as it can mimic features that may be expected of asset loss series. The model allows the observed mark values to affect the conditional intensity. The choice of $\omega(\cdot, \cdot)$ is usually such that the increase in the conditional intensity following an exceedance is greater for larger excesses. The intuitive appeal of this is that shortly after an extreme loss one may expect that the likelihood of further extreme losses will increase, and for this increase to be greater when the observed extreme loss was large. The model can thus mimic the clustering of extreme losses that may be expected in times of market excitement (McNeil *et al.*, 2005, p. 308, Chavez-Demoulin and McGill, 2012). The conditional mark distribution is dependent on the past of the process through the scale parameter $\beta + \alpha v(t|\widetilde{\mathcal{H}}_t)$. The effect of using such a scale parameter is that as the self-excitement function $v(t|\widetilde{\mathcal{H}}_t)$ increases, the scale parameter of the conditional mark distribution increases, and the magnitude of any resulting excess is more likely to be large. The intuitive appeal of this is that one may expect that during periods of market excitement, the magnitudes of losses are more likely to be large (McNeil *et al.*, 2005, p. 308, Chavez-Demoulin and McGill, 2012).

The scale parameter in the conditional GPD given by (5.7) is not the only means by which effects from the past of the process can be included in the conditional mark distribution. In fact, Embrechts *et al.* (1997, p. 367) note that non-stationary effects may be included in all of the parameters of a GPD, which makes it attractive. In the context of marked Hawkes processes, an alternative method suggested by Chavez-Demoulin *et al.* (2005, p. 231), but found not to be useful for the loss series that they consider, is to have $\log \beta_i = a + b m_{i-1}$, where β_i is the scale parameter of the GPD for the ith excess. The marks in this case are a first order Markov process, and, for $b > 0$, the model has similar intuitive appeal to that of Model h. As for the parameter ξ, Chavez-Demoulin *et al.* (2005) and Chavez-Demoulin and McGill (2012) conclude from empirical analyses that leaving ξ as a constant is reasonable. Herrera (2013, p. 67) attempts to replace ξ by a dynamic

alternative in his applications, but notes that the estimation of the model is 'severely affected', and as a consequence reverts to a model with constant ξ.

In the literature, the marked Hawkes processes used to model the timing and magnitudes of exceedances are estimated via DNM of the ODLL. The models used in our applications in Chapter 7 are also estimated via DNM of the ODLL. For the observations $(t_1, m_1), (t_2, m_2), \ldots, (t_{N_g(T)}, m_{N_g(T)})$ from a period $[0, T]$, the ODLL for Model h is given by

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N_g(T)} \log \lambda(t_i|\widetilde{\mathcal{H}}_{t_i}) - \tau T - \psi \sum_{i=1}^{N_g(T)} \int_{t_i}^{T} \omega(u - t_i, m_i)\,\mathrm{d}u + \sum_{i=1}^{N_g(T)} \log \left(\frac{1}{\beta + \alpha v(t_i|\widetilde{\mathcal{H}}_{t_i})} \left(1 + \xi \frac{m_i}{\beta + \alpha v(t_i|\widetilde{\mathcal{H}}_{t_i})} \right)^{-1/\xi - 1} \right). \quad (5.8)$$

As there are parameters shared between the conditional intensity and the conditional mark distribution, the numerical maximisation of (5.8) cannot be broken into two separate parts for this general model as is suggested in Section 3.2.2.

5.4 Forecasting market risk measures

Once the parameters have been estimated for the marked Hawkes process models, they can be used to forecast conditional measures of market risk. McNeil *et al.* (2005, pp. 2–3) define market risk as

> the risk of a change in the value of a financial position due to changes in the value of the underlying components on which that position depends, such as stock and bond prices, exchange rates, commodity prices, etc.

Measuring market risk is important to many financial institutions and investors. Some financial institutions are governed by regulation which requires that they submit measures of their market risk on a regular basis. For example, banks that are required to follow the Basel Capital Adequacy Framework (Basel Framework) have to assess and submit measures of their market risk on a regular basis; see the requirements set out by the Basel Committee on Banking Supervision (BCBS, 2011, pp. 13–16, 25–26). These market risk measures are also used to determine the capital requirements

for such banks; see the capital calculation outlined by the BCBS (2011, p. 15). In addition, market risk measures can be used in setting limits and controlling risk within a financial institution (McNeil *et al.*, 2005, p. 34).

Two popular market risk measures are conditional VaR and conditional ES. Marked Hawkes process models have been shown to provide adequate forecasts of conditional VaR in several applications in the literature, e.g. in the work of Chavez-Demoulin *et al.* (2005) and Chavez-Demoulin and McGill (2012). The models have also been used to forecast conditional ES, e.g. in the work of Herrera and Schipp (2009, pp. 209–231) and Chavez-Demoulin and McGill (2012).

5.4.1 Definitions of conditional VaR and ES

VaR is a measure that is often used to illustrate the market risk associated with holding a portfolio of assets over a future period. Informally, VaR is the maximum loss on a portfolio that will occur over a given future period for a given level of confidence. Formally, conditional VaR for one day ahead can be defined as follows. Consider the real-valued discrete-time stochastic process $\{X_t, t = 1, 2, \ldots\}$ representing the daily losses on an asset or portfolio of assets. Let $F_{X_{t+1}|\widetilde{\mathcal{H}}_t^x}(\cdot)$ be the continuous conditional forecast distribution of the loss on day $t+1$, given the observed history of the losses up to and including day t, $\widetilde{\mathcal{H}}_t^x$. The conditional VaR for this portfolio for day $t+1$ at the $100\phi\%$ confidence level is given by

$$\mathrm{VaR}_\phi(X_{t+1}) = \inf\{x \in \mathbb{R} : F_{X_{t+1}|\widetilde{\mathcal{H}}_t^x}(x) \geq \phi\},$$

where $\phi \in (0, 1)$. The conditional VaR is therefore a quantile of the conditional forecast distribution being used. VaR is the market risk measure that the BCBS (2011, p. 13) currently requires banks to use, and we will interpret this as a requirement to use conditional VaR.

VaR has several drawbacks as a measure of market risk. VaR is not a coherent risk measure in general, as it lacks subadditivity in certain circumstances (Artzner *et al.*, 1999, McNeil *et al.*, 2005, p. 239). A risk measure is described as coherent if it satisfies the axioms set out by Artzner *et al.* (1999). These axioms are properties that are 'desired' of risk measures. The lack of subadditivity has some important implications if VaR is used as a

measure of market risk. For example, one such implication is that diversifying a portfolio of assets may not result in a lower overall risk level when risk is judged by VaR. Several other implications are discussed by Artzner *et al.* (1999). Another problem with VaR is that it does not give an indication of the likely size of a loss given that the loss exceeds the VaR level. These and other deficiencies of VaR have led to alternative market risk measures being considered.

ES is one such alternative risk measure. ES is an attractive risk measure as it provides a measure of the size of the loss given that it exceeds the VaR level and as it is a coherent risk measure (Acerbi and Tasche, 2002). As a result, ES has become the preferred risk measure amongst many risk managers (McNeil *et al.*, 2005, p. 44) and the BCBS (2012, p. 20) is considering moving to ES as their required market risk measure.

The conditional ES for day $t+1$ at the $100\phi\%$ confidence level is given by

$$\mathrm{ES}_\phi(X_{t+1}) = \frac{\int_\phi^1 \mathrm{VaR}_u(X_{t+1})\,\mathrm{d}u}{1-\phi},$$

where $\mathrm{VaR}_u(X_{t+1})$ is defined above and $\phi \in (0,1)$. For a continuous conditional distribution function $F_{X_{t+1}|\widetilde{\mathcal{H}}_t^x}(\cdot)$,

$$\mathrm{ES}_\phi(X_{t+1}) = \mathrm{E}\left(X_{t+1}\,\middle|\,X_{t+1} \geq \mathrm{VaR}_\phi(X_{t+1}), \widetilde{\mathcal{H}}_t^x\right); \tag{5.9}$$

see Acerbi and Tasche (2002, p. 1498).

5.4.2 Forecasting conditional VaR and ES

In this section we discuss how Model h can be used to forecast conditional VaR and ES for a period of one day. The formulae for the other marked Hawkes process models that we consider in our applications are specialisations of those presented here.

Conditional VaR

To find the conditional $100\phi\%$ VaR, we need to solve for $\mathrm{VaR}_\phi(X_{t+1})$ in

$$\Pr\{X_{t+1} \geq \mathrm{VaR}_\phi(X_{t+1})|\widetilde{\mathcal{H}}_t^x\} = 1-\phi. \tag{5.10}$$

The following approach to finding $\mathrm{VaR}_\phi(X_{t+1})$ is close to that taken by Chavez-Demoulin *et al.* (2005), and is slightly different to the approaches

of McNeil *et al.* (2005, p. 309), Herrera and Schipp (2009, p. 217), Chavez-Demoulin and McGill (2012), and Herrera (2013). Provided ϕ is sufficiently large, the left-hand side of Equation (5.10) can be manipulated as follows:

$$\begin{aligned} & \Pr\{X_{t+1} \geq \mathrm{VaR}_\phi(X_{t+1}) | \widetilde{\mathcal{H}}_t^x\} \\ = \; & \Pr\{X_{t+1} - u \geq \mathrm{VaR}_\phi(X_{t+1}) - u | X_{t+1} > u, \, \widetilde{\mathcal{H}}_t^x\} \\ & \qquad \times \Pr\{X_{t+1} > u | \widetilde{\mathcal{H}}_t^x\} \qquad (5.11) \\ \approx \; & \Pr\{M_{i+1} \geq \mathrm{VaR}_\phi(X_{t+1}) - u | \, \widetilde{\mathcal{H}}_{t+}, t_{i+1}\} \\ & \qquad \times \Pr\{X_{t+1} > u | \widetilde{\mathcal{H}}_t^x\}. \qquad (5.12) \end{aligned}$$

The line marked (5.11) follows if $\mathrm{VaR}_\phi(X_{t+1}) \geq u$, which is the case if ϕ is large enough. In the line marked (5.12), we substitute the excess $X_{t+1} - u$ for the mark random variable M_{i+1} with density given by (5.7), and suppose that it is the $(i+1)$st observed exceedance. The equality is approximate as we are moving from the discrete-time process for the asset losses to the continuous-time marked Hawkes process for the exceedances. The time $t+$ denotes a time just after time t so that $\widetilde{\mathcal{H}}_{t+}$ is the observed history of the marked point process up to and including time t. This allows an extreme loss on day t to impact the forecast conditional VaR for day $t+1$.

The second factor in line (5.12) is the conditional probability of the loss over day $t+1$ exceeding u. This conditional probability can be approximated under the marked Hawkes process model by the conditional probability of at least one point event occurring over the period $(t, t+1]$;

$$\begin{aligned} \text{i.e. } \Pr\{X_{t+1} > u | \widetilde{\mathcal{H}}_t^x\} \; & \approx \; 1 - \Pr\{N_g(t, t+1] = 0 | \widetilde{\mathcal{H}}_{t+}\} \qquad (5.13) \\ & = \; 1 - \exp\left(- \int_0^1 \lambda(t+s | \widetilde{\mathcal{H}}_{t+}) \, \mathrm{d}s \right). \qquad (5.14) \end{aligned}$$

In (5.13) we are again moving from the discrete-time process for the asset losses to the continuous-time marked Hawkes process for the extreme losses. The approximation arises because under the marked Hawkes process model we can have more than one exceedance per day. The approximation is justified by arguing that for a sufficiently short time interval, the conditional probability of a point event occurring in that interval is approximately equal to the conditional probability of at least one point event occurring.

By using Equations (5.12) and (5.14), and solving for $\mathrm{VaR}_\phi(X_{t+1})$ in Equation (5.10), we can show that the approximate conditional $100\phi\%$ VaR

is given by

$$\text{VaR}_\phi(X_{t+1}) \approx F^{-1}\left(1 - \frac{1-\phi}{1-\exp\left(-\int_0^1 \lambda(t+s|\widetilde{\mathcal{H}}_{t+})\,\mathrm{d}s\right)}\right) + u, \quad (5.15)$$

where $F^{-1}(\cdot)$ is the inverse of the conditional mark distribution function. The approximate $\text{VaR}_\phi(X_{t+1})$ in (5.15) is defined for

$$\exp\left(-\int_0^1 \lambda(t+s|\widetilde{\mathcal{H}}_{t+})\,\mathrm{d}s\right) \leq \phi. \quad (5.16)$$

This places an additional lower bound on the available confidence levels, i.e. in addition to the lower bound required in line (5.11).

If we use the conditional mark distribution with density given by (5.7), the approximate conditional $100\phi\%$ VaR under Model h is given by

$$\text{VaR}_\phi(X_{t+1}) \approx \frac{\beta + \alpha v(t+|\widetilde{\mathcal{H}}_{t+})}{\xi}\left[\left(\frac{1-\exp\left(-\int_0^1 \lambda(t+s|\widetilde{\mathcal{H}}_{t+})\,\mathrm{d}s\right)}{1-\phi}\right)^{\xi} - 1\right] + u. \quad (5.17)$$

The above approach to finding $\text{VaR}_\phi(X_{t+1})$ is slightly different from that of McNeil *et al.* (2005, p. 309), Herrera and Schipp (2009, p. 217), Chavez-Demoulin and McGill (2012, p. 3422)[1], and Herrera (2013, p. 67). The above authors effectively approximate the conditional probability $\Pr\{X_{t+1} > u|\widetilde{\mathcal{H}}_t^x\}$ by $\lambda(t+|\widetilde{\mathcal{H}}_{t+})$. A disadvantage of using such an approximation is that $\lambda(\cdot|\widetilde{\mathcal{H}})$ can be greater than one. Practical difficulties which might arise when $\lambda(\cdot|\widetilde{\mathcal{H}})$ is greater than one are not explicitly mentioned in the literature.

Conditional ES

The derivation of conditional $100\phi\%$ ES given here is similar to that outlined by McNeil *et al.* (2005, p. 309).

The conditional $100\phi\%$ ES can be found by using the approximate value for $\text{VaR}_\phi(X_{t+1})$ above and expression (5.9) for the conditional ES. If ϕ is

[1] Chavez-Demoulin and McGill (2012, p. 3422) appear to switch between the two approaches to approximating $\Pr\{X_{t+1} > u|\widetilde{\mathcal{H}}_t^x\}$ in their derivation of $\text{VaR}_\phi(X_{t+1})$. Their final expression for $\text{VaR}_\phi(X_{t+1})$ is consistent with them using $\Pr\{X_{t+1} > u|\widetilde{\mathcal{H}}_t^x\} \approx \lambda(t+|\widetilde{\mathcal{H}}_{t+})$.

such that $\mathrm{VaR}_\phi(X_{t+1}) \geq u$, we can manipulate the right-hand side of (5.9) as follows:

$$\begin{aligned} \mathrm{ES}_\phi(X_{t+1}) &= \mathrm{E}(X_{t+1}|X_{t+1} \geq \mathrm{VaR}_\phi(X_{t+1}), \widetilde{\mathcal{H}}_t^x) \\ &\approx \mathrm{E}(M_{i+1}|M_{i+1} \geq \mathrm{VaR}_\phi(X_{t+1}) - u, \widetilde{\mathcal{H}}_{t+}, t_{i+1}) + u. \end{aligned}$$

In the final line above, we have substituted $X_{t+1} - u$ for M_{i+1} as is done for conditional VaR. This is possible as $\mathrm{VaR}_\phi(X_{t+1}) \geq u$, and the equality is again approximate as we are moving from the discrete-time process for the asset losses to the continuous-time marked Hawkes process for the extreme losses.

Under Model h, the conditional survivor function of M_{i+1}, given $M_{i+1} \geq \mathrm{VaR}_\phi(X_{t+1}) - u$ and the history of the marked point process up to and including time t, is given by

$$\begin{aligned} &\Pr(M_{i+1} \geq x|M_{i+1} \geq \mathrm{VaR}_\phi(X_{t+1}) - u, \widetilde{\mathcal{H}}_{t+}, t_{i+1}) \\ &\qquad = \left(1 + \frac{\xi[x - (\mathrm{VaR}_\phi(X_{t+1}) - u)]}{\beta + \alpha v(t + |\widetilde{\mathcal{H}}_{t+}) + \xi(\mathrm{VaR}_\phi(X_{t+1}) - u)}\right)^{-1/\xi}, \end{aligned} \quad (5.18)$$

for $x \geq \mathrm{VaR}_\phi(X_{t+1}) - u$. This is the survivor or tail function of a GPD with location parameter $\mathrm{VaR}_\phi(X_{t+1}) - u$, scale parameter $\beta + \alpha v(t + |\widetilde{\mathcal{H}}_{t+}) + \xi(\mathrm{VaR}_\phi(X_{t+1}) - u)$ and shape parameter ξ. As the distribution in (5.18) is a GPD, we can use the expression for the mean of a GPD to find

$$\mathrm{E}\left(M_{i+1} \,\middle|\, M_{i+1} \geq \mathrm{VaR}_\phi(X_{t+1}) - u, \widetilde{\mathcal{H}}_{t+}, t_{i+1}\right) = \frac{\beta + \alpha v(t + |\widetilde{\mathcal{H}}_{t+}) + \mathrm{VaR}_\phi(X_{t+1}) - u}{1 - \xi},$$

provided $\xi < 1$. Thus, the conditional $100\phi\%$ ES under Model h is given by

$$\mathrm{ES}_\phi(X_{t+1}) \approx \frac{\beta + \alpha v(t + |\widetilde{\mathcal{H}}_{t+}) + \mathrm{VaR}_\phi(X_{t+1}) - \xi u}{1 - \xi},$$

provided $\xi < 1$.

5.5 Models used in Chapter 7

Here we describe the response functions that we use in our applications, as well as the models that we consider. The models described in the Sections

5.5.2 and 5.5.3 are separated according to their conditional mark distributions. The first group have conditional mark distributions that are generalised Pareto and the second group have conditional mark distributions that are exponential. These models, including Model h, are the models that we investigate in our applications in Chapter 7. Figure 5.4 at end of the chapter provides a summary of the models.

5.5.1 Response functions

The two main decay functions that we consider in our applications are the exponential and power decay functions, each of which include the mark impact $e^{\delta m}$. Both of these decay functions are popular in the literature. The exponential decay function is given by

$$\omega(t, m) = \exp(\delta m - \gamma t),$$

where $\delta \geq 0$, and $\gamma > 0$, and the power decay function is given by

$$\omega(t, m) = \frac{e^{\delta m}}{(t + \gamma)^{\eta+1}}, \tag{5.19}$$

where $\delta \geq 0$, and $\gamma, \eta > 0$. To identify the models that use these different decay functions, we attach '-exp' and '-pow' to the names of the models. For example, Model h-exp and Model h-pow are the two versions of Model h.

In our applications, we experienced difficulty when trying to estimate models with conditional intensities that use the power decay function (5.19). The problems appear to arise from a lack of identifiability. The apparent lack of identifiability relates to the parameters η, ψ, and γ. Attempting to estimate all three of these parameters for such models typically resulted in the DNM routines not converging. In most cases, the routines reached the maximum number of iterations allowed, and increasing this maximum did not solve the problem. The strategy that we adopt is to set η equal to a constant and then estimate the remaining parameters. This strategy aids the estimation of the models. In our applications, we set η equal to 0 and 0.5. The value of 0 is considered in other applications, e.g. those of Chavez-Demoulin *et al.* (2005) and Chavez-Demoulin and McGill (2012), and the value of 0.5 was chosen after trying several different values. We find that

the models with $\eta = 0.5$ perform better than the models with $\eta = 0$; see the results presented in Table 7.3. However, it is likely that other values of η could produce better results. To identify the models that use different values of η, we attach the modified identifiers '-pow-0' and '-pow-0.5' to the names of the models.

In addition to the two decay functions above, we also consider the response function

$$\omega(t, m) = t^{\zeta-1} \exp(\delta m - \gamma t), \tag{5.20}$$

where $\gamma, \zeta > 0$ and $\delta \geq 0$. The conditional intensity with this response function is the same as (3.18) in Section 3.5. We refer to this as the gamma response function and use the identifier '-gamma'. This response function is not necessarily monotonic, as seen in Figure 5.3. By using this response function, the estimated models could display a 'delay' in reaction following an extreme loss, i.e. this response function allows for the possibility that it is not the most recent point events which contribute the most to the conditional intensity. The response functions considered in applications in the literature of marked Hawkes processes to extreme losses are, as far as we know, all monotonic decreasing functions, and we include the gamma response function to see if using a nonmonotonic $\omega(\cdot, \cdot)$ is worthwhile. In our applications, the gamma response function is used only with Model h and not with all of the models outlined in the following subsections.

Figure 5.3 presents examples of the exponential and power decay functions, and the gamma response function.

5.5.2 Models with generalised Pareto distributed marks

Model h

For convenience, the form of Model h is restated here. Model h has a conditional intensity of the form

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi \sum_{j:t_j \in (0,t)} \omega(t - t_j, m_j),$$

and a conditional mark density of the form

$$f(m) = \frac{1}{\beta + \alpha v(t|\widetilde{\mathcal{H}}_t)} \left(1 + \xi \frac{m}{\beta + \alpha v(t|\widetilde{\mathcal{H}}_t)}\right)^{-1/\xi - 1} \quad \text{for } m \geq 0,$$

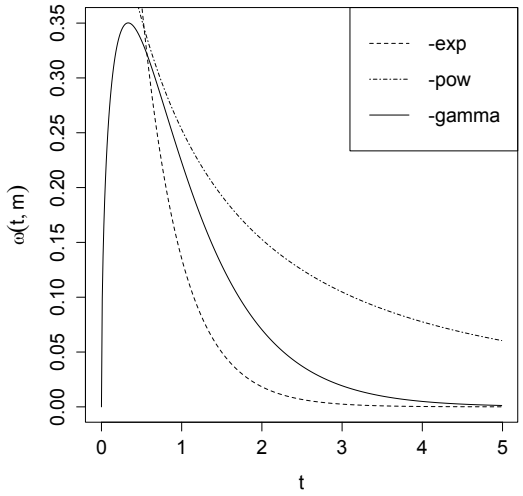


Figure 5.3: *Examples of the exponential and power decay functions, and the gamma response function. In all of the examples* $\delta = 0$*. For the exponential decay function,* $\gamma = 2$*; for the power decay function,* $\gamma = 1.5$ *and* $\eta = 0.5$*; and for the gamma response function,* $\gamma = 1.5$ *and* $\zeta = 1.5$*.*

where

$$v(t|\widetilde{\mathcal{H}}_t) = \sum_{j:t_j \in (0,t)} \omega(t - t_j, m_j).$$

Model g

For Model g, α is set equal to zero in Model h. The resulting model has the same conditional intensity as Model h, but has a mark distribution that is independent of the past of the process. The marks are unpredictable, and have the density

$$f(m) = \frac{1}{\beta}\left(1 + \xi\frac{m}{\beta}\right)^{-1/\xi - 1} \quad \text{for } m \geq 0,$$

where the scale parameter β is now constant. By including a model with $\alpha = 0$, we can assess whether the observed marks are best modelled by a marked Hawkes process with predictable marks or unpredictable marks. That is, we can assess whether the generalisation of the GPD suggested by McNeil *et al.* (2005, pp. 306–309) improves the fit of the models for the loss series that we consider. McNeil *et al.* (2005, p. 309) find that, for the loss

series and model they consider, the inclusion of α significantly improves the fit of the model.

Chavez-Demoulin *et al.* (2005) also consider models with the same form as Model g in their applications.

Model f

For Model f, α is set equal to zero in Model h and δ is set equal to zero in the response functions. The resulting model has a conditional intensity that is independent of the observed mark values. The change in the conditional intensity just after an extreme loss is now equal to ψ. The conditional intensity of Model f is given by

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi \sum_{j:t_j \in (0,t)} \omega(t - t_j).$$

By comparing the results for Model f and Model g, we can assess, for the loss series that we consider, whether the clustering of extreme losses is explained in part by their magnitudes. The conditional mark distribution for this model is the same as that of Model g. In their applications, Herrera and Schipp (2009, p. 223) find models with forms similar to that of Model f can be worthwhile, i.e. they find that models with unpredictable marks and conditional intensities that do not allow for effects from the marks can outperform more complex models.

Model e

For Model e, α and ψ are set equal to zero in Model h. The conditional intensity of this model is constant, i.e. $\lambda(t|\widetilde{\mathcal{H}}_t) = \tau$, and the mark density function of this model is the same as that given for Model g, i.e. the marks are iid GPD random variables. This model is analogous to the marked Poisson process model described in Section 5.2.2. Such models are considered by Chavez-Demoulin *et al.* (2005) and Chavez-Demoulin and McGill (2012) in their applications. We include Model e and Model a (defined below) in our applications so as to investigate, for the loss series that we consider, whether including the self-exciting components in the conditional intensity and the mark distribution are worthwhile.

5.5.3 Models with exponentially distributed marks

All of the following models have conditional mark distributions that are exponential. They are special cases of the models presented in Section 5.5.2, with $\xi = 0$. Most applications in the literature do not explicitly consider models with conditional exponential marks. An exception is the application considered by Embrechts *et al.* (2011). They consider a marked multitype Hawkes process with exponential marks in their application, but do not consider whether such a mark distribution is suitable. The purpose of including these models is to assess whether the simpler conditional exponential distributions for the marks are useful.

Each of the following models corresponds to a model presented in the previous subsection, where the only difference between the models that correspond is their conditional mark distributions. For example, Model d corresponds with Model h. The only difference between these two models is their conditional mark distributions.

Model d

Model d is the most general model that we consider with a conditional mark distribution that is exponential. The conditional intensity of Model d is the same as that of Model h, and the conditional mark density of Model d is given by

$$f(m) = \mu(t|\widetilde{\mathcal{H}}_t) \exp\left(-m \times \mu(t|\widetilde{\mathcal{H}}_t)\right) \quad \text{for } m \geq 0,$$

where $\mu(t|\widetilde{\mathcal{H}}_t) = (\beta + \alpha v(t|\widetilde{\mathcal{H}}_t))^{-1}$, $\beta > 0$, $\alpha \geq 0$, and

$$v(t|\widetilde{\mathcal{H}}_t) = \sum_{j:t_j \in (0,t)} \omega(t - t_j, m_j).$$

This model corresponds to Model h.

Model c

For model c, the parameter α is set equal to zero in Model d. This results in a model with unpredictable marks that are iid exponential random variables with mean β. The conditional intensity of this model is the same as that of Model h. This model corresponds to Model g.

Model b

For Model b, α is set equal to zero in Model d and δ is set equal to zero in the response functions. The resulting model has unpredictable marks that are iid exponential random variables with mean β and a conditional intensity that is the same as that of Model f. This model corresponds to Model f.

Model a

For this model, the parameters α and ψ in Model d are set equal to zero. The conditional intensity of this model is constant, i.e. $\lambda(t|\widetilde{\mathcal{H}}_t) = \tau$, and the marks are iid exponential random variables with mean β. This model corresponds to Model e.

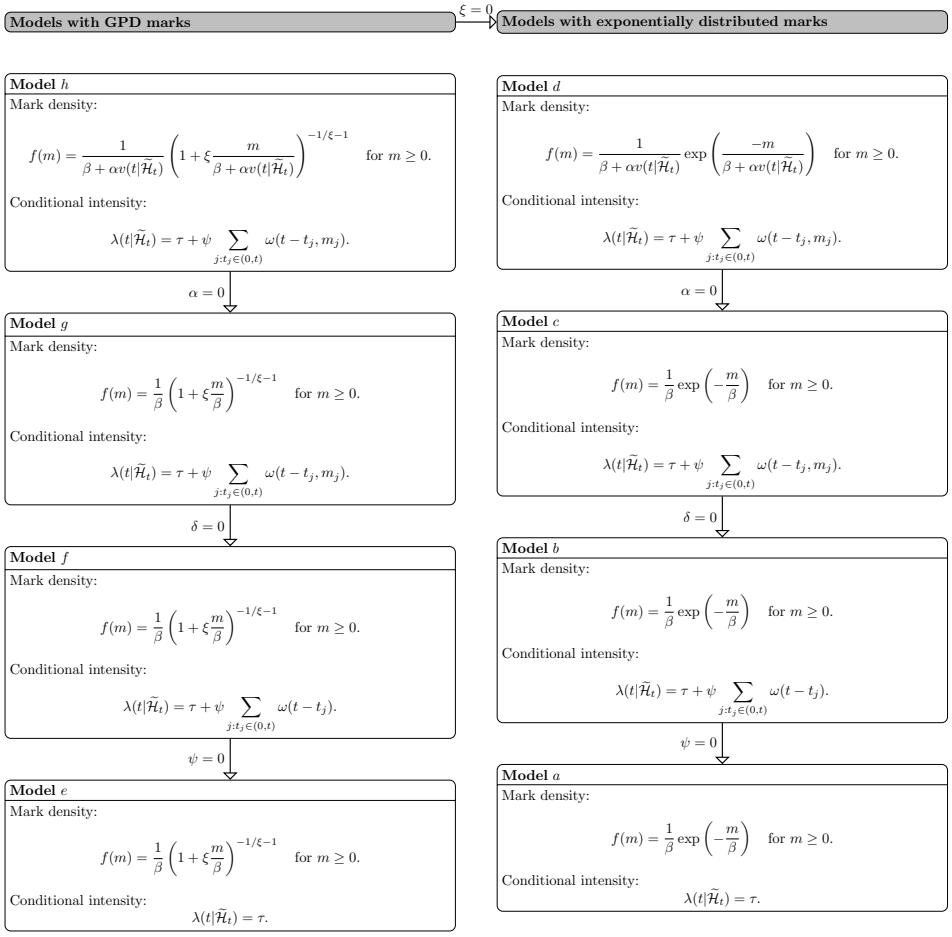


Figure 5.4: *The relationships between the various models for a general response function $\omega(t, m)$. The decay functions used with Models b–d, and f–h, can take one of two forms, either the exponential or power decay forms. The gamma response function is also used with Model h. The function $v(t|\widetilde{\mathcal{H}}_t)$ is given by $v(t|\widetilde{\mathcal{H}}_t) = \sum_{j:t_j \in (0,t)} \omega(t - t_j, m_j)$.*

CHAPTER 6

Some stochastic volatility models and backtesting

Discrete-time stochastic volatility (SV) models are an alternative class of models to the marked Hawkes process models discussed in Chapter 5 — in the sense that they can also be used to forecast conditional VaR and ES. In this chapter we describe two nonstandard SV models introduced by Langrock *et al.* (2012). This description of the SV models includes a brief account of the parameter estimation methods which make use of a structured hidden Markov model (HMM) approach. The HMM structure provides access to conditional forecast distributions, and these are used to forecast conditional VaR and ES.

In our applications, the conditional VaR and ES forecasts found by using the marked Hawkes process models and the SV models will be assessed by using the backtesting methods described in the last section of this chapter.

6.1 Some nonstandard stochastic volatility models

Again we consider the real-valued discrete-time stochastic process $\{X_t, t = 1, 2, \ldots\}$ which represents the daily losses on an asset or portfolio of assets. The standard SV model for losses[1] can be defined as

$$X_t = \epsilon_t \beta \exp(G_t/2), \quad \text{where} \quad G_t = \phi G_{t-1} + \sigma \eta_{t-1}, \tag{6.1}$$

[1] We use losses to be consistent with the previous chapter, but typically the models are defined for returns.

$|\phi| < 1$, $\beta, \sigma > 0$, and $\{\epsilon_t\}$ and $\{\eta_t\}$ are independent sequences of independent standard normal random variables. This model is labelled SV_0.

6.1.1 SVt model with baseline volatility

In the literature, the SV_0 model has been generalised in several ways; see, for example, the generalisations proposed by Nakajima and Omori (2009). Such generalisations of the SV_0 model typically attempt to mimic features that one may expect daily loss series to display. The generalisation in this section involves using a t distribution for ϵ_t and introducing a lower bound on the volatility of the observed process. Langrock *et al.* argue that a lower bound on the volatility of the observed process is plausible as some level of volatility is always present. They also find in their empirical applications that including the lower bound is worthwhile for most of the models and returns series that they consider. The use of a t distribution for ϵ_t is intended to mimic the heavy tails associated with the empirical distribution of daily asset losses (Taylor, 2005, pp. 69–76; Nakajima and Omori, 2009).

The model takes the form

$$X_t = \epsilon_t(\beta \exp(G_t/2) + \xi), \tag{6.2}$$

where ϵ_t now has a t distribution with $\nu > 0$ degrees of freedom and the additional parameter for the lower bound on volatility is $\xi \geq 0$. The Gaussian AR(1) process for $\{G_t\}$ is the same as given above. This model is labelled SVt, but note that it is slightly different from the model labelled SVt by Langrock *et al.*, as they do not include ξ in their SVt model.

6.1.2 $SVMt$ model

The generalisation considered in this section involves changing the latent log-volatility process $\{G_t\}$ in (6.2) to an independent mixture of two Gaussian AR(1) processes. This generalisation is intended to mimic the abrupt changes that the level of volatility may display over time (Langrock *et al.*, 2012). The latent log-volatility process is given by

$$G_t = \begin{cases} \phi_1 G_{t-1} + \sigma_1 \eta_{t-1} & \text{with probability } \alpha, \\ \phi_2 G_{t-1} + \sigma_2 \eta_{t-1} & \text{with probability } 1 - \alpha, \end{cases} \tag{6.3}$$

where $\sigma_1, \sigma_2 > 0$, $0 \leq \alpha \leq 1$, and $\{\eta_t\}$ is a sequence of independent standard normal random variables. As is done by Langrock *et al.*, we label this model $SVMt$.

The following necessary and sufficient condition for the second-order stationarity of $\{G_t\}$ is given by Wong and Li (2000),

$$\left|\alpha\phi_1^2 + (1-\alpha)\phi_2^2\right| < 1.$$

We enforce this condition in our applications of this model. It is noted by Wong and Li (2000) that one of the component AR(1) processes can be explosive, e.g. $\phi_1 = 1.1$, without the second-order stationarity of the mixture process being violated. In the empirical applications carried out by Langrock *et al.*, it can be seen that one of the AR(1) processes is invariably explosive for the $SVMt$ model.

The $SVMt$ model is one of the models that performs well on the basis of AIC in the empirical applications carried out by Langrock *et al.*.

6.1.3 Parameter estimation

The likelihood function for the observed losses $x_1, x_2, \ldots, x_T$ is given by

$$\begin{aligned} L(\boldsymbol{\theta}) &= \int \cdots \int f(\mathbf{x}, \mathbf{g}) \, \mathrm{d}\mathbf{g} \\ &= \int \cdots \int f_{G_1}(g_1) f_{X_1|G_1=g_1}(x_1) \\ &\qquad \times \prod_{t=2}^{T} f_{G_t|G_{t-1}=g_{t-1}}(g_t) f_{X_t|G_t=g_t}(x_t) \, \mathrm{d}g_T \ldots \mathrm{d}g_1, \end{aligned} \tag{6.4}$$

where the dimension of the integral is T, the number of observations. The exact evaluation of the likelihood function (6.4) is difficult, if not impossible. This has led to several innovative methods for fitting SV models; see Broto and Ruiz (2004) for a survey of some of these methods. The method that we use to fit the SV models is that described by Zucchini and MacDonald (2009, pp. 190–192) and used by Langrock *et al.*. This method involves discretising the latent log-volatility process $\{G_t\}$ so as to structure the observed process $\{X_t\}$ as an HMM; see Chapter 2 of Zucchini and MacDonald (2009) for an introduction to HMMs. The likelihood of the SV model can then be approximated by the likelihood of an HMM. The details of the method are described as follows.

Let the range of the latent log-volatility process $\{G_t\}$ be discretised into m equally-sized intervals $B_i = (b_{i-1}, b_i)$ for $i = 1, \ldots, m$, and let $b_i^\star$ be the midpoint in B_i. Define $\{C_t\}$ to be a (discrete-time, homogeneous) Markov chain which is said to be in state i at step t if $G_t \in B_i$. Then the above likelihood function can be approximated by

$$L(\boldsymbol{\theta}) \approx \sum_{i_1=1}^{m} \ldots \sum_{i_T=1}^{m} \Pr\{C_1 = i_1\} f_{X_1|G_1=b_{i_1}^\star}(x_1) \times \prod_{t=2}^{T} \Pr\{C_t = i_t | C_{t-1} = i_{t-1}\} f_{X_t|G_t=b_{i_t}^\star}(x_t). \quad (6.5)$$

The transition probability $\Pr\{C_t = j | C_{t-1} = i\}$ can be approximated by

$$\gamma_{ij} = F_{G_t|G_{t-1}=b_i^\star}(b_j) - F_{G_t|G_{t-1}=b_i^\star}(b_{j-1}),$$

where $F_{G_t|G_{t-1}}(\cdot)$ is the conditional distribution function of G_t given G_{t-1}. As the value of m increases the closer the approximate likelihood (6.5) will be to the likelihood (6.4), but increasing m will increase the computational burden of evaluating the approximate likelihood.

Let $\boldsymbol{\Gamma}$ be the approximate transition probability matrix associated with $\{C_t\}$. The matrix $\boldsymbol{\Gamma}$ is an $m \times m$ matrix with (i, j) element equal to γ_{ij}. Let the row vector $\boldsymbol{\delta}$ be the distribution of C_1, where the ith element of $\boldsymbol{\delta}$ is equal to $\Pr\{C_1 = i\} = \Pr\{G_1 \in B_i\}$. Then the multiple sum in (6.5) can be written as the matrix product

$$L(\boldsymbol{\theta}) \approx \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\ldots\boldsymbol{\Gamma}\mathbf{P}(x_{T-1})\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}', \quad (6.6)$$

where $\mathbf{1}'$ is a column vector of ones and $\mathbf{P}(x_t)$ is a diagonal matrix with ith entry equal to $f_{X_t|G_t=b_i^\star}(x)$. This approximate likelihood function has the same form as the likelihood function of an HMM; see Zucchini and MacDonald (2009, p. 37).

We can then find the (approximate) MLEs by maximising the approximate log-likelihood function. This is done by using R and the `nlm` routine. The constraints on the parameter values are enforced by transforming the parameters, and as there may be multiple local maxima in the likelihood surface, we use several sets of starting values for the maximisation routine.

There are several values that need to be specified before proceeding; the range of the latent process $\{G_t\}$ and the total number of intervals m need

to be chosen. For the range of $\{G_t\}$, we use $(-5, 5)$ for the SVt model and $(-8, 8)$ for the $SVMt$ model, as done by Langrock *et al.*. For m, we use 100, as parameter estimation is fast and as the parameter estimates do not change markedly for larger values of m; see the examples presented by Zucchini and MacDonald (2009, p. 193) and Langrock *et al.*. The vector $\boldsymbol{\delta}$ is set to the uniform distribution.

6.1.4 Forecast distributions and market risk measures

To forecast conditional VaR and ES, we need to find expressions for the conditional forecast distributions associated with the SV models described above. The approximating HMM used to estimate the SV models provides access to the forecast distributions available for HMMs. Langrock *et al.* make use of the HMM forecast distributions when forecasting conditional VaR in their applications. Zucchini and MacDonald (2009, pp. 77–79) derive the forecast distributions for an HMM and we use their results here.

The conditional forecast distribution $F_{X_{t+1}|\widetilde{\mathcal{H}}_t^x}(\cdot)$ for the loss on day $t+1$, given the observed loss history up to and including time t, $\widetilde{\mathcal{H}}_t^x$, can be approximated by

$$F_{X_{t+1}|\widetilde{\mathcal{H}}_t^x}(x) \approx \sum_{i=1}^{m} w_i F_{X_{t+1}|G_{t+1}=b_i^\star}(x), \tag{6.7}$$

where w_i is the ith element of the vector

$$\frac{\boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\ldots\boldsymbol{\Gamma}\mathbf{P}(x_t)\boldsymbol{\Gamma}}{\boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\ldots\boldsymbol{\Gamma}\mathbf{P}(x_t)\mathbf{1}'}.$$

The right-hand side of (6.7) follows from the forecast distribution of an HMM as given by Zucchini and MacDonald (2009, p. 79), except here we are considering a continuous observation process and a forecast horizon of one period.

Conditional VaR

The approximate forecast distribution (6.7) can then be used to forecast the approximate conditional $100\phi\%$ VaR. This is done by solving for $\mathrm{VaR}_\phi(X_{t+1})$ in

$$\Pr\{X_{t+1} \geq \mathrm{VaR}_\phi(X_{t+1})|\widetilde{\mathcal{H}}_t^x\} = 1 - \phi,$$

where the left-hand side can be approximated by

$$\Pr\{X_{t+1} \geq \mathrm{VaR}_\phi(X_{t+1}) | \widetilde{\mathcal{H}}_t^x\} \approx 1 - \sum_{i=1}^{m} w_i F_{X_{t+1}|G_{t+1}=b_i^\star}\left(\mathrm{VaR}_\phi(X_{t+1})\right).$$

As a result, to find an estimate for $\mathrm{VaR}_\phi(X_{t+1})$, we need to solve for $\mathrm{VaR}_\phi(X_{t+1})$ in

$$\phi = \sum_{i=1}^{m} w_i F_{X_{t+1}|G_{t+1}=b_i^\star}(\mathrm{VaR}_\phi(X_{t+1})).$$

This is done by using the `multiroot` root-finding algorithm in R.

Langrock *et al.* forecast conditional 99% VaR for several returns series in their empirical applications. They perform a backtest to assess these forecasts, and they find that the $SVMt$ model performs fairly well and that it outperforms the SVt model.

Conditional ES

To find the conditional $100\phi\%$ ES, we need to evaluate

$$\begin{aligned} \mathrm{ES}_\phi(X_{t+1}) &= \mathrm{E}\left(X_{t+1} \,\middle|\, X_{t+1} \geq \mathrm{VaR}_\phi(X_{t+1}), \widetilde{\mathcal{H}}_t^x\right) \\ &= \frac{\int_{\mathrm{VaR}_\phi(X_{t+1})}^{\infty} x f_{X_{t+1}|\widetilde{\mathcal{H}}_t^x}(x)\,\mathrm{d}x}{1-\phi}, \end{aligned} \tag{6.8}$$

where $f_{X_{t+1}|\widetilde{\mathcal{H}}_t^x}(\cdot)$ is the conditional forecast density. The conditional forecast density in (6.8) can be approximated by

$$f_{X_{t+1}|\widetilde{\mathcal{H}}_t^x}(x) \approx \sum_{i=1}^{m} w_i f_{X_{t+1}|G_{t+1}=b_i^\star}(x),$$

and the expression for the approximate conditional $100\phi\%$ ES is given by

$$\mathrm{ES}_\phi(X_{t+1}) \approx \frac{\sum_{i=1}^{m} w_i \int_{\mathrm{VaR}_\phi(X_{t+1})}^{\infty} x f_{X_{t+1}|G_{t+1}=b_i^\star}(x)\,\mathrm{d}x}{1-\phi}.$$

The integrals in the above expression are evaluated numerically to find an approximate value for $\mathrm{ES}_\phi(X_{t+1})$. The integration is performed by using the `integrate` routine in R.

It is worth noting here that the SV models can be used to forecast conditional VaR and ES for all confidence levels $\phi \in (0, 1)$. In contrast,

the marked Hawkes process models can only be used to forecast conditional VaR and ES for confidence levels greater than the larger of the lower bound given in (5.16) and the bound required in (5.11). This is a disadvantage of using the marked Hawkes process models to forecast conditional VaR and ES, as they are not as flexible as the SV models in this respect.

6.2 Backtesting

Backtesting refers here to the assessment of the forecasts of the conditional market risk measures. The backtesting methods involve two non-overlapping periods. The first period is referred to as the in-sample period, and the losses which occur during this period, $x_1, x_2, \ldots, x_T$, are used to fit the models. The fitted models are then used to forecast conditional VaR and ES for the second period. The second period is called the out-of-sample period. The forecasts of conditional VaR and ES are then compared to the losses, $x_{T+1}, x_{T+2}, \ldots, x_{T+w}$, which occurred in the out-of-sample period to evaluate the forecasts.

6.2.1 Backtesting conditional VaR

To backtest the conditional VaR forecasts from a particular model, we consider the number of VaR exceptions in the out-of-sample period. A VaR exception occurs when the observed loss for a particular day exceeds the conditional VaR forecast for that day. If the number of exceptions is large, the model underestimates the true conditional VaR, and vice versa. Given the observed number of exceptions, we can perform a hypothesis test. The null hypothesis is that the model correctly forecasts the conditional VaR, and the alternate hypothesis is that the model underestimates the conditional VaR. The hypothesis test here is similar to that outlined by McNeil and Frey (2000), and the details are as follows.

Let $\mathrm{VaR}_\phi(X_t)$ be the true value of the conditional $100\phi\%$ VaR and $\widehat{\mathrm{VaR}}_\phi(X_t)$ be the forecast found by using the fitted model. Then define the indicator

$$I_t = \begin{cases} 1 & \text{if } X_t > \mathrm{VaR}_\phi(X_t) \\ 0 & \text{if } X_t \leq \mathrm{VaR}_\phi(X_t) \end{cases}$$

for $t = T+1, T+2 \ldots, T+w$. The I_ts are iid Bernoulli random variables with $p = 1-\phi$, as

$$\Pr\{I_t = 1\} = 1 - \phi.$$

The total number of exceptions over the out-of-sample period is given by

$$\sum_{t=T+1}^{T+w} I_t,$$

and this has a Binomial distribution with parameters $n = w$ and $p = 1-\phi$.

We can then construct a test by using the observed indicator values for the out-of-sample period,

$$\widehat{I}_t = \begin{cases} 1 & \text{if } x_t > \widehat{\text{VaR}}_\phi(X_t) \\ 0 & \text{if } x_t \leq \widehat{\text{VaR}}_\phi(X_t) \end{cases},$$

and the sum

$$Q = \sum_{t=T+1}^{T+w} \widehat{I}_t.$$

Under the null hypothesis, Q should be a realisation from a Binomial random variable with parameters $n = w$ and $p = 1-\phi$. The alternate hypothesis is that $p > 1-\phi$, as the concern is that the model underestimates the conditional VaR. Then for an observed Q, we can calculate the associated one-sided p-value.

In addition to performing the above hypothesis test, the model can be classified according to the three-zone classification system of the BCBS (2006, pp. 313–321). A model is said to be in the 'green zone' if the observed number of exceptions Q is less than the 95th percentile of its distribution under the null hypothesis. A model in the green zone is regarded as suitable. A model is said to be in the 'red zone' if the observed number of exceptions Q is greater than or equal to the 99.99th percentile of its distribution under the null hypothesis. A model in the red zone is regarded as inaccurate. A model is said to be in the 'yellow zone' if the observed number of exceptions Q is between the above percentiles. If a bank's model is in the yellow zone, the bank is very likely to be required to hold additional capital, and if the model is in the red zone, the bank is required to hold additional capital (BCBS, 2006, pp. 313–321, 2011, pp. 15–16). The classification of the

models is mainly illustrative, as in most circumstances the hypothesis test conveys much of the same information.

Tests which consider only the number of exceptions, as is the case with the test above, have drawn criticism; see, for example, Christoffersen (1998). One criticism made of such tests is that they do not test whether the exceptions are independent. As a result, a model which has clusters of VaR exceptions may still be regarded as accurate, provided that the total number of exceptions is low. We consider the above test as banks regulated under the Basel Framework are required to perform a similar test, and the results of the test are used to determine the cost to a bank in terms of additional capital requirements. See Annex 10a of BCBS (2006, pp. 310–321).

Alternative tests which may be used to assess conditional VaR forecasts include that suggested by Christoffersen (1998), and the test for quantile forecasts outlined by Gneiting (2011) may be used to compare forecasts from different models.

6.2.2 Backtesting conditional ES

The appropriate method for backtesting conditional ES forecasts is not clear. The recent review by Embrechts and Hofert (2014) highlights the fact that ES cannot be 'properly' backtested in light of the work published by Gneiting (2011). However, there are methods in the literature that have been used to measure the accuracy of ES forecasts. These methods may not produce the correct decisions about the model that best forecasts conditional ES (see the work of Gneiting (2011) for an example illustrating the outcomes of using incorrect measures to assess forecasts), but in the absence of any clear-cut 'correct' technique we use here the statistics discussed by Embrechts *et al.* (2005) as possible measures of the accuracy of ES. As such, the ES backtest results in Chapter 7 are not the 'last word' and have to be treated with some caution.

The test used by Embrechts *et al.* (2005) to assess the ES forecasts from several different models involves the two statistics V_1^{ES} and V_2^{ES}, and is described as follows.

The first statistic is the average of the differences between the actual losses and the conditional ES forecasts for the losses that exceed the condi-

tional VaR forecasts. It is calculated as follows:

$$V_1^{\mathrm{ES}} = \frac{\sum_{t=T+1}^{T+w}(x_t - \widehat{\mathrm{ES}}_\phi(X_t))\widehat{I}_t}{\sum_{t=T+1}^{T+w}\widehat{I}_t}.$$

A model which provides good forecasts of conditional ES should have $|V_1^{\mathrm{ES}}|$ close to zero, as $\mathrm{E}[(X_t - \mathrm{ES}_\phi(X_t))I_t|\widetilde{\mathcal{H}}_{t-1}^x] = 0$.

The weakness of the first statistic is that it depends heavily on the $\mathrm{VaR}_\phi(X_t)$ forecasts (Embrechts *et al.*, 2005). The second statistic depends less heavily on the $\mathrm{VaR}_\phi(X_t)$ forecasts, and is calculated as follows:

$$V_2^{\mathrm{ES}} = \frac{\sum_{t=T+1}^{T+w}\widehat{D}_t\widehat{I}_t^D}{\sum_{t=T+1}^{T+w}\widehat{I}_t^D},$$

where $\widehat{D}_t = x_t - \widehat{\mathrm{ES}}_\phi(X_t)$ and

$$\widehat{I}_t^D = \begin{cases} 1 & \text{if } \widehat{D}_t > \widehat{D}^\phi \\ 0 & \text{if } \widehat{D}_t \leq \widehat{D}^\phi \end{cases},$$

where $\widehat{D}^\phi$ is the empirical 100ϕth percentile of the $\widehat{D}_t$s. A model which provides good forecasts of conditional ES should have a low value for $|V_2^{\mathrm{ES}}|$ (Embrechts *et al.*, 2005).

The final statistic is a combination of V_1^{ES} and V_2^{ES}, specifically

$$V^{\mathrm{ES}} = \frac{|V_1^{\mathrm{ES}}| + |V_2^{\mathrm{ES}}|}{2}.$$

This statistic gives an indication of the model's ability to forecast conditional ES — a low value indicates that forecasts of conditional ES from a particular model are good.

CHAPTER 7

Applications

In this chapter we present applications of the marked Hawkes process models and the SV models. The models are applied to South African asset loss data. The objectives of our applications are to investigate the performance of the marked Hawkes process models for South African asset loss data and to determine whether the marked Hawkes process models are competitive relative to the SV models. In addition to these objectives, we also attempt to identify the marked Hawkes process model which performs best.

7.1 Loss data and preliminary analysis

The data used in the applications were downloaded via Bloomberg L.P. from a terminal in the UCT Chancellor Oppenheimer Library on 13 June 2012. The data are daily closing values for the ALSI, the South African Rand to United States dollar exchange rate (ZAR/USD)[1], and the MTN Group Limited share price (MTN). The daily losses are calculated as $x_t = 100 \log(s_{t-1}/s_t)$, where s_t is the index value, exchange rate, or share price at the end of day t. The positive ZAR/USD exchange rate losses are devaluations of the United States dollar relative to the South African Rand. These losses, and the associated risk measures, would therefore be of interest to a party who held assets denominated in United States dollars and who had liabilities denominated in South African Rand, for example.

The daily losses used in the applications are for the period from 8 De-

[1]Exchange rate values recorded on South African public holidays are ignored so that the dates and numbers of observations are similar for all three loss series.

cember 1999 (4 January 2000 for the ZAR/USD loss series) to 12 June 2012, which is about twelve and a half years of data. The daily losses are split into losses from two non-overlapping periods. The first period, from 8 December 1999 (4 January 2000 for the ZAR/USD loss series) to 7 December 2007, is the in-sample period. Plots of the in-sample data are presented in Figure 7.1. The second period, from 10 December 2007 to 12 June 2012, is the out-of-sample period. The losses from the in-sample period are used to fit the models, and the losses from the out-of-sample period are used to backtest the forecast risk measures. The out-of-sample period includes the 2008 financial crisis.

Summary statistics for the in-sample daily losses are reported in Table 7.1. The kurtoses for the in-sample daily losses are all greater than three which strongly suggests that the loss distributions are non-normal. Histograms of the loss data are presented in the lower panels of Figure 7.1. Each of the histograms displays greater clustering around the mean than does the superimposed normal distribution. Figure 7.1 also contains plots of the levels of the ALSI, ZAR/USD exchange rate, and the MTN share price for the in-sample period, as well as plots of each of the daily loss series.

Table 7.1: *Some summary statistics for the in-sample loss series.*

Series	ALSI	ZAR/USD	MTN
no. obs.	1999	1984	1999
max.	7.948	4.213	12.770
min.	−5.889	−6.167	−21.711
mean	−0.068	−0.005	−0.106
std. dev.	1.178	1.070	2.717
kurtosis	6.036	5.657	6.744

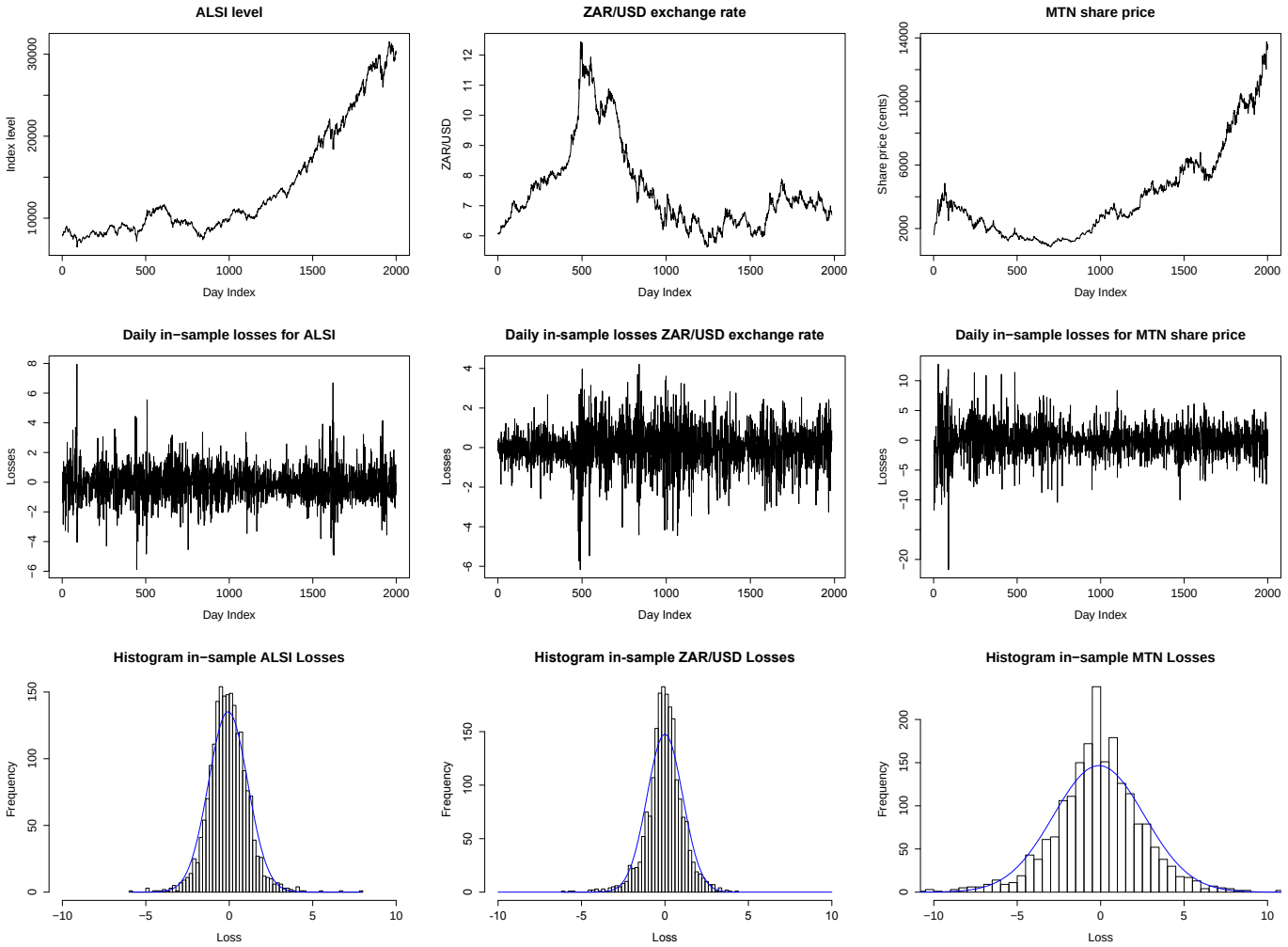


Figure 7.1: *Top row: ALSI value, ZAR/USD exchange rate, and MTN share price for the in-sample period. Middle row: daily losses for each loss series for the in-sample period. Bottom row: histograms of the daily losses for the in-sample period of each loss series with superimposed normal distributions (blue curves). The horizontal axis of the histogram for the MTN losses has been truncated.*

The in-sample marked point process realisations for each loss series are extracted by using a threshold equal to the 90th percentile of the losses for the in-sample period. The marked Hawkes process models are fitted to these marked point process realisations. Figure 7.2 presents the empirical mean excess plots for the positive in-sample losses for each loss series. The thresholds used to extract the marked point processes are indicated by the vertical dashed lines. The empirical mean excess functions appear to be approximately linear above the chosen thresholds for all three loss series if we discount the increased variability at high thresholds. The same vertical axes are used for the three empirical mean excess plots so as to highlight the near-horizontal empirical mean excess function for the ZAR/USD loss series. Table 7.2 presents some information about the marked point process realisations extracted from the in-sample losses.

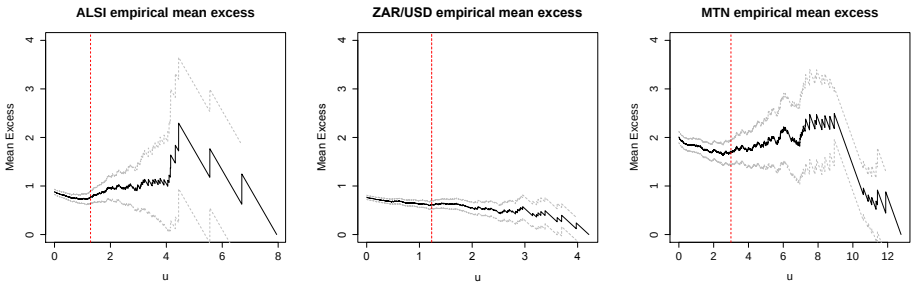


Figure 7.2: *Plots of the empirical mean excess functions for the positive in-sample losses for each loss series considered. The vertical dashed lines indicate the thresholds used to extract the marked point process realisations. The grey dashed curves are approximate Wald-type 95% confidence intervals.*

Table 7.2: *Some summary statistics for the in-sample marked point process realisations.*

Series	ALSI	ZAR/USD	MTN
threshold	1.289	1.232	2.993
no. of point events	200	199	200
mean mag. of the excesses	0.772	0.618	1.716

7.2 Model fitting results

The marked Hawkes process models and the SV models are fitted to the in-sample loss data by using the methods described in Sections 3.2 and 6.1.3, respectively. The AIC values for all of the models are reported in Table 7.3. The AIC values are for the in-sample data, and the best-performing marked Hawkes process model for each loss series has its AIC value in bold. The BIC values are reported in Table B.1 in Appendix B.1.

We highlight several features apparent in Table 7.3 for the marked Hawkes process models.

- Models a and e, both of which have constant intensity functions, do not perform well when compared to the models with self-exciting intensities. This is not surprising, as in Figure 5.2 we saw that the timing of extreme losses on the ALSI is not well modelled by a homogeneous Poisson process. The superior performance of Models b-... and f-..., which are the simplest models with self-exciting intensities, demonstrates that the models with constant intensities are not suitable, and can be rejected.

- The models with the power decay function (5.19) and $\eta = 0.5$ (i.e. the models labelled ...-pow-0.5) perform better across all of the loss series when compared to the corresponding models with the power decay function and $\eta = 0$ (i.e. the models labelled ...-pow-0). In particular, Models c-pow-0.5 and d-pow-0.5 rank in the top three models for the ZAR/USD loss series, and Model d-pow-0.5 ranks in the top two models for the MTN loss series. This demonstrates that including η as a strictly positive constant can be worthwhile.

- The inclusion of the mark impact $e^{\delta m}$ in the conditional intensity, which is a common practice in the literature, appears to be useful in most, but not all, cases. This can be seen by comparing the AIC values for the models labelled b-... with those of the corresponding models labelled c-..., or by comparing the AIC values for the models labelled f-... with those of the corresponding models labelled g-.... This result gives some credibility to the intuitive arguments for including effects from the marks in the intensities — the magnitudes of

Table 7.3: *AIC values for all of the models. The models with identifiers '-exp' and '-pow' have exponential and power type decay functions respectively. The model with the identifier '-gamma' has the gamma response function. The best-performing marked Hawkes process model for each loss series has its AIC value in bold. A summary of the marked Hawkes process models is given at the foot of the table; see Section 5.3 and Figure 5.4 for a more complete overview of the marked Hawkes process models. The AIC values for the two SV models are not comparable to those of the marked Hawkes process models.*

Model (no. parameters)	ALSI	ZAR/USD	MTN
a (2)	1621.6	1523.6	1940.9
b-exp (4)	1598.5	1470.1	1912.9
b-pow-0 (4)	1611.1	1478.5	1927.2
b-pow-0.5 (4)	1603.2	1472.5	1917.3
c-exp (5)	1600.4	1470.6	1910.6
c-pow-0 (5)	1612.6	1471.1	1917.0
c-pow-0.5 (5)	1605.0	1469.4	1912.4
d-exp (6)	1578.6	1469.0	**1894.2**
d-pow-0 (6)	1596.2	1471.6	1902.5
d-pow-0.5 (6)	1584.4	**1468.6**	1895.7
e (3)	1617.8	1525.6	1942.1
f-exp (5)	1594.8	1472.1	1914.1
f-pow-0 (5)	1607.3	1480.5	1928.3
f-pow-0.5 (5)	1599.5	1474.5	1918.4
g-exp (6)	1596.7	1472.6	1911.7
g-pow-0 (6)	1608.9	1473.1	1918.1
g-pow-0.5 (6)	1601.3	1471.4	1913.5
h-exp (7)	1580.1	1471.0	1896.0
h-pow-0 (7)	1596.9	1473.6	1904.3
h-pow-0.5 (7)	1585.7	1470.6	1897.6
h-gamma (8)	**1575.5**	1470.3	1897.7
SVt (5)	6037.0	5505.7	9248.8
$SVMt$ (8)	6039.2	5508.8	9253.7

Descriptions of the marked Hawkes process models :

Model 1:	**mark df for 1**	**Model 2:**	**mark df for 2**	**intensity for 1 and 2**
$a-\ldots:$	iid exp.	$e-\ldots:$	iid GPD	constant
$b-\ldots:$	iid exp.	$f-\ldots:$	iid GPD	self-exciting, no mark impacts
$c-\ldots:$	iid exp.	$g-\ldots:$	iid GPD	self-exciting, mark impacts
$d-\ldots:$	predictable exp.	$h-\ldots:$	predictable GPD	self-exciting, mark impacts

Abbreviations: df: distribution function, exp.: exponential.

the extreme losses seem able to explain some of the temporal clustering of extreme losses for some of the loss series. The exceptions are the ALSI loss series and the models with exponential decay functions for the ZAR/USD loss series. In these cases including effects from the observed marks in the conditional intensities does not improve the AIC values.

- The predictable mark distributions, as proposed by McNeil *et al.* (2005, pp. 308–309), are worthwhile in almost all cases. This can be seen by comparing the AIC values for the models labelled *c*-... with those of the corresponding models labelled *d*-..., and similarly by comparing the models labelled *g*-... with the corresponding models labelled *h*-.... In particular, for each loss series the best-performing model has predictable marks; Model *h*-gamma is the best-performing model for the ALSI loss series, Model *d*-pow-0.5 is the best-performing model for the ZAR/USD loss series, and Model *d*-exp is the best-performing model for the MTN loss series. However, the generalisation involved when moving from a model with unpredictable marks to a model with predictable marks, which is achieved by including the additional parameter α, does not always lead to an improvement in the fit of the models. For example, for the ZAR/USD loss series, Model *c*-pow-0 outperforms Model *d*-pow-0.

- Including the models with conditional exponential marks, which are special cases of the models with conditional GPD marks, proved useful. The more general models with conditional GPD marks, Models *e*–*h*-..., did not always perform better. For example, Model *d*-exp, which is a special case of Model *h*-exp, outperformed Model *h*-exp for all of the loss series, and performed well across all three loss series when compared to the other models.

- For the ZAR/USD loss series, the differences between the AIC values of Models *a*–*d*-... and the AIC values of the corresponding Models *e*–*h*-... are all equal to minus two, e.g. the difference between the AIC values of Model *c*-exp and Model *g*-exp is $1470.6 - 1472.6 = -2$. This is because the conditional GPDs for the marks in Models *e*–*h*-... degenerated to conditional exponential distributions. Such a de-

generation is suggested by the near-horizontal empirical mean excess function for the ZAR/USD loss series in Figure 7.2. The estimates of ξ, which are reported in Table B.2 in Appendix B.2, are (very) close to zero. In effect, Models *a*–*d*-. . . and Model *h*-gamma are the models that we are considering for the ZAR/USD loss series.

- The 'experimental' model, Model *h*-gamma, which can have a non-monotonic response function, showed varied performance across the three loss series. It was the best-performing model for the ALSI loss series, but did not perform as well for the other two loss series. This suggests that the generalisation may be useful in some cases.

The models that we carry forward are: Models *d*-exp, *d*-pow-0.5, *h*-exp and *h*-gamma. Model *h*-exp is a model that has been considered in several applications in the literature, and it is included here because its AIC value is in the top three models for both the ALSI and MTN loss series. We perform goodness-of-fit tests for these models and use them to forecast conditional VaR and ES for the out-of-sample period of each loss series.

As regards the AIC values for the two SV models, the SVt model performs better for all of the loss series. As our SVt model includes the baseline volatility parameter ξ and the SVt model of Langrock *et al.* (2012) does not, our results are not comparable to those of Langrock *et al.* (2012). However, it is interesting to note that Langrock *et al.* (2012) find that the $SVMt$ model mostly outperforms their SVt model.

The MLEs for the marked Hawkes process models fitted to the marked point process realisation extracted from the in-sample ALSI losses are presented in Table 7.4. The MLEs for the other two loss series are presented in Tables B.2–B.5 in Appendix B.2.

There are several interesting features to note about the MLEs presented in Table 7.4.

- The estimates of τ found for Models *a* and *e* are larger than the estimates found for the models with self-exciting intensities. This suggests that a high proportion of the point events during the in-sample period are offspring point events as opposed to immigrant point events.

  The estimates of τ also vary across the different marked Hawkes process models. This indicates that the proportion of observed point

Table 7.4: *MLEs for the marked Hawkes process models fitted to the in-sample marked point process realisation extracted from the ALSI losses.*

Model	$\widehat{\tau}$	$\widehat{\gamma}$	$\widehat{\psi}$	$\widehat{\delta}$	$\widehat{\beta}$	$\widehat{\xi}$	$\widehat{\alpha}$	$\widehat{\zeta}$
a	0.1000	–	–	–	0.7721	–	–	–
b-exp	0.0518	0.0730	0.0354	–	0.7721	–	–	–
b-pow-0	0.0385	2.7025	0.1089	–	0.7721	–	–	–
b-pow-0.5	0.0407	7.8381	0.9413	–	0.7721	–	–	–
c-exp	0.0526	0.0750	0.0337	0.0728	0.7721	–	–	–
c-pow-0	0.0391	2.6423	0.0946	0.1510	0.7721	–	–	–
c-pow-0.5	0.0416	7.6626	0.8487	0.0927	0.7721	–	–	–
d-exp	0.0587	0.1071	0.0388	0.1573	0.4233	–	0.2414	–
d-pow-0	0.0440	1.7584	0.0771	0.2054	0.3027	–	0.5608	–
d-pow-0.5	0.0475	5.5192	0.6008	0.1546	0.3390	–	4.0158	–
e	0.1000	–	–	–	0.6400	0.1732	–	–
f-exp	0.0518	0.0730	0.0354	–	0.6400	0.1732	–	–
f-pow-0	0.0385	2.7025	0.1089	–	0.6400	0.1732	–	–
f-pow-0.5	0.0407	7.8381	0.9413	–	0.6400	0.1732	–	–
g-exp	0.0526	0.0750	0.0337	0.0728	0.6400	0.1732	–	–
g-pow-0	0.0391	2.6423	0.0946	0.1510	0.6400	0.1732	–	–
g-pow-0.5	0.0416	7.6626	0.8487	0.0927	0.6400	0.1732	–	–
h-exp	0.0581	0.1033	0.0381	0.1543	0.3961	0.0539	0.2251	–
h-pow-0	0.0431	1.8769	0.0797	0.1996	0.2625	0.0905	0.5379	–
h-pow-0.5	0.0469	5.7174	0.6229	0.1503	0.3083	0.0653	3.9468	–
h-gamma	0.0630	0.3547	0.0409	0.1547	0.4418	0.0476	0.2170	1.9990

events that would likely be classified as immigrant point events varies across the models. The estimated probability distribution of the branching structure for a particular model would give us the estimated probabilities that the model assigns to a particular point event being an immigrant or offspring point event.

For example, Table 7.5 presents the estimated probability distribution of the branching structure found by using the estimated Model d-exp. The probability distribution is found by using (3.8) and is for the first 13 point events extracted from the in-sample ALSI losses. The (i, j) element of the table, for $i \neq j$, is the estimated probability that the ith point event is an immediate offspring of the jth point event, and the (i, j) element, for $i = j$, is the estimated probability that the ith point event is an immigrant point event. For example, from the table it is clear that under the estimated Model d-exp, the eleventh point event is more likely to be an offspring point event, as the diagonal entry (11,11) is less than 500(/1000); that is, at the risk of over-interpreting such results, the eleventh extreme loss in the in-sample period is more likely to be caused by market excitement than by external factors.

Table 7.5: *Estimated probability distribution of the branching structure for the first 13 point events in the ALSI in-sample data. The distribution was found by using the estimated Model d-exp. The probabilities have been multiplied by 1000.*

1000	0	0	0	0	0	0	0	0	0	0	0	0
60	940	0	0	0	0	0	0	0	0	0	0	0
32	412	556	0	0	0	0	0	0	0	0	0	0
21	274	247	458	0	0	0	0	0	0	0	0	0
10	134	121	151	584	0	0	0	0	0	0	0	0
7	91	82	103	273	443	0	0	0	0	0	0	0
5	68	61	77	204	219	367	0	0	0	0	0	0
3	45	41	51	135	145	162	417	0	0	0	0	0
3	33	30	37	99	106	118	235	339	0	0	0	0
2	24	21	27	71	76	85	169	189	336	0	0	0
1	19	17	21	56	60	67	134	149	180	295	0	0
1	11	10	12	32	34	38	76	85	102	114	487	0
0	4	4	5	13	14	16	32	35	42	48	191	594

The availability of the probability distribution of the branching structure is not really an advantage of using the EM algorithm to find the

MLEs — contrary to the claim of Halpin and De Boeck (2013, p. 801). The distribution is readily computed once one has found the MLEs via DNM, and this is what was done here.

- The estimates of ψ are all different from zero which suggests that the models with the self-exciting intensities are worthwhile. The parameter ψ controls the size of the increase in the intensity following an extreme loss.

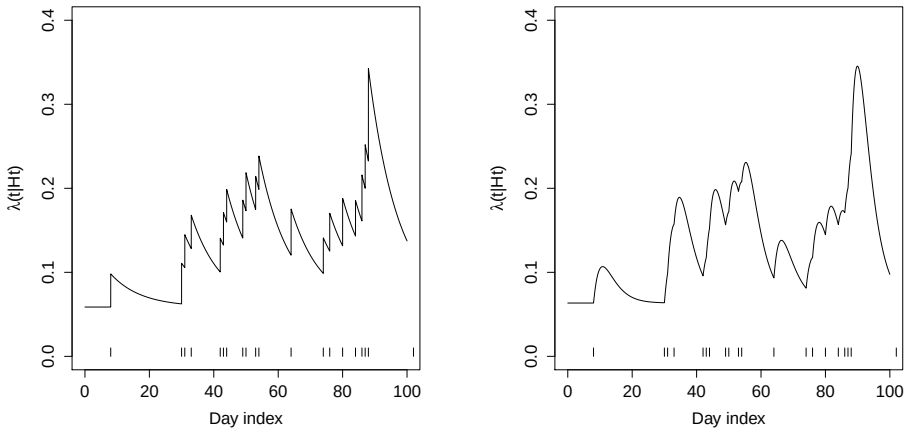
- The estimates of α, which is the parameter that controls the influence of the observed marked point process on the conditional mark distributions, are all different from zero. This suggests that the predictable mark distributions are useful for this loss series.

- An interesting feature to note is that the estimate of ζ for Model h-gamma is greater than one. As a result, the estimated response function is humped, i.e. $\widehat{\omega}(t, m)$ increases and then decreases for increasing t and a given m. It is not the latest point events which contribute the most to the conditional intensity, i.e. there is a delay in the increase in the conditional intensity following a point event. This delay could be interpreted as a delay in the transmission or synthesis of the information carried by an extreme loss.

  Figure 7.3 presents plots of the estimated conditional intensities for Models d-exp and h-gamma for some of the ALSI in-sample losses. The effect of the humped response function of Model h-gamma can be seen in the rounded peaks of the estimated conditional intensity. This is different from the peaks seen for the estimated conditional intensity of Model d-exp, which are sharp, but the shapes of the two intensities are otherwise similar.

  The estimates of ζ for the other two loss series are also greater than one.

- Standard errors and confidence intervals for the estimates can be found by using the methods outlined in Section 3.4.4. For example, the approximate likelihood-based 95% confidence intervals for the MLEs



(a) *Model d-exp* **(b)** *Model h-gamma*

Figure 7.3: *Extracts of the estimated conditional intensities for (a) Models d-exp and (b) h-gamma for the in-sample ALSI losses. The extracts are for the first 100 days of the in-sample period and the times of the extreme losses are indicated by the rug on each plot.*

of Model h-gamma for the ALSI loss series are:

$$\widehat{\tau} : (0.044, 0.083), \quad \widehat{\gamma} : (0.124, 1.076), \quad \widehat{\psi} : (0.016, 0.105), \quad \widehat{\delta} : (0, 0.386),$$
$$\widehat{\beta} : (0.307, 0.593), \quad \widehat{\xi} : (0, 0.231), \quad \widehat{\alpha} : (0.074, 0.578), \quad \widehat{\zeta} : (1.177, 3.718).$$

The lower bounds of the intervals for $\widehat{\delta}$ and $\widehat{\xi}$ are truncated at zero because the solutions of the relevant equations are negative (the solutions are -0.178 for δ and -0.0840 for ξ) and we are concerned with nonnegative values for these parameters. The lower bounds of zero for the confidence intervals for $\widehat{\delta}$ and $\widehat{\xi}$ suggest that a simpler model with both of these parameters set equal to zero may be useful. Such a model would not allow for effects from the observed marks on the conditional intensity and would have a conditional mark distribution that is exponential.

It should be noted that several of the likelihood-based confidence intervals are asymmetric, and so considering only standard errors and Wald-type confidence intervals may be misleading. For example, the Wald-type 95% confidence interval for $\widehat{\zeta}$ is $(0.882, 3.116)$. From this Wald-type interval, we would conclude that $\widehat{\zeta}$ is not significantly dif-

ferent from one at the 5% level, but when the likelihood-based 95% confidence interval is considered, we see that the interval does not include one. As the two intervals differ, we use the likelihood-based interval and conclude that $\widehat{\zeta}$ is different from one at the 5% level. Provisionally, the inclusion of ζ in the conditional intensity of Model h-gamma appears to be worthwhile.

The MLEs of the SV models fitted to the in-sample ALSI loss data are presented in Table 7.6. We can see that the t distributions degenerated to normal distributions for the ALSI loss series, as the estimates of ν are large. This does not occur for the other two loss series that we consider. The estimates of ϕ for the SVt model are just less than one for all of the loss series. For each of the loss series, the fitted $SVMt$ models has a latent log-volatility process with one explosive AR(1) component, but in each case the mixture of the two AR(1) processes is stationary. The estimates of ξ are all greater than zero which suggests that the baseline volatility parameter is worthwhile. These results are mostly consistent with the estimates found by Langrock *et al.* (2012) in their empirical applications. The main difference is that for the returns series considered by Langrock *et al.* (2012), none of the estimated t distributions degenerated to normal distributions.

Table 7.6: *MLEs for the SV models fitted to the in-sample ALSI losses.*

Model	$\widehat{\nu}$	$\widehat{\beta}$	$\widehat{\phi}$ ($\widehat{\phi_1}$)	$\widehat{\sigma}$ ($\widehat{\sigma_1}$)	$\widehat{\xi}$	$\widehat{\alpha}$	$\widehat{\phi_2}$	$\widehat{\sigma_2}$
SVt	> 100	0.4024	0.9676	0.3789	0.5651	–	–	–
$SVMt$	> 100	0.4650	0.6556	0.8681	0.5620	0.1438	1.0355	0.0115

7.3 Results of goodness-of-fit tests

We now investigate the fit of the chosen marked Hawkes process models for both the in-sample and out-of-sample periods. The goodness-of-fit tests for the in-sample period are used to check whether the models fail to capture any features of the data to which they are fitted. The goodness-of-fit tests for the out-of-sample period give us an indication of whether the models are suitable for data to which they are not fitted, but are applied.

Figures 7.4–7.6 on pp. 103–105 present the graphical goodness-of-fit tests for the models. Subfigures (a) and (b) of each figure present the goodness-of-fit tests for the in-sample and out-of-sample periods, respectively. The tests presented for the temporal component of each model are: plots of the scaled residual process with 95% and 99% confidence lines (first row), and plots of the points (U_i, U_{i+1}) (second row). For the conditional mark distributions, we present QQ-plots for the $\widehat{F}(m_i)$ quantiles, which are plotted against uniform $(0, 1)$ quantiles, and we present the p-value of a two-sided Kolmogorov–Smirnov test (KS p-value) on each QQ-plot (third and fourth rows). If a plot suggests that a model fits the data poorly, it is marked by an asterisk to the upper right of the plot.

As the estimated temporal components of Models d-exp and h-exp are very similar for all three loss series, we present only the goodness-of-fit tests for the temporal component of Model d-exp.

7.3.1 Results of goodness-of-fit tests for the in-sample period

The goodness-of-fit tests for the in-sample period of each loss series provide no significant indication that the models fit the data poorly. For some of the loss series and models, the scaled residual processes do deviate noticeably from the identity line, but these deviations are not significant. The fit of the conditional mark distributions appears suitable in all cases — the points in the QQ-plots lie close to the diagonal line and the Kolmogorov–Smirnov test statistics are far from being significant.

7.3.2 Results of goodness-of-fit tests for the out-of-sample period

Models d-exp and h-exp

Figure 7.4(b) presents the out-of-sample goodness-of-fit tests for Models d-exp and h-exp. The plots show a deterioration in the fit of the models when compared to the plots for the in-sample period. This deterioration is significant for the ALSI and MTN loss series — for both of these loss series, the scaled residual processes breach the confidence lines indicating that the models fail to capture the temporal behaviour of the out-of-sample extreme losses satisfactorily. For the ALSI loss series, there is also a lack of points

towards the upper right-hand corner of the plot of the points (U_i, U_{i+1}). This absence of points suggests a lack of independence in the consecutive inter-arrival times of the residual process, and adds further evidence that the temporal components of the models are not suitable for the out-of-sample extreme losses on the ALSI. From the QQ-plots, there is some evidence that the conditional mark distribution of Model h-exp may not be suitable for the out-of-sample ALSI excesses. The Kolmogorov–Smirnov test statistic is significant at the 10% level in this case.

For the ZAR/USD loss series, there is some deterioration in the fit of the models for the out-of-sample period, but it is not significant.

Model d-pow-0.5

Figure 7.5(b) presents the out-of-sample goodness-of-fit tests for Model d-pow-0.5. Again we can see that there is a deterioration in the fit of the models when we consider the out-of-sample data. The scaled residual processes for the ALSI and MTN loss series breach the 95% confidence lines. This indicates that Model d-pow-0.5 fails to model the temporal behaviour of the extreme losses occurring in the out-of-sample period satisfactorily for each of these loss series. For the ALSI loss series, there is also a lack of points towards the upper right-hand corner of the plot of the points (U_i, U_{i+1}). This absence points adds further evidence that the temporal component of Model d-pow-0.5 is not suitable for the ALSI loss series.

For the ZAR/USD loss series, there is some deterioration in the fit of the model. The QQ-plot and Kolmogorov–Smirnov test indicate that the conditional mark distribution of Model-d-pow-0.5 is not suitable for the out-of-sample ZAR/USD excesses.

Model h-gamma

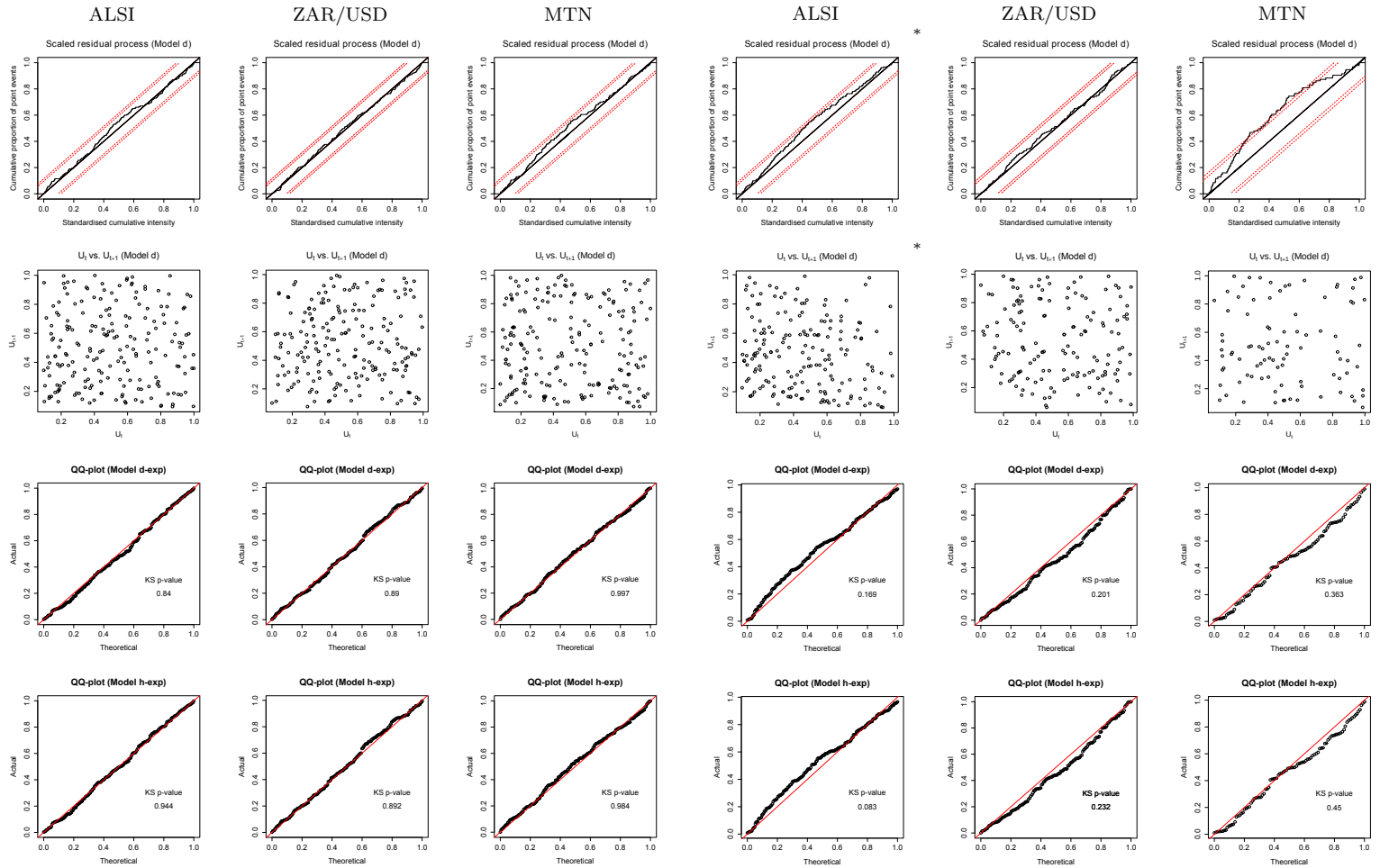
Figure 7.6(b) presents the out-of-sample goodness-of-fit tests for Model h-gamma. We see that once again there is some deterioration in the fit of the models for the out-of-sample data. The temporal component of Model h-gamma fails to model the timing of the extreme losses from the ALSI and MTN loss series satisfactorily, as the scaled residual processes breach the upper 95% confidence lines. In addition for the ALSI loss series, the

conditional mark distribution of Model h-gamma does not appear to be suitable, as the Kolmogorov–Smirnov test statistic is significant at the 5% level.

For the ZAR/USD loss series, there is some deterioration in the fit of the model, but it is not significant.

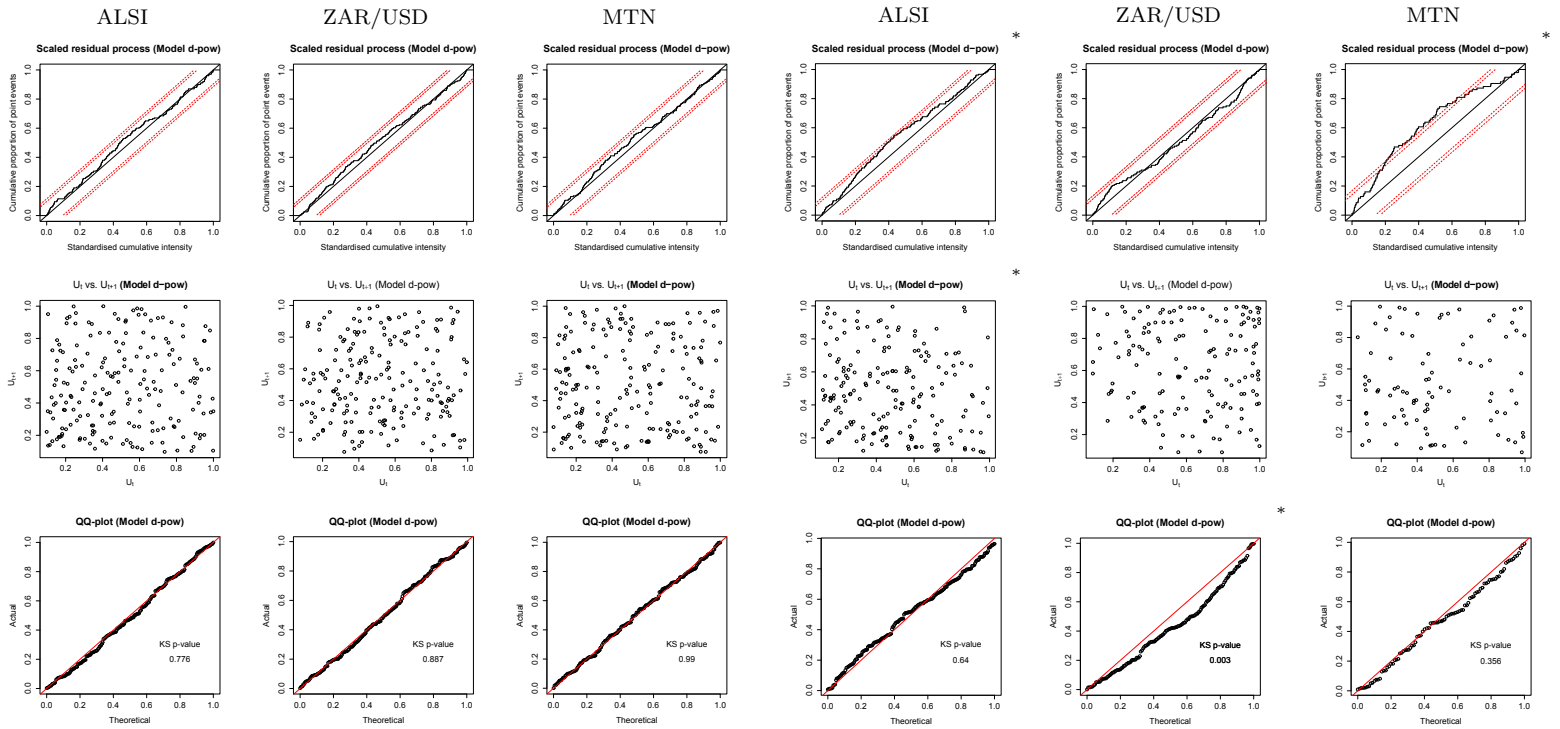
7.3.3 Summary

In summary, the goodness-of-fit tests give no significant indication that Models d-exp, h-exp, d-pow-0.5, and h-gamma fit the in-sample data poorly. However, we see that in most cases there is a significant deterioration in the fit of the models for the out-of-sample data. The 2008 financial crisis, and the increased volatility associated with it, is one likely reason why most of the models perform poorly for the out-of-sample period. The exceptions are Models d-exp, h-exp, and h-gamma, for which there are no significant indications that the models fit the out-of-sample ZAR/USD data poorly.



(a) *Goodness-of-fit tests for the in-sample data.* **(b)** *Goodness-of-fit tests for the out-of-sample data.*

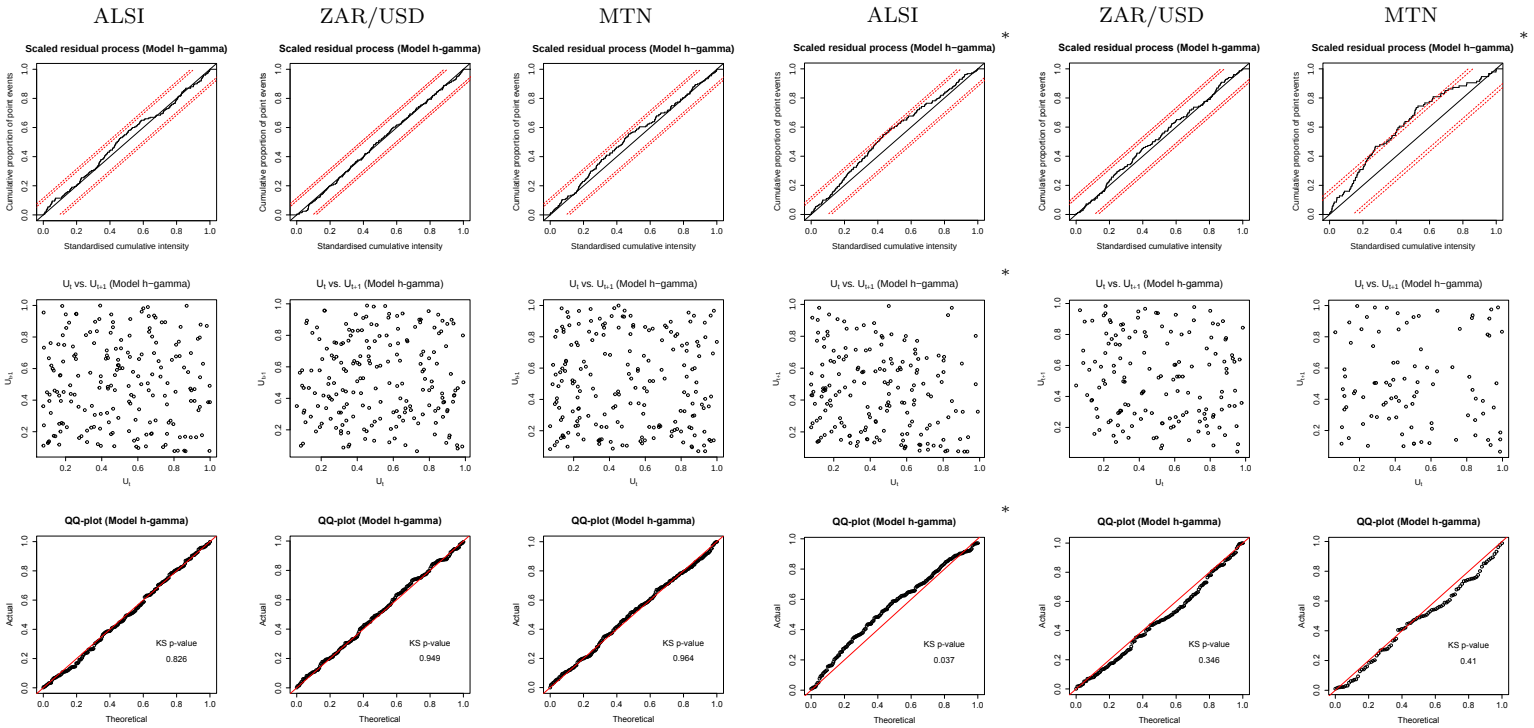
Figure 7.4: *Graphical goodness-of-fit tests for Models d-exp and h-exp. The first row contains plots of the scaled residual processes with 95% and 99% confidence lines. The second row contains plots of the points* (U_i, U_{i+1})*. The plots in the first two rows are for the temporal component of Model d-exp. The third and fourth rows contain the QQ-plots for the conditional mark distributions of Models d-exp and h-exp, respectively. If a plot suggests that a model fits the data poorly, it is marked by an asterisk to the upper right of the plot.*



(a) *Goodness-of-fit tests for the in-sample data.*

(b) *Goodness-of-fit tests for the out-of-sample data.*

Figure 7.5: *Graphical goodness-of-fit tests for Model d-pow-0.5. The first row contains plots of the scaled residual processes with 95% and 99% confidence lines. The second row contains plots of the points* (U_i, U_{i+1})*. The third row contains the QQ-plots for the conditional mark distribution. If a plot suggests that a model fits the data poorly, it is marked by an asterisk to the upper right of the plot.*



(a) *Goodness-of-fit tests for the in-sample data.*

(b) *Goodness-of-fit tests for the out-of-sample data.*

Figure 7.6: *Graphical goodness-of-fit tests for Model h-gamma. The first row contains plots of the scaled residual processes with 95% and 99% confidence lines. The second row contains plots of the points* (U_i, U_{i+1})*. The third row contains the QQ-plots for the conditional mark distribution. If a plot suggests that a model fits the data poorly, it is marked by an asterisk to the upper right of the plot.*

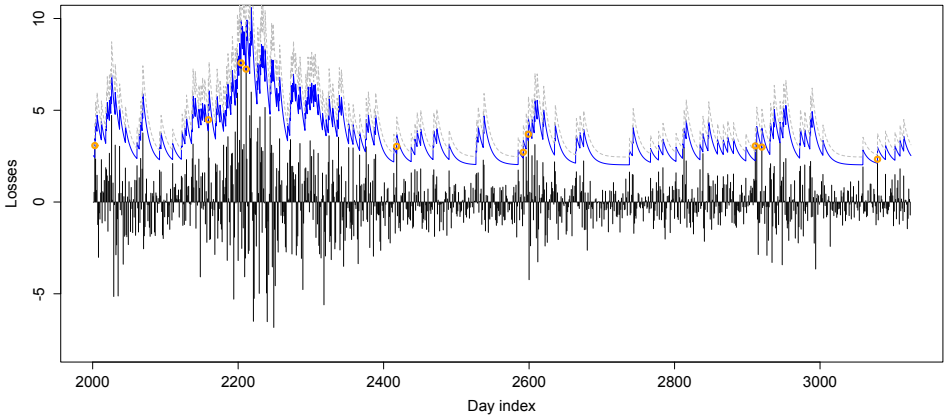
7.4 Backtesting results

The marked Hawkes process models chosen, the two SV models, and Models a and e are used to forecast conditional 99% and 99.9% VaR, and conditional 99% ES, for the out-of-sample period of each loss series considered. The parameter values used for the models are the MLEs found for the in-sample data. The 95% confidence level is not considered here, as it is not large enough to satisfy the lower bound on the available confidence levels for all of the marked Hawkes process models and loss series considered.

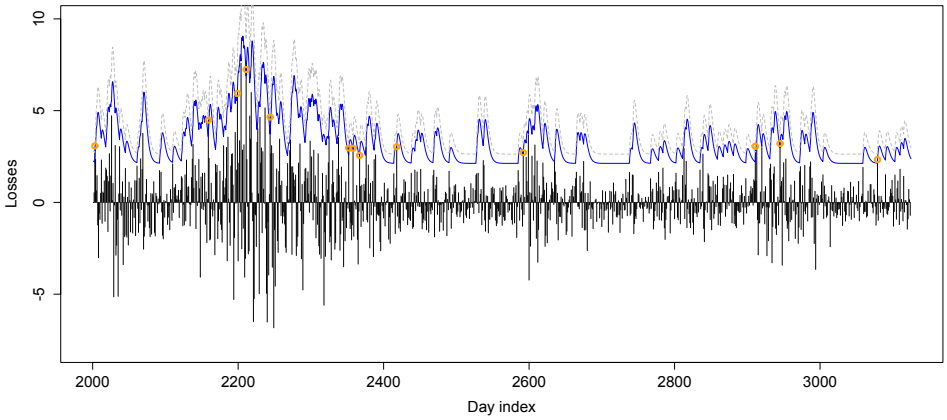
Models a and e, both of which provide forecasts of VaR and ES which are independent of the history of the marked point process, are included here so as to provide a comparison for the other models whose forecasts of VaR and ES are conditional on the history of the relevant processes. To make this distinction, Models a and e are referred to as unconditional models and the other models are referred to as conditional models.

Figures 7.7 and 7.8 present plots of the conditional 99% VaR (solid blue curve) and 99% ES (dashed grey curve) for the out-of-sample ALSI losses. The forecasts are found by using Models d-exp, h-gamma, and the SVt model. The VaR exceptions are marked on each plot by orange circles. The volatile VaR and ES forecasts reflect the conditional nature of the forecasts provided by these models. The increased volatility associated with the 2008 financial crisis, which occurs between days 2100 and 2400, results in noticeable increases in the conditional VaR and ES forecasts.

The forms of the response functions of Models d-exp and h-gamma carry through to their conditional VaR and ES forecasts. Specifically, the conditional 99% VaR and ES forecasts from Model d-exp decrease exponentially after a peak associated with an extreme loss, and the forecasts from Model h-gamma have a hump associated with each extreme loss. The conditional VaR and ES forecasts from the SVt model do not show such functional form. However, it is worth noting that the forecast conditional VaR and ES from the SV models react to the overall volatility of the losses, i.e. they react to the volatility of both positive and negative losses. This can be seen in Figure 7.8. By react, we mean that the conditional VaR and ES forecasts increase when volatility is high and decrease when volatility is low. The conditional VaR and ES forecasts from the marked Hawkes process models only react



(a) *Model d-exp*



(b) *Model h-gamma*

Figure 7.7: *Conditional 99% VaR (solid blue curve) and 99% ES (dashed grey curve) forecasts from (a) Model d-exp and (b) Model h-gamma for the out-of-sample ALSI losses. For both of the panels, the losses are given by the vertical black lines and the VaR exceptions are marked with orange circles.*

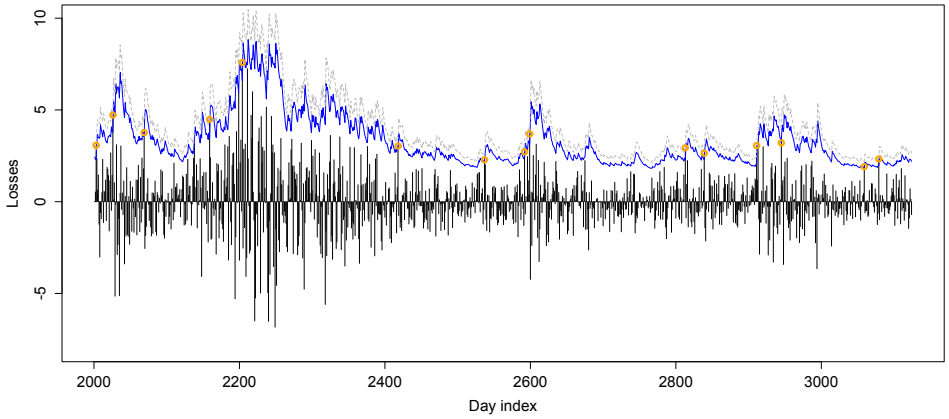


Figure 7.8: *Conditional 99% VaR (solid blue curve) and 99% ES (dashed grey curve) forecasts from the SVt model for the ALSI out-of-sample losses. The losses are given by the vertical black lines and the VaR exceptions are marked with orange circles.*

to volatility associated with positive losses as the models ignore losses below the threshold. This difference in the structure of the models may be a reason why the two classes of models show markedly different results for some of the backtests.

The conditional VaR and ES forecasts are assessed by using the backtests described in Section 6.2, and the results are presented in the following subsections. The out-of-sample period includes the 2008 financial crisis, and so this backtesting exercise is a stern test of the models.

7.4.1 Conditional VaR forecasts

Tables 7.7 and 7.8 present the numbers of VaR exceptions for the 99% and 99.9% confidence levels, respectively. For each loss series, the number of out-of-sample losses w and the expected number of exceptions, under the null hypothesis that the models correctly forecast the conditional VaR, are given in the tables. The p-values associated with one-sided Binomial tests are presented in parentheses next to the numbers of exceptions. A p-value less than or equal to 0.05 is interpreted as evidence against the null hypothesis. The classification of the model according to the BCBS's three-zone system is indicated by asterisks. For a particular loss series, a model in the green

zone has no asterisks next to its p-value, a model in the yellow zone has one asterisk, and a model in the red zone has two asterisks.

Conditional 99% VaR

From Table 7.7, we can see that all of the conditional models perform well; that is, the numbers of exceptions are not significant at the 5% level. The two unconditional models, Models a and e, do not perform well for the ALSI and ZAR/USD loss series. Both models have significant numbers of exceptions for these two loss series, and are placed in the red and yellow zones. Models a and e do not allow for changes in volatility over time and so they fail to allow for the increased volatility associated with the 2008 financial crisis. This results in significant numbers of exceptions for these two models.

Table 7.7: *Numbers of exceptions for the conditional 99% VaR for each of the models and loss series considered. The number of out-of-sample losses for each loss series is given in the second row, and the expected number of exceptions is given in the row below. The p-values of one-sided Binomial tests are reported in parentheses. Models in the green zone have no asterisks, models in the yellow zone have one asterisk, and models in the red zone have two asterisks.*

Model	ALSI	ZAR/USD	MTN
w	1124	1124	1124
Expected no.	11	11	11
a	35 (0.00)**	22 (0.00)*	6 (0.97)
d-exp	10 (0.69)	10 (0.69)	8 (0.87)
d-pow-0.5	9 (0.79)	7 (0.93)	8 (0.87)
e	33 (0.00)**	22 (0.00)*	6 (0.97)
h-exp	13 (0.34)	10 (0.69)	8 (0.87)
h-gamma	13 (0.34)	13 (0.34)	8 (0.87)
SVt	15 (0.16)	0 (1.00)	8 (0.87)
$SVMt$	15 (0.16)	1 (1.00)	8 (0.87)

The marked Hawkes process models chosen (not including Models a and e) perform reasonably well across all three loss series. The numbers of

exceptions are not too different from the expected numbers of exceptions.

The two SV models show mixed performance across the three loss series. The observed numbers of exceptions are above the expected number for the ALSI loss series, but are not significant. In contrast, the observed numbers of exceptions for the ZAR/USD loss series are considerably smaller than those of the other conditional models and the expected number. A potential reason for the low numbers of exceptions for the ZAR/USD loss series is that there was greater volatility in the direction of negative losses (devaluations of the Rand relative to the US dollar) than positive losses during the out-of-sample period; the standard deviations of the negative and positive losses are 1.0957 and 0.7830, respectively. It is suspected that the higher volatility associated with the negative losses increased the forecasts of conditional VaR and this, coupled with the lower volatility in the direction of the positive losses, resulted in low numbers of VaR exceptions. This can happen for the SV models, and not the marked Hawkes process models, because the SV models react to the overall volatility of the losses and not only volatility in the direction of positive losses. The models are not penalised for low numbers of exceptions under the one-sided Binomial tests and the BCBS's three-zone classification system, but this would be of concern to a bank, as overestimating VaR may result in excessive amounts of capital being put aside for the risk.

All of the models perform well for the MTN loss series. The likely reason for this is that the magnitudes of the extreme losses are larger on average for the in-sample period than for the out-of-sample period. These larger losses for the in-sample period may result in the estimated models producing large conditional VaR and ES forecasts for the out-of-sample period, and as such, we see few VaR exceptions for the out-of-sample period.

Conditional 99.9% VaR

Table 7.8 presents the observed numbers of exceptions for the conditional 99.9% VaR for each of the models and loss series considered. For the most part, the results are consistent with those presented in Table 7.7. The most striking difference for the 99.9% confidence level is that Models d-exp, h-exp, and h-gamma have significant numbers of exceptions for the ZAR/USD loss series. For the other two loss series, Models d-exp, h-exp,

and h-gamma performed well with observed numbers of exceptions close to the expected numbers of exceptions. Model d-pow-0.5, and the two SV models, have observed numbers of exceptions for all three loss series that are not significant, and so, in this sense, they outperformed all of the other models.

Table 7.8: *Numbers of exceptions for the conditional 99.9% VaR for each of the models and loss series considered. The number of out-of-sample losses for each loss series is given in the second row, and the expected number of exceptions is given in the row below. The p-values of one-sided Binomial tests are reported in parentheses. Models in the green zone have no asterisks and models in the yellow zone have one asterisk.*

Model	ALSI	ZAR/USD	MTN
w	1124	1124	1124
Expected no.	1	1	1
a	5 (0.01)*	5 (0.01)*	2 (0.31)
d-exp	0 (1.00)	4 (0.03)*	0 (1.00)
d-pow-0.5	0 (1.00)	2 (0.31)	0 (1.00)
e	4 (0.03)*	5 (0.01)*	1 (0.68)
h-exp	0 (1.00)	4 (0.03)*	0 (1.00)
h-gamma	0 (1.00)	4 (0.03)*	0 (1.00)
SVt	1 (0.68)	0 (1.00)	1 (0.68)
$SVMt$	0 (1.00)	0 (1.00)	1 (0.68)

7.4.2 Conditional ES forecasts

Table 7.9 presents the backtesting results for the conditional 99% ES forecasts. A model is deemed to provide good forecasts of conditional ES for a particular loss series if its V^{ES} statistic is close to zero. As the SVt model had no VaR exceptions for the ZAR/USD loss series at the 99% confidence level, we set $V^{\text{ES}} = |V_2^{\text{ES}}|$ in this case. The best model for each loss series, i.e. the model with V^{ES} statistic closest to zero, has its V^{ES} statistic in bold.

There is no one model which consistently outperformed all of the other models. The SVt model performed best for both the ALSI and MTN loss

series, and Model d-pow-0.5 performed best for the ZAR/USD loss series.

If we consider the conditional models, one may expect them to outperform the unconditional models. To a large extent this is true — most of the conditional models outperformed Models a and e, but not all of them. Model d-pow-0.5 fails to outperform Model e for the ALSI loss series, and the two SV models fail to outperform the unconditional models for the ZAR/USD loss series. The poor performance of the SV models for the ZAR/USD loss series is consistent with them overestimating the conditional ES. The inconsistent performance of Model d-pow-0.5 and the two SV models makes it difficult to point out a single best model.

If we consider only the marked Hawkes process models, we can see that there is no clear best model. Model d-pow-0.5 performs best for the ZAR/USD and MTN loss series, and Model h-gamma performs best for the ALSI loss series. If we consider only the SV models, the SVt model outperformed the $SVMt$ model for two of the three loss series. If we consider the marked Hawkes process models and the SV models together, the SV models outperformed all of the marked Hawkes process models for the ALSI and MTN loss series, and all of the marked Hawkes process models outperformed the SV models for the ZAR/USD loss series.

As there are very few or no VaR exceptions for most of the models at the 99.9% confidence level, we do not consider backtesting the conditional 99.9% ES. The need for large amounts of data, and the lack of suitable data, to backtest ES, particularly at high confidence levels, are concerns raised by several banks and banking associations in their responses[2] to the BCBS's proposal to use ES instead of VaR in the future; see, for example, the responses from UBS (Lofts, 2012, p. 3) and the Canadian Bankers Association (2012, p. 12).

[2]The responses to the consultative document issued by the BCBS (2012) are available at `http://www.bis.org/publ/bcbs219/cacomments.htm`.

Table 7.9: *Backtesting results for the conditional 99% ES forecasts. See Section 6.2.2 for a discussion of the criterion used here. The figures are percentage losses and the best model for each loss series has its V^{ES} statistic in bold.*

	ALSI			ZAR/USD			MTN		
Model	V_1^{ES}	V_2^{ES}	V^{ES}	V_1^{ES}	V_2^{ES}	V^{ES}	V_1^{ES}	V_2^{ES}	V^{ES}
Optimal	0	0	0	0	0	0	0	0	0
a	0.1630	1.4452	0.8041	0.2360	0.8152	0.5256	1.2637	−0.4752	0.8695
d-exp	−0.5046	−0.3759	0.4403	0.5070	0.3314	0.4192	−0.2295	−0.6749	0.4522
d-pow-0.5	−0.7038	−0.6224	0.6631	0.3583	−0.1416	**0.2499**	−0.2404	−0.6312	0.4358
e	−0.1769	1.0490	0.6129	0.2364	0.8156	0.5260	0.9397	−0.7992	0.8695
h-exp	−0.5696	−0.3911	0.4803	0.5294	0.3513	0.4403	−0.2513	−0.6917	0.4715
h-gamma	−0.4822	−0.2983	0.3902	0.4656	0.5456	0.5056	−0.3295	−0.7593	0.5444
SVt	−0.1498	−0.0201	**0.0849**	–	−0.8640	0.8640	0.0698	−0.4655	**0.2676**
$SVMt$	−0.1585	−0.0361	0.0973	−0.7874	−0.8311	0.8093	0.1916	−0.4636	0.3276

7.5 Remarks

We have investigated applications of a range of marked Hawkes processes to South African asset loss data. The models that we have investigated all have forms derived from Model h and which use two decay functions and a potentially nonmonotonic response function. These models include most of the models considered in the literature, and some models which appear to be new. The results presented here of course may be particular to the loss series and time periods considered; there is no guarantee that they will be transferable to other loss series and time periods. That said, we make the following remarks about our applications.

There is no single marked Hawkes process model which performs best across all of the criteria that we have considered. Instead, there are several models which perform well when their AIC values and backtesting results are considered. Model h-exp is one of the models that performs reasonably well, and is a model that has received much attention in the marked Hawkes process literature; see, for example, the work of McNeil *et al.* (2005, pp. 306–311) and Chavez-Demoulin and McGill (2012). If we are, from the results presented here, to recommend models for future applications to extreme losses, it appears that marked Hawkes process models with forms close to those of Models d-exp, h-gamma, and d-pow-0.5 are likely to perform well.

Model h-gamma is a generalisation of Model h-exp. It performs well for the ALSI loss series on the basis of its AIC value and its ES backtest results. It is different from most, if not all, the models considered in applications in the literature of marked Hawkes process models to extreme losses, as it has a nonmonotonic response function. The results presented here demonstrate that it is not necessarily a model with a monotonic response function that performs best, and that allowing models to have nonmonotonic response functions can be worthwhile. If one is to interpret such models, then nonmonotonic response functions have the attractive interpretation of indicating a potential lack of market efficiency. The confidence intervals presented in Section 7.2 suggest that in future applications it may be worthwhile to consider a simpler version of Model h-gamma.

The models with conditional exponential marks also proved worthwhile. Model d-exp and Model d-pow-0.5, both of which have conditional expo-

nential marks, performed well on the basis of their AIC values and backtest results. Models with conditional exponential marks are simpler than most of the models considered in the literature, which almost invariably have conditional GPD marks. The simplification is worthwhile for the loss series that we have considered here.

When we compare the backtest results for the marked Hawkes process models (not including Models a and e) to those of the SV models, we see that the marked Hawkes process models are mostly competitive and do not show consistent underperformance relative to the SV models. In particular, we see that the marked Hawkes process models perform well for the conditional 99% VaR, where the SV models show some inconsistent performance. However, the SV models perform better than most, but not all, of the marked Hawkes process models when we consider the conditional 99% ES backtest results. The SV models also perform better than several of the marked Hawkes process models for the conditional 99.9% VaR backtest. Model d-pow-0.5 is the model that appears to be the most competitive of the marked Hawkes process models, as it performs well for most or all of the loss series in each of the backtests.

The structural differences between the two classes of models are a likely reason for some of the differences in performance. The marked Hawkes process models could be extended so that both extreme positive and extreme negative losses contribute to the conditional intensities. For example, a conditional intensity of the following form would allow for effects from both extreme positive and negative losses:

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi_1 \sum_{t_i \in (0,t)} \omega_1(t - t_i, m_i) + \psi_2 \sum_{s_i \in (0,t)} \omega_2(t - s_i, k_i),$$

where $(s_1, k_1), (s_2, k_2), \ldots$ are the observed times and magnitudes of the excesses for the extreme negative losses, $\tau > 0$, $\psi_i \geq 0$ for $i = 1, 2$, and $\omega_i(t, m) \geq 0$ for $t \geq 0$, $m \geq 0$ and zero otherwise for $i = 1, 2$. Such an extension is suggested in passing by McNeil *et al.* (2005, p. 307), but whether or not it is worthwhile needs investigation.

The backtesting methods used in the applications here may not be the best for choosing between different models. The results we have presented here could be of interest to a bank, but the Binomial test has drawn criticism. In further work it may be worthwhile to consider alternative backtesting or

forecast evaluation methods such as those presented by Christoffersen (1998) and Gneiting (2011). These methods would provide a better assessment of a model's ability to forecast conditional VaR.

There is a lack of literature on the evaluation of conditional ES forecasts (Gneiting, 2011), and the appropriate backtesting method is not clear. If the BCBS decides to move to ES, it would be useful to revisit the backtesting of conditional ES and to use the BCBS's chosen backtests. The BCBS's chosen backtests and penalties for a poor model may give a different view of which is the best model for a bank to adopt.

CHAPTER 8

Concluding remarks and suggestions for further work

This dissertation has focussed mainly on the application of existing methods and models to South African asset data. The main results of our research are the following. We demonstrated that marked Hawkes processes with predictable marks may be suitable models for extreme losses from South African assets, and we have shown that some of the models provide competitive forecasts of market risk measures when compared to those of some nonstandard SV models. We have also highlighted that marked Hawkes process models do have a disadvantage when compared to the SV models. The constraint that is placed on the available confidence levels when forecasting conditional VaR and ES is a disadvantage of using the marked Hawkes process models.

In carrying out this research, several areas for possible future work have become apparent. Some of these areas have been mentioned in earlier chapters, such as those in the last chapter.

An area of further research, which may be of interest to those seeking to apply complex Hawkes process models in practice, would involve extending the investigation of the EM algorithm for Hawkes process models to more complex models. This would make the work in Section 3.4 more comparable to that of other researchers, e.g. that of Halpin and De Boeck (2013) and Olson and Carley (2013). The results presented here are for a simple Hawkes process model, and suggest that DNM of the ODLL is preferable to an EM algorithm. An extension to more complex Hawkes process models, such as

multitype Hawkes processes, would allow us to clarify the exact difficulties that other researchers apparently experienced when using DNM of the ODLL to find the MLEs, and to establish whether the EM algorithm does have advantages in such cases.

The focus here was the univariate marked Hawkes process. A possible extension would be to consider multitype marked Hawkes process models similar to those of Embrechts *et al.* (2011) and their application to South African asset data.

The modelling of extreme returns from several South African assets could be a worthwhile application of multitype marked Hawkes process models. By using a multitype marked Hawkes process, we could allow for the possibility of 'contagious' extreme returns that spread between assets. From such an application, it may possible to identify the assets with contagious extreme returns and to which assets these contagious returns spread. Such models and application may be of interest to practitioners looking to manage their market-risk, as well as those seeking to profit from such market reactions.

An extension of the above proposed application would involve jointly modelling the extreme returns from developed markets and those from an emerging market, such as South Africa. By using a multitype marked Hawkes process to model the extreme returns from the different markets, we could investigate whether the extreme returns in the emerging market follow those in developed markets, or vice versa. Such an application would be similar to that considered by Aït-Sahalia *et al.* (2013), who investigate the clustering and contagion of extreme returns across several financial markets, which include some emerging markets, but not South Africa. If extreme returns in the South African market do follow another market's extreme returns, then identifying the markets which South African returns follow would be useful to investors exposed to South African markets.

APPENDIX A

Simulation and parameter estimation code

A.1 Simulation

The classic method for simulating a non-homogeneous Poisson process is the thinning method of Lewis and Shedler (1979). This method requires that the conditional intensity be bounded above, i.e. there is a finite M such that for all t, $\lambda(t|\widetilde{\mathcal{H}}_t) \leq M$. The method of Lewis and Shedler (1979) was generalised by Ogata (1981). This generalised thinning algorithm, referred to as 'Ogata's modified thinning algorithm' in the literature, only requires that the intensity be locally bounded. It is a general simulation algorithm that can be used to simulate a Hawkes process, and can be extended to simulate a marked Hawkes process. The attention in this section is on Ogata's modified thinning algorithm as it is easy to implement and as it is used extensively in Section 3.4. There are other simulation methods available. For example, Ozaki (1979) uses the conditional survivor probability (2.8) to simulate a Hawkes process with exponential decay; Møller and Rasmussen (2005, 2006) present perfect and approximate simulation algorithms based on the Poisson cluster process representation; and recently Dassios and Zhao (2013) presented an exact simulation algorithm for Hawkes processes with exponential decay functions.

Ogata's modified thinning algorithm is described as follows; see Ogata (1981) for the original description. Suppose that we can find a piecewise constant process $M(\cdot|\widetilde{\mathcal{H}})$, conditional on the history of the point process of

interest $\widetilde{\mathcal{H}}$, such that for $t \in [0, T)$,

$$\lambda(t|\widetilde{\mathcal{H}}_t) \leq M(t|\widetilde{\mathcal{H}}_t), \tag{A.1}$$

where $\lambda(\cdot|\widetilde{\mathcal{H}})$ is the conditional intensity of the point process of interest.

Given that we can find a suitable $M(\cdot|\widetilde{\mathcal{H}})$, we can simulate a realisation of the point process of interest as follows. Define a nonhomogeneous Poisson process N^* which has a piecewise constant intensity $M(\cdot|\widetilde{\mathcal{H}})$ that changes value according to the history $\widetilde{\mathcal{H}}$. Decide on the termination condition, e.g. the simulation interval is $[0, T)$, and then simulate the points $0 \leq t_1^* < t_2^* < \ldots < t_{N^*[0,T)}^* < T$ from the process N^*. Each t_i^* is then selected with probability $\lambda(t_i^*|\widetilde{\mathcal{H}}_{t_i^*})/M(t_i^*|\widetilde{\mathcal{H}}_{t_i^*})$ to form part of the simulated realisation of the point process of interest, where the history $\widetilde{\mathcal{H}}_{t_i^*}$ gives the simulated history of the point process of interest up to time t_i^*. The process of selecting points from $t_1^*, t_2^*, \ldots, t_{N^*[0,T)}^*$ is the thinning procedure whereby the realisation of the point process of interest is thinned from the simulated realisation of N^*. Suppose that the selected points are $t_1, t_2, \ldots, t_{N[0,T)}$. These selected points then form a simulated realisation of the point process with conditional intensity $\lambda(\cdot|\widetilde{\mathcal{H}})$ on the interval $[0, T)$. The proof that the simulated realisation is from the point process of interest involves showing that the conditional intensity of the simulated process is equal to that of the point process of interest; see Ogata (1981, p. 24) for the details.

In practice, the function $M(\cdot|\widetilde{\mathcal{H}})$ changes value each time a point event is added to the simulated realisation of the process of interest, and so it will not be known before carrying out the simulation. To implement this simulation algorithm, the selection of the point events simulated from N^* is performed as each point event t_i^* is simulated, and the value of $M(\cdot|\widetilde{\mathcal{H}})$ is updated when a point event is selected; i.e. a point event t_i^* is simulated by using the current value of $M(t|\widetilde{\mathcal{H}}_t)$ and chosen with probability $\lambda(t_i^*|\widetilde{\mathcal{H}}_{t_i^*})/M(t_i^*|\widetilde{\mathcal{H}}_{t_i^*})$ to form part of the simulated realisation; if the point event t_i is added to the simulated realisation, the value of $M(\cdot|\widetilde{\mathcal{H}})$ is updated to ensure $\lambda(t|\widetilde{\mathcal{H}}_{t_i+}) \leq M(t|\widetilde{\mathcal{H}}_{t_i+})$ for $t_i < t$. In this way the form of $M(\cdot|\widetilde{\mathcal{H}})$ is 'uncovered' as point events are added to the simulated realisation, and does not need to be known before the simulation is carried out.

This description is for an unmarked point process and is easily modified to simulate a marked point process. To simulate a marked point process,

we would simulate the t_is as above, but in addition, we would simulate an m_i for each point t_i that is added to the simulated realisation of the point process of interest when the point t_i is added. This is done by using the conditional mark distribution $F(m)$, i.e. $m_i = F^{-1}(U_k)$, where $F^{-1}(\cdot)$ is the inverse of the conditional mark distribution function and U_k is a simulated uniform $(0, 1)$ random variable. This is implemented in Algorithm A.1.

If the conditional intensity of the Hawkes process is monotonically decreasing between point events, we can set $M(t|\widetilde{\mathcal{H}}_{t_i+}) = \lambda(t_i + |\widetilde{\mathcal{H}}_{t_i+})$ when the point event t_i is added to the simulated realisation (Ogata, 1981). This value for $M(\cdot|\widetilde{\mathcal{H}})$ is then only updated when the next point event is added to the simulated realisation. For such a Hawkes process, Daley and Vere-Jones (2003, p. 271) describe a means of making the simulation algorithm more efficient when the conditional intensity decays rapidly between point events. The increased efficiency can be achieved by finding functions $M(\cdot|\widetilde{\mathcal{H}})$ and $L(\cdot|\widetilde{\mathcal{H}})$ such that

$$\lambda(t + u|\widetilde{\mathcal{H}}_{t+}) \leq M(t|\widetilde{\mathcal{H}}_{t+}) \quad \text{for } 0 < u \leq L(t|\widetilde{\mathcal{H}}_{t+}).$$

The value of $M(\cdot|\widetilde{\mathcal{H}})$ is now updated at least when a point event is added to the simulated realisation or once the time period $L(\cdot|\widetilde{\mathcal{H}})$ has elapsed. Daley and Vere-Jones (2003, p. 271) suggest setting $M(t|\widetilde{\mathcal{H}}_{t+}) = \lambda(t + |\widetilde{\mathcal{H}}_{t+})$ and $L(t|\widetilde{\mathcal{H}}_{t+}) = \kappa\lambda(t+|\widetilde{\mathcal{H}}_{t+})$, where t is not necessarily the time of a point event and Daley and Vere-Jones recommend using $\kappa = 0.5$. The parameter $\kappa > 0$ is a tuning parameter which controls how frequently the value of $M(\cdot|\widetilde{\mathcal{H}})$ is updated. The idea being that when $\lambda(\cdot|\widetilde{\mathcal{H}})$ decays rapidly between point events, by using a small value for κ, $M(\cdot|\widetilde{\mathcal{H}})$ will be updated more frequently than before, and as a result fewer of the points simulated from N^* will be rejected. This reduces the computational cost of simulating points from N^*. However, if κ is too small, the computational cost of repeatedly updating the value of $M(\cdot|\widetilde{\mathcal{H}})$, which involves evaluating the conditional intensity, can quickly outweigh the efficiency gains from rejecting fewer points of N^*. The exact choice for the value of κ can be made after experimenting with different values for short simulations and comparing the computation times.

Algorithm A.1 is an implementation of the thinning algorithm for a marked Hawkes process with a conditional intensity which is monotonically decreasing between point events. The algorithm is based on Algorithm

7.5.V. of Daley and Vere-Jones (2003, p. 273), and simulates a realisation of a marked Hawkes process on the interval $[0, T)$.

begin
  Set $t \leftarrow 0$, $i \leftarrow 0$, and $\widetilde{\mathcal{H}} \leftarrow \emptyset$;
  while $t < T$ **do**
    Calculate $M(t|\widetilde{\mathcal{H}}) = \lambda(t + |\widetilde{\mathcal{H}})$ and $L(t|\widetilde{\mathcal{H}}) = \kappa\lambda(t + |\widetilde{\mathcal{H}})$ for a chosen κ;
    Simulate an exponential r.v. R with mean $1/M(t|\widetilde{\mathcal{H}})$;
    if $R > L(t|\widetilde{\mathcal{H}})$ **then**
      Set $t \leftarrow t + L(t|\widetilde{\mathcal{H}})$;
    else
      Simulate a uniform $(0, 1)$ r.v. U;
      if $U > \lambda(t + R|\widetilde{\mathcal{H}})/M(t|\widetilde{\mathcal{H}})$ **then**
        Set $t \leftarrow t + R$;
      else
        Set $i \leftarrow i + 1$, $t \leftarrow t + R$ and $t_i = t$;
        Simulate m_i from the conditional mark distribution $F(\cdot)$;
        Set $\widetilde{\mathcal{H}} \leftarrow \widetilde{\mathcal{H}} \cup \{(t_i, m_i)\}$;
  return $\{(t_1, m_1), (t_2, m_2), \ldots, (t_i, m_i)\}$.

Algorithm A.1: Thinning algorithm for a marked Hawkes process with conditional intensity $\lambda(\cdot|\widetilde{\mathcal{H}})$, which is monotonically decreasing between point events, and conditional mark distribution $F(\cdot)$.

The R code presented in Figures A.1–A.3 simulates a marked version of the Hawkes process described in Section 3.4. The marks are iid exponential random variables with mean μ^{-1}. The code can be run directly in R. The code in Figure A.3 simulates the realisation used in the simulation study in Section 3.4. The marks are ignored in the simulation study.

A.2 ODLL and parameter estimation code

The code presented in Figures A.4 and A.5 evaluates the ODLL in Section 3.4 and finds the MLEs via DNM of the ODLL. If the code presented in

```
lmdt.nimp = function( tim, Ht, tau, gamma, psi){
  Htt = Ht[Ht<tim]
  if(length(Htt)==0){ lbt = tau}
  else{ lbt = tau + sum( psi * exp( - gamma * ( tim - Htt)))}
  return( lbt)
}
```

Figure A.1: R *code to evaluate the conditional intensity* $\lambda(\cdot|\widetilde{\mathcal{H}})$ *in Section 3.4. The argument* `Ht` *of the function is a vector that contains the ordered point event times.*

```
sim.exp.nimp.t = function( S, tau, gamma, psi, mu, kappa){
    tim = 0
    Ht = c()
    Xt = c()
    while(tim < S)
    {
        Mt = lmdt.nimp( tim + 1e-10, Ht, tau, gamma, psi)
        Lt = kappa*Mt
        R  = rexp( 1, Mt)
            if( R > Lt){ tim = tim + Lt}else{
            cond = lmdt.nimp( (tim + R), Ht, tau, gamma, psi)/Mt
            U = runif( 1, min=0, max=1)
            if( U[1] > cond[1]) {
                tim = tim + R
            } else {
                M   = rexp( 1, mu)
                tim = tim + R
                Xt  = c( Xt, M)
                Ht  = c( Ht, tim)
               }
        }
    }
    Ht = Ht[Ht<S]
    Xt = Xt[1:length(Ht)]
    return(list(Ht=Ht,Xt=Xt))
}
```

Figure A.2: R *code which implements Algorithm A.1 for a marked version of the Hawkes process described in Section 3.4. The marks are iid exponential random variables with mean* μ^{-1}*. The argument* `S` *of the function is the upper bound for the length of the simulation interval. The tuning parameter* κ *is one of the arguments of the function and is denoted by* `kappa`*. The algorithm returns a list containing two equal length vectors* `Ht` *and* `Xt`*, where* `Ht` *contains the ordered times of the simulated point events and* `Xt` *contains the corresponding mark values.*

```
S = 10000
tau = 0.05
gamma = 0.07
psi = 0.035
mu = 1.3
kappa = 10
set.seed(500)
Sim = sim.exp.nimp.t( S, tau, gamma, psi, mu, kappa)
Sim
```

Figure A.3: R *code to generate the simulated realisation used in the simulation study in Section 3.4. The last simulated point event time should be* $t_{976} = 9993.052269$. *The marks are ignored in the simulation study.*

Figures A.1–A.5 is run sequentially in R, most of the results in line three of Table 3.1 should be reproduced. The exception is the Time (s) result. The reason being that the code presented in Figure A.4 makes use of only R code, and this code is slower than R code which makes use of a C subroutine. The time taken will also depend on the computer being used.

```
Rnll.nimp = function( param, tim, Hst){
  param = exp( param)
  Htt = Hst$Ht[Hst$Ht<tim]
  nllp1 = param[1] * tim + ( param[3] / param[2] ) *
        sum(( 1 - exp( -param[2] *( tim-Htt))))
  llp2 = sum( log( sapply( Htt, lmdt.nimp,
        Htt, param[1], param[2], param[3])))
  nll = nllp1 - llp2
  return( nll)
}
```

Figure A.4: R *code for the minus ODLL function used in Section 3.4.*

The R code presented in Figure A.6 implements part of the ODLL in Section 3.4 in C++. The code can be run in R if the `Rcpp` package is loaded and a C++ compiler is available, e.g. the C++ compiler available in Rtools[1]. If the code in Figures A.1–A.3, and Figure A.6, is run sequentially in R, the results in line three of Table 3.1 should mostly be reproduced. The median time taken should be comparable to that reported in line three of Table 3.1 when using a relatively new computer.

[1]Rtools is available from `http://cran.r-project.org/bin/windows/Rtools/`.

```
R.par.est.exp = function( Hst, tim,
                start.par = log( c( 0.04, 0.05, 0.025))){
  fit = nlm(Rnll.nimp, start.par, Hst = Hst, tim = tim)
  par = exp( fit$estimate)
  output = list( nll=fit$minimum, tau.est = par[1],
        gamma.est = par[2], psi.est=par[3],
        conv = fit$code, iterations = fit$iterations)
  return( output)
}
R.par.est.exp(Sim,tim=S,start.par=log(c(0.08,0.035,0.025)))
require(microbenchmark)
microbenchmark( R.par.est.exp( Sim, S, start.par=c(0.08,0.035,
        0.025)), unit = "s", times = 50)
```

Figure A.5: R *code to find the MLEs via DNM of the ODLL. The last three lines of the code call the* `microbenchmark` *package and use the* `microbenchmark` *routine to measure the time taken to find the MLEs.*

A.3 Simulation results

A smaller set of simulated data was used while checking the exact and approximate EM algorithms in Section 3.4, and it is these data which were used to produce Figure 3.1. The Hawkes process model used is the same as that in Section 3.4; i.e. it has conditional intensity

$$\lambda(t|\widetilde{\mathcal{H}}_t) = \tau + \psi \sum_{j:t_j \in (0,t)} \exp(-\gamma(t - t_j)),$$

and parameter values:

$$\tau = 0.05, \quad \psi = 0.035, \quad \text{and} \quad \gamma = 0.07.$$

The data are simulated by using Ogata's modified thinning algorithm. The simulated data consists of 195 point events over the interval $[0, 2\,000)$. Table A.1 presents the results for this smaller simulation study.

```
library(Rcpp)
cppFunction('
  NumericVector sumintvec(NumericVector Hs,double tau,
        double gamma, double psi) {
  int n = Hs.size();
  NumericVector out(n);
  for(int i = 0; i < n; ++i) {
    double temp = 0.0;
    double thetime = Hs[i];
      for(int j = 0; j < i; ++j) {
        temp = temp + psi*exp(-gamma*(thetime - Hs[j]));
      }
    out[i] = temp + tau;
    }
  return out;
            }'
)
Cnll.nimp = function( param, tim, Hst){
  param = exp(param)
  Htt = c(Hst$Ht[Hst$Ht<tim])
  nllp1 = param[1] * tim + (param[3]/param[2]) *
        sum(( 1 - exp( - param[2] * (tim - Htt))))
  llp2 = sum(log(sumintvec(Htt,param[1],param[2],param[3])))
  nll = nllp1 - llp2
  return(nll)
}
C.par.est.exp = function(Hst,tim,
                         start.par = log(c(0.04,0.05,0.025))){
  fit=nlm(Cnll.nimp,start.par,Hst=Hst,tim=tim)
  par=exp(fit$estimate)
  output = list(nll=fit$minimum,tau.est=par[1],
                gamma.est=par[2],psi.est=par[3],
                conv=fit$code,iterations=fit$iterations)
  return(output)
}
C.par.est.exp(Sim,tim=S,start.par = log(c(0.08,0.035,0.025)))
require(microbenchmark)
microbenchmark(C.par.est.exp(Sim,S,start.par=c(0.08,0.035,
0.025)),list=NULL,unit="s",times=50)
```

Figure A.6: R *code to find the MLEs via DNM of the ODLL. The last three lines of the code call the* `microbenchmark` *package and use the* `microbenchmark` *routine to measure the time taken to find the MLEs. Part of the ODLL code presented here is implemented in* C++ *and the time taken to find the MLEs should be comparable to that reported in line three of Table 3.1. Note: in addition to having the* `Rcpp` *and* `microbenchmark` *packages available in* R*, a* C++ *compiler needs to be available, e.g. the* C++ *compiler available in Rtools.*

Table A.1: *Results from the small simulation study used to produce Figure 3.1. The estimates (est.) found via the three parameter-estimation methods are presented in the table along with the median time taken to find the estimates in seconds. Confidence intervals found by using the approximate Hessian matrix, profile likelihoods, and bootstrap (btst.) routines are also presented in the table. For the bootstrap routines, the confidence intervals are found by using the BC_a method; see Efron and Tibshirani (1994, pp. 184–188) for details.*

Parameter	τ	ψ	γ		
True value	0.0500	0.0350	0.0700	$-\ell(\widehat{\boldsymbol{\theta}})$	Time (s)
DNM est. (`nlm`)	0.0429	0.0434	0.0762	627.684	0.1
Exact EM est.	0.0429	0.0434	0.0762	627.684	2.8
App. EM est.	0.0459	0.0451	0.0853	629.443	2.7
Wald-type 95% CI[1]	$(0.025, 0.061)$	$(0.022, 0.065)$	$(0.037, 0.115)$		
s.e.[1]	0.0091	0.0112	0.0198		
Likelihood-based 95% CI[2]	$(0.027, 0.062)$	$(0.025, 0.070)$	$(0.046, 0.131)$		
Btst. 95% CI[3]	$(0.027, 0.065)$	$(0.021, 0.075)$	$(0.032, 0.152)$		
Bootstrap mean[3]	0.0452	0.0446	0.0879		
Bootstrap s.e.[3]	0.0103	0.0143	0.0521		
Btst. 95% CI[4]	$(0.027, 0.065)$	$(0.020, 0.074)$	$(0.032, 0.151)$		
Bootstrap mean[4]	0.0451	0.0445	0.0872		
Bootstrap s.e.[4]	0.0103	0.0142	0.0468		
Btst. 95% CI[5]	$(0.025, 0.062)$	$(0.020, 0.071)$	$(0.034, 0.124)$		
Bootstrap mean[5]	0.0470	0.0465	0.0949		
Bootstrap s.e.[5]	0.0105	0.0141	0.0514		

[1] Found by using the approximate Hessian matrix supplied by the `nlm` routine in R.

[2] Found by using the profile likelihood functions and DNM.

[3,4,5] Found by using a bootstrap routine where the parameter estimation was carried out via: (3) DNM of the ODLL function, (4) the exact EM algorithm, and (5) the approximate EM algorithm.

APPENDIX B

Model fitting results and parameter estimates

B.1 BIC values

Table B.1: *BIC values for all of the models. The models with identifiers '-exp' and '-pow' have exponential and power type decay functions respectively. The model with the identifier '-gamma' has the gamma response function. The best-performing marked Hawkes process model for each loss series has its BIC value in bold. A summary of the marked Hawkes process models is given at the foot of the table; see Section 5.3 and Figure 5.4 for a more complete overview of the marked Hawkes process models. The BIC values for the two SV models are not comparable to those of the marked Hawkes process models.*

Model (no. parameters)	ALSI	ZAR/USD	MTN
a (2)	1628.2	1530.2	1947.5
b-exp (4)	1611.7	**1483.2**	1926.1
b-pow-0 (4)	1624.3	1491.7	1940.4
b-pow-0.5 (4)	1616.4	1485.7	1930.5
c-exp (5)	1616.9	1487.0	1927.1
c-pow-0 (5)	1629.1	1487.6	1933.5
c-pow-0.5 (5)	1621.5	1485.9	1928.9
d-exp (6)	**1598.4**	1488.8	**1914.0**
d-pow-0 (6)	1616.0	1491.3	1922.3
d-pow-0.5 (6)	1604.2	1488.4	1915.5
e (3)	1627.7	1535.4	1952.0
f-exp (5)	1611.3	1488.5	1930.6
f-pow-0 (5)	1623.8	1497.0	1944.8
f-pow-0.5 (5)	1616.0	1491.0	1934.9
g-exp (6)	1616.5	1492.3	1931.5
g-pow-0 (6)	1628.7	1492.9	1937.9
g-pow-0.5 (6)	1621.1	1491.2	1933.3
h-exp (7)	1603.2	1494.1	1919.1
h-pow-0 (7)	1620.0	1496.6	1927.4
h-pow-0.5 (7)	1608.8	1493.7	1920.7
h-gamma (8)	1601.9	1496.7	1924.1
SVt (5)	6065.0	5533.7	9276.8
$SVMt$ (8)	6084.0	5553.6	9298.5

Descriptions of the marked Hawkes process models:

Model 1:	**mark df for 1**	**Model 2:**	**mark df for 2**	**intensity for 1 and 2**
$a - \ldots$:	iid exp.	$e - \ldots$:	iid GPD	constant
$b - \ldots$:	iid exp.	$f - \ldots$:	iid GPD	self-exciting, no mark impacts
$c - \ldots$:	iid exp.	$g - \ldots$:	iid GPD	self-exciting, mark impacts
$d - \ldots$:	predictable exp.	$h - \ldots$:	predictable GPD	self-exciting, mark impacts

Abbreviations: df: distribution function, exp.: exponential.

B.2 Parameter estimates

The following tables present the MLEs for the models applied in Chapter 7 to the in-sample ZAR/USD and MTN loss data.

Table B.2: *MLEs for the marked Hawkes process models fitted to the in-sample marked point process realisation extracted from the ZAR/USD losses.*

Model	$\widehat{\tau}$	$\widehat{\gamma}$	$\widehat{\psi}$	$\widehat{\delta}$	$\widehat{\beta}$	$\widehat{\xi}$	$\widehat{\alpha}$	$\widehat{\zeta}$
a	0.1003	–	–	–	0.6175	–	–	–
b-exp	0.0143	0.0136	0.0122	–	0.6175	–	–	–
b-pow-0	0.0151	9.8799	0.1991	–	0.6175	–	–	–
b-pow-0.5	0.0136	29.9873	3.0494	–	0.6175	–	–	–
c-exp	0.0165	0.0106	0.0057	0.6462	0.6175	–	–	–
c-pow-0	0.0168	15.2635	0.0646	1.2652	0.6175	–	–	–
c-pow-0.5	0.0159	43.7659	1.6032	0.9941	0.6175	–	–	–
d-exp	0.0166	0.0179	0.0127	0.2940	0.4037	–	0.0245	–
d-pow-0	0.0168	12.8591	0.0651	1.2299	0.4969	–	0.0712	–
d-pow-0.5	0.0158	31.9834	1.5167	0.8668	0.4463	–	2.3227	–
e	0.1003	–	–	–	0.6175	0.0000	–	–
f-exp	0.0143	0.0136	0.0122	–	0.6175	0.0000	–	–
f-pow-0	0.0151	9.8799	0.1991	–	0.6175	0.0000	–	–
f-pow-0.5	0.0136	29.9873	3.0494	–	0.6175	0.0000	–	–
g-exp	0.0165	0.0106	0.0057	0.6462	0.6175	0.0000	–	–
g-pow-0	0.0168	15.2635	0.0646	1.2652	0.6175	0.0000	–	–
g-pow-0.5	0.0159	43.7660	1.6032	0.9941	0.6175	0.0000	–	–
h-exp	0.0166	0.0180	0.0129	0.2861	0.4023	0.0000	0.0251	–
h-pow-0	0.0168	12.8595	0.0649	1.2330	0.4971	0.0000	0.0708	–
h-pow-0.5	0.0158	31.9690	1.5170	0.8663	0.4463	0.0000	2.3231	–
h-gamma	0.0209	0.0586	0.0105	0.0156	0.3330	0.0000	0.0285	1.5658

Table B.3: *MLEs for the marked Hawkes process models fitted to the in-sample marked point process realisation extracted from the MTN losses.*

Model	$\widehat{\tau}$	$\widehat{\gamma}$	$\widehat{\psi}$	$\widehat{\delta}$	$\widehat{\beta}$	$\widehat{\xi}$	$\widehat{\alpha}$	$\widehat{\zeta}$
a	0.1000	–	–	–	1.7156	–	–	–
b-exp	0.0456	0.0517	0.0284	–	1.7156	–	–	–
b-pow-0	0.0293	2.6028	0.1206	–	1.7156	–	–	–
b-pow-0.5	0.0341	9.0555	1.1183	–	1.7156	–	–	–
c-exp	0.0496	0.0516	0.0175	0.1903	1.7156	–	–	–
c-pow-0	0.0281	3.8931	0.0592	0.3055	1.7156	–	–	
c-pow-0.5	0.0378	10.6670	0.6772	0.2334	1.7156	–	–	–
d-exp	0.0535	0.0584	0.0161	0.2353	1.0845	–	0.1489	–
d-pow-0	0.0376	2.4944	0.0402	0.3490	0.9035	–	0.4090	–
d-pow-0.5	0.0451	8.0601	0.4423	0.2788	0.9686	–	4.4055	–
e	0.1000	–	–	–	1.5961	0.0700	–	–
f-exp	0.0456	0.0517	0.0284	–	1.5961	0.0700	–	–
f-pow-0	0.0293	2.6028	0.1206	–	1.5961	0.0700	–	–
f-pow-0.5	0.0341	9.0555	1.1183	–	1.5961	0.0700	–	–
g-exp	0.0496	0.0516	0.0175	0.1903	1.5961	0.0700	–	–
g-pow-0	0.0281	3.8931	0.0592	0.3055	1.5961	0.0700	–	–
g-pow-0.5	0.0378	10.6670	0.6772	0.2334	1.5961	0.0700	–	–
h-exp	0.0533	0.0577	0.0160	0.2347	1.0399	0.0293	0.1468	–
h-pow-0	0.0373	2.6074	0.0406	0.3494	0.8584	0.0308	0.4094	–
h-pow-0.5	0.0449	8.2051	0.4483	0.2788	0.9430	0.0185	4.4240	–
h-gamma	0.0533	0.0709	0.0151	0.2276	1.0528	0.0324	0.1338	1.1295

Table B.4: *MLEs for the SV models fitted to the in-sample ZAR/USD losses.*

Model	$\widehat{\nu}$	$\widehat{\beta}$	$\widehat{\phi}$ $(\widehat{\phi_1})$	$\widehat{\sigma}$ $(\widehat{\sigma_1})$	$\widehat{\xi}$	$\widehat{\alpha}$	$\widehat{\phi_2}$	$\widehat{\sigma_2}$
SVt	12.1748	0.5784	0.9834	0.1870	0.2798	–	–	–
$SVMt$	12.2211	0.4795	0.9103	0.0022	0.3418	0.5892	1.0830	0.2898

Table B.5: *MLEs for the SV models fitted to the in-sample MTN losses.*

Model	$\widehat{\nu}$	$\widehat{\beta}$	$\widehat{\phi}$ $(\widehat{\phi_1})$	$\widehat{\sigma}$ $(\widehat{\sigma_1})$	$\widehat{\xi}$	$\widehat{\alpha}$	$\widehat{\phi_2}$	$\widehat{\sigma_2}$
SVt	18.0862	1.5859	0.9686	0.2307	0.5502	–	–	–
$SVMt$	17.6112	2.0441	0.9470	0.1900	0.0265	0.7543	1.0612	0.0001

References with index

The unique resource locators (URLs) provided for some of the references have been shortened by using the `goo.gl` URL shortener. The '[p., ...]' at the end of each reference gives the page number(s) on which the particular reference is cited.

Acerbi, C. and Tasche, D. (2002). On the coherence of expected shortfall, *Journal of Banking & Finance* **26**(7): 1487–1503. [66]

Aït-Sahalia, Y., Cacho-Diaz, J. and Laeven, R. J. A. (2013). Modeling financial contagion using mutually exciting jump processes.
URL: *http://goo.gl/5Fq3im* [118]

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**(6): 716–723. [47]

Ardia, D., Mullen, K. M., Peterson, B. G. and Ulrich, J. (2013). `DEoptim`*: Differential Evolution in* R. version 2.2-2.
URL: *http://CRAN.R-project.org/package=DEoptim* [26]

Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D. (1999). Coherent measures of risk, *Mathematical Finance* **9**(3): 203–228. [65, 66]

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns, *Australian & New Zealand Journal of Statistics* **42**(3): 283–322. [22]

Balderama, E., Schoenberg, F. P., Murray, E. and Rundel, P. W. (2012). Application of branching models in the study of invasive species, *Journal of the American Statistical Association* **107**(498): 467–476. [11]

Basel Committee on Banking Supervision (2006). International Convergence of Capital Measurement and Capital Standards: A Revised Framework,

Technical report, Bank for International Settlements.
URL: *http://goo.gl/yMcty* [84, 85]

Basel Committee on Banking Supervision (2011). Revisions to the Basel II market risk framework, *Technical report*, Bank for International Settlements.
URL: *http://goo.gl/PqC1n* [64, 65]

Basel Committee on Banking Supervision (2012). Consultative document: Fundamental review of the trading book, *Technical report*, Bank for International Settlements.
URL: *http://goo.gl/bRy0Z7* [66, 112]

Bebbington, M. and Harte, D. S. (2001). On the statistics of the linked stress release model, *Journal of Applied Probability* **38A**: 176–187. [22]

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management, *Journal of Business & Economic Statistics* **19**(4): 465–474. [52]

Berman, M. (1983). Comment on "Likelihood analysis of point processes and its applications to seismological data" by Y. Ogata, *Bulletin International Statistics Institute* **50**(3): 412–418. [51]

Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes, *The Annals of Probability* **24**(3): 1563–1588. [11]

Broto, C. and Ruiz, E. (2004). Estimation methods for stochastic volatility models: a survey, *Journal of Economic Surveys* **18**(5): 613–649. [79]

Canadian Bankers Association (2012). CBA Comments on the Basel Committee on Banking Supervison's Consultative Document: Fundamental Review of the Trading Book.
URL: *http://www.bis.org/publ/bcbs219/canadianbankers.pdf* [112]

Chambers, J. (2008). *Software for Data Analysis: Programming with* R, Springer. [25, 26]

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. [1, 17, 54, 55, 56, 60, 61, 62, 63, 65, 66, 70, 73]

Chavez-Demoulin, V. and McGill, J. A. (2012). High-frequency financial data modeling using Hawkes processes, *Journal of Banking & Finance* **36**(12): 3415–3426. [1, 19, 47, 54, 62, 63, 65, 67, 68, 70, 73, 114]

Chornoboy, E. S., Schramm, L. P. and Karr, A. F. (1988). Maximum likelihood identification of neural point process systems, *Biological Cybernetics* **59**(4–5): 265–275. [27]

Christoffersen, P. F. (1998). Evaluating interval forecasts, *International Economic Review* **39**(4): 841–862. [85, 116]

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer, London. [56, 59, 60]

Cox, D. R. and Isham, V. (1980). *Point Processes*, Chapman and Hall, London. [6, 7, 8, 11]

Daley, D. J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes: Elementary Theory and Methods*, Vol. 1, 2nd edn, Springer, New York. [3, 6, 7, 8, 9, 10, 11, 14, 21, 22, 23, 33, 49, 122, 123]

Dassios, A. and Zhao, H. (2011). A dynamic contagion process, *Advances in Applied Probability* **43**(3): 814–846. [11]

Dassios, A. and Zhao, H. (2013). Exact simulation of Hawkes process with exponentially decaying intensity, *Electronic Communications in Probability* **18**(62): 1–13. [120]

Davison, A. C. (2003). *Statistical Models*, Cambridge University Press, Cambridge. [44, 51]

Davison, A. C. (2013). `SMPracticals`*: Practicals for use with Davison (2003) Statistical Models*. R package version 1.4-2.
URL: *http://CRAN.R-project.org/package=SMPracticals* [44]

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. [27, 37]

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*, Chapman & Hall/CRC, London and Boca Raton. [38, 41, 128]

Embrechts, P. and Hofert, M. (2014). Statistics and quantitative risk management for banking and insurance, *Annual Review of Statistics and its Applications* **1**: To appear. [85]

Embrechts, P., Kaufmann, R. and Patie, P. (2005). Strategic long-term financial risk: single risk factors, *Computational Optimization and Applications* **32**(1–2): 61–90. [85, 86]

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin, Heidelberg. [56, 58, 59, 60, 63]

Embrechts, P., Liniger, T. J. and Lin, L. (2011). Multivariate Hawkes processes: an application to financial data, *Journal of Applied Probability* **48A**: 367–378. [19, 26, 51, 54, 74, 118]

Gneiting, T. (2011). Making and evaluating point forecasts, *Journal of the American Statistical Association* **106**(494): 746–762. [85, 116]

Guo, C., Luk, W., Vinkovskaya, E. and Cont, R. (2013). Customisable pipelined engine for intensity evaluation in multivariate Hawkes point processes.
URL: *http://goo.gl/D9piZ* [26]

Guttorp, P. and Thorarinsdottir, T. L. (2010). Bayesian inference for non-Markovian point processes, *Technical report*, University of Washington, Nowegian Computing Centre and Heidelberg University.
URL: *http://goo.gl/OzVgm* [47]

Halpin, P. F. (2013). A scalable EM algorithm for Hawkes processes, *Psychometrika* **Submitted**. [17, 32]

Halpin, P. F. and De Boeck, P. (2013). Modelling dyadic interaction with Hawkes processes, *Psychometrika* **78**(4): 793–814. [17, 28, 37, 40, 41, 45, 97, 117]

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes, *Biometrika* **58**(1): 83–90. [i, 3, 11, 12]

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process, *Journal of Applied Probability* **11**(3): 493–503. [14]

Hegemann, R. A., Lewis, E. A. and Bertozzi, A. L. (2013). An "Estimate & Score Algorithm" for simultaneous parameter estimation and reconstruction of missing data on social networks, *Security Informatics* **2**(1): 1–13. [28]

Herrera, R. (2013). Energy risk management through self-exciting marked point processes, *Energy Economics* **38**: 64–76. [1, 47, 54, 61, 62, 63, 67, 68]

Herrera, R. and Schipp, B. (2009). *Statistical Inference, Econometric Analysis and Matrix Algebra*, Springer, chapter Self-exciting Extreme Value Models for Stock Market Crashes, pp. 209–231. [54, 62, 65, 67, 68, 73]

Jalilian, A. (2012). `ETAS`*: Modeling earthquake data using Epidemic Type Aftershock Sequence model.* R package version 0.0-1.
URL: *http://CRAN.R-project.org/package=ETAS* [25]

Langrock, R., MacDonald, I. L. and Zucchini, W. (2012). Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models, *Journal of Empirical Finance* **19**(1): 147–161. [2, 77, 78, 79, 81, 82, 94, 99]

Lewis, E. A. and Mohler, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes.
URL: *http://goo.gl/WnEdbj* [17, 28, 32, 33]

Lewis, P. A. W. and Shedler, G. S. (1979). Simulation of nonhomogeneous Poisson processes by thinning, *Naval Research Logistics Quarterly* **26**(3): 403–413. [120]

Liniger, T. J. (2009). *Multivariate Hawkes Processes*, PhD thesis, Eidgenössische Technische Hochschule Zürich.
URL: *http://goo.gl/TfTNH* [12, 13, 17, 22, 23, 24, 51, 54]

Lofts, P. J. (2012). UBS's response to: Consultative document: Fundamental review of the trading book.
URL: *http://www.bis.org/publ/bcbs219/ubs.pdf* [112]

Lomnitz, C. (1974). *Global Tectonics and Earthquake Risk*, Elsevier Scientific Publishing Company, Amsterdam. [24]

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn, Chapman & Hall/CRC, London and Boca Raton. [40]

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*, 2nd edn, John Wiley & Sons, Hoboken, New Jersey. [27]

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**(3–4): 271–300. [83]

McNeil, A. J., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton University Press, Princeton. [25, 54, 59, 60, 62, 63, 64, 65, 66, 67, 68, 72, 93, 114, 115]

Meng, X. L. (1997). The EM algorithm and medical studies: a historical linik, *Statistical Methods in Medical Research* **6**(1): 3–23. [37]

Mersmann, O. (2013). `microbenchmark`*: Sub microsecond accurate timing functions.* R package version 1.3-0.
URL: *http://CRAN.R-project.org/package=microbenchmark* [43]

Mino, H. (2001). Parameter estimation of the intensity process of self-exciting point processes using the EM algorithm, *IEEE Transactions on Instrumentation and Measurement* **50**(3): 658–664. [28]

Møller, J. and Rasmussen, J. G. (2005). Perfect simulation of Hawkes processes, *Advances in Applied Probability* **37**(3): 629–646. [16, 120]

Møller, J. and Rasmussen, J. G. (2006). Approximate simulation of Hawkes processes, *Methodology and Computing in Applied Probability* **8**(1): 53–64. [120]

Mullen, K. M., Ardia, D., Gil, D. L., Windover, D. and Cline, J. (2011). `DEoptim`: An R package for global optimization by differential evolution, *Journal of Statistical Software* **40**(6): 1–26. [26]

Nakajima, J. and Omori, Y. (2009). Leverage, heavy-tails and correlated jumps in stochastic volatility models, *Computational Statistics and Data Analysis* **53**(6): 2335–2353. [78]

Oakes, D. (1975). The Markovian self-exciting process, *Journal of Applied Probability* **12**(1): 69–77. [12]

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes, *Annals of the Institute of Statistical Mathematics* **30**(2): 243–261. [27]

Ogata, Y. (1981). On Lewis' simulation method for point processes, *IEEE Transactions on Information Theory* **27**(1): 23–31. [23, 30, 33, 120, 121, 122]

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processess, *Journal of the American Statistical Association* **83**(401): 9–27. [12, 17, 19, 20, 23, 24, 43, 44, 45, 46, 47, 48, 49, 50, 51]

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences, *Annals of the Institute of Statistical Mathematics* **50**(2): 379–402. [11, 25]

Ogata, Y. (1999). Seismicity analysis through point-process modeling: a review, *Pure and Applied Geophysics* **155**(2–4): 471–507. [27]

Ogata, Y. and Akaike, H. (1982). On linear intensity models for mixed doubly stochastic Poisson and self-exciting point processes, *Journal of the Royal Statistical Society. Series B (Methodological)* **44**(1): 102–107. [11, 27]

Ogata, Y., Matsu'ura, R. S. and Katsura, K. (1993). Fast likelihood computation of epidemic type aftershock-sequence model, *Geophysical Research Letters* **20**(19): 2143–2146. [24]

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. [17, 19, 20, 28, 32, 33, 37, 40, 41, 43, 117]

Omori, F. (1894). On the after-shocks of earthquakes, *The Journal of the College of Science, Imperial University, Japan* **7**(2): 111–200. [12]

Otsuka, M. (1985). Studies on aftershock sequences—Part 1. Physical interpretation of Omori's formula, *Technical report*, Shimabara Earthquake and Volcano Observatory. [45]

Otsuka, M. (1987). A simulation of earthquake occurence—Part 8. On Omori's law to express aftershock seismicity, *Zisin, Series 2* **40**: 65–75. [45]

Ozaki, T. (1979). Maximum likelihood estimation of Hawkes' self-exciting point processes, *Annals of the Institute of Statistical Mathematics* **31**(1): 145–155. [19, 120]

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford Science Publications, Oxford. [40]

Pfaff, B. and McNeil, A. J. (2012). `QRM`*: Provides* R*-language code to examine Quantitative Risk Management concepts.* R package version 0.4-8.
URL: *http://CRAN.R-project.org/package=QRM* [25]

Pouzat, C. (2012). `STAR`*: Spike Train Analysis with* R. R package version 0.3-7.
URL: *http://CRAN.R-project.org/package=STAR* [44]

Pouzat, C. (2013). Email to author, 6 June 2013. [44]

Rasmussen, J. G. (2011). Bayesian inference for Hawkes processes, *Methodology and Computing in Applied Probability* **15**(3): 623–642. [21]

Rosenblatt, M. (1952). Remarks on a multivariate transformation, *The Annals of Mathematical Statistics* **23**(3): 470–472. [52]

Rubin, I. (1972). Regular point processes and their detection, *IEEE Transactions on Information Theory* **18**(5): 547–557. [8]

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statsitics* **6**(2): 461–464. [48]

R Core Team (2012). R*: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org* [20, 26]

Soetaert, K. (2013). `rootSolve`*: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations.* R package version 1.6.4.
URL: *http://cran.r-project.org/web/packages/rootSolve/index.html* [35]

Soetaert, K. and Herman, P. M. J. (2009). *A Practical Guide to Ecological Modelling: Using* R *as a Simulation Platform*, Springer. [35]

Storn, R. and Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces, *Journal of Global Optimization* **11**(4): 341–359. [27]

Taylor, S. J. (2005). *Asset Price Dynamics, Volatility, and Prediction*, Princeton University Press, Princeton. [60, 78]

Tsukakoshi, Y. and Shimazaki, K. (2006). Temporal behavior of the background seismicity rate in central Japan, 1998 to mid-2003, *Tectonophysics* **417**(1–2): 155–168. [28]

Utsu, T. (1961). A statistical study on the occurrence of aftershocks, *Geophysical Magazine* **30**: 521–605. [12]

Utsu, T. (1982). Catalog of large earthquakes in the region of Japan from 1885 through 1980, *Bulletin of Earthquake Research Institute* **57**: 401–463. [43]

Utsu, T., Ogata, Y. and Matsu'ura, R. S. (1995). The centenary of the Omori formula for a decay law of aftershock activity, *Journal of Physics of the Earth* **43**(1): 1–33. [12, 45]

Veen, A. and Schoenberg, F. P. (2008). Estimation of space-time branching process models in seismology using an EM-type algorithmm, *Journal of the American Statistical Association* **103**(482): 614–624. [11, 16, 17, 19, 27, 28, 32]

Vere-Jones, D. (1978). Earthquake prediction – a statistician's view, *Journal of Physics of the Earth* **26**(2): 129–146. [19]

Wang, T., Bebbington, M. and Harte, D. S. (2012). Markov-modulated Hawkes process with stepwise decay, *Annals of the Institute of Statistical Mathematics* **64**(3): 521–544. [11, 47]

Wong, C. S. and Li, W. K. (2000). On a mixture autoregressive model, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **62**(1): 95–115. [79]

Zucchini, W. (2000). An introduction to model selection, *Journal of Mathematical Psychology* **44**(1): 41–61. [47]

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*, Chapman & Hall/CRC Press, London and Boca Raton. [79, 80, 81]